THE NEW SPACE ERA

RECENT ADVANCES IN APPLIED PROBABILITY

Edited by

Ricardo Baeza-Yates

Joseph Glaz

Henryk Gzyl

Jürgen Hüsler

José Luis Palacios

Springer

Recent Advances in Applied Probability

This page intentionally left blank

Recent Advances in Applied Probability

Edited by

RICARDO BAEZA-YATES
Universidad de Chile, Chile

JOSEPH GLAZ
University of Connecticut, USA

HENRYK GZYL
Universidad Simón Bolívar, Venezuela

JÜRGEN HÜSLER
University of Bern, Switzerland

JOSÉ LUIS PALACIOS
Universidad Simón Bolívar, Venezuela

Springer

Visit Springer's eBookstore at: http://ebooks.kluweronline.com
and the Springer Global Website Online at: http://www.springeronline.com

Contents

This page intentionally left blank

Preface

The possibility of the present collection of review papers came up the last day of IWAP 2002. The idea was to gather in a single volume a sample of the many applications of probability.

As a glance at the table of contents shows, the range of covered topics is wide, but it sure is far away of being close to exhaustive.

Picking up a name for this collection not easier than deciding on a criterion for ordering the different contributions. As the word 'advances'' suggests, each paper represents a further step toward understanding a class of problems. No last word on any problem is said, no subject is closed.

Even though there are some overlaps in subject matter, it does not seem sensible to order this eclectic collection except by chance, and such an order is already implicit in a lexicographic ordering by first author's last name: Nobody (usually, that is) chooses a last name, does she/he? So that is how we settled the matter of ordering the papers.

We thank the authors for their contribution to this volume.

We also thank John Martindale, Editor, Kluwer Academic Publishers, for inviting us to edit this volume and for providing continual support and encouragement.

This page intentionally left blank

Acknowledgments

This page intentionally left blank

MODELING TEXT DATABASES

Ricardo Baeza-Yates
Depto. de Ciencias de la Computación
Universidad de Chile
Casilla 2777, Santiago, Chile
rbaeza@dcc.uchile.cl

Gonzalo Navarro
Depto. de Ciencias de la Computación
Universidad de Chile
Casilla 2777, Santiago, Chile
gnavarro@dcc.uchile.cl

Abstract   We present a unified view to models for text databases, proving new relations between empirical and theoretical models. A particular case that we cover is the Web. We also introduce a simple model for random queries and the size of their answers, giving experimental results that support them. As an example of the importance of text modeling, we analyze time and space overhead of inverted files for the Web.

1.1   Introduction

Text databases are becoming larger and larger, the best example being the World Wide Web (or just Web). For this reason, the importance of the information retrieval (IR) and related topics such as text mining, is increasing every day [Baeza-Yates & Ribeiro-Neto, 1999]. However, doing experiments in large text collections is not easy, unless the Web is used. In fact, although reference collections such as TREC [Harman, 1995] are very useful, their size are several orders of magnitude smaller than large databases. Therefore, scaling is an important issue. One partial solution to this problem is to have good models of text databases to be able to analyze new indices and searching algorithms before making the effort of trying them in a large scale. In particular if our application is searching the Web. The goals of this article are two fold: (1) to present in an integrated manner many different results on how to model nat-

ural language text and document collections, and (2) to show their relations, consequences, advantages, and drawbacks.

We can distinguish three types of models: (1) models for static databases, (2) models for dynamic databases, and (3) models for queries and their answers. Models for static databases are the classical ones for natural language text. They are based in empirical evidence and include the number of different words or vocabulary (Heaps' law), word distribution (Zipf's law), word length, distribution of document sizes, and distribution of words in documents. We formally relate the Heaps' and Zipf's empirical laws and show that they can be explained from a simple finite state model.

Dynamic databases can be handled by extensions of static models, but there are several issues that have to be considered. The models for queries and their answers have not been formally developed until now. Which are the correct assumptions? What is a random query? How many occurrences of a query are found? We propose specific models to answer these questions.

As an example of the use of the models that we review and propose, we give a detailed analysis of inverted files for the Web (the index used in most Web search engines currently available), including their space overhead and retrieval time for exact and approximate word queries. In particular, we compare the trade-off between document addressing (that is, the index references Web pages) and block addressing (that is, the index references fixed size logical blocks), showing that having documents of different sizes reduces space requirements in the index but increases search times if the blocks/documents have to be traversed. As it is very difficult to do experiments on the Web as a whole, any insight from analytical models has an important value on its own.

For the experiments done to backup our hypotheses, we use the collections contained in TREC-2 [Harman, 1995], especially the Wall Street Journal (WSJ) collection, which contains 278 files of almost 1 Mb each, with a total of 250 Mb of text. To mimic common IR scenarios, all the texts were transformed to lower-case, all separators to single spaces (except line breaks); and stopwords were eliminated (words that are not usually part of query, like prepositions, adverbs, etc.). We are left with almost 200 Mb of filtered text. Throughout the article we talk in terms of the size of the filtered text, which takes 80% of the original text. To measure the behavior of the index as n grows, we index the first 20 Mb of the collection, then the first 40 Mb, and so on, up to 200 Mb. For the Web results mentioned, we used about 730 thousand pages from the Chilean Web comprising 2.3Gb of text with a vocabulary of 1.9 million words.

This article is organized as follows. In Section 2 we survey the main empirical models for natural language texts, including experimental results and a discussion of their validity. In Section 3 we relate and derive the two main empirical laws using a simple finite state model to generate words. In Sections 4 and 5 we survey models for document collections and introduce new models

for random user queries and their answers, respectively. In Section 6 we use all these models to analyze the space overhead and retrieval time of different variants of inverted files applied to the Web. The last section contains some conclusions and future work directions.

1.2   Modeling a Document

In this section we present distributions for different objects in a document. They include characters, words (unique and total) and their length.

1.2.1   Distribution of Characters

Text is composed of symbols from a finite alphabet. We can divide the symbols in two disjoint subsets: symbols that separate words and symbols that belong to words. It is well known that symbols are not uniformly distributed. If we consider just letters (a to z), we observe that vowels are usually more frequent than most consonants (e.g., in English, the letter 'e' has the highest frequency.) A simple model to generate text is the Binomial model. In it, each symbol is generated with certain fixed probability. However, natural language has a dependency on previous symbols. For example, in English, a letter 'f' cannot appear after a letter 'c' and vowels, or certain consonants, have a higher probability of occurring after 'c'. Therefore, the probability of a symbol depends on previous symbols. We can use a finite-context or Markovian model to reflect this dependency. The model can consider one, two or more letters to generate the next symbol. If we use k letters, we say that it is a k-order model (so the Binomial model is considered a 0-order model). We can use these models taking words as symbols. For example, text generated by a 5-order model using the distribution of words in the Bible might make sense (that is, it can be grammatically correct), but will be different from the original [Bell, Cleary & Witten, 1990, chapter 4]. More complex models include finite-state models (which define regular languages), and grammar models (which define context free and other languages). However, finding the correct complete grammar for natural languages is still an open problem.

For most cases, it is better to use a Binomial distribution because it is simpler (Markovian models are very difficult to analyze) and is close enough to reality. For example, the distribution of characters in English has the same average value of a uniform distribution with 15 symbols (that is, the probability of two letters being equal is about 1/15 for filtered lowercase text, as shown in Table 1).

1.2.2   Vocabulary Size

What is the number of distinct words in a document? This set of words is referred to as the document *vocabulary*. To predict the growth of the vocabulary

size in natural language text, we use the so called *Heaps' Law* [Heaps, 1978], which is based on empirical results. This is a very precise law which states that the vocabulary of a text of n words is of size $V = Kn^\beta = \Theta(n^\beta)$, where K and β depend on the particular text. The value of K is normally between 10 and 100, and β is a positive value less than one. Some experiments [Araújo et al, 1997; Baeza-Yates & Navarro,1999] on the TREC-2 collection show that the most common values for β are between 0.4 and 0.6 (see Table 1). Hence, the vocabulary of a text grows sub-linearly with the text size, in a proportion close to its square root. We can also express this law in terms of the number of words, which would change K.

Notice that the set of different words of a language is fixed by a constant (for example, the number of different English words is finite). However, the limit is so high that it is much more accurate to assume that the size of the vocabulary is $O(n^\beta)$ instead of $O(1)$ although the number should stabilize for huge enough texts. On the other hand, many authors argue that the number keeps growing anyway because of the typing or spelling errors.

How valid is the Heaps' law for small documents? Figure 1 shows the evolution of the β value as the text collection grows. We show its value for up to 1 Mb (counting words). As it can be seen, β starts at a higher value and converges to the definitive value as the text grows. For 1 Mb it has almost reached its definitive value. Hence, the Heaps' law holds for smaller documents but the β value is higher than its asymptotic limit.
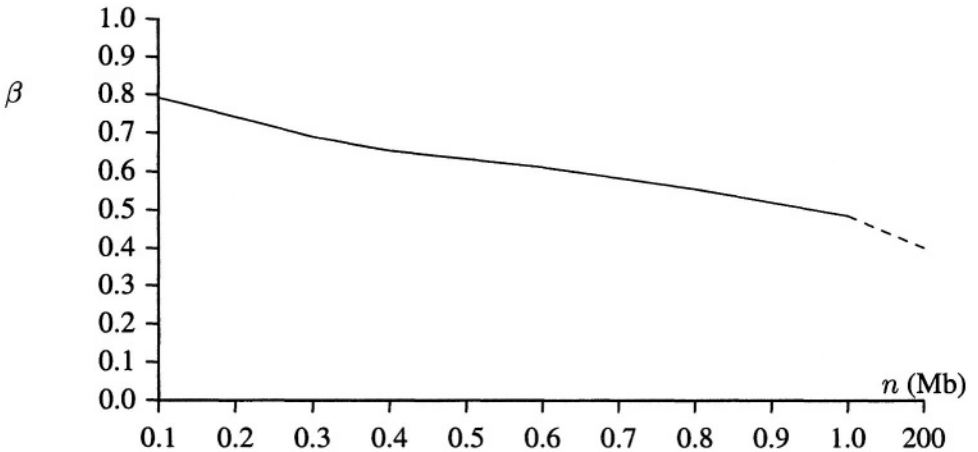
Figure 1.   Value of β as the text grows. We added at the end the value for the 200 Mb collection.

For our Web data, the value of β is around 0.63. This is larger than for English text for several reasons. Some of them are spelling mistakes, multiple languages, etc.

1.2.3 Distribution of Words

How are the different words distributed inside each document?. An approximate model is the *Zipf's Law* [Zipf, 1949; Gonnet & Baeza-Yates, 1991], which attempts to capture the distribution of the frequencies (that is, number of occurrences) of the words in the text. The rule states that the frequency of the i-th most frequent word is $1/i^\theta$ times that of the most frequent word. This implies that in a text of n words with a vocabulary of V words, the i-th most frequent word appears $n/(i^\theta H_V(\theta))$ times, where $H_V(\theta)$ is the harmonic number of order θ of V, defined as

$$H_V(\theta) = \sum_{j=1}^{V} \frac{1}{j^\theta}$$

so that the sum of all frequencies is n. The value of θ depends on the text. In the most simple formulation, $\theta = 1$, and therefore $H_V(\theta) = O(\log n)$. However, this simplified version is very inexact, and the case $\theta > 1$ (more precisely, between 1.7 and 2.0, see Table 1) fits better the real data [Araújo et al, 1997]. This case is very different, since the distribution is much more skewed, and $H_V(\theta) = O(1)$. Experimental data suggests that a better model is $k/(c+i)^\theta$ where c is an additional parameter and k is such that all frequencies add to n. This is called a Mandelbrot distribution [Miller, Newman & Friedman, 1957; Miller, Newman & Friedman, 1958]. This distribution is not used because its asymptotical effect is negligible and it is much harder to deal with mathematically.

It is interesting to observe that if, instead of taking text words, we take n-grams, no Zipf-like distribution is observed. Moreover, no good model is known for this case [Bell, Cleary & Witten, 1990, chapter 4]. On the other hand, Li [Li, 1992] shows that a text composed of random characters (separators included) also exhibits a Zipf-like distribution with smaller θ, and argues that the Zipf distribution appears because the rank is chosen as an independent variable. Our results relating the Zipf's and Heaps' law (see next section), agree with that argument, which in fact had been mentioned well before [Miller, Newman & Friedman, 1957].

Since the distribution of words is very skewed (that is, there are a few hundred words which take up 50% of the text), words that are too frequent, such as *stopwords,* can be disregarded. A stopword is a word which does not carry meaning in natural language and therefore can be ignored (that is, made not searchable), such as `"a"`, `"the"`, `"by"`, etc. Fortunately the most frequent words are stopwords, and therefore half of the words appearing in a text do not need to be considered. This allows, for instance, to significantly reduce the space overhead of indices for natural language texts. Nevertheless, there are very frequent words that cannot be considered as stopwords.

For our Web data, $\theta = 1.59$, which is smaller than for English text. This what we expect if the vocabulary is larger. Also, to capture well the central part of the distribution, we did not take in account very frequent and unfrequent words when fitting the model. A related problem is the distribution of k-grams (strings of exactly k characters), which follow a similar distribution [Egghe, 2000].

1.2.4   Average Length of Words

A last issue is the average length of words. This relates the text size in words with the text size in bytes (without accounting for punctuation and other extra symbols). For example, in the different sub-collections of TREC-2 collection, the average word length is very close to 5 letters, and the range of variation of this average in each sub-collection is small (from 4.8 to 5.3). If we remove the stopwords, the average length of a word increases to little more than 6 letters (see Table 1). If we take the average length in the vocabulary, the value is higher (between 7 and 8 as shown in Table 1). This defines the total space needed for the vocabulary. Figure 2 shows how the average length of the vocabulary words and the text words evolve as the filtered text grows for the WSJ collection.
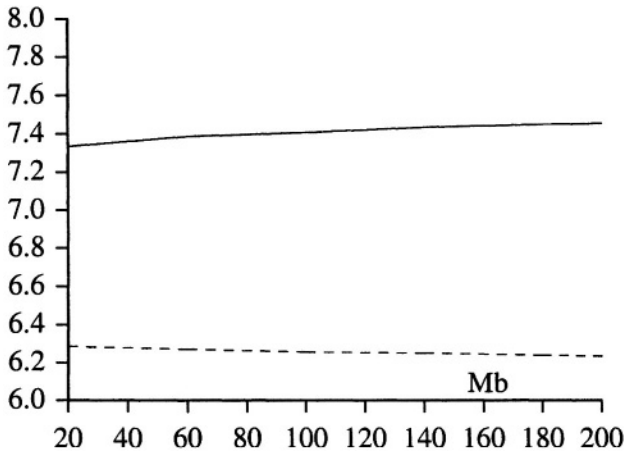
Figure 2.   Average length of the words in the vocabulary (solid line) and in the text (dashed line).

Heaps' law implies that the length of the words of the vocabulary increase logarithmically as the text size increases, and longer and longer words should appear as the text grows. This is because if for large n there are n^β different words, then their average length must be $\log_\sigma(n^\beta) = \beta \log_\sigma n$ at least (counting once each different word). However, the average length of the words in the overall text should be constant because shorter words are common enough (e.g.

stopwords). Our experiment of Figure 2 shows that the length is almost constant, although decreases slowly. This balance between short and long words, such that the average word length remains constant, has been noticed many times in different contexts. It can be explained by a simple finite-state model where the separators have a fixed probability of occurrence, since this implies that the average word length is one over that probability. Such a model is considered in [Miller, Newman & Friedman, 1957; Miller, Newman & Friedman, 1958], where: (a) the space character has probability close to 0.2, (b) the space character cannot appear twice subsequently, and (c) there are 26 letters.

1.3 Relating the Heaps' and Zipf's Law

In this section we relate and explain the two main empirical laws: Heaps' and Zipf's. In particular, if both are valid, then a simple relation between their parameters holds. This result is from [Baeza-Yates & Navarro,1999].

Assume that the least frequent word appears $O(1)$ times in the text (this is more than reasonable in practice, since a large number of words appear only once). Since there are $\Theta(n^\beta)$ different words, then the least frequent word has rank $i = \Theta(n^\beta)$. The number of occurrences of this word is, by Zipf's law,

$$\frac{n}{i^\theta H_V(\theta)} = \Theta\left(\frac{n}{n^{\beta\theta} H_V(\theta)}\right)$$

and this must be $O(1)$. This implies that, as n grows, $\beta = 1/\theta$. This equality may not hold exactly for real collections. This is because the relation is asymptotical and hence is valid for sufficiently large n, and because Heaps' and Zipf's rules are approximations. Considering each collection of TREC-2 separately, $\beta\theta$ is between 0.80 and 1.00. Table 1 shows specific values for K and β (Heaps' law) and θ (Zipf's law), without filtering the text. Notice that $1/\beta$ is always larger than θ. On the other hand, for our Web data, the match is almost perfect, as $\beta\theta \approx 1$.

Text	K	β	$1/\beta$	θ	Len. (text)	Len. (vocab.)	Eq. σ
AP	26.8	0.46	2.17	1.87	6.328	8.012	15.44
DOE	10.8	0.52	1.92	1.70	6.429	8.423	15.41
FR	13.2	0.48	2.08	1.94	6.096	6.827	15.64
WSJ	43.5	0.43	2.33	1.87	6.233	7.453	15.37
ZIFF	11.3	0.51	1.96	1.79	6.441	7.181	15.79

Table 1. Experimental results for the parameters of Heaps' and Zipf's laws, as well as the average length of words and equivalent alphabet size.

The relation of the Heapst' and Zipt's Laws is mentioned in a line of a paper by Mandelbrot [Mandelbrot, 1954], but no proof is given. In the Appendix

we give a non trivial proof based in a simple finite-state model for generating words.

1.4   Modeling a Document Collection

The Heaps' and Zipf's laws are also valid for whole collections. In particular, the vocabulary should grow faster (larger β) and the word distribution could be more biased (larger θ). That would match better the relation $\beta\theta = 1$, which in TREC-2 is less than 1. However, there are no experiments on large collections to measure these parameters (for example, in the Web). In addition, as the total text size grows, the predictions of these models become more accurate.

1.4.1   Word Distribution Within Documents

The next issue is the distribution of words in the documents of a collection. The simplest assumption is that each word is uniformly distributed in the text. However, this rule is not always true in practice, since words tend to appear repeated in small areas of the text (locality of reference). A uniform distribution in the text is a pessimistic assumption since it implies that queries appear in more documents. However, a uniform distribution can have different interpretations. For example, we could say that each word appears the same number of times in every document. However, this is not fair if the document sizes are different. In that case, we should have occurrences proportional to the document size. A better model is to use a Binomial distribution. That is, if f is the frequency of a word in a set of D documents with n words overall, the probability of finding the word k times in a document having w words ($w \leq f$) is

$$Pr(k, n, w) = \binom{w}{k} p^k (1-p)^{w-k}, \quad p = \frac{f}{n}$$

For large w, we can use the Poisson approximation $Pr(k, n, w) = \frac{\lambda^k}{k!} e^{-\lambda}$ with $\lambda = w f/n$. Some people apply these formulas using the average for all the documents, which is unfair if document sizes are very different.

A model that approximates better what is seen in real text collections is to consider a negative binomial distribution, which says that the fraction of documents containing a word k times is

$$F(k) = \binom{\alpha + k - 1}{k} p^k (1+p)^{-\alpha-k}$$

where p and α are parameters that depend on the word and the document collection. Notice that $F(k) = D \, Pr(k, n, w)$ if we use $w = n/D$, the average number of words per document, so this distribution also has the problem of being unfair if document sizes are different. For example, for the Brown Corpus

[Francis & Kucera, 1982] and the word "said", we have $p = 9.24$ and $\alpha = 0.42$ [Church & Gale, 1995]. The latter reference gives other models derived from a Poisson distribution. Another model related to Poisson which takes in account locality of reference is the Clustering Model [Thom & Zobel, 1992].

1.4.2 Distribution of Document Sizes

Static databases will have a fixed document size distribution. Moreover, depending on the database format, the distribution can be very simple. However, this is very different for databases that grow fast and in a chaotic manner, such as the Web. The results that we present next are based in the Web.

The document sizes are self-similar [Crovella & Bestavros, 1996], that is, the probability distribution remains unchanged if we change the size scale. The same behavior appears in Web traffic. This can be modeled by two different distributions. The main body of the distribution follows a Logarithmic Normal curve, such that the probability of finding a Web page of x bytes is given by

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \; e^{-(\ln x - \mu)^2/2\sigma^2}$$

where the average (μ) and standard deviation (σ) are 9.357 and 1.318, respectively [Barford & Crovella, 1998]. See figure of an example in 3 (from [Crovella & Bestavros, 1996]).
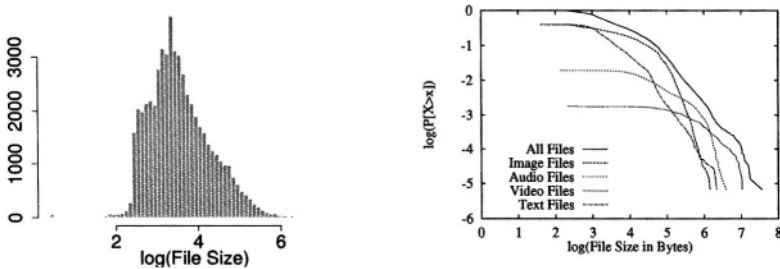
Figure 3. Left: Distribution for all file sizes. Right: Right tail distribution for different file types. All logarithms are in base 10. (Both figures are courtesy of Mark Crovella).

The right tail of the distribution is "heavy-tailed". That is, the majority of documents are small, but there is a non trivial number of large documents. This is intuitive for image or video files, but it is also true for textual pages. A good fit is obtained with the Pareto distribution, that says that the probability of finding a Web page of x bytes is

$$p(x) = \frac{\lambda k^\lambda}{x^{1+\lambda}}$$

for $x \geq k$, and zero otherwise. The cumulative distribution is

$$F(x) \;=\; 1 - \left(\frac{k}{x}\right)^{\lambda}$$

where k and λ are constants dependent on the particular collection [Barford & Crovella, 1998]. The parameter k is the minimum document size, and λ is about 1.36 for textual data, being smaller for images and other binary formats [Crovella & Bestavros, 1996; Willinger & Paxson, 1998] (see the right side of Figure 3). Taking all Web documents into account, using $k = 9.3\mathbf{Kb,}$ we get $\lambda = 1.1,$ and 93% of all the files have a size below this value. The parameters of these distributions were obtained from a sample of more than 50 thousand Web pages requested by several users in a period of two months. Recent results show that these distributions are still valid [Barford et al, 1999], but the exact parameters for the distribution of all textual documents is not known, although average page size is estimated in 6Kb including markup (which is traditionally not indexed).

1.5   Models for Queries and Answers

1.5.1   Motivation

When analyzing or simulating text retrieval algorithms, a recurrent problem is how to model the queries. The best solution is to use real users or to extract information from query logs. There are a few surveys and analyses of query logs with respect to the usage of Web search engines [Pollock & Hockley, 1997; Jensen et al, 1998; Silverstein et al, 1998]. The later reference is the study of 285 million AltaVista user sessions containing 575 million queries. Table 2 gives some results from that study, done in September of 1998. Another recent study on Excite, shows similar statistics, and also the queries topics [Spink et al, 2002]. Nevertheless, these studies give little information about the exact distribution of the queries. In the following we give simple models to select a random query and the corresponding average number of answers that will be retrieved. We consider exact queries and approximate queries. An approximate query finds a word allowing up to k errors, where we count the minimal number of insertions, deletions, and substitutions.

1.5.2   Random Queries

As half of the text words are stopwords, and they are not typical user queries, stopwords are not considered. The simplest assumption is that user queries are distributed uniformly in the vocabulary, i.e. every word in the vocabulary can be searched with the same probability. This is not true in practice, since unfrequent words are searched with higher probability. On the other hand,

Measure	Average value	Range
Number of words	2.35	0 to 393
Number of operators	0.41	0 to 958
Repetitions of each query	3.97	1 to 1.5 million

Table 2. Queries on the Web: average number of words, Boolean operations, and query repetitions.

approximate searching makes this distribution more uniform, since unfrequent words may match with k errors with other words, with little relation to the frequencies of the matched words. In general, however, the assumption of uniform distribution in the vocabulary is pessimistic, at least because a match is always found.

Looking at the results in the AltaVista log analysis [Silverstein et al, 1998], there are some queries much more popular than others and the range is quite large. Hence, a better model would be to consider that the queries also follow a Zipf's like distribution, perhaps with θ larger than 2 (the log data is not available to fit the best value). However, the actual frequency order of the words in the queries is completely different from the words in the text (for example, "sex" and "xxx" appear between the top most frequent word queries), which makes a formal analysis very difficult. An open problem, which is related to the models of term distribution in documents, is whether the distribution for query terms appearing in a collection of documents is similar to that of document terms. This is very important as these two distributions are the base for relevance ranking in the vector model [Baeza-Yates & Ribeiro-Neto, 1999]. Recent results show that although queries also follow a Zipf distribution (with parameter θ from 1.24 to 1.42 [Baeza-Yates & Castillo, 2001; Baeza-Yates & Saint-Jean, 2002]), the correlation to the word distribution of the text is low (0.2) [Baeza-Yates & Saint-Jean, 2002]. This implies that choosing queries at random from the vocabulary is reasonable and even pessimistic.

Previous work by DeFazio [DeFazio, 1993] divided the query vocabulary in three segments: high (words representing the most used 90% of the queries), moderate (next 5% of the queries), and low use (words representing the least used 5% of the queries). Words are then generated by first randomly choosing the segment, the randomly picking a token within that segment. Queries are formed by choosing randomly one to 50 words. According to currently available data, real queries are much shorter, and the generation algorithm does not produce the original query distribution. Another problem is that the query vocabulary must be known to use this model. However, in our model, we can generate queries from the text collection.

1.5.3   Number of Answers

Now we analyze the expected number of answers that will be obtained using the simple model of the previous section. For a simple word search, we will find just one entry in the vocabulary matching it. Using Heaps' law, the average number of occurrences of each word in the text is $n/V = \Theta(n^{1-\beta})$. Hence, the average number of occurrences of the query in the text is $O(n^{1-\beta})$. This fact is surprising, since one can think in the process of traversing the text word by word, where each word of the vocabulary has a fixed probability of being the next text word. Under this model the number of matching words is a fixed proportion of the text size (this is equivalent to say that a word of length ℓ should appear about $O(n/\sigma^{\ell})$ times). The fact that this is not the case (demonstrated experimentally later) shows that this model does not really hold on natural language text.

The root of this fact is not in that a given word does not appear with a fixed probability. Indeed, the Heaps' law is compatible with a model where each word appears at fixed text intervals. For instance, imagine that Zipf's law stated that the i-th word appeared $n/2^i$ times. Then, the first word could appear in all the odd positions, the second word in all the positions multiple of 4 plus 2, the third word in all the multiples of 8 plus 4, and so on. The real reason for the sublinearity is that, as the text grows, there are more words, and one selects randomly among them. Asymptotically, this means that the length of the vocabulary words must be $\ell = \Omega(\log n)$, and therefore, as the text grows, we search on average longer and longer words. This allows that even in the model where there are n/σ^{ℓ} matches, this number is indeed $o(n)$ [Navarro, 1998]. Note that this means that users search for longer words when they query larger text collections, which seems awkward but may be true, as the queries are related to the vocabulary of the collection.

How many words of the vocabulary will match an approximate query? In principle, there is a constant bound to the number of distinct words which match a given query with k errors, and therefore we can say that $O(1)$ words in the vocabulary match the query. However, not all those words will appear in the vocabulary. Instead, while the vocabulary size increases, the number of matching words that appear increases too, at a lower rate. This is the same phenomenon observed in the size of the vocabulary. In theory, the total number of words is finite and therefore $V = O(1)$, but in practice that limit is never reached and the model $V = O(n^{\beta})$ describes reality much better. We show experimentally that a good model for the number of matching words in the vocabulary is $O(n^{\nu})$ (with $\nu < \beta$). Hence, the average number of occurrences of the query in the text is $O(n^{1-\beta+\nu})$ [Baeza-Yates & Navarro, 1999].

1.5.4   Experiments

We present in this section empirical evidence supporting our previous statements. We first measure V, the number of words in the vocabulary in terms of n (the text size). Figure 4 (left side) shows the growth of the vocabulary. Using least squares we fit the curve $V = 78.81n^{0.40}$. The relative error is very small (0.84%). Therefore, $\beta = 0.4$ for the WSJ collection.
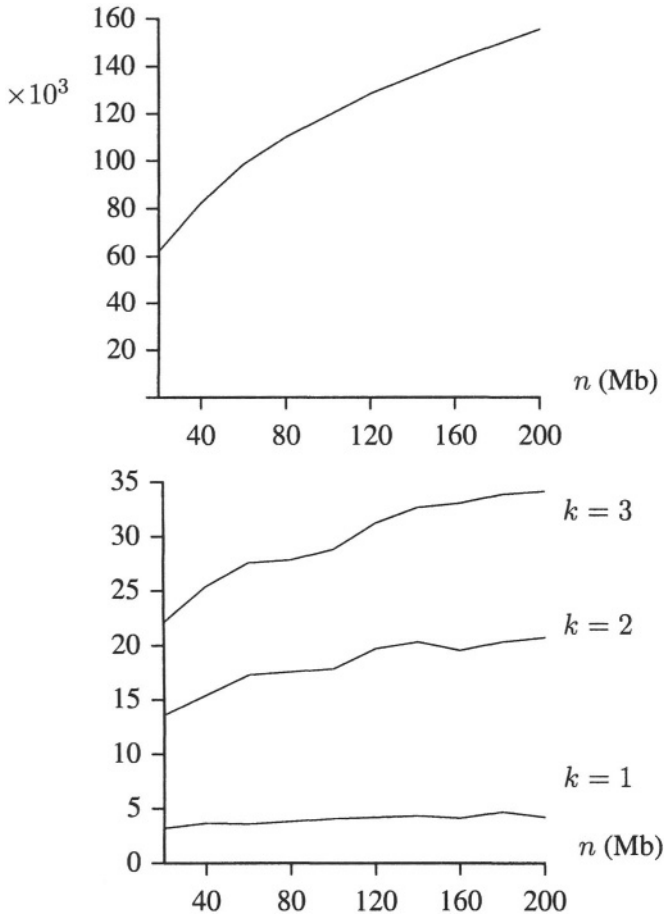
Figure 4. Vocabulary tests for the WSJ collection. On the left, the number of words in the vocabulary. On the right, number of matching words in the vocabulary.

We measure now the number of words that match a given pattern in the vocabulary. For each text size, we select words at random from the vocabulary allowing repetitions. In fact, not all user queries are found in the vocabulary in

practice, which reduces the number of matches. Hence, this test is pessimistic in that sense.

We test $k = 1, 2$ and 3 errors. To avoid taking into account queries with very low precision (e.g. searching a 3-letter word with 2 errors may match too many words), we impose limits on the length of words selected: only words of length 4 or more are searched with one error, length 6 or more with two errors, and 8 or more with three errors.

We perform a number of queries which is large enough to ensure a relative error smaller than 5% with a 95% confidence interval. Figure 4 (right side) shows the results. We use least squares to fit the curves $0.31n^{0.14}$ for $k = 1$, $0.61n^{0.18}$ for $k = 2$ and $0.88n^{0.19}$ for $k = 3$. In all cases the relative error of the approximation is under 4%. The exponents are the ν values mentioned later in this article. One possible model for ν is $\beta(1 - e^{-\alpha\,k})$, because for $k = 0$ we have $\nu = 0$ and when $k \to \infty$, $\nu \to \beta$, as expected.

We could reduce the variance in the experiments by selecting once the set of queries from the index of the first 20 Mb. However, our experiments have shown that this is not a good policy. The reason is that the first 20 Mb will contain almost all common words, whose occurrence lists grow faster than the average. Most uncommon words will not be included. Therefore, the result would be unfair, making the results to look linear when they are in fact sublinear.

1.6   Application: Inverted Files for the Web

1.6.1   Motivation

Web search engines currently available use inverted files that reference Web pages [Baeza-Yates & Ribeiro-Neto, 1999]. So, reference pointers should have as many bits as needed to reference all Web pages (currently, about 3 billion). The number and size of pointers is directly related with the space overhead of the inverted file. For the whole Web, this implies at least 600 GB. Some search engines also index word locations, so the space needed is increased. One way to reduce the size of the index is to use fixed logical blocks as reference units, trading the reduction of space obtained with an extra cost at search time. The block mechanism is a logical layer and the files do not need to be physically split or concatenated. In which follows we explain this technique in more detail.

Assume that the text is logically divided into "blocks". The index stores all the different words of the text (the vocabulary). For each word, the list of the blocks where the word appears is kept. We call b the size of the blocks and r the number of blocks, so that $n \approx rb$. The exact organization is shown in Figure 5. This idea was first used in *Glimpse* [Manber & Sun Wu, 1994].
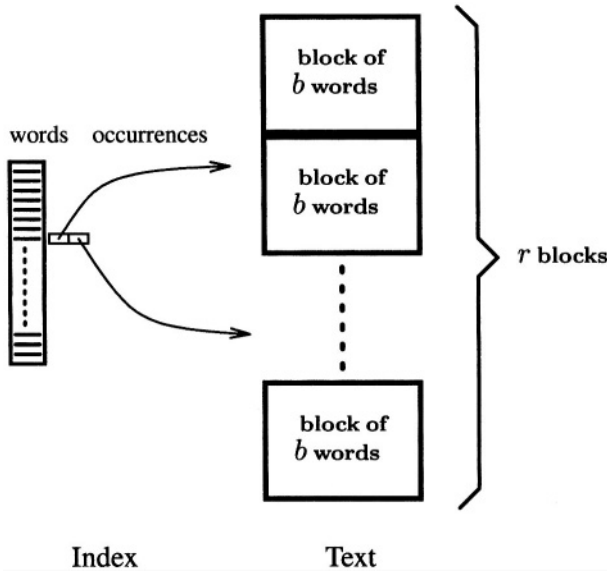
Figure 5.   The block-addressing indexing scheme.

At this point the reader may wonder which is the advantage of pointing to artificial blocks instead of pointing to documents (or files), this way following the natural divisions of the text collection. If we consider the case of simple queries (say, one word), where we are required to return only the list of matching documents, then pointing to documents is a very adequate choice. Moreover, as we see later, it may reduce space requirements with respect to using blocks of the same size. Moreover, if we pack many short documents in a logical block, we will have to traverse the matching blocks (even for these simple queries) to determine which documents inside the block actually matched.

However, consider the case where we are required to deliver the exact positions which match a pattern. In this case we need to sequentially traverse the matching blocks or documents to find the exact positions. Moreover, in some types of queries such as phrases or proximity queries, the index can only tell that two words are in the same block, and we need to traverse it in order to determine if they form a phrase.

In this case, pointing to documents of different sizes is not a good idea because larger documents are searched with higher probability and searching them costs more. In fact, the expected cost of the search is directly related to the variance in the size of the pointed documents. This suggests that if the documents have different sizes it may be a good idea to (logically) partition

large documents into blocks and to put together small documents, such that blocks of the same size are used.

In [Baeza-Yates & Navarro,1999], we show analytically and experimentally that using fixed size blocks it is possible to have a sublinear-size index with sublinear search times, even for approximate word queries. A practical example shows that the index can be $O(n^{0.94})$ in space and in retrieval time for approximate queries with at most two errors. For exact queries the exponent lowers to 0.85. This is a very important analytical result which is experimentally validated and makes a very good case for the practical use of this kind of index. Moreover, these indices are amenable to compression. Block-addressing indices can be reduced to 10% of their original size [Bell et al, 1993], and the first works on searching the text blocks directly in their compressed form are just appearing [Moura et al, 1998a; Moura et al, 1998] with very good performance in time *and* space.

Resorting to sequential searching to solve a query may seem unrealistic for current Web search engine architectures, but makes perfect sense in a near future when a remote access could be as fast as a local access. Another practical scenario is a distributed architecture where each logical block is a part of a Web server or a small set of Web servers locally connected, sharing a local index.

As explained before, pointing to documents instead of blocks may or may not be convenient in terms of query times. We analyze now the space and later the time requirements when we point to Web pages or to logical blocks of fixed size. Recall that the distribution has a main body which is log-normal (that we approximate with a uniform distribution) and a Pareto tail.

We start by relating the free parameters of the distribution. We call C the cut point between both distributions and f the fraction of documents smaller than C. Since Then the integral over the tail (from C to infinity) must be $(1-f)$, which implies that $k = (1-f)^{1/\lambda}C$. We also need to know the value of the distribution in the uniform part, which we call t, and it holds $tC = f$. For the occurrences of a word inside a document we use the uniform distribution taking into account the size of the document.

1.6.2   Space Overhead

As the Heaps' law states that a document with x words has x^{β} different words, we have that each new document of size x added to the collection will insert x^{β} new references to the lists of occurrences (since each different word of each different document has an entry in the index). Hence, an index of r blocks of size b takes $O(rb^{\beta})$ space. If, on the other hand, we consider the Web document size distribution, we have that the average number of new entries in

the occurrence list per document is

$$\int_0^\infty p(x)x^\beta dx \;=\; \int_0^C tx^\beta dx + \int_C^\infty \lambda k^\lambda x^{\beta-\lambda-1} \;=\; \frac{tC^{1+\beta}}{1+\beta} + \frac{\lambda k^\lambda}{(\lambda-\beta)C^{\lambda-\beta}}$$
(6.1)

where $p(x)$ was defined in Section 1.4.2.

To determine the total size of the collection, we consider that r documents exist, whose average length is b^* given by

$$b^* \;=\; \int_0^\infty p(x)x\,dx \;=\; \frac{tC^2}{2} + \frac{\lambda k^\lambda}{(\lambda-1)C^{\lambda-1}}$$
(6.2)

and therefore the total size of the collection is

$$n \;=\; rb^* \;=\; r\left(\frac{tC^2}{2} + \frac{\lambda k^\lambda}{(\lambda-1)C^{\lambda-1}}\right)$$
(6.3)

The final size of the occurrence lists is (using Eq. (6.1))

$$r\left(\frac{tC^{1+\beta}}{1+\beta} + \frac{\lambda k^\lambda}{(\lambda-\beta)C^{\lambda-\beta}}\right)$$
(6.4)

We consider now what happens if we take the average document length and use blocks of that fixed size (splitting long documents and putting short documents together as explained). In this case, the size of the vocabulary is $O(n^\beta)$ as before, and we assume that each block is of a fixed size $b = zb^*$. We have introduced a constant z to control the size of our blocks. In particular, if we use the same number of blocks as Web pages, then $z = 1$. Then the size of the lists of occurrences is

$$(r/z)b^\beta \;=\; \frac{r}{z^{1-\beta}}\left(\frac{tC^2}{2} + \frac{\lambda k^\lambda}{(\lambda-1)C^{\lambda-1}}\right)^\beta$$

(using Eq. (6.3)). Now, if we divide the space taken by the index of documents by the space taken by the index of blocks (using the previous equation and Eq. (6.4)), the ratio is

$$\frac{\text{document index}}{\text{block index}} \;=\; \frac{r\left(\frac{tC^{1+\beta}}{1+\beta} + \frac{\lambda k^\lambda}{(\lambda-\beta)C^{\lambda-\beta}}\right)}{\frac{r}{z^{1-\beta}}\left(\frac{tC^2}{2} + \frac{\lambda k^\lambda}{(\lambda-1)C^{\lambda-1}}\right)^\beta} \;=\; z^{1-\beta}\frac{\frac{tC}{1+\beta} + \frac{\lambda k^\lambda}{(\lambda-\beta)C^\lambda}}{\left(\frac{tC}{2} + \frac{\lambda k^\lambda}{(\lambda-1)C^\lambda}\right)^\beta}$$

$$=\; z^{1-\beta}\frac{\frac{f}{\beta+1} + \frac{\lambda(1-f)}{\lambda-\beta}}{\left(\frac{f}{2} + \frac{\lambda(1-f)}{\lambda-1}\right)^\beta}$$
(6.5)

which is independent of r, n, k and C; and is about 85% for $z = 1$, $f = 0.93$ and $\beta = 0.4..0.6$. We approximated $f = 0.93$, which corresponds to all the Web pages, because the value for textual pages is not known. This shows that indexing documents yields an index which takes 85% of the space of a block addressing index, if we have as many blocks as documents. Figure 6 shows the ratio as a function of λ and β. As it can be seen, the result varies slowly with β, while it depends more on λ (tending to 1 as the document size distribution is more uniform).

The fact that the ratio varies so slowly with β is good because we already know that the β value is quite different for small documents. As a curiosity, see that if the documents sizes were uniformly distributed in all the range (that is, letting $f \rightarrow 1$) the ratio would become $2^{\beta}/(1 + \beta)$, which is close to 0.94 for intermediate β values. On the other hand, letting $f \rightarrow 0$ (as in the simplified model [Crovella & Bestavros, 1996]) we have a ratio near 0.83. As another curiosity, notice that there is a β value which gives the minimum ratio for document versus block index (that is, the worst behavior for the block index). This is $\beta = .57$ for $z = 1$, quite close to the real values (0.63 in our Web experiments).

If we want to have the same space overhead for the document and the block indices, we simply make the expression of Eq. (6.5) equal to 1 and obtain $z \approx 1.27..1.48$ for $\beta = 0.4..0.6$, that is, we need to make the blocks larger than the average of the Web pages. This translates into worse search times. By paying more at search time we can obtain smaller indices (letting z grow over 1.48).

1.6.3   Retrieval Time

We analyze the case of approximate queries, given that for exact queries the result is the same by using $\nu = 0$. The probability of a given word to be selected by a query is $O(n^{\nu - \beta})$. The probability that none of the words in a block is selected is therefore $(1 - O(n^{\nu - \beta}))^b$. The total amount of work of an index of fixed blocks is obtained by multiplying the number of blocks (r) times the work to do per selected block (b) times the probability that some word in the block is selected. This is

$$\Theta \left(rb \left(1 - \left(1 - n^{\nu - \beta} \right)^b \right) \right) \; = \; \Theta \left(n \left(1 - e^{-\Theta(b/n^{\beta - \nu})} \right) \right) \qquad (6.6)$$

where for the last step we used that $(1 - x)^y = e^{y \ln(1 - x)} = e^{y(-x + O(x^2))} = \Theta(e^{-\Theta(yx)})$ provided $x = o(1)$.

We are interested in determining in which cases the above formula is sublinear in n. Expressions of the form "$1 - e^{-x}$" are $O(x)$ whenever $x = o(1)$ (since $e^{-x} = 1 - x + O(x^2)$). On the other hand, if $x = \Omega(1)$, then e^{-x} is far away from 1, and therefore "$1 - e^{-x}$" is $\Omega(1)$.
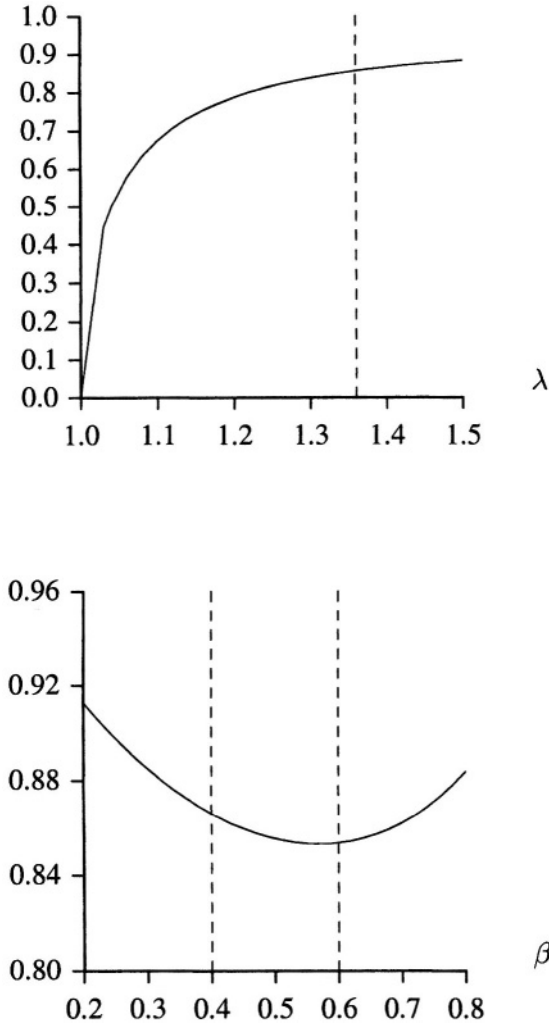
Figure 6. On the left, ratio between block and document index as a function of λ for fixed $\beta = 0.5$ (the dashed line shows the actual λ value for the Web). On the right, the same as a function of β for $\lambda = 1.36$ (the dashed lines enclose the typical β values). In both cases we use $f = 0.93$ and the standard $z = 1$.

For the search cost to be sublinear, it is thus necessary that $b = o(n^{\beta - \nu})$. When this condition holds, we derive from Eq. (6.6) that

$$Time = \Theta\left(n^{\beta} + bn^{1-\beta+\nu}\right) \tag{6.7}$$

We consider now the case of an index that references Web pages. As we have shown, if a block has size x then the probability that it has to be traversed is $(1 - e^{-\Theta(x/n^{\beta-\nu})})$. We multiply this by the cost x to traverse it and integrate over all the possible sizes, so as to obtain its expected traversal cost (recall Eq. (6.6))

$$\int_k^\infty x(1 - e^{-\Theta(x/n^{\beta-\nu})})p(x)dx$$

which we cannot solve. However, we can separate the integral in two parts, (*a*) $x = o(n^{\beta-\nu})$ and (*b*) $x = \Omega(n^{\beta-\nu})$. In the first case the traversal probability is $O(x/n^{\beta-\nu})$ and in the second case it is $\Omega(1)$. Splitting the integral in two parts and multiplying the result by $r = n/b^*$ we obtain the total amount of work:

$$\Theta\left(\frac{n}{\frac{1}{2} + \frac{\lambda}{\lambda-1}}\left(\left(\frac{C}{3} - \frac{\lambda f}{2-\lambda}\right)n^{\nu-\beta} + \frac{\lambda C^{\lambda-1}}{(2-\lambda)(\lambda-1)}n^{(\nu-\beta)(\lambda-1)}\right)\right)$$

where since this is an asymptotic analysis we have considered $C = o(n^{\beta-\nu})$, as C is constant.

On the other hand, if we used blocks of fixed size, the time complexity (using Eq. (6.7)) would be $O(bn^{1-\beta+\nu})$, where $b = zb^*$. The ratio between both search times is

$$\frac{\text{doc. index traversal}}{\text{block index traversal}} = \Theta\left(n^{(\beta-\nu)(2-\lambda)}\right)$$

which shows that the document index would be asymptotically slower than a block index as the text collection grows. In practice, the ratio is between $O(n^{0.2})$ and $O(n^{0.4})$. The value of z is not important here since it is a constant, but notice that k is usually quite large, which favors the block index.

1.7   Concluding Remarks

The models presented here are common to other processes related to human behavior [Zipf, 1949] and algorithms. For example, a Zipf like distribution also appears for the popularity of Web pages with $\theta < 1$ [Barford et al, 1999]. On the other hand, the phenomenon of sublinear vocabulary growing is not exclusive of natural language words. It appears as well in many other scenarios, such as the number of different words in the vocabulary that match a given query allowing errors as shown in Section 5, the number of states of the deterministic automaton that recognizes a string allowing errors [Navarro, 1998], and the number of suffix tree nodes traversed to solve an approximate query [Navarro & Baeza-Yates, 1999]. We believe that in fact the finite state model for generating words used in Section 3 could be changed for a more general

one that could explain why is this behavior so extended in apparently very dissimilar processes.

By the Heaps' law, more and more words appear as the text grows. Hence, $\Theta(\log n)$ bits are necessary in principle to distinguish among them. However, as proved in [Moura et al, 1998], the entropy of the words of the text remains constant. This is related to Zipf's law: the word distribution is very skewed and therefore they can be referenced with a constant number of average bits. This is used in [Moura et al, 1998] to prove that a Huffman code to compress words will not degrade as the text grows, even if new words with longer and longer codes appear. This resembles the fact that although longer and longer words appear, their average length in the text remains constant.

Regarding the number of answers of other type of queries, like prefix searching, regular expressions and other multiple-matching queries, we conjecture that the set of matching words grows also as $O(n^{\nu})$ if the query is going to be useful in terms of precision. This issue is being considered for future work.

With respect to our analysis of inverted files for the Web, our results say that using blocks we can reduce the space requirements by increasing slightly the retrieval time, keeping both of them sublinear. Fine tuning of these ideas is matter of further study. On the other hand, the fact that the average Web page remains constant even while the Web grows shows that sublinear space is not possible unless block addressing is used. Hence, future work includes the design of distributed architectures for search engines that can use these ideas.

Finally, as it is very difficult to do meaningful experiments in the Web, we believe that careful modeling of Web pages statistics may help in the final design of search engines. This can be done not only for inverted files, but also for more difficult design problems, such as techniques for evaluating Boolean operations in large answers and the design of distributed search architectures, where Web traffic and caching become an issue as well.

Acknowledgments

Appendix

Deducing the Heaps' Law

We show now that the Heaps' law can be deduced from the simple finite state model mentioned before. Let us assume that a person hits the space with probability $(1-p)$ and any other letter (uniformly distributed over an alphabet of size σ) with probability p, without hitting the space bar twice in a row (see Figure A.1).

Since there are no words of length zero, the probability that a produced word is of length ℓ is $p^{\ell-1}(1-p)$, since we have a geometric distribution. The expected word length is $1/(1-p)$, from where $p = 0.84$ can be approximated since the average word length is close to 6.3 as
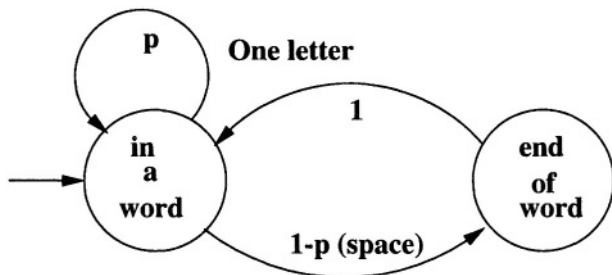
Figure A.1.   Simple finite-state model for generating words.

shown later, for text without stopwords. For this case, we use $\sigma = 15$, which would be the equivalent number of letters for text generated using a uniformly distributed alphabet.

On average, if n words are written, $p^{\ell-1}(1-p)n$ of them are of length ℓ. We count now how many of these are different, considering only those of length ℓ. Each of the σ^ℓ strings of length ℓ is different from each written word of length ℓ with probability $(1 - 1/\sigma^\ell)$, and therefore it is never written in the whole process with probability

$$\left(1 - \frac{1}{\sigma^\ell}\right)^{p^{\ell-1}(1-p)n} \quad = \quad e^{-\frac{p^{\ell-1}(1-p)n}{\sigma^\ell}(1+O(1/\sigma^\ell))}$$

from where we obtain that the total number of different words that are written is

$$\sigma^\ell \left(1 - e^{-\frac{p^{\ell-1}(1-p)n}{\sigma^\ell}(1+O(1/\sigma^\ell))}\right)$$

Now we consider two possible cases

(a) $x = p^{\ell-1}(1-p)n/\sigma^\ell = o(1)$

The condition is equivalent to $\ell = \omega(L)$, where $L = \ln((1-p)n/p)/\ln(\sigma/p)$, i.e. large ℓ. In this case, $e^{-x} = 1 - x + o(x)$, and hence the number of strings is

$$\sigma^\ell \left(\frac{p^{\ell-1}(1-p)n}{\sigma^\ell}\right)(1+O(1/\sigma^\ell)) \quad = \quad p^{\ell-1}(1-p)n\,(1+O(1/\sigma^\ell))$$

that is, basically all the written words are different.

(b) $x = p^{\ell-1}(1-p)n/\sigma^\ell = \Omega(1)$

In this case, e^{-x} is far away from 1, and therefore $\sigma^\ell(1 - e^{-x}) = \Theta(\sigma^\ell)$. That is, ℓ is small and all the different words are generated.

We sum now all the different words of each possible length generated,

$$\sum_{\ell=1}^{\lfloor L \rfloor} \sigma^\ell \quad + \quad \sum_{\ell=\lfloor L+1 \rfloor}^{\infty} p^{\ell-1}(1-p)n$$

and obtain that both summations are

$$O\left(n^{\frac{1}{1+\log_\sigma(1/p)}}\right)$$

which is of the form $O(n^\beta)$.

The value obtained with $p = 0.84$ and $\sigma = 15$ is $n^{0.94}$, which is much higher than reality. Consider, however, that it is unrealistic to assume that all the 15 or 26 letters are equally probable and to ignore the dependencies among consecutive letters. In fact, not all possible combinations of letters are valid words. Even in this unfavorable case, we have shown that the number of different words follows Heaps' law. More accurate models should yield the empirically observed values between 0.4 and 0.6.

Deducing the Zipf's Law

We show now that also the Zipf's law can be deduced from the same model. From the previous Heaps' result, we know that if we consider words of length $\ell = O(L)$ then all the different σ^ℓ combinations appear, while if $\ell = \omega(L)$ then all the $p^{\ell-1}(1-p)n$ words generated are basically different.

Since shorter words are more probable than longer words, we know that, if we sort the vocabulary by frequency (from most to least frequent), all the words of length smaller than ℓ will appear before those of length ℓ.

In the case $\ell = O(L)$, the number of different words shorter than ℓ is

$$\sum_{i=1}^{\ell-1} \sigma^i \;=\; \frac{\sigma^\ell - \sigma}{\sigma - 1} \;=\; \Theta(\sigma^\ell)$$

while, on the other hand, if $\ell = \omega(L)$, the summation is split in all those smaller than L and those between L and ℓ:

$$\sum_{i=1}^{\lfloor L \rfloor} \sigma^i \;+\; \sum_{i=\lfloor L+1 \rfloor}^{\ell-1} p^{i-1}(1-p)n \;=\; \Theta(\sigma^{L+1} + n(p^L - p^{\ell-1}))$$

which, since $L = \ln((1-p)n/p)/\ln(\sigma/p)$, is $\Theta(((1-p)n/p)^{1/(1+\log_\sigma(1/p))})$.

We relate now the result with Zipf's law. In the case of small ℓ, we have that the rank of the first word of length ℓ is $i = \Theta(\sigma^\ell)$. We also know that, since all the σ^ℓ different words of length ℓ appear, they are uniformly distributed, and $p^{\ell-1}(1-p)n$ words of length ℓ are written, then the number of times each different word appears is

$$\frac{p^{\ell-1}(1-p)n}{\sigma^\ell} \;=\; \frac{(1-p)n/p}{(\sigma/p)^\ell} \;=\; \frac{(1-p)n/p}{(\sigma^\ell)^{\log_\sigma(\sigma/p)}} \;\approx\; \frac{(1-p)n/p}{i^{1+\log_\sigma(1/p)}}$$

which, under the light of Zipf's law, shows that $\theta = 1 + \log_\sigma(1/p)$.

We consider the case of large ℓ now. As said, basically every typed word of this length is different, and therefore its frequency is 1. Since this must be $O(n/i^\theta)$, we have

$$O(n) \;=\; i^\theta \;=\; ((1-p)n/p)^{\theta/(1+\log_\sigma(1/p))}$$

where the last step considered that, as found before, the rank i of this word is $((1-p)n/p)^{1/(1+\log_\sigma(1/p))}$. Equating the first and last term yields again $\theta = 1 + \log_\sigma(1/p)$.

Hence, the finite state model implies Zipf's law, moreover, the θ value found is precisely $1/\beta$, where β is the value for Heaps' law. As we have shown, this relation must hold when both rules are valid. The numerical value we obtain for θ assuming $p = 0.84$ and a uniform model over 15 letters is $\theta = 1.06$, which is also far from reality but is close to the Mandelbrot distribution fitting obtained by Miller *et al* [Miller, Newman & Friedman, 1957] (they use $p = 0.82$). Note also that the development of Li [Li, 1992] is similar to ours regarding the Zipf's law, although he uses different techniques and argues that this law appears because the frequency rank is used as independent variable. However, we have been able to relate β and θ.

References

M. Araújo, G. Navarro, and N. Ziviani. Large text searching allowing errors. In *Proc. WSP'97,* pages 2–20, Valparaíso, Chile, 1997. Carleton University Press.

R. Baeza-Yates and G. Navarro. Block-addressing indices for approximate text retrieval. *Journal of the American Society for Information Science* 51 (1), pages 69–82, 1999.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley, 1999.

R. Baeza-Yates and C. Castillo. Relating Web Structure and User Search Behavior. Poster in *Proc. of the WWW Conference,* Hong-Kong, 2001.

R. Baeza-Yates and F. Saint-Jean. A Three Level Search Index and Its Analysis. CS Technical Report, Univ. of Chile, 2002.

P. Barford, A. Bestavros, A. Bradley, and M. E. Crovella. Changes in web client access patterns: Characteristics and caching implications. *World Wide Web* 2, pages 15–28, 1999.

P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. In *ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems,* pages 151–160, July 1998.

T.C. Bell, J. Cleary, and I.H. Witten. *Text Compression.* Prentice-Hall, 1990.

T. C. Bell, A. Moffat, C. Nevill-Manning, I. H. Witten, and J. Zobel. Data compression in full-text retrieval systems. *Journal of the American Society for Information Science,* 44:508–531, 1993.

M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. In *ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems,* pages 160–169, May 1996.

K. Church and W. Gale. Poisson mixtures. *Natural Language Engineering,* 1(2):163–190, 1995.

S. DeFazio. Overview of the Full-Text Document Retrieval Benchmark. In The Benchmark Handbook for Database and Transaction Processing Systems, J. Gray (ed.), Morgan Kaufmann, pages 435–487, 1993.

L. Egghe. The distribution of N-grams. *Scientometrics* 47(2), pages 237-252, 2000.

W. Francis and H. Kucera. *Frequency Analysis of English Usage.* Houghton Mifflin Co., 1982.

G. Gonnet and R. Baeza-Yates. *Handbook of Algorithms and Data Structures.* Addison-Wesley, Wokingham, England, 2nd edition, 1991.

D. K. Harman. Overview of the third text retrieval conference. In *Proc. Third Text REtrieval Conference (TREC-3),* pages 1–19, Gaithersburg, USA, 1995. National Institute of Standards and Technology Special Publication.

H.S. Heaps. *Information Retrieval - Computational and Theoretical Aspects.* Academic Press, 1978.

B.J. Jensen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the Web. *ACM SIGIR Forum,* 32(1):5–17, 1998.

W. Li. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Trans.on Information Theory,* 38(6): 1842–45, 1992.

Udi Manber and Sun Wu. GLIMPSE: A tool to search through entire file systems. In *Proc. of USENIX Technical Conference,* pages 23–32, San Francisco, USA, January 1994. `ftp://cs.arizona.edu/glimpse/glimpse.ps.Z`.

B. Mandelbrot. On recurrent noise limiting coding. In *Symp. on Information Networks,* pages 205-221, 1954.

G. Miller, E. Newman, and E. Friedman. Some effects of intermittent silence. *American J. of Psychology,* 70:311–312, 1957.

G.A. Miller, E.B. Newman, and E.A. Friedman. Length-frequency statistics for written English. *Information and Control,* 1:370–389, 1958.

E. S. Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates. Direct pattern matching on compressed text. In *Proc. of the 5th Symposium on String Processing and Information Retrieval,* pages 90–95, Santa Cruz, Bolivia, September 1998.

E. S. Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates. Fast searching on compressed text allowing errors. In *Proc. of the ACM-SIGIR International Conference on Research and Development in Information Retrieval,* pages 298–306, Melbourne, Australia, August 1998.

G. Navarro. *Approximate Text Searching.* PhD thesis, Dept. of Computer Science, Univ. of Chile, December 1998. Tech. Report TR/DCC-98-14. `ftp://ftp.dcc.uchile.cl/pub/-users/gnavarro/thesis98.ps.gz`.

G. Navarro and R. Baeza-Yates. A new indexing method for approximate string matching. In *Proc. 10th Symposium on Combinatorial Pattern Matching (CPM'99),* pages 163–185, Warwick, England, July 1999.
`ftp://ftp.dcc.uchile.cl/pub/users/gnavarro/stindex.ps.gz`.

A. Pollock and A. Hockley. What's wrong with Internet searching. *D-Lib Magazine,* March 1997.

C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large AltaVista query tog. Technical Report 1998-014, COMPAQ Systems Research Center, Palo Alto, CA, USA, 1998.

A. Spink, B.J. Jansen, D. Wolfram, and T. Saracevic. From E-Sex to E-Commerce: Web Search Changes. *IEEE Computer* 35, pages 107–109, 2002.

J. Thom and J. Zobel. A model for word clustering. *Journal of American Society for Information Science,* 43(9):616–627, 1992.

W. Willinger and V. Paxson. Where mathematics meets the Internet. *Notices of the AMS,* 45(8):961–970, 1998.

G. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, 1949.

This page intentionally left blank

AN OVERVIEW OF PROBABILISTIC AND TIME SERIES MODELS IN FINANCE

Alejandro Balbás
Dept. of Business Administration, Universidad Carlos III de Madrid

Rosario Romera
Dept. of Statistics and Econometrics, Universidad Carlos III de Madrid

Esther Ruiz
Dept. of Statistics and Econometrics, Universidad Carlos III de Madrid

Abstract   In this paper, we partially review probabilistic and time series models in finance. Both discrete and continuous-time models are described. The characterization of the No-Arbitrage paradigm is extensively studied in several financial market contexts. As the probabilistic models become more and more complex to be realistic, the Econometrics needed to estimate them are more difficult. Consequently, there is still much research to be done on the link between probabilistic and time series models.

2.1   Introduction

Uncertainty plays a central role in financial theory and its empirical implementation. The objective of this paper is to review the connection between the theory and the empirical analysis in the area of Finance. It is obvious that the scope of the subject is too wide and, consequently, we will not be able to cover all contributions in the area. Therefore, in the framework of probabilistic models, we focus on those pricing models reflecting the absence of arbitrage and free-lunch. The problem of valuation and hedging of contingent claims (risks) presents important difficulties when markets imperfections are met. The characterization of No-Arbitrage (NA) is extensively studied in section 2. Pricing of contingent claims when markets are subject to portfolio constraints, transactions costs and taxes as well as new results for nonlinear pricing along with

a universal framework for pricing financial and insurance risks are reviewed in this section.

Section 3 reviews the main time series models devoted to the analysis of financial returns. We start describing models for the conditional mean usually fitted to test whether financial prices are predictable. In this sense, it is generally accepted that asset returns are close to be martingale difference processes. However, they are not independent because of the often observed dependence of some transformations related with second moments. Consequently, we then describe models to represent the dynamic evolution of conditional variances and covariances of high frequency returns. Finally, section 3 reviews the models recently proposed to represent the main empirical properties of ultra high frequency (intra-daily) returns.

In section 4, we focus on the link between probabilistic models and Financial Econometrics. We show that the estimation of realistic financial models for asset prices are, in general, difficult and much research remains to be done in this area. In particular, in this section, we describe the empirical implementation of the CAPM as well as the estimation procedures of the term structure, the VaR and continuous time diffusions.

The paper finishes in section 5 with a summary of the main conclusions.

2.2   Probabilistic models for finance

A classical problem in mathematical finance is the pricing of financial assets. The usual solution of this problem involves the so-called Fundamental Theorem of Asset Pricing. This result ensures that the assumption of NA is essentially equivalent to the existence of an equivalent martingale measure, in a perfect financial market. The NA assumption amounts to saying that there is no plan yielding some profit without a countervailing threat of loss. It prevents the existence of zero cost portfolios with positive return. The problem of fair pricing of financial assets is then reduced to taking their expected values with respect to equivalent martingale measures. Initial results on the Fundamental Theorem of Asset Pricing hold in the case of finite number of assets and a finite discrete time models; see Harrison and Kreps (1979) and Harrison and Pliska (1981).

Various generalizations are now available in the literature. For discrete infinite or continuous time, the notion of "no free lunch" or "no free lunch with bounded (vanishing) risk" is needed, which is a slightly stronger version of the non-arbitrage condition; see, for example, Dalang *et al.* (1989), Back and Pliska (1991) and Schachermayer (1992). In these generalizations, securities markets are assumed to be frictionless, i.e. without considering transaction costs. For discrete infinite case see Schachermayer (1994). For continuous time models see Delbaen (1992) or Delbaen and Schachermayer (1994, 1998);

see also Duffie and Huang (1986), Striker (1990) and Kabanov and Kramkov (1994).

2.2.1   The Fundamental Theorem of Asset Pricing

The mathematical translation of this concept uses martingale theory and stochastic analysis. Under the assumption that the R^d-**valued** price process $\{S_t\}_{t \in R^+}$ reflect economically meaningful ideas and does not generate arbitrage profits, the Fundamental Theorem of Asset Pricing allows the probability P on the underlying probability space (Ω, F, P) to be replaced by an equivalent measure Q such that $\{S_t\}_{t \in R^+}$ becomes a (local) martingale under the new measure. The information structure is given by a filtration $(F_t)_{t \in T}$. Following Delbaen and Schachermayer (1994, 1998), there should be no trading strategy H for the process S, such that the final payoff described by the stochastic integral $(H.S)_\infty$, is a nonnegative function, strictly positive with positive probability.

A buy-and-hold strategy can be described, from the mathematical point of view, as an integrand of the form $H = f.1_{(T_1, T_2]}$, where $T_1 \leq T_2$ are stopping times and f is F_{T_1}-**measurable**. The interpretation of this integrands is clear: when time $T_1(w)$ comes up, buy $f(w)$ units of the financial asset, keep them until time $T_2(w)$ and sell. Stopping times are interpreted as signals coming from available information and this is one reason why, in mathematical finance, the filtration and further concepts such as predictable processes, are so relevant. Even if the process S is not a semi-martingale, the stochastic integral *(H.S)* for a buy-and-hold strategy H can be defined as the process $(H.S)_t = (S_{\min(t, T_2)} - S_{\min(t, T_1)})$. A linear combination of buy-and-hold strategies is called a simple integrand. In the general case simple integrands are not sufficient to characterize these processes that admit an equivalent martingale measure. On the other hand the use of general integrands leads the problem of the existence of *(H.S)*. The so called admissible integrands avoid all of these pathologies.

Formally, if S denotes an R^d-**valued** semi-martingale, defined on the filtered probability space $(\Omega, \{F_t\}_{t \in R^+}, P)$, an R^d-**valued** predictable process H is called α-**admissible** if it is S-integrable, if $H_0 = 0$, if the stochastic integral satisfies $H.S \geq -\alpha$ and if the $\lim_{t \to \infty} (H.S)_t$ exists *a.s.* If H is admissible for some α, then is simply call admissible.

In order to characterize mathematically the NA and the No Free Lunch (NFL) properties, we need to consider the following vector spaces. Let us denote by L^0 the vector space of all real-valued measurable functions defined on Ω. Endowed with the topology of convergence in probability, this space becomes a Fréchet space (i.e. a complete and metrisable vector space). L^∞ denotes the subspace of L^0 of all bounded functions. It is remarkable that the two

spaces L^0 and L^∞ are, among the L^p spaces, the only two spaces that remain the same when the original probability measure is replaced by an equivalent one. Let us to introduce the following sets:

$$\Phi = \{(H.S)_\infty \ / \ H \text{ is admissible}\},$$
$$\Phi_\alpha = \{(H.S)_\alpha \ / \ H \text{ is } \alpha\text{-admissible }\},$$
$$\Upsilon_0 = \Phi - L^0_+,$$
$$\Upsilon = \Phi_0 \cap L^\infty.$$

In all papers dealing with the Fundamental Theorem of Asset Pricing (with simple integrands), the assumption of NA or NFL essentially amounts to saying that the set Φ does not contain any non-negative random variable except the null one.

Formally, we say that the process S satisfies the NA property if:

$$\Phi \cap L^0_+ = \{0\} \tag{2.1}$$

which is equivalent to the expression

$$\Upsilon \cap L^\infty_+ = \{0\}.$$

The process S satisfies the NFL property if

$$\bar{\Upsilon} \cap L^\infty_+ = \{0\}, \tag{2.2}$$

where the bar denotes closure in the norm topology of L^∞.

The NFL is an old expression used in the early days of the finance literature. The NA postulates that the set of random variables which can be achieved by a zero cost portfolio does not include any positive random variable. The NFL condition, postulates the same on the topological closure of the previous set. The following technical definition is due to Kreps (1981). Let S be a bounded process and let us denote by Φ^* the set of all outcomes with respect to bounded simple integrands. Υ^* is defined in the same way $\Upsilon^* = (\Phi^* - L^0_+) \cap L^\infty$.

Then, an adapted process S satisfies the NFL property, as above, if the corresponding set of outcomes does not contain any non-negative random variable except the null, $\tilde{\Upsilon}^* \cap L^\infty_+ = \{0\}$, where the tilde denotes *weak* closure. Dealing with the *weak* closure it may happen that an element of this set can only be obtained by an *unbounded* generalized sequence. Unfortunately the economic interpretation of this unbounded objects is unclear. However requirements of NA and NFL in expressions (1) and (2) are very strong. We assume that S is a semi-martingale and there is an equivalent martingale measure for the process S. On the other hand we need a definition for the set of outcomes with respect to *general* admissible integrands. The following theorem from Delbaen and Schachermayer (1998), characterizes the NFL concept through a boundedness property in L^0.

THEOREM 1 *The process S satisfies the property NFL (2) if and only if it satisfies*

    1 the NA property (1) and

    2 Φ_1 is bounded in the space L^0

They remark that the boundedness of the set Φ_1 has the following economic interpretation: for outcomes that have a maximal loss bounded by 1, the profit is bounded in probability, this means that the probability of making a big profit can be estimated from above, uniformly over all such outcomes.

For further characterization of the NFL property and related results for locally bounded semi-martingales *S,* see Delbaen and Schachermayer (1994, 1998).

A recent projective system approach to the martingale characterization of the absence of arbitrage is provided by Balbás *et al.* (2002). The equivalence between the absence of arbitrage and the existence of an equivalent martingale measure fails when an infinite number of trading dates is considered. Thus, enlarging the set of states of nature and the probability measure through a projective system of perfect measure space, the authors characterize the absence of arbitrage when the time set is countable.

The martingale characterization can be extended in the context of imperfect financial models, mainly financial models with proportional transaction costs, short sale constraints, convex cone constraints, etc.

We can observe three main lines of research generalizing these initial results. The first one applies in the context of imperfect financial markets for a model with transaction costs. The second line of research expands the restricted feasible portfolio case, usually cone constraints. The third research direction and the most recent one is based on the assumption that the price is non-linear with respect to the portfolio. Then the subadditivity property is needed and the Choquet integral is a powerful tool to be used in this context. The asset pricing problem is then solved as a Choquet integral of the future returns with respect to a new capacity introduced by Chateaunef *et al.* (1994,1996).

Currently there is a pressing need for a universal framework for the determination of the fair value of financial and insurance risks. In the financial services industry, this pressing need is evidenced by the recent Basel Accords on regulatory risk management that require fair value, analogous to market prices, to be applied to all assets or losses, whether traded or not. More recently Wang (2000, 2001) presents a universal framework for pricing financial and insurance risks.

2.2.2   Asset Pricing in Imperfect Financial Markets

In the classical setting, the financial market is modeled in a "frictionless" way which is a clear idealization of the real world. Therefore models with transaction costs have been increasingly studied in the literature; see Davis and Norman (1990) or Striker (1990). Jouini and Kallal (1995a) characterize the assumption of NFL in a model with transaction costs and give fair pricing intervals for contingent claims in such a model. As for other imperfection, Jouini and Kallal (1995b, 1999) consider the case of short sale constraints or shortselling costs with possibly different rates for borrowing and lending rates. The problem of hedging contingent claims, in continuous time, is study by Cvitanic and Karatzas (1996). They propose a diffusion model (with one bond and one risky asset) with proportional transaction costs, and give a dual formulation for the so-called super-replication price of a contingent claim (i.e. the minimum initial wealth needed to hedge the contingent claim, or in other words, to obtain, through the investment opportunities available on the market, at least the contingent claim). Delbaen *et al.* (1998) generalize this result to the multivariate case, in discrete as well as in continuous time, and with a semi-martingale price process. In these models too, typically there is a "bond" which serves as numeraire asset. The usual assumption is that, at final date *T*, all the positions in the other traded assets are liquidated, i.e., converted into units of the bond.

More recently, Jouini and Napp (2002) generalize existing results in the following ways: first, they do not assume that there exists a numéraire available to investors and allowing them to transfer money from one date to another; this enables to consider any type of friction on the numéraire-like no borrowing, different borrowing and lending rates, bonds with default risk, etc. These setting also take into account the fact that all investors are not equal with regard to borrowing and lending, namely some investors may enjoy special borrowing facilities while others may not; second, they are led to introduce a new notion of NFL, which is the classical concept in finite time but does not exclude a free lunch at infinite and is therefore may be more economically meaningful; last, they characterize the NFL assumption for very general investments, which enables to consider investment opportunities that are not necessary related to a market model and, to generalize the results obtained for imperfect markets and to obtain them all in a unified way. Technically, all investment opportunities are described in terms of cash flow. Therefore, separation techniques in more complex spaces to obtain the Fundamental Theorem of Asset Pricing are needed. Let consider their main *Assumption A.*

DEFINITION 2 *An investment is an* $(F_t)_{t \in T}$-*adapted process* $H = (H_t)_{t \in T}$, *null outside a finite number of dates, i.e. there exists* $(t_1^H, \ldots t_N^H)$ *such that* $H_t = 0$ *for all* $t \notin (t_i^H)_{i=1}^N$, *and such that* H_t *is in* $L^1(\Omega, F_t, P)$ *for all* $t \in T$.

DEFINITION 3 (ASSUMPTION A) *There exists a sequence $d = (d_n)_{n \geq 0}$ such that for all $t^* \geq 0$, for all B_{t^*} in F_{t^*} of positive probability, there exists H in the convex cone of investment opportunities J, of the form $H_{t^*} = 0$ outside B_{t^*}, $H_t = 0$ for all $t < t^*$, $H_t \geq 0$ for all $t > t^*$, and there exists $d_n \in d$, $P[H_{d_n} > 0] > 0$.*

Roughly, Assumption A corresponds to the possibility of transferring "some money" from any date and event to some particular date. This assumption is not too restrictive: it is satisfied if we can buy at every date and event a bond with a given maturity even if this bond is defaultable and even if there is no secondary market for that bond (i.e. we have to wait until maturity in order to recover any money with a positive probability, which may be different from 1); this includes market models with frictions on the numeraire like no borrowing, different borrowing and lending rates, bonds with default risk, different borrowing facilities among the investors. More generally, it is satisfied if there is at least one asset whose price cannot be negative (which is usually the case for stocks or for options, defaultable bonds, etc.).

Then a characterization of the NA property in a model with flows is given by Jouini and Napp (2002) in the following theorem.

THEOREM 4 *Let J denote a convex cone of investments satisfying Assumption A. There is NFL for J if and only if there exists a process $g = (g_t)_{t \in T}$ satisfying for all t in T, $P[0 < g_t < M]$ for some M in R_+, and such that*

$$E[\sum_{t \in T} g_t H_t] \leq 0 \quad \text{for all} \quad H = (H_t)_{t \in T} \in J.$$

Moreover, the process g can be taken $(F_t)_{t \in T}$-adapted.

In other words, there is NFL for a convex cone of available investments satisfying Assumption A if and only if a given convex set of "admissible" discount processes is non-void. The theorem ensures the existence of a "discount process" such that, using this process as deflator, all available investments have non-positive present value; this means that there exists a term structure such that the market consisting of the primitive investment opportunities and of the additional borrowing and lending facilities is still "arbitrage-free". Besides, the existence of such a discount process prevents from any arbitrage opportunity. Notice that Assumption A is not needed to obtain this result if the set of investment opportunities is related to a countable set of dates.

Since most market models with frictions can fit in the model with flows for a specific convex cone of available investments, the model in Jouini and Napp (2002) provides a unified framework for the study of the characterization of the absence of FL in such imperfect market models. However this model with flows does not stand for economies with fixed transaction costs, since the set of available investments is not a cone.

Kabanov (1999, 2001) develops a mathematical theory of currency markets with transaction costs based on ideas of convex geometry. He proposed an

appealing framework to model financial markets in a numeraire-free way for both frictionless markets and markets with transaction costs. This approach turns out to be conceptually interesting, even in the frictionless case, as it allows for a new look on the wealth processes, arising in financial modelling, without explicitly using stochastic integration: expressing portfolios in terms of the number of *physical units* of the assets, as opposed to the *values of the assets in terms of some numéraire,* opens new perspectives. Basically, the financial market is modelled by a *dxd* matrix-valued stochastic process specifying the mutual *bid and ask* prices between *d*-assets. The terms of trade at time t are modeled via an F_t-**measurable** non-negative *dxd* matrix -valued map $\omega \to \Sigma_t(\omega)$ denoting the bid and ask prices for the exchange between the *d* assets. The entry σ^{ij} of Σ_t denotes the number of units of asset i from which an agent can trade in one unit of asset j in terms of the asset i bid-ask processes are defined as adapted processes taking values in the set of bid-ask-matrices. $\sigma_t^{ij}\sigma_t^{ji} = 1$ a.s. for all $1 \le i, j \le d$ and $t = 0, ..., T$ in the frictionless case.

Kabanov *et al.* (2001) introduce the bid-ask process in a somewhat indirect way. They start with a *d*-**dimensional** price process which models the prices of the *d* assets without transaction cost in terms of some numeraire (it may be a traded asset or not). One then defines a non-negative *dxd* -matrix $\Lambda = (\lambda^{ij})_{1 \le i,j \le d}$ of transaction cost non-negative coefficients λ^{ij}, modelling the proportionally factor one has to pay in transaction costs, when exchanging the i'th into the j'th asset. Then the bid-ask process is obtained as

$$\sum_t(\omega) = Diag(S_t(\omega))^{-1}(1 + \Lambda_t(\omega))Diag(S_t(\omega)), \qquad (2.3)$$

where *1* denotes the unit matrix (not to be confused with the identity matrix).

Schachermayer (2002) presents a direct modelization of the bid-ask process $\Sigma = (\sum_t)_{t=0}^T$ without first defining $(S_t)_{t=0}^T$ and Λ. It seems more natural, from an economic point of view, as in a market with friction an agent is certainly faced with a bid-and an ask-price. But these prices are not necessarily decomposed into a "frictionless" price and additional transaction costs.

The notion of *consistent price system* (resp. *strictly consistent*) introduced by Kabanov and his co-authors extends the notion of equivalent martingale measures. Similar notions are in Schachermaver (2002).

DEFINITION 5 *An adapted* R_+^d *valued-process* $Z = (Z_t)_{t=0}^T$ *is called a consistent (resp. strictly consistent) price process for the bid-ask process* Σ, *if Z is a martingale under P, and* $Z_t(\omega)$ *lies in* $K_t^*(\omega)\backslash\{0\}$ *(resp. in the relative interior of* $K_t^*(\omega))$ *a.s., for each t=0,...,T,.*

$K_t^*(\Sigma) = \{ \omega \in R^d : \; < v, \omega > \; \geq 0, \; for \; v \in K(\Sigma)\}$ is the polar of. $-K(\Sigma)$, and $K(\Sigma)$ is the solvency cone, i.e., the convex cone in $\mathbf{R}^d$ spanned by the unit vectors $e^i, 1 \leq i \leq n$, and the vectors $\sigma^{ij}e^i - e^j, 1 \leq i, j \leq d$.

The cone $K^*(\Sigma)$ has a nice economic interpretation, eluded by the term "consistent price system". A vector $\omega \neq 0$ is in $K^*(\Sigma)$ if it defines a friction-less pricing system for the assets 1,...,d which is consistent with the bid-ask-matrix Σ in the following sense: if the price of asset i (denoted in terms of some numéraire) equals ω^i, then the friction-less exchange rates, denoted by τ^{ij}, clearly equal

$$\tau^{ij} = \frac{\omega^j}{\omega^i}, 1 \leq i, j \leq d.$$

>From the economical point of view, a consistent price system $\omega = (\omega^i)_{i=1}^d$ is strictly consistent if, for all $1 \leq i, j \leq d$, the exchange rate $\tau^{ij} = \frac{\omega^j}{\omega^i}$ is in the relative interior of the bid-ask spread $[\frac{1}{\sigma^{ji}}, \sigma^{ij}]$.

The main theorem in Kabanov *et al.* (2001) is the following version of the Fundamental Theorem of Asset Pricing: under an additional assumption, a bid-ask process Σ satisfies the strict NA condition, if there is a strictly consistent price system Z for Σ. The additional assumption is called "efficient friction" and requires that $F_t(\omega) = \{0\}$, a.s., for all $t = 0, ..., T$. It was asked by these authors whether this additional assumption can be dropped. Schachermayer (2002) gives an example of a bid-ask process Σ, with $d = 5$ and $T = 2$, showing that, in general, the answer to this question is no. In the same paper a slight strengthening of the notion *strict NA*, called the *robust no arbitrage NA^r* is introduced. A subsequent Fundamental Theorem of Asset Pricing as a main result is then formulated.

2.2.3   Asset Pricing with Cone Constraints

Pham and Touzi (1999) addresses the problem of characterization of NA in the presence of frictions in a discrete-time financial market model. They extend the Fundamental Theorem of Asset Pricing with cone constraints on the trading strategies under a nondegeneracy assumption. In the presence of transaction costs and under a nondegeneracy condition on the risky assets price process, they also prove that the NFL and the NA conditions are locally equivalent i.e. when trading is restricted to some period $[t - 1, t]$. Their main result states the equivalence of the no local arbitrage condition and the existence of an equivalent probability measure satisfying a further generalization of the martingale property. They do not provide a multiperiod version of this result. For a more general setting of convex constraints see Brannath (1997).

2.2.4   **Nonlinear Asset Pricing**

On financial markets without frictions, no-arbitrage pricing allows to price non-marketed redundant assets using the equilibrium prices of the marketed assets. Assets are then valued by a linear function of their payoffs (mathematical expectation). The equilibrium prices of the marketed assets determine a set of *risk neutral* probability distributions such that the equilibrium price of a redundant asset equals the mathematical expectation of its discounted payoff with respect these probability distributions. This pricing rule is consistent with equilibrium in the sense that, introducing a redundant asset at its no-arbitrage price does not affect the equilibrium allocation; see, for example, Harrison and Kreps (1979). In markets with frictions, pricing rules may be non-linear. Two portfolios yielding the same payoffs need not have the same formation cost (net of transaction cost), but the difference may not imply the existence of a free lunch because of frictions. Consider for example bid-ask spreads or transaction costs. Then clearly prices (as a function of asset payoffs) are non-linear, since the price an agent has to pay for buying an asset is strictly larger than the price an agent receives for selling it. Therefore equilibrium asset prices cannot be represented by the mathematical expectation of their discounted payoff with respect to a probability measure.

Asset valuation by a Choquet integral is introduced in Chateauneuf *et al.* (1996). They introduce a nonlinear valuation formula similar to the usual expectation with respect to the risk-adjusted probability measure. This formula expresses the asset's selling and buying prices set by dealers as the Choquet integrals of their random payoffs. In this paper bid-ask spreads are considered. Bid-ask spreads is one of many types of friction prevailing in financial markets which differs from the traditional formalization of proportional transaction costs.

Let consider the following situation pointed out by Chateauneuf *et al. (*1996): assumed that a dealer sells an asset Y (defined by its flow of payoffs) at a price $q(Y)$ and that she buys it a price $-q(-Y)$ such that she makes the positive profit $q(Y) + q(-Y) > 0$. Then, because $0 = q(0) = q(Y - Y)$, q cannot be linear, hence it cannot be calculated as $q(Y) = \int_S Y d\mu$, where S is the set of random states and μ is some risk-adjusted probability over S. In these settings, the paper imposes certain axioms on prices (generalizing the usual no-arbitrage conditions) and deduces from them a result on the structure of prices (representation as Choquet integral: an expectation with respect to a concave capacity). Capacities were introduced by Schmeidler (1989) in individual decision theory. Formally, a capacity on a measurable space (S, F_S) is a set of functions $\nu : F_S \rightarrow [0, 1]$ satisfying $\nu(F_S) = 1, \nu(\Phi) = 0$. Furthermore ν is said to be convex (resp. concave or supermodular) if

$$\nu(A \cup B) + \nu(A \cap B) \geq (resp. \leq) \nu(A) + \nu(B), \ for \ all \ A, B \in F_S.$$

In this context, a convex capacity is interpreted as a representation of risk (uncertainty) aversion. This characterization of uncertainty aversion has been used in single-agents models for which convex capacities are representations of individual behaviors. In contrast, Chateauneuf *et al.* (1996) use a model for which agents are price takers and the concave capacity is derived from prices.

Formally, the model uncertainty they consider is described by the measurable state space (S, F_S) where F_S is a given σ-algebra of events of S. An asset is defined by the random variable X of its payoffs. Bounded assets are considered. These assets are sold and bought by a dealer to agents. Hence, all traded assets have a bid and an ask price fixed by the dealer. These prices are described by $q(Y)$ and $-q(-Y)$ respectively, i.e., the prices at which the dealer sells asset Y to agents and buy asset Y from agents. Three axioms on prices which generalize the usual NA conditions to market with a dealer are then imposed. The first is the usual NFL. The second one, as is usually done in pricing models, assumes no transaction costs on riskless assets. The third axiom replaces the (usually implicit) tight markets condition. Traditionally, two portfolios yielding the same payoffs must have the same price, implying that price functional is linear. Taking into account potential reduction of risks when portfolio $X + Y$ is sold instead of X or Y alone induces the dealer to sell $X + Y$ at a discount to X and Y.

A typical example where hedging effects occur and X and Y are not comonotone (*comonotonicity := for all $s, s' \in S$, $[X(s) - X(s')][Y(s) - Y(s')] \geq 0$*), is the following one from Chateneauf *et al.* (1996). Suppose that X offers 1000 if even B occurs, 5000 otherwise, Y offers 5000 if B occurs, 1000 otherwise. Clearly X and Y are not comonotone and X (resp Y) is a hedge against Y (resp. X) since $X + Y$ is riskless: it offers 6000 with certainty. So, subadditivity for q : $q(X + Y) \leq q(X) + q(Y)$ is required. Notice that, consequently, no discount will be offered by the dealer when X and Y are comonotone; i.e., $q(X + Y) = q(X) + q(Y)$ if X and Y are comonotone. Then the third axiom (Comonotonicity Premium) expresses for all $X, Y \in A$: $q(X + Y) \leq q(X) + q(Y)$ equality holds if X and Y are comonotone. Their main result is the so-called Choquet Sublinear Pricing Theorem. Under the three axioms as above this theorem asserts that there exists a unique concave capacity ν on the set of states S such that the value of an asset X is defined by $q(X) = Max\{\int_S X \, d\mu; \mu$ is an additive probability s.t. $\mu \leq \nu\}$. The price of X is the Choquet integral of its payoffs: $q(X) = \int_S X \, d\nu$, where

$$\int_S X d\nu = \int_{R_-} [\nu(X \geq t) - 1] dt + \int_{R_+} [\nu(X \geq t) dt,$$

and q is sublinear (i.e. subadditive and positively homogeneous, and indeed q is concave).

Application to pricing "primes" and "scores" are given in the paper of Chateauneuf *et al.* (1996).

In these settings De Waegenaere *et al.* (1996) propose a pricing rule for the valuation of assets on financial markets with intermediaries. They assume that the non-linearity arises from the fact that dealers charge a price for their intermediation between buyer and seller. The price of an asset equals the signed Choquet integral of its discounted payoff with respect to a concave signed capacity. Furthermore, they show that this pricing rule is consistent with equilibrium and equilibria satisfy a notion of constrained Pareto optimality.

On the other hand, a universal framework for pricing financial and insurance risks has been introduced recently by Wang (2000) who proposes a pricing method based on the following transformation $F^*(x) = \phi[\phi^{-1}(F(x)) + \lambda]$, where ϕ is the standard normal cumulative distribution. The key parameter λ is called the *market price of risk,* reflecting *the level of systematic risk.* For a given asset X with $F(x) = \Pr\{X \leq x\}$, the Wang transform will produce a "risk-adjusted" cumulative probability distribution $F^*(x)$. The mean value under $F^*(x)$ will define a risk-adjusted "fair value" of X at time T, which can be further discounted to time zero, using the risk-free interest rate. This approach is partly inspired in the work of Venter (1991) and Butsic (1999).

2.3   Time series models

In this section, we revise the literature on the time series models usually fitted to financial data. As this is a very broad area, the focus is only on the main branches of the literature with special attention to the most recent developments. Campbell *et al.* (1997) and Tsay (2002) present excellent textbook reviews of Financial Econometrics and Bollerslev (2001) and Engel (2001, 2002a) have very interesting discussions on past developments and future perspectives in this area.

Traditionally, the two main motivations to use time series models to analyze financial data are to represent the empirical properties often observed in real prices and to estimate and test the financial models described in section 2. In this section, we describe models proposed mainly to represent the empirical properties of financial prices while section 4 is devoted to the relationship between time series models and Finance theory.

The empirical properties of financial prices depend crucially on the frequency of observation. We consider three main classes of frequencies. First, it is possible to observe prices at very high frequencies as, for example, tick by tick or hourly prices. These observations are called Ultra-high-frequency (UHF) data by Engle (2000) and they are usually characterized by unequally

spaced and discrete-value observations. Another important property is the presence of strong daily patterns with highest volatility at the open and toward the close of the day. On top of this intraday volatility pattern, UHF returns are characterized by highly persistent conditionally heteroscedastic components along with discrete information arrival effects; see Andersen and Bollerslev (1997a, 1997b, 1998), Müller *et al.* (1997) and Andersen *et al.* (2001). Finally, it is possible to have multiple transactions within a single second.

Prices can also be observed at high frequencies, as for example, daily or weekly. This frequency is the most extensively analyzed in the empirical literature. There is a vast number of papers that show that high frequency returns are nearly non-correlated although they are not independent because there are non-linear transformations, as squares or absolute values, that have significant autocorrelations. Furthermore, these autocorrelations are usually small and decay very slowly towards zero. The significant autocorrelations of squared returns are often related with the presence of volatility clustering, i.e. periods of low volatility are usually followed by periods of low volatility and viceversa. Furthermore, the slow decay is usually interpreted as the presence of long-memory in the volatility; see Lobato and Savin (1998) and Granger *et al.* (2000) and the references therein. On the other hand, high frequency returns are often leptokurtic and, consequently, non-Gaussian. The heavy tails property of returns can also be related with the dynamic evolution of volatility.

Finally, prices are sometimes observed at very low frequencies as, for example, monthly. Tsay (2002) shows that monthly returns still have excess kurtosis although smaller than in lower frequencies. On the other hand, monthly returns seem to have more serial correlations than daily returns. Given that low frequencies are not in general of interest for asset pricing models, the focus in this section is on UHF and high frequency observations.

The rest of the section is organized as follows. Subsections 3.1 to 3.3 deal with models for high frequency observations. In subsections 3.1 and 3.2, we describe the models usually fitted to represent expected returns and volatilities respectively. In subsection 3.3, we consider multivariate models for systems of returns. Finally, in subsection 3.4, we describe models for UHF data.

2.3.1   Models for the conditional mean

One of the central questions in the Financial Econometrics literature is whether financial prices are predictable and this is still a topic of controversy; see, for example, the special issue of the *Journal of Empirical Finance, 8 (2001)*. In this section we describe univariate models and, consequently, the problem is whether future prices can be predicted with information contained in their own past. The main hypothesis that have often been tested are the martingale

and the random walk hypothesis. The martingale hypothesis can be expressed as follows:

$$E[P_t|P_{t-1}, P_{t-2}, ...] = P_{t-1} \qquad (3.1)$$

Therefore, given the prices up to time $t-1$, the price at time t is expected to be equal to the price at time $t-1$. The martingale hypothesis places a restriction on expected returns but does not take into account the risk. However, as said in section 2, once asset returns are properly adjusted for risk, the martingale hypothesis holds for rationally determined asset prices; see Harrison and Kreps (1979). It is known that, the risk-adjusted martingale property is the basis of many financial derivatives as, for example, options and swaps; see, for example, Merton (1990) and Campbell *et al.* (1997).

The second hypothesis often tested in the financial literature is whether prices are generated by a random walk plus drift model given by:

$$P_t = \mu + P_{t-1} + \varepsilon_t \qquad (3.2)$$

where ε_t is an independent process with zero mean and variance σ^2 and μ is the expected price change. In model (5), if the distribution of the errors ε_t is, for example, Gaussian, there is a positive probability that prices can be negative, violating limited liability. Therefore, it is usual to assume the random walk model not for prices but for logarithmic prices, i.e.

$$\log(P_t) = \mu + \log(P_{t-1}) + \varepsilon_t \qquad (3.3)$$

In model (6) any arbitrary transformation of prices is unforecastable using any arbitrary transformation of past prices. However, it is usual to assume that the errors ε_t are merely uncorrelated instead of independent allowing, for example, for the presence of conditional Heteroscedasticity.. As we have mentioned before, this is a property often observed in high frequency returns. Consequently, we will focus on tests of the random walk hypothesis where ε_t is uncorrelated.

When testing the null hypothesis that the autocorrelation coefficients of returns, $r_t = \triangle \log(P_t)$, are all zero, it is important to take into account that ε_t is not independent because, usually, ε_t^2 is correlated. Therefore, the traditional tests for uncorrelatedness should be adequately modified; see Romano and Thombs (1996) and Lobato *et al.* (2001) among others.

Alternatively, the random walk hypothesis can be tested using the Variance Ratio (VR) statistic. This test is based on the property that the variance of random walk increments is a linear function of time interval; see Campbell *et al.* (1997) for a detailed description of the VR test.

The implementation of the previous tests to financial prices, seems to suggest that financial asset returns are predictable; see the special issue of the

Journal of Empirical Finance, 8 (2001) and the references therein. There are several alternative explanations for this predictability. For example, Campbell *et al.* (1997) and Lo and MacKinlay (1990) show that nonsynchronous trading can introduce negative autocorrelations in returns. The bid-ask spread can also introduce negative autocorrelations in asset returns; see, among others, Campbell *et al.* (1997). Other possible explanations are time-varying risk premiums as in Harvey (2001) and Bekaert *et al.* (2001), irrational behavior of market participants in Hong and Stein (1999), Benartzi and Thaler (1995), Barberis *et al.* (2001) and Epsein and Zin (2001), market frictions as transaction costs or agency problems or fluke due to statistical inference.

2.3.2 Models for the conditional variance

Although, it is generally accepted that asset returns appear to be close to a martingale difference process, there is an overwhelming evidence that they are not independent due to autocorrelated squares. Assuming that returns have zero mean and are serially uncorrelated, they can be represented by the following model:

$$r_t = \sigma_t \varepsilon_t \qquad (3.4)$$

where ε_t is an independent and identically distributed (i.i.d.) process with zero mean and unity variance independent of the volatility, σ_t. There are two main proposals in the literature to represent the dynamic evolution of σ_t: Generalized Autoregressive Conditional Heteroscedasticity (GARCH) and Stochastic Volatility (SV) models.

GARCH models, originally proposed by Engle (1982) and Bollerslev (1986), are based on modelling the volatility as the variance of returns conditional on past observations. There is a pleyade of papers where GARCH models are investigated from a theoretical point of view or are applied to the empirical analysis of financial time series. The main properties of GARCH models have been reviewed, among others, by Bollerslev *et al.* (1995) and Carnero *et al.* (2001a). Although the original motivation of GARCH models was mainly empirical, Nelson (1992) shows that even when mispecified, ARCH models may serve as consistent filters for the continuous–time stochastic volatility diffusions often employed in the asset pricing literature. Furthermore, Nelson (1990, 1994) and Nelson and Foster (1994) provide some important links between GARCH and the corresponding continuous–time models.

The original GARCH model has been extended in a huge number of directions. Two of the main extensions from the empirical point of view, are models to represent the asymmetric response of volatility to positive and negative returns and to represent the effect of the volatility on the return of a stock. The first effect is known as *leverage effect* and was introduced by Black (1986).

The first model proposed to represent the leverage effect was the Exponential GARCH (EGARCH) model of Nelson (1991). Later, Hentschel (1995), Duan (1997) and He and Terasvirta (1999) have proposed models general enough to unify many of the main previous ARCH-type models. With respect to the effect of volatility on the expected return, Engle *et al.* (1987) introduced the GARCH in mean (GARCH-M) model given by

$$
\begin{aligned}
r_t &= \mu + c\sigma_t^2 + a_t \\
a_t &= \sigma_t \varepsilon_t \\
\sigma_t^2 &= \omega + \alpha a_{t-1}^2 + \beta \sigma_{t-1}^2
\end{aligned}
\tag{3.5}
$$

The parameter c is known as the *riskpremium* parameter. Returns generated by the GARCH-M model are autocorrelated because of the autocorrelations of the volatility, σ_t^2.

There are many other generalizations of the original GARCH model. For example, Zakodian (1994) allows for regime switching where volatility persistence can take different values depending on whether returns are in a high or a low volatility regime. To represent the long memory property of squared returns, Baillie *et al.* (1996) introduce the Fractionally Integrated GARCH (FIGARCH) model. Although the FIGARCH model has been fitted in several empirical applications, it is not stationary in covariance and, consequently, the properties of the corresponding estimators and tests are generally unknown. Finally, Engle and Lee (1999) have proposed a GARCH model with two components in volatility: one which is nearly nonstationary and another that is much less persistent.

All GARCH models have the attractive that can be easily estimated by Maximum Likelihood techniques. However, Terasvirta (1996) and Carnero *et al.* (2001b) show that the basic GARCH(1.1) model is not flexible enough to represent adequately the properties often observed in real time series of returns.

Alternatively, the volatility, σ_t^2, can be modelled using SV models that introduce an additional noise in its equation. Therefore, the volatility is a latent variable composed of a predictable component, that depends on past returns, plus an unexpected component. SV models were originally proposed by Taylor (1986) and their properties have been reviewed by Taylor (1994), Ghysels *et al.* (1996) and Shephard (1996). The introduction of the unobserved component in the representation of the volatility, gives more flexibility to SV models to represent the empirical properties often observed in real time series of returns; see Carnero *et al.* (2001b). However, the estimation of these models present some added difficulties over the estimation of GARCH models. The likelihood function has not a close form and, consequently, most estimation methods proposed in the literature are based on numerical approximations of the likelihood or on transformations of the observations. Although, there is

not still a consensus about which are the most adequate methods to estimate SV models, recently there has been important progress towards methods that are computationally feasible and, at the same time, have properties similar to the Maximum Likelihood estimators; see Broto and Ruiz (2002) for a detailed description of estimation methods for SV models.

Recently, Chib *et al.* (2002) have proposed the following SV model where returns can contain a jump component to allow for large, transient movements,

$$
\begin{aligned}
r_t &= x_t'\beta + k_t q_t + w_t^\gamma \sigma_t \varepsilon_t \\
\log \sigma_t^2 &= \mu + z_t'\alpha + \phi(\log \sigma_{t-1}^2 - \mu) + \eta_t
\end{aligned}
\tag{3.6}
$$

where x_t, w_t and z_t are covariates and γ denotes the level effect. The covariate w_t is a non-negative process as, for example, lagged interest rates; see Andersen and Lund (1997). The noises ε_t and η_t are mutually independent Student-t and Gaussian white noise processes respectively, both with zero mean and variances one and σ_η^2. Finally, with respect to the jump component, q_t is a Bernoulli random variable that takes value one with probability κ and k_t is the size of the jump distributed as $\log(1 + k_t) \sim N(-0.5\delta^2, \delta^2)$. They argue that model (9) without the jump component can be thought of as an Euler discretization of a Student-t Lévy process with additional stochastic volatility effects. This process has been used in the continuous time options and risk assessment literature; see, for example, Barndorff-Nielsen and Shephard (2002b), Eberlein (2002) and Eberlein and Prause (2002). On the other hand, models with jumps have also been frequently applied in continuous time models of financial asset pricing; see, for example, Merton (1976), Ball and Torous (1985), Bates (1996), Duffie *et al.* (2000) and Barndorff-Nielsen and Shephard (2001). From the point of view of the Financial Econometrics literature, SV models with jumps have been previously considered by Chernov *et al.* (2000), Barndorff-Nielsen and Shephard (2002a) and Eraker *et al.* (2003).

As in the case of GARCH models, SV models have also been extended to represent the asymmetric response of volatility to negative and positive returns and the response of expected returns to volatility by Harvey and Shephard (1996) and Koopman and Uspensky (2002) respectively. Another extension of SV models considered in the literature is to allow for long memory in volatility; see Harvey (1998) and Breidt *et al.* (1998).

2.3.3   Models for conditional covariances

Multivariate models have been often used to represent financial series of returns related, for example, with the Asset Pricing Theory (APT), asset allocation, estimation of time-varying betas or Value at Risk (VaR). However, although numerous multivariate models for returns have been proposed, there

is not jet a consensus about which models are better mainly due to a dimension-
ality problem. The literature on multivariate GARCH models is often related
with the lack of parsimony of these models and the constraints needed to guar-
antee that the conditional covariance matrix, Σ_t, is positive definite; see Engle
(2002a,b) who revises the most popular multivariate models proposed in the
context of GARCH. The dimensionality becomes very quickly a problem be-
cause the conditional covariance matrix of a k-dimensional return series has
k(k+1)/2 distinct quantities. To keep the number of parameters low, Boller-
slev (1990) considers a multivariate GARCH model with constant correlations
that always satisfies the positive-definite condition of Σ_t. The constant cor-
relation hypothesis can be tested using the Lagrange multiplier test proposed
by Tse (2000). Because of its computational simplicity, the constant correla-
tion model of Bollerslev (1990) has been widely used in the empirical analysis
of financial data. However, if the correlations evolve over time, this model is
inadequate and can give incorrect inferences. Very recently, there have been
different proposals of multivariate GARCH models with time varying condi-
tional correlations. For example, Tsay (2002) proposes two alternative ways of
dealing with the conditional covariance matrix. The first one consists of model-
ing directly the evolution of the autocorrelation and the second is based on the
Cholesky decomposition of Σ_t. The attractive of the second alternative is that
it does not require any constraint to ensure the positive definiteness of Σ_t. Al-
ternatively, Tse and Tsui (2002) propose a multivariate GARCH (MGARCH)
model with time-varying correlations where the constraints required to ensure
positive definite covariance matrix can be imposed during the optimization
procedure. Finally, Engle (2002b) proposes a nonlinear Dynamic Conditional
Correlation (DCC) model that can be estimated in two steps from univariate
GARCH models. Alternatively, Ledoit *et al.* (2003a) also propose a two step
estimation procedure of the original unrestricted diagonal-Vech multivariate
GARCH(1,1) model of Bollerslev *et al.* (1988) given by

$$Cov(r_{it}, r_{jt} \mid \Omega_{t-1}) = h_{ij,t} = c_{ij} + a_{ij} r_{it-1} r_{jt-1} + b_{ij} h_{ij,t-1} \qquad (3.7)$$

In the first step, the parameters are estimated separately by estimating the
two-dimensional or one-dimensional equations in (10). Then, the estimated
matrices are transformed to guarantee positive semi-definiteness.

An extensive and detailed comparison between the alternative models to
represent time-varying correlations is still to be done.

Another completely different approach to simplify the dynamic structure of
a multivariate volatility process is to use factor models. Multivariate factor
models provide a way of dealing with the APT; see, for example, Campbell
et al. (1997) for a very simple exposition. Denoting by y_t the $N x 1$ vector of
returns at time t, it is given by

$$\begin{aligned} r_t &= \alpha + B f_t + \varepsilon_t & (3.8) \\ (\varepsilon_t^{'} \, f_t^{'})^{'} &\sim NID(0, D) \end{aligned}$$

where D is a diagonal matrix, B is the matrix of factor loadings and f_t is a K dimensional vector of factors. The APT says that, as the dimension of r_t increases (approximating the market), then $\alpha \simeq \iota r + B\lambda$, where r is the riskless interest rate, ι is a vector of ones and λ is a vector representing the factor risk premium associated with the factors often identified as the variances of the factors. However, the normality assumption in (11) is usually inadequate for high frequency series of returns. Consequently, this assumption has been relaxed in the consequent literature. Diebold and Nerlove (1989) and King *et al.* (1994) analyze factor models where the factors and idiosyncratic errors follow their own ARCH process. Sentana and Fiorentini (2001) show that the identifiability restrictions for conditionally heteroscedastic factor models are less severe than in static factor models.

In the context of SV models, the first multivariate model was originally proposed by Harvey *et al.* (1994) who allow the variances and covariances to evolve through time with possibly common trends. Later, Ray and Tsay (2000) used the same model to study common long memory components in daily stock volatilities of groups of companies. However, the multivariate SV model of Harvey *et al.* (1994) restricts the correlations to be constant over time. Later, Jacquier *et al.* (1995) propose a factor SV model given by

$$\begin{aligned} r_t &= B f_t + \varepsilon_t & (3.9) \\ \varepsilon_i &\sim NID(0, I) \\ f_i &\sim SV(\phi^{f_i}; \sigma_\eta^{f_i}; 0), i = 1, ..., K \end{aligned}$$

Kim *et al.* (1998) generalize model (12) by allowing the idiosyncratic noises to follow independent univariate SV models. Then, Aguilar and West (2000) and Pitt and Shephard (1999) implement the model using two alternative Monte Carlo Markov Chain (MCMC) techniques. Finally, Tsay (2002) presents a MCMC estimation of the multivariate SV model based on the Cholesky decomposition.

2.3.4 Models for intradaily data

The analysis of UHF data is closely related with what is known as Market Microstructure and is one of the most active research areas in Financial Econometrics. However, traditional econometric tools may not be appropriate as tick by tick observations are not equally spaced and discrete valued. In this case, it is possible to use market point processes or continuous time methods in which the sampling frequency is determined by some notion of time

deformation; see, for example, Andersen (1996). With respect to using UHF data to estimate the volatility, Andersen and Bollerslev (1998) show that the precision of volatility forecast is improved if the data are sampled more frequently. However, UFH data are affected by problems as the bid-ask spread or non-synchronous trading that, as previously mentioned, can generate autocorrelations in returns. Andersen *et al.* (2001) develop new robust methods for inference in the UHF data setting. Their approach is based on an extension of the Fourier Flexible Form (FFF) regression framework.

Hausman *et al.* (1992) proposes an ordered probit model to study price movements in transactions data where the explanatory variables are the duration between trades, the bid-ask spread, the lagged values of price change and volume, the return of the S&P500 index and an indicator variable that depends on the bid and ask prices. Alternatively, Rydberg and Shephard (2003) propose to decompose the price change into three components: an indicator for the price change, the direction of the change and the size of the change.

Finally, when analyzing UHF data, it is important to model not only the trades but also the timing between trades. In this sense, Engle and Russell (1998) propose the Autoregressive Conditional Duration (ACD) model that estimates the distribution of the time between events conditional on past information. Later, Dufour and Engle (2000) show that the more frequent the transactions, the greater the volatility. Furthermore, they show that transaction arrivals are predictable based on economic variables as the bid-ask spread. Zhang *et al.* (2001) extend the ACD model to account for nonlinearity and structural breaks in the data. Finally, Tsay (2002) introduces the Price Change and Duration (PCD) model to describe the multivariate dynamics of prices changes and associate durations.

2.4   Applications of time series to financial models

Summarizing the literature described in sections 2 and 3, it seems rather clear that there is a gap between the theoretical asset pricing and the Financial Econometrics literature. First, although continuous time methods and no-arbitrage arguments are prominent in the asset pricing literature, most influential contributions have been derived under very restrictive assumptions about the underlying process. For example, the Black-Scholes option valuation formula assumes constant volatility when, it is generally accepted empirically, that volatility evolves over time. However, recently, some authors have proposed more realistic continuous time processes with time varying volatilities; see, for example, Hull and White (1987), Heston (1993), Duffie and Kan (1996) and Dai and Singleton (2000). Engle (2001) suggests that the use of UHF data potentially could provide information on the more appropriate class of diffusion models to use for pricing both underlying and derivative assets.

On the other hand, the Financial Econometrics literature has many challenges to provide instruments adequate to represent the behavior of asset prices. The econometrics of, for example, jump diffusion or affine models are difficult. Bollerslev (2001) points out that recent research on the link between the probability distributions of actual asset prices and the corresponding risk-neutral probability distributions implied by derivative prices has just started and that much research remains to be done. Some relevant references in this sense are Aït-Sahaila and Lo (2000), Andersen *et al.* (2002), Chernov and Ghysels (2000) and Duffie *et al.* (2000). Also, it is very useful the guest editorial by Ghysels and Tauchen (2003) and all the papers within the special issue of the *Journal of Econometrics* on the intersection between Financial Econometrics and Financial Engineering.

2.4.1 Estimation of the CAPM

Two classical pricing models arise in the financial literature. *Capital Asset Pricing Model* (CAPM) is a set of predictions concerning equilibrium expected return on assets; see, for example, Sharpe (1964) or Lintner (1965). Classic CAPM assumes that all investors have the same one-period horizon, and asset returns have multivariate normal distributions. For a fixed time horizon, let R_i and R_M be the returns of asset i and of the market portfolio M, respectively. Classic CAPM, sometimes called Sharpe-Lintner CAPM, asserts that

$$E[R_i] = r + \beta_i \{E[R_M] - r\} \tag{4.1}$$

where r is the risk-free return and $\beta_i = \dfrac{cov(R_i, R_M)}{\sigma_M^2}$ is the *beta* of asset i.

Assuming that asset returns are normally distributed and the time horizon is one period (e.g., one year), a key concept in financial economics is the *market price of risk,* given by $\lambda_i = \dfrac{E[R_i] - r}{\sigma_i}$. In asset portfolio management, this is also called the *Sharpe Ratio,* after William Sharpe.

In terms of market price of risk, CAPM can be restated as follows:

$$\lambda_i = \frac{E[R_i] - r}{\sigma_i} = \frac{cov(R_i, R_M)}{\sigma_i \sigma_M} \cdot \frac{E[R_M] - r}{\sigma_M} = \rho_{i,M} \lambda_M, \tag{4.2}$$

where $\rho_{i,M}$ is the linear correlation coefficient between R_i and R_M. In other words, the market price of risk for asset i is directly proportional to the correlation coefficient between asset i and the market portfolio M.

CAPM automatically prices assets in the set of all linear combinations of basic assets according to this linearity rule, as long as the market portfolio used in the CAPM is the mean-variance efficient portfolio of risky assets (alternative termed the Markowitz portfolio). CAPM provides a powerful insight regarding

the risk-return relationship, where only systematic risk deserves an extra risk premium in an efficient market. However, CAPM and the concept of "market price of risk" were developed under the assumption of normal multivariate distributions for asset returns, and in practice the underwriting beta can be difficult to estimate.

On the other hand, a common practice pricing non-marketed assets is to infer the price applying the CAPM formula to this asset as well, by simply entering the random payoff B corresponding to the non-marketed asset into the CAPM formula. Technically, the new price has a systematic relationship to the prices of the basic assets, more precisely, it is the price of the marketed asset that best approximates the random payoff B in the sense of minimum expected squared error. Following geometric and statistical considerations, Luenberger (2002a) proposes a correlation pricing formula similar to the CAPM formula, which expresses the price of a non-marketed asset in terms of a priced asset that is the most correlated with the non-marketed asset, rather than in terms of the marked portfolio. The method has accuracy advantages when values in the formula must be estimated. Beyond the NA principle, Luenberger (2002b) derives a pricing method for non-marketed assets determining the price such that an investor with a specific utility function will elect to include the new asset in his/her portfolio at the zero level. The idea of zero-level pricing of a non-marketed payoff is to find the price such that a certain investor will elect to neither purchase nor short it. At this price the investor is indifferent to the inclusion of the considered payoff. Conditions ensuring for such a price to be unique are given in Luenberger (2002b).

Besides CAPM, another major financial pricing paradigm is *modern option pricing theory,* first developed by Black and Scholes (1973). Unfortunately, the Black-Scholes formula only applies to lognormal distributions of market returns. Options pricing is performed in a world of Q-measure, where the available data consists of observed market prices for related financial assets. On the other hand, actuarial pricing takes place in a world of P-measure, where the available data consists of projected losses, whose amounts and likelihood need to be converted to a "fair value" price; see Panjer (1998). Because of this difference in types of data available, modern option pricing is mostly concerned with the minimal cost of setting up a hedging portfolio, whereas actuarial pricing is based on actuarial present value of costs, with additional adjustments for correlation risk, parameter uncertainty and cost of capital. In these setting new research directions are proposed in the recent literature.

The statistical framework for estimation and testing for the classical CAPM is the Maximum Likelihood (ML) approach; see Campbell *et al.* (1997), Gibbons *et al.* (1989) and Bollerslev *et al.* (1988).

Inferences when there are deviations from the assumption that returns are jointly normal and iid through time have been developed. Tests which accom-

modate non-normality, heteroscedasticity, and temporal dependence returns are of interest for two reasons. First, while the normality assumption is sufficient, it is not necessary to derive the CAPM as a theoretical model. Rather, the normality assumption is adopted for statistical purposes. Without this assumption, the finite sample properties of asset pricing model tests are difficult to derive. Second, departures of monthly security returns from normality have been documented. As we have pointed out in this review, there is also abundant evidence of heteroscedasticity and temporal dependence in stock returns. It is therefore of interest to consider the effects of relaxing these statistical hypothesis. Robust tests of the CAPM can be constructed using a Generalized Method of Moments (GMM). Within the GMM framework, the distribution of returns conditional on the market return can be both serially dependent and conditionally heteroscedastic. The only assumption is that excess asset returns are stationary and ergodic with finite fourth moments. GMM procedure to estimate time-varying term premia and a consumption based asset pricing model are used in Hansen and Singleton (1982) and Hansen and Scheikman (1995).

Other lines of research are also of interest. One important topic is the extension of the framework to test conditional versions of the CAPM, in which the model holds conditional on state variables that describe the state of the economy. Econometric methods from section 3 are suitable for testing the conditional CAPM.

Another important subject is Bayesian analysis of mean-variance efficiency and the CAPM. Bayesian analysis allows the introduction of prior information. Harvey and Zhou (1990) and Kandel *et al.* (1995) are examples of work with this perspective.

There is a controversy about the statistical evidence against the CAPM in the past 30 years. Some authors argue that the CAPM should be replaced by multifactor models with several sources of risk; others argue that the evidence against the CAPM is overstated because of mismeasurement of the market portfolio, improper neglect of conditional information, data snooping, or sample-selection bias; and yet others claim that no risk-based model can explain the anomalies of stock-market behavior. Campbell *et al.* (1997) explore multifactor asset pricing models.

2.4.2   Estimation of the term structure

There is a vast literature devoted to the estimation of dynamic models of the term structure that describe the evolution of yields at all maturities. One of the main problems in this area is that the theoretical models need to be complex enough as to represent adequately the empirical complexity often observed. However, as the complexity of the models increases, their estimation becomes more difficult.

Models of the term structure focus mainly on affine models, characterized originally by Duffie and Kan (1996), that assume that the market price of risk is a multiple of the interest rate volatility and that the state variables are independent. Under these assumptions, ML estimation of the parameters is feasible. However, many empirical studies have shown that this model has fundamental limitations; see, for example, Ghysels and Ng (1998) and Dai and Singleton (2000) between many others. To overcome these limitations, Dai and Singleton (2000) propose the multivariate affine term structure models while Ahn *et al.* (2002) propose the quadratic term structure models. However, neither of these models is able to track adequately the dynamic evolution of volatility. Recently, Ahn *et al.* (2003) investigates whether an hybrid model between affine, quadratic and nonlinear models is able to outperform each of the individual models. However, they conclude that, in general, this is not the case. Dai and Singleton (2003) is an excellent review on models of the term structure described from the point of view of their empirical implementation. They focus on the fit of the theoretical specifications of dynamic structure models to the historical shapes of the yield curves.

On the other hand, as we mentioned before, the estimation of these more complex models becomes difficult as the likelihood does not have, in general, a close form. One of the most popular methods in this context is the Efficient Method of Moments (EMM) of Gallant and Tauchen (1996). Duffee and Stanton (2003) estimate a multifactor term structure model with correlated factors, nonlinear dynamics and flexible price of interest rate risk, using both the EMM and an approximate Kalman filter. They conclude that the best results are obtained when the latter procedure is used to estimate the model although it is not asymptotically optimal. However, their results reveal severe biases in the parameter estimates regardless of the estimation method; see also Duan and Simonato (1999) and Chen and Scott (2002) for other authors that have also used the Kalman filter to estimate the term structure.

2.4.3   Estimation of the VaR

Regulators and risk managers are interested in obtaining measures of the Value at Risk (VaR), defined as the expected loss of a portfolio after a given period of time (usually 10 days) corresponding to the $\alpha\%$ quantile (usually 1%). This interest has motivate new methods designed to estimate the tails of the distribution of returns. There are several methods to estimate the VaR. The early VaR parametric models impose a known theoretical distribution to price changes. Usually it is assumed that the density function of risk factors influencing asset returns is a multivariate normal distribution as, for example, in J.P. Morgan (1996). The most popular parametric methods are variance-covariance models and Monte Carlo simulation. However, excess kurtosis of

these factors will cause losses greater than VaR to occur more frequently and be more extreme than those predicted by the Gaussian distribution. Consequently, several authors propose to use nonparametric (historical simulation) and semiparametric models that avoid to assume a particular distribution of price increments although they usually assume independent increments; see, for example, Danielsson and de Vries (1998). Finally, some authors propose to use extreme value theory estimation of tail shapes to estimate the VaR; see, for example, Embrechts *et al.* (1997) and McNeil and Frey (2000). In relation with these methods, Pearson and Smithson (2002) describe refinements which increase computational speed and improve accuracy.

However, as described in previous sections, financial returns are often characterized by volatility clustering and non-Gaussianity. Therefore, several authors have considered extensions of the previous approaches that allow for time-varying volatilities. The most popular approach is to estimate the VaR based on Conditional Gaussian GARCH models; see, for example, Christoffersen and Diebold (2000) and Christoffersen *et al.* (2001). Guermat and Harris (2002) even extend further the GARCH approach to allow for kurtosis clustering.

Recently, Engle and Manganelli (1999) have proposed a conditional quantile estimation based on the CaViar model given by

$$VaR_t = \beta_0 + \beta_1 VaR_{t-1} + \beta_2 |y_{t-1}| \qquad (4.3)$$

Gourieroux and Jasiak (2001) describe several alternative methods to estimate the VaR, focusing on their main advantages and limitations. Tsay (2002) also describe several of these methods and compare their performance to estimate the VaR of daily returns of IBM stocks. In particular, he compares the RiskMetrics methodology developed by J.P. Morgan, GARCH models, nonparametric estimation, quantile regression and extreme value, finding substantial differences among the approaches.

Given that, as we have mentioned already, the distribution of high frequency price increments is non-Gaussian, and even in many cases the conditional distribution of GARCH models is not Gaussian, many authors suggest using bootstrap techniques to avoid particular assumptions on the distribution of factors beyond stationarity of the distribution of returns; see, for example, Barone-Adessi *et al.* (1999), Barone-Adessi and Giannopoulos (2001) and Vlaar (2000). Ruiz and Pascual (2002) review the use of bootstrap methods to estimate the VaR.

Although there is a huge number of papers devoted to analyze methods to estimate the VaR as a measure of financial risk, this measure is not without criticisms; see, for example, Szego (2002) and the papers contained in the especial number of the *Journal of Banking and Finance, 26*. There are several

new measures of risk proposed as remedy for the deficiencies of VaR as, for example, Conditional VaR (CVaR) and Expected Shortfall.

2.4.4   Estimation of diffusion processes

There are two relatively independent lines in financial modeling: continuous–time models typically used in theoretical finance and discrete-time models favored for empirical work. The continuous–time models are dominated by the diffusion approach. In contrast to stochastic differential equations used in discrete-time models, stochastic differential equations are widely used to describe continuous–time models in the theoretical finance literature. The stochastic processes characterized by the stochastic differential equations are Itô processes, and continuous–time model assumes that a security price S_t follows the stochastic differential equation:

$$dS_t = \mu_t S_t dt + \sigma_t S_t dW_t \ \ t \in [0, T] \tag{4.4}$$

where W_t is a standard Wiener process, μ_t is called diffusion drift in probability or instantaneous mean rate of return in finance and σ_t^2 is called diffusion variance in probability or instantaneous conditional variance (or volatility). The celebrated Black-Scholes model corresponds to (16) with constants μ_t and σ_t. Given that financial time series tend to be highly heteroscedastic, the general modelization assumes that σ_t^2 is random and itself is governed by another stochastic differential equation.

For continuous–time models, the "no arbitrage" condition, as we have extensively developed in section 2, can be characterized by a martingale measure, that is, a probability law under which S_t is a martingale. Prices of options and derivatives are then the conditional expectation of certain functionals of S under this measure. The calculations and derivations can be manipulated by tools as the Itô lemma and Girsanov theorem; see Karatzas and Shreve (1991) or the overviews in Dixit (1993) and Merton (1990).

The log price process $X_t = \log(S_t)$ after the Itô lemma and from (16) follows the diffusion model

$$dX_t = (\mu_t + \sigma_t^2/2)dt + \sigma_t dW_t, \tag{4.5}$$

where the drift for X_t has a term σ_t^2. GARCH models are used to represent statistically the increments of the log price process, so from the diffusion point of view, (17) is also a natural parametrization of the GARCH drift μ_k.

While the models are written in continuous–time, the available data are mostly sampled discretely in time. Ignoring this difference can result in inconsistent estimators (see, e.g., Merton (1980)). A number of statistical/econometric methods have been recently developed to estimate the parameters of a

continuous–time diffusion without requiring that a continuous record of observations be available.

The methods of moments together with simulation estimations have been used by Gouriéroux *et al.* (1993) and Gallant and Tauchen (1996). A forceful criticism of simulation-based method-of-moments estimation has been that this method does not provide a representation of the observables in terms of their own past as do maximum likelihood based on a conditional density and time series methods such as ARIMA, ARCH and GARCH modeling; see Jacquier *et al.* (1994). Gallant and Tauchen (1998) use the notion of reprojection to let a representation of the observed process in terms of observables that incorporates the dynamics implied by the possibly nonlinear system under consideration. They propose a methodology for estimation and diagnostic assessment of several diffusion models of the short rate expressed as a partially observed system of stochastic differential equations. The theoretical support of the projection method was provided by Gallant and Long (1997) who showed that it achieves the same efficiency as ML.

Nonparametric density-matching methods have been applied in Aït-Sahalia (1996a, 1996b). Discretely observed diffusions have also been fit by estimating functions; see Kessler and Sørensen (1999) and Kessler (2000). A Monte Carlo Markov Chain (MCMC) based method is proposed in Eraker (2001). The method is applied to the estimation of parameters in one-factor interest-rate models and a two-factor model with a latent stochastic volatility component.

Elerian *et al.* (2001) propose a new method for dealing with the estimation problem of stochastic differential equations that is likelihood based, can handle nonstationarity, and is not dependent on finding an appropriate auxiliary model. As they point out, their idea is simply to treat the values of the diffusion between any two discrete measurements as missing data and then to apply tuned MCMC methods based on the Metropolis-Hasting algorithm to learn about the missing data and the parameters.

As in most contexts, provided one trusts the parametric specification in the diffusion, ML is the method of choice. The major caveat in the present context is that the likelihood function for discrete observations generated by the parametric stochastic differential equation cannot be determined explicitly for most models. Since the transition density is generally unknown, one is forced to approximate it. The simulation-based approach suggested by Pedersen (1995), has great theoretical appeal but its implementation is computationally costly. Durham and Gallant (2002) examine a variety of numerical techniques designed to improve the performance of this approach.

If sampling of the process were continuous, the situation would be simpler. First, the likelihood function for a continuous record can be obtained by means of a classical absolutely continuous change of measure. Second, when the sam-

pling interval goes to zero, expansions of the transition function "in small time" are available in the statistical literature and some calculate expressions for the transition function in terms of functionals of a Brownian Bridge. Available methods to compute the likelihood function in the case of discrete-time sampling, involve either solving numerically the Fokker-Plank-Kolmogorov partial differential equation (see Lo (1988)) or simulating a large number of sample paths along with the process is sampled very finely (see Pedersen (1995)). Neither methods produces a closed-form expression to be maximized over the parameter: the criterion function takes either the form of an implicit solution to a partial differential equation, that could be approximated by a sum over the outcome of the simulations. Using Hermite polynomials, Aït-Sahalia (2002) provides an explicit sequence of closed-form functions. It is shown that it converges to the true (but unknown) likelihood function. It is also documented that maximizing the sequence results in an estimator that converges to the true ML estimator and shares its asymptotic properties.

As we have pointed out in section 3, high-frequency financial data are not only discretely sampled in time but the time separating successive observations is often random. Aït-Sahalia and Mykland (2003) analyzes the consequences of this dual feature of the data when estimating a continuous–time model. More precisely, they measure the additional effect of the randomness of the sampling intervals over and beyond those due to the discreteness of the data. They also examine the effect of simply ignoring the sampling randomness and find that in many situations the randomness of the sampling has larger impact than the discreteness of the data.

As we have described previously, continuous–time models, dominated by the diffusion approach, are typically favored in the theoretical finance while discrete-time models, mainly of the ARCH type, are the focus of empirical research. Nelson (1990) tried for the first time to reconcile both approaches, showing that GARCH processes weakly converge to some bivariate diffusions as the length of the discrete time interval goes to zero. Later, Duan (1997) proposed an augmented GARCH model and derived its diffusion limit. These authors link the two types of models by weak convergence. Consequently, it is rather common to apply the statistical inferences derived under the GARCH model to its diffusion limit. However, recently Wang (2002), using the Le Cam's deficiency distance, shows that the GARCH model and its diffusion limit are asymptotically equivalent only under deterministic volatility. He concludes that, for modelling stochastic volatility, if a diffusion model is preferred, it is statistically more efficient to fit data directly to the diffusion model and carry out the inference.

2.5   Conclusions

Throughout the paper we have summarized several applications of probabilistic and time series models in finance. We have specially focused on those pricing models reflecting the absence of arbitrage and free-lunch. Almost all of them are characterized by the existence of equivalent martingale probability measures (or risk-neutral measures). Thus the martingale property permits to price, hedge, speculate or compose efficient portfolios since future prices must verify the random walk assumption.

However, there are still many open problems that will merit future research. So, the absence of arbitrage (free-lunch) does not always lead to martingales, even it one focuses on perfect markets. When dealing with incomplete markets there are infinitely many risk-neutral measures and it is necessary to establish coherent criteria in order to choose the adequate one. For imperfect markets we will never have a unique risk-neutral measure and it is also necessary to find appropriate instruments in order to relate risk-neutral measures and hedging or efficient strategies.

Most of the concrete pricing models applied in practice are characterized by stochastic differential equations reflecting the market dynamic behavior. By manipulating the stochastic equation it is possible to obtain the partial differential equation or the risk-neutral measure leading to pricing or hedging rules, as well as, to those usual topics of asset pricing theory. Time Series and Econometric Models are the key when designing these pricing models and calibrating or evaluating its empirical possibilities. Furthermore, the growing complexity of real markets, characterized by more and more connections amongst them all, higher and higher volatilities, more and more complex risks and securities, and a increasing number of investors, make it rather necessary to improve those models usually applied when dealing with pricing issues or interest-rate linked topics.

Summarizing, probabilistic and time series approaches play a crucial role in finance, and it is emphasized if one focuses on arbitrage pricing theory. Moreover, the level of development of current markets makes it essential to improve and enlarge our knowledge about all the involved fields, from theoretical foundations to empirical applications.

References

Aguilar, O. and M. West (2000), Bayesian dynamic factor models and variance matrix discounting for portfolio allocation, *Journal of Business and Economic Statistics,* 18, 338-357.

Ahn, D.-H., R.F. Dittmar and A.R. Gallant (2002), Quadratic term structure models: theory and evidence, *The Review of Financial Studies,* 15, 243-288.

Ahn, D.-A., R.F. Dittmar, A.R. Gallant and B. Gao (2003), Purebred of hybrid?: Reproducing the volatility in term structure dynamics, *Journal of Econometrics,* forthcoming.

Aït-Sahalia, Y. (1996a), Nonparametric Pricing of Interest Rate Derivative Securities, *Econometrica,* 64, 527-560.

Aït-Sahalia, Y. (1996b), Testing Continuous-Time Models of the Spot Interest Rate, *Review of Financial Studies,* 9, 385-426.

Aït-Sahalia, Y. (2002), Maximum likelihood estimation of discretely sample diffusions: a closed form approach, *Econometrica,* 70, 223-262.

Aït-Sahalia, Y. and P.A. Mykland (2003), The effect of random and discrete sampling when estimating continuous-time diffusions, *Econometrica,* 71, 483-549.

Aït-Sahaila, Y. and A.W. Lo (2000), Nonparametric risk management and implied risk aversion, *Journal of Econometrics,* 94, 9-51.

Andersen, T.G. (1996), Return volatility and trading volume: an information flow interpretation of stochastic volatility, *Journal of Finance,* 51, 169-204.

Andersen, T.G. and T. Bollerslev (1997a), Intraday periodicity and volatility persistence in financial markets, *Journal of Empirical Finance,* 4, 115-158.

Andersen, T.G. and T. Bollerslev (1997b), Heterogenous information arrivals and return volatility dynamics: uncovering the long-run in high frequency returns, *Journal of Finance,* 52, 975-1005.

Andersen, T.G. and T. Bollerslev (1998), Deutsche mark-dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer-run dependencies, *Journal of Finance,* 53, 219-265.

Andersen, T.G. and J. Lund (1997), Estimating continuous-time stochastic volatility models of the short-term interest rate, *Journal of Econometrics,* 77, 343-377.

Andersen, T.G., L. Benzoni and J. Lund (2002), An empirical investigation of continuous-time equity return models, *Journal of Finance,* 57, 1239-1284.

Andersen, T.G., T. Bollerslev and A. Das (2001), Variance-ratio statistics and high-frequency data: testing for changes in intraday volatility patterns, *Journal of Finance,* 56, 305-327.

Back, K. and S.R. Pliska (1991), On the fundamental theorem of asset pricing with infinite state space*, Journal of Mathematical Economics,* 20, 1-18.

Baillie, R.T., T. Bollerslev and H.O. Mikkelsen (1996), Fractionally integrated generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics,* 74, 3-30.

Balbás. A., M.A. Mirás and M.J. Muñoz-Bouzo (2002), Projective system approach to the martingale characterization of the absence of arbitrage, *Journal of Mathematical Economics,* 37, 311-323.

Ball, C. and W. Torous (1985), On jumps in common stock prices and their impact on call option pricing, *Journal of Finance,* 40, 155-173.

Barberis, N., M. Huang and T. Santos (2001), Prospect theory and asset prices, *The Quarterly Journal of Economics,* 116, 1-53.

Barndorff-Nielsen, O.E. and N.G. Shephard (2001), Non-Gaussian OU processes and some of their uses in financial economics (with discussion), *Journal of the Royal Statistical Society,* series B, 63,167-241.

Barndorff-Nielsen, O.E. and N.G. Shephard (2002a), Econometric analysis of realized volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society,* series B, 64, 253-280.

Barndorff-Nielsen, O.E. and N.G. Shephard (2002b), Lèvy based dynamic models for financial economics, Unpublished book manuscript.

Barone-Adessi, G., K. Giannopoulos and L. Vosper (1999), VaR without correlations for nonlinear portfolios, *Journal of Future Markets,* 19, 583-602.

Barone-Adessi, G. and K. Giannopoulos (2001), Non-parametric VaR techniques. Myths and realities. *Economic Notes,* 30, 167-181.

Bates, D.S. (1996), Jumps and stochastic volatility: exchange rate processes implicit in Deutsche mark options, *The Review of Financial Studies,* 9, 69-107.

Bekaert, G., R. Hodrick and D. Marshall (2001), Peso Problem explanations for term structure anomalies, *Journal of Monetary Economics,* 48, 241-270.

Benartzi, S. and R. Thaler (1995), Myopic loss aversion and equity premium puzzle, *Quarterly Journal of Economics,* 110, 73-92.

Black, F. (1986), Noise, *Journal of Finance,* 3, 529-543.

Black, F. and M. Scholes (1973), The pricing of options and corporate liabilities, *Journal of Political Economy,* 81, 637-650.

Bollerslev, T. (1986), Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics,* 31, 307-327.

Bollerslev, T. (1990), Modelling the coherence in short-run nominal exchange rates: a Multivariate generalized ARCH approach, *Review of Economics and Statistics,* 72,498-505.

Bollerslev, T. (2001), Financial econometrics: Past developments and future challenges, *Journal of Econometrics,* 100, 41-51.

Bollerslev, T., R. F. Engle and J. Wooldridge (1988), A capital asset pricing model with time varying covariances, *Journal of Political Economy,* 96,116-131.

Bollerslev, T., R.F. Engle and D.B. Nelson (1995), ARCH models, in R.F. Engle and D. McFadden (eds.), *The Handbook of Econometrics,* vol. 4, North-Holland, Amsterdam.

Brannath, W. (1997), No arbitrage and martingale measures in option pricing, Dissertation zur Erlangung des akademischen Grades. Univertsität Wien.

Breidt, F.J., N. Crato and P.J.F. de Lima (1998), The detection and estimation of long memory in stochastic volatility, *Journal of Econometrics,* 83, 325-348.

Broto, C. and E. Ruiz (2002), Estimation methods for stochastic volatility: A survey, Working Paper.

Butsic, R.P. (1999), Capital allocation for property-liability insures: a catastrophe reinsurance application, Casualty Actuarial Society Forum, Spring 1999.

Campbell, J.Y., A.W. Lo and A.C. MacKinlay (1997), *The Econometrics of Financial Markets,* Princeton University Press, Princeton, New Jersey.

Carnero, M.A., D. Peña and E. Ruiz (2001a), Outliers and conditional autoregressive heteroscedasticity in time series, *Estadística,* 53, 143-213.

Carnero, M.A., D. Peña and E. Ruiz (2001b), Is stochastic volatility more flexible than GARCH?, Working Paper 01 -08(05), Serie Estadística y Econometría, Universidad Carlos III de Madrid.

Cvitanic, J. and I. Karatzas (1996), Hedging and portfolio optimization under transaction costs: a martingale approach, *Math. Fin.* 6, 133-166.

Chateaunef, A., R. Kast and A. Lapied (1994), Market preferences revealed by prices: non-linear pricing in slack markets, in Machina, M. and B. Munier (eds), *Models and experiments in risk and rationality,* Kluwer, Dordrecht.

Chateaunef, A., R. Kast and A. Lapied (1996), Choquet pricing for financial markets with frictions, *Mathematical Finance,* 6, 323-330.

Chen, R.-R. and L. Scott (2002), Multi-factor Cox-Ingersoll-Ross models of the term structure: Estimates and tests from a Kalman filter, *Journal of Real State Finance and Economics,* forthcoming.

Chernov, M. and E. Ghysels (2000), A study towards a unified approach to the joint estimation of objective risk neutral measures for the purpose of option valuation, *Journal of Financial Economics,* 57, 407-458.

Chernov, M., A.R. Gallant, E. Ghysels and G. Tauchen (2000), A new class of stochastic volatility models with jumps. Theory and estimation, Working paper, Columbia University.

Chib, S., F. Nardari and N. Shephard (2002), Markov chain Monte Carlo methods for stochastic volatility models, *Journal of Econometrics,* 108, 281-316.

Christoffersen, P.F. and F. Diebold (2000), How relevant is volatility forecasting for financial risk management?, *Review of Economics and Statistics,* 82, 12-22.

Christoffersen, P.F., J. Hahn and A. Inuoe (2001), Testing and comparing Value-at-Risk measures, *Journal of Empirical Finance,* 8, 325-342.

Dai, Q. and K. Singleton (2000), Specification analysis of affine term structure models, *Journal of Finance,* 55, 1943-1978.

Dai, Q. and K. Singleton (2003), Term structure dynamics in theory and reality, *Review of Financial Studies,* forthcoming.

Dalang, R.C., A. Morton, W. Willinger (1989), Equivalent martingale measure and no arbitrage in stochastic securities market model, Stochastics and Stochastic Rep. 29, 185-202.

Danielsson, J. and G. de Vries (1998), Beyond the sample: Extreme quantile and probability estimations, Discussion paper 298, London School of Economics, London.

Davis, M.H.A. and A. Norman (1990), Portfolio selection with transaction costs, *Math. Operation Research,* 15, 676-713.

De Waegenaere A.M.B., R. Kast and A. Lapied (1996), Non-linear Asset Valuation on Markets with Frictions, CentER Discussion Paper 96112.

Delbaen, F. (1992), Representing martingale measures when asset prices are continuous and bounded, *Math. Fin. 2,* 107-130.

Delbaen, F. and W. Schachermayer (1994), A general version of the Fundamental Theorem of Asset Pricing, *Math. Annalen* 300 ,463-520.

Delbaen, F. and W. Schachermayer (1998), The fundamental theorem of asset pricing for unbounded stochastic process, *Mathematische Annalen,* 312, 215-250.

Delbaen, F., Y. Kabanov and E. Valkeila (1998), Hedging under transaction costs in currency markets: a discrete-time model. Preprint.

Diebold, F.X. and M. Nerlove (1989), The dynamics of exchange rate volatility: a multivariate latent factor ARCH models, *Journal of Applied Econometrics, 4,* 1-21.

Dixit, A. (1993), *The Art of Smooth Pasting,* Harwood, Switzerland.

Duan, J.C. (1997), Augmented GARCH(p.q) process and its diffusion limit, *Journal of Econometrics,* 79, 97-127.

Duan, J.-C. and J.-G. Simonato (1999), Estimating and testing exponential-affine term structure models by Kalman filter, *Review of Quantitative Finance and Accounting,* 13, 111-135.

Duffee, G.R. and R.H. Stanton (2003), Estimation of dynamic term structure models, mimeo, U.C. Berkeley.

Duffie, D. and C.F. Huang (1986), Multiperiod security markets with differential information; martingales and resolution times, *J. Math. Econom.,* 15, 283-303.

Duffie, D. and R. Kan (1996), A yield-factor model of interest rates, *Mathematical Finance,* 6, 379-406.

Duffie, D., J. Pan and K.J. Singleton (2000), Transform analysis and asset pricing for affine jump-diffusions, *Econometrica,* 68, 1343-1376.

Dufour, A. and R. Engle (2000), Time and the price impact of a trade, *Journal of Finance,* 55, 2467-2498.

Durham, G.B. and A.R. Gallant (2002), Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes, *Journal of Business and Economic Statistics,* 20, 297-316.

Eberlein, E. (2001), Application of generalized hyperbolic Lévy motions to finance, in Barndorff-Nilesen, O.E., T. Mikosch and S. Resnick (eds.), *Lévy Processes-Theory and Applications,* 319-337, Birkhauser, Boston.

Eberlein, E. and K. Prause (2002), The Generalized hyperbolic model: Financial derivatives and risk measures, in *Mathematical Finance-Bachelier Congress 2000,* Springer Verlag, forthcoming.

Elerian, O., S. Chib and N.G. Shephard (2001), Likelihood inference for discretely observed nonlinear diffusions, *Econometrica,* 69, 959-993.

Embrechts, P., C. Kluppelberg and T. Mikosch (1997), *Modelling Extremal Events,* Springer, Berlin.

Engle, R. (1982), Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica,* 50, 987-1007.

Engle, R. (2000), The econometrics of ultra high frequency data, *Econometrica,* 68,1-22.

Engle, R. (2001), Financial econometrics - A new discipline with new methods, *Journal of Econometrics,* 100, 53-56.

Engle, R. (2002a), New frontiers for ARCH models, forthcoming in *Journal of Applied Econometrics*

Engle, R. (2002b), Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models, *Journal of Business and Economic Statistics,* 20, 339-350.

Engle, R. and G.G.J. Lee (1999), A long run and short run component model of stock return volatility, in Engle, H. and H. White (eds), *Cointegration, Causality and Forecasting,* Oxford University Press, Oxford.

Engle, R. and S. Manganelli (1999), CaViar: Conditional Autoregressive Value at Risk by regression quantiles. Discussion paper, University of California, San Diego.

Engle, R.F. and J. R. Russell (1998), Autoregressive conditional duration: A new model for irregularly spaced transaction data, *Econometrica,* 66, 1127-1162.

Engle, R., D. Lilien and R. Robins (1987), Estimating time-varying risk premia in the term structure: the ARCH-M model, *Econometrica,* 55, 391-407.

Epstein, L.G. and S.E. Zin (2001), The independence axiom and asset returns, *Journal of Empirical Finance,* 8, 537-572.

Eraker, B. (2001), MCMC Analysis of Diffusion Models with Applications to Finance, *Journal of Business and Economic Statistics,* 19, 177-191.

Eraker, B., M. Johannes and N. Polson (2003), The impact of jumps in volatility and returns, *Journal of Finance,* forthcoming.

Gallant, A.R. and J.R. Long (1997), Estimating Stochastic Differential Equations Efficiently by Minimum Chi-Squared, *Biometrika,* 84, 125-141.

Gallant, A.R. and G. Tauchen (1996), Which moments to match?, *Econometric Theory,* 12, 657-681.

Gallant, A.R. and G. Tauchen (1998), Reprojecting partially observed systems with application to interest rate diffusion, *Journal of the American Statistical Association,* 93,10-24.

Ghysels, E. and S. Ng (1998), A semiparametric factor model of interest rates and tests of the affine term structure, *Review of Economics and Statistics,* 80, 535-548.

Ghysels, E. and G. Tauchen (2003), Frontiers of financial econometrics and financial engineering, *Journal of Econometrics,* forthcoming.

Ghysels, E., A.C. Harvey and E. Reanult (1996), Stochastic Volatility, in J.Knight and S. Satchell (eds.), *Handbook of Statistics,* 14, 119-191, North-Holland, Amsterdam.

Gibbons, M., S. Ross and J. Shanken (1989), A test of the efficiency of a given portfolio, *Econometrica,* 57, 1121-1152.

Gouriéroux, C.A., A. Monfort and E. Renault (1993), Indirect Inference, *Journal of Applied Econometrics,* 8, S85-S118.

Gourieroux, C.A. and J. Jasiak (2001), *Financial Econometrics,* Princeton University Press, Princeton.

Granger, C.W.J., Z. Ding and S. Spear (2000), Stylized facts on the temporal and distributional properties of absolute returns: An update, Working paper, University of California, San Diego.

Guermat, C. and R.D.F. Harris (2002), Forecasting value at risk allowing for time variation in the variance and kurtosis of portfolio returns, *International Journal of Forecasting,* 18, 409-419.

Hansen, L. and K. Singleton (1982), Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations models, *Econometrica,* 50, 1269-1288.

Hansen, L. and J. Scheikman (1995), Back to the future: Generating moment implications for continuous-time Markov processes, *Econometrica,* 63, 767-804.

Harrison, M. and D. Kreps (1979), Martingales and arbitrage in multiperiod security markets, *Journal of Economic Theory,* 20, 381-408.

Harrison, M. and S. Pliska (1981), Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes Appl.,* 11, 215-260.

Harvey, A.C. (1998), Long memory in stochastic volatility, in J. Knight and S. Satchell (eds.), *Forecasting Volatility in Financial Markets,* Butterworth-Haineman, Oxford.

Harvey, A.C. and N.G. Shephard (1996), Estimation of an asymmetric stochastic volatility model for asset returns, *Journal of Business and Economic Statistics,* 14, 429-434.

Harvey, A.C., E. Ruiz and N.G. Shephard (1994), Multivariate stochastic variance models, *Review of Economic Studies,* 247-264.

Harvey, C.R. (2001), The specification of conditional expectations, *Journal of Empirical Finance,* 8, 573-637.

Harvey, C.R. and G. Zhou (1990), Bayesian inference in asset pricing tests, *Journal of Financial Economics,* 26, 221-254.

Hausman, J., A. Lo and C. MacKinlay (1992), An ordered probit analysis of transaction stock prices, *Journal of Financial Economics,* 31, 319-379.

He, C. and T. Terasvirta (1999), Properties of moments of a family of GARCH processes, *Journal of Econometrics,* 92, 173-192.

Hentschel, L. (1995), All in the family: Nesting symmetric and asymmetric GARCH models, *Journal of Financial Economics,* 39, 71-104.

Heston, S. (1993), A closed-form solution for options with stochastic volatility with applications to bond currency options, *Review of Financial Studies, 6,* 327-343.

Hong, H. and J.C. Stein (1999), A unified theory of underreaction, momentum trading and overreaction in asset markets, *Journal of Finance,* 54, 2143-2184.

Hull, J. and A. White (1987), The pricing of options on assets with stochastic volatilities, *Journal of Finance,* 42, 281-300.

Jacquier, E., N.G. Polson and P.E. Rossi (1994), Bayesian analysis of stochastic volatility models, *Journal of Business and Economic Statistics,* 12, 371-417.

Jacquier, E., N.G. Polson and P.E. Rossi (1995), Models and prior distributions for multivariate stochastic volatility, Technical report, Graduate School of Business, University of Chicago.

Jouini, E. and H. Kallal (1995a), Martingales and arbitrage in securities markets with transaction costs, *Journal of Economic Theory,* 66, 178-197.

Jouini, E. and H. Kallal (1995b), Arbitrage in securities markets with short-sales constraints, *Math Fin.,* 5,197-232.

Jouini, E. and H. Kallal (1999), Viability and Equilibrium in Securities Market with Frictions, *Math Fin., 9,* 275-292.

Jouini, E. and C. Napp (2002), Arbitrage and Investment Opportunities, to appear in *Finance and Stochastics.*

J.P. Morgan (1996), *Risk Metrics Technical Document,* 4th edition, J.P. Morgan, New York.

Kabanov, Y. (1999), Hedging and liquidation under transaction costs in currency markets, *Finance and Stochastics,* 3, 237-248.

Kabanov, Y. (2001), Arbitrage theory. Handbooks in Mathematical Finance, *Option Pricing: Theory and Practice,* 3-42.

Kabanov, Y. and D. Kramkov (1994), No arbitrage and equivalent martingale measures: An elementary proof of the Harrison-Pliska theorem, *Theory Prob. Appl.,* 39

Kabanov, Y., M. Rasonyi and C. Striker (2001), No-arbitrage criteria for financial markets with efficient friction, *Finance and Stochastics,* 6, 371-382.

Kandel, S., R. McCulloch and R. Stambaugh (1995), Bayesian inference and portfolio efficiency, *Review of Financial Studies,* 8, 1-53.

Karatzas, I. and S.E. Shreve (1991), *Brownian Motion and Stochastic Calculus,* 2nd ed., Springer, New York.

Kessler, M. (2000), Estimation of an ergodic diffusion from discrete observations, *Scandinavian Journal of Statistics,* 24

Kessler, M. and M. Sørensen (1999), Estimating equations based on eigenfunctions for discretely observed diffusion process, *Bernoulli,* 5, 299-314.

Kim, S., N.G. Shephard and S. Chib (1998), Stochastic volatility: likelihood inference and comparison with ARCH models, *Review of Economic Studies,* 65, 361-393.

King, M., E. Sentana and S. Wadhwany (1994), Volatility and links between national stock markets, *Econometrica,* 62, 901-933.

Koopman, S.J. and E.H. Uspensky (2002), The stochastic volatility in mean model: Empirical evidence from international stock markets, *Journal of Applied Econometrics,*

Kreps, D. (1981), Arbitrage and equilibrium in economies with infinitely many commodities, *J. Math. Econ.* 8, 15-35.

Ledoit, O., P. Santa-Clara and M. Wolf (2003a), Flexible multivariate GARCH modeling with an application to international stock markets, *Review of Economics and Statistics,* forthcoming.

Lintner, J. (1965), The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics,* 47, 13-37.

Lo, A. W. (1988), Maximum likelihood estimation of generalized Itô processes with discretely sampled data, *Econometric Theory,* 4, 231-247.

Lo, A. and A.C. MacKinlay (1990), An econometric analysis of nonsynchronous trading, *Journal of Econometrics,* 45, 181-212.

Lobato, I.N. and N.E. Savin (1998), Real and spurious long-memory properties of stock-market data, with discussion, *Journal of Business and Economic Statistics,* 16, 261-283.

Lobato, I., J.C. Nanverkis and N.E. Savin (2001), Testing for autocorrelation using a modified Box-Pierce Q test, *International Economic Review,* 42, 187-205.

Luenberger, D.G. (2002a), A correlation pricing formula, *Journal of Economics Dynamics & Control,* 26, 1113-1126.

Luenberger, D.G. (2002b), Arbitrage and universal pricing, *Journal of Economics Dynamics & Control,* 26, 1613-1628.

McNeil, A. and R. Frey (2000), Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach, *Journal of Empirical Finance*

Merton, R. C.(1990), *Continuous-time Finance,* Blackwell, Cambridge.

Merton, R.C. (1976), Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics,* 3, 125-144.

Merton, R.C. (1980), On estimating the expected return on the market: An exploratory investigation, *Journal of Financial Economics,* 8, 323-361.

Müller, U.A., M.M. Dacorogna, R.D. Davé, R.B. Olsen, O.V. Pictet and J.E. von Weizsächer (1997), Volatilities at different time resolutions-analysing the dynamics of market components, *Journal of Empirical Finance,* 4, 213-239.

Nelson, D.B. (1990), ARCH models as diffusion approximations, *Journal of Econometrics,* 45, 7-38.

Nelson, D.B. (1991), Conditional Heteroskedasticity in asset returns: A new approach, *Econometrica,* 59, 347-370.

Nelson, D.B. (1992), Filtering and forecasting with misspecified ARCH models I: Getting the right variance with the wrong model, *Journal of Econometrics,* 52, 61-90.

Nelson, D.B. (1994), Asymptotically optimal smoothing with ARCH models, *Econometrica,* 63,

Nelson, D.B. and D.P. Foster (1994), Asymptotic filtering theory for univariate ARCH models, *Econometrica,* 62, 1-41.

Panjer, H.H. (editor) (1998), *Financial Economics,* The Actuarial Foundation, Schaumburg, IL.

Pearson, N.D. and C. Smithson (2002), VaR. The state of play, *Review of Financial Economics,* 11,175-189.

Pedersen, A.R., (1995), A new approach to maximum-likelihood estimation for stochastic differential equations based on discrete observations, *Scandinavian Journal of Statistics,* 22, 55-71.

Pham H. and N. Touzi (1999), The fundamental theorem of asset pricing with cone constraints, *Journal of Mathematical Economics,* 31, 265-279.

Pitt, M. and N. Shephard (1999), Time varying covariances: a factor stochastic volatility approach (with discussion), in. Bernardo, J., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.), *Bayesian Statistics,* 6, 547-570, Oxford University Press, Oxford.

Ray, B.K. and R.S. Tsay (2000), Long-range dependence in daily stock volatilities, *Journal of Business and Economic Statistics,* 18, 254-262.

Romano, J.L. and L.A. Thombs (1996), Inference for autocorrelations under weak assumptions, *Journal of American Statistical Association,* 91, 590-600.

Rydberg and N.G. Shephard (2003), Dynamics of trade-by-trade price movements: decomposition and models, *Journal of Financial Econometrics,* forthcoming.

Ruiz, E. and L. Pascual (2002), Bootstrapping financial time series, *Journal of Economic Surveys,* 16, 271-300.

Schachermayer, W. (1992), A Hilbert space proof of the fundamental theorem of asset pricing in finite discrete time, *Insurance: Mathematics and Economics,* 11,4, 249-257.

Schachermayer, W. (1994), Martingale measures for discrete time processes with infinite horizon, *Math. Finance,* 4, 25-55.

Schachermayer, W. (2002), The Fundamental Theorem of Asset Pricing under proportional transaction costs in finite discrete time, Working paper.

Schmeidler, D. (1989), Subjective probability and expected utility without additivity, *Econometrica,* 52, 571-587.

Sentana, E. and G. Fiorentini (2001), Identification, estimation and testing of conditionally heteroskedastic factor models, *Journal of Econometrics,* 102, 143-164.

Sharpe, W.F. (1964), Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance* 19, 425-442.

Shephard, N.G. (1996), Statistical aspects of ARCH and stochastic volatility, in Cox, D.R., D.V. Hinkley and O.E. Barndorff-Nielsen (eds), *Time Series Models In Econometrics, Finance and other Fields,* Chapman & Hall, London.

Striker, C. (1990), Arbitrage et lois de martingale, *Ann. Inst. H. Poincaré Prob. Statist.,* 26, 451-460.

Szego, G. (2002), Measures of risk, *Journal of Banking and Finance,* 26, 1253-1272.

Taylor, S.J. (1986), *Modeling Financial Time Series,* John Wiley, Chichester.

Taylor, S. J. (1994), Modeling stochastic volatility, *Mathematical Finance,* 4, 183-204.

Terasvirta, T. (1996), Two stylized facts and the GARCH(1,1) model, Stockholm School of Economics, Working Paper 96.

Tsay, R.S. (2002), *Analysis of Financial Time Series,* Wiley, New York.

Tse, Y.K. (2000), A test for constant correlations in a multivariate GARCH model, *Journal of Econometrics,* 98, 107-127.

Tse, Y.K. and A.K.C. Tsui (2002), A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlation, *Journal of Business and Economic Statistics,* 20, 351-362.

Venter, G.G. (1991), Premium implications of reinsurance without arbitrage, *ASTIN Bulletin,* 21, 223-230.

Vlaar, P.J.G. (2000), Value at Risk models for Dutch bond portfolios, *Journal of banking and Finance,* 24, 1131-1154.

Wang, S. S. (2000), A class of distortion operators for pricing financial and insurance risks, *Journal of Risk and Insurance,* 67, 15-36.

Wang, S. S. (2001), A two factor model for pricing of risks, Working Paper, June 2001.

Wang, Y. (2002), Asymptotic nonequivalence of GARCH models and diffusions, *The Annals of Statistics,* 30, 754-783.

Zakodian, J.M. (1994), Threshold heteroskedastic models, *Journal of Economic Dynamics and Control,* 18, 931-955.

Zhang, M.Y., J.R. Russell and R.S. Tsay (2001), A nonlinear autoregressive conditional duration model with applications to financial transaction data, *Journal of Econometrics*

This page intentionally left blank

STEREOLOGICAL ESTIMATION OF THE ROSE OF DIRECTIONS FROM THE ROSE OF INTERSECTIONS

Viktor Beneš

Charles University, Dept. of Probability and Statistics, Sokolovská 83, CZ 186 75 Praha 8, Czech Republic

Ivan Sax

Mathematical Institute, Academy of Sciences of the Czech Republic, Žitná 25, CZ 115 67 Praha 1, Czech Republic

Abstract

      The paper is a review on the problem from stochastic geometry stated in the title. This problem concerns anisotropy quantification of fibre and surface processes. The stereological equation connecting the rose of directions and the rose of intersections (for a specific test system) was first attacked by means of analytical methods. Later on, an analogue from convex geometry lead to a deeper investigation using the notion of a Steiner compact. Various estimators of the rose of directions and their properties are reviewed in the planar and spatial case. The methods are important for practice when quantifying real structures in material science, biomedicine, etc.

Introduction

In the model based approach of stochastic geometry, objects are modelled by means of random sets [Matheron, 1975]. The isotropy of a random set can be defined by means of the invariance of its distribution with respect to any rotation operator. The deviance from this property is called anisotropy. Anisotropy is thus a rather broad notion. One can imagine the anisotropy of spatial distribution of objects which may form chains of preferred orientation violating thus the isotropy assumption. This type of anisotropy is formalized and studied e.g. in [Stoyan & Beneš, 1991]. Special models of random sets are fibre and surface processes where besides anisotropy of spatial distribution a simpler type of anisotropy may be described by means of the distribution of tangent, normal orientations of the fibres, surfaces at each point where it

is defined, respectively. This probability distribution $\mathcal{R}$ is called the rose of directions and will be of main interest in this paper.

In classical stereology, the information on geometrical objects is derived from observations on lower dimensional probes (test systems). A well-known stereological inverse problem (first formulated in [Hilliard, 1962]) relates the rose of directions to the rose of intersections between the process and a test system. In its simplest form it can be derived from the Buffon needle problem formulated in geometrical probability in 1777. The rose of intersections $P_L(u)$ is defined as the mean number of intersections between the process and a unit test system of orientation u. Given observed intersection numbers the stereological relation is used to the estimation of the rose of directions. There are several approaches to the solution of this problem. An analytical solution of the integral equation leads to various difficulties. We review estimators of the rose of directions separately in the planar and spatial case since the background is qualitatively different. Probably the most promising is the approach which makes use of an analogy from convex geometry which relates the support function of a zonoid to its generating measure. Statistical properties of the estimators such as consistency are reviewed and a comparison of methods and models is done by means of the simulated distribution of the Prohorov distance between the estimated and true rose of directions. Various test systems are investigated and demonstrating examples added.

3.1   An analytical approach

Consider a stationary planar fibre process Φ which is a random element in the measurable space $\mathcal{N}$ of fibre systems (collections of smooth fibres), see [Stoyan, Kendall & Mecke, 1995]. Let P be the distribution of Φ, L_A the intensity (mean fibre length per unit area) and $\mathcal{R}$ the rose of directions. A realization $\phi \in \mathcal{N}$ of Φ is alternatively interpreted as a locally finite length measure on $\mathbb{R}^2$, i.e. $\phi(B)$ is the length of fibres from ϕ in a Borel set B. Denote $w(x)$ the tangent orientation at a fibre point x. Axial orientations from $\Pi = [0, \pi)$ are considered.

3.1.1   A general stereological relation

First a more general stereological relation in $\mathbb{R}^2$ is derived, cf. [Mecke & Stoyan, 1980]. Let ν_d denote the $d-$**dimensional** Lebesgue measure. From the Campbell theorem [Stoyan, Kendall & Mecke, 1995] it follows immediately for an arbitrary non-negative measurable function f on $\mathbb{R}^2 \times \Pi$

$$\int_{\mathcal{N}} \int_{\mathbb{R}^2} f(x, w(x))\phi(dx)P(d\phi) = L_A \int_{\mathbb{R}^2} \int_{\Pi} f(x, \alpha)\mathcal{R}(d\alpha)\nu_2(dx). \quad (1.1)$$

LEMMA 1 *Let $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a non-negative measurable function, $\phi \in \mathcal{N}$. Then*

$$\int \sum_{x_1:(x_1,x_2)\in\phi} g(x_1,x_2)\nu_1(dx_2) = \int g(x)\sin(w(x))\phi(dx), \qquad (1.2)$$

where $x = (x_1, x_2)$ in $\mathbb{R}^2$.

Proof: A simple argument based on the total projection is used. For a Borel set $B \subset \mathbb{R}^2$ and $g = 1_B$ it holds

$$\int \sum_{x_1:(x_1,x_2)\in\phi} 1_B(x_1,x_2)\nu_1(dx_2) = \int_B \sin(w(x))\phi(dx),$$

since the both sides correspond to the length of the total projection of ϕ onto x_2-**axis**. Using the standard measure theoretic argument, formula (1.2) is obtained. $\qquad\square$

THEOREM 2 *Let $f : \mathbb{R} \times \Pi \rightarrow \mathbb{R}$ be a measurable non-negative function, Φ a stationary fibre process in $\mathbb{R}^2$. For the intersection of Φ with x_1-axis it holds*

$$E \sum_{x_1:(x_1,0)\in\Phi} f(x_1, w(x_1,0)) = L_A \int\int f(x_1,\alpha)\sin\alpha \mathcal{R}(d\alpha)\nu_1(dx_1).$$

$$(1.3)$$

Proof: Let $g : \mathbb{R} \rightarrow \mathbb{R}_+$ be a measurable function such that $\int g(t)\nu_1(dt) = 1$. It holds using (1.1), (1.2) and stationarity

$$L_A \int\int\int g(x_2)f(x_1,\alpha)\sin\alpha \mathcal{R}(d\alpha)\nu_2(dx) =$$

$$= E \int\int g(x_2)f(x_1,w(x))\sin w(x)\Phi(dx) =$$

$$= E \int \sum_{x_1:(x_1,x_2)\in\Phi} g(x_2)f(x_1,w(x_1,x_2))\nu_1(dx_2) =$$

$$= E \int g(x_2)\nu_1(dx_2) \sum_{x_1:(x_1,0)\in\Phi} f(x_1, w(x_1,0)).$$

$$\square$$

    The intersection of Φ with x_1-**axis** forms a stationary point process Ψ, denote its intensity P_L. Using special forms of f in Theorem 2, the relations are obtained between the fibre process and the induced structure on the test line (here x_1-**axis**).

COROLLARY 3 *In the situation of Theorem 2 let θ be the distribution of the fibre tangent orientation at the point of intersection with x_1-axis. Then it holds for $\beta \in \Pi$*

$$P_L \theta([0, \beta)) = L_A \int_0^\beta \sin \alpha \mathcal{R}(d\alpha), \tag{1.4}$$

thus

$$\theta([0, \beta)) = \frac{\int_0^\beta \sin \alpha \mathcal{R}(d\alpha)}{\int_0^\pi \sin \alpha \mathcal{R}(d\alpha)}, \tag{1.5}$$

if $\mathcal{R}(\{0\}) < 1$.

Proof: Putting $f(x_1, \alpha) = 1_{[0,1]}(x_1)1_{[0,\beta)}(\alpha)$ in (1.3) one obtains (1.4) and using this with $\beta = \pi$ finally (1.5) is concluded.   □

EXAMPLE 4 : *If $\mathcal{R}(\{0\}) = 0$ one can get L_A and $\mathcal{R}$ from P_L and θ (the latter pair of quantities can be estimated from the observation in the neighbourhood of a linear section). From (1.4) it holds*

$$\mathcal{R}([0, \beta)) = \frac{P_L}{L_A} \int_0^\beta (\sin \alpha)^{-1} \theta(d\alpha)$$

and for $\beta = \pi$ specially

$$L_A = P_L \int_0^\pi (\sin \alpha)^{-1} \theta(d\alpha).$$

A simpler choice of $f(x_1, \alpha) = 1_{[0,1]}(x_1)$ in (1.3) leads to the well-known formula

$$P_L = L_A \int \sin \alpha \mathcal{R}(d\alpha) \tag{1.6}$$

which corresponds to the frequent case that the information on intersection angles is not available. This case is in fact the main object of our paper.

3.1.2   Relation between roses of directions and intersections

Let Φ be a stationary fibre process in $\mathbb{R}^2$ as in the previous paragraph. Let $P_L(\beta)$, $\beta \in \Pi$, be the rose of intersections, i.e. the mean number of points $\Phi \cap l(\beta)$ per unit length of a test straight line $l(\beta)$ with orientation β. The basic integral equation relating the rose of directions of Φ to its rose of intersections is obtained by a simple generalization of (1.6). Consider Π with addition modulo π. The addition may be interpreted as a rotation of straight lines around origin in the plane $\mathbb{R}^2$.

It holds from (1.6)

$$P_L(\beta) = L_A \mathcal{G}_{\mathcal{R}}(\beta), \tag{1.7}$$

where we denote the sine transform

$$\mathcal{G}_{\mathcal{R}}(\beta) = \int_0^\pi |\sin(\beta - \alpha)| \mathcal{R}(d\alpha). \tag{1.8}$$

In the following text an equivalent expression of formula (1.7) is used. By S^{d-1} the unit sphere in $\mathbb{R}^d$ is denoted. Characterize a test line in $\mathbb{R}^2$ by its pair of unit normal vectors $\pm u \in S^1$, define $P_L(-u) = P_L(u)$. Denote $\langle .,. \rangle$ the scalar product.

Then it holds

$$P_L(u) = L_A \mathcal{F}_{\mathcal{R}}(u), \tag{1.9}$$

where the cosine transform

$$\mathcal{F}_{\mathcal{R}}(u) = \int_{S^1} |\langle u, v \rangle| \mathcal{R}(dv). \tag{1.10}$$

Note that here $\mathcal{R}$ represents a centrally symmetric probability measure on S^1. Further by $\mathcal{M}_d$, $\mathcal{P}_d$ the space of finite measures, probability measures on S^d, respectively, is denoted. If there is no danger of confusion we write $\mathcal{M}_d = \mathcal{M}$, $\mathcal{P}_d = \mathcal{P}$.

Let the test system for a fibre process in $\mathbb{R}^3$ be a plane or its subset characterized by a unit normal $u \in S^2$. Denoting by L_V the length intensity of a stationary fibre process Φ in $\mathbb{R}^3$ and by $P_A(u)$ the intensity of the point process induced by Φ in the test plane, we have

$$P_A(u) = L_V \int_{S^2} |\langle u, v \rangle| \mathcal{R}(dv). \tag{1.11}$$

By symmetry, a stationary surface process [Stoyan, Kendall & Mecke, 1995] of intensity S_V (mean surface area per unit volume) with a local normal $v \in S^2$ having an orientation distribution $\mathcal{R}$ induces on a test line of direction $u \in S^2$ a point process with intensity $P_L(u)$ and similarly

$$P_L(u) = S_V \int_{S^2} |\langle u, v \rangle| \mathcal{R}(dv). \tag{1.12}$$

The generalization to $\mathbb{R}^d$ for stationary fibre and hypersurface processes with intensity λ is straightforward; the form of the integral equations (1.11), (1.12) remains intact and only the integration region S^2 is replaced by S^{d-1}.

Denote by $\mathcal{U}$ a uniform probability measure on S^{d-1}. Note that for unknown $\mathcal{R}$ it is possible to estimate $\lambda = P_L/\mathcal{O}^{d-1}$, where $P_L = \int P_L(u)\mathcal{U}(du)$ can be approximated by an average of observations $P_L(u_j)$ systematically spread on S^{d-1} and $\mathcal{O}^{d-1} = \int \mathcal{F}_{\mathcal{R}}(u)\mathcal{U}(du)$ is a known constant ($\mathcal{O}^2 = 2/\pi$, $\mathcal{O}^3 = 1/2$). Therefore in the following the problem of estimating $\mathcal{R}$ can be considered equivalent to the problem of estimating $\lambda\mathcal{R}$.

3.1.3   Estimation of the rose of directions

Several methods based on formula (1.9) have been suggested for the estima-
tion of the rose of directions of a planar fibre process , cf. [Hilliard, 1962], [Di-
gabel, 1976], [Mecke, 1981], [Kanatani, 1984], [Rataj & Saxl, 1989], [Beneš
& Gokhale, 2000]. The aim is to estimate $\mathcal{R}$ given estimators $\eta_j = \eta(u_j)$
of $P_L(u_j)$, $u_j \in S^1$, $j = 1, ...n$, where η_j is the observed number of inter-
sections per unit test probe of orientation u_j. This was done basically in three
ways.

First, if a continuous probability density ρ of $\mathcal{R}$ exists we have

$$P_L''(u) + P_L(u) = 2L_A\rho(u), \tag{1.13}$$

which yields an explicit solution. This is in practice hardly tractable since the
second derivative P_L'' has to be evaluated from discrete data. However, the
formula is useful when a parametric model for $\mathcal{R}$ is available, cf. [Digabel,
1976].

Another natural approach to the solution of (1.7) is the Fourier analysis.
Hilliard [Hilliard, 1962] showed that for the Fourier images

$$\hat{\mathcal{R}}(k) = \int_0^\pi e^{2iku}\mathcal{R}(du), \quad k = ... -1,0,1..., \tag{1.14}$$

and $\hat{P}_L(k) = \int_0^\pi P_L(v)e^{2ikv}dv$, it holds

$$\hat{\mathcal{R}}(k) = \frac{1}{2L_A}(1 - 4k^2)\hat{P}_L(k), \quad k = ..,-1,0,1,... \tag{1.15}$$

When getting $\hat{P}_L(k)$ from the data and using (1.15), the variances of $\hat{\mathcal{R}}(k)$ may
tend to infinity.

The third approach is based on the convex geometry and will be described
in a separate section.

EXAMPLE 5 *Consider a fibre system in Fig. 1 with four test lines of equal
length 1 and the orientations* $u_i = i\pi/4$, $i = 0,1,2,3$, *respectively. The
intersection counts* $\eta_i = 6,3,7,7$, $i = 0,1,2,3$. *First a parametric approach
is used for the estimation of the rose of directions. Using a cardioidal model
[Rataj & Saxl, 1992] for ρ:*

$$\rho(v) = \frac{1}{\pi}(1 - k\cos 2(v - v_0)), \tag{1.16}$$

we obtain from (1.13)

$$P_L(u) = \frac{2L_A}{\pi}(1 - \frac{k}{3}\cos 2(u - u_0)). \tag{1.17}$$

Using the least squares method a fitted curve is obtained for $P_L(u)$, see Fig. 2 and the estimated rose of directions in Fig. 3. Since the parameter k was estimated by a value $k = 1.2$ which is greater than 1, the model density of the rose of directions yields also negative values which are presented in Fig. 3 along the orientation $\frac{3\pi}{4}$. The presence of negative values is a common problem of analytical estimators (also those based on Fourier expansions).
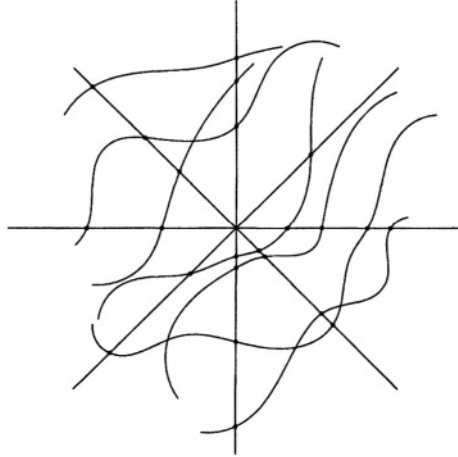
Figure 1.   A fibre system intersected by a system of test lines of unit lengths.
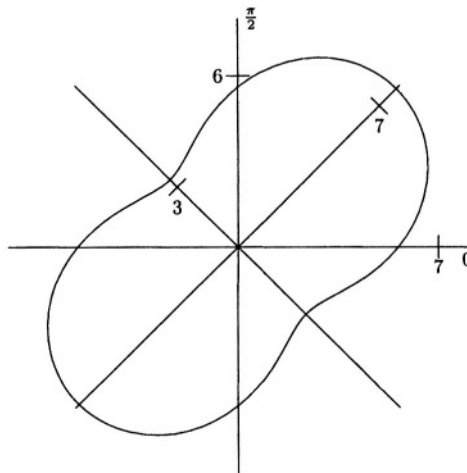
Figure 2.   Polar plot of intersection counts η_i from Fig.1 and the rose of intersections fitted by means of the cardioidal model.
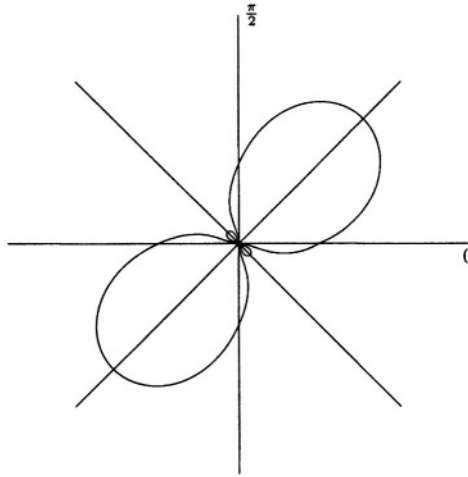
Figure 3.   Polar plot of the rose of directions estimated from data on Fig.2 using the cardioidal model. Two small loops along $\frac{3\pi}{4}$ have negative values of radii.

Consider further the three-dimensional situation. Because of an equal structure of integral equations (1.11), (1.12) for fibre and surface processes in $\mathbb{R}^3$, we restrict ourselves to the case of a stationary fibre process Φ. The problem is again to estimate the rose of directions $\mathcal{R}$ given a sample of test directions $u_1, ..., u_k \in S^2$ and estimators of $\eta_i = n_i/A$, where n_i is the number of intersections between Φ and a planar test probe with an area A and a normal orientation u_i. Similarly to the planar case and leaving aside the procedure based on convex geometry, there are basically two other approaches to the solution.

First a parametric approach means that a parametric type of the distribution on the sphere is suggested and the parameters estimated from the data using (1.11). In [Cruz-Orive et al, 1985] the axial Dimroth-Watson distribution was used

$$\mathcal{R}(du) = const. \exp(\kappa \cos(2\vartheta))du,$$

where $u = (\vartheta, \varphi)$ in spherical coordinates, $\vartheta \in [0, \pi/2]$ being the colatitude and $\varphi \in [0, 2\pi)$ the longitude. The parameter $\kappa \in \mathbb{R}^1$ is estimated.

Secondly an inversion formula to (1.11) is available ([Hilliard, 1962], [Mecke & Nagel, 1980]) using spherical harmonics. It is based on the fact that spherical harmonics are eigenfunctions of the cosine transform (1.10). The method in [Kanatani, 1984] approximates η_i by a finite series of even spherical harmonics and the inverse is then evaluated directly. An explicit inverse formula

from [Mecke & Nagel, 1980] says

$$\rho(v) = \frac{1}{L_V} \sum_{n=0}^{\infty} \frac{4n+1}{c_n} \int_{S^2} P_A(u) Q_{2n}(\langle u, v \rangle) \mathcal{U}(du), \qquad (1.18)$$

where Q_m is a Legendre polynomial of order m, ρ the probability density of $\mathcal{R}$ (with respect to $\mathcal{U}$). The constants c_n are

$$c_n = \frac{(-1)^{n+1}}{4n^2 - 1} \frac{1.3.5.\ldots.(2n-1)}{2.4.6.\ldots.(2n-2)}, \ n = 0, 1, \ldots$$

To conclude, analytical solutions of the inverse problem (1.9) in both two and three dimensions may lead to estimators of the rose of directions which are not non-negative densities. Typically these methods are not useful for sharp or multimodal anisotropies.

3.2   Convex geometry approach

In this section first some notions from convex geometry will be recalled (see e.g. [Schneider, 1993]). Let $\mathcal{K}$, $\mathcal{K}'$ be the system of all compact convex sets, nonempty compact convex sets in $\mathbb{R}^d$, respectively. If $K \in \mathcal{K}'$ then for each $u \in S^{d-1}$ there is exactly one number $h(K, u)$ such that the hyperplane (line in $\mathbb{R}^2$, plane in $\mathbb{R}^3$)

$$\{x \in R^d : \langle x, u \rangle - h(K, u) = 0\} \qquad (2.1)$$

intersects K and $\langle x, u \rangle - h(K, u) \leq 0$ for each $x \in K$. This hyperplane is called the support hyperplane and the function $h(K, u)$, $u \in S^{d-1}$, is the support function (restricted to S^{d-1}) of K. Equivalently, $h(K, u) = \sup\{\langle x, u \rangle, x \in K\}$. Its geometrical meaning is the signed distance of the support hyperplane from the origin of coordinates, $h(K, u) + h(K, -u) = w(K, u)$ is the width of K - the distance of the parallel support hyperplanes. The important property of $h(K, u)$ is its additivity in the first argument: $h(K_1 + K_2, u) = h(K_1, u) + h(K_2, u)$ (the addition of sets on the left hand side is in the Minkowski sense). Convex bodies with the centre of symmetry will be considered mostly in what follows. They will be shortly called *centred* if this centre is in the origin of $\mathcal{R}^d$.

A Minkowski sum of finitely many line segments is called a *zonotope*. Besides its being centrally symmetric, also its two-dimensional faces are centrally symmetric. Consequently, regular octahedron, icosahedron and pentagonal dodecahedron are not zonotopes. On the other hand in $\mathbb{R}^2$, all centrally symmetric polygons are zonotopes.

Consider a centred zonotope

$$Z = \sum_{i=1}^{k} a_i [v_i, -v_i], \qquad (2.2)$$

where $a_i > 0$, $v_i \in S^{d-1}$. Its support function is given by

$$h(Z, u) = \sum_{i=1}^{k} a_i \mid \langle u, v_i \rangle \mid \qquad (2.3)$$

and, conversely, a body $Z \in \mathcal{K}'$ with the support function (2.3) is a zonotope with the centre in the origin.

Consider the Hausdorff metric on $\mathcal{K}'$

$$H(K, L) = \max(\sup_{x \in L} d(K, x), \sup_{y \in K} d(y, L)),$$

the corresponding convergence is denoted as H–convergence. A set $Z \in \mathcal{K}'$ is called a *zonoid* if it is a H–limit of a sequence of zonotopes.

$Z \in \mathcal{K}'$ is a centred zonoid if and only if its support function has a representation

$$h(Z, u) = \int_{S^{d-1}} |\langle u, v \rangle| \mu(dv), \qquad (2.4)$$

for an even measure μ on S^{d-1}. μ is called the *generating measure* of Z and it is unique as shown in [Goodey & Weil, 1993]. For the zonotope (2.2) we have the generating measure

$$\mu = \sum_{i=1}^{k} a_i \epsilon_{v_i}, \qquad (2.5)$$

where $\epsilon_{v_i} = \frac{1}{2}(\delta_{v_i} + \delta_{-v_i})$ and δ_u is the Dirac measure concentrated at u.

Zonotopes and zonoids have several interesting properties and wide applications (see [Goodey & Weil, 1993], [Schneider & Weil, 1983]), e.g. the polytopes filling (tiling) $\mathbb{R}^3$ by translations are obligatory zonotopes (cubes, rhombic dodecahedrons, tetrakaidecahedrons). The roses of intersections $P_L(u)$, $P_A(u)$ are proportional to $\int_{S^2} |\langle u, v \rangle| \mathcal{R}(dv)$, cf. (1.11), (1.12). Consequently, they can be considered as support functions of certain zonoids the generation measures of which are proportional to the corresponding roses of directions. This idea has been put forward first by Matheron [Matheron, 1975] and the corresponding zonoid Z associated to $\mathcal{R}$ was called *the Steiner compact*. Because of the uniqueness of the generating measure of zonoids, the association is unique. The problem is, as before, to estimate (in atomic form) the generating measure μ or its normalized version $\mathcal{R}$ (rose of directions) from η_i assumed to be the support function values $h(Z_n, u_i)$ of a zonotope Z_n estimating Z in (2.4). The following theorem can serve as a basis of the procedure.

THEOREM 6 *For a zonoid $Z \subset \mathbb{R}^d$ and unit vectors $u_1, \ldots, u_k$ there always exists a zonotope Z_k which is the sum of at most k segments and fulfills*

$$h(Z, \pm u_1) = h(Z_k, \pm u_1), \ldots, h(Z, \pm u_k) = h(Z_k, \pm u_k). \qquad (2.6)$$

If a zonotope Z_k satisfying (2.6) is found its generating measure of the type (2.5) yields after normalizing to a probability measure the desired estimator $\mathcal{R}_k$ of the rose of directions $\mathcal{R}$. Generating measures belong to the space $\mathcal{M}$. The H–convergence on $\mathcal{K}'$ is equivalent to the weak convergence on $\mathcal{M}$ with respect to the transformation (2.4). Since the weak convergence on $\mathcal{M}$ is metrized by the Prohorov metric, it is possible to describe theoretically the quality of the estimator by means of the Prohorov distance between $\mathcal{R}_k$ and $\mathcal{R}$.

The Prohorov distance between measures $Q, T \in \mathcal{M}$ is defined as

$$r(Q,T) = \inf\{\varepsilon > 0; \ Q(C) \le T(C^\varepsilon) + \varepsilon, \ T(C) \le Q(C^\varepsilon) + \varepsilon$$
$$\text{for all closed } C \subset S^{d-1}\}.$$

This definition is for probability measures and therefore also in our situation equivalent to a restricted condition which is used in the form

$$r(\mathcal{R}_k, \mathcal{R}) = \inf\{\varepsilon > 0; \ \mathcal{R}_k(C) \le \mathcal{R}(C^\varepsilon) + \varepsilon \text{ for all closed } C \subset S^{d-1}\}. \tag{2.7}$$

Because of (2.5) the estimator $\mathcal{R}_k$ is discrete with finite support $supp\,\mathcal{R}_k \subset \{z_1, ..., z_k\}$ so there is the following reduction to finitely many conditions, cf. [Beneš & Gokhale, 2000]. It holds

$$r(\mathcal{R}_k, \mathcal{R}) = \inf\{\varepsilon > 0; \ \mathcal{R}_k(C) \le \mathcal{R}(C^\varepsilon) + \varepsilon \text{ for all } C \subset supp\,\mathcal{R}_k\}. \tag{2.8}$$

This enables to compute the Prohorov distance which will be used in the following for a comparison of estimators.

The construction of a zonotope or of a sequence of zonotopes Z_k such that $H(Z_k, Z) \to 0$ when $k \to \infty$ is simple only in $\mathbb{R}^2$. It is sufficient to set

$$Z_k = \bigcap_{i=1}^{k} \{x \in \mathbb{R}^2; \ \langle x, u \rangle \le h_i\}, \tag{2.9}$$

since every centred polygon is a zonotope in $\mathbb{R}^2$. In $\mathbb{R}^3$ this is not the case thus the situation is more complicated and an optimization procedure based on the constructive proof of Theorem 6 in [Campi, Haas & Weil, 1994] is a partial solution. Recently the paper [Kiderlen, 2001] makes a substantial step forwards in this problem.

Consequently, the estimation of $\mathcal{R}$ by means of the Steiner compact will be treated separately for the planar and spatial cases as follows.

3.2.1   Steiner compact in $\mathbb{R}^2$

The relation between a measure $\mu \in \mathcal{M}$ and the zonoid Z generated by it has a direct consequence of geometrical nature. Let $T_Z(u)$ be the intersection point of the support line (corresponding to u) with Z (if the intersection is a

line segment, $T_Z(u)$ will be the endpoint with respect to the anti-clockwise orientation of the boundary ∂Z of Z). If x, y are two points of ∂Z by $l_Z(x, y)$ the length of the corresponding arc of ∂Z is denoted. The following result comes from [Rataj & Saxl, 1989] and it was obtained in [Matheron, 1975] in a more general setting.

THEOREM 7 *There is a one-to-one correspondence between symmetric elements $\mu \in \mathcal{M}$ and $Z \in \mathcal{K}'$ centrally symmetric given by*

$$\mu((s, t]) = l_Z(T_Z(s), T_Z(t)), \quad s, t \in S^1.$$

Consequently, the length (per unit area) of fibres with tangents within an interval of directions $(v_1, v_2]$ is proportional to the length of the boundary ∂Z bounded by the pair of equally oriented tangents.

For a stationary fibre process Φ and the zonoid (Steiner compact) Z associated to the rose of directions $\mathcal{R}$ of Φ it holds

$$h(Z, u) = \frac{1}{2} L_A \mathcal{F}_\mathcal{R}(u), \quad u \in S^1, \tag{2.10}$$

i.e. comparing with (1.7) $2h(Z, u) = P_L(u), \ u \in S^1$.

[Rataj & Saxl, 1989] suggested a graphical method of estimation of the rose of directions by means of its related Steiner compact set. Let

$$p_i = \frac{1}{2} \eta_i = \frac{1}{2} \frac{n_i}{l} \tag{2.11}$$

be the estimators of the support function values at orientations (axial) $u_i \in S^1, \ i = 1, ..., k$, where n_i is the number of intersections of the studied fibre system (realization of a fibre process) with a test segment of length l and orientation u_i. Then by (2.9), the convex polygon (2k-gon, $p_{i+k} = p_i$, $i = 1, \dots, k$)

$$Z_k = \{x : \langle x, u_i \rangle \le p_i, \ i = 1, ..., 2k\} \tag{2.12}$$

provides a basis to the estimation of the Steiner compact Z related to $\mathcal{R}$. The measure μ_k corresponding to Z_k according to Theorem 7 is

$$\mu_k = \sum_{i=1}^{k} h_i \delta_{u_i}, \tag{2.13}$$

where h_i are the lengths of edges of the polygon Z_k. The h_i's have outer normals u_i, in fact Z_k may have less edges than $2k$ if $h_i = 0$ for some i. The relation between p_i and h_i follows (cf.[Beneš & Gokhale, 2000], we denote $a_+ = \max(a, 0)$):

$$h_i = \left(\min_{-\pi < \beta_{ij} < 0} \frac{p_i \cos \beta_{ij} - p_j}{\sin \beta_{ij}} - \max_{0 < \beta_{ij} < \pi} \frac{p_i \cos \beta_{ij} - p_j}{\sin \beta_{ij}} \right)_+, \quad i = 1, ..., k, \tag{2.14}$$

where β_{ij} are anticlockwise oriented angles between u_i and u_j. Finally, after normalization

$$h_i' = \frac{h_i}{\sum_i h_i} \qquad (2.15)$$

we obtain the desired estimator $\mathcal{R}_k$ of the rose of directions $\mathcal{R}$:

$$\mathcal{R}_k = \sum_{i=1}^{k} h_i' \delta_{u_i}. \qquad (2.16)$$

The *H*-convergence of Z_k is investigated by [Rataj & Saxl, 1989].

EXAMPLE 8 *We continue in Example 5. This time the data from Fig.1 are evaluated by means of the Steiner compact method. Using formula (2.12) the zonotope in Fig.4 (left) is constructed (recall that the test lines are character-ized by its unit normal vectors) and from (2.14) the estimator (2.13) is obtained and plotted in Fig.4 (right). The dominant direction is recognized, however, the second largest atom at $\frac{3\pi}{4}$ is unrealistic as a consequence of the sparse test system.*
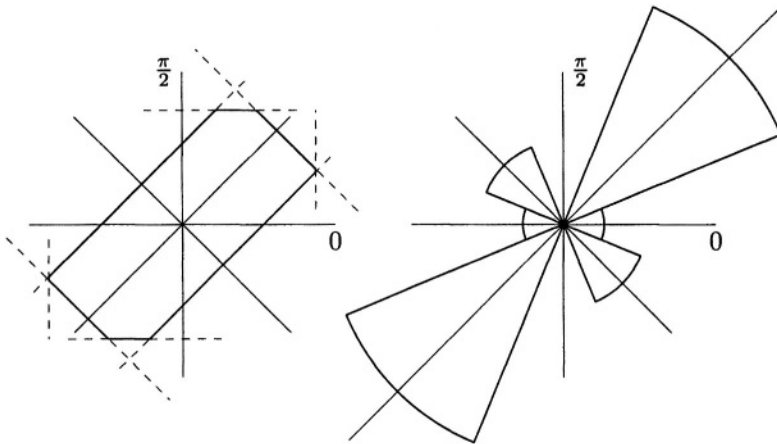
Figure 4.   A Steiner compact Z_k (left) and the estimated rose of directions (right) for data from Example 5. On the right a circular plot is used where h_i' in (2.16) correspond to the radii of classes.

[Rataj & Saxl, 1989] developed a modification of Steiner compact estima-tors of $\mathcal{R}$ by means of the following smoothing. For integer n and orientations $0 < u_1 < u_2 < ... < u_n \leq \pi$, for integer r and weights

$$\{c_j : j = -r, .., 0, .., r\}, \; c_{-j} = c_j \geq 0, \; j = 0, .., r, \; \sum_j c_j = 1 \qquad (2.17)$$

they construct polygons

$$\bar{K}_n = \{x : \langle x, u_i \rangle \le \bar{p}_i, \; i = 1, ..., n\}, \text{where} \quad \bar{p}_i = \sum_{j=-r}^{r} c_j p_{i+j}, \; i = 1, .., n$$

(2.18)

and pi are as in (2.11). Let h_i be the lengths of edges of $\bar{K}_n$ and h_i' as in (2.15). Then the estimator of $\mathcal{R}$ is $\mathcal{R}_n(B) = \sum_{i=1}^{n} h_i' 1_B(u_i)$, for a Borel set $B \subset \mathbb{R}^2$, cf. (2.16).

EXAMPLE 9 *Again for the data from Example 5 we use the modified Steiner compact estimator with $r = 1$ and $(c_{-1}, c_0, c_1) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. In Fig.5 (left) the Steiner compact estimated from the smoothed rose of intersections is drawn, the estimator of the rose of directions in Fig.5 (right) corresponds better to the data at the first sight.*
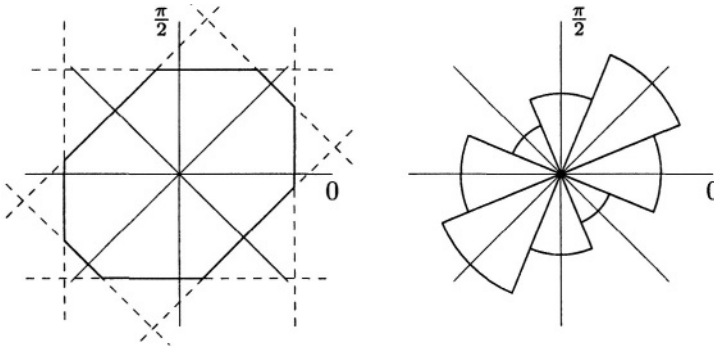
Figure 5.   A Steiner compact (left) and the estimated rose of directions (right) for data from Example 5 using the modified method with smoothing described in Example 9.

There is a theorem in [Rataj & Saxl, 1989] concerning the properties of the modified Steiner compact estimator.

THEOREM 10 *Let $\varepsilon > 0$ and $\alpha \in (0, 1)$. Then there is a plan of experiment, i.e. integers $n, r; u_1, \ldots u_n \in S^1$ and c_j as in (2.17), such that for a planar fibre system and $\bar{p}_i$, $\bar{K}_n$ in (2.18) we have probability*

$$P(H(\bar{K}_n, K) \le \varepsilon L_A) \ge \alpha$$

under the condition that $\bar{p}_i - p_i$, $i = 1, \ldots n$ is a family of independent, centred normally distributed random variables with variances bounded by a constant $\sigma^2 > 0$.

The normality assumption seems to be quite appropriate when using independent test lines, which can be achieved when independent realizations of a fibre process are available.

3.2.2   Poisson line process

Any straight line $l(x)$ in the plane can be represented by a point $x = (v, y)$ in the parametric space formed by a set $C_1 = (0, \pi] \times (-\infty, \infty)$. Here v is the orientation of the line and y its signed distance from the origin. We have y positive, negative for lines intersecting the positive, negative horizontal semi-axis in $\mathbb{R}^2$. respectively. If $v = \pi$, y is positive for lines in the upper half plane. We can thus represent a stationary line process Φ by means of a point process Ψ on C_1, such that the intensity measure Λ of the process Ψ is (see [Stoyan, Kendall & Mecke, 1995])

$$\Lambda(d(v, y)) = L_A dy \mathcal{R}(dv). \tag{2.19}$$

If the stationary line process Φ is Poisson then the point process Ψ is Poisson stationary with respect to y coordinate. Conversely, a random point process on C_1 stationary in $y-$coordinate defines a stationary line process in $\mathbb{R}^2$.

We will investigate the intersections of a line process with test segments of constant length l and of varying orientations. Consider the unit semicircle $x = \cos \beta$, $y = \sin \beta$, $\beta \in [-\pi, \pi]$. Denote $\alpha_n = \frac{\pi}{2n}$ and define the test system $\mathcal{T}$ of n segments s_i inscribed in the semicircle, see Fig. 6a. The segments have centres (x_j, y_j), $x_j = \cos \beta_j \cos \alpha_n$, $y_j = \sin \beta_j \cos \alpha_n$, normal orientations $\beta_j = (2j - n - 1)\alpha_n$, $j = 1, ..., n$. The segments have equal lengths $l = 2 \sin \alpha_n$. The total length of $\mathcal{T}$ converges to π with $n \to \infty$. Any straight line in the plane has at most two intersections with the test system $\mathcal{T}$. Denote by A_i, A_{ij} the subsets of C_1 corresponding to lines which intersect exactly one, two segments, respectively. In Fig. 6b these subsets are drawn in the case of $n = 3$.
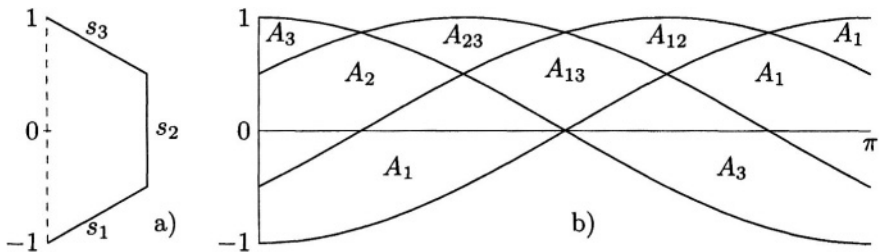
Figure 6.   The test system $\mathcal{T}$ for $n = 3$ (a), the corresponding subsets $A_i, A_{ij}, i, j = 1, ..., n$, $i < j$, (b).

Consider a stationary Poisson line process Φ with intensity L_A and a rose of directions $\mathcal{R}$. Denote N_{ij}, N_i the independent Poisson distributed random variables with parameters λ_i, λ_{ij}, respectively, corresponding to numbers of intersections of Φ with given $i-$th, $i-$th and $j-$th segment, respectively. It

holds

$$\lambda_{ij} = L_A \int_{A_{ij}} dy \mathcal{R}(dv), \quad \lambda_i = L_A \int_{A_i} dy \mathcal{R}(dv).$$

From a realization of the process Φ we get estimators of support function values

$$p_j = \frac{1}{2l}\left(N_j + \sum_i N_{ij}\right) \quad j = 1, ..., n.$$

Observe that

$$cov(p_i, p_j) = \frac{1}{4l^2} var N_{ij}, \quad i \neq j.$$

EXAMPLE 11 *The aim is to obtain the probability distribution of the Prohorov distance between the estimator $\mathcal{R}_n$ in (2.16) and a theoretical $\mathcal{R}$. For a stationary Poisson line process and a special test system in Fig.6 this can be achieved by just simulating the data N_i, N_{ij} from the Poisson distribution, evaluating the estimators and finally the Prohorov distance. The results from 1000 independent simulations for $\mathcal{R} = \mathcal{U}$ uniform yield approximations of probability density of the Prohorov distance $r(\mathcal{R}_n, \mathcal{R})$ in Fig. 7 (without smoothing), Fig. 8 (with smoothing (2.18)), respectively.*
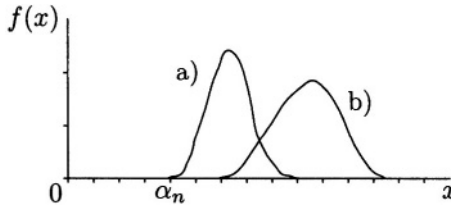
Figure 7.   Estimated probability densities functions of the Prohorov distance $r(\mathcal{R}_n, \mathcal{U})$, $n = 8$, $\alpha_n = 0.19$, for $L_A = 50$ (a), $L_A = 1000$ (b).

3.2.3   Theoretical properties of the Prohorov distance distribution

If the distance between a discrete and continuous distribution is measured we observe that the distribution of the Prohorov distance (cf. Figs.7, 8) is not concentrated near zero. Among the discrete distributions $\mathcal{R}_n \in \mathcal{P}$ with a support T of cardinality at most n the uniform discrete distribution $\mathcal{U}_n$ (with exactly n equidistant atoms) is the nearest to $\mathcal{U}$ in the sense of Prohorov distance. It holds $r(\mathcal{U}_n, \mathcal{U}) = \frac{\pi}{2n+\pi}$ since the worst case in (2.8) is

$$1 = \mathcal{U}_n(\tau) \leq \mathcal{U}(\tau^\varepsilon) + \varepsilon = \frac{2n\varepsilon}{\pi} + \varepsilon.$$
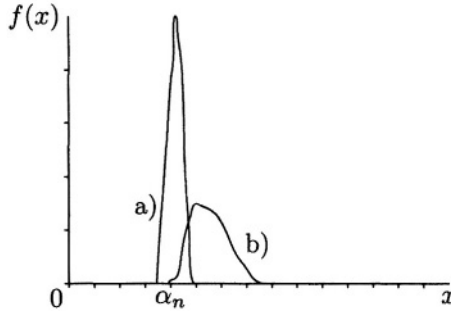
Figure 8. The same case as in Fig.7 after smoothing with $r = 2$, $c_j = \frac{1}{2r+1}$, $j = -r, ..., r$.

A larger lower bound can be obtained under a supplementary condition [Beneš & Gokhale, 2000]:

PROPOSITION 1 *For the test system* $\mathcal{T}$, *an isotropic fibre process and the Steiner compact estimator* $\mathcal{R}_n$ *of* $\mathcal{R} = \mathcal{U}$ *it holds that the Prohorov distance*

$$r(\mathcal{R}_n, \mathcal{R}) \geq \frac{4\alpha_n}{\pi + 2}$$

under the condition $A = [h_i = 0 \text{ for some } i]$.

Proof: Let i be the index which satisfies A, assume that $r(\mathcal{R}_n, \mathcal{U}) < \frac{4\alpha_n}{\pi+2}$. Then there is a $\delta > 0$ such that $r(\mathcal{R}_n, \mathcal{U}) = \frac{4\alpha_n}{\pi+2} - \delta$. We use an equivalent definition of the Prohorov distance

$$r(\mathcal{R}_n, \mathcal{U}) = \inf\{\varepsilon > 0; \ \mathcal{U}(C) \leq \mathcal{R}_n(C^\varepsilon) + \varepsilon, \ C \text{ closed}\}.$$

Put

$$C = [\beta_i - \frac{2\pi\alpha_n}{\pi + 2}, \ \beta_i + \frac{2\pi\alpha_n}{\pi + 2}],$$

then $\mathcal{U}(C) = \frac{4\alpha_n}{\pi+2}$ and for $\varepsilon = \frac{4\alpha_n}{\pi+2} - \delta$ we have $C^\varepsilon = [\beta_i - 2\alpha_n + \delta, \beta_i + 2\alpha_n - \delta]$ and $\mathcal{R}_n(C^\varepsilon) = 0$. Altogether $\mathcal{R}_n(C^\varepsilon) + \varepsilon = \frac{4\alpha_n}{\pi+2} - \delta < \mathcal{U}(C)$, which leads to a contradiction. $\qquad\square$

A lower bound for $Pr(A)$ is $\sum_i Pr(B_i) - \sum_{i<j} Pr(B_i \cap B_j)$, where the event $B_i = [p_{i-1} + p_{i+1} - 2p_i \cos\frac{\pi}{n} < 0]$.

3.2.4   Simulation study

In this section, the Steiner compact estimation procedure for more complex models of fibre systems is investigated which needs a simulation of a realization together with a chosen test system.
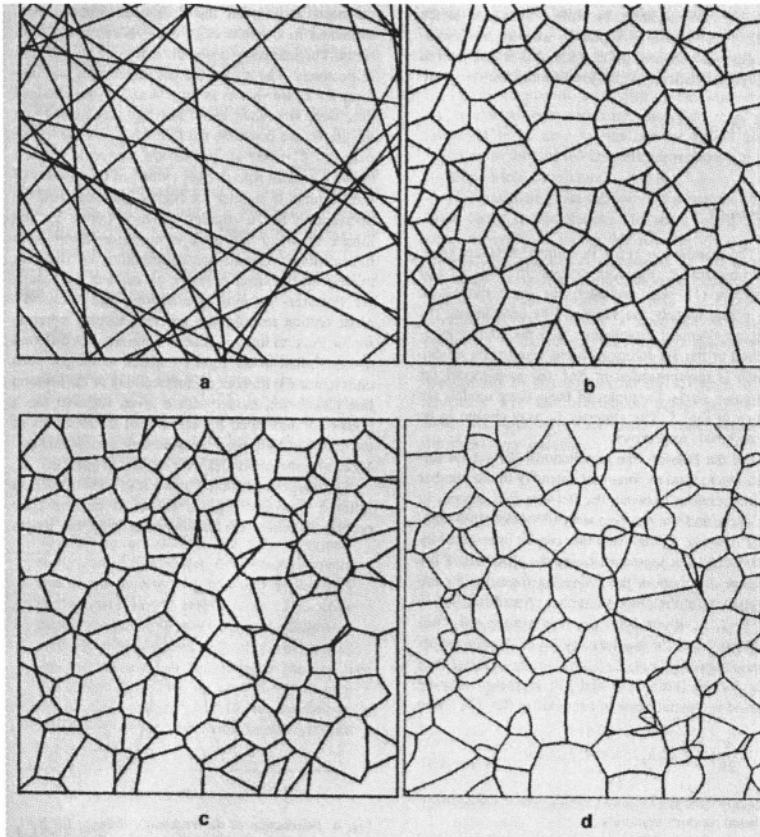
Figure 9.   Realizations of tessellations with the approximate intensity $L_A = 22$: (a) a Poisson
line process, (b) a 2D Poisson-Voronoi tessellation, (c) a planar section of a 3D Poisson-Voronoi
tessellation (not used in simulations), (d) a planar section of a 3D Johnson-Mehl tessellation.

The distribution of the Prohorov distance, given the uniform rose of direc-
tions, the test system in Fig. 6 and estimator (2.13), was evaluated for three
models in the plane, see Fig. 9. Namely they are the Poisson line process
, the Poisson-Voronoi tessellation [Stoyan, Kendall & Mecke, 1995] and the
planar intersection of the three-dimensional Johnson-Mehl tessellation [Ohser
& Mücklich, 2000] model. Using the algorithm for the Prohorov distance esti-
mation and 1000 repeated simulations the distribution of Prohorov distance is
obtained in Fig. 10. It follows that for more regular fibre processes (formed by
tessellations) the estimator is more precise.
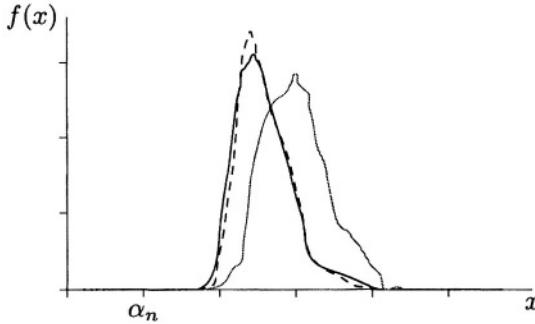
Figure 10.   Estimated probability densities functions of the Prohorov distance $r(\mathcal{R}_n, \mathcal{U})$, $n = 12$, $\alpha_n = 0.131$, for $L_A = 100$. The result for the Poisson line process is marked by the gray dotted line, for the Poisson-Voronoi tessellation by solid line, and for the Johnson-Mehl tessellation by the dashed line.

3.2.5   Curved test systems

We shall investigate the role of curved test systems in the estimation of the rose of directions of a planar fibre process following [Beneš & Gokhale, 2000]. Consider a test system T' of arcs with finite total length l and $T'(B)$, $B \in \mathcal{B}^2$, the corresponding length measure of T' in B. Assume that almost surely (w.r.t. the length measure) the tangent orientation $w(x)$ of T' at x is defined. Then the orientation distribution Q of T' on S^1 is given by

$$\int f(\alpha)Q(d\alpha) = \frac{1}{l} \int f(w(x))T'(dx)$$

valid for any $f \geq 0$ measurable on S^1. Denote also by $T'(u)$ the rotation of $T' = T'(0)$ by an angle of $u \in S^1$ with $x-$axis.

[Mecke, 1981] points out that if the test system is formed by curved lines with tangent orientation distribution $Q \in \mathcal{P}$, then

$$P_L^Q(u) = L_A \mathcal{G}_{\mathcal{R}*Q_-}(u), \qquad (2.20)$$

where $P_L^Q(u)$ is the rose of intersections $\Phi \cap T'(u)$. Further Q_- is the reflection of Q, i.e. $\int f(u)Q_-(du) = \int f(\pi - u)Q(du)$ for any non-negative measurable function f on S^1, and $\mathcal{R} * Q_-$ is the convolution of measures defined by $\int f(x)\mathcal{R} * Q_-(dx) = \int \int f(x+y)\mathcal{R}(dx)Q_-(dy)$. In particular for $Q = \mathcal{U}$ uniform it follows from (2.20) that $P_L^{\mathcal{U}}(u) = \frac{2}{\pi}L_A$, $u \in S^1$, is a constant denoted $P_L^{\mathcal{U}}(u) = P_L$.

Generally, comparing (1.7) and (2.20) we see that if there is a statistical method for estimating $\mathcal{R}$ from (1.7), the same method estimates $\mathcal{R} * Q_-$ from (2.20) when using a curved test system. Unfortunately, the system $\mathcal{P}$ with

convolution operation does not posses natural inverse element to solve equation $\mathcal{R} * Q = Q_1$ for an unknown $\mathcal{R}$, cf. [Heyer, 1977].

Elements $\delta_u \in \mathcal{P}$, $u \in S^1$ provide rotation $Q(u) = Q * \delta_u$ of a given measure $Q \in \mathcal{P}$. The effect of the convolution operation of measures on Steiner compact sets may be observed most easily when the both measures are discrete: $\mathcal{R} = \sum_{i=1}^{n} a_i \delta_{u_i}$, $Q = \sum_{j=1}^{m} b_j \delta_{v_j}$, $\sum a_i = \sum b_j = 1$, $a_i, b_j > 0$, $u_i, v_j \in S^1$. Then the convolution $\mathcal{R} * Q$ is again a measure with finite support $\{u = u_i + v_j; \ i = 1, .., n, \ j = 1, .., m\}$. The atom in $u_i + v_j$ has size $a_i b_j$. Now the Steiner compact associated with a discrete measure has form $Z = \sum_{i=1}^{k} [-c_{ij}, c_{ij}]$, cf. (2.2), where c_{ij} are vectors in $\mathbb{R}^2$ with orientations $u_i + v_j$ and lengths $a_i b_j$.

The following result comes from [Hilliard, 1962], [Mecke, 1981].

PROPOSITION 2 *For the Fourier images* $\hat{R}(k)$, $\hat{Q}(k)$ *defined by (1.14) and for* $\hat{P}_L^Q(k) = \int_0^\pi P_L^Q(u) e^{2iku} du$ *it holds*

$$\hat{\mathcal{R}}(k)\hat{Q}(-k) = \frac{1}{2L_A}(1 - 4k^2)\hat{P}_L^Q(k), \quad k = .., -1, 0, 1, ... \qquad (2.21)$$

Proof: Let f be a π-**periodic** twice continuously differentiable function. Then $\int_0^\pi f(u)\mathcal{R}(du) = \frac{1}{2}\int_0^\pi \mathcal{G}_{\mathcal{R}}(u)[f(u) + f''(u)]du$ using two-fold integration by parts. Then putting $f(u) = e^{2iku}$ we get formula (1.15). Using the same idea to $\mathcal{R} * Q_-$ and using the fact that the Fourier transform of a convolution is a product of Fourier transforms we get (2.21).   □

Further we observe that the local smoothing in (2.18) can be expressed in terms of the convolution with a discrete measure Q representing the orientation distribution of a test system.

PROPOSITION 3 *Let* $Q = \sum_{i=1}^{m} b_i \delta_{v_i}$, $b_i > 0$, $\sum b_i = 1$, $v_i \in S^1$, $i = 1, .., n$. *Then*

$$P_L^Q(u) = \sum_{i=1}^{m} b_i P_L(u - \pi + v_i), \ u \in S^1.$$

Proof: We have $Q_- = \sum_i b_i \delta_{\pi - v_i}$ and $\mathcal{G}_{\mathcal{R}*Q_-}(w) = \int_0^\pi |\sin(u - w)|\mathcal{R} * Q_-(du) =$
$= \sum_{i=1}^{m} b_i \int_0^\pi |\sin(u + \pi - v_i - w)|\mathcal{R}(du) = \sum_{i=1}^{m} b_i \mathcal{G}_{\mathcal{R}}(w - \pi + v_i)$.
Then

$$P_L^Q(w) = L_A \mathcal{G}_{\mathcal{R}*Q_-}(w) = L_A \sum_{i=1}^{m} b_i \mathcal{G}_{\mathcal{R}}(w - \pi + v_i)$$

$$= \sum_{i=1}^{m} b_i P_L(w - \pi + v_i).$$

□

Naturally it is not necessary to restrict to atomic measures Q for local smoothing; diffuse measures correspond to curved test systems.

EXAMPLE 12 *Let* $\mathcal{R} = \delta_0$ *and* Q_- *has probability density* $q(w) = \frac{1}{a}$ *for* $w \in [0, a)$ *and* $q(w) = 0$ *elsewhere for some* a, $0 < a < \frac{\pi}{2}$. *Then* $\mathcal{G}_{\mathcal{R}}(w) = \sin w$ *and* $\mathcal{G}_{\mathcal{R}*Q_-}(w) = \frac{\cos w - \cos(w+a)}{a}$, $w \in [0, \pi - a)$, *with apparent smoothing effect for* $a > 0$ *small.*

It is concluded that curved test systems present an alternative to local smoothing in (2.18) when estimating the Steiner compact. It should be kept in mind that using the rose of intersections $P_L^Q(u)$ (i.e. using local smoothing) we get estimators of $\mathcal{R} * Q_-$ which is not exactly $\mathcal{R}$. In $\mathbb{R}^3$, the convolution operation does not exists in a simple form because of the complexity of the space of rotations on S^2.

3.2.6   Steiner compact in $\mathbb{R}^d$ and in $\mathbb{R}^3$

The complications in approximating the zonoid associated to the rose of direction $\mathcal{R}$ in $\mathbb{R}^d$, $d \geq 3$, are consequences of the special nature of zonotopes and zonoids. Thus the intersection of supporting halfspaces (2.9) produces a centrally symmetric polytope but it is not a zonotope in general because its two-dimensional faces need not be centrally symmetric. Also the interpolation and smoothing procedures do not produce zonoids but only generalized zonoids. They are centrally symmetric but their even generating measures are not non-negative as required but only signed ones [Schneider, 1993]. Consequently, the inversion of the integral equation (1.11) proposed in [Hilliard, 1962], [Kanatani, 1984] need not give a non-negative estimator of the rose of direction $\mathcal{R}$ as pointed out by [Goodey & Weil, 1993].

More correct solutions are based on the Theorem 6 as shown in [Kiderlen, 2001]. The basic idea is an approximation of the generating measure $\mu \in \mathcal{M}$ by a measure concentrated on a finite support

$$T_m = \{v_1, \ldots, v_m, -v_1, \ldots, -v_m\} \subset S^{d-1}, \tag{2.22}$$

such that $Z_m = \sum_{i=1}^{m} \alpha_i [v_i, -v_i]$ is a zonotope estimating a zonoid Z corresponding to μ. The problem is a suitable choice of T_m and of the weights α_i such that $Z_m \to Z$ in H–convergence.

Let Φ be a stationary fibre process in $\mathbb{R}^d$ with intensity λ (specially in $\mathbb{R}^3$ we denote λ by L_V) and the rose of directions $\mathcal{R}$. Consider k fixed test hyperplanes $u_i^\perp$ with normals $u_i \in S^{d-1}$, $i = 1, \ldots, k$, such that they do not contain a common line. Denote $\eta_i = \#(\Phi \cap W_i)$ the number of intersection points counted in $\Phi \cap W_i$, where $W_i \subset u_i^\perp$ are the observation windows of unit areas $\nu_{d-1}(W_i) = 1$ in the test hyperplanes. The set of all η_i

then constitutes a random vector $\eta(\mu) = \{\eta_1, \ldots, \eta_k\}$ with the mean value $\mathbf{E}\eta(\mu) = \{\mathbf{E}\eta_1, \ldots, \mathbf{E}\eta_k\}$, where in $\mathbb{R}^3$ we have $\mathbf{E}\eta_i = P_A(u_i)$, $i = 1, \ldots, n$. In contrast to the test system $\mathcal{T}'$ in the planar case, we assume here that η_i are independent which can be ensured by examining independent realizations of Φ for different planes $u_i^\perp$. This assumption is violated in the next section where curved or polytopal probes are used for investigation of a single realization.

The idea of a maximum likelihood (ML) estimator of the measure μ was formulated in [Mair, Rao & Anderson, 1996] and is further developed in [Kiderlen, 2001]. Assume that the fibre process Φ is a stationary Poisson line process, η_i are Poisson distributed. Further assume that the observed realization $\hat{\eta}$ of $\eta(\mu)$ is a non-zero vector. The ML estimator $\hat{\mu}$ maximizes the log-likelihood function $L(\mu) : \mu \mapsto \log P(\eta(\mu) = \hat{\eta})$, i.e.

$$L(\mu) = \sum_{i=1}^{k} (\hat{\eta}_i \log(\mathbf{E}\eta_i) - \mathbf{E}\eta_i) \qquad (2.23)$$

The convex optimization problem
(i) to minimize $-L(\mu)$ with respect to $\mu \in \mathcal{M}$
is shown to have a solution in [Mair, Rao & Anderson, 1996]. It is not unique but any two solutions μ_1, μ_2 are tomographically equivalent, i.e. they satisfy

$$\mathbf{E}\eta_i(\mu_1) = \mathbf{E}\eta_i(\mu_2)$$

for all $i = 1, \ldots, k$. For large k and regularly distributed u_i on S^{d-1}, the Prohorov distance of tomographically equivalent measures is small.

To solve the problem (i) numerical methods must be used searching for a solution in the finite-dimensional subcone $\mathcal{M}(T_m) \subset \mathcal{M}$ of measures with support in T_m. Then the optimization problem (i) reduces to
(ii) to minimize $-L(\mu)$ with respect to $\mu \in \mathcal{M}(T_m)$.
There is a choice of T_m which is optimal in the sense of the following theorem. We will specify this just for $d = 3$, for general formulation see [Kiderlen, 2001], where the theorem is proved under assumption that Φ is the Poisson line process and, consequently, $\eta(\mu)$ is multivariate Poisson distributed.

THEOREM 13 *Under the above assumptions concerning the choice of test planes and $\hat{\eta}$, the problem (ii) has a solution. If T_m is the set of all unit vectors orthogonal to the all linearly independent pairs in $\{u_1, \ldots, u_k\}$ then any solution of (ii) is a solution of (i).*

Clearly $m \leq k(k-1)/2$ for $d = 3$. Denote $\mathcal{R}_{m,k}$ the ML estimator of the rose of direction based on k test orientations and T_m as introduced in the Theorem 13: $\hat{\mu} = \lambda \mathcal{R}_{m,k}$. It can be shown that Theorem 13 holds for general stationary fibre processes, too. It need not be a maximum likelihood estimator then (the

Poisson property of η_i may fail), but it is consistent in the following sense [Kiderlen, 2001]. An asymptotically smooth sequence $\{u_1, u_2, \ldots\} \in \mathbb{R}^{d-1}$ is such that the sequence of measures $\tau_k = \frac{1}{k} \sum_{i=1}^{k} \delta_{u_i}$ converges weakly in $\mathcal{M}$ and the limit has a positive density.

THEOREM 14 *Let Φ be a stationary fibre process in $\mathbb{R}^3$ with $\mu = L_V \mathcal{R}$ which is not supported by any great circle in S^2 and $\{u_1, u_2, \ldots\}$ be an asymptotically smooth sequence in S^2. Let $\eta_1, \ldots, \eta_k$ be non-correlated intersection counts in unit windows in $\{u_1^\perp, \ldots, u_k^\perp\}$, respectively, and there exists a constant $c \in \mathbb{R}$ such that $\mathbf{E}(\#(\Phi \cap B^{d-1})^2) \leq c$ for all unit $(d-1)-$dimensional balls B^{d-1}.*

Then $\mathcal{R}$ is estimated consistently by the ML estimator in the strong sense, i.e. we have

$$\lim_{k \to \infty} \mathcal{R}_{m,k} = \mathcal{R}$$

almost surely.

For the numerical solution $\hat{\mu}$ of problem (ii) the EM algorithm is proposed in [Kiderlen, 2001].

The second approach to the estimation of $\mathcal{R}$ [Kiderlen, 2001] is based on an idea of [Campi, Haas & Weil, 1994] and it generalizes the 2D approach based on (2.9). Theorem 6 implies the possibility of approximating zonoids by zonotopes in fixed directions $u_1, \ldots, u_k$. Next we are looking for a zonotope Z which is contained in a polytope

$$Q_k = \bigcap_{i=1}^{k} \{x \in \mathbb{R}^d; \ \langle x, u \rangle \leq h(Z, u_i)\}. \tag{2.24}$$

Q_k need not be a zonotope in dimension $d \geq 3$. Theorem 13 suggests the choice of T_m which should contain the set of orientations of line segments forming the zonotope Z. Then only the lengths of its line segments have to be determined. Using $Z_m = \sum_{i=1}^{m} \alpha_j [-v_j, v_j]$ we get a linear program

minimize : $\sum_{i=1}^{k} (h(Z, u_i) - \sum_{j=1}^{m} \alpha_j \mid \langle v_j, u_i \rangle \mid)$,

subject to : $\sum_{j=1}^{m} \alpha_j \mid \langle v_j, u_i \rangle \mid \leq h(Z, u_i), \ i = 1, \ldots, k,$
$\alpha_j \geq 0, \quad j = 1, \ldots, m.$

It can be derived from Theorem 13 that there exists a solution of this linear program with objective function value 0, which yields the desired zonotope and, by optimization theory, at most k of $\alpha_j > 0$. However, the substitution of $\hat{\eta}_i$ for $h(Z, u_i)$ is dangerous in this case because the values of $\hat{\eta}_i$ substantially lower then $\mathbf{E}\eta_i$ (their presence cannot be excluded) can produce an estimate $\mathcal{R}_m = 0$

with a positive probability. Consequently, it is recommended to replace $\hat{\eta}_i$ by their arithmetic averages obtained by independent replicated sampling. Using a numerical optimization procedure to the solution of linear program (LP) the estimator of the rose of directions is obtained and a consistency theorem analogous to Theorem 14 can be formulated, see [Kiderlen, 2001], where also both estimators (EM and LP) are compared. It is concluded that for a smaller sample size the maximum likelihood estimator is slightly better while for larger sample sizes the linear programming should be preferred because the slightly worse performance of the LP estimator is well compensated by its being less time consuming.

3.2.7   Estimation of 3D fibre anisotropy; computer simulation

A 3D analogy of the arc and polygonal test systems for the anisotropy estimation in $\mathbb{R}^2$ are polyhedral probes. In this subsection, the situation frequently used in practice is examined in detail, namely that only a single realization of the fibre process is available. Then the assumptions of the Theorem 14 are not satisfied because of correlated intersection counts η_i. Three isotropic fibre processes (edges of various Voronoi tessellations) were examined by means of cubic and octahedral probes and the distribution of the Prohorov distance was estimated in [Hlawiczková, 2001]. Its variance decreases with the growing number of probe faces (similarly as with the number of random testing planes in [Kiderlen, 2001]) and increases with a growing local inhomogeneity of the process as characterized e.g. by the distribution of the tessellation cell volume (compare with the 2D results in [V. Beneš et al, 2001]. For a more detailed study [Hlawiczková, Ponížil & Saxl, 2001], again the processes of Voronoi cell edges have been selected. They represent a continuous passage from a pronounced anisotropy of linear and planar types to the complete isotropy. Beside these processes with diffuse roses of directions, also three processes with atomic roses have been theoretically considered for the comparison.

The characteristics of the examined fibre processes are as follows:

i. The monoclinic point lattice $\mathcal{H}_0$ with the lattice vectors $|a_1| = |a_2| = |a_3|/q$, $\langle a_1, a_2 \rangle = 0.5|a_1|^2$ generates the *isohedral tiling* $\mathcal{T}'_0$ by regular hexagonal prisms with the four-valent base edges (the relative weights of their three orientations are $1/(3 + 2q)$) of lengths b and three-valent vertical edges (the relative weight of their orientation is $2q/(3+2q)$) of length qb. The edge process $\Phi_0(q)$ with atomic measures was examined in three particular cases: $q = 0.2$ (thin plates producing nearly planar anisotropy), 10 (long rods producing nearly linear anisotropy) and 1 (intermediate case).

ii. Let ξ_x be i.i.d. random vectors with the Gaussian $N(0, \Xi^2)$ distribution, $\Xi^2 = a^2\mathcal{I}$, $\mathcal{I}$ is a unit matrix and $x \in \mathcal{H}_0$ denotes the lattice points. The

displaced lattice or the *Bookstein model* on $\mathcal{H}_0$ [Stoyan & Stoyan, 1994] is $\mathcal{H}_a = \bigcup_{x \in \mathcal{H}_0}(x + \xi_x)$. The tessellation $\mathcal{T}'_a$ generated by $\mathcal{H}_a$ is a *normal* random tessellation with three-valent edges and several its characteristics are discontinuous at $a = 0_+$ (for details see [Hlawiczková, Donížil & Saxl, 2001]). The edge process $\Phi_a(q)$ with a diffuse anisotropy measure was examined for $a = 0.005, 0.2, 0.5, 2$ (in the units of the nearest neighbour distance in $\mathcal{H}_0$) at the values of q chosen above for $\Phi_0(q)$. For high a, $\mathcal{T}'_a$ approaches the stationary Poisson-Voronoi tessellation and Φ_a is isotropic for an arbitrary q.
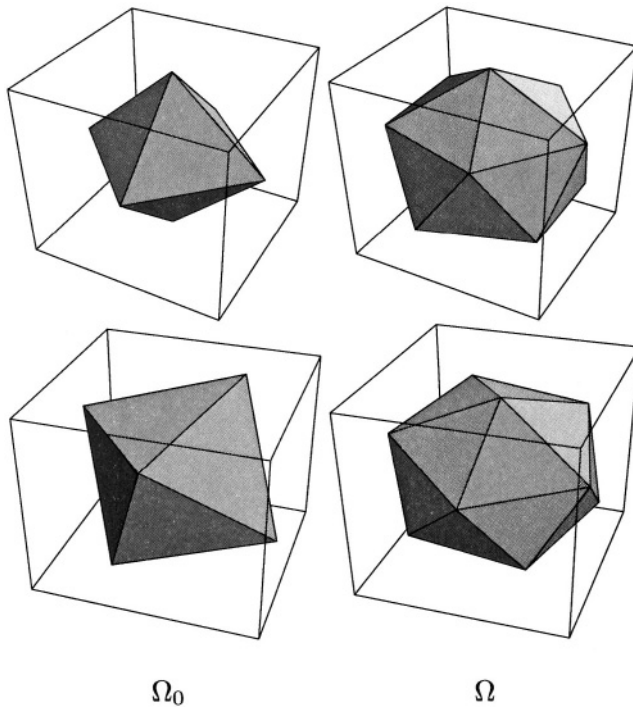
Figure 11. Enlarged probes in the Ω_0 and Ω orientations; the embedding cubes show the mutual orientations of the probes and of the tessellated cube but not its true size.

The tessellations have been constructed in a unit cube by means of the incremental method with the nearest neighbour algorithm [Okabe, Boots & Sugihara, 1977]. The number of process realizations was between 500 and 1000. Centrally symmetric polyhedral probes (icosahedron, octahedron, dodecahedron and cube; the results for the first two of them only are shown in what follows) of the same surface area ($A = 0.8617$) have been placed in the centre of the tessellated unit cube - Fig. 11. In order to suppress a possible positional bias between the tessellation and the probes, each realization was randomly shifted as a whole with respect to the cube centre by a random vector η with
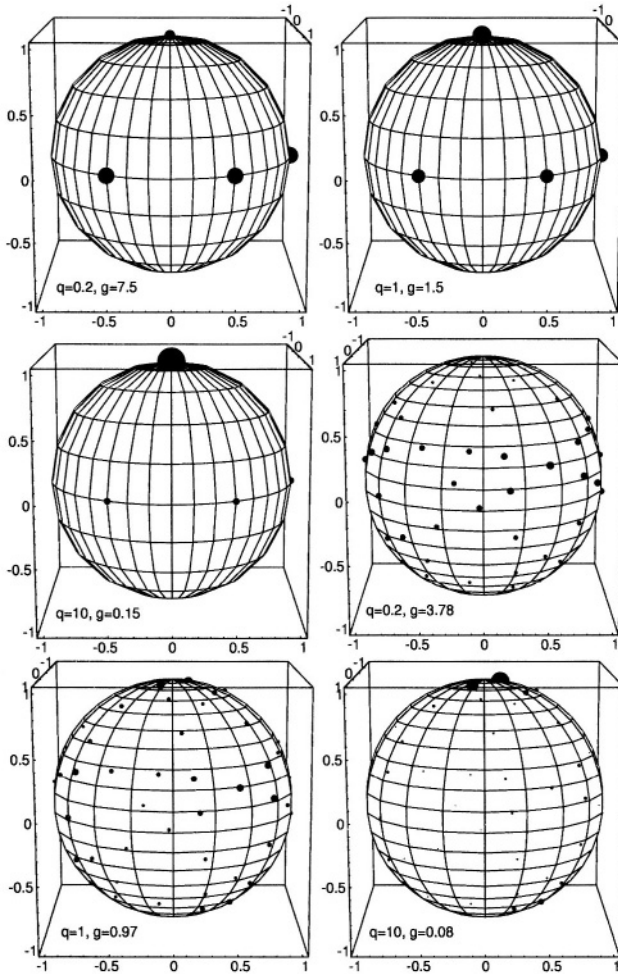
Figure 12. The true discrete roses of directions $\mathcal{R}$ (orientations and weights) and the corresponding factors g (see below, Eq. (2.25)) for Φ_0 fibre processes (upper row), the calculated estimates $\mathcal{R}_{45}$ and g values for the icosahedral probe in the Ω orientation (lower row) as obtained by the EM algorithm.

the Gaussian $N(0, \Xi^2)$ distribution, $\Xi^2 = \tau^2 \mathcal{I}$ and the value of τ was comparable with the lattice constants of $\mathcal{H}_0$.

Two orientations of the probes were examined, namely Ω_0 (all octahedron diagonals parallel to the coordinate axes, one icosahedral diagonal perpendicular to the $\{x, y\}$−plane and two icosahedral edges parallel with the x−axis) and Ω obtained by rotations from Ω_0 (octahedron rotated by Euler angles $(\phi, \psi, \chi) = (\pi/7, \pi/3, \pi/2)$, icosahedron rotated by $(\pi/7, \pi/9, \pi/2)$ - see

Fig. 11. Their size and the intensities of tessellations λ were chosen in such a way that the expected total number of intersections per the whole probe $\mathbf{E}N$ was approximately constant ($\mathbf{E}N = 1840$) in all considered cases and the edge effects were considerably suppressed by confining the examination to the central part of the unit cube.
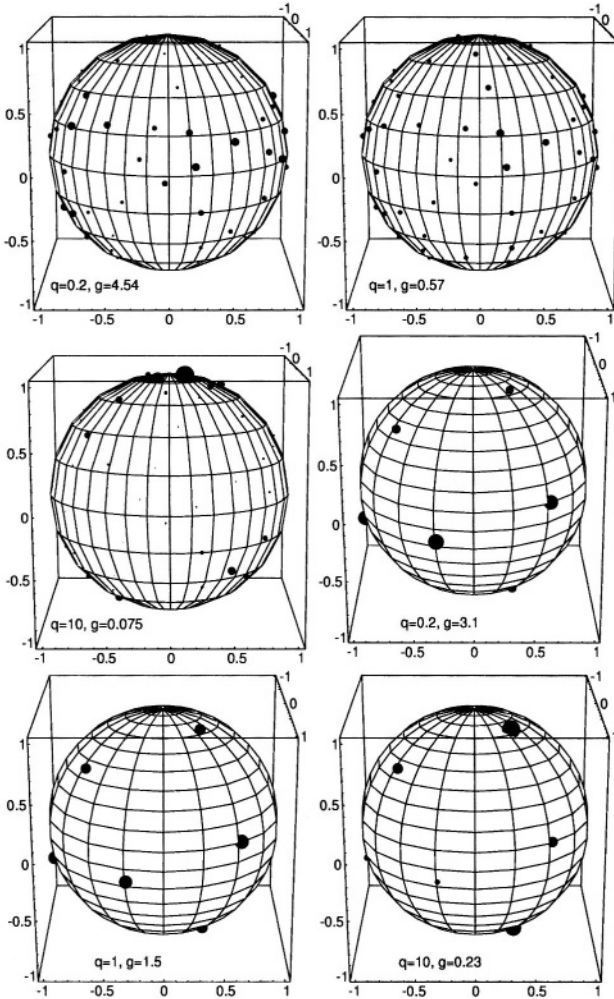
Figure 13. The discrete roses of directions $\mathcal{R}_m$ and g values estimating the fibre processes $\Phi_{0.005}$ at various values of q by means of icosahedral ($\mathcal{R}_{45}$, upper row) and octahedral ($\mathcal{R}_6$, lower row) probes in the Ω orientations. Note the considerably weaker performance octahedral probe.

The rose of directions $\mathcal{R}_m$ approximating the rose $\mathcal{R}$ of the examined fibre process Φ is estimated by the ML procedure described above and the weights α_i are found by the iterative EM algorithm.

The atomic measures $\mathcal{R}$ of Φ_0 are shown and compared with the positions and weights of the estimate $\mathcal{R}_{45}$ as calculated for the icosahedral probe in Fig. 12, circle areas are proportional to the weights α_i and their total area is 1% of the projected sphere area). It is clearly seen that the description of the true atomic measures by $\mathcal{R}_{45}$ is rather unsatisfactory; discrete measures concentrated in the polar region at $q = 0.2$ and in the equatorial strip at $q = 10$ are clearly underestimated. Moreover, the atomic measures in the equatorial plane at $q = 0.2$ are approximated by a broad layer of many weaker atoms. The result would be perhaps better for another probe orientation.

The estimation is more successful in the case of $\Phi_{0.005}$ processes – Fig. 13 – where the diffuse planar anisotropy is reflected much better even when the lack of equatorial directions in the case of linear anisotropy in the estimate by the icosahedral probe is again surprising. The estimation improves considerably when Φ_a approaches isotropy.

The effect of the probe orientation with respect to the fibre orientations may be quite substantial, in particular when the number of probe faces is small. It is shown in [Hlawiczková, Poní žil & Saxl, 2001] that cubic and octahedral probes in the orientation Ω_0 are completely "blind" to the changes in anisotropy of $\Phi_{0.005}$ and the estimated roses of directions $\mathcal{R}_3$ and $\mathcal{R}_6$ are identical for $q = 0.2, 1, 10$. Consequently, if there is no preliminary knowledge of the type of the examined anisotropy, the combination of several probe orientations is always unavoidable.

Frequently, a simple numerical characteristic of the degree of anisotropy is required in practice. If there is some preliminary knowledge concerning the type of the examined anisotropy as in the examined case (linear anisotropy in the direction of z-axis), a suitable numerical factor describing the measure arrangement and strength can be the ratio

$$g = \sum_{\{i|v_i \in S_e\}} \alpha_i / \sum_{\{j|v_j \in S_c\}} \alpha_j. \tag{2.25}$$

where $S_e \in S^2$ is the equatorial strip of area 2π and S_c its complement in S^2. For Φ_0, $g = 3/(2q)$, hence g is high in the case of a quasi-planar anisotropy and low if the linear anisotropy prevails. $g \approx 1$ when approaching the isotropic case. The estimated values of g for Φ_0 and $\Phi_{0.005}$ are given in Fig's 12 and 13. For details see [Hlawiczková, Poní žil & Saxl, 2001].

3.2.8 Approach to the isotropy, Prohorov distance

Prohorov distance was used as a characteristic of the estimation quality of the rose of direction $\mathcal{R}$ in the above citied papers [V. Beneš et al, 2001], [Hlawiczková, 2001], [Kiderlen, 2001]. In [Hlawiczková, Ponížil & Saxl, 2001], a different goal is followed by means of its estimation, namely the approach to the isotropy of the examined $\Phi_a(q)$ with growing standard deviation a of lattice point shifts. It will be described by the decrease of the Prohorov distance between $\mathcal{R}_6$ as estimated by the octahedral probe and uniform rose of directions $\mathcal{U} = 1/4\pi$ with growing a. The estimates of the corresponding pdf's of $r(\mathcal{R}_m, \mathcal{R})$ are shown in Fig. 14 (Epanechnikov kernel estimator with the band width $h = 0.02$ was used).
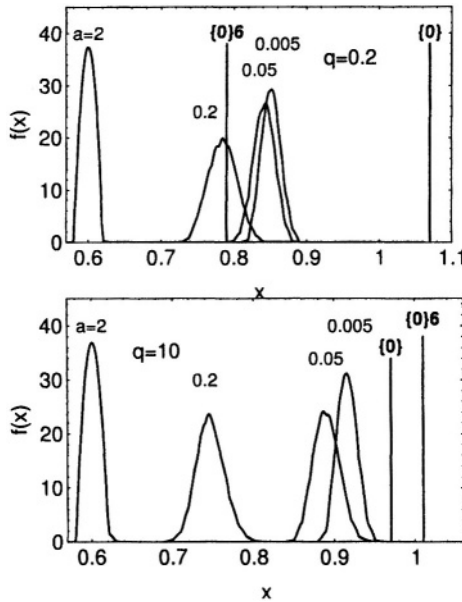
Figure 14. The probability density functions of the Prohorov distance $r(\mathcal{R}_6, \mathcal{U})$ for Φ_a as determined by the octahedral probe.

The approach of all $\Phi_a(q)$ to an isotropic fibre process with increasing standard deviation a is clearly documented by pdf's of the corresponding $r(\mathcal{R}_6, \mathcal{U})$; they shift to smaller values (slowly at $a \leq 0.1$) and coincide at $a = 2$ – Fig. 14. The standard deviations of distance distributions are comparable as the local inhomogeneity of the examined tessellations is similar. The Prohorov distances are rather high as the approximation of a quasi-linear and quasi-planar anisotropies is difficult with a generally oriented probe. A similar result presents the consideration of the g factor (see Fig. 6 in [Hlawiczková, Ponížil

& Saxl, 2001]), namely nearly constant values in the interval $a \in [0.005, 0.1]$ and then a quick approach to the isotropic value $g = 1$. However, the values of g for the process Φ_0 as calculated from the corresponding $\mathcal{R}_m$ are biased. The correct values are {7.5, 1.5, 0.15} and their estimates by $\mathcal{R}_{45}$ are {3.78, 0.97, 0.08} at $q = 0.2, 1, 10$, resp., (see Fig. 12). The smaller is the number of probe faces the greater is the bias. Note that the negative bias would describe a more pronounced linear anisotropy at $q = 10$, whereas in the remaining two cases would the estimated planar anisotropy be weaker. A further examination should elucidate whether a greater number of probe faces or a combination of several probe orientations would give better and more reliable results.

Conclusion

The problem of the estimation of the rose of directions of fibre and surface processes from the rose of intersections has a long history but it is not yet satisfactorily solved. It is unpleasant that while the basic integral equation has the same form for any dimension, surprisingly the properties of theoretical tools for the solution of this equation differ substantially from the planar to the spatial case.

For analytical methods is this difference not so essential but this confirms the fact that analytical methods are not deep enough to produce reliable solutions. Typically we obtain negative values of densities in the solution. In the stochastic approach are statistical properties of the estimators poor. This concerns even the planar situation and problems increase when dealing with the spatial case.

Convex geometry yields excellent tools for the investigation of the basic integral equation. The analogy between the rose of intersections and the support function of a zonoid is striking. The zonotopes converging to the zonoid corresponding to the desired rose of directions are thus already the desired estimators. Their construction in the plane is simple and we can say that this approach leads to good estimators even for sharp or multimodal anisotropies.

Problems arise when applying the Steiner compact method of estimation of the rose of directions in the space. A natural extension of the planar estimator is not available because of the properties of zonoids and zonotopes in higher dimensions. Still two constructions were suggested based either on linear programming techniques or EM-algorithm for the maximum likelihood estimation.

The Prohorov distance between the true and estimated rose of directions is used as a measure of quality of the estimator. It enables comparison between various methods. Since the estimator is typically a discrete measure (based on observations from a finite set of test line orientations) this distance is not concentrated near zero. Simulation methods are used to verify new estimators

and distribution of the Prohorov distance is plotted. The results presented here are the first systematic trial and there is still a lot of work to be done in order to understand the properties of estimators (especially in the spatial case) properly.

The survey is concentrated on a single complex problem, there are also related problems concerning anisotropy. The anisotropy of spatial distribution of objects is mentioned in the Introduction. A more general Stereological formula derived in Section 1 makes possible the use of a local angular information around the Stereological probe. For surfaces of particles, there is a variant of the rose of normal directions considering only outer normal vectors to the particles. This rose of directions is examined in several papers, e.g. [Rataj, 1996], [Weil, 1997], [Schneider, 2001]. Several authors considered also the anisotropy estimation for thick fibre systems modelled by Boolean models [Molchanov & Stoyan, 1994], [Kärkkäinen, Vedel Jensen & Jeulin, 2001]. These problems, however, lead to different concepts of stochastic geometry and were not aimed to be discussed here.

Acknowledgments

References

V. Beneš and A. M. Gokhale, Planar anisotropy revisited, *Kybernetika,* **36/2,** 149-164, 2000.

V. Beneš, M. Hlawiczková, A. M. Gokhale and G. F. Vander Voort, Anisotropy estimation properties for microstructural models, *Materials Characterization,* **46/2-3**, 93-98, 2001.

S. Campi and D. Haas and W. Weil, Approximation of zonoids by zonotopes in fixed directions, *Comput. Geom.,* **11,** 419-431, 1994.

L. M. Cruz-Orive, H. Hoppeler, O. Mathieu and E. R. Weibel, Stereology analysis of anisotropic structures using directional statistics, *JRSS, Series C,* **34/1**, 14-32, 1985.

H. Digabel, Détermination practique de la rose des directions. Technical report, 15 fascicules de morphologic mathématique appliquée (6), Fontainebleau, 1976.

P. Goodey and W. Weil, Zonoids and Generalizations, In P. M. Gruber and J. M. Wills, editors, *Handbook of Convex Geometry,* 1297-1326, North-Holland, Amsterdam, 1993.

H. Heyer, *Probability Measures on Locally Compact Groups,* Springer, Berlin, 1977.

J. E. Hilliard, Specification and measurement of microstructural anisotropy, *Trans. Metall. Soc. AIME,* **224**, 1201-1211, 1962.

M. Hlawiczková, Estimating fibre process anisotropy, In: *Presented at Conference on Inversion Problems,* Aalborg University, March, 2001.

M. Hlawiczková, P. Ponížil and I. Saxl, Estimating 3D fibre process anisotropy, In V. V. Kluev and N. E. Mastorakis, editors, *Topics in Applied and Theoretical Mathematics and Computer Science,* 214-219, WSEAS Press, 2001.

K. Kanatani, Stereological determination of structural anisotropy, *Int. J. Eng. Sci.,* **22,** 531-546, 1984.

M. Kiderlen, Non-parametric estimation of the directional distribution, *Adv. Appl. Probab. (SGSA),* **33**, 6-24, 2001.

B. A. Mair, M. Rao and J. M. M. Anderson, Positron emission tomography, Borel measures and weak convergence, *Inverse Problems,* **12,** 965-976, 1996.

G. Matheron, *Random Sets and Integral Geometry,* Wiley, New York, 1975.

J. Mecke, Formulas for stationary planar fibre processes III - Intersections with fibre systems, *Math. Oper. Statist., Ser. Statist.,* **12**, 201-210, 1981.

J. Mecke and D. Stoyan, Formulas for stationary planar fibre processes. I. General theory., *Math. Oper. Stat.,* **12**, 267-279, 1980.

J. Mecke and W. Nagel, Stationäre raumliche Faserprozesse und ihre Schnittzahlrozen, *Elektron. Inf. Kybernet.,* **16**, 475-483, 1980.

J. Ohser and F. Mücklich, *Statistical Analysis of Microstructures in Materials Science,* Wiley, New York, 2000.

A. Okabe, B. Boots and K. Sugihara, *Spatial tessellations ,* J.Wiley & Sons, Chichester, 1977.

J. Rataj and I. Saxl, Analysis of planar anisotropy by means of the Steiner compact, *J. Appl. Probab.,* **26**, 490-502, 1989.

R. Schneider, *Convex Bodies: the Brunn–Minkowski Theory,* Cambridge University Press, Cambridge, 1993.

R. Schneider and W. Weil, Zonoids and related topics. In P. M. Gruber and J. M. Wills editors, *Convexity and its Applications,* 296-317, Birkhäuser, Basel, 1983.

D. Stoyan and V. Beneš, Anisotropy analysis for particle systems, J. *Microscopy,* **164/2**, 159-168, 1991.

D. Stoyan, S. Kendall and J. Mecke, *Stochastic Geometry and Its Applications. 2nd Edit.,* Wiley, New York, 1995.

D. Stoyan and H. Stoyan, *Fractals, Random Shapes and Point Fields,* J. Wiley & Sons, New York, 1994.

J. Rataj and I. Saxl, Estimation of direction distribution of a planar fibre system, *Acta Stereol.,* **11/I**, 631-637, 1992.

I. Molchanov and D. Stoyan, Directional analysis of fibre processes related to Boolean models, *Metrika,* **41**, 183-199, 1994.

S. Kärkkäinen and E.B. Vedel Jensen and D. Jeulin, *On the orientation analysis of Boolean fibres from digital images. Res. Rep. 15.,* Laboratory for Computational Statistics, University of Aarhus, 2001.

J. Rataj, Estimation of oriented direction distribution of a planar body, *Adv. Appl. Prob.,* **28**, 294-304, 1996.

R. Schneider, On the mean normal measures of a particle process, *Adv. Appl. Prob.,* **33,** 25-38, 2001.

W. Weil, The mean normal distribution of stationary random sets and particle process, In D. Jeulin, editor, *Advances in Theory and Applications of Random Sets,* 21-33, Singapore, World Scientific, 1997.

APPROXIMATIONS FOR MULTIPLE SCAN STATISTICS

Jie Chen
Department of Computing Services University of Massachusetts, Boston
Boston, MA 02125, USA

Joseph Glaz
Department of Statistics University of Connecticut
Storrs, CT 06269, USA
glaz@uconnvm.uconn.edu

Abstract   In this article Poisson-type and compound Poisson approximations are discussed for a multiple scan statistic for Binomial and Poisson data in one and two dimensions. Numerical results are presented to evaluate the performance of these approximations. Direction for future research and open problems are also stated.

4.1   Introduction

In this article we discuss Poisson-type and compound Poisson approximations for multiple scan statistics for independent and identically distributed (iid) integer valued random variables from a binomial or a Poisson distribution. Both one dimensional and two dimensional scan statistics are considered. The multiple scan statistics are discussed both for the unconditional case and for the case when the total number of the observed events is known (conditional case). One dimensional multiple scan statistics for iid Bernoulli random have been discussed in Chen and Glaz (1996) and Balakrishnan and Koutras (2002).

One dimensional multiple scan statistics for continuous data are discussed in Glaz, Naus and Wallenstein (2001, Ch. 17). Approximations for multidimensional multiple scan statistics are discussed in Barbour and Mansson (2000) and Mansson (1999a, 1999b and 2000).

This article is organized as follows. In Section 2, we present Poisson-type and compound Poisson approximations for the one dimensional multiple scan statistic, both conditional and unconditional case. We also derived Bonferroni-

type inequalities for the binomial model in the unconditional case. Since these inequalities have not performed well, we have not derived them for other cases. In Section 3, we present Poisson-type and compound Poisson approximations for the two dimensional multiple scan statistic, both conditional and unconditional case. In Section 4 numerical results are discussed for the approximations derived in this article. Concluding remarks are presented in Section 5.

4.2   The One Dimensional Case

Let $X_1, \cdots, X_N$ be iid nonnegative integer valued random variables following a binomial or a Poisson distribution. First we consider the unconditional case, when the total number of events $\sum_{i=1}^{N} X_i$ is unknown. For integers $1 \leq j \leq N - m + 1$ and $k \geq 2$ let

$$A_j = (X_j + \cdots + X_{j+m-1} \geq k) \tag{2.1}$$

and

$$I_j = \begin{cases} 1, & \text{if } A_j \text{ occurs} \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

For integers $2 \leq m < N$ define a discrete *scan statistic.*

$$S_m = S_m(N) = \max\{X_i + \cdots + X_{i+m-1}; 1 \leq i \leq N - m + 1\}. \tag{2.3}$$

We say that a scan statistic of size k has been observed if S_m exceeds the value $k - 1$. Approximations for the distribution of S_m, applications and references are given in Glaz, Naus and Wallenstein (2001, Ch. 13). In this article we are interested in approximations for the distribution of a *multiple scan statistic* of size k defined as:

$$\xi = \sum_{j=1}^{N-m+1} I_j, \tag{2.4}$$

where I_j is given in Equation (2.2). For $1 \leq n \leq N - m + 1$, a Poisson,approximation for $P(\xi \geq n)$ is given by

$$P(\xi \geq n) \approx 1 - \sum_{i=0}^{n-1} \frac{e^{-\lambda}\lambda^i}{i!}, \tag{2.5}$$

where

$$\lambda = E(\xi) = (N - m + 1)P(A_1).$$

Since the events $A_j, 1 \leq j \leq N - m + 1$, tend to clump, the Poisson approximation given in Equation (2.5) performed poorly for $P(\xi \geq 1)$ (Chen and Glaz 1999). Following the approach in Chen and Glaz (1997) the following Poisson-type approximation will be investigated. For $1 \leq j \leq N - m$, let

$$I_j^* = \begin{cases} 1, & \text{if } A_j \cap A_{j+1}^c \cap \ldots \cap A_{\min(j+m-1,N-m+1)}^c \text{ occurs} \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

and

$$I^*_{N-m+1} = \begin{cases} 1, & \text{if } A_{N-m+1} \text{ occurs} \\ 0, & \text{otherwise.} \end{cases}$$

By defining the indicators I^*_j we are not allowing the events A_j to clump. A Poisson-type approximation for $P(\xi \geq n)$ to be examined here is given by

$$P(\xi \geq n) \approx 1 - \sum_{i=0}^{n-1} \frac{e^{-\lambda^*}\lambda^{*i}}{i!}, \tag{2.7}$$

where

$$\lambda^* = E\left(\sum_{j=1}^{N-m+1} I^*_j\right) = 1 - q_{2m-2} + (N - 2m + 2)(q_{2m-2} - q_{2m-1}) \tag{2.8}$$

and for $m \leq j \leq N$

$$q_j = P(S_m(j) \leq k - 1) = P\left(\bigcap_{j=1}^{j-m+1} A^c_j\right). \tag{2.9}$$

Numerical results for this Poisson-type approximation are given in Section 4, Tables 1 and 2.

A compound Poisson approximation for ξ based on the approach in Roos (1993, Lemma 3.3.4) is given by:

$$P(\xi \geq n) \approx 1 - \sum_{j=0}^{n-1}\left(\sum_{\beta_1+2\beta_2+....+(2m-1)\beta_{2m-1}=j} \prod_{i=1}^{2m-1} \frac{\lambda_i^{\beta_i}}{\beta_i!}\right)\exp\left(-\sum_{i=1}^{2m-1}\lambda_i\right), \tag{2.10}$$

where β_i are non-negative integers,

$$\lambda_i = (N - m + 1)\pi(1 - p)^2 p^{i-1}, \ i = 1, \cdots, m - 1,$$

$$\lambda_i = \frac{(N - m + 1)\pi}{i}\left[2(1 - p)p^{i-1} + (2m - i - 2)(1 - p)^2 p^{i-1}\right],$$

$$m \leq i \leq 2m - 2,$$

$$\lambda_{2m-1} = \frac{(N - m + 1)(1 - q_m)p^{2m-2}}{2m - 1},$$

and

$$\pi = P(I_1 = 1) = 1 - q_m,$$

$$p = P(I_1 = 1, I_2 = 2)/P(I_1 = 1) = (1 - 2q_m + q_{m+1})/(1 - q_m).$$

Numerical results for this compound Poisson approximation are presented in Section 4, Tables 1 and 2.

We now discuss Bonferroni-type inequalities for $P(\xi \geq n)$. Consider the events A_j, $1 \leq j \leq N - m + 1$, defined in Equation (2.1). Let ν_{N-m+1} be the number of $A'_j s$ that have occurred. Then

$$P(\nu_{N-m+1} \geq n) = P(\xi \geq n).$$

For $1 \leq j \leq 3$, let

$$s_j = \sum_{1 \leq i_1 < \dots < i_j \leq N-m+1} P\left(\bigcap_{t=1}^{j} A_{i_t}\right).$$

It follows from Galambos and Simonelli (1996, pages 118-119) that:

$$
\begin{aligned}
P(\xi \geq n) \geq &\ \frac{6s_3 - 2(2i + N - m - 1)s_2 + i(i + 2N - 2m + 1)s_1}{(N - m - n + 2)(i + 1 - n)(i + 2 - n)} \\
&+ \frac{(n - 1)[(2i - n)(N - m + 1) + 2(N - m + 1)]}{(N - m - n + 2)(i + 1 - n)(i + 2 - n)} \\
&+ \frac{(n - 1)[n^2 - 2in - 3n + i^2 + 3i + 2]}{(N - m - n + 2)(i + 1 - n)(i + 2 - n)},
\end{aligned}
$$

for $1 \leq n \leq N - m - 1$, $n \leq i \leq N - m - 1$ and

$$P(\xi \geq n) \leq \frac{i(2n + i - 1)s_1 - 2(n + 2i - 2)s_2 + 6s_3}{i(i + 1)n},$$

for $1 \leq n \leq N - m - 1$, $m + 1 \leq i \leq n - 1$. Numerical results for these Bonferroni-type inequalities are presented in Section 4, Table 1.

We now discuss approximations for a multiple scan statistic when the total number of events $\sum_{i=1}^{N} X_i = a$ is known. For $1 \leq j \leq N - m + 1$ and $k \geq 2$, let

$$A_j^* = (X_j + \dots + X_{j+m-1} \geq k \mid \sum_{i=1}^{N} X_i = a)$$

and

$$I_j(a) = \begin{cases} 1, & \text{if } A^*_j \text{ occurs} \\ 0, & \text{otherwise.} \end{cases}$$

The conditional multiple scan statistic is defined as:

$$\xi(a) = \sum_{j=1}^{N-m+1} I_j(a).$$

For $m \le j \le N$, set

$$q_j(a) = P(S_m(j) \le k - 1 \mid \sum_{j=1}^{N-m+1} X_j = a) = P\left(\bigcap_{j=1}^{j-m+1} A_j^{*c}\right).$$

A Poisson-type approximations for $P(\xi(a) \ge n)$ can be obtained from Equations (2.7) and (2.8) by replacing the terms q_{2m-2} and q_{2m-1} with $q_{2m-2}(a)$ and $q_{2m-1}(a)$, respectively. Let

$$\lambda_i(a) = (N - m + 1)\pi(a)(1 - p(a))^2 p(a)^{i-1},$$

for $1 \le i \le m - 1$,

$$\lambda_i(a) = \frac{(N - m + 1)\pi}{i}\left[2(1 - p(a))p(a)^{i-1} + (2m - i - 2)(1 - p(a))^2 p(a)^{i-1}\right],$$

for $m \le i \le 2m - 2$,

$$\lambda_{2m-1}(a) = \frac{(N - m + 1)(1 - q_m(a))p(a)^{2m-2}}{2m - 1},$$

and

$$\pi(a) = P(I_1(a) = 1) = 1 - q_m(a),$$

$$p(a) = \frac{P(I_1(a) = 1, I_2(a) = 1)}{P(I_1(a) = 1)} = \frac{1 - 2q_m(a) + q_{m+1}(a)}{(1 - q_m(a))}.$$

A compound Poisson approximation for $P(\xi(a) \ge n)$ can be obtained from Equation (2.10) by replacing the terms λ_i, π and p with $\lambda_i(a), \pi(a)$ and $p(a)$, respectively. Numerical results for these approximations are presented in Section 4, Tables 3 and 4.

4.3 The Two Dimensional Case

Let $X_{i,j}, i = 1, \cdots, N_1$ and $j = 1, \cdots, N_2$, be iid nonnegative integer valued random variables with a binomial or a Poisson distribution. Let

$$Y_{i_1,i_2} = \sum_{j=i_2}^{i_2+m_2-1} \sum_{i=i_1}^{i_1+m_1-1} X_{i,j}, \tag{3.1}$$

where $1 \le i_1 \le N_1 - m_1 + 1$ and $1 \le i_2 \le N_2 - m_2 + 1$. The two-dimension scan statistic is defined as:

$$S_{m_1,m_2} = max\left\{Y_{i_1,i_2}; 1 \le i_1 \le N_1 - m_1 + 1, 1 \le i_2 \le N_2 - m_2 + 1\right\}. \tag{3.2}$$

Approximations for the distribution of S_{m_1,m_2}, applications and references are given in Glaz, Naus and Wallenstein (2001, Ch. 16). For simplicity we assume here that $N_1 = N_2 = N$ and $m_1 = m_2 = m$. For $1 \leq i_1, i_2 \leq N - m + 1$ define the events

$$A_{i_1,i_2} = \left(\sum_{i=i_1}^{i_1+m-1} \sum_{j=i_2}^{i_2+m-1} X_{i,j} \geq k \right). \tag{3.3}$$

Let $\Gamma = \{(i_1, i_2); 1 \leq i_1 \leq N - m + 1, 1 \leq i_2 \leq N - m + 1\}$, denote the index set of a collection of the integer valued random variables $\{I_\alpha; \alpha \in \Gamma\}$, where

$$I_\alpha = \begin{cases} 1, & \text{if } Y_\alpha \geq k \\ 0, & \text{otherwise..} \end{cases} \tag{3.4}$$

We are interested in approximating the distribution of a two dimensional multiple scan statistic

$$\xi_{m,m} = \sum_{\alpha \in \Gamma} I_\alpha.$$

For $1 \leq j \leq m + 1$, let

$$q'_{m+j-1} = P\left(\bigcap_{i=1}^{j} A^c_{1,i} \right). \tag{3.5}$$

Under quite general conditions the distribution of $\sum_{\alpha \in \Gamma} I_\alpha$ converges to the Poisson distribution with mean λ_1, where

$$\lambda_1 = E(\sum_{\alpha \in \Gamma} I_\alpha) = (N - m + 1)^2 (1 - q'_m). \tag{3.6}$$

(Darling and Waterman 1986). This Poisson.approximation for the special case of $k = m^2$ has been discussed in Barbour, Chryssaphinou and Roos (1995), Koutras, Papadopoulos and Papastavridis (1993), and Roos (1994). The Poisson approximation is not expected to perform well when $k < m^2$, since the events $\{(S(\alpha) \geq k); \alpha \in \Gamma\}$ tend to clump. Employing a local declumping approach, Chen and Glaz (1996) derived a more accurate Poisson-type approximation:

$$P(\xi_{m,m} \geq 1) = P(S_{m,m} \geq k) \approx 1 - \exp(-\lambda_1^*), \tag{3.7}$$

where

$$\lambda_1^* = 1 - q'_{2m-2} + (N - 2m + 2)(N - m + 1)(q'_{2m-2} - q'_{2m-1}). \tag{3.8}$$

In this article we investigate the performance of the following Poisson-type approximation:

$$P(\xi_{m,m} \geq n) \approx 1 - \sum_{i=0}^{n-1} \frac{e^{-\lambda_1^*} \lambda_1^{*i}}{i!}. \tag{3.9}$$

A compound Poisson approximation for $P(\xi_{m,m} \geq n)$ presented below is based on Roos (1993 and 1994):

$$P(\xi \geq n) \approx 1 - \sum_{j=0}^{n-1} \left(\sum_{\beta_1+2\beta_2+3\beta_3+4\beta_4+5\beta_5=j} \prod_{i=1}^{5} \frac{\lambda_{1i}^{\beta_i}}{\beta_i!} \right) \exp(-\sum_{i=1}^{5} \lambda_{1i}),$$
(3.10)

where for $1 \leq i \leq 5$:

$$\lambda_{1i} = \frac{1}{i}(1 - q'_m) \left\{ 4\pi_{1,i} + 4(N - m - 1)\pi_{2,i} + (N - m + 1)^2\pi_{3,i} \right\},$$

$$\pi_{1,i} = P\left\{ I_{1,2} + I_{2,1} = i - 1|I_{1,1} = 1 \right\},$$

$$\pi_{2,i} = P\left\{ I_{1,1} + I_{2,2} + I_{3,1} = i - 1|I_{2,1} = 1 \right\},$$

and

$$\pi_{3,i} = P\left\{ I_{1,2} + I_{2,1} + I_{2,3} + I_{3,2} = i - 1|I_{2,2} = 1 \right\}.$$

In Section 4, Tables 5 and 6, we present numerical results for these Poisson-type and compound Poisson approximations.

We know present approximations for the multiple scan statistic given that the total number of observed events $\sum_{j=1}^{N} \sum_{i=1}^{N} X_{i,j} = a$ is known. For $1 \leq i_1, i_2 \leq N - m + 1$ define the events

$$A^*_{i_1,i_2} = \left(\sum_{i=i_1}^{i_1+m-1} \sum_{j=i_2}^{i_2+m-1} X_{i,j} \geq k \,\middle|\, \sum_{j=1}^{N} \sum_{i=1}^{N} X_{i,j} = a \right).$$
(3.11)

Let

$$I_\alpha(a) = \begin{cases} 1, & \text{if } A^*_{i_1,i_2} \text{ occurs} \\ 0, & \text{otherwise}, \end{cases}$$

where $\alpha \in \Gamma$. The conditional multiple scan statistic considered here is given by

$$\xi_{m,m}(a) = \sum_{\alpha \in \Gamma} I_\alpha(a).$$

For $1 \leq j \leq m + 1$, let

$$q'_{m+j-1}(a) = P\left(\bigcap_{i=1}^{j} A^{*c}_{1,i} \right).$$

A Poisson-type approximation for $P(\xi_{m,m}(a) \geq n)$ is obtained from Equations (3.9) and (3.8) by replacing the terms q'_{2m-1} and q'_{2m-2} with $q'_{2m-1}(a)$ and $q'_{2m-2}(a)$, respectively.

For $1 \leq i \leq 5$ let

$$\lambda_{1i}(a) = \frac{1}{i}(1 - q'_m(a))\{4\pi_{1,i}(a) + 4(N - m - 1)\pi_{2,i}(a)$$
$$+ (N - m + 1)^2\pi_{3,i}(a)\},$$

$$\pi_{1,i}(a) = P\{I_{1,2}(a) + I_{2,1}(a) = i - 1 | I_{1,1}(a) = 1\},$$

$$\pi_{2,i}(a) = P\{I_{1,1} + I_{2,2} + I_{3,1} = i - 1 | I_{2,1}(a) = 1\},$$

and

$$\pi_{3,i}(a) = P\{I_{1,2}(a) + I_{2,1}(a) + I_{2,3}(a) + I_{3,2}(a) = i - 1 | I_{2,2}(a) = 1\}.$$

A compound Poisson approximation for $P(\xi_{m,m}(a) \geq n)$ is obtained from Equation (3.10) by replacing the terms λ_{1i}, $\pi_{1,i}$, $\pi_{2,i}$ and $\pi_{3,i}$ with $\lambda_{1i}(a)$, $\pi_{1,i}(a)$, $\pi_{2,i}(a)$ and $\pi_{3,i}(a)$, respectively. Numerical results for these approximations are given in Section 4, Tables 7 and 8.

4.4 Numerical Results

In this section we present numerical results for approximations and inequalities for the multiple scan statistics discussed in this article. In Tables 1-8 the improved Poisson-type approximations are denoted by *ImPoi,* while the compound Poisson approximations are denoted by *ComPoi.* The Bonferroni-type inequalities considered in this article have not performed well. Numerical results for these inequalities are presented in Table 1 and they are denoted by *LBound* and *UBound,* respectively. In Tables 1-8, $\widehat{P}(\circ \geq n)$ is an approximation for the tail probability of an appropriate multiple scan statistic based on a simulation with 10,000 trials. In Tables 1-2, Poisson-type and compound Poisson approximations, as well as the Bonferroni-type inequalities, are evaluated using an algorithm discussed in Glaz and Naus (1991). The quantities q_j, $m \leq j \leq 2m - 1$, needed for evaluating Poisson-type and compound Poisson-type approximations for the multiple scan statistic $\xi(a)$, Tables 3-4, are obtained from a simulation with 100,000 trials of sequences of $2m - 1$ iid binomial or Poisson random variables. The quantities q'_j, $m \leq j \leq 2m - 1$, needed for evaluating Poisson-type and compound Poisson-type approximations for the multiple scan statistics $\xi_{m,m}$ and $\xi_{m,m}(a)$, Tables 5-8, are obtained from a simulation with 100,000 trials of sequences of $(2m - 1) \times (2m - 1)$ iid binomial or Poisson random variables.

Table 1. Approximations and Bounds for ξ. Binomial Model.

N	m	p	L	k	n	$\hat{P}(\xi \geq n)$	ImPoi	ComPoi	LBound	UBound
100	10	.05	5	5	1	0.8840	0.7978	0.8973	0.6135	1.0000
					2	0.8371	0.4746	0.8420	0.1768	1.0000
					3	0.7923	0.2163	0.7851	0.1593	1.0000
					4	0.7446	0.0786	0.7277	0.1414	1.0000
					5	0.6970	0.0236	0.6710	0.1232	1.0000
				6	1	0.5972	0.5521	0.6450	0.3818	1.0000
					2	0.5123	0.1923	0.5350	0.0189	1.0000
					3	0.4383	0.0479	0.4409	0.0137	1.0000
					4	0.3659	0.0092	0.3614	0.0093	0.7832
					5	0.3121	0.0014	0.2948	0.0058	0.6472
				7	1	0.2847	0.2738	0.3201	0.1808	1.0000
					2	0.2143	0.0415	0.2259	0.0010	0.5128
					3	0.1605	0.0043	0.1591	0.0005	0.3495
					4	0.1192	0.0003	0.1118	0.0001	0.2679
					5	0.0891	0.0000	0.0784	0.0000	0.2189
	.10		5	10	1	0.4971	0.4650	0.5540	0.3079	1.0000
					2	0.3932	0.1304	0.4240	0.0026	1.0000
					3	0.3073	0.0257	0.3221	0.0016	0.7125
					4	0.2415	0.0039	0.2430	0.0010	0.5399
					5	0.1925	0.0005	0.1823	0.0008	0.4364
				11	1	0.2635	0.2489	0.2947	0.1612	0.8145
					2	0.1821	0.0339	0.1938	0.0004	0.4115
					3	0.1331	0.0032	0.1270	0.0003	0.2772
					4	0.0967	0.0002	0.0830	0.0002	0.2100
					5	0.0680	0.0000	0.0541	0.0001	0.1697
				12	1	0.1110	0.1082	0.1253	0.0702	0.2836
					2	0.0695	0.0061	0.0718	0.0001	0.1433
					3	0.0452	0.0002	0.0411	0.0001	0.0965
					4	0.0295	0.0000	0.0235	0.0000	0.0731
					5	0.0187	0.0000	0.0135	0.0000	0.0470

From Tables 1-8 it is evident that compound Poisson approximations are more accurate than the Poisson-type approximations investigated in this article. The compound Poisson approximations have performed well, especially in the one dimensional case. In the two dimensional case these approximations were not as accurate as one would like them to be. There is a need for further research to derive accurate approximations for multiple scan statistics.

Table 2.   Approximations for ξ. Poisson Model.

N	m	θ	k	n	$\hat{P}(\xi \geq n)$	ImPoi	ComPoi
100	10	.10	3	1	0.7665	0.6928	0.7728
				2	0.7170	0.3303	0.7074
				3	0.6707	0.1163	0.6455
				4	0.6273	0.0321	0.5874
				5	0.5776	0.0072	0.5333
			4	1	0.3520	0.3309	0.3718
				2	0.2954	0.0621	0.2941
				3	0.2469	0.0080	0.2325
				4	0.2024	0.0008	0.1836
				5	0.1633	0.0001	0.1449
			5	1	0.0960	0.0953	0.1048
				2	0.0720	0.0047	0.0720
				3	0.0538	0.0002	0.0494
				4	0.0399	0.0000	0.0340
				5	0.0282	0.0000	0.0234
			6	1	0.0208	0.0195	0.0209
				2	0.0137	0.0002	0.0128
				3	0.0099	0.0000	0.0079
				4	0.0061	0.0000	0.0048
				5	0.0044	0.0000	0.0030

4.5   Concluding Remarks

From the numerical results presented in this article it is evident that further research has to be conducted in the area of multiple scan statistics, especially in the multi-dimensional case. Approximations for the distribution of multiple scan statistics for continuous data also presents many challenging problems. Modeling and statistical inference of spatial data is one the most active research areas in probability and statistics. It has many applications in science and technology including: anthropology, archaeology, astronomy, ecology, environmental science, epidemiology, geology, image analysis, meteorology, reconnaissance and urban and regional planning. The use of spatial scan statistics in two or higher dimensional regions have been discussed among others in Wallenstein, Gould and Kleinman (1989), Priebe, Olson, Healy (1997), Kulldorff (1999), Chan and Lai (2000), Siegmund and Yakir (2000), Glaz, Naus and Wallenstein (2001), Priebe and Chen (2001), Priebe, Naiman and Cope (2001). Multiple scan statistics are of great importance in this area of research as well. More work is needed to be done for deriving accurate approximations

for the distribution of multiple scan statistics used in statistical inference for spatial data.

Table 3. Approximations for $\xi(a)$. Binomial Model.

N	m	L	a	k	n	$\hat{P}(\xi(a) \geq n)$	ImPoi	ComPoi
100	5	5	25	4	1	0.6654	0.6887	0.6982
					2	0.5207	0.3254	0.5165
					3	0.3742	0.1134	0.3668
					4	0.2477	0.0310	0.2512
					5	0.0901	0.0069	0.1352
				5	1	0.2702	0.2605	0.2635
					2	0.1493	0.0373	0.1352
					3	0.0730	0.0037	0.0689
					4	0.0302	0.0003	0.0351
					5	0.0100	0.0000	0.0179
	20	5	25	6	1	0.6268	0.5167	0.5903
					2	0.5080	0.1653	0.4672
					3	0.4042	0.0375	0.3672
					4	0.3215	0.0066	0.2868
					5	0.2448	0.0009	0.2228
				7	1	0.2351	0.2121	0.2349
					2	0.1590	0.0243	0.1588
					3	0.1084	0.0019	0.1072
					4	0.0721	0.0001	0.0723
					5	0.0449	0.0000	0.0487
	10	5	50	10	1	0.4845	0.3988	0.4636
					2	0.3544	0.0929	0.3391
					3	0.2565	0.0151	0.2466
					4	0.1870	0.0019	0.1784
					5	0.1347	0.0002	0.1285
				11	1	0.2068	0.1841	0.2064
					2	0.1314	0.0181	0.1279
					3	0.0847	0.0012	0.0790
					4	0.0555	0.0001	0.0488
					5	0.0326	0.0000	0.0301

Table 4. Approximations for $\xi(a)$. Poisson Model.

N	m	a	k	n	$\hat{P}(\xi(a) \geq n)$	ImPoi	ComPoi
100	5	5	2	1	0.6330	0.5155	0.6179
				2	0.5319	0.1644	0.5211
				3	0.4234	0.0372	0.3855
				4	0.3044	0.0065	0.2772
				5	0.1780	0.0009	0.2231
	10	5	2	1	0.9116	0.6779	0.7204
				2	0.8755	0.3130	0.6664
				3	0.8281	0.1063	0.6157
				4	0.7795	0.0282	0.5683
				5	0.7254	0.0061	0.5241
			3	1	0.2023	0.1697	0.1728
				2	0.1683	0.0153	0.1403
				3	0.1367	0.0009	0.1141
				4	0.1030	0.0000	0.0928
				5	0.0784	0.0000	0.0756
	10	10	4	1	0.3123	0.2270	0.2805
				2	0.2490	0.0280	0.2207
				3	0.1947	0.0023	0.1736
				4	0.1469	0.0001	0.1366
				5	0.1051	0.0000	0.1075
			5	1	0.0472	0.0441	0.0498
				2	0.0350	0.0010	0.0346
				3	0.0232	0.0000	0.0240
				4	0.0152	0.0000	0.0167
				5	0.0098	0.0000	0.0116

Table 5. Approximations $\xi_{m,m}$. Binomial Mode

N	m	L	p	k	n	$\hat{P}(\xi_{m,m} \geq n)$	ImPoi	ComPoi
25	5	5	.05	15	1	0.2461	0.2954	0.3008
					2	0.1447	0.0487	0.1646
					3	0.0890	0.0055	0.0552
					4	0.0591	0.0005	0.0233
					5	0.0375	0.0000	0.0094
				16	1	0.1027	0.1232	0.1319
					2	0.0527	0.0079	0.0710
					3	0.0301	0.0003	0.0235
					4	0.0196	0.0000	0.0083
					5	0.0121	0.0000	0.0029
	10	5	.05	39	1	0.1846	0.1976	0.2645
					2	0.1373	0.0210	0.2033
					3	0.1098	0.0015	0.1470
					4	0.0917	0.0001	0.0897
					5	0.0767	0.0000	0.0595
100	10	5	.05	46	1	0.1585	0.2449	0.2334
					2	0.0998	0.0328	0.1012
					3	0.0698	0.0030	0.0455
					4	0.0507	0.0002	0.0168
					5	0.0360	0.0000	0.0075
				47	1	0.0876	0.1351	0.1022
					2	0.0526	0.0096	0.0715
					3	0.0325	0.0005	0.0559
					4	0.0231	0.0000	0.0261
					5	0.0163	0.0000	0.0132

Table 6. Approximations for $\xi_{m,m}$. Poisson Model.

N	m	θ	k	n	$\hat{P}(\xi_{m,m} \geq n)$	ImPoi	ComPoi
25	5	.25	14	1	0.5274	0.6526	0.6546
				2	0.3854	0.2853	0.5013
				3	0.2910	0.0911	0.3440
				4	0.2258	0.0227	0.2247
				5	0.1744	0.0046	0.1672
			15	1	0.2964	0.3698	0.3754
				2	0.1825	0.0788	0.2511
				3	0.1173	0.0116	0.0930
				4	0.0809	0.0013	0.0451
				5	0.0577	0.0001	0.0184
			16	1	0.1448	0.1708	0.1418
				2	0.0805	0.0155	0.0654
				3	0.0445	0.0010	0.0273
				4	0.0277	0.0000	0.0103
				5	0.0181	0.0000	0.0019
100	5	.50	29	1	0.2307	0.2715	0.3012
				2	0.1064	0.0407	0.1174
				3	0.0546	0.0042	0.0285
				4	0.0297	0.0003	0.0187
				5	0.0157	0.0000	0.0031
			30	1	0.1061	0.1246	0.0968
				2	0.0402	0.0081	0.0233
				3	0.0154	0.0004	0.0021
				4	0.0071	0.0000	0.0015

Table 7. Approximations for $\xi_{m,m}(a)$. Binomial Model..

N	m	L	a	k	n	$\hat{P}(\xi_{m,m}(a) \geq n)$	ImPoi	ComPoi
25	5	5	25	4	1	0.8659	0.9263	0.9438
					2	0.7946	0.7341	0.9223
					3	0.7108	0.4835	0.8386
					4	0.6347	0.2656	0.7707
					5	0.5542	0.1236	0.6763
	5			5	1	0.3182	0.3525	0.2906
					2	0.2323	0.0711	0.2099
					3	0.1522	0.0099	0.1462
					4	0.1137	0.0011	0.0690
					5	0.0813	0.0001	0.0298
				6	1	0.0598	0.1337	0.0468
					2	0.0337	0.0016	0.0305
					3	0.0179	0.0000	0.0144
					4	0.0113	0.0000	0.0055
					5	0.0075	0.0000	0.0001
25	5	5	50	7	1	0.4098	0.4441	0.4958
					2	0.2817	0.1177	0.3788
					3	0.1899	0.0219	0.2155
					4	0.1363	0.0031	0.1255
					5	0.0949	0.0004	0.0666
				8	1	0.1172	0.1284	0.1234
					2	0.0645	0.0086	0.0545
					3	0.0347	0.0004	0.0195
					4	0.0200	0.0000	0.0076
					5	0.0124	0.0000	0.0025
100	5	5	100	5	1	0.8695	0.9442	0.8716
					2	0.7895	0.7830	0.8078
					3	0.6944	0.5506	0.6992
					4	0.6117	0.3270	0.5867
					5	0.5263	0.1658	0.4759
				6	1	0.3052	0.3595	0.3210
					2	0.2041	0.0741	0.2286
					3	0.1290	0.0106	0.1037
					4	0.0894	0.0012	0.0541
					5	0.0585	0.0001	0.0247

Table 8.   Approximations for $\xi_{m,m}(a)$. Poisson Model.

N	m	a	k	n	$\hat{P}(\xi_{m,m}(a) \geq n)$	ImPoi	ComPoi
25	5	300	22	1	0.5937	0.6718	0.6586
				2	0.4241	0.3061	0.4853
				3	0.3052	0.1024	0.3168
				4	0.2231	0.0268	0.2009
				5	0.1635	0.0057	0.1186
			23	1	0.3810	0.4291	0.3916
				2	0.2148	0.1091	0.2336
				3	0.0000	0.0194	0.1076
				4	0.0000	0.0026	0.0519
				5	0.0000	0.0003	0.0221
			24	1	0.2314	0.2427	0.2785
				2	0.1122	0.0322	0.1667
				3	0.0000	0.0029	0.0474
				4	0.0000	0.0002	0.0206
				5	0.0000	0.0000	0.0065
100	10	1000	24	1	0.2197	0.3119	0.2868
				2	0.1021	0.0547	0.1660
				3	0.0000	0.0066	0.0364
				4	0.0000	0.0006	0.0147
				5	0.0000	0.0000	0.0028
			25	1	0.1467	0.0722	0.1776
				2	0.0628	0.0027	0.1745
				3	0.0000	0.0001	0.1715
				4	0.0000	0.0000	0.0826
				5	0.0000	0.0000	0.0164

References

Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications,* John Wiley & Sons, Inc., New York.

Barbour, A.D., Chryssaphinou, O. and Roos, M. (1995). Compound Poisson approximation in systems reliability. *Naval Research Logistics* **43,** 251-264.

Barbour, A.D. and Mansson, M. (2000). Compound Poisson approximation and the clustering of random points. *Advances in Applied Probability* **32,** 19-38.

Chan, H. P., and Lai, T. L. (2000). Saddlepoint approximations for Markov random walks and nonlinear boundary crossing probabilities for scan statistics. Technical Report, Stanford University.

Chen, J. and J. Glaz (1996). Two dimensional discrete scan statistics. textitStatistics & Probability Letters **31,** 59-68.

Chen, J. and Glaz, J. (1999). Approximations for discrete scan statistics on the circle. *Statistics & Probability Letters* **44,** 167-176.

Darling, R.W.R. and Waterman, M.S. (1986). Extreme value distribution for the largest cube in a random lattice. *SIAM Journal on Applied Mathematics* **46,** 118-132.

Galambos, J. and Simonelli, I. (1996). *Bonferroni-type Inequalities with Applications,* Springer-Verlag, New York.

Glaz, J. and Balakrishnan, N. (Eds.) (1999). *Scan Statistics and Applications.* Birkhauser, Boston.

Glaz, J. and Naus, J. (1983). Multiple clusters on the line. *Communications in Statistics - Theory and Methods,* **12,** 1961-1986.

Glaz, J. and Naus, J. I. (1991). Tight bounds and approximations for scan statistic probabilities for discrete data, *Annals of Applied Probability,* **1,** 306-318.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics,* New York: Springer-Verlag.

Huffer, F. W. and Lin, C. T. (1997). Computing the exact distribution of the extremes of sums of consecutive spacings. *Computational Statistics and Data Analysis* **26,** 117-132.

Huffer, F. and Lin, C. T. (1997). Approximating the distribution of the scan statistic using moments of the number of clumps. *Journal of the American Statistical Association* **92 ,** 1466-1475.

Koutras, M. V., Papadopoulos, G. K. and Papastavridis, S. G. (1993). Reliability of 2-dimensionaI consecutive-k-out-of n: F systems, *IEEE Transaction on Reliability,* **R-42,** 658-661.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics A - Theory and Methods* **26,** 1481-1496.

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations and applications. In Glaz, J. and Balakrishnan, N., eds. *Scan Statistics and Applications.* Birkhauser, Boston, 303-322.

Kulldorff, M., Fang, Z. and Walsh, S. J. (2003). A tree-based scan statistic for data base disease surveillance. *Biometrics* (to appear).

Mansson, M. (1999a). Poisson approximation in connection with clustering of random points. *Annals of Applied Probability* **9,** 465-492.

Mansson, M. (1999b). On Poisson approximation for continuous multiple scan statistics in two dimensions. In Glaz, J. and Balakrishnan, N., eds. *Scan Statistics and Applications,* Birkhauser, Boston.

Mansson, M. (2000). On compound Poisson approximation for sequence matching. *Combinatorics, Probability, and Computing* **9,** 529-548.

Nagarwalla, N. (1996). A scan statistic with a variable window. *Statistics in Medicine* **15,** 845-50.

Naiman, D. Q. and Priebe, C. (2001) Computing scan statistic p-values using importance sampling, with applications to genetics and medical image analysis. *Journal of Computational and Graphical Statistics* **10,** 296-328.

Patil, G. P., Bishop, J., Myers, W. L., Vraney, R. and Wardrop, D. (2003). Detection and delineation of critical areas using echelons and spatial scan statistics with synoptic data. *Environmental and Ecological Statistics: Special Issue on Multiscale Advanced Raster Map Analysis System* (to appear).

Priebe, C. E., Olson, T. and Healy, Jr., D. M. (1997). A spatial scan statistic for stochastic scan partitions. *Journal of the American Statistical Association* **92,** 1476-1484.

Priebe, C. E. and Chen, D. (2001). Spatial scan density estimates. *Technometrics* **43,** 73.

Priebe, C. E., Naiman, D. Q. and Cope, L. M. (2001) Importance sampling for spatial scan analysis: computing scan statistic p-values for marked point processes. *Computational Statistics and Data Analysis* **35,** 475-485.

Roos, M. (1993). *Stein-Chen Method for compound Poisson Approximation.* Ph. D. dissertation, University of Zurich, Zurich.

Roos, M. (1994). Stein's method for compound Poisson approximation, *Annals of Applied Probability,* **4,** 1177-1187.

Siegmund, D. and Yakir, B. (2000). Tail probabilities for the null distribution of scanning statistics, *Bernoulli* **6**, 191-213.

Su, X. and Wallenstein, S.(1999). New approximations for the distribution of the r-scan statistic, *Statistics and Probability Letters* **46,** 411-19.

Wallenstein, S., Gould, M. S. and Kleinman, M. (1989). Use of the scan statistic to detect time-space clustering. *American Journal of Epidemiology* **130,** 1057-1064.

KRAWTCHOUK POLYNOMIALS AND KRAWTCHOUK MATRICES

Philip Feinsilver
Department of Mathematics Southern Illinois University Carbondale, IL 62901

Jerzy Kocik
Department of Mathematics Southern Illinois University Carbondale, IL 62901
jkocik@siu.edu

Abstract   Krawtchouk matrices have as entries values of the Krawtchouk polynomials for nonnegative integer arguments. We show how they arise as condensed Sylvester-Hadamard matrices via a binary shuffling function. The underlying symmetric tensor algebra is then presented.

To advertise the breadth and depth of the field of Krawtchouk polynomials / matrices through connections with various parts of mathematics, some topics that are being developed into a Krawtchouk Encyclopedia are listed in the concluding section. Interested folks are encouraged to visit the website

    http://chanoir.math.siu.edu/Kravchuk/index.html

which is currently in a state of development.

5.1   What are Krawtchouk matrices

Of Sylvester-Hadamard matrices and Krawtchouk matrices, the latter are less familiar, hence we start with them.

DEFINITION   1 The Nth-*order Krawtchouk matrix* $K^{(N)}$ is an $(N+1)\times(N+1)$ matrix, the entries of which are determined by the expansion:

$$(1+v)^{N-j}\,(1-v)^{j} = \sum_{i=0}^{N} v^{i} K_{ij}^{(N)} \tag{1.1}$$

Thus, the polynomial $G(v) = (1+v)^{N-j}\,(1-v)^{j}$ is the *generating function* for the row entries of the jth column of $K^{(N)}$. Expanding gives the explicit

values of the matrix entries:

$$K_{ij}^{(N)} = \sum_k (-1)^k \binom{j}{k} \binom{N-j}{i-k}.$$

where matrix indices run from 0 to N.

Here are the Krawtchouk matrices of order zero, one, and two:

$$K^{(0)} = \begin{bmatrix} 1 \end{bmatrix} \qquad K^{(1)} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad K^{(2)} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \\ 1 & -1 & 1 \end{bmatrix}$$

The reader is invited to see more examples in Table 1 of the Appendix.

The columns of Krawtchouk matrices may be considered *generalized binomial coefficients.* The rows define Krawtchouk *polynomials:* for fixed order N, the ith *Krawtchouk polynomial* takes its corresponding values from the ith row:

$$k_i(j, N) = K_{ij}^{(N)} \tag{1.2}$$

One can easily show that $k_i(j, N)$ can be given as a polynomial of degree i in the variable j. For fixed N, one has a system of $N+1$ polynomials orthogonal with respect to the symmetric binomial distribution.

A fundamental fact is that the square of a Krawtchouk matrix is proportional to the identity matrix.

$$(K^{(N)})^2 = 2^N \cdot I$$

This property allows one to define a Fourier-like *Krawtchouk transform* on integer vectors. For more properties we refer the reader to [Feinsilver, 2001]. In the present article, we focus on Krawtchouk matrices as they arise from corresponding Sylvester-Hadamard matrices. More structure is revealed through consideration of symmetric tensor algebra.

Symmetric Krawtchouk matrices.   When each column of a Krawtchouk matrix is multiplied by the corresponding binomial coefficient, the matrix becomes symmetric. In other words, define the **symmetric Krawtchouk matrix** as

$$S^{(N)} = K^{(N)} B^{(N)}$$

where $B^{(N)}$ denotes the $(N + 1) \times (N + 1)$ diagonal matrix with binomial coefficients, $B_{ii}^{(N)} = \binom{N}{i}$, as its non-zero entries.

Example. For $N = 3$, we have

$$S^{(3)} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & 1 & -1 & -3 \\ 3 & -1 & -1 & 3 \\ 1 & -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 3 & 1 \\ 3 & 3 & -3 & -3 \\ 3 & -3 & -3 & 3 \\ 1 & -3 & 3 & -1 \end{bmatrix}$$

Some symmetric Krawtchouk matrices are displayed in Table 2 of the Appendix. A study of the spectral properties of the symmetric Krawtchouk matrices was initiated in work with Fitzgerald [Feinsilver & Fitzgerald, 1996].
Background note. Krawtchouk's polynomials Krawtchouk polynomial were introduced by Mikhail Krawtchouk in the late 20's [Krawtchouk, 1929; Krawtchouk, 1933]. The idea of setting them in a matrix form appeared in the 1985 work of N. Bose [Bose, 1985] on digital filtering in the context of the Cayley transform on the complex plane. For some further development of this idea, see [Feinsilver, 2001].

The Krawtchouk polynomials play an important role in many areas of mathematics. Here are some examples:

- **Harmonic analysis.** As orthogonal polynomials, they appear in the classic work by Szëgo [Sze, 1959]. They have been studied from the point of view of harmonic analysis and special functions, e.g., in work of Dunkl [Dunkl, 1976; Dunkl, 1974]. Krawtchouk polynomials maybe viewed as the discrete version of Hermite polynomials (see, e.g., [Atakishiyev, 1997]).

- **Statistics.** Among the statistics literature we note particularly Eagleson [Eagelson, 1969] and Vere-Jones [Vere-Jones, 1971].

- **Combinatorics and coding theory.** Krawtchouk polynomials are essential in MacWilliams' theorem on weight enumerators [Levenstein, 1995; MacWilliams & Sloane, 1977], and are a fundamental example in association schemes [Delsarte, 1972; Delsarte, 1973; Delsarte, 1973a].

- **Probability theory.** In the context of the classical symmetric random walk, it is recognized that Krawtchouk's polynomials are elementary symmetric functions in variables taking values ±1. It turns out that the generating function (1.1) is a martingale in the parameter N [Feinsilver & Schott, 1991].

■ **Quantum theory.** Krawtchouk matrices interpreted as operators give
rise to two new interpretations in the context of both classical and quan-
tum random walks [Feinsilver, 2001]. The significance of the latter in-
terpretation lies at the basis of quantum computing.

Let us proceed to show the relationship between Krawtchouk matrices and
Sylvester-Hadamard matrices.

5.2   Krawtchouk matrices from Hadamard matrices

Taking the Kronecker (tensor) product of the initial matrix

$$H = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

with itself N times defines the family of *Sylvester-Hadamard matrices.*

(For a review of Hadamard matrices, see Yarlagadda and Hershey [Rao &
Hershey, 1997].)

NOTATION 2 *Denote the Sylvester-Hadamard matrices, tensor (Kronecker)
powers of the fundamental matrix H, by*

$$H^{(N)} = H^{\otimes N} = \underbrace{H \otimes H \otimes \cdots \otimes H}_{N\ times}$$

The first three Sylvester-Hadamard matrices are $H^{(1)}, H^{(2)}$ and $H^{(3)}$ given
by:

where, to emphasize the patterns, we use ● for 1 and ○ for -1. See Table 3 of
the Appendix for these matrices up to order 5.

For $N = 1$, the Hadamard matrix coincides with the Krawtchouk matrix:
$H^{(1)} = K^{(1)}$. Now we wish to see how the two classes of matrices are re-
lated for higher N. It turns out that appropriately contracting (condensing)

Hadamard-Sylvester matrices yields corresponding symmetric Krawtchouk matrices.

The problem is that the tensor products disperse the columns and rows that have to be summed up to do the contraction. We need to identify the right sets of indices.

DEFINITION 3 Define the *binary shuffling function* as the function

$$w\colon \mathbf{N} \to \mathbf{N}$$

giving the "binary weight" of an integer. That is, let $n = \sum_k d_k 2^k$ be the binary expansion of the number n. Then $w(n) = \sum_k d_k$, the number of ones in the representation.

Notice that, as sets,

$$w(\{0, 1, \ldots, 2^N - 1\}) = \{0, 1, \ldots, N\}$$

Here are the first 16 values of w listed for the integers running from 0 through $2^4 - 1 = 15$:

$$0 \quad 1 \quad 1 \quad 2 \quad 1 \quad 2 \quad 2 \quad 3 \quad 1 \quad 2 \quad 2 \quad 3 \quad 2 \quad 3 \quad 3 \quad 4$$

The shuffling function can be defined recursively. Set $w(0) = 0$ and

$$w(2^N + k) = w(k) + 1 \qquad (2.1)$$

for $0 \le k < 2^N$. One can thus create the sequence of values of the shuffling function by starting with 0 and then appending to the current string of values a copy of itself with values increased by 1:

$$0 \;\to\; 01 \;\to\; 0112 \;\to\; 01121223 \;\to\; \ldots$$

Now we can state the result;

THEOREM 4 *Symmetric Krawtchouk matrices are reductions of Hadamard matrices as follows:*

$$S_{ij}^{(N)} = \sum_{\substack{w(a)=i \\ w(b)=j}} H_{ab}^{(N)}$$

Example. Let us see the transformation for $H^{(4)} \to S^{(4)}$ (recall that $\bullet$ stands for 1, and $\circ$ for -1). Applying the binary shuffling function to $H^{(4)}$, mark the rows and columns accordingly:

	0	1	1	2	1	2	2	3	1	2	2	3	2	3	3	4
0	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
1	●	○	●	○	●	○	●	○	●	○	●	○	●	○	●	○
1	●	●	○	○	●	●	○	○	●	●	○	○	●	●	○	○
2	●	○	○	●	●	○	○	●	●	○	○	●	●	○	○	●
1	●	●	●	●	○	○	○	○	●	●	●	●	○	○	○	○
2	●	○	●	○	○	●	○	●	●	○	●	○	○	●	○	●
2	●	●	○	○	○	○	●	●	●	●	○	○	○	○	●	●
3	●	○	○	●	○	●	●	○	●	○	○	●	○	●	●	○
1	●	●	●	●	●	●	●	●	○	○	○	○	○	○	○	○
2	●	○	●	○	●	○	●	○	○	●	○	●	○	●	○	●
2	●	●	○	○	●	●	○	○	○	○	●	●	○	○	●	●
3	●	○	○	●	●	○	○	●	○	●	●	○	○	●	●	○
2	●	●	●	●	○	○	○	○	○	○	○	○	●	●	●	●
3	●	○	●	○	○	●	○	●	○	●	○	●	●	○	●	○
3	●	●	○	○	○	○	●	●	○	○	●	●	●	●	○	○
4	●	○	○	●	○	●	●	○	○	●	●	○	●	○	○	●

The contraction is performed by summing columns with the same index, then summing rows in similar fashion. One checks from the given matrix that indeed this procedure gives the symmetric Krawtchouk matrix $S^{(4)}$:

$$S^{(4)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \\ \left(\begin{array}{ccccc} 1 & 4 & 6 & 4 & 1 \\ 4 & 8 & 0 & -8 & -4 \\ 6 & 0 & -12 & 0 & 6 \\ 4 & -8 & 0 & 8 & -4 \\ 1 & -4 & 6 & -4 & 1 \end{array}\right) \end{array}$$

Now we give a method for transforming the Nth (symmetric) Krawtchouk matrix into the $N + 1^{\text{st}}$.

DEFINITION 5 The *square contraction* $r(M)$ of a $2n \times 2n$ matrix M_{ab}, $1 \le a, b \le 2n$, is the $(n + 1) \times (n + 1)$ matrix with entries

$$(rM)_{ij} = \sum_{\substack{a=2i,\, 2i+1 \\ b=2j,\, 2j+1}} M_{ab}$$

$0 \le i, j \le n$, where the values of M_{ab} with a or b outside of the range $(1, \ldots, 2n)$ are taken as zero.

THEOREM 6 *Symmetric Krawtchouk matrices satisfy:*

$$S^{(N+1)} = r(S^{(N)} \otimes H)$$

with $S^{(1)} = H$.

Example. Start with symmetric Krawtchouk matrix of order 2:

$$S^{(2)} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 0 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

Take the tensor product with H:

$$S^{(2)} \otimes H = \begin{bmatrix} 1 & 1 & 2 & 2 & 1 & 1 \\ 1 & -1 & 2 & -2 & 1 & -1 \\ 2 & 2 & 0 & 0 & -2 & -2 \\ 2 & -2 & 0 & 0 & -2 & 2 \\ 1 & 1 & -2 & -2 & 1 & 1 \\ 1 & -1 & -2 & 2 & 1 & -1 \end{bmatrix}$$

surround with zeros and contract:

$$r(S^{(2)} \otimes H) = r \left[\begin{array}{cc|cc|cc|cc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 & 2 & 1 & 1 & 0 \\ \hline 0 & 1 & -1 & 2 & -2 & 1 & -1 & 0 \\ 0 & 2 & 2 & 0 & 0 & -2 & -2 & 0 \\ \hline 0 & 2 & -2 & 0 & 0 & -2 & 2 & 0 \\ 0 & 1 & 1 & -2 & -2 & 1 & 1 & 0 \\ \hline 0 & 1 & -1 & -2 & 2 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

$$= \begin{bmatrix} 1 & 3 & 3 & 1 \\ 3 & 3 & -3 & -3 \\ 3 & -3 & -3 & 3 \\ 1 & -3 & 3 & -1 \end{bmatrix}$$

COROLLARY 7 *Krawtchouk matrices satisfy:*

$$K^{(N+1)} = r(K^{(N)} B^{(N)} \otimes H)(B^{(N+1)})^{-1}$$

where B is the diagonal binomial matrix.

Note that starting with the 2×2 identity matrix, I, set $I^{(1)} = I$, $I^{(N+1)} = r(I^{(N)} \otimes I)$. Then, in fact, $I^{(N)} = B^{(N)}$.

Next, we present the algebraic structure underlying these remarkable properties.

5.3　　Krawtchouk matrices and symmetric tensors

Given a d-**dimensional** vector space V over $\mathbf{R}$, one may construct a d^N-**dimensional** space $V^{\otimes N}$, the N-fold tensor product of V, and, as well, a $\binom{d+N-1}{N}$-**dimensional** *symmetric tensor space* $V^{\otimes_s N}$. There is a natural map

$$\mathbf{symm}\colon V^{\otimes N} \longrightarrow V^{\otimes_s N}$$

which, for homogeneous tensors, is defined via

$$\mathbf{symm}\,(v \otimes w \otimes \ldots) = \text{symmetrization of } (v \otimes w \otimes \ldots)$$

For computational purposes, it is convenient to use the fact that the symmetric tensor space of order N of a d-**dimensional** vector space is isomorphic to the space of polynomials in d variables homogeneous of degree N.

Let $\{e_1, e_2, \ldots e_d\}$ be a basis of V. Map e_i to x_i, replace tensor products by multiplication of the variables, and extend by linearity. For example,

$$2e_1 \otimes e_2 + 3e_2 \otimes e_1 - 7e_3 \otimes e_2 \longrightarrow 5x_1 x_2 - 7x_2 x_3$$

thus identifying basis (elementary) tensors in $V^{\otimes N}$ that are equivalent under any permutation.

This map induces a map on certain linear operators. Suppose $A \in \mathrm{End}(V)$ is a linear transformation on V. This induces a linear transformation $A_N = A^{\otimes N} \in \mathrm{End}(V^{\otimes N})$ defined on elementary tensors by:

$$A_N(v \otimes w \otimes \ldots) = A(v) \otimes A(w) \otimes \ldots$$

Similarly, a linear operator on the symmetric tensor spaces is induced so that the following diagram commutes:

$$
\begin{array}{ccc}
V^{\otimes N} & \xrightarrow{\;A_N\;} & V^{\otimes N} \\
{\scriptstyle \mathbf{symm}}\downarrow & & \downarrow{\scriptstyle \mathbf{symm}} \\
V^{\otimes_s N} & \xrightarrow{\;\overline{A}_N\;} & V^{\otimes_s N}
\end{array}
$$

This can be understood by examining the action on polynomials. We call $\overline{A}_N$ the *symmetric representation of A in degree N.* Denote the matrix elements of $\overline{A}_N$ by $\overline{A}_{mn}$. If A has matrix entries A_{ij}, let

$$y_i = \sum_j A_{ij}\, x^j$$

It is convenient to label variables with indices from 0 to $\delta = d - 1$. Then the matrix elements of the symmetric representation are defined by the expansion:

$$y_0^{m_0} \cdots y_\delta^{m_\delta} = \sum_n \overline{A}_{mn} \, x_0^{n_0} \cdots x_\delta^{n_\delta}$$

with multi-indices m and n homogeneous of degree N.

Mapping to the symmetric representation is an algebra homomorphism, i.e.,

$$\overline{AB} = \overline{A}\,\overline{B}$$

Explicitly, in matrix notation, $\overline{(AB)}_{mn} = \sum_r (\overline{A})_{mr} (\overline{B})_{rn}$.

Now we are ready to state our result

PROPOSITION 4 *For each $N > 0$, the symmetric representation of the Nth Sylvester-Hadamard matrix equals the transposed N^{th} Krawtchouk matrix:*

$$(\overline{H}_N)_{ij} = K_{ji}^{(N)} .$$

Proof. Writing (x, y) for (x_0, x_1), we have in degree N for the kth component:

$$(x + y)^{N-k}(x - y)^k = \sum_l \overline{H}_{kl} \, x^{N-l} y^l$$

Substituting $x = 1$ yields the generating function (1.1) for the Krawtchouk matrices with the coefficient of y^l equal to $K_{lk}^{(N)}$. Thus the result. ∎

Insight into these correspondences can be gained by splitting the fundamental Hadamard matrix $H \,(= K^{(1)})$ into two special symmetric 2×2 operators:

$$F = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

so that

$$H = F + G = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

One can readily check that

$$F^2 = G^2 = I$$
$$FH = HG \quad \text{and} \quad GH = HF \tag{3.1}$$

The first of the second pair of equations may be viewed as the spectral decomposition of F and we can interpret the Hadamard matrix as diagonalizing F into G. Taking transposes gives the second equation of (3.1).

Now we proceed to the interpretation leading to a symmetric Bernoulli quantum random walk ([Feinsilver, 2001]). For this interpretation, the Hilbert space of states is represented by the Nth tensor power of the original 2-dimensional space V, that is, by the 2^N-dimensional Hilbert space $V^{\otimes N}$. Define the following linear operator on $V^{\otimes N}$:

$$
\begin{aligned}
X_F \;=\; & F \otimes I \otimes \cdots \otimes I \\
& + I \otimes F \otimes I \otimes \cdots \otimes I \\
& + \ldots \\
& + I \otimes I \otimes \cdots \otimes F \\
=\; & f_1 + f_2 + \ldots + f_i + \ldots + f_N
\end{aligned}
$$

each term describing a "flip" at the ith position (cf. [Hess, 1954; Siegert, 1949]). Analogously, we define:

$$
\begin{aligned}
X_G \;=\; & G \otimes I \otimes \cdots \otimes I \\
& + I \otimes G \otimes I \otimes \cdots \otimes I \\
& + \ldots \\
& + I \otimes I \otimes \cdots \otimes G \\
=\; & g_1 + g_2 + \ldots + g_i + \ldots + g_N
\end{aligned}
$$

From equations (3.1) we see that our X-operators intertwine the Sylvester-Hadamard matrices:

$$
X_F H^{(N)} = H^{(N)} X_G \qquad \text{and} \qquad X_G H^{(N)} = H^{(N)} X_F
$$

Since products are preserved in the process of passing to the symmetric tensor space, we get

$$
\overline{X}_F \overline{H}_N = \overline{H}_N \overline{X}_G \qquad \text{and} \qquad \overline{X}_G \overline{H}_N = \overline{H}_N \overline{X}_F \qquad (3.2)
$$

the bars indicating the corresponding induced maps.

We have seen in Proposition 4 how to calculate $\overline{H}_N$ from the action of H on polynomials in degree N. For symmetric tensors we have the components in degree N, namely $x^{N-k} y^k$, for $0 \le k \le N$, where for convenience we write x for x_0 and y for x_1. Now consider the generating function for the elementary symmetric functions in the quantum variables f_j. This is the N-fold tensor power

$$
\mathcal{F}_N(t) = (I + tF)^{\otimes N} = I^{\otimes N} + t\, X_F + \cdots
$$

noting that the coefficient of t is X_F. Similarly, define

$$\mathcal{G}_N(t) = (I + tG)^{\otimes N} = I^{\otimes N} + tX_G + \cdots$$

From $(I + tF)H = H(I + tG)$ we have

$$\mathcal{F}_N H^{(N)} = H^{(N)}\mathcal{G}_N \qquad \text{and} \qquad \overline{\mathcal{F}}_N \overline{H}_N = \overline{H}_N \overline{\mathcal{G}}_N$$

The difficulty is to calculate the action on the symmetric tensors for operators, such as X_F, that are not pure tensor powers. However, from $\mathcal{F}_N(t)$ and $\mathcal{G}_N(t)$ we can recover X_F and X_G via

$$X_F = \left.\frac{d}{dt}\right|_{t=0} (I + tF)^{\otimes N}, \qquad X_G = \left.\frac{d}{dt}\right|_{t=0} (I + tG)^{\otimes N}$$

with corresponding relations for the barred operators. Calculating on polynomials yields the desired results as follows.

$$I + tF = \begin{bmatrix} 1 & t \\ t & 1 \end{bmatrix}, \qquad I + tG = \begin{bmatrix} 1+t & 0 \\ 0 & 1-t \end{bmatrix}$$

In degree N, using x and y as variables, we get the kth component for $\overline{X}_F$ and $\overline{X}_G$ via

$$\left.\frac{d}{dt}\right|_{t=0} (x + ty)^{N-k}(tx + y)^k = (N - k)x^{N-(k+1)}y^{k+1} + kx^{N-(k-1)}y^{k-1}$$

and since $I + tG$ is diagonal,

$$\left.\frac{d}{dt}\right|_{t=0} (1 + t)^{N-k}(1 - t)^k \, x^{N-k}y^k = (N - 2k) \, x^{N-k}y^k .$$

For example, calculations for $N = 4$ result in

$$\overline{X}_F = \begin{bmatrix} 0 & 4 & 0 & 0 & 0 \\ 1 & 0 & 3 & 0 & 0 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & 3 & 0 & 1 \\ 0 & 0 & 0 & 4 & 0 \end{bmatrix} \tag{3.3}$$

$$\overline{X}_G = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & -4 \end{bmatrix} \tag{3.4}$$

$$\overline{H}_4 = \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 1 & 2 & 0 & -2 & -1 \\ 1 & 0 & -2 & 0 & 1 \\ 1 & -2 & 0 & 2 & -1 \\ 1 & -4 & 6 & -4 & 1 \end{bmatrix} \tag{3.5}$$

Since $\overline{X}_G$ is the result of diagonalizing $\overline{X}_F$, we observe that

COROLLARY 8 *The spectrum of $\overline{X}_F$ is N, N − 2,…,2 − N, −N, coinciding with the support of the classical random walk.*

Remark on the shuffling map. Notice that the top row of $(I + tF)^{\otimes N}$ is exactly $t^{w(k)}$, where $w(k)$ is the binary shuffling function of section §5.2. Each time one tensors with $I + tF$, the original top row is reproduced, then concatenated with a replica of itself modified in that each entry picks up a factor of t (compare with equation (2.1)). And, collapsing to the symmetric tensor space, the top row will have entries $\binom{N}{k}t^k$. This follows as well by direct calculation of the 0th component matrix elements in degree N, namely by expanding $(x + ty)^N$.

We continue with some areas where Krawtchouk polynomials/matrices play a rôle, very often not explicitly recognized in the original contexts.

5.4   Ehrenfest urn model

Ehrenfest urn model In order to explain how the apparent irreversibility of the second law of thermodynamics arises from reversible statistical physics, the Ehrenfests introduced a so-called urn model, variations of which have been considered by many authors [Kac, 1947; Karlin & McGregor, 1965; Voit, 1996].

We have an urn with N balls. Each ball can be in two states represented by, say, being lead or gold. At each time $k \in \mathbb{N}$, a ball is drawn at random, changed by a Midas-like touch into the opposite state (gold ↔ lead) and placed back in the urn. The question is of course about the distribution of states — and this leads to Krawtchouk matrices.

Represent the states of the model by vectors in $\mathbb{R}^{n+1}$, namely by the state of k gold balls by

$$\mathbf{v}_k = \begin{bmatrix} 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}^\top$$
$$\underset{k^{\text{th}} \text{ position}}{\uparrow}$$

(4.1)

In the case of, say, $N = 3$, we have 4 states

$$\begin{matrix} \text{0 gold balls} \\ \text{3 lead balls} \end{matrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \begin{matrix} \text{1 gold ball} \\ \text{2 lead balls} \end{matrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \cdots \begin{matrix} \text{3 gold balls} \\ \text{0 lead balls} \end{matrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

It is easy to see that the matrix of elementary state change in this case is

$$
\begin{bmatrix} 0 & \frac{1}{3} & 0 & 0 \\ 1 & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & 1 \\ 0 & 0 & \frac{1}{3} & 0 \end{bmatrix}
= \frac{1}{3} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3 & 0 & 2 & 0 \\ 0 & 2 & 0 & 3 \\ 0 & 0 & 1 & 0 \end{bmatrix}
= \frac{1}{3} A^{(3)} \ ,
$$

and in general, we have the **Kac matrix** with off-diagonals in arithmetic progression 1,2,3, ... descending and ascending, respectively:

$$
A^{(N)} = \begin{bmatrix}
0 & 1 & 0 & 0 & \cdots & 0 & 0 \\
N & 0 & 2 & 0 & \cdots & 0 & 0 \\
0 & N-1 & 0 & 3 & \vdots & 0 & 0 \\
0 & 0 & N-2 & 0 & \ddots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\
0 & 0 & 0 & 0 & \ddots & 0 & N \\
0 & 0 & 0 & 0 & \cdots & 1 & 0
\end{bmatrix}
$$

It turns out that the spectral properties of the Kac matrix involve Krawtchouk matrices, namely, the *collective solution* to the eigenvalue problem $Av = \lambda v$ is

$$
A^{(N)} K^{(N)} = K^{(N)} \Lambda^{(N)}
$$

where $\Lambda^{(N)}$ is the $(N+1) \times (N+1)$ diagonal matrix with entries $\Lambda_{ii}^{(N)} = N - 2i$

$$
\Lambda^{(N)} = \begin{bmatrix}
N & & & & & \\
& N-2 & & & (*) & \\
& & N-4 & & & \\
& & & \ddots & & \\
& (*) & & & 2-N & \\
& & & & & -N
\end{bmatrix}
$$

the $(*)$'s denoting blocks of zeros.

To illustrate, for $N = 3$ we have

$$
\begin{bmatrix} 0 & 1 & 0 & 0 \\ 3 & 0 & 2 & 0 \\ 0 & 2 & 0 & 3 \\ 0 & 0 & 1 & 0 \end{bmatrix}
\begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & 1 & -1 & -3 \\ 3 & -1 & -1 & 3 \\ 1 & -1 & 1 & -1 \end{bmatrix}
$$

$$
= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & 1 & -1 & -3 \\ 3 & -1 & -1 & 3 \\ 1 & -1 & 1 & -1 \end{bmatrix}
\begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}
$$

To see this in general, we note that, cf. equations (3.3–3.5), these are the same operators appearing in the quantum random walk model, namely, we discover that $\Lambda^{(N)} = \overline{X}_G$, $A^{(N)} = \overline{X}_F^\top$. Now, recalling $K^{(N)} = \overline{H}_N^\top$, taking transposes in equation (3.2) yields

$$A^{(N)} K^{(N)} = K^{(N)} \Lambda^{(N)} \qquad \text{and} \qquad K^{(N)} A^{(N)} = \Lambda^{(N)} K^{(N)}$$

which is the spectral analysis of $A^{(N)}$ from both the left and the right. Thus, e.g., the columns of the Krawtchouk matrix are eigenvectors of the Ehrenfest model with N balls where the kth column $\mathbf{v}_k := (K._k)$ has corresponding eigenvalue $\lambda_k = (N - 2k)/N$.

Remarks

1 Clearly, the Ehrenfest urn problem can be expressed in other terms. For instance, it can be reformulated as a random walk on an N-dimensional cube. Suppose an ant walks on the cube, choosing at random an edge to progress to the next vertex. Represent the states by vectors in $Z = \mathbb{Z}_2 \times \cdots \times \mathbb{Z}_2$, N factors. The equivalence of the two problems comes via the correspondence of states

$$Z \ni [\, a_1 \, a_2 \ldots a_N \,] \longrightarrow \mathbf{v}_w \in \mathbb{R}^{N+1}$$

where $w = \sum a_i$ is the weight of the vector calculated in $\mathbb{N}$, see (4.1).

2 The urn model in the appropriate limit as $N \to \infty$ leads to a diffusion model on the line, the discrete distributions converging to the diffusion densities. See Kac' article ([Kac, 1947]).

3 There is a rather unexpected connection of the urn model with finite-dimensional representations of the Lie algebra $sl(2) \cong so(2, 1)$. Indeed, introduce a new matrix by the commutator:

$$\overline{A} = \frac{1}{2} [A, \Lambda]$$

The matrix $\overline{A}$ is a skew-symmetric version of A. For $N = 3$, it is

$$\overline{A} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 3 & 0 & -2 & 0 \\ 0 & 2 & 0 & -3 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

It turns out that the triple A, $\overline{A}$ and Λ is closed under commutation, thus forms a Lie algebra, namely

$$\text{span}\,\{\, A,\ \overline{A},\ \Lambda \,\} \cong so(2, 1) \cong sl(2, \mathbb{R})$$

with commutation relations

$$[A, \overline{A}] = 2\Lambda, \qquad [\overline{A}, \Lambda] = 2A, \qquad [\Lambda, A] = -2\overline{A}$$

5.5 Krawtchouk matrices and classical random walks

In this section we will give a probabilistic meaning to the Krawtchouk matrices and illustrate some connections with classical random walks.

5.5.1 Bernoulli random walk

Let X_i be independent symmetric Bernoulli random variables taking values ± 1. Let $x_N = X_1 + \cdots + X_N$ be the associated random walk starting from 0. Now observe that the generating function of the elementary symmetric functions in the X_i is a martingale, in fact a discrete exponential martingale:

$$M_N = \prod_{i=1}^{N}(1 + vX_i) = \sum_k v^k a_k(X_1, \ldots, X_N)$$

where a_k denotes the kth elementary symmetric function. The martingale property is immediate since each X_i has mean 0. Refining the notation by setting $a_k^{(N)}$ to denote the kth elementary symmetric function in the variables $X_1, \ldots, X_N$, multiplying M_N by $1 + vX_{N+1}$ yields the recurrence

$$a_k^{(N+1)} = a_k^{(N)} + a_{k-1}^{(N)} X_{N+1}$$

which, with the boundary conditions $a_k^{(0)} = 0$, for $k > 0$, $a_0^{(n)} = 1$ for all $n \geq 0$, yields, for $k > 0$,

$$a_k^{(N+1)} = \sum_{j=0}^{N} a_{k-1}^{(j)} X_{j+1}$$

that is, these are discrete or *prototypical iterated stochastic integrals* and thus the simplest example of Wiener's homogeneous chaoses.

Suppose that at time N, the number of the X_i that are equal to -1 is j_N, with the rest equal to $+1$. Then $j_N = (N - x_N)/2$ and M_N can be expressed solely in terms of N and x_N, or, equivalently, of N and j_N

$$M_N = (1 + v)^{N - j_N}(1 - v)^{j_N} = (1 + v)^{(N + x_N)/2}(1 - v)^{(N - x_N)/2}$$

From the generating function for the Krawtchouk matrices, equation (1.1), follows

$$M_N = \sum_i v^i K_{i, j_N}^{(N)}$$

so that as functions on the Bernoulli space, each sequence of random variables $K_{i,j_N}^{(N)}$ is a martingale.

Now we can derive two basic recurrences. From a given column of $K^{(N)}$, to get the corresponding column in $K^{(N+1)}$, we have the Pascal's triangle recurrence:

$$K_{i-1\,j}^{(N)} + K_{i\,j}^{(N)} = K_{i\,j}^{(N+1)}$$

This follows in the probabilistic setting by writing $M_{N+1} = (1 + vX_N)M_N$ and remarking that for j to remain constant, X_N must take the value $+1$. The martingale property is more interesting in the present context. We have

$$K_{ij_N}^{(N)} = E(K_{ij_{N+1}}^{(N)}|X_1, \ldots, X_N) = \frac{1}{2}\left(K_{ij_N+1}^{(N+1)} + K_{ij_N}^{(N+1)}\right)$$

since half the time X_{N+1} is -1, increasing j_N by 1, and half the time j_N is unchanged. Thus, writing j for j_N,

$$K_{ij}^{(N)} = \frac{1}{2}\left(K_{ij+1}^{(N+1)} + K_{ij}^{(N+1)}\right)$$

which may be considered as a 'reverse Pascal'.

5.5.1.1   Orthogonality.   As noted above — here with a slightly simplified notation — it is natural to use variables (x, N), with x denoting the position of the random walk after N steps. Writing $K_\alpha(x, N)$ for the Krawtchouk polynomials in these variables, cf. equation (1.2), we have the generating function

$$G(v) = \sum_{\alpha=0}^{N} v^\alpha K_\alpha(x, N) = (1 + v)^{(N+x)/2}(1 - v)^{(N-x)/2}$$

The expansion

$$(1 - v)^{y-a}(1 - (1 - R)v)^{-y} = \sum_{n=0}^{\infty} \frac{v^n}{n!}(a)_n\,{}_2F_1\left({-n, y \atop a}\,\middle|\,R\right) \qquad (5.1)$$

with $(a)_n = \Gamma(a + n)/\Gamma(a)$, yields the identification as hypergeometric functions

$$K_\alpha(x, N) = \binom{N}{\alpha}\,{}_2F_1\left({-\alpha, (x - N)/2 \atop -N}\,\middle|\,2\right)$$

The calculation

$$\langle G(v)\,G(w)\rangle = \prod(1 + (v + w)X_j + vwX_j^2) = (1 + vw)^N$$

exhibits the orthogonality of the K_α if one observes that after taking expectations only terms in the product vw remain. Thus, the K_α are notable for two important features:

  1 They are the iterated integrals (sums) of the Bernoulli process.

  2 They are orthogonal polynomials with respect to the binomial distribution.

5.5.2   Multivariate Krawtchouk polynomials

The probabilistic approach may be carried out for general finite probability spaces. Fix an integer $d > 0$ and d values $\{\xi_0, \ldots, \xi_\delta\}$, with the convention $\delta = d - 1$. Take a sequence of independent identically distributed random variables having distribution $P(X = \xi_j) = p_j$, $0 \le j \le \delta$. Denote the mean and variance of the X_i by μ and σ^2 as usual.

For $N > 0$, we have the martingale

$$M_N = \prod_{j=1}^{N} (1 + v(X_j - \mu))$$

We now switch to the multiplicities as variables. Set

$$n_j = \sum_{k=1}^{N} 1_{\{X_k = \xi_j\}}$$

the number of times the value ξ_j is taken. Thus the generating function

$$G(v) = \prod_{j=0}^{\delta} (1 + v(\xi_j - \mu))^{n_j} = \sum_{\alpha=0}^{N} v^\alpha K_\alpha(n_0, \ldots, n_\delta)$$

defines our generalized Krawtchouk polynomials. One quickly gets

PROPOSITION 5 *Denoting the multi-index* $\mathbf{n} = (n_0, \ldots, n_\delta)$ *and by* $\mathbf{e_j}$ *the standard basis on* $\mathbb{Z}^d$, *Krawtchouk polynomials satisfy the recurrence*

$$K_\alpha(\mathbf{n} + \mathbf{e_j}) = K_\alpha(\mathbf{n}) + (\xi_j - \mu)K_{\alpha-1}(\mathbf{n})$$

We also find by binomial expansion

PROPOSITION 6

$$K_\alpha(n_0, \ldots, n_\delta) = \sum_{|\mathbf{k}|=\alpha} \prod_j \binom{n_j}{k_j} (\xi_j - \mu)^{k_j}$$

where $|\mathbf{k}| = \sum_{j=0}^{\delta} k_j.$

There is an interesting connection with the multivariate hypergeometric functions of Appell and Lauricella. The Lauricella polynomials F_B are defined by

$$F_B \left(\begin{matrix} -\mathbf{r}, \mathbf{b} \\ t \end{matrix} \,\middle|\, \mathbf{s} \right) = \sum_{\mathbf{k} \in \mathbb{N}^{\delta}} \frac{(-\mathbf{r})_{\mathbf{k}}(\mathbf{b})_{\mathbf{k}}}{(t)_{|\mathbf{k}|} \mathbf{k}!} \, \mathbf{s}^{\mathbf{k}}$$

with, e.g., $\mathbf{r} = (r_1, \ldots, r_{\delta})$, $(\mathbf{r})_{\mathbf{k}} = (r_1)_{k_1}(r_2)_{k_2} \cdots (r_{\delta})_{k_{\delta}}$ for multi-index $\mathbf{k}$, also $\mathbf{s}^{\mathbf{k}} = s_1^{k_1} \cdots s_{\delta}^{k_{\delta}}$, and $\mathbf{k}! = k_1! \cdots k_{\delta}!$. Note that t is a single variable. The generating function of interest here is

$$(1 - \sum v_i)^{\sum b_j - t} \prod_j (1 - \sum v_i + s_j v_j)^{-b_j} = \sum_{\mathbf{r} \in \mathbb{N}^{\delta}} \frac{\mathbf{v}^{\mathbf{r}}(t)_{|\mathbf{r}|}}{\mathbf{r}!} F_B \left(\begin{matrix} -\mathbf{r}, \mathbf{b} \\ t \end{matrix} \,\middle|\, \mathbf{s} \right)$$

$$(5.2)$$

a multivariate version of (5.1).

PROPOSITION 7   *Let $N = |\mathbf{n}|$. If $\xi_0 = 0$, then,*

$$K_{\alpha}(\mathbf{n}) = (-N)_{\alpha} \sum_{|\mathbf{r}|=\alpha} \frac{\prod (p_j \xi_j)^{r_j}}{\mathbf{r}!} F_B \left(\begin{matrix} -\mathbf{r}, -\mathbf{n} \\ -N \end{matrix} \,\middle|\, \frac{1}{p_1}, \ldots, \frac{1}{p_{\delta}} \right)$$

Proof   Let $v_j = v p_j \xi_j$, $b_j = -n_j$, $t = -N$, $s_j = p_j^{-1}$ in (5.2), for $1 \le j \le \delta$. Note that $\sum v_j = v\mu$, $\sum b_j - t = N - (\sum_{1 \le j \le \delta} n_j) = n_0$.   □

Orthogonality follows similar to the binomial case:

PROPOSITION 8 *The Krawtchouk polynomials $K_{\alpha}(n_0, \ldots, n_{\delta})$ are orthogonal with respect to the induced multinomial distribution. In fact, with $N = |\mathbf{n}|$,*

$$\langle K_{\alpha} K_{\beta} \rangle = \delta_{\alpha\beta} \, \sigma^{2\alpha} \binom{N}{\alpha}$$

Proof

$$\begin{aligned} \langle G(v) \, G(w) \rangle &= \sum \binom{N}{n_0, \ldots, n_{\delta}} p_0^{n_0} \cdots p_{\delta}^{n_{\delta}} \prod (1 + (v+w)(\xi_j - \mu) \\ &\qquad\qquad\qquad\qquad\qquad + vw(\xi_j - \mu)^2)^{n_j} \\ &= \left(\sum (p_j + (v+w)p_j(\xi_j - \mu) + vw p_j(\xi_j - \mu)^2) \right)^N \end{aligned}$$

Thus, $\langle G(v) \, G(w) \rangle = (1 + vw\sigma^2)^N$. This shows orthogonality and yields the squared norms as well.   □

5.6 "Kravchukiana" or the World of Krawtchouk Polynomials

About the year 1995, we held a seminar on Krawtchouk polynomials at Southern Illinois University. As we continued, we found more and more properties and connections with various areas of mathematics.

Eventually, by the year 2000 the theory of quantum computing had been developing with serious interest in the possibility of implementation, at the present time of MUCH interest. Sure enough, right in the middle of everything there are our flip operators, su(2), etc., etc. — same ingredients making up the Krawtchouk universe. Well, we can only report that how this all fits together is still quite open. Of special note is the idea of a hardware implementation of a Krawtchouk transform. A beginning in this direction may be found in the just-published article with Schott, Botros, and Yang [Botros et al, 2002].

At any rate, for the present we list below the topics which are central to our program. They are the basis of the **Krawtchouk Encyclopedia,** still in development; we are in the process of filling in the blanks. An extensive web resource for Krawtchouk polynomials we recommend is Zelenkov's site:

```
http://www.geocities.com/orthpol/
```

Note that we do not mention work in areas less familiar to us, notably that relating to q-Krawtchouk polynomials, such as in [Steele, 1997].

We welcome contributions. If you wish either to send a reference to your paper(s) on Krawtchouk polynomials or contribute an article, please contact one of us !

Our email: `pfeinsil@math.siu.edu` or `jkocik@math.siu.edu`.

5.6.1 Krawtchouk Encyclopedia

Here is a list of topics currently in the Krawtchouk Encyclopedia.

 1 Pascal's Triangle

 2 Random Walks

- Path integrals
- A, K, and Λ
- Nonsymmetric Walks

- Symmetric Krawtchouk matrices and binomial expectations

3 Urn Model

- Markov chains
- Initial and invariant distributions

4 Symmetric Functions. Energy

- Elementary symmetric functions and determinants
- Traces on Grassman algebras

5 Martingales

- Iterated integrals
- Orthogonal functionals
- Krawtchouk polynomials and multinomial distribution

6 Lie algebras and Krawtchouk polynomials

- so(2,1) explained
- so(2,1) spinors
- Quaternions and Clifford algebras
- S and so(2,1) tensors
- Three-dimensional simple Lie algebras

14 Krawtchouk Transforms

- Orthogonal transformation associated to K

- Exponential function in Krawtchouk basis

- Krawtchouk transform

15 Hypergeometric Functions

- Krawtchouk polynomials as hypergeometric functions

- Addition formulas

16 Symmetric Krawtchouk Matrices

- The matrix T

- S-squared and trace formulas

- Spectrum of S

17 Gaussian Quadrature

- Zeros of Krawtchouk polynomials

- Gaussian-Krawtchouk summation

18 Coding Theory

- Mac Williams' theorem

- Association schemes

19 Appendices

- K and S matrices for N from 1 to 14

- Krawtchouk polynomials in the variables x,N/i,j/j,N for N from 1 to 20

- Eigenvalues of S

- Remarks on the multivariate case

- Time-zero polynomials

- Mikhail Philippovitch Krawtchouk: a biographical sketch

5.7   Appendix

5.7.1   Krawtchouk matrices

$$K^{(0)} = \begin{bmatrix} 1 \end{bmatrix}$$

$$K^{(1)} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$K^{(2)} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & -2 \\ 1 & -1 & 1 \end{bmatrix}$$

$$K^{(3)} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & 1 & -1 & -3 \\ 3 & -1 & -1 & 3 \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

$$K^{(4)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 4 & 2 & 0 & -2 & -4 \\ 6 & 0 & -2 & 0 & 6 \\ 4 & -2 & 0 & 2 & -4 \\ 1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

$$K^{(5)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 5 & 3 & 1 & -1 & -3 & -5 \\ 10 & 2 & -2 & -2 & 2 & 10 \\ 10 & -2 & -2 & 2 & 2 & -10 \\ 5 & -3 & 1 & 1 & -3 & 5 \\ 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix}$$

$$K^{(6)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 4 & 2 & 0 & -2 & -4 & -6 \\ 15 & 5 & -1 & -3 & -1 & 5 & 15 \\ 20 & 0 & -4 & 0 & 4 & 0 & -20 \\ 15 & -5 & -1 & 3 & -1 & -5 & 15 \\ 6 & -4 & 2 & 0 & -2 & 4 & -6 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

Table 1

5.7.2   Symmetric Krawtchouk matrices

$$S^{(0)} = \begin{bmatrix} 1 \end{bmatrix}$$

$$S^{(1)} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$S^{(2)} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 0 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

$$S^{(3)} = \begin{bmatrix} 1 & 3 & 3 & 1 \\ 3 & 3 & -3 & -3 \\ 3 & -3 & -3 & 3 \\ 1 & -3 & 3 & -1 \end{bmatrix}$$

$$S^{(4)} = \begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 8 & 0 & -8 & -4 \\ 6 & 0 & -12 & 0 & 6 \\ 4 & -8 & 0 & 8 & -4 \\ 1 & -4 & 6 & -4 & 1 \end{bmatrix}$$

$$S^{(5)} = \begin{bmatrix} 1 & 5 & 10 & 10 & 5 & 1 \\ 5 & 15 & 10 & -10 & -15 & -5 \\ 10 & 10 & -20 & -20 & 10 & 10 \\ 10 & -10 & -20 & 20 & 10 & -10 \\ 5 & -15 & 10 & 10 & -15 & 5 \\ 1 & -5 & 10 & -10 & 5 & -1 \end{bmatrix}$$

$$S^{(6)} = \begin{bmatrix} 1 & 6 & 15 & 20 & 15 & 6 & 1 \\ 6 & 24 & 30 & 0 & -30 & -24 & -6 \\ 15 & 30 & -15 & -60 & -15 & 30 & 15 \\ 20 & 0 & -60 & 0 & 60 & 0 & -20 \\ 15 & -30 & -15 & 60 & -15 & -30 & 15 \\ 6 & -24 & 30 & 0 & -30 & 24 & -6 \\ 1 & -6 & 15 & -20 & 15 & -6 & 1 \end{bmatrix}$$

Table 2

5.7.3 Sylvester-Hadamard matrices

$$H^{(0)} = \begin{bmatrix} \bullet \end{bmatrix}$$

$$H^{(1)} = \begin{bmatrix} \bullet & \bullet \\ \bullet & \circ \end{bmatrix}$$

$$H^{(3)} = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \circ & \bullet & \circ \\ \bullet & \bullet & \circ & \circ \\ \bullet & \circ & \circ & \bullet \end{bmatrix}$$

$$H^{(4)} = \begin{bmatrix}
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ \\
\bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ & \circ \\
\bullet & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet \\
\bullet & \bullet & \bullet & \bullet & \circ & \circ & \circ & \circ \\
\bullet & \circ & \bullet & \circ & \circ & \bullet & \circ & \bullet \\
\bullet & \bullet & \circ & \circ & \circ & \circ & \bullet & \bullet \\
\bullet & \circ & \circ & \bullet & \circ & \bullet & \bullet & \circ
\end{bmatrix}$$

$$H^{(5)} = \begin{bmatrix}
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\
\bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ \\
\bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ & \circ \\
\bullet & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet \\
\bullet & \bullet & \bullet & \bullet & \circ & \circ & \circ & \circ & \bullet & \bullet & \bullet & \bullet & \circ & \circ & \circ & \circ \\
\bullet & \circ & \bullet & \circ & \circ & \bullet & \circ & \bullet & \bullet & \circ & \bullet & \circ & \circ & \bullet & \circ & \bullet \\
\bullet & \bullet & \circ & \circ & \circ & \circ & \bullet & \bullet & \bullet & \bullet & \circ & \circ & \circ & \circ & \bullet & \bullet \\
\bullet & \circ & \circ & \bullet & \circ & \bullet & \bullet & \circ & \bullet & \circ & \circ & \bullet & \circ & \bullet & \bullet & \circ \\
\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \circ \\
\bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet \\
\bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ & \circ & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet & \bullet \\
\bullet & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet & \circ & \bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ \\
\bullet & \bullet & \bullet & \bullet & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \circ & \bullet & \bullet & \bullet & \bullet \\
\bullet & \circ & \bullet & \circ & \circ & \bullet & \circ & \bullet & \circ & \bullet & \circ & \bullet & \bullet & \circ & \bullet & \circ \\
\bullet & \bullet & \circ & \circ & \circ & \circ & \bullet & \bullet & \circ & \circ & \bullet & \bullet & \bullet & \bullet & \circ & \circ \\
\bullet & \circ & \circ & \bullet & \circ & \bullet & \bullet & \circ & \circ & \bullet & \bullet & \circ & \bullet & \circ & \circ & \bullet
\end{bmatrix}$$

Table 3

Replace • with 1 and o with −1 to obtain Sylvester-Hadamard matrices.

References

N.M. Atakishiyev and K.B. Wolf, *Fractional Fourier-Kravchuk transform* J. Opt. Soc. Amer. A,**14** 7 (1997) 1467–1477.

N. Bose, *Digital filters: theory and applications,* North-Holland, 1985.

N. Botros, J. Yang, P. Feinsilver, and R. Schott, *Hardware Realization of Krawtchouk Transform using VHDL Modeling and FPGAs,* IEEE Transactions on Industrial Electronics, **49** 6 (2002)1306–1312.

W.Y.C. Chen and J.D. Louck, *The combinatorics of a class of representation functions,* Adv. in Math., **140** (1998) 207–236.

P. Delsarte, *Bounds for restricted codes, by linear programming,* Philips Res. Reports, **27** (1972) 272–289.

P. Delsarte, *Four fundamental parameters of a code and their combinatorial significance,* Info. & Control, **23** (1973) 407–438.

P. Delsarte, *An algebraic approach to the association schemes of coding theory,* Philips Research Reports Supplements, No. 10, 1973.

C.F. Dunkl, *A Krawtchouk polynomial addition theorem and wreath products of symmetric groups,* Indiana Univ. Math. J., **25** (1976) 335–358.

C.F. Dunkl and D.F. Ramirez, *Krawtchouk polynomials and the symmetrization of hypergraphs,* SIAM J. Math. Anal., **5** (1974) 351–366.

G.K. Eagelson, *A characterization theorem for positive definite sequences of the Krawtchouk polynomials,* Australian J. Stat, **11** (1969) 29–38.

P. Feinsilver and R. Fitzgerald, *The spectrum of symmetric Krawtchouk matrices,* Lin. Alg. & Appl., **235** (1996) 121–139.

P. Feinsilver and J. Kocik, *Krawtchouk matrices from classical and quantum random walks,* Contemporary Mathematics, **287** (2001) 83–96.

P. Feinsilver and R. Schott, *Krawtchouk polynomials and finite probability theory,* Probability Measures on Groups X, Plenum (1991) 129–135.

F.G. Hess, *Alternative solution to the Ehrenfest problem,* Amer. Math. Monthly, **61** (1954) 323–328.

M. Kac, *Random Walks and the theory of Brownian motion,* Amer. Math. Monthly 54, 369–391, 1947.

S. Karlin, J. McGregor, *Ehrenfest Urn Model,* J. Appl. Prob. 2, 352–376, 1965.

M. Krawtchouk, *Sur une generalisation des polynomes d'Hermite,* Comptes Rendus, **189** (1929) 620–622.

M. Krawtchouk, *Sur la distribution des racines des polynomes orthogonaux,* Comptes Rendus, **196** (1933) 739–741.

V.I. Levenstein, *Krawtchouk polynomials and universal bounds for codes and design in Hamming spaces,* IEEE Transactions on Information Theory, **41** 5 (1995) 1303–1321.

S.J. Lomonaco, Jr., *A Rosetta Stone for quantum mechanics with an introduction to quantum computation,*
http://www.arXiv.org/abs/quant-ph/0007045

F.J. MacWilliams and N.J.A. Sloane, *The theory of Error-Correcting Codes,* The Netherlands, North Holland, 1977.

A.J.F. Siegert, *On the approach to statistical equilibrium,* Phys. Rev., **76** (1949), 1708–1714.

D. Stanton, *Some q-Krawtchouk polynomials on Chevalley groups,* Amer. J. Math., **102** (1980) 625–662.

G. Szegö, *Orthogonal Polynomials,* Colloquium Publications, Vol. 23, New York, AMS, revised eddition 1959, 35–37.

D. Vere-Jones, *Finite bivariate distributions and semi-groups of nonnegative matrices,* Q. J. Math. Oxford, **22** 2 (1971) 247–270.

M. Voit, *Asymptotic distributions for the Ehrenfest urn and related random walks,* J. Appl. Probab., **33** (1996) 340–356.

R.K. Rao Yarlagadda and J.E. Hershey, *Hadamard matrix analysis and synthesis: with applications to communications and signal/image processing,* Kluwer Academic Publishers, 1997.

This page intentionally left blank

AN ELEMENTARY RIGOROUS INTRODUCTION TO EXACT SAMPLING

F. Friedrich
Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany
friedrich@gsf.de

G. Winkler
Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany
gwinkler@gsf.de

O. Wittich
Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany
wittich@gsf.de

V. Liebscher
Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany
liebscher@gsf.de
  http://www.gsf.de/institute/ibb/

Abstract We introduce coupling from the past, a recently developed method for exact sampling from a given distribution. Focus is on rigour and thorough proofs. We stay on an elementary level which requires little or no prior knowledge from probability theory. This should fill an obvious gap between innumerable intuitive and incomplete reviews, and few precise derivations on an abstract level.

6.1   Introduction

We introduce a recently developed method for exact sampling from a given distribution. It is called *coupling from the past.* This is in contrast to Markov chain Monte Carlo samplers like the Gibbs, sampler or the family of Metropolis-Hastings samplers which return samples from a distribution approximating the target distribution. The drawback is that MCMC methods apply generally and exact sampling works in special cases only. On the other hand, it is the object of current research and the list of possible applications increases rapidly. Another advantage is that problems like burn in and convergence diagnostics do not arise where exact sampling works. Exact sampling was proposed in the seminal paper [J.G. PROPP & D.B. WILSON, 1996]. Whereas these authors called the method *exact sampling,* some prefer the term *perfect sampling* since random sampling never is exact. For background in Markov chains and sampling, and for examples, we refer to [G. WINKLER, 1995; G. WINKLER, 2003].   The aim of the present paper is a rigorous derivation and a thorough analysis at an elementary level. Nothing is really new; the paper consists of a combination of ideas, examples, and techniques from various recent papers, basically along the lines in [F. FRIEDRICH, 2003]. Hopefully, we can single out the basic conditions under which the method works theoretically, and what has to be added for a practicable implementation.

Coupling from the past is closely related to Markov Chain Monte Carlo sampling (MCMC), which nowadays is a widespread and commonly accepted statistical tool, especially in Bayesian statistical analysis. Hence we premise the discussion of coupling to the past with some remarks on Markov Chain Monte Carlo sampling. Let us first introduce the general framework which simultaneously gives us the basis for coupling from the past. For background and a detailed discussion see [G. WINKLER, 1995].

Let X be a finite set of generic elements $x, y, \ldots$. A *probability distribution* ν on X is a function on X taking values in the unit interval $[0,1]$ such that $\sum_{x \in X} \nu(x) = 1$. A *Markov kernel* or *transition probability* on X is a function $P : X \times X \to [0,1]$ such that for each $x \in X$ the function $P(x, \cdot) : X \to [0, 1]$, $y \mapsto P(x,y)$ is a probability distribution on X. A probability distribution ν on X can be interpreted as a row vector $(\nu(x))_{x \in X}$ and a Markov kernel P as a stochastic matrix $(P(x,y))_{x,y \in X}$. A *right Markov chain* with initial distribution ν and transition probability P is a sequence $(\xi_i)_{i \geq 0}$ of random variables the law of which is determined by ν and P via the finite-dimensional marginal distributions given by

$$\mathbb{P}(\xi_0 = x_0, \xi_1 = x_1, \ldots, \xi_n = x_n) = \nu(x_0) P(x_0, x_1) \cdot \cdots \cdot P(x_{n-1}, x_n).$$

P is called *primitive* if there is a natural number τ such that $P^\tau(x, y) > 0$ for all $x, y \in X$. This means that the τ-step probability from state x to state y is strictly positive for arbitrary x and y. If P is primitive then there is a unique

probability distribution μ which is *invariant* w.r.t. P, i.e. $\mu P = \mu$ where μP is the matrix product of the (left) row vector μ and the matrix P, and this invariant probability distribution μ is strictly positive.

The laws or distributions of the variables ξ_n of such a process converge to the invariant distribution, i.e.

$$\nu P \cdot \cdots \cdot P(y) \longrightarrow \mu(y), \quad y \in X, \tag{1.1}$$

cf. [G. WINKLER, 1995], Theorem 4.3.1. Perhaps the most important statistical features to be estimated are expectation values of functions on the state space X, and the most common estimators are empirical means. Fortunately, such stochastic processes fulfill the law of large numbers, which in its most elementary version reads: For each function f on X, the empirical means along time converge in probability (and in L^2) to the expectation of f with respect to the invariant distribution; in formulae this reads

$$\frac{1}{n} \sum_{i=0}^{n-1} f(\xi_i) \longrightarrow \mathbb{E}(f; \mu) \text{ as } n \to \infty, \quad \text{in probability}, \tag{1.2}$$

(cf. [G. WINKLER, 1995], Theorem 4.3.2). The symbol $\mathbb{E}(f; \mu)$ denotes the expectation

$$\mathbb{E}(f; \mu) = \sum_{x \in X} f(x)\mu(x)$$

of f with respect to μ. A sequence of random variables ξ_i *converges* to the random variable ξ *in probability* if for each $\varepsilon > 0$ the probability $\mathbb{P}(|\xi_i - \xi| > \varepsilon)$ tends to 0 as n tends to ∞. Plainly, (1.2) implies that for every natural number m, averaging may be started from m without destroying convergence in probability; more precisely for each $m \geq 0$ one has

$$\frac{1}{n-m} \sum_{i=m+1}^{n} f(\xi_i) \longrightarrow \mathbb{E}(f; \mu) \text{ as } n \to \infty, \quad \text{in probability}. \tag{1.3}$$

In view of the law of large numbers for identically distributed and independent variables, the step number m should be large enough such that the distributions of the variables $\xi_{m+1}, \ldots, \xi_n$ are close to the invariant distribution μ in order to estimate the expectation of f with respect to μ properly from the samples $f(\xi_{m+1}), \ldots, f(\xi_n)$.

In fact, according to (1.1), after some time m the laws of the ξ_i should be close to the invariant distribution μ although they may be far from μ during the initial period. The values during this *burn in* period are usually discarded and an average $(\sum_{m+1}^{n} f(\xi_i))/(n-m)$ like in (1.3) is computed. In general, the burn in time can hardly be determined. There are a lot of suggestions ranging from visual inspection of the time series $(f(\xi_i))_{i \geq 0}$ to more formal tools, called *convergence diagnostics*convergence diagnostics. In this text we are not concerned with burn in and restrict ourselves to the illustration in Fig. 1. A Gibbs, sampler (introduced in Section 6.4) for the Ising model is started with a pepper and salt configuration in the left picture. A typical sample of the invariant distribution is the right one which appears after about 8000 steps. The pictures in-between show intermediate configurations which are pretty improbable given the invariant distribution but which are quite stable with respect to the Gibbs sampler. In physical terms, the right middle configuration is close to a 'meta-stable' state. Since we are interested in a typical configuration of the invariant distribution μ, we should consider the burn in to be completed if the sample from the Markov chain looks like the right hand side of Fig. 1, i.e. after about 8000 steps of the Gibbs sampler. The curve in the next figure
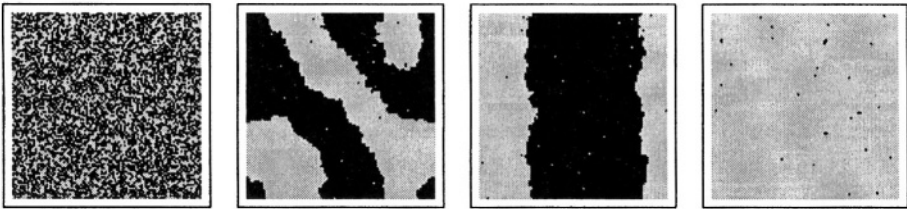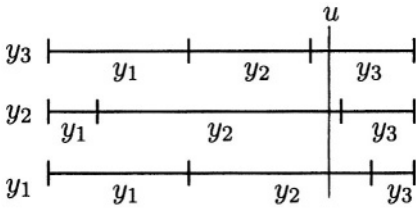

Figure 1.   Configurations for Ising Gibbs Sampler with $\beta = 0.8$ starting in a pepper and salt-configuration (left), after 150 steps (left middle), after 350 steps (right middle) and after 8000 steps (right).

Fig. 2 displays the relative frequency of equal neighbour pairs. Superficial visual inspection of this plot suggests that the sampler should be in equilibrium after about 300 steps. On the other hand, comparison with Fig. 1 reveals that the slight ascent at about 7800 steps presumably is much more relevant for the decision whether burn is completed or not. This indicates that primitive diagnostic tools may be misleading. The interested reader is referred to the references in [W.R. GILKS ET AL., 1996; A. GELMAN, 1996; A.E. RAFTERY & S.M. LEWIS, 1996], see [W.R. GILKS ET AL., 1996b]. If initial samples from μ itself are available, then there is no need for a burn in, and one can average from the beginning. This is one of the most valuable advantages of exact sampling.

First, we indicate how a Markov chain can be simulated.

EXAMPLE 1 (SIMULATING A MARKOV CHAIN) We denote by P the transition probability of a homogeneous Markov chain. At each time $n \geq 1$, given the previous state x_{n-1}, we want to pick a state x_n at random from $P(x_{n-1}, \cdot)$. For each x, we partition the unit interval $(0,1]$ into intervals I_y^x of length $P(x, y)$, and pick u_n uniformly at random from $(0,1]$. Given the present state x_{n-1}, we search for the state y with $u_n \in I_y^{x_{n-1}}$ and set $x_n = y$. The picture on the left illustrates this procedure for $|\boldsymbol{X}| = 3$, where $x_n = y_2$ if x_{n-1} was y_1 or y_2 and $x_n = y_3$ if $x_{n-1} = y_3$. In general, the procedure can be rephrased as follows: Define a *transition rule* for P by

$$f : \boldsymbol{X} \times (0,1] \longrightarrow \boldsymbol{X}, \quad f(x, u) = y \quad \text{if and only if} \quad u \in I_y^x.$$

More explicitly, enumerate $\boldsymbol{X} = \{y_1, \ldots, y_N\}$ and set $f(x, u) = F_x^-(u)$ where $F_x(u) = P(x, \{y_i : i \leq u\})$ is the cumulative distribution function of $P(x, \cdot)$ and $F_x^-(u) = \min\{t : F_x(t) \geq u\}$ its generalized inverse. Let $U_1, U_2, \ldots$ be independent random variables uniformly distributed over $(0,1]$, and set $\xi_0 := x_0$, and $\xi_n := f(\xi_{n-1}, U_n)$. Then $(\xi_n)_{n \geq 0}$ is a homogeneous Markov chain starting at x_0 with transition probability P. For inhomogeneous chains, replace f by f_n varying in time. Note that the exclusive source of randomness are the independent random variables U_i.
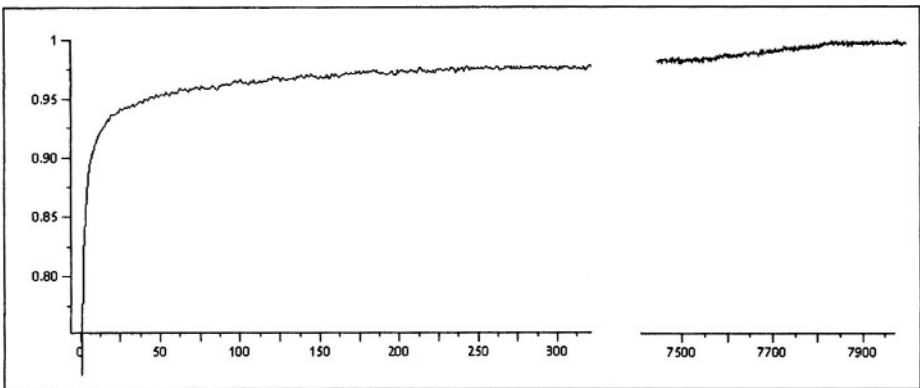

Figure 2.   Convergence Diagnostics for Ising Gibbs Sampler

6.2   Exact Sampling

The basic idea of coupling from the past is closely related to the law of large numbers (1.2). According to (1.1), for primitive P with invariant distribution μ the corresponding Markov chain converges to μ; more precisely

$$\nu P^n \longrightarrow \mu, \quad \text{as } n \to \infty, \tag{2.1}$$

uniformly in all initial distributions ν, and with respect to any norm on $\mathbb{R}^X$.

Generalizing the concept of right Markov chains, let us consider now *two-sided Markov chains* with transition probabilities given by a Markov kernel P, i.e. double sequences $(\xi_i)_{i \in \mathbb{Z}}$ of random variables taking values in X, and with law determined by the marginal distributions

$$\mathbb{P}\big(\xi_m = x_m, \ \dots \ , \xi_n = x_n\big) = \nu_m\big(x_m\big) P\big(x_m, x_{m+1}\big) \cdot \dots \cdot P\big(x_{n-1}, x_n\big), \tag{2.2}$$

for $m, n \in \mathbb{Z}, n > m$, where ν_k denotes the law of ξ_k.

If P is primitive, or more generally, if (2.1) holds uniformly, these two-sided chains are automatically *stationary*. This important concept means that a time shift does not change the law of the chain; in terms of the marginal distributions this reads

$$\mathbb{P}\big(\xi_m = x_m, \ \dots \ , \xi_n = x_n\big) = \mathbb{P}\big(\xi_{m+\tau} = x_m, \ \dots \ , \xi_{n+\tau} = x_n\big) \tag{2.3}$$

for all $m \in \mathbb{Z}$ and $\tau \in \mathbb{Z}$, and in particular, that all ν_m in (2.2) are equal to μ. In fact, because of (2.2) one has $\nu_0 = \nu_{-k} P^k$ for all $k \in \mathbb{N}$. By uniformity in (2.1), this implies $\nu_0 = \mu$ and hence in view of (2.2) the process $(\xi_i)_{i \in \mathbb{Z}}$ is stationary.

At a first glance, this does not seem to be helpful since we cannot simulate the two-sided chain starting at time $-\infty$. On the other hand, if we want to start sampling at some (large negative) time n, there is no distinguished state to start in, since stationarity of the chain implies that the initial state necessarily is already distributed according to μ. The main idea to overcome this problem is to start chains simultaneously at all states in X and at each time. This means that a lot of Markov chains are *coupled* together. The coupling will be constructed in such a fashion that if two of the chains happen to be in the same state in X at some (random) time, they will afterwards follow the same trajectory forever. This phenomenon is called *coalescence* of trajectories. Our definite aim is to couple the chains in a cooperative way such that after a large time it is very likely that any two of the chains have met each other at time 0. Then, at time 0, all chains started simultaneously at sufficiently large negative time have coalesced, and therefore their common state at time 0 does not dependent on the starting points in the far past anymore. We will show that after complete coalescence the unique random state at time 0 *is distributed according to the invariant distribution μ.*

To make this precise we consider the following setup: Let X be a finite space and let μ be a strictly positive probability distribution on X. The aim is to realize a random variable which *exactly* has law μ, or - in other words - to sample from μ. Since Markov chains have to be started at each time $k < 0$ and at each state $x \in X$ simultaneously, a formal framework is needed into which all these processes can be embedded. The appropriate concept is that of *iterated random maps* or *stochastic flows,* systematically exploited in [P. DI-ACONIS & D. FREEDMAN, 1999].

Let μ be the strictly positive distribution on X from which we want to sample and let P be a Markov kernel on X for which μ is the unique invariant distribution. Let Φ be the set of all maps from X to itself:

$$\Phi = \{\varphi : X \longrightarrow X\} = X^X = \operatorname{Map}(X, X).$$

On this space we consider distributions p reflecting the action of P on X in the sense that the p-**probability** that some point x is mapped by the random function φ to some y is given by $P(x, y)$. This connection between p and P is formalized by the condition

(P) $\qquad\qquad p\left(\{\varphi : \varphi(x) = y\}\right) = P(x, y),\ x, y \in X.$

EXAMPLE 2 Such a distribution does always exist. A synchronous one is given by $q(\varphi) = \prod_{x \in X} P(x, \varphi(x))$. It is a probability distribution since it can be written as a product of the distributions $P(x, \cdot)$. It also fulfills Condition (P): Let Φ' be the set of all maps from $X \backslash \{x\}$ to X. Then

$$q\big(\varphi : \varphi(x) = y\big)$$
$$= \sum_{\{\varphi : \varphi(x) = y\}} \prod_{z \in X} P(z, \varphi(z)) = P(x, y) \sum_{\varphi \in \Phi'} \prod_{z \neq x} P(z, \varphi(z)) = P(x, y);$$

the sum over Φ' equals 1 since the summands again define a product measure.

Since we want to mimic Markov processes, we need measures on sets of paths, and since we will proceed from time $-\infty$ to finite times we introduce measures on the set $\Omega = \Phi^{\mathbb{Z}}$ with one-dimensional marginal measures p. The simplest choice are product measures $\mathbb{P} = p^{\mathbb{Z}}$. The space $\Omega = \Phi^{\mathbb{Z}}$ consists of double sequences

$$\underline{\varphi} = (\varphi_j)_{j \in \mathbb{Z}} = (\ldots, \varphi_{-1}, \varphi_0, \varphi_1, \ldots) \in \operatorname{Map}(X, X)^{\mathbb{Z}}.$$

If J is a finite subset of $\mathbb{Z}$ then for each choice $\psi_j,\ j \in J$, we have

$$\mathbb{P}(\{\underline{\varphi} \in \Omega : \varphi_j = \psi_j,\ j \in J\}) = \prod_{j \in J} p(\psi_j).$$

Given a double sequence $\underline{\varphi}$ of maps φ_j, $j \in \mathbb{Z}$, we consider compositions of the components φ_j over time intervals. For each $\underline{\varphi} \in \Omega$ and $x \in X$, set

$$\varphi_j^k(x) = \varphi_k \circ \cdots \circ \varphi_j(x) = \varphi_k\big(\varphi_{k-1}(\cdots(\varphi_j(x)))\big), \quad j \le k.$$

Note that $\varphi_i^i = \varphi_i$.

REMARK   Given Condition (P), for each $n \in \mathbb{Z}$ and $x \in X$, the process $\xi_n \equiv x$, $\xi_{n+k} = \varphi_{n+1}^{n+k}(x)$, $k \ge 1$, is a Markov chain starting at x and with transition probability P. Hence the stochastic flow is a common representation of Markov chains starting at all initial states and at all times; we shall say that they are *coupled from the past.*

Coupling from the past at time n will work as follows: Pick a double sequence

$$\ldots, \varphi_m, \ldots, \varphi_n, \ldots$$

of maps at random, and fix a number $n \in \mathbb{Z}$. Then decrease m until $\varphi_m^n(x) = w$ hopefully does not depend on x anymore. If we are successful and this happens then we say that all trajectories

$$\varphi_m(x), \varphi_{m+1} \circ \varphi_m(x), \ldots, \varphi_m^n(x), \quad x \in X,$$

have *coalesced.* We shall also say that for $\underline{\varphi}$ there is *complete coalescence* at time n. This works if sufficiently many of the φ_j map different elements x to the same image. Going further backwards does not change anything since $\varphi_{m-k}^n(x) = \varphi_m^n(\varphi_{m-k}^{m-1}(x)) = w$ holds as well for all x. This may be rephrased in terms of sets as follows: Let $\varphi : X \to X$ be a map and $\mathrm{Im}\varphi = \{\varphi(x) : x \in X\}$ the image of X under φ. For fixed n the sets $\mathrm{Im}\varphi_m^n$ decrease as m decreases. Complete coalescence means that $\mathrm{Im}\varphi_m^n$ is a singleton $\{w\}$. Then there is a unique $W_n(\underline{\varphi}) \in X$ with

$$\{W_n(\underline{\varphi})\} := \bigcap_{m \le n} \mathrm{Im}\varphi_m^n. \tag{2.4}$$

If there is no coalescence then $W_n(\underline{\varphi})$ is not defined. Let us set

$$F_n = \{\underline{\varphi} : W_n(\underline{\varphi}) \text{ exists}\}, \quad F = \bigcap_{n \in \mathbb{Z}} F_n.$$

Then all W_n are well defined on F; to complete the definition let $W_n(\underline{\varphi}) = z_0$ for some fixed $z_0 \in X$ if $\underline{\varphi} \notin F$. Obviously, independent of the choice of $x \in X$,

$$\begin{aligned} W_{n+k}(\underline{\varphi}) &= \lim_{m \to -\infty} \varphi_m^{n+k}(x) = \varphi_{n+1}^{n+k} \circ \lim_{m \to -\infty} \varphi_m^n(x) \\ &= \varphi_{n+1}^{n+k} \circ W_n(\underline{\varphi}), \qquad \underline{\varphi} \in F, \ n \in \mathbb{Z}, \ k > 0. \end{aligned} \tag{2.5}$$

This indicates that the random variables $W_n(\underline{\varphi})$ have law μ. To exploit this observation for a sampling algorithm we need almost sure complete coalescence in finite time. We enforce this by the formal condition

(F) $$\mathbb{P}(F) = 1.$$

Provided that (F) holds, we call $\mathbb{P}$ *successful*. Condition (F) will be verified below under natural conditions.

LEMMA 3 *Under the hypothesis* (P) *and* (F) *the process* $(W_m)_{m \in \mathbb{Z}}$ *is a stationary homogeneous Markov process with Markov kernel P.*

Proof. Recall that $\mathbb{P}$ is a homogeneous product measure, and hence for each $\tau \in \mathbb{Z}$ all random sequences $\varphi_m, \dots, \varphi_{m+\tau}$, $m \in \mathbb{Z}$, have the same law. Hence the stochastic flow is stationary, and the process $(W_m)_{m \in \mathbb{Z}}$ is stationary as well. Moreover, φ_{n+1}^{n+k} depends on $\varphi_{n+1}, \dots, \varphi_{n+k}$ only and each W_m depends only on $\dots, \varphi_{m-1}, \varphi_m$. Again, since $\mathbb{P} = p^{\mathbb{Z}}$ is a product measure, the variables φ_{n+1}^{n+k} and W_m, $m \leq n$, are independent. By (2.4) and (P),

$$\mathbb{P}\big(W_{n+1} = x_{n+1}, W_n = x_n, \dots, W_{n-k} = x_{n-k}\big)$$
$$= \mathbb{P}\big(\varphi_{n+1}^{n+1}(x_n) = x_{n+1}, W_n = x_n, \dots, W_{n-k} = x_{n-k}\big)$$
$$= \mathbb{P}\big(\varphi_{n+1}(x_n) = x_{n+1}\big)\mathbb{P}\big(W_n = x_n, \dots, W_{n-k} = x_{n-k}\big)$$
$$= P\big(x_n, x_{n+1}\big)\mathbb{P}\big(W_n = x_n, \dots, W_{n-k} = x_{n-k}\big),$$

which shows

$$\mathbb{P}\big(W_{n+1} = x_{n+1} \big| W_n = x_n, \dots, W_{n-k} = x_{n-k}\big) = P\big(x_n, x_{n+1}\big).$$

Hence P is the transition probability of the process $(W_m)_{m \in \mathbb{Z}}$. ∎ Let us put things together in the first main theorem.

THEOREM 4 (EXACT SAMPLING) *Suppose that is μ a strictly positive probability distribution and P a primitive Markov kernel on X such that $\mu P = \mu$. Assume further that $p\left(\{\varphi : \varphi(x) = y\}\right) = P(x, y)$ for all $x, y \in X$, and that $\mathbb{P}$ is successful. Then each random variable W_n has law μ; more precisely:*

$$\mathbb{P}\left(\{\underline{\varphi} \in \Omega : W_n(\underline{\varphi}) = x\}\right) = \mu(x), \quad x \in X. \tag{2.6}$$

Proof. By stationarity from Lemma 3, all one-dimensional marginal distributions coincide, and P is the transition probability of $(W_n)_{n \in \mathbb{Z}}$. If P is primitive then by [G. WINKLER, 1995], Theorem 4.3.1, its unique invariant distribution is μ. ∎ To sample from μ, only one of the W_m is needed.

COROLLARY 5 *Under the assumptions of Theorem 4, the random variable W_0 has law μ.*

The next natural question concerns the waiting time for complete coalescence at time zero. The random times T_n of latest coalescence before n are given by

$$T_n(\underline{\varphi}) = \sup\{m \le n : \text{there is } w \in \mathbf{X} \text{ such that } \varphi^n_m(x) = w \text{ for every } x \in \mathbf{X}\}.$$

The numbers $T_n(\underline{\varphi})$ definitely are finite if $\underline{\varphi} \in F$; outside F they may be finite or equal $-\infty$. Condition (F) is equivalent to

$$\mathbb{P}(\{\underline{\varphi} \in \Omega : T_n(\underline{\varphi}) > -\infty\}) = 1 \quad \text{for every} \quad n \in \mathbb{Z}. \tag{2.7}$$

Such a random time is also called *successful.* To realize W_0 one subsequently and independently picks maps $\varphi_0, \varphi_{-1}, \ldots, \varphi_m$ until there is coalescence say in $w \in \mathbf{X}$. This element w is a sample from μ. For computational reasons, one usually goes back in time by powers of 2. Clearly, choosing $k_0(\underline{\varphi})$ such that $-2^{k_0(\underline{\varphi})} \le T_n(\underline{\varphi})$ assures coalescence at time 0. Recall that such a $k_0(\underline{\varphi})$ exists for each $\underline{\varphi} \in \overline{F}$. An example of a stochastic flow coalescing completely at time $m = 0$ is shown in Fig. 3. We are going now to discuss a condition
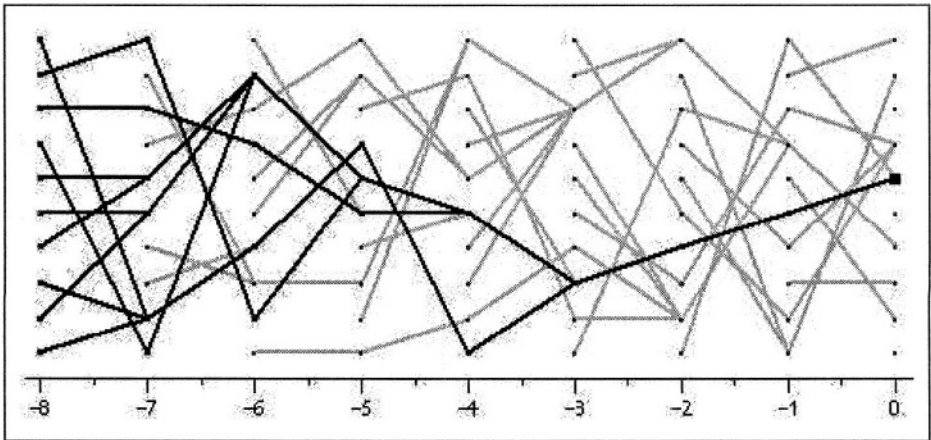
Figure 3.   Latest complete coalescence time before time 0

for (F) to hold. *Pairwise coalescence with positive probability* is perhaps the most natural condition and easy to check:

(C) For each pair $x, y \in \mathbf{X}$ there is an integer $n(x, y)$ such that

$$p^{n(x,y)}(\{(\varphi_1, \ldots, \varphi_{n(x,y)}) \in \Phi^{n(x,y)} : \varphi_1^{n(x,y)}(x) = \varphi_1^{n(x,y)}(y)\}) > 0.$$

We shall show in Theorem 9 below that (C) and (F) are equivalent. We give now a simple example where coupling fails.

EXAMPLE 6 Consider P with invariant μ on $X = \{1,2\}$ given by

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}, \quad \mu = (1/2, 1/2).$$

Let $p(\iota) = 1/2 = p(\psi)$ for the identity map $\iota(1) = 1$, $\iota(2) = 2$, and the flip map $\psi(1) = 2$, $\psi(2) = 1$. Compositions of ι and ψ never will couple. On the other hand the flow is associated to P since $p(\{\varphi : \varphi(x) = y\}) = 1/2 = P(x, y)$, regardless of x and y, and Condition (P) holds.

We shall show now that the coupling condition (C) implies complete coalescence (F) (and the converse). The latter condition may be rephrased as follows: All random times T_n are finite almost surely. By stationarity this boils down to: The random time T_0 is finite almost surely. The simplest, but fairly abstract way to verify (F) is to use shift invariance of F and ergodicity of $\mathbb{P}$. We will argue along these lines but in a more explicit and elementary way. The first step is to ensure existence of a finite τ such that the flow coalesces completely in less than τ steps with positive probability.

LEMMA 7 *Under condition* (C) *there is a natural number* τ *such that*

$$\mathbb{P}(\{\underline{\varphi} : T_0(\underline{\varphi}) > -\tau\}) > 0.$$

Proof. Let $n_c = \max\{n(x, y) : x, y \in X\}$. If $\varphi_1^n(x) = \varphi_1^n(y)$ for some $n < n_c$ then $\varphi_1^{n_c}(x) = \varphi_{n+1}^{n_c} \circ \varphi_1^n(x) = \varphi_1^{n_c}(y)$ as well. Hence Condition (C) implies

$$q = \min \left\{ p^{n_c} \{ (\varphi_1, \dots, \varphi_{n_c}) : \varphi_1^{n_c}(x) = \varphi_1^{n_c}(y) \} : x, y \in X \right\} > 0.$$

Therefore $|X| > |\mathrm{Im}\varphi_1^{n_c}|$ at least with probability $q > 0$ if $|X| \geq 2$. Similarly, $|\mathrm{Im}\varphi_1^{n_c}| > |\mathrm{Im}\varphi_1^{2n_c}|$ with probability at least q^2 if the left set is no singleton. This holds because $\varphi_1^{2n_c} = \varphi_{n_c+1}^{2n_c} \circ \varphi_1^{n_c}$ and the variables $\varphi_1, \dots, \varphi_{n_c}$ and $\varphi_{n_c+1}, \dots, \varphi_{2n_c}$ are independent and identically distributed. By induction,

$$|X| > |\mathrm{Im}\varphi_1^{n_c}| > |\mathrm{Im}\varphi_1^{2n_c}| > \cdots > |\mathrm{Im}\varphi_1^{kn_c}|$$

at least with probability q^k until the last cardinality becomes 1; this happens after at most $|X| - 1$ steps. Let $\tau = (|X| - 1)n_c - 1$. Nothing changes if we renumber the maps as $\varphi_{-\tau}, \dots, \varphi_0$, $m < 0$. Hence $\mathbb{P}(\{|\mathrm{Im}\varphi_{-\tau}^0| = 1\}) \geq q^\tau$ and the lemma is proved. ∎

The next step is a sub-multiplicativity property of probabilities for coalescence times.

LEMMA 8 *Let* $n, m < 0$ *be negative integers. Then*

$$\mathbb{P}(T_0 \leq m + n) \leq \mathbb{P}(T_0 \leq m)\mathbb{P}(T_m \leq m + n) = \mathbb{P}(T_0 \leq m)\mathbb{P}(T_0 \leq n).$$

Proof. Suppose that $T_0(\varphi) \leq m + n$. This holds if and only if $\mathrm{Im}\varphi_{m+n+1}^0$ has more than one element. Then both, $\mathrm{Im}\varphi_{m+1}^0$ and $\mathrm{Im}\varphi_{m+n+1}^m$, have more than one element. Hence

$$\mathbb{P}\big(\varphi : T_0(\varphi) \leq m + n\big) \leq \mathbb{P}\big(\varphi : T_0(\varphi) \leq m \text{ and } T_m(\varphi) \leq m + n\big).$$

To check whether $T_0(\varphi) \leq m$ holds true it is sufficient to know the maps $\varphi_{m+1}, \ldots, \varphi_0$, and similarly, to check $T_m(\varphi) \leq n + m$ only $\varphi_{m+n+1}, \ldots, \varphi_m$ are needed. Hence the respective sets are independent and the inequality holds. The remaining identity follows from stationarity. ■ In combination with Theorem 4, the next result completes the derivation of exact sampling.

THEOREM 9 *The Conditions* (F) *and* (C) *are equivalent. In particular, the process governed by* $\mathbb{P}$ *is successful under* (C), *and almost sure coalescence in Theorem 4 is assured.*

Proof. Suppose that (C) holds. By Lemma 7, we have $\mathbb{P}\big(T_0 > -\tau\big) > 0$ and Lemma 8 implies

$$\mathbb{P}\big(T_0 \leq -n\tau\big) \leq \mathbb{P}\big(T_0 \leq -\tau\big)^n = \big(1 - \mathbb{P}\big(T_0 > -\tau\big)\big)^n \longrightarrow 0 \text{ as } n \to \infty.$$

By stationarity, this implies (F). Conversely, suppose that (F) holds, i.e. that $\mathbb{P}(F) = 1$. Since F is the intersection of the sets

$$F_n = \big\{\varphi : \text{ there is } m \leq n \text{ such that } |\mathrm{Im}\varphi_m^n| = 1\big\}$$

each of these sets has full measure 1 as well. Fix n now. Plainly, the sets

$$F_m^n = \big\{\varphi : |\mathrm{Im}\varphi_m^n| = 1\big\}$$

increase to F_n as m decreases to $-\infty$. Hence there is $m < n$ such that $\mathbb{P}(F_m^n) > 0$. Choose now $x \neq y$ in X. Since φ_1^{m-n+1} and φ_m^n are equal in law, for $\tau = n - m + 1$ one has

$$p^\tau\big(\{\varphi_1, \ldots, \varphi_\tau\} : \varphi_1^\tau(x) = \varphi_1^\tau(y)\big) = \mathbb{P}\big(\varphi : \varphi_m^n(x) = \varphi_m^n(y)\big) \geq \mathbb{P}\big(F_m^n\big) > 0,$$

and (C) holds. ■ This shows that any derivation of coupling from the past which does not explicitly or implicitly use a hypothesis like (C) or a suitable substitute is necessarily incomplete or incorrect.

REMARK   It is tempting to transfer the same idea to 'coupling to the future'. Unfortunately, starting at zero and returning the first state of complete coalescence after zero, in general does not give a sample from μ.

The reader may want to check the following simple example from [F. FRIEDRICH, 2003].

EXAMPLE 10 Let $X = \{1,2\}$. Positive transition probabilities P and their invariant distributions μ have the form

$$P := \begin{pmatrix} 1-\lambda & \lambda \\ \kappa & 1-\kappa \end{pmatrix}, \quad 0 < \lambda, \kappa < 1, \quad \mu = \left(\frac{\kappa}{\lambda+\kappa}, \frac{\lambda}{\lambda+\kappa} \right).$$

Start two independent chains η and ξ with transition probability P at time 0 from 1 and 2, respectively. The time of first coalescence in the future is

$$T := \min\{m \in \mathbb{N} : \eta_m = \xi_m\}.$$

Denote the common law of η_T and ξ_T by ϱ. We will shortly verify that $\varrho = \mu$ if and only if $\lambda = \kappa$. Compute first

$$\mathbb{P}\big(\eta_n = \xi_n = 1, \eta_m \neq \xi_m, m < n\big)$$

$$= \kappa(1-\lambda) \sum_{k=0}^{n} \binom{n}{k} \big((1-\lambda)(1-\kappa)\big)^k (\lambda\kappa)^{n-k}$$

$$= \kappa(1-\lambda)\big((1-\lambda)(1-\kappa) + \lambda\kappa\big)^n = \kappa(1-\lambda)\big(1 - (\lambda+\kappa - 2\lambda\kappa)\big)^n.$$

and

$$\varrho(1) = \kappa(1-\lambda) \sum_{n=0}^{\infty} \big(1 - (\lambda+\kappa - 2\lambda\kappa)\big)^n = \frac{\kappa(1-\lambda)}{\kappa(1-\lambda) + \lambda(1-\kappa)}.$$

Hence

$$\varrho = \left(\frac{\kappa(1-\lambda)}{\kappa(1-\lambda) + \lambda(1-\kappa)}, \frac{\lambda(1-\kappa)}{\kappa(1-\lambda) + \lambda(1-\kappa)} \right).$$

This is the invariant distribution μ if and only if $\lambda = \kappa$.

The representation of Markov chains by stochastic flows is closely connected to the actual implementation of coupling from the past. Extending previous notation, a *transition rule* will be a map $f : X \times \Theta \to X$, with some set Θ to be specified. Let now V_i, $i \in \mathbb{Z}$, be independent identically distributed random variables taking values in Θ. Then $\varphi_i = f(\cdot, V_i)$, $i \in \mathbb{Z}$, is a stochastic flow. If, moreover, $\mathbb{P}(f(x, V_i) = y) = P(x,y)$ then the flow fulfills Condition (P). The remaining problem is to construct a transition rule such that the associated flow fulfills Condition (C) too.

EXAMPLE 11 Recall from Example 1 how a Markov chain was realized there. Let again $f(x,u)$ be a deterministic transition rule taking values in X, such that for a random variable U with uniform distribution on $\Theta = [0,1]$ the variable $f(x,U)$ has law $P(x, \cdot)$. This way we - theoretically - may for an $m \leq 0$ realize all values $\varphi_m^0(x)$, $x \in X$, and check coalescence. If we go back k more steps in time we need all $\varphi_m^0 \circ \varphi_{m-k}^{m-1}(x)$. Since the maps $\varphi_0, \ldots, \varphi_m$ are kept,

we must work with the same random numbers $u_0, \ldots, u_m$, i.e. realizations of the $U_0, \ldots, U_m$, as in the preceding run, and only independently generate additional random numbers $u_{m-1}, \ldots, u_{m-k}$. For this special coupling there is complete coalescence at time 0 in finite time. The strength of coupling depends on the special form of f which in turn depends on the concrete implementation. In Example 1, for each $x \in X$, we partitioned [0,1] into intervals I_y^x of length $P(x, y)$ and in step n took that y with $U_n \in I_y^x$. The intervals $I_{y^*}^x$ with left end at 0 have an intersection I_{y^*} of length at least $\min_{x,y} P(x, y)$.

This simultaneously is the probability that U falls into I_{y^*} and all states coalesce in y^* in one single step, irrespective of x. We may improve coupling by a clever arrangement of the intervals. If we put the intervals $I_{y^*}^x$ for which
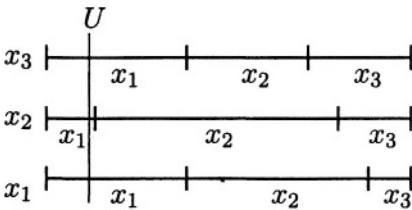
$\min\{|I_{y^*}^x| : x \in X|\}$ is maximal, to the left end of [0,1] then we get the lower bound $\max_y \min_x P(x, y)$ for the coalescence probability. We can improve coupling even further, splitting the intervals into pieces of length $\min\{|I_y^x| : x \in X\}$ and their rest, and arrange the equal pieces on the left of [0,1]. This gives a bound $\sum_y \min_x P(x, y)$.

Note that although all these procedures *realize the same Markov kernel P* they correspond to *different transition rules,* to *different stochastic flows,* and to *different couplings.* Apart from all these modifications, we can summarize:

PROPOSITION 9 *Suppose that $P > 0$. Then all stochastic flows $\varphi_i = f(\cdot, U_i)$ from the present Example 11 fulfill Condition* (C).

Note that the distribution of all these random maps definitely is not the synchronous one from Example 2. For this distribution, set $\Theta = [0, 1]^{|X|}$, use independent copies $U_k^z, z \in X$, of U_k, and let $\varphi_k(x) = f(x, (U_k^z)_{z \in X}) = g(x, U_k^x)$ for g on $X \times [0,1]$ constructed like above. Condition (C) is obviously fulfilled and coupling from the past works also for this method.

REMARK In Example 11 we found several lower bounds for the probability that states coalesce in one step. An upper bound is given by

$$\mathbb{P}\big(\varphi(x) = \varphi(y)\big) = \sum_z \mathbb{P}\big(\varphi(x) = z, \varphi(y) = z\big)$$

$$\leq \sum_z \mathbb{P}\big(\varphi(x) = z\big) \wedge \mathbb{P}\big(\varphi(y) = z\big) = \sum_z P(x, z) \wedge P(y, z).$$

This is closely related to DOBRUSHIN'S contraction technique, which in the finite case is based on *Dobrushin's contraction coefficient $c(P) = 1 - \sum_z P(x, z) \wedge P(y, z)$*, cf. [G. WINKLER, 1995], Chapter 4. The relation is

$$\mathbb{P}\big(\varphi(x) = \varphi(y)\big) \leq 1 - c(P).$$

This upper bound is not sharp.

6.3 Monotonicity

Checking directly whether there is complete coalescence at time 0 starting at more and more remote past times and at all possible states is time consuming, and even impossible if the state space is large (as it is in the applications we have in mind). If coalescence of very few states enforces coalescence of all other states then the procedure becomes feasible. One of the concepts to make this precise is *monotonicity.* We are now going to introduce this concept on an elementary level.

DEFINITION 12 *A **partial order** on a set X is a relation $x \preceq y$ between elements $x, y \in X$ with the two properties*

(i) $x \preceq x$ *for each $x \in X$ (**reflexivity**)*

(ii) $x \preceq y$ *and $y \preceq z$ implies $x \preceq z$ (**transitivity**).*

Recall that a *total order* requires the additional condition that any two elements $x, y \in X$ are *comparable,* i.e $x \preceq y$ or $y \preceq x$.

EXAMPLE 13 (a) The usual relation $x \leq y$ on $\mathbb{R}$ is a *total order.* In the *component-wise order* on $\mathbb{R}^d$, $(x_1, \ldots, x_d) \preceq (y_1, \ldots, y_d)$ if and only if $x_i \leq y_i$ for each i. It is a partial but no total order since elements like $(0,1)$ and $(1,0)$ are not related, (b) If $X = \{\pm 1\}^S$, then in the component-wise order from (a), the constant configurations $b \equiv 1$ and $w \equiv -1$ are *maximal* and *minimal,* respectively, i.e. $x \preceq b$ and $w \preceq x$ for every $x \in X$. This will be exploited in exact sampling for the Ising field in Section 6.4.

Next we want to lift partial orderings to the level of probability distributions. Call a subset I of X an *order ideal* if $x \in I$ and $y \preceq x$ imply $y \in I$.

EXAMPLE 14 (a) The order ideals in $\mathbb{R}$ with the usual order are the rays $(-\infty, u]$ and $(-\infty, u)$, $u \in \mathbb{R}$.
(b) In the binary setting of Example 13(b), $x \preceq y$ if each black pixel of x is also black in y (if we agree that $x_s = +1$ means that the colour of pixel s is black). The order ideals are of the form $\{x \in X : x \preceq y\}$.

DEFINITION 15 *Let $(X, \preceq)$ be a finite partially ordered set, and let ν and μ be probability distributions on X. Then $\nu \preceq \mu$ in **stochastic order,** if and only if $\nu(I) \geq \mu(I)$ for each order ideal I.*

EXAMPLE 16 Let ν and μ be distributions on $\mathbb{R}$ with cumulative distribution functions F_ν and F_μ, respectively. Then $\nu \preceq \mu$ if and only if $\mu((-\infty, u]) \leq \nu((-\infty, u])$ if and only if $F_\mu(u) \leq F_\nu(u)$ for every $u \in \mathbb{R}$.

This means that 'the mass of ν is more on the left than the mass of μ'. For Dirac distributions $\varepsilon_u \preceq \varepsilon_v$ if and only if $u \leq v$.

The natural extension to Markov kernels reads

DEFINITION 17 *We call a Markov kernel P on a partially ordered space* $(X, \preceq)$ **stochastically monotone,** *if and only if* $P(x, \cdot) \preceq P(y, \cdot)$ *whenever* $x \preceq y$.

In Example 11 we constructed transition rules f for homogeneous Markov chains, or rather Markov kernels P. A transition rule is called *monotone* if $f(x, u) \preceq f(y, u)$ for each u whenever $x \preceq y$. Plainly, a monotone transition rule induces a monotone Markov kernel. Conversely, a monotone kernel is not necessarily induced by a monotone transition rule, even in very simple situations. [D.A. Ross, 1993], see [J.A. FILL & M. MACHIDA, 2001], p. 2., gives a simple counterexample:

EXAMPLE 18 Consider the space $X = \{u, v, a, b\}$ and let $u \preceq a, b$, and $a, b \preceq v$. Define a Markov kernel P by
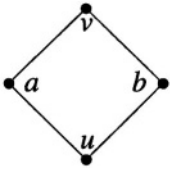
$$P(u, u) = 1/2 = P(u, a), \quad P(a, u) = 1/2 = P(a, v)$$
$$P(b, a) = 1/2 = P(b, b), \quad P(v, a) = 1/2 = P(v, v)$$

The order ideals are $\emptyset, \{u\}, \{a, u\}, \{b, u\}$ and X, and it is readily checked that P is monotone. Suppose now that there are random variables with $\xi_u \preceq \xi_a, \xi_b \preceq \xi_v$ almost surely and with laws $P(u, \cdot), P(a, \cdot), P(b, \cdot)$, and $P(v, \cdot)$, respectively. We shall argue that

$$\mathbb{P}(\xi_u = a) = \mathbb{P}(\xi_u = a, \xi_a = v, \xi_b = a, \xi_v = v) = 1/2$$
$$\mathbb{P}(\xi_b = b) = \quad \mathbb{P}(\xi_u = u, \xi_b = b, \xi_v = v) \quad = 1/2.$$

The two events are disjoint and hence $\mathbb{P}(\xi_v = v) = 1$ in contradiction to $\mathbb{P}(\xi_v = v) = 1/2$. We finally indicate how for example the first identity can be verified: Since $\xi_u \preceq \xi_a$ one has $\mathbb{P}(\xi_u = a) = \mathbb{P}(\xi_u = a, \xi_a \in \{a, v\})$. Since $\mathbb{P}(\xi_a = a) = 0$, we conclude $\mathbb{P}(\xi_u = a) = \mathbb{P}(\xi_u = a, \xi_a = v)$. Now repeat this argument two times.

Suppose now that the partially ordered space $(X, \preceq)$ contains a *minimal element* u and a *maximal element* v, i.e. $u \preceq x \preceq v$ for every $x \in X$. Suppose further that the stochastic flow is induced by a monotone transition rule, i.e. $\varphi_i(x) = f(x, U_i)$ and $f(x, u) \preceq f(y, u)$ if $x \preceq y$. Then

$$\varphi_m^n(u) \preceq \varphi_m^n(x) \preceq \varphi_m^n(v) \text{ for every } x \in X, \ m \leq n,$$

and $\varphi_m^0(x) = w$, $m \leq 0$, for each $x \in \boldsymbol{X}$, as soon as $\varphi_m^0(u) = w = \varphi_m^0(v)$. The previous findings can be turned into practicable algorithms.

PROPOSITION 10 *Suppose that P is monotone and* $(\boldsymbol{X}, \preceq)$ *has a minimum* $\boldsymbol{u}$ *and maximum* $\boldsymbol{v}$. *Then coalescence for* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *enforces complete coalescence.*

6.4 Random Fields and the Ising Model

Random fields serve as flexible models in image analysis and spatial statistics. In particular, any full probabilistic model of textures with random fluctuations necessarily is a random field. Recursive (auto-associative) neural networks can be reinterpreted in this framework as well, cf. e.g. [G. WINKLER, 1995]. To understand the phenomenology of these models, sampling from their *Gibbs distribution* provides an important tool. In the sequel we want to show how the concepts developed above serve to establish exact sampling from the Gibbs distribution of a well known random field - the Ising model.

Let a *pattern* or *configuration* be represented by an array $x = (x_s)_{s \in S}$ of 'intensities' $x_s \in G_s$ in 'pixels' or 'sites' $s \in S$ with finite sets G_s and *S*. *S* might be a finite square grid or - in case of neural networks - an undirected finite graph. A (finite) *random field* is a strictly positive probability measure Π on the space $\boldsymbol{X} = \prod_{s \in S} G_s$ of all configurations x. Taking logarithms shows that Π is of the *Gibbsian form*

$$\Pi(x) = Z^{-1} \exp(-K(x)), \ Z = \sum_z \exp(-K(z)), \qquad (4.1)$$

with a function *K* on *X*. It is called a *Gibbs fields* with *energy function K* and *partition function Z*. These names remind of their roots in statistical physics.

For convenience we restrict ourselves to the Gibbs,sampler with random visiting scheme. Otherwise we had slightly to modify the setup of Section 6.2. Let pr_t be the projection $\boldsymbol{X} \rightarrow G_t$, $x \mapsto x_t$. For a Gibbs field Π let

$$\Pi(y_s \mid x_t, t \neq s) = \Pi(pr_s = y_s \mid pr_t = x_t, t \neq s) \qquad (4.2)$$

denote the single-site conditional probabilities. The *Gibbs sampler with random visiting scheme* first picks a site $s \in S$ at random from a probability distribution *D* on *S*, and then picks an intensity at random from the conditional distribution (4.2) on G_s. Given a configuration $x = (x_t)$ this results in a new configuration $y = (y_t)$ which equals x everywhere except possibly at site s. The procedure is repeated with the *new* configuration y, and so on and so on. This defines a homogeneous Markov chain on *X* with Markov kernel

$$P(x, y) = \sum_{s \in S} D(s)\Pi_{\{s\}}(x, y), \ \ x, y \in \boldsymbol{X}, \qquad (4.3)$$

where $\Pi_{\{s\}}(x,y) = \Pi(y_s \mid x_t, t \neq s)$ if x and y are equal off s and
$\Pi_{\{s\}}(x,y) = 0$ otherwise. These transition probabilities $\Pi_{\{s\}}$ are called the
local characteristics. D is called the *proposal* or *exploration distribution.*

We assume that D is strictly positive; frequently it is the uniform distri-
bution on S. Then P is primitive since $P^{|S|}$ is strictly positive. In fact, in
each step each site and each intensity in the site has positive probability to be
chosen, and thus each y can be reached from each x in $|S|$ steps with positive
probability. It is easily checked - verifying the detailed balance equations - that
Π is the invariant distribution of $P,$ and thus the invariant distribution of the
homogeneous Markov chain generated by $P.$

EXAMPLE 19 (THE ISING MODEL) Let us give an example for exact sam-
pling by way of the Ising model. *The ferromagnetic Ising model with magnetic
field* $h := (h_s)_{s\in S}$ is a binary random field with $G_s = \{-1,1\}$ and energy
function

$$K(x) = \beta \sum_{s\sim t} x_s x_t - \sum_s h_s x_s,$$

where $\beta > 0, h_s \in \mathbb{R}$ and $s \sim t$ indicates that s and t are neighbours. For
the random visiting scheme in (4.3) the Markov chain is homogeneous and fits
perfectly into the setting of Section 6.2. The formula from [G. WINKLER,
1995], Proposition 3.2.1 (see also [G. WINKLER, 1995], Example 3.1.1) for
the local characteristics boils down to

$$p^+(x) = \Pi\big(X_s = 1 \mid X_t = x_t,\ t \neq s\big) = \Big(1 + \exp\big(-2\beta \sum_{t\sim s} x_t - h_s\big)\Big)^{-1}.$$

This probability increases with the set $\{t \in S : x_t = 1\}$. Hence
$p^+(y) \geq p^+(x)$ if $x \preceq y$ in the component-wise partial order introduced in
Example 13. The updates x' and y' preserve all the black sites off $s,$ and pos-
sibly create an additional black one at $s.$ We conclude that P from (4.3) is
monotone and fulfills the hypotheses of Proposition 10. Hence for complete
coalescence one only has to check whether the completely black and the com-
pletely white patterns coalesce. For transition rules like in Example 11 the
Condition (C) on page 152 is also fulfilled and coupling from the past works.

6.5   Conclusion

The authors are not aware of other mathematical fields, where so many in-
sufficient arguments, ranging from incomplete or misleading, to completely
wrong, have been published (mainly in the Internet). In particular, Condi-
tion (C) or a substitute for it, are missing in a lot of presently available texts.
A rigorous treatment is [S.G. FOSS & R.L. TWEEDIE, 1998]. These au-
thors do not use iterated random maps. These are exploited systematically in
[P. DIACONIS & D. FREEDMAN, 1999]. [J.A. FILL, 1998] introduces 'in-

terruptible' perfect sampling based on acceptance/rejection sampling. Meanwhile there is a body of papers on exact sampling. On the other hand, the field still is in the state of flux and hence it does not make sense to give further references; a rich and up to date source is the home-page of D.B. WILSON, http://www.dbwilson.com/exact/. The connection between transition probabilities and random maps was clarified in [H.V. WEIZSCKER, 1974].

Acknowledgment

We thank H.V. WEIZSCKER, Kaiserslautern, for helpful discussions during the initial phase of the work.

References

P. Diaconis and D. Freedman. Iterated random functions. *SIAM Rev.,* 41(l):45–76, 1999.

J.A. Fill. An interuptible algorithm for perfect sampling via Markov chains. *The Ann. of Appl. Probab.* 8(1)*:*131–162*, 1998.*

J.A. Fill and M. Machida. Stochastic monotonicity and realizable monotonicity. *Ann. Probab.,* 29:938–978, 2001.

S.G. Foss and R.L. Tweedie. Perfect simulation and backward coupling. *Stoch. Models,* 14(1-2): 187–204, 1998.

F. Friedrich. *Sampling and statistical inference for Gibbs fields.* PhD thesis, University of Heidelberg, Munich, Germany, 2003. draft.

A. Gelman. Inference and monitoring convergence. In [W.R. GILKS ET AL., 1996b], chapter 8, pages 131-143.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. Introducing Markov chain Monte Carlo. In [W.R. GILKS ET AL., 1996b], chapter 1, pages 1–19.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice.* Interdisciplinary Statistics. Chapman *&* Hall, London, Weinheim, New York, Tokyo, Melbourne, Madras, 1996b.

J.G. Propp and D.B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms,* 9:223–252, 1996.

A.E. Raftery and S.M. Lewis. Implementing MCMC. In [W.R. GILKS ET AL., 1996b], chapter 7, pages 115–130.

D.A. Ross. A coherence theorem for ordered families of probability measures on a partially ordered space. Unpublished manuscript, 1993.

H.v. Weizsäcker. Zur Gleichwertigkeit zweier Arten der Randomisierung. *Manuscripta Mathematika,* 11:91–94, 1974.

G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods,* volume 27 of *Applications of Mathematics.* Springer Verlag, Berlin, Heidelberg, New York, 1995.

G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods,* volume 27 of *Applications of Mathematics.* Springer Verlag, Berlin, Heidelberg, New York, second edition, 2003.

This page intentionally left blank

ON THE DIFFERENT EXTENSIONS OF THE ERGODIC THEOREM OF INFORMATION THEORY

Valerie Girardin

Mathématiques, Campus II, Université de Caen, BP 5186, 14032 Caen, France

girardin@math.unicaen.fr

Abstract   The purpose of this paper is to review the different generalizations and extensions of the ergodic theorem of information theory in terms of reference measure, state space, index set and required properties (ergodicity, stationarity, etc.) of the process, from the original Shannon-McMillan-Breiman version to its latest developments.

7.1   Introduction

The statement of convergence of the entropy at time n of a random process divided by n to a constant limit called the entropy rate of the process is known as the ergodic theorem of information theory or asymptotic equirepartition property (AEP). Its original version proven in the 50's for ergodic stationary processes with a finite state space, is known as Shannon-McMillan theorem for the convergence in mean and as Shannon-McMillan-Breiman theorem for the almost sure convergence.   Since then, numerous extensions have been made in direction of weakening the hypothesis on the reference measure (from the counting or product measure to Markovian or semi-Markovian measures), state space (from a finite set to any Borel set), index set (from discrete-time to continuous-time, product sets or groups) and required properties (ergodicity, stationarity, etc.) of the process.

The purpose of this paper is to review these different generalizations and extensions. Some necessary basics are given in Section 7.2 concerning entropy definition, ergodicity, stationarity and Markovian measures and processes. General statement and applications of the AEP are given too. The original AEP with hints of proof is presented in Section 7.3.1. Extensions in terms of state

space are considered in Section 6.4, in terms of measures in Section 7.3.3 and to continuous-time processes in Section 7.3.4. Finally, explicit expressions of the entropy rate for Markovian and Gaussian processes are given in Section 7.4.

7.2   Basics

7.2.1   Definition of entropy

The concept of entropy is the basis of information theory. It has first been introduced in the field of probability by Boltzman in the XIX–th century in statistical mechanics and then by Shannon (1948) for studying communication systems.

DEFINITION 1 *The entropy of a probability distribution P with density p with respect to a reference measure μ is defined as Boltzman's H-function, that is to say*

$$\mathbb{S}(P) = -\int p(x) \log p(x) dx = H_p,$$

with the convention $0 \log 0 = 0$.

It inspired Shannon (1948) to define and study the entropy of a discrete distribution taking n values as

$$\mathbb{S}(P) = -\sum_{i=1}^{n} p_i \log p_i.$$

The function $\mathbb{S}$ has interesting properties as measure of uncertainty in communication theory, see Reza (1961), Ash (1965), Cover & Thomas (1991). Actually, in the continuous case, these properties cannot be derived from the discrete case. For example, the entropy of the uniform distribution U on an interval $[a, b]$ equals $\log(b - a)$ but the entropy of the uniform distribution U on a partition in n values of the same interval equals $\log n$. The link between these two separate notions was made by Kullback & Leibler (1951), see also Kullback (1978).

DEFINITION 2 *Let P and Q be two distributions on the same measurable space. The Kullback-Leibler information of P relative to Q is defined as*

$$\mathbb{S}(P \mid Q) = \sum_{i} p_i \log \frac{p_i}{q_i},$$

for discrete distributions and as

$$\mathbb{S}(P \mid Q) = \mathbb{E}_P \left(\log \frac{dP}{dQ} \right) = \int \log \frac{dP(x)}{dQ(x)} dP(x),$$

if P is absolutely continuous with respect to Q (and as +∞ if not).

The definition can be extended to two positive measures on the same measurable space.

In both discrete and absolutely continuous cases, we have

$$\mathbb{S}(P \mid Q) = \sup \sum_{i=1}^{n} P(A_i) \log \frac{P(A_i)}{Q(A_i)},$$

where the supremum is taken on all finite partitions of the space, and

$$\mathbb{S}(P) = \mathbb{S}(U) - \mathbb{S}(P \mid U).$$

The meaning of entropy appears thus as well in information theory as in statistical mechanics. In the former, it measures the variation of information from the uniform distribution to *P*, hence has a meaning as a measure of uncertainty of the system. In the latter, a system is in equilibrium if the probability density (or number of particles in an infinitesimal volume) is close to the uniform repartition.

The entropy methods can also be justified by purely probabilistic or statistic arguments (large deviations principle, Bayesian statistics, properties of the induced estimates, etc.), see Csizár (1996), Garret (2001), Grendar & Grendar (2001), and particularly for Markov chains, Moran (1961).

Basic properties of entropy and links with communication theory are given in Girardin & Limnios (200la). For a detailed study, see Reza (1961), Ash (1965), Guiasu (1977), Cover & Thomas (1991).

DEFINITION 3 *The entropy at time n of a discrete-time stochastic process* **X** = $(X_n)_{n \in \mathbb{N}}$ *taking values in* $(E, \mathcal{E})$ *is by definition the entropy of its n-dimensional marginal distribution, namely*

$$\begin{aligned}
\mathbb{H}_n(\mathbf{X}) &= -\mathbb{E}[\log p_n^{\mathbf{X}}(x_1, \ldots, x_n)] \\
&= -\int p^{\mathbf{X}}(x_1, \ldots, x_n) \log p^{\mathbf{X}}(x_1, \ldots, x_n) d\mu_n(x_1, \ldots, x_n),
\end{aligned}$$

where $p^{\mathbf{X}}$ *is the density of the random vector* $X = (X_1, \ldots, X_n)$ *with respect to the n-th marginal* μ_n *of a reference measure* μ *on the infinite product space* $(E^{\mathbb{N}}, \mathcal{E}^{\mathbb{N}})$.

The entropy at time n is a nondecreasing nonnegative function of n. It can also be seen as the Kullback information $\mathbb{H}_n(\mathbf{X} \mid \mathbf{Y})$ of the marginal distribution of **X** relative to the marginal distribution μ_n of a process **Y**, also called relative entropy between **X** and **Y**. For this point of view, see especially Pinsker (1960) and Perez (1964).

Under suitable conditions, the entropy at time n divided by n converges,

$$\frac{\mathbb{H}_n(\mathbf{X})}{n} \longrightarrow \mathbb{H}(\mathbf{X}), \quad n \to +\infty. \tag{2.1}$$

If the limit $\mathbb{H}(\mathbf{X})$ (or $\mathbb{H}(\mathbf{X} \mid \mathbf{Y})$) exists and is finite, it is called the Shannon entropy rate of the process and we have $\mathbb{H}(\mathbf{X}) = \inf \mathbb{H}_n(\mathbf{X})/n$.

For simplification, let us set $x_n^m = (x_n, \ldots, x_m)$. Set $h_n(x) = -\log p_n(x_1^n)$ and let $g_{m,n}(x) = p_n(x_m^n)/p_n(x_m^{n-1})$ be the conditional density of X_n relative to $(X_m, \ldots, X_{n-1})$. If $-\mathbb{E} \log g_{0,n}(X)$ converges to some limit, then $\mathbb{H}_n(\mathbf{X})/n$ is the Cezáro's sum of the sequence and hence converges to the same limit. The entropy rate is sometimes defined in this way, see for example Reza (1961).

The convergence in (2.1) appears as the consequence of the convergence in mean of the sequence of random variables $(-\log p_n(X_1, \ldots, X_n)/n)$. The almost sure convergence is also of interest. They constitute together the ergodic theorem of information theory also called Shannon-McMillan-Breiman theorem, or Asymptotic Equirepartition Property.

THEOREM 1 (ERGODIC THEOREM OF INFORMATION THEORY)
Under suitable conditions on the process $\mathbf{X}$, *its index-space* $\mathbb{T}$, *its state space* E *and the reference measure* μ, *the sequence* $(-\log p_n(X_1, \ldots, X_n)/n)$ *converges in mean or almost surely to the entropy rate of the process.*

In the following, for simplification, we will call mean AEP the convergence in mean and strong AEP the almost sure convergence.

Similarly, for continuous-time processes, we get the following definition.

DEFINITION 4 *The entropy at time T of a continuous-time process* $\mathbf{X} = (X_t)_{t\in\mathbb{R}_+}$ *is defined as*

$$\mathbb{H}_T(\mathbf{X}) = -\int p_T(x) \log p_T(x) d\mu_T(x),$$

where $p_T(x)$ *is the likelihood of* $(X_t)_{0\le t\le T}$ *with respect to the restriction* μ_T *to* [0, T] *of some reference measure* μ.

The definition of the entropy rate and the statement of the corresponding AEP derive immediately.

This theorem has many applications. Let us list some of them.

First of all, the application which made M. McMillan call it AEP. The typical set of a process is defined as the set of sequences $(x_1, \ldots, x_n)$ such that

$$2^{-n\mathbb{H}(\mathbf{X})+\varepsilon} \le p_n(x_1^n) \le 2^{-n\mathbb{H}(\mathbf{X})-\varepsilon},$$

and a sequence $(X_1, \ldots, X_n)$ is said to be typical if its density satisfies the above relation. From the mean AEP for a finite state space, the probability of the typical set is proven to be nearly one; all its elements are nearly equiprobable and this set contains nearly $2^{n\mathbb{H}(\mathbf{X})}$ elements, see Cover & Thomas (1991) (with application to data compression). For a Borel space, the distribution of $(X_1, \ldots, X_n)$ is proven to be asymptotically uniform on the typical set, which has the least asymptotic volume (equal to $2^{-n\mathbb{H}(\mathbf{X})}$) among sets of high probability. And from the almost sure convergence, the sequences $(X_1, \ldots, X_n)$ are proven to be almost surely typical for large n, see Barron (1985).

The AEP has thus a prominent role in information theory together with the linked Shannon channel coding theorem. This theory has been presented in many books since the original exposition of Shannon (1948), see for example Khinchin (1957), Feinstein (1958), Gallager (1968) and more recently Guiasu (1977), Cover & Thomas (1991).

It also plays a role in finance, see for example Algoet & Cover (1988b) and Algoet (1994) and the reference therein.

Many applications of the maximum entropy methods involving the entropy rate exist in the literature, see, e.g., Girardin (2002) and the references therein for applications involving Markov chains and processes.

Application to statistical inference derives too, for example through likelihood maximization, often equivalent to maximization of the entropy rate.

Large deviations results derive too. See Gallager (1968) or Cover & Thomas (1991) for results in information theory, and Ellis (1985) for a statistical mechanics point of view.

Linnik (1959) initiated the use of entropy for proving limit theorems in a proof of the central limit theorem. For other examples and recent developments, see Johnson (1999) and the references therein.

7.2.2   Ergodicity, stationarity and Markov properties

The AEP involves the notions of ergodicity and stationarity. Let us recall their definitions.

DEFINITION 5 *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space. A process* (X_n) *taking values in E can be defined as* $X_n(\omega) = X(S^n\omega)$, *where X is a random variable with values in E and S is a shift from* Ω *to itself.*

The shift (and thus the process) is said to be

- *stationary if* $\mathbb{P}(SA) = \mathbb{P}(A)$ *for all* $A \in \mathcal{A}$;

- *ergodic if SA = A implies* $\mathbb{P}(A) = 0$ *or 1.*

For an ergodic shift, the strong law of large numbers takes the form

$$\frac{1}{n}\sum_{k=0}^{n-1}\mathbf{1}_A(S^k\omega) \longrightarrow \mathbb{P}(A), \quad A \in \mathcal{A},$$

and implies the following Birkhoff (or individual) ergodic theorem, see Doob (1953).

THEOREM 2 *If T is an ergodic shift of* $(\Omega, \mathcal{A}, \mathbb{P})$ *and F is an integrable function on* Ω, *then*

$$\frac{1}{n}\sum_{k=0}^{n-1}F(S^k\omega) \longrightarrow \mathbb{E}F, \quad \textit{a.s. and in mean.}$$

The following extension is due to Breiman (1957).

THEOREM 3 *If* (F_k) *is a uniformly* L^1-**bounded** *(i.e., such that* $\mathbb{E}\sup_k F_k < +\infty$*) sequence of measurable functions converging almost surely to some function F, then*

$$\frac{1}{n}\sum_{k=0}^{n-1}F_k(S^k\omega) \longrightarrow \mathbb{E}F, \quad \textit{a.s..}$$

Stationarity and ergodicity can equivalently be defined by considering the state probability space, i.e., $E^{\mathbb{N}}$ endowed with the σ-algebra $\mathcal{E}^{\mathbb{N}}$ and the law of the process, say $\mathbb{P}_{\mathbf{X}}$, defined on $(E^{\mathbb{N}}, \mathcal{F} = \mathcal{E}^{\mathbb{N}})$. The process is stationary or ergodic if the translation shift θ defined by $(\theta x)_n = x_{n+1}$ (where $x = (x_0, \ldots, x_n, \ldots)$) is thus for $\mathbb{P}_{\mathbf{X}}$. The ergodic theorems involve then $f(\theta^k x)$ instead of $f(T^k\omega)$ for any integrable function f defined on $E^{\mathbb{N}}$.

The same notions can be defined and considered for a continuous-time process $\mathbf{X} = (X_t)_{t \in \mathbb{T}}$, with a group of shifts $\{T^t, t \in \mathbb{T}\}$, or a translation of the state space $E^{\mathbb{T}}$.

The entropy rate can also be defined in terms of shifts of the finite measure space $(\Omega, \mathcal{A}, \mathbb{P})$ as follows, see Billingsley (1978) –giving many connections between information theory and ergodic theory, or Guiasu (1977). The entropy of a finite σ-field $\mathcal{B} \subset \mathcal{A}$ is defined as

$$h(\mathcal{B}) = \sum_{B \in \mathcal{B}}\mathbb{P}(B)\log\mathbb{P}(B),$$

the entropy of $\mathcal{B}$ relative to a shift S is

$$h(S, \mathcal{B}) = \underline{\lim}_{n \to +\infty}\frac{1}{n}h\Big(\bigvee_{k=0}^{n-1}S^{-k}\mathcal{B}\Big),$$

where $\mathcal{B} \vee \mathcal{B}'$ denotes the σ-field generated by $\mathcal{B} \cup \mathcal{B}'$, and finally the entropy (rate) of S is

$$h(S) = \sup_{\mathcal{B} \subset \mathcal{A}} h(S, \mathcal{B}).$$

See Krengel (1967) for a generalization to σ-finite measure spaces.

Markovian measures and processes play a prominent part in entropy theory. Let us recall some definitions.

The theory of Markov processes and its extensions was initiated by A. Markov (1856-1922) through the property which bears his name: the future of a Markov process depends on its past only through its present, see Anderson (1991). Several generalizations were proposed since, all of them in the aim of weakening the Markov property, as for example the semi-Markov processes introduced by P. Lévy (1954) and W. Smith (1955). The latter generalize in a natural way the pure jump Markov processes and the renewal processes. The future evolution of a semi-Markov process depends on its present state and on the time elapsed since the latest transition, while the evolution of a pure jump Markov process depends only on its present state, see Limnios & Oprişan (2001).

DEFINITION 6 *A probability measure on the product space* $(E^{\mathbb{N}}, \mathcal{F})$ *is Markovian of order* $r \in \mathbb{N}$ *if*

$$\mu(x_n \in F \mid x_m, \dots, x_{n-1}) \;=\; \mu(x_n \in F \mid x_{n-r}, \dots, x_{n-1}),$$
$$F \in \mathcal{E}, \; m < n - r \in \mathbb{N}.$$

It is homogeneous (or has stationary transition probabilities) if

$$\mu(x_n \in F \mid x_{n-1}) = \theta^n \mu(x_1 \in F \mid x_0), \quad a.s., \; F \in \mathcal{E}, \; n \in \mathbb{N}.$$

A Markovian measure of order one is just said to be Markovian. If the order is zero, the measure is just the product of independent measures.

A process whose distribution is a Markovian measure is a Markov chain or discrete Markov process, $P = (\mathbb{P}(X_k = j \mid X_{n-1} = j))_{i,j \in E}$ is its transition kernel and a probability π such that $\pi P = \pi$ is its stationary distribution. General continuous-time Markov processes are defined in a similar way.

Continuous-time jump Markov processes can also be seen as special semi-Markov processes. The definition of semi-Markov processes is easier in terms of Markov renewal processes, here only with a countable state space.

DEFINITION 7 *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *denote a probability space and let E be a finite or countable set. A process* $(J_n, S_n)_{n \geq 0}$ *is a Markov renewal process with*

semi-Markov kernel $Q(t) = (Q_{ij}(t); i, j \in E)$, *for* $t \geq 0$, *if*

$$\mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_1, \ldots, J_{n-1}, J_n = i, S_1, \ldots, S_n) =$$
$$= \mathbb{P}(J_{n+1} = j, S_{n+1} - S_n \leq t \mid J_n = i) = Q_{ij}(t).$$

The process (J_n) is an *E-valued* Markov chain with transition kernel $P = (P(i, j))_{i,j \in E}$, where $P(i, j) = \lim_{t \to +\infty} Q_{ij}(t)$, see for example Limnios & Oprişan (2001). And the times $0 = S_0 \leq \cdots \leq S_n \leq \cdots$ are the $\mathbb{R}_+$-valued jump times of the corresponding *E*-valued semi-Markov process $\mathbf{X} = (X_t)_{t \geq 0}$ defined by

$$X_t = J_n \text{ if } S_n \leq t < S_{n+1}. \tag{2.2}$$

A jump Markov process is a semi-Markov process with semi-Markov kernel $Q_{ij}(t) = a_{ij}(1 - e^{-a_i t})/a_i$, where $a_i = -\sum_{j \neq i} a_{ij}$. The matrix $A = (a_{ij})$ is called its infinitesimal generator and a probability $\pi = (\pi_i)$ such that $\pi A = 0$ is called its stationary distribution.

7.3   The theorem and its extensions

First, let us see conditions for (2.1) to hold. If E is finite, $\mathbb{E}[h_n(X)]/n$ is bounded and hence by Fatou's lemma $h(X) = \underline{\lim} h_n(X)/n$ is integrable. If $\mathbf{Y} = \theta \mathbf{X}$, then, by entropy properties, $h_n(X) \leq h_n(Y)$. If X is stationary, then $\mathbb{E}h(X) = \mathbb{E}h(Y)$ so $h(X) = h(Y)$ and h e nh(X) s aninvariantfinite random variable. Thus, the limit of $\mathbb{H}_n(\mathbf{X})/n$ is a random variable which is invariant by θ; the ergodicity of the process ensures that almost surely this entropy rate is constant. If E is not finite, finiteness of $\mathbb{H}(\mathbf{X})$ (or equivalently up-boundedness of the sequence $\mathbb{H}_n(\mathbf{X})/n$) will be a necessary condition for the AEP to hold.

7.3.1   The original AEP

Shannon (1948) stated the convergence in probability of $(-\log p_n(X_1^n)/n)$ for ergodic finite processes, and proves it for i.i.d. sequences and for Markov chains, using the law of large numbers.

McMillan (1953) proved the convergence in mean for stationary ergodic processes with a finite state space. This constitutes the Shannon-McMillan theorem. He writes

$$-\log p_n(x_0^{n-1}) = -\frac{1}{n} \sum_{k=0}^{n-1} \log g_{0,k}(\theta^k x),$$

which is the basis of proof of most of the different extensions of the AEP. Here, the reference measure is the counting measure, as for all finite or countable

valued discrete-time process, and so $p_n(x_0^{n-1}) = \mathbb{P}(X_0 = x_0, \ldots, X_{n-1} = x_{n-1})$. He uses a martingale argument to prove the almost sure convergence of $(-\log g_{0,n}(\mathbf{X}))$ to a limit $-\log g(\mathbf{X})$ and derives its convergence in mean from the finiteness of E. The AEP is then given by the mean ergodic theorem applied to $-\log g(\mathbf{X})$. Gallager (1968) gave a simpler proof avoiding martingale arguments.

The almost sure convergence proven by Breiman (1957,1960) constitutes the Shannon-McMillan-Breiman theorem, also called ergodic theorem of information theory or strong AEP. He proves the almost sure convergence of $(-\log g_{0,n}(X))$ as a nonnegative lower semi-martingale and uses then Theorem 3. See also Shields (1987) for an alternative proof using a sample path covering argument.

author	date	limit	**X**	T	E	reference measure
Shannon	(1948)	probability	Markov	$\mathbb{N}$	finite	counting
McMillan	(1953)	mean	ergodic	$\mathbb{N}$	finite	counting
Breiman	(1957,60)	a.s.	erg. station.	$\mathbb{N}$	finite	counting

Fig. 1: The original AEP.

7.3.2   Extensions in terms of state space

The extension to a countable state space was made by Carleson (1958) for the convergence in mean (see also Parthasarathy (1964) for a simple proof) and by Chung (1961) for the almost sure convergence by proving that the uniform L^1 boundedness of the sequence $(-\log g_{0,n}(X))$ still holds in this case provided that the entropy rate $\mathbb{H}(\mathbf{X})$ is finite.

Perez (1957) made the first extension of the theorem to an arbitrary state space. He proved that if μ is the infinite product measure and X is stationary, then under finiteness of $\mathbb{H}(\mathbf{X})$, convergence in mean of $(h_n(X)/n)$ holds.

Moy (1960,1961) extended it to a homogeneous Markovian measure μ, first under finiteness of $\mathbb{H}_2(\mathbf{X})$, and then under finiteness of $\mathbb{H}_1(\mathbf{X})$ and upboundedness of the sequence $-\mathbb{E}\log g_{0,n}(x)$ (equivalent to finiteness of $\mathbb{H}(\mathbf{X})$ if the reference measure is a product of independent measures).

Both proofs follow the lines of McMillan's proof, using Doob's martingale theorem and embedding the process in a bilateral $(X_n)_{n\in\mathbb{Z}}$ process. Let π_i denote the n-th coordinate function defined on $E^{\mathbb{N}}$ by $\pi_i(x) = x_n$ and let

$\mathcal{F}_{m,n}$ be the σ-field generated by $(\pi_m, \ldots, \pi_n)$ for $m < n$. The density of $\mathbb{P}_{\mathbf{X}}^{m,n}$ with respect to the measure $\kappa^{m,n}$ defined on $\mathcal{F}_{m,n}$ by

$$\kappa^{m,n}(A \times B) = \int_B \mu(A \mid \mathcal{F}_{m,n-1}) d\mathbb{P}_{\mathbf{X}}, \quad A \in \mathcal{E}, \ B \in \mathcal{F}_{m,n-1}$$

is proven to be $g_{m,n}(x)$. If μ is Markovian, then $\kappa^{m',0}$ is an extension of $\kappa^{m',0}$ to $\mathcal{F}_{m',0}$ for all $m' < m$. And if μ is homogeneous, then $g_{m,n}(x) = \theta^n g_{m-n,0}(x)$.

Following Gallager's method, Kieffer (1974) gave a simpler proof of the same result.

Perez (1964) reviewed, applied and generalized the previous extensions of the AEP in terms of relative entropy between processes.

author	date	limit	$\mathbf{X}$	$\mathbb{T}$	E	reference measure
Gallager	(1968)	mean	ergodic	$\mathbb{N}$	finite	counting
Perez	(1957)	mean	erg. station.	$\mathbb{N}$	Borel	product
Moy	(1960,61)	mean	erg. station.	$\mathbb{N}$	Borel	hmg Markov
Kieffer	(1974)	mean	erg. station.	$\mathbb{N}$	Borel	hmg Markov
Shields	(1987)	a.s.	erg. station.	$\mathbb{N}$	finite	counting
Carleson	(1958)	a.s.	erg. station.	$\mathbb{N}$	countable	counting
Chung	(1961)	a.s.	erg. station.	$\mathbb{N}$	countable	counting
Barron	(1985)	a.s.	erg. station.	$\mathbb{N}$	Borel	hmg Markov
Orey	(1985)	a.s.	erg. station.	$\mathbb{N}$	Borel	hmg Markov
Algoet & Cover	(1988)	a.s.	erg. station.	$\mathbb{N}$	Borel	hmg Markov

Fig.2: Extensions of the AEP in terms of state space; discrete-time processes.

7.3.3   Extensions in terms of measures

Other extensions have been made in the direction of weakening the assumptions of stationarity of the process and of Markovian type of the reference measure.

Let us set $\nu = \mathbb{P}_{\mathbf{X}}$ for simplification. Jacobs (1962) proved that if $\nu \ll \nu'$ and if the AEP holds for ν', then it holds for ν too, for a finite state space. Gray & Kieffer (1980) extended it to the case where ν is asymptotically dominated by a stationary measure ν', in the sense that

$$\nu'(F) = 0 \Longrightarrow \nu(\theta^{-n}F) \longrightarrow 0, \quad n \to +\infty,$$

and the strong AEP holds for ν'. It allows them to use a generalized version of the ergodic theorem and to prove the AEP for ν both in mean and almost surely. Barron (1985) extended it to a Borel state space for the almost sure convergence, with a Markov of order $m \geq 0$ reference measure.

Klimko & Sucheston (1968) proved the mean AEP for an irreducible Markov chain with a countable state space and an infinite invariant measure, under several additional conditions.

Wen & Weiguo (1995,1996) proved the AEP for a non-homogeneous Markov chain with a finite state space, using the particular form of $\mathbb{H}_n$ for this case and proving that

$$\frac{1}{n}\left[h_n(\mathbf{X}) + \sum_{k=1}^{n}\sum_{j=1}^{|E|} p_k(X_{k-1},j)\log p_k(X_{k-1},j)\right] \longrightarrow 0, \quad n \to +\infty,$$

where $p_k(i,j) = \mathbb{P}(X_k = j \mid X_{n-1} = j)$ are the transition probabilities of the chain.

author	date	limit	**X**	T	E	reference measure
Jacobs	(1962)	mean	$\ll$ station.	$\mathbb{N}$	finite	counting
Gray & Kieffer	(1980)	a.s. mean	asympt. $\ll$ station.	$\mathbb{N}$	finite	counting
Barron	(1985)	a.s.	asympt. $\ll$ station.	$\mathbb{N}$	Borel	homog. Markov
Orey	(1985)	a.s.	erg. station.	$\mathbb{N}$	Borel	homog. nearly Markov

author	date	limit	**X**	$\mathbb{T}$	E	reference measure
Klimko & Sucheston	(1968)	a.s.	Markov with ∞ station. measure	$\mathbb{N}$	countable	counting
Wen & Weiguo	(1995) (1996)	a.s.	nonhomog. Markov	$\mathbb{N}$	finite	counting

Fig.3: Extensions of the AEP in terms of measures; discrete-time processes.

7.3.4   Extensions to continuous-time processes

Perez (1957) showed the mean AEP for ergodic stationary for discrete as well as for continuous time processes with a measurable state space and for the product reference measure, under finiteness of $\lim_{T\to+\infty}\mathbb{H}_T(\mathbf{X})/T$ (or up-boundedness of the sequence $\mathbb{H}_T(\mathbf{X})/T$).

Pinsker (1960) extended it (via a discretization procedure and using McMillan's proof) to conditions amounting to homogeneity and Markovian properties of the reference measure for a finite state space.

Kieffer (1974) extended it to a Borel state space, using Gallager's method.

Bad Dumitrescu (1988) showed the mean convergence for a pure jump Markov process with a finite state space by using Perez (1957) and a convergence result of Albert (1962) on the number of transitions from one state to another. She proved the finiteness of $\lim_{T\to+\infty}\mathbb{H}_T(\mathbf{X})/T$ by writing explicitly the likelihood of the associated renewal Markov process with respect to the product of the Lebesgue measure and the counting measure on $E^{\mathbb{N}}$, say μ^*.

Girardin & Limnios (2001b) extended the mean and strong AEP to an irreducible positive recurrent semi-Markov process with a finite state space and a semi-Markov kernel absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_+$ with derivative q_{ij} such that $(\log q_{ij})$ is uniformly $L^1(\mathbb{R}_+)$-bounded and with $m_i < +\infty$, where m_i denotes the mean sojourn time in state i, i.e.,

$$m_i = \int_0^{+\infty} \left[1 - \sum_j Q_{ij}(t)\right] dt, \quad i \in E.$$

The proof uses the likelihood of the associated renewal Markov process with respect to μ^* via a generalization to these processes of the convergence result

of Albert (1962). Note that any irreducible positive recurrent semi-Markov process with a finite state space is ergodic.The case of a pure jump Markov process is derived as a particular case.

The generalization to a countable state space is straightforward under finiteness conditions.

Under similar hypothesis, the strong and mean AEP for the entropy of a semi-Markov processes relative to another is proven too in Girardin& Limnios (2001b). The reference measure μ is then the distribution of a semi-Markov process too, that is to say a semi-Markovian measure.

author	date	limit	**X**	$\mathbb{T}$	E	reference measure
Perez	(1957)	mean	erg. station.	$\mathbb{R}_+$	Borel	product
Pinsker	(1960)	mean	erg. station.	$\mathbb{R}_+$	finite	homog. Markov
Kieffer	(1974)	mean	erg. station.	$\mathbb{R}_+$	Borel	Markov
Bad Dumitrescu	(1988)	mean	ergodic Markov	$\mathbb{R}_+$	finite	Lebesgue counting
Girardin & Limnios	(2001)	a.s.	ergodic semi-Markov	$\mathbb{R}_+$	finite (countable)	Lebesgue ×Markov
Girardin & Limnios	(2001)	a.s.	ergodic semi-Markov	$\mathbb{R}_+$	finite (countable)	ergodic semi-Markov

Fig.4: Extensions of the AEP; continuous-time processes.

7.4   Explicit expressions of the entropy rate

For some kinds of processes, as the Markovian or gaussian processes, the entropy rate $\mathbb{H}(\mathbf{X})$ has an explicit form.

It has been first defined by Shannon (1948) for an ergodic Markov chain with a finite state set as the sum of the entropies of the transition probabilities $(p_{ij})_j$ weighted by the probability of occurrence of each state according to the stationary distribution, namely

$$\mathbb{H}(\mathbf{X}) = -\sum_i \pi_i \sum_j p_{ij} \log p_{ij}, \tag{4.1}$$

and he proved the AEP then. The entropy rate of a positive recurrent chain with a countable state space takes this form too.

Krengel (1967) proved that the entropy rate of a null recurrent chain with a countable state space is still given by (4.1), if π denotes an invariant measure of the chain (with $\sum_i \pi_i = +\infty$).

For a semi-Markov process, under suitable hypothesis, Girardin & Limnios (2001b) showed that

$$\mathbb{H}(\mathbf{X}) = \frac{\sum_{i,j} \nu_i \mathbb{S}(Q_{ij} \mid \Lambda)}{\sum_\ell \nu_\ell m_\ell},$$

where ν denotes the stationary distribution of (J_n).

The relative entropy rate between two semi-Markov processes $\mathbf{X}$ and $\mathbf{Z}$ is

$$\mathbb{H}(\mathbf{X} \mid \mathbf{Y}) = \frac{\sum_{i,j} \nu_i \mathbb{S}(Q_{ij} \mid R_{ij})}{\sum_\ell \nu_\ell m_\ell},$$

where X and Y are semi-Markov processes as above, with R denoting the semi-Markov kernel of Y.

The entropy and relative entropy rates of irreducible ergodic finite pure jump Markov processes X and Y defined in Bad Dumitrescu (1988) are obtained as special cases of semi-Markov processes, namely

$$\mathbb{H}(\mathbf{X}) = -\sum_i \pi_i \sum_{j \neq i} a_{ij} \log a_{ij} + \sum_i \pi_i \sum_{j \neq i} a_{ij}$$

and

$$\mathbb{H}(\mathbf{X} \mid \mathbf{Y}) = \sum_i \pi_i \sum_{j \neq i} \left(a_{ij} \log \frac{a_{ij}}{b_{ij}} + a_{ij} - b_{ij} \right),$$

where A and B denote the respective infinitesimal generators of X and Z, and π is the stationary distribution of X (i.e., $\pi A = 0$).

An L^2 process $\mathbf{X} = (X_n)_{n \in \mathbb{N}}$ is weakly stationary if its covariance function is invariant with respect to shifts of time. Its entropy at time n is then less than the entropy of the Gaussian process Y with the same n-dimensional covariance matrix Γ_n. To be specific, $\mathbb{H}_n(\mathbf{X}) \leq \mathbb{H}_n(\mathbf{Y})$, with

$$\mathbb{H}_n(\mathbf{Y}) = \frac{1}{2} \log \text{Det} \Gamma_n + \frac{n}{2} \log(2\pi e),$$

see for example Choi (1987), and

$$\frac{\text{Det} \Gamma_n}{n} \longrightarrow \exp \int \log h(\lambda) d\lambda = \exp \mathcal{I}(\mathbf{Y}),$$

where h is the spectral density of the Gaussian process. Hence for any Gaussian stationary process $\mathbf{Y}$,

$$\mathbb{H}(\mathbf{Y}) = \log \sqrt{2\pi(e+1)} + \frac{1}{4\pi}\mathcal{I}(\mathbf{Y}).$$

The quantity

$$\mathcal{I}(\mathbf{Y}) = \int \log h(\lambda) d\lambda$$

is the Burg entropy of Y. It is also the limit of another sequence. Indeed, if σ_N^2 (resp. σ^2) denotes the variance of the linear prediction error of X_n knowing the finite past $Y_{n-1}, \dots, Y_{n-N}$ (resp. infinite), then $\sigma_N^2 = \mathrm{Det}\Gamma_N/\mathrm{Det}\Gamma_{N-1}$. Due to the projection properties and to Szëgo's theorem (see Grenander & Szëgö (1955)), this yields

$$\sigma_N^2 \longrightarrow \sigma^2 = \exp\mathcal{I}(\mathbf{Y}).$$

Conclusion

Extensions to group index sets, different from $\mathbb{N}$ or $\mathbb{R}_+$ is possible. The AEP is proven to hold for the same time spaces as the individual ergodic theorem is known to hold when the state space is finite, see Ornstein & Weiss (1983) through a proof avoiding martingale arguments.

The case of a general Borel state space is still to be studied for semi-Markov processes. Extension of the AEP to other families of non-stationary processes could be considered.

The Markov nature of the reference measure seems necessary; different attempts to get ride of it have failed, see both Perez (1964)'s statement and counter-example by Kieffer (1974,1976) and Perez (1980) commented by Orey (1985). Orey (1985) extends the strong AEP to "nearly Markovian" measures, a notion too complicated to be developed here, but which seems to constitute the limit of extension in this direction.

The real minimal hypothesis on the process and the reference measure for the AEP to hold is still an open question.

References

Albert, A. *Estimating the infinitesimal generator of a continuous time finite state Markov process.* Ann. Math. Stat. V38, pp727–53 (1962).

Algoet, P. H. *The strong law of large numbers for sequential decisions under uncertainty.* IEEE Trans. Inform. Theory, V40, pp609–33 (1994).

Algoet, P. H. & Cover, T. M. *A sandwich proof of the Shannon-McMillan-Breiman theorem.* Annals Prob., V16, pp899-909 (1988a).

Algoet, P. H. & Cover, T. M. *Asymptotic optimality and asymptotic equirepartition properties of log-optimum investment.* Annals Prob., V16, pp876–898 (1988b).

Anderson, W. J. CONTINUOUS-TIME MARKOV CHAINS. Springer-Verlag, New-York (1991).

Ash, R. A. INFORMATION THEORY. Intersciences, New York (1965) republication: Dover, New York (1994).

Bad Dumitrescu, M. *Some informational properties of Markov purejump processes.* Cas. Pestovani Mat. V113, pp429–34 (1988).

Barren, A. *The strong ergodic theorem for densities: generalized Shannon-McMillan-Breiman theorem.* Ann. Probab., V13, pp1292–1303 (1985).

Billingsley, P. *Ergodic Theory and Information.* R. E. Krieger Publishing Co, Huntington (1978).

Breiman, L. *The individual ergodic theorem of information theory.* Ann. Math. Stat., V28, pp809–11 (1957).

Breiman, L. *Correction to: the individual ergodic theorem of information theory.* Ann. Math. Stat., V31, pp809–10 (1960).

Carleson, L. *Two remarks on the basic theorems of information theory.* Math. Scand., V6, pp175–80 (1958).

Choi, B. S. *A proof of Burg's theorem,* in MAXIMUM ENTROPY AND BAYESIAN SPECTRAL ANALYSIS AND ESTIMATION PROBLEMS. Eds C.R. Smith & G.J. Erickson pp75–84 (1987).

Chung, K. L. *A note on the ergodic theorem of information theory.* Ann. Math. Stat., V32, pp612–14 (1961).

Cover, L. & Thomas, J. ELEMENTS OF INFORMATION THEORY. Wiley series in telecommunications, New-York (1991).

Csizár, I. *Maxent, mathematics, and information theory.* in MAXIMUM ENTROPY AND BAYESIAN METHODS. Kluwer Academic Publishers, pp35–50 (1996).

Donsker, M. D. & Varadhan, S. R. *Asymptotic evaluation of certain Markov process expectations for large time I.* Comm. Pure Appl. Math., V18, pp1–47 (1975).

Doob, J. L. STOCHASTIC PROCESSES. John Wiley & sons, New York, 1953.

Ellis, ENTROPY, LARGE DEVIATIONS AND STATISTICAL MECHANICS., Springer-Verlag, New-York (1985).

Feinstein, A. FOUNDATIONS OF INFORMATION THEORY. McGraw-Hill, New York (1958).

Gallager, R. G. INFORMATION THEORY AND RELIABLE COMMUNICATION. Wiley (1968).

Garret A. *Maximum entropy from the laws of probability.* in BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING. M. Mohammad-Djafari (Ed.), AIPCP, pp3–22 (2001).

Girardin, V. *Entropy maximization for Markov and semi-Markov processes.,* submitted (2002).

Girardin, V. & Limnios, N. PROBABILITS EN VUE DES APPLICATIONS. Vuibert, Paris (2001a).

Girardin, V. & Limnios, N. *Entropy of semi-Markov and Markov processes.* Prépublication Paris-Sud Orsay (2001b).

Gray, R. M. & Kieffer, J. C. *Asymptotically mean stationary measures.* Annals Prob., V8, pp962–73 (1980).

Grenander & Szëgo TOEPLITZ FORMS AND THEIR APPLICATIONS. Chelsea Pub. Co., New York (1955).

Grendar, M. & Grendar, M. *What is the question MaxEnt answears? A probabilistic interpretation.* in BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING. M. Mohammad-Djafari (Ed.), AIPCP, pp83–93 (2001).

Guiasu, S. INFORMATION THEORY WITH APPLICATIONS. McGraw-Hill, New York (1977).

Jacobs, K. LECTURE NOTES ON ERGODIC THEORY. Matematik Institut, Aarhus Univ., Denmark, V1 (1962).

Johnson, O. *Entropy and limit theorems.* PhD Thesis, Cambridge (1999).

Khinchin, A. MATHEMATICAL FOUNDATIONS OF INFORMATION THEORY. Dover, New York (1957).

Kieffer, J. C. *A simple proof of the Moy-Perez generalization of the Shannon-McMillan theorem.* Pacific J. Math. V51, pp 203–06 (1974).

Kieffer, J. C. *A counterexample to Perez's generalization of the Shannon-McMillan theorem.* Annals Prob., V1, pp362–64 (1973) and V4, pp153–54 (1976).

Krengel, U. *Entropy of conservative transformations.* Z. Wahrsch. verw. Geb., V7, pp161-81 (1967).

Kullback, S. INFORMATION THEORY AND STATISTICS. Peter Smith (1978).

Kullback, S. & Leibler, R. A. *On information and sufficiency.* Ann. Math. Stat., V29, pp79–86 (1951).

Lévy, P. *Processus semi-markoviens.* Proc. Int. Cong. Math. Amsterdam, pp416–26 (1954).

Limnios, N. & Oprişan, G. SEMI-MARKOV PROCESSES AND RELIABILITY. Birkhauser, Boston (2001).

Linnik Y. V. *An information-theoretic proof of the central limit theorem with the Lindeberg condition.* Theory Prob. Appl., V4, pp288–299 (1959).

McMillan, M. *The basic theorems of information theory.* Ann. Math. Stat., V24, pp196–219 (1953).

Moran, P. *Entropy, Markov processes and Boltzmann's H-theorem.* Proc. Camb. Philos. Soc. V57, pp833–42 (1961).

Moy, S.-T., *Asymptotic properties of derivatives of stationary measures.* Pacific J. Math., V10, pp1371–83 (1960).

Moy, S.-T, *Generalisations of Shannon-McMillan theorem.* Pacific J. Math., V11, pp705–14 (1961).

Orey, S. *On the Shannon-Perez-Moy theorem.* Contemp. Math. V41, pp319–27 (1985).

Ornstein, D. & Weiss B. *The Shannon-McMillan-Breiman theorem for a class of amenable groups.* Israel J. Math., V44, pp53–60 (1983).

Parthasarathy, K. R. *A note on McMillan's theorem for countable alphabets,* in Inf. Theory, Stat. Decision Functions, Random Processes, pp541–543 Prague (1964).

Perez, A. *Sur la convergence des incertitudes, entropies et informations échantillon (sample) vers leur vraies.* Trans. First Prague Conf. Inf. Theory, Stat. Decision Functions, Random Processes, pp209–243, Prague (1957).

Perez, A. *Extensions of Shannon-McMillan's limit theorem to more general stochastic processes.* in Inf. Theory, Stat. Decision Functions, Random Processes, pp545–574 (1964).

Perez, A. *On Shannon-McMillan's limit theorem for pairs of stationary random processes.* Kybernetika, V19, pp301–14 (1980).

Pinsker, M. S. INFORMATION AND INFORMATION STABILITY OF RANDOM VARIABLES AND PROCESSES. Moscow (1960), Holden-Day, New York (1964).

Reza, F. AN INTRODUCTION TO INFORMATION THEORY. McGraw-Hill, New York (1961), republication: Dover, New-York (1994).

Shannon, C. *A mathematical theory of communication.* Bell Syst., Techn. J., V27, pp379–423, 623–656 (1948).

Shields, P. C. *The ergodic and entropy theorems revisited.* IEEE Trans. Inf. Theory, V33, pp263–66 (1987).

Smith, W.L. *Regenerative stochastic processes.* Proc. Roy. Soc. London, Ser. A, V232, pp6–31 (1955).

Wen, L. & Weiguo, Y. A *limit theorem for the entropy density of nonhomogeneous Markov information source.* Stat. Prob. Letters, V22, pp295–301 (1995).

Wen, L. & Weiguo, Y. *An extension of Shannon-McMillan theorem and some limit properties for nonhomogeneous Markov chains.* Stoch. Proc. Appl., V61, pp129–45 (1996).

This page intentionally left blank

DYNAMIC STOCHASTIC MODELS FOR INDEXES AND THESAURI, IDENTIFICATION CLOUDS, AND INFORMATION RETRIEVAL AND STORAGE

Michiel Hazewinkel
CWI
P.O. Box 94079
1090GB Amsterdam
The Netherlands
mich@cwi.nl

Abstract   The first topic of this partial survey paper is that of the growth of adequate lists of key phrase terms for a given field of science or thesauri for such a field. A very rough 'taking averages' deterministic analysis predicts monotonic growth with saturation effects. A much more sophisticated realistic stochatic model confirms that.

The second, and possibly more important, concept in this paper is that of an identification cloud of a keyphrase (or of other things such as formulas or classification numbers). Very roughly this is (textual) context information that indicates whether a standard keyphrase is present, or, better, should be present, whether it is linguistically recognizable or not (or even totally absent). Identification clouds capture a certain amount of expert information for a given field. Applications include automatic keyphrase assignment and dialogue mediated information retrieval (as discussed in this paper). The problem arises how to generate (semi-)automatically identification clouds and a corresponding enriched weak thesaurus for a given field. A possible (updatable and adaptive) solution is described.

8.1   Introduction

The first topic of this paper is concerned among others with the following question. Suppose one has made an index or thesaurus for a given (super)specialism like for instance discrete mathematics (understood as combinatorics) on the basis of a given corpus, like the two (leading?) journals 'Discrete Mathematics' and 'Applied Discrete Mathematics'. How does one tell that the index made is more or less complete, i.e. more or less good enough to describe the field in question. And, arising from that, are we really dealing with leading journals (as the publisher, in this case Elsevier, believes). As a matter of fact, indexes for the two journals named have been made, [Hazewinkel, 2000; Hazewinkel, 2001] and a very preliminary analysis, [Rudzkis, 2002], indicates that they go some way towards completeness.

One way to tackle this is to test the collection obtained against another corpus. However, such a second corpus may not be available. And if it were available one would like to use it also for key phrase extraction in order to obtain an index/thesaurus that is as complete as possible and the same problem comes back for the new index/thesaurus based on all material available.

Another way to try to deal with the question is to watch how the index/thesaurus grows as more and more material is processed. If, as one would intuitively expect, eventually saturation phenomena appear, that is a good indicator, that some sort of completeness has been reached. To deal with this not only qualitatively but also quantitatively, a dynamic stochastic model is needed, together with appropriate estimators. This is the first topic addressed in this paper.

The second topic deals with information retrieval and automatic indexing. These matters seem to have reached a certain plateau. As I have argued at some length elsewhere, see, e.g., [Marcantognini, 2000; Marcantognini, 2001; Woerdeman, 1989; Hazewinkel, 1999b] there is only so much that can be done with linguistic and statistical means only. To go beyond, it could be necessary to build in some expert knowledge into search engines and the like. This has led to the idea of identification clouds, which is one of the topics of this paper.

The same idea grew out of a rather different (though related) concern. It is known and widely acknowledged, that a thesaurus for a given field of inquiry is a very valuable something to have. However, a classical thesaurus according to ISO standard 2788, see [Arocena, 1990], and various national and international multilingual standards, is not an easily incrementally updatable structure. Indeed, keeping up to date the well known thesaurus EMBASE, [Burg, 1975; Castro, 1986], which is at the basis of Excerpta Medica, takes the full time efforts of four people. This problem of semi-automatic incremental up-

dating of a thesaurus has lead to the idea of an enriched weak thesaurus, [Marcantognini, 2001; Hazewinkel, 1999b], and identification clouds are a central part of that kind of structure.

   In the second part of this paper I try to give some idea of what ID clouds are and how they can be used. More applications can be found in the papers quoted. The idea has meanwhile evolved, largely because of the use of ID clouds in the EC project TRIAL SOLUTION, [Dahn, 1999], and in this paper I also sketch the refinements that have emerged, and indicate some open problems that need to be solved if this approach is to be really useful.

   This paper is an outgrowth of the lecture I gave on (some of) these matters at the IWAP 2002 meeting in Caracas, Venzuela, January 2002. I thank the organizers of that meeting for that opportunity.

8.2   A First Preliminary Model for the Growth of Indexes

   The problem considered in this section is how a global index, a list of terms supposed to describe a given field of enquiry, evolves as indexing proceeds and, simultaneously, the field develops (at a far from trivial pace). The questions arises how does such an index evolve chronologically (assuming, for simplicity, that the indexing is also done chronologically), and, most important, how does one judge on the basis of these data whether the index generated is adequate for the field in question or not.

   Here is a very simple (and naive) stochastic model for this situation and a preliminary (deterministic) analysis of it. At starting time (time zero) there is an (unknown) collection, $K(0)$, of key phrases that is adequate for the field in question. In addition there is an infinite universe of potential terms that can be dreamed up by authors and others of new (important) key phrases. Thus, from the point of view of indexing and thesauri the field grows as:

$$K(t+1) = K(t) \cup B(t),$$

where the union is disjoint and $B(t)$ is the collection of new terms generated in period t. These are not yet known (i.e. identified/recognized), but they do exist in one form or another in the corpus as it exists at time t.

   Now let indexing start. At time zero no terms have been identified. Let $X(t)$ stand for the set of terms recognized (found) at time t, $X(t) \subset K(t)$. Hence $X(0) = \emptyset$. A generalization would be that one starts with an existing thesaurus and tries to bring it up-to-date; then $X(0)$ is a known subset of $K(0)$.

   The indexing proceeds as follows. At time t a set of terms $S(t)$ is selected (found, recognized) and added to $X(t)$. This set $S(t)$ consists of two parts, $S(t) = A(t) \cup C(t)$, $A(t) \subset K(t)$, $C(t) \subset B(t)$, $A(t) \cap C(t) = \emptyset$. Thus

$$X(t+1) = X(t) \cup S(t) \subset K(t+1).$$

As a rule, of course, part of $A(t)$ is already in $X(t)$. The main problem is to have criteria or estimates to decide whether eventually $X(t)$ exhausts $K(t)$ or, $K(t - \tau)$ for a suitable dealy τ, or not. For instance in the form

$$y(t) = \frac{x(t)}{k(t)} \to 1, \quad \text{as } t \to \infty,$$

where $x(t)$ is the cardinality of $X(t)$ and similarly for $k(t)$. The (only) basic observable is $S(t)$ and deriving from that $X(t)$.

Let us do some rather crude average reasoning. First, let us assume linear growth of the field of science in question:

$$k(t) = k(0) + tv$$

for some constant v. Also on average u terms are selected (per period) with a fraction $x(t)/k(t)$ coming from known stuff, and a fraction $(k(t) - x(t))/k(t)$ new terms. There results a recursion equation for $x(t)$:

$$x(t + 1) = x(t) + u\left(1 - \frac{x(t)}{k(t)}\right).$$

Let $y(t) = x(t)/k(t)$ be the fraction of terms covered by the thesaurus at this time. Then

$$y(t + 1) - y(t) = \frac{u}{k(t + 1)} - \frac{(u + v)y(t)}{k(t + 1)}.$$

Assume that the differential equation

$$y' = \frac{u}{k(t + 1)} - \frac{(u + v)y(t)}{k(t + 1)}$$

approximates the difference equation above well enough (which is certainly the case). This differential equation is actually explicitly solvable and the solution is:

$$y(t) = \frac{u}{u + v} - \frac{u(k + v)^{1+(u/v)}}{(u + v)(k + (t + 1)v)^{1+(u/v)}},$$

where $k = k(0)$. So

$$\lim_{t \to \infty} y(t) = \frac{u}{u + v} \tag{2.1}$$

and $y(t)$ grows monotonically from 0 to the asymptotic limit value $u/u + v$.

In particular the recognized fraction of relevant (latent) index terms does not approach one as long as the field keeps growing, and it grows slowly (compared to the indexing rate) once one gets very close to the asymptotic limit. Note also that the saturation phenomenon alluded to in the introduction does indeed occur.
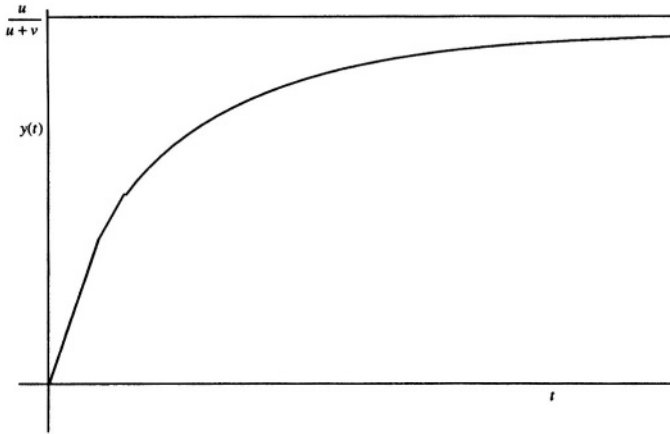
Of course this is quite primitive. Frequently, replacing stochastic phenomena with averages (in a nonlinear case) does not work. So a more sophisticated anlysis of this kind of stochastic processes – apparently a new kind – is needed. This is described in the next section.

8.3   A Dynamic Stochastic Model for the Growth of Indexes

Using the same notations as above the basic assumptions of the model are as follows.

- The $x(t)$, the cardinalities of the sets of key phrases identified up to and including time t, form a random Poisson process. That is, the increments $\Delta x(t) = x(t) - x(t-1)$ are independent random variables with a Poisson distribution P_{λ_t}. For simplicity $x(0)$ is assumed to be a deterministic quantity. Let $n(t) = \mathbf{E}x(t)$, then $\lambda_t = \Delta n(t)$.

- The key phrases are numbered consecutively as they appear in time. A key phrase $w_k \in K(t)$ at the time of its emergence has attached to it a random weight W_k that reflects its relevance (= importance) at that time. The W_k are supposed to be i.i.d. positive random variables with a distribution function F independent of the sequence $x(t)$, and $\mathbf{E}W_k = 1$.

- As before let $S(t)$ be the set of key phrases that were observed at time t and let $A_{k,t} = \{w_k \in S(t)\}$. The probabilities of the random events $A_{k,t}$ depend on the random weights W_k and the history so far, I_t, of the system considered. Assume that for fixed W_k and I_t, the events $A_{k,t}$, $k = 1, \mathrm{L}, x(t)$ are conditionally independent and that the following

equalities hold

$$\mathbf{P}\{A_{k,t} \mid I_t; W_{(.)}\} = \min\left\{\frac{u_t W_k}{x(t)}\right\} \overset{\text{def}}{=} \pi_{k,t}. \qquad (3.1)$$

Here $u_t = \mathbf{E}S(t)$ is a deterministic function that reflects the importance of the corpus used. This (3.1) is quite a weak assumption, practically dictated by the way indexes and thesauri grow in practice.

The results to be quoted below are some of the ones in [Hazewinkel & Rudzkis, 2001] and concentrate on the case that $W_k \equiv 1$. Obviously, much more general models should be examined. For one thing the importance of a key phrases is certainly not a constant and, moreover, is likely to change in time.

Set

$$h(t) = \mathbf{E}x(t), \qquad a = \mathbf{E}\frac{Wu\lambda}{Wu + \lambda},$$

then, besides other asymptotic results, assuming $\lambda_t \equiv \lambda$, $u_t = u$

$$\lim_{t\to\infty} \mathbf{E}\left|\frac{x(t)}{k(t)} - \frac{a}{\lambda}\right| = 0$$

which in the case that $W_{(.)} \equiv 1$ is precisely the result (2.1) of the crude "taking averages" analysis of Section 2 above. It remains to be sorted out what happens in more general circumstances.

There is also an exhaustion result:

$$\lim_{t\to\infty} \mathbf{P}\{K(0) \subset S(t)\} = 1 \Leftrightarrow \sum_t \frac{u_t}{n(t)} = \infty$$

which means that if the observation rate is not too small compared to the growth rate of the field then, eventually, the (latent) key phrases at time zero will all be found.

Shifting time this means that for any time t a certain amount of time later all potential key phrases $K(t)$ will have been recognized with probability 1. What is still needed is an estimate of how much time that will take (depending of course on growth and observation rates).

For a number of statistical estimators of the parameters of the model see loc. cit.

8.4   Identification Clouds

Now suppose that we have a near perfect list of key phrases for, say, mathematics. That is not the case, but adequate lists do exist for certain subfields, [Kailath, 1986; Sz-Nagy, 1970; Schur, 1986; Hazewinkel, 2000; Hazewinkel, 2001; Hazewinkel, 2001a; Hazewinkel, 2002].

Even then there remain most serious open problems of information storage and retrieval. To start lets look at an example. Here is a phrase that occurred in an abstract that came my way for indexing purposes some 6 years ago:

> "… using the Darboux process the complete structure of the solutions of the equation can be obtained,"

At first sight, speaking linguistically, it looks like there is here a perfect natural key phrase to be assigned, viz. "Darboux process". Presumably, some sort of stochastic process like "Cox process", "Gallon–Watson process", "Dirichlet process", or "Poisson process".

However, there is no concept, or result, or anything else in mathematics that goes by the name "Darboux processs". Also the context did not look like having anything to do with stochastics and/or statistics. Had the abstract been classified – it wasn't – using the MSCS (Mathematics Subject Classification Scheme) it would have carried a number like 58F07 (1991 version) or 37J35 (2000 version), neither of which have anything to do with stochastics.

The proper name "Darboux" is also not sufficient to identify what is meant; there are too many terms with "Darboux" in them: "Darboux surface", "Darboux Baire 1 function", "Darboux property", "Darboux function", "Darboux transformation", "Darboux theorem", "Darboux equation",… (these all come from the indexes of [Landau, 1987]).

Or take the following example from [Smeaton, 1992]. Suppose a querier is interested in "prenatal ultrasonic diagnosis". Then texts containing phrases like "in utero sonographic diagnosis", "sonographic detection of fetal ureteral obstruction", "obstetric ultrasound", "ultrasonics in pregnancy", "midwife's experience with ultrasound screening" should also be picked up. Or, inversely, when assigning key-phrase metadata to documents, the documents containing these phrases should also receive the standard controlled key phrase "prenatal ultrasonic diagnosis".

One way to handle such problems (and a number of other problems, see below) is by means of the idea of identification clouds.

Basically the *"identification cloud"* of an item from a controlled list of standardized key phrases is a list of words and possibly other (very short) phrases that are more or less likely to be found near that key phrase in a scientific text treating of the topic described by the key phrase under consideration.

For instance the key phrase

Darboux transformation

could have as (part of its) identification cloud the list

> soliton
> dressing transformation
> Liouville integrable
> completely integrable
> Hamiltonian system
> inverse spectral transform
> Bäcklund transformation
> KdV equation
> KP equation
> Toda lattice
> conservation law
>
> inverse spectral method
> exactly solvable
> …
> (37J35, 37K (the two MSC2000 classification codes for this area of mathematics))
> …

And in fact this particular identification cloud solves the "Darboux process" problem above. The surrounding text contained such words as 'soliton', 'completely integrable', and others from the list above. The appropriate index phrase to be attached was "Darboux transformation".

What the authors of the abstract meant was something like "repeated use of the process 'apply a Darboux transformation' will give all solutions".

A human mathematician, more or less expert in the area of completely integrable systems of differential equations, would have no difficulty in recognizing the phrase "Darboux process" in this sense. Thus what identification clouds do is to add some human expertise to the thesaurus (list of key phrases) used by an automatic system.

The idea of an identification cloud is part of the concept of an enriched weak thesaurus as defined and discussed in [Marcantognini, 2001; Rudin, 1979; Hazewinkel, 1999b].

8.5   Application 1: Automatic Key Phrase Assignment

A first application of the idea of identification clouds is the automatic assignment of key phrases to scientific documents or suitable chunks of scientific texts.

It is simply a fact that it often happens that in an abstract or chunk of text a perfectly good key phrase for the matter being discussed is simply not present

or so well hidden that linguistic and/or statistical techniques do not suffice to recognize it automatically.

The idea here is simple. If enough of the identification cloud of a term (= standard keyphrase) is present than that key phrase is a good candidate at least for being assigned to the document under consideration.

Here are two examples.

8.5.1.   Example

Two-dimensional iterative arrays: characterizations and applications.

We analyse some properties of two-dimensional iterative and **cellular arrays.** For example, we show that **arrays** operating in $T(n)$ time can be sped up to operate in time $n+(T(n)-n)/k$.

...

computation. Unlike previous approaches, we carry out our analyses using *sequential machine characterizations of the iterative and cellular arrays.* Consequently, we are able to prove our results on the much simpler **sequential machine models.**

iterative array
sequential characterization of cellular arrays
sequential characterization of iterative arrays
characterization of cellular arrays
characterization of iterative arrays
arrays of processors

speed-up theorem

Here the available data consisted of an abstract (which is only partially reproduced here). In bold, in the abstract itself, are indicated the index (thesaurus) phrases which can be picked-out directly from the text. Below the original text are five more phrases, that can be obtained from the available data by relatively simple linguistic means, assuming that one has an adequate list of standard key phrases available. For instance "sequential characterization of cellular arrays" and "sequential characterization of iterative arrays" result from the phrase in italics in the abstract fragment above. Note that instead of doing (more or less complicated) linguistic transformations, these could also have been obtained by means of identification clouds. There are advantages in this because there are so very many possible linguistic transformations.

Then, in shadow, there is the term "array of processors". This one is more complicated to find. But, given an adequate standard list, and with "array", "processors" and "machine" all in the available text, it is recognizable, using identification clouds, as a term that belongs to this document.

Finally, in bold-shadow, there is the key phrase "speed-up theorem" a well known type of result in complexity theory. In the text there just occurs "sped up". Certainly, unless one has a good list of (standard) key phrases available, this would be missed. Also purely linguistic means plus such a very good list are clearly still not sufficient; there is no way that one can have a key phrase extraction rule like 'if "sped up" occurs "speed-up theorem" is a likely key phrase'. However, "sped up" plus supporting evidence from the context in the form of a sufficient number of terms from the identification cloud of "speed-up theorem" being present, would do the job.

8.5.2.   Example

Sequential and **concurrent behaviour in Petri net theory.**

Two ways of describing the **behaviour of concurrent systems** have widely been suggested: arbitrary **interleaving and partial orders.** Sometimes the latter has been claimed superior because **concurrency** is represented in a 'true' way; on the other hand, some authors have claimed that the former is sufficient for all practical purposes. **Petri net** theory offers a framework in which both kinds of **semantics** can be defined formally and hence compared with each other. Occurrence sequences correspond to **interleaved behaviour** while the notion of a process is used to capture **partial-order semantics.** This paper aims at obtaining formal results about the

. . .

more powerful than **inductive semantics** using

. . .

of **nets** which are of **finite synchronization** and **1-safe.**

sequential behaviour in Petri net theory
Petri net theory
axiomatic definition of processes

interleaving semantics
1-safe nets

The style coding is the same as in the previous example. Here, the constituents "1-safe" and "nets" of "1-safe nets" actually occur in the text. But they are so far apart that without standard lists and identification clouds the phrase would probably not be picked up. The same holds for the key phrase "interleaving semantics".

Afterwards, I checked against the full text whether these extra key phrases were indeed appropriate. They were. Two more examples can be found in [Woerdeman, 1989] or [Hazewinkel, 1999b]. These are all actual examples which occurred in the corpora used to produce the indexes [Sz-Nagy, 1970; Schur, 1986].

A C-program that takes as input a keyphrase list with identification clouds and a suitably prepared corpus of documents (chunks of text or abstracts) and that gives as output the same corpus with each item enriched with automatically assigned keyphrases has been written in the context of the EC project "TRIAL SOLUTION" (Febr. 2000–Febr. 2003), [Dahn, 1999]. It also outputs an html file for human use which can used to check how well the program worked. This validation test is currently (2002) under way.

It is already clear, that the idea of identification clouds needs refinements; certainly when used on rather elementary material (as in TRIAL SOLUTION). Two of these will be briefly touched on below.

8.6   Application 2: Dialogue Mediated Information Retrieval

Given a keyphrase list with identification clouds, or, better, an enriched weak thesaurus, it is possible to use a dialogue with the machine to refine and sharpen queries. Here is an example of how part of such a dialogue could look:

(Query:)   I am interested in spectral analysis of transformations?

(Answer:)   I have:

- spectral decompositions of operators in Hilbert space (in domain 47, operator theory, 201 hits)

- spectral analysis (in domain 46, functional analysis, 26 hits)

- spectrum of a map (in domain 28, measure theory, 62 hits)

- spectral transform (in domain 58, global analysis, 42 hits)

- inverse spectral transform (in domain 58, global analysis, 405 hits)

Please indicate which are of interest to you by selecting up to five of the above and indicating, if desired, other additional words or key phrases.

The way this works is that the machine scans the query against the available identification clouds (using some (approximate) string matching algorithm, e.g., Boyer–Moore) and returns those keyphrases whose ID clouds match best, together with some additional information to help the querier make up his mind.

8.7   Application 3: Distances in Information Spaces

As it is, the collection of standard keyphrases is just a set. It is a good idea to have a notion of distance on this set: are two selected standard key phrases near, i.e. closely related, or are they quite far from each other. Identification clouds provide one way to get at this idea: two phrases which have large overlap in their identification clouds are near to each other.

A use of this, again dialogue mediated, is as follows.

(Query:)   I am interested in something related to <StandardKeyPhrase 1>. Please give me all standard keyphrases that are within distance x of this one.

For other ways to define distances on information spaces (such as the space of standard key phrases) and other potential uses of distance, see [Hazewinkel, 1999b].

A distance on the space of key phrases is related to a distance on the space of documents, see loc. cit. This is also most useful in dialogue mediated querying. Suppose a really good document for a given query has been found. Than a very useful option is

(Query:)   I am interested in documents close to <Document 1>. Please give me all standard documents that are within distance x of this one and which have two or more of the following key phrases in their key phrase metadata field.

Some search engines have a facility like this in the form of a button like 'similar results' in SCIRUS of Elsevier. But not based on distances in information spaces.

8.8   Application 4: Disambiguation

Ambiguous terms are a perennial problem in (automatic) indexing and thesaurus building.

Identification clouds can serve to distinguish linguistically identical terms from very different areas of the field of inquiry in question. E.g., "regular ring" in mathematics, or the technical term "net" which has at least five completely different meanings in various parts of mathematics and theoretical computer science. For instance 'transportation net' in optimization and operations research, 'net of lines' in differential geometry, 'net' in topology (which replaces the concept of a sequence in topological spaces where the notion of sequence is not good enough), 'communication net', 'net(work) of automata',....

Identification clouds also serve to distinguish rather different instances of the same basic idea in different specializations. E.g., *spectrum* of a commutative algebra in mathematics, *spectrum* of an operator in a different part of
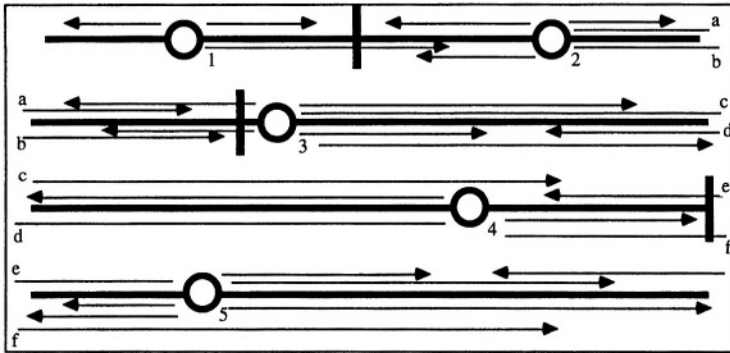
mathematics, and *spectrum* (of a substance) in physics or chemistry are distantly related and ultimately based on the same idea but are in practice completely different terms.

Possibly an even worse problem is caused by phrases and words which have very specific technical meanings but also occur in scientific texts in everyday language meanings. A nice example is the technical concept "end" as it occurs in group theory, topology and complex function theory (three technically different though related concepts). Searching for "end" in a large database such as MATH of FIZ/STN (Berlin, Karlsruhe) is completely hopeless. Searching for "end" together with its ID cloud for its technical meaning in group theory would be a completely different matter. Note that specifying group theory as well in the query would not help much; there are simply too many ways in which the word 'end' occurs (end of a section, to this end, end of the argument, end of proof, …). There are many more words like this; also phrases. For instance 'sort' (as in many sorted languages or sorting theory) and 'bar' (as in bar construction). For more about the 'story of ends', see [Woerdeman, 1989].

8.9   Application 5. Slicing Texts

One important thing made possible by modern electronic technology, i.e. computers and the internet, is the systematic reuse of (educational) material and the composing of books and documents exactly taylored to the needs of an individual user. For instance a teacher may like the introduction to the idea of a topological space from book 1, consider the formal definition of book2 better and may want to use some examples from book3, some exercises from book4, and some historical comments from book5.

The question arises how to chop up a longer text into chunks (slices) that can be efficiently recombined to form such individually taylored texts. This is the subject of the EC Framework 5 project TRIAL SOLUTION (Febr. 2000–

Febr. 2003), [Dahn, 1999]. If the to be sliced document is well structured, for instance composed using LaTeX2e, the structure imposed by the author is a good guide where to slice and this is what TRIAL has so far concentrated on.

Now suppose we have a long section (slices should be relatively short; certainly not more than one computer screen) or an unstructured text, i.e. no clear markings indicating sections, subsections, etc., the exact opposite of a good LaTeX2e document. Suppose also that key phrases have been found and marked in the text and that for each key phrase the evidence for including that key phrase has also been marked; i.e. for each key phrase the corresponding items from its identification cloud have been marked. Treating the text as a long linear string we get a picture like the following.

The numbered fat hollow circles are key phrases in the text which is depicted as a fat horizontal line running over four lines; the arrows connect a key phrase to a member of its identification cloud. If the key phrase is not actually present, the fat circle is the centre of mass of the terms indicating its virtual presence. An arrow can run over more than one line; then labels are used to indicate how it continues.

It is now natural to cut the text at those spots where the number of arrow lines is smallest. For instance, at the three points indicated by fat vertical lines. This can be done at several levels to get a hierarchical slicing. To be able to do this optimally one needs a good stochastic model for the distribution of key phrases through a text and also for the distribution of identification cloud items for a key phrase.

The problem of slicing a text into suitable chunks also comes up in other contexts. For instance in the matter of automatic generation of indexes and identification clouds, see Section 17 below, and in the topic of text mining, see [Visa, 2001], p. 7.

8.10   Weights

One thing that emerged out of the use of identification clouds in the project TRIAL SOLUTION was that it is wise to give weights (numbers between 0 and 1 adding up to 1) to the elements making up an identification cloud.

Here is an example:

```
<KEYPHRASE NAME=<Burgers-Gleichung> THRESHOLD=<0.67>>
    <WORD VALUE=<Burgers-Gleichung> WEIGHT=<0.7>>
    <WORD VALUE=<Burgers> WEIGHT=<0.4>>
    <WORD VALUE=<Gleichung> WEIGHT=<0.2>>
    <WORD VALUE=<Boussinesq> WEIGHT=<0.025>>
    <WORD VALUE=<nichtlinear> WEIGHT=<0.025>>
    <WORD VALUE=<Evolutionsgleichung> WEIGHT=<0.025>>
    <WORD VALUE=<Solitonlösung> WEIGHT=<0.025>>
    <WORD VALUE=<Transformation> WEIGHT=<0.025>>
    <WORD VALUE=<KdV> WEIGHT=<0.025>>
    <WORD VALUE=<sinh> WEIGHT=<0.025>>
    <WORD VALUE=<Gordon> WEIGHT=<0.025>>
    <WORD VALUE=<Hirota> WEIGHT=<0.025>>
    <WORD VALUE=<Kadomzev> WEIGHT=<0.025>>
    <WORD VALUE=<Pedviashwili> WEIGHT=<0.025>>
    <WORD VALUE=<Soliton> WEIGHT=<0.025>>
    <WORD VALUE=<Bäcklund> WEIGHT=<0.025>>
    <WORD VALUE=<inverse spektral> WEIGHT=<0.025>>
    <WORD VALUE=<HOPF> WEIGHT=<0.025>>
    <WORD VALUE=<COLE> WEIGHT=<0.025>>
<\KEYPHRASE>
```

This particular identification cloud is designed to find occurences of the Burgers equation as it occurs in the area of completely integrable dynamical systems (soliton equations, Liouville integrable systems). There are other areas where it occurs; a matter which is further discussed in Section 18 below.

Of course if the phrase itself occurs that is enough as reflected by the first item in the 'WORD VALUE list'. Note further that the occurrence of "Burgers" and of "equation" is not quite enough. There is a good reason for that. For one thing there is also a concept called "Burgers vector" (in connection with torsion in differential geometry); also "Burgers" is a fairly common surname. Further "equation" is of such frequent occurence (in mathematics) that it can turn up just about anywhere. Thus the occurence of both "Burgers" and "equation" in a chunk of text is not enough to decide that "Burgers equation" is a suitable key phrase for that chunk. But if three or more of the sort of words that belong to completely integrable dynamical systems are also present one can be quite sure that it is indeed a suitable key phrase.

Of course if formula recognition, see Section 14 below, were available one would add to the list above

$$< \text{WORD VALUE} =< u_t - u_{xx} - uu_x = 0 > \text{WEIGHT} = 0.7 >$$

(which is the Burgers equation in formula form).

How to assign weights optimally is a large problem. Obviously this cannot be done by hand: a more or less adequate list of standard key phrases for mathematics needs at least 150 000 terms. I propose to use, amoung other things, something like the following adaptive procedure.

Suppose one has an identification cloud of a term consisting of items 1, L, n with weights p_1, p_2, L, p_n adding up to 1. Let a subset $S \subset \{1, 2, \text{L}, n\}$ be successful in identifying the phrase involved. Then the new weights are:

$$\text{For } i \in S, \; p_i' = p_i \left(\frac{\sum_{i \in S} p_i + r(1 - \sum_{i \in S} p_i)}{\sum_{i \in S} p_i} \right).$$

$$\text{For } i \notin S, \; p_i' = p_i - r p_i,$$

where r is a fixed number to be chosen, $0 < r < 1$. (Note that the new weights again add up to 1; note also that the $i \in S$ increase in relative importance and the $i \notin S$ decrease in relative importance; if $S = \{1, \text{L}, n\}$ nothing happens.) This is an adaptation of a reasonably well known algorithm for communication (telephone call) routing that works well in practice but is otherwise still quite fairly mysterious, [Azencott, 1986; Srikantakumar & Narendra, 1982].

8.11   Application 6. Synonyms

There are a variety of things one can do with identification clouds to handle the well known problem of synonyms.

Suppose there are two synonymous key phrases. Then providing both of them with the same identification clouds (including both phrases themselves also as items) will cause both of them to be assigned to those documents where that is appropriate. This would probably the best way to handle this in most circumstances.

Should, however, one prefer to have have just one standardized key phrase this can be handled by having the alternative key phrases in the identification cloud of the standardized one with a weight equal or higher than the threshold value of the selected standardized key phrase; see Section 10 above for how these weights would work.

8.12   Application 7. Crosslingual IR

There are a variety of applications of the idea of identification clouds when dealing with multilingual situations in information retrieval and storage. Suppose for instance one has English language key phrases supplied with German language identification cloud items. One bit of use one can make of this is to attach English language key phrases to German language papers and chunks of text.

Another one is as follows. Suppose we have a German speaking querier who is looking for English language documents as in dialogue mediated search (Section 6 above). Then the same German identification clouds attached to English key phrases permit the machine to handle a German language query.

8.13　Application 8. Automatic Classification

Here "automatic classification" means assigning to a document one or more classification numbers from the MSC2000 (Mathematics Subject Classification Scheme, [MSC2000, 1998]), or its precursor MSC1991. For instance

14M06: linkage
54B35: spectra
55M10: dimension theory

In this setting, instead of key phrases, it is the classification numbers from MSC2000 which are provided with information clouds. This also give these classification numbers substance and meaning. The terse describtions like the three above are far from sufficient to indicate adequately what is meant (even to experts on occasion).

Certainly the mere occurrence of the word "linkage" should not be considered sufficient to assign a paper or chunk of text the classification number 14M06. First of all one would like to be sure that the document in question is about algebraic geometry, this can be done by referring to the identification cloud of the parent node 14 (Algebraic geometry), and second one would like additional evidence like the presence of such supporting phrases as "complete intersection", "determinantal variety", "determinantal ideal", ….

Inversely, a paper may wery well be about the rather technical group of ideas "linkage" without ever mentioning that particular word.

The other two examples just given also need more complete descriptions as to what is really meant (disambiguation and more). For instance there are notions of spectrum in many parts of mathematics: combinatorics, number theory (two different ones at least), homological algebra, ordinary and partial differential equations, dynamical system theory, harmonic analysis, operator theory, general topology, algebraic topology, global analysis, statistics, mechanics, quantum theory, …. Most are somehow related to the original idea of the spectrum of a substance as in physics/chemistry; but some others are completely different.

The exact phrase "dimension theory" occurs four times in MSC2000 while the stem "dimension" occurs no less than 94 times.

8.14　Application 9. Formula Recognition

Recognizing (or finding) formulas in scientific texts is (in any case at first sight) a completely different matter from recognising or finding key phrases. First because formulas are two dimensional and second because the symbols occurring in formulas are not standardized (except a few like the integral sign and the summation sign). Even a standard symbol like π for the number 3.1415 … that gives the radius of the circumference of a circle to its diam-

eter, is not a reliable guide. The Greek letter π is also often used for, for instance, all kinds of mappings in various kinds of geometry, for partitions in combinatorics, and for permutations in group theory.

For instance the two expressions

$$\int_0^1 \frac{\sin x}{x}\,\mathrm{d}x \quad \text{and} \quad \int_0^1 \frac{\sin t}{t}\,\mathrm{d}t$$

mean exactly the same thing. It is the pattern rather than the actual glyphs that occur which determine what a formula means.

And even the patterns are not all that fixed. For instance here are a few versions of that very well known concept in mathematics and engineering, the (one dimensional) Fourier transform (there quite a few more):

$$\hat{f}(\xi) = \int f(x)e^{-i\xi x}\,\mathrm{d}x \qquad \text{see [Katznelson, 1968], p. 120}$$

$$\tilde{f}(p) = (2\pi)^{-1/2}\int_{-\infty}^{\infty} \mathrm{d}q\, f(q)\exp(-ipq) \qquad \text{see [Wolf, 1979], p. 134}$$

$$g(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-iux}\,\mathrm{d}x \qquad \text{see [Wiener, 1933], p. 34}$$

$$C(\lambda) = \int_{-\infty}^{+\infty} e^{-2i\pi\lambda x} f(x)\,\mathrm{d}x \qquad \text{see [Schwartzm 1966], p. 176}$$

$$F(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,\mathrm{d}t \qquad \text{see [Levich, 1970], p. 376}$$

$$F(\omega) = F[f(t)] = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,\mathrm{d}t \qquad \text{see [Hsu, 1967], p. 103}$$

$$F(y) = \int_{-\infty}^{\infty} f(\chi)\exp(-2\pi i x y)\,\mathrm{d}x \qquad \text{see [Bakonyi, 1992], p. 45}$$

$$\hat{f}(\chi) = \int_G \bar{\chi} f\,\mathrm{d}\lambda \qquad \text{see [Hewitt \& Ross, 1963], p. 359}$$

Most of the variations come from different notations for the exponential, the insertion or deletion of normalizing factors involving π, the engineering tradition of writing $\sqrt{-1}$ as j instead of i (as in most of mathematics and physics), different notations for integrands, and putting in or leaving out the integration limits.

Still, it is not easy to define formally what kind of transformations are allowed. On the other hand, trained mathematicians have no difficulty in recognizing any of the above (except possibly the last) as instances of a Fourier transform. Quite generally trained mathematicians can look at a text in their fields of expertise in a language totally unknown to them and still decide what topics the text deals with and at what level things are treated just by looking

at the formulas. Whether that sort of expertise can be taught to machines is an open question. The field of formula recognition is still in its infancy – I would say it is still in a foetal stage.

Identification clouds can help. The idea is the same as before. But instead of a key phrase it is now a (standardized) formula which has an identification cloud attached to it. In the present case one can imagine that the (obligatory) presence of an integral sign, the (also obligatory) presence of the function symbols 'exp(.)' or '$e^{-(\cdot)}$' and an integration variable 'd' in the formula, plus supporting evidence in the form of the occurrence of (some of the) words like "transform", "Fourier", "spectral analysis", "harmonic", … in the surrounding text would do not a bad job in identifying Fourier transform formulas.

Some preliminary work on formula recognition using identification clouds is planned in the EC project [Choi, 1986].

8.15   Context Sensitive IR

In a very real sense the idea of identification clouds is that of context sensitive approximate string recognition. Even if the string itself, that is the key phrase in question, is not recognized the context may provide sufficient supporting evidence to conclude that string should be there as a key phrase. But the way the context is used is very much nonsophisticated. There is no (complicated) grammatical analysis or anything like that. I believe that this is how trained scientists function. They just look casually at the surrounding text of, say, a formula, and on the basis of what they see there decide what it is all about. I do not believe they really do any kind of grammatical analysis or transformations. Indeed, many of us are incapable of doing anything like that, for very often we have to work in foreign languages which are far from perfectly known to us.

8.16   Models for ID Clouds

So far there has been no worry about just how the supporting evidence coming from identification clouds is distributed. This does not matter too much if one is dealing with the problem of assigning key phrases to short chunks of text or to abstracts. Say, to documents of the size of one computer screen or one A4 page maximal.

Things change drastically if one has to deal with longer chunks of text and expecially if one has to assign key phrases, classifications, and other metadata to complete, full text documents. Obviously if the items of an identification cloud for some key phrase of classification of formula or … are spread around very far, are very diffuse, or if they are concentrated in jsut a few lines of text, makes an enormous difference.

Thus what is needed for many applications touched upon in this paper is an experimentally justified stochastic model on how the items of an identification cloud are distributed. And for that matter, how key phrases, whether actually present or not, are distributed over a document. This is of particular importance for the application "slicing of documents" discussed in Section 9 above.

8.17   Automatic Generation of Identification Clouds

Take a large enough, well indexed corpus, and divide it into suitable chunks called documents. For instance take the 700 000 abstracts of articles in the STN/FIZ database Math (ZMG data)[1], or take as documents the sections or pages of a large handbook or encyclopaedia such as the Handbook of Theoretical Computer Science, [van Leeuwen, 1990] or the Encyclopaedia of Mathematics, [Landau, 1987], or an index like [Schur, 1986; Hazewinkel, 2001]. Now use a parser for prepositional noun phrases (PNP's) (or an automaton recognizing PNP's) or a software indexing program like TExTract or CLARIT, [Arocena, 1990A; Arov, 1983; Dym, 1988; Foias, 1990; Gabardo, 1993], to generate from these documents a list of key phrases, keeping track of what phrases come from what document. Now assign, as ID clouds, to the items of the list of keyphrases, those words and phrases found by, say, the software indexing program, which occur in the same document as the key phrase under consideration.

8.18   Multiple Identification Clouds

Picture the set of all documents (chunks of text) in mathematics as a space. For instance a discrete metric space as in [Hazewinkel, 1999b]. There may then very well be several distinct regions in this space where a given key phrase, like "Burgers equation" occurs with some frequency. In this case one may well need several different identification clouds for the same key phrase, even though there is no ambiguity involved. This happens in fact in the case at hand. The Burgers equation has relations with the field of completely integrable systems: it itself has soliton solutions and it is also related to what is probably the most famous soliton equation, the KdV equation (Korteweg–de Vries equation). The identification cloud above in Section 10 was designed to catch this type of occurence of the concept. On the other hand it is the simplest nonlinear diffusion equation and plays a role as such and in discussions of turbulence. To catch those occurrences a rather different set of supporting words and phrases is needed (like diffusion, turbulence, eddy, nonlinearity, ...). Just combining

---

[1] Though this one is not really well indexed in the sense that the key phrases assigned are not from a controlled list. However, if the intention would be to generate the controlled list at the same time as the correponding ID clouds, this material would be most suitable.

the two identification clouds is dangerous because then, by accident, the various different collections of supporting evidence phrases together may combine to give a spurious assignment. One can also not concentrate too much on the proper name "Burgers" for the reasons mentioned in Section 10 above.

8.19   More about Weights. Negative Weights

Another refinement that came out of the experiences with the TRIAL SOLUTION project is that it could be a very good idea to allow negative weights. Let's look at an example.

> "The next topic to be discussed is that of the Fibonacci *numbers.* The generating formula is very simple. But all in all these numbers and their surprisingly many applications are sufficiently *complex* to make the topic very interesting. Similar things happen in the study of fractals."

Or even worse:

> "These mixed spectrum solutions must be *numbered* amoung the more *complex* ones of the KdV equation. Still they can be not neglected."

Both 'complex' and 'numbers' occur in the first fragment of text above (italized). But, obviously it would be totally inappropriate to assign the technical keyphrase 'complex numbers' to this fragment. A negative weight on 'Fibonacci' in the ID cloud of 'complex numbers' will prevent that.

For the second text fragment the technique of stemming, which needs to be used, will give "number", and "complex" also occurs. But here also it would be totally inappropriate to assign the key phrase "complex numbers". It is not so easy to see how to avoid this.

There are still other possible sources of difficulties because "complex" is also a technical term in algebraic topology and homological algebra so one can have a fragment like

> "The Betti numbers of this cell complex are..."

or still worse:

> "The idea of a simplicial complex numbers amoung the most versatile notions that..."

Here even the exact phrase "complex numbers" occurs and negative weights are a must to avoid a spurious assignment.

Quite generally it seems fairly clear that the presence of the constituents of a standard key phrase in a given chunk of text is by no means sufficient to be sure that key phrase is indeed appropriate. This is especially the case for concepts that are made up out of frequently occurring words like "complex numbers" or "boundary value formula". But we have also seen this in the case of the "Burgers equation" above in Section 10. For the case of the phrase "complex numbers" one needs an identification cloud like

```
<KEYPHRASE NAME=<complex numbers> THRESHOLD=<0.4>>
    <WORD VALUE=<complex numbers> WEIGHT=<0.5>>
    <WORD VALUE=<complex> WEIGHT=<0.2>>
    <WORD VALUE=<numbers> WEIGHT=<0.2>>
    <WORD VALUE=<field> WEIGHT=<0.06>>
    <WORD VALUE=<imaginary part> WEIGHT=<0.06>>
    <WORD VALUE=<real part> WEIGHT=<0.06>>
    <WORD VALUE=<absolute value> WEIGHT=<0.06>>
    <WORD VALUE=<Gauss> WEIGHT=<0.06>>
    <WORD VALUE=<argument> WEIGHT=<0.06>>
    <WORD VALUE=<principal value> WEIGHT=<0.06>>
    <WORD VALUE=<vector representation> WEIGHT=<0.06>>
    <WORD VALUE=<addition> WEIGHT=<0.06>>
    <WORD VALUE=<multiplication> WEIGHT=<0.06>>
    <WORD VALUE=<Fibonacci> WEIGHT=<-0.5>>
    <WORD VALUE=<Betti> WEIGHT=<-0.5>>
<\KEYPHRASE>
```

So that besides "complex" and "number" one needs at least 2 more bits of supporting evidence to have a reasonable chance that the fragment in question is indeed has to do with the field of complex numbers. On the other hand if at least 8 of the last ten positive weight terms of the identification cloud above are present one is also rather sure that the fragment in question has to do with the field of complex numbers. The tentative identification cloud given above reflects this. But it is clear that assigning weights properly is a delicate matter; it is also clear that much can be done with weights.

Thus also in the case of occurrences of the same concept in the same part of mathematics, more than one identification cloud may be a good idea, reflecting different styles of presentation and different terminological traditions.

The concrete examples of Section 2 above also illustrates the possible value of negative information.

8.20   Further Refinements and Issues

There are a good many other issues to be addressed. Here is one. It is more or less obvious that making one keyphrase list with ID clouds for all of science and technology is a hopeless task. What one aims at is instead an Atlas of Science and Technology consisting of many weak thesauri that partially overlap, may have different levels of detail, and may focus on different kinds of interest. Much like a geographical atlas which has charts of many different levels of detail and many different kinds (mineralogical, roads and train lines, soil types, height, type of terrain, demographical, climatological, ...). Here the problem arises of how to match the different 'charts'.

Another one is how to adapt the adaptive scheme of Section 10 to a situation with negative weights and how to handle insertion and deletion of ID cloud members.

In an enriched weak thesaurus a key phrase has not only words in its identification cloud but also one or more classification numbers from MSC2000. In turn these classification numbers have identification clouds. The idea is that once a candidate key phrase has been found these are used to check that indeed the paper is related to the topics described by those classification numbers. This idea of referring to other (secondary) identification clouds can be used in all of the various applications described above. For instance it is needed of one uses a formula to identify a key phrase as suggested at the end of Section 10. Such referring to other identification clouds was also briefly mentioned in Section 13 above. Just how this should be implemented stil needs to be worked out.

Probably the most crucial issue to be addressed at this stage is the formulation of a good probabilistic model of ID clouds complete with statistical estimators, see Section 16. A project in this direction has been started by the CWI, Amsterdam together with the IMI, Lithuanian Acad. of Sciences, Vilnius.

References

Jean Aitchison and Alan Gilchrist, *Thesaurus construction,* 2nd edn, Aslib, 1990.

H. Bego, *TExtract: snelle en eenvoudige 'back of the book index' generatie.* In: L. G. M. Noordman and W. A. M. de Vroomen (ed.), Derde STINFON conferentie, 1993, 214.

H. Bego, *TExtraxt. Back-of-the-book index creation system,* TEXYZ, Utrecht, 1997.

G. Bel, P. Chemouil, J. M. Garsia, F. Le Gall and J. Bernusso, *Adaptive traffic routing in telephone networks,* Large Scale Systems **8** (1985), 267–282.

D. C. Champeney, *A handbook of Fourier theorems,* Cambridge Univ. Press, 1987.

Ian Crowlesmith, *Creating a treasure trove of words,* Elsevier Science World, 14–15, 1993.

Ian Crowlesmith, *The development of a biomedical thesaurus,* NBBI Thesaurus Seminar, 1993a.

J. Davenport, a.o., *MKMNET. Mathematical knowledge management network,* Project IST-2001-37057. September 2002–December 2003, 2001.

David A. Evans, *Snapshots of the Clarit text retrieval,* Preprint, copies of slides, Carnegie Mellon University, 1994.

D. A. Evans, K. Ginther-Webster, M. Hart, R. G. Lefferts and I. A. Monarch, *Automatic indexing using selective NLP and first-order thesauri.* In: A. Lichnérowicz (ed.), Intelligent text and image handling, Elsevier, 1991, 524–643.

David M. Evans and Robert C. Lefferts, *Clarit–Trec experiments,* Preprint, Carnegie Mellon University, 1994.

Revaz V. Gamkrelidze, Franz Guenthner, Michiel Hazewinkel and Arkady I. Onishchik, *ERETIMA: English Russian bilingual thesaurus for Invariant theory, Lie groups, Algebraic geometry, Dynamical systems, Optimal control, Commutative algebra. INTAS project 96-0741,* 2001.

Michiel Hazewinkel (ed.), *Encyclopaedia of mathematics; 13 volumes including three supplements,* KAP, 1988–2002.

Michiel Hazewinkel, *Classification in mathematics, discrete metric spaces, and approximation by trees,* Nieuw Archief voor Wiskunde **13** (1995), 325–361.

Michiel Hazewinkel, *Enriched thesauri and their uses in information storage and retrieval.* In: C. Thanos (ed.), Proceedings of the first DELOS workshop, Sophia Antipolis, March 1996, INRIA, 1997, 27–32.

Michiel Hazewinkel, *Index "Artificial Intelligence", Volumes 1–89,* Elsevier, 1997. Large size.

Michiel Hazewinkel, *Topologies and metrics on information spaces.* In: J. Plümer and R. Schwänzl (ed.), Proceedings of the workshop: "Metadata: qualifying web objects", http://www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html, 1997a.

Michiel Hazewinkel, *Index "Theoretical Computer Science", Volumes 1–200,* Theoretical Computer Science **213/214** (1999), 1–699.

Michiel Hazewinkel, *Key words and key phrases in scientific databases. Aspects of guaranteeing output quality for databases of information.* In: Proceedings of the ISI conference on Statistical Publishing, Warsaw, August 1999, ISI, 1999a, 44–48.

Michiel Hazewinkel, *Topologies and metrics on information spaces,* CWI Quarterly **12**:2 (1999b), 93–110. Preliminary version: http://www.mathematik.uni-osnabrueck.de/projects/workshop97/proc.html.

Michiel Hazewinkel, *Index Discrete Applied Mathematics Vols 1–95,* Discrete Applied Mathematics **106** (2000), 1–261.

Michiel Hazewinkel, *Index Discrete Mathematics Vols 1–200,* Discrete Mathematics **227/228** (2001), 1–648.

Michiel Hazewinkel, *Index Information processing letters Vols 1–75,* Information Processing Letters, **78**:1–6 (2001a), 1–448.

Michiel Hazewinkel, *Index journal of logic and algebraic programming volumes 1–45* 68, J. Logic and Algebraic Programming **50**:1–2 (2002), 1–103.

Michiel Hazewinkel and R. Rudzkis, *A probabilistic model for the growth of thesauri,* Acta Appl. Math. **67** (2001), 237–252.

Edwin Hewitt and Kenneth A. Ross, *Abstract harmonic analysis. Volume 1,* Springer, 1963.

Hwei P. Hsu, *Outline of Fourier analysis,* Unitech, 1967.

Yitzak Katznelson, *An introduction to harmonic analysis,* Dover reprint, 1976. Original edition: Wiley, 1968.

Benjamin G. Levich, *Theoretical physics. Volume 1,* North Holland, 1970.

Editors of Mathematical Reviews and Zentralblat für Mathematik, *MSC2000 classification scheme,* 1998.

R. Rudzkis, *Letter to M. Hazewinkel,* 2002.

Laurent Schwartz, *Mathematics for the physical sciences,* Hermann, 1966.

Alan F. Smeaton, *Progress in the application of natural language processing to information retrieval tasks,* The Computer Journal **35**:3 (1992), 268–278.

P. R. Srikantakumar and K. S. Narendra, *A learning model for routing in telephone networks,* SIAM J. Control and Optimization **20**:1 (1982), 34–57.

Jan van Leeuwen (ed.), *Handbook of theoretical computer science,* Elsevier, 1990.

Ari Visa, *Technology of text mining.* In: Petra Perner (ed.), Machine learning and data mining in pattern recognition. Second international workshop, Leipzig, 2001, Springer, 2001, 1–11.

Norbert Wiener, *The Fourier integral and certain of its applications,* Cambridge Univ. Press, 1933.

Kurt Bernardo Wolf, *Integral transforms in science and engineering,* Plenum, 1979.

B. Ingo Dahn, TRIAL SOLUTION. Tools for reusable integrated adaptable learning systems; standards for open learning using tested interoperable objects and networking, Project IST-1999-11397: Febr. 2000–May 2003, 1999.

STABILITY AND OPTIMAL CONTROL FOR SEMI-MARKOV JUMP PARAMETER LINEAR SYSTEMS

Kenneth J. Hochberg

Department of Mathematics and Computer Science, Bar-Ilan University, 52900 Ramat-Gan, Israel

Department of Mathematics, College of Judea and Samaria, 44837 Ariel, Israel

hochbergmacs.biu.ac.il

Efraim Shmerling

Department of Mathematics and Computer Science, Bar-Ilan University, 52900 Ramat-Gan, Israel

Abstract   We consider continuous-time and discrete-time jump parameter linear control systems with semi-Markov coefficients and solution jumps that coincide with jumps of a semi-Markov random process. First, we derive stability conditions for semi-Markov systems of differential equations. We then determine necessary optimality conditions for the solutions of continuous-time and discrete-time control systems.

Keywords: Random polynomials, Jump parameter linear system, semi-Markov process, stability, optimal control

9.1   Introduction

Jump parameter linear control systems with Markov coefficients have been examined in many recent publications. To date, systems of equations defining optimal control have been derived, and recent research in this field now focuses on developing effective numerical methods for solving these systems ([Arov, 1983]-[Castro, 1986]).

In this article, we consider continuous-time and discrete-time jump parameter linear systems with semi-Markov coefficients. These systems represent a generalization of those systems described above, since a semi-Markov process that satisfies certain conditions is Markov. The well-known systems of equations which define optimal control for Markov jump parameter systems can be

obtained utilizing the systems of equations for semi-Markov control systems that we will derive in this paper, and they can be viewed as a particular case of these systems.

The problem of obtaining optimal control for semi-Markov control systems is closely correlated with the problem of finding necessary and sufficient L_2-stability conditions for semi-Markov systems of differential equations. We therefore also consider this other problem in this article.

In order to formulate the problems that we are going to study, we need to introduce some notation and review some well-known facts concerning finite-valued semi-Markov processes.

Consider a finite-valued semi-Markov process $\xi(t)$ with n possible states $\theta_1, \ldots, \theta_n$ which jumps from some state θ_k to some state θ_s at consecutive times t_j, $j = 1, 2, \ldots$, $t_0 = 0$. The random chain $\xi(t_j)$ is a Markov chain, the transition-probabilities matrix of which

$$\Pi = (\pi_{sk})_1^n,$$

$$\pi_{sk} = P\{\xi(t_{j+1}) = \theta_s \mid \xi(t_j) = \theta_k\} \quad (k, s = 1, \ldots, n)$$

is given. (Note the order of the indices s, k here.)

Jump times t_j for the semi-Markov process $\xi(t)$ are defined by distribution functions $F_{sk}(t) = P\{T_{sk} < t\}$ of random variables T_{sk} $(s, k = 1, \ldots, n)$ — the duration of time in which the process belongs to state θ_k before it jumps to state θ_s, provided that such a jump takes place.

The behavior of the process $\xi(t)$ after any time t_j is completely defined by Π and the probability-functions matrix

$$F(t) = (F_{sk}(t))_{s,k=1}^n$$

or the corresponding probability-density-functions matrix

$$f(t) = (f_{sk}(t))_{s,k=1}^n.$$

The intensities $q_{sk}(t)$ are then defined by the formulas

$$q_{sk} = \pi_{sk} f_{sk}(t) \quad (k, s = 1, \ldots, n), \tag{1.1}$$

and we define

$$q_k(t) = \sum_{s=1}^n q_{sk}(t) \quad (k = 1, \ldots, n).$$

Finally, we let T_k denote the duration of time between two consecutive jump times t_j and t_{j+1}, provided that at time t_j the process jumps to θ_k.

Obviously, $q_k(t)$ is the probability density of T_k. Let $F_k(t)$ denote the probability distribution function of T_k, and let $\psi_k(t)$ denote the probability of the

event that no jumps take place during the time interval $(t_j, t_j + t)$ provided that at time t_j the process jumps to θ_k. Clearly,

$$\psi_k(t) = 1 - F_k(t) = \int_t^\infty q_k(t)dt. \tag{1.2}$$

Now, let n different functions $u_k(t)$ $(k = 1, \dots, n)$ be defined at $t \geq 0$. We will call a random process $u(t, \xi(t))$ a *semi-Markov function* if at $t_j \leq t \leq t_{j+1}$, $\xi(t) = \theta_k$, we have

$$u(t, \xi(t)) = u_k(t - t_j); \tag{1.3}$$

i.e., between two jumps of the random process $\xi(t)$, when $\xi(t) = \theta_k$, the semi-Markov function coincides with the deterministic function $u_k(t - t_j)$. In the special case when $u_k(t) \equiv \theta_k (k = 1, \dots, n)$, the semi-Markov function coincides with the semi-Markov finite-valued process $\xi(t)$.

Let $\langle u(t, \xi(t)) \rangle$ denote the mathematical expectation of a semi-Markov function, and denote the conditional mathematical expectations by

$$v_k(t) = \langle u(t, \xi(t)) \mid \xi(0) = \theta_k \rangle \quad (k = 1, \dots, n). \tag{1.4}$$

We thus have the system of integral equations

$$v_k(t) = \psi_k(t)u_k(t) + \int_0^t \sum_{s=1}^n v_k(t - \tau) \, q_{sk}(\tau) \, d\tau \quad (k = 1, \dots, n). \tag{1.5}$$

Let $A_k(t)$ $(k = 1, 2, \dots, n)$ be some given deterministic matrix functions, and let $A(t, \xi(t))$ denote a semi-Markov matrix function that takes values $A_k(t - t_j)$ for $t_j \leq t < t_{j+1}$, provided that $\xi(t)$ belongs to state θ_k during the time period $[t_j, t_{j+1})$.

We consider the system of differential equations

$$\frac{dX(t)}{dt} = A(t, \xi(t))X(t). \tag{1.6}$$

Assume that the solutions of the system have jumps which take place simultaneously with the jumps of $\xi(t)$. These jumps are defined by the formulas

$$X(t_j + 0) = T_{sk}X(t_j - 0), \quad \det T_{sk} \neq 0, \tag{1.7}$$

where T_{sk} $(s, k = 1, \dots, n)$ are some given matrices.

Next, we introduce the notion of L_2-stability for semi-Markov systems given by (6)–(7). First, let $\langle X \rangle$ denote the mathematical expectation $E(X)$. Then, the system (6)–(7) is called L_2-*stable* if, for arbitrary $X(0)$, we have

$$I = \int_0^\infty D(t)dt < \infty, \tag{1.8}$$

where $D(t) \equiv \langle X(t) X^T(t) \rangle$ and $X(t)$ is the solution of (6).

Linear continuous-time semi-Markov control systems are introduced in a similar way in Section 3.

In Section 4, we consider discrete-time semi-Markov control systems, the coefficients of which depend on a discrete semi-Markov process ξ_k, the jumps of which can take place at times $t = 0, 1, 2, \dots$.

The notations $q_s(k)$ $(s = 1, \dots, n;\ k = 1, 2, \dots)$ and $\psi_s(k)$ will be analogous to $q_k(t)$ $(k = 1, \dots, n)$ and $\psi_k(t)$ given earlier. Obviously, the following equalities hold:

$$\psi_s(k) = \sum_{j=k+1}^{\infty} q_s(j),$$

$$q_s(k) = \sum_{\ell=1}^{n} q_{\ell s}(k),$$

where the intensities $q_{\ell s}(k)$ are analogous to the intensities $q_{sk}(t)$ introduced earlier.

9.2   Stability conditions for semi-Markov systems

We introduce the quadratic form

$$w(X) = X^T B X, \qquad B = B^T > 0 \tag{2.1}$$

and define the Lyapunov function by the formula

$$V = \int_0^\infty < w(X(t)) > dt, \tag{2.2}$$

where $X(t)$ is the random solution of system (6) with solution jumps (7). In order to find V, we introduce conditional stochastic Lyapunov functions

$$V_k(X) \equiv X^T C_k X \equiv \int_0^\infty < w(X(t)) \mid X(0) = X, \xi(0) = \theta_k > dt \tag{2.3}$$
$$(k = 1, \dots, n).$$

If the functions $V_k(X)$ $(k = 1, \dots, n)$ are known, then the function V in (10) can be found from the formula

$$V = \int_{Em} \sum_{k=1}^{n} V_k(X) f_k(0, X) dX = \sum_{k=1}^{n} \int_{Em} C_k \circ X X^T f_k(0, X)\, dX \tag{2.4}$$
$$= \sum_{k=1}^{n} C_k \circ D_k(0),$$

where the symbol "∘" denotes the scalar product of matrices, and the functions $f_k(0, X)$ $(k = 1, \ldots, n)$ are defined by the formula

$$P\{\xi(0) = \theta_k, \ X \in \Delta\} = \int_\Delta f_k(0, X) dX,$$

where Δ is an arbitrary domain in the Euclidean space E_m.

In order to form a system of equations which defines the functions $V_k(X)$ $(k = 1, \ldots, n)$, we introduce auxiliary quadratic forms

$$u_k(t, X) \equiv X^T U_k(t) \, X = \langle w(X(t)) \mid X(0) = X, \xi(0) = \theta_k \rangle. \qquad (2.5)$$

We denote by $N_k(t)$ the fundamental-solutions matrices for the systems of linear differential equations

$$\frac{dX_k(t)}{dt} = A_k(t) X_k(t) \qquad (k = 1, \ldots, n). \qquad (2.6)$$

The solutions of system (14) can then be expressed in the form

$$X_k(t) = N_k(t) X_k(0).$$

Utilizing formulas (5), we derive the system of equations

$$u_k(t, X) = \psi_k(t) w\left(N_k(t)X\right) + \int_0^t \sum_{s=1}^n u_s\left(t - \tau, T_{sk}N_k(\tau)X\right) q_{sk}(\tau) \, d\tau$$

$$(k = 1, \ldots, n).$$
$$(2.7)$$

This system can be rewritten as

$$X^T U_k(t) \, X = \psi_k(t) \, X^T \, N_k^T(t) \, B \, N_k(t) \, X$$

$$+ \int_0^t \sum_{s=1}^n X^T \, N_k^T(\tau) \, T_{sk}^T \, U_s(t - \tau) \, T_{sk} \, N_k(\tau) \, X \, q_{sk}(\tau) \, d\tau \quad (k = 1, \ldots, n)$$
$$(2.8)$$

and as

$$U_k(t) = \psi_k(t) N_k^T(t) B N_k(t) +$$

$$\int_0^t \sum_{s=1}^n N_k(\tau) T_{sk}^T U_s(t - \tau) \, T_{sk} N_k(\tau) q_{sk}(\tau) d\tau. \qquad (2.9)$$

Integrating the system of equations, we find the following equations for the matrices C_k :

$$C_k = \int_0^\infty U_k(t)dt.$$

$$= \int_0^\infty \psi_k(t)N_k^T(t)B \, N_k(t)dt \tag{2.10}$$

$$+ \int_0^\infty \sum_{s=1}^n N_k^T(t)T_{sk}^T C_s T_{sk} N_k(t) \, q_{sk}(t) \, dt \quad (k = 1, \ldots, n).$$

The monotonicity of the operators Q_{sk}^* defined by the formula

$$Q_{sk}^* C \equiv \int_0^\infty N_k^T(t) \, T_{sk}^T \, C \, T_{sk} \, N_k(t) \, q_{sk}(t) \, dt \qquad (s, k = 1, \ldots, n) \tag{2.11}$$

enables us to formulate a theorem on the L_2-**stability** of the system (6).

First, we formulate (without proof) the following lemma, which asserts that here, all norms are equivalent:

LEMMA 1 *The integral*

$$\int_0^\infty \psi_k(t)N_k^T(t)BN_k(t)dt$$

converges iff the integral

$$J_k = \int_0^\infty \psi_k(t)\|N_k(t)\|^2 dt$$

converges, where $\|N_k(t)\|$ *designates the Euclidean norm of* $N_k(t)$.

We then have the following theorem:

THEOREM 1 *Assume that for the system of linear differential equations (6) with random semi-Markov coefficients and solution jumps (7), the necessary stability conditions* $J_k < \infty$ $(k = 1, \ldots, n)$ *are satisfied. Then the zero solution of the system is* L_2-**stable** *iff for some positive definite matrices* $B_k > 0$, *the system of equations*

$$C_k = B_k + \sum_{s=1}^n Q_{sk}^* C_s \qquad (k = 1, \ldots, n) \tag{2.12}$$

has a positive-definite solution $C_k > 0$ $(k = 1, \ldots, n)$.

Proof. The proof follows from the fact that the existence of a positive-definite solution of (20) is equivalent to the convergence of the successive approximations

$$C_k^{(j+1)} = B_k + \sum_{s=1}^{n} Q_{sk}^* C_k^{(j)}, \quad C_k^{(0)} = 0 \quad (k = 1, \ldots, n; \ j = 0, 1, 2, \ldots).$$

It can easily be shown that if for some matrices $B_k > 0$ $(k = 1, \ldots, n)$ the successive approximations converge, then they converge also for any arbitrarily chosen positive-definite matrices B_k $(k = 1, \ldots, n)$. ∎

9.3 Optimization of continuous control systems with semi-Markov coefficients

In this section, we find necessary optimality conditions for solutions of linear continuous control systems with semi-Markov coefficients and solution jumps coinciding with jumps of a semi-Markov random process. Values of a quadratic functional are obtained with the help of equations for Lyapunov functions and minimized by choosing control coefficients. The necessary optimality conditions can be utilized in determining the optimal control.

We consider the linear control system

$$\frac{dX(t)}{dt} = A(t, \xi(t))X(t) + B(t, \xi(t))U(t) \tag{3.1}$$

with random semi-Markov coefficients. We seek a control vector $U(t)$ which minimizes the quadratic functional

$$V = \left\langle \int_0^\infty \left(X^T(t)Q(t, \xi(t))X(t) + U^T(t)L(t, \xi(t))U(t) \right) dt \right\rangle, \tag{3.2}$$

where $Q(t, \xi(t))$ and $L(t, \xi(t))$ are symmetric positive definite matrices. Suppose that a semi-Markov process $\xi(t)$ has jumps at times t_j $(j = 0, 1, 2, \ldots,)$, where $t_0 = 0 < t_1 < t_2 < \ldots$. Assume that at $t_j \le t < t_{j+1}$, $\xi(t) = \theta_s$, the following equalities hold:

$$\begin{aligned} A(t, \xi(t)) &= A_s(t - t_j), \quad B(t, \xi(t)) = B_s(t - t_j); \\ Q(t, \xi(t)) &= Q_s(t - t_j), \quad L(t, \xi(t)) = L_s(t - t_j) \quad (s = 1, \ldots, n), \end{aligned}$$

where $A_s(t), B_s(t), Q_s(t), L_s(t)$ are deterministic matrices. Assume that the optimal control has the form

$$U(t) = S(t, \xi(t)) X(t), \tag{3.3}$$

where $S(t, \xi(t))$ is a matrix with semi-Markov coefficients which, at $t \in [t_j, t_{j+1})$, $\xi(t) = \theta_s$, takes values

$$S(t, \xi(t)) = S_s(t - t_j) \qquad (s = 1, \ldots, n).$$

We introduce the following notation:

$$
\begin{aligned}
G(t, \xi(t)) &\equiv A(t, \xi(t)) + B(t, \xi(t))S(t, \xi(t)); \\
H(t, \xi(t)) &\equiv Q(t, \xi(t)) + S^T(t, \xi(t))L(t, \xi(t))S(t, \xi(t)).
\end{aligned}
\qquad (3.4)
$$

We then obtain the system of linear differential equations with semi-Markov coefficients

$$\frac{dX(t)}{dt} = G(t, \xi(t)) \, X(t), \qquad (3.5)$$

for which we seek the value of the quadratic functional

$$V = \left\langle \int_0^\infty X^T(t) H(t, \xi(t)) \, X(t) dt \right\rangle. \qquad (3.6)$$

Assume that if there is a jump of the random process $\xi(t)$ at time t_j, then the solution of (25) also has a jump

$$X(t_j + 0) = T_{sk} X(t_j - 0), \qquad (3.7)$$

if $\xi(t_j + 0) = \theta_s$, $\xi(t_j - 0) = \theta_k$.

For calculating the functional V, we utilize formula (12):

$$V = \sum_{k=1}^n C_k \circ D_k(0) = \sum_{k=1}^n \int_{E_m} V_k(X) f_k(0, X) dX, \qquad (3.8)$$

where $V_k(X) = X^T C_k X$ are partial stochastic Lyapunov functions

$$
V_k(X) \equiv X^T C_k X = \int_0^\infty \langle X^T(t) H(t, \xi(t)) X(t) \mid X(0) = X, \xi(0) = \theta_k \rangle dt
$$
$$
(k = 1, \ldots, n).
$$
$$(3.9)$$

We can now use the expression for C_k obtained in equation (18):

$$
C_k = \int_0^\infty \psi_k(t) N_k^T(t) \left(Q_k(t) + S_k^T(t) L_k(t) S_k(t) \right) N_k(t) dt
$$
$$
+ \sum_{s=1}^n \int_0^\infty q_{sk}(t) N_k^T(t) T_{sk}^T C_s T_{sk} N_k(t) dt \quad (k = 1, \ldots, n).
$$
$$(3.10)$$

Here $N_k(t)$ are fundamental-solutions matrices for the system of linear differential equations

$$\frac{dX_k(t)}{dt} = G_k(t)X_k(t), \quad G_k(t) \equiv A_k(t) + B_k(t)S_k(t)$$

$$X_k(t) = N_k(t)X \quad (k = 1, \dots, n). \tag{3.11}$$

Next we find an expression for the partial stochastic Lyapunov functions

$$V_k(X) \equiv X^T C_k X = \int_0^\infty \left(X_k^T(t) \left(\psi_k(t) Q_k(t) + \sum_{s=1}^n q_{sk}(t) T_{sk}^T C_s T_{sk} \right) \cdot \right.$$

$$\left. \cdot X_k(t) + U_k^T(t) \psi_k(t) L_k(t) U_k(t) \right) dt,$$

$$X_k(0) = X \quad (k = 1, \dots, n). \tag{3.12}$$

The system of equations (31) can be written as

$$\frac{dX_k(t)}{dt} = A_k(t)X_k(t) + B_k(t)U_k(t), \quad U_k(t) \equiv S_k(t)X_k(t). \tag{3.13}$$

Suppose that there exists an optimal control (in the form (23)) for the system (21) that minimizes the functional (22) and does not depend on the initial value $X(0)$. We seek values for the symmetric matrices C_k $(k = 1, \dots, n)$ which minimize the functional V. The problem of finding minimum values of $V_k(X)$ $(k = 1, \dots, n)$ by choosing controls $U_k(t)$ has been thoroughly investigated; see, for example, [Arocena, 1990] and [Arocena, 1990A]. For our purposes, it is important that all matrices C_k $(k = 1, \dots, n)$ in formula (32) are constants.

Thus, the problem of obtaining optimal control (23) for a continuous control system with semi-Markov coefficients is reduced to n independent problems of obtaining optimal control for deterministic systems (33) with minimized functionals (32).

We now apply some well-known results on finding optimal control for the system of equations

$$\frac{dX(t)}{dt} = A(t)X(t) + B(t)U(t), \quad X(0) = X, \tag{3.14}$$

where we seek an optimal control $U(t)$ which minimizes the quadratic functional

$$X^T C X = \int_0^\infty \left(X^T(t)Q(t)X(t) + U^T(t)L(t)U(t) \right) dt. \tag{3.15}$$

Optimal control $U(t)$ is defined by the formula

$$U(t) = -L^{-1}(t)B^T(t)Y(t), \quad Y(t) = K(t)X(t),$$

where the matrix $K(t)$ satisfies the following matrix differential Riccati equation:

$$\frac{dK(t)}{dt} = -Q(t) - A^T(t)K(t) - K(t)A(t) + K(t)B(t)L^{-1}(t)B^T(t)K(t),$$

$$K(\infty) = 0.$$

(3.16)

Methods for solving equation (36) are described, for example, in [Arocena, 1990].

In view of these known results, we obtain the following expression for the optimal control $U_k(t)$ which minimizes the functional $V_k(X)$ for the system of equations (32):

$$U_k(t) = -\psi_k^{-1}(t)L_k^{-1}(t)B_k^T(t)K_k(t)X_k(t) \quad (k = 1, \ldots, n), \qquad (3.17)$$

where matrices $K_k(t)$ $(k = 1, \ldots, n)$ satisfy the following Riccati-type system of equations:

$$\begin{aligned}
\frac{dK_k(t)}{dt} = & -\psi_k(t)Q_k(t) - \sum_{s=1}^{n} q_{sk}(t)T_{sk}^T C_s T_{sk} - A_k^T(t)K_k(t) \\
& -K_k(t)A_k(t) + K_k(t)B_k(t)\psi_k^{-1}(t)L_k^{-1}(t)B_k(t)K_k(t), \\
& K_k(\infty) = 0, \qquad (k = 1, \ldots, n).
\end{aligned}$$

(3.18)

The systems of equations (37)–(38) define necessary optimality conditions for solutions of the system of equations (21). Matrices $S_k(t)$ $(k = 1, \ldots, n)$ defining the optimal control (23) are defined by the matrix equations

$$S_k(t) = -\psi_k^{-1}(t)L_k^{-1}(t)B_k^T(t)K_k(t) \qquad (k = 1, \ldots, n),$$

and matrices C_k are defined by the equalities

$$C_k = K_k(0) \qquad (k = 1, \ldots, n).$$

We solve each equation of the system (38) as a parameter equation with parameter matrices $K_k(0)$ $(k = 1, \ldots, n)$, utilizing numerical methods developed for systems of type (36).

Thus, we obtain a system of n matrix equations with n unknown matrices $K_k(0)$ $(k = 1, \ldots, n)$, which enables us to find the values of $K_k(0)$ $(k = 1, \ldots, n)$, and then the values of $K_k(t)$, $t > 0$.

Now introduce new matrices

$$R_k(t) = \psi_k^{-1}(t)K_k(t) \ , \quad \psi_s(0) = 1 \quad (k = 1, \ldots, n). \qquad (3.19)$$

The system of equations (38) then takes the form

$$\frac{dR_k(t)}{dt} = -Q_k(t) - A_k^T(t)R_k(t) - R_k(t)A_k(t)$$

$$+ R_k(t)B_k(t)L_k^{-1}(t)B_k^T(t)R_k(t)$$

$$- \left(\frac{\psi_k'(t)}{\psi_k(t)} R_k(t) + \sum_{s=1}^{n} \frac{q_{sk}(t)}{\psi_k(t)} T_{sk}^T R_k(0) T_{sk} \right),$$

$$R_k(\infty) = 0 \ (k = 1, \dots, n), \tag{3.20}$$

and optimal control is given by the formulas

$$U_k(t) = -L_k^{-1}(t)B_k^T(t)R_k(t)X_k(t) \qquad (k = 1, \dots, n). \tag{3.21}$$

The necessary optimality conditions (40) and (41) generalize previously obtained optimality conditions for control systems with coefficients dependent on a Markov random process.

The following particular case is important in many applications. Suppose that the semi-Markov process $\xi(t)$ cannot remain in any state θ_s for a time period greater than $T_s > 0$ $(s = 1, \dots, n)$. Assume that

$$q_{ks}(t) \equiv 0 \ (t > T_s), \quad \psi_s(t) \equiv 0 \ (t > T_s). \tag{3.22}$$

We obtain the system of equations

$$C_k = \int_0^{T_s} \psi_k(t) N_k^T(t) \left(Q_k(t) + S_k^T(t) L_k(t) S_k(t) \right) N_k(t) dt$$

$$+ \sum_{s=1}^{n} \int_0^{T_k} q_{sk}(t) N_k^T(t) T_{sk}^T C_s T_{sk} N_k(t) dt \qquad (k = 1, \dots, n),$$

and also the system of equations for the functions $V_k(X)$:

$$V_k(X) = \int_0^{T_k} \left(X_k^T(t) \left(\psi_k(t) \right) Q_k(t) + \sum_{s=1}^{n} q_{sk}(t) T_{sk}^T C_s T_{sk} \right) X_k(t)$$

$$+ U_k^T(t)\psi_k(t)L_k(t)U_k(t) \right) dt, \ X(0) = X \ (k = 1, \dots, n).$$

$$\tag{3.23}$$

In the system (38), we assume that

$$K_s(T_s) = 0 \qquad (s = 1, \dots, n). \tag{3.24}$$

Since $\psi_s(T_s) = 0$ $(s = 1, \dots, n)$, conditions (44) will be satisfied if the matrices $R_s(T_s)$ are bounded, in view of (39).

In order to find matrices $R_s(t)$ $(s = 1, \ldots, n)$, we have to integrate the system of nonlinear matrix differential equations (40). Each equation belonging to the system can have a singular point $t = T_s$, where $\psi_s(T_s) = 0$ $(s = 1, \ldots, n)$. We can obtain necessary conditions for boundedness of matrices $R_s(t)$ at singular points $t = T_s$:

$$\psi'(T_s)R_s(T_s) + \sum_{k=1}^{n} q_{ks}(T_s)T_{ks}^T R_k(0))T_{ks} = 0 \quad (s = 1, \ldots, n). \quad (3.25)$$

We formulate these results as a theorem.

THEOREM 2 *Assume that the optimal control $U(t)$ in the form* (23) *for a control system* (21) *exists. Then the optimal control $U(t)$ that minimizes the quadratic functional* (22) *is defined by the system* (41), *where matrices $R_s(t)$ $(s = 1, \ldots, n)$ satisfy the Riccati-type system of nonlinear differential equations* (40).

9.4   Optimization of discrete control systems with semi-Markov coefficients

We consider the discrete control system

$$X_{k+1} = A(k, \xi_k)X_k + B(k, \xi_k)U_k \qquad (k = 0, 1, 2, \ldots) \qquad (4.1)$$

with semi-Markov coefficients. We seek the control vector U_k which minimizes the quadratic functional

$$V = \left\langle \sum_{k=0}^{\infty} \left(X_k^T Q(k, \xi_k)X_k + U_k^T L(k, \xi_k)U_k \right) \right\rangle, \qquad (4.2)$$

where $Q(k, \xi_k), L(k, \xi_k)$ are symmetric positive definite matrices. Let k_j $(j = 0, 1, 2, \ldots)$, $k_0 = 0$, be jump times of a semi-Markov process which takes a finite number of distinct values $\theta_1, \ldots, \theta_n$. Assume that at $k \in [k_j, k_{j+1})$, $\xi_k = \theta_s$, the matrix coefficients in system (46) and in formula (47) are defined by the following expressions:

$$\begin{aligned} A(k, \xi_k) &= A_s(k - k_j), B(k, \xi_k) = B_s(k - k_j), \\ Q(k, \xi_k) &= Q_s(k - k_j), L(k, \xi_k) = L_s(k - k_j) \quad (s = 1, \ldots, n), \end{aligned} \qquad (4.3)$$

where $A_s(k), B_s(k), Q_s(k), L_s(k)$ are deterministic matrices.
Assume that the optimal control has the form

$$U_k = S(k, \xi_k)X_k \qquad (k = 0, 1, 2, \ldots), \qquad (4.4)$$

where $S(k, \xi_k)$ is a matrix with semi-Markov coefficients, and that at $k \in [k_j, k_{j+1}), \xi_k = \theta_s$ we have the equalities

$$S(k, \xi_k) = S_s(k - k_j) \qquad (s = 1, \dots, n). \tag{4.5}$$

Now we introduce the matrices

$$\begin{aligned} G(k, \xi_k) &= A(k, \xi_k) + B(k, \xi_k)S(k, \xi_k) \\ H(k, \xi_k) &= Q(k, \xi_k) + S^T(k, \xi_k)L(k, \xi_k)S(k, \xi_k). \end{aligned} \tag{4.6}$$

We obtain the system of linear difference equations

$$X_{k+1} = G(k, \xi_k)X_k \qquad (k = 0, 1, 2, \dots), \tag{4.7}$$

with the minimized quadratic functional

$$V = \left\langle \sum_{k=\ell}^{\infty} X_k^T H(k, \xi_k) X_k \right\rangle. \tag{4.8}$$

Next, we introduce partial stochastic Lyapunov functions

$$V_s(X) \equiv X^T C_s X = \sum_{k=0}^{\infty} \langle X_k^T H(k, \xi_k) X_k \mid X_0 = X, \ \xi_0 = \theta_s \rangle \ (s = 1, \dots, n). \tag{4.9}$$

If the functions $V_s(X)$ $(s = 1, \dots, n)$ are calculated, the value of V in (53) can be obtained by the formula

$$V = \int_{Em} \sum_{s=1}^{n} X^T C_s X f_s(0, X) dX = \sum_{s=1}^{n} C_s \circ D_s(0). \tag{4.10}$$

Now, consider the system of linear difference equations (52). Assume that the solution of this system is multiplied from the left by constant matrices $T_{s\ell}$, $\det T_{s\ell} \neq 0$ $(s, \ell = 1, \dots, n)$ at times when the random process ξ_k has jumps from state θ_ℓ to state θ_s.

Let $k \in [k_j, k_{j+1}), \xi_k = \theta_\ell$. The system of equations (52) takes the form

$$X_{k+1} = G_\ell(k - k_j)X_k \qquad (\ell = 1, \dots, n),$$

where

$$G_\ell(k) \equiv A_\ell(k) + B_\ell(k)S_\ell(k) \qquad (\ell = 1, \dots, n).$$

Let systems of linear difference equations

$$X_{k+1}^{(s)} = G_s(k)X_k^{(s)} \qquad (k = 0, 1, 2, \dots; \ s = 1, \dots, n)$$

have fundamental-solutions matrices $N_s(k)$ $(k = 0, 1, \ldots; \; s = 1, \ldots, n)$, which implies that

$$X_k^{(s)} = N_s(k) X_0^{(s)} \qquad (k = 0, 1, 2, \ldots; \; s = 1, \ldots, n). \qquad (4.11)$$

Assume that if the conditions

$$\xi_k = \theta_\ell \quad \text{at} \quad k \in [k_{j-1}, k_j), \quad \xi_k = \theta_s \quad \text{at} \quad k \in [k_j, k_{j+1})$$

are satisfied, then the following equalities hold:

$$
\begin{aligned}
X_k &= N_\ell(k - k_{j-1}) X_{k_{j-1}} && (k_{j-1} \le k < k_j) \\
X_{k_j} &= T_{s\ell} N_\ell(k - k_{j-1}) X_{k_{j-1}}, && \det T_{s\ell} \ne 0 \\
X_k &= N_s(k - k_j) X_{k_{j-1}} && (k_j \le k < k_{j+1}), \qquad (4.12)
\end{aligned}
$$

i.e., at jump times, the solution of (52) is multiplied by a nonsingular matrix $T_{s\ell}$. The system of equalities

$$
\begin{aligned}
V_s(X) \; &\equiv \; X^T C_s X = \sum_{k=0}^{\infty} \Big((X_k^{(s)})^T (\psi_s(k) Q_s(k) \\
&+ \; \sum_{\ell=1}^{n} q_{\ell s}(k) T_{\ell s}^T C_\ell T_{\ell s}) X_k^{(s)} + (U_k^{(s)})^T \psi_s(k) L_s(k) U_k^{(s)} \Big) \\
& \hspace{6cm} (s = 1, \ldots, n) \hspace{2cm} (4.13)
\end{aligned}
$$

is analogous to the system (32) and can be derived in a similar way. Assuming that

$$
\begin{aligned}
H_s(k) &= Q_s(k) + S_s^T L_s(k) S_s(k) && (s = 1, \ldots, n) \\
U_k^{(s)} &= S_s(k) X_k^{(s)} && (s = 1, \ldots, n),
\end{aligned}
$$

we can rewrite equalities (58) as

$$
\begin{aligned}
V_s(X) \equiv X^T C_s X \; &= \; \sum_{k=0}^{\infty} X^T N_s^T(k) \Big(\psi_s(k) H_s(k) \\
&+ \; \sum_{\ell=1}^{n} q_{\ell s}(k) T_{\ell s}^T C_\ell T_{\ell s} \Big) N_s(k) X \\
& \hspace{5cm} (s = 1, \ldots, n). \hspace{1.5cm} (4.14)
\end{aligned}
$$

Minimization of the functional V in (53) is reduced to the minimization of the functions $V_s(X)$ in (54). Thus, the problem of finding optimal control is reduced to n problems of optimizing the deterministic control systems

$$X_{k+1}^{(s)} = A_s(k) X_k^{(s)} + B_s(k) U_k^{(s)} \qquad (s = 1, \ldots, n), \qquad (4.15)$$

where the optimal control $U_k^{(s)}$ minimizes the quadratic functional $V_s(X)$.

Now, we state a well-known result on optimizing systems of linear difference equations with variable coefficients

$$X_{k+1} = A(k)X_k + B(k)U_k \qquad (k = 0, 1, 2, \dots), \qquad (4.16)$$

where the optimal control U_k minimizes the quadratic functional

$$V = \sum_{k=0}^{\infty} \left(X_k^T Q(k)X_k + U_k^T L(k)U_k \right). \qquad (4.17)$$

If an optimal control exists, it is defined by the formula

$$
\begin{aligned}
U_k &= -L^{-1}(k)B^T(k) \left(E + K(k+1)B(k)L^{-1}(k)B^T(k) \right)^{-1} \\
&\quad K(k+1)A(k)X_k \\
&\equiv -\left(L(k) + B^T(k)K(k+1)B(k) \right)^{-1} B^T(k)K(k+1)A(k)X_k,
\end{aligned}
$$
$$(4.18)$$

where the matrices $K(k)$ $(k = 0, 1, 2, \dots)$ satisfy the system of equations

$$
\begin{aligned}
K(k) = \quad & Q(k) + A^T(k)K(k+1)A(k) - A^T(k)K(k+1)B(k) \cdot \\
& \cdot (L(k) + B^T(k)K(k+1)B(k))^{-1}B^T(k)K(k+1)A(k) \\
& \qquad\qquad (k = 0, 1, 2, \dots).
\end{aligned}
$$
$$(4.19)$$

Next, we find an optimal control for the system of linear difference equations (46) with minimized functional (47) by finding $U_k^{(s)}$ which minimize the functionals (59) for the systems of difference equations (60). We obtain the following formulas:

$$
\begin{aligned}
U_k^{(s)} = -\left(L_s(k)\psi_s(k) + B_s^T(k)K_s(k+1)B_s(k) \right)^{-1} \\
B_s^T(k)K_s(k+1)A_s(k)X_k^{(s)} \\
(k = 0, 1, 2, \dots; \;\; s = 1, \dots, n)
\end{aligned}
$$
$$(4.20)$$

$$
\begin{aligned}
K_s(k) = \quad & \psi_s(k)Q_s(k) + \sum_{\ell=1}^{n} q_{\ell s}(k)T_{\ell s}^T C_\ell T_{\ell s} + A_s^T(k)X_s(k+1)A_s(k) \\
& -A_s^T(k)K_s(k+1)B_s(k)(L_s(k)\psi_s(k) \\
& +B_s^T(k)K_s(k+1)B_s(k))^{-1}B_s^T(k)K_s(k+1)A_s(k) \\
& \qquad\qquad (k = 0, 1, 2, \dots; \;\; s = 1, \dots, n).
\end{aligned}
$$
$$(4.21)$$

These equations can be simplified by setting

$$P_s(k-1) = \frac{1}{\psi_s(k-1)} K_s(k) \quad (k=0,1,2,\dots;\ s=1,\dots,n) \quad (4.22)$$

$$U_k^{(s)} = S_s(k)X_k^s \qquad (k=0,1,2,\dots;\ s=1,\dots,n), \qquad (4.23)$$

where $\psi_s(-1)$ is defined to equal 1.

We thus obtain the following system of matrix equations:

$$S_s(k) = -\left(L_s(k) + B_s^T(k)P_s(k)B_s(k)\right)^{-1} B_s^T(k)P_s(k)A_s(k)$$
$$(k=0,1,2\dots;\ s=1,2,\dots); \quad (4.24)$$

$$P_s(k-1) = \frac{\psi_s(k)}{\psi_s(k-1)}\left(Q_s(k) + \sum_{\ell=1}^{n} \frac{q_{\ell s}(k)}{\psi_s(k)} T_{\ell s}^T C_\ell T_{\ell s}\right.$$
$$+A_s^T(k)P_s(k)A_s(k) - A_s^T(k)P_s(k)(L_s(k)$$
$$\left.+B_s^T(k)P_s(k)B_s(k))^{-1}B_s^T(k)P_s(k)A_s(k)\right)$$
$$(k=1,2,\dots;\ s=1,\dots,n), \quad (4.25)$$

which define necessary optimality conditions for solutions of the system (46). The system of equations (66) contains unknown matrices $C_\ell(\ell=1,\dots,n)$.

Now, we utilize the known auxiliary formula

$$X_k^T K(k)X_k = \sum_{j=k}^{\infty}\left(X_j^T Q(j)X_k + U_j^T L(j)U_j\right)$$
$$(k=0,1,2,\dots) \qquad (4.26)$$

for the control system (61), where $X_j, U_j\ (j=k,k+1,k+2,\dots)$ are optimal solutions and optimal control which minimize the functional (62). From this formula, it follows that the matrices $K(k)$ are symmetric and positive semi-definite. From equality (71) and formulas (59), it follows that

$$C_s = K_s(0) \qquad (s=1,\dots,n). \qquad (4.27)$$

We formulate the obtained result in a theorem.

THEOREM 3 *Assume that the optimal control in the form* (49)

$$U_k = S(k,\xi_k)X_k \qquad (k=0,1,2,\dots)$$

for a control system (46)

$$X_{k+1} = A(k,\xi_k)X_k + B(k,\xi_k)U_k \qquad (k=0,1,2,\dots)$$

exists. Then the optimal control is defined by the system (69), *where matrices* $P_s(k)\ (s=1,\dots,n;\ k=0,1,2,\dots)$ *satisfy the Riccati-type system of difference equations* (70).

References

Anderson, B.D., and Moore, J.B. LINEAR OPTIMAL CONTROL, Prentice-Hall, New York (1971).

Barnett, S. POLYNOMIALS AND LINEAR CONTROL SYSTEMS, Marcel Dekker, New York - Basel (1983).

Borno, I. and Gajic, A. *Parallel algorithm for solving coupled algebraic equations of discrete-time jump linear systems.* Computers and Math, with Appl. **29** (1995).

Gajic, Z. and Borno, L. *Lyapunov iterations for optimal control of jump linear systems at steady state.* IEEE Trans. Auto. Cont. (1995).

Gajic, Z. and Qureshi, M.T.J. LYAPUNOV MATRIX EQUATION IN SYSTEM STABILITY AND CONTROL, Academic Press, New York (1995)

Mariton, M. JUMP LINEAR SYSTEMS IN AUTOMATIC CONTROL, Marcel Dekker, New York - Basel (1990).

Mariton, M., and Bertrand, P. *A homotopy algorithm for solving coupled Riccati equations* Optimal Contr. Appl. and Methods **6**, (1985), 351-357.

Shmerling, E. LINEAR SYSTEMS WITH RANDOM COEFFICIENTS, Ph.D. dissertation, Bar-Ilan University, Israel (2000).

This page intentionally left blank

STATISTICAL DISTANCES BASED ON EUCLIDEAN GRAPHS

R. Jiménez

Departamento de Estadística, Universidad Simón Bolívar.
AP. 89000 Caracas 1080, Venezuela.
rjimenez@usb.ve

J. E. Yukich

Department of Mathematics, Lehigh University Bethlehem PA 18015, USA.
joseph.yukich@lehigh.edu

Abstract   A general approach, based on covering by cells, induced by Euclidean graphs, is developed to provide asymptotic characterizations of multivariate sample densities. This approach provides high dimensional analogs of basic results for random partitions based on one-dimensional sample spacings. The methods used in the proofs yield asymptotics for empirical ϕ-divergences based on k-spacings and also for the total edge length of the graphs involved.

10.1   Introduction and background

Statistics in the form of ϕ-divergences are used for several purposes including, among others, goodness-of-fit tests and parametric estimation. The Pearson χ^2 is a well known statistic of this type. They are in general designed for discrete or one dimensional continuous data. Although χ^2 and related methods can be used for continuous multivariate data, they are virtually useless in high dimensions. How to deal with empirical ϕ-divergences when the observations are continuous and multivariate has been a long-time need. Basically, the difficulty is to define suitable analogues in $\mathbb{R}^d$ of spacings on the line. In this work, we use random Euclidean graphs as adaptive schemes to define statistical distances of continuous samples in $\mathbb{R}^d$, $d \geq 1$. Formally, we prove strong laws for empirical ϕ-divergences based on multidimensional spacings induced by Euclidean graphs. In particular, these laws extend some basic results of sample

---

spacing theory on the line. While this work is related to [Jiménez, 2002], the approach taken here is considerably simpler and more general; it relies heavily on the objective method developed in [Aizenman, 1982; Ahmed, 2000] and more recently [Penrose, 2002A]. The methods also yield strong laws for the empirical ϕ-divergence for k-spacings.

10.1.1.　ϕ-divergence statistics for discrete data

Given a strictly convex function $\phi : \mathbb{R}^+ \to \mathbb{R}$, [Csiszár, 1978] ϕ-divergence between two nonnegative n-dimensional vectors $\mathbf{p} := (p_1, ..., p_n)$ and $\mathbf{q} := (q_1, ..., q_n)$ is

$$I_\phi(\mathbf{p}, \mathbf{q}) := \sum_{i=1}^n q_i \, \phi\left(\frac{p_i}{q_i}\right).$$

As in [Csiszár, 1967], we interpret undefined expressions by

$$\phi(0) = \lim_{t \to 0^+} \phi(t),$$

$$0 \, \phi\left(\frac{0}{0}\right) = 0,$$

$$0 \, \phi\left(\frac{a}{0}\right) = \lim_{\varepsilon \to 0^+} \varepsilon \, \phi\left(\frac{a}{\varepsilon}\right) = a \lim_{t \to \infty} \frac{\phi(t)}{t}. \tag{1.1}$$

Assuming that ϕ is normalized (that is $\phi(1) = 0$), and that $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i$, then Jensen's inequality implies

$$I_\phi(\mathbf{p}, \mathbf{q}) \geq 0 \text{ and } I_\phi(\mathbf{p}, \mathbf{q}) = 0 \text{ iff } \mathbf{p} = \mathbf{q}. \tag{1.2}$$

These are properties of a distance. However, I_ϕ is not a distance: the triangle inequality does not hold and I_ϕ is not symmetric, i.e., in general $I_\phi(\mathbf{p}, \mathbf{q}) \neq I_\phi(\mathbf{q}, \mathbf{p})$.

If we additionally assume that ϕ is nonnegative, then (1.2) holds even if $\sum_{i=1}^n p_i \neq \sum_{i=1}^n q_i$. On the other hand, for any strictly convex and normalized ϕ, the function ϕ^* defined by

$$\phi^*(x) := \phi(x) - (x-1) \lim_{h \to 0^+} \frac{\phi(1+h) + \phi(1-h) - 2\phi(1)}{2h}$$

is strictly convex, normalized, and nonnegative.
Moreover, if $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i$ then

$$I_{\phi^*}(\mathbf{p}, \mathbf{q}) = I_\phi(\mathbf{p}, \mathbf{q}).$$

Thus we can and will assume without loss of generality that ϕ is strictly convex, normalized, nonnegative, and that (1.2) holds whether $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i$ or not.

Property (1.2) makes ϕ-divergence useful in various fields. [Csiszár, 1978] reviews how ϕ-divergence can be used in statistics. Roughly speaking, if $\hat{\mathbf{p}}$ are observed frequencies then $I_\phi(\mathbf{p}, \hat{\mathbf{p}})$ or $I_\phi(\hat{\mathbf{p}}, \mathbf{p})$ can be used as loss-function in statistical inference. Frequently used ϕ-divergences in statistics involve the power-divergence family introduced by [Cressie, 1984]

$$\Psi := \left\{ \psi_\beta(t) = \frac{t^\beta - \beta t + \beta - 1}{\beta(\beta - 1)} : \ -\infty < \beta < \infty \right\}.$$

For example, when $\beta \to 0$ then $\psi_\beta(x) \to \psi_0(x) = -\log x + x - 1$. Thus

$$I_{\psi_0}(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{i=1}^n \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i}\right) \ \text{ and } \ I_{\psi_0}(\hat{\mathbf{p}}, \mathbf{p}) = \sum_{i=1}^n p_i \log\left(\frac{p_i}{\hat{p}_i}\right)$$

are the log-likelihood ratio and the Kullback-Leibler divergence respectively. When $\beta = 1/2$, $\psi_{1/2}(x) = 2(\sqrt{t} - 1)^2$ and

$$I_{\psi_{1/2}}(\mathbf{p}, \hat{\mathbf{p}}) = I_{\psi_{1/2}}(\hat{\mathbf{p}}, \mathbf{p}) = 2 \sum_{i=1}^n (\sqrt{p_i} - \sqrt{\hat{p}_i})^2$$

is the Hellinger distance. When $\beta = 2$, $\psi_2(x) = (t-1)^2/2$ and the statistics

$$I_{\psi_2}(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{i=1}^n \frac{(p_i - \hat{p}_i)^2}{2\hat{p}_i} \ \text{ and } \ I_{\psi_2}(\hat{\mathbf{p}}, \mathbf{p}) = \sum_{i=1}^n \frac{(p_i - \hat{p}_i)^2}{2p_i}$$

yields the χ^2 statistics of Neyman and Pearson respectively. The statistics $I_{\psi_\beta}(\hat{\mathbf{p}}, \mathbf{p})$ and $I_{\psi_\beta}(\mathbf{p}, \hat{\mathbf{p}})$ are one of the more important cases of statistical distances and have been used for several purposes including, among others, goodness-of-fit tests of discrete data ([Cressie, 1984]) and parametric estimation ([Lindsay, 1994]).

For any strictly convex, normalized, and nonnegative function $\phi(x)$ defined on $(0, \infty)$, its adjoint function $\phi^\diamond(x) = x\phi(1/x)$ is also strictly convex, normalized, and nonnegative. In particular, if $\psi_\beta \in \Psi$ then $\psi_\beta^\diamond \equiv \psi_{1-\beta}$. Since $I_{\phi^\diamond}(\mathbf{p}, \hat{\mathbf{p}}) = I_\phi(\hat{\mathbf{p}}, \mathbf{p})$, without loss of generality we will only consider the statistical distance $I_\phi(\mathbf{p}, \hat{\mathbf{p}})$ and we will omitted in the sequel its adjoint statistical distance $I_\phi(\hat{\mathbf{p}}, \mathbf{p})$. See [Jiménez, 2001] for some aspects related with divergence statistics and its adjoints.

10.1.2. Empirical ϕ-divergences based on spacings

The use of empirical ϕ-divergences with one dimensional continuous data is related with spacing theory as follows.   Consider the order statistics $X_1, X_2, \ldots, X_n$ of n independent random variables with common distribution F. Let $X_0 \equiv -\infty$. Then, the empirical estimate of the one dimensional

transformed spacing $F(X_i) - F(X_{i-1})$ is $1/n$. Thus,

$$I_\phi(F, n) \quad := \quad I_\phi(\{F(X_i) - F(X_{i-1}), 1 \le i \le n\}, \{\mathbf{1/n}\})$$

$$:= \quad \frac{1}{n} \sum_{i=1}^{n} \phi\left(n \cdot (F(X_i) - F(X_{i-1}))\right) \qquad (1.3)$$

can be viewed as a statistical distance between the sample distribution F and the empirical distribution. The importance of statistical distances based on spacings dates from the classic paper of [Pyke, 1965]. When F is unknown, the statistic $I_\phi(G, n)$ has been used to test the hypothesis $H_0 : G = F$. [Darling, 1953] provided the first systematic study of this statistic. If we assume that F is in some family of distributions $\mathcal{G}$, then F can be estimated by minimizing $I_\phi(G, n)$ over $G \in \mathcal{G}$. A remarkable case is given by $\psi_0(x) = -\log x + x - 1$ which corresponds to the maximum product of spacing method, introduced by [Cheng, 1983] and later by [Ranneby, 1984]. The strong consistency of the maximum product of spacing method and the strong consistency of the goodness-of-fit test based on $I_{\psi_0}(F, n)$ relies on the following strong law, proved by [Shao, 1995] under mild conditions on G and F,

$$\lim_{n \to \infty} I_{\psi_0}(G, n) = \int \mathbb{E}\left[\psi_0\left(\varepsilon \frac{dG}{dF}(x)\right)\right] dF(x) \quad a.s. \qquad (1.4)$$

Here and elsewhere ε is an exponential random variable with mean one.

The main result of [Holst, 1979] implies, for general ϕ, the asymptotic normality of the empirical ϕ-divergence based on $k-$spacings

$$S_\phi^k(G, n) = \frac{1}{n} \sum_{i=k}^{n} \phi(n[G(X_i) - G(X_{i-k})]),$$

under the hypothesis $H_0 : G = F$. This includes, for the particular case $k = 1$, the asymptotic normality of $I_\phi(F, n)$. Also the asymptotic normality of $S_\phi^k(G_n, n)$ has been studied for special sequences of alternatives G_n such that $G_n \to F$ when $n \to +\infty$; see [Hall, 1986] and its references. Under stringent regularity conditions on $G, F,$ and ϕ, [Holst, 1981] proved a central limit theorem for $I_\phi(G, n)$. However the asymptotic normality of $S_\phi^k(G, n)$ for fixed $G \ne F$ has been an open problem dating from the 1950's [Pyke, 1965].

10.2   The nearest neighbor ϕ-divergence and main results

ϕ-divergence We show in this work that random Euclidean graphs with a locally defined structure provide a natural scheme for generalizing one dimensional results based on spacings. We will first consider a scheme based on nearest neighbors.

For every sample point X_i consider the cell $C_i := C_i(X_1, ..., X_n)$ centered at X_i with radius equal to the distance to the nearest neighbor in the sample $\{X_1, ..., X_n\}$. We will use these cells to define a high dimensional spacing statistic analogous to the classical one-dimensional statistic. The cell C_i is of course a ball, but we prefer to call it a cell, since this anticipates more general spacing statistics described in the sequel. An attractive feature of these spacings is a *monotonicity* property identical to that for the classic one dimensional spacings: the cell around a given point decreases in volume as the number of points increases.

Throughout $X_1, X_2, ...$ are independent random variables in $\mathbb{R}^d$ with common probability density f, and g is an arbitrary probability density function.

DEFINITION 1 *For each $n \geq 1$, we define for $1 \leq i \leq n$ the sample spacings*

$$D_{i,n} := D_i(X_1, ..., X_n) := \int_{C_i(X_1, ..., X_n)} dx \qquad (2.1)$$

and the transformed spacings

$$D_{i,n}^g := D_i^g(X_1, ..., X_n) := \int_{C_i(X_1, ..., X_n)} g(x)dx. \qquad (2.2)$$

For all $1 \leq i \leq n$ we have $D_{i,n}^{g_1}(x_1, ..., x_n) \leq D_{i,n}^{g_2}(x_1, ..., x_n, y_1, ..., y_k)$ for any functions $0 \leq g_1 \leq g_2$. We will measure the discrepancy between g and the sample density f by comparing the transformed spacings $\{D_{i,n}^g, 1 \leq i \leq n\}$ with $\{D_{i,n}^f, 1 \leq i \leq n\}$.

We will use

$$N_\phi(\{D_{i,n}^g\}, \{D_{i,n}^f\}) := \sum_{i=1}^n D_{i,n}^f \phi\left(\frac{D_{i,n}^g}{D_{i,n}^f}\right) \qquad (2.3)$$

as a measure of the "distance" between g and f; we term this the "nearest neighbors ϕ-divergence". It is a discrete version induced by the balls of the nearest neighbors graph of [Csiszár, 1967] ϕ-divergence between g and f on B, namely

$$\int f(x)\phi\left(\frac{g(x)}{f(x)}\right) dx. \qquad (2.4)$$

If f is unknown, we can replace $D_{i,n}^f$ in (2.3) by its empirical estimate

$$\hat{D}_{i,n}^f := \frac{1}{n}.$$

In this manner, we obtain the following statistic, which we call the "empirical nearest neighbor ϕ-divergence", and which forms one of our central objects of

interest:

$$N_\phi^g := N_\phi^g(X_1, ..., X_n) := I_\phi(\{D_{i,n}^g\}, \{\hat{D}_{i,n}^f\}) = \frac{1}{n}\sum_{i=1}^{n} \phi(n \cdot D_{i,n}^g). \quad (2.5)$$

Our main purpose is to describe the a.s. behavior of $N_\phi^g(X_1, ..., X_n)$. We first introduce some notation.

DEFINITION 2 *Let Φ be the class of all normalized and strictly convex functions $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ such that there exists $\gamma > 0$ such that for all $\alpha > 0$ we have $\int_0^\infty \phi^{4+\gamma}(\alpha t)\exp(-t)dt < \infty$.*

It is easy to check that the frequently used ϕ's in statistics, including the power divergence family, are in the class Φ.

The following limit theorem, the main result of this section, establishes the a.s. consistency of the empirical nearest neighbor ϕ-**divergence**. We let A denote the support of f.

THEOREM 1 *Let $X_1, X_2, ...$ be independent random variables with a density f and let g be a continuous density. If f and g are bounded away from zero and infinity on A and if $\phi \in \Phi$, then*

$$\lim_{n\to\infty} N_\phi^g(X_1, ..., X_n) = \int_A f(x)\mathbb{E}\left[\phi\left(\varepsilon\frac{g(x)}{f(x)}\right)\right] dx \quad a.s. \quad (2.6)$$

The integral in (2.6) represents a divergence between f and g, which by Jensen's inequality and the identity $\mathbb{E}[\varepsilon] = 1$, exceeds the Csiszár divergence (2.4). Thus a small empirical nearest neighbors ϕ-**divergence** implies a small Csiszár divergence.

If $g(x) = f(x)$ a.e., then the right hand side of (2.6) equals $\mathbb{E}[\phi(\varepsilon)]$. On the other hand, if $g(x) \neq f(x)$ on some subset with positive Lebesgue measure, a combined application of Fubini's theorem and Jensen's inequality gives

$$\int_A f(x)\mathbb{E}\left[\phi\left(\varepsilon\frac{g(x)}{f(x)}\right)\right] dx = \mathbb{E}\left[\int_A f(x)\phi\left(\varepsilon\frac{g(x)}{f(x)}\right)dx\right] > \mathbb{E}[\phi(\varepsilon)].$$

Thus, using this notation we obtain the following corollary.

COROLLARY 1 *Under the same conditions of Theorem 1,*

$$\lim_{n\to\infty} N_\phi^g(X_1, ..., X_n) \geq \mathbb{E}[\phi(\varepsilon)] \quad a.s. \quad (2.7)$$

Moreover, there is strict inequality in (2.7) except for the case $g(x) = f(x)$ a.e.

In dimension $d = 1$, Theorem 1 is closely related with the empirical ϕ-**divergence** for k-**spacings**. The next theorem extends (1.4) to the context of k-**spacings** and general ϕ.

THEOREM 2 *Let $X_1, X_2, \ldots$ be independent real valued random variables with common density f. Let g be a continuous density and $G(x) = \int_{-\infty}^{x} g(u)du$. Let $\Gamma(k, 1)$ be a gamma random variable with parameters k and 1. If f and g are bounded away from zero and infinity on A and if $\phi \in \Phi$, then*

$$\lim_{n \to \infty} S_\phi^k(G, n) = \int_A f(x) \mathbb{E}\left[\phi\left(\Gamma(k, 1)\frac{g(x)}{f(x)}\right)\right] dx \quad a.s. \qquad (2.8)$$

Remark 2.1 It is a simple consequence of the uniform integrability of the left-hand side of (2.6) and (2.8) that the limits there also hold in L^1.

Remark 2.2 [Bickel, 1983] develop central limit theorems for statistics based on nearest neighbor distances. They consider the special case $\phi(x) = \exp(-x)$ and use the approximation

$$\int_{C_i(X_1, \ldots, X_n)} f(x)dx \sim f(X_i)|C_i(X_1, \ldots, X_n)|$$

and confine attention to sums $\sum_i \exp(-nf(X_i)|C_i|)$, where here and elsewhere $|C|$ denotes the volume of a set C. The strong consistency established by Theorem 1 can be viewed as an initial step in extending [Bickel, 1983] to more general ϕ. From the standpoint of goodness of fit tests, it would be desirable to supplement Theorem 1 with a central limit theorem for the empirical nearest neighbors ϕ-divergence and to provide an explicit formula for the limiting variance.

Remark 2.3. (a Shannon entropy estimate) The proof of Theorem 1 describes the large sample behavior of the sum-function of nearest neighbor spacings

$$\frac{1}{n} \sum_{i=1}^{n} \phi(n \cdot D_{i,n}).$$

These statistics provide estimates for entropy-type functionals of the sample density. To fix this idea consider $\phi(t) := \psi_0(t) = -\log t + t - 1$ and $g \equiv 1$. An elementary computation involving Theorem 1 and convention (1.1) imply

$$\frac{1}{n} \sum_{i=1}^{n} \log(n \cdot D_{i,n}) \to H(f) + \mathbb{E}\log(\varepsilon) \quad a.s.,$$

where $H(f) := -\int f(x) \log f(x)dx$ is the well-known Shannon entropy. Estimates of $H(f)$ are of general interest; see [Dudewicz, 1987] for a review of the one-dimensional case. They can be used in the context of the maximum entropy method which has wide applications in several fields.

Remark 2.4. (equivalence with maximum likelihood) Suppose that the sample density $f(x)$, belongs to the parametric family $\mathcal{F} := \{f_\theta(x) : \theta \in \Theta\}$. Let θ_0 be such that $f(x) = f_{\theta_0}(x)$. The maximum likelihood (ML) estimate of θ_0 is obtained by maximizing the log-likelihood function

$$l_n(\theta) := \sum_{i=1}^{n} \log f_\theta(X_i).$$

Let $K(f, g)$ denote the Kullback-Leibler relative entropy, that is

$$K(f, g) := \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx.$$

By the strong law of large numbers, $K(f_{\theta_0}, f_\theta) < \infty$ implies

$$\frac{1}{n}(l_n(\theta_0) - l_n(\theta)) \to K(f_{\theta_0}, f_\theta) \quad a.s.$$

On the other hand, if f_θ is bounded away from zero and infinity on the support of f_{θ_0}, then by Theorem 1 we have

$$\frac{1}{n} \sum_{i=1}^{n} \log(n \cdot D_{i,n}^{f_\theta}) \to -K(f_{\theta_0}, f_\theta) + \mathbb{E}[\log(\varepsilon)] \quad a.s. \qquad (2.9)$$

Thus, under general conditions, maximizing the log-likelihood function is asymptotically equivalent to maximizing the left-hand side of (2.9). We will call

$$\hat{\theta}_\phi := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \phi(n \cdot D_{i,n}^{f_\theta}) \qquad (2.10)$$

the minimum nearest neighbors ϕ-**divergence** (MϕD) estimate. Roughly speaking, the ML estimate and $\hat{\theta}_{-\log}$ are asymptotically equivalent.

Remark 2.5. (multivariate version of maximum spacing method) Under general conditions, the ML estimate can have optimal asymptotic properties and thus $\hat{\theta}_{-\log}$ must have the same type of asymptotic properties. However, when the likelihood function is unbounded, the ML estimate can be inconsistent. The MϕD method is a multivariate version of the maximum product of spacing (MPS) method, which is an alternative to the ML method when the likelihood function is unbounded. Since the sum of the logarithm of spacings is always upper bounded, even in the cases where the ML method fails, the MPS method can generate asymptotically optimal estimates. This feature can be observed for example in many mixture models, which are not necessarily restricted to the one dimensional case. Similarly to the one dimensional case, the empirical nearest neighbors ϕ-**divergence** is always lower bounded. Thus, the MϕD method can generate consistent estimates even when the ML method fails.

Remark 2.6. (consistency of MϕD estimates) For $\phi(t) := -\log t$, Theorem 1 resembles the asymptotics (1.4) for the logarithm sum of one-dimensional spacings obtained by [Shao, 1999]. Information-type inequalities such as Corollary 1 play a key role in proving strong consistency of the MPS method ([Shao, 1999]) and related one-dimensional methods. In the same way, our results can be applied to prove strong consistency of the estimate $\hat{\theta}_\phi$ defined in (2.10). For example, Corollary 1 implies that the estimate $\hat{\theta}_\phi$ is always consistent for any $\phi \in \Phi$, if Θ is finite. General consistency theorems may be obtained assuming regularity conditions on $\mathcal{F}$.

10.3 Statistical distances based on Voronoi cells

Theorem 1 shows the efficacy of using random graphs based on nearest neighbor distances to define statistical distances which generalize consistency results for one dimensional spacings to higher dimensions. Nearest neighbor graphs are easy to generate but in some cases it may be advantageous to consider statistical distances using other graphs which have a strong locally defined structure. We illustrate the possibilities by considering graphs involving Voronoi tessellations.

Voronoi tessellations generated by random sets of points are of general interest and have been used in many diverse fields ([Aurenhammer, 1991], [Obake, 1992], [Möller, 1994]). Much like nearest neighbors graphs, Voronoi tessellations may be used as an adaptive scheme to compare probability densities on $\mathbb{R}^d$, $d \geq 1$.

Given a set of points $\mathcal{X} := \{x_1, ..., x_n\} \subset \mathbb{R}^d$ and a Borel subset B of $\mathbb{R}^d$, consider for any $x_i \in \mathcal{X} \cap B$ the locus of points closer to x_i than to any other point of $\mathcal{X} \cap B$. The intersection of this set of points with B is a Voronoi cell and is denoted by $V_i(B) := V_i(B; x_1, ..., x_n)$, that is

$$V_i(B; x_1, ..., x_n) := \{y \in B : \|y - x_i\| \leq \|y - x_j\|, \ \forall x_j \in \mathcal{X} \cap B\},$$

where $\|\cdot\|$ denotes the Euclidean distance. If $x_i \notin B$ then we define $V_i(B) = \emptyset$. Thus, $\{V_i(B), 1 \leq i \leq n\}$ is a partition of B which is called the *Voronoi tessellation* of B generated by $\mathcal{X}$ and is denoted by $\mathcal{V}(B; \mathcal{X})$. It is understood that if $\mathcal{X} \cap B = \emptyset$ then $\mathcal{V}(B; \mathcal{X}) = B$. Also, if $X_1, X_2, ...$ are i.i.d. with a density whose support is A, then we reserve the notation $V_i(X_1, ..., X_n)$ for $V_i(A; X_1, ..., X_n)$. Figure 1 shows the Voronoi tessellation generated by a uniform random sample on the unit square.

We may use the Voronoi cells to define high dimensional sample spacings as follows.

DEFINITION 3 *For each $n \geq 1$, we define for $1 \leq i \leq n$ the sample spacings*

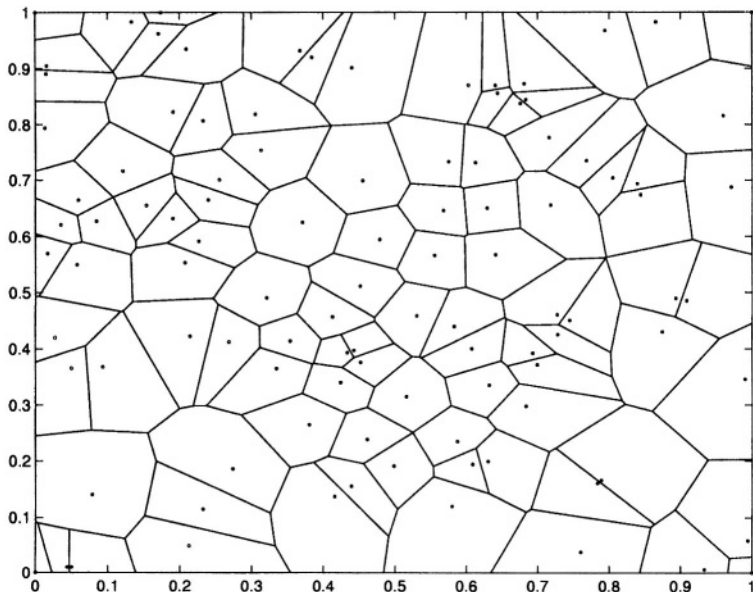$$D_{i,n} := D_i(X_1, ..., X_n) := \int_{V_i(X_1, ..., X_n)} dx \tag{3.1}$$

Figure 1.   Voronoi tessellation of the unit square.

and the transformed spacings

$$D^g_{i,n} := D^g_i(X_1, ..., X_n) := \int_{V_i(X_1,...,X_n)} g(x)dx. \qquad (3.2)$$

Exactly as in the context of the nearest neighbors graph, we will measure the discrepancy between g and the sample density f by comparing the transformed spacings $\{D^g_{i,n}, 1 \le i \le n\}$ with $\{D^f_{i,n}, 1 \le i \le n\}$.

We thus obtain the following statistic, which we call the "empirical Voronoi ϕ-divergence", and which forms the natural analog of the empirical nearest neighbors ϕ-divergence (2.5):

$$V^g_\phi := V^g_\phi(X_1, ..., X_n) := I_\phi(\{D^g_{i,n}\}, \{\hat{D}^f_{i,n}\}) = \frac{1}{n} \sum_{i=1}^n \phi(n \cdot D^g_{i,n}). \qquad (3.3)$$

The following main result is the Voronoi analog of Theorem 1. Let $\mathcal{P}_1$ denote a homogeneous Poisson point process of constant intensity 1 on $\mathbb{R}^d$, let **0** denote the origin of $\mathbb{R}^d$, and let ε_V denote the volume of the Voronoi cell around **0** in the Voronoi tessellation on $\mathcal{P}_1 \cup \mathbf{0}$. While Theorem 3 is similar to Theorem 1.1 of [Jiménez, 2002], which assumes continuity of f, the method of proof is much easier and follows the relatively simple proof of Theorem 1.

THEOREM 3 *Let $X_1, X_2, \ldots$ be independent random variables with a density f and let g be a continuous density. If f and g are bounded away from zero and infinity on A and if $\phi \in \Phi$, then*

$$\lim_{n \to \infty} V_{\phi}^{g}(X_1, \ldots, X_n) = \int_A f(x) \mathbb{E}\left[\phi\left(\varepsilon_V \frac{g(x)}{f(x)}\right)\right] dx \quad a.s. \qquad (3.4)$$

Other Euclidean graphs may also be used as adaptive schemes to compare probability densities. For this, we must define the cells around the sample points according to the geometric characteristics of the considered graph. Thus, empirical **ϕ-divergences** can be defined analogously to (3.3) and in general they satisfy a.s. asymptotics of the form (3.4), with ε_V replaced by the volume of the related cell around the origin induced by the graph on $\mathcal{P}_1 \cup \mathbf{0}$.

10.4   The objective method

Theorem 1 is anticipated by Theorems 2.2 and 2.4 of [Penrose, 2002A], which uses the objective method to establish a weak law of large numbers for stabilizing functionals of random variables. Similarly Theorem 3 is anticipated by Theorem 2.5 of [Penrose, 2002A]. However, neither Theorem 1 nor Theorem 3 is a consequence of [Penrose, 2002A] since neither the nearest neighbors nor Voronoi statistic is translational invariant (translating the sample points changes the statistic according to the density g). Thus one needs to modify existing methods in order to establish Theorems 1 and 3. In the first part of this section we prove Theorem 1. Completely similar methods may be used to prove Theorem 3.

Let A denote the support of f and for all $\langle > 0$, let $\mathcal{P}_{\langle}^{f}$ denote a Poisson point process with intensity measure $\langle f : A \to \mathbb{R}$. To prove Theorem 1, we start by showing that a Poissonized version of (2.6) holds in expectation, namely we show that if we only assume $\mathbb{E}[\phi(\alpha\varepsilon)] < \infty$ for all $\alpha > 0$, then

$$\lim_{\langle \to \infty} \int_A \mathbb{E}\left[\phi(\lambda \int_{C(x, \mathcal{P}_{\langle}^{f})} g(u)du\right] f(x)dx = \int_A \mathbb{E}\left[\phi\left(\varepsilon \frac{g(x)}{f(x)}\right)\right] f(x)dx.$$

$$(4.1)$$

The proof of (4.1) may be established using lengthy and somewhat cumbersome methods, as in [Jiménez, 2002], which actually requires continuity of f. It is more instructive and much easier to use the following key lemma, which further illustrates the power of the objective method [Aizenman, 1982; Ahmed, 2000].

To set the stage, we note that for fixed x and for large $\langle$, the volume of the cell $C(x, \mathcal{P}_{\langle}^{f})$ when multiplied by $\langle$, is roughly the same as the volume of the cell $C(x, \mathcal{P}_{f(x)})$. Here and elsewhere $\mathcal{P}_{\tau}$ denotes a homogeneous Poisson

point process on $\mathbb{R}^d$ with intensity τ. Since for all τ we have $\tau|C(x, \mathcal{P}_\tau)| \stackrel{\mathcal{D}}{=} |C(\mathbf{0}, \mathcal{P}_1)| = \varepsilon$, this suggests the following lemma, where here and elsewhere, $\stackrel{P}{\longrightarrow}$ denotes convergence in probability.

LEMMA 1 *For almost all $x \in A$ we have as $\langle \to \infty$*

$$\langle f(x)|C(x, \mathcal{P}_\langle^f)| \stackrel{P}{\longrightarrow} \varepsilon. \tag{4.2}$$

We defer the proof of Lemma 1 and show how to use it to deduce Theorem 1. By hypothesis we have positive finite constants k_1 and k_2 such that for all $x \in A$

$$k_1 \le f(x) \le k_2 \text{ and } k_1 \le g(x) \le k_2. \tag{4.3}$$

Since $\varepsilon \frac{g(x)}{f(x)}$ has the same distribution as $\langle \int_{C(x, \mathcal{P}_{\langle f(x)\rangle})} g(x)du$, we need only to show for almost all $x \in A$ that

$$\lim_{\langle \to \infty} \left| \mathbb{E}\left[\phi\left(\lambda \int_{C(x, \mathcal{P}_\langle^f)} g(u)du \right) - \phi\left(\langle \int_{C(x, \mathcal{P}_{\langle f(x)\rangle})} g(x)du \right) \right] \right| = 0 \tag{4.4}$$

as $\langle \to \infty$. Letting $x' := x'(\langle)$ denote a point in the cell $C(x, \mathcal{P}_\langle^f)$ such that $g(x')|C(x, \mathcal{P}_\langle^f)| = \lambda \int_{C(x, \mathcal{P}_\langle^f)} g(u)du$ we equivalently only need to show that

$$\left| \mathbb{E}\left[\phi\left(\langle g(x')|C(x, \mathcal{P}_\langle^f| \right) - \phi\left(\varepsilon \frac{g(x)}{f(x)} \right) \right] \right| \to 0 \tag{4.5}$$

as $\langle \to \infty$.

Now (4.5) is bounded by the sum of

$$\left| \mathbb{E}\left[\phi\left(\langle f(x) \frac{g(x')}{f(x)}|C(x, \mathcal{P}_\langle^f)| \right) - \phi\left(\frac{g(x')}{f(x)}\varepsilon \right) \right] \right| \tag{4.6}$$

and

$$\left| \mathbb{E}\left[\phi\left(\frac{g(x')}{f(x)}\varepsilon \right) - \phi\left(\frac{g(x)}{f(x)}\varepsilon \right) \right] \right|. \tag{4.7}$$

Given a convex function $\phi \in \Phi$, let $\phi_1(x) := \phi(x)1_{(0,1)}(x)$ be its decreasing part and let $\phi_2(x) := 1_{[1,+\infty)}(x)$ be its increasing part. By Lemma 1 and the continuity of ϕ,

$$\phi_2\left(\langle f(x) \frac{g(x')}{f(x)}|C(x, \mathcal{P}_\langle^f)| \right) - \phi_2\left(\frac{g(x')}{f(x)}\varepsilon \right)$$

tends to zero in probability. Since ϕ_2 is increasing we have for all $\langle > 0$

$$\mathbb{E}\left[\phi_2\left(\langle f(x) \frac{g(x')}{f(x)}|C(x, \mathcal{P}_\langle^f)| \right) \right] \le \mathbb{E}\left[\phi_2\left(\frac{k_2\varepsilon}{k_1} \right) \right]$$

and thus the assumed integrability of $\phi_2(\alpha\varepsilon)$, $\alpha > 0$, shows that

$$\phi_2\left(\langle f(x)\frac{g(x')}{f(x)}|C(x,\mathcal{P}_\langle^f)|\right) - \phi_2\left(\frac{g(x')}{f(x)}\varepsilon\right), \quad \langle > 0,$$

are uniformly integrable. Thus, ([Dudley, 1989], Thm 10.3.6)

$$\mathbb{E}\left|\phi_2\left(\langle f(x)\frac{g(x')}{f(x)}|C(x,\mathcal{P}_\langle^f)|\right) - \phi_2\left(\frac{g(x')}{f(x)}\varepsilon\right)\right| \to 0. \qquad (4.8)$$

Similarly, since ϕ_1 is decreasing

$$\mathbb{E}\left[\phi_1\left(\langle f(x)\frac{g(x')}{f(x)}|C(x,\mathcal{P}_\langle^f)|\right)\right] \leq \mathbb{E}\left[\phi_1\left(\frac{k_1\varepsilon}{k_2}\right)\right],$$

showing that

$$\phi_1\left(\langle f(x)\frac{g(x')}{f(x)}|C(x,\mathcal{P}_\langle^f)|\right) - \phi_1\left(\frac{g(x')}{f(x)}\varepsilon\right), \quad \langle > 0,$$

are also uniformly integrable. Thus splitting ϕ as $\phi_1 + \phi_2$, we see that the difference (4.6) tends to zero as $\langle \to \infty$. Similarly, by the continuity of g we have as $\langle \to \infty$

$$\frac{g(x')}{f(x)}\varepsilon \xrightarrow{\quad P \quad} \frac{g(x)}{f(x)}\varepsilon$$

and together with uniform integrability arguments, this shows that the difference (4.7) also tends to zero as $\langle \to \infty$. Thus (4.4) tends to zero as desired.

Since the cell $C := C(x,\mathcal{P}_\langle^f)$ is a nearest neighbors cell, it depends only locally on the surrounding points and this localization, together with the moment condition $\int_0^\infty \phi^{4+\gamma}(\alpha t)\exp(-t)dt < \infty$, makes it straightforward to de-Poissonize the mean limit (4.1). This can be accomplished by following verbatim Lemma 2.5 of [Jiménez, 2002].

Since the density f is assumed bounded away from zero and infinity and since the volume of the nearest neighbor cell around x with high probability depends on sample points distant $C(\log n/n)^{1/d}$ from x, we may follow the proof of Lemma 3.1 of [Jiménez, 2002] and use isoperimetric, arguments to establish that the difference of our de-Poissonized statistic with its mean, namely $|\frac{1}{n}\sum_{i=1}^n \phi(n \cdot D_{i,n}^g) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\phi(n \cdot D_{i,n}^g)|$, is almost surely of order $o(1)$, showing that convergence of the mean is equivalent to a.s. convergence. We leave these details to the reader.

It only remains to prove Lemma 1. For $r > 0$, let $B_r(x)$ denote the Euclidean ball $\{y \in \mathbb{R}^d : |y - x| \leq r\}$ of radius r centered at x.

Proof of Lemma 1. Given $\tau > 0$, recall that $\mathcal{P}_\tau$ denotes a homogeneous Poisson point process on $\mathbb{R}^d$ with intensity τ. For all $z \in \mathbb{R}^d$, let $C(z, \mathcal{P}_\tau)$ denote the nearest neighbors cell around z with respect to $\mathcal{P}_\tau$. Note that for all $\tau > 0$ we have $\tau|C(\mathbf{0}, \mathcal{P}_\tau)| \overset{\mathcal{D}}{=} |C(\mathbf{0}, \mathcal{P}_1)| \overset{\mathcal{D}}{=} \varepsilon$, since the volume of the nearest neighbors cell around the origin is a mean one exponential random variable.

$C(z, \mathcal{P}_\tau)$ is locally defined in the sense (section 6 of [Penrose, 2001]) that there is a random variable $R := R(z, \tau)$ with exponentially decaying tails and an a.s. finite random variable $C_\infty(z, \mathcal{P}_\tau)$ such that

$$C_\infty(z, \mathcal{P}_\tau) = C(z, \mathcal{P}_\tau \cap B_R(z) \cup \mathcal{A})$$

for all locally finite $\mathcal{A}$ outside $B_R(z)$.

Given $\mathcal{P}_\langle^f$, the Poisson point process with intensity $\langle f : A \to \mathbb{R}^+$, for all $x \in A$ let $\mathcal{P}_{\langle f(x)}$ be a homogeneous Poisson point process with constant intensity $\langle f(x)$. We may assume that $\mathcal{P}_{\langle f(x)}$ is *coupled* to $\mathcal{P}_\langle^f$ in such a way that for all Borel sets $B \subset A$ we have

$$P[\mathcal{P}_\langle^f \neq \mathcal{P}_{\langle f(x)}] \leq \langle \int_B |f(x) - f(y)| dy. \tag{4.9}$$

Next, for any Lebesgue point x for f, $x \in A$, for all $\langle,\, t \in \mathbb{R}^+$ consider the event

$$E(x, \langle, t) := \{R(\langle^{1/d}x, \langle^{1/d}\mathcal{P}_{\langle f(x)}) < t,\ \langle^{1/d}\mathcal{P}_\langle^f = \langle^{1/d}\mathcal{P}_{\langle f(x)} \text{ on } B_t(\langle^{1/d}x)\}.$$

By (4.9) we have that $P[E(x, \langle, t)^c]$ is bounded by

$$P[R(\langle^{1/d}x, \langle^{1/d}\mathcal{P}_{\langle f(x)}) > t] + \langle \int_{B_t(\langle^{1/d}x)} |f(x) - f(y)| dy. \tag{4.10}$$

Since f is Lebesgue integrable and since x is a Lebesgue point for f, the integral in (4.10) tends to zero as $\lambda \to \infty$. Since $R(\langle^{1/d}x, \langle^{1/d}\mathcal{P}_{\langle f(x)})$ has the same distribution as $R(\mathbf{0}, \mathcal{P}_{f(x)})$, which is finite a.s., it follows that if t is large enough, then the first term in (4.10) tends to zero as $\langle \to \infty$. Therefore, for all $\delta > 0$ and for $\langle$ and t large enough,

$$P[E(x, \langle, t)^c] < \delta$$

Now we can prove Lemma 1 as follows. We observe

$$\langle f(x) \cdot |C(x, \mathcal{P}_\langle^f)|$$

$$= \langle f(x) \cdot |C(x, \mathcal{P}_\langle^f)| \cdot 1_{E(x,\langle,t)} + \langle f(x) \cdot |C(x, \mathcal{P}_\langle^f)| \cdot 1_{E(x,\langle,t)^c}$$

$$= f(x) \cdot |C(\langle^{1/d}x, \langle^{1/d}\mathcal{P}_\langle^f)| \cdot 1_{E(x,\langle,t)} + \langle f(x) \cdot |C(x, \mathcal{P}_\langle^f)| \cdot 1_{E(x,\langle,t)^c}$$

$$= f(x)|C(\langle^{1/d}x, \langle^{1/d}\mathcal{P}_\langle^f \cap B_{R((\langle^{1/d}x,\langle f(x)\rangle))}| \cdot 1_{E(x,\langle,t)}$$
$$+ \langle f(x) \cdot |C(x, \mathcal{P}_\langle^f)| \cdot 1_{E(x,\langle,t)^c}$$

$$= f(x) \cdot C(\langle^{1/d}x, \langle^{1/d}\mathcal{P}_{\langle f(x)}) \cdot 1_{E(x,\langle,t)} + \langle f(x) \cdot C(x, \mathcal{P}_\langle^f| \cdot 1_{E(x,\langle,t)^c},$$

where the last equality holds since on the set $E(x,\langle,t)$ we have $\langle^{1/d}\mathcal{P}_\langle^f = \langle^{1/d}\mathcal{P}_{\langle f(x)}$.

The above is equal in distribution to

$$f(x) \cdot |C(\mathbf{0}, \mathcal{P}_{f(x)})| - f(x)|C(\mathbf{0}, \mathcal{P}_{f(x)})| \cdot 1_{E(x,\langle,t)^c} + \langle f(x) \cdot |C(x, \mathcal{P}_\langle^f)| \cdot 1_{E(x,\langle,t)^c} \tag{4.11}$$

The first term is equal in distribution to $|C(\mathbf{0}, \mathcal{P}_1)|$ and the last two terms in (4.11) tend to zero in probability as $\langle \to \infty$ and $t \to \infty$. This follows from the probability estimate $P[E(x,\langle,t)^c] < \delta$ as well as the bounds $\mathbb{E}\left[\langle |C(x, \mathcal{P}_\langle^f)| \right] \leq \mathbb{E}[|C(x, \mathcal{P}_{k_1})|]$ and $\mathbb{E}[|C(x, \mathcal{P}_{k_1})|^p] < \infty$ for all $x \in \mathbb{R}^d$ and all $p > 1$.

This completes the proof of Lemma 1. ∎

It only remains to give the proof of Theorem 2. Since the methods are very similar, we only give a sketch.

Proof of Theorem 2. We will follow the proof of Theorem 1 closely. Let $\mathcal{P}_\tau$ be a homogeneous Poisson point process on $\mathbb{R}$ with constant intensity τ. Let $\{X_i\}$ be the realization of $\mathcal{P}_\tau$ and let $X_{(i)}$ be the usual order statistics. For any $x \in \mathbb{R}$, let $C_k(x, \mathcal{P}_\tau) = X_{k(x)} - x$ denote the length of the associated k-spacing, where $X_{k(x)}$ is the k-th point in $\mathcal{P}_\tau$ to the right of x. The proof of Theorem 2 depends upon the following lemma.

LEMMA 2 *For almost all $x \in A$ we have as $\langle \to \infty$*

$$\langle f(x)|C_k(x, \mathcal{P}_\langle^f)| \xrightarrow{P} \Gamma(k, 1). \tag{4.12}$$

To prove this lemma, we simply follow the proof of Lemma 1 with $C_k(x, \mathcal{P}_\langle^f)$ replacing $C(x, \mathcal{P}_\langle^f)$ and note that for all $\tau > 0$ we have

$$\tau|C_k(x, \mathcal{P}_\tau)| \overset{\mathcal{D}}{=} C_k(0, \mathcal{P}_1) \overset{\mathcal{D}}{=} \Gamma(k, 1).$$

Now just follow the proof of Theorem 1. ∎

References

Aldous, D. and J.M. Steele (1992). Asymptotics for Euclidean minimal spanning trees on random points. *Probab. Theory Related Fields,* **92**, 247-258.

Aldous, D. and J.M. Steele (2002). The objective method: probabilistic combinatorial optimization and local weak convergence. *Encyclopedia of Mathematics,* to appear.

Aurenhammer, F. (1991). Voronoi diagrams - A survey of a fundamental geometric data structure. *ACM Computing Surveys,* **23**, 3, 345-405.

Beardwood, J., Halton, J. H., and J. M. Hammersley (1959). The shortest path through many points. *Proc. Camb. Philos. Soc.,* **55**, 299-327.

Bickel, P. and L. Breiman (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Annals of Prob.,* **11**, 185-214.

Cheng, R. C. H. and Amin N. A. K. (1983). Estimating parameters in continuous univariate distributions with shifted origin. *J. R. Statist. Soc. B,* **45**, 394-403.

Cressie, N and T. R. C. Read (1984). Multinomial goodness-of-fit tests. *J. R. Statist. Soc. B,* **46**, 440-464.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations *Studia Sci. Math. Hungarica,* **2**, 299-318.

Csiszár, I. (1978). Information measures: A critical survey. *Transaction 7th Prague Conf. on Info. Th Statist., Decis. Funct., Random Process and 8th European Meeting of Statist.* Academia, Prague, 73-86.

Darling, D. A. (1953). On a class of problems related to the random division of an interval. *Ann. Math. Statist.,* **24**, 239-253.

Dudewicz, E. J. and E. C. Van der Meulen (1987). The empiric entropy, a new approach to nonparametric density estimation. *New perspectives in theoretical and applied statistics. Eds. M. I. Puri, J. Vilaplana and M. Wertz.*Wiley, New York, 202-227.

Dudley, R. M. (1989). *Real Analysis and Probability.* Wadsworth and Brooks/Cole.

Hall, P. (1986). On powerful distributional tests based on sample spacings. *J. Multi. Anal.* **19**, 201-224.

Holst, L. (1979). Asymptotic normality of sum-functions of spacings. *Annals of Prob.,* **7**, 1066-1072.

Holst, L. and J. S. Rao (1981). Asymptotic spacings theory with applications to the two-sample problem. *Canadian J. Statist.,* **9**, 79-89.

Jiménez, R. and Y. Shao (2001). On robustness and efficiency of minimum divergence estimators, *Test,* **10**, 2, 241-248.

Jiménez, R. and J. E. Yukich (2002). Asymptotics for statistical distances based on Voronoi tessellations, *Journal of Theoretical Probability,* **15**, 2, 503-541.

Jimenez, R. and J. E. Yukich (2002). Strong laws for Euclidean graphs with general edge weights, *Statist. Probab. Lett.,* **56**, 251-259.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Annals of Statistics,* **22**, 1081-1114.

McGivney, K. and J.E. Yukich (1999). Asymptotics for Voronoi tessellations on random samples. *Stochastic Process. Appl.,* **83**, 273-288.

Möller, J. (1994). *Lectures on Random Voronoi Tessellations. Lecture Notes in Statistics,* **87**, Springer-Verlag.

Okabe, A., Boots, B., and K. Sugihara (1992). *Spatial Tessellations - Concepts and Applications of Voronoi Diagrams.* J. Wiley and Sons, England.

Penrose, M. D. and J.E. Yukich (2001). Central limit theorems for some graphs in computational geometry. *Ann. Appl. Probab.* **11,** 1005-1041.

Penrose, M. D. and J.E. Yukich (2002). Weak laws of large numbers in geometric probability. *Ann. Appl. Probab.,* to appear.

Pyke, R. (1965). Spacings. *Royal Statist. Soc. B,* **27**, 395-436.

Ranneby, B. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. *Scand. J. Statist.,* **11,** 93-112.

Shao, Y. and M. G. Hahn (1995). Limit theorems for logarithm of sample spacings. *Statistics and Probability Letters,* **24,** 121-132.

Shao, Y. and M. G. Hahn (1999). Strong consistency of the maximum product of spacings estimates with applications in nonparametrics and in estimation of unimodal densities. *Ann. Inst. Statist.,* **51,** Math. 31-49.

Steele, J. M. (1997). *Probability Theory and Combinatorial Optimization.* SIAM.

Yukich, J. E. (1998). *Probability Theory of Classical Euclidean Optimization Problems. Lecture Notes in Mathematics,* **1675,** Springer, Berlin.

This page intentionally left blank

IMPLIED VOLATILITY: STATICS, DYNAMICS, AND PROBABILISTIC INTERPRETATION

Roger W. Lee
Department of Mathematics, Stanford University;
and Courant Institute of Mathematical Sciences, NYU.

Abstract Given the price of a call or put option, the Black-Scholes *implied volatility* is the unique volatility parameter for which the Black-Scholes formula recovers the option price. This article surveys research activity relating to three theoretical questions: First, does implied volatility admit a probabilistic interpretation? Second, how does implied volatility behave as a function of strike and expiry? Here one seeks to characterize the shapes of the implied volatility skew (or smile) and term structure, which together constitute what can be termed the *statics* of the implied volatility surface. Third, how does implied volatility evolve as time rolls forward? Here one seeks to characterize the *dynamics* of implied volatility.

11.1   Introduction

11.1.1.   Implied volatility

Assuming that an underlying asset in a frictionless market follows geometric Brownian motion, which has constant volatility, the Black-Scholes formula gives the no-arbitrage price of an option on that underlying. Inverting this formula, take as given the price of a call or put option. The Black-Scholes *implied volatility* is the unique volatility parameter for which the Black-Scholes formula recovers the price of that option.

This article surveys research activity in the theory of implied volatility. In light of the compelling empirical evidence that volatility is *not* constant, it is natural to question why the inversion of option prices in an "incorrect" formula should deserve such attention.

To answer this, it is helpful to regard the Black-Scholes implied volatility as a *language* in which to express an option price. Use of this language does not entail any belief that volatility is actually constant. A relevant analogy is the quotation of a discount bond price by giving its yield to maturity, which is the interest rate such that the observed bond price is recovered by the usual

constant interest rate bond pricing formula. In no way does the use or study of bond yields entail a belief that interest rates are actually constant. As YTM is just an alternative way of expressing a bond price, so is implied volatility is just an alternative way of expressing an option price.

The language of implied volatility is, moreover, a *useful* alternative to raw prices. It gives a metric by which option prices can be compared across different strikes, maturities, and underlyings, and by which market prices can be compared to assessments of fair value. It is a standard in industry, to the extent that traders quote option prices in "vol" points, and exchanges update implied volatilty indices in real time.

Furthermore, to whatever extent implied volatility has a simple interpretation as an average future volatility , it becomes not only useful, but also *natural.* Indeed, understanding implied volatility as an average will be one of the focal points of this article.

11.1.2.   Outline

Under one interpretation, implied volatility is the market's expectation of future volatility, time-averaged over the term of the option. In what sense does this interpretation admit mathematical justification? In section 2 we review the progress on this question, in two contexts: first, under the assumption that instantaneous volatility is a deterministic function of the underlying and time; and second, under the assumption that instantaneous volatility is stochastic in the sense that it depends on a second random factor.

If instantaneous volatility is not constant, then implied volatilities will exhibit variation with respect to strike (described graphically as a *smile* or *skew*) and with respect to expiry (the *term structure*); the variation jointly in strike and expiry can be described graphically as a *surface.* In section 3, we review the work on characterizing or approximating the shape of this surface under various sets of assumptions. Assuming only absence of arbitrage, one finds bounds on the slope of the volatility surface, and characterizations of the tail growth of the volatility skew. Assuming stochastic volatility dynamics for the underlying, one finds perturbation approximations for the implied volatility surface, in any of a number of different regimes, including long maturity, short maturity, fast mean reversion, and slow mean reversion.

Whereas sections 2 and 3 examine how implied volatility behaves under certain assumptions on the spot process, section 4 directly takes as primitive the implied volatility, with a view toward modelling accurately its time-evolution. We begin with the no-arbitrage approach to the direct modelling of stochastic implied volatility. Then we review the statistical approach, Whereas the focus of section 3 is cross-sectional (taking a "snapshot" of all strikes and expiries)

hence the term *statics,* the focus of section 4 is instead time-series oriented, hence the term *dynamics.*

11.1.3. Definitions

Our underlying asset will be a non-dividend paying stock or index with non-negative price process S_t. Generalization to non-zero dividends is straightforward.

A call option on *S,* with strike *K* and expiry *T,* pays $(S_T - K)^+$ at time *T.* The price of this option is a function *C* of the contract variables *(K, T)*, today's date $t,$ the underlying $S_t,$ and any other state variables in the economy. We will suppress some or all of these arguments. Moreover, sections 2 and 3 will for notational convenience assume $t = 0$ unless otherwise stated; but section 4, in which the time-evolution of option prices becomes more important, will not assume $t = 0$.

Let the risk-free interest rate be a constant r. Write

$$x := \log\left(\frac{K}{S_t e^{r(T-t)}}\right)$$

for *log-moneyness* of an option at time t. Note that both of the possible choices of sign convention appear in the literature; we have chosen to define log-moneyness to be such that x has a *positive* relationship with *K*.

Assuming frictionless markets, Black and Scholes [Black & Scholes, 1973] showed that if *S* follows geometric Brownian motion

$$dS_t = \mu S_t dt + \sigma S_t d\tilde{W}_t$$

then the no-arbitrage call price satisfies

$$C = C^{BS}(\sigma),$$

where the Black-Scholes formula is defined by

$$C^{BS}(\sigma) := C^{BS}(S_t, t, K, T, \sigma) := S_t N(d_1) - K e^{-r(T-t)} N(d_2).$$

Here

$$d_{1,2} := \frac{\log(S_t e^{r(T-t)}/K)}{\sigma\sqrt{T-t}} \pm \frac{\sigma\sqrt{T-t}}{2},$$

and *N* is the cumulative normal distribution function.

On the other hand, given *C(K,T)*, the *implied [Black-Scholes] volatility* for strike *K* and expiry *T* is defined as the *I(K, T)* that solves

$$C(K,T) = C^{BS}(K, T, I(K, T)).$$

The solution is unique because C^{BS} is strictly increasing in σ, and as $\sigma \to 0$ (resp. ∞), the Black-Scholes function $C^{BS}(\sigma)$ approaches the lower (resp. upper) no-arbitrage bounds on a call.

Implied volatility can also be written as a function $\tilde{I}$ of log-moneyness and time, so $\tilde{I}(x,T) := I(S_t e^{x+r(T-t)}, T)$. Abusing notation, we will drop the tilde on $\tilde{I}$, because the context will make clear whether I is to be viewed as a function of K or x.

The derivation of the Black-Scholes formula can proceed by means of a hedging argument that yields a PDE to be solved for $C(S, t)$:

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} - rC = 0, \tag{1.1}$$

with terminal condition $C(S,T) = (S - K)^+$. Alternatively, one can appeal to martingale pricing theory, which guarantees that in the absence of arbitrage (appropriately defined – see for example [Delbaen & Schachermayer, 1994]), there exists a "risk-neutral" probability measure under which the discounted prices of all tradeable assets are martingales. We assume such conditions, and unless otherwise stated, our references to probabilities, distributions, and expectations will be with respect to such a pricing measure, not the statistical measure. In the constant-volatility case, changing from the statistical to the pricing measure yields

$$dS_t = rS_t dt + \sigma S_t dW_t.$$

So log S_T is normal with mean $(r - \sigma^2/2)(T - t)$ and variance $\sigma^2(T - t)$, and the Black-Scholes formula follows from $C = e^{-r(T-t)}\mathbb{E}(S_T - K)^+$.

11.2 Probabilistic Interpretation

In what sense is implied volatility an average expected volatility? Some econometric studies [Canina & Figlewski, 1993; Christensen & Prabhala, 1998] test whether or not implied volatility is an "unbiased" predictor of future volatility, but they have limited relevance to our question, because they address the empirics of a far narrower question in which "expected" future volatility is with respect to the *statistical* probability measure. Our focus, instead, is the theoretical question of whether there exist natural definitions of "average" and "expected" such that implied volatility can indeed be understood – provably – as an average expected volatility.

11.2.1.   Time-dependent volatility

In the case of time-dependent but nonrandom volatility, a simple formula exists for Black-Scholes implied volatility.

Suppose that

$$dS_t = rS_t dt + \sigma(t)S_t dW_t$$

where σ is a deterministic function. Define

$$\tilde{\sigma} := \left(\frac{1}{T} \int_0^T \sigma^2(u)du \right)^{1/2}.$$

Then one can show that $\log S_T$ is normal with mean $(r - \bar{\sigma}^2/2)T$ and variance $\bar{\sigma}^2 T$, from which it follows that

$$C = C^{BS}(\bar{\sigma}).$$

and hence

$$I = \bar{\sigma}.$$

Thus implied volatility is equal to the quadratic mean volatility from 0 to *T*.

11.2.2.   Time-and-spot-dependent Volatility

Now assume that

$$dS_t = rS_t dt + \sigma(S_t, t)S_t dW_t \qquad (2.1)$$

where σ is a deterministic function, usually called the *local volatility*. We will also treat local volatility as a function $\tilde{\sigma}$ of time-0 moneyness x, via the definition $\tilde{\sigma}(x, T) := \sigma(S_0 e^{x+rT}, T)$; but abusing notation, we will suppress the tildes.

11.2.2..1   Local volatility and implied local volatility.   Under local volatility dynamics, call prices satisfy (1.1), but with variable coefficients:

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2(S, t)S^2 \frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} - rC = 0, \qquad (2.2)$$

and also with terminal condition $C(S, T) = (S - K)^+$.

Dupire [Dupire, 1994] showed that instead of fixing *(K, T)* and obtaining the backward PDE for *C(S,t)*, one can fix *(S, t)* and obtain a forward PDE for *C(K, T)*. *A* derivation (also in [Bouchouev & Isakov, 1997]) proceeds as follows.

Differentiating (2.2) twice with respect to strike shows that $G := \partial^2 C/\partial K^2$ satisfies the same PDE, but with terminal data $\delta(S - K)$. Thus *G* is the Green's function of (2.2), and it is the transition density of *S*. By a standard result (in

[Friedman, 1964], for example), it follows that G as a function of the variables *(K, T)* satisfies the adjoint equation, which is the Fokker-Planck PDE

$$\frac{\partial G}{\partial T} - \frac{\partial^2}{\partial K^2}\left(\frac{1}{2}\sigma^2(K,T)K^2 G\right) + r\frac{\partial}{\partial K}(KG) + rG = 0.$$

Integrating twice with respect to K and applying the appropriate boundary conditions, one obtains the Dupire equation:

$$\frac{\partial C}{\partial T} - \frac{1}{2}K^2\sigma^2(K,T)\frac{\partial^2 C}{\partial K^2} + rK\frac{\partial C}{\partial K} = 0, \tag{2.3}$$

with initial condition $C(K,0) = (S_0 - K)^+$.

Given call prices at all strikes and maturities up to some horizon, define *implied local volatility* as

$$L(K,T) := \left(\frac{\frac{\partial C}{\partial T} + rK\frac{\partial C}{\partial K}}{\frac{1}{2}K^2\frac{\partial^2 C}{\partial K^2}}\right)^{1/2}. \tag{2.4}$$

According to (2.3), this is the local volatility function consistent with the given prices of options. Define *implied local variance* as L^2.

Following standard terminology, our use of the term *implied volatility* will, in the absence of other modifiers, refer to implied *Black-Scholes* volatility, not implied *local* volatility. The two concepts are related as follows: Substituting

$$C = C^{BS}(I(S_0 e^{x+rT}, T)) \tag{2.5}$$

into (2.4) yields

$$L^2(x,T) = \frac{2TI\frac{\partial I}{\partial T} + I^2}{\left(1 - x\frac{\partial I}{\partial x}/I\right)^2 + TI\frac{\partial^2 I}{\partial x^2} - \frac{1}{4}T^2 I^2\frac{\partial^2 I}{\partial x^2}}$$

See, for example, Andersen and Brotherton-Ratcliffe [Andersen & Brotherton-Ratcliffe, 1998]. Whereas the computation of I from market data poses no numerical difficulties, the recovery of L is an ill-posed problem that requires careful treatment; see also [Avellaneda et al, 1997; Bouchouev & Isakov, 1997; Coleman, Li & Verma, 1999; Gzyl & Villasana, 2003]. These issues will not concern us here, because our use of implied local volatility L will be strictly as a theoretical device to link local volatility results to stochastic volatility results, in section 11.2.3..1.

11.2.2..2   Short-dated implied volatility as harmonic mean local volatility.
In certain regimes, the representation of implied volatility as an average expected volatility can be made precise. Specifically, Berestycki, Busca,

and Florent ([Berestycki, Busca & Florent, 2002]; BBF henceforth) show that in the short-maturity limit, implied volatility is the *harmonic mean* of *local volatility.*

The PDE that relates implied volatility $I(x, T)$ to local volatility $\sigma(x, T)$ is, by substituting (2.5) into (2.3),

$$2TI\frac{\partial I}{\partial T} + I^2 - \sigma^2(x, T)\left(1 - x\frac{\partial I}{\partial x}\Big/I\right)^2$$
$$- \sigma^2(x, T)TI\frac{\partial^2 I}{\partial x^2} + \frac{1}{4}\sigma^2(x, T)T^2I^2\frac{\partial^2 I}{\partial x^2} = 0$$

Let $I_0(x)$ be the solution to the ODE generated by taking $T = 0$ in the PDE. Thus

$$I_0^2 - \sigma^2(x, 0)\left(1 - x\frac{\partial I_0}{\partial x}\Big/I_0\right)^2 = 0.$$

Elementary calculations show that the ODE is solved by

$$I_0(x) = \left(\int_0^1 \frac{ds}{\sigma(sx, 0)}\right)^{-1},$$

A natural conjecture is that the convergence $I_0 = \lim_{T\to 0} I(x, T)$ holds. Indeed this is what Berestycki, Busca, and Florent [Berestycki, Busca & Florent, 2002] prove. Therefore, short-dated implied volatility is approximately the harmonic mean of local volatility, where the mean is taken "spatially," along the line segment on $T = 0$, from moneyness 0 to moneyness x.

The *harmonic* mean here stands in contrast to arithmetic or quadratic means that have been proposed in the literature as rules of thumb. As BBF argue, probabilistic considerations rule out the arithmetic and quadratic means; for example, consider a local volatility diffusion in which there exists a price level $H \in (S_0, K)$ above which the local volatility vanishes, but below which it is positive. Then the option must have zero premium, hence zero implied volatility. This is inconsistent with taking a spatial mean of σ arithmetically or quadratically, but *is* consistent with taking a spatial mean of σ harmonically.

11.2.2..3   Deep in/out-of-the-money implied volatility as quadratic mean local volatility. BBF also show that if local volatility is uniformly continuous and bounded by constants so that

$$0 < \underline{\sigma} \leqslant \sigma(x, T) \leqslant \overline{\sigma},$$

and if local volatility has continuous limit(s)

$$\sigma_\pm(t) = \lim_{x\to\pm\infty} \sigma(x, t)$$

locally uniformly in t, then deep in/out-of-the-money implied volatility approximates the *quadratic* mean of local volatility, in the following sense:

$$\lim_{x\to\pm\infty} I(x,T) = \left(\frac{1}{T}\int_0^T \sigma_\pm^2(s)ds\right)^{1/2}.$$

The idea of the proof is as follows. Considering by symmetry only the $x\to\infty$ limit, let $I_\infty(T) := (\frac{1}{T}\int_0^T \sigma_+^2(s)ds)^{1/2}$. Note that I_∞ induces, via definition (2.4), a local variance L^2 that has the correct behavior at $x=\infty$, because the denominator is 1 while the numerator is $\sigma_+^2(T)$.

To turn this into a proof, BBF show that for any ε one can construct a function $\overline{\psi}(x)$ such that $1 < \overline{\psi}(\infty) < 1 + \varepsilon$ and such that $I_\infty(T)\overline{\psi}(x)$ induces via (2.4) a local volatility that dominates L. By a comparison result of BBF,

$$\limsup_{x\to\infty} I(x,T) < I_\infty(T)\overline{\psi}(\infty) < (1+\varepsilon)I_\infty(T).$$

On the other hand, one can construct $\underline{\psi}$ such that

$$\liminf_{x\to\infty} I(x,T) > I_\infty(T)\underline{\psi}(\infty) > (1-\varepsilon)I_\infty(T).$$

Taking ε to 0 yields the result.

11.2.3.   Stochastic volatility

Now suppose that
$$dS_t = rS_t dt + \sigma_t S_t dW_t,$$

where σ_t is stochastic. In contrast to local volatility models, σ_t is not determined by S_t and t.

Intuition from the case of time-dependent volatility does not apply directly to stochastic volatility. For example, one can define the random variable

$$\bar{\sigma} := \left(\frac{1}{T}\int_0^T \sigma_t^2 dt\right)^{1/2},$$

but note that in general

$$I \neq \mathbb{E}\bar{\sigma}.$$

For example, in the case where the σ process is independent of W, the mixing argument of Hull and White [Hull & White, 1987] shows that

$$\begin{aligned}C_0 &= \mathbb{E}e^{-rT}(S_T - K)^+ \\ &= \mathbb{E}(\mathbb{E}[e^{-rT}(S_T - K)^+|\{\sigma_t\}_{0\leqslant t\leqslant T}]) = \mathbb{E}C^{BS}(\bar{\sigma}).\end{aligned}$$

(2.6)

However, this is *not* equal to $C^{BS}(\mathbb{E}\bar{\sigma})$ because C^{BS} is not a linear function of its volatility argument. What we can say is that for the at-the-money-forward

strike, C^{BS} is nearly linear in σ, because its second σ derivative is negative but typically small; so by Jensen $I < \mathbb{E}\bar{\sigma}$, but equality nearly holds.

Note that this $I \approx \mathbb{E}\bar{\sigma}$ heuristic is specific to one particular strike, that it assumes independence of σ_t and W_t, and that the expectation is under a risk-neutral pricing measure, not the statistical measure. We caution against the improper application of this rule outside of its limited context.

So is there some time-averaged volatility interpretation of I, that *does* hold in contexts where $I \approx \mathbb{E}\bar{\sigma}$ fails?

11.2.3..1   Relation to local-volatility results.   Under stochastic volatility dynamics, implied local variance at *(K, T)* is the risk-neutral conditional expectation of σ_t^2, given $S_T = K$. The argument of Derman and Kani [Derman & Kani, 1998] is as follows. Let $f(S) = (S - K)^+$. Now take, formally, an Ito differential with respect to *T*:

$$d_T C = d_T[e^{-rT}\mathbb{E}(S_T - K)^+] = \mathbb{E}d_T[e^{-rT}(S_T - K)^+]$$

$$= e^{-rT}\mathbb{E}\left[f'(S_T)dS_T + \frac{1}{2}\sigma_T^2 S_T^2 \delta(S_T - K)dT - (S_T - K)^+ dT\right]$$

$$= e^{-rT}\mathbb{E}\left[rS_T H(S_T - K) + \frac{1}{2}\sigma_T^2 S_T^2 \delta(S_T - K) - (S_T - K)^+\right]dT$$

$$= e^{-rT}\mathbb{E}\left[-rKH(S_T - K) + \frac{1}{2}\sigma_T^2 S_T^2 \delta(S_T - K)\right]dT,$$

where H denotes the Heaviside function. Assuming that (S_T, σ_T^2) has a joint density p_{S_T, V_T}, let p_{S_T} denote the marginal density of S_T. Continuing, we have

$$\frac{\partial C}{\partial T} = -rK\frac{\partial C}{\partial K} + \frac{1}{2}e^{-rT}\iint vs^2\delta(s - K)p_{S_T, V_T}(s, v)dsdv$$

$$= -rK\frac{\partial C}{\partial K} + \frac{1}{2}e^{-rT}K^2\int vp_{S_T, V_T}(K, v)dv.$$

So, by definition of implied local variance,

$$L^2(K, T) = \frac{\frac{\partial C}{\partial T} + rK\frac{\partial C}{\partial K}}{\frac{1}{2}K^2\frac{\partial^2 C}{\partial K^2}} = \frac{\int vp_{S_T, V_T}(K, v)dv}{p_{S_T}(K)} = \mathbb{E}(\sigma_t^2 | S_T = K).$$

Consequently, any characterization of I as an average expected local volatility becomes tantamount to a characterization of I as an average conditional expectation of stochastic volatility.

APPLICATION 11.2.1 *The BBF results in sections 11.2.2..2 and 11.2.2..3 can be interpreted, under* stochastic *volatility, as expressions of implied volatility*

as *[harmonic or quadratic] average* conditional *expectations of future volatil-ity.*

11.2.3..2   The path-from-spot-to-strike approach.   The following rea-soning by Gatheral [Gatheral, 2001] provides an interpretation of implied volatil-ity as average expected stochastic volatility, without assuming short times to maturity or strikes deep in/out of the money.

Fix K and T. Let

$$\Gamma^{BS} := \frac{\partial^2 C^{BS}}{\partial S^2}$$

be the Black-Scholes gamma function.

Assume there exists a nonrandom nonnegative function $v(t)$ such that for all t in $(0,T)$,

$$v(t) = \frac{\mathbb{E}[\sigma_t^2 S_t^2 \Gamma^{BS}(S_t, t, \bar{\sigma}(t))]}{\mathbb{E}[S_t^2 \Gamma^{BS}(S_t, t, \bar{\sigma}(t))]} \tag{2.7}$$

where

$$\bar{\sigma}(t) := \left(\frac{1}{T-t}\int_t^T v(u)du\right)^{1/2}.$$

Note that σ_t need not be a deterministic function of spot and time.

Define the function

$$c(S, t) := C^{BS}(S, t, \bar{\sigma}(t)),$$

which solves the following PDE for $(S, t) \in (0, \infty) \times (0, T)$:

$$\frac{\partial c}{\partial t} = -\frac{1}{2}v(t)S^2 \frac{\partial^2 c}{\partial S^2} - rs\frac{\partial C}{\partial S} + rC. \tag{2.8}$$

We have

$$C(K, T) = \mathbb{E}[e^{-rT}(S_T - K)^+] = \mathbb{E}[e^{-rT}c(S_T, T)]$$

$$= c(S_0, 0) + e^{-rT}\mathbb{E}\left[\int_0^T \frac{\partial c}{\partial t}(S_t, t)dt + \frac{1}{2}\sigma_t^2 S_t^2 \frac{\partial^2 c}{\partial S^2}(S_t, t)dt\right.$$

$$\left. + \frac{\partial c}{\partial S}(S_t, t)dS_t - rc(S_t, t)dt\right]$$

$$= c(S_0, 0) + e^{-rT}\mathbb{E}\left[\int_0^T \frac{1}{2}(\sigma_t^2 - v(t))S_t^2 \frac{\partial^2 c}{\partial S^2}(S_t, t)dt\right]$$

$$= c(S_0, 0) = C^{bs}(S_0, 0, \bar{\sigma}(0)).$$

using Ito's rule, then (2.8), then (2.7). Therefore

$$I^2 = \bar{\sigma}^2(0) = \frac{1}{T}\int_0^T v(t)dt = \frac{1}{T}\int_0^T \mathbb{E}^{\mathbb{G}_t}\sigma_t^2 dt, \qquad (2.9)$$

where the final step re-interprets the definition (2.7) of $v(t)$ as the expectation of σ_t^2 with respect to the probability measure $\mathbb{G}_t$ defined, relative to the pricing measure $\mathbb{P}$, by the Radon-Nikodym derivative

$$\frac{d\mathbb{G}_t}{d\mathbb{P}} := \frac{S_t^2 \Gamma^{BS}(S_t, t, \bar{\sigma}(t))}{\mathbb{E}[S_t^2 \Gamma^{BS}(S_t, t, \bar{\sigma}(t))]}.$$

So (2.9) interprets implied volatility as an average expected variance. More-over, this expectation with respect to $\mathbb{G}_t$ can be visualized as follows. Write

$$\mathbb{E}^{\mathbb{G}_t}\sigma_t^2 = \int_0^\infty \mathbb{E}(\sigma_t^2 | S_t = s)\kappa_t(s)ds, \qquad (2.10)$$

where the nonrandom function κ_t is defined by

$$\kappa_t(s) := \frac{s^2 \Gamma^{BS}(s, t, \bar{\sigma}(t)) p_{S_t}(s)}{\int_0^\infty s^2 \Gamma^{BS}(s, t, \bar{\sigma}(t)) p_{S_t}(s)ds},$$

and p_{S_t} denotes the density of S_t.

Thus $\mathbb{E}(\sigma_t^2 | S_t = s)$ is integrated against a kernel $\kappa(s)$ which has the fol-lowing behavior. For $t \downarrow 0$, the κ approaches the Dirac function $\delta(s - S_0)$, because the p_{S_t} factor has that behavior, while the $s^2 \Gamma^{BS}$ factor approaches an ordinary function. For $t \uparrow T$, the κ approaches the Dirac function $\delta(s - K)$, because the $s^2 \Gamma^{BS}$ factor has that behavior, while the p_{S_t} factor approaches an ordinary function. At each time t intermediate between 0 and T, the kernel has a finite peak, which moves from S_0 to K, as t moves from 0 to T.

This leads to two observations. First, one has the conjectural approximation

$$\mathbb{E}^{\mathbb{G}_t}\sigma_t^2 \approx \mathbb{E}(\sigma_t^2 | S_t = s^*(t)),$$

where the non-random point $s^*(t)$ is the s that maximizes the kernel κ_t. By (2.10), therefore,

$$I^2 \approx \frac{1}{T}\int_0^T \mathbb{E}(\sigma_t^2 | S_t = s^*(t))dt.$$

Second, the kernel's concentration of "mass" initially (for $t = 0$) at S_0, and terminally (for $t = T$) at K resembles the marginal densities of the S diffusion, *pinned* by conditioning on $S_T = K$. This leads to Gatheral's observation that implied variance is, to a first approximation, the time integral of the expected instantaneous variance *along the most likely path from S_0 to K.* We leave

open the questions of how to make these observations more precise, and how to justify the original assumption.

APPLICATION 11.2.2 *Given an approximation for local volatility, such as in [Gatheral, 2001], one can usually compute explicitly an approximation for a spot-to-strike average, thus yielding an approximation to implied volatility.*

For example, given an approximation for local volatility linear in x, the spot-to-strike averaging argument can be used to justify a rule of thumb (as in [Derman, Kani & Zou, 1996]) that approximates implied volatility also linearly in x, but with one-half the slope of local volatility.

11.3   Statics

We examine here the implications of various assumptions on the shape of the implied volatility surface, beginning in section 11.3.1. with only minimal assumptions of no-arbitrage, and then specializing in 11.3.2. and 11.3.3. to the cases of local volatility and stochastic volatility diffusions. The term "statics" refers to the analysis of $I(x, T)$ or $I(K, T)$ for t *fixed.*

As reference points, let us review some of the empirical facts about the shape of the volatility surface; see, for example, [Rebonato, 1999] for further discussion. A plot of I is not constant with respect to K (or x). It can take the shape of a *smile,* in which $I(K)$ is greater for K away-from-the-money than it is for K near-the-money. The more typical pattern in post-1987 equity markets, however, is a *skew* (or skewed smile) in which at-the-money I slopes downward, and the smile is far more pronounced for small K than for large K. Empirically the smile or skew *flattens* as T increases. In particular, a popular rule-of-thumb (which we will revisit) states that skew slopes decay with maturity approximately as $1/\sqrt{T}$; indeed, when comparing skew slopes across different maturities, practitioners often define "moneyness" as $x/\sqrt{T}$ instead of x.

The theory of how I behaves under various model specifications has at least three applications. First, to the extent that a model generates a theoretical I shape that differs qualitatively from empirical facts, we have evidence of model misspecification. Second, given an observed volatility skew, analytical expressions approximating $I(x, T)$ in terms of model parameters can be useful in calibrating those parameters. Third, necessary conditions on I for the absence of arbitrage provide consistency checks that can help to reject unsound proposals for volatility skew parameterizations.

Part of the challenge for future research will be to extend this list of models and regimes for which we understand the behavior of implied volatility.

11.3.1.   Statics under absence of arbitrage

Assuming only the absence of arbitrage, one obtains bounds on the slope of the implied volatility surface, as well as a characterization of how fast I grows at extreme strikes.

11.3.1..1   Slope bounds.   Hodges [Hodges, 1996] gives bounds on implied volatility based on the nonnegativity of call spreads and put spreads. Specifically, if $K_1 < K_2$ then

$$C(K_1) \geqslant C(K_2) \qquad P(K_1) \leqslant P(K_2) \qquad (3.1)$$

Gatheral [Gatheral, 1999] improves this observation to

$$C(K_1) \geqslant C(K_2) \qquad \frac{P(K_1)}{K_1} \leqslant \frac{P(K_2)}{K_2}, \qquad (3.2)$$

which is evident from a comparison of the respective payoff functions. Assuming the differentiability of option prices in K,

$$\frac{\partial C}{\partial K} \leqslant 0 \qquad \frac{\partial}{\partial K}\left(\frac{P}{K}\right) \geqslant 0.$$

Substituting $C = C^{BS}(I)$ and $P = Schönbucher(I)$ and simplifying, we have

$$-\frac{N(-d_1)}{\sqrt{T}N'(d_1)} \leqslant \frac{\partial I}{\partial x} \leqslant \frac{N(d_2)}{\sqrt{T}N'(d_2)},$$

where the upper and lower bounds come from the call and put constraints, respectively.

Using (as in [Carr & Wu, 2002]), the *Mill's Ratio* $R(d) := (1-N(d))/N'(d)$ to simplify notation, we rewrite the inequality as

$$-\frac{R(d_1)}{\sqrt{T}} \leqslant \frac{\partial I}{\partial x} \leqslant \frac{R(-d_2)}{\sqrt{T}}$$

Note that proceeding from (3.1) without Gatheral's refinement (3.2) yields the significantly weaker lower bound $-R(d_2)/\sqrt{T}$.

Of particular interest is the behavior at-the-money, where $x = 0$. In the short-dated limit, as $T \to 0$, assume that $I(0, T)$ is bounded above. Then

$$d_{1,2}(x = 0) = \pm I(0, T)\sqrt{T}/2 \longrightarrow 0.$$

Since $R(0)$ is a positive constant, the at-the-money skew slope must have the short-dated behavior

$$\left|\frac{\partial I}{\partial x}(0, T)\right| = O\left(\frac{1}{\sqrt{T}}\right), \qquad T \to 0. \qquad (3.3)$$

In the long-dated limit, as $T \to \infty$, assume that $I(0, T)$ is bounded away from 0. Then

$$d_{1,2}(x = 0) = \pm I(0, T)\sqrt{T}/2 \longrightarrow \pm\infty.$$

Since $R(d) \sim d^{-1}$ as $d \to \infty$, the at-the-money skew slope must have the long-dated behavior

$$\left| \frac{\partial I}{\partial x}(0, T) \right| = O\left(\frac{1}{T} \right), \qquad T \to \infty. \tag{3.4}$$

REMARK 11.3.1 *According to* (3.4), *the rule of thumb that approximates the skew slope decay rate as* $T^{-1/2}$ *cannot maintain validity into long-dated expiries.*

11.3.1..2   The moment formula.   Lee [Lee, 2002] proves the moment formula for implied volatility at extreme strikes. Previous work, in Avellaneda and Zhu [Avellaneda & Zhu, 1998], had produced asymptotic calculations for one specific stochastic volatility model, but the moment formula is entirely general, and it uncovers the key role of finite moments.

At any given expiry *T,* the tails of the implied volatility skew can grow no faster than $x^{1/2}$. Specifically, in the right-hand tail, for $|x|$ sufficiently large, the Black-Scholes implied variance satisfies

$$I^2(x, T) \leqslant 2|x|/T \tag{3.5}$$

and a similar relationship holds in the left-hand tail.

For proof, write $I^* := (2|x|/T)^{1/2}$, and show that $C^{BS}(I) < C^{BS}(I^*)$ for large $|x|$. This holds because the left-hand side approaches 0 but the right-hand side approaches a positive limit as $x \to \infty$.

APPLICATION 11.3.2 *This bound has implications for choosing functional forms of splines to extrapolate volatility skews. Specifically, it advises against fitting the skew's tails with any function that grows more quickly than* $x^{1/2}$.

Moreover, the tails cannot grow more slowly *than* $x^{1/2}$, *unless* S_T *has finite moments of all orders. This further restricts the advisable choices for parameterizing a volatility skew. To prove this fact, note that it is a consequence of the moment formula, which we now describe.*

The smallest (infimal) coefficient that can replace the 2 in (3.5) depends, of course, on the distribution of S_T, but the form of the dependence is notably simple. This sharpest possible coefficient is entirely determined by $\tilde{p}$ in the right-hand tail, and $\tilde{q}$ in the left-hand tail, where the real numbers

$$\tilde{p} := \sup\{p : \mathbb{E}S_T^{1+p} < \infty\}$$
$$\tilde{q} := \sup\{q : \mathbb{E}S_T^{-q} < \infty\},$$

can be considered, by abuse of language, the "number" of finite moments in underlying distribution. The moment formula makes explicit these relationships.

Specifically, let us write I^2 as a variable coefficient times $|x|/T$, the ratio of absolute-log-moneyness to maturity. Consider the limsups of this coefficient as $x \to \pm\infty$:

$$\beta_R(T) := \limsup_{x \to \infty} \frac{I^2(x, T)}{|x|/T}$$

$$\beta_L(T) := \limsup_{x \to -\infty} \frac{I^2(x, T)}{|x|/T}.$$

One can think of β_R and β_L as the right-hand and absolute left-hand slopes of the linear "asymptotes" to implied variance.

The main theorem in [Lee, 2002] establishes that β_R and β_L both belong to the interval [0, 2], and that their values depend only on the moment counts $\tilde{p}$ and $\tilde{q}$, according to the *moment formula:*

$$\tilde{p} = \frac{1}{2\beta_R} + \frac{\beta_R}{8} - \frac{1}{2}$$

$$\tilde{q} = \frac{1}{2\beta_L} + \frac{\beta_L}{8} - \frac{1}{2}.$$

One can invert the moment formula, by solving for β_R and β_L:

$$\beta_R = 2 - 4(\sqrt{\tilde{p}^2 + \tilde{p}} - \tilde{p}),$$

$$\beta_L = 2 - 4(\sqrt{\tilde{q}^2 + \tilde{q}} - \tilde{q}).$$

The idea of the proof is as follows. By the Black-Scholes formula, the tail behavior of the implied volatility skew carries the same information as the tail behavior of option prices. In turn, the tail growth of option prices carries the same information as the number of finite moments – intuitively, option prices are bounded by moments, because a call or put payoff can be dominated by a power payoff; on the other hand, moments are bounded by option prices, because a power payoff can be dominated by a mixture, across a continuum of strikes, of call or put payoffs.

In a wide class of specifications for the dynamics of *S,* the moment counts $\tilde{p}$ and $\tilde{q}$ are readily computable functions of the model's parameters. This occurs whenever $\log S_T$ has a distribution whose characteristic function f is explicitly known. In such cases, one calculates $\mathbb{E}S_T^{p+1}$ simply by extending f analytically to a strip in $\mathbb{C}$ containing $-i(p + 1)$, and evaluating f there; if no such extension exists, then $\mathbb{E}S_T^{p+1} = \infty$. In particular, among affine jump-diffusions and Levy processes, one finds many instances of such models. See, for example, [Duffie, Pan & Singleton, 2000; Lee, 2001].

APPLICATION 11.3.3 *The moment formula may speed up the calibration of model parameters to observed skews. By observing the tail slopes of the volatility skew, and applying the moment formula, one obtains $\tilde{p}$ and $\tilde{q}$. Combined with analysis of the characteristic function, this produces two constraints on the model parameters, and in models such as the examples below, actually* determines *two of the model's parameters. We do not claim that the moment formula alone can replace a full optimization procedure, but it could facilitate the process by providing a highly accurate initial guess of the optimal parameters.*

EXAMPLE 11.3.4 *In the double-exponential jump-diffusion model of [Kou, 2002; Kou & Wang, 2001], the asset price follows a geometric Brownian motion between jumps, which occur at event times of a Poisson process. Up-jumps and down-jumps are exponentially distributed with the parameters η_1 and η_2 respectively, and hence the means $1/\eta_1$ and $1/\eta_2$ respectively. Using the characteristic function, one computes*

$$\tilde{q} = \eta_2 \qquad \tilde{p} = \eta_1 - 1. \tag{3.6}$$

Thus η_1 and η_2 can be inferred from $\tilde{p}$ and $\tilde{q}$, which in turn come from the slopes of the volatility skew, via the moment formula.

The intuition of (3.6) is as follows: the larger the expected size of an up-jump, the fatter the S_T distribution's right-hand tail, and the fewer the number of positive moments. Similar intuition holds for down-jumps. Note that the jump frequency *has no effect on the asymptotic slopes.*

EXAMPLE 11.3.5 *In the normal inverse gaussian model of Barndorff-Nielsen [Barndorff-Nielsen, 1998], returns have a distribution defined as follows: consider two dimensional Brownian motion with constant drift (δ, γ), and let α be the Euclidean magnitude of this drift. The NIG distribution is the distribution of the first coordinate of the Brownian motion at the stopping time when the second coordinate hits a specified constant barrier. Then one can calculate*

$$\tilde{q} = \alpha + \delta \qquad \tilde{p} = \alpha - \delta - 1, \tag{3.7}$$

which also has intuitive content: larger α implies earlier stopping, hence thinner tails and more moments (of both positive and negative order); larger δ fattens the right-hand tail and thins the left-hand tail, decreasing the number of positive moments and increasing the number of negative moments.

11.3.2. Statics under local volatility

Assume that the underlying follows a local volatility diffusion of the form (2.1). Writing $F := Se^{r(T-t)}$ for the forward price, suppose that local volatility can be expressed as a function h of F alone:

$$\sigma(S, t) = h(Se^{r(T-t)}).$$

Hagan and Woodward (in [Hagan & Woodward, 1999], and with Kumar and Lesniewski in [Hagan et al, 2002]), develop regular perturbation solutions to (2.2) in powers of $\varepsilon := h(K)$, assumed to be small. The resulting call price formula then yields the implied volatility approximation

$$I(K,T) \approx h(\bar{F}) + \frac{1}{24}h''(\bar{F})(F_0 - K)^2, \qquad (3.8)$$

where $\bar{F} := (F_0 + K)/2$ is the midpoint between forward and strike. The same sources also discuss alternative assumptions and more refined approximations.

REMARK 11.3.6 *The reasoning of section 11.2.3..2 suggests an interpretation of the leading term $h(\bar{F})$ in (3.8) as a midpoint approximation to the average local volatility along a path from $(F_0, 0)$ to (K, T).*

11.3.3.   Statics under stochastic volatility

Now assume that the underlying follows a stochastic volatility diffusion of the form

$$dS_t = rS_t dt + \sqrt{V_t}S_t dW_t$$
$$dV_t = \alpha(V_t)dt + \beta(V_t)dZ_t$$

where Brownian motions W and Z have correlation ρ. From here one obtains, typically via perturbation methods, approximations to the implied volatility skew I. Our coverage will emphasize those approximations which apply to entire *classes* of stochastic volatility models, not specific to one particular choice of α and β. We label each approximation according to the regime in which it prevails.

11.3.3..1   Zero correlation.   Renault and Touzi [Renault & Touzi, 1996] prove that in the case $\rho = 0$, implied volatility is a symmetric smile – symmetric in the sense that

$$I(x, T) = I(-x, T)$$

and a smile in the sense that I is increasing in x for $x > 0$.

Moreover, as shown in [Ball & Roma, 1994], the parabolic shape of I is apparent from Taylor approximations. Expanding the function $C^{bs}(v) := C^{BS}(\sqrt{v})$ about $v = \mathbb{E}\overline{V}$, we have

$$C = C^{bs}(I) \approx C^{bs}(\mathbb{E}\overline{V}) + (I^2 - \mathbb{E}\overline{V})\frac{\partial C^{bs}}{\partial V}.$$

Comparing this to a Taylor expansion of the mixing formula

$$C = \mathbb{E}C^{bs}(\overline{V}) \approx C^{bs}(\mathbb{E}\overline{V}) + \frac{1}{2}\text{Var}(\overline{V})\frac{\partial^2 C^{bs}}{\partial V^2}$$

yields the approximation

$$I^2 \approx \mathbb{E}\overline{V} + \frac{1}{4}\frac{\text{Var}(\overline{V})}{(\mathbb{E}\overline{V})^2}\left(\frac{x^2}{T} - \mathbb{E}\overline{V} - \frac{1}{4}(\mathbb{E}\overline{V})^2 T\right),$$

which is quadratic in x, with minimum at $x = 0$.

REMARK 11.3.7 *To the extent that implied volatility skews are empirically not symmetric in equity markets, stochastic volatility models with zero correlation will not be consistent with market data.*

11.3.3..2 **Small volatility of volatility, and the short-dated limit.** Lewis [Lewis, 2000] shows that the forward call price, viewed as a function of x, has a complex Fourier transform given by $\hat{H}(k, V, T)/(k^2 - ik)$, where k is the transform variable and $\hat{H}$ solves the PDE

$$\frac{\partial \hat{H}}{\partial T} = \frac{1}{2}b^2 \frac{\partial^2 \hat{H}}{\partial V^2} + (a - ik\rho b V^{1/2})\frac{\partial \hat{H}}{\partial V} - \frac{k^2 - ik}{2}V\hat{H},$$

with initial condition $\hat{H}(k, V, 0) = 1$. In our setting, $\hat{H}$ can be viewed as the characteristic function of the negative of the log-return on the forward price of S.

Assuming that $b(V) = \eta B(V)$ for some constant parameter η, one finds a perturbation solution for $\hat{H}$ in powers of η. The transform can be inverted to produce a call price, by a formula such as

$$C = S - \frac{Ke^{-rT}}{2\pi}\int_{i/2-\infty}^{i/2+\infty} e^{ikx}\frac{\hat{H}(k, V, T)}{k^2 - ik}dk, \qquad (3.9)$$

yielding a series for C in powers of η. From the C series and the Black-Schole formula, Lewis derives the implied variance expansion

$$I^2 = \mathbb{E}\overline{V} + \eta\frac{J^{(1)}}{T}\left(\frac{x}{T\mathbb{E}\overline{V}} + \frac{1}{2}\right)$$
$$+ \eta^2\left[\frac{J^{(2)}}{T} + \frac{J^{(3)}}{T}\left(\frac{x^2}{2(\mathbb{E}\overline{V})^2 T^2} - \frac{1}{2T\mathbb{E}\overline{V}} - \frac{1}{8}\right)\right.$$
$$+ \frac{J^{(4)}}{T}\left(\frac{x^2}{T^2(\mathbb{E}\overline{V})^2} + \frac{x}{T\mathbb{E}\overline{V}} - \frac{4 - T\mathbb{E}\overline{V}}{4T\mathbb{E}\overline{V}}\right)$$
$$\left. + \frac{(J^{(1)})^2}{2T}\left(-\frac{5x^2}{2T^3(\mathbb{E}\overline{V})^3} - \frac{x}{T\mathbb{E}\overline{V}} + \frac{12 + T\mathbb{E}\overline{V}}{8T^2(\mathbb{E}\overline{V})^2}\right)\right] + O(\eta^3),$$

where $J^{()}$ are integrals of known functions.

EXAMPLE 11.3.8 *The short-time-to-expiry limit is*

$$I^2(x, 0) = V_0 + \frac{1}{2}\frac{\rho b}{\sqrt{V_0}}x + \left[\left(\frac{1}{12} - \frac{11}{48}\rho^2\right)\frac{b^2}{V_0^2} + \frac{1}{6}\frac{\rho b}{V_0}\frac{\partial(\rho b)}{\partial V}\right]x^2 + O(\eta^3).$$

(3.10)

The leading terms agree to $O(\eta)$ with the slow-mean-reversion result of section 11.3.3..5. We defer further commentary until there.

EXAMPLE 11.3.9 *In the case where*

$$dV_t = \kappa(\theta - V_t)dt + \eta V_t^\varphi dW_t,$$

(3.11)

we have

$$\mathbb{E}\overline{V} = \theta + \frac{1 - e^{-\kappa T}}{\kappa T}(V_0 - \theta)$$

and $J^{(2)} = 0$, while

$$J^{(1)} = \frac{\rho}{\kappa}\int_0^T (1 - e^{-\kappa(T-s)})\left[\theta + e^{-\kappa s}(V_0 - \theta)\right]^{\varphi + 1/2}ds$$

$$J^{(3)} = \frac{1}{2\kappa^2}\int_0^T (1 - e^{-\kappa(T-s)})^2\left[\theta + e^{-\kappa s}(V_0 - \theta)\right]^{2\varphi}ds$$

$$J^{(4)} = \left(\varphi + \frac{1}{2}\right)\frac{\rho^2}{\kappa}\int_0^T \left[\theta + e^{-\kappa(T-s)}(V_0 - \theta)\right]^{\varphi + 1/2}J^{(6)}(T, s)ds$$

$$J^{(6)} = \int_0^s (e^{-\kappa(s-u)} - e^{-\kappa s})\left[\theta + e^{-\kappa(T-u)}(V_0 - \theta)\right]^{\varphi - 1/2}du.$$

In particular, taking $\varphi = 1/2$ produces the Heston [Heston, 1993] square-root model. In the special case where $V_0 = \theta$, the slope of the implied variance skew is, to leading order in η,

$$\frac{\partial I^2}{\partial x} = \frac{\rho\eta}{\kappa T}\left(1 - \frac{1 - e^{-\kappa T}}{\kappa T}\right),$$

which agrees with a computation, by Gatheral [Gatheral, 2001], that uses the expectations interpretation of local volatility.

11.3.3..3   The long-dated limit.   Given a stochastic volatility model with a known transform $\hat{H}$, Lewis solves for $\lambda(k)$ and $u(k, T)$ such that $\hat{H}$ separates multiplicatively, for large T, into T-dependent and V-dependent factors:

$$\hat{H}(k, V, T) \approx e^{-\lambda(k)T}u(k, V), \qquad T \to \infty.$$

Suppose that $\lambda(k)$ has a saddle point at $k_0 \in \mathbb{C}$ where $\lambda'(k_0) = 0$. Applying classical saddle-point methods to (3.9) yields

$$C(S, V, T) \approx S - Ke^{-rT}\frac{u(k_0, V)}{k_0^2 - ik_0}\frac{\exp[-\lambda(k_0)T + ik_0 x]}{\sqrt{2\pi\lambda''(k_0)T}}.$$

By comparing this to the corresponding approximation of $C^{BS}(I)$, Lewis obtains the implied variance approximation

$$I^2(x) \approx 8\lambda(k_0) + (8\mathrm{Im}(k_0) - 4)\frac{x}{T} - \frac{x^2}{2\lambda(k_0)T^2} + O(T^{-3}), \qquad T \to \infty.$$

The fact that $I(x,T)$ is linear to first order in x/T agrees with the fast-mean-reversion result of Fouque, Papanicolaou, and Sircar [Fouque, Papanicolaou & Sircar, 2000]. We defer further commentary until section 11.3.3..4.

EXAMPLE 11.3.10 *In the case* (3.11) *with* $\varphi = 1/2$ *(the square-root model), Lewis finds*

$$k_0 = \frac{i}{1-\rho^2}\left[\frac{1}{2} - \frac{\rho}{\eta}\left(\kappa - \frac{1}{2}\sqrt{4\kappa^2 + \eta^2 - 4\rho\kappa\eta}\right)\right]$$

$$\lambda(k_0) = \frac{\kappa\eta}{2(1-\rho^2)\eta^2}\left[\sqrt{(2\kappa - \rho\eta)^2 + (1-\rho^2)\eta^2} - (2\kappa - \rho\eta)\right].$$

The sign of the leading-order at-the-money skew slope $(8Im(k_0)-4)/T$ *agrees with the sign of the correlation* ρ.

11.3.3..4 **Fast mean reversion.** Fouque-Papanicolaou-Sircar ([Fouque, Papanicolaou & Sircar, 2000]; FPS henceforth) model stochastic volatility as a function f of a state variable Y_t that follows a rapidly mean-reverting diffusion process. In the case of Ornstein-Uhlenbeck Y, this means that for some large α,

$$dS_t = \mu_t S_t dt + f(Y_t)S_t d\tilde{W}_t$$
$$dY_t = \alpha(\theta - Y_t)dt + \beta d\tilde{Z}_t$$

under the statistical measure, where the Brownian motions $\tilde{W}$ and $\tilde{Z}$ have correlation ρ.

Rewriting this under a pricing measure,

$$dS_t = rS_t dt + f(Y_t)S_t dW_t$$
$$dY_t = [\alpha(\theta - Y_t) - \beta\Lambda(Y_t)]dt + \beta dZ_t,$$

where the volatility risk premium Λ is assumed to depend only on Y. Let p_Y denote the invariant density (under the statistical probability measure) of Y, which is normal with mean θ and variance $\beta^2/(2\alpha)$. Let angle brackets denote average with respect to that density. Write

$$\bar{\sigma}_\infty^2 := \langle f^2 \rangle,$$

so that $\bar{\sigma}_\infty$ is the quadratic average of volatility with respect to the invariant distribution.

By a singular perturbation analysis of the PDE for call price, FPS show that implied volatility has an expansion with leading terms

$$I(x,T) = A\frac{x}{T} + B + O(1/\alpha),$$

where

$$A := -\frac{V_3}{\bar{\sigma}_\infty^3}$$

$$B := \bar{\sigma}_\infty + \frac{3V_3/2 - V_2}{\bar{\sigma}_\infty},$$

(3.12)

and

$$V_2 := \frac{\beta}{2\alpha}\langle(2\rho f - \Lambda)\phi\rangle$$

$$V_3 := \frac{\beta}{2\alpha}\langle\rho f\phi\rangle$$

$$\phi(y) := \frac{2\alpha}{\beta^2 p_Y(y)}\int_{-\infty}^{y}(f^2(z) - \langle f^2\rangle)p_Y(z)dz.$$

REMARK 11.3.11 *The fast-mean-reversion approximation is particularly suited for pricing long-dated options; in that long time horizon, volatility has time to undergo much activity, so relative to the time scale of the option's lifetime, volatility can indeed be considered to mean-revert rapidly.*

Note that $I(x,T)$ is, to first order, linear in x/T. This functional form agrees with Lewis's long-dated *skew approximation* (11.3.3..3).

REMARK 11.3.12 *Today's volatility plays no role in the leading-order coefficients A and B. Instead, the dominant effects depend only on ergodic means. Intuitively, the assumption of large mean-reversion rapidly erodes the influence of today's volatility, leaving the long-run averages to determine A and B.*

REMARK 11.3.13 *The slope of the long-dated implied volatility skew satisfies*

$$\left|\frac{\partial I}{\partial x}(0,T)\right| \sim \frac{1}{T} \qquad T \to \infty.$$

As a consistency check, note that the long-dated asymptotics are consistent with the no-arbitrage constraint (3.4). Specifically, the $T \to \infty$ skew slope decay of these stochastic volatility models achieves the $O(T^{-1})$ bound.

APPLICATION 11.3.14 *FPS give approximations to prices of certain path-dependent derivatives under fast-mean-reverting stochastic volatility. Typically, such approximations involve the Black-Scholes price for that derivative, corrected by some term that depends on V_2 and V_3.*

To evaluate this correction term, note that the formulas (3.12) *can be solved for V_2 and V_3 in terms of A, B, and $\bar{\sigma}$. FPS calibrate A and B to the implied volatility skew, and estimate $\bar{\sigma}$ from historical data, producing estimates of V_2 and V_3, which become the basis for an approximation of the derivative price.*

For example, in the case of uncorrelated volatility where $\rho = 0$, FPS find that the price of an American put is approximated by the Black Scholes American put price, evaluated at the volatility parameter

$$\sqrt{\bar{\sigma}^2 - 2V_2},$$

which can be considered an "effective volatility."

11.3.3..5 **Slow mean reversion.** Assuming that for a constant parameter ε,

$$d\sigma_t = \varepsilon\alpha(V_t)dt + \sqrt{\varepsilon}\beta(V_t)dW_t,$$

Sircar and Papanicolaou [Sircar & Papanicolaou, 1999] develop, and Lee [Lee, 2001a] extends, a regular perturbation analysis of the PDE

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + \sqrt{\varepsilon}\rho S\sigma\beta\frac{\partial^2 C}{\partial S\partial\sigma} + \frac{1}{2}\varepsilon\beta^2\frac{\partial^2 C}{\partial\sigma^2} + \varepsilon\alpha\frac{\partial C}{\partial\sigma} + rS\frac{\partial C}{\partial S} = rC$$

satisfied by the call price under stochastic volatility. This leads to an expansion for C in powers of ε, which in turn leads to the implied volatility expansion

$$I \approx \sigma_0 + \sqrt{\varepsilon}\left[\frac{\rho\beta}{2\sigma_0}x + \frac{\rho\sigma_0\beta}{4}T\right]$$

$$+ \varepsilon\left[\left(\left(\frac{\beta\beta'}{6\sigma^2} - \frac{5\beta^2}{12\sigma^3}\right)\rho^2 + \frac{\beta^2}{6\sigma^3}\right)x^2 + \left(\left(\frac{\sigma\beta^2}{12} + \frac{\sigma^2\beta\beta'}{24}\right)\rho^2 - \frac{\sigma\beta^2}{24}\right)T^2\right.$$

$$\left. - \left(\frac{\beta^2}{24\sigma} - \frac{\beta\beta'}{6}\right)\rho^2 Tx + \left(\left(\frac{\beta^2}{24\sigma} - \frac{\beta\beta'}{6}\right)\rho^2 + \frac{\alpha}{2} - \frac{\beta^2}{12\sigma}\right)T\right],$$

where $\beta' := \partial\beta/\partial\sigma$. In particular, short-dated implied volatility satisfies

$$I(x,0) \approx \sigma_0 + \sqrt{\varepsilon}\frac{\rho\beta}{2\sigma_0}x. \qquad (3.13)$$

REMARK 11.3.15 *The slow-mean-reversion approximation is particularly suited for pricing short-dated options; in that short time horizon, volatility has little time in which to vary, so relative to the time scale of the option's lifetime, volatility can indeed be considered to mean-revert slowly.*

Note that (3.13) *agrees precisely with the leading terms of Lewis's* short-dated *skew approximation* (3.10).

REMARK 11.3.16 *In contrast to the case of rapid mean-reversion, the level to which volatility reverts here plays no role in the leading-order coefficients.*

With a small rate of mean-reversion, today's volatility will have the dominant effect.

REMARK 11.3.17 *For $\rho \neq 0$, the at-the-money skew exhibits a slope whose sign agrees with ρ. For $\rho = 0$ the skew has a parabolic shape.*

REMARK 11.3.18 *In agreement with a result of Ledoit, Santa-Clara, and Yan [Ledoit, Santa-Clara & Yan, 2001], we have $I(x, T) \to \sigma_0$ as $(x, T) \to (0, 0)$.*

APPLICATION 11.3.19 *In principle, given a parametric form for b, the fact that the short-dated skew has slope ρb gives information that can simplify parameter calibration. For example, if the modelling assumption is that $b = \beta f(V)$ for some constant parameter β and known function f, then directly from the short-dated skew and its slope, one obtains the product of the parameters ρ and β.*

APPLICATION 11.3.20 *Lewis observes, moreover, that this tool facilitates the* inference *of the functional form of b. Specifically, observe time-series of the short-dated at-the-money data pair: (implied volatility, skew slope). As implied volatility ranges over its support, the functional form of b is, in principle, revealed.*

REMARK 11.3.21 *Note that the $T \to 0$ skew slope is $O(1)$, which is strictly smaller than the $O(T^{-1/2})$ constraint. To the extent that the short-dated volatility skew slope empirically seems to* attain *the $O(T^{-1/2})$ upper bound instead of the $O(1)$ diffusion behavior, this observed skew will not be easily captured by standard diffusion models. Two approaches to this problem, and subjects for further research, are to remain in the stochastic-volatility diffusion framework but introduce time-varying coefficients (as in [Fouque, Papanicolaou, Sircar & Solna, 2002]); or alternatively to go outside the diffusion framework entirely and introduce jump dynamics, such as in [Carr & Wu, 2002].*

11.4 Dynamics

While traditional diffusion models specify the dynamics of the spot price and its instantaneous volatility, a newer class of models seeks to specify directly the dynamics of one or more implied volatilities. One reason to take I as primitive is that it enjoys wide acceptance as a descriptor of the state of an options market. A second reason is that the observability of I makes calibration trivial.

In this section, today's date t is not fixed at 0, because we are now concerned with the time evolution of I.

11.4.1.   No-arbitrage approach

11.4.1..1   **One implied volatility.**   Consider the time-evolution of a single implied volatility I at some fixed strike K and maturity date T. Schönbucher [Schönbucher, 1998] models directly its dynamics as

$$dI_t = u_t dt + \gamma_t dW_t^{(0)} + v_t dW_t,$$

where W and $W^{(0)}$ are independent Brownian motions. The spot price has dynamics

$$dS_t = rS_t dt + \sigma_t S_t dW_t^{(0)},$$

where σ_t is yet to be specified.

Since the discounted call price $e^{-r(T-t)} C^{BS}(t, S_t, I_t)$ must be a martingale under the pricing measure, we have for all $I > 0$ the following drift restriction on the call price:

$$\frac{\partial C^{BS}}{\partial t} + rS\frac{\partial C^{BS}}{\partial S} + u\frac{\partial C^{BS}}{\partial I} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C^{BS}}{\partial S^2}$$
$$+ \gamma\sigma S\frac{\partial^2 C^{BS}}{\partial I \partial S} + \frac{1}{2}v^2 \frac{\partial^2 C^{BS}}{\partial I^2} = rC^{BS}.$$

This reduces to a joint restriction on the diffusion coefficients of I, the drift of I, and the instantaneous volatility σ:

$$Iu = \frac{I^2 - \sigma^2}{2(T-t)} - \frac{1}{2}d_1 d_2 v^2 + \frac{d_2}{\sqrt{T-t}}\sigma\gamma. \tag{4.1}$$

Since S, t, and T are observable, we have that the volatility of I, together with the drift of I, determines the spot volatility. Other papers [Brace, Goldys, Klebaner & Womersley, 2000; Ledoit, Santa-Clara & Yan, 2001] have arrived at analogous results in which one fixes not (strike, expiry), but instead some other specification of exactly which implied volatility is to be modelled, such as (moneyness, time to maturity).

Schönbucher imposes a further constraint to ensure that I does not blow up as $t \to T$. He requires that

$$(I^2 - \sigma^2) - d_1 d_2 (T-t)v^2 + 2d_2\sqrt{T-t}\sigma\gamma = O(T-t) \qquad t \to T, \tag{4.2}$$

which simplifies to

$$I^2\sigma^2 + 2\gamma x I\sigma - I^4 + x^2 v^2 = 0.$$

This can be solved to get expiration-date implied volatility in terms of expiration-date spot volatility. The solution is particularly simple in the zero-correlation case, where $\gamma = 0$. Then, suppressing subscripts T,

$$I^2 = \frac{1}{2}\sigma^2 + \sqrt{\frac{\sigma^2}{4} + x^2 v^2}.$$

Under condition (4.2), therefore, implied volatility behaves as $\sigma + O(x^2)$ for x small, but $O(|x|^{1/2})$ for x large. Both limits are consistent with the statics of sections 11.3.1..2 and 11.3.3..1.

APPLICATION 11.4.1 *Schönbucher applies this model to the pricing of other derivatives as follows. Subject to condition (4.2), the modeller specifies the drift and volatility of I, and infers the dependence of instantaneous volatility σ on the state variables (S, t, I) according to (4.1). Then the price $C(S, t, I)$ of a non-strongly-path-dependent derivative satisfies the usual two-factor pricing equation*

$$\frac{\partial C}{\partial t} + rS\frac{\partial C}{\partial S} + u\frac{\partial C}{\partial I} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C}{\partial S^2} + \gamma\sigma S\frac{\partial^2 C}{\partial S\partial I} + \frac{1}{2}v^2\frac{\partial^2 C}{\partial I^2} = rC$$

with boundary conditions depending on the particular contract. Finite differ-ence methods can solve such a PDE.

    Care should be taken to ensure that I does not become negative.

11.4.1..2   Term structure of implied volatility. Schönbucher extends this model M different maturities. The implied volatilities to be modelled are $I_t(K_m, T_m)$ for $m = 1, \ldots, M$, where $T_1 < T_2 < \cdots < T_M$. Let

$$V_t^{(m)} := I_t^2(K_m, T_m)$$

be the implied variance. One specifies the dynamics for the shortest-dated variance $V^{(1)}$, as well as all "forward" variances

$$V^{(m,m+1)} := \frac{(T_{m+1} - t)V^{(m+1)} - (T_m - t)V^{(m)}}{T_{m+1} - T_m}.$$

The spot volatility σ_t and the drift and diffusion coefficients of $V_t^{(1)}$ are jointly subject to the drift restriction (4.1) and the no-explosion condition (4.2). Then, given the σ_t and $V_t^{(1)}$ dynamics, specifying each $V^{(m,m+1)}$ diffusion coeffi-cient determines the corresponding drift coefficient, by applying (4.1) to $V^{(m+1)}$.

APPLICATION 11.4.2 *To price exotic contracts under these multi-factor dy-namics, Schönbucher recommends Monte Carlo simulation of the spot price (which depends on simulation of implied volatilities). Upon expiry of the T_1 option, the T_2 option becomes the "front" contract; at that time $V^{(2)}$ coincides with $V^{(1,2)}$, and at later times its evolution is linked to spot volatility via the drift and the no-explosion conditions. Similar transitions occur at each later expiry.*

    Care should be taken to avoid negative forward variances.

11.4.2.   Statistical approach

Direct modelling of arbitrage-free evolution of an entire implied volatility *surface* remains largely unresolved. Unlike traditional models of spot dynamics, direct implied volatility models face increasing difficulty in enforcing no-arbitrage conditions, when multiple strikes are introduced at a maturity.

Instead of demanding no-arbitrage, the modeller may have a goal more statistical in nature, namely to describe the empirical movements of the implied volatility surface. According to Cont and da Fonseca's [Cont & da Fonseca, 2002] analysis of SP500 and FTSE data, the empirical features of implied volatility include the following:

Three principal components explain most of the daily variations in implied volatility: one eigenmode reflecting an overall (parallel) shift in the level, another eigenmode reflecting opposite movements (skew) in low and high strike volatilties, and a third eigenmode reflecting convexity changes. Variations of implied volatility along each principal component are autocorrelated, mean-reverting, and correlated with the underlying.

To quantify these features, Cont and da Fonseca introduce and estimate a d-factor model of the volatility surface, viewed as a function of moneyness m and time-to-maturity τ. The following model is specified under the statistical probability measure:

$$\log I_t(m, \tau) = \log I_0(m, \tau) + \sum_{k=1}^{d} y_t^{(k)} f^{(k)}(m, \tau),$$

where the eigenmodes $f^{(k)}$, such as the three described above, can be estimated by principal component analysis; the coefficients $y^{(k)}$ are specified as mean-reverting Ornstein-Uhlenbeck processes

$$dy_t^{(k)} = -\lambda^{(k)}(y_t^{(k)} - \bar{y}^{(k)})dt + v^{(k)}dW_t^{(k)}.$$

REMARK 11.4.3 *If one takes $y_t^{(k)} = 0$ for all k, then $I(m, \tau)$ does not vary in time. This corresponds to an ad-hoc model known to practitioners as "sticky delta." Balland [Balland, 2002] proves that if the dynamics of S are consistent with such a model (or even a generalized sticky delta model in which $I_t(m, \tau)$ is time-varying but determinstic), then assuming no arbitrage, S must be the exponential of a process with independent increments.*

APPLICATION 11.4.4 *A natural application is the Monte Carlo simulation of implied volatility, for the purpose of risk management.*

However, this model, unlike the theory of section 11.4.1., is not intended to determine the consistent volatility drifts needed for martingale pricing of exotic derivatives. How best to introduce the ideas from this model into a no-arbitrage theory remains an open question.

Acknowledgments

References

Marco Avellaneda and Yingzi Zhu, A Risk-Neutral Stochastic Volatility Model, *International Journal of Theoretical and Applied Finance,* **1** 2, 289–310,1998

Patrick Hagan and Deep Kumar and Andrew Lesniewski and Diana Woodward, Managing Smile Risk, Preprint, 2002.

Patrick Hagan and Diana Woodward, Equivalent Black Volatilities, *Applied Mathematical Finance,* **6** 3, 147–157, 1999.

Leif B. G. Andersen and Rupert Brotherton-Ratcliffe, The Equity Option Volatility Smile: an Implicit Finite-Difference Approach, *Journal of Computational Finance,* **1** 2, 5–37,1998.

Peter Carr and Liuren Wu, The Finite Moment Logstable Process and Option Pricing, *Journal of Finance,* Forthcoming, 2002.

Bent Christensen and Nagpurnanand Prabhala, The Relation between Implied and Realized Volatility, *Journal of Financial Economics,* **50** 2, 125–150, 1998.

Linda Canina and Stephen Figlewski, The Informational Content of Implied Volatility,*Review of Financial Studies,* **6** 2, 659–681, 1993.

Philipp L. Schönbucher, A Market Model for Stochastic Implied Volatility, Bonn University, 1998.

Roger W. Lee, The Moment Formula for Implied Volatility at Extreme Strikes, Stanford University and Courant Institute, 2002.

Jim Gatheral, The Volatility Skew: Arbitrage Constraints and Asymptotic Behaviour, Merrill Lynch, 1999.

Jim Gatheral, Lecture 2: Fitting the Volatility Skew, Case Studies in Financial Modelling course notes, Courant Institute, 2001.

Hardy M. Hodges, Arbitrage Bounds on the Implied Volatility Strike and Term Structures of European-Style Options, *Journal of Derivatives,* 23–35, 1996.

Roger W. Lee,Option Pricing by Transform Methods: Extensions, Unification, and Error Control, Stanford University and Courant Institute, 2001.

Roger W. Lee, Implied and Local Volatilities under Stochastic Volatility, *International Journal of Theoretical and Applied Finance,* **4** 1,45–89, 200la.

Gurdip Bakshi and Charles Cao and Zhiwu Chen, Empirical Performance of Alternative Option Pricing Models, *Journal of Finance,* **52**, 2003–2049, 1997.

Riccardo Rebonato, *Volatility and Correlation in the Pricing of Equity, FX and Interest Rate Options,* John Wiley & Sons, 1999.

Alan L. Lewis, *Option Valuation under Stochastic Volatility,* Finance Press, 2000.

Jean-Pierre Fouque, George Papanicolaou and K. Ronnie Sircar, *Derivatives in Financial Markets with Stochastic Volatility,* Cambridge University Press, 2000.

Steven L. Heston, A Closed-Form Solution for Options with Stochastic Volatility and Applications to Bond and Currency Options, *Review of Financial Studies,* **6** 2, 327–343, 1993.

Jean-Pierre Fouque and George Papanicolaou and Ronnie Sircar and Knut Solna, Maturity Cycles in Implied Volatility, Preprint, 2002.

Steven G. Kou, A Jump Diffusion Model for Option Pricing, Preprint, 2002.

Steven G. Kou and Hui Wang, Option Pricing under a Double Exponential Jump Diffusion Model, Preprint, 2001.

Freddy Delbaen and Walter Schachermayer, A General Version of the Fundamental Theorem of Asset Pricing, *Mathematische Annalen,* **2** 4, 61–73, 1994.

Henri Berestycki, Jérôme Busca and Igor Florent, Asymptotics and Calibration of Local Volatility Models, *Quantitative Finance,* **2** 2, 61–69, 2002.

Rama Cont and Jose da Fonseca, Dynamics of implied volatility surfaces, *Quantitative Finance,* **2** 2, 45–60, 2002.

Philippe Balland, Deterministic Implied Volatility Models, *Quantitative Finance,* **2** 1, 31–44, 2002.

Ole E. Barndorff-Nielsen, Processes of Normal Inverse Type,*Finance and Stochastics,* **2** 2, 1998.

Olivier Ledoit, Pedro Santa-Clara and Shu Yan, Relative Pricing of Options with Stochastic Volatility, Preprint, 2001.

Thomas F. Coleman, Yuying Li and Arun Verma, Reconstructing the Unknown Local Volatility Function, *Journal of Computational Finance,* **2** 3, 77-102, 1999.

Henryk Gzyl and Minaya Villasana, A Perturbative Approach for recostructing Diffusion Coefficients, Preprint, To appear in *Applied Mathematics and Computation,* 2003.

Fisher Black and Myron Scholes, The Pricing of Options and Corporate Liabilities, *Journal of Political Economy,* **81,** 637–659, 1973.

Bruno Dupire, Pricing with a Smile, *RISK,* **7**, 18–20, 1994.

Ilia Bouchouev and Victor Isakov, The Inverse Problem of Option Pricing, *Inverse Problems,* **13** 5, L11–L17, 1997.

Avner Friedman, *Partial Differential Equations of Parabolic Type,* Prentice-Hall, 1964.

Marco Avellaneda, Craig Friedman, Richard Holmes and Dominick Samperi, Calibrating Volatility Surfaces via Relative-Entropy Minimization, *Applied Mathematical Finance,* **4** 1, 37–64, 1997.

John Hull and Alan White, The Pricing of Options on Assets with Stochastic Volatilities, *Journal of Finance,* **42** 2, 281–300, 1987.

Emanuel Derman and Iraj Kani, Stochastic Implied Trees: Arbitrage Pricing with Stochastic Term and Strike Structure of Volatility, *International Journal of Theoretical and Applied Finance,* **1** 1, 61–110, 1998.

Darrell Duffie, Jun Pan and Kenneth Singleton, Transform Analysis and Option Pricing for Affine Jump-Diffusions, *Econometrica,* **68** 6, 1343–1376, 2000.

Alan Brace, Ben Goldys, Fima Klebaner and Bob Womersley, Market Model of Stochastic Implied Volatility with Application to the BGM Model, Working Paper S01-1, Department of Statistics, University of New South Wales, 2000.

Emanuel Derman, Iraj Kani and Joseph Z. Zou, The Local Volatility Surface: Unlocking the Information in Index Option Prices, *Financial Analysts Journal,* 25–36, 1996.

Eric Renault and Nizar Touzi, Option Hedging and Implied Volatilities in a Stochastic Volatility Model, *Mathematical Finance,* **6** 3, 279–302, 1996.

Clifford A. Ball and Antonio Roma, Stochastic Volatility Option Pricing, *Journal of Financial and Quantitative Analysis,* **29** 4, 589–607 1994.

K. Ronnie Sircar and George Papanicolaou, Stochastic Volatility, Smile & Asymptotics, *Applied Mathematical Finance,* **6** 2, 107–145, 1999.

ON THE INCREMENTS OF THE BROWNIAN SHEET

José R. León
U.C.V. Facultad de Ciencias. Departamento de Matemáticas.
Apartado Postal 47197. Los Chaguaramos. Caracas, Venezuela

Oscar Rondón
Departamento de Cómputo y Estadística.
Univesidad Simón Bolívar, Caracas, Venezuela

Abstract

   Let $\{W_{st} : s, t \in [0,1]\}$ be the Brownian sheet. We define the regularized process W_{st}^{ε} as the convolution of W_{st} and $\varphi_{\varepsilon}(s,t) = \frac{1}{\varepsilon^2}\varphi\left(\frac{s}{\varepsilon}\right)\varphi\left(\frac{t}{\varepsilon}\right)$ where φ is a function satisfying some conditions. For ω fixed we prove that

$$\lambda\left(\left\{(s,t) \in [0,1] \times [0,1] : \frac{\varepsilon}{\|\varphi\|_2^2}\frac{\partial^2 W_{st}^{\varepsilon}}{\partial s \partial t} \leq x\right\}\right) \xrightarrow[\epsilon \to 0]{} \Phi(x)$$

almost surely, where λ is the Lebesgue measure in R^2, Φ is the standard Gaussian distribution and $\|\cdot\|_2$ is the usual norm in $L^2([-1,1], dx)$. These results are generalized to two parameter martingales M given by stochastic integrals of the Cairoli & Walsh type. Finally, as a consequence of our method we also obtain similar results for the normalized double increment of the processes W and M. These results constitute a generalisation of those obtained by Wschebor for Brownian stochastic integrals.

Keywords: Wiener process, Brownian sheet, double increment

12.1   Introduction

   Several works have been recently devoted to study the problem of estimation of a process $\{X_t\}$ when one observes the process at discrete times i.e. $X_{\frac{k}{n}}$ or the observation is the smoothed process $X_t^{\varepsilon} = \int_{-\infty}^{\infty} \varphi(\frac{t-s}{\varepsilon})X_s ds$, where φ is a smooth kernel and ε is a window parameter. In each case the asymptotic behavior of the estimators is established when the step of observation $1/n$ or the window ε respectively, tend towards zero. This type of problems are important when the observation device allows improving the resolution. The case where the observed process is a Brownian diffusion has been studied by Genon-Catalot and Jacod in the discrete case and in Wschebor and Perera

and Wschebor in the other one. In this work we consider the same type of problems when we observe a regularization by convolution of a random field, which is solution of a stochastic differential equation driven by a Brownian sheet or more generally a Carioli-Walsh stochastic integral with respect to the Brownian sheet. We restrict our study to the law large number type result, the CLT will be considered elsewhere.

Let us introduce the problem. Wschebor has shown that, for almost every ω, the increments of the Wiener process $B = \{B_t, t \in [0,1]\}$ as a function of time converge in distribution towards a standard Gaussian distribution Φ. Namely, he proved that if $\Delta_\varepsilon(t) = \varepsilon^{-1/2}(B_{t+\varepsilon} - B_t)$ denotes the normalized increments of such a process and m is the Lebesgue measure in R then almost surely $m(\{t \in I : \Delta_\varepsilon(t) \leq x\}) \to m(I)\Phi(x)$ when $\varepsilon \to 0$ where I is any interval in $[0,1]$ and $x \in R$. Moreover he defined the process $B_t^\varepsilon = B * \zeta_\varepsilon(t)$ as the convolution of B_t and $\zeta_\varepsilon(t) = \varepsilon^{-1}\varphi(t/\varepsilon)$, a convolution kernel that approaches Dirac's delta function as $\varepsilon \to 0$, and he showed that almost surely $m(\{t \in I : \frac{\varepsilon^{1/2}}{\|\varphi\|_2}\}) \to m(I)\Phi(x)$ when $\varepsilon \to 0$ where $\|\cdot\|_2$ is the usual norm of $L^2([-1,1], dm)$. By taking $\varphi(x) = 1_{[-1,0]}(x)$ the result for the normalized increments is a particular case of this. Finally, Wschebor generalized these results to the class of stochastic processes N given by $N_t = \int_0^t \psi_s \, dB_s$ where ψ satisfies certain regularity conditions, and obtained that almost surely,

$$\lim_{\varepsilon \to 0} m\left(\left\{t \in I : \frac{\varepsilon^{1/2}}{\|\varphi\|_2} \frac{dN_t^\varepsilon}{dt} \leq x\right\}\right) = \int_I F_s(x) \, dx$$

where N_t^ε denotes the convolution of N_t and $\zeta_\varepsilon(t)$ and F_s is the distribution function of a centered normal variable with random variance equal to ψ_s^2.

In this article we follow Wschebor's method to generalize the above results to the case of the Brownian sheet instead of the Brownian motion and to the case of strong martingales, i.e., stochastic processes M given by

$$M_{st} = \int_0^s \int_0^t \Psi_{uv} dW_{uv} \tag{1.1}$$

where the integral considered is the stochastic integral of Cairoli and Walsh instead of the stochastic integral of Ito type. Note however that in this case the procedure is a little more involved due to the dimensional nature of the time parameter.

These results are interesting because they give a way to obtain nonparametric estimators of the coefficient a for two parameter stochastic differential equations:

$$dX_{st} = a(X_{st}) dW_{st} + b(X_{st}) dsdt$$

These models have been studied, for example, in Carmona and Nualart. We apply our results to this case, see the Remark of Theorem 3 and Corollary 2 bellow.

12.2 Assumptions and Notations

1 On the process W: $\{W_{st} : s, t \in [0, 1]\}$ is a Brownian sheet. In what follows, we shall suppose that W_{st} is defined for all $s, t \in R$ setting $W_{st} = 0$ if $s \notin [0, 1]$ or $t \notin [0, 1]$.

   For a rectangle $A = (s, s'] \times (t, t']$, $W(A)$ will denote the double increment over A, i.e. $W(A) = W_{s't'} - W_{st'} - W_{s't} + W_{st}$.

2 On the process M: $\{M_{st} : s, t \in [0, 1]\}$ is a two parameter strong martingale given by (1.1) where Ψ is a process satisfying the conditions of Cairoli and Walsh for this kind of integral. Also we suppose that M_{st} is defined for all $s, t \in R$ setting $M_{st} = 0$ if $s \notin [0, 1]$ or $t \notin [0, 1]$. Finally, $M(A)$ will denote the double increment of M over the rectangle A defined as before.

3 On the kernel φ: $supp\varphi \subset [-1, 1]$, φ is the distribution function of a (signed) measure $d\varphi(x)$ which has bounded total variation and $\int_{-1}^{1} \varphi(x)\, dx = 1$.

   Throughout the paper we shall consider W_{st}^{ε} and M_{st}^{ε} the regularization by convolution of $\varphi_{\varepsilon}(s, t) = \varepsilon^{-2}\varphi\left(\frac{s}{\varepsilon}\right)\varphi\left(\frac{t}{\varepsilon}\right)$ with W_{st} and M_{st} respectively and $Z_{\varepsilon}(s, t) = \frac{\varepsilon}{\|\varphi\|_2^2}\frac{\partial^2 W_{st}^{\varepsilon}}{\partial s \partial t}$ where

$$\frac{\partial^2 W_{st}^{\varepsilon}}{\partial s \partial t} = \frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W_{s-\varepsilon u, t-\varepsilon v}\, d\varphi(u)\, d\varphi(v)$$

   Note that $Z_{\varepsilon}(s, t)$ has standard normal distribution for each $s, t \in [\varepsilon, 1-\varepsilon]$ and that $Z_{\varepsilon}(s, t) = 0$ if $s, t \notin [\varepsilon, 1-\varepsilon]$. Finally, ξ will denote a standard normal variable, C shall stand for a generic constant whose value change during a proof, $A_{\varepsilon}(s, t)$ will denote the square $(s, s+\varepsilon] \times (t, t+\varepsilon]$ and I, J will be arbitrary intervals in $[0, 1)$.

12.3 Results

THEOREM 4 *If $\varepsilon \to 0$, then*

$$\lambda\left(\{(s, t) \in I \times J : Z_{\varepsilon}(s, t) \leq x\}\right) \to \lambda(I \times J)\,\Phi(x)$$

almost surely for all $x \in R$.

Remark. Taking $\varphi(x) = 1_{[-1,0]}(x)$ we have that $Z_{\varepsilon}(s, t) = \frac{W(A_{\varepsilon}(s,t))}{\varepsilon}$, so Theorem 1 holds for the normalized double increment of the Brownian sheet over $A_{\varepsilon}(s, t)$.

COROLLARY 2 *If* $f, g : R \rightarrow R$ *are continuous functions and* g *satisfies* $E \left| g(\xi) \right|^{1+\delta} \leq C$ *then*

$$\int_0^1 \int_0^1 f(W_{st}^\varepsilon) g(Z_\varepsilon(s,t)) \, dt ds \rightarrow E(g(\xi)) \int_0^1 \int_0^1 f(W_{st}) \, dt ds$$

a.s when $\varepsilon \rightarrow 0$.

THEOREM 5 *Let* $T_\Psi(\varepsilon) = \sup \{ |\Psi_{st} - \Psi_{s't'}| : |s - s'| \leq \varepsilon, |t - t'| \leq \varepsilon \}$. *If* $T_\Psi(\varepsilon) \left(\log \log (1/\varepsilon^2) \right)^{1/2}$ *tends to zero a.s when* $\varepsilon \rightarrow 0$ *then*

$$\lambda \left(\left\{ (s,t) \in I \times J : \frac{M(A_\varepsilon(s,t))}{\varepsilon} \leq x \right\} \right) \rightarrow \int_I \int_J F_{st}(x) \, ds dt$$

a.s when $\varepsilon \rightarrow 0$, *where* F_{st} *is the distribution function of a centered normal variable with random variance* Ψ_{st}^2.

THEOREM 6 *Let* $\Upsilon_j(\varepsilon) = \sup \left\{ E \left| \Psi_{st} - \Psi_{s't'} \right|^{2j} : |s - s'| \leq \varepsilon, |t - t'| \leq \varepsilon \right\}$. *If* Ψ *has continuous paths and for each* j *there exist positive constants* c_j *and* γ_j *such that* $\Upsilon_j(\varepsilon) \leq c_j \varepsilon^{\gamma_j}$ *then*

$$\lambda \left(\left\{ (s,t) \in I \times J : \frac{\varepsilon}{\|\varphi\|_2^2} \frac{\partial^2 M_{st}^\varepsilon}{\partial s \partial t} \leq x \right\} \right) \rightarrow \int_I \int_J F_{st}(x) \, ds dt$$

a.s when $\varepsilon \rightarrow 0$.

COROLLARY 3 *Under the assumptions of Theorem 6 and Corollary 1,*

$$\int_0^1 \int_0^1 f(M_{st}^\varepsilon) g\left(\frac{\varepsilon}{\|\varphi\|_2^2} \frac{\partial^2 M_{st}^\varepsilon}{\partial s \partial t} \right) ds dt$$

$$\rightarrow \int_0^1 \int_0^1 f(M_{st}) \left(\frac{1}{\sqrt{2\pi} \Psi_{st}} \int_{-\infty}^\infty g(x) e^{-\frac{x^2}{2\Psi_{st}^2}} dx \right) dt ds$$

a.s when $\varepsilon \rightarrow 0$.

Remark. Theorem 6 can be applied to X_{st}^ε that is a regularization of the process solution of the equation:

$$dX_{st} = a(X_{st}) \, dW_{st} + b(X_{st}) \, ds dt$$

obtaining

$$\lambda \left(\left\{ (s,t) \in I \times J : \frac{\varepsilon}{\|\varphi\|_2^2} \frac{\partial^2 X_{st}^\varepsilon}{\partial s \partial t} \leq x \right\} \right) \rightarrow \int_I \int_J G_{st}(x) \, ds dt$$

a.s. where

$$G_{st}(x) = \frac{1}{\sqrt{2\pi}a(X_{st})} \int_{-\infty}^{x} e^{-\frac{u^2}{2a^2(X_{st})}} du.$$

Moreover taking $f \equiv 1$ and $g(x) = x^2$ Corollary 2 gives

$$\int_0^1 \int_0^1 \left(\frac{\varepsilon}{\|\varphi\|_2^2} \frac{\partial^2 X_{st}^{\varepsilon}}{\partial s \partial t} \right)^2 ds dt \to \int_0^1 \int_0^1 a^2(X_{st}) ds dt$$

a.s.

12.4 Proofs

We can assume for simplicity sake that $I = J = [0, 1]$. The proof for general intervals can be treated in a similar fashion, with some minor modifications.

12.4.1. Proof of Theorem 4

First, we observe that it is sufficient to prove the convergence of the moments of $Z_\varepsilon(s, t)$, as a random variable in the time parameters, to the moments of a standard normal variable. i.e. to prove that $V_k(\varepsilon) = \int_0^1 \int_0^1 Z_\varepsilon(s, t)^k \, ds dt$ tends to $E(\xi^k)$ a.s when $\varepsilon \to 0$, for all $k \geq 1$.

Computing covariances we can show that $Z_\varepsilon(s, t)$ and $Z_\varepsilon(s', t')$ are independent if $|s - s'| \geq 2\varepsilon$ or $|t - t'| \geq 2\varepsilon$. Using this fact we can see that

$$\text{var}(V_k(\varepsilon)) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 \text{cov}(Z_\varepsilon(s, t), Z_\varepsilon(s', t')) \, dt' ds' ds dt \leq C \varepsilon$$

splitting conveniently the integrals. Therefore, if $\varepsilon_\nu = \nu^{-a}$ $a > 1$, the Borel-Cantelli Lemma implies that $V_k(\varepsilon_\nu) \to E(\xi^k)$ a.s when $\nu \to +\infty$. To finish the demonstration we have to show that $\sup_{\varepsilon_{\nu+1} \leq \varepsilon \leq \varepsilon_\nu} |V_k(\varepsilon_\nu) - V_k(\varepsilon)| \to 0$ when $\nu \to +\infty$.

Start with $|V_k(\varepsilon_\nu) - V_k(\varepsilon)| \leq J_1 + J_2$ where

$$J_1 = \frac{|\varepsilon^k - \varepsilon_\nu^k|}{\|\varphi\|_2^{2k}} \left| \int_0^1 \int_0^1 \left(\frac{\partial^2 W_{st}^{\varepsilon_\nu}}{\partial s \partial t} \right)^k ds dt \right|$$

and

$$J_2 = \frac{\varepsilon^k}{\|\varphi\|_2^{2k}} \left| \int_0^1 \int_0^1 \left(\frac{\partial^2 W_{st}^{\varepsilon}}{\partial s \partial t} \right)^k ds dt - \int_0^1 \int_0^1 \left(\frac{\partial^2 W_{st}^{\varepsilon_\nu}}{\partial s \partial t} \right)^k ds dt \right|$$

As $\varepsilon_{\nu+1} \leq \varepsilon \leq \varepsilon_\nu$ for term J_1 we have $J_1 \leq \left| 1 - \frac{\varepsilon_{\nu+1}^k}{\varepsilon_\nu^k} \right| |V_k(\varepsilon_\nu)| \rightarrow 0$ a.s

when $\nu \rightarrow +\infty$. Next we define $A(s,t,\varepsilon) = \frac{\partial^2 W_{st}^\varepsilon}{\partial s \partial t}$ and $B(s,t,\varepsilon,\varepsilon_\nu) = A(s,t,\varepsilon) - A(s,t,\varepsilon_\nu)$ and using the identity

$$(A+B)^k - A^k = \sum_{j=0}^{k-1} \binom{k}{j} A^j B^{k-j}$$

we obtain that

$$J_2 \leq C \frac{\varepsilon_\nu^k}{\|\varphi\|_2^{2k}} \sum_{j=0}^{k-1} \int_0^1 \int_0^1 |A(s,t,\varepsilon_\nu)|^j |B(s,t,\varepsilon,\varepsilon_\nu)|^{k-j} \, dsdt$$

By the appendix with $\varepsilon_\nu \leq s,t \leq 1 - \varepsilon_\nu$ we conclude that $|A(s,t,\varepsilon_\nu)|^j \leq \varepsilon_\nu^{-j(1+\delta)} K_\varphi^j$ and $|B(s,t,\varepsilon,\varepsilon_\nu)|^{k-j} \leq \varepsilon_{\nu+1}^{-(k-j)} H_\delta^{(k-j)}(\nu)$ where, for $0 < \delta < 1/2$,

$$H_\delta(\nu) = 2K_\varphi \frac{\varepsilon_\nu^{2(1-\delta)}}{\varepsilon_{\nu+1}^2} \left| 1 - \frac{\varepsilon_{\nu+1}}{\varepsilon_\nu} \right|^{\frac{1}{2} - \delta}$$

and K_φ is a constant dependent on φ. Therefore,

$$J_2 \leq \frac{C}{\|\varphi\|_2^{2k}} \sum_{j=0}^{k-1} \frac{\varepsilon_\nu^k}{\varepsilon_{\nu+1}^k} \frac{\varepsilon_{\nu+1}^j}{\varepsilon_\nu^j} \varepsilon_\nu^{-j\delta} H_\delta^{(k-j)}(\nu)$$

For δ small enough $\varepsilon_\nu^{-j\delta} H_\delta^{(k-j)}(\nu) \rightarrow 0$ when $\nu \rightarrow +\infty$ for all $0 \leq j \leq k-1$. So J_2 tends to zero when $\nu \rightarrow +\infty$ and this completes the proof of Theorem 1.

12.4.2.   Proof of Corollary 1

Note that

$$\left| \int_0^1 \int_0^1 f(W_{st}^\varepsilon) g(Z_\varepsilon(s,t)) \, dsdt - E(g(\xi)) \int_0^1 \int_0^1 f(W_{st}) \, dsdt \right| \leq Q_1 + Q_2$$

where

$$Q_1 = \sup_{s,t \in [0,1]} |f(W_{st}^\varepsilon) - f(W_{st})| \int_0^1 \int_0^1 |g(Z_\varepsilon(s,t))| \, dsdt$$

and

$$Q_2 = \left| \int_0^1 \int_0^1 f(W_{st}) g(Z_\varepsilon(s,t)) \, dsdt - E(g(\xi)) \int_0^1 \int_0^1 f(W_{st}) \, dsdt \right|$$

Using the Dominated Convergence Theorem we have that $\lim_{\varepsilon \to 0} W_{st}^{\varepsilon} = W_{st}$. Therefore, the continuity of f and the boundedness of W_{st}^{ε} and W_{st} imply that

$$\sup_{s,t \in [0,1]} |f(W_{st}^{\varepsilon}) - f(W_{st})| \to 0$$

when $\varepsilon \to 0$. To see that Q_1 tends to zero when $\varepsilon \to 0$ it is sufficient to observe that by Theorem 1 and the assumption on g we have that

$$\int_0^1 \int_0^1 |g(Z_\varepsilon(s,t))| \, dsdt \to E|g(\xi)|$$

when $\varepsilon \to 0$. The convergence to zero of Q_2 can be obtained from Theorem 4 by a standard approximation argument.

Remark. Following the proof of Corollary 1 we can show that

$$\int_0^s \int_0^t uvf(W_{uv}^{\varepsilon}) g(Z_\varepsilon(u,v)) \, dvdu \to E(g(\xi)) \int_0^s \int_0^t uvf(W_{uv}) \, dvdu$$

a.s when $\varepsilon \to 0$. Therefore, if f is bounded we have by Theorem 6.1 of Cairoli and Walsh [1] that there exists a process $\{\phi(x; s,t) : x \in R, s, t \in [0,1]\}$ which is a.s. jointly continuous in x, s and t such that

$$\int_0^s \int_0^t uvf(W_{uv}) \, dvdu = \int_{-\infty}^{+\infty} \phi(x; s,t) f(x) \, dx$$

almost surely. Hence, we obtain an a.s. approximation of this kind of local time for the Brownian sheet.

12.4.3.   Proof of Theorem 5

We have

$$\frac{M(A_\varepsilon(s,t))}{\varepsilon} = \frac{W(A_\varepsilon(s,t))}{\varepsilon} \Psi_{st} + \frac{1}{\varepsilon} \int_s^{s+\varepsilon} \int_t^{t+\varepsilon} [\Psi_{s't'} - \Psi_{st}] \, dW_{s't'}$$

$$(4.1)$$

Using Theorem 4 and the remark at the end of it, we have that

$$\lim_{\varepsilon \to 0} \lambda \left(\left\{ (s,t) \in [0,1] \times [0,1] : \frac{W(A_\varepsilon(s,t))}{\varepsilon} \Psi_{st} \leq x \right\} \right)$$

$$= \int_0^1 \int_0^1 F_{st}(x) \, dsdt$$

a.s. Hence, if we denote by Q_ε^{st} the second term in the right hand side of (4.1), it is enough to prove that Q_ε^{st} tends to zero when $\varepsilon \to 0$ for almost all w and $s, t \in [0,1]$ to finish the proof of Theorem 2.

First, notice that $U_\varepsilon^{st} = \varepsilon Q_\varepsilon^{st}$ is a stochastic integral in the plane. Therefore, it is a martingale and a i-martingale, $i = 1, 2$ (see Cairoli and Walsh). Hence, for fixed $s, t \in [0, 1]$, $\{U_\varepsilon^{st}, \ \varepsilon > 0\}$ is a martingale with increasing process

$$C_\varepsilon^{st} = \int_s^{s+\varepsilon} \int_t^{t+\varepsilon} [\Psi_{s't'} - \Psi_{st}]^2 \ dt'ds' \leq \varepsilon^2 T_\Psi^2(\varepsilon)$$

Therefore, using the time change theorem, the law of iterated logarithm and our assumptions we have the desired result.

12.4.4.   Proof of Theorem 6

As in the proof of Theorem 4, it is sufficient to show the a.s convergence of moments of order $k \geq 1$ to $E\left(\xi^k\right) \int_0^1 \int_0^1 \Psi_{st}^k \ dsdt$.
Using the differentiation formulas of pages 224 and 226 of Farré and Nualart with $\zeta_\varepsilon(s) = \varepsilon^{-1}\varphi(s/\varepsilon)$ we obtain that

$$\frac{\partial^2 M_{st}^\varepsilon}{\partial s \partial t} = \frac{\partial^2 W_{st}^\varepsilon}{\partial s \partial t} \Psi_{st} + \int_{R^2} \zeta_\varepsilon(s - s') \zeta_\varepsilon(t - t') [\Psi_{s't'} - \Psi_{st}] \ dW_{s't'}$$

Taking

$$I_\varepsilon(s, t) = \frac{\varepsilon}{\|\varphi\|_2^2} \frac{\partial^2 W_{st}^\varepsilon}{\partial s \partial t} \Psi_{st}$$

and

$$J_\varepsilon(s, t) = \frac{\varepsilon}{\|\varphi\|_2^2} \int_{R^2} \zeta_\varepsilon(s - s') \zeta_\varepsilon(t - t') [\Psi_{s't'} - \Psi_{st}] dW_{s', t'}$$

we have that

$$\int_0^1 \int_0^1 \left(\frac{\varepsilon}{\|\varphi\|_2^2} \frac{\partial^2 M_{st}^\varepsilon}{\partial s \partial t}\right)^k dsdt = \sum_{j=0}^{k-1} \binom{k}{j} U_\varepsilon(j, k - j) + U_\varepsilon(k, 0) \quad (4.2)$$

with $U_\varepsilon(i, j) = \int_0^1 \int_0^1 I_\varepsilon(s, t)^i J_\varepsilon(s, t)^j \ dsdt$.
Theorem 4 implies that $U_\varepsilon(k, 0) \to E\left(\xi^k\right) \int_0^1 \int_0^1 \Psi_{st}^k \ dsdt$ a.s when $\varepsilon \to 0$. So, to finish the proof we have to show that the first term in the right hand side of (4.2) tends to zero a.s when $\varepsilon \to 0$. Using the inequality from Theorem 2.1 of Guyon and Prum we have that for any positive integer h

$$E\left|J_\varepsilon(s, t)\right|^{2h} \leq$$

$$C \frac{\varepsilon^{4h-2}}{\|\varphi\|_2^{4h}} \int_{s-\varepsilon}^{s+\varepsilon} \int_{t-\varepsilon}^{t+\varepsilon} \zeta_\varepsilon^{2h}(s - s') \zeta_\varepsilon^{2h}(t - t') E[\Psi_{s't'} - \Psi_{st}]^{2h} \ dt'ds'$$

Because of the hypothesis on Υ we obtain that $E\,|J_\varepsilon\,(s,t)|^{2h} \leq C\ c_h\varepsilon^{\gamma h}$. Therefore

$$E\left(\int_0^1 \int_0^1 |J_\varepsilon\,(s,t)|^{2h}\right)\,dsdt \leq C\ c_h\varepsilon^{\gamma h}$$

Using the Borel-Cantelli Lemma with $\varepsilon_\nu = \nu^{-a}$ as in Theorem 1 we obtain that $U_{\varepsilon_\nu}\,(0,2h) \to 0$ a.s when $\nu \to +\infty$ for any positive integer h. So,

$$U_{\varepsilon_\nu}\,(j,k-j)^2 \leq U_{\varepsilon_\nu}\,(2j,0)\,U_{\varepsilon_\nu}\,(0,2\,(k-j)) \to 0$$

almost surely when $\nu \to +\infty$. Hence

$$\int_0^1 \int_0^1 \left(\frac{\varepsilon_\nu}{\|\varphi\|_2^2}\frac{\partial^2 M_{st}^{\varepsilon_\nu}}{\partial s\partial t}\right)^k\,dsdt \to E\left(\xi^k\right)\int_0^1 \int_0^1 \Psi_{st}^k\,dsdt$$

a.s. when $\nu \to +\infty$.

Finally, we can obtain analogous results to those of the appendix for the process M, and proceeding as in Theorem 1 we have the result.

Appendix

In this appendix, we show how to obtain the bounds for terms $|A\,(s,t,\varepsilon_\nu)|$ and $|B\,(s,t,\varepsilon,\varepsilon_\nu)|$ used in the proof of Theorem 1.

Recall that $A\,(s,t,\varepsilon) = \frac{\partial^2 W_{st}^\varepsilon}{\partial s\partial t}$ and $B\,(s,t,\varepsilon,\varepsilon_\nu) = A\,(s,t,\varepsilon) - A\,(s,t,\varepsilon_\nu)$. Using $\int_{-\infty}^{+\infty} d\varphi\,(x) = 0$ we have that

$$\varepsilon A\,(s,t,\varepsilon) = \frac{1}{\varepsilon}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W\,(D_\varepsilon\,(s,t,u,v))\,d\varphi\,(u)\,d\varphi\,(v)$$

where $\varepsilon < s,t < 1 - \varepsilon$ and $D_\varepsilon\,(s,t,u,v) = (s - \varepsilon, s - \varepsilon u] \times (t - \varepsilon, t - \varepsilon v]$. So, because of the modulus of continuity of the Brownian sheet (see Csörgó and Révész or Orey and Pruitt) and the hypothesis on φ we obtain that

$$|\varepsilon A\,(s,t,\varepsilon)| \leq C\,\frac{1}{\varepsilon}\varepsilon^{(1-\delta)}\left(\int_{-\infty}^{+\infty} d|\varphi|(x)\right)^2 = \varepsilon^{-\delta}\,K_\varphi$$

with $|\varphi|$ the total variation of φ. Thus

$$|A\,(s,t,\varepsilon)| \leq \varepsilon^{-(1+\delta)}\,K_\varphi \tag{A.1}$$

Regarding the other term, we observe first that

$$|\varepsilon B\,(s,t,\varepsilon,\varepsilon_\nu)| \leq \left|\varepsilon\frac{\partial^2 W_{st}^\varepsilon}{\partial s\partial t} - \varepsilon_\nu\frac{\partial^2 W_{st}^{\varepsilon_\nu}}{\partial s\partial t}\right| + \frac{|\varepsilon - \varepsilon_\nu|}{\varepsilon_\nu}\left|\varepsilon_\nu\frac{\partial^2 W_{st}^{\varepsilon_\nu}}{\partial s\partial t}\right|$$

the second term in the right hand side of the above equation can be bounded, using (A. 1), by

$$\frac{|\varepsilon - \varepsilon_\nu|}{\varepsilon_\nu}\varepsilon_\nu^{-\delta}\,K_\varphi \leq \left|1 - \frac{\varepsilon_{\nu+1}}{\varepsilon_\nu}\right|\varepsilon_\nu^{-\delta}\,K_\varphi$$

and the first by

$$\frac{1}{\varepsilon_{\nu+1}^2} \int_{R^2} \left[f_1\left(s,t,u,v,\varepsilon,\varepsilon_\nu\right) + f_2\left(s,t,u,v,\varepsilon,\varepsilon_\nu\right) \right] \, d|\varphi|(v)d|\varphi|(u)$$

with

$$f_1\left(s,t,u,v,\varepsilon,\varepsilon_\nu\right) = |\varepsilon - \varepsilon_\nu| \, |W\left(D_{\varepsilon_\nu}\left(s,t,u,v\right)\right)|$$

and

$$f_2\left(s,t,u,v,\varepsilon,\varepsilon_\nu\right) = \varepsilon_\nu \left| W\left(D_{\varepsilon_\nu}\left(s,t,u,v\right)\right) - W\left(\overline{D}_{\varepsilon,\varepsilon_\nu}\left(s,t,u,v\right)\right) \right|$$

where $\overline{D}_{\varepsilon,\varepsilon_\nu}\left(s,t,u,v\right) = (s - \varepsilon_\nu, s - \varepsilon u] \times (t - \varepsilon_\nu, t - \varepsilon v]$ for $\varepsilon_\nu < s, t < 1 - \varepsilon_\nu$. Using again the modulus of continuity of the Brownian sheet, the function f_1 can be bounded by $C \left| 1 - \frac{\varepsilon_{\nu+1}}{\varepsilon_\nu} \right| \varepsilon_\nu^{(2-\delta)}$ for $0 < \delta < 1/2$. With respect to function f_2, it is enough to study the shape of the rectangles $D_{\varepsilon_\nu}\left(s,t,u,v\right)$ and $\overline{D}_{\varepsilon,\varepsilon_\nu}\left(s,t,u,v\right)$ for distinct values (positive or negatives) of u and v and to use the modulus of continuity to obtain that

$$\left| W\left(D_{\varepsilon_\nu}\left(s,t,u,v\right)\right) - W\left(\overline{D}_{\varepsilon,\varepsilon_\nu}\left(s,t,u,v\right)\right) \right| \le C \left[\varepsilon_\nu \left(\varepsilon_\nu - \varepsilon_{\nu+1}\right) \right]^{\frac{1}{2}-\delta}$$

Hence,

$$\begin{aligned}
\varepsilon B\left(s,t,\varepsilon,\varepsilon_\nu\right) &\le K_\varphi \frac{\varepsilon_\nu^{(2-2\delta)}}{\varepsilon_{\nu+1}^2} \left| 1 - \frac{\varepsilon_{\nu+1}}{\varepsilon_\nu} \right|^{\frac{1}{2}-\delta} + \left| 1 - \frac{\varepsilon_{\nu+1}}{\varepsilon_\nu} \right| K_\varphi \, \varepsilon_\nu^{-\delta} \\
&\le 2K_\varphi \left| 1 - \frac{\varepsilon_{\nu+1}}{\varepsilon_\nu} \right|^{\frac{1}{2}-\delta} \frac{\varepsilon_\nu^{2(1-\delta)}}{\varepsilon_{\nu+1}^2} = H_\delta\left(\nu\right)
\end{aligned}$$

or equivalently $\left| B\left(s,t,\varepsilon,\varepsilon_\nu\right) \right| \le \varepsilon_{\nu+1}^{-1} H_\delta\left(\nu\right)$.

References

Cairoli, R. and Walsh, J. Stochastic Integral in the Plane, *Acta Math.* **134**, 1975, 111-83.

Carmona, R and Nualart,D. Random Non-Linear Wave Equations: Smoothness of the solutions, *Probab.Th. Rel. Fields,* **79,** 1988, 469-508

Csörgó, M and Révész, P. *Strong Approximations in Probability and Statistic (Academic Press, New York),* 1981.

Farré, M. and Nualart, D. Nonlinear Stochastic Integral Equations in the Plane, *Stochastic Processes and their Applications.* **46**, 1993, 219-239.

Genon-Catalot, V. and Jacod, J. On the estimation of the diffusion coefficient from discrete observations, *Ann. Int. Henri Poincaré.* **28**, 1992, 119-151.

Guyon, X. and Prum, B. Variations Produit et Formule de Ito pour les Semi-Martingales Représentable a Deux Paramètres, *Z. Wahrscheinlichkeitstheorie verw.* **56**, 1981, 361-369.

Orey, S. and Pruitt, W. Sample Functions for the N-parameter Wiener Process, *The Annals of Probability.* **1**, 1973, 138-163.

Perera, G. and Wschebor, M. Crossings and occupation measures for a class of semimartingales *The Annals of Probability* **26**, 1998, 253-266.

Wschebor, M. Sur les Accroissements du Processus de Wiener. *C.R.A.S.* **315**, 1992, 1293-1296.

COMPOUND POISSON APPROXIMATION WITH DRIFT FOR STOCHASTIC ADDITIVE FUNCTIONALS WITH MARKOV AND SEMI-MARKOV SWITCHING

Vladimir S. Korolyuk
Ukrainian National Academy of Science, Ukraine

Nikolaos Limnios
Université de Technologie de Compiègne, France

Abstract   We present Poisson,approximation results for additive functionals switched by Markov and semi-Markov processes.   The weak convergence results are obtained via semimartingale representations of additive functionals and the convergence of generators for Markov processes and of compensative operator of the extended Markov renewal processes. This is a review paper of our previous results given in [Korolyuk, 2002; Korolyuk, 2002A].

Keywords: Additive functional, Poisson approximation, Compound Poisson approximation with drift, Markov, semi-Markov switching, semimartingale, compensative operator, extended Markov renewal process.

13.1   Introduction

Poisson approximation is a very active research field [Aldous, 1989; Barhour, 1992; Barbour, 2002]. Three kind of Poisson processes approximation exist: standard Poisson process [Aldous, 1989; Barbour, 1992], compound Poisson process [Barbour, 2002], and compound Poisson process with drift [Korolyuk, 2000; Korolyuk, 2001A; Korolyuk, 2002].

A compound Poisson process with drift (CPPD) is defined as follows

$$\xi(t) = at + \sum_{k=1}^{\nu(t)} \alpha_k, \quad t \geq 0, \qquad (1.1)$$

where (α_k) is a real i.i.d. sequence, $\nu(t), t \geq 0$ is a time-homogeneous Poisson process and $a \in \mathbb{R}$.

The results we present here are a review of our previous results [Korolyuk, 2000; Korolyuk, 2001A; Korolyuk, 2002; Korolyuk, 2002A], and concern approximation of additive functionals by CPPDs like (1.1).

Additive functionals of stochastic processes play an important part in theory and in many applications [Korolyuk, 1999; Korolyuk, 1999A; Korolyuk, 1999B; Korolyuk, 2001; Korolyuk, 2000A; Korolyuk, 2000B; Korolyuk, 2000C; Korolyuk, 2002]. We have obtained diffusion approximation of additive functionals with Markov switching with and without balance condition in [Korolyuk, 2000A; Korolyuk, 2000B; Korolyuk, 2000C], and Poisson approximation for increment processes and their stochastic exponentials with Markov switching in [Korolyuk, 2000]. In the above cases, we have worked in the settings of the books [Jacod, 1987] and [Ethier, 1986], where the martingale characterization is used. We have also obtained results of CPPD approximation for integral functionals with semi-Markov switching [Korolyuk, 2002]. In the latter case, due to the semi-Markov process, the martingale characterization does not further works, hence a need for more adapted tools. In fact, we make use of the compensative operator for extended Markov renewal processes, introduced by Wentzel & Sviridenko [Sviridenko, 1989], from which we derive the martingale characterization.

Consider a sequence of r.v.s $\alpha_k, \ k \geq 0,$ and a multivariate point process $\tau_k, x_k, \ k \geq 1,$ [Anisimov, 1995; Borovskikh, 1997; Jacod, 1987; Korolyuk, 1999; Liptser, 1989], with counting process $\nu(t) := \inf\{k \geq 1 : \tau_k \leq t\}.$ The stochastic process $\zeta(t), t \geq 0,$ defined by

$$\zeta(t) := \sum_{k=1}^{\nu(t)} \alpha_k(\theta_k, x_{k-1}), \quad t \geq 0, \tag{1.2}$$

is called an *increment process* [Borovskikh, 1997; Jacod, 1987]. We study the increment process with Markov switching as an additive semimartingale [Çinlar, 1980]. If the r.v.s $\alpha_k, \ k \geq 1,$ are iid and the multivariate point process $\tau_k, x_k, \ k \geq 0, \theta_{k+1} = \tau_{k+1} - \tau_k,$ is just a renewal point process on $\mathbb{R}_+,$ then (1.2) is called a *compound process* or a *renewal reward process* [Osaki, 1985]. If the r.v.s $\alpha_k, \ k \geq 0,$ are iid and the multivariate point process $\tau_k, x_k, \ k \geq 0,$ is just a Poisson point process on $\mathbb{R}_+,$ then (1.2) is called a *compound Poisson process* [Osaki, 1985]. If α_k is a fixed function defined on $\mathbb{R} \times E,$ then (1.2) is a shot noise process [Parzen, 1999], which play an important role in the theory of noise of physical devices. In [Kluppelberg, 1995] the authors consider the random measure $\alpha_k(\tau_k)$ and a Poisson process $\nu(t),$ and they derive asymptotic results for (1.2) with application in insurance. For a semimartingale representation, see [Borovskikh, 1997; Çinlar, 1980; Jacod, 1987; Liptser, 1994; Liptser, 1989; Liptser, 1991].

The *Additive functionals* that we consider are of the following form

$$\xi(t) := \int_0^t \eta(ds; x(s)), \quad t \geq 0,$$ (1.3)

where the switched process $\eta(t, x), x \in E$ is a Markov process with locally independent increments and the switching process $x(t)$ is a semi-Markov process, with state space E. This additive functional is a continuous functional.

In fact, an additive functional can also be represented by the sum of an increment process and of another term, i.e.,

$$\xi(t) = \sum_{k=1}^{\nu(t)-1} \eta(\theta_k; x_{k-1}) + \eta(t - \tau_{\nu(t)}; x(t)).$$

This kind of processes are widely used in applications, i.e., risk and storage theory [Prabhu, 1980], reliability and maintenance theory [Osaki, 1985], finance and insurance [Kluppelberg, 1995], noise of physical device [Parzen, 1999], etc. In applications, θ_{k+1} is the acting time of the kth event and α_k is its magnitude, its cost, etc..

In many applied problems the r.v.s α_k depend on the environment. For example, the cost of a damage for an insurance company depends on which place, time, weather, etc. it happens. In the case where we have a multistate environment, E say, we suppose that the r.v.s α_k depend on the state $x \in E$, denoted $\alpha_k(x)$.

The increment process considered here is based on a multivariate point process which corresponds to a Markov renewal representation of a Markov process and the r.v.s α_k depend on the states of that Markov process. The convergence of the increment process towards a compound Poisson process with drift is due to the fact that we assume the r.v.s α_k, take small values with big probabilities and big values with small probabilities. Small jumps are transformed into deterministic drift.

In Section 2, we define continuous and discontinuous additive functionals. In Section 3, we give weak convergence results of the increment processes towards compound Poisson processes with drift. In Section 4, we consider an asymptotic split phase space for the switching Markov process and give Poisson,approximation results of the increment processes. In Section 5, we give Poisson approximation results for an additive functional with semi-Markov switching process. Finally, in Section 6, we give the main steps of proof of the theorems.

13.2   Preliminaries

Let us consider a time-homogeneous cadlag stochastic process $x(t), t \geq 0$ with values in a Polish space $(E, \mathcal{E})$. Times $0 = \tau_0 \leq \tau_1 \leq \cdots$ denote the jump times and define the embedded process $x_n = x(\tau_n), \ n \geq 0$.

Let $\eta^\varepsilon(t; x), \ t \geq 0, \ x \in E, \ \varepsilon > 0$, be a family of homogeneous Markov jump processes in the Euclidean space $\mathbb{R}^d, \ d \geq 1$, defined by the generators

$$\mathbb{\Gamma}^\varepsilon(x)\varphi(u) = \varepsilon^{-1} \int_{\mathbb{R}^d} [\varphi(u+v) - \varphi(u)]\Gamma_\varepsilon(dv; x), \ x \in E. \qquad (2.1)$$

The results presented here concern the following additive functionals:

$$\zeta^\varepsilon(t) := \sum_{k=1}^{\nu(t/\varepsilon)} \alpha_k^\varepsilon(x_k), \quad t \geq 0, \qquad (2.2)$$

and

$$\xi^\varepsilon(t) := \int_0^t \eta^\varepsilon(ds; x(s/\varepsilon)), \quad t \geq 0. \qquad (2.3)$$

In the first case (2.2) we suppose that the process $x(t), t \geq 0$ is a Markov process, and that it is uniformly ergodic with stationary distribution $\pi(B), \ B \in \mathcal{E}$. Thus the embedded Markov chain $x_k, \ k \geq 0$, is uniformly ergodic too, with stationary distribution $\rho(B), \ B \in \mathcal{E}$, related by the following relation

$$\pi(dx)q(x) = q\rho(dx), \quad q := \int_E \pi(dx)q(x). \qquad (2.4)$$

In the sequel we will suppose that

$$0 < q_0 \leq q(x) \leq q_1 < +\infty, \quad x \in E. \qquad (2.5)$$

In the second case (2.3), we suppose that $x(t), t \geq 0$ is a semi-Markov process with semi-Markov kernel

$$Q(x, B, t) = P(x, B)G_x(t), \quad x \in E, \ B \in \mathcal{E}, \ t \geq 0, \qquad (2.6)$$

which defines the associated Markov renewal process $(x_n, \tau_n; \ n \geq 0)$ by :

$$\begin{aligned} Q(x, B, t) &= \mathbb{P}(x_{n+1} \in B, \theta_{n+1} \leq t \mid x_n = x) \\ &= \mathbb{P}(x_{n+1} \in B \mid x_n = x)\mathbb{P}(\theta_{n+1} \leq t \mid x_n = x). \end{aligned} \qquad (2.7)$$

The semi-Markov process defined by the semi-Markov kernel (2.6) is a special case whose $G_x(t)$ does not depend on the next visited state. Nevertheless, this is not restrictive since any semi-Markov process can be transformed into the above form, see [Limnios, 2001].

Here $\theta_{n+1} := \tau_{n+1} - \tau_n$, $n \geq 0$, are the sojourn times given by the distribution functions

$$G_x(t) = \mathbb{P}(\theta_{n+1} \leq t \mid x_n = x) =: \mathbb{P}(\theta_x \leq t). \qquad (2.8)$$

The embedded Markov chain $(x_n, \ n \geq 0)$ is defined by the stochastic kernel

$$P(x, B) = \mathbb{P}(x_{n+1} \in B \mid x_n = x). \qquad (2.9)$$

We suppose that the semi-Markov process $x(t)$, $t \geq 0$, is regular [Limnios, 2001], that is to say

$$\mathbb{P}_x(\nu(t) < \infty) = 1, \quad \text{for all } x \in E, \ \text{and} \ t \in \mathbb{R}_+, \qquad (2.10)$$

with the counting process

$$\nu(t) = \max\{n \geq 0 : \tau_n \leq t\}. \qquad (2.11)$$

The additive functional (2.3) can be represented by the sum

$$\xi^\varepsilon(t) = \sum_{n=0}^{\nu(t/\varepsilon)-1} \eta^\varepsilon(\varepsilon\theta_{n+1}^\varepsilon; x_n) + \eta^\varepsilon(\theta^\varepsilon(t); x(t/\varepsilon)), \qquad (2.12)$$

where $\theta^\varepsilon(t) := t/\varepsilon - \tau(t)$, $\tau(t) := \tau_{\nu(t)}$.

We will present Poisson,approximation results of functionals (2.2) and (2.3) by a semimartingale approach. In both cases, the limit processes are compound Poisson processes with drift.

The semimartingale approach used here is interesting not only because it offers a general framework for convergence of stochastic processes but also because the semimartingale representation of additive functionals is obtained by using the Poisson approximation conditions for distribution functions of jumps.

13.3   Increment Process

Let us introduce the convergence-determining class of functions $C_3(\mathbb{R})$, (see [Jacod, 1987], VII.2.7). This class is characterized by the following condition: $g \in C_3(\mathbb{R})$ is a real-valued bounded continuous function with $g(u)/u^2 \to 0$ as $|u| \to 0$.

Let us consider the additive functional $\zeta^\varepsilon(t)$ given in (2.3).

Assumptions (A)

(A1:)   The switching Markov jump process $x(t)$, $t \geq 0$, is uniformly ergodic with the stationary distribution (2.4).

(A2:)   The family of random variables $\alpha_k^\varepsilon(x)$, $k \geq 0$, $x \in E$ is uniformly square integrable, i.e.,

$$\sup_{\varepsilon > 0} \sup_{x \in E} \int_{|u| > c} u^2 \Phi_x^\varepsilon(du) \longrightarrow 0, \quad \text{as } c \to \infty.$$

(A3:)   Approximation of mean value

$$\int_{\mathbb{R}} u \Phi_x^\varepsilon(du) = \varepsilon[a(x) + \theta^\varepsilon(x)],$$

and $\sup_{x \in E} |a(x)| \leq a < \infty$.

(A4:)   Poisson,approximation condition

$$\int_{\mathbb{R}} g(u) \Phi_x^\varepsilon(du) = \varepsilon[\Phi_x(g) + \theta_g^\varepsilon(x)], \quad g \in C_3(\mathbb{R}),$$

and $\sup_{x \in E} |\Phi_x(g)| \leq \Phi(g) < \infty$.

(A5:)   Square-integrability condition

$$\sup_{x \in E} \int_{\mathbb{R}} u^2 \Phi_x(du) < +\infty.$$

where the measure $\Phi_x(du)$ is defined by the relation (see [Jacod, 1987])

$$\Phi_x(g) = \int_{\mathbb{R}} g(u) \Phi_x(du), \quad g \in C_3(\mathbb{R}).$$

The negligible terms $\theta^\varepsilon(x)$ and $\theta_g^\varepsilon(x)$ in the above conditions satisfy:

$$\sup_{x \in E} |\theta^\varepsilon(x)| \to 0, \quad \sup_{x \in E} |\theta_g^\varepsilon(x)| \to 0, \quad \varepsilon \to 0.$$

THEOREM 1 *Under Assuptions A1-A5, the increment process (2.2) converges weakly to the compound Poisson process with drift*

$$\zeta_0(t) := \sum_{k=1}^{\nu_0(t)} \alpha_k^0 + tqa_0, \quad t \geq 0. \tag{3.1}$$

The distribution function $\Phi^0(u)$ *of the* iid *random variables* α_k^0, $k \geq 0$, *is defined on the measure-determining class* $C_3(\mathbb{R})$ *of functions* g *by the relation*

$$\mathbb{E}g(\alpha_k^0) = \int_{\mathbb{R}} g(u) \Phi^0(du) = \hat{\Phi}(g)/\hat{\Phi}(1), \quad g \in C_3(\mathbb{R}), \tag{3.2}$$

where

$$\hat{\Phi}(g) := \int_E \rho(dx)\Phi_x(g), \quad \hat{\Phi}(1) := \int_E \rho(dx)\Phi_x(1). \tag{3.3}$$

The counting Poisson process $\nu_0(t)$ is defined by the intensity

$$q_0 := q\hat{\Phi}(1). \tag{3.4}$$

The drift parameter a_0 is defined by

$$a_0 = \hat{a} - \hat{\Phi}(1)\mathbb{E}\alpha_1^0, \quad \hat{a} := \int_E \rho(dx)a(x). \tag{3.5}$$

The following corollary concerns the case where the state space E is finite.

COROLLARY 1 *The increment process (2.2) with a finite number of jump values:*

$$\mathbb{P}(\alpha_k^\varepsilon(x) = a_m) = \varepsilon p_m(x), \quad 1 \le m \le M,$$
$$\mathbb{P}(\alpha_k^\varepsilon(x) = \varepsilon a_0) = 1 - \varepsilon p_0(x), \tag{3.6}$$
$$p_0(x) = \sum_{m=1}^M p_m(x),$$

converges weakly to the compound Poisson process (3.1) determined by the distribution function of jumps:

$$\mathbb{P}(\alpha_k^0 = a_m) = p_m^0, \quad 1 \le m \le M,$$

$$p_m^0 = \hat{p}_m/\hat{p}_0, \quad \hat{p}_m = \int_E \rho(dx)p_m(x), \quad 1 \le m \le M. \tag{3.7}$$

The intensity of the counting Poisson process $\nu_0(t)$, $t \ge 0$, is defined by

$$q_0 := q\hat{p}_0, \tag{3.8}$$

and the drift parameter a_0 is given in (3.6).

Example. Let us assume that the ergodic process $x(t), t \ge 0$ takes values in $E = \{1,2\}$, and has generator matrix Q,

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$$

The transition matrix of the embedded Markov chain is

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Thus, the stationary distributions of $(x(t))$ and (x_n) are respectively:

$$\pi = \left(\frac{\mu}{\lambda + \mu}, \frac{\lambda}{\lambda + \mu}\right), \qquad \rho = \left(\frac{1}{2}, \frac{1}{2}\right).$$

Now, suppose that, for each $\varepsilon > 0$, the random variables $\alpha_k^\varepsilon(x), x = 1, 2, \ k \geq 1$ take values in $\{\varepsilon a_0, a_1\}$ with probabilities depending on the state x, i.e., $\Phi_x^\varepsilon(\varepsilon a_0) = \mathbb{P}(\alpha_k^\varepsilon = \varepsilon a_0) = 1 - \varepsilon p_x$ and $\Phi_x^\varepsilon(a_1) = \mathbb{P}(\alpha_k^\varepsilon = a_1) = \varepsilon p_x$, for $x \in E$.

We have

$$\int g(u)\Phi_x^\varepsilon(du) = \varepsilon[g(a_1)p_x + \theta_g^\varepsilon(x)],$$

where $\theta_g^\varepsilon(x) := \varepsilon a_0^2 g(\varepsilon a_0)/\varepsilon^2 a_0^2 = \varepsilon a_0^2 \cdot o(1) = o(\varepsilon)$, for $\varepsilon \to 0$, and

$$\int u\Phi_x^\varepsilon(du) = \varepsilon[(a_0 + a_1 p_x) + \theta^\varepsilon(x)],$$

where $\theta^\varepsilon(x) = -\varepsilon a_0 p_x$.

For the limit process, we have $\mathbb{P}(\hat{a}^0 = a_1) = 1$, thus

$$A^0(t) = qa_0 t + a_1 \nu^0(t),$$

with $\mathbb{E}\nu^0(t) = q_0 t, q = \lambda + \mu, q_0 = q\hat{p}_0 = q(p_1 + p_2)/2$.

Let us now take: $\lambda = \mu = 0.01; p_1 = 0.5; p_2 = 0.6; a_1 = 100; a_0 = -2; \varepsilon = 0.1$. Then we get $q_0 = 0.0165$, and figure 1 gives two trajectories in the time interval $[0, 4500]$, one for the initial process and the other for the limit process.

13.4   Increment Process in an Asymptotic Split Phase Space

The switching Markov process $x^\varepsilon(t), \ t \geq 0$, is here considered in the series scheme with a small series parameter $\varepsilon > 0$, on an asymptotic split phase space:

$$E = \bigcup_{v \in V} E_v, \quad E_v \bigcap E_{v'} = \emptyset, \quad v \neq v', \qquad (4.1)$$

where $(V, \mathcal{V})$ is a compact measurable space. The case where V is a finite set is of particular interest in applications.

The generator is given by the relation

$$Q^\varepsilon \varphi(x) = \int_E Q_\varepsilon(x, dy)[\varphi(y) - \varphi(x)]. \qquad (4.2)$$

The transition kernel Q_ε has the following representation

$$Q_\varepsilon(x, B) = q(x)P^\varepsilon(x, B) = Q(x, B) + \varepsilon Q_1(x, B), \qquad (4.3)$$
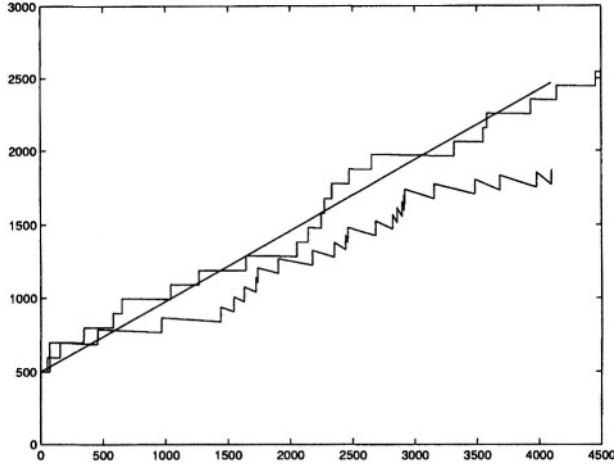
Figure 1. Trajectories of initial and limit processes, and of the drift.

with the stochastic kernel P^ε representation

$$P^\varepsilon(x, B) = P(x, B) + \varepsilon P_1(x, B). \tag{4.4}$$

The stochastic kernel $P(x, B)$ is linked with the split phase space (4.1) as follows

$$P(x, E_v) = 1_v(x) := \begin{cases} 1 & x \in E_v \\ 0 & x \notin E_v. \end{cases} \tag{4.5}$$

In the sequel we suppose that the kernel P_1 is of bounded variation, i.e.,

$$|P_1|(x, E) < +\infty. \tag{4.6}$$

According to (4.4) and (4.5), the Markov process $x^\varepsilon(t)$, $t \geq 0$, spends a long time in every class E_v and the probability of transition from one class to another is in $O(\varepsilon)$.

The phase merging scheme [Korolyuk, 1999] is realized under the condition that the support Markov process $x^0(t)$, $t \geq 0$, defined by the kernel $Q(x, dy) = q(x)P(x, dy)$ is uniformly ergodic in every class E_v, $v \in V$, with the stationary distributions

$$\pi_v(dx)q(x) = q_v\rho_v(dx), \quad q_v := \int_{E_v} \pi_v(dx)q(x). \tag{4.7}$$

Let us define the merged function

$$v(x) = v, \quad x \in E_v. \tag{4.8}$$

By the phase merging scheme [Korolyuk, 1999], the merged Markov process converges weakly

$$v(x^\varepsilon(t/\varepsilon)) \Longrightarrow \hat{x}(t), \quad t \geq 0, \text{ as } \varepsilon \to 0, \tag{4.9}$$

to the merged Markov process $\hat{x}(t)$, $t \geq 0$, defined on the merged phase space V by the generating kernel

$$\hat{Q}(v, B_\Gamma) = \int_{E_v} \pi_v(dx) Q(x, B_\Gamma), \quad B_\Gamma = \bigcup_{v \in \Gamma} E_v, \ \Gamma \in \mathcal{V}. \tag{4.10}$$

The counting process of jumps, noted $\hat{\nu}(t)$, can be obtained as the following limit [Korolyuk, 1995]

$$\varepsilon \nu^\varepsilon(t/\varepsilon) \Longrightarrow \hat{\nu}(t), \text{ as } \varepsilon \to 0, \ t \geq 0.$$

THEOREM 2 *Under the Assumptions A2-A5, in the phase merging scheme the increment process with Markov switching in series scheme*

$$\xi^\varepsilon(t) := \sum_{k=1}^{\nu^\varepsilon(t/\varepsilon)} \alpha_k^\varepsilon(x_k^\varepsilon), \quad t \geq 0, \tag{4.11}$$

converges weakly to the additive semimartingale $\zeta_0(t)$, $t \geq 0$, which is defined by its predictable characteristics,

$$B(t) = \int_0^t b(\hat{x}(s))ds, \quad b(v) = q_v \hat{a}(v), \quad \hat{a}(v) := \int_{E_v} \rho_v(dx) a(x); \tag{4.12}$$

the modified second characteristic $\tilde{C}^\varepsilon$ converges to

$$\hat{C}(t) = \int_0^t \hat{C}(\hat{x}(s))ds,$$

where

$$\hat{C}(v) = q_v \int_{E_v} \rho_v(dx) C_0(x), \quad v \in V \quad and \quad C_0(x) = \int_{\mathbf{R}} u^2 \Phi_x(du). \tag{4.13}$$

And the predictable measure is

$$\nu_t(g) = \int_0^t \tilde{\Phi}_{\hat{x}(s)}(g)ds, \quad \tilde{\Phi}_v(g) = q_v \hat{\Phi}_v(g), \tag{4.14}$$

where

$$\hat{\Phi}_v(g) := \int_{E_v} \rho_v(dx) \Phi_x(g).$$

The semimartingale $\zeta_0(t),\ t \geq 0,$ with predictable characteristics (4.12) and (4.14), can be represented in the following form

$$\zeta_0(t) = \int_0^t A^0(ds; \hat{x}(s)),\qquad\qquad (4.15)$$

or, in the equivalent increment form

$$\zeta_0(t) = \sum_{k=1}^{\hat{\nu}(t)} A^0_{\hat{x}_{k-1}}(\hat{\theta}_k) + A^0_{\hat{x}(t)}(\hat{\gamma}(t)).\qquad\qquad (4.16)$$

The compound Poisson processes $A^0_v(t)$ are defined by the generators

$$A(v)\varphi(u) = q^0_v \int_{\mathbf{R}} [\varphi(u+z) - \varphi(u)]\Phi_v(dz),$$

and $\nu^0_v(t)$ are the counting Poisson processes characterized by the intensity $q^0_v = q_v \hat{\Phi}_v(1).$ It is also defined explicitly by

$$A^0_v(t) = \sum_{\ell=1}^{\nu^0_v(t)} \alpha^0_{v\ell} + t q_v a^0_v, \quad v \in V,$$

for fixed $v \in V,$ where $\alpha^0_{v\ell}, \ell \geq 1,$ are iid r.v.s with common distribution function defined by the measure

$$\Phi^0_v(g) = \hat{\Phi}_v(g)/\hat{\Phi}_v(1).$$

The drift parameter is given by

$$a^0_v = \hat{a}(v) - \hat{\Phi}_v(1)\mathbb{E}\alpha^0_{v1}.$$

In applications, the limit semimartingale (4.15) can be considered in the following form

$$\zeta_0(t) = \sum_{k=1}^{\hat{\nu}(t)} \hat{\theta}_k c(x_{k-1}) + \hat{\gamma}(t)c(\hat{x}(t)) + \mu_0(t),\qquad\qquad (4.17)$$

where $\mu_0(t)$ is a martingale fluctuation. The predictable term in (4.17) is a linear deterministic drift between jumps of the merged switching Markov process $\hat{x}(t),\ t \geq 0.$

13.5   Continuous Additive Functional

We consider an additive jump functional with semi-Markov switching in a Poisson approximation scheme depending on the small series parameter $\varepsilon > 0$, namely

$$\xi^\varepsilon(t) = \int_0^t \eta^\varepsilon(ds; x(s/\varepsilon)), \tag{5.1}$$

where $\eta^\varepsilon(t; x)$, $t \geq 0, x \in E, \varepsilon > 0$, is a family of Markov jump processes in the series scheme defined by the generators

$$\mathbb{\Gamma}^\varepsilon(x)\varphi(u) = \varepsilon^{-1} \int_{\mathbf{R}^d} [\varphi(u+v) - \varphi(u)]\Gamma_\varepsilon(dv; x), \ x \in E.$$

Assumptions (C)

(C1:) The switching semi-Markov process $x(t)$, $t \geq 0$, is uniformly ergodic with the stationary distribution

$$\pi(dx) = \rho(dx)m(x)/m, \tag{5.2}$$

$$m(x) := \mathbb{E}\theta_x = \int_0^\infty \overline{G}_x(t)dt, \quad m := \int_E \rho(dx)m(x), \tag{5.3}$$

$$\rho(B) = \int_E \rho(dx)P(x,B), \quad \rho(E) = 1. \tag{5.4}$$

(C2:) Approximation of the mean jump

$$a_\varepsilon(x) = \int_{\mathbf{R}} v\Gamma_\varepsilon(dv; x) = \varepsilon[a(x) + \theta^\varepsilon(x)] \tag{5.5}$$

and $a(x)$ is bounded, i.e., $|a(x)| \leq a < +\infty$.

(C3:) Poisson,approximation condition

$$\Gamma_g^\varepsilon(x) = \int_{\mathbf{R}} g(v)\Gamma_\varepsilon(dv; x) = \varepsilon[\Gamma_g(x) + \theta_g^\varepsilon(x)] \tag{5.6}$$

for all $g \in C_3(\mathbb{R})$, and the kernel $\Gamma_g(x)$ is bounded for all $g \in C_3(\mathbb{R})$, i.e.,

$$|\Gamma_g(x)| \leq \Gamma_g.$$

The negligible terms in (5.5) and (5.6) satisfy the conditions

$$\sup_{x \in E} |\theta^\varepsilon(x)| \to 0 \quad \text{and} \quad \sup_{x \in E} |\theta_g^\varepsilon(x)| \to 0, \quad as \ \varepsilon \to 0, \tag{5.7}$$

for all $g \in C_3(\mathbb{R})$.

(C4:) Uniform square-integrability

$$\lim_{c \to \infty} \sup_{x \in E} \int_{|v| \geq c} vv^*\Gamma(dv; x) = 0, \tag{5.8}$$

where the kernel $\Gamma(dv; x)$ is defined on the measure-determining class $C_3(\mathbb{R})$ by the relation

$$\Gamma_g(x) = \int_{\mathbb{R}} g(v)\Gamma(dv; x), \quad g \in C_3(\mathbb{R}). \tag{5.9}$$

(C5:) Cramér's condition

$$\int_0^\infty e^{hs}\overline{G}_x(s)ds < H < \infty. \tag{5.10}$$

THEOREM 3 *Under Assumptions C1-C5, the additive functional (5.1) converges weakly to the CPPD $\xi_0(t)$, $t \geq 0$, defined by the generator*

$$\hat{\mathbb{\Gamma}}\varphi(u) = \hat{a}\varphi'(u) + \int_{\mathbb{R}} [\varphi(u + v) - \varphi(u) - v\varphi'(u)]\hat{\Gamma}(dv), \tag{5.11}$$

where

$$\hat{a} = \int_E \pi(dx)a(x), \tag{5.12}$$

and

$$\hat{\Gamma}(dv) = \int_E \pi(dx)\Gamma(dv; x). \tag{5.13}$$

The additive jump functional (5.1) in the Poisson approximation scheme can be considered with the semi-Markov switching in the split state space (see Section 4, Theorem 2).

Due to both the representation (5.11)–(5.13) of the limit generator, and the approximation conditions C2 and C3, the small jumps of the initial functional are transformed into the deterministic drift $\hat{U}_0(t) = \hat{a}_0 t$,

$$\hat{a}_0 = \hat{a} - \hat{b}, \quad \hat{b} := \int_{\mathbb{R}} v\hat{\Gamma}(dv). \tag{5.14}$$

The big jumps of the initial functional (5.1) are distributed following the averaged distribution function

$$\hat{F}(dv) := \hat{\Gamma}(dv)/\hat{\Gamma}(\mathbb{R}), \tag{5.15}$$

with the intensity of jump moments $\hat{\gamma} := \hat{\Gamma}(\mathbb{R})$. The limit Markov process has the representation $\xi^0(t) = U^0(t) + \zeta^0(t)$, where the Markov process $\zeta^0(t)$ has the following generator

$$\hat{\Gamma}_0\varphi(u) = \hat{\gamma} \int_{\mathbb{R}} [\varphi(u+v) - \varphi(u)]\hat{F}(dv).$$

13.6   Scheme of Proofs

Let us give here the main steps of the proofs in the case of the continuous additive functional $\xi^\varepsilon(t)$ given in (2.3).

The weak convergence for additive functionals with semi-Markov switching is considered here as in our previous paper [Korolyuk, 2002] in the setting of the books by Jacod & Shiryaev [Jacod, 1987] and Ethier & Kurtz [Ethier, 1986].

The semi-Markov switching requires new approach based on the compensative operator of the Markov renewal process, see [Sviridenko, 1989]. The additive jump functional (5.1) is first considered as an additive semimartingale defined by its predictable characteristics [Jacod, 1987; Liptser, 1989; Liptser, 1991; Borovskikh, 1997; Çinlar, 1980].

The main steps of proofs include: the construction of the predictable characteristics of the semimartingale $\xi^\varepsilon(t)$, the construction of compensative operator of the extended Markov renewal process, convergence of predictable characteristics, and identification of the limit process.

LEMMA 1 *Under the assumptions of Theorem 3, the predictable character-istics $(B^\varepsilon(t), C^\varepsilon(t), \gamma^\varepsilon(t))$ (see [Jacod, 1987], Theorem VI.3.31) of the semi-martingale*

$$\xi^\varepsilon(t) = \int_0^t \eta^\varepsilon(ds; x(s/\varepsilon)) + \xi_0^\varepsilon, \tag{6.1}$$

are defined by the following relations:

$$B^\varepsilon(t) = \varepsilon\Big[\int_0^{t/\varepsilon} a(x(s))ds + \theta_b^\varepsilon(t)\Big], \quad t \geq 0. \tag{6.2}$$

The modified second characteristic is

$$C^\varepsilon(t) = \varepsilon\Big[\int_0^{t/\varepsilon} C(x(s))ds + \theta_c^\varepsilon(t)\Big], \quad t \geq 0. \tag{6.3}$$

The predictable measure is

$$\gamma_g^\varepsilon(t) = \varepsilon\Big[\int_0^{t/\varepsilon} \Gamma_g(x(s))ds + \theta_\gamma^\varepsilon(t)\Big], \quad t \geq 0, \tag{6.4}$$

where $C(x) = \int_{\mathbf{R}^d} vv^* \Gamma(dv; x)$. v^* *means the transpose of vector* v.

Note that θ_b^ε, θ_c^ε, and $\theta_\gamma^\varepsilon$ satisfy the negligible condition (see [Jacod, 1987], Lemma VI.3.31],

$$\sup_{t < T} |\theta_\cdot^\varepsilon(t)| \to 0, \quad \text{as } \varepsilon \to 0, \quad \text{for all } T > 0.$$

In what follows, it is sufficient to study only the convergence of $(B_0^\varepsilon(t), C_0^\varepsilon(t), \gamma_0^\varepsilon(t; g))$, where

$$B_0^\varepsilon(t) = \int_0^t a(x(s/\varepsilon)) ds, \quad t \geq 0, \tag{6.5}$$

$$C_0^\varepsilon(t) = \int_0^t C(x(s/\varepsilon)) ds, \quad t \geq 0, \tag{6.6}$$

$$\gamma_0^\varepsilon(t; g) = \int_0^t \Gamma_g(x(s/\varepsilon)) ds, \quad t \geq 0. \tag{6.7}$$

In the sequel the process $A^\varepsilon(t)$ will denote one of the above predictable characteristics $B_0^\varepsilon(t), C_0^\varepsilon(t), \gamma_0^\varepsilon(t)$.

The following auxiliary processes will be used:

$$\nu^\varepsilon(t) := \max\{n : \tau_n^\varepsilon \leq t\} = \max\{n : \tau_n \leq t/\varepsilon\},$$

$$\tau^\varepsilon(t) = \tau_{\nu^\varepsilon(t)}^\varepsilon, \quad \tau_+^\varepsilon(t) = \tau_{\nu_+^\varepsilon(t)}^\varepsilon, \quad \nu_+^\varepsilon(t) = \nu^\varepsilon(t) + 1,$$

$$\theta_-^\varepsilon(t) = t - \tau^\varepsilon(t), \quad \theta_+^\varepsilon(t) = \tau_+^\varepsilon(t) - t.$$

The extended Markov renewal process is considered as a three component Markov chain

$$A_n^\varepsilon = A^\varepsilon(\tau_n^\varepsilon), \quad x_n^\varepsilon, \quad \tau_n^\varepsilon, \quad n \geq 0, \tag{6.8}$$

where $x_n^\varepsilon = x^\varepsilon(\tau_n^\varepsilon)$, $x^\varepsilon(t) := x(t/\varepsilon)$ and $\tau_{n+1}^\varepsilon = \tau_n^\varepsilon + \varepsilon\theta_n^\varepsilon, n \geq 0$, and

$$\mathbb{P}(\theta_{n+1}^\varepsilon \leq t \mid x_n^\varepsilon = x) = G_x(t) = \mathbb{P}(\theta_x \leq t). \tag{6.9}$$

We are using here the notion of compensative operator introduced by Wentzel & Sviridenko (see [Sviridenko, 1989]).

DEFINITION 1 *([Sviridenko, 1989]) The compensative operator* $\mathbb{L}^\varepsilon$ *of the extended Markov renewal process (6.8) is defined by the following relation*

$$\mathbb{L}^\varepsilon \varphi(u, x, t) = \{\mathbb{E}[\varphi(A_1^\varepsilon, x_1^\varepsilon, \tau_1^\varepsilon) - \varphi(u, x, t) \mid \mathcal{F}_t^\varepsilon]\} / \varepsilon m(x), \tag{6.10}$$

where $m(x) = \mathbb{E}\theta_x = \int_0^\infty \overline{G}_x(t) dt$,

$$\mathcal{F}_t^\varepsilon := \sigma(A^\varepsilon(s), x^\varepsilon(s), \tau^\varepsilon(s); 0 \leq s \leq t). \tag{6.11}$$

Let $\mathcal{A}_t(x)$, $t \geq 0$, $x \in E$, be a family of semigroups determined by the generators

$$A(x)\varphi(u) = a(x)\varphi'(u). \tag{6.12}$$

LEMMA 2 *The compensative operator (6.10) of the extended Markov renewal process (6.8) can be defined by the relation*

$$\mathbb{L}^\varepsilon \varphi(u, x, t) =$$
$$\left[\int_0^\infty G_x(ds) \mathcal{A}_{\varepsilon s}(x) \int_E P(x, dy) \varphi(u, y, t + \varepsilon s) - \varphi(u, x, t) \right] / \varepsilon m(x). \tag{6.13}$$

The proof of Lemma 2 follows directly from Definition 1.

LEMMA 3 *The extended Markov renewal process (6.8) is characterized by the martingale*

$$\mu_{n+1}^\varepsilon = \varphi(A_{n+1}^\varepsilon, x_{n+1}^\varepsilon, \tau_{n+1}^\varepsilon) - \sum_{k=0}^n \varepsilon \theta_{k+1}^\varepsilon \mathbb{L}^\varepsilon \varphi(A_k^\varepsilon, x_k^\varepsilon, \tau_k^\varepsilon), \quad n \geq 0. \tag{6.14}$$

In what follows the martingale property will be used for the process

$$\zeta^\varepsilon(t) = \varphi \left(A^\varepsilon(\tau_+^\varepsilon(t)), x^\varepsilon(\tau_+^\varepsilon(t)), \tau_+^\varepsilon(t) \right) -$$
$$\int_0^{\tau_+^\varepsilon(t)} \mathbb{L}^\varepsilon \varphi \left(A^\varepsilon(\tau^\varepsilon(s)), x^\varepsilon(s), \tau^\varepsilon(s) \right) ds, \tag{6.15}$$

where $\tau_+^\varepsilon(t) := \tau_{\nu_+^\varepsilon(t)}$, $\nu_+^\varepsilon(t) := \nu^\varepsilon(t) + 1$.

Note that the following relations hold:

$$\zeta^\varepsilon(\tau_n) = \mu_{n+1}^\varepsilon, \quad n \geq 0, \tag{6.16}$$

and

$$\zeta^\varepsilon(t) = \zeta^\varepsilon(\tau^\varepsilon(t)), \quad \text{for} \quad \tau^\varepsilon(t) \leq t < \tau_+^\varepsilon(t). \tag{6.17}$$

The random numbers $\nu_+^\varepsilon(t) = \nu^\varepsilon(t) + 1$ are Markov moments for

$$\mathcal{F}_n^\varepsilon = \sigma(A_k^\varepsilon, x_k^\varepsilon, \tau_k^\varepsilon; 0 \leq k \leq n).$$

LEMMA 4 *The process (6.17) has the martingale property*

$$\mathbb{E}[\zeta^\varepsilon(t) - \zeta^\varepsilon(s) \mid \mathcal{F}_s^\varepsilon] = 0, \quad \text{for} \quad 0 \leq s < t \leq T. \tag{6.18}$$

Note that the process $\zeta^\varepsilon(t)$, $t \geq 0$, is not a martingale since it is not $\mathcal{F}_t^\varepsilon$-adapted. The next lemma is basic in the proof of the *compact containment condition* for the additive functionals $A^\varepsilon(t)$, $t \geq 0$. (Compare with Lemma 3.2 [Ethier, 1986]).

LEMMA 5 *The process*

$$\zeta_c^\varepsilon(t) = e^{-c\tau_+^\varepsilon(t)}\varphi\left(A^\varepsilon(\tau_+^\varepsilon(t)), x^\varepsilon(\tau_+^\varepsilon(t)), \tau_+^\varepsilon(t)\right)$$
$$+ \int_0^{\tau_+^\varepsilon(t)} \left[e^{-cs}c\varphi\left(A^\varepsilon(\tau_+^\varepsilon(s)), x^\varepsilon(\tau_+^\varepsilon(s)), \tau_+^\varepsilon(s)\right)\right.$$
$$\left. - e^{-c\tau^\varepsilon(s)}\mathbb{L}^\varepsilon\varphi\left(A^\varepsilon(\tau^\varepsilon(s)), x^\varepsilon(s), \tau^\varepsilon(s)\right)\right]ds \qquad (6.19)$$

has the martingale property for every $c \in \mathbb{R}$, i.e.,

$$\mathbb{E}[\zeta_c^\varepsilon(t) - \zeta_c^\varepsilon(s) \mid \mathcal{F}_s^\varepsilon] = 0, \quad for \quad 0 \le s < t \le T. \qquad (6.20)$$

The algorithm of Poisson, approximation given in Theorem 1 provides the asymptotic representation of the compensative operator.

LEMMA 6 *The compensative operator (6.13) applied to function $\varphi \in C^2(\mathbb{R}) \times B(E)$ has the asymptotic representation*

$$\mathbb{L}^\varepsilon\varphi(u, x) = \varepsilon^{-1}Q\varphi(u, x) + A(x)P\varphi(u, x) + \varepsilon\theta^\varepsilon(x)P\varphi(u, x), \qquad (6.21)$$

where

$$Q\varphi(\cdot, x) = q(x)\int_E P(x, dy)[\varphi(\cdot, y) - \varphi(\cdot, x)], \quad q(x) := 1/m(x), \qquad (6.22)$$

$$A(x)\varphi(u, \cdot) = a(x)\varphi_u'(u, \cdot). \qquad (6.23)$$

And the negligible operator is defined as follows

$$\theta^\varepsilon(x)\varphi(u, \cdot) = A^2(x)A^\varepsilon(x)\varphi(u, \cdot), \qquad (6.24)$$

where

$$A^\varepsilon(x) = \int_0^\infty \mathcal{A}_{\varepsilon s}(x)\overline{G}_x^{(2)}(s)ds, \qquad (6.25)$$

$$\overline{G}_x^{(2)}(s) := \int_s^\infty \overline{G}_x(t)dt. \qquad (6.26)$$

Note that the remaining term in (6.21) is computed by using the relation

$$A^\varepsilon(x) = \int_0^\infty G_x(ds)A_s^\varepsilon(x) = \int_0^\infty \mathcal{A}_{\varepsilon s}(x)\overline{G}_x^{(2)}(s)ds. \qquad (6.27)$$

LEMMA 7 *A solution of the singular perturbation problem*

$$\mathbb{L}^\varepsilon[\varphi(u) + \varepsilon\varphi_1(u, x)] = L\varphi(u) + \varepsilon\theta_0^\varepsilon(x)\varphi(u) \qquad (6.28)$$

is given by the generator

$$L\varphi(u) = \hat{a}\varphi'(u). \qquad (6.29)$$

The negligible term in (6.28) is represented as follows

$$\theta_0^\varepsilon(x)\varphi(u) = [(A(x) + \varepsilon\theta^\varepsilon(x))R_0\tilde{A}(x) + \varepsilon\theta^\varepsilon(x)]\varphi(u). \qquad (6.30)$$

The following compact containment condition together with the submartingale condition (see [Korolyuk, 2000]) provides the compactness of the family $(A^\varepsilon(t), \ t \geq 0, \ \varepsilon > 0)$.

LEMMA 8 *The family of processes $(A^\varepsilon(t), \ t \geq 0, \ 0 < \varepsilon \leq \varepsilon_0)$ with bounded initial value $\mathbb{E}\,|A^\varepsilon(0)| \leq b < +\infty$, satisfies the compact containment condition (see [Ethier, 1986])*

$$\lim_{\ell \to \infty} \sup_{0 < \varepsilon \leq \varepsilon_0} \mathbb{P}\left(\sup_{0 \leq t \leq T} |A^\varepsilon(t)| \geq \ell\right) = 0. \qquad (6.31)$$

The completion of the proof of theorem is realized by the scheme described in our previous paper [Korolyuk, 2000], by using Theorem VIII.2.18, in Jacod & Shiryaev [Jacod, 1987].

■

Acknowledgments

References

D. Aldous (1989). *Probability Approximations via the Poisson Clumping Heuristic,* Springer-Verlag, New York.

V.V. Anisimov (1995). Switching processes: averaging principle, diffusion approximation and applications, *Acta Aplicandae Mathematica,* **40,** 95–141.

A.D. Barbour, L. Holst and S. Janson (1992). *Poisson Approximation,* Clarendon Press, Oxford.

A.D. Barbour and O. Chryssaphinou (2002). Compound Poisson approximation: a user's guide, *Ann. Appl. Probab.,* **11,** No 3, 964–1002.

P. Billingsley (1968). *Convergence of Probability Measures,* J. Wiley & Sons, New York.

Y. V. Borovskikh and V. S. Korolyuk (1997). *Martingale Approximation,* VSP, Utrecht, The Netherlands.

E. Çinlar, J. Jacod, P. Protter and M.J. Sharpe (1980). Semimartingale and Markov processes, *Z. Wahrschein. verw. Gebiete,* **54,** 161–219.

S.N. Ethier and T.G. Kurtz (1986). *Markov Processes: Characterization and convergence,* J. Wiley & Sons, New York.

A. Gut (1988). *Stopped Random Walks,* Springer-Verlag, N.Y..

J. Jacod and A.N. Shiryaev (1987). *Limit Theorems for Stochastic Processes,* Springer-Verlang, Berlin.

C. Kluppelberg and T. Mikosch (1995). Explosive Poisson shot processes with applications to risk reserves, *Bernoulli,* **1,** (1 & 2), pp 125-147.

V.S. Korolyuk and V. V. Korolyuk (1999). *Stochastic Models of Systems,* Kluwer Academic Publishers.

V.S. Korolyuk and N. Limnios (1999a). A singular perturbation approach for Liptser's functional limit theorem and some extensions, *Theory Probab. and Math. Statist.,* **58**, 83–88.

V.S. Korolyuk and N. Limnios (1999b). Diffusion approximation of integral functionals in merging and averaging scheme, *Theory Probab. and Math. Statist.,* **59***, pp 101–108.

V.S. Korolyuk and N. Limnios (2001). Diffusion approximation of integral functionals in double merging and averaging scheme, *Theory Probab. and Math. Statist.* **60**, pp 87–94.

V.S. Korolyuk and N. Limnios (2000). Poisson approximation of increment processes with Markov switching, *submitted.*

V.S. Korolyuk and N. Limnios (2000a). Evolutionary systems in an asymptotic split state space, in *Recent Advances in Reliability Theory: Methodology, Practice and Inference,* N. Limnios & M. Nikulin (Eds), Birkhäuser, Boston.

V.S. Korolyuk and N. Limnios (2000b). Average and diffusion approximation of evolutionary systems in an asymptotic split state space, *submitted.*

V.S. Korolyuk and N. Limnios (2000c). Diffusion approximation of evolutionary systems without balance condition, *submitted.*

V.S. Korolyuk, N. Limnios, (2001a). Poisson approximation of integral functionals of semi-Markov processes, 10th ASMDA International Symposium, Compiègne, 12-15 June 2001.

V.S. Korolyuk and N. Limnios (2002). Poisson approximation of homogeneous stochastic additive functionals with semi-Markov switching, *Theory Probab. and Math. Statist.* **64**, pp 75–84.

V.S. Korolyuk and N. Limnios (2002a). Increment processes and its stochastic exponential with Markov switching in Poisson approximation scheme, *Computers Mathematics Appl.,* (to appear).

Korolyuk, V.S. and Swishchuk, A. (1995). *Evolution of System in Random Media,* CRC Press.

N. Limnios and **G. Oprişan** (2001). *Semi-Markov Processes and Reliability,* Birkhäuser, Boston.

R. S. Liptser (1994). The Bogolubov averaging principle for semimartingales, Proceedings of the Steklov Institute of Mathematics, Moscow, N4, pp 1–12.

R. S. Liptser and A. N. Shiryayev (1989). *Theory of Martingales,* Kluwer Academic Publishers, Dordrecht, The Netherlands.

R. Sh. Liptser and A. N. Shiryayev (1991). Martingales and limit theorems for stochastic processes, in *Encyclopaedia of Mathematical Sciences. Probability Theory III,* Yu. Prokhorov and A.N. Shiryaev (Eds), Springer, pp. 158–247.

S. Osaki (1985). *Stochastic System Reliability Modeling,* Word Scientific, Singapore.

E. Parzen (1999). *Stochastic Processes,* SIAM Classics, Philadelphia.

N.U. Prabhu (1980). *Stochastic Storage Processes,* Springer-Verlag, Berlin.

D.W. Stroock and S.R.S. Varadhan (1979), *Multidimensional Diffusion Processes,* Springer-Verlag, Berlin.

M.N. Sviridenko (1989). Martingale approach to limit theorems for semi-Markov processes. *Theory of Probability and Applications,* N 3, pp 540-545.

This page intentionally left blank

PENALIZED MODEL SELECTION FOR ILL-POSED LINEAR PROBLEMS

Carenne Ludeña
Instituto Venezolano de Investigaciones Científicas.
cludena@ivic.ve

Ricardo Ríos
Universidad Central de Venezuela. Facutad de Ciencias. Escuela de Matemáticas,
rrios@euler.ciens.ucv.ve

Abstract   In this article we review the problem of discretization-regularization for inverse linear ill-posed problems from a statistical point of view. We discuss the problem in the context of adaptive model selection and relate these results to Bayesian estimation.

Keywords: Model selection, penalized estimation, Rosenthal type inequalities, ill-posed problems.

14.1   Introduction

In many situations we require estimating a certain function $f \in H$, a given Hilbert space, based on indirect observations $y_i = (Af)(x_i) + \eta_i, i = 1, ..., n,$ when A is an ill posed operator. That is, when A does not have an inverse or when its inverse is not continuous.

Here η_i is assumed to be a zero mean i.i.d. sequence of generally non bounded random variables which accounts for a perturbation of the true value $A(f)(x_i)$ and $x_i, i = 1, \ldots, n$ is assumed to be a fixed set of observation points.

As A is ill posed, searching for the solution f based on the noise corrupted observations $y_i, i = 1, \ldots, n$ is useless. It is usual to look instead at solutions that not only adjust to the observations but are *regular* as defined by a given functional $J(f)$. Thus we search for the solution f of the minimization problem

$$\min_H (\|y - A(f)(x)\|_{(n)} + J(\lambda f)) \tag{1.1}$$

here $\| \cdot \|_{(n)}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - A(f)(x_i))^2$, and λ is a *regularization parameter* which must be chosen according to some criteria. Typically methods such as the L-curve methods [Engl & Grever, 1994] or Morozov's discrepancy principle (see for example [Frommer & Maass, 1999]) in the least squares case, or statistical methods such as PMSE (Predictive mean square error) or cross-validation are used (for example see [O'sullivan, 1996]). Equation (1.1) also includes most estimation schemes based on entropy methods [Gamboa & Gassiat, 1997], [Gamboa, 1999].

In practice however (1.1) is hardly ever considered. Indeed solutions are seeked for in finite dimensional closed subspaces S_m ($dim(S_m) = d_m$) of H. Normally, the restricted problem is also ill conditioned and must be regularized. This yields a sequence of closed subspaces S_m indexed by $m \in M_n$, a collection of index sets, and a sequence of regularization parameters λ_m. An important problem is thus how to choose a "correct" subspace S_m based on the data and how to interpret the sequence of λ_m in such a choice.

We shall refer to the *penalized model selection* framework developed in a series of works by Birgé and Massart [Barron, Birgé & Massart, 1999] (see also [Birge & Massart 2001], [Birgé & Massart, 1998],[Massart, 2000]) based on the idea of sieves due originally to Grenander [Grenander, 1981]. Related ideas are also developed by Vapnik [Vapnik, 1998] in his Structural Risk Minimization setting. This is a statistical point of view and solution choice is compared to optimal rate estimation over certain classes of functions.

Basically, the idea is to penalize high dimensional spaces. Intuitively, estimation will be better if d_m is large, but then A will be harder to control (this will be true even if the operator is not ill posed). Penalization should be chosen in such a way as to obtain almost optimal results. That is, the chosen solution should be the (almost) best among all possible choices of subspaces S_m, for $m \in M_n$.

Many authors have addressed the problem of simultaneous discretization-projection and regularization for ill posed problems (see for example [Kilmer & O'Leary, 2001], [Maass et al, 2001], [Neubauer, 1998], [Solodky, 1999]). If regularization is done by projection (truncated S.V.D.), the problem is essentially that of determining a "good" subspace. This can be done by selecting a cutoff point or by threshold methods. As will be seen this amounts to an appropriate selection of a penalization term for the dimension. Choosing the right subspace will be called model selection.

If regularization is done by Tikhonov [Tikhonov & Arsenin, 1998], a problem cited by many authors is whether the appropriate regularization parameter for the projected solution is also appropriate for non projected one. From a model selection point of view, the problem is stated as minimizing a certain contrast function over a certain parameter space for each d_m dimensional

subspace and then choosing the best subspace, corrected by the appropriate penalization term.

The main goal of this article is to present simultaneous discretization-regularization in an adaptive fashion, based on the ideas of model selection and to interpret solutions in a Bayesian point of view, considering a prior distribution over the family of models and a suitable prior over all possible solutions in a given model.

We start with a short review of the main ideas of model selection, as developed by Birgé and Massart. We then describe minimax estimation bounds for ill posed problems, and finally apply model selection techniques to ill posed problems.

In Section 14.5 penalized minimum contrast estimation is presented in a Bayesian framework.

In the last Section we discuss a different choice of the regularization functional, namely $J(\cdot)$ such as $J(f) = \|f\|_1$ as proposed by Aluffi-Pentini et al. [Aluffi-Pentini et al, 1999]. It can be seen that this estimator, for an L^2 loss function is actually soft thresholding [Kaliffa & Mallat, 2001].

14.2 Penalized model selection [Barron, Birgé & Massart, 1999]

Consider the direct problem of estimating a function $f \in L^2(T, \mu)$ based on observations

$$y_i = f(x_i) + \eta_i, \quad i = 1, \ldots, n$$

where as before we assume $x_i, i = 1, \ldots, n$ to be a fixed design (actually we could consider the more general problem of the white noise framework). Although not specified, usually we associate the above problem to an orthonormal basis $\{\phi_j\}_{j \in \mathbb{Z}}$. The problem is then analogous to selecting the correct parameters of function f over this basis. This is usually done by minimizing a certain discrepancy functional l over the n observations. Function l is called the loss function and $\gamma_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(y_i - f(x_i))$, is called the empirical risk function, so actually the idea is to find $\hat{f}$ that minimizes the empirical risk. If we assume $f(x) = \sum_{j \in m} f_j \phi_j(x)$, the above strategy leads to $\hat{f} = \sum_{j \in m} \hat{f}_j \phi_j(x)$, with $\hat{f}_j = \frac{1}{n} \sum_{i=1}^{n} y_i \phi_j(x_i)$. If m is known beforehand then $\mathbb{E}\|f - \hat{f}\| = \frac{\sigma^2 m}{n}$, where $\sigma^2 = var(\zeta_i)$, for each $i = 1, \ldots, n$. What happens if we do no know m? One possibility is estimating $\hat{f}_m$ for m in a certain subset and compare $\mathbb{E}\|f - \hat{f}_m\| = \|f - \Pi_m f\| + \frac{\sigma^2 m}{n}$. As f is not necessarily equal to $\Pi_m(f)$ the first term controls this error. This is the typical bias-variance decomposition. If the function is sufficiently regular, for example if

$$\sum_j f_j^2 j^{2\alpha} \le Q,$$

then $\|f - \Pi_m f\|$ is known and it turns out that $\mathbb{E}\|f - \hat{f}_m\| = O((Qn^\alpha)^{2/(2\alpha+1)})$. This suggests searching m until this bound is obtained (if m is too big the variance term will be too big, if m is too small the bias term will be too big). However, if α is not known, underestimating this parameter will lead to a bad choice of m and the risk will be too big. If we overestimate α we will force the risk to be too small which is known as overfitting (we adjust our data only).

Adaptive model selection is a technique which penalizes the dimension of the estimating subset (considering a set $\mathcal{S}_n = \{S_m | m \in M_n\}$, with $dim(S_m) = d_m$), in such a way that if we choose $\hat{f}_{\hat{m}}$ by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} l(y_i - \hat{f}(x_i)) + pen(m),$$

for l a certain loss function, then, there exists a constant K such that

$$\mathbb{E}\|f - \hat{f}_{\hat{m}}\|^2 \leq C \inf_{m \in M_n} (\inf_{f_m \in S_m} \|f - f_m\|^2 + pen(m)) + \frac{K}{n}. \qquad (2.1)$$

Usually $pen(m) = \kappa L_m d_m / n$, where $\kappa > 1$ and L_m is a sequence which is incorporated in order to control the complexity of $\mathcal{S}_n$. It is chosen so that

$$\sum_{m \in M_n} e^{-L_m d_m} < K < \infty.$$

If the number of subsets S_m with equal dimension is small, i.e. if $\sup_j(\frac{1}{j} \log |\{m | d_m = j\}|) \leq \nu$, then it is enough to choose $L_m = L > \nu$.

If the number of subsets with equal dimension is big, for example in the problem of complete model selection, L_m must be chosen non constant. Indeed, following [Barron, Birgé & Massart, 1999] assume we choose among all subsets of the set $\{\phi_j\}_{j \in \{1,...,N\}}$. In this case the cardinality of all models with dimension d_m is equal to

$$\binom{N}{d_m} \leq (eN/d_m)^{d_m}$$

as the cited authors show in Lemma 6. This implies that a good choice is $L_m = c(1 + \log N)$. The authors further show that in fact this choice yields the hard threshold estimator of Donoho and Johnstone [Donoho, 1995].

Actually, model selection allows for much more general contrast functions γ_n based on the empirical distribution, which may yield the problem non linear. Think, for example, of the correct choice of a neural network, or maximum likelihood estimation for non Gaussian error distribution. The results cited in equation (2.1) are quite general and include these cases provided the contrast

and the family of subspaces S_m satisfy certain conditions (Theorem 7.1 [Barron, Birgé & Massart, 1999]).

An important issue is that equation (2.1) is non asymptotic and is useful if the choice of the penalization term yields optimal results, i.e. minmax estimation rates and constants. Based on these ideas we shall see that the bounds in (2.1) can be obtained in the ill-posed case, and compare penalized estimation to optimal linear and minimax estimation in this case.

As we mentioned in Section 1, we refer to the fixed point design problem. Optimal results for this problem based on adaptive model selection in the well posed case are given by [Baraud, 2000].

14.3   Minimax estimation for ill posed problems

Assume A to be a known, linear operator $A : H_1 \rightarrow H_2$, H_i Hilbert spaces with inner product $<,>_{H_i}$ and norm $\| \; \|_{H_i}$.

Our aim is estimating $f \in H_1$ given the set of indirect observations $(x_i, y_i)_{i=1}^n$, for a fixed point design $(x_i)_{i=1}^n$, which are assumed to follow the model

$$y_i = Af(x_i) + \eta_i. \tag{3.1}$$

Here $\{\eta_i\}_{i=1}^n$ is a centered and i.i.d sequences of r.v. with finite p^{th} moment and variance σ^2. As $f \in H_1$, we approximate $f(x) = \sum_{j\in m} f_j\phi_j(x)$ in terms of some orthonormal basis $\{\phi_j\}$ of H_1 ($M = |m|$ can be ∞), where $f_j = \langle f, \phi_j \rangle$ stand for the Fourier coefficients of f with respect to the given basis. The choice of a finite M in a data driven fashion is part of the problem we address here.

We assume also that there exists a basis $\{\psi_j\}$ of H_2 such that $(Af, \psi_j) = f_j b_j$, with $b_j > 0$ and $b_j \rightarrow 0$. This happens if, for example, A admits a Singular Value Decomposition.

Let M_n be a colection of index sets ($m \in M_n, m = \{j_1, \ldots, j_{d_m}\}$), and let $(S_m)_{M_n}$ be the sequence of closed linear subspaces of H_1, $S_m = \text{span}\{\phi_j, j \in m\}$, with dimension $d_m < \infty$.

We also need some notation concerning the fixed point setting. For $g, h \in H_2$, set $< g, h >_{(n)} = \frac{1}{n} \sum_{j=1}^n g(x_i)h(x_i)$ and $\|g\|_{(n)}^2 =< g, g >_{(n)}$. Also let

$$\Sigma_m = \left[\frac{1}{n} \sum_{i=1}^n \psi_j(x_i)\psi_k(x_i) \right]_{(j,k\in m)}$$

be the Gram matrix associated to $\{\psi_j\}$ over S_m. Set $B = Diag(b_k)_{k\in m}$ and define $A_m = B\Sigma_m B$. Let

$$\tilde{y}_k = \frac{1}{n} \sum_{i=1}^n y_i b_k \psi_k(x_i).$$

Finally set $\tilde{x} = A_m^{-1}(\tilde{y}_j)_{j \in m}$ and

$$\zeta = A_m^{-1}(\frac{1}{n}b_j \sum_{i=1}^n \eta_i \psi_j(x_i))_{j \in m}.$$

Then the estimation problem is equivalent to estimating f from

$$\tilde{x}_j = f_j + \zeta_j \qquad\qquad (3.2)$$

In this problem the noise is not white. If Σ_m for all $m \in M_n$ is diagonal (which occurs if the basis is orthogonal for the fixed design), then it will be uncorrelated, so that if the original noise is Gaussian then ζ will be an independent sequence. Let $\sigma_k = 1/b_k$ then we also have $var\zeta_j = \sigma_j^2$, which tends to infinity as $j \to \infty$. Thus the problem is transformed into a noisy problem with dependent and growing variance noise.

An estimator for $f|_{S_m}$ will be called linear if $\hat{f} = C\tilde{y}$, with C a given matrix.

If $J(f)$, for f restricted to S_m is a quadratic functional, the resulting estimator is linear. This relates linear estimators to quadratic regularization functionals. In the rest of this Section we discuss efficient estimation for linear estimators in the case Σ_m is diagonal (for all $m \in M_n$). This means assuming $\sup_{m \in M_n} d_m < n$.

In this case $\tilde{x}_j = \sigma_j \tilde{y}_j$ and we will say the estimator is linear if $\hat{f}_j = h_j \tilde{y}_j$. We have

$$R(m, h, f) = \mathbb{E}\|f - \hat{f}\|_{H_1}^2 = \sum_{k \in m}(1 - b_k h_k)^2 f_k^2 + \sum_{k \in m^c} f_k^2 + \sigma^2/n \sum_{k \in m} h_k^2$$
$$(3.3)$$

For fixed f, the minimum risk is attained at [Tsybakov, 2000]

$$h_k = \frac{b_k f_k^2}{\sigma^2/n + b_k^2 f_k^2} = \sigma_k \frac{f_k^2}{\sigma_k^2 \sigma^2/n + f_k^2} \qquad (3.4)$$

This factor cannot be calculated since it depends on f.

14.3.1.   Minimax estimation over ellipsoids

An important problem is thus giving the minimum risk over a family of functions f with prescribed regularity. The next example, develops these ideas for a specific family of functions. Set $\Theta = \Theta(a, Q)$ equal to the set of functions f such that

$$\sum_j f_j^2 a_j^2 \leq Q.$$

The linear minimax risk $RL(m, Q)$ over Θ, is the minimum linear risk over the worst case in this set for each case

$$RL(m, Q) = \inf_{h} \sup_{f \in \Theta} R(h, m, f)$$

and the minimax risk over Θ, considering all estimators linear and non linear, is

$$R(m, Q) = \inf_{\hat{f} \in S_m} \sup_{f \in \Theta} \mathbb{E} \|\hat{f} - f\|_{H_1}^2.$$

An estimator that achieves the lower bounds is called a linear minimax estimator (respectively a minimax estimator).

Let

$$k_n = \frac{\sum_{j=1}^{t} \sigma_j^2 a_j}{Qn/\sigma^2 + \sum_{j=1}^{t} \sigma_j^2 a_j^2},$$

where $t = \max\{\ell | \sigma^2/n \sum_{j=1}^{\ell} \sigma_j^2 a_j (a_\ell - a_j) \leq Q\}$.

The next result is due to Pinsker [Pinsker, 1980].

THEOREM 7 *Let* $\{a_j\}$ *be a non-decreasing sequence of non-negative numbers such that* $a_j \to \infty$ *and let* $b_j > 0$ *for each* $j = 1, \ldots$ *Then the linear minimax estimator is given by* $\hat{f}_j = h_j^\star \tilde{y}_j$, $h_j^\star = max(\sigma_j(1 - k_n a_j), 0)$, *and*

$$RL(m, Q) = \frac{\sigma^2}{n} \sum_{j \in m} (h_j^\star)^2 + \sum_{j \in m^c} f_j^2$$

Also, if

$$\frac{\max_{j \leq d} \sigma_j^2}{\sum_{j \leq d} \sigma_j^2} \to 0 \quad \text{as } d \to \infty,$$

then

$$R(m, Q) = RL(m, Q)(1 + o(1)).$$

This result gives minimax rates for linear estimators for this family of functions and gives conditions under which minimax linear rates are asymptotically the best possible rates. However, the above results depend on a known sequence $\{a_j\}$. In general, when dealing with real data this kind of information is not available. The problem is to develop strategies based only on the data.

We may consider it convenient to restrict our attention to linear estimators over a certain subset Λ. For example, when considering Pinsker weights

$$h_j = \sigma_j (1 - (\frac{j}{w})^\alpha)^+,$$

with w, α in a certain set. Or

$$h_j = \frac{\sigma_j}{1 + (\frac{j}{w})^\alpha},$$

which corresponds to the Tikhonov-Phillips weights.

In this case, in the spirit of the above results, a linear estimator $h^\star$ will be asymptotically efficient if

$$R(m, h^\star, f) \leq (1 + o(1)) \min_{h \in \Lambda} R(m, h, f).$$

Of course, when estimating, we do not know how to choose good weights, or for that matter a good estimating set $m \in M_n$. Estimators are adaptive if, only based on the data, they are able to achieve efficient rates.

14.4   Penalized model selection for ill posed linear problems

To get a flavor of penalized model selection, in this Section we develop two examples. The proofs are rather technical so they are given in the Appendix. These results say how penalization terms must be chosen, in terms of the dimension of the underlying subspace. They also say, that under additional technical conditions these results are good, in the sense they achieve optimal rates. We stress that what we are doing is controlling complexity by means of dimension. However, the ill posedness, as measured by the sequence $\{b_j\}$ must be considered in this control.

The proof of Theorems 8 and 9 below are based on Ronsenthal type inequalities. These results can be improved by giving exponential rates and controlling the complexity of the spaces S_m in terms of a "covering" number for the L^2 and L^∞ norms, but we have rather not included this additional complication. Our proofs follow closely those of [Baraud, 2000].

We will study two situations:

(A.) When the Gram matrix $\Sigma_m = I$ (that is, when the basis $\{\phi_j\}$ is or-thonormal for the given fixed point design). In this case $\lambda_m \in \Lambda_m$, a given parameter space. Although the general case with a nondiagonal Gram matrix could be dealt with in this situation, it complicates notation and doesn't really add any further insights to the problem. This case is studied in the simpler setting below.

(B.) When the estimation scheme corresponds to the projection estimator. In this case, we do not require Σ_m to be diagonal, but a certain restriction is imposed on its eigenvalues. This kind of restriction is also found in [Kaliffa & Mallat, 2001] when discussing almost diagonal estimation.

14.4.1. First case

As in the above setting, consider linear estimators defined by the expression $\hat{f}_j = q_j(\lambda)(\tilde{y}_k/b_k)$, for $j \in m$. Of course then $\hat{f}(x) = \sum_{j \in m} \hat{f}_j \phi(x)$. For this family of estimators we have, for fixed λ and m, that

$$\mathbb{E}\|\hat{f} - f\|_{H_1}^2 = \sum_{j \in m}(1 - q_j(\lambda))^2 f_j^2 + \sum_{j \in m^c} f_j^2 + \frac{\sigma^2}{n}\sum_{j \in m}\sigma_j^2 q_j(\lambda)^2.$$

The goal is then finding m and λ_m such that the solution is optimal over a given set of parameters Λ. We assume $\Lambda = \cup_{m \in M_n}\Lambda_m$ for a given index set M_n and that $\Lambda_m \subset \Lambda_{m'}$ if $m \subset m'$. We also assume that for each m, Λ_m is a subset of $\mathbb{R}^{d_m}$.

For $\lambda_m \in \Lambda_m$, set $q_j(\lambda) = \frac{1}{1+\sigma_j^2\lambda_j^2}, j \in m$.

As in [Tsybakov, 2000] we shall consider the following contrast, based on the risk function:

$$\gamma_n(\lambda, m) = \sum_{j \in m}(q_j^2(\lambda) - 2q_j(\lambda))((\tilde{y}_j\sigma_j)^2 - \sigma^2\sigma_j^2/n) + \sigma^2/n\sum_{j \in m}q_j^2(\lambda)\sigma_j^2.$$

(4.1)

For each fixed λ, set

$$R(\lambda, m) = \sum_{j \in m}(1 - q_j(\lambda))^2 f_j^2 + \sum_{j \in m^c} f_j^2 + \frac{\sigma^2}{n}\sum_{j \in m}\sigma_j^2 q_j(\lambda)^2$$

We have $\mathbb{E}(\gamma(m, \lambda)) = R(\lambda, m) - \|f\|_2$.

We now introduce the following penalized version of γ_n given by

$$\gamma_n^{pen}(\lambda_m, m) = \gamma_n(\lambda, m) + pen(m),$$

(4.2)

where $pen(m)$ will be defined below.

Set $\hat{f}_{\hat{m},\hat{\lambda}} = \arg\min \gamma_n^{pen}(\lambda_m, m)$. We have the following result

THEOREM 8

- *Assume η is such that there exists $p > 6$ with $\mathbb{E}|\eta|^p < \infty$.*

- *Assume $\sup_{m \in M_n}\sup_{\lambda_m \in \Lambda_m}|q_j(\lambda)| < a_j$.*

- *Assume*

$$\sup_{m \in M_n} \sup_{\lambda_m \in \Lambda_m} \frac{\sum_{j \in m} \sigma_j^4 q_j^2(\lambda)}{\sum_{j \in m} \sigma_j^4 q_j^4(\lambda)} \leq A.$$

- *Assume* $\sup_{m \in M_n} \sum_{j \in m} \sigma_j^2 a_j^2 / n \leq B.$

- *Set* $C_m = \sum_{j \in m} \sigma_j^2 a_j^2 \vee 1.$

- *Set* $s_m = \sup_{j \in m} \sigma_j^2 a_j^2 \vee 1.$

- *Let* L_m *be such that*

$$\sum_{m \in M_n} s_m (\frac{s_m}{L_m C_m})^{p/2-1} + \sum_{m \in M_n} s_m d_m (\frac{s_m}{L_m C_m})^{p-1} < C.$$

- *Assume*

$$pen(m) > \sigma^2 \frac{C_m}{n} (1 + \kappa + (2 + \frac{1}{\kappa} + \mu)L_m),$$

  for $\kappa > 0, \mu > 0.$

Then for a certain $K = K(p)$ *which depends on the distribution of* η*, and on constants A, B, κ and μ,*

$$\mathbb{E}R(\hat{\lambda}, \hat{m}) \leq \inf_{m \in M_n} (\inf_{\lambda \in \Lambda_m} R(\lambda, m) + pen(m)) + CK(p)/n. \qquad (4.3)$$

REMARK 14.4.1   *The assumptions over the regularizing coefficients* $q_j(\lambda)$ *are technical and are given in order to control fluctuations over set* Λ.

REMARK 14.4.2   *The inclusion of term* L_m *is necessary as the number of terms in the sum over* $m \in M_n$ *with the same dimension might be big.*

REMARK 14.4.3   *The contrast can be written as*

$$\gamma_n(\lambda, m) = \sum_{j \in m} (q_j^2(\lambda) - 2q_j(\lambda))((\tilde{y}_j \sigma_j)^2 + c/n \sum_{j \in m} 2q_j(\lambda)\sigma_j^2, \qquad (4.4)$$

$c > \sigma^2$ *and with* $pen(m) > c\frac{C_m}{n}(1+\kappa+(2+\frac{1}{\kappa}+\mu)L_m)$ *for c an given constant as σ may be unknown. In this case however we might be over penalizing. If Λ is finite and* $\sup q_j(\lambda)/\inf q_j(\lambda) < S$*, then c, κ and μ can be chosen from the data (see equation (4.12)).*

REMARK 14.4.4   *If* $w_m \leq C$*, the bounds are as in the usual regression problem. Moreover if* $\sup q_j(\lambda)/\inf q_j(\lambda) < S$*,* $\max_{m \in M_n} d_m < N$ *and* $L_m < c$ *(ordered selection) the bounds are as in [Tsybakov, 2000].*

REMARK 14.4.5   *If* $\sup q_j(\lambda)/\inf q_j(\lambda) < S$*, the penalization term is comparable to the minimum risk, this yields the estimation is efficient modulo a constant.*

14.4.2.   Second case

In this section we are interested in studying projection estimators: that is $q_{j,m}(\lambda) = 1$ if $j \in m$ and $q_{j,m}(\lambda) = 0$ if $j \in m^c$.

For a given linear operator $A : \mathbb{R}^{dm} \rightarrow \mathbb{R}^d$, define, for $\|\|$ the usual euclidean norm

$$\rho(A) = \sup_{s,s \neq 0} \frac{\|As\|}{\|s\|} \tag{4.5}$$

Let $t_m = \frac{\rho(\Sigma_m^{-1})^2}{\rho(\Sigma_m)}$. This term will play an important role in the proof of a result analogous to Theorem 8. Basically, we will assume that $\sup_{m \in M_n} t_m < C$. Heuristically, we can argue that as the number of observations $x_i, i = 1, \ldots, n$ grows, the associated Gram matrix tends to the identity matrix, for, as we recall ψ is an orthonormal basis for H_2. In fact, we shall require for the proof a stronger condition than the one suggested above, namely that

$$\sup_{m \in M_n} \sup_{k,j} |\Sigma_m(k,j) - \delta_{k,j}| < \frac{B}{n}. \tag{4.6}$$

This condition, once again can be argued by the above heuristics, as asking tnat the Gram matrix be "almost" diagonal.

We also require some additional notation. Set $C_m = tr(A_m^{-1})$, $R_m = \frac{\sup_{j \in m} \sigma_j^2}{\sum_{j \in m} \sigma_j^2}$ and $r_m = \frac{\sum_{j \in m} \sigma_j^2}{n}$.

As before, we may consider the estimation scheme in terms of contrasts. Let $m \in M_n$, and $g \in \mathbb{R}^{dm}$. Set,

$$\Gamma_n(g,m) = \|\tilde{x}_m - g\|^2, \tag{4.7}$$

where $\tilde{x}_m$ is defined in Section 3. Of course, minimizing Γ_n is setting $g_j = \tilde{x}_{m,j}$ and then $\Gamma_n(g) = 0$. It will be more convenient however to consider $\gamma_n(g,m) = \Gamma_n(g,m) - \|\tilde{x}_m\|^2$ instead, and in this case the minimum will be $\gamma_n(g,m) = -\sum_{j \in m} \tilde{x}_{m,j}^2$. Now consider,

$$\gamma_n^{pen}(m) = \inf_{g \in \mathbb{R}^{dm}} (\gamma_n(g,m) + pen(m)) = \inf_{g \in \mathbb{R}^{dm}} \left(-\sum_{j \in m} \tilde{x}_{m,j}^2 + pen(m)\right) \tag{4.8}$$

and define

$$\hat{m} = \arg \min_{m \in M_n} \gamma_n^{pen}(m). \tag{4.9}$$

If we identify $f \in H_1$ with the sequence of its Fourier coefficients over basis ϕ, the estimator of γ^{pen} will be $\hat{f}_{\hat{m}} = (\tilde{x}_{\hat{m},j}^2)_{j \in \hat{m}}$. We have the following result:

THEOREM 9 *Assume* $\|f\|_{H_1} < B.$ *Assume (4.6) is satisfied.* *Set* $pen(m) = \kappa\sigma^2 C_m(1 + L_m)/n$, *with* $\kappa > 1$. *Assume* η *is such that there exists* $p > 2$ *with* $\mathbb{E}|\eta|^p < \infty$.
Assume $\sum_{m \in M_n} \left(\frac{R_m r_m t_m}{L_m}\right)^{p/2-1} C_m(1 + L_m) < C(p)$. *Then,*

$$\mathbb{E}\|\hat{f}_{\hat{m}} - f\|_{H_1}^2 \leq c(\kappa)(\inf_{m \in M_n}(\inf_{s \in S_m}\|s - f\|_{H_1}^2 + pen(m)) + C(\kappa, p)\sigma^2 \frac{\mathbb{E}|\eta_1^p|}{n\sigma^p}.$$
$$(4.10)$$

    The proof of the last result follows very much as in [Baraud, 2000], and is given in the Appendix.

REMARK 14.4.6 *The inclusion of term* L_m *is in order to assure that* $\sum_{m \in M_n}\left(\frac{R_m r_m t_m}{L_m}\right)^{p/2} < C.$ *Usually for non ordered selection over a finite set of possibilities this term is chosen as* $L_m = \log N$ *(see Section 2).*

REMARK 14.4.7 *If* Σ_m *is diagonal, the penalization can be writen as* $C(1 + L_m)\sum_{j \in m}\sigma_j^2$. *This case is simpler than the one considered in Theorem 8 as the problem is really discrete, so constants can be estimated. Departure of the penalization from the one given above depends on the eigenvalues of* Σ_m^{-1}. *This introduces the idea of almost diagonal estimation as described in [Kaliffa & Mallat, 2001]. In the diagonal case, the problem is equivalent to hard thresholding estimation [Barron, Birgé & Massart, 1999], which yields the choice of index* j *if* $\tilde{x}_j > \sqrt{C\sigma_j^2 \log N/n}$. *These rates are optimal in the Gaussian case [Kaliffa & Mallat, 2001].*

REMARK 14.4.8 *Assume that we look at the problem (in the diagonal case)*

$$\min_{m \in M_n}[\min_{f \in S_m}(\|y - Af(x)\|_{(n)}^2 - \|y\|_{(n)}^2) + pen(m)], \qquad (4.11)$$

with $pen(m) = \kappa\sigma^2 L_m d_m/n$ *and* $L_m = (1 + log(N))$ $(N = \max_{m \in M_n} d_m)$. *As above, it can be seen that the minimum is obtained ([Barron, Birgé & Massart, 1999]) for* $\tilde{y}_j > \sqrt{\frac{\kappa(1+\log N)}{n}}$. *In other words,* $\tilde{x}_j > \sigma_j\sqrt{\frac{\kappa(1+\log N)}{n}}$, *which is the solution of the problem as defined above. It is remarked that in (4.11) the penalization is just as in the problem with direct observations. However, although we can see that both problemas are equivalent we do not have an equivalent to Theorem 9 for the contrast*

$$\|y - Af(x)\|_{(n)}^2 - \|y\|_{(n)}^2 + pen(m)$$

in the general case

REMARK 14.4.9 *If we assume certain regularity conditions over f, namely $f \in \Theta(a, Q)$, both the ordered selection and the truncated selection yield efficient rates. In the first case, the choice of the penalty yields the quadratic risk smaller than*

$$\inf_{m \in M_n} [\sup_{f \in \Theta} \sum_{k > d_m} f_k^2 + \frac{1}{n} \sum_{j \in m} \sigma_k^2] + C/n$$

$$\leq O(\sum_{k > \tau_n} \frac{1}{a_k^2}),$$

where $\tau_n = \inf\{d_m \ s.t. \ \sum_{k > d_m} \frac{1}{a_k^2} < \frac{1}{n} \sum_{j \in m} \sigma_k^2\}$

14.4.3. Choosing the penalty

Following [Birgé & Massart, 2001a], [Lavielle, 2001], in the ordered selection case we can choose κ in the penalization function from a discrete family. Indeed, we have the following result

LEMMA 1 *There exists two sequences $m_1 = 1 < m_2 < \dots$ and $c_0 = \infty > c_1 > \dots$ defined by*

$$c_i = \frac{\gamma_n(m_{i+1}, \hat{f}_{m_{i+1}}) - \gamma_n(m_i, \hat{f}_{m_i})}{\sum_{j=1}^{m_{i+1}} \sigma_j^2 - \sum_{j=1}^{m_i} \sigma_j^2} \tag{4.12}$$

such that

$$\forall c \in (c_i, c_{i+1}), \hat{m} = m_i.$$

In order to choose the "correct" dimension $\hat{m}$ we inspect the longest intervals (c_i, c_{i+1}), in a sense the most robust as they depend less on small changes of the penalization parameter.

14.5 Bayesian interpretation

Assume, η_i is **Gaussian**$(0, \frac{\sigma^2}{n})$. That is to say each y_i is $\mathcal{N}(Af(x_i), \frac{\sigma^2}{n})$. If we look at the likelihood of $\mathbf{y} = (y_1, \dots, y_n)$ given f we have

$$p(\mathbf{y}|f) \sim p(\mathbf{y}|f, m)p(f|m)p(m)$$

where $\sim$ stands for proportional.

In terms of the discussion of Section 2, we have $\log(p(m)) \sim -\kappa L_m d_m / n$. So that minimizing

$$\|y - A(f)\|_{(n)}^2 + \kappa L_m d_m / n, \tag{5.1}$$

is equivalent to maximizing the likelihood of the observations under an improper uniform prior $p(f|m)$, as suggested by Birgé and Massart [Birgé & Massart 2001].

In a Bayesian framework selecting between two models is achieved by looking at the Bayes factor (see, for example[Han & Carlin, 2000]), that is

$$ B_{ji} = \frac{P(m=j|y)/P(m=i|y)}{P(m=j)/P(m=i)}. $$

As shown in [Han & Carlin, 2000], improper priors for $s|m = k$ will render improper priors for s as well, so that the Bayes factors are not defined. However, a reasonable proposition seems to look instead at the ratio

$$ \tilde{B}_{ji} = \frac{\sup_{s\in m=j} P(s|y, m=j)/\sup_{s\in m=i} P(s|y, m=i)}{P(m=j)/P(m=i)}. $$

Choosing j such that $\tilde{B}_{ji} \le \tilde{B}_{ki}$ for $k \in M_n$ is exactly penalized estimation as in (5.1). It is interesting to remark that in the above setting, model priors are selected solely on the basis of their dimension: $p(m) \sim f(d_m)$. In ordered selection typically $\log p(m)/n \sim C\frac{d_m}{n}$, which corresponds to a Geometric prior. Binomial type priors, yield a heavier penalization, of order $d_m(1 + \log(K))$, for $K = \max_{m\in M_n} d_m$, which corresponds to non ordered selection. Poisson type priors yield penalizations of order $d_m \log d_m$.

In the ill posed case, we look at the renormalized problem associated to $\tilde{x}$ instead of the original problem associated to the observations y. This is done in order to show the estimation scheme is correct. The priors then become functions of $\sum_{j\in m} \sigma_j^2$ instead of functions of the dimension d_m. In terms of the contrasts, rather than looking at the discrepancy measure $\|y - Af(x)\|^2$ we look rather at the empirical risk function associated to a linear estimator $h_k\tilde{x}_k$. That is, the contrast is chosen in such a way that its expectation is the risk function plus a constant. Penalized version of these contrasts must take into account the variance of the renormalized errors.

If we consider additionally a regularization term $J(\lambda f)$, this amounts to selecting $p(f|m) = J(\lambda_m f)$, that is, assuming the priors are not improper. If λ_m must be chosen also we obtain

$$ p(\mathbf{y}|f) \sim p(\mathbf{y}|f, \lambda, m)p(f|\lambda, m)p(\lambda|m)p(m). $$

This is what is done in Section 14.4, Theorem 8.

We remark that in certain cases (see Remark 14.4.8) these penalizations are equivalent to the ones given for the well posed problem based on the original observations, that is, for the contrast $\|y - Af(x)\|^2$. Also see the discussion in Section 14.6 below.

If λ doesn't have to be chosen (other than its dimension which is controlled by $p(m)$), the problem amounts to minimizing

$$p(\mathbf{y}|f) \sim p(\mathbf{y}|f,m)p(f|m)p(m).$$

If $J(f)$ is a quadratic functional of the unknown f_j the resulting estimator will be linear (including projection estimators). This of course corresponds to a Gaussian prior distribution for these coefficients. From a numeric point of view, quadratic functionals "boost" eigenvalues ν_j of A_m by a factor of λ_j if $j \in m$ or by ∞ if $j \notin m$. Choosing the right λ_j is thus choosing the variance $(1/\lambda_j)$ of f_j in such a way that the rates of optimal estimation are achieved.

This Bayesian point of view is also developed in [Loubés, 2001]. In this work the author is interested in obtaining correct rates over ellipsoids of prescribed regularity and thus chooses $\lambda_j, j = 1, \ldots, n$ in order to obtain optimal rates assuming known regularity. As regularity is not known beforehand, he must consider a prior distribution over the set of possible regularities. This prior, $q(s)$ is again chosen in such a way as to assure convergence at optimal rates.

In the next Section 14.6 we discuss $J(f) = \|f\|_1$ and relate this with soft thresholding estimators as in [Kaliffa & Mallat, 2001], although in this case we consider a uniform prior over the set of all possible models $S_m, m \in M_n$.

14.6 L^1 penalization

Consider, as in [Aluffi-Pentini et al, 1999] the problem of regularizing functionals other than quadratic. These authors consider the problem

$$\min_f \|y - Af(x)\|_p + \lambda\|f\|_p, \ p = 1, \infty.$$

In the penalization context, the latter contrast assumes a uniform distribution over the set of all possible $m \in M_n$, and penalizes rather on the coefficients f_k associated to the Fourier expansion of f over the basis $\{\phi\}$.

For the case $p = 1$, this problem has an interpretation in terms of soft threshold estimators [Kaliffa & Mallat, 2001].

As above, consider minimizing

$$\min m \in M_n[\min_{f \in S_m}(\|y - Af\|_{(n)}^2 + J(\lambda f)) + pen(m)]. \qquad (6.1)$$

for a certain λ which will be chosen below.

The solution to this problem for λ given is ([Loubés, 2001], [Loubés & Van de Geer, 2001])

$$\begin{array}{ll} \tilde{x}_j - \lambda_j \sigma_j^2 & \text{if } \tilde{x}_j > \frac{1}{2}\lambda_j \sigma_j^2 \\ \tilde{x}_j + \lambda_j \sigma_j^2 & \text{if } \tilde{x}_j < \frac{1}{2}\lambda_j \sigma_j^2 \\ \quad 0 & \text{if not.} \end{array}$$

Assume f belongs to a set S, such that $\sup_{f \in S} |f_j| \le s_j$. It can be seen [Kaliffa & Mallat, 2001; Kaliffa & Mallat, 2001a] that in order to obtain asymptotically minimax rates, in the Gaussian case,

$$\lambda_j = \left[\begin{array}{ll} b_j/n & \text{if } (\sigma_j \sqrt{1 + N_j})/\sqrt{n} \le s_j \\ \infty & \text{if } (\sigma_j \sqrt{1 + N_j})/\sqrt{n} > s_j, \end{array} \right.$$

where N_j is the total number of coefficients $\tilde{x}$ whose variance satisfy a certain condition [Kaliffa & Mallat, 2001].

In the penalization setup, the above is equivalent to considering $pen(m) = \kappa \frac{1}{n} \sum_{j \in m} \log(N_j)$ and $\lambda_j = 2b_j/n$. Penalization over the dimension thus acts to prevent indexes with big coefficients $\sigma_j \log(N_j)$ to appear. Again, in Bayesian terms, the prior $p(f|m)$ is an L^1 prior, weighted by b_j/n in such a way that it gives less weight to higher dimensions.

14.7   Numerical examples

We next show some numerical results for the projection estimator for ordered model selection. Examples are developed with the cosine basis over [0,1] and the operator is defined by the sequence $b_j = 1/\sqrt{j}$. Noise is gaussian with variance one.

In each case a series of coefficients are randomly selected for a fixed order and then both order and coefficients are estimated from the data. The experiment is repeated for order 5, 10, 15 and 20. In each case the algorithm is allowed to select up to order 40. The number of observations is $n = 100$.

The order is selected as discussed in section 4.3. The sequence of constants is generated as in equation (4.12) for the whole sequence $m = 1, \ldots, 40$. Then the local maxima subsequence is chosen and the lower extreme of the longest interval is selected as the appropriate constant. The selected order is the index corresponding to the selected value. Another way is looking at the sequence of index related to the local maxima. Typically index increase slowly and then jump abruptly. The jump point is a good order pointer.

The figures show the original function, the observations and the reconstruction at the selected order. In the last example, the selected order is 13 although the correct order is 20. The reconstruction for order 20 is also given. Clearly, the reconstruction for the chosen order is better: the illposedness of the operator yields a not as good reconstruction for the correct order as for the chosen order.
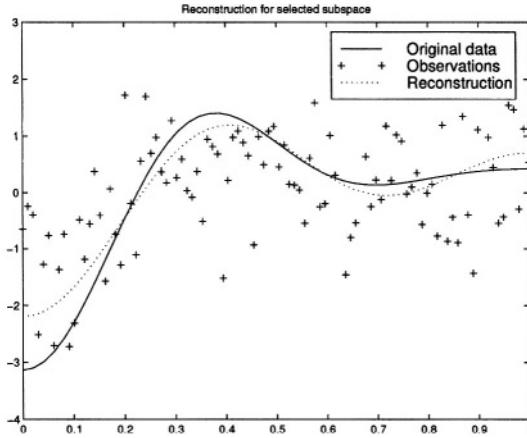
Figure 1. Original function, observations and reconstruction. Original order is 5, selected order is 4
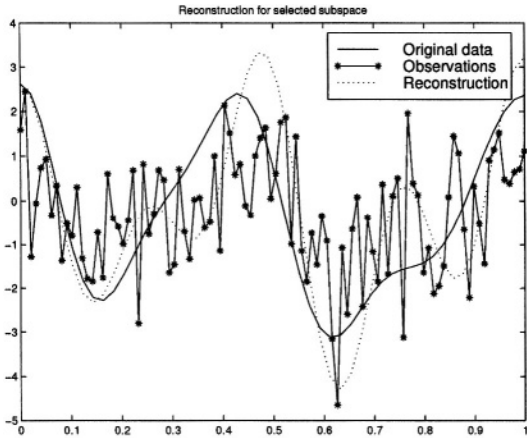
Figure 2. Original function, observations and reconstruction. Original order is 10, selected order is 9
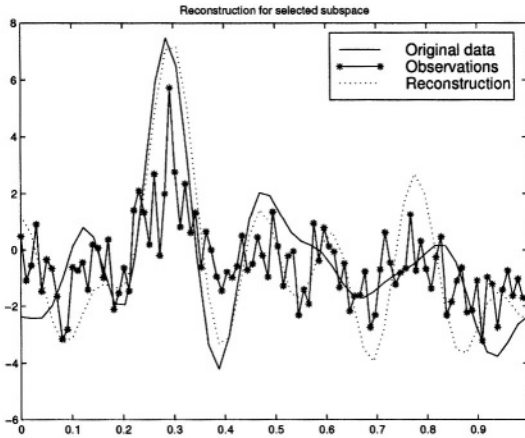
Figure 3.   Original function, observations and reconstruction. Original order is 15, selected order is 14
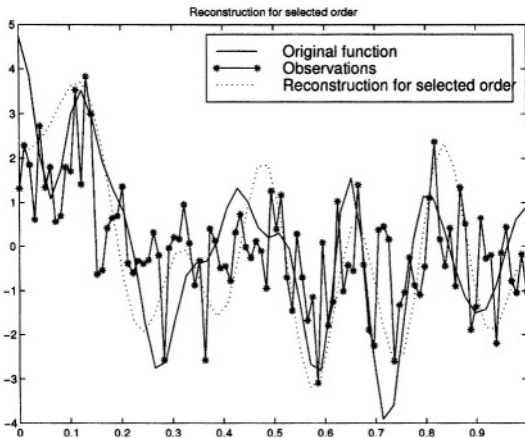
Figure 4.   Original function, observations and reconstruction. Original order is 20, selected order is 13
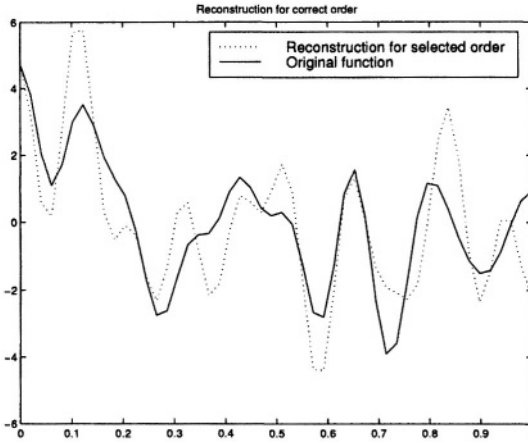
Figure 5. Comparing original function to reconstruction using the correct order. Original order is 20 (same example as Figure 3).

14.8   Appendix

Proof of Theorem 8:
Using standard arguments we have for any λ and m,

$$
\begin{aligned}
R(\hat{\lambda}, \hat{m}) \;\; &\leq R(\lambda, m) + pen(m) - pen(\hat{m}) \\
&+ ((\gamma_n(\lambda, m) - R(\lambda, m)) - (\gamma_n(\hat{\lambda}, \hat{m}) - R(\hat{\lambda}, \hat{m}))).
\end{aligned}
$$

The basic idea behind the proof, is to bound (in probability) the fluctuations of the random part of the contrast using adequate inequalities, which in our case are Rosenthal type inequalities as in [Baraud, 2000].
Set $T(\lambda, \lambda', m, m') = (\gamma_n(\lambda, m) - R(\lambda, m)) - (\gamma_n(\lambda', m') - R(\hat{\lambda}, \hat{m}))$, following [Baraud, 2000] we shall bound this expression for all λ, λ', m and m'.
Let $z_k = 1/n \sum_{i=1}^{n} \eta_i \psi_k(x_i)$. So that

$$
\begin{aligned}
&T(\lambda, \lambda', m, m') \\
= \;\; &2 \sum_{k \in m \cup m'} z_k \sigma_k f_k ((q_k(\lambda)^2 - 2q_k(\lambda)) - (q_k(\lambda)'^2 - 2q_k(\lambda)')) \\
+ \;\; &\sum_{k \in m \cup m'} (z_k^2 - \sigma^2/n) \sigma_k^2 ((q_k(\lambda)^2 - 2q_k(\lambda))^2 - (q_k(\lambda)'^2 - 2q_k(\lambda)')^2) \\
= \;\; &Z_1 + Z_2.
\end{aligned}
$$

First we deal with Z_1.

$$Z_1 = 2 \sum_{k \in m \cap m'} \sigma_k f_k z_k ((q_k(\lambda) - 1)^2 - (q_k(\lambda)' - 1)^2)$$

$$+ \ 2 \sum_{k \in m \cup m' \backslash m \cap m'} \sigma_k f_k z_k ((q_k(\lambda)^2 - 2q_k(\lambda)) - (q_k(\lambda')^2 - 2q_k(\lambda')))$$

$$= \ Z_{11} + Z_{12}$$

It is shown in [Tsybakov, 2000] that

$$((q_k(\lambda) - 1)^2 - (q_k(\lambda') - 1)^2)^2$$
$$= \ [(q_k(\lambda) - 1) + (q_k(\lambda)' - 1)]^2 (q_k(\lambda) - q_k(\lambda)')^2$$
$$\leq \ 2[(q_k(\lambda) - 1)^2 + (q_k(\lambda)' - 1)^2](q_k(\lambda)^2 + q_k(\lambda')^2).$$

Also, we have $2xy \leq ax^2 + \frac{1}{a}y^2$ for all $a > 0, x, y$.
Recall also that $\sup_{m \in M_n} \sup_{\lambda \in \Lambda_m} q_k(\lambda) \leq a_k$.
Thus, for $0 < \alpha < 1$

$$Z_{11} \leq \sum_{k \in m \cap m'} [\alpha f_k^2 [(q_k(\lambda) - 1)^2 + (q_k(\lambda') - 1)^2] + \frac{4}{\alpha} z_k^2 \sigma_k^2 a_k^2]$$

On the other hand,

$$Z_{12} \leq \sum_{k \in m \cup m' \backslash m \cap m'} [\alpha f_k^2 + \frac{2}{\alpha} z_k^2 \sigma_k^2 a_k^2].$$

So that

$$Z_1 \ \leq \ \alpha (\sum_{k \in m} f_k^2 (q_k(\lambda) - 1)^2 + \sum_{k \notin m} f_k^2)$$

$$+ \ \alpha (\sum_{k \in m'} f_k^2 (q_k(\lambda') - 1)^2 + \sum_{k \notin m'} f_k^2)$$

$$+ \ \frac{2}{\alpha} \sum_{k \in m} z_k^2 \sigma_k^2 a_k^2 + \frac{2}{\alpha} \sum_{k \in m'} z_k^2 \sigma_k^2 a_k^2.$$

The latter term is equal to

$$\frac{1}{n^2} \eta^t D_m \eta + \frac{1}{n^2} \eta^t D_{m'} \eta,$$

where, for any given m, D_m is the $n \times n$ matrix

$$D_m = [\sum_{k \in m} \psi_k(x_i)\, \psi_k(x_\ell)\sigma_k^2 a_k^2]_{i,\ell=1,\dots,n}.$$

It is straightforward to see that

$$Tr(D_m) = \frac{1}{n}\sum_{k \in m}\sigma_k^2 a_k^2,$$

and for $\|\|\|$ the usual Euclidean norm over $\mathbb{R}^n$,

$$\rho(D_m) = \sup_{x,x\neq 0}\frac{\|D_m x\|}{\|x\|} \leq \sqrt{Tr(D_m^t D_m)} = (\frac{1}{n^2}\sum_{k \in m}\sigma_k^4 a_k^4)^{1/2}.$$

In Corollary 5.1, [Baraud, 2000] it is shown that for any $n \times n$ matrix M and $p > 2$,

$$P(\eta^t M\eta > \sigma^2 Tr(M) + 2\sigma^2\sqrt{Tr(M)t} + \sigma^2 t) \qquad (8.1)$$
$$\leq \tau(p)C(p)t^{-p/2}\rho(M)^{p-2}Tr(M^t M),$$

where $\tau(p) = \mathbb{E}|\eta_1|^p/\sigma^p$.

Let $a > 0$ and set $u = (1+1/a)L - mTr(D_m) + x/n$. We have, for $p > 2$,

$$P(\eta^t D_m\eta > (1+a)\sigma^2 Tr(D_m) + \sigma^2 u)$$
$$\leq P(\eta^t D_m\eta > \sigma^2 Tr(D_m) + 2\sigma^2\sqrt{Tr(D_m)u} + \sigma^2 u)$$
$$\leq \tau(p)C(p)u^{-p/2}\rho(D_m)^{p-2}Tr(D_m^t D_m)$$
$$\leq \tau(p)C(p)(\frac{Tr(D_m^t D_m)}{u})^{p/2}.$$

Now we bound Z_2. Set $w_j = z_j^2 - \sigma^2/n$ and call $g_j(\lambda) = \sigma_j^2(q_j(\lambda)^2 - 2q_j(\lambda))$.
Set

$$b_m = \max(\sup_{\lambda \in \Lambda_m}\frac{\sum_{k \in m}\sigma_k^4 q_j^2(\lambda)}{\sum_{k \in m}\sigma_k^2 q_j^2(\lambda)}, 1).$$

So that for any $0 < \beta_i < 1, i = 1, 2$

$$
\begin{aligned}
Z_2 \;\leq\; & \;|\sum_{j\in m} g_j(\lambda)w_j| + |\sum_{j\in m'} g_j(\lambda')w_j| \\[2mm]
\leq\; & \;\beta_1/n \sum_{j\in m} q_j^2(\lambda)\sigma_j^4 + \beta_2/n \sum_{j\in m'} q_j^2(\lambda')\sigma_j^4 + \frac{n}{\beta_1}\sum_{j\in m} w_j^2 + \frac{n}{\beta_2}\sum_{j\in m'} w_j^2 \\[2mm]
\leq\; & \;\frac{b_m\beta_1}{n}\sup_{j\in m}\sigma_j^2 \sum_{j\in m} q_j^2(\lambda)\sigma_j^2 + \frac{b_{m'}\beta_1}{n}\sup_{j\in m'}\sigma_j^2 \sum_{j\in m'} q_j^2(\lambda')\sigma_j^2 \\[2mm]
+\; & \;\frac{n}{\beta_1}\sum_{j\in m} w_j^2 + \frac{n}{\beta_2}\sum_{j\in m'} w_j^2.
\end{aligned}
$$

It remains to bound the latter terms.

To begin with, set $d = Var(\eta_1^2)$,

$$
\mathbb{E}w_k^2 = \mathbb{E}(z_k^2 - \frac{\sigma^2}{n})^2 \leq \frac{d}{n^4}\sum_{i=1}^{n}\psi_k^4(x_i) \leq \frac{d}{n^2},
$$

the last inequality because ψ is orthonormal under $<>_n$.

Now set $u_k = w_k^2 - \mathbb{E}w_k^2$. Assume p is even and set $D = \mathbb{E}((\eta_1^2-\sigma^2)^2-d)^2$. We have

$$
\mathbb{E}(\sum_{j\in m}(u_k))^p \leq p^{p-1}\frac{D^p d_m}{n^{4p}}(\sum_{i=1}^{n}\psi_k^8(x_i))^{p/2}
$$

$$
\leq C(p)\frac{D^p d_m}{n^{4p}}(\sum_{i=1}^{n}\psi_k^2)^{2p} = \frac{D^p d_m}{n^{2p}}.
$$

Which allows us to deduce for each $m \in M_n$

$$
P(\sum_{j\in m} w_j^2 > \sigma^2\frac{L_m}{b_m n^2}\sum_{k\in m}\sigma_k^2 a_k^2 + \frac{u}{b_m n^2}) \qquad (8.2)
$$

$$
\leq\; C(p)d_m b_m^p(\sigma^2 L_m \sum_{k\in m}\sigma_k^2 a_k^2 + u)^{-p}.
$$

With the above bounds we are ready to continue the proof. By our choice, we have $pen(m) > \sigma^2(1 + \kappa + (2 + 1/\kappa + \mu)L_m)/n$. Let $\nu < 1$ be such that $(1 + \kappa)\nu/2 > 1$, $\nu(1 + \mu)$ and set $\alpha = \nu$. Let $\delta = (1 + \kappa)\nu/2 - 1$. Choose β_i such that $\beta_1 b_m = \nu$ and $\beta_2 b_{m'} = \nu$

By our choice, we have $pen(m) > \frac{1}{\nu}((C_m \vee 1)\sigma^2 + (b_{m,n} \vee 1)D)d_m/n$. Then,

$$
\begin{aligned}
(1 - \nu)R(\hat{\lambda}, \hat{m}) \;\leq\; & (1 + \nu)(R(\lambda, m) + pen(m)) \\
+\; & 2/\nu[\eta^t D_{\hat{m}} \eta - \frac{\sigma^2(1 + \delta + (1 + 1/\delta)L_{\hat{m}})}{n} C_{\hat{m}}] \\[2mm]
+\; & 2/\nu[\eta^t D_m \eta - \frac{\sigma^2(1 + \delta + (1 + 1/\delta)L_m)}{n} C_m] \\[2mm]
+\; & 1/\nu[b_{\hat{m}} nn \sum_{k \in \hat{m}} w_k^2 - \sigma^2 \frac{L_{\hat{m}}}{n} C_{\hat{m}}] \\
+\; & 1/\nu[b_m nn \sum_{k \in m} w_k^2 - \sigma^2 \frac{L_m}{n} C_m] \\
=\; & (1 + \nu)(R(\lambda, m) + pen(m)) \\
+\; & T_1(\hat{m}) + T_1(m) + T_2(\hat{m}) + T_2(m)
\end{aligned}
$$

Thus, if we set

$$
\Delta = |(1 - \nu)R(\hat{\lambda}, \hat{m}) - (1 + \nu)(\inf_{m \in M_n} (\inf_{\lambda \in \Lambda_m} R(\lambda, m)) + pen(m))|
$$

$$
\begin{aligned}
& P(\Delta > \sigma^2 x/n) \\
\leq\; & P(\cup_{m \in M_n} T_1(m) > \sigma^2 x/4n) + P(\cup_{m \in M_n} T_2(m) > \sigma^2 x/4n) \\
\leq\; & \sum_{m \in M_n} P(T_1(m) > \sigma^2 x/4n) + \sum_{m \in M_n} P(T_2(m) > \sigma^2 x/4n).
\end{aligned}
$$

By (8.1) and (8.2), for any given $m \in M_n$

$$
\begin{aligned}
& P(T_1(m) > \sigma^2 x/4n) \\
=\; & P(\eta^t D_m \eta > \frac{\sigma^2(1 + \delta + (1 + 1/\delta)L_{\hat{m}})}{n} C_{\hat{m}} + \frac{\sigma^2 \nu x}{4n}) \\
\leq\; & C(p, \nu)\tau(p)\left(\frac{Tr(D_m^t D_m)}{(1 + 1/\delta)L_m Tr(D_m) + \nu x/(4n)}\right)^{p/2} \\
\leq\; & C(p, \nu)\tau(p)\left(\frac{1/n \sum_{k \in m} \sigma_k^4 a_k^4)}{(1 + 1/\delta)L_m C_m + \nu x/4}\right)^{p/2} \\
\leq\; & C(p, \nu)\tau(p)B^{p/2}\left(\frac{s_m}{(1 + 1/\delta)L_m C_m + \nu x/4}\right)^{p/2},
\end{aligned}
$$

and,

$$P(T_2(m) > \sigma^2 x/4n) = P(\sum_{j \in m} w_j^2 > \frac{\sigma^2 L_m C_m}{b_m n^2} + \frac{\nu \sigma^2 x}{4n^2 b_m})$$

$$\leq \quad (\frac{D}{\sigma^2})^p C(p, \nu) d_m (\frac{b_m}{L_m C_m + \nu x/4})^p$$

$$\leq \quad (\frac{D}{\sigma^2})^p C(p, \nu) d_m A^p (\frac{s_m}{L_m C_m + \nu x/4})^p.$$

Adding up, we have

$$P(\Delta > \sigma^2 x/n)$$

$$\leq 4C_1(p, \nu)(A^p + B^{p/2}) \sum_{m \in M_n} [(\frac{s_m}{(1 + 1/\delta)L_m C_m + x})^{p/2}$$

$$+ d_m (\frac{s_m}{L_m C_m + x})^p].$$

Since for X positive $\mathbb{E}X = \int_0^\infty P(X > u)du$, we then have that

$$\mathbb{E}[R(\hat{\lambda}, \hat{m}) - \frac{(1 + \nu)}{1 - \nu} (\inf_{m \in M_n} (\inf_{\lambda \in \Lambda_m} R(\lambda, m)) + pen(m))]_+$$

$$\leq \quad C(p, \nu) \frac{1}{n} \sum_{m \in M_n} [s_m (\frac{s_m}{(1 + 1/\delta)L_m C_m})^{p/2 - 1} + d_m s_m (\frac{s_m}{L_m C_m + x})^{p-1}]$$

$$\leq \quad \frac{C(p, \nu) K(p)}{n}$$

which yields the desired result.

Proof of Theorem 9:
The proof of this theorem is essentially as that of Theorem 8. Since there are no weights to be chosen, the proofs are actually simpler. We follow closely the proof of Theorem 3.1 in [Baraud, 2000].
Recall $\gamma_n^{pen} = -\sum_{j \in m} \tilde{x}_{m,j}^2 + pen(m)$. Also recall that $\tilde{x}_{m,j} = f_{m,j} + \zeta_{m,j}$ as defined in Section 3, where $f_{m,j}$ corresponds to the respective Fourier coefficient of f (m is an index set) in terms of the orthonormal basis $\{\phi\}$. Or, in vector notation $\tilde{x}_m = \Pi_m f + \zeta_m$, where $\Pi_m f$ is the projection of f over the subset S_m. Identify $f \in H_1$ with its Fourier coefficients $(f_j)_{j=1}^\infty \in l_2$. If $g \in \mathbb{R}^{d_m}$, for some $m \in M_n$,

$$\gamma_n(m) \quad = \quad -\|\Pi_m f\|^2 - 2 < \Pi_m f, \zeta_m > -\|\zeta_m\|^2$$
$$= \quad \|f - \Pi_m f\|^2 - 2 < \Pi_m f, \zeta_m > -\|\zeta_m\|^2 - \|f\|^2.$$

Thus if $\hat{m}$ is the minimizer of γ_n^{pen} we have, for any other m

$$\gamma_n^{pen}(\hat{m}) \leq \gamma_n^{pen}(m),$$

so that

$$
\begin{aligned}
\|f - \hat{f}_{\hat{m}}\|^2 \leq\ & \|f - \Pi_m f\|^2 + 2 < \Pi_{\hat{m}} f, \zeta_{\hat{m}} > -2 < \Pi_m f, \zeta_m > \\
& +\|\zeta_{\hat{m}}\|^2 - \|\zeta_m\|^2 + pen(m) - pen(\hat{m}).
\end{aligned}
$$

Now,

$$
\begin{aligned}
& 2(< \Pi_{\hat{m}} f, \zeta_{\hat{m}} > - < \Pi_m f, \zeta_m >) \\
=\ & 2(\sum_{k \in \hat{m} \setminus m} f_k \zeta_{\hat{m},k} - \sum_{k \in m \setminus \hat{m}} f_k \zeta_{\hat{m},k} + \sum_{k \in \hat{m} \cap m} f_k(\zeta_{\hat{m},k} - \zeta_{m,k})) \\
\leq\ & \alpha(\sum_{k \in \hat{m} \setminus m} f_k^2 + \sum_{k \in m \setminus \hat{m}} f_k^2) + \frac{1}{\alpha}(\sum_{k \in \hat{m} \setminus m} \zeta_{\hat{m},k}^2 + \sum_{k \in m \setminus \hat{m}} \zeta_{m,k}^2) \\
& +2|\sum_{k \in \hat{m} \cap m} f_k(\zeta_{\hat{m},k} - \zeta_{m,k})| \\
\leq\ & \alpha\|\Pi_{\hat{m}} f - \Pi_m f\|^2 + \frac{1}{\alpha}\|\zeta_{\hat{m}}\|^2 + \frac{1}{\alpha}\|\zeta_m\|^2 \\
& +2|\sum_{k \in \hat{m} \cap m} f_k(\zeta_{\hat{m},k} - \zeta_{m,k})| \\
\leq\ & \alpha\|\Pi_{\hat{m}} f - f\|^2 + \alpha\|\Pi_m f - f\|^2 \\
+\ & \frac{1}{\alpha}\|\zeta_{\hat{m}}\|^2 + \frac{1}{\alpha}\|\zeta_m\|^2 \\
& +2|\sum_{k \in \hat{m} \cap m} f_k(\zeta_{\hat{m},k} - \zeta_{m,k})|
\end{aligned}
$$

where $0 < \alpha < 1$. In the proof above, if Σ_m is the identity for all $m \in M_n$, $\zeta_{m,k} = \zeta_{m',k}$ so that the last term does not appear. Set $c = 1/\alpha + 1$, we have

$$
\begin{aligned}
& (1 - \alpha)\|\Pi_{\hat{m}} f - f\|^2 \leq (1 + 1/\alpha)\|\Pi_m f - f\|^2 \\
& +c\|\zeta_{\hat{m}}\|^2 + c\|\zeta_m\|^2 \\
& +2|\sum_{k \in \hat{m} \cap m} f_k(\zeta_{\hat{m},k} - \zeta_{m,k})| \\
& +pen(m) - pen(\hat{m})
\end{aligned}
$$

Thus,

$$\Delta = (1 - 1\alpha)(\|\Pi_{\hat{m}} f - f\|^2 - K(\|\Pi_m f - f\|^2 + pen(m)))$$
$$\leq \quad c\|\zeta_{\hat{m}}\|^2$$
$$+ \quad c\|\zeta_m\|^2 + 2| \sum_{k \in \hat{m} \cap m} f_k(\zeta_{\hat{m},k} - \zeta_{m,k})| + cpen(\hat{m}) - cpen(m)$$

with $K = (1 + 1/\alpha)/(1 - \alpha)$ and

$$P(\Delta > \frac{\sigma^2 x}{n})$$

$$P\{c\|\zeta_{\hat{m}}\|^2 + c\|\zeta_m\|^2 + 2|\sum_{k \in \hat{m} \cap m} f_k(\zeta_{\hat{m},k} - \zeta_{m,k})| + pen(\hat{m}) + pen(m) > \frac{\sigma^2 x}{n}$$

In order to bound this last probability we have to bound $\|\zeta_m\|^2$ and $|\sum_{k \in m' \cap m} f_k(\zeta_{m',k} - \zeta_{m,k})|$ for any m, m'.

First, remark that for any m

$$\|\zeta_m\|^2 = \frac{1}{n^2} \sum_{i,\ell} \sum_k \eta_i \eta_\ell \sigma_k^2 t_{m,k}(x_i) t_{m,k}(x_\ell) = \frac{1}{n^2} \eta^t D_m \eta,$$

where $t_{m,k}(x) = \Sigma_m^{-1}(\psi_{m,j}(x))_{j \in m}$, D_m is the corresponding $n \times n$ matrix and η is the original error vector. It is straightforward to check that $Tr(D_m) = 1/n Tr(A_m^{-1})$. And because B_m is diagonal, we have

$$Tr(A_m^{-1}) \geq \inf_{k \in m} |\Sigma_m(k, k)| Tr((B^{-1})^2) \geq \rho(\Sigma_m) Tr((B^{-1})^2).$$

On the other hand,

$$\rho^2(D_m) \leq Tr(D_m^t D_m)$$
$$= \quad \frac{1}{n^2} Tr(\Sigma_m^{-1}(B_m^{-1})^4 \Sigma_m^{-1})$$
$$\leq \quad \frac{1}{n^2} (\rho(\Sigma_m^{-1}))^2 Tr((B_m^{-1})^4)$$
$$\leq \quad \frac{1}{n^2} (\rho(\Sigma_m^{-1}))^2 r_m Tr((B_m^{-1})^2).$$

Set for any $n \times n$ matrix M, $v = \eta^t M \eta$. In Corollary 5.1, [Baraud, 2000] it is shown that

$$P(v > \sigma^2 tr(M) + 2\sigma^2 \sqrt{Tr(M)t} + \sigma^2 t) \leq C(p)\tau(p)t^{-p/2}\rho(M)^{p-2}Tr(M^t M),$$

where $\tau(p) = \mathbb{E}|\eta_1|^p/\sigma^p$. Now, set $u = c_2 L_m Tr(D_m) + ct/n$

$$
\begin{aligned}
&P(\zeta^t D_m \zeta - pen(m) > Ct/n) \\
=\ &P(\zeta^t D_m \zeta > c_1 Tr(D_m) + c_2 L_m Tr(D_m) + ct/n) \\
\leq\ &P(\zeta^t D_m \zeta > (c_1 - 1/\alpha)Tr(D_m) + 2\sqrt{Tr(D_m)u} + (1-\alpha)u)
\end{aligned}
$$

The inequalities for $Tr(D_m)$ and $\rho(D_m)$ end the proof, very much as in [Baraud, 2000] (see the proof ot Theorem 8).

Second, we must bound $R = |\sum_{k\in m'\cap m} f_k(\zeta_{m',k} - \zeta_{m,k})|$. This we shall do in several steps.

First, set $t_j = \frac{1}{n}\sum_{i=1}^n \psi_j(x_i)\eta_i$. Recalling the definition of $\zeta_{m,k}$ rewrite

$$
\begin{aligned}
R &= 2\sum_{j\in m} t_j \Sigma_m^{-1} B_m^{-1} \Pi_{m\cap m'} f - 2\sum_{j\in m'} t_j \Sigma_{m'}^{-1} B_{m'}^{-1} \Pi_{m\cap m'} f \\
&= 2\sum_{j\in m\cup m'} t_j(\Sigma_m^{-1} B_m^{-1} \Pi_{m\cap m'} - \Sigma_{m'}^{-1} B_{m'}^{-1} \Pi_{m\cap m'})f \\
&\leq 2(\sum_{j\in m\cup m'} t_j^2)^{1/2}\|(\Sigma_m^{-1} B_m^{-1} \Pi_{m\cap m'} - \Sigma_{m'}^{-1} B_{m'}^{-1} \Pi_{m\cap m'})f\|_2 \\
&\leq \alpha \sum_{j\in m\cup m'} t_j^2 + \frac{1}{\alpha}\|(\Sigma_m^{-1} B_m^{-1} \Pi_{m\cap m'} - \Sigma_{m'}^{-1} B_{m'}^{-1} \Pi_{m\cap m'})f\|_2^2.
\end{aligned}
$$

The first term in the above sum

$$
\begin{aligned}
\sum_{j\in m\cup m'} t_j^2 &\leq \sum_{j\in m} t_j^2 + \sum_{j\in m'} t_j^2 \\
&= \frac{1}{n^2}\eta^t C_m \eta + \frac{1}{n^2}\eta^t C_{m'}\eta
\end{aligned}
$$

with $C_m = (\sum_{j\in m} \psi_j(x_\ell)\psi_j(x_i))$. As before, we have $Tr(C_m) = Tr(\Sigma_m)$ and also that $\rho(C_m) \leq \rho(\Sigma_m)$.

For the second term

$$
\begin{aligned}
&\|(\Sigma_m^{-1} B_m^{-1} \Pi_{m\cap m'} - \Sigma_{m'}^{-1} B_{m'}^{-1} \Pi_{m\cap m'})f\|_2 \\
\leq\ &\|(\Sigma_m^{-1} B_m^{-1} \Pi_{m\cap m'} - B_m^{-1} \Pi_{m\cap m'})f\|_2 + \|B_{m'}^{-1}(\Pi_{m\cap m'}\Sigma_{m'}^{-1} B_{m'}^{-1} \Pi_{m\cap m'})f\|_2 \\
\leq\ &\|(\Sigma_m^{-1} B_m^{-1} - B_m^{-1})\Pi_{m\cap m'})f\|_2 + \|(B_{m'}^{-1}\Sigma_{m'}^{-1} B_{m'}^{-1})\Pi_{m\cap m'})f\|_2 \\
\leq\ &\rho(\Sigma_m^{-1} B_m^{-1} - B_m^{-1})\|f\|_2 + \rho(B_{m'}^{-1}\Sigma_{m'}^{-1} B_{m'}^{-1})f\|_2 \\
\leq\ &\rho(\Sigma_m^{-1})\rho(\Sigma_m B_m^{-1} - B_m^{-1})\|f\|_2 + \rho(\Sigma_{m'}^{-1})\rho(\Sigma_{m'} B_{m'}^{-1} - B_{m'}^{-1})\|f\|_2.
\end{aligned}
$$

By assumption, we have $|\Sigma_m(k,j) - \delta_{k,j}| = |\frac{1}{n}\sum_{i=1}^n \psi_k(x_i)\psi_j(x_i) - \delta_{k,j}| \leq B/n$. On the other hand, for any matrix $M = (a_{i,j})_{i,j=1^d}$. Then

$\rho(M)^2 \leq \sum_{i=1}^{d} \sum_{j=1}^{d} a_{i,j}^2$. Thus,

$$\rho(\Sigma_m B_m^{-1} - B_m^{-1})^2 \leq \frac{B^2}{n^2} \sum_{j \in m} \sum_{k \in m} \sigma_j^2 \leq \frac{B^2 d_m}{n} \frac{1}{n} \sum_{j \in m} \sigma_j^2.$$

Finally, assuming $\sup_{m \in M_n} d_m/n \leq 1$

$$\|(\Sigma_m^{-1} B_m^{-1} \Pi_{m \cap m'} - \Sigma_{m'}^{-1} B_{m'}^{-1} \Pi_{m \cap m'})f\|_2^2$$
$$\leq \quad K(\frac{1}{n} \sum_{j \in m} \sigma_j^2).$$

The rest of the proof now follows as in the proof of Theorem 3.1 in [Baraud, 2000] (pg. 484-485) (see the proof of Theorem 8).

Acknowledgments

References

Aluffi-Pentini et al. (1999) J Optim. Th. & Appl. Vol 103,p 45-64.

Baraud, Y. Model selection for regression on a fixed design.*Probab. Theory Relat. Fields* 117,467-493(2000)

A. Barron, L. Birgé & P. Massart. (1999). Risk bounds for model selection via penalization. Probab. Theory Related Fields. Vol. 113(3), p. 301-413.

L. Birgé & P. Massart. (1998) Minimum contrast estimators on sieves: exponential bounds and rates of convergence. Bernoulli. Vol. 4(3), p. 329-375.

L. Birgé & P. Massart.(2001) Gaussian model selection. J. Eur. math Soc. 3, p. 203-268 .

L. Birgé & P. Massart. (2001a). A generalized C_p criterion for Gaussian model selection. Preprint.

Cavalier, L and Tsybakov, A.B. Sharp adaptation for inverse problems with random noise. Preprint. (2000).

Cavalier,L., Golubev, G.K., Picard, D. and Tsybakov, A.B. Oracle inequalities for inverse problems. Preprint (2000).

[Donoho, 1995, Dohono & Johnstone, 1994]D. Donoho & I. Johnstone. (1994) ideal spatial adaptation via wavelet shrinkage. Biometrika, vol 81, p. 425-455.

D. Donoho. (1995) Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. J. of Appl. and Comput. Harmonic Analysis. Vol. 2(2), p. 101-126.

H.W. Engl & W. Grever. (1994) Using the L-curve foe determining optimal regularization parameters, Numer. Math. Vol. 69, p. 25-31.

A. Frommer & P. Maass. (1999) Fast CG-based methods for Tikhonov-Phillips regularization. SIAM J. Sci. Comput. Vol 20(5), p. 1831-1850.

F. Gamboa & E. Gassiat. (1997) Bayesian methods for ill-posed problems. The Annals of Statistics. Vol. 25, p. 328-350.

F. Gamboa. (1999) New Bayesian Methods for Ill Posed problems. Statistics & Decisions, 17, p. 315-337

Goldenshluger,A. and Tsybakov, A. Adaptive prediction and estimation in linear regression with infinitely many parameters. Preprint (2001).

Grenander, Ulf. (1981) Abstract inference. Wiley Series in Probability and Mathematical Statistics. New York etc.: John Wiley & Sons.

C. Han & B. Carlin. (2000) MCMC methods for computing Bayes factors. A comparative review. Preprint.

J. Kaliffa; S. Mallat. (2001) Thresholding estimators for inverse problems and deconvolutions. To appear in Annals of Stat.

J. Kaliffa; S. Mallat. (2001a) Thresholding estimators for inverse problems and deconvolutions. To appear in Annals of Stat.

M. Kilmer & D. O'Leary. (2001) Choosing regularization parameters in iterative methods for ill posed problems. SIAM-J. Matrix-Anal.-Appl. Vol 22(4),p. 1204-1221.

Lavielle, M. On the use of penalized contrasts for solving inverse problems. Application to the DDC problem. Preprint (2001).

J.M. Loubés. (2001) Adaptive bayesian estimation. Preprint.

J.M. Loubés & S. Van de Geer. (2001). Adaptive estimation in regression, using soft thresholding type penalties. Preprint.

P. Maass, S. Pereverzev, R. Ramlau & S. Solodky. (2001). An adaptive discretization forTikhonov - Phillips regularization with a posteriori parameter selection. Numer. Math.. Vol. 87(3), p. 485-502.

Massart, P. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse* Vol. IX(2), 245-303 (2000).

A. Neubauer (1988). An a posteriori parameter choice for Tikhonov regularization in the presence of modelling error. Appl. Numer. Math. Vol. 4(6),p. 507-519.

F. O'sullivan. (1986) A statistical perspective on ill-posed inverse problems. Statistical Science. Vol. 1(4), p. 502-527.

Tsybakov, A.B. Adaptive estimation for inverse problems: a logarithmic effect in L^2. Preprint (2000).

S.G. Solodky. (1999) Optimization of Projection methods for linear ill-posed problems. Computational Mathematics and Mathematical Physics. Vol 39(2), p. 185-193.

M. Pinsker. (1980) Optimal filtering of square integrable signals in Gaussian white noise. Problems Inform. Transmission. Vol. 16, p. 120-133.

Tikhonov, Andrey N.; Arsenin, Vasiliy Y. Solutions of ill-posed problems. Translation editor Frity John. (English) Scripta Series in Mathematics. New York etc.: John Wiley & Sons; Washington, D.C. 1977

V. Vapnik. (1998) Statistical Learning Theory, John Wiley, NY.

This page intentionally left blank

THE AROV-GROSSMAN MODEL AND BURG'S ENTROPY

J.G. Marcano
Facultad de Ciencias, Universidad de Carabobo

M.D. Morán
Facultad de Ciencias, Universidad Central de Venezuela

Abstract   In this paper, we use the connection between the classic trigonometric Cara-theodory problem and the maximum entropy Burg problem for a stationary processes to obtain from an Operator Theory point of view: Levinson's algorithm, Schur's recursions and the Christoffel-Darboux formula. We deal with a functional model due to Arov and Grossman, which provides a complete description of all minimal unitary extensions of an isometry by the Schur class, in order to describe all the solutions of the Covariance Extension Problem and then we obtain the density that solves the maximum entropy problem of Burg.

15.1   Introduction

A common problem in practice, is to obtain, as a result of any collecting data process in time series studies, a finite complex sequence $\{c_k\}_{-p}^{p}$, p a natural number and try to known when such a sequence constitutes the first p covariance function coefficients. The mathematical formulation of the problem is: Under what conditions over $\{c_k\}_{-p}^{p}$ there is at least a measure on the unit circle μ such that:

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} e^{-ikt} d\mu(t) \qquad k = -p, \ldots, p. \qquad (1.1)$$

We realize that in such a case we have that $c_{-k} = \bar{c}_k$, $k = -p, \ldots, p$. This problem has a long history. In 1911, Toeplitz dealt with the case that the data sequence is of the form $\{c_k\}_{k=0}^{\infty}$, he proved that if a solution exists then it is unique. Problem (1.1), can be seen as a generalization of Toeplitz's problem, but now the solution, if it exists doesn't has to be unique, therefore we have two additional problems: conditions for the uniqueness and the description of all the solutions. In 1940, Nairmark, studied the existence of the covariance

extension problem, using Operator Theory techniques (cf. [Sz-Nagy, 1970]). In 1988, Dym (cf. [Dym, 1988]) and 1989, Woerdeman (cf. [Woerdeman, 1989]) described partially the solutions. Since the solution of Problem (1.1) isn't unique, it is important to find the one which maximizes the Burg maximum entropy functional (cf. [Burg, 1975], [Castro, 1986], [Choi, 1986] y [Landau, 1987]) defined as:

$$\varepsilon(f) = \frac{1}{2\pi} \int_0^{2\pi} \log f(t) dt,$$

where f is the density of a measure μ, which is a solution of Problem (1.1).

In this paper, we approach to this problem from the point of view of Operator Theory: we use Arov-Grossman's model. We associate to the given finite set $\{c_k\}_{k=0}^p$ of autocorrelation coefficients of a second order centered stationary process X, an isometry V acting on a Hilbert space, and we prove that some minimal unitary extension of V, generate a process X such that the spectrum f verifies

$$\frac{1}{2\pi} \int_0^{2\pi} e^{-ikt} f(t) dt = c_k, \quad k = -p, \ldots, p. \qquad (1.2)$$

We use the Arov-Grossman's model (cf. [Arov, 1983]) to describe all different spectrum f of X, verifying (1.2). The description is given by the 1-1 correspondence between such set and a subset of the open unitary ball of $H^\infty(\mathbb{D})$, the set of all analytic and essentially bounded functions. We use some ideas of Marcantognini, Morán and Octavio (cf. ([Marcantognini, 2000], [Marcantognini, 2001]).

Furthermore, the density which solves the maximum entropy problem corresponds to $H \equiv 0$.

We describe all the densities in the Wiener class that are solution to the problem obtained by Dym (cf. [Dym, 1988]) and Woerdeman (cf. [Woerdeman, 1989]).

The same approach is used to obtain Levinson's algorithm, Schur's algorithm and the Christoffel-Darboux formula (cf. [Arocena, 1990], [Bakonyi, 1992], [Castro, 1986], [Foias, 1990], [Kailath, 1986], [Landau, 1987], [Schur, 1986]).

15.2   Notations and preliminaries

Let $\mathbb{C}$ denote the set of complex numbers and let $\mathbb{T}$ denote the complex unit circle $\{z \in \mathbb{C} : |z| = 1\}$, which is the boundary of $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$ the open unit disk of $\mathbb{C}$. We write

$$L^\infty := \{f : \mathbb{T} \to \mathbb{C} : f \text{ is Lebesgue measurable and } \operatorname*{ess\,sup}_{\zeta \in \mathbb{T}} |f(\zeta)| < \infty\}$$

and $L^2 = L^2(\mathbb{T})$ denotes the square integrable (with respect to Lebesgue's measure, dt, on $\mathbb{T}$) Lebesgue measurable functions from $\mathbb{T}$ to $\mathbb{C}$, with the usual norm and inner product denoted by $\|.\|$ and $\langle,\rangle$, respectively. Define e_n by $e_n(\zeta) := \zeta^n$, $\zeta \in \mathbb{T}$, $n \in \mathbb{Z}$, and recall that they form a complete orthonormal basis for the Hilbert space L^2. As usual, $\widehat{f}(n) = \frac{1}{2\pi}\int_0^{2\pi} e^{-int} f(e^{it})dt$ (respectively $\widehat{\mu}(n) = \frac{1}{2\pi}\int_0^{2\pi} e^{-int}d\mu(t)$) denotes the Fourier coefficient of the function f (respectively of the finite measure μ). Also, $H^\infty(\mathbb{D})$ is the set of analytic functions, f, on $\mathbb{D}$ such that its norm $\|f\|_\infty = \sup_{z\in\mathbb{D}} |f(z)|$ is finite. For $q = 2, \infty$, we set $H^q = H^q(\mathbb{T}) := \{f \in L^q(\mathbb{T}) : \widehat{f}(n) = 0, \ n < 0\}$. Finally, we recall that the Wiener algebra $\mathcal{W}$ on the unit circle consists of all complex valued functions f on the unit circle $\mathbb{T}$ of the form $f(\zeta) = \sum_{k=-\infty}^{\infty} \zeta^k \widehat{f}(k)$, $\zeta \in \mathbb{T}$ where $\sum_{k=-\infty}^{\infty} |\widehat{f}(k)| < \infty$. Let $\mathcal{E}_p$ be the manifold spanned by $\{e_0, e_1, \cdots, e_p\}$.

A sequence $\{c_k\}_{k=-p}^{p} \subset \mathbb{C}$ is said to be strictly positive definite if and only if

$$\sum_{n=0}^{p} \sum_{m=0}^{p} \lambda_n \overline{\lambda_m} c_{n-m} > 0, \quad \{\lambda_n\}_{n=0}^{p} \subset \mathbb{C} - \{0\} \tag{2.1}$$

If $\{c_k\}_{k=-p}^{p}$ is a strictly positive definite sequence of complex numbers, we can introduce an inner product in $\mathcal{E}_p$ by setting, for $f = \sum_{k=0}^{p} a_k e_k$ and $g = \sum_{k=0}^{p} b_k e_k$

$$\langle f, g \rangle_p = \sum_{n=0}^{p} \sum_{m=0}^{p} a_n \overline{b_m} c_{m-n}. \tag{2.2}$$

As a consequence of (2.1) we have that $(\mathcal{E}_p, \langle,\rangle_p)$ is a $(p+1)$–dimensional Hilbert space. We define $\Gamma_p : (\mathcal{E}_p, \langle,\rangle) \longrightarrow (\mathcal{E}_p, \langle,\rangle)$ by

$$\langle \Gamma_p f, g \rangle := \langle f, g \rangle_p, \quad f, g \in \mathcal{E}_p. \tag{2.3}$$

Clearly, Γ_p is a linear operator and $\|\Gamma_p\| \leq 1$. We, also conclude:

LEMMA 15.2.1 *Let* $p \in \mathbb{N}$, $\{c_k\}_{k=-p}^{p}$ *a strictly positive definite sequence of complex numbers and* $(\mathcal{E}_p, \langle.,.\rangle_p)$ *be the* $(p+1)$-*dimensional Hilbert space defined in (2.2). Let* $\mathcal{D}_p = Span\{e_k\}_{k=0}^{p-1}$, $\mathcal{R}_p = Span\{e_k\}_{k=1}^{p}$ *be subspaces of* $\mathcal{E}_p$ *and set* $V_p : \mathcal{D}_p \longrightarrow \mathcal{R}_p$ *defined by* $(V_p f)(\zeta) = \zeta f(\zeta), \zeta \in \mathbb{T}, f \in \mathcal{D}_p$. *Then,*

(a) *V_p in an isometry acting on the space* $(\mathcal{E}_p, \langle,\rangle_p)$.

(b) *The orthogonal complement of* $\mathcal{D}_p$, $\mathcal{N}_p = \mathcal{E}_p \ominus \mathcal{D}_p$ *and the orthogonal complement of* $\mathcal{R}_p$, $\mathcal{M}_p = \mathcal{E}_p \ominus \mathcal{R}_p$ *have dimension 1. Furthermore,*

$\mathcal{N}_p$ and $\mathcal{M}_p$ are spanned by $n_p := \dfrac{\Gamma_p^{-1} e_p}{\|\Gamma_p^{-1} e_p\|_p}$ and $m_p := \dfrac{\Gamma_p^{-1} e_0}{\|\Gamma_p^{-1} e_0\|_p}$ respectively.

(c) $P_{\mathcal{N}_p}^{\mathcal{E}_p} e_p = \dfrac{n_p}{\widehat{n_p}(p)} = \left(1 - P_{\mathcal{D}_p}^{\mathcal{E}_p}\right) e_p$, where $\widehat{n_p}(p)$ is the *p-th Fourier coefficient of n_p.*

Proof: (a) is immediate from (2.2). In order to prove (b), we recall that the operator Γ_p defined in (2.3) verifies $\langle x, y \rangle_p = \langle \Gamma_p x, y \rangle$ and since $\langle f, f \rangle_p > 0$ if $f \in \mathcal{E}_p - \{0\}$, we have that if $f \in \mathcal{E}_p$ and $\Gamma_p f = 0$, i.e. $\langle f, f \rangle_p = 0$, whence $f = 0$. This shows that Γ_p is injective. Finally, since $\mathcal{E}_p$ is finite dimensional we obtain Γ_p is invertible. Set $\mathcal{N}_p = \mathcal{E}_p \ominus \mathcal{D}_p$. Let $f \in \mathcal{N}_p$ and $k \in \{0, \cdots, p-1\}$, then

$$\langle \Gamma_p f, e_k \rangle = \langle f, e_k \rangle_p = 0.$$

Since $\Gamma_p f \in (\mathcal{E}_p, \langle ., . \rangle)$ there exists $\lambda \in \mathbb{C}$ such that $\Gamma_p f = \lambda e_p$, so $f = \lambda \Gamma_p^{-1} e_p$. Therefore, $\mathcal{N}_p$ is a 1-dimensional subspace of $\mathcal{E}_p$, moreover, $\mathcal{N}_p$ is spanned by $n_p := \dfrac{\Gamma^{-1} e_p}{\|\Gamma^{-1} e_p\|_p}$, that is,

$$\mathcal{N}_p = Span\{n_p\}.$$

The result concerning $\mathcal{M}_p$ can be proved in a similar fashion.

In order to prove (c), we realize that

$$\widehat{n_p}(p) = \langle \dfrac{\Gamma_p^{-1} e_p}{\|\Gamma_p^{-1} e_p\|_p}, e_p \rangle = \dfrac{1}{\|\Gamma_p^{-1} e_p\|_p} \langle \Gamma_p^{-1} e_p, \Gamma_p^{-1} e_p \rangle_p = \|\Gamma_p^{-1} e_p\|_p$$

and therefore

$$P_{\mathcal{N}_p}^{\mathcal{E}_p} e_p = \langle n_p, e_p \rangle_p n_p = \dfrac{1}{\|\Gamma_p^{-1} e_p\|_p} n_p = \dfrac{n_p}{\widehat{n_p}(p)}.$$

$\square$

REMARK 1 *Let $p \in \mathbb{N}$. We remark that $(\mathcal{E}_{p-1}, \langle ., . \rangle)$ is a subspace of $(\mathcal{E}_p, \langle ., . \rangle)$ and*

$$P_{\mathcal{E}_{p-1}}^{\mathcal{E}_p} \Gamma_p |_{\mathcal{E}_{p-1}} = \Gamma_{p-1}$$

so Γ_{p-1} is the compression of Γ_p to $\mathcal{E}_{p-1}$. Thus if $x, y \in \mathcal{E}_{p-1}$

$$\langle x, y \rangle_p = \langle \Gamma_p x, y \rangle = \langle \Gamma_{p-1} x, y \rangle = \langle x, y \rangle_{p-1}.$$

Also, with the notation of lemma 15.2.1, it is easy to check that

- ▪ $\mathcal{D}_p = \mathcal{E}_{p-1} = \mathcal{D}_{p-1} \oplus \mathcal{N}_{p-1} = \mathcal{R}_{p-1} \oplus \mathcal{M}_{p-1} = V_p \mathcal{D}_{p-1} \oplus \mathcal{M}_{p-1},$

- $\mathcal{R}_p = V_p \mathcal{D}_p = \mathcal{R}_{p-1} \oplus V_p \mathcal{N}_{p-1}$

- $\mathcal{R}_p \ominus \{V_p n_{p-1}\} = Span\{e_k\}_{k=1}^{p-1}$

- $\mathcal{E}_p = \mathcal{D}_1 \oplus \mathcal{N}_p \oplus \mathcal{N}_{p-1} \oplus \cdots \oplus \mathcal{N}_1.$

The following lemma establishes a connection between n_p and m_p and also shows where the zeros of both functions lie.

LEMMA 15.2.2 *Given $p \in \mathbb{N}$, let $\{c_k\}_{k=-p}^p$ be a strictly positive definite sequence of complex numbers and $(\mathcal{E}_p, \langle ., . \rangle_p)$ the $(p+1)$-dimensional Hilbert space defined in (2.2). If Γ_p is the operator defined in (2.3) then*

(a) $\Gamma_p^{-1} e_p = e_p \overline{\Gamma_p^{-1} e_0}$

(b) $n_p = e_p \overline{m_p}$, *that is;* $n_p = \widehat{m_p}(p) + \widehat{m_p}(p-1) e_1 + \cdots + \widehat{m_p}(0) e_p$ *where* $\widehat{m_p}(j)$ *is the j-th Fourier coefficient of m_p.*

(c) *All the zeros of $n_p(z)$ and $m_p(z)$ lie in $|z| < 1$ and $|z| > 1$, respectively.*

Proof: First, (a) follows from the assumption that $\Gamma_p^{-1} e_0 \in \mathcal{E}_p$ and it can be written as $\Gamma_p^{-1} e_0 = \sum_{n=0}^p a_n e_n$, so $\langle \Gamma_p^{-1} e_p, e_k \rangle_p = \delta_p(k) = \langle e_p \overline{\Gamma_p^{-1} e_0}, e_k \rangle_p$. (b) can be obtained as a consequence of (a) and lemma 15.2.1. Finally, let us prove (c). Suppose γ is a zero of n_p. There exists $S_{p-1} \in \mathcal{E}_{p-1}$, such that $n_p(z) = (z - \gamma) S_{p-1}(z)$, $z \in \mathbb{C}$ or equivalently,

$$n_p(z) + \gamma S_{p-1}(z) = z S_{p-1}(z), \quad z \in \mathbb{T}.$$

Since n_p is orthogonal to $\mathcal{E}_{p-1}$, and V_p is an isometry,

$$\|n_p + \gamma S_{p-1}\|_p^2 = \langle n_p + \gamma S_{p-1}, n_p + \gamma S_{p-1} \rangle_p = \|n_p\|_p^2 + |\gamma|^2 \|S_{p-1}\|_p^2$$

which yields

$$\|n_p\|_p^2 + |\gamma|^2 \|S_{p-1}\|_p^2 = \|V_p S_{p-1}\|_p^2 = \|S_{p-1}\|_p^2$$

whence, $1 - |\gamma|^2 = 1/\|S_{p-1}\|_p^2 > 0$, as required. The result concerning to the zeros of m_p can be proved in a similar fashion. $\square$

15.3 Levinson's Algorithm and Schur's Algorithm

Let μ be a positive finite measure on $\mathbb{T}$, and $L^2(\mu)$ be the space of all μ-measurable and square μ-integrable functions. Let $\{\tilde{q}_k\}_{k=0}^p$ be the orthonormal system obtained by applying the Gram-Schmidt process to $\{e_k\}_{k=0}^p$. It is a classic result that for $k = 1, \cdots, p$, the following recurrence equations due to

Szegö (cf. [Bakonyi, 1992], [Castro, 1986], [Choi, 1986], [Kailath, 1986]) are verified:

$$q_k = e_1 q_{k-1} + \gamma_k p_{k-1}$$
$$p_k = q_{k-1} + \overline{\gamma_k} e_1 q_{k-1}, \tag{3.1}$$

where q_k is the monic polynomial associated to $\tilde{q}_k$, $p_k = e_k \overline{q_k}$, $q_0 = p_0 = 1$ and $|\gamma_k| \leq 1$.

The following result is analogous to (3.1).

PROPOSITION 15.3.1 *For each* $p \in \mathbb{N}$, $p \geq 2$ *there exists* $\gamma_p \in \mathbb{C}$ *such that*

$$\frac{n_p}{\widehat{n_p}(p)} = \frac{\zeta n_{p-1}}{\widehat{n_{p-1}}(p-1)} - \gamma_p \frac{m_{p-1}}{\widehat{m_{p-1}}(p-1)}$$
$$\frac{m_p}{\widehat{m_p}(p)} = \frac{m_{p-1}}{\widehat{m_{p-1}}(p-1)} - \overline{\gamma_p} \frac{\zeta n_{p-1}}{\widehat{n_{p-1}}(p-1)}. \tag{3.2}$$

where $\frac{n_1}{\widehat{n_1}(1)} = e_1 e_0 - \overline{c_1} e_0$, $m_1 = e_1 \overline{n_1}$ *and all members in the formulas are the same as in lemma 15.2.1. Furthermore,*

$$(\widehat{n_p}(p))^2 = \frac{(\widehat{n_{p-1}}(p-1))^2}{1 - |\gamma_p|^2} \tag{3.3}$$

Proof: Following remark 1 we have the following decompositions

$$\begin{aligned} P_{\mathcal{D}_p}^{\mathcal{E}_p} e_p &= P_{\mathcal{D}_p}^{\mathcal{E}_p} V_p P_{\mathcal{D}_{p-1}}^{\mathcal{E}_{p-1}} e_{p-1} + P_{\mathcal{D}_p}^{\mathcal{E}_p} V_p P_{\mathcal{N}_{p-1}}^{\mathcal{E}_{p-1}} e_{p-1} \\ &= V_p P_{\mathcal{D}_{p-1}}^{\mathcal{E}_{p-1}} e_{p-1} + P_{\mathcal{M}_{p-1}}^{\mathcal{E}_{p-1}} V_p P_{\mathcal{N}_{p-1}}^{\mathcal{E}_{p-1}} e_{p-1}, \end{aligned}$$

and,

$$\begin{aligned} P_{\mathcal{D}_p}^{\mathcal{E}_p} V_p P_{\mathcal{N}_{p-1}}^{\mathcal{E}_{p-1}} e_{p-1} &= P_{\mathcal{M}_{p-1}}^{\mathcal{E}_p} V_p \langle n_{p-1}, e_{p-1} \rangle_{p-1} n_{p-1} \\ &= \langle n_{p-1}, e_{p-1} \rangle_{p-1} \langle m_{p-1}, V_p n_{p-1} \rangle_p m_{p-1} \end{aligned}$$

hence,

$$P_{\mathcal{D}_p}^{\mathcal{E}_p} e_p = V_p P_{\mathcal{D}_{p-1}}^{\mathcal{E}_{p-1}} e_{p-1} + \langle n_{p-1}, e_{p-1} \rangle_{p-1} \langle m_{p-1}, V_p n_{p-1} \rangle_p m_{p-1}.$$

Thus,

$$e_p - P_{\mathcal{D}_p}^{\mathcal{E}_p} e_p = V_p e_{p-1} - V_p P_{\mathcal{D}_{p-1}}^{\mathcal{E}_{p-1}} e_{p-1} - \frac{\langle m_{p-1}, V_p n_{p-1} \rangle_p m_{p-1}}{\widehat{m_{p-1}}(p-1)},$$

which leads the desired result:

$$\frac{n_p}{\widehat{n_p}(p)} = \frac{\zeta n_{p-1}}{\widehat{n_{p-1}}(p-1)} - \gamma_p \frac{m_{p-1}}{\widehat{m_{p-1}}(p-1)}$$

where $\gamma_p = \langle m_{p-1}, V_p n_{p-1} \rangle_p$. The other recursion of (3.2) follows easily from the equality $n_p = e_p \overline{m_p}$.

To obtain γ_p, let us rewrite (3.2) in the form

$$\frac{n_p}{\widehat{n_p}(p)} + \gamma_p \frac{m_{p-1}}{\widehat{m_{p-1}}(p-1)} = \frac{\zeta n_{p-1}}{\widehat{n_{p-1}}(p-1)},$$

thus

$$\langle \frac{n_p}{\widehat{n_p}(p)} + \gamma_p \frac{m_{p-1}}{\widehat{m_{p-1}}(p-1)}, \frac{n_p}{\widehat{n_p}(p)} + \gamma_p \frac{m_{p-1}}{\widehat{m_{p-1}}(p-1)} \rangle_p$$

$$= \langle \frac{V_p n_{p-1}}{\widehat{n_{p-1}}(p-1)}, \frac{V_p n_{p-1}}{\widehat{n_{p-1}}(p-1)} \rangle_p$$

using the fact that V_p is an isometry and that n_p is orthogonal to m_{p-1}, we find

$$\frac{1}{(\widehat{n_p}(p))^2} = \frac{1 - |\gamma_p|^2}{(\widehat{n_{p-1}}(p-1))^2}.$$

☐ The coefficients γ_p are called the Schur parameters , this name comes from the classical Schur algorithm (cf. [Bakonyi, 1992], [Kailath, 1986], [Landau, 1987], [Schur, 1986]). Indeed, setting $G_p(z) = \frac{n_p(z)}{m_p(z)}$ and using Levinson's algorithm we can rewrite $G_p(z)$ as

$$G_p(z) = \frac{z n_{p-1} - \gamma_p m_{p-1}}{m_{p-1} - \overline{\gamma_p} z n_{p-1}} = \frac{z G_{p-1} - \gamma_p}{1 - \overline{\gamma_p} z G_{p-1}}.$$

15.4 The Christoffel-Darboux formula

If $P \in \mathcal{E}_p$ then P is a polynomial function and we can evaluate $P(z)$ for $z \in \mathbb{D}$. Let $z \in \mathbb{D}$ and consider $f^z : (\mathcal{E}_p, \langle \cdot, \cdot \rangle_p) \to \mathbb{C}$, be defined by $f^z(P) = P(z)$. Clearly, f^z is a linear function. The next proposition shows that f^z is continuous, which implies that $(\mathcal{E}_p, \langle \cdot, \cdot \rangle_p)$ can be considered a reproducing kernel space.

PROPOSITION 15.4.1 *Let $z \in \mathbb{D}$ and $E_p^z = \Gamma_p^{-1} \sum_{k=0}^{p} \overline{z}^k e_k$ then*

$$\langle P, E_p^z \rangle_p = P(z), \quad E_p^z = \sum_{k=0}^{p} \langle E_p^z, n_k \rangle_p n_k, \tag{4.1}$$

where $n_0 = \frac{e_0}{\|e_0\|_p}$ and n_k, $k = 1, \ldots, p$ are as in lemma 15.2.1.

Proof: If $f \in L^2$ we have $\langle \frac{f}{1 - z e_{-1}}, e_0 \rangle = \sum_{k=0}^{\infty} z^k \widehat{f}(k)$. It is easy to check that

$$P(z) = \langle P, \Gamma_p^{-1} \sum_{k=0}^{p} \overline{z}^k e_k \rangle_p \text{ and } |P(z)| \leq \|P\|_p \|E_p^z\|_p$$

which shows that the linear function that associates to $P \in \mathcal{E}_p$ its value at $z \in \mathbb{D}$ is continuous. The second equation is an easy consequence of the fact that $\{n_0, n_1, \ldots, n_p\}$ is an orthonormal system for $\mathcal{E}_p$. □

As a consequence of Proposition 15.4.1, we have:

THEOREM 15.4.2 *(The Christoffel-Darboux formula)*
If $z, \xi \in \mathbb{D}$ then

$$E_p^\xi(z) = \frac{\overline{m_p(\xi)} m_p(z) - \overline{\xi n_p(\xi)} z n_p(z)}{1 - \bar{\xi} z}$$

where n_p, m_p are defined as in lemma 15.2.1 and E_p^ξ is defined as in proposition 15.4.1.

Proof: Let $Q \in \mathcal{E}_{p-1}$ and $z, \xi \in \mathbb{D}$. Using the definition of E_p^ξ and the fact that V_{p+1} is an isometry, we obtain that

$$0 = \langle (e_1 - \xi)Q, E_p^\xi \rangle_p = \langle e_1 Q, (1 - \bar{\xi} e_1) E_p^\xi \rangle_{p+1}.$$

Using the fact that $\mathcal{R}_p = Span\{e_1 Q : Q \in \mathcal{E}_{p-1}\}$ we obtain $(1 - \bar{\xi} e_1) E_p^\xi \in \mathcal{E}_{p+1} \ominus \mathcal{R}_p$ the orthogonal complement of the subspace $\mathcal{R}_p$ with respect to the $(p + 1)$-dimensional space $\mathcal{E}_{p+1}$. On the other hand, $e_1 n_p, E_p^0 \in \mathcal{E}_{p+1} \ominus \mathcal{R}_p$ and since both polynomials have different degree they generate the at most 2-dimensional space $\mathcal{E}_{p+1} \ominus \mathcal{R}_p$. Therefore,

$$(1 - \bar{\xi} e_1) E_p^\xi = a E_p^0 + b e_1 n_p.$$

By (4.1), $b = -\overline{\bar{\xi} n_p(\xi)}$ and also,

$$(1 - \bar{\xi} e_1) E_p^\xi = a E_p^0 - \overline{\bar{\xi} n_p(\xi)} e_1 n_p,$$

which yields

$$(1 - \bar{\xi} w) E_p^\xi(w) = a E_p^0(w) - \overline{\bar{\xi} n_p(\xi)} w n_p(w), \quad w \in \mathbb{C}.$$

The desired result comes easily from the fact $n_p = e_p \overline{m_p}$. □

15.5   Description of all spectrums of a stationary process

The main result of this section is the description of the set of all measures μ absolutely continuous with respect Lebesgue's measure on $\mathbb{T}$ and such that $\hat{\mu}(k) = c_k, \quad k = -p, \ldots, p$, where $\{c_k\}_{k=-p}^p$ is a given strictly positive definite complex sequence.

We require some notions of Harmonic Analysis of Operators on Hilbert Spaces (cf. [Sz-Nagy, 1970]).

Let $\mathcal{H}$ be a Hilbert space, $\mathcal{D}, \mathcal{R}$ two closed subspaces of $\mathcal{H}$ and $V : \mathcal{D} \to \mathcal{R}$ an isometry acting on $\mathcal{H}$. We say that a unitary operator U acting on a Hilbert space $\mathcal{F}$ is a unitary extension of the isometry V if and only if $\mathcal{H}$ is a closed subspace of $\mathcal{F}$ and $U|_{\mathcal{D}} = V$. If in addition, $\mathcal{F} = \bigvee_{n \in \mathbb{Z}} U^n(\mathcal{H})$, we say that U is a minimal unitary extension of V. We identify two minimal unitary extensions of V, U and U' acting respectively, on the Hilbert spaces $\mathcal{F}$ and $\mathcal{F}'$ if and only if there exists a unitary operator $\Phi : \mathcal{F} \to \mathcal{F}'$ such that $\Phi|_{\mathcal{H}} = I_{\mathcal{H}}$ and $\Phi U = U'\Phi$. Let $\mathcal{N}, \mathcal{M}$ be two closed subspaces of the Hilbert space $\mathcal{H}$, $\mathcal{L}(\mathcal{N}, \mathcal{M})$ denotes as usual the set of all bounded linear operators from $\mathcal{N}$ to $\mathcal{M}$. An operator valued function $\Theta : \mathbb{D} \to \mathcal{L}(\mathcal{N}, \mathcal{M})$ is a contractive analytic function if and only if $\sup_{z \in \mathbb{D}} \|\Theta(z)\| \leq 1$ and there exists a sequence $\{\Theta_k\}_{k \geq 0} \subset \mathcal{L}(\mathcal{N}, \mathcal{M})$ such that $\Theta(z) = \sum_{k > 0} z^k \Theta_k$, $z \in \mathbb{D}$, where the convergence is in the operator norm. The Schur's class, $\mathcal{S}(\mathcal{N}, \mathcal{M})$ is the set of all contractive analytic function $\Theta : \mathbb{D} \longrightarrow \mathcal{L}(\mathcal{N}, \mathcal{M})$. The Arov and Grossman functional model (cf. [Arov, 1983], [Marcantognini, 2000]) establishes the existence of a bijection between the unitary extension of an isometry $V : \mathcal{D} \longrightarrow \mathcal{R}$ acting on $\mathcal{H}$, indistinguishable from the geometric point of view, and the class of Schur $\mathcal{S}(\mathcal{N}, \mathcal{M})$, where $\mathcal{N}, \mathcal{M}$ are the defect spaces of V. Given $U \in \mathcal{L}(\mathcal{F})$ a minimal unitary extension of the isometry V, $\Theta : \mathbb{D} \longrightarrow \mathcal{L}(\mathcal{N}, \mathcal{M})$ defined by

$$\Theta(z) = P_{\mathcal{M}}^{\mathcal{F}} U (I - z P_{\mathcal{F} \ominus \mathcal{H}}^{\mathcal{F}} U)^{-1}|_{\mathcal{N}}, \quad z \in \mathbb{D}$$

is a function in the Schur class, and the relation is bijective. When U and Θ are related as above, we denote $U = U_{\Theta}$ and $\mathcal{F} = \mathcal{F}_{\Theta}$.

We use this theory for the particular case when $\mathcal{H} = (\mathcal{E}_p, \langle, \rangle_p)$, $\mathcal{D} = \mathcal{D}_p$, $\mathcal{R} = \mathcal{R}_p$, $V = V_p$, $\mathcal{M} = \mathcal{M}_p$ and $\mathcal{N} = \mathcal{N}_p$. We recall that in this case $\mathcal{M}_p$ and $\mathcal{N}_p$ are 1–dimensional subspaces of $\mathcal{E}_p$ and therefore there exists a bijection between the Schur class $\mathcal{S}(\mathcal{M}_p, \mathcal{N}_p)$ and the closed unitary ball of $H^{\infty}(\mathbb{D})$. In the other hand $U \in \mathcal{L}(F)$ is a minimal unitary extension of V_p if and only if $U^{-1} \in \mathcal{L}(F)$ is a minimal unitary extension of V_p^*. Consequently, there exists a one to one correspondence between the minimal unitary extension U^{-1} of the isometry V_p^* and the functions of $H^{\infty}(\mathbb{D})$, such that $\|H\|_{\infty} \leq 1$, in order to recall the relation between a fixed H and a minimal unitary extension $U^{-1} \in \mathcal{L}(\mathcal{F})$ we set $U^{-1} = U_H^{-1}$ and $\mathcal{F} = \mathcal{F}_H$.

LEMMA 15.5.1 *For each $p \in \mathbb{N}$ and $P \in (\mathcal{E}_p, \langle \cdot, \cdot \rangle_p)$ then*

$$\left\langle \left(I - z V_p^* P_{\mathcal{R}_p}^{\mathcal{E}_p} \right)^{-1} P, m_p \right\rangle_p = \frac{P(z)}{m_p(z)}$$

where V_p is the isometry and m_p is the function given in lemma 15.2.1.

Proof: Let $Q := Q(z, \zeta) = \left(I - zV_p^* P_{\mathcal{R}_p}^{\mathcal{E}_p} \right)^{-1} P$.

Then $Q - zV_p^* P_{\mathcal{R}_p}^{\mathcal{E}_p} Q = P$ thus

$$Q - ze_{-1}Q + ze_{-1}\langle Q, m_p\rangle_p m_p = P$$

and

$$Q = \frac{P - ze_{-1}\langle Q, m_p\rangle_p m_p}{1 - ze_{-1}}.$$

The result follows from $\langle Q, m_p\rangle_p = \frac{\widehat{Q}(0)}{\widehat{m}_p(0)}$ and

$$
\begin{aligned}
\widehat{Q}(0) &= \langle Q, e_0\rangle = \left\langle \frac{P - e_{-1}\langle Q, m_p\rangle_p m_p}{1 - ze_{-1}}, e_0 \right\rangle \\
&= \left\langle \frac{P}{1 - ze_{-1}}, e_0 \right\rangle - z\langle Q, m_p\rangle_p \left\langle \frac{m_p}{1 - ze_{-1}}, e_1 \right\rangle \\
&= P(z) - \langle Q, m_p\rangle_p m_p(z) + \langle Q, m_p\rangle_p \widehat{m_p}(0).
\end{aligned}
$$

$\square$

We will use the following lemma, the proof can be seen in [Arov, 1983] or [Marcantognini, 2000].

LEMMA 15.5.2 *Given* $H \in H^\infty(\mathbb{D})$ *such that* $\|H\|_\infty \leq 1$. *If* $U_H^{-1} \in \mathcal{L}(\mathcal{F}_H)$ *is the minimal unitary extension of* $V_p^* : \mathcal{R}_p \longrightarrow \mathcal{D}_p$ *related to H then,*

$$P_{\mathcal{E}_p}^{\mathcal{F}_H} \left(I - zU_H^{-1} \right)^{-1} |_{\mathcal{E}_p} = \left(I - z\left(V_p^* P_{\mathcal{R}_p}^{\mathcal{E}_p} + \beta_H(z) P_{\mathcal{M}_p}^{\mathcal{E}_p} \right) \right)^{-1}$$

The following lemma establishes a useful relation between $\beta_H(z)$ and H.

LEMMA 15.5.3 *If* $\beta_H \in \mathcal{S}(\mathcal{M}_p, \mathcal{N}_p)$, *then*

$$\left(I - z\beta_H(z) P_{\mathcal{M}_p}^{\mathcal{E}_p} \left(I - zV_p^* P_{\mathcal{R}_p}^{\mathcal{E}_p} \right)^{-1} \right)^{-1} e_0 = e_0 + \frac{zH(z)}{m_p(z) - zH(z)n_p(z)} n_p$$

Proof: We use that: if $A, B \in \mathcal{L}(\mathcal{F})$ and $\|A\|, \|B\| < 1$ and $\|A + B\| < 1$ then,

$$\left(I - (A + B) \right)^{-1} = (I - A)^{-1} \left(I - B(I - A)^{-1} \right)^{-1}, \qquad (5.1)$$

to check that

$$Q_H := Q_H(\zeta, z) = \left(I - z\beta_H(z) P_{\mathcal{M}_p}^{\mathcal{E}_p} \left(I - zV_p^* P_{\mathcal{R}_p}^{\mathcal{E}_p} \right)^{-1} \right)^{-1} e_0,$$

exists. Hence

$$Q_H - z\beta_H(z)P_{\mathcal{M}_p}^{\mathcal{E}_p}\left(I - zV_p^*P_{\mathcal{R}_p}^{\mathcal{E}_p}\right)^{-1}Q_H = e_0,$$

thus, we obtain

$$Q_H - zH(z)\left\langle\left(I - zV_p^*P_{\mathcal{R}_p}^{\mathcal{E}_p}\right)^{-1}Q_H, m_p\right\rangle_p n_p = e_0. \tag{5.2}$$

Let

$$d = \left\langle\left(I - zV_p^*P_{\mathcal{R}_p}^{\mathcal{E}_p}\right)^{-1}Q_H, m_p\right\rangle_p$$

therefore, if we apply to both members of equality (5.2) the operator

$$\left(I - zV_p^*P_{\mathcal{R}_p}^{\mathcal{E}_p}\right)^{-1}$$

and take the scalar product with m_p, we obtain

$$d - zH(z)d\left\langle\left(I - zV_p^*P_{\mathcal{R}_p}^{\mathcal{E}_p}\right)^{-1}n_p, m_p\right\rangle_p = \left\langle\left(I - zV_p^*P_{\mathcal{R}_p}^{\mathcal{E}_p}\right)^{-1}e_0, m_p\right\rangle_p.$$

>From lemma 15.5.1 we have

$$d - zH(z)d\frac{n_p(z)}{m_p(z)} = \frac{e_0}{m_p(z)} = \frac{1}{m_p(z)}.$$

Therefore $d[m_p(z) - zH(z)n_p(z)] = 1$ and

$$m_p(z) - zH(z)n_p(z) \neq 0 \tag{5.3}$$

thus

$$d = \frac{1}{m_p(z) - zH(z)n_p(z)}.$$

The result follows easily. $\qquad\square$

As seen in the previous lemma, $H \equiv 0$ is simpler than the others cases. We study such a case in the next proposition.

PROPOSITION 15.5.4 *If μ^0 is the spectral measure related to U_0 the minimal unitary extension of V_p associated to $H \equiv 0$ then,*

$$d\mu_0(t) = \frac{1}{|m_p(e^{it})|^2}dt$$

where $\mu_0(s) = \langle\mu^0(s)e_0, e_0\rangle, s \in [0, 2\pi]$.

Proof: As consequence of lemma 15.5.1 and lemma 15.5.2 and the Spectral Theorem we have that

$$
\begin{aligned}
1 &= \left\langle \left(I - zV_p^* P_{\mathcal{R}_p}^{\mathcal{E}_p}\right)^{-1} m_p, m_p \right\rangle_p = \left\langle \left(I - zU_0^{-1}\right)^{-1} m_p, m_p \right\rangle_{\mathcal{F}_0} \\
&= \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{1 - ze^{-it}} d\langle \mu^0(t)m_p, m_p \rangle \\
&= \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{1 - ze^{-it}} |m_p(e^{it})|^2 d\langle \mu^0(t)e_0, e_0 \rangle.
\end{aligned}
$$

Whence $\frac{1}{2\pi} \int_0^{2\pi} e^{-ikt} |m_p(e^{it})|^2 d\langle \mu^0(t)e_0, e_0 \rangle = \delta_0(k)$ and therefore $|m_p(e^{it})|^2 d\langle \mu^0(t)e_0, e_0 \rangle = dt$.

$\square$

The next proposition shows that there are some spectral measures of $U_H \in \mathcal{L}(\mathcal{F}_H)$ the minimal unitary extension of V_p that are absolutely continuous with respect to dt.

PROPOSITION 15.5.5 *Given* $H \in H^\infty(\mathbb{D})$ *such that* $\|H\|_\infty \le 1$, *let* μ^H *be the spectral measure of* $U_H \in \mathcal{L}(\mathcal{F}_H)$ *the minimal unitary extension of* $V_p : \mathcal{D}_p \longrightarrow \mathcal{R}_p$ *associated to* H. *If* $\mu_H(t) = \langle \mu^H(t)e_0, e_0 \rangle$ *then* μ_H *verifies:*

$$
\begin{aligned}
\frac{1}{2\pi} \int_0^{2\pi} \frac{e^{it} + z}{e^{it} - z} d\mu_H(t) &= \frac{2zH(z)}{m_p(z) - zH(z)n_p(z)} \frac{z^p}{m_p(z)} \\
&\quad + \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{it} + z}{e^{it} - z} \frac{1}{|m_p(e^{it})|^2} dt
\end{aligned}
$$

Proof: Let $A_H(z) = \left\langle \left(I + zU_H^{-1}\right)\left(I - zU_H^{-1}\right)^{-1} e_0, e_0 \right\rangle_{\mathcal{F}_H}$, then by lemma 15.5.2 we have that

$$
\begin{aligned}
A_H(z) &= 2\left\langle \left(I - zU_H^{-1}\right)^{-1} e_0, e_0 \right\rangle_{\mathcal{F}_H} - \left\langle e_0, e_0 \right\rangle_{\mathcal{F}_H} \\
&= 2\left\langle \left(I - z\left(V_p^* P_{\mathcal{R}_p}^{\mathcal{E}_p} + \beta_H(z)P_{\mathcal{M}_p}^{\mathcal{E}_p}\right)\right)^{-1} e_0, e_0 \right\rangle_{\mathcal{F}_H} - \left\langle e_0, e_0 \right\rangle_{\mathcal{F}_H}.
\end{aligned}
$$

We recall (5.1) and we obtain

$$
A_H(z) = 2\left\langle \left(I - zV_p^* P_{\mathcal{R}_p}^{\mathcal{E}_P}\right)^{-1} \left(I - z\beta_H(z)P_{\mathcal{M}_p}^{\mathcal{E}_P}\left(I - zV_p^* P_{\mathcal{R}_p}^{\mathcal{E}_P}\right)^{-1}\right)^{-1} e_0, e_0 \right\rangle_{\mathcal{F}_H}
$$

$$
- \left\langle e_0, e_0 \right\rangle_{\mathcal{F}_H}.
$$

From lemma 15.5.3 and the Spectral Theorem we conclude that

$$A_H(z) = 2\left\langle \left(I - zV_p^* P_{\mathcal{R}_p}^{\mathcal{E}_P}\right)^{-1} \left(e_0 + \frac{zH(z)}{m_p(z) - zH(z)n_p(z)}n_p\right), e_0 \right\rangle_{\mathcal{F}_H}$$
$$- \left\langle e_0, e_0 \right\rangle_{\mathcal{F}_H}$$

$$= 2\left\langle \left(I - zU_0^{-1}\right)^{-1} \left(e_0 + \frac{zH(z)}{m_p(z) - zH(z)n_p(z)}n_p\right), e_0 \right\rangle_{\mathcal{F}_H}$$
$$- \left\langle e_0, e_0 \right\rangle_{\mathcal{F}_H}$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \left\{\frac{1}{1 - ze^{-it}}\left[2 + \frac{2zH(z)}{m_p(z) - zH(z)n_p(z)}n_p(e^{it})\right] - 1\right\}$$
$$\frac{1}{|m_p(e^{it})|^2}dt$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{1 - ze^{-it}}\left[(1 + ze^{-it}) + \frac{2zH(z)}{m_p(z) - zH(z)n_p(z)}n_p(e^{it})\right]$$
$$\frac{1}{|m_p(e^{it})|^2}dt$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \frac{e^{it} + z}{e^{it} - z}\frac{1}{|m_p(e^{it})|^2}dt + \frac{2zH(z)}{m_p(z) - zH(z)n_p(z)}\frac{z^p}{m_p(z)}.$$

$\square$

The following corollary gives a necessary and sufficient condition in order that μ_H be absolutely continuous with respect to Lebesgue's measure.

COROLLARY 15.5.6 *Given $H \in H^\infty(\mathbb{D})$ with $\|H\|_\infty \leq 1$, let μ_H the measure defined in the previous proposition. The measure μ_H is absolutely continuous with respect to the Lebesgue measure on $\mathbb{T}$, with density f_H given by*

$$f_H(\zeta) = \frac{1}{|m_p(\zeta)|^2}Re\left[\frac{m_p(\zeta) + \zeta H(\zeta)n_p(\zeta)}{m_p(\zeta) - \zeta H(\zeta)n_p(\zeta)}\right].$$

if and only if the set $\{\zeta \in \mathbb{T} : |H(\zeta)| = 1\}$ has Lebesgue measure zero and $1/(m_p - \zeta H n_p) \in L^2$.

Furthermore, if H is inner then μ_H is singular with respect to Lebesgue's measure on $\mathbb{T}$.

Proof: Let $H \in H^\infty(\mathbb{D})$ with $\|H\|_\infty \leq 1$. We know from (5.3) that $m_p(z) - zH(z)n_p(z) \neq 0$, $z \in \mathbb{D}$ and then the set

$$\{\zeta \in \mathbb{T} : m_p(\zeta) - \zeta H(\zeta)n_p(\zeta) = 0\}$$

has Lebesgue measure zero. Therefore, from the last proposition

$$\lim_{z \to \zeta} \frac{1}{2\pi} \int_0^{2\pi} Re\frac{e^{it}+z}{e^{it}-z} d\mu_H(t) = Re\left[\frac{2H(\zeta)\zeta^{p+1}}{m_p(\zeta)[m_p(\zeta)-\zeta H(\zeta)n_p(\zeta)]}\right]$$

$$+ \frac{1}{|m_p(\zeta)|^2}$$

$$= \frac{1}{|m_p(\zeta)|^2} Re\left[\frac{m_p(\zeta)+\zeta H(\zeta)n_p(\zeta)}{m_p(\zeta)-\zeta H(\zeta)n_p(\zeta)}\right]$$

$$= \frac{1-|H(\zeta)|^2}{|m_p(\zeta)-\zeta H(\zeta)n_p(\zeta)|^2} = f_H(\zeta) \text{ a.e.}$$

From the fact; $1-\|H\|_\infty^2 \le 1-|H|^2 \le 1$, we obtain $f_H \in L^1$ if and only if $\frac{1}{m_p-\zeta Hn_p} \in L^2$, in fact,

$$\frac{1-\|H\|_\infty^2}{|m_p-\zeta Hn_p|^2} \le f_H \le \frac{1}{|m_p-\zeta Hn_p|^2}.$$

Moreover, f_H is positive a.e. if and only if the set $\{\zeta \in \mathbb{T} : |H(\zeta)| = 1\}$ has Lebesgue measure zero.

If H is inner, let

$$M(z) := e^{-\frac{1}{2\pi} \int_0^{2\pi} \frac{e^{it}+z}{e^{it}-z} d\mu_H(t)}$$

Since $Re\left[\frac{m_p(\zeta)+\zeta H(\zeta)n_p(\zeta)}{m_p(\zeta)-\zeta H(\zeta)n_p(\zeta)}\right] = 0$, it results

$$\lim_{z \to \zeta} |M(z)| = 1$$

and so μ_H is singular respect to Lebesgue's measure on $\mathbb{T}$ (cf. [Rudin, 1979]). □

REMARK 2 *If $|H| \le a < 1$, a.e. then is very easy to check that the Lebesgue's measure of the set $\{\zeta \in \mathbb{T} : |H(\zeta)| = 1\}$ is null and $1/(m_p - \zeta Hn_p) \in L^2$.*

15.6   On covariance's extension problem

First, we state the covariance's extension problem: Given $p \in \mathbb{N}$, and $c_0, c_1, \cdots, c_p$, complex numbers with $c_0 > 0$ and $c_{-k} = \overline{c_k}$, $k = 1, \ldots, p$ find a nonnegative finite measure μ on $\mathbb{T}$ such that

$$\frac{1}{2\pi} \int_0^{2\pi} e^{-ikt} d\mu(t) = c_k, \quad k = -p, \cdots, p. \tag{6.1}$$

The following proposition gives the conditions on $c_0, c_1, \cdots, c_p$, in order that there exists of a nonnegative finite measure μ on $\mathbb{T}$ such that (6.1) is satisfied.

PROPOSITION 15.6.1 *Let* $p \in \mathbb{N}$. *If* $\{c_k\}_{k=-p}^{p} \subseteq \mathbb{C}$ *with* $c_0 > 0$ *and* $c_{-k} = \overline{c_k}$, *such that there exists* μ, *a positive measure absolutely continuous with respect to Lebesgue's measure on* $\mathbb{T}$ *with* $c_k = \widehat{\mu}(k)$, $k = -p, \cdots, p$. *Then* $\{c_k\}_{k=-p}^{p}$ *is strictly positive definite sequence.*

Proof: Since there exists μ a positive measure absolutely continuous with respect to Lebesgue's measure on $\mathbb{T}$ such that $c_k = \widehat{\mu}(k)$, $k = -p, \cdots, p$, for $\{\lambda_n\}_{n=0}^{p} \subset \mathbb{C}$

$$\sum_{n=0}^{p} \sum_{m=0}^{p} \overline{c_{n-m}} \lambda_n \overline{\lambda_m} = \frac{1}{2\pi} \sum_{n=0}^{p} \sum_{m=0}^{p} \lambda_n \overline{\lambda_m} \overline{\int_0^{2\pi} e^{-i(n-m)t} d\mu}$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{p} \lambda_n e^{int} \right|^2 d\mu \geq 0$$

On the other hand, denote $\frac{d\mu}{dt} = f$, where f is a positive Lebesgue's integrable function, let $\{\lambda_n\}_{n=0}^{p} \subset \mathbb{C} - \{0\}$ and assume

$$0 = \sum_{n=0}^{p} \sum_{m=0}^{p} \overline{c_{n-m}} \lambda_n \overline{\lambda_m} = \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{0}^{p} \lambda_n e^{int} \right|^2 f(e^{it}) dt.$$

If $Q(e^{it}) := \sum_{n=0}^{p} \lambda_n e^{int}$ then $|Q(\zeta)|^2 f(\zeta) = 0$, a.e., that is there exists a Lebesgue measurable set A such that $|A| = 1$, (where $|A|$ denotes the Lebesgue measure of A) and $|Q(\zeta)|^2 f(\zeta) = 0$, $\zeta \in A$. Let $X = \{\zeta \in \mathbb{T} : Q(\zeta) = 0\}$, if Q is not the null polynomial then $| X | = 0$ and so, $f(\zeta) = 0$, $\zeta \in A - X$, with $| A - X | = 1$ then $0 < c_0 = \widehat{\mu}(0) = \widehat{f}(0) = 0$. $\square$

The main result of this section is the following theorem. It is important, since characterizes a strictly positive definite sequence as a finite number of Fourier's coefficients of a measure μ which is absolutely continuous with respect to Lebesgue's measure. However, the theorem also gives the Radon-Nikodym derivate of μ, establishing a 1-1 correspondence between the densities and a subset of the open unitary ball of $H^{\infty}(\mathbb{D})$. Finally, we present a factorization formula.

THEOREM 15.6.2 *Let* $p \in \mathbb{N}$ *and* $\{c_n\}_{n=-p}^{p} \subseteq \mathbb{C}$, *the following conditions are equivalent:*

i) $\sum_{n=0}^{p} \sum_{m=0}^{p} \lambda_n \overline{\lambda_m} c_{m-n} > 0$, $\{\lambda_n\}_{n=0}^{p} \subseteq \mathbb{C} - \{0\}$,

ii) There exists a positive Lebesgue's integrable function f on $\mathbb{T}$ such that

$$c_k = \widehat{f}(k), \quad k = -p, \cdots, p.$$

Moreover, given $H \in H^\infty(\mathbb{D})$ such that $\|H\|_\infty \leq 1$, the set $\{\zeta \in \mathbb{T} : |H(\zeta)| = 1\}$ has Lebesgue measure zero and $1/(m_p - \zeta H n_p) \in L^2$, we define

$$f_H(\zeta) = \frac{1}{|m_p(\zeta)|^2} Re \left[\frac{m_p(\zeta) + \zeta H(\zeta) n_p(\zeta)}{m_p(\zeta) - \zeta H(\zeta) n_p(\zeta)} \right], \zeta \in \mathbb{T} \qquad (6.2)$$

then

$$\widehat{f_H}(k) = c_k, \quad k = -p, \cdots, p.$$

Furthermore, the relation (6.2) establishes a bijection between all the power spectrum that solves the covariance extension problem and the $H \in H^\infty(\mathbb{D})$ verifying that the set $\{\zeta \in \mathbb{T} : |H(\zeta)| = 1\}$ has Lebesgue measure zero and $1/(m_p - \zeta H n_p) \in L^2$. Finally, the following factorization formula holds:

$$f_H = |F_H|^2 \text{ for some } F_H \in H^\infty.$$

Proof: As a consequence of proposition 15.6.1 if statement (ii) is valid then (i) is true. Assume that (i) holds and let $V_p : \mathcal{D}_p \longrightarrow \mathcal{R}_p$ be the isometry defined in lemma 15.2.1. Given $H \in H^\infty(\mathbb{D})$ such that $\|H\|_\infty \leq 1$, the set $\{\zeta \in \mathbb{T} : |H(\zeta)| = 1\}$ has Lebesgue measure zero and $1/(m_p - \zeta H n_p) \in L^2$, let $U_H \in \mathcal{L}(\mathcal{F}_H)$ be a minimal unitary extension of V_p associated to H. Clearly, if $k = 0, 1 \cdots p$ and μ^H is the spectral measure of the unitary operator U_H then, as a consequence of the Spectral Theorem

$$c_k = \overline{c_{-k}} = \langle e_0, e_k \rangle_p = \langle e_0, V_p^k e_0 \rangle_p = \langle e_0, U_H^k e_0 \rangle_{\mathcal{F}_H} = \frac{1}{2\pi} \int_0^{2\pi} e^{-ikt} d\mu_H(t)$$

where $\mu_H(t) = \langle \mu^H(t) e_0, e_0 \rangle$. The desired result is a consequence of corollary 15.5.6. The others statements of the theorem can be easily proved.   □

The following corollary shows that the set of all solutions of the Covariance Extension Problem that we have obtained contains strictly the set of all densities in the Wiener class (cf. [Dym, 1988], [Woerdeman, 1989]). The proof is very easy.

COROLLARY 15.6.3 *Given $H \in H^\infty(\mathbb{D})$ such that $\|H\|_\infty \leq 1$, the set $\{\zeta \in \mathbb{T} : |H(\zeta)| = 1\}$ has Lebesgue measure zero and $1/(m_p - \zeta H n_p) \in L^2$, define*

$$f_H(\zeta) = \frac{1}{|m_p(\zeta)|^2} Re \left[\frac{m_p(\zeta) + \zeta H(\zeta) n_p(\zeta)}{m_p(\zeta) - \zeta H(\zeta) n_p(\zeta)} \right], \zeta \in \mathbb{T}.$$

Then, $f_H \in \mathcal{W}$ *if, and only if,*

$$Re\frac{\zeta n_p(\zeta)H(\zeta)}{m_p(\zeta) - \zeta n_p(\zeta)H(\zeta)} \in \mathcal{W}.$$

In 1993, Gabardo(cf. [Gabardo, 1993]) defines the function

$$W_\alpha(e^{i\theta}) = \frac{(1 - |\alpha|^2)\|E_p^\alpha\|_p^2}{|1 - \alpha e^{-i\theta}|^2|E_p^\alpha(e^{i\theta})|^2}$$

where $\alpha \in \mathbb{D}$ and E_p^α is defined as in proposition 15.4.1. He proves that $\widehat{W_\alpha}(k) = c_k$, $k = 0, 1, \cdots, p$. Furthermore, he shows that when $\alpha = 0$ the function W_α maximizes the Burg maximum entropy functional. The following corollary shows that the function W_α can be obtained from (6.2) for some H.

COROLLARY 15.6.4 *Given* $\alpha \in \mathbb{D}$, *the functions* W_α *can be obtain from (6.2), in the particular case, when H is the constant function*

$$H_\alpha(z) = \frac{\overline{\alpha}\overline{n_p(\alpha)}}{\overline{m_p(\alpha)}}$$

Proof: Let $\alpha \in \mathbb{D}$. From proposition 15.4.1 and the Christoffel-Darboux formula we obtain that

$$\|E_p^\alpha\|_p^2 = \langle E_p^\alpha, E_p^\alpha \rangle_p = E_p^\alpha(\alpha) = \frac{|m_p(\alpha)|^2 - |\alpha|^2|n_p(\alpha)|}{1 - |\alpha|^2}$$

and

$$W_\alpha(e^{i\theta}) = \frac{(1 - |\alpha|^2)\frac{|m_p(\alpha)|^2 - |\alpha|^2|n_p(\alpha)|}{1-|\alpha|^2}}{|1 - \overline{\alpha}e^{i\theta}|^2|\frac{\overline{m_p(\alpha)}m_p(e^{i\theta}) - e^{i\theta}\overline{\alpha}\overline{n_p(\alpha)}n_p(e^{i\theta})}{1-\overline{\alpha}e^{i\theta}}|^2}$$

$$= \frac{1 - \frac{|\alpha n_p(\alpha)|^2}{|m_p(\alpha)|^2}}{|m_p(e^{i\theta}) - e^{i\theta}\frac{\overline{\alpha}\overline{n_p(\alpha)}}{\overline{m_p(\alpha)}}n_p(e^{i\theta})|^2}$$

$\square$

If we assume that $1 = c_0, \ldots, c_p$ are the correlations of a second order stationary process $X = \{X_k\}_{k\in\mathbb{Z}}$, then

$$\sum_{j=0}^{p}\sum_{k=0}^{p} a_j\overline{a_k}c_{k-j} = \sum_{j=0}^{p}\sum_{k=0}^{p} a_j\overline{a_k}E(X_j\overline{X_k}) = E\left(\left|\sum_{k=0}^{p} a_kX_k\right|^2\right) > 0 \quad (6.3)$$

that is , the sequence $c_0 = 1, \ldots, c_p$ is strictly positive definite. As a consequence of the previous theorem there exists an integrable function f such that

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} e^{-ikt} f(e^{it}) dt, \quad k = -p, \cdots, p. \tag{6.4}$$

In this case, f is called the spectrum of the process X. According to (6.4) is immediate that the application $X_{-j} \to e_j$, $j \in \mathbb{Z}$ establishes a unitary isomorphism between $H_X = \overline{Span\{X_j : j \in \mathbb{Z}\}}$ and $L^2(fdt)$. Whence,

$$\langle e_j, e_k \rangle_p = c_{k-j} = \langle X_k, X_j \rangle_{H^X} = \langle X_{-j}, X_{-k} \rangle_{H^X} = \langle e_j, e_k \rangle_{L^2(fdt)} \tag{6.5}$$

Let $l, k \in \mathbb{N}$, $k > l$ and $H_{l,k} := Span\{X_{-n}\}_{n=l}^k$ be a subspace of H^X. The innovations are defined by

$$\varepsilon_p = X_{-p} - P_{H_{1,p-1}} X_{-p}, \quad \varepsilon_p^* = X_0 - P_{H_{1,p-1}} X_0,$$

and they verify on account of (6.5) it is readily obtained that

$$\frac{\langle \varepsilon_p, \varepsilon_p^* \rangle_{H^X}}{\langle \varepsilon_p, \varepsilon_p \rangle_{H^X}^{1/2} \langle \varepsilon_p, \varepsilon_p \rangle_{H^X}^{1/2}} = \frac{\langle V_p P_{\mathcal{N}_{p-1}}^{\mathcal{E}_p} e_{p-1}, P_{\mathcal{M}_{p-1}}^{\mathcal{E}_p} e_0 \rangle_p}{\|P_{\mathcal{N}_{p-1}}^{\mathcal{E}_p} e_{p-1}\|_p \|P_{\mathcal{M}_{p-1}}^{\mathcal{E}_p} e_0\|_p}$$

$$= \langle V_p n_{p-1}, m_{p-1} \rangle_p = \gamma_p. \tag{6.6}$$

The last equality is clear that the γ_p are called partial autocorrelation coefficients when they are as in formula (3.2), known as Levinson's algorithm. We set $\sigma_p^2 := \|\varepsilon_p\|_{H^X}^2$ then

$$\sigma_p^2 = \|V_p n_{p-1}\|_p^2 = \frac{1}{\|\Gamma_{p-1}^{-1} e_{p-1}\|_p^2} = \frac{1}{(\widehat{n_{p-1}}(p-1))^2}.$$

Using formula (3.3) it follows that $\sigma_p^2 = \sigma_{p-1}^2(1 - |\gamma_p|^2)$.

15.7   Burg's Entropy

In this section we use the functional model of Arov-Grossman (cf. [Arov, 1983]) to find the density of a second order stationary process that solves the maximum entropy Burg's problem (cf. [Burg, 1975]).

The next theorem gives the solution of the main problem stated in the introduction of this paper.

THEOREM 15.7.1 *Let $p \in \mathbb{N}$ and $\{c_k\}_{k=0}^p$ be the first $(p+1)$ autocorrelations of a second order stationary process $X = \{X_k\}_{k \in \mathbb{Z}}$ then the density f_0 of X which maximizes Burg's functional $\varepsilon(f)$ restricted to the conditions*

$$\frac{1}{2\pi}\int_0^{2\pi} e^{-ikt} f(e^{it})dt = c_k, \quad k = 0, \cdots, p$$

is

$$f_0(e^{it}) = \frac{1}{|n_p(e^{it})|^2}, \ t \in [0, 2\pi]$$

Proof: Let $p \in \mathbb{N}$ and $\{c_k\}_{k=0}^p$, be the first $(p+1)$ autocorrelations of a second order stationary process $X = \{X_k\}_{k\in\mathbb{Z}}$. We use theorem 15.6.2 to conclude that there exists a measure μ, absolutely continuous with respect to the Lebesgue measure on $\mathbb{T}$ that satisfies the conditions

$$\frac{1}{2\pi}\int_0^{2\pi} e^{-ikt} d\mu(t) = c_k, \ k = -p, \cdots, p.$$

Then it has a density $\frac{d\mu}{dt} = f_H$ where the density f_H is the one stated in the last theorem and $H \in H^\infty(\mathbb{D})$ such that $\|H\|_\infty \leq 1$, the set $\{\zeta \in \mathbb{T} : |H(\zeta)| = 1\}$ has Lebesgue measure zero and $1/(m_p - \zeta H n_p) \in L^2$. Therefore, if there exists a maximum of ε it has to be of form $\varepsilon(f_H)$ with H verifying the previous conditions. Thus, we have that

$$f_H(\zeta) = \frac{1}{|m_p(\zeta)|^2} Re\left[\frac{m_p(\zeta) + \zeta H(\zeta)n_p(\zeta)}{m_p(\zeta) - \zeta H(\zeta)n_p(\zeta)}\right]$$

therefore

$$\frac{1}{2\pi}\int_0^{2\pi} \log f_H(e^{it})dt = \frac{1}{2\pi}\int_0^{2\pi}\left(\log Re\left[\frac{m_p(e^{it}) + e^{it}H(e^{it})n_p(e^{it})}{m_p(e^{it}) - e^{it}H(e^{it})n_p(e^{it})}\right]\right.$$

$$\left. - \log|m_p(e^{it})|^2\right)dt$$

$$= \frac{1}{2\pi}\int_0^{2\pi}\log Re\left[\frac{m_p(e^{it}) + e^{it}H(e^{it})n_p(e^{it})}{m_p(e^{it}) - e^{it}H(e^{it})n_p(e^{it})}\right]dt$$

$$+ \frac{1}{2\pi}\int_0^{2\pi}\log f_0(e^{it})dt$$

where $f_0(e^{it}) = \frac{1}{|m_p(e^{it})|^2}$.

>From Jensen's inequality and Cauchy's formula we obtain

$$\frac{1}{2\pi}\int_0^{2\pi}\log Re\left[\frac{m_p(e^{it}) + e^{it}H(e^{it})n_p(e^{it})}{m_p(e^{it}) - e^{it}H(e^{it})n_p(e^{it})}\right]dt$$

$$\leq \log\frac{1}{2\pi}\int_0^{2\pi} Re\left[\frac{m_p(e^{it}) + e^{it}H(e^{it})n_p(e^{it})}{m_p(e^{it}) - e^{it}H(e^{it})n_p(e^{it})}\right]dt$$

$$= \log Re\frac{1}{2\pi i}\int_{|z|=1}\frac{m_p(z) + zH(z)n_p(z)}{m_p(z) - zH(z)n_p(z)}\frac{dz}{z} = 0.$$

□

REMARK 3 *Other entropy functional different to the Burg was used by Gabardo (cf. [Gabardo, 1993]). He proves that if $\mu = f_H + \mu_s$ (μ_s is a singular measure) satisfies (1.1), then*

$$\frac{1}{2\pi}\int_0^{2\pi} \log[f_H(e^{it})]\frac{1-|\alpha|^2}{|1-\alpha e^{-it}|^2}dt \leq \frac{1}{2\pi}\int_0^{2\pi} \log[W_\alpha(e^{it})]\frac{1-|\alpha|^2}{|1-\alpha e^{-it}|^2}dt$$

Another way to characterize the solution of the maximum entropy problem, is the one given by Arocena (cf. ([Arocena, 1990], [Arocena, 1990A]).) We obtain such result as an easy consequence of the fact that if $\widetilde{U} \in \mathcal{L}(\widetilde{F})$ is the minimal unitary extension of the isometry V_p associated to $H \equiv 0$, then,

$$\widetilde{U}^k \mathcal{N}_p \perp \mathcal{E}_p, \quad k \in \mathbb{N}. \tag{7.1}$$

Therefore if $\widetilde{\mu}(t) = \langle\mu(t)e_0, e_0\rangle$, where $\mu(t)$ is the spectral measure of the unitary operator $\widetilde{U}$ then

$$d\widetilde{\mu}(t) = \frac{1}{|n_p(e^{it})|^2}dt.$$

Conversely, if (7.1) is true for a minimal unitary extension U of V_p then U corresponds to $H \equiv 0$.

We know from [Azencott, 1986] that the application $X_{-j} \longrightarrow e_j, j \in \mathbb{Z}$ is a unitary isomorphism from $H_X = \overline{Span\{X_j : j \in \mathbb{Z}\}}$ to $L^2(\mathbb{T}, \mu_X)$, where $\mu_X(t)$ is the spectral measure of the process X. It is a known result that if $\mu_X(t) = f_0(e^{it})dt$, then there exists an autoregressive process $AR(p)$ given by

$$a_0 X_0 + \cdots a_p X_p = \varepsilon_p$$

with $\{\varepsilon_n\}_{n\in\mathbb{Z}}$ a white noise. The latter shows that the maximum entropy solution which is obtained when $H \equiv 0$, and has the form $f_0(e^{it}) = \frac{1}{|n_p(e^{it})|^2}$ and this is the spectral density of an autoregressive process of p order $AR(p)$.

References

R. Arocena, Extensiones Unitarias de Isometrías, Entropía Máxima y Parámetros de Schur de un Proceso Estacionario, Spanish, *Actas III Congreso Latinoamericano: Prob. y Est. Mat.,* (1990), 1-8.

R. Arocena, Schur Analysis of a class of Traslation Invariant Forms, *Lectures notes in pure and applied mathematic,* **122** (1990), 355-369.

D.Z. Arov and L.Z. Grossman, Scattering Matrices in the Theory of Dilations of Isometric Operators, *Soviet Math. Dokl.* **27** (1983), 518-522.

R. Azencott and D. Dacunha-Castelle, *Series of irregular Observations,* Springer-Verlag, (1986).

M. Bakonyi and T. Constantinescu, *Schur's algorithm and several applications,* Pitman Research Notes in Mathematic Series, **261**, Longman Scientific and Technical, Harllow (1992).

J.P. Burg, *Maximun entropy spectral analysis,* PhD Dissertation, Stanford University, Stanford, CA, 1975.

G. Castro, *Coeficientes de Réflexion Géneralisés. Extension de Covariances Multidimensionalles et autres* Applications, PhD Dissertation, Université d'Orsay.

B.S. Choi, On the Relation between the Maximun Entropy Probability Density Function and the Autoregressive Model, *IEEE Trans. Acoust. Speech Signal Process. ASSP-34,* (1986), 1659-1661.

H. Dym, Hermitiam block Toeplitz matrices, orthogonal polynomial, reproducing kernel pontryagin spaces, interpolation and extensión (I. Gohberg, ed.), *Operator Theory: Advances and Applications.* **34** (1988).

C. Foias and A. E. Frazho, *The commutant lifting approach to interpolation problem* (I. Gohberg, ed.), Operator Theory: Advances and Applications, **44** (1990).

J.P. Gabardo, Extension of positive definite distribution and maximun entropy, *Mem. Am. Math. Soc.,* **102,** No. 489, (1993).

T. Kailath, A theorem of I. Schur and its impact on modern signal processing, I. Schur method in operator theory and signal processing (I. Gohberg, ed.), *Op. Theory: Adv. and Appl.* **18** (1986), 9-30.

H. J. Landau, Maximun entropy and the moment problem, *Bull. Am. Math. Soc.* **16,** No. 1, (1987), 47-77.

S.A.M. Marcantognini y M.D. Morán , *El modelo de Arov y Grossman y sus aplicaciones,* Spanish, Decimotercera Escuela Venezolana de Matemática (2000).

S.A.M. Marcantognini, M.D. Morán and A. Octavio, On Nehari's problem for Wiener functions, *Acta Científica Venezolana,* **52,** No 3, (2001), 180-185.

B. Sz-Nagy and C. Foias, *Harmonic Analysis of Operator on Hilbert Spaces,* North-Holland, Amsterdan, (1970).

W. Rudin, *Real and Complex Analysis,* Tata McGraw-Hill, New Delhi, (1979).

I. Schur, On power series which are bounded in the interior of the unit circle I, II, first published in German in 1918-1919, English Translation in I. Schur method in operator theory and signal processing (I. Gohberg, ed.), *Op. theory: Adv. and Appl.* **18** (1986), 31-88.

H.J. Woerdeman, *Matrix and Operator Extension,* PhD Dissertation, Amsterdan (1989)

This page intentionally left blank

RECENT RESULTS IN GEOMETRIC ANALYSIS INVOLVING PROBABILITY

Patrick McDonald

New College of Florida, Sarasota, FL 34243

ptm@virtu.sar.usf.edu

Abstract   We survey recent results in geometric analysis which explicitly involve both the geometry of Riemannian manifolds and probability. We include developments in spectral geometry, the study of isoperimetric phenomena, comparison geometry, minimal varieties, harmonic functions, and Hodge theory.

Keywords: Spectral geometry, isoperimetric conditions, comparison theorems, minimal varieties, harmonic functions, Hodge theory

16.1   Introduction

The first task of a survey concerning results in geometric analysis is to limit the scope of the project by creating a theme which provides a focus and is of interest to a reasonably large audience. The theme which runs throughout this paper can be concisely stated: the material reviewed in this survey explicitly involves both the geometry of (finite dimensional) Riemannian manifolds and probability.

The second task of a survey concerning results which bridge a number of topics is to choose a perspective from which to work. We choose to treat geometric phenomena as primary in our organization of the material. Thus, the paper is broken up into sections, each of which focusses on a specific category of geometric problems. Inside each of these categories we discuss a variety of related probabilistic results.

It is now common knowledge that there are a number of important constructions which tie together analysis, probability and geometry. For example, associated to a Riemannian manifold there is a natural differential operator (the Laplace-Beltrami operator), which is defined in terms of the underlying geometry of the manifold, and in turn serves as the infinitesimal generator for the natural diffusion process on the manifold (Brownian motion). Because the Laplace operator is closely related to the metric, solutions of the fundamental

partial differential equations and boundary value problems (Dirichlet problem, heat equation, etc) and the associated constructions (spectrum, eigenfunctions, etc) contain a great deal of geometric information related to the underlying manifold. Because it is possible to use the path properties of Brownian motion to give probabilistic representations of the objects constructed to study the fundamental boundary value problems, there is hope that the techniques of modern probability can be brought to bear on questions involving the geometry of the underlying manifold. History bears this out; the results which follow are part of this record.

There are many connections between analysis, probability and geometry in addition to those described above. All of these connections are united by a common thread: The metric gives rise to objects belonging to each of the three categories (eg, the Laplace-Beltrami operator, Brownian motion, the Riemann curvature tensor). One moves between categories by constructing identities/inequalities in one category using the objects of another. We have organized the material to reflect this fundamental logic. More precisely, in each of the sections that follow, we define a collection of geometric/analytic problems by reference to a Riemannian metric. Citing relationships between the problems of a given section and modern probability (relationships usually afforded by the metric), we sketch results which occur as corollaries, with implications in both directions.

Given that all results depend on familiarity with the basic construction in each of the categories, we include a short exposition of the material common to all topics. It is hoped that in addition to fixing notation, this exposition makes the paper relatively self-contained. Given that this is a survey, proofs are for the most part omitted, with appropriate references sufficing.

The paper is organized as follows. In section 2 we establish notation that will be used throughout the paper while reviewing the background material in analysis, probability, and geometry. In section 3 we study the geometry of balls and tubes in Riemannian manifolds. Much of section 3 revolves around the study of the asymptotics of exit time moments of Brownian motion, although we also review results invovling cover times and principal curves. In section 4 we review results related to spectral geometry. While much of section 4 is related to the relationship between Dirichlet spectrum and various norms of exit time moments of Brownian motion, we also review material involving coupling techniques and estimates for a variety of problems involving a spectral gap. In section 5 we focus on topics related to isoperimetric phenomena and comparison geometry. Again, we study results involving exit time moments for Brownian motion, as well comparison phenomena involving transience/recurrence of Brownian motion. In section 6 we study minimal varieties (ie varieties which arise as solutions to geometric variational problems). In sec-

tion 7 we review material involving harmonic functions. Much of this section is devoted to results involving the study of Martin boundaries and natural extensions to the theory of harmonic maps. Finally, in section 8 we review work involving Hodge theory.

Because we have chosen to limit the scope and organize the material as sketched above, we do not include many results which could certainly be counted as explicitly involving modern probability and geometric analysis. In particular, we have not included material involving the largely parallel theory of random walks on graphs, nor have we included results which involve the (infinite dimensional) geometry of path spaces. We have not reviewed results using the Malliavin calculus, nor have we included material which involves processes on Euclidean domains when that material does not clearly indicate that there is an underlying geometric phenomena being studied. Most regretably, we have not included material involving index theory where Bismut's probabilistic techniques have led to important results for both the geometry of Riemannian manifolds and the geometry of their loop spaces (for those interested in this material, see the survey [Bismut, 1986], the article [Jones, 1997] and references therein).

As is clear from the outline of the paper, one could devote several volumes to any one of the topics we survey (and others have). This survey is not intended as a comprehensive review of any of the topics, let alone all of the topics. Rather, we have attempted to provide enough information on each topic to give the reader a feel for new results in the context of specific developmental trends. Given limitations of space and time, decisions concerning what material to include must be made. Given imperfect knowledge, there are bound to be, in addition to the choices dictated by our choice of focus and obvious constraints, a number of unintentional sins of ommision for which we apologize in advance.

16.2   Notation and Background Material

Throughout this paper, M will denote a smooth n-dimensional manifold with Riemannian structure g. We will write $C^\infty(M)$ for the space of smooth functions on M. We will denote by TM the tangent bundle of M. As a point set,

$$TM = \bigsqcup_{x \in M} T_x M$$

where $T_x M$ is the space of tangent vectors to M at x, a vector space of dimension n. There is a natural (projection) map $\pi : TM \to M$ which associates to a tangent vector $V \in T_x M$ the point at which it is a tangent vector $\pi(V) = x$. The space TM carries a natural smooth structure for which the projection map is smooth; it is a manifold of dimension $2n$, a vector bundle over M with fiber at $x \in M$ the vector space $T_x M$. Smooth sections of the bundle TM are

smooth maps $s : M \to TM$ which satisfy $\pi(s(x)) = x$. A smooth section of the tangent bundle is just a smooth vectorfield on the manifold *M*.

If $X = (X_1, X_2, \ldots X_n)$ gives local coordinates near a point $x \in M$, the tangent space at x is spanned by $\{\frac{\partial}{\partial X_i}\}_{i=1}^n$ and the cotangent space at x, denoted $T_x^* M$, is the dual space to $T_x M$ and is spanned by $\{dX_i\}_{i=1}^n$ (the collection of objects dual to $\{\frac{\partial}{\partial X_i}\}_{i=1}^n$). We denote the cotangent bundle by $T^* M$; it is constructed as was the tangent bundle as a disjoint union of vector spaces:

$$T^* M \quad = \quad \bigsqcup_{x \in M} T_x^* M$$

The tangent bundle also carries a natural smooth structure; it si a manifold of dimension $2n$.

For $k \leq n$, we will denote by $\Lambda^k T_x^* M$, the *k*th exterior power of $T_x^* M$. If I is a *k*-multinomial, $I = (i_1, i_2, \ldots, i_k)$, and $dX_I = dX_{i_1} \wedge dX_{i_2} \wedge \cdots \wedge dX_{i_k}$, then $\{dX_I : I \text{ increasing}\}$ is a basis of $\Lambda^k T_x^* M$. As in the construction of the tangent bundle, we can endow the disjoint union

$$\Lambda^k M \quad = \quad \bigsqcup_{x \in M} \Lambda^k T_x^* M$$

making it a vector bundle of dimension $\binom{n}{k}$. We denote by $\Omega^k = C^\infty(M, \Lambda^k)$ the smooth sections of the bundle of *k*th exterior powers (the *k*-forms on *M*). Those interested in the details should consult any one of the many references to this material, eg [Dubrovin, 1984].

Given a point $x \in M$, the Riemannian metric g is a nondegenerate quadratic form on the space of tangent vectors at x which varies smoothly in x. We will often write the metric as (g_{ij}) by which we intend to communicate that it can be viewed locally as an $n \times n$ matrix relative to a choice of local coordinates.

Given two Riemannian manifolds (M, g) and (N, h), and a smooth map $\phi : M \to N$, we will denote by $D\phi$ the induced map (derivative) on tangent spaces: $D\phi : T_x M \to T_{\phi(x)} N$. We say that *M* and *N* are isometric if there is a diffeomorphism $\phi : M \to N$ satisfying $g = \phi^* h$, where ϕ^* is the pullback operation:

$$g_x(V_1, V_2) \quad = \quad h_{\phi(x)}(D\phi(V_1), D\phi(V_2)).$$

We say that *M* and *N* are locally isometric if at each point we can find neighborhoods of *M* and *N* which are isometric. We say that a Riemannian manifold is locally flat if it is locally isometric to $\mathbb{R}^n$.

Given a function $f \in C^\infty(M)$ and local coordinates as above, we can define a 1-form by the local formula $df = \sum \frac{\partial f}{\partial X_i} dX_i$. This map, the exterior derivative on functions, is defined similarly on all form bundles and denoted

by $d : \Omega^k \to \Omega^{k+1}$. The adjoint map (defined via the metric) will be denoted by $d^* : \Omega^{k+1} \to \Omega^k$. The Laplace-Beltrami operator is then invariantly defined by

$$\Delta^{(k)} \quad = \quad dd^* + d^*d : \Omega^k \to \Omega^k. \tag{2.1}$$

When the form dimension is understood, we will denote the Laplace-Beltrami operator by Δ.

Acting on functions, with local coordinates as above, the Laplace operator is given in terms of the metric by

$$\Delta \quad = \quad \frac{1}{\sqrt{g}} \sum_{i,j} \frac{\partial}{\partial X_i} \left(\sqrt{g} g^{ij} \frac{\partial}{\partial X_j} \right) \tag{2.2}$$

where g is the determinant of the metric and g^{ij} is the ijth entry of the matrix of the inverse of Riemannian metric (g_{ij}). There is a similar form for the Laplace-Beltrami operator on forms (locally, the Laplace-Beltrami operator on forms is given as a system).

The metric on M induces a volume form, denoted dg, which in turn induces a pairing on the space of compactly supported k-forms. Let $L^2(\Omega^k, dg)$ denote the L^2-completion of the compactly supported k-forms with respect to dg. When M is compact, the Laplace-Beltrami operator is essentially self-adjoint and thus admits a unique self-adjoint extension to $L^2(\Omega, dg)$. When M is not compact, the situation is more complicated. For those interested in the general details the reference [Reed, 1978] provides the requisite functional analysis.

Letting Δ act on the space of compactly supported smooth function on M, denoted $C_0^\infty(M)$, we will denote by $p_M(t, x, y)$ the heat kernel on $(0, \infty) \times M \times M$. We recall that $p_M(t, x, y)$ is the smallest positive solution of the initial value problem

$$\partial_t p_M - \frac{1}{2} \Delta p_M \quad = \quad 0 \text{ on } (0, \infty) \times M \times M$$
$$\lim_{t \to 0^+} p_M(\cdot, y, t) \quad = \quad \delta_y(\cdot) \tag{2.3}$$

where δ_y is the Dirac distribution with mass at y.

As is well known, Brownian motion is the diffusion process with transition densities given by p_M. We will denote by $\mathbb{P}^x$, $x \in M$, the probability measure weighting Brownian paths beginning at x and by $\mathbb{E}^x$ the corresponding expectation operators. We denote by $P_t = e^{-t\Delta}$ the operator semigroup acting on continuous functions on M :

$$P_t f(x) \quad = \quad \int_M p_M(t, x, y) f(y) dg(y)$$

where dg is the metric density. We note that $P_t f(x)$ gives the solution to the Cauchy problem:

$$\partial_t u - \frac{1}{2}\Delta u = 0 \text{ on } (0,\infty) \times M$$
$$\lim_{t \to 0+} u(x,t) = f(x) \text{ on } M. \qquad (2.4)$$

Analogous remarks hold in the case of k-**forms** for the operator semigroup $P_t = e^{-t\Delta^{(k)}}$

Given a domain $D \subset M$ with sufficient boundary regularity, we can construct the heat kernel associated to D, denoted $p_D(x,y,t)$, and an associated Brownian motion on D (Brownian motion absorbed at the boundary). Following Kakutani, we can use properties of Brownian motion to solve the fundamental boundary value problems associated to D. More precisely, let X_t be Brownian motion on M and let τ be the first exit time of X_t from D :

$$\tau = \inf\{t \ge 0 : X_t \notin D\}. \qquad (2.5)$$

If $f \in C^\infty(\partial D)$ and $g \in C^\infty(D)$, then the solution of the Dirichlet problem

$$\frac{1}{2}\Delta u = 0 \text{ on } M$$
$$u(x) = f(x) \text{ on } M \qquad (2.6)$$

is given by

$$u(x) = \mathbb{E}^x f(X_\tau) \qquad (2.7)$$

while the solution of the Poisson problem

$$\frac{1}{2}\Delta u = g \text{ on } D$$
$$u(x) = 0 \text{ on } \partial D \qquad (2.8)$$

is given by

$$u(x) = -\mathbb{E}^x \left[\int_0^\tau g(X_t) dt \right]. \qquad (2.9)$$

More generally, if $c(x)$ is sufficiently regular, the solution of

$$\frac{1}{2}\Delta u - cu = g \text{ on } D$$
$$u = f \text{ on } \partial D \qquad (2.10)$$

is given by the Feynman-Kac formula:

$$u(x) = -\mathbb{E}^x \left[\int_0^\tau g(X_t) \exp\left\{ -\int_0^t c(X_s)ds \right\} dt \right]$$
$$+ \mathbb{E}^x \left[f(X_\tau)) \exp\left\{ -\int_0^\tau c(X_s)ds \right\} \right]. \quad (2.11)$$

There are, of course, similar formulae for the solution of boundary value problems involving the heat operator.

By choosing $g(x) = -1$ in (2.8) and (2.9) we obtain

$$\mathbb{E}^x[\tau] = u(x). \quad (2.12)$$

There are similar expressions for the higher moments given by recursive solution of the Poisson problems: Writing $u_1(x) = u(x)$ for u as in (2.12), let $u_n(x)$ be the solution of

$$\frac{1}{2}\Delta u_n = -n u_{n-1} \text{ on } D$$
$$u_n(x) = 0 \text{ on } \partial D. \quad (2.13)$$

Then, as in [Kinateder, 1998] and [McDonald, 2002],

$$\mathbb{E}^x[\tau^n] = u_n(x). \quad (2.14)$$

There are closely related parabolic results: consider the special case of (2.4) with f taken to be the constant function 1 on the interior of D, 0 on the boundary of D, and the boundary held at 0 for all time. With $p_D(t, x, y)$ the heat kernel and dg the volume form, we set

$$u_D(t, x) = \int_D p_D(t, x, y)dg(y). \quad (2.15)$$

Then $u_D(t, x)$ is the solution to the initial value problem

$$\frac{1}{2}\Delta u_D = \frac{\partial u_D}{\partial t} \text{ on } (0, \infty) \times D$$
$$u_D(x, 0) = \begin{cases} 1 & \text{if } x \in D \\ 0 & \text{if } x \in \partial D \end{cases}$$
$$u_D(t, x) = 0 \text{ if } x \in \partial D. \quad (2.16)$$

In addition, u_D gives the distribution of the exit time:

$$u_D(x, t) = \mathbb{P}^x(\tau > t). \quad (2.17)$$

These observations provide a well-studied means of moving between PDE and probability.

While properly speaking it is the Riemannian metric which defines the category of Riemannian manifolds, it is the Riemannian curvature tensor (which measures the obstruction to the Riemannian manifold being locally isometric to Euclidean space), and the notion of geodesic upon which much interest is focused. Both of these objects are most easily described using the language of connections. We recall the basic facts:

A connection on a manifold M is a differential operator

$$\nabla : C^\infty(M, TM) \times C^\infty(M, TM) \to C^\infty(M, TM)$$

which for any $f \in C^\infty(M)$ satisfies

$$\nabla_{Y_2} f Y_1 \;=\; f \nabla_{Y_2} Y_1 + Y_2(f) Y_1. \tag{2.18}$$

A connection which satisfies

$$\nabla_{Y_1} Y_2 - \nabla_{Y_2} Y_1 \;=\; [Y_1, Y_2] \tag{2.19}$$

is said to be torsion free. Given a Riemannian metric g, a straightforward computation establishes that there always exists a unique torsion free connection, $\nabla : C^\infty(M, TM) \times C^\infty(M, TM) \to C^\infty(M, TM)$, compatible with the metric in the sense that

$$Y_3 \langle Y_1, Y_2 \rangle \;=\; \langle \nabla_{Y_3} Y_1, Y_2 \rangle + \langle Y_1, \nabla_{Y_3} Y_2 \rangle \tag{2.20}$$

where the pairing is defined by the metric. The torsion free connection satisfying (2.20) is called the Levi-Civita connection. The Levi-Civita connection defines, for Y_1, $Y_2 \in C^\infty(M, TM)$, a curvature operator $R(Y_1, Y_2) : C^\infty(M, TM) \to C^\infty(M, TM)$:

$$R(Y_1, Y_2) \;=\; \nabla_{Y_1} \nabla_{Y_2} - \nabla_{Y_2} \nabla_{Y_1} - \nabla_{[Y_1, Y_2]}. \tag{2.21}$$

From (2.21) it is clear that $R(Y_1, Y_2) = -R(Y_2, Y_1)$ and thus the curvature operator is a tensor that takes values in the skew-symmetric endomorphisms of the tangent bundle. The curvature operator defines the Riemann curvature tensor R_{jklm} whose components relative to a basis $\{Y_i\}$ of the tangent space $T_x M$ are given by

$$R_{jklm} \;=\; \langle R(Y_j, Y_k) Y_l, Y_m \rangle \tag{2.22}$$

where once again the pairing is given by the metric.

We can use the Levi-Civita connection to express the Laplacian on k-forms (the Weitzenbock decomposition):

$$\Delta^{(k)} \;=\; \nabla^* \nabla + \mathcal{R}^k \tag{2.23}$$

where $\mathcal{R}^k$, the Weitzenbock curvature term, is given by certain components of the Riemann curvature tensor (for $k = 1$, $\mathcal{R}^1 = \mathbf{Ric}$, the Ricci curvature (2.24)). Such a decomposition was exploited by Bochner to relate the structure of the space of harmonic forms and the underlying geometry and topology of the manifold (cf [Goldberg, 1962] for a variety of examples).

Taking appropriate contractions of the Riemann curvature tensor, we obtain well-studied invariants of the Riemannian metric. For example, the Ricci tensor is the 2-form defined by

$$\mathrm{Ric}(Y_j, Y_l) = \sum_k R_{jklk} \tag{2.24}$$

while the scalar curvature is defined by

$$\mathcal{S} = \sum_k \mathrm{Ric}(Y_k, Y_k). \tag{2.25}$$

The sectional curvature associated to a two-plane in $T_x M$ is given by choosing a spanning set for the two plane, say $\{Y_j, Y_k\}$, and defining

$$K(Y_j, Y_k) = \frac{R_{jkjk}}{\langle Y_j, Y_j \rangle \langle Y_k, Y_k \rangle - (\langle Y_j, Y_k \rangle)^2}. \tag{2.26}$$

Sectional curvature generalizes the notion of Gauss curvature for a surface in three space, and one can recover the Riemann curvature tensor from knowledge of all the corresponding sectional curvatures. The relationship of sectional curvatures to the Ricci curvature is particularly useful: Suppose that $V \in T_x M$ is a unit vector and suppose that $\{Y_i\}_{i=1}^n$ is an orthonormal basis of $T_x M$ with $Y_n = V$. Then, from (2.24),

$$\mathrm{Ric}(V, V) = \sum_{i=1}^{n-1} K(Y_i, V) \tag{2.27}$$

from which we conclude that, for any unit vector V, $\mathrm{Ric}(V, V)/(n - 1)$ is the average of the sectional curvature of all the two-planes containing V.

Given two points x, $y \in M$, we denote by $\mathcal{C}_{xy}$ the collection of smooth curves $\gamma : [0, 1] \to M$ satisfying $\gamma(0) = x$ and $\gamma(1) = y$. In local coordinates we will write $\gamma(t) = (\gamma_1(t), \dots, \gamma_n(t))$. Denoting the tangent vector to γ at t by $\dot{\gamma}(t)$, the length of gamma is given by

$$l(\gamma) = \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle} \, dt$$

where the pairing is given by the metric acting on the tangent space $T_{\gamma(t)} M$. The distance between x and y is defined by

$$\mathrm{dist}(x, y) = \inf_{\gamma \in \mathcal{C}_{xy}} l(\gamma). \tag{2.28}$$

Fixing x, if y is near x, the distance between x and y is realized by a smooth curve $\gamma = \gamma_{xy}$ which minimizes the length function. To obtain γ one can compute the Euler-Lagrange equation associated to the length functional. This gives a system of second order ODEs for the components of γ :

$$\frac{d^2\gamma_k}{dt^2} + \Gamma_{ij}^k \frac{d\gamma_i}{dt}\frac{d\gamma_j}{dt} = 0 \qquad (2.29)$$

where the functions Γ_{ij}^k define the Christoffel symbols. The Christoffel symbols can be written in terms of the connection and, in turn, the Christoffel symbols give a local expression for the connection (cf [Chavel, 1984]). In particular, the Christoffel symbols can be used to define the Riemann curvature tensor:

$$R_{jkl}^i = \frac{\partial \Gamma_{jk}^i}{\partial X_l} - \frac{\partial \Gamma_{jl}^i}{\partial X_k} + \Gamma_{sl}^i\Gamma_{jk}^s - \Gamma_{sk}^i\Gamma_{jl}^s.$$

Returing to (2.29) if we require that the curve be parameterized by arclength, we note that, for small times, the associated initial value problem has a unique solution. This solution is called a geodesic. We say that a Riemannian manifold is *complete* if the (small time) solution of the initial value problem for (2.29) does not explode; that is, the solution of (2.29) exists for all $t \in [0, \infty)$.

Given $x \in M$ and $v \in T_xM$, we will denote the geodesic with initial data (x, v) by $\gamma(x, v, t)$. Given v of small norm, the exponential map, $\exp_x : T_xM \to M$ defined by

$$\exp_x(v) = \gamma(x, v, 1) \qquad (2.30)$$

is a diffeomorphism onto its image. Using the exponential map we obtain an important set of local coordinates (geodesic normal coordinates) defined by

$$(r, \theta) = (\|v\|, v/\|v\|) \longrightarrow \gamma(x, v, 1).$$

It is often the case that computations in geodesic normal coordinates facilitate an understanding of both the analysis and the geometry of a given problem. For example, if we fix $x \in M$ and use geodesic normal coordinates near x we can expand components of the Riemannian metric. For $v \in T_xM$ of small norm and $\{Y_j\}$ an orthonormal basis of T_xM,

$$g_{jk}(v) = \delta_{jk} - \frac{1}{3}\langle R(v, Y_j)v, Y_k\rangle + O(|v|^3) \qquad (2.31)$$

where R is the curvature operator (this exhibits the Riemann curvature tensor as the second order obstruction to the metric being locally Euclidean). Similarly, there is an expression for the volume form:

$$\det(g_{jk}(v)) = 1 - \frac{1}{3}\text{Ric}(v, v) + O(|v|^3) \qquad (2.32)$$

where Ric is the Ricci curvature (this exhibits the Ricci curvature as the second order obstruction to the volume form being locally Euclidean).

16.3   The geometry of small balls and tubes

Let $H \subset \mathbb{R}^n$ be a compact embedded submanifold of $\mathbb{R}^n$ of dimension p and for $\varepsilon > 0$, let $T(H, \varepsilon)$ be the tube of radius ε around H :

$$T(H, \varepsilon) \;=\; \{y \in \mathbb{R}^n : \text{dist}(y, H) \leq \varepsilon\} \tag{3.1}$$

where $\text{dist}(y, x)$ is the Euclidean distance between the points y and x and $\text{dist}(y, H) = \inf_{x \in H} \text{dist}(y, x)$. In a remarkable 1939 paper which arose to address a problem in statistics, Herman Weyl developed a formula for the volume of $T(H, \varepsilon)$ for ε small:

$$\text{Vol}(T(H, \varepsilon)) = \frac{(\pi \varepsilon^2)^{\frac{n-p}{2}}}{\frac{1}{2}(n-p)!} \sum_{j=0}^{[\frac{p}{2}]} \frac{k_{2j}(H) \varepsilon^{2j}}{(n-p+2)(n-p+4) \cdots (n-p+2j)} \tag{3.2}$$

where k_{2j} denote certain curvature invariants of the submanifold H. Weyl's formula inspired a great many developments in geometry, statistics and probability (the book [Gray, 1990] is devoted to the topic). In this section we focus on those developments related to probability.

We begin by noting that there is an invariant description of the tube around H which can be obtained using the normal bundle of H. More precisely, let (M, g) be a Riemannian manifold, $H \subset M$ an embedded compact submanifold of dimension p. Let NH be the normal bundle of H in M, that is, the bundle over H whose fibre at $x \in H$ is the vector space

$$N_x H \;=\; \{v \in T_x M : \langle v, w \rangle = 0 \text{ for all } w \in T_x H\}.$$

Given a point $x \in H$ and a unit tangent vector $v \in N_x H$, the small time solution of the second order ODE for length minimizing curves (see (2.29)) gives a unique geodesic starting at x with tangent vector at x given by v. We denote this geodesic by $\gamma(x, v, t)$, where $0 \leq t \leq \varepsilon(x, v)$. Allowing x to vary in H and v to vary in the unit sphere of $N_x H$, we obtain a family of geodesics, all defined up to some time $\varepsilon > 0$. For ε small enough, the pointset $T(H, \varepsilon)$ defined by

$$T(H, \varepsilon) \;=\; \{y \in M : \exists x \in H, \, v \in N_x H, \, \varepsilon_y < \varepsilon \text{ such that } y = \gamma(x, v, \varepsilon_y)\} \tag{3.3}$$

is open in M and diffeomorphic to the zero section of NH (this is the tubular neighborhood theorem and $T(H, \varepsilon)$ is called a tubular neighborhood of H in M; the corresponding system of coordinates are called Fermi coordinates). In this setting there is a result corresponding to Weyl's formula [Gray, 1981].

16.3.1.   Exit time for Brownian motion

Given that the construction of a tube is completely geometric, it is possible to view Weyls' formula as a special case of a more general program in which one studies the asymptotic behavior of various geometric analogs of "volume" of a tube. This idea was carried out by Gray and Pinsky who studied the behavior of the mean exit time of Brownian motion (integrated over starting points in the given submanifold) from a tube of radius ε. There are by now a number of surveys of this material ([Pinsky, 1991], [Pinsky, 1995]). We sketch the main ideas and a few of the main results when the submanifold is a point.

Thus, let (M, g) be a Riemannian manifold, $x \in M$, and $T(x, \varepsilon)$ the geodesic ball of radius ε centered at x. Let X_t be Brownian motion on M, τ_ε the exit time of Brownian motion from $T(x, \varepsilon)$:

$$\tau_\varepsilon = \inf\{t \geq 0 : X_t \notin T(x, \varepsilon)\}.$$

THEOREM 1 *(cf [Gray, 1983]) As $\varepsilon \to 0^+$,*

$$\mathbb{E}^x[\tau_\varepsilon] = c_0 \varepsilon^2 + c_1 \mathcal{S} \varepsilon^4$$
$$+ \left[c_2 |R_{jklm}| + c_3 |\text{Ric}_{ij}| + c_4 \mathcal{S}^2 + c_5 \Delta \mathcal{S} \right] \varepsilon^6 + O(\varepsilon^8) \quad (3.4)$$

where the constants c_i depend only on dimension, $\mathcal{S}$ is the scalar curvature at x, $|\text{Ric}_{ij}|$ is the norm of the Ricci curvature at x, $|R_{jklm}|$ is the norm of the Riemann curvature at x and Δ is the Laplace operator.

Using expansion (3.4) one has

THEOREM 2 *(cf[Gray, 1983]) Suppose that (M, g) is Riemannian of dimension $n < 6$. Suppose that for all $x \in M$,*

$$\mathbb{E}^x[\tau_\varepsilon] = \frac{1}{n} \varepsilon^2 + O(\varepsilon^8).$$

Then M is locally flat.

The condition $n < 6$ suggests that one can do no better. This is a result of Hughes:

THEOREM 3 *(cf[Hughes, 1992]) Let S^3 be the unit sphere in $\mathbb{R}^4$ and let H^3 be three dimensional hyperbolic space. **Let** (M, g) be the product Riemannian*

manifold given by $S^3 \times H^3$ and let $x \in M$. For any $\varepsilon < \frac{\pi}{k}$, the probability law of τ_ε coincides with the probability law of the exit time of
   Brownian motion from a ball of radius ε in $\mathbb{R}^6$.

The negative result of Theorem 3.3 indicates that to obtain more geometric information from Brownian motion in a small ball, one should consider something other than higher moments. A natural choice is the exit place of Brownian motion. Using the exponential map, there is a simple representation of the exit place distribution as a measure on S^{n-1}. More precisely, we have

THEOREM 4 *(cf [Liao, 1988], [Pinsky, 1995]) Let (M, g) be Riemannian, $x \in M$, and $\exp_x$ the exponential map at x. Suppose that $f : S^{n-1} \to \mathbb{R}$ is a continuous map. Define $S_\varepsilon f(x)$ by*

$$S_\varepsilon f(x) \;=\; \mathbb{E}^x\left[f(\varepsilon^{-1}\exp_x^{-1}(X_{\tau_\varepsilon}))\right].$$

Then, as $\varepsilon \to 0^+$,

$$S_\varepsilon f(x) = \int_{S^{n-1}} \left[1 - \frac{1}{12}\varepsilon^2\left(\text{Ric}_{ij} - \frac{\delta_{ij}S}{n}\right)\theta_i\theta_j\right] f(\theta)d\theta$$

$$+ \frac{1}{24}\varepsilon^3 \int_{S^{n-1}} \left[\frac{\partial \text{Ric}_{ij}}{\partial x_k}\theta_i\theta_j\theta_k - \frac{\theta_k}{n+2}\frac{\partial S}{\partial x_k}\right] f(\theta)d\theta \quad (3.5)$$

where $d\theta$ is Lebesgue measure, Ric_{ij} is the Ricci curvature, and S is the scalar curvature.

This expansion gives the following result:

THEOREM 5 *(cf [Liao, 1988]) Suppose that S_ε and f are as in Theorem 4. Suppose that for all $x \in M$,*

$$S_\varepsilon f(x) = \int_{S^{n-1}} f(\theta)d\theta + O(\varepsilon^2).$$

Then M is Einstein. If, in addition, $\mathbb{E}^x[\tau_\varepsilon] = \frac{\varepsilon^2}{n} + O(\varepsilon^8)$, then M is locally flat.

16.3.2.   Cover times

   Let G be a finite graph, X_n a random walk on G. Define the cover time of G by X_n, denoted T_G, by

$$T_G \;=\; \min\{n : \{X_j\}_{j=0}^n = \text{vertices}(G)\}. \quad (3.6)$$

Cover times appear in a variety of applications in computer science, physics and statistics (for a survey, see [Aldous, 1989]). For many such applications,

understanding how cover time is related to the underlying structure of the walk is an important problem. An example of particular interest is the two dimensional torus $\mathbb{Z}_n^2 = \mathbb{Z}^2/n\mathbb{Z}^2$ with a simple random walk. In this case, there is a conjecture of Aldous (1989) for the asymptotic behavior of the cover time for large n :

$$\lim_{n\to\infty} \frac{T_n}{(n \log n)^2} \;=\; \frac{4}{\pi} \qquad \textbf{almost surely.} \qquad (3.7)$$

This conjecture has recently been settled by Dembo, Peres, Rosen and Zeituni using a careful analysis of Brownian excursion on the two-torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ [Dembo, 2001]. More precisely, suppose that X_t is Brownian motion on $\mathbb{T}^2$ and let $\varepsilon > 0$. Let $B(x,\varepsilon)$ be the ball of radius ε centered at x. Let $T_{x,\varepsilon}$ be the time required for Brownian motion to come within ε of x :

$$T_{x,\varepsilon} \;=\; \inf\{t > 0 : X_t \in B(x,\varepsilon)\}.$$

Let C_ε be the time it takes for Brownian motion to come within distance ε of every point of $\mathbb{T}^2$:

$$C_\varepsilon \;=\; \sup\{T_{x,\varepsilon} : x \in \mathbb{T}^2\}. \qquad (3.8)$$

Thus, C_ε is the time it takes the Wiener sausage (the ε-**tube** around Brownian motion) to cover $\mathbb{T}^2$. The main result of [Dembo, 2001] is the following

THEOREM 6 *([Dembo, 2001]) Let X_t be Brownian motion on $\mathbb{T}^2$. Then*

$$\lim_{\varepsilon\to 0} \frac{C_\varepsilon}{|\log \varepsilon|^2} \;=\; \frac{2}{\pi}. \qquad (3.9)$$

The result generalizes to two-dimensional, compact, connected Riemannian manifolds.

To establish the theorem, the authors control ε-**hitting** using excursions between concentric disks. Their techniques as well as their results are of interest and will be useful for attacking a wide variety of related problems; for example, the *Erdos-Taylor conjecture.*

Given a simple random walk on $\mathbb{Z}^2$ and a point $x \in \mathbb{Z}^2$, let $T_n(x)$ be the number of times that the walk visits x up to time n. Let

$$T_n^* = \max_{x\in\mathbb{Z}^2} T_n(x)$$

be the number of times that the walk visits the most frequently visited position. It is a longstanding conjecture of Erdos and Taylor that

$$\lim_{n\to\infty} \frac{T_n^*}{(\log(n))^2} \;=\; \frac{1}{\pi} \qquad \textbf{almost surely}$$

Using techniques closely related to those developed in [Dembo, 2001], the conjecture is established in [Dembo, 2001 A].

16.3.3.   Principal curves

Suppose that X is a random vector in $\mathbb{R}^2$ with distribution given by a smooth density p. Suppose that $\Gamma \subset \mathbb{R}^2$ is a smooth embedded compact curve and let $\mathrm{dist}(x, \Gamma)$ be the Euclidean distance from $x \in \mathbb{R}^2$ to the curve Γ. Define an exceptional set, E, by

$$E \; = \; \{x \in \mathbb{R}^2 : \exists y_1 \neq y_2 \in \Gamma, \; \mathrm{dist}(x, \Gamma) = \mathrm{dist}(x, y_i)\}.$$

Then E is a set of Lebesgue measure zero. Let $\pi_\Gamma : \mathbb{R}^2 \setminus E \to \Gamma$ be the map which associates to each $x \in \mathbb{R}^2 \setminus E$ the point on Γ nearest to x. A curve Γ is called *principal* for the random variable X if Γ is self-consistent:

$$\mathbb{E}[X|\pi_\Gamma(X) = x] = x$$

for almost every $x \in \Gamma$ Principal curves, first studied by Hastie-Stuetzle [Hastie, 1989], generalize the statistical notion of linear principal components and are designed to give meaning to the idea of a "curve passing through a data set."

Given a random vector X as above one can formulate a natural variational problem for the "best fit principal curve Γ" by minimizing the expected distance squared between Γ and the vector X, ie by minimizing $F(\Gamma)$ where F is given by

$$F(\Gamma) \; = \; \mathbb{E}[\|X - \pi_\Gamma(X)\|^2] \tag{3.10}$$

where the norm is given by the Euclidean distance. Such a program was carried out by Duchamp-Stuetzle [Duchamp, 1996] who computed the corresponding Euler-Lagrange equation for the functional, finding the critical curves are constrained to have their curvatures given in terms of the first and second moments of the induced transverse densities along the normal fibres of the curve Γ. In addition, they found that none of these curves are minima.

It is possible to formulate the notion of principal submanifolds for a random vector in a Riemannian manifold. The corresponding variational problem for the expected distant to the principal submanifold leads to constraints on the curvature components appearing in (3.2) in terms of moments of the induced densities along normal fibers. At present it is unclear whether the notion of a principal submanifold can be used to effectively address problems involving "statistical shape." What is clear is that these calculations give rise to geometric and probabilistic objects which warrant further study.

16.4   Spectral Geometry

Let (M, g) be a closed Riemannian manifold (ie M is compact without boundary), $\Delta : C^\infty(M) \to C^\infty(M)$ the Laplace-Beltrami operator acting on functions. Then Δ is essentially self-adjoint (ie it has a unique self-adjoint extension to $L^2(M, dg)$) and it spectrum is real and nonnegative.   Let

$R(\lambda) = (\Delta - \lambda)^{-1}$ be the resolvent of Δ at λ. By Rellich's theorem $R(\lambda)$ is a compact operator on $L^2(M, dg)$ and it follows from the machinery of functional analysis ([Reed, 1978]) that the spectrum of Δ consists of discrete eigenvalues of finite multiplicity with a unique accumulation point at infinity. We write the spectrum of Δ as

$$\mathrm{spec}(M) \;=\; \{\lambda \in \mathbb{R} : \exists f \in L^2(M, dg),\ \Delta f + \lambda f = 0\}. \quad (4.1)$$

Since the Laplace-Beltrami operator on k-forms can be treated in the same fashion as the Laplace operator on functions, the spectrum of the Laplace-Beltrami operator on k-forms consists of discrete eigenvalues of finite multiplicity with a unique accumulation point at infinity.

When $D \subset M$ is a smoothly bounded domain with compact closure and we impose Dirichlet boundary conditions, it is again true that the spectrum of D, denoted $\mathrm{spec}(D)$, will behave as it does when M is closed. When M is not compact, the behavior of the spectrum of the Laplacian is considerably more involved. The majority of our comments are restricted to the case of smoothly bounded domains with compact closure.

In both the closed case and the case of a smoothly bounded domain, the fundamental problem of spectral geometry can be stated as follows:

> What is the precise relationship between $\mathrm{spec}(M)$ (respectively, $\mathrm{spec}(D)$) and the geometry of M (respectively, D)?

There are a number of good surveys of spectral geometry available (cf [Anderson, 1997], [Bérard, 1986] and references therein, [Bérard, 1986] contains an extensive bibliography for results prior to 1985). In addition, there are a number of texts which discuss the connections between geometry and spectral data (cf [Chavel, 1984], [Schoen, 1994]). We focus on those topics related to probability. Our results fall roughly into two classes: (1) results involving the use of exit time moments to study spectral geometric objects and (2) techniques involving the notion of coupling for studying spectral geometric objects.

16.4.1.   Principal eigenvalue for planar domains and torsional rigidity

Interest in the connection between the geometry of a Euclidean domain and the associated Dirichlet spectrum first arose during the 19th century in studies involving elastic bodies. In these studies the Dirichlet spectrum of a plane domain indexed the allowable modes of vibration of a homogeneous planar membrane with boundary held fixed. For such a model, the first Dirichlet eigenvalue, giving the lowest allowable energy of vibration, plays a special role as it is the dominant factor in studies involving small perturbations of the membrane. Counted among the first results of the field is the conjecture of Rayleigh (later proved by Faber and Krahn - Theorem 16):

THEOREM 7 *Let v be a positive real number. Then for all domains $D \subset \mathbb{R}^2$,*

$$\text{Vol}(D) = v \quad \Rightarrow \quad \lambda_1(D) \geq \lambda_1(B) \tag{4.2}$$

where B is a disk of volume v.

The Raleigh conjecture can be viewed from a variety of perspectives. For the present we note that (4.2) provides a lower bound for the Dirichlet spectrum in terms of geometric data associated to the domain. In this sense the Rayleigh conjecture is prototypical of a great many estimates for the principal eigenvalue (the idea being to bound λ_1 in terms of natural geometric parameters associated to the underlying domain). We review results for which the bounds are probabilistic.

Let X_t be Brownian motion on $\mathbb{R}^2$, let $D \subset \mathbb{R}^2$ be smoothly bounded with compact closure and let $\tau = \tau_D$ be the first exit time from D. Motivated in part by Hayman's bound for λ_1 for planar domains in terms of the inradius of the domain [Hayman, 1978], Banuelos and Carroll prove

THEOREM 8 *(cf[Banuelos, 1994]) Let $D \subset \mathbb{R}^2$ and suppose that τ is the exit time of Brownian motion. Then*

$$\frac{2}{\sup_{x \in D} \mathbb{E}^x[\tau]} \leq \lambda_1 \leq \frac{7\zeta(3)j_0^2}{8\sup_{x \in D} \mathbb{E}^x[\tau]} \tag{4.3}$$

where $\zeta(s)$ is the Riemann zeta-function and j_0 is the first positive zero of the Bessel function of the first type, $J_0(x)$. If σ_D is the Schlict-Landau-Bloch constant of D, then

$$\frac{1}{2\sigma_D^2} \leq \sup_{x \in D} \mathbb{E}^x[\tau] \leq \frac{7\zeta(3)}{8\sigma_D^2}. \tag{4.4}$$

Moreover, the left hand side of (4.3) is sharp.

Inequality (4.3) of Theorem 8 states that λ_1 can be estimated by the L^∞-norm of $\mathbb{E}^x[\tau]$ (and inequality (4.4) indicates that there are geometric estimates for the L^∞-norm of $\mathbb{E}^x[\tau]$). There are similar statements for all L^p-norms of $\mathbb{E}^x[\tau]$, denoted $\|\mathbb{E}^x[\tau]\|_p$, as well as estimates involving the higher moments of τ. It should be clear that these norms are all geometric invariants; they do not change under the action of the isometry group of the ambient space.

The L^1-norm of the first moment of the exit time plays an interesting role in the theory. Historically, interest in the $\|\mathbb{E}^x[\tau]\|_1$ first arose in the 19th century, again in the theory of planar elastic bodies, where it is proportional to the *torsional rigidity* associated to a homogeneous cylinder with defining cross section D. The St. Venant Torsion Conjecture, first proved by Polya [Polya, 1948], gives a natural geometric bound:

THEOREM 9 *Let v be a positive real number. Then for all domains $D \subset \mathbb{R}^2$,*

$$\text{Vol}(D) = v \quad \Rightarrow \quad \|\mathbb{E}^x[\tau_D]\|_1 \leq \|\mathbb{E}^x[\tau_B]\|_1 \tag{4.5}$$

where B is a disk of volume v.

The Torsion Conjecture inspired a great deal of analysis and the corresponding literature is extensive (cf [Bandle, 1986], [Iesan, 1980], and references therein). The vast majority of the literature is written from the point of view of elastica. Thus, there are a variety of techniques and results for dealing with torsional rigidity which may be brought to bear on problems involving L^1-norms of exit time and vice-versa. We provide an example concerning the fundamental result of [Serrin, 1971] in the section 6 below.

16.4.2. Dirichlet spectrum for domains with compact closure in complete Riemannian manifolds and exit time moments

Suppose that M is a complete Riemannian manifold, $D \subset M$ a smoothly bounded domain with compact closure. Let τ be the exit time of Brownian motion from D. For $\lambda \in \text{spec}(D)$, let $\mathcal{E}_\lambda(1)$ denote the projection of the constant function 1 on the eigenspace of the Dirichlet Laplacian corresponding to λ. Set

$$a_\lambda^2 \ = \ \int_D |\mathcal{E}_\lambda(1)|^2 dg \tag{4.6}$$

where dg is the volume form associated to the metric g. Let

$$\text{spec}^*(D) \ = \ \{\lambda \in \text{spec}(D) : a_\lambda^2 \neq 0\} \tag{4.7}$$

and define

$$\text{vp}(D) \ = \ \{a_\lambda^2 : \lambda \in \text{spec}^*(D)\}. \tag{4.8}$$

Then $\text{vp}(D)$ describes how the volume of the domain D is partitioned amongst eigenspaces and, in particular, $\sum_{\lambda \in \text{spec}(D)} a_\lambda^2 = \text{Vol}(D)$. Moreover, denoting the L^1-norm of the kth moment of the exit time by

$$\|\mathbb{E}^x[\tau^k]\|_1 \ = \ \int_D \mathbb{E}^x[\tau^k] dg, \tag{4.9}$$

we have (cf [McDonald, (to appear)])

$$\|\mathbb{E}^x[\tau^k]\|_1 \ = \ \Gamma(k+1) \sum_{\lambda \in \text{spec}^*(D)} a_\lambda^2 \left(\frac{2}{\lambda}\right)^k \tag{4.10}$$

where Γ is the gamma-function (in fact, (4.10) holds for all real $k > 0$). A straightforward computation gives the estimate

$$2\left(\frac{k!a_{\lambda_1}^2}{\|\mathbb{E}^x[\tau^k]\|_1}\right)^{\frac{1}{k}} \le \lambda_1 \le 2\left(\frac{k!\text{Vol}(D)}{\|\mathbb{E}^x[\tau^k]\|_1}\right)^{\frac{1}{k}} \qquad (4.11)$$

Estimate (4.11) holds for arbitrary compact manifolds with nonempty boundary and suggest that in this context, the L^1-**norms** of the exit time moments behave like the reciprocal of the principal eigenvalue. This observation is consistent with the relationship between Theorem 7 and Theorem 9, as well as with the results of Theorem 8. The same relationship appears for a great number of *comparison geometry* results and will be developed below (cf Theorem 22). That the relationship holds also provides a means of studying the behavior of the first Dirichlet eigenvalue using techniques developed for studying first exit time moments. We provide an example:

Let $p_D(t, x, y)$ be the heat kernel associated to D, $\{\phi_\lambda\}$ a complete set of orthonormal eigenfunctions for the Dirichlet Laplacian. Write

$$p_D(t, x, y) = \sum \phi_\lambda(x)\phi_\lambda(y)e^{-\lambda t}. \qquad (4.12)$$

Let dg be the volume form and, as in (2.15), let $u_D(t, x)$ be defined by

$$u_D(t, x) = \int_D p_D(t, x, y)dg(y). \qquad (4.13)$$

Then $u_D(t, x)$ is the distribution of the exit time

$$u_D(t, x) = \mathbb{P}^x(\tau > t) \qquad (4.14)$$

and using (4.12), (4.13) and (4.14) we see that the first Dirichlet eigenvalue characterizes large deviations of τ :

$$\mathbb{P}^x(\tau > t) \simeq Ce^{-\lambda_1 t}. \qquad (4.15)$$

When D is a small geodesic ball of radius ε, this observation and the corresponding analysis of the small ε asymptotics of the first exit time led Karp and Pinsky to the small ε asymptotics for the first Dirichlet eigenvalue. More precisely,

THEOREM 10 *(cf [Karp, 1987]) Suppose that (M, g) is a Riemannian manifold, that $x \in M$ and that $B_M(x, \varepsilon)$ is a geodesic ball of radius ε centered at x. Let $\lambda_1(x, \varepsilon)$ be the corresponding first Dirichlet eigenvalue. Then, as $\varepsilon \to 0^+$, there is an expansion of the form*

$$\lambda_1(x, \varepsilon) = c_0\varepsilon^{-2} + c_1\mathcal{S}$$
$$+ c_2\left[|R_{jklm}| - |\text{Ric}_{ij}|^2 + 6\Delta\mathcal{S}\right]\varepsilon^2 + O(\varepsilon^4) \quad (4.16)$$

where the constants c_i depend only on dimension, $\mathcal{S}$ is the scalar curvature at x, $|\text{Ric}_{ij}|$ is the norm of the Ricci curvature at x, $|R_{jklm}|$ is the norm of the Riemann curvature at x and Δ is the Laplace operator.

If M is compact, one can consider the asymptotics of the Dirichlet spectrum for the complement of a small ball, $M \setminus B_M(x, \varepsilon)$. This problem was studied probabilistically by Kac [Kac, 1974], who considered the first time Brownian motion hits the small ball and obtained partial results on the asymptotics of the jth eigenvalue. These results were refined by Chavel and Feldman [Chavel, 1988]. The problem continues to define an active area of research.

Returning to the study of moments, we will write

$$\mathbf{mspec}(D) = \{\|\mathbb{E}^x[\tau^k]\|_1\}_{k=0}^{\infty}.$$

Then (4.10) says that the set $\mathbf{spec}^*(D) \cup \mathbf{vp}(D)$ determines the set mspec(D). It turns out that the converse is also true:

THEOREM 11 *(cf [McDonald, (to appear)]) Suppose D, D' are smoothly bounded domains with compact closure in M. Then*

$$\mathbf{mspec}(D) = \mathbf{mspec}(D') \Rightarrow \mathbf{spec}^*(D) = \mathbf{spec}^*(D')$$
$$\text{and } \mathbf{vp}(D) = \mathbf{vp}(D'). \quad (4.17)$$

The proof of this result uses the solution of the classical Stieltjes moment problem [Akhiezer, 1965] and suggests that the techniques developed in the context of the moment problem might be useful in the context of spectral geometry.

As a corollary of Theorem 11, we obtain that the first Dirichlet eigenvalue is determined by mspec(D).

COROLLARY 12 *([McDonald, (to appear)]) Let $D \subset M$ be a smoothly bounded domain with compact closure. Let $\mathbf{spec}^*(D) = \{\nu_i\}_{i=1}^{\infty}$ enumerate elements of $\mathbf{spec}^*(D)$ in increasing order. Then*

$$\nu_1 = \sup\left\{\nu \geq 0; \limsup_{n\to\infty}(\nu/2)^n \frac{\|\mathbb{E}^x[\tau^n]\|_1}{\Gamma(n+1)} < \infty\right\} \quad (4.18)$$

and

$$a_{\nu_1}^2 = \limsup_{n\to\infty}(\nu_1/2)^n \frac{\|\mathbb{E}^x[\tau^n]\|_1}{\Gamma(n+1)}. \quad (4.19)$$

In fact, from Corollary 12 and (4.10) it is clear that the tail of the moment spectrum gives a recursion for the elements of $\mathbf{spec}^*(D)$ and $\mathbf{vp}(D)$ (cf [McDonald, (to appear)]).

Given (4.19), it is clear that the exit time moments are closely tied to the integrals of normalized eigenfunctions. Such objects have received attention for

a variety of reasons, including their relationship to asymptotics for the spectral counting function, the asymptotics of the spectral heat function, and the *heat content asymptotics* of D. Focussing our attention on heat content, we recall the neccesary facts:

Let $u_D(t, x)$ be as defined in (4.13). Then $u_D(t, x)$ is the solution of the initial value problem

$$
\begin{aligned}
\frac{1}{2}\Delta u_D &= \frac{\partial u_D}{\partial t} \text{ on } (0, \infty) \times D \\
u_D(x, 0) &= \begin{cases} 1 & \text{if } x \in D \\ 0 & \text{if } x \in \partial D \end{cases} \\
u_D(t, x) &= 0 \text{ if } x \in \partial D.
\end{aligned} \tag{4.20}
$$

Let $q(t)$ be the heat content of D at time t:

$$
q(t) = \int_D u_D(t, x)dg. \tag{4.21}
$$

We note that $q(t)$ is the Laplace-Stieltjes transform of the spectral heat function, $h : \mathbb{R}^+ \to \mathbb{R}^+$ defined by

$$
h(\sigma) = \sum_{\lambda \in \text{spec}(D),\ \lambda < \sigma} a_\lambda^2
$$

where a_λ^2 is as in (4.6). Using a Tauberian theorem, van den Berg and Watson have determined the first two terms in an asymptotic expansion of $h(\sigma)$ and used this to obtain an estimate on the rate at which the a_λ^2 converge to zero [van den Berg, 1999A].

It is a theorem of van den Berg and Gilkey [van den Berg, 1994] that $q(t)$ admits a small time asymptotic expansion:

$$
q(t) \simeq \sum_{n=0}^{\infty} q_n t^{n/2} \tag{4.22}
$$

where the coefficients q_n are locally computable geometric invariants of D (that is, every q_n is given as an integral over the boundary of the domain or an integral over the interior of the domain of a finite number of derivatives of components of the Riemannian metric). We will refer to the coefficients occuring on right hand side of (4.22) as the *heat content asymptotics of D* and we write

$$
\text{hca}(D) = \{q_n\}_{n=0}^{\infty}. \tag{4.23}
$$

We note that the invariants hca(D) are *not* spectral.

The probabilistic study of heat content is by now well developed in a variety of contexts (piecewise smooth domains, fractals domains, etc) and the identification of a number of the coefficients in the expansion has been carried out (cf [van den Berg, 1994A], [van den Berg, 1994]; cf [Gilkey, 1999] for a recent survey of results concerning heat content). For example, it is known that the first coefficient is given by the volume of the domain (this is clear from (4.20) and (4.21)), while the second coefficient is given by a constant multiple of the area of the boundary of the domain, suggesting that heat content might be useful in the study of isoperimetric phenomena (cf section 5.1 below and [Burchard, 2002]). For polygonal domains, it is known that the asymptotics terminate after 3 terms (cf [Burchard, 2002]); it would be interesting to know whether similar phenomena exist in higher dimensions.

From Corollary 12 it is clear that heat content is closely related to mspec(D). We have:

THEOREM 13 *([McDonald, (to appear)]) Let M be a complete Riemannian manifold, $D \subset M$ a smoothly bounded domain with compact closure. Then* mspec(D) *determines* $q(t)$ *(and thus* hca(D)*)*

Using Theorem 4.4 and Theorem 4.5, we see that $\mathbf{spec}^*(D) \cup \mathbf{vp}(D)$ determines hca(D), a geometric result proved via the analysis of a probabilistic object (mspec(D)). This result suggests that the invariants mspec(D) may be useful tools in studying questions involving the fine structure of isospectral domains. To formulate a more precise statement, we again recall the basic facts:

In his often cited 1965 paper, Mark Kac popularized a fundamental problem of planar spectral geometry: Does spec(D) determine D up to isometry? The problem was settled (at least in the piecewise smooth category) by Gordan, Webb, and Wolpert [Gordon, 1992], who constructed a pair of nonisometric, isospectral planar polygons. In 1994 Buser, Conway, Doyle and Semmler [Buser, 1994] gave an elegant and straightforward construction of families of isospectral nonisometric planar polygonal pairs (we will abbreviate reference to such pairs by INIPP). Their constructions include a simplified version of the example of [Gordon, 1992] as a special case, as well as the first example of a pair of isospectral planar domains all of whose normalized eigenfunctions agree at a pair of distinguished interior points (so called *homophonic domains*). These examples are generated by a "seed" triangle together with a collection of congruent "reflection progeny" triangles produced by a sequence of reflections across edges. In particular, the construction is essentially combinatorial and by focussing on the vertices and edges of the corresponding triangles, the construction can be taken to occur in the category of planar graphs.

One might summarize the work of [Buser, 1994] by saying that, for piecewise smooth planar domains, the Dirichlet spectrum provides an incomplete collection of geometric invariants. Such a summary suggests that to construct

a good collection of geometric invariants, one might be well served by finding invariants which distinguish INIPPs. In [McDonald, (to appear)A] we show that in the category of weighted graphs and their associated combinatorial Laplacians, there exist natural weighted graph analogs of INIPPs which are isospectral but not isomorphic, and that these graph pairs are distinguished by their heat content asymptotics (and thus by their moment spectra). The natural conjecture is that heat content distinguishes the isospectral domains of [Buser, 1994].

16.4.3.   Spectral gap and coupling

In the previous two subsections we have considered results which involve exit time moments of Brownian motion and the Dirichlet spectrum. In this section we consider estimates of the spectral gap obtained via coupling methods. We begin by recalling the requisite material involving spectral gaps.

For clarity of exposition, suppose that a_{ij}, $b_i : \mathbb{R}^n \to \mathbb{R}$ are smooth with (a_{ij}) positive definite as an $n \times n$ matrix. Suppose there is a smooth function V satisfying

1 $b_i = \sum_j a_{ij} \frac{\partial V}{\partial X_j} + \frac{\partial a_{ij}}{\partial X_j}$

2 $\int \exp(V(x)) dx < \infty$.

Let

$$L = \sum_{i,j} a_{ij}(x) \frac{\partial}{\partial X_i} \frac{\partial}{\partial X_j} + \sum_i b_i(x) \frac{\partial}{\partial X_i}. \qquad (4.24)$$

Let μ be the measure defined by

$$d\mu = \frac{\exp(V(x))}{\int \exp(V(x)) dx} dx$$

and note that L is symmetric with respect to the measure μ. Let $\|f\|$ be the norm of f in $L^2(\mathbb{R}^n, d\mu)$.

Let $P_t = e^{-tL}$ be the heat operator for L. Fixing f in the domain of L, for ε small we have

$$\left\| P_t f - \int f d\mu \right\| \leq \left\| f - \int f d\mu \right\| e^{-\varepsilon t} \qquad (4.25)$$

for all $t \geq 0$. We are interested in studying the maximal ε for which (4.25) holds (ie the rate at which $P_t f$ converges to $\int f d\mu$ in L^2). To this effect, we define the spectral gap associated to L by the variational principle

$$\mathrm{sg}(L) = \inf \left\{ - \langle f, Lf \rangle ; f \in L^2(\mathbb{R}^n, d\mu), \int f d\mu = 0, \|f\| = 1 \right\}. \qquad (4.26)$$

Under mild assumptions on L (eg the Dirichlet form is regular), it follows that for all $\varepsilon < \text{sg}(L)$, (4.25) holds.

It should be clear that the development sketched above can be carried out in the context of ambient spaces other than $\mathbb{R}^n$. It is also the case that analogous statements hold for processes which are not diffusions (eg general reversible Markov processes [Chen, 1994A]).

If we restrict our attention to the Dirichlet Laplacian on a compact Riemannian manifold, it is clear that the spectral gap coincides with the first nonzero Dirichlet eigenvalue, the variational principle being equivalent to the Raleigh quotient. Thus, general results for estimates of the spectral gap give rise to estimates for principle eigenvalues. It is in this context that we develop the notion of coupling.

Coupling was originally introduced by Doeblin [Doob, 1983] to study the rate of convergence to stationarity of a Markov chain. Lindvall is responsible for adapting coupling techniques to Brownian motion (cf [Linvall, 1983], [Linvall, 1986]). There are a number of surveys of coupling techniques available (eg [Brin, 2001]) as well as a text ([Linvall, 1992]). We recall the basic facts:

Again, for clarity of exposition let X_t be a diffusion process on $\mathbb{R}^n$ with generator the operator L given in (4.24). By a coupling for the process X_t we mean two copies of the process, denoted by (X_t^1, X_t^2), which are taken to begin at different points. More precisely, the processes X_t^1 and X_t^2 have the same distribution as X_t and the processes X_t^1, X_t^2, and (X_t^1, X_t^2) are all Markov with respect to the filtration generated jointly by X_t^1 and X_t^2. Define the coupling time, T, by

$$T \;=\; \inf\{t \geq 0 : X_t^1 = X_t^2\}. \tag{4.27}$$

Suppose that

1 it is possible to construct X^1 and X^2 such that for all $t \geq T$, $X_t^1 = X_t^2$.

2 there is a constant v such that for generic starting points x_1, x_2, $\mathbb{P}(T > t | X_t^1 = x_1, X_t^2 = x_2) \simeq e^{-\nu t}$.

Then one can prove that v is a lower bound for $\text{sg}(L)$.

Thus, to apply coupling to estimate the sepctral gap we must check the above and arrange for v to be close to $\text{sg}(L)$ (ie the coupling should be *efficient* in the language of [Brin, 2001]). This program has been carried out in a number of interesting geometric contexts in which it produces general lower bounds on the spectral gap (cf [Chen, 1997], [Chen, 1994]). We restrict our attention to examples of special interest; those involving the Dirichlet spectral gap and coupling.

Suppose that M is a complete Riemannian manifold, $D \subset M$ a smoothly bounded domain with compact closure. Let $\{\phi_\lambda\}_{\lambda \in \text{spec}(D)}$ be a complete orthonormal family of eigenfunctions for the Dirichlet Laplacian and write the

heat kernel as

$$p_D(t, x, y) = \left[\phi_{\lambda_1}(x)\phi_{\lambda_1}(y) + \sum_{\lambda \neq \lambda_1} \phi_\lambda(x)\phi_\lambda(y)e^{-(\lambda-\lambda_1)t} \right] e^{-\lambda_1 t}.$$

We consider the Dirchlet spectral gap $\lambda_2 - \lambda_1$.

There is a long history of estimates for the Dirichlet spectral gap in terms of the underlying geometry of the domain. When $M = \mathbb{R}^n$ and D is a convex regular domain with diameter d, Singer, Wong, Yau and Yau [Singer, 1985] established

$$\lambda_2 - \lambda_1 \geq \frac{\pi^2}{4d^2}. \tag{4.28}$$

On the other hand, when D is a rectangle it is easy to check that

$$\lambda_2 - \lambda_1 \geq \frac{3\pi^2}{d^2} \tag{4.29}$$

and thus one expects improvements of the [Singer, 1985] estimate (4.28). For Euclidean domains as above, such an improvement was given by Yu-Zhang [Yu, 1986] who established the estimate

$$\lambda_2 - \lambda_1 \geq \frac{\pi^2}{d^2}. \tag{4.30}$$

Realizing that the Dirichlet spectral gap can be considered as the first eigenvalue of Brownian motion conditioned to remain forever in the domain, R. Smits [Smits, 1996] gave a second (probabilistic) proof of the estimate (4.30). Combining the ideas of Smits, comparison and the powerful general estimates of [Chen, 1997], Wang has considered the analog of the problem for general ambient manifolds. His recent results [Wang, 2000] recover and improve the known results involving Dirichlet spectral gaps and suggest that the technique will continue to produce improvements and new directions for further research.

16.5   Isoperimetric Conditions and Comparison Geometry

The Rayleigh Conjecture (Theorem 7 above) was established in the early twentieth century by Faber and by Krahn who both realized that the feature of fundamental importance in establishing a proof is the isoperimetric property of planar domains (among planar domains of fixed area, a disk has minimum perimeter). The first rigorous proof of the isoperimetric property for Euclidean domains was given by Steiner in the nineteenth century using rearrangement techniques pioneered for just this purpose. That the isoperimetric property

holds for domains when the ambient space is a Euclidean sphere or hyperbolic space was established by Schmidt [Schmidt, 1943]. Using Euclidean space, Euclidean spheres, and hyperbolic space as models we can study analogs of the isoperimetric property and other geometric phenomena in more general ambient spaces.

16.5.1.   Isoperimetric phenomena and moments of exit times

We begin by formalizing our notion of a model:

DEFINITION 14 *Let κ be a real number. The constant curvature space form with curvature κ, denoted M_κ, is*

1 A sphere in Euclidean space if $\kappa > 0$,

2 Euclidean space if $\kappa = 0$,

3 A hyperbolic space if $\kappa < 0$.

DEFINITION 15 *Suppose that M is a Riemannian manifold. We say that M satisfies an isoperimetric condition with constant curvature comparison space M_κ if, for all Borel $D \subset M$,*

$$\mathrm{Vol}(D) = v \quad \Rightarrow \quad \mathrm{Area}_M(\partial D) \geq \mathrm{Area}_{M_\kappa}(\partial B)$$

where $B \subset M_\kappa$ is a geodesic ball of volume v and "Area" denotes the Minkowski measure induced by the corresponding Riemannian metrics.

We note that there is a great deal of literature devoted to determining precise regularity requirements for isoperimetric phenomena. For the purpose of this section, all domains are taken to be smoothly bounded unless otherwise indicated. In this case, all reasonable definitions of area will coincide.

We can now state the result of Faber-Krahn:

THEOREM 16 *Suppose that M is a Riemannian manifold which satisfies an isoperimetric condition with constant curvature comparison space M_κ. Then, for all $D \subset M$,*

$$\mathrm{Vol}(D) = v \quad \Rightarrow \quad \lambda_1(D) \geq \lambda_1(B)$$

where $B \subset M_\kappa$ is a geodesic ball of volume v and λ_1 is the first Dirichlet eigenvalue.

The proof of Theorem 16 uses *symmetric rearrangement*. As this will play a role in much of this section, we recall the basic facts.

Given $D \subset M$, a Borel set of finite volume, we denote by D^* the ball in M_κ (centered at an appropriate origin) of volume equal to that of D. Suppose that $f : D \to [0, \infty)$ and suppose that the positive level sets of f all have finite volume. Suppose $\mu > 0$, and let

$$D(\mu) = \{y \in D : f(y) \geq \mu\}.$$

We define the spherically symmetric decreasing rearrangement of f, denoted $f^* : D^* \to [0, \infty)$, as the radial function

$$f^*(|x|) \;=\; \sup\{\mu : x \in D(\mu)^*\}$$

It follows from the definition that f and f^* are equimeasurable. It follows from the co-area formula (see [Chavel, 1984], [Chavel, 2001]) that symmetric rearrangement is nonincreasing for the H^1-**Sobolev** norm. Applying this to the Rayleigh quotients which compute the first Dirichlet eigenvalue, we have

$$\frac{\int_D |\nabla f|^2 dg}{\int_D f^2 dg} \;\geq\; \frac{\int_{D^*} |\nabla f^*|^2 dg_\kappa}{\int_{D^*} (f^*)^2 dg_\kappa}$$

from whence Theorem 16 follows.

The results of the previous section (Theorem 8, (4.11)) suggest that the L^p-norms of the exit time moments behave like the reciprocal of the principal Dirichlet eigenvalue. This suggests the following analog of the Faber-Krahn result:

THEOREM 17 *Suppose that M is a Riemannian manifold which satisfies an isoperimetric condition with constant curvature comparison space M_κ. Let τ be the first exit time of Brownian motion. Then, for all $D \subset M$, for all $k \in \mathbb{N}$,*

$$\mathrm{Vol}(D) = v \;\Rightarrow\; \|\mathbb{E}^x(\tau_D^k)\|_p \leq \|\mathbb{E}^x(\tau_B^k)\|_p$$

where $B \subset M_\kappa$ is a geodesic ball of volume v.

This theorem is essentially due to Aizenman and Simon in the Euclidean case [Aizenman, 1982] (see also [Kinateder, 1998]). The general result can be found in [McDonald, 2002].

In fact, the argument of Aizenman-Simon establishes a more general conclusion than the estimate on moments. Their precise theorem is

THEOREM 18 *([Aizenman, 1982]) Suppose that D is a domain in $\mathbb{R}^n$ of finite volume and let τ be the exit time of Brownian motion. Suppose that $f : [0, \infty) \to \mathbb{R}$ is nonnegative and nondecreasing. Then, if D^* is the ball centered at the origin with the same volume as D, we have*

$$\mathbb{E}^x[f(\tau_D)] \leq \mathbb{E}^0[f(\tau_{D^*})].$$

The proof of this result uses a deep result of Brascamp, Lieb, and Luttinger [Brascamp, 1974] involving symmetric rearrangement of multiple integrals. The result of Brascamp, Lieb and Luttinger and the theorem of Aizenman and Simon have been further refined by Burchard and Schmuckenschläger. Using rearrangement techniques at the level of Brownian paths and a Trotter product formula, they prove

THEOREM 19 *(cf [Burchard, 2002]) Let M_κ be a constant curvature space form, $D \subset M_\kappa$ a Borel set of finite volume, D^* an open disk of volume equal to that of D. Let τ_D be the exit time of Brownian motion and let $u_D(t,x) = \mathbb{P}^x(\tau_D > t)$. Then, for all $t > 0$, the exit time from D is dominated by the exit time from D^* in the sense that for every convex increasing function F,*

$$\int_D F(u_D(t,x))dx \quad \leq \quad \int_{D^*} F(u_{D^*}(t,x))dx \tag{5.1}$$

where dx is uniform measure. In particular, if x^* is the center of the disk D^*, then

$$\sup_{x \in D} u_D(t,x) \quad \leq \quad u_{D^*}(t,x^*). \tag{5.2}$$

Equality in (5.1) when $F(u_D(t,x))$ is nonconstant or equality in (5.2) occurs if and only if there is a ball B where $D \backslash B$ has zero volume and $B \backslash D$ is polar.

16.5.2.   Comparison and exit time moments

The structure of Theorem 16 can be abstracted to the following form: given a geometric restriction on a Riemannian manifold (ie it satisfies an isoperimetric condition), the geometry is further constrained (ie there is a lower bound on the principal eigenvalue of any domain of a given volume). Such structure defines those results which comprise the field of *Comparison Geometry*. There are a number of such comparison results which involve probability.

We begin with a result of Debiard, Gaveau and Mazet [Debiard, 1976] who use path properties of Brownian motion to prove

THEOREM 20 *(cf [Debiard, 1976]) Suppose that M is a Riemannian manifold with sectional curvatures denoted by K. Suppose that $x_0 \in M$ and that ρ_0 is a positive constant that is less than the injectivity radius of M at x_0. Let $B_M(x_0, \rho_0)$ be the geodesic ball of radius ρ_0 centered at x_0. Let $B_\kappa(x'_0, \rho_0)$ be the geodesic ball of radius ρ_0 in the constant curvature space form M_κ centered at some origin x'_0. Let $p_B(t, x_0, x)$ be the heat kernel on $B_M(x_0, \rho_0)$ and denote by $p_\kappa(t, r)$ the heat kernel on $B_\kappa(x'_0, \rho_0)$, where r is the distance from the origin x'_0 to the second variable. Then, $\forall t > 0$, $x \in B_M(x_0, \rho_0)$,*

$$K \leq \kappa \quad \Rightarrow \quad p_B(t, x_0, x) \leq p_\kappa(t, \text{dist}(x_0, x)) \tag{5.3}$$

$$K \geq \kappa \quad \Rightarrow \quad p_B(t, x_0, x) \geq p_\kappa(t, \text{dist}(x_0, x)). \tag{5.4}$$

There is a corresponding result for Ricci curvature due to Cheeger-Yau

THEOREM 21 *(cf [Cheeger, 1981]) Suppose that M is an n-dimensional Riemannian manifold with Ricci curvatures denoted by* Ric. *With the notation of Theorem 20,* $\forall t > 0,\ x \in B_M(x_0, \rho_0),$

$$\text{Ric} \geq (n-1)\kappa \quad \Rightarrow \quad p_B(t, x_0, x) \geq p_\kappa(t, \text{dist}(x_0, x)). \qquad (5.5)$$

From the heat kernel comparison theorems (Theorem 20 and Theorem 19) and standard comparison techniques (eg Bishop's volume comparison [Chavel, 1984]), it is possible to derive a number of comparison results for norms of exit time moments. For example, the following is an analog of a well-known comparison result of Cheng [Cheng, 1975]:

THEOREM 22 *Suppose that M is an n-dimensional Riemannian manifold with sectional curvatures denoted by K and Ricci curvatures denoted by* Ric. *Let τ denote the first exit time of Brownian motion. With the notation of Theorem 20, for all p and all m,*

$$K \leq \kappa \quad \Rightarrow \quad \|\mathbb{E}^x(\tau_B^m)\|_p \leq \|\mathbb{E}^x(\tau_\kappa^m)\|_p \qquad (5.6)$$

$$K \geq \kappa \quad \Rightarrow \quad \|\mathbb{E}^x(\tau_B^m)\|_p \geq \|\mathbb{E}^x(\tau_\kappa^m)\|_p \qquad (5.7)$$

$$\text{Ric} \geq (n-1)\kappa \quad \Rightarrow \quad \|\mathbb{E}^x(\tau_B^m)\|_p \geq \|\mathbb{E}^x(\tau_\kappa^m)\|_p \qquad (5.8)$$

16.5.3. Comparison and transience/recurrence

In addition to the above results concerning the relationship of exit time to isoperimetric phenomena and comparison geometry, there is a deep and beautiful connection between isoperimetric and comparison phenomena for noncompact Riemannian manifolds on the one hand and the transience or recurrence of Brownian motion on the other. There is an excellent recent survey of this material [Grigorýan, 1999] and we remark that the deep work of Varopoulos has been of fundamental importance, especially in the context of groups (cf [Varopoulos, 1992] and references therein). We present a few of the more striking results. Let *M* be a complete non-compact Riemannian manifold and let X_t denote Brownian motion on *M*. Recall,

DEFINITION 23 *Brownian motion on M is* transient *if for some open set U and some point x, Brownian motion eventually leaves U with positive probability:*

$$\mathbb{P}^x\{\exists T : \forall t > T,\ X_t \notin U\} > 0.$$

It is a classical result that Brownian motion in $\mathbb{R}^n$ is recurrent for $n \leq 2$ and transient for $n \geq 3$. Straightforward comparison results allow one to extend this to spaces with variable curvature: In dimension 2 all nonnegatively curved manifolds have recurrent Brownian motion while in dimension 3 and above, all

nonpositively curved manifolds have transient Brownian motion (cf [Kendall, 1987] and references therein).

It is a result of classical potential theory (cf [Doeblin, 1938]) that transience of Brownian motion is equivalent to M being non-parabolic:

DEFINITION 24 *We say that a complete manifold M is* non-parabolic *if M admits a non-constant positive superharmonic function. Otherwise, we say that M is* parabolic.

A recent typical result tying transience of Brownian motion to the geometry of a non-compact manifold involves establishing sufficiency conditions for parabolicity in terms of volume growth (for a survey containing results on volume growth and geometry, see [Li, 2000]):

THEOREM 25 *([Grigorýan, 1999], [Karp, 1982], [Varopoulos, 1983]) Suppose that M is complete and that $x \in M$. Let $B(x, \rho)$ be*
the ball of radius ρ centered at x. Suppose that

$$\int_0^\infty \frac{\rho}{\text{Vol}(B(x, \rho))} d\rho \;=\; \infty. \tag{5.9}$$

Then M is parabolic.

Similar results hold for manifolds which admit a Faber-Krahn type inequality with isoperimetric function Λ :

DEFINITION 26 *Suppose that $\Lambda : (0, \infty) \to \mathbb{R}$ is a positive decreasing function. We say that a complete manifold M satisfies a Faber-Krahn type inequality with isoperimetric function Λ if for all precompact $\Omega \subset M$,*

$$\text{Area}(\partial\Omega) \geq \Lambda(\text{Vol}(\Omega)). \tag{5.10}$$

The following is a theorem of Grigoryan [Grigorýan, 1994]:

THEOREM 27 *([Grigorýan, 1994]) Suppose that M is complete and that for all precompact open sets of large enough volume, M satisfies a Faber-Krahn type inequality with isoperimetric function Λ satisfying*

$$\int_0^\infty \frac{1}{v^2 \Lambda(v)} dv \;<\; \infty. \tag{5.11}$$

Then M is non-parabolic.

One can also estimate the heat kernel [Grigorýan, 1994]:

THEOREM 28 *([Grigorýan, 1994]) Suppose that M is complete and that M satisfies a Faber-Krahn type inequality with isoperimetric function Λ. Fix $x \in$*

$M, t_0 \geq 0, \delta \in (0,1)$ *and suppose that there exists a non-negative function* $\Phi : [t_0, \infty) \to \mathbb{R}$ *such that*

$$\Phi(t_0) \leq \frac{2}{\delta p_M(2t_0, x, x)}$$

$$\int_{\Phi(t_0)}^{\Phi(t)} \frac{1}{v\Lambda(v)} dv = (1-\delta)(t-t_0). \tag{5.12}$$

Then for all $t > t_0$,

$$p_M(2t, x, x) \leq \frac{2}{\delta\Phi(t)}. \tag{5.13}$$

These results follow via estimates for capacity and can be further refined [Grigorýan, 1999 A].

Estimates for the long time behavior of the heat kernel on a complete Riemannian manifold can often be parlayed into information concerning the geometry of the manifold at infinity (to make this precise, see section 6.2 below). There are a number of excellent surveys of this theme available (cf [Grigorýan, 1999 B]). That we consider a single recent result should in no way be taken to represent activity in the field; the associated literature is volumnious.

Intuitively, given a ball $B(x, r)$ of radius r centered at $x \in M$, one expects that the faster the volume $\text{Vol}(B(x, r))$ grows as a function of the radius, the faster the heat kernel $p(t, x, x)$ should decay. In fact, it is possible to give a bound for the decay of the heat kernel in terms of volume growth. More precisely, Barlow, Coulhon and Grigoryan prove [Barlow, 2001]

THEOREM 29 *Let M be a geodesically complete noncompact Riemannian manifold with bounded geometry and let* $r_0 > 0$ *be its injectivity radius. Suppose that for all points* $x \in M$ *and all* $r \geq r_0$,

$$\text{Vol}(B(x, r)) \geq v(r)$$

where $v : [r_0, \infty) \to \mathbb{R}^+$ *is a continuous positive strictly increasing function. Then, for all* $t \geq t_0 = r_0^2$,

$$\sup_{x \in M} p(t, x, x) \leq \frac{C}{\gamma(ct)} \tag{5.14}$$

where γ *is defined by*

$$t - t_0 = \int_{v(r_0)}^{\gamma(t)} v^{-1}(s) ds$$

where v^{-1} *is the inverse function, and* c, C *are positive constants.*

To prove the theorem, the authors first note that their volume growth hypothesis implies a Faber-Krahn type inequality

$$\lambda_1(D) \;\geq\; \Lambda(\mathrm{Vol}(D)) \tag{5.15}$$

where D is a large enough precompact set, λ_1 is the principle Dirichlet eigenvalue and Λ is a function determined by $v(r)$. They then establish that the inequality (5.15) is equivalent to the required heat kernel estimates.

16.6   Minimal Varieties

For the purpose of this section, minimal varieties are geometric objects which arise as solutions to geometric variational problems. In this section, we review minimal varieties with ties to probability.

We begin with the proto-typical example given by the St. Venant Torsion Problem (Theorem 9). In this case the minimal varieties are domains which maximize the L^1-norm of the first exit

time moment of Brownian motion, given a volume constraint. In the category of smooth domains, that maximizers must be spheres follows from the work of Serrin [Serrin, 1971]. In more detail, suppose we consider the collection of all smoothly bounded domains $D \subset \mathbb{R}^2$ with compact closure:

$$\mathcal{D} \;=\; \{D \subset \mathbb{R}^2 : \bar{D} \text{ compact}, \partial D \text{ smooth}\}. \tag{6.1}$$

This space has a natural smooth structure with the tangent space at each $D \in \mathcal{D}$ identified with smooth functions on $\partial D : T_D\mathcal{D} \simeq C^\infty(\partial D)$. Consider the smooth function $F : \mathcal{D} \to \mathbb{R}$ defined by

$$F(D) \;=\; \|\mathbb{E}^x[\tau_D]\|_1 \tag{6.2}$$

Smoothly perturbing D, we obtain a characterization of critical points of $F : D$ is critical for F if one can solve the overdetermined boundary value problem:

$$\begin{aligned}
\Delta u + 1 &= 0 \text{ on } D \\
u &= 0 \text{ on } \partial D \\
\frac{\partial u}{\partial \nu} &= c \text{ on } \partial D
\end{aligned} \tag{6.3}$$

where $\frac{\partial u}{\partial \nu}$ is the normal derivative along the boundary and $c = -\frac{\mathrm{Vol}(D)}{\mathrm{Area}(\partial D)}$ is a constant. Serrin's result states that it is possible to solve the overdetermined boundary value problem (6.3) if and only if the domain is a ball. As pointed out by Serrin, his result holds when one replaces the Laplace operator by the Laplace operator with certain types of lower order nonlinearity. Serrin's result, as well as his technique, led to a great deal of progress in nonlinear PDE (cf [Gidas, 1979], [Gidas, 1981], [Berestycki, 1991], [Berestycki, 1993]).

If one considers in place of $\mathbb{R}^2$ a constant curvature space form and in place of F the L^p-**norm** of the **k**th moment of the exit time, one can run the same variational argument, obtaining a characterization of critical points by overdetermined boundary value problems(with nonlinearity in the boundary condition as opposed to the operator). It turns out that the boundary value problems have solutions if and only if the domain is a ball (cf [McDonald, 2002]), thus characterizing the minimal varieties for the L^p-**norm** of the exit time moments for smooth domains in constant curvature space forms. For Borel sets, the case of equality is settled by Burchard and Schmuckenschläger [Burchard, 2002] (cf Theorem 19).

In addition to controlling volume there are a number of other geometric constraints which one can impose on domains in an ambient space when studying L^p-**norms** of exit time moments of Brownian motion. One such constraint, important in a number of applications, involves fixing the *inradius* of a domain. We recall the definition:

DEFINITION 30 *Suppose that M is Riemannian and that* $D \subset M$. *Then* inradius *of D is the extended real number*

$$\sup\{r \in \mathbb{R} : \exists x \in D, \ B(x,r) \subset D\}$$

where $B(x,r)$ *is the ball of radius* r *centered at* x.

Using conformal techniques, the following result is contained in the work of Banuelos, Carroll, and Housworth [Banuelos, 1998]:

THEOREM 31 *Suppose* $D \subset \mathbb{R}^2$ *and let* τ *be the first exit time of Brownian motion. Then*

$$inradius(D) = 1 \ \Rightarrow \ \|\mathbb{E}^x(\tau_D)\|_\infty \le \|\mathbb{E}^x(\tau_S)\|_\infty$$

where S *is the infinite rectangular strip* $S = \{(x,y) : 0 \le y \le 1\}$.

There are a variety of related recent results for unbounded domains.

16.7   Harmonic Functions

Let M be a complete Riemannian manifold, Δ the Laplace operator acting on functions on M. Recall, a function $f : M \to \mathbb{R}$ is *harmonic* if it satisfies

$$\Delta f = 0.$$

Equivalently, f is harmonic if and only if f is stationary for the Dirichlet form

$$E(f) \ = \ \int_M |\nabla f|^2 dg. \tag{7.1}$$

It is well known that there is a deep relationship between Brownian motion and harmonic functions. An example of this relationship is given by the Kakutani's representation of the solution to the Dirichlet problem (2.6) in terms of Brownian motion (2.7). In this section we survey such probabilistic

representations and their connection to the geometry of noncompact, complete manifolds.

The representation (2.7) is but one instance of an extensive body of work devoted to the representation of harmonic functions via boundary geometry. Another such representation of harmonic function is given by the Poisson kernel. More precisely, suppose for concreteness that D is a smoothly bounded domain with compact closure in $\mathbb{R}^n$. Let $G(x, y)$ be the Green's function for D, dy surface measure on ∂D. Define a function u on D by

$$u(x) \quad = \quad \int_{\partial D} k_y(x) d\mu(y) \qquad\qquad (7.2)$$

where $k_y(x) = \frac{\partial G(x,y)}{\partial \nu}$ is the Poisson kernel and $d\mu(y) = f(y)dy$ for some positive function f on the boundary of D. Then (7.2) defines a positive harmonic function: the solution of the Dirichlet problem with boundary data f. In fact, allowing the measure $d\mu$ to be supported and finite on ∂D (with no other constraints) provides a representation of every positive harmonic function on D.

It was an idea of Martin [Martin, 1941] that such a representation should be possible for bounded but otherwise arbitrary domains in $\mathbb{R}^n$, given the appropriate definition of "boundary." This idea came to play an important role in potential theory, both from a probabilistic and from an analytic point of view. The material was developed by both schools (cf [Dynkin, 1965], [Dynkin, 1982], [Doeblin, 1938], [Pinsky, 1995]).

16.7.1.   Martin boundaries

To define the Martin boundary, let $D \subset \mathbb{R}^n$ be an arbitrary bounded domain and fix $p \in D$. Let G be the minimal Green's function for D and define

$$h_y(x) \quad = \quad \frac{G(x, y)}{G(p, y)} \qquad\qquad (7.3)$$

Let $\{y_i\}$ be a nonconvergent sequence of points in D and consider the harmonic functions $h_i(x) = h_{y_i}(x)$. Then the sequence $\{h_i\}$ is uniformly bounded on compact subsets of D and for all i, $h_i(p) = 1$. By Harnack's inequality, there exists a convergent subsequence, denoted $\{h_{i_j}\}$, which converges uniformly on compact subsets of D to a positive harmonic function $h(x)$. We call the sequence of points $\{y_i\}$ a *Martin sequence.* We say that two Martin sequences are equivalent if and only if the have the same limiting harmonic functions.

DEFINITION 32 *The* Martin boundary *of D, denoted* $\mathcal{M}$, *is the collection of equivalence classes of Martin sequences. We say that a point* $[h] \in \mathcal{M}$ *is* minimal *if the corresponding harmonic limit* h *satisfies*

> *If* h' *is a positive harmonic function on D and* $h' \leq h$, *then* $h' = ch$ *for some* $c \in (0, 1]$.

The minimal Martin boundary *of D, denoted* $\mathcal{M}_0$, *is the collection of all minimal points.*

The results of [Martin, 1941] contain the Martin representation theorem:

THEOREM 33 *Suppose* $D \subset \mathbb{R}^n$ *is bounded and that* $\mathcal{M}_0$ *is the minimal Martin boundary of D. For* $y \in \mathcal{M}_0$, *let* $k_y(x)$ *denote the corresponding positive harmonic function. Then for each positive harmonic function* u *there exists a unique finite measure* μ_u *supported on* $\mathcal{M}_0$ *such that*

$$u(x) \;=\; \int_{\mathcal{M}_0} k_y(x) d\mu_u(y). \tag{7.4}$$

Conversely, for every finite measure supported on $\mathcal{M}_0$, *(7.4) defines a positive harmonic function on D.*

In providing the above representation theorem, the Martin boundary provides a means of employing analytic techniques to study the geometry of the underlying domain. To see that this is the case, note that there is a natural metric topology on $D \cup \mathcal{M}$ (cf[Pinsky, 1995]) for which $\mathcal{M}$ becomes a compactification of D. When D is sufficiently regular, this coincides with the Euclidean compactification of D. We have the following theorem of Hunt-Wheedon:

THEOREM 34 *(cf [Hunt, 1970]) Suppose* $D \subset \mathbb{R}^n$. *Suppose that for each* $y \in \partial D$, *there is a ball,* $B(y)$, *centered at* y *such that* $B(y) \cap \partial D$ *is the graph of a Lipschitz function. Then the Martin boundary of D, the minimal Martin boundary of D and the Euclidean boundary of D all coincide.*

This result strongly suggests that the ideas surrounding the notion of a Martin boundary might be useful in the study of the geometry of complete noncompact Riemannian manifolds near their "boundary." Obviously, the first step in such a program is to establish precisely what is meant by "geometry of the boundary" in this context. There is a natural geometric approach:

DEFINITION 35 *Let M be a complete Riemannian manifold. Given two geodesic rays,* γ_1 *and* γ_2, *in M we say that* $\gamma_1(t)$ *and* $\gamma_2(t)$ *are asymptotic if* $\text{dist}(\gamma_1(t), \gamma_2(t))$ *is a bounded function of* t.

It is clear that the notion of asymptotic defines an equivalence relation on the collection of geodesic rays.

DEFINITION 36 *Let M be a complete Riemannian manifold. We define the sphere at infinity, denoted $S(\infty)$, as the collection of equivalence classes of geodesic rays in M.*

There is a natural topology on $M \cup S(\infty)$ (the cone topology) and, with respect to this topology, $S(\infty)$ gives a topological compactification of M. Given this, we refer to $S(\infty)$ as the *geometric boundary* of M.

To define the Martin boundary of a (class of) complete Riemannian manifolds, we model the development on Definition 32 and its motivating discussion:

DEFINITION 37 *Suppose that M is a complete Riemannian manifold admitting a Green's function, $G(x, y)$. Let $p \in M$ and, for $x, y \in M$, let $h_y(x)$ be defined by (7.3). Let y_i be a nonconvergent sequence of points, $h_{y_i}(x)$ the corresponding harmonic functions, and $h_{y_{i_j}}$ a subsequence converging uniformly on compacts to a harmonic limit $h(x)$. We call the sequence y_i a Martin sequence and we say that two Martin sequences are equivalent if they have the same harmonic limit. The Martin boundary of M is the collection of equivalence classes of Martin sequences.*

The question of which non-compact Riemannian manifolds should be studied via Martin's approach was clarified by the seminal work of Yau [Yau, 1975], [Yau, 1976]. Before proceeding, we need a fundamental definition:

DEFINITION 38 *A manifold is said to have the* Liouville property *if it does not admit any nonconstant bounded harmonic functions. A manifold is said to have the* strong Liouville property *if it does not admit any nonconstant positive harmonic functions.*

Yau proved

THEOREM 39 *([Yau, 1975]) If M has nonnegative Ricci curvature, then M has the strong Liouville property.*

As the Martin boundary construction requires a rich structure of positive harmonic functions, Yau's result suggests that if Martin boundaries are to play a role in the study of the geometry of a non-compact Riemannian manifold, negative curvature will be necessary (cf also [Dynkin, 1965]). Given that the definition of the Martin boundary involves the existence of a Green's function, we must further restrict to a class of manifolds andmitting Green's functions; for example, manifolds with pinched negative curvature. This is the setting of the work of Anderson-Schoen [Anderson, 1985] who proved

THEOREM 40 *(cf [Anderson, 1985]) Let M be a complete simply connected manifold with sectional curvature K_M satisfying*

$$-b^2 \leq K_M \leq -a^2 < 0.$$

Then there is a natural homeomorphism between the Martin boundary of M
and the geometric boundary of M (the sphere at infinity).

Since the publication of [Anderson, 1985], there has been an explosion in the
study of harmonic functions on complete Riemannian manifolds, their cor-
responding Martin boundaries, and the geometry of such manifolds at infin-
ity. There are a number of informative surveys available (cf [Li, 2000]), most
focussing on the geometric/function theoretic aspects of the material. There
has also been a roughly concurrent probabilistic development of the mate-
rial (a survey can be found in [Pinsky, 1995]), with results largely parallel-
ing those obtained function theoretically (cf [Doeblin, 1938], [Dynkin, 1965],
[Kifer, 1992], [Hsu, 1985], [Cranston, 1993] [Grigorýan, 1999] and references
therein). Many such results can be inferred in the context of volume compar-
ison and potential theory (cf section 5.3 above). Reference [Grigorýan, 1999]
contains an excellent review of this material. We focus our remarks on material
of independent interest.

The probabilistic approach to Martin boundaries involves the study of the
asymptotic behavior of Brownian motion and the existence, given appropriate
assumptions on the ambient manifold, of almost sure limiting directions. Be-
cause the probabilistic approach does not require the existence of a uniquely
defined Laplace operator, it is possible to formulate a theory of Martin bound-
aries for spaces which are not manifolds, for example simplicial complexes
whose simplices are Euclidean (ie Euclidean complexes). Such a program has
recently been carried out in part by Brin and Kifer, who prove the appropriate
analog of the Anderson-Schoen result for Euclidean complexes [Brin, 2001].
This development provides a framework for a geometric function theory for
large classes of singular spaces.

Along a similar vein, given that the probabilistic development of Martin
boundaries involves specific path properties of an underlying process, one
might choose to fix the ambient manifold and investigate analogs of the Martin
constructions for processes other than Brownian motion. This has recently
been carried out by Chen and Song, who investigate the appropriate analogs of
Martin boundary for symmetric stable processes [Chen, 1998].

16.7.2.　　Harmonic maps

Suppose that M and N are Riemannian manifolds and $f : M \to N$. We say
that f is a harmonic map if f is a stationary point of the Dirichlet form

$$E(f) \;=\; \int_M \|Df\|^2 dg \qquad (7.5)$$

where Df is the derivative of f (the induced map between tangent spaces).
Given the relationship between Brownian motion and harmonic functions, it is
natural to expect that probability will play an interesting role in the theory of

harmonic maps (this seems to have been first suggested by Eells and Lemaire [Eells, 1983]). This is indeed the case; an interesting survey of recent developments can be found in [Kendall, 1998].

16.8   Hodge Theory

Let M be an n-**dimensional** differentiable manifold, $\Omega^k = \Omega^k(M)$ the bundle of smooth k-**forms** on M, $d : \Omega^k \rightarrow \Omega^{k+1}$ the exterior derivative (see section 2). The *kth* de Rham cohomology of M is the quotient space of closed k-**forms** by exact k-**forms**:

$$H^k_{dR}(M) \quad = \quad \frac{\{\omega \in \Omega^k : d\omega = 0\}}{\{\omega \in \Omega^k : \exists \alpha \in \Omega^{k-1} d\alpha = \omega\}}. \tag{8.1}$$

The celebrated work of de Rham provides an isomorphism between $H^k_{dR}(M)$ and the **k**th Cech cohomology group of M with real coefficients. Thus, the spaces $H^k_{dR}(M)$, $0 \leq p \leq n$ are topological invariants of M. In this section we survey probabilistic results which provide a means of studying $H^k_{dR}(M)$ under appropriate conditions on M. These results revolve around heat flow and the work of Hodge.

Suppose that M is a compact Riemannian manifold and let Δ be the Laplace-Betrami operator acting on k-**forms** on M. Let $L^2(\Omega^k, dg)$ be the L^2 completion of the sections of the k-**form** bundle with respect to the induced volume dg. As discussed in section 2 and section 4, the Laplace-Beltrami operator is essentially self-adjoint and thus admits a unique self-adjoint extension to an operator on L^2-**sections** of the k-**form** bundle. It is elliptic, and thus its kernel consists of smooth k-**forms** which we suggestively denote by

$$H^k_{Hodge}(M) \quad = \quad \{\omega \in \Omega^k : \Delta\omega = 0\}. \tag{8.2}$$

Let $P_t = e^{-t\Delta}$ be the heat operator acting on k-**forms**. Let $\omega \in \Omega^k$. Then the solution of the Cauchy initial value problem

$$\partial_t \omega - \Delta\omega_t \quad = \quad 0 \text{ on } (0, \infty) \times M$$
$$\omega_0 \quad = \quad \omega \tag{8.3}$$

is given by

$$\omega_t \quad = \quad P_t\omega_0. \tag{8.4}$$

The operator P_t is compact for all $t > 0$ and admits a unique self-adjoint extension to $L^2(\Omega^k, dg)$. In fact, P_t is a contraction for all $t > 0$ which converges in norm to orthogonal projection on $H^k_{Hodge}(M)$ as $t \rightarrow \infty$.

Suppose $\omega_0 \in \Omega^k$ and write $\omega_0 = (I - P_t)\omega_0 + P_t\omega_0$. Then

$$
\begin{aligned}
(I - P_t)\omega_0 &= \int_0^t \partial_s(P_s\omega_0)ds \\
&= \Delta \int_0^t P_s\omega_0 ds \\
&= d\alpha + d^*\beta
\end{aligned}
\tag{8.5}
$$

where

$$
\alpha = d^* \int_0^t P_s\omega_0 ds \in \Omega^{k-1}
\tag{8.6}
$$

$$
\beta = d \int_0^t P_s\omega_0 ds \in \Omega^{k+1}.
\tag{8.7}
$$

Taking a limit, we obtain the celebrated *Hodge decomposition:*

$$
\omega_0 = d\alpha + d^*\beta + \gamma
\tag{8.8}
$$

where $\alpha \in \Omega^{k-1}$, $\beta \in \Omega^{k+1}$ and $\gamma \in H^k_{Hodge}(M)$, the decomposition being orthogonal. Given $[\omega] \in H^k_{dR}(M)$, let ω_0 represent $[\omega]$ and write ω_0 as in (8.8). Then since ω_0 is closed, $\beta = 0$. Moreover, if $\tilde{\omega} = d\tilde{\alpha} + \tilde{\gamma}$ is another representation of $[\omega]$, then $\gamma = \tilde{\gamma}$. We conclude that the evolution $\omega_t = P_t\omega_0$ smoothly deforms every representative of the class $[\omega]$ to its harmonic projection γ, which is the element of minimal norm representing $[\omega]$ as an element of $H^k_{dR}(M)$. Thus, we obtain the celebrated result of Hodge:

THEOREM 41 *If M is a compact Riemannian manifold, there is a natural isomorphism between the de Rham cohomology of M and the harmonic forms of M:*

$$
H^k_{dR}(M) \simeq H^k_{Hodge}(M).
\tag{8.9}
$$

The isomorphism is given by identifying each de Rham class with its representative of minimal norm.

When M is not compact one cannot expect the operators P_t to be compact and the above approach must be modified if it is to have any hope of producing an analog of the Hodge theorem. To see how one might go about constructing an analog, consider the case of $\mathbb{R}^n$. The DeRham cohomology of $\mathbb{R}^n$ is well known:

$$
H^k_{dR}(\mathbb{R}^n) = \begin{cases} \mathbb{R} & \text{if } k = 0 \\ 0 & \text{else} \end{cases}
$$

the zero dimensional cohomology being represented by constant functions which are not L^2 with respect to the measure induced by the volume form (ie Lebesgue measure). To produce a reasonable candidate for the Hodge Laplacian, we rescale the volume element appropriately: **Fix $t > 0$ and $x_0 \in \mathbb{R}^n$ and** let $\rho_t(x_0, x)$ be the fundamental solution of the heat equation on $\mathbb{R}^n$ at time t :

$$\rho_t(x_0, x) = \frac{1}{(4\pi t)^{\frac{n}{2}}} e^{-\frac{|x-x_0|^2}{4t}}.$$

Consider the heat kernel weighted measure

$$d\mu = \rho_t(x_0, x)dx$$

and let $L^2(\mathbb{R}^n, d\mu)$ be the corresponding weighted L^2-space. The measure $d\mu$ induces an adjoint of the exterior derivative, denoted d_μ^*, and a corresponding Laplace-Beltrami operator $\Delta_\mu^{(k)}$ acting on the appropriately weighted L^2-form bundles. Writing

$$H_{Hodge}^{k,\mu} = \text{kernel}(\Delta_\mu^{(k)})$$

one can compute directly [Bueler, 1999] that

$$H_{Hodge}^{k,\mu}(\mathbb{R}^n) \simeq H_{dR}^k(\mathbb{R}^n).$$

In the context of Hodge theorems for finite dimensional Riemannian manifolds these ideas seem to have been introduced by Bueler [Bueler, 1999] and further developed by Ahmed-Stroock [Ahmed, 2000]. We sketch the results of the latter.

Suppose that M is a complete, oriented connected Riemannian manifold with Ricci curvature bounded below and the Riemann curvature operator bounded above. Suppose that $U : M \to [0, \infty)$ is a smooth function satisfying

1 U has compact level sets

2 There exists $C < \infty$ and $\theta \in (0, 1)$ such that $\Delta U \leq C(1 + U)$ and $\|\text{grad}U\|^2 \leq Ce^{\theta U}$.

3 There exists an $\varepsilon > 0$ such that $\varepsilon U^{1+\varepsilon} \leq 1 + \|\text{grad}U\|^2$.

4 There exists a $B < \infty$ such that for all $x \in M$ and all $V \in T_x M$,

$$\langle V, \text{Hess}_U V \rangle \geq -B\|V\|^2$$

where Hess_U denotes the Hessian of U and the pairing is given by the metric.

Then (cf [Ahmed, 2000] Lemma 6.2), for each $x \in M$, there is a unique path $F_t^U(x) : [0, \infty) \to M$ satisfying $F_0(x) = x$ and $\frac{d}{dt}F_t^U(x) = -\text{grad}_{F_t^U(x)}U$. Moreover, $F_t^U : [0, \infty) \times M \to M$ is a smooth map which is a diffeomorphism onto its image with differential a linear map everywhere bounded by e^{Bt}. In particular, given a smooth form ω, the pullback $(F_1^U)^*\omega$ is bounded and if ω is exact so is the pullback. Let $\Phi^U\omega$ be the orthogonal projection of $(F_1^U)^*\omega$ onto the space of L^2 U-weighted harmonic forms on M. Then (cf [Ahmed, 2000] Theorem 6.4)

THEOREM 42 *With M, U and Φ^U as above, the map Φ^U induces a linear isomorphism between the U-weighted L^2 cohomology of M and the deRham cohomology of M. The map is natural in the sense that $\Phi^U\omega$ is the element of minimal U-weighted L^2 norm.*

In addition to the work of Ahmed-Stroock, recent work of Gong-Wang [Gong, 2001] involving heat kernel estimates for a class of complete Riemannian manifolds containing those manifolds with Ricci curvature bounded below can be used to compute Hodge cohomology for Witten-deformed Laplacian in the top dimension.

Finally, we mention the work of Elworthy, Li and Rosenberg on L^2 harmonic forms [Elworthy, 1998].

Recall, if M is Riemannian, the Weitzenbock decomposition of the Laplace-Beltrami operator on k-forms expresses the Laplacian in terms of the Levi-Civita connection and certain curvature invariants (2.23). When M is complete and the curvature term is positive, it is a theorem of Bochner that the corresponding cohomology in dimension k vanishes.

In [Elworthy, 1998], the authors consider Riemannian manifolds whose Weitzenbock curvature term is *strongly stochastically positive* (when M is compact, this allows the curvature term to be negative on a set of small volume). They establish a number of vanishing theorems and a variety of curvature pinching results; for example, they prove that a compact manifold cannot admit both a strongly stochastically

positive $\mathcal{R}^2$ term and a metric with pinched negative curvature. Many of their results apply to the Witten Laplacian. The approach should yield a number of additional results.

References

N. Akhiezer *The Classical Moment Problem,* Hafner, New York (1965).

D. Aldous *An introduction to covering problems for random walks on graphs* J. Theoret. Prob. **2** (1989) 87–89.

S. I. Anderson and M. L Lapidus *Progress in Inverse Spectral Geometry* Trends in Mathematics, Birkhäuser, Basel, (1997)

M. Aizenman and B. Simon *Brownian motion and Harnack's inequalities for Schrodinger operators* Comm. Pure Appl. Math. **35** (1982) 209–273.

Z. M. Ahmed and D. W. Stroock *A Hodge theory for some non-compact manifolds* Jour. Diff. Geom. **54** (2000) 177–225.

M. Anderson and R. Schoen *Positive harmonic functions on complete manifolds of negative curvature* Ann. Math. **121** (1985) 429–461.

C. Bandle *Isoperimetric Inequalities and Applications,* Pitman, Boston, (1986).

R. Banuelos and T. Carroll *Brownian motion and the fundamental frequency of a drum* Duke **75** (1994) 575–602.

R. Banuelos, T. Carroll and E. Housworth *Inradius and integral means for the Green's functions and conformal mappings* Proc AMS **126** (1998) 577–585.

M. T, Barlow, T. Coulhon and A. Grigoryan *Manifolds and graphs with slow heat kernel decay* Invent. Math. **144** (2001) 609–649.

H. Berestycki and L. Nirenberg *On the method of moving planes and the sliding method* Bol. Soc. Brasil. Mat. **22** (1991) 1–37.

H. Berestycki, L. Caffarelli and L. Nirenberg *Symmetry for elliptic equations in a half plane* In: Boundary value problems for partial differential equations and applications, RMA Res. Notes Appl. Math **29** Masson, Paris (1993) 27–42.

M. van den Berg and E. Bolthausen *Estimates for Dirichlet eigenfunctions* J. London Math. Soc. **59** (1999) 607–619 .

M. van den Berg and P. Gilkey *Heat content asymptotics of a Riemannian manifold with boundary* Jour. Funct. Anal. **120** (1994) 48–71.

M. van den Berg and J. F. Le Gall *Mean curvature and the heat equation* Math. Zeit. **215** (1994) 437–464.

M. van den Berg and S. Srisatkunarajah *Heat flow and Brownian motion for a region in $\mathbb{R}^2$ with polygonal boundary* Prob. Theor. Rel. Fields **86** (1990) 41–52.

M. van den Berg and S. P. Watson *Asymptotics for the spectral heat function and bounds for integrals of Dirichlet eigenfunctions* Proc. Royal Soc. of Edin. **129** (1999) 841–854.

P. Bérard *Spectral Geometry: direct and inverse problems* with appendices by G. Besson, B. Berger and M. Berger, Springer Lecture Notes in Mathematics, **1207** Berlin (1986).

J. M. Bismut *Probability and geometry* In: Probability and analysis (Varenna, 1985) Lecture Notes in Math, vol 1206, Springer, Berlin (1986) 1–60.

H. J. Brascamp, E. H. Lieb and J. M. Luttinger *A general rearrangement inequality for multiple integrals* Jour. Funct. Anal. **17** (1974) 227–237.

M. Brin and Y. Kifer *Brownian motion, harmonic functions and hyperbolicity for Euclidean complexes* Math. Zeit. **237** (2001) 421–168.

E. L. Bueler *The heat kernel weighted Hodge Laplacian on noncompact manifolds* Trans. AMS **351** (1999) 683–713.

K. Burdzy and W. Kendall *Efficient Markov couplings: examples and counterexamples* Ann. Appl. Prob. **10** (2000) 362–409.

A. Burchard and M. Schmuckenschläger *Comparison theorems for exit times* GAFA **11** (2002) 651–692.

P. Buser, J. Conway, P. Doyle, and D. Semmler *Some planar isospectral domains,* Intern. Math. Res. Notices **9** (1994) 391–400.

I. Chavel *Eigenvalues in Riemannian geometry,* Academic Press, New York (1984).

I. Chavel *Isoperimetric Inequalities ,* Cambridge University Press, Cambridge, (2001).

I. Chavel and E. A. Feldman *Spectra of manifolds less a small domain* Duke Math. J. **56** (1988) 399–414.

S. Y. Cheng *Eigenvalue comparison theorems and its geometric applications* Math. Zeit. **143** (1975) 289–297.

M. F. Chen *Optimal couplings and applications to Riemannian geometry* In: Probability Theory and Mathematical Statistics, B. Grigelionis et al, eds. VPS/TEV (1994) 121–142.

Z. Q. Chen and R. Song *Martin boundary and integral representation for harmonic functions of symmetric stable processes* J. Funct. Anal. **159** (1998) 267–294.

M. F. Chen and F. Y. Wang *General formula for the lower bound of the first eigenvalue on a Riemannian manifold* Sci. Sin. (A) **40** (1997) 384–394.

M. F. Chen and F. Y. Wang *Applications of the coupling method to the first eigenvalue on a manifold* Sci. Sin. (A) **40** (1994) 384–394.

J. Cheeger and S. T. Yau *A lower bound for the heat kernel* Comm. Pure Appl. Math. **34** (1981) 465–480.

M. Cranston *A probabilistic approach to Martin boundaries for manifolds with ends* Prob. Th. and Rel. **96** (1993) 319–334.

A. Debiard, B. Gavaeu, and E. Mazet *Théorèrems de comparison en géométrie riemannienne* Publ. Res. Inst. Math. Sci **12** (1976/77) 391–425.

A. Dembo, Y. Peres, J. Rosen, and O Zeituni *Cover times for Brownian motion and random walks in two dimensions* preprint (2001).

A. Dembo, Y. Peres, J. Rosen, and O Zeituni *Thick Points of Planar Brownian Motion and the Erods-Taylor Conjecture on Random Walk* preprint (2001).

W. Doeblin *Exposé de la theorie des chaines simple constantes de Markov à un nombre fini d'etats* Rev. Math. Union Interbakanique **2** (1938) 77–105.

J. Doob *Classical Potential Theory and Its Probabilistic Counterpart* Springer, Berlin (1983).

B. A. Dubrovin, A. T. Fomenko and S. P. Novikov *Modern Geometry - Methods and Applications, Part 1* Springer, Berlin (1984).

T. Duchamp and W. Stuetzle *Extremal properties of principal curves in the plane* Ann. Statist. **24** (1996)1511-1520.

E. Dynkin *Markov Processes* Vol 1, 2, Springer, Berlin (1965).

E. Dynkin *Markov Processes and Related Problems of Analysis* London Math. Soc. Let. Notes, Vol **54**, Cambridge University Press, Cambridge, UK (1982).

E. Dynkin *The space of exits of a Markov process* Russian Math. Surveys **XXIV** (1969) 89–157.

J. Eells and L. Lemaire *Selected Topics in Harmonic Maps* American Math. Soc. Providence, RI, (1983).

K. D. Elworthy, X.-M. Li and S .Rosenberg *Bounded and L^2 harmonic forms on universal covers* Geom. Funct. Anal **8** (1998) 283–303.

B. Gidas, W. M. Ni and L. Nirenberg *Symmetry and related properties via the maximum principal* Comm. Math. Phys. **68** (1979) 209–243.

B. Gidas, W. M. Ni and L. Nirenberg *Symmetry of positive solutions of nonlinear elliptic equations in $\mathbb{R}^n$* In: Mathematical Analysis and Applications, Part A, Adv. in Math, Suppl. Stud. 7a, Academic Press, New York, (1981) 369–342.

P. Gilkey *Heat content asymptotics* In: Geometric aspects of partial differential equations (Roskilde, 1998), Contemp. Math **242** AMS, Providence, RI (1999) 125–133.

S. Goldberg *Curvature and Homology* Dover, New York, NY (1962).

F. Z. Gong and F. Y. Wang *Heat kernel estimates with application to compactness of manifolds* Q. J. Math **52** (2001), 171–180.

C. Gordon, D. Webb, and S. Wolpert *Isospectral plane domains and surfaces via Riemannian orbifolds,* Invent. Math. **110** (1992), 1–22.

A. Gray *Tubes* Addison Wesley, Redwood City, CA (1990).

A. Gray and M. Pinsky *Mean exit time from a geodesic ball in Riemannian manifolds* Bull. des Sci. Math. **107** (1983) 345–370.

A. Gray and L. Vanhecke *The volumes of tubes in a Riemannian manifold* Rend. Sem. Math. Politec. Torino **39** (1981) 1–50.

A. Grigorýan *Analytic and geometric background of recurrence and non-explosion of the Brownian motion on Riemannian manifolds* Bull. AMS **36** (1999) 135–249.

A. Grigorýan *Heat kernel upper bounds on a complete non-compact manifold* Rev. Math. Iberoamer. **10** (1994) 395–452.

A. Grigorýan *Isoperimetric inequalities and capacities on Riemannian manifolds* In: The Mazýa anniversary collection, Vol 1 (Rostock, 1998), Oper Theory Adv. Appl., 109 Birhauser, Basel (1999) 139–153.

A. Grigorýan *Estimates of heat kernels on Riemannian manifolds* In: "Spectral Theory and Geometry. ICMS Instructional Conference, Edinburgh 1998" London Math. Soc. Lecture Note Series 273, Cambridge Univ. Press, 1999 140–225.

W. K. Hayman *Some bounds for principal frequencies* Appl. Anal. **7** (1978) 247–254.

T. Hastie and W. Stuetzle *Principal curves* J. Amer. Statist. Assoc. **84** (1989) 502–516.

P. Hsu and P. March *The limiting angle of certain Riemannian Brownian motions* Comm. Pure and Appl. **38** (1985) 755–768.

H. R. Hughes *Brownian exit distributions from normal balls in $S^3 \times H^3$* Ann. Prob. **20** (1992) 655–659.

G. A. Hunt and R. L. Wheedon *Positive harmonic functions on Lipschitz domains* Trans AMS **132** (1970) 307–322.

D. Iesan *Saint Venant's Problem* Springer Lecture Notes in Mathematics, **1279** Berlin (1980).

J. D. S. Jones and R. Léandre *A stochastic approach to the Dirac operaotr over the free loop space* Proc. Steklov Inst. Math. **217** (1997) 253–282.

L. Karp *Subharmonic functions, harmonic mappings and isometric immersions* In: Seminar on Differential Geometry, ed. S. T. Yau **102** Princeton University Press, Princeton (1982).

L. Karp and M. Pinsky *First eigenvalue of a small geodesic ball in a Riemannian manifold* Bull. Sci. Math. **111** (1987) 222–239.

M. Kac *Probabilistic methods in some problems of scattering theory* Rocky Mountain J. Math. **4** (1974) 511–537.

W. S. Kendall *From stochastic parallel transport to harmonic maps* In: "New Directions in Dirichlet Forms" AMS/IP Stud. Adv. Math. **8** AMS, Providence, RI (1998) 49–115.

W. S. Kendall *Stochastic differential geometry: an introduction* Acta Appl. Math. **9** (1987) 29–60.

Y. Kifer *Brownian motion and positive harmonic functions on complete manifolds of nonpositive curvature* In; Pitman Res. Notes in Math. **150** (1992) 187–232.

K. J. Kinateder, P. McDonald and D. Miller *Exit time moments, boundary value problems and the geometry of domains in Euclidean space* Prob. Th. and Rel. **111** (1998) 469–487.

M. Liao *Hitting distributions of small geodesic spheres* Ann. Prob. **16** (1988) 1029–1050.

P. Li *Curvature and function theory on Riemannian manifolds* In: Surveys in Diff. Geom. **VII** (2000) 1–58.

T. Linvall *On coupling for diffusion processes* J. Appl. Prob. **20** (1983) 82–93.

T. Linvall *Lectures on the coupling method,* John Wiley and Sons, New York, (1992).

T. Linvall and L. C. G. Rogers *Coupling of multidimensional diffusions by reflection* Ann. Prob. **14** (1986) 860–872.

R. S. Martin *Minimal positive harmonic functions* Trans. AMS **49** (1941) 137–172.

P. McDonald *Isoperimetric conditions, Poisson problems and diffusions in Riemannian manifolds* Potential Analysis **16** (2002) 115–138.

P. McDonald and R. Meyers *Dirichlet spectrum and heat content* Jour. Funct. Anal. (to appear).

P. McDonald and R. Meyers *Isospectral polygons, planar graphs and heat content* Proc. AMS (to appear).

M. Pinsky *Feeling the shape of a manifold with Brownian motion - the last word in 1990* In: Stochastic Analysis, Cambridge University Press (1991) 305–320.

M. Pinsky *Can you feel the shape of a manifold with Brownian motion?* In: Topics in Contemporary Probability and Its Applications, CRC Press, Inc (1995) 89–102.

R. Pinsky *Positive Harmonic Functions and Diffusion,* Cambridge University Press, Cambridge, UK (1995).

G. Polya *Torsional rigidity, principle frequency, electrostatic capacity and symmetrization* Quart. Appl. Math. **6** (1948) 267–277.

M. Reed and B. Simon *Methods of Modem Mathematical Physics ,* Academic Press, Orlando (1978).

E. Schmidt *Beweis der isoperimetrischen Eigenschaft der Kugel im hyperbolischen und sphräris-cheschen Raum jeder Dimensionzahl* Math Z., **49** (1943) 1–109.

J. Serrin.,*A symmetry problem in potential theory* Archiv. Rat. Mech. Anal. (1971) 304–318.

R. Smits *Spectral gaps and rates to equilibrium for diffusions in convex domains* Michigan Math. J. **43** (1996) 141–157.

R. Schoen and S. T. Yau *Lectures on Differential Geometry,* International Press, Redwood City, CA (1994).

I. M. Singer, B. Wong, S. T. Yau and S. S. T. Yau *An estimate of the gap of the first two eigenvalues in the Schrodinger operator* Ann. Scula Norm. Sup. Pisa **12** (1985) 319–333.

N. Th. Varopoulos *Potential theory and diffusions in Riemannian manifolds* In: Conference on Harmonic Analysis in Honor of Antonio Zygmund, Wadsworth Math Series, Wadsworth, Belmont. CA (1983) 821–837.

N. Th. Varopoulos, L. Saloffe-Coste and T. Coulhon *Analysis And Geometry On Groups,* Cambridge University Press, Cambridge (1992).

F. Y. Wang *On estimation of the Dirichlet spectral gap* Arch. Math. **75** (2000) 450–455.

S. T. Yau *Harmonic functions on complete Riemannian manifolds* Comm. Pure and Appl. **28** (1975) 201–228.

S. T. Yau *Some function-theoretic properties of complete Riemannian manifolds and their applications to geometry* Ind. Univ. Math. J. **25** (1976) 659–670.

Q. H. Yu and J. Q. Zhong *Lower bounds of the gap between the first and the second eigenvalues of the Schrodinger operator* Trans AMS **294** (1986) 341–349.

This page intentionally left blank

DEPENDENCE OR INDEPENDENCE OF THE SAMPLE MEAN AND VARIANCE IN NON-IID OR NON-NORMAL CASES AND THE ROLE OF SOME TESTS OF INDEPENDENCE

Nitis Mukhopadhyay

Department of Statistics, UBox4120, University of Connecticut, Storrs, CT 06269-4120, U.S.A.

mukhop@uconnvm.uconn.edu

Abstract　　Let $X_1, ..., X_n$ be independent and identically distributed (iid) random variables. We denote the sample mean $\overline{X} = n^{-1}\Sigma_{i=1}^n X_i$ and the sample variance $S^2 = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X})^2$ for $n \geq 2$. Then, it is well-known that if the underlying common probability model for the X's is $N(\mu, \sigma^2)$, the sample mean $\overline{X}$ and the sample variance S^2 are independently distributed. On the other hand, it is also known that if $\overline{X}$ and S^2 are independently distributed, then the underlying common probability model for the X's must be normal (Zinger (1958)). Theorem 1.1 summarizes these. But, what can one expect regarding the status of independence or dependence between $\overline{X}$ and S^2 when the random variables X's are allowed to be non-iid or non-normal? In a direct contrast with the message from Theorem 1.1, what we find interesting is that the sample mean $\overline{X}$ and the variance S^2 may or may not follow independent probability models when the observations X_i's are not iid or when these follow non-normal probability laws. With the help of examples, we highlight a number of interesting scenarios. These examples point toward an opening for the development of important characterization results and we hope to see some progress on this in the future. Illustrations are provided where we have applied the t-test based on Pearson-sample correlation coefficient, a traditional non-parametric test based on Spearman-rank correlation coefficient, and the Chi-square test to "validate" independence or dependence between the appropriate $\overline{x}, s$ data. In a number of occasions, the t-test and the traditional non-parametric test unfortunately arrived at conflicting conclusions based on same data. We raise the potential of a major problem in implementing either a t-test or the nonparametric test as exploratory data analytic (EDA) tools to examine dependence or association for paired data in practice! The Chi-square test, however, correctly validated dependence whenever $(\overline{x}, s)$ data were dependent. Also, the Chi-square test never sided against a correct conclusion that the paired data $(\overline{x}, s)$ were independent whenever the paired variables were in fact independent. It is safe to say that among three contenders, the Chi-square test stood out as the most reliable EDA tool in validating the true state of nature of dependence (or independence) between $\overline{X}, S^2$ as ev-

idenced by the observed paired data $\overline{x}, s$ whether the observations $X_1, ..., X_n$ were assumed iid, not iid or these were non-normal.

Keywords: Frequency histogram, tests for independence, *P*-value, Chi-square test, Pearson-sample correlation test, *t*-test, Spearman-rank correlation test, nonparametric test.

17.1   Introduction

Let us suppose that $X_1, ..., X_n$ are independent and identically distributed (iid) random variables governed by a common distribution function $F(x)$, $x \in \Re$. We denote the sample mean $\overline{X} = n^{-1}\Sigma_{i=1}^{n} X_i$ and the sample variance $S^2 = (n - 1)^{-1}\Sigma_{i=1}^{n}(X_i - \overline{X})^2$ where the sample size $n(\geq 2)$ is held fixed. Now, the two statistics $\overline{X}$ and S^2 would be independently distributed if and only if we can write

$$P_F\{\overline{X} \in A \cap S^2 \in B\} = P_F\{\overline{X} \in A\}P_F\{S^2 \in B\} \text{ for } all \text{ sets } A \subseteq \Re,$$

$B \subseteq \Re^+$ such that $A \times B$ belongs to the Borel sigma-field over $\Re \times \Re^+$.

$$(1.1)$$

Now, we present two illustrations successively through data analyses.

DATA ILLUSTRATION 1.1 In order to examine whether the dependence or independence between $\overline{X}, S^2$ can be checked out when we had some available data, we decided to generate random samples, each of size $n = 5$, from Normal(5,100) population. From each sample, we obtained the values of $\overline{x}, s$ thereby leading to the observed pairs $(\overline{x}_i, s_i), i = 1, ..., 500(= k, \text{ say})$. The respective frequency histograms for $\overline{x}$ and $u = (n - 1)s^2/\sigma^2$ are given in Figure 1. A joint plot of $\overline{x}$ and u is given in Figure 2.
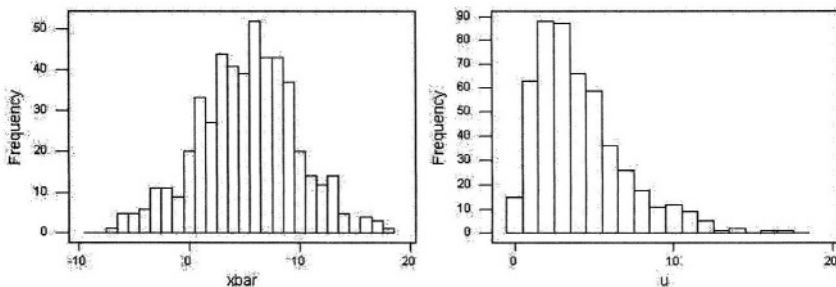
Figure 1.   Marginal frequency histograms of $\overline{x}$ and $u(= s^2/25)$ based on 500 observations from *N*(5,100) distribution
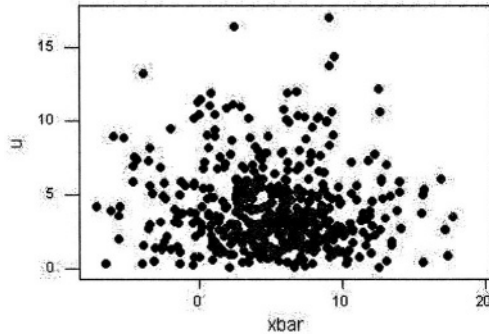
Figure 2. A plot of $\overline{x}$ and $u(= s^2/25)$ based on 500 observations from $N(5,100)$ distribution

In Figure 1, the $\overline{x}$ (or u) frequency histogram looks fairly symmetric (or skewed to the right). From the scatter plot in Figure 2, the $\overline{x}$ and u values seem to disperse independently of each other!

For a more formal test of significance, however, we formed a 4×3 table (Table 1) of count data based on the observations, indicating how many from 500 pairs $(\overline{x}_i, u_i)$ fell in each cell. Then, we simply used the customary Chi-square test of independence for the cell categories chosen for $\overline{x}, u$.

Table 1. Frequency and Expected Frequency of $(\overline{x}, u)$
In 500 Random Samples

$\overline{x}$	$(0,5)$	$u = s^2/25$ $[5,7]$	$(7,\infty)$	Total
$(-\infty,0)$	$38(= O_1)$	$10(= O_2)$	$12(= O_3)$	60
Exp. Freq.	$42.24(= E_1)$	$9.12(= E_2)$	$8.64(= E_3)$	
$[0,5]$	$117(= O_4)$	$29(= O_5)$	$27(= O_6)$	173
Exp. Freq.	$121.79(= E_4)$	$26.30(= E_5)$	$24.91(= E_6)$	
$(5,10]$	$150(= O_7)$	$28(= O_8)$	$27(= O_9)$	205
Exp. Freq.	$144.32(= E_7)$	$31.16(= E_8)$	$29.52(= E_9)$	
$(10,\infty)$	$47(= O_{10})$	$9(= O_{11})$	$6(= O_{12})$	62
Exp. Freq.	$43.65(= E_{10})$	$9.42(= E_{11})$	$8.93(= E_{12})$	
Total	352	76	72	500

At this point, we like to test the null hypotheses H_0 : Categories based on $\overline{x}, u$ are independent against the alternative hypotheses H_1 : Categories based on $\overline{x}, u$ are dependent, with the level of significance $\alpha = .05$. Let O_j and E_j respectively denote the observed and expected frequencies (under H_0) in the jth cell, $j = 1, ..., 12$. Then, the test statistic is given by

$$\chi^2_{calc} = \Sigma_{j=1}^{rc} \frac{(O_j - E_j)^2}{E_j} \text{ with the degree of freedom } \nu = (r-1)(c-1) \tag{1.2}$$

where r, c respectively denote the number of rows and columns.

Now, the test statistic is

$$
\begin{aligned}
\chi^2_{calc} = {} & \tfrac{(38-42.24)^2}{42.24} + \tfrac{(10-9.12)^2}{9.12} + \tfrac{(12-8.64)^2}{8.64} + \tfrac{(117-121.79)^2}{121.79} + \tfrac{(29-26.30)^2}{26.30} \\
& + \tfrac{(27-24.91)^2}{24.91} + \tfrac{(150-144.32)^2}{144.32} + \tfrac{(28-31.16)^2}{31.16} + \tfrac{(27-29.52)^2}{29.52} + \tfrac{(47-43.65)^2}{43.65} \\
& + \tfrac{(9-9.42)^2}{9.42} + \tfrac{(6-8.93)^2}{8.93} = 4.4545 \text{ with } \nu = (4-1)(3-1) = 6 \\
& \qquad \text{degrees of freedom; } P\text{-}value = 0.61535
\end{aligned}
\tag{1.3}
$$

$$\Rightarrow \text{We do not reject the null hypotheses } H_0 \text{ at 5\% level.}$$

That is, the observed data does not violate the postulate of independence between $\overline{x}, s^2$ values at 5% level. Incidentally, the *P-value* is calculated as follows:

$$
\begin{aligned}
P\text{-}value &= P\{\text{Observing more extreme data when } H_0 \text{ is true}\} \\
&= P\{\chi^2_\nu > 4.4545, \text{ the observed } \chi^2_{calc}, \text{ when } H_0 \text{ is true}\} \text{ with } \nu = 6.
\end{aligned}
\tag{1.4}
$$

We reject (do not reject) H_0 with the level of significance α if and only if the *P-value* is less (not less) than α. A "large" *P-value* indicates less evidence against the null hypothesis H_0.

REMARK 1.1. Before one applies the Chi-square test (1.2), one needs to make sure that the expected frequency in each cell, that is each $E_j, j = 1, ..., rc$, is five or more. Sometimes this restriction may severely impact on the number of cells that can be chosen.

REMARK 1.2. The sample correlation coefficient leading to a *t*-test is frequently used in practice to choose between the two hypotheses H_0, H_1 if $(\overline{X}, U)$ could be treated as a bivariate normal random variable. We had the Pearson-sample correlation coefficient $r^P_{\overline{x},u} = -0.080$ with the *P-value* $= 0.072$ which exceeded α, indicating that we should not reject the null hypotheses H_0 at 5% level. But, we may not rely upon this test because the underlying assumption of bivariate normality of $(\overline{X}, U)$ does not hold here (see Figure 1). On top of that, the *P-value barely* exceeded α!

REMARK 1.3. One may opt for a nonparametric approach to test H_0 versus H_1 by using the Spearman-rank correlation coefficient between the $\overline{x}, u$ data. Refer to Noether (1991, pp. 236-237), Lehmann (1986, pp. 350-351), or Gibbons and Chakraborti (1992, Chapter 12) for details. What one does first is to rank all k observations on $\overline{x}$ and u separately. Then, the Spearman-rank correlation coefficient between the $\overline{x}, u$ data, denoted by $r^S_{\overline{x},u}$, is simply the Pearson-sample correlation coefficient between the k two-dimensional vectors of ranks. For the observed data, we found $r^S_{\overline{x},u} = -0.091$. Under H_0, the probability distribution of the test statistic $Z = \sqrt{k-1}\, r^S_{\overline{x},u}$ is approximated by a standard normal distribution. One may refer to Noether (1991, pp. 236-237). We obtain $z_{calc} = -2.0328$, that is the associated *P-value* ≈ 0.042073 which unfortunately falls below the nominal 5% level, indicating that we should re-

ject the null hypotheses of independence at 5% level. Thus, for the same data, the *t*-test and the nonparametric test came up with opposite conclusions!

What will be different if the data is generated using a non-normal probability model? In order to get a feel for this, we provide the following illustration. DATA ILLUSTRATION 1.2 We generated 500 random samples, each of size $n = 5$, from a **Gamma**($\alpha = 4, \beta = 8$) model. From each sample, we obtained the values of $\bar{x}, s$ and the observed vectors $(\bar{x}_i, s_i), i = 1, ..., 500(= k)$. The respective frequency histograms for $\bar{x}$ and $u = (n-1)s^2/(\alpha\beta^2)$ are given in Figure 3. A joint plot of $\bar{x}$ and u is given in Figure 4.

In Figure 3, both $\bar{x}, u$ frequency histograms look very skewed to the right, particularly in comparison with Figure 1. From the scatter plot in Figure 4, the $\bar{x}, u$ values seem to disperse in a dependent fashion. For example, if we observe a "small" value of $\bar{x}$, then it seems unlikely that we will also observe a "large" value of u or equivalently a "large" value of s! For a more formal test of significance, however, we formed a 3×3 table (Table 2) of count data in each cell. Then, we simply used the Chi-square test (1.2).

We may like to test if the categories based on $\bar{x}, s$ are independent 5% level. The test statistic from (1.2) is given by

$$\chi^2_{calc} = \Sigma_{j=1}^9 \frac{(O_j - E_j)^2}{E_j} = 91.627 \text{ with 4 degrees of freedom; } P\text{-}value \approx 0$$
$$\Rightarrow \text{We reject the null hypotheses } H_0 \text{ at } 5\% \text{ level.}$$

We reject independence between $\bar{x}, s$ values at 5% level. In order to claim that $\bar{x}, s$ values are dependent, note that one simply needs to contradict (1.1) for some Borel sets *A, B*. In Table 2, we constructed a precise system of nine Borel sets for which the multiplicative probability rule quoted in (1.1) does not hold!
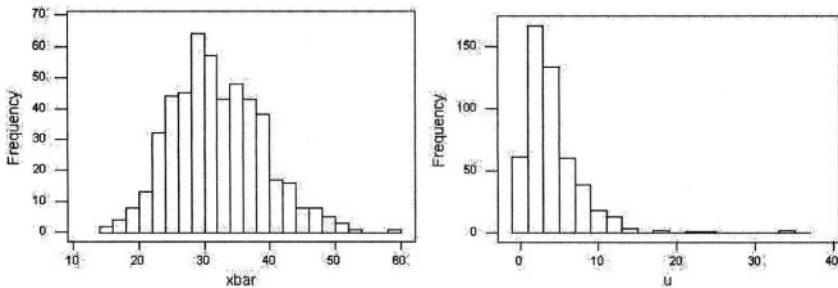
Figure 3. Marginal frequency histograms of $\bar{x}$ and $u(= s^2/64)$ based on 500 observations from Gamma(4,8) distribution
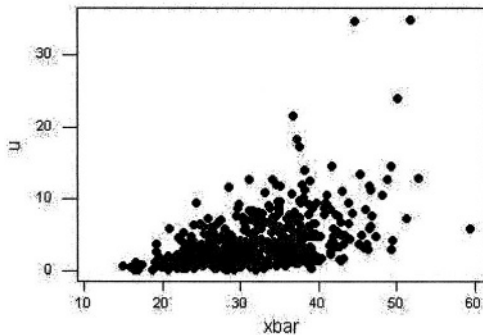
Figure 4.   A plot of $\bar{x}$ and $u(= s^2/64)$ based on 500 observations from Gamma(4,8) distribution

Table 2. Frequency and Expected Frequency of $(\bar{x}, s)$
In 500 Random Samples

$\bar{x}$	$(0, 30)$	$[30, 40]$	$(40, \infty)$	Total
		s		
$(0, 10)$	$75 (= O_1)$	$119 (= O_2)$	$9 (= O_3)$	203
Exp. Freq.	$47.10 (= E_1)$	$125.05 (= E_2)$	$30.86 (= E_3)$	
$[10, 20]$	$36 (= O_4)$	$162 (= O_5)$	$38 (= O_6)$	236
Exp. Freq.	$54.75 (= E_4)$	$145.38 (= E_5)$	$35.87 (= E_6)$	
$[20, \infty)$	$5 (= O_7)$	$27 (= O_8)$	$29 (= O_9)$	61
Exp. Freq.	$14.15 (= E_7)$	$37.58 (= E_8)$	$9.27 (= E_9)$	
Total	116	308	76	**500**

We had $r^P_{\bar{x},s} = 0.514$ between the $\bar{x}, s$ data with the *P-value* ≈ 0. That is, the *t-test* sides with the earlier conclusion to reject the null hypotheses of independence between the $\bar{x}, s$ data at 5% level. But, also see Remark 1.2.

One may again explore the nonparametric test. We found $r^S_{\bar{x},s} = 0.488$, that is the test statistic $z_{calc} = \sqrt{k-1} r^S_{\bar{x},s} \approx 10.901$ with the associated *P-value* ≈ 0. That is, we would reject the null hypotheses of independence between the $\bar{x}, s$ data at 5% level.

In these illustrations, one may want to know which of the two hypothesis was true?   The following result will address this.   Let us denote $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ and $\Phi(x) = \int_{-\infty}^{x} \phi(y) dy$ for $x \in \Re$.

THEOREM 1.1 *Suppose that $X_1, ..., X_n$ are iid random variables governed by a common distribution function $F(x), x \in \Re$. Then, the sample mean $\overline{X}$ and the sample variance S^2 are independently distributed if and only if the common distribution of the X's is $N(\mu, \sigma^2)$, that is $F(x) = \Phi((x - \mu)/\sigma)$ for some $\mu \in \Re, \sigma \in \Re^+$ and for all $x \in \Re$.*

The *if part* is a well-known result that may be verified easily with the help of Helmert transformations (Mukhopadhyay (2000), pp. 197-201). See Section 2 and Remark 2.1. The *only if part,* however, provides a characterization of a normal probability law which is quite hard to prove. Zinger's (1958) proof of the *only if part* requires deep analyses with an interplay of Cramér's (1946, pp. 151-165) fundamental results involving characteristic functions. Important historical notes may be found in Lukacs (1960), Ramachandran (1967, Chapter 8), and Kagan et al. (1973).

In Illustration 1.1, we generated data from a normal probability model, and hence we would have expected to favor H_0 with the help of a "large" *P-value.* On the other hand, in Illustration 1.2, we generated data from a gamma probability model, and hence we would have expected to favor H_1 with the help of a "small" *P-value* In the first illustration, both Chi-square and *t*-tests came up with the correct answer, but the nonparametric test gave a wrong answer. In the second illustration, all three tests came up with the correct answer. But, one needs to keep in mind that in situations like ours, a *t*-test is not reasonable any way! See Remark 1.2.

It is safe to say that among three contenders, the Chi-square test (1.2) thus far stands out as the most reliable exploratory data analytic (EDA) tool in validating the true state of nature of dependence (or independence) between $\overline{X}, S^2$ as evidenced by the observed paired data $\overline{x}, s$ when the observations $X_1, ..., X_n$ are assumed iid. As the story unfolds, one will see that the Chi-square test would remain most reliable in the same sense when the observations $X_1, ..., X_n$ are not iid or these are non-normal.

17.1.1.   What If the Observations Are Not IID or They Are Non-Normal?

In a direct contrast with the message from Theorem 1.1, what we find interesting is that the sample mean $\overline{X}$ and the variance S^2 may or may not follow independent probability models when the observations X_i's are not iid or when these follow some non-normal probability laws. We highlight examples depicting a number of interesting scenarios including the following:

(*i*) $\overline{X}, S^2$ follow independent probability models, each X's follows the same normal probability law, $\overline{X}$ has a normal probability model, S^2 has a Chi-square probability model, but the X's are dependent (Section 2);

(*ii*) $\overline{X}, S^2$ follow independent probability models, the X's follow non-identical but dependent normal probability laws, $\overline{X}$ has a normal probability model, S^2 has a (non-central) Chi-square probability model, when $n = 2$ (Section 3);

(*iii*) $\overline{X}, S^2$ follow dependent and uncorrelated probability models, $\overline{X}$ has a non-normal probability model, but X_1, X_2 both follow standard normal probability laws and they are dependent when $n = 2$ (Section 4);

(iv) $\overline{X}, S^2$ follow dependent and uncorrelated probability laws, X_1 follows a standard normal probability law, X_2 follows a mixture-normal symmetric probability law, $\overline{X}$ has a normal probability model, but X_1, X_2 are dependent when $n = 2$ (Section 5);

(v) $\overline{X}, S^2$ follow independent probability laws, $\overline{X}$ has a normal probability model, S^2 does not have a Chi-square probability model, even if the observations X_1, X_2 are governed by one common bi-modal mixture-normal symmetric probability law when $n = 2$ (Section 6);

(vi) $\overline{X}, S^2$ follow independent probability laws, $\overline{X}$ does not have a normal probability model, S^2 has a Chi-square probability model, even if the observations X_1, X_2 are governed by one common bi-modal mixture-normal symmetric probability law when $n = 2$ (Section 6.1);

(vii) $\overline{X}, S^2$ follow independent probability laws, $\overline{X}$ has a normal probability model, S^2 does not have a Chi-square probability model, even if the observations $X_1, ..., X_n$ are governed by one common mixture-normal symmetric probability law when $n \geq 2$ (Section 7, Example 7.1); and

(viii) $\overline{X}, S^2$ follow independent probability laws, $\overline{X}$ does not have a normal probability model, S^2 has a Chi-square probability model, even if the observations $X_1, ..., X_n$ are governed by one common mixture-normal symmetric probability law when $n \geq 2$ (Section 7, Example 7.2).

Each example, except the one mentioned in Section 2, is new as far as we know. The example cited in Section 2 was described in Rao (1973, pp. 196-197). In the abstract, we asked the following question: What can one expect regarding the status of independence or dependence between $\overline{X}, S^2$ when the random variables X's are allowed to be non-iid or non-normal? The specific examples described in Sections 2-7 should clearly highlight the point that there is a large array of interesting possibilities when the random variables X's are allowed to be non-iid or non-normal.

In order to formulate a general result, in our opinion, one has to focus on some particular nature of non-iid or non-normal probability model for the observations $X_1, ..., X_n$ and explore necessary and/or sufficient conditions for the independence between $\overline{X}, S^2$ to hold. The examples here show that one may expect contrasting results even within scenarios which are "close" to each other. In other words, one would necessarily proceed on a case by case basis with regard to differing aspects of how non-iid or how non-normal the joint probability models are. This article points toward an opening for the development of important characterization results and we hope to see some progress on this in the future.

In the case of Examples (iii)-(v), illustrations through simulated data are provided in our attempt to examine the performances of the Chi-square test, *t*-test, and the nonparametric test in detecting dependence of x_1, x_2 data as well as the dependence (or independence) of $\overline{x}, s^2$ data. In a number of occasions,

the *t*-test and the nonparametric test arrived at conflicting conclusions based on same data. We raise the potential of a major problem in implementing either a *t*-test or the nonparametric test as EDA tools to examine dependence or association for paired data in practice! The Chi-square test, however, correctly validated dependence under consideration in every single case, and this test never sided against a correct conclusion when the paired variables were in fact independent. We conclude that in this sense, the Chi-square test (1.2) stands out as the most reliable EDA tool whether the observations $X_1, ..., X_n$ are assumed iid, or not iid, or non-normal.

17.2   A Multivariate Normal Probability Model

This interesting situation in the context of a multivariate normal probability model was described in Rao (1973, pp. 196-197). Consider n-dimensional vector-valued random variable X where $\mathbf{X}' = (X_1, ..., X_n)$. We assume that X has the n-dimensional normal distribution $N_n(\mu\mathbf{1}, \sigma^2\mathbf{\Sigma})$ with $\mathbf{1}' = (1, ..., 1)_{1 \times n}, \mathbf{\Sigma}_{n \times n} = (1 - \rho)\mathbf{I}_{n \times n} + \rho\mathbf{1}\mathbf{1}'$ where $-\infty < \mu < \infty$, $0 < \sigma < \infty, \rho \in (-(n - 1)^{-1}, 1) - \{0\}$, and I is the $n \times n$ identity matrix.

We may define the associated Helmert variables (Mukhopadhyay (2000), pp. 197-201) $Y_1, Y_2, ..., Y_n$ where

$$Y_1 = (X_1 + ... + X_n)/\sqrt{n},$$
$$Y_i = \{X_1 + ... + X_{i-1} - (i - 1)X_i\}/\sqrt{i(i - 1)}, i = 2, ..., n. \tag{2.1}$$

This constitutes an orthogonal transformation from $(X_1, X_2, ..., X_n)$ to $(Y_1, Y_2, ..., Y_n)$. One can easily derive the joint probability model for $Y_1, Y_2, ..., Y_n$ from the assumed joint probability model of $X_1, X_2, ..., X_n$, and hence conclude in a straightforward manner that

the random variables $Y_1, Y_2, ..., Y_n$ are independent,
$$Y_1 \sim N(\mu\sqrt{n}, [1 + (n - 1)\rho]\sigma^2), \text{ and } Y_2, ..., Y_n \text{ are iid } N(0, (1 - \rho)\sigma^2). \tag{2.2}$$

Obviously, we also have

$$(n - 1)S^2 = \Sigma_{i=1}^n (X_i - \overline{X})^2 = \Sigma_{i=1}^n X_i^2 - n\overline{X}^2 = \Sigma_{i=1}^n Y_i^2 - Y_1^2 = \Sigma_{i=2}^n Y_i^2, \quad (2.3)$$

since $\Sigma_{i=1}^n X_i^2 = \Sigma_{i=1}^n Y_i^2$. Now, we note that $\overline{X}$ depends only on Y_1 whereas S^2 depends only on $(Y_2, ..., Y_n)$, but Y_1 is independent of $(Y_2, ..., Y_n)$. Hence, $\overline{X}$ and S^2 are independently distributed statistics. It is now quite straightforward to check that $\overline{X}(= Y_1/\sqrt{n})$ is distributed as $N(\mu, n^{-1}\sigma^2)$ and $(n - 1)S^2\sigma^{-2}$ is distributed as $(1 - \rho)\chi_{n-1}^2$. We may summarize the findings as follows:

$$\overline{X} \text{ is distributed as } N(\mu, n^{-1}[1 + (n - 1)\rho]\sigma^2),$$
$$(n - 1)S^2 \text{ is distributed as } (1 - \rho)\sigma^2\chi_{n-1}^2, \tag{2.4}$$
$$\text{and } \overline{X}, S^2 \text{ are independently distributed.}$$

In this example, one readily notices that the *if part* in Theorem 1.1 does hold, that is $\overline{X}$ and S^2 are *independent,* when each observation $X_1, ..., X_n$ follows the same $N(\mu, \sigma^2)$ probability law so that they are *identically distributed,* but $X_1, ..., X_n$ are *dependent* as ρ is assumed different from zero!

REMARK 2.1. A proof of the *if part* in Theorem 1.1 follows from the derivation given above if we assume that $\rho = 0$.

17.3   A Bivariate Normal Probability Model

Let us start with a two-dimensional random variable **X** where $\mathbf{X}' = (X_1, X_2)$ and let **X** be governed by the probability model $N_2(\mu, \sigma^2 \Sigma)$ where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} c+1 & c-1 \\ c-1 & c+1 \end{pmatrix}$$

with $-\infty < \mu_1, \mu_2 < \infty, \mu_1 \neq \mu_2, 0 < \sigma < \infty$, and $0 < c < 1$. In other words, we have X_1, X_2 respectively distributed as $N(\mu_1, (c+1)\sigma^2)$ and $N(\mu_2, (c+1)\sigma^2)$, but they are dependent.

Now, let us define two random variables $Y_1 = X_1 + X_2, Y_2 = X_1 - X_2$ and denote $\mathbf{Y}' = (Y_1, Y_2)$. Observe that any arbitrary linear function U of Y_1, Y_2 is clearly a linear function of X. Now, since X is distributed as N_2, the random variable U must have a univariate normal distribution. Thus, the random vector **Y** would have a bivariate normal probability model, say $N_2(\theta, \sigma^2 \Sigma^*)$ where $\theta' = (\theta_1, \theta_2), \theta_1 = \mu_1 + \mu_2, \theta_2 = \mu_1 - \mu_2, \Sigma^* = diag(4c, 4)$.

Thus, we have $\overline{X} = \frac{1}{2}Y_1 \sim N(\frac{1}{2}(\mu_1 + \mu_2), c\sigma^2)$ and $Y_2 \sim N(\mu_1 - \mu_2, 4\sigma^2)$. Obviously, Y_1, Y_2 have independent probability models since Σ^* is a diagonal matrix so that $\overline{X}$ and $S^2 = \frac{1}{2}(X_1 - X_2)^2 = \frac{1}{2}Y_2^2$ are independently distributed.

This example provides a different scenario from the one described in Section 2 when we fix $n = 2$. Here, we note that marginally X_1, X_2 have *non-identical* and *dependent normal* probability models. But, $\overline{X}$ and S^2 are *independent*!

17.4   Bivariate Non-Normal Probability Models: Case I

Let us denote the probability density function (pdf) of a bivariate normal probability model $N_2(\theta, \Sigma)$ with

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

by $g(w_1, w_2; \theta_1, \theta_2, \sigma_1, \sigma_2, \rho)$ where $-\infty < \theta_1, \theta_2 < \infty, 0 < \sigma_1, \sigma_2 < \infty$, $-1 < \rho < 1$, and $-\infty < w_1, w_2 < \infty$. In other words, let us denote

$$g(w_1, w_2; \theta_1, \theta_2, \sigma_1, \sigma_2, \rho) = c \exp\left[-\frac{1}{2}(1-\rho^2)^{-1}\{u_1^2 - 2\rho u_1 u_2 + u_2^2\}\right]$$

$$(4.1)$$

where

$$u_1 = (w_1 - \theta_1)/\sigma_1, u_2 = (w_2 - \theta_2)/\sigma_2,$$

$$c = \{2\pi\sigma_1\sigma_2(1 - \rho^2)^{\frac{1}{2}}\}^{-1}, -\infty < w_1, w_2 < \infty. \tag{4.2}$$

Now, we construct an example of a two-dimensional random variable $\mathbf{X}$ with $\mathbf{X}' = (X_1, X_2)$ where X_1, X_2 both follow the $N(0,1)$ probability law, X_1, X_2 have dependent probability models, neither $X_1 + X_2$ nor $X_1 - X_2$ follows a normal probability law, but $\overline{X}, S^2$ have *dependent* probability models. Here, one finds an example where X_1, X_2 have *identical* but *dependent normal* probability laws, and yet $\overline{X}, S^2$ have *dependent* probability models.

To be specific, we consider the joint probability model for an observation $\mathbf{X}_{2\times 1}$ governed by the pdf

$$f(x_1, x_2; \alpha, \rho) = \alpha g(x_1, x_2; 0, 0, 1, 1, \rho) + (1 - \alpha)g(x_1, x_2; 0, 0, 1, 1, -\rho) \tag{4.3}$$

for $-\infty < x_1, x_2 < \infty, 0 < \alpha, \rho < 1$. The pdf given in (4.3) is a mixture of two bivariate normal models.

THEOREM 4.1 *Suppose that (X_1, X_2) has the joint pdf from (4.3). Let us denote $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Then, for all $0 < \alpha, \rho < 1$, we have the following:*

   (i)   *Both X_1, X_2 have a standard normal probability model, but these are dependent;*
   (ii)   *The joint probability model of Y_1, Y_2 is governed by the pdf from (4.4), but $\overline{X}$ has a mixture normal probability model with its pdf from (4.7), and Y_2 has analogous mixture normal probability model with its pdf from (4.6);*
   (iii) *Y_1, Y_2 are dependent, and so are $\overline{X}, S^2$;*
   (iv)   *Y_1, Y_2^q are uncorrelated, $q = 1, 2$;*
   (v)   *$\overline{X}, S^2$ are uncorrelated.*

PROOF (i) From the joint pdf $f(x_1, x_2; \alpha, \rho)$, by integrating x_1 or x_2 out, one easily verifies that marginally both X_1, X_2 have a standard normal probability model so that their common pdf is $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$ with $-\infty < x < \infty$. Now, observe that $f(0, 0; \alpha, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \neq \phi^2(0) = \frac{1}{2\pi}$, whatever be $0 < \alpha, \rho < 1$. Thus, the random variables X_1, X_2 have *identical* but *dependent* probability models.

(ii) Next, we consider Y_1, Y_2 and then with the help of the one-to-one transformation from $(x_1, x_2) \to (y_1, y_2)$ we can write down the joint pdf of Y_1, Y_2. Toward this end, we begin with

$$f(x_1, x_2; \alpha, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}}\Big\{\alpha\exp[-\frac{1}{2(1-\rho^2)}(x_1^2 + x_2^2 - 2\rho x_1 x_2)]$$

$$+ (1 - \alpha)\exp[-\frac{1}{2(1-\rho^2)}(x_1^2 + x_2^2 + 2\rho x_1 x_2)]\Big\},$$

for $-\infty < x_1, x_2 < \infty$. Observe that $x_1 = \frac{1}{2}(y_1 + y_2), x_2 = \frac{1}{2}(y_1 - y_2)$ and hence the Jacobian matrix amounts to

$$J = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \Rightarrow |\det(J)| = \frac{1}{2}.$$

Thus, the joint probability model of Y_1, Y_2 is governed by the following pdf:

$$h(y_1, y_2; \alpha, \rho)$$
$$= \tfrac{1}{2\pi\sqrt{1-\rho^2}}\left\{\alpha\exp[-\tfrac{1}{4(1-\rho^2)}(y_1^2 + y_2^2 - \rho y_1^2 + \rho y_2^2)]\right.$$
$$\left.+ (1-\alpha)\exp[-\tfrac{1}{4(1-\rho^2)}(y_1^2 + y_2^2 + \rho y_1^2 - \rho y_2^2)]\right\} \times \tfrac{1}{2},$$

which simplifies to the expression

$$h(y_1, y_2; \alpha, \rho) = \frac{1}{4\pi\sqrt{1-\rho^2}}\left[\alpha e^{-\frac{1}{4}y_1^2/(1+\rho)}e^{-\frac{1}{4}y_2^2/(1-\rho)}\right.$$
$$\left.+(1-\alpha)e^{-\frac{1}{4}y_1^2/(1-\rho)}e^{-\frac{1}{4}y_2^2/(1+\rho)}\right]$$

$$\text{for } -\infty < y_1, y_2 < \infty. \tag{4.4}$$

Now, by integrating y_1 or y_2 out from the joint pdf $h(y_1, y_2; \alpha, \rho)$ one easily verifies that the marginal pdf's of Y_1, Y_2 are respectively given by

$$h_1(y_1; \alpha, \rho) = \alpha\frac{1}{2\sqrt{\pi(1+\rho)}}e^{-\frac{1}{4}y_1^2/(1+\rho)} + (1-\alpha)\frac{1}{2\sqrt{\pi(1-\rho)}}e^{-\frac{1}{4}y_1^2/(1-\rho)}$$

$$\text{for } -\infty < y_1 < \infty,$$

$$\tag{4.5}$$

$$h_2(y_2; \alpha, \rho) = \alpha\frac{1}{2\sqrt{\pi(1-\rho)}}e^{-\frac{1}{4}y_2^2/(1-\rho)} + (1-\alpha)\frac{1}{2\sqrt{\pi(1+\rho)}}e^{-\frac{1}{4}y_2^2/(1+\rho)}$$

$$\text{for } -\infty < y_2 < \infty.$$

$$\tag{4.6}$$

Both $h_1(y_1; \alpha, \rho), h_2(y_2; \alpha, \rho)$ happen to be mixtures of $N(0, 2 - 2\rho)$ and $N(0, 2 + 2\rho)$ distributions. From (4.5), it is obvious that the probability model for $U = \bar{X}(= \frac{1}{2}Y_1)$ will be governed by the pdf

$$h^*(u; \alpha, \rho) = \alpha\frac{1}{\sqrt{\pi(1+\rho)}}e^{-u^2/(1+\rho)} + (1-\alpha)\frac{1}{\sqrt{\pi(1-\rho)}}e^{-u^2/(1-\rho)}$$

$$\text{for } -\infty < u < \infty. \tag{4.7}$$

The pdf $h^*(u; \alpha, \rho)$ happens to be a mixture of $N(0, \frac{1}{2}(1 - \rho))$ and $N(0, \frac{1}{2}(1 + \rho))$ distributions.

(*iii*) Next, by combining (4.4)-(4.6) we observe that $h(0, 0; \alpha, \rho) = \frac{1}{4\pi\sqrt{1-\rho^2}}$ whereas

$$h_1(0; \alpha, \rho) = \frac{1}{2\sqrt{\pi}}\left[\frac{\alpha}{\sqrt{1+\rho}} + \frac{1-\alpha}{\sqrt{1-\rho}}\right],$$

and

$$h_2(0; \alpha, \rho) = \frac{1}{2\sqrt{\pi}}\left[\frac{\alpha}{\sqrt{1-\rho}} + \frac{1-\alpha}{\sqrt{1+\rho}}\right]$$

In other words, we would conclude that $h(0,0;\alpha,\rho) = h_1(0;\alpha,\rho)h_2(0;\alpha,\rho)$ if and only if

$$\frac{1}{\sqrt{1-\rho^2}} = [\tfrac{\alpha}{\sqrt{1+\rho}} + \tfrac{1-\alpha}{\sqrt{1-\rho}}][\tfrac{\alpha}{\sqrt{1-\rho}} + \tfrac{1-\alpha}{\sqrt{1+\rho}}]$$

$$\Leftrightarrow 1 = [\alpha\sqrt{1-\rho} + (1-\alpha)\sqrt{1+\rho}][\alpha\sqrt{1+\rho} + (1-\alpha)\sqrt{1-\rho}]$$

$$\Leftrightarrow \rho^2 = 1 - [1 - 2\alpha(1-\alpha)]^2[\alpha^2 + (1-\alpha)^2]^{-2} \equiv 0,$$

since $\alpha^2 + (1-\alpha)^2 = 1 - 2\alpha(1-\alpha)$. But, we have assumed that ρ is positive! That is, we have $h(0,0;\alpha,\rho) \neq h_1(0;\alpha,\rho)h_2(0;\alpha,\rho)$, whatever be $0 < \alpha < 1$. Hence, whatever be $0 < \alpha < 1$, the random variables Y_1, Y_2 are *dependent*, that is $\overline{X}, S^2$ have *dependent* probability models since $\overline{X} = \tfrac{1}{2}Y_1, S^2 = \tfrac{1}{2}Y_2^2$.

(*iv*) We obviously have $E[Y_1] = E[Y_2] = 0, V[X_1] = V[X_2] = 1,$ $V[Y_1] = V[Y_2] = 2[1 - \rho(1 - 2\alpha)]$.

Also, note that $Cov(Y_1, Y_2) = V[X_1] - V[X_2] = 1 - 1 = 0$, so that the Pearson correlation coefficient between the observations Y_1, Y_2 is given by $\rho_{Y_1,Y_2} = Cov(Y_1, Y_2)/\sqrt{V[Y_1]V[Y_2]} = 0$. That is, the observations Y_1, Y_2 are *uncorrelated* whatever be $0 < \alpha, \rho < 1$.

Next, using (4.4), let us evaluate the covariance between the random variables Y_1, Y_2 and express

$$Cov(Y_1, Y_2^2) = E[Y_1 Y_2^2] - E[Y_1]E[Y_2^2] = E[Y_1 Y_2^2],$$

whereas $E[Y_1 Y_2^2]$ may be found as follows:

$$\frac{1}{4\pi\sqrt{1-\rho^2}}\Big[\alpha \int_{y_2=-\infty}^{\infty}\int_{y_1=-\infty}^{\infty} y_1 y_2^2 e^{-\frac{1}{4}y_1^2/(1+\rho)} e^{-\frac{1}{4}y_2^2/(1-\rho)} dy_1 dy_2$$

$$+ (1-\alpha)\int_{y_2=-\infty}^{\infty}\int_{y_1=-\infty}^{\infty} y_1 y_2^2 e^{-\frac{1}{4}y_1^2/(1-\rho)} e^{-\frac{1}{4}y_2^2/(1+\rho)} dy_1 dy_2\Big]$$

$$= \frac{1}{4\pi\sqrt{1-\rho^2}}\Big[\alpha \int_{y_2=-\infty}^{\infty} y_2^2 e^{-\frac{1}{4}y_2^2/(1-\rho)} dy_2 \int_{y_1=-\infty}^{\infty} y_1 e^{-\frac{1}{4}y_1^2/(1+\rho)} dy_1 \qquad (4.8)$$

$$+ (1-\alpha)\int_{y_2=-\infty}^{\infty} y_2^2 e^{-\frac{1}{4}y_2^2/(1+\rho)} dy_2 \int_{y_1=-\infty}^{\infty} y_1 e^{-\frac{1}{4}y_1^2/(1-\rho)} dy_1\Big]$$

$$= 0.$$

Also, $V[Y_1]$ and $V[Y_2^2]$ are both finite so that the Pearson correlation coefficient between the observations Y_1, Y_2^2 is given by

$$\rho_{Y_1,Y_2^2} = Cov(Y_1, Y_2^2)/\sqrt{V[Y_1]V[Y_2^2]} = 0.$$

Thus, Y_1, Y_2^2 are *uncorrelated*.

(*v*) This follows from part (*iv*) since $\overline{X} = \tfrac{1}{2}Y_1$ and $S^2 = \tfrac{1}{2}Y_2^2$. ∎

REMARK 4.1 Recall that $E[X_1] = E[X_2] = 0$ and $V[X_1] = V[X_2] = 1$. Also, we can easily write

$$E[X_1 X_2] = \alpha\rho + (1-\alpha)(-\rho) = (2\alpha - 1)\rho.$$

Thus, we have

$$Cov(X_1, X_2) = E[X_1 X_2] = (2\alpha - 1)\rho,$$

so that the Pearson correlation coefficient between the observations X_1, X_2 is given by $\rho_{X_1,X_2} = Cov(X_1, X_2)/\sqrt{V[X_1]V[X_2]} = (2\alpha - 1)\rho$. Hence, whatever be $0 < \rho < 1$, the observations X_1, X_2 are *uncorrelated* if and only if $\alpha = \frac{1}{2}$.

DATA ILLUSTRATION 4.1 We focus on working under the pdf from (4.3) when $\alpha = \rho = \frac{1}{2}$ and compare performances of the Chi-square test, *t*-test, and the nonparametric test in detecting dependence within (x_1, x_2) data and within (y_1, y_2) data. Thus, we generated 500 random pairs $(x_{1i}, x_{2i}), i = 1, ..., 500$ $(= k)$ governed by the joint probability model (4.3) with $\alpha = \rho = \frac{1}{2}$. Subsequently, we obtained (y_{1i}, y_{2i}) where $y_{1i} = x_{1i} + x_{2i}, y_{2i} = x_{1i} - x_{2i}$, $i = 1, ..., k$. The frequency histograms for x_1, x_2 and y_1, y_2 are given in Figures 5-6. The plots of x_1 vs x_2 and y_1 vs y_2 are given in Figure 7.
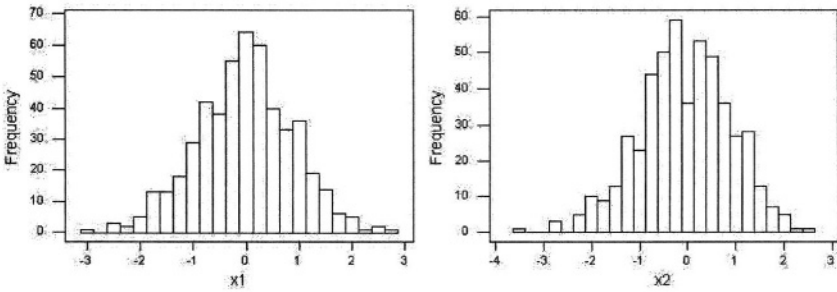
Figure 5.   Marginal frequency histograms of x_1 and x_2 obtained from observations with the joint distribution (4.3), $\alpha = \rho = \frac{1}{2}$
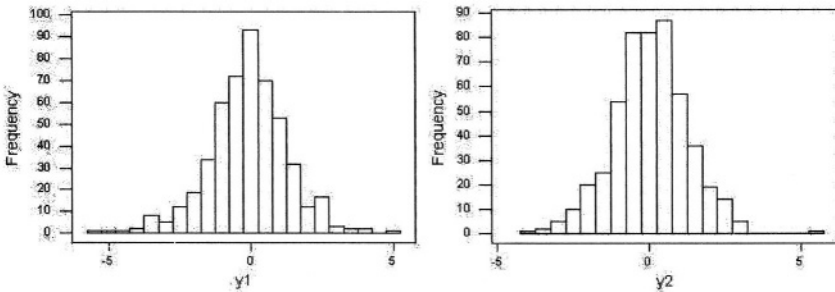
Figure 6.   Marginal frequency histograms of $y_1 = x_1 + x_2$ and $y_2 = x_1 - x_2$ obtained from (x_1, x_2) observations with the joint distribution (4.3), $\alpha = \rho = \frac{1}{2}$

From Figures 5-6, we observe that the frequency histograms for x_1, x_2 and y_1, y_2 have heavy tails on either side. In Figure 7, the two scatter plots seem to indicate that both x_1, x_2 and y_1, y_2 data are dependent as they are expected to be so.

For a test of significance, however, we formed a 4×4 table (Table 3) of count data of how many pairs (x_{1i}, x_{2i}) fell in each cell. Then, we used the Chi-square test (1.2).

Next, we test whether the categories based on the x_1, x_2 data are independent at 5% level, and the test statistic from (1.2) is given by

$$\chi^2_{calc} = \Sigma_{j=1}^{16} \frac{(O_j - E_j)^2}{E_j} = 34.611 \text{ with 9 degrees of freedom; } P\text{-value} = 0.0000698$$
$$\Rightarrow \text{We reject the null hypotheses } H_0 \text{ at 5% level.}$$

Since the *P-value* is "small", we reject the hypothesis of independence between x_1, x_2 values at 5% level.
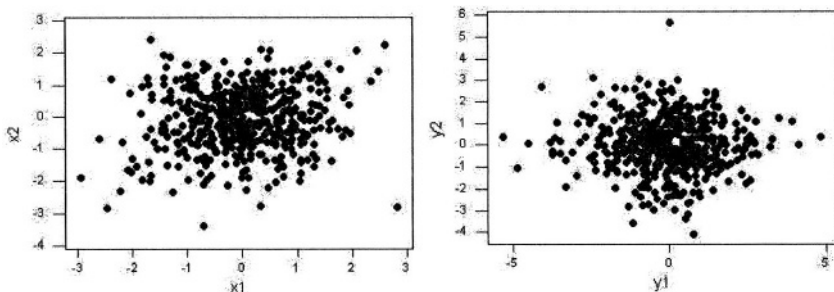
Figure 7. Plots of x_1 vs x_2 and y_1 vs y_2 obtained from (x_1, x_2) observations with the joint distribution (4.3), $\alpha = \rho = \frac{1}{2}$

Table 3. Frequency and Expected Frequency of (x_1, x_2)
In 500 Random Samples

x_1	$(-\infty, -1)$	$[-1, 0]$	$(0, 1]$	$[1, \infty)$	Total
			x_2		
$(-\infty, -1)$	$17(= O_1)$	$20(= O_2)$	$14(= O_3)$	$13(= O_4)$	64
Exp. Freq.	$9.60(= E_1)$	$23.68(= E_2)$	$22.40(= E_3)$	$8.32(= E_4)$	
$[-1, 0]$	$27(= O_5)$	$72(= O_6)$	$74(= O_7)$	$14(= O_8)$	187
Exp. Freq.	$28.05(= E_5)$	$69.19(= E_6)$	$65.45(= E_7)$	$24.31(= E_8)$	
$(0, 1]$	$21(= O_9)$	$73(= O_{10})$	$71(= O_{11})$	$20(= O_{12})$	185
Exp. Freq.	$27.75(= E_9)$	$68.45(= E_{10})$	$64.75(= E_{11})$	$24.05(= E_{12})$	
$[1, \infty)$	$10(= O_{13})$	$20(= O_{14})$	$16(= O_{15})$	$18(= O_{16})$	64
Exp. Freq.	$9.60(= E_{13})$	$23.68(= E_{14})$	$22.40(= E_{15})$	$8.32(= E_{16})$	
Total	75	185	175	65	**500**

Similarly, we formed a 4×4 table (Table 4) of count data of pairs (y_{1i}, y_{2i}) that fell in each cell and proceeded to use the Chi-square test (1.2).

Now, for testing the independent of the categories based on y_1, y_2 values at 5% level, the test from (1.2) gives

$$\chi^2_{calc} = \Sigma_{j=1}^{16} \frac{(O_j - E_j)^2}{E_j} = 17.722 \text{ with 9 degrees of freedom; } P\text{-}value = 0.038539$$

$$\Rightarrow \text{We reject the null hypotheses } H_0 \text{ at 5\% level.}$$

We reject independence between y_1, y_2 values at 5% level since we observe a "small" *P-value.*

With regard to *t-tests,* we respectively found $r^P_{x_1, x_2} = 0.124$ and $r^P_{y_1, y_2} = -0.045$ with associated *P-values* $= 0.005$ and 0.315. That is, the *t-test* based on $r^P_{x_1, x_2}$ will side with the conclusion that x_1, x_2 data are dependent at 5% level, but an analogous *t-test* based on $r^P_{y_1, y_2}$ unfortunately gives a wrong message at 5% level!

Table 4. Frequency and Expected Frequency of (y_1, y_2)
In 500 Random Samples

			y_2		
y_1	$(-\infty, -1)$	$[-1, 0]$	$(0, 1]$	$[1, \infty)$	Total
$(-\infty, -1)$	$14 (= O_1)$	$31 (= O_2)$	$35 (= O_3)$	$25 (= O_4)$	105
Exp. Freq.	$17.43 (= E_1)$	$33.18 (= E_2)$	$31.71 (= E_3)$	$22.68 (= E_4)$	
$[-1, 0]$	$32 (= O_5)$	$40 (= O_6)$	$56 (= O_7)$	$31 (= O_8)$	159
Exp. Freq.	$26.39 (= E_5)$	$50.24 (= E_6)$	$48.02 (= E_7)$	$34.34 (= E_8)$	
$(0, 1]$	$27 (= O_9)$	$56 (= O_{10})$	$28 (= O_{11})$	$29 (= O_{12})$	140
Exp. Freq.	$23.24 (= E_9)$	$44.24 (= E_{10})$	$42.28 (= E_{11})$	$30.24 (= E_{12})$	
$[1, \infty)$	$10 (= O_{13})$	$31 (= O_{14})$	$32 (= O_{15})$	$23 (= O_{16})$	96
Exp. Freq.	$15.94 (= E_{13})$	$30.34 (= E_{14})$	$28.99 (= E_{15})$	$20.74 (= E_{16})$	
Total	83	158	151	108	500

Next, with regard to the nonparametric test, we found $r^S_{x_1, x_2} = 0.090$ and $r^S_{y_1, y_2} = -0.058$, along with test statistics $z_{calc} = \sqrt{k-1} r^S_{x_1, x_2} \approx 2.0104$ and $\sqrt{k-1} r^S_{y_1, y_2} \approx -1.2956$ respectively. The associated *P-values* were 0.044389 and 0.19511 respectively, indicating that we would (would not) reject the hypotheses of independence between x_1, x_2 values (y_1, y_2 values) at 5% level. That is, the nonparametric test for y_1, y_2 data leads to an incorrect inference in this example!

17.5   Bivariate Non-Normal Probability Models: Case II

Let us repeat the earlier notation from Section 4. Now, we give an example of a two-dimensional random variable X with $\mathbf{X}' = (X_1, X_2)$ where X_1 is

normally distributed, X_2 is not normally distributed, $\overline{X}$ is normally distributed, $X_1 - X_2$ is not normally distributed, but $\overline{X}$ and S^2 are *dependent* random variables.

Recall the function $g(w_1, w_2; \theta_1, \theta_2, \sigma_1, \sigma_2, \rho)$ from (4.1). Suppose that $\mathbf{X}_{2 \times 1}$ has its pdf given by

$$f(x_1, x_2; \alpha) = \alpha g(x_1, x_2; 0, 0, 1, 2, 0.5) + (1 - \alpha)g(x_1, x_2; 0, 0, 1, 3, -0.5) \tag{5.1}$$

for $-\infty < x_1, x_2 < \infty, 0 < \alpha < 1$.

THEOREM 5.1 *Suppose that (X_1, X_2) has the joint pdf from (5.1). Let us denote $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Then, for all $0 < \alpha < 1$, we have the following:*

(i) X_1 *has the standard normal probability model, X_2 has a mixture normal probability model governed by the pdf from (5.2), and they are dependent;*

(ii) *The joint probability model of Y_1, Y_2 is governed by the pdf from (5.3), but $\overline{X}$ has $N(0, \frac{7}{4})$ distribution with pdf from (5.4), and Y_2 has a mixture normal probability model with its pdf from (5.5);*

(iii) Y_1, Y_2 *are dependent, and so are $\overline{X}, S^2$;*

(iv) Y_1, Y_2 *are correlated, but Y_1, Y_2^2 are uncorrelated;*

(v) $\overline{X}, S^2$ *are uncorrelated.*

PROOF (i) From the joint pdf $f(x_1, x_2; \alpha)$, by integrating x_1 or x_2 out, one easily verifies that X_1 has the $N(0,1)$ distribution with its pdf $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ for $-\infty < x < \infty$, but the marginal pdf of X_2 is given by

$$f_2(x_2; \alpha) = \alpha \frac{1}{2\sqrt{2\pi}} e^{-x_2^2/8} + (1 - \alpha) \frac{1}{3\sqrt{2\pi}} e^{-x_2^2/18} \text{ for } -\infty < x_2 < \infty, \tag{5.2}$$

whatever be $0 < \alpha < 1$. It is clear that $f_2(x_2; \alpha)$ happens to be a mixture of $N(0,4)$ and $N(0,9)$ probability models.

Now, observe that $f(0,0; \alpha) = \frac{1}{6\sqrt{3}\pi}(2 + \alpha)$ whereas $\phi(0)f_2(0; \alpha) = \frac{1}{12\pi}(2 + \alpha)$, so that we have $f(0,0; \alpha) \neq \phi(0)f_2(0; \alpha)$. Hence, the random variables X_1, X_2 have the *dependent* probability models.

(ii) We have $Y_1 = X_1 + X_2, Y_2 = X_1 - X_2$. Then, along the line of derivation for Theorem 4.1 part *(ii)*, we can again use transformation techniques to express the joint pdf of Y_1, Y_2 as follows:

$$h(y_1, y_2; \alpha) = \alpha g\left(y_1, y_2; 0, 0, \sqrt{7}, \sqrt{3}, -\frac{3}{\sqrt{21}}\right)$$
$$+ (1 - \alpha)g\left(y_1, y_2; 0, 0, \sqrt{7}, \sqrt{13}, -\frac{8}{\sqrt{91}}\right) \tag{5.3}$$

for $-\infty < y_1, y_2 < \infty$. From (5.3), it is obvious that marginally, Y_1 is distributed as $N(0,7)$ with its pdf

$$h_1(y_1) = \frac{1}{\sqrt{7}\sqrt{2\pi}} e^{-y_1^2/14} \text{ for } -\infty < y_1 < \infty, \tag{5.4}$$

whatever be $0 < \alpha < 1$. However, the marginal pdf of Y_2 is given by

$$h_2(y_2; \alpha) = \alpha \frac{1}{\sqrt{3}\sqrt{2\pi}} e^{-y_2^2/6} + (1-\alpha) \frac{1}{\sqrt{13}\sqrt{2\pi}} e^{-y_2^2/26} \text{ for } -\infty < y_2 < \infty, \tag{5.5}$$

which happens to be a mixture of $N(0,3)$ and $N(0,13)$ probability models. Obviously, $\overline{X} = \frac{1}{2}Y_1$ is distributed as $N(0, \frac{7}{4})$.

(iii) Next, by combining (5.3)-(5.5) we observe that $h(0,0;\alpha) = \frac{1}{2\pi}(\frac{\alpha}{\sqrt{12}} + \frac{1-\alpha}{\sqrt{27}})$ whereas $h_1(0) = \frac{1}{\sqrt{14\pi}}$, and $h_2(0;\alpha) = \frac{1}{\sqrt{2\pi}}[\frac{\alpha}{\sqrt{3}} + \frac{1-\alpha}{\sqrt{13}}]$. In other words, we would conclude that $h(0,0;\alpha) = h_1(0)h_2(0;\alpha)$ if and only if

$$\frac{1}{2\pi}(\frac{\alpha}{\sqrt{12}} + \frac{1-\alpha}{\sqrt{27}}) = \frac{1}{\sqrt{14\pi}}\frac{1}{\sqrt{2\pi}}[\frac{\alpha}{\sqrt{3}} + \frac{1-\alpha}{\sqrt{13}}]$$
$$\Leftrightarrow \frac{\alpha}{1-\alpha} = (\frac{1}{\sqrt{91}} - \frac{1}{\sqrt{27}})/(\frac{1}{\sqrt{12}} - \frac{1}{\sqrt{21}}),$$

which is a negative number! But, we have assumed that $\alpha \in (0,1)$ so that we immediately conclude that $h(0,0;\alpha) \neq h_1(0)h_2(0;\alpha)$ whatever be $0 < \alpha < 1$. Hence, for all $0 < \alpha < 1$, the random variables Y_1, Y_2 are *dependent,* that is $\overline{X}, S^2$ also have *dependent* probability models since $\overline{X} = \frac{1}{2}Y_1, S^2 = \frac{1}{2}Y_2^2$.

(iv) From (5.4)-(5.5), we obviously have $E[Y_1] = E[Y_2] = 0, V[Y_1] = 7$, and $V[Y_2] = 13 - 10\alpha$. From (5.3), we note that

$$Cov(Y_1, Y_2) = E[Y_1 Y_2]$$
$$= \alpha(-\frac{3}{\sqrt{21}})(\sqrt{7})(\sqrt{3}) + (1-\alpha)(-\frac{8}{\sqrt{91}})(\sqrt{7})(\sqrt{13})$$
$$= 5\alpha - 8,$$

which is certainly non-zero. Hence, the Pearson correlation coefficient between the observations Y_1, Y_2 is given by

$$\rho_{Y_1, Y_2} = Cov(Y_1, Y_2)/\sqrt{V[Y_1]V[Y_2]} = (5\alpha - 8)/\sqrt{7(13 - 10\alpha)}.$$

Hence, the observations Y_1, Y_2 are *correlated* whatever be $0 < \alpha < 1$.

Next, using (5.3) again, let us evaluate the covariance between the random variables Y_1, Y_2^2 and express

$$Cov(Y_1, Y_2^2) = E[Y_1 Y_2^2] - E[Y_1]E[Y_2^2] = E[Y_1 Y_2^2],$$

whereas $E[Y_1 Y_2^2]$ may be found as follows:

$$E[Y_1 E\{Y_2^2 \mid Y_1\}]$$
$$= \alpha E[Y_1\{\tfrac{12}{7} + \tfrac{9}{49}Y_1^2\}] + (1-\alpha)E[Y_1\{\tfrac{27}{7} + \tfrac{64}{49}Y_1^2\}] \tag{5.6}$$
$$= 0, \text{ since } E[Y_1] = E[Y_1^3] = 0.$$

Again, both $V[Y_1], V[Y_2^2]$ are finite so that the Pearson correlation coefficient between the observations Y_1, Y_2^2 is given by

$$\rho_{Y_1,Y_2^2} = Cov(Y_1, Y_2^2)/\sqrt{V[Y_1]V[Y_2^2]} = 0$$

Hence, Y_1, Y_2^2 are uncorrelated.

(v) This follows from part (iv) since $\overline{X} = \frac{1}{2}Y_1$ and $S^2 = \frac{1}{2}Y_2^2$. ∎

REMARK 5.1 We note that $E[X_1] = E[X_2] = 0, V[X_1] = 1$, $V[X_2] = E[X_2^2] = 4\alpha + 9(1 - \alpha) = 9 - 5\alpha$. We can also express

$$E[X_1 X_2] = \alpha(1)(2)\tfrac{1}{2} + (1 - \alpha)(1)(3)(-\tfrac{1}{2}) = \tfrac{1}{2}(5\alpha - 3).$$

Thus, we have

$$Cov(X_1, X_2) = E[X_1 X_2] = \tfrac{1}{2}(5\alpha - 3),$$

so that $\rho_{X_1,X_2} = \frac{1}{2}(5\alpha - 3)/\sqrt{(9 - 5\alpha)}$. That is, the observations X_1, X_2 are *uncorrelated* if and only if $\alpha = \frac{3}{5}$.

DATA ILLUSTRATION 5.1 We focus on working under the pdf from (5.1) when $\alpha = \frac{1}{2}$ and compare performances of the Chi-square test, *t*-tests, and the non-parametric test in detecting dependence for (x_1, x_2) and (y_1, y_2) data. Thus, we generated random pairs $(x_{1i}, x_{2i}), i = 1, ..., 500(= k)$ governed by the joint probability model (5.1) with $\alpha = \frac{1}{2}$. Subsequently, we obtained (y_{1i}, y_{2i}) where $y_{1i} = x_{1i} + x_{2i}, y_{2i} = x_{1i} - x_{2i}, i = 1, ..., k$. The frequency histograms for x_1, x_2 and y_1, y_2 are given in Figures 8-9. The plots of x_1 vs x_2 and y_1 vs y_2 are given in Figure 10.
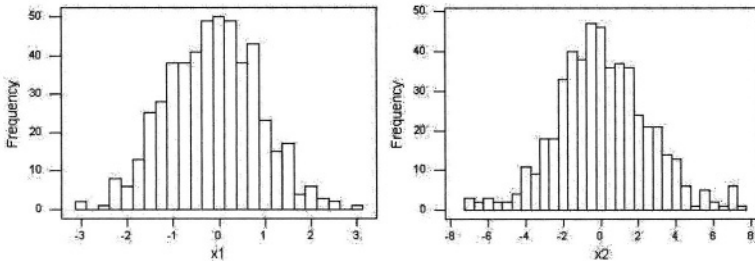
Figure 8. Marginal frequency histograms of x_1 and x_2 obtained from observations with the joint distribution (5.1), $\alpha = \frac{1}{2}$

From Figure 8, we observe that the frequency histograms for both x_1, x_2 are skewed, whereas from Figure 9, the frequency histograms for both y_1, y_2 have heavy tails on either side. In Figure 10, the two scatter plots seem to indicate that both x_1, x_2 and y_1, y_2 are dependent as they are expected to be.

For a more formal test of significance, however, we formed a 5×3 table (Table 4) of count data of how many pairs (x_{1i}, x_{2i}) fell in each cell and used
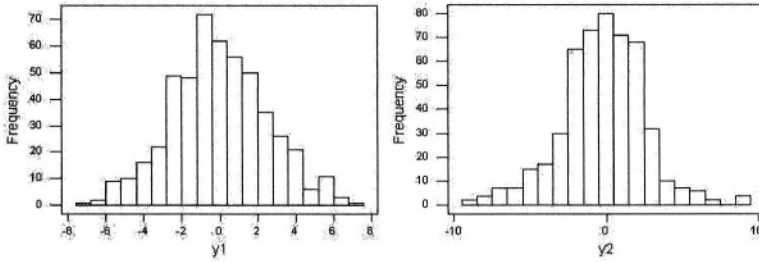
Figure 9. Marginal frequency histograms of $y_1 = x_1 + x_2$ and $y_2 = x_1 - x_2$ obtained from (x_1, x_2) observations with the joint distribution (5.1), $\alpha = \frac{1}{2}$
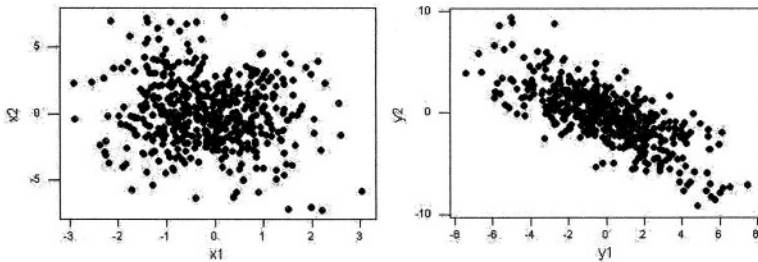
Figure 10. Plots of x_1 vs x_2 and y_1 vs y_2 obtained from (x_1, x_2) observations with the joint distribution (5.1), $\alpha = \frac{1}{2}$

the Chi-square test (1.2).

Table 5. Frequency and Expected Frequency of (x_1, x_2)
In 500 Random Samples

x_1	$(-\infty, -1)$	x_1 $[-1, 0]$	$(0, \infty)$	Total
$(-\infty, -2)$	$19(= O_1)$	$27(= O_2)$	$44(= O_3)$	90
Exp. Freq.	$18.36(= E_1)$	$29.70(= E_2)$	$41.94(= E_3)$	
$[-2, 0]$	$30(= O_4)$	$53(= O_5)$	$85(= O_6)$	168
Exp. Freq.	$34.27(= E_4)$	$55.44(= E_5)$	$78.29(= E_6)$	
$(0, 2]$	$20(= O_7)$	$50(= O_8)$	$62(= O_9)$	132
Exp. Freq.	$26.93(= E_7)$	$43.56(= E_8)$	$61.51(= E_9)$	
$(2, 4]$	$20(= O_{10})$	$26(= O_{11})$	$37(= O_{12})$	83
Exp. Freq.	$16.93(= E_{10})$	$27.39(= E_{11})$	$38.68(= E_{12})$	
$(4, \infty)$	$13(= O_{13})$	$9(= O_{14})$	$5(= O_{15})$	27
Exp. Freq.	$5.51(= E_{13})$	$8.91(= E_{14})$	$12.58(= E_{15})$	
Total	102	165	233	**500**

Next, we test whether the categories based on the x_1, x_2 data are independent at 5% level, and the test statistic from (1.2) is given by

$$\chi^2_{calc} = \Sigma_{j=1}^{15} \frac{(O_j - E_j)^2}{E_j} = 19.782 \text{ with 8 degrees of freedom; } P\text{-}value = 0.011193$$
$$\Rightarrow \text{We reject the null hypotheses } H_0 \text{ at 5% level.}$$

In this example, we should have expected to see a "small" *P-value* which we did. Thus, we reject independence between x_1, x_2 values at 5% level.

Table 6. Frequency and Expected Frequency of (y_1, y_2)
In 500 Random Samples

y_1	$(-\infty, -1)$	$[-1, 0]$	$(0, 1]$	$(1, \infty]$	Total
		y_2			
$(-\infty, -1)$	$16(= O_1)$	$18(= O_2)$	$23(= O_3)$	$114(= O_4)$	171
Exp. Freq.	$60.53(= E_1)$	$29.75(= E_2)$	$23.26(= E_3)$	$57.46(= E_4)$	
$[-1, 0]$	$18(= O_5)$	$20(= O_6)$	$24(= O_7)$	$32(= O_8)$	94
Exp. Freq.	$33.28(= E_5)$	$16.36(= E_6)$	$12.78(= E_7)$	$31.58(= E_8)$	
$(0, 1]$	$24(= O_9)$	$19(= O_{10})$	$10(= O_{11})$	$11(= O_{12})$	64
Exp. Freq.	$22.66(= E_9)$	$11.14(= E_{10})$	$8.70(= E_{11})$	$21.50(= E_{12})$	
$(1, \infty]$	$119(= O_{13})$	$30(= O_{14})$	$11(= O_{15})$	$11(= O_{16})$	171
Exp. Freq.	$60.53(= E_{13})$	$29.75(= E_{14})$	$23.26(= E_{15})$	$57.46(= E_{16})$	
Total	177	87	68	168	**500**

Also, we carry out similar analysis with the y_1, y_2 values by forming a 4×4 table (Table 6) of count data of how many pairs (y_{1i}, y_{2i}) fell in each cell and used the Chi-square test (1.2) to check whether the categories based on the y_1, y_2 data were independent at 5% level. The test statistic from (1.2) is given by

$$\chi^2_{calc} = \Sigma_{j=1}^{16} \frac{(O_j - E_j)^2}{E_j} = 222.175 \text{ with 9 degrees of freedom; } P\text{-}value \approx 0$$
$$\Rightarrow \text{We reject the null hypotheses } H_0 \text{ at 5% level.}$$

Here, we may have expected to see a "small" *P-value* which we do. Thus, we reject independence between y_1, y_2 values at 5% level.

We mention that we found $r^P_{x_1, x_2} = -0.123$ and $r^P_{y_1, y_2} = -0.728$ with the associated *P-value* $= 0.006$ and *P-value* ≈ 0 respectively. So, the *t*-test based on $r^P_{x_1, x_2}$ and $r^P_{y_1, y_2}$ respectively sided with the conclusions that the x_1, x_2 data and y_1, y_2 data were dependent at 5% level.

With regard to the nonparametric test, we observed $r^S_{x_1, x_2} = -0.082$ and $r^S_{y_1, y_2} = -0.712$ along with the test statistics $z_{calc} = \sqrt{k-1} r^S_{x_1, x_2} \approx -1.8317$ and $\sqrt{k-1} r^S_{y_1, y_2} \approx -15.905$ respectively. The associated *P-values* were 0 .066996 and nearly zero respectively for the two datasets, indicating that the test would not (would) reject independence between x_1, x_2 values $(y_1, y_2$ values) at 5% level. That is, the test using the Spearman-rank correlation coef-

ficient between the y_1, y_2 data led to correct inference, but an analogous test gave an incorrect decision for the x_1, x_2 data.

17.6   A Bivariate Non-Normal Population: Case III

We repeat the notation from Section 4 and give an example of a two-dimensional random variable X with $\mathbf{X}' = (X_1, X_2)$ where X_1, X_2 are identically distributed with a common non-normal distribution, $X_1 + X_2$ is normally distributed, $X_1 - X_2$ is not normally distributed, but $\overline{X}, S^2$ are *independent* random variables.

Recall the function $g(w_1, w_2; \theta_1, \theta_2, \sigma_1, \sigma_2, \rho)$ from (4.1). Suppose that X has its pdf given by

$$f(x_1, x_2; \alpha) = \alpha g(x_1, x_2; -5, 5, 1, 1, .5) + (1 - \alpha)g(x_1, x_2; 5, -5, 1, 1, .5)$$
$$(6.1)$$

for $-\infty < x_1, x_2 < \infty, 0 < \alpha < 1$.

THEOREM 6.1 *Suppose that* (X_1, X_2) *has the joint pdf from (6.1). Let us denote* $Y_1 = X_1 + X_2$ *and* $Y_2 = X_1 - X_2$. *Then, for all* $0 < \alpha < 1$, *we have the following:*

  (i)   *Both* X_1, X_2 *have mixture normal probability models governed by the pdf from (6.2), and they are dependent;*
  (ii)   *The joint probability   model of* Y_1, Y_2 *is governed by the pdf from (6.3), but* $\overline{X}$ *has* $N(0, \frac{3}{4})$ *distribution,   and* Y_2 *has a mixture normal probability model with its pdf from (6.5);*
  (iii) Y_1, Y_2 *are independent, and so are* $\overline{X}, S^2$.

PROOF (i) From the joint pdf $f(x_1, x_2; \alpha)$, by integrating x_1 or x_2 out, one easily verifies that X_1 and X_2 respectively have the marginal pdf's

$$f_1(x_1; \alpha) = \alpha \frac{1}{\sqrt{2\pi}} e^{-(x_1+5)^2/2} + (1-\alpha)\frac{1}{\sqrt{2\pi}}e^{-(x_1-5)^2/2}$$
$$\text{for } -\infty < x_1 < \infty,$$
$$(6.2)$$
$$f_2(x_2; \alpha) = \alpha \frac{1}{\sqrt{2\pi}} e^{-(x_2-5)^2/2} + (1-\alpha)\frac{1}{\sqrt{2\pi}}e^{-(x_2+5)^2/2}$$
$$\text{for } -\infty < x_2 < \infty,$$

whatever be $0 < \alpha < 1$. It is clear that both $f_1(x_1; \alpha), f_2(x_2; \alpha)$ happen to be mixtures of $N(-5,1)$ and $N(5,1)$ probability models.

Next, observe that $f(0, 0; \alpha) = \frac{1}{\pi\sqrt{3}}e^{-50}$ whereas $f_1(0; \alpha)f_2(0; \alpha) = \frac{1}{2\pi}e^{-25}$ so that we have $f(0, 0; \alpha) \neq f_1(0; \alpha)f_2(0; \alpha)$. Hence, the random variables X_1, X_2 have *dependent* probability models.

  (ii) We have $Y_1 = X_1 + X_2, Y_2 = X_1 - X_2$. Then, along the line of derivation for Theorem 4.1 part (ii), we can again use transformation techniques to

express the joint pdf of Y_1, Y_2 as follows:

$$h(y_1, y_2; \alpha) = \alpha g\left(y_1, y_2; 0, -10, \sqrt{3}, 1, 0\right)$$
$$+ (1 - \alpha) g\left(y_1, y_2; 0, 10, \sqrt{3}, 1, 0\right) \tag{6.3}$$

for $-\infty < y_1, y_2 < \infty$. From (6.3), it is obvious that marginally, Y_1 is distributed as $N(0,3)$ with its pdf

$$h_1(y_1) = \frac{1}{\sqrt{3}\sqrt{2\pi}} e^{-y_1^2/6} \text{ for } -\infty < y_1 < \infty, \tag{6.4}$$

whatever be $0 < \alpha < 1$. However, the marginal pdf of Y_2 is given by

$$h_2(y_2; \alpha) = \alpha \frac{1}{\sqrt{2\pi}} e^{-(y_2+10)^2/2} + (1 - \alpha) \frac{1}{\sqrt{2\pi}} e^{-(y_2-10)^2/2}$$
$$\text{for } -\infty < y_2 < \infty, \tag{6.5}$$

which happens to be a mixture of $N(-10,1)$ and $N(10,1)$ probability models. Obviously, $\overline{X} = \frac{1}{2}Y_1$ is distributed as $N(0, \frac{3}{4})$.

(iii) From (6.3) it is clear that for all $0 < \alpha < 1$, the random variables Y_1, Y_2 are *independent*, that is $\overline{X}, S^2$ also have *independent* probability models since $\overline{X} = \frac{1}{2}Y_1, S^2 = \frac{1}{2}Y_2^2$. ∎

REMARK 6.1 It is clear that both X_1, X_2 have identical mixture normal and bi-modal probability models governed by the pdf

$$p(x; \alpha) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-(x+5)^2/2} + \frac{1}{\sqrt{2\pi}} e^{-(x-5)^2/2}\right] \text{ for } -\infty < x < \infty, \tag{6.6}$$

when $\alpha = \frac{1}{2}$.

REMARK 6.2 We note that $E[X_1] = 5 - 10\alpha, E[X_2] = -5 + 10\alpha$, $V[X_1] = V[X_2] = 1 + 100\alpha - 100\alpha^2$. We can also express

$$E[X_1 X_2] = \alpha\{(1)(1)\tfrac{1}{2} + (-25)\} + (1 - \alpha)\{(1)(1)(\tfrac{1}{2}) + (-25)\} = -\tfrac{49}{2}.$$

Thus, we have

$$Cov(X_1, X_2) = E[X_1 X_2] - E[X_1]E[X_2] = -\tfrac{49}{2} - (5 - 10\alpha)(-5 + 10\alpha)$$
$$= \tfrac{1}{2} - 100\alpha + 100\alpha^2,$$

so that $\rho_{X_1,X_2} = (\frac{1}{2} - 100\alpha + 100\alpha^2)/(1 + 100\alpha - 100\alpha^2)$. That is, the observations X_1, X_2 are *uncorrelated* if and only if $\alpha = \frac{1}{2} - \frac{7}{10\sqrt{2}} \approx 0.0050253$ or $\frac{1}{2} + \frac{7}{10\sqrt{2}} \approx 0.99497$.

DATA ILLUSTRATION 6.1 We focus on working under the pdf from (6.1) when $\alpha = \frac{1}{2}$ and compare performances of the Chi-square test, *t*-tests, and the non-parametric test in detecting dependence for (x_1, x_2) and (y_1, y_2) data. Thus,

we generated random pairs $(x_{1i}, x_{2i}), i = 1, ..., 500 (= k)$ governed by the joint probability model (6.1). Subsequently, we obtained (y_{1i}, y_{2i}) where $y_{1i} = x_{1i} + x_{2i}, y_{2i} = x_{1i} - x_{2i}, i = 1, ..., k$. The frequency histograms for x_1, x_2 and y_1, y_2 are given in Figures 11-12. The plots of x_1 vs x_2 and y_1 vs y_2 are given in Figure 13.
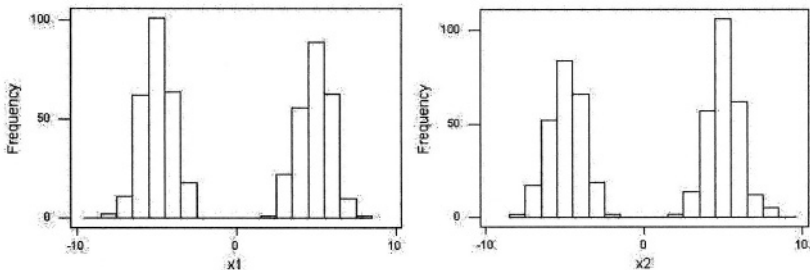
Figure 11.   Marginal frequency histograms of x_1 and x_2 obtained from observations with the joint distribution (6.1), $\alpha = \frac{1}{2}$
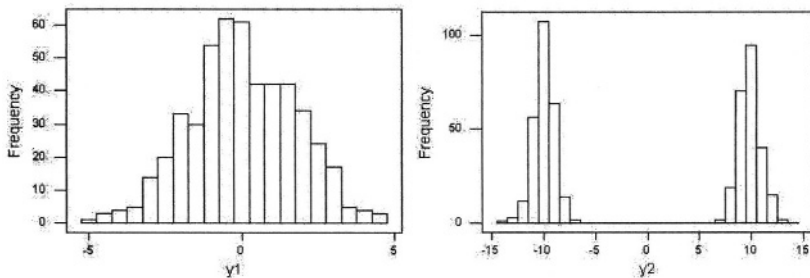
Figure 12.   Marginal frequency histograms of $y_1 = x_1 + x_2$ and $y_2 = x_1 - x_2$ obtained from (x_1, x_2) observations with the joint distribution (6.1), $\alpha = \frac{1}{2}$
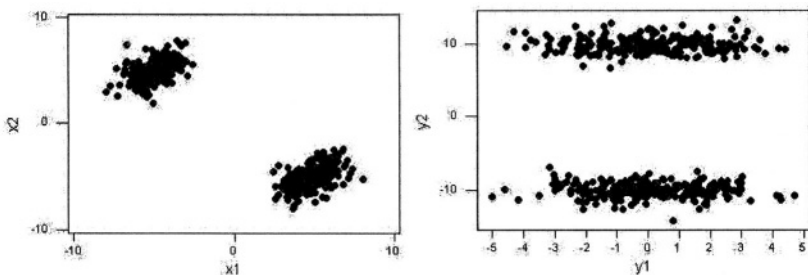
Figure 13.   Plots of x_1 vs x_2 and y_1 vs y_2 obtained from (x_1, x_2) observations with the joint distribution (6.1), $\alpha = \frac{1}{2}$

From Figure 11, we observe that the frequency histograms for both x_1, x_2 are fairly similar and these are bi-modal. Refer to Remark 6.1. We also note

from Figure 12 that the frequency histogram for y_1 resembles the shape of the $N(0,3)$ probability model (6.4), and that for y_2 resembles the shape of $h_2(y_2; \alpha)$ given by (6.5) which is clearly bi-modal. In Figure 13, the scatter plot for x_1, x_2 seem to indicate that x_1, x_2 data are *dependent* whereas the scatter plot for y_1, y_2 indicates that y_1, y_2 data are *independent*.

For a test of significance, however, we formed a 2×3 table (Table 7) of count data of how many pairs (x_{1i}, x_{2i}) fell in each cell and used the Chi-square test (1.2).

Table 7. Frequency and Expected Frequency of (x_1, x_2)
In 500 Random Samples

x_1	x_2			Total
	$(-\infty, 3)$	$[3, 5]$	$(5, \infty)$	
$(-\infty, -5)$	$6(= O_1)$	$84(= O_2)$	$38(= O_3)$	128
Exp. Freq.	$63.49(= E_1)$	$31.49(= E_2)$	$33.02(= E_3)$	
$[-5, \infty)$	$242(= O_4)$	$39(= O_5)$	$91(= O_6)$	372
Exp. Freq.	$184.51(= E_4)$	$91.51(= E_5)$	$95.98(= E_6)$	
Total	248	123	129	500

For testing whether the categories based on x_1, x_2 data are independent at 5% level, the test statistic from (1.2) is given by

$$\chi^2_{calc} = \Sigma^6_{j=1} \frac{(O_j - E_j)^2}{E_j} = 188.68 \text{ with 2 degrees of freedom; } P\text{-}value \approx 0$$
$$\Rightarrow \text{We reject the null hypotheses } H_0 \text{ at 5\% level.}$$

Since the *P-value* is "small", we reject independence between x_1, x_2 values at 5% level.

Table 8. Frequency and Expected Frequency of (y_1, y_2)
In 500 Random Samples

y_1	y_2				Total
	$(-\infty, -10)$	$[-10, -5]$	$(5, 10]$	$(10, \infty)$	
$(-\infty, -2)$	$12(= O_1)$	$18(= O_2)$	$19(= O_3)$	$16(= O_4)$	65
Exp. Freq.	$16.25(= E_1)$	$17.29(= E_2)$	$17.94(= E_3)$	$13.52(= E_4)$	
$[-2, -1]$	$23(= O_5)$	$17(= O_6)$	$18(= O_7)$	$12(= O_8)$	70
Exp. Freq.	$17.50(= E_5)$	$18.62(= E_6)$	$19.32(= E_7)$	$14.56(= E_8)$	
$(-1, 0]$	$35(= O_9)$	$32(= O_{10})$	$37(= O_{11})$	$23(= O_{12})$	127
Exp. Freq.	$31.75(= E_9)$	$33.78(= E_{10})$	$35.05(= E_{11})$	$26.42(= E_{12})$	
$(0, 1]$	$18(= O_{13})$	$26(= O_{14})$	$25(= O_{15})$	$19(= O_{16})$	88
Exp. Freq.	$22.00(= E_{13})$	$23.41(= E_{14})$	$24.29(= E_{15})$	$18.30(= E_{16})$	
$(1, 2]$	$23(= O_{17})$	$20(= O_{18})$	$18(= O_{19})$	$22(= O_{20})$	83
Exp. Freq.	$20.75(= E_{17})$	$22.08(= E_{18})$	$22.91(= E_{19})$	$17.26(= E_{20})$	
$(2, \infty)$	$14(= O_{21})$	$20(= O_{22})$	$21(= O_{23})$	$12(= O_{24})$	67
Exp. Freq.	$16.75(= E_{21})$	$17.82(= E_{22})$	$18.49(= E_{23})$	$13.94(= E_{24})$	
Total	73	162	180	85	500

Similarly, we formed a 6×4 table (Table 8) of count data of pairs (y_{1i}, y_{2i}) that fell in each cell and proceeded with the Chi-square test (1.2) for the cells. For testing whether the categories based on y_1, y_2 data are independent at 5% level, the test statistic from (1.2) is given by

$$\chi^2_{calc} = \Sigma^{24}_{j=1} \frac{(O_j - E_j)^2}{E_j} = 10.223 \text{ with 15 degrees of freedom; } P\text{-}value = 0.80548$$
$$\Rightarrow \text{ We do not reject the null hypotheses } H_0 \text{ at 5\% level.}$$

Thus, we do not reject independence between y_1, y_2 values at 5% level. In this example, we should have expected to see a "large" *P-value* which we did.

With regard to *t*-tests, we respectively found $r^P_{x_1,x_2} = -0.941$ and $r^P_{y_1,y_2} = -0.015$ with associated *P-value* ≈ 0 and *P-value* $= 0.734$. That is, the *t*-test based on $r^P_{x_1,x_2}$ and $r^P_{y_1,y_2}$ respectively sides with the conclusions that the x_1, x_2 data are dependent at 5% level, and that the y_1, y_2 data are independent at 5% level. See Remark 1.2.

Next, with regard to the nonparametric test, we found $r^S_{x_1,x_2} = -0.624$ and $r^S_{y_1,y_2} = -0.013$, along with the test statistics $z_{calc} = \sqrt{k-1} r^S_{x_1,x_2} \approx -13.939$ and $\sqrt{k-1} r^S_{y_1 y_2} \approx -0.2904$ respectively. The associated *P-value* amounts to nearly zero and 0.77151 respectively for the two datasets, indicating that we would (would not) reject the hypotheses of independence between x_1, x_2 values (y_1, y_2 values) at 5% level. That is, the nonparametric test leads to correct inferences in this example for both the x_1, x_2 and y_1, y_2 data. See Remark 1.3.

17.6.1.   Another Example

Here, we list a slightly different example. Instead of (6.1), suppose that X has its pdf given by

$$f(x_1, x_2; \alpha) = \alpha g(x_1, x_2; 5, 5, \sigma, \sigma, .5) + (1 - \alpha) g(x_1, x_2; 10, 10, \sigma, \sigma, .5)$$
(6.7)

for $-\infty < x_1, x_2 < \infty, 0 < \sigma < \infty, 0 < \alpha < 1$. Now, whatever be $0 < \sigma < \infty, 0 < \alpha < 1$, we have the following:

(i) X_1, X_2 are identically distributed with a common mixture normal distribution,
(ii) $Y_1 = X_1 + X_2$ is not normally distributed,
(iii) $Y_2 = X_1 - X_2$ is normally distributed with mean zero and variance σ^2.

Again, we can conclude that

(iv) $\overline{X} (= \frac{1}{2} Y_1)$ and $S^2 (= \frac{1}{2} Y_2^2)$ are *independent* random variables, and
(v) $2S^2$ distributed as Chi-square with one degree of freedom.

17.7   Multivariate Non-Normal Probability Models

EXAMPLE 7.1 Let us recall the *n*-dimensional normal distribution $N_n(\mu 1, \sigma^2 \Sigma)$ that was used in Section 2. Suppose that the associated pdf is denoted by

$a_n(\mathbf{x}; \mu, \sigma, \rho)$ for $\mathbf{x} \in \Re^n$. Next, let us define a n-**dimensional** random vector $\mathbf{X}$ whose pdf is given by

$$f(\mathbf{x};\alpha) = \alpha a_n(\mathbf{x}; 0, \sigma, 0) + (1 - \alpha)a_n(\mathbf{x}; 0, \tfrac{2\sigma}{\sqrt{5}}, \tfrac{1}{4(n-1)}) \text{ for } \mathbf{x} \in \Re^n,$$

(7.1)

$0 < \sigma < \infty, 0 < \alpha < 1$ with some fixed $n(\geq 2)$. Clearly, each $X_1, ..., X_n$ has a common non-normal distribution which happens to be a mixture of $N(0, \sigma^2)$ and $N(0, \tfrac{2}{5}\sigma^2)$ probability models.

Now, we may visualize the Helmert variables $Y_1, Y_2, ..., Y_n$ from (2.1) and *pretend* applying that orthogonal transformation $\mathbf{x} \to \mathbf{y}$ *separately* under the probability models $a_n(\mathbf{x}; 0, \sigma, 0)$ and $a_n(\mathbf{x}; 0, \tfrac{2\sigma}{\sqrt{5}}, \tfrac{1}{4(n-1)})$ for $\mathbf{x}$. From the summary results stated in (2.2), under the probability model $a_n(\mathbf{x}; 0, \sigma, 0)$ for $\mathbf{x}$, we conclude that $Y_1, ..., Y_n$ are iid with the common $N(0, \sigma^2)$ probability model.

Similarly, under the probability model $a_n(\mathbf{x}; 0, \tfrac{2\sigma}{\sqrt{5}}, \tfrac{1}{4(n-1)})$ for $\mathbf{x}$, we conclude that $Y_1 \sim N(0, \sigma^2)$ and $Y_2, ..., Y_n$ are iid with the common $N(0, \tfrac{4n-5}{5n-5}\sigma^2)$ distribution.

Hence, whatever be $0 < \sigma < \infty, 0 < \alpha < 1$, once we implement the transformation $\mathbf{x} \to \mathbf{y}$ under the probability model $f(\mathbf{x};\alpha)$ for $\mathbf{x}$ from (7.1), we can immediately claim that

(i) $Y_1 \sim N(0, \sigma^2)$ so that $\overline{X}(= \tfrac{1}{\sqrt{n}}Y_1) \sim N(0, \tfrac{1}{n}\sigma^2)$,

(ii) $Y_2, ..., Y_n$ are iid with the common pdf which is a mixture of $N(0, \sigma^2)$ and $N(0, \tfrac{4n-5}{5n-5}\sigma^2)$ probability models,

(iii) Y_1 is distributed *independently* of the random vector $(Y_2, ..., Y_n)$.

But, refennnr to (2.3) and recall that we can express $S^2 = \tfrac{1}{(n-1)}\Sigma_{i=2}^n Y_i^2$ which clearly implies that

(iv) $\overline{X}, S^2$ have *independent* probability models.

EXAMPLE 7.2 Here is another example. Along the line of (7.1), suppose that we have a slightly different n-**dimensional** random vector $\mathbf{X}$ whose pdf given by

$$f(\mathbf{x};\alpha) = \alpha a_n(\mathbf{x}; 0, \sigma, 0) + (1 - \alpha)a_n(\mathbf{x}; 0, \sqrt{2}\sigma, \tfrac{1}{2}) \qquad (7.2)$$

for $\mathbf{x} \in \Re^n, 0 < \sigma < \infty, 0 < \alpha < 1$. Clearly, each $X_1, ..., X_n$ has a common non-normal distribution which happens to be a mixture of $N(0, \sigma^2)$ and $N(0, 2\sigma^2)$ probability models.

Again, we may visualize the Helmert variables $Y_1, Y_2, ..., Y_n$ from (2.1) and *pretend* applying that transformation $\mathbf{x} \to \mathbf{y}$ *separately* under the probability models $a_n(\mathbf{x}; 0, \sigma, 0)$ and $a_n(\mathbf{x}; 0, \sqrt{2}\sigma, \tfrac{1}{2})$ for $\mathbf{x}$. From summary results in (2.2), under the probability model $a_n(\mathbf{x}; 0, \sigma, 0)$ for $\mathbf{x}$, recall that $Y_1, ..., Y_n$ are iid with the common $N(0, \sigma^2)$ probability model. Also, under the probability model $a_n(\mathbf{x}; 0, \sqrt{2}\sigma, \tfrac{1}{2})$ for $\mathbf{x}$, we conclude that $Y_1 \sim N(0, (n + 1)\sigma^2)$ and $Y_2, ..., Y_n$ are iid with the common $N(0, \sigma^2)$ probability model.

Hence, once we implement the transformation $\mathbf{x} \to \mathbf{y}$ under the probability model $f(\mathbf{x};\alpha)$ for $\mathbf{x}$ from (7.2), we can immediately claim that

(*i*) Y_1 has the pdf $g(y;\alpha) = \alpha \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2/(2\sigma^2)} + (1-\alpha) \frac{1}{\sigma\sqrt{2(n+1)\pi}} e^{-y^2/(2(n+1)\sigma^2)}$

for $-\infty < y < \infty$, which is a mixture of $N(0,\sigma^2)$ and $N(0,(n+1)\sigma^2)$ probability models, and

(*ii*) $Y_2, ..., Y_n$ are iid $N(0,\sigma^2)$.

Then, obviously we also have

(*iii*) $\overline{X}(= Y_1/\sqrt{n})$ has a mixture normal probability model,

(*iv*) $(n-1)S^2/\sigma^2$ has the χ_{n-1}^2 distribution,

(*v*) Y_1 is distributed independently of the random vector $(Y_2, ..., Y_n)$ so that $\overline{X}, S^2$ have *independent* probability models.

17.8   Concluding Thoughts

By allowing the observations $X_1, ..., X_n$ to be non-iid or non-normal, we have provided a number of specific examples where different scenarios developed with regard to dependence or independence between the sample mean $\overline{X}$ and the sample variance S^2. In these examples, we assigned fixed values for some of the "parameters" primarily because they made the analyses simpler and yet they drove the point home. In Sections 4-7, we could clearly envision population models $f(x_1, x_2)$ (or $f(\mathbf{x})$) defined as mixtures of three or more appropriate bivariate (or multivariate) normal probability models instead of focusing only on mixtures of simply two bivariate (or multivariate) normal probability models time after time. But, we must admit that we have deliberately stayed away from "generalizing" the examples too much because such additional frills, in our opinion, will harm both beauty and simplicity of the message.

Five major illustrations through simulated data have been provided where we applied the customary *t*-test based on Pearson-sample correlation coefficient as well as the traditional nonparametric test based on Spearman-rank correlation coefficient and the Chi-square test to "validate" independence or dependence between the two variables under consideration. We have included the *t*-test because practitioners often rely upon some routine statistical packages to come up with Pearson-sample correlation coefficient and the associated *t*-test with the intent to check "dependence" or "association" for paired data. A succinct summary of our findings follows.

Data Illustration 1.1: $\overline{X}, S^2$ were independent. The Chi-square and *t*-tests did not side against the correct conclusion that the $\overline{x}, s^2$ data were independent. The nonparametric test came up with a wrong conclusion.

Data Illustration 1.2: $\overline{X}, S^2$ were dependent. The Chi-square test, *t*-test,* and nonparametric test sided with the correct conclusion that the $\overline{x}, s$ data were dependent.

Data Illustration 4.1: X_1, X_2 were dependent with $\rho_{X_1, X_2} = 0$. The Chi-square and *t*-tests sided with the correct conclusion that the x_1, x_2 data were dependent. The nonparametric test came up with a wrong conclusion.

Y_1, Y_2 were dependent with $\rho_{Y_1, Y_2} = 0$. The Chi-square test sided with the correct conclusion that the y_1, y_2 data were dependent. Both *t*-test and nonparametric test came up with wrong conclusions.

Data Illustration 5.1: X_1, X_2 were dependent with $\rho_{X_1, X_2} = -0.098058$. The Chi-square and *t*-tests sided with the correct conclusion that the x_1, x_2 data were dependent. The nonparametric test came up with a wrong conclusion.

Y_1, Y_2 were dependent with $\rho_{Y_1, Y_2} = -0.73497$. The Chi-square test, *t*-test, and nonparametric test sided with the correct conclusion that the y_1, y_2 data were dependent.

Data Illustration 6.1: X_1, X_2 were dependent with $\rho_{X_1, X_2} = -0.94231$. The Chi-square test, *t*-test, and nonparametric test sided with the correct conclusion that the y_1, y_2 data were dependent.

Y_1, Y_2 were independent. The Chi-square test, *t*-test, and nonparametric test did not side against the correct conclusion that the y_1, y_2 data were independent.

From this summary, it is clear that in some instances the *t*-test and the nonparametric test behaved erratically in their "validation" of independence or dependence in question. In a number of occasions, the *t*-test and the nonparametric test unfortunately arrived at conflicting conclusions based on same data. When we had ρ_{X_1, X_2} or ρ_{Y_1, Y_2} significantly away from zero, we noted correct decisions regardless of which test was used for the x_1, x_2 and y_1, y_2 data. On the other hand, whenever we found that ρ_{X_1, X_2} or ρ_{Y_1, Y_2} was zero or nearly zero, we noted that these tests using the x_1, x_2 and y_1, y_2 data gave mixed signals. We realize that if the paired data were independent, then ρ would be zero, whereas even if the paired data were dependent, again ρ might be zero or nearly zero. Given this, the present investigation raises the potential of a major problem in implementing either a *t*-test or the nonparametric test as EDA tools to examine dependence or association for paired data in practice!

The Chi-square test, however, correctly validated dependence under consideration in every case, and the same test never sided against the correct conclusion that the paired data were independent when the paired variables were in fact independent. This exercise suggests that among three contenders, the Chi-square test is certainly more reliable.

Acknowledgments

References

H. Cramér, *Mathematical Methods of Statistics,* Princeton Univ. Press: Princeton, 1946.

J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference,* second edition, Marcel Dekker: New York, 1992.

A. Kagan, Yu. V. Linnik, and C. R. Rao, *Characterization Problems of Mathematical Statistics,* John Wiley & Sons: New York, 1973.

E. L. Lehmann, *Testing Statistical Hypotheses,* second edition, Springer-Verlag: New York, 1986.

E. Lukacs, *Characteristic Functions,* Charles Griffin: London, 1960.

N. Mukhopadhyay, *Probability and Statistical Inference,* Marcel Dekker: New York, 2000.

G. E. Noether, *Introduction to Statistics: The Nonparametric Way,* Springer-Verlag: New York, 1991.

B. R. Ramachandran, *Advanced Theory of Characteristic Functions,* Statistical Publishing Society: Calcutta, 1967.

C. R. Rao, *Linear Statistical Inference and Its Applications,* second edition, John Wiley & Sons: New York, 1973.

A. A. Zinger, "The independence of quasi-polynomial statistics and analytical properties of distributions," *Theory Probab. Appl.* vol. 3 pp. 247-265, 1958.

OPTIMAL STOPPING PROBLEMS FOR TIME-HOMOGENEOUS DIFFUSIONS: A REVIEW

Jesper Lund Pedersen

RiskLab, Department of Mathematics, ETH-Zentrum
CH-8092 Zürich, Switzerland

pedersen@math.ethz.ch

Abstract   The first part of this paper summarizes the essential facts on general optimal stopping theory for time-homogeneous diffusion processes in $\mathbb{R}^n$. The results displayed are stated in a little greater generality, but in such a way that they are neither too restrictive nor too complicated. The second part presents equations for the value function and the optimal stopping boundary as a free-boundary (Stefan) problem and further presents the principle of smooth fit. This part is illustrated by examples where the focus is on optimal stopping problems for the maximum process associated with a one-dimensional diffusion.

18.1   Introduction

This paper reviews some methodologies used in optimal stopping problems for diffusion processes in $\mathbb{R}^n$. The first aim is to give a quick review of the general optimal stopping theory by introducing the fundamental concepts of excessive and superharmonic functions. The second aim is to introduce the common technique to transform the optimal stopping into a free-boundary (Stefan) problem, such that explicit or numerical computations of the value function and the optimal stopping boundary are possible in specific problems.

Problems of optimal stopping have a long history in probability theory and have been widely studied by many authors. Results on optimal stopping were first developed in the discrete case. The first formulations of optimal stopping problems for discrete time stochastic processes were in connection with sequential analysis in mathematical statistics, where the number of observations is not fixed in advance (that is a random number) but terminated by the behaviour of the observed data. The results can be found in [Wald, 1947]. [Snell, 1952] obtained the first general results of optimal stopping theory for stochas-

tic processes in discrete time. For a survey of optimal stopping for Markov sequences see [Shiryaev, 1978] and the references therein. The first general results on optimal stopping problems for continuous time Markov processes were obtained by [Dynkin, 1963] using the fundamental concepts of excessive and superharmonic functions. There is an abundance of work in general optimal stopping theory using these concepts, but one of the standard and master reference is the monograph of [Shiryaev, 1978] where the definite results of general optimal stopping theory are stated and it also contains an extensive list of references to this topic. (Another thorough exposition is founded in [Karoui, 1981]). This method gives results on the existence and uniqueness of an optimal stopping time, under very general conditions, of the gain function and the Markov process. Generally, for solving a specific problem the method is very difficult to apply. In a concrete problem with a smooth gain function and a continuous Markov process, it is a common technique to formulate the optimal stopping problem as a free-boundary problem for the value function and the optimal stopping boundary along with the non-trivial boundary condition the principle of smooth fit (also called smooth pasting ([Shiryaev, 1978]) or high contact principle ([Øksendal, 1998])). The principle of smooth fit says that the first derivatives of the value function and the gain function agree at the optimal stopping boundary (the boundary of the domain of continued observation). The principle was first applied by [Mikhalevich, 1958] (under leadership of Kolmogorov) for concrete problems in sequential analysis and later independently by [Chernoff, 1961] and [Lindley, 1961]. [McKean, 1965] applied the principle to the American option problem. Other important papers in this respect are [Grigelionis & Shiryaev, 1966] and [van Moerbeke, 1974]. For a complete account of the subject and an extensive bibliography see [Shiryaev, 1978]. [Peskir, 2000] introduced the principle of continuous fit solving sequential testing problems for Poisson processes (processes with jumps).

The background for solving concrete optimal stopping problems is the following. Before and in the seventies the investigated concrete optimal stopping problems were for one-dimensional diffusions where the gain process contained two terms: a function of the time and the process, and a path-dependent integral of the process (see, among others, [Taylor, 1968], [Shepp, 1969] and [Davis, 1976]). In the nineties the maximum process (path-dependent functional) associated with a one-dimensional diffusion was studied in optimal stopping. [Jacka, 1991] treated the case of reflected Brownian motion and later [Dubins et al, 1993] treated the case of Bessel processes. In both papers the motivation was to obtain sharp maximal inequalities and the problem was solved by guessing the nature of the optimal stopping boundary. [Graversen & Peskir, 1998] formulated the maximality principle for the optimal stopping boundary in the context of geometric Brownian motion. [Peskir, 1998] showed that the maximality principle is equivalent to the superharmonic characteriza-

tion of the value function from the general optimal stopping theory and led to the solution of the problem for a general diffusion ([Peskir, 1998] also contains many references to this subject). In recent work, Graversen, [Graversen, Peskir & Shiryaev, 2001] formulated and solved an optimal stopping problem where the gain process was not adapted to the filtration.

Optimal stopping problems appear in many connections and have a wide range of applications from theoretical to applied problems. The following applications illustrate this point.

Mathematical finance

The valuation of American options is based on solving optimal stopping problems and is prominent in the modern optimal stopping theory. The literature devoted to pricing American options is extensive; for an account of the subject see the survey of Myneni [Myneni, 1992] and the references therein. The most famous result in this direction is that of McKean [McKean, 1965] solving the standard American option in the Black-Scholes model. This example can further serve to determine the right time to sell the stocks ([Øksendal, 1998]). In [Shepp & Shiryaev, 1993] the valuation of the Russian option is computed in the Black-Scholes model (see Example 7). The payoff of the option is the maximum value of the asset between the purchase time and the exercise time.

Optimal prediction

The development of optimal prediction of an anticipated functional of a continuous time process was recently initiated in [Graversen, Peskir & Shiryaev, 2001] (see Example 8). The general optimal stopping theory cannot be applied in this case since, due to the anticipated variable, the gain process is not adapted to the filtration. The problem under consideration in [Graversen, Peskir & Shiryaev, 2001] is to stop a Brownian path as close as possible to the unknown ultimate maximum height of the path. The closeness is measured by a mean-square distance. This problem was extended in [Pedersen, 2003] to cases where the closeness is measured by a L^q distance and a probability distance. These problems can be viewed as an optimal decision that needs to be based on a prediction of the future behaviour of the observable motion. For example, when a trader is faced with a decision on anticipated market movements without knowing the exact date of the optimal occurrence. The argument can be carried over to other applied problems where such a prediction plays a role.

Sharp inequalities

Optimal stopping problems are a natural tool to derive sharp versions of known inequalities, as well as to deduce new sharp inequalities. By this method

Davis [Davis, 1976] derived sharp inequalities for a reflected Brownian motion. [Jacka, 1991] and [Dubins et al, 1993] derived sharp maximal inequalities for a reflected Brownian motion and for Bessel processes, respectively. In the same direction see [Graversen & Peskir, 1997] and [Graversen & Peskir, 1998a] (Doob's inequality for Brownian motion and Hardy-Littlewood inequality, respectively) and [Pedersen, 2000] (Doob's inequality for Bessel processes).

Mathematical statistics

The Bayesian approach to sequential analysis of problems on testing two statistical hypotheses can be solved by reducing the initial problems to optimal stopping problems. Testing two hypotheses about the mean value of a Wiener process with drift was solved by [Mikhalevich, 1958] and [Shiryaev, 1969]. Peskir & Shiryaev [Peskir, 2000] solved the problem of testing two hypotheses on the intensity of a Poisson process. Another problem in this direction is the quickest detection problem (disruption problem). Shiryaev [Shiryaev, 1961] investigated the problem of detecting (alarm) a change in the mean value of a Brownian motion with drift with a minimal error (false alarm). Again, a thorough exposition of the subject can be found in [Shiryaev, 1978].

The remainder of this paper is structured as follows. The next section introduces the formulation of the optimal stopping problem under consideration. The concepts of excessive and superharmonic functions with some basic results can be found in Section 18.3. The main theorem on optimal stopping of diffusions is the point of discussion in Section 18.4. In Section 18.5, the optimal stopping problem is transformed into a free-boundary problem and the principle of smooth fit is introduced. The paper concludes with some examples in Section 18.6, where the focus is on optimal stopping problems for the maximum process associated with a diffusion.

18.2   Formulation of the problem

Let $(X_t)_{t \geq 0}$ be a time-homogeneous diffusion process with state space $\mathbb{R}^n$ associated with the infinitesimal generator

$$\mathbf{L_X} = \sum_{i=1}^{n} \mu_i(x) \, \frac{\partial}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{n} (\sigma\sigma^t)_{i,j}(x) \, \frac{\partial^2}{\partial x_i \, \partial x_j}$$

for $x \in \mathbb{R}^n$ where $\mu : \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ are continuous and further $\sigma\sigma^t$ is non-negative definite. See [Øksendal, 1998] for conditions on $\mu(\cdot)$ and $\sigma(\cdot)$ that ensure existence and uniqueness of the diffusion process. Let (Z_t) be a diffusion process depending on both time and space (and hence is not time-homogeneous diffusion) given by $(Z_t) = (t, X_t)$ which under $\mathbf{P}_z$ starts at $z = (t, x)$. Thus (Z_t) is a diffusion process in $\mathbb{R}_+ \times \mathbb{R}^n$ associated

with the infinitesimal generator

$$\mathbf{L_Z} = \frac{\partial}{\partial t} + \mathbf{L_X}$$

for $z = (t, x) \in \mathbb{R}_+ \times \mathbb{R}^n$.

The optimal stopping problem to be studied in later sections is of the following kind. Let $G : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}$ be a *gain function,* which will be specified later. Consider the *optimal stopping problem* for the diffusion (Z_t) with the *value function* given by

$$V_*(z) = \sup_\tau \mathbf{E}_z\big(G(Z_\tau)\big) \tag{2.1}$$

where the supremum is taken over all stopping times τ for (Z_t) . At the elements $\omega \in \Omega$ where $\tau(\omega) = \infty$ set $G(Z_\tau)$ to be $-\infty$. There are two problems to be solved in connection with the problem (2.1). The first problem is to compute the value function V_* and the second problem is to find an optimal stopping time τ_* , that is, a stopping time for (Z_t) such that $V_*(z) = \mathbf{E}_z(G(Z_{\tau_*}))$. Note that optimal stopping times may not exist, or be unique if they do.

18.3   Excessive and superharmonic functions

This section introduces the two fundamental concepts of excessive and superharmonic functions that are the basic concepts in the next section for a characterization of the value function in (2.1). For the facts presented here and a complete account (including proofs) of this subject, consult [Shiryaev, 1978].

In the main theorem in the next section it is assumed that the gain function belongs to the following class of functions. Let $\mathcal{L}(Z)$ be the class consisting of all lower semicontinuous functions $H : \mathbb{R}_+ \times \mathbb{R}^n \to (-\infty, \infty]$ satisfying either of the following two conditions

$$\mathbf{E}_z\big(\sup_{s \geq 0} H(Z_s)\big) < \infty \tag{3.1}$$
$$\mathbf{E}_z\big(\inf_{s \geq 0} H(Z_s)\big) > -\infty \tag{3.2}$$

for all $z = (t, x)$. If the function H is bounded from below then condition (3.2) is trivial fulfilled. The following two families of functions are crucial in the sequel presentation of the general optimal stopping theory.

DEFINITION 1 (**Excessive functions**). *A function* $H \in \mathcal{L}(Z)$ *is called excessive for* (Z_t) *if*

$$\mathbf{E}_z\big(H(Z_s)\big) \leq H(z)$$

for all $s \geq 0$ *and all* $z = (t, x)$.

DEFINITION 2 (**Superharmonic functions**). *A function* $H \in \mathcal{L}(Z)$ *is called superharmonic for* (Z_t) *if*

$$\mathbf{E}_z\big(H(Z_\tau)\big) \leq H(z)$$

for all stopping times τ *for* (Z_t) *and all* $z = (t, x)$.

The basic and useful properties of excessive and superharmonic functions are stated in [Shiryaev, 1978] and [Øksendal, 1998]. It is clear from the two definitions that a superharmonic function is excessive. Moreover, in some cases, the converse also holds – which is not obvious. The result is stated in the next proposition.

PROPOSITION 1 *Let* $H \in \mathcal{L}(Z)$ *satisfy condition* (3.2). *Then H is excessive for* (Z_t) *if and only if H is superharmonic for* (Z_t).

The above definitions play a definite role in describing the structure of the value function in (2.1). The following definition is important in this direction.

DEFINITION 3 *(***The least superharmonic (excessive) majorant***). Let* $G \in \mathcal{L}(Z)$ *be finite. A superharmonic (excessive) function* H *is called a superharmonic (excessive) majorant of* G *if* $H \geq G$. *A function* $\widehat{G}$ *is called the least superharmonic (excessive) majorant of* G *if*

(i) $\widehat{G}$ *is a superharmonic (excessive) majorant of* G.

(ii) *If* H *is an arbitrary superharmonic (excessive) majorant of* G *then* $\widehat{G} \leq H$.

To complete this section, a general iterative procedure is presented for constructing the least superharmonic majorant under the condition (3.2).

PROPOSITION 2 *Let* $G \in \mathcal{L}(Z)$ *satisfy condition* (3.2) *and* $G < \infty$. *Define the operator*

$$Q_j[G](z) = G(z) \vee \mathbf{E}_z\big(G(Z_{2^{-j}})\big)$$

and set

$$G_{j,n}(z) = Q_j^n[G](z)$$

where Q_j^n *is the n'te power of the operator* Q_j. *Then the function*

$$\widehat{G}(z) = \lim_{j\to\infty} \lim_{n\to\infty} G_{j,n}(z)$$

is the least superharmonic majorant of G.

There is a simple iterative procedure for the construction of } $\widehat{G}$ when the Markov process and the gain function are "nice".

COROLLARY 1 *Let* (Z_t) *be a Feller process and let* $G \in \mathcal{L}(Z)$ *be continuous and bounded from below. Set*

$$G_j(z) = \sup_{t \geq 0} \mathbf{E}_z\big(G_{j-1}(Z_t)\big)$$

for $j \geq 1$ *and* $G_0 = G$ *. Then*

$$\widehat{G}(z) = \lim_{j \to \infty} G_j(z)$$

is the least superharmonic majorant of G *.*

REMARK 1 *Proposition 2 and Corollary 1 are both valid under condition (3.2) and excessive and superharmonic functions are the same in this case, according to Proposition 1. When condition (3.2) is violated, the least excessive majorant may differ from the least superharmonic majorant. In this case, the least excessive majorant is smaller than the least superharmonic majorant, since there are more excessive functions than superharmonic functions. The construction of the least superharmonic majorant follows a similar pattern but is generally more complicated (see [Shiryaev, 1978]).*

REMARK 2 *The iterative procedures to construct the least superharmonic majorant are difficult to apply to concrete problems. This makes it necessary to search for explicit or numerical computations of the least superharmonic majorant.*

18.4   Characterization of the value function

The main theorem of general optimal stopping theory of diffusion processes is contained in the next theorem. The result gives existence and uniqueness of an optimal stopping time in problem (2.1). The result could have been stated in a more general setting, but is stated with a minimum of technical assumptions. For instance, the theorem also holds for a larger class of Markov process such as Lévy processes. For details of this and the main theorem consult [Shiryaev, 1978].

THEOREM 1 *Consider the optimal stopping problem* (2.1) *where the gain function* G *is lower semicontinuous and satisfies either* (3.1) *or* (3.2)*.*

   (I). *The value function* V_* *is the least superharmonic majorant of the gain function* G *with respect to the process* $(Z_t)_{t \geq 0}$ *, that is,*

$$V_*(z) = \widehat{G}(z)$$

for all $z = (t, x)$ *.*
   (II). *Define the domain of continued observation*

$$C = \{ z \in \mathbb{R}_+ \times \mathbb{R}^n \,|\, G(z) < V_*(z) \}$$

and let τ_* be the first exit time of (Z_t) from C, that is,

$$\tau_* = \inf\{t > 0 : Z_t \notin C\}.$$

If $\tau_* < \infty$ $\mathbf{P}_z$-*a.s. for all* z, *then* τ_* *is an optimal stopping time for the problem* (2.1), *at least when* G *is continuous and satisfies both* (3.1) *and* (3.2).

   (III). *If there exists an optimal stopping time* σ *in problem* (2.1), *then* $\tau_* \leq \sigma$ $\mathbf{P}_z$-*a.s. for all* z *and* τ_* *is also an optimal stopping time for problem* (2.1).

REMARK 3 *Part* **(II)** *of the theorem gives the existence of an optimal stopping time. The conditions could have been stated with a little greater generality; again, for more details cf.* [Shiryaev, 1978].

   Part **(III)** *of the theorem says that if there exists an optimal stopping time* σ *then* τ_* *is also an optimal stopping time and is the smallest among all optimal stopping times for problem* (2.1). *This extremal property of the optimal stopping time* τ_* *characterizes it uniquely.*

REMARK 4 *Sometimes it is convenient to consider "approximate" optimal stopping times. An example is given in the setting of Theorem 1* **(II)**, *if the stopping time* τ_* *does not satisfy* $\tau_* < \infty$ $\mathbf{P}_z$-*a.s. Then the following approximate stopping times are available. For* $\varepsilon > 0$ *let* $C_\varepsilon = \{z \in \mathbb{R}_+ \times \mathbb{R}^n \mid G(z) < V_*(z) - \varepsilon$. *Let* τ_ε *be the first exit time of* (Z_t) *from* C_ε, *that is,* $\tau_\varepsilon = \inf\{t > 0 : Z_t \notin C_\varepsilon$. *Then* $\tau_\varepsilon < \infty$ $\mathbf{P}_z$-*a.s. and* τ_ε *is approximated optimal in the following sense* $\lim_{\varepsilon \downarrow 0} \mathbf{E}_z(G(Z_{\tau_\varepsilon})) = V_*(z)$ *for all* $z = (t, x)$. *Furthermore,* $\tau_\varepsilon \uparrow \tau_*$ *as* $\varepsilon \downarrow 0$.

   At first glance, it seems that the initial setting of the optimal stopping problem (2.1) and Theorem 1 only cover the cases where the gain process is a function of time and the state of the process (X_t). But the next two examples illustrate that Theorem 1 also covers some cases where the gain process contains path-dependent functional of (X_t), where it is a matter of properly defining (Z_t).

   For simplicity, let $n = 1$ in the examples below and assume, moreover, that (X_t) solves the stochastic differential equation

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dB_t$$

where (B_t) is a standard Brownian motion.

EXAMPLE 1 *(**Optimal stopping problems involving an integral**). **Let** F* : $\mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ *and* $c : \mathbb{R} \to \mathbb{R}_+$ *be continuous functions. Consider the optimal stopping problem*

$$W_*(t, x) = \sup_\tau \mathbf{E}_x\left(F(t + \tau, X_\tau) - \int_0^\tau c(X_u)\,du\right). \tag{4.1}$$

The integral term might be interpreted as an accumulated cost. This problem can be reformulated to fit in the setting of problem (2.1) and Theorem 1 by the following simple observations.

Set $A_t = \int_0^t c(X_u)\,du$ and denote (Z_t) by $Z_t = (t, X_t, A_t)$. Thus (Z_t) is a diffusion process in $\mathbb{R}^3$ associated with the infinitesimal generator

$$\mathbf{L_Z} = \frac{\partial}{\partial t} + \mathbf{L_X} - c(x)\frac{\partial}{\partial a}$$

for $z = (t, x, a)$. Let $G(z) = F(t, x) - a$ be a gain function and consider the new optimal stopping problem

$$V_*(z) = \sup_\tau \mathbf{E}_z\big(G(Z_\tau)\big)\,.$$

This problem fits into the setting of Theorem 1 and it is clear that $W_*(t, x) = V_*(t, x, 0)$. Note that the gain function G is linear in a

Another approach is by Itô formula to reduce the problem (4.1) to the setting of the initial problem (2.1). Assume that the function $x \mapsto D(x)$ is smooth and satisfies $\mathbf{L_X}D(x) = c(x)$. Itô formula yields that

$$D(X_t) = D(x) + \int_0^t \mathbf{L_X}D(X_u)\,du + M_t$$

where $M_t = \int_0^t D'(X_u)\,\sigma(X_u)\,dB_u$ is a continuous local martingale. The optional sampling implies that $\mathbf{E}_x(M_\tau) = 0$ (by localization and some uniform integrable conditions) and hence

$$\mathbf{E}_x\big(D(X_\tau)\big) = D(x) + \mathbf{E}_x\left(\int_0^\tau c(X_u)\,du\right)\,.$$

Therefore, the problem (4.1) is equivalent to solving the initial problem (2.1) with the gain function $\widetilde{G}(t, x) = F(t, x) - D(x)$.

EXAMPLE 2 *(**Optimal stopping problems for the maximum process**). Peskir [Peskir, 1998] made the following observation. Denote the maximum process associated with (X_t) by $S_t = \max_{0 \leq u \leq t} X_u$. It can be verified that the two-dimensional process $(Z_t) = (X_t, S_t)$ with state space $\{(x, s) \in \mathbb{R}^2 \,|\, x \leq s\}$ (see Figure 1) is a continuous Markov process associated with the infinitesimal generator*

$$\mathbf{L_Z} = \mathbf{L_X} \ \textit{for } x < s$$
$$\frac{\partial}{\partial s} = 0 \quad \textit{for } x = s$$
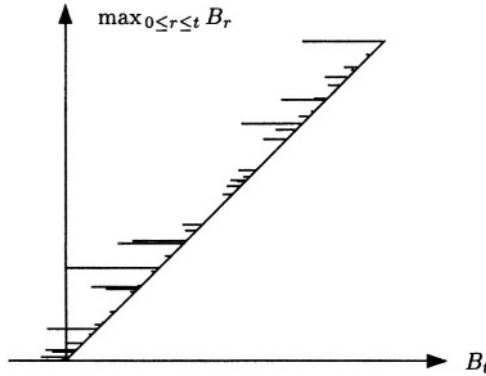
Figure 1.   A   simulation   of   a   path   of   the   two-dimensional   process
$(B_t(\omega), \max_{0 \le u \le t} B_u(\omega))$ where (B_t) is a Brownian motion.

with **L**$_\mathbf{X}$ *given in Section 18.2. Hence the optimal stopping problem*

$$V_*(x, s) = \sup_\tau \mathbf{E}_{x,s}\big(G(X_\tau, S_\tau)\big)$$

for $x \le s$ *fits in the setting of Theorem 1.*

18.5   The free-boundary problem and the principle of smooth fit

For solving a specific optimal stopping problem the superharmonic charac-
terization is not easy to apply. To carry out explicit computations of the value
function another methodology therefore is needed. This section considers the
optimal stopping problem as a free-boundary (Stefan) problem. This is also
important for computations of the value function from a numerical point of
view. First, the notation of characteristic generator (see [Øksendal, 1998])
is introduced and is an extension of the infinitesimal generator. Let (Z_t) be
the diffusion process given in Section 18.2. For any open set $U \subseteq \mathbb{R}_+ \times \mathbb{R}^n$,
associate $\tau_U = \inf\{t > 0 : Z_t \notin U\}$ to be the first exit time from U of
(Z_t).

DEFINITION 4 (**Characteristic generator**). *The characteristic generator* $\mathcal{A}_{\mathbf{Z}}$
of (Z_t) *is defined by*

$$\mathcal{A}_{\mathbf{Z}} f(z) = \lim_{U \downarrow z} \frac{\mathbf{E}_z\big(f(Z_{\tau_U})\big) - f(z)}{\mathbf{E}_z(\tau_U)}$$

where the limit is to be understood in the following sense.   The open sets
U_j *decrease to the point* z, *that is,* $U_{j+1} \subseteq U_j$ *and* $\cap_{j \ge 1} U_j = \{z\}$. *If*

$\mathbf{E}_z(\tau_U) = \infty$ *for all open sets* $U \ni z$ *, then set* $\mathcal{A}_{\mathbf{Z}} f(z) = 0$ *. Let* $\mathcal{D}(\mathcal{A}_{\mathbf{Z}})$ *be the family of Borel functions* f *for which the limit exists.*

REMARK 5 *As already mentioned above the characteristic generator is an extension of the infinitesimal generator in the following sense that* $\mathcal{D}(\mathbf{L_Z}) \subseteq \mathcal{D}(\mathcal{A_Z})$ *and* $\mathbf{L_Z} f = \mathcal{A_Z} f$ *for any* $f \in \mathcal{D}(\mathbf{L_Z})$ *.*

Assume in the sequel that the value function V_* in (2.1) is finite. Let $C = \{ z \in \mathbb{R}_+ \times \mathbb{R}^n \mid V_*(z) > G(z) \}$ be the domain of continued observation (see Theorem 1). Then the following result gives equations for the value function in the domain of continued observation.

THEOREM 2 *Let the gain function G be continuous and satisfy both conditions (3.1) and (3.2). Then the value function* $V_*(z)$ *for* $z \in C$ *belongs to* $\mathcal{D}(\mathcal{A_Z})$ *and solves the equation*

$$\mathcal{A_Z} V_*(z) = 0 \tag{5.1}$$

for $z \in C$

REMARK 6 *Since the gain function G is continuous and the value function* V_* *is lower semicontinuous, the domain of continued observation* C *is an open set in* $\mathbb{R}_+ \times \mathbb{R}^n$ *. If* $\tau_C < \infty$ $\mathbf{P}_z$*-a.s. then it follows from Theorem 1 that*

$$V_*(z) = \mathbf{E}_z\big(G(Z_{\tau_C})\big) .$$

Then the general Markov process theory yields that the value function solves the equation (5.1) and Theorem 2 follows directly. In other words, one is led to formulate equation (5.1).

If the value function is C^2 in the domain of continued observation, the characteristic generator can be replaced by the infinitesimal generator according to Remark 5. This has the advantage that the infinitesimal generator is explicitly given.

Equation (5.1) is referred to as a free-boundary problem. The domain of continued observation C is not known a priori but must be found along with unknown value function V_* . Usually, a free-boundary problem has many solutions and further conditions must be added (e.g. the principle of smooth fit) which the value function V_* satisfies. These additional conditions are not always enough to determine V_* . In that case, one must either guess or find more sophisticated conditions (e.g. the maximality principle, see Example 5 in the next section).

The famous principle of smooth fit is one of the most frequently used nontrivial boundary conditions in optimal stopping. The principle is often applied in the literature (see, among others, [McKean, 1965], [Jacka, 1991] and [Dubins et al, 1993]).

The principle of smooth fit

If the gain function G is smooth then a non-trivial boundary condition for the free-boundary problem for $i = 1, \ldots, n$ might be the following

$$\frac{\partial V_*}{\partial t}(z)\Big|_{z \in \partial C} = \frac{\partial G}{\partial t}(z)\Big|_{z \in \partial C}$$

$$\frac{\partial V_*}{\partial x_i}(z)\Big|_{z \in \partial C} = \frac{\partial G}{\partial x_i}(z)\Big|_{z \in \partial C}.$$

A result in [Shiryaev, 1978] states that the principle of smooth fit holds under fairly general assumptions. The principle of smooth fit is a very fine condition in the sense that the value function often is often precisely C^1 at the boundary of the domain of continued observation. This is demonstrated in the examples in the next section.

The above results can be used to formulate the following method for solving a particular stopping problem.

A recipe to solve optimal stopping problems

Step 1. First one tries to guess the nature of the optimal stopping boundary and then, by using ad hoc arguments, to formulate a free-boundary problem with the infinitesimal generator and some boundary conditions. The boundary conditions can be trivial ones (e.g. the value function is continuous, odd/even, normal reflection etc.) or non-trivial, such as the principle of smooth fit and the maximality principle.

Step 2. One solves the formulated free-boundary system and maximizes over the family of solutions if there is no unique solution.

Step 3. Finally, one must verify that the guessed at candidates for the value function and the optimal stopping time are indeed correct, (e.g., using Itô formula).

The methodology has been used in, among others, [Dubins et al, 1993], [Graversen & Peskir, 1998], [Pedersen, 2000] and [Shepp & Shiryaev, 1993].

It is generally difficult to find the appropriate solution of the (partial) differential equation $\mathbf{L_Z} V(z) = 0$. It is therefore of most interest to formulate the free-boundary problem such that the dimension of the problem is as small as possible. The two examples below present cases where the dimension can be reduced. For simplicity let $n = 1$ and assume, moreover, that (X_t) solves the stochastic differential equation

$$dX_t = \mu(X_t) \, dt + \sigma(X_t) \, dB_t$$

where (B_t) is a standard Brownian motion.

EXAMPLE 3 (**Integral and discounted problem**). *The general Markov process theory states that the free-boundary problem is one-dimensional in some special cases.*

1. *Let* $F : \mathbb{R} \to \mathbb{R}$ *and* $c : \mathbb{R} \to \mathbb{R}_+$ *be continuous functions and let the gain function be given by* $G(x,a) = F(x) - a$ *which is linear in* a *(see Example 1). Let* $(Z_t) = (X_t, A_t)$ *where* $A_t = \int_0^t c(X_u)\, du$ *and consider the two-dimensional optimal stopping problem*

$$V_*(x) = \sup_\tau \mathbf{E}_x \left(F(X_\tau) - \int_0^\tau c(X_u)\, du \right).$$

At first glance, it seems to be a two-dimensional problem, but the Markov process theory yields that the free-boundary problem can formulated as

$$\mathbf{L_X} V_*(x) = -c(x)$$

for x *in the domain of continued observation, which is also clear from the last part of Example 1. This is a one-dimensional problem.*

2. *Given the gain function* $G(t,x) = e^{-\lambda t} F(x)$ *where* $\lambda > 0$ *is a constant. Let* $(Z_t) = (t, X_t)$ *and consider the "two-dimensional" optimal stopping problem*

$$V_*(x) = \sup_\tau \mathbf{E}_x \left(e^{-\lambda \tau} F(X_\tau) \right).$$

In this case, the free-boundary problem can be formulated as

$$\mathbf{L_X} V_*(x) = \lambda V_*(x)$$

for x *in the domain of continued observation. Again, this is a one-dimensional problem.*

EXAMPLE 4 *(**Deterministic time-change method**). This example uses a deterministic time-change to reduce the problem. The method is described in [Pedersen & Peskir, 2000]. Consider the optimal stopping problem*

$$V_*(t, x) = \sup_\tau \mathbf{E}_x \left(\alpha(t + \tau)\, X_\tau \right)$$

where α *is a smooth non-linear function. Thus, the value function* V_* *might solve the following partial differential equation*

$$\frac{\partial V_*}{\partial t}(t, x) + \mathbf{L_X} V_*(t, x) = 0$$

for (t, x) *in the domain of continued observation.*

The time-change method transforms the original problem into a new optimal stopping problem, such that the new value function solves an ordinary differential equation. The problem is to find a deterministic time-change $t \mapsto \sigma_t$ which satisfies following two conditions:

(i) $t \mapsto \sigma_t$ *is continuous and strictly increasing.*

(ii) *There exists a one-dimensional time-homogeneous diffusion (Y_t) with infinitesimal generator $\mathbf{L_Y}$ such that $\alpha(\sigma_t) X_{\sigma_t} = e^{-\lambda t} Y_t$ for some $\lambda \in \mathbb{R}$.*

The condition (i) ensures that τ is a stopping time for (Y_t) if and only if σ_τ is a stopping time for (X_t). Substituting (ii) in the problem, the new (time-changed) value function becomes

$$W_*(y) = \sup_\tau \mathbf{E}_y\big(e^{-\lambda \tau} Y_\tau\big) .$$

As in Example 3 the new problem might solve the ordinary differential equation

$$\mathbf{L_Y} W_*(y) = \lambda W_*(y)$$

in the domain of continued observation. Given the diffusion (X_t) the crucial point is to find the process (Y_t) and the time-change σ_t fulfilling the two conditions above. By Itô calculus it can be shown that the time-change given by

$$\sigma_t = \inf\left\{ r > 0 \ \middle|\ \int_0^r \rho(u)\, du > t \right\}$$

where $\rho(\cdot)$ satisfies that the two terms

$$\left(\frac{\alpha'(t)}{\alpha(t)} y + \alpha(t)\, \mu\big(y/\beta(t)\big)\right)\frac{1}{\rho(t)} \ \ and \ \ \alpha(t)^2\, \sigma^2\big(y/\alpha(t)\big)\,\frac{1}{\rho(t)}$$

do not depend on t, will fulfill the above two conditions. This clearly imposes the following conditions on $\alpha(\cdot)$ to make the method applicable

$$\mu\big(y/\alpha(t)\big) = \gamma(t)\, G_1(y) \ \ and \ \ \sigma^2\big(y/\alpha(t)\big) = \frac{\gamma(t)}{\alpha(t)}\, G_2(y)$$

where $\gamma(t)$, $G_1(y)$ and $G_2(y)$ are functions required to exist. For more information and remaining details of this method see [Pedersen & Peskir, 2000] (see also [Graversen, Peskir & Shiryaev, 2001]).

18.6   Examples and applications

This section presents the solutions of three examples of stopping problems which illustrate the method established in the previous section and some applications. The focus will be on optimal stopping problems for the maximum process associated with a one-dimensional diffusion.

Let $n = 1$. Assume that (X_t) is a non-singular diffusion with state space $\mathbb{R}$, that is $x \mapsto \sigma(x) > 0$ and (X_t) solves the stochastic differential equation

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dB_t$$

where (B_t) is a standard Brownian motion. The infinitesimal generator of (X_t) is given by

$$\mathbf{L_X} = \mu(x)\,\frac{\partial}{\partial x} + \tfrac{1}{2}\sigma^2(x)\,\frac{\partial^2}{\partial x^2}\,. \tag{6.1}$$

Let $S_t = \max_{0 \le u \le t} X_u \vee s$ denote the maximum process associated with (X_t) and let it start at $s \ge x$ under $\mathbf{P}_{x,s}$. The scale function and speed measure of (X_t) are given by

$$S(x) = \int_0^x \exp\left(-2\int_0^u \frac{\mu(r)}{\sigma^2(r)}\,dr\right)du \ \text{and}\ m(dx) = \frac{2}{S'(x)\,\sigma^2(x)}\,dx$$

for $x \in \mathbb{R}$.

The first example is important from the general optimal stopping theory point of view.

EXAMPLE 5 (**The maximality principle**). *The results of this example are found in [Peskir, 1998]. Let $x \mapsto c(x) > 0$ be a continuous (cost) function. Consider the optimal stopping problem with the value function*

$$V_*(x, s) = \sup_\tau \mathbf{E}_{x,s}\left(S_\tau - \int_0^\tau c(X_u)\,du\right) \tag{6.2}$$

where the supremum is taken over all stopping times τ for (X_t) satisfying

$$\mathbf{E}_{x,s}\left(\int_0^\tau c(X_u)\,du\right) < \infty \tag{6.3}$$

for all $x \le s$. The recipe from the previous section is applied to solve the problem.

1. *The process (X_t, S_t) with state space $\{(x, s) \in \mathbb{R}^2 \,|\, x \le s\}$ changes only in the second coordinate when it hits the diagonal $x = s$ in $\mathbb{R}^2$ (see Figure 1). It can be shown that it is not optimal to stop on the diagonal. Due to the positive cost function $c(\cdot)$ the optimal stopping boundary might be a function which stays below the diagonal. Thus, the stopping time might be on*

the form $\tau_* = \inf\{t > 0 : X_t \leq g_*(S_t)\}$ for some function $s \mapsto g_*(s) < s$
to be found. In other words, the domain of continued observation is on the
form $C = \{(x, s) \in \mathbb{R}^2 \mid g_*(s) < x \leq s\}$. It is now natural to formulate
the following free-boundary problem that the value function and the optimal
stopping boundary is a solution of

$$\mathbf{L_X} V(x, s) = c(x) \qquad \text{for } g(s) < x < s \text{ with } s \text{ fixed} \qquad (6.4)$$

$$\frac{\partial V}{\partial s}(x, s)\bigg|_{x=s-} = 0 \qquad \text{(normal reflection)} \qquad\qquad (6.5)$$

$$V(x, s)\bigg|_{x=g(s)+} = s \qquad \text{(instantaneous stopping)} \qquad (6.6)$$

$$\frac{\partial V}{\partial x}(x, s)\bigg|_{x=g(s)+} = 0 \quad \text{(smooth fit)} . \qquad\qquad (6.7)$$

Note that (6.4) and (6.5) follow from Example 2 and Example 3. The condition
(6.6) is clear and since the setting is smooth the principle of smooth fit should
be satisfied, that is condition (6.7) holds. (The theorem below shows that the
guessed system is indeed correct).
  2. *Define the function*

$$V_g(x, s) = s + \int_{g(s)}^{x} \big(S(x) - S(u)\big)c(u)\, m(du) \qquad (6.8)$$

for $g(s) \leq x \leq s$ and set $V_g(x, s) = s$ for $x \leq g(s)$. Further, define the
first order non-linear differential equation

$$g'(s) = \frac{\sigma^2(g(s))\, S'(g(s))}{2c\big(g(s)\big)\big(S(s) - S(g(s))\big)} . \qquad (6.9)$$

For a solution $s \mapsto g(s) < s$ of equation (6.9) the corresponding function
$V_g(x, s)$ in (6.8) solves the free-boundary problem in the region $g(s) < x < s$.

  The problem now is to choose the right optimal stopping boundary $s \mapsto$
$g(s) < s$. To do this a new principle is needed and it will be the maximality
principle. The main observations in [Peskir, 1998] are the following.

  (i) $g \mapsto V_g(x, s)$ is increasing.

  (ii) The function $(x, s, a) \mapsto V_g(x, s) - a$ is superharmonic for the Markov
      process $(Z_t) = (X_t, S_t, A_t)$ (for stopping times τ satisfying (6.3))
      where $A_t = \int_0^t c(X_u)\, du + a$.

The superharmonic characterization of the value function in Theorem 1 and
the above two observations lead to the formulation of the following principle
for determining the optimal stopping boundary.

The maximality principle

The optimal stopping boundary $s \mapsto g_*(s)$ for the problem (6.2) is the maximal solution of the differential equation (6.9) which stays strictly below the diagonal in $\mathbb{R}^2$ (and is simply called the maximal solution in the sequel).

3. In [Peskir, 1998] it was proved that this principle is equivalent to the superharmonic characterization of the value function. The result is formulated in the next theorem and is motivated by Theorem 1.

THEOREM 3 *Consider the optimal stopping problem* (6.2).

(I). *Let* $s \mapsto g_*(s)$ *denote the maximal solution of* (6.9) *which stays below the diagonal in* $\mathbb{R}^2$. *Then the value function is given by*

$$
V_*(x, s) = \begin{cases} s + \int_{g_*(s)}^x \big(S(x) - S(u)\big) c(u)\, m(du) & \text{for } g_*(s) < x \le s \\ s & \text{for } x \le g_*(s) \,. \end{cases}
$$

(II). *The stopping time* $\tau_* = \inf\{t > 0 : X_t \le g_*(S_t)\}$ *is optimal whenever it satisfies condition* (6.3).

(III). *If there exists an optimal stopping time* σ *in* (6.2) *satisfying* (6.3), *then* $\tau_* \le \sigma$ $\mathbf{P}_{x,s}$-*a.s. for all* (x, s) , *and* τ_* *is also an optimal stopping time.*

(IV). *If there is no maximal solution of* (6.9) *which stays strictly below the diagonal in* $\mathbb{R}^2$, *then* $V_*(x, s) = \infty$ *for all* (x, s) , *and there is no optimal stopping time.*

For more information and details see [Peskir, 1998]. A similar approach was used in [Pedersen & Peskir, 1998] to compute expectation of Azéma-Yor stopping times.

The theorem extends to diffusions with other state spaces in $\mathbb{R}$. The non-negative diffusion version of the theorem is particularly interesting to derive sharp maximal inequalities, which will be applied in the next example.

Peskir [Peskir, 1998] conjectured that the maximality principle holds for the discounted version of problem (6.2). In Shepp & Shiryaev [Shepp & Shiryaev, 1993] and Pedersen [Pedersen, 2000a] the principle is shown to hold in specific cases. A technical difficulty arises in verifying the conjecture because the corresponding free-boundary problem may have no simple solution and the (optimal) boundary function is thus implicitly defined.

EXAMPLE 6 **(Doob's inequality for Brownian motion).** *This example is an application of the previous example (see also [Graversen & Peskir, 1997]). Consider the optimal stopping problem* (6.2) *with* $X_t = |B_t + x|^p$ *and* $c(x) = c\, x^{(p-2)/p}$ *for* $p > 1$. *Then* (X_t) *is a non-negative diffusion having* 0

as an instantaneously reflecting boundary point and the infinitesimal generator of (X_t) in $(0, \infty)$ is given in (6.1) with $\mu(x) = \frac{1}{2} p(p-1) x^{1-2/p}$ and $\sigma^2(x) = p^2 x^{2-2/p}$. If $c > \frac{1}{2} p^{p+1}/(p-1)^{p-1}$, it follows from Theorem 3 that the value function is given by

$$V_*(x, s) = s - \frac{2c}{p-1} g_*(s)^{1-1/p} x^{1/p} + \frac{2c}{p} g_*(s) + \frac{2c}{p(p-1)} x$$

where $s \mapsto g_*(s) < s$ is the maximal solution of the differential equation

$$g'(s) = \frac{p}{2c} \frac{g(s)^{1/p}}{s^{1/p} - g(s)^{1/p}} . \tag{6.10}$$

The maximal solution (see Figure 2) can be found to be $g_*(s) = \alpha_*^c s$ where $0 < \alpha_*^c < 1$ is the greater root of the equation (the maximality principle)

$$\alpha - \alpha^{1-1/p} + p/(2c) = 0 .$$

The equation admits two roots if and only if $c > \frac{1}{2} p^{p+1}/(p-1)^{p-1}$. Further, the stopping time

$$\tau_*(c) = \inf \{ t > 0 : X_t \leq \alpha_*^c S_t \}$$

satisfies $\mathbf{E}_{x,s}(\tau_*(c)^{p/2}) < \infty$ if and only if $c > \frac{1}{2} p^{p+1}/(p-1)^{p-1}$. By an extended version of Theorem 3 for non-negative diffusions and an observation in Example 3, it follows by the definition of the value function for $c > \frac{1}{2} p^{p+1}/(p-1)^{p-1}$ that

$$\mathbf{E}_x \big(\max_{0 \leq t \leq \tau} |B_t|^p \big) \leq \frac{c}{p(p-1)} \mathbf{E}_x \big(|B_\tau|^p \big) + V_*(x, x) - \frac{c}{p(p-1)} x^p$$

for all stopping times τ for (B_t) satisfying $\mathbf{E}(\tau^{p/2}) < \infty$. Letting $c \downarrow \frac{1}{2} p^{p+1}/(p-1)^{p-1}$, the Doob's inequality follows.

THEOREM 4 *Let (B_t) be a standard Brownian motion started at x under $\mathbf{P}_x$ for $x \geq 0$, let $p > 1$ be given and fixed, and let τ be any stopping time for (B_t) such that $\mathbf{E}_x(\tau^{p/2}) < \infty$. Then the following inequality is sharp*

$$\mathbf{E}_x \big(\max_{0 \leq t \leq \tau} |B_t|^p \big) \leq \big(\tfrac{p}{p-1} \big)^p \mathbf{E}_x \big(|B_\tau|^p \big) - \tfrac{p}{p-1} x^p .$$

The constants $(p/(p-1))^p$ and $p/(p-1)$ are the best possible and the equality is attained through the stopping times $\tau_* = \inf \{ t > 0 : X_t \leq \alpha_*^c S_t \}$ for $c \downarrow \frac{1}{2} p^{p+1}/(p-1)^{p-1}$.

For details see [Graversen & Peskir, 1997]. The results are extended to Bessel processes in [Dubins et al, 1993] and [Pedersen, 2000].
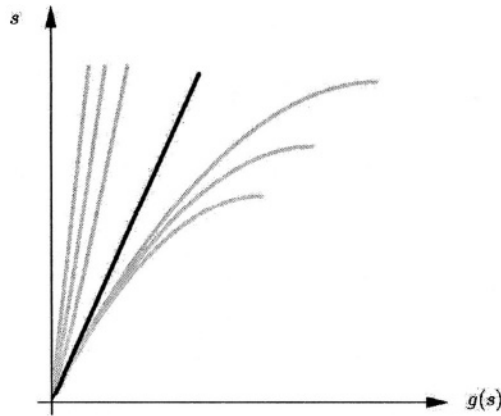
Figure 2.   A computer drawing of solutions of the differential equation (6.10). The bold line is the maximal solution which stays below and never hits the diagonal in $\mathbb{R}^2$. By the maximality principle, this solution equals g_* .

EXAMPLE 7 **(Russian option).** *This is an example of pricing an American option with infinite time horizon in the framework of the standard Black-Scholes model.   The option under consideration is the Russian option (see [Shepp & Shiryaev, 1993]). If (X_t) is the price process of a stock then the payment function of the Russian option is given by*

$$f_t = \max_{0 \le u \le t} X_u$$

where the expiration time is infinity.   Thus, it is a perpetual Lookback option (see [Conze & Viswanathan, 1991]). Assume a standard Black-Scholes model with a dividend paying stock; under the equivalent martingale measure the price process is thus the geometric Brownian motion

$$dX_t = (r - \lambda)X_t\, dt + \sigma X_t\, dB_t$$

with $\lambda > 0$ the dividend yield, $r > 0$ the interest rate and $\sigma > 0$ the volatility. The infinitesimal generator of (X_t) on $(0, \infty)$ is given in (6.1) with $\mu(x) = (r - \lambda)x$ and $\sigma(x) = \sigma x$.

    Under these assumptions, the fair price of the Russian option is – according to the general pricing theory – is the value of the optimal stopping problem

$$C_*(x, s) = \sup_{\tau} \mathbf{E}_{x,s}\left(e^{-r\tau}\, f_\tau\right) = \sup_{\tau} \mathbf{E}_{x,s}\left(e^{-r\tau}\, S_\tau\right)$$

where the supremum is taken over all stopping times τ for (X_t) . To solve this problem, the idea is to apply Example 3 and the maximality principle for

this discounted optimal stopping problem. The recipe from the previous section is applied to solve the problem.

1. *As in Example 5, and using an observation in Example 3, it is natural to formulate the following free-boundary problem that the value function and the optimal stopping boundary is a solution of*

$$\mathbf{L_X} C(x, s) = r\, C(x, s) \ \text{ for } g(s) < x < s$$

$$\frac{\partial C}{\partial s}(x, s)\Big|_{x=s-} = 0 \qquad \textit{(normal reflection)}$$

$$C(x, s)\Big|_{x=g(s)+} = s \qquad \textit{(instantaneous stopping)}$$

$$\frac{\partial C}{\partial x}(x, s)\Big|_{x=g(s)+} = 0 \quad \textit{(smooth fit).}$$

Since the setting is smooth, the principle of smooth fit should be satisfied. The theorem below shows that this system is indeed correct.

2. *Let* $\gamma_1 < 0$ *and* $\gamma_2 > 1$ *be the two roots of the quadratic equation*

$$\tfrac{1}{2}\sigma^2\gamma^2 + (r - \lambda - \tfrac{1}{2}\sigma^2)\gamma - r = 0$$

and set

$$\beta_* = \left(\frac{1 - 1/\gamma_2}{1 - 1/\gamma_1}\right)^{1/(\gamma_2-\gamma_1)} < 1\,.$$

The solutions to the free-boundary problem are

$$C(x, s) = \frac{s}{\gamma_2 - \gamma_1}\left[\gamma_2\left(\frac{x}{g(s)}\right)^{\gamma_1} - \gamma_1\left(\frac{x}{g(s)}\right)^{\gamma_2}\right]$$

where $s \mapsto g(s)$ *satisfies the nonlinear differential equation*

$$g'(s) = \left(\frac{1}{\gamma_2}\left(\frac{s}{g(s)}\right)^{\gamma_2} - \frac{1}{\gamma_1}\left(\frac{s}{g(s)}\right)^{\gamma_1}\right) \Big/ \left(\left(\frac{s}{g(s)}\right)^{\gamma_2+1} - \left(\frac{s}{g(s)}\right)^{\gamma_1+1}\right).$$

The maximality principle says that maximal solution of the differential equation is the optimal stopping boundary. It can be shown that $g_*(s) = \beta_* s$ *is the one.*

3. *The standard procedure of applying Itô formula, Fatou's lemma etc. can be used to verify that the estimated candidates are indeed correct. The result on the fair price of the Russian option is stated below.*

THEOREM 5 *The fair price of the Russian option is given by*

$$C_*(x, s) = \frac{s}{\gamma_2 - \gamma_1}\left(\gamma_2\left(\frac{x}{\beta_* s}\right)^{\gamma_1} - \gamma_1\left(\frac{x}{\beta_* s}\right)^{\gamma_2}\right)$$

and the optimal stopping time is given by

$$\tau_* = \inf\{t > 0 \,:\, X_t \le \beta_* S_t\}.$$

The fair price of the Russian option was calculated by [Shepp & Shiryaev, 1993] which also should be consulted for more information and details. The result is extended in [Pedersen, 2000a] to Lookback options with fixed and floating strike.

EXAMPLE 8 (**Optimal prediction of the ultimate maximum of Brownian motion**). *This example presents solutions to the problem of stopping a Brownian path as close as possible to the unknown ultimate maximum height of the path. The closeness is first measured by a mean-square distance and next by a probability distance. The optimal stopping strategies can also be viewed as selling strategies for stock trading in the idealized Bachelier model. These problems do not fall under the general optimal stopping theory, since the gain process is not adapted to the natural filtration of the process.*
In this example the diffusion $X_t = B_t$. Let

$$\Phi(x) = \int_{-\infty}^{x} \tfrac{1}{\sqrt{2\pi}}\, e^{-u^2/2}\, du$$

for $x \in \mathbb{R}$ denote the distribution function of a standard normal variable. Let $\mathcal{S}_T$ be the family of all stopping times τ for (B_t) satisfying $\tau \le T$.

Mean-square distance

This problem was formulated and solved by [Graversen, Peskir & Shiryaev, 2001] and in [Pedersen, 2003] the problem is solved for all L^q-*distances*. Consider the optimal stopping problem with value function

$$V_* = \inf_{\tau \in \mathcal{S}_1} \mathbf{E}\big((S_1 - B_\tau)^2\big). \tag{6.11}$$

The idea is to transform problem (6.11) into an equivalent problem that can be solved by the recipe presented in the previous section.
To follow the above plan, note that S_1 is square integrable; then in accordance with Itô-Clark representation theorem formula

$$S_1 = \mathbf{E}(S_1) + \int_0^1 H_u\, dB_u$$

where (H_t) is a unique adapted process satisfying $\mathbf{E}\big(\int_0^1 H_u^2\, du\big) < \infty$. Furthermore, it is known that

$$H_t = 2\left(1 - \Phi\left(\frac{S_t - B_t}{\sqrt{1-t}}\right)\right).$$

If (M_t) denote the square integrable martingale $M_t = \int_0^t H_u\,dB_u$, then the martingale theory gives that $\mathbf{E}(S_1 - B_\tau)^2 = \mathbf{E}(\int_0^\tau (1 - 2H_u)\,du) + 1$ for all $\tau \in \mathcal{S}_1$. Problem (6.11) can therefore be represented as

$$V_* = \inf_{\tau \in \mathcal{S}_1}\ \mathbf{E}\left(\int_0^\tau c\left(i\frac{S_u - B_u}{\sqrt{1 - u}} \right) du \right) + 1$$

where $c(x) = 4\Phi(x) - 3$. By Lévy's theorem and general optimal stopping theory, the problem (6.11) is equivalent to

$$V_* = \inf_{\tau \in \mathcal{S}_1}\ \mathbf{E}\left(\int_0^\tau c\left(\frac{|B_u|}{\sqrt{1 - u}} \right) du \right) + 1 \,.$$

The form of the gain function indicates that the deterministic time-change method introduced in Example 4 can be applied successfully. Let $\sigma_t = 1 - e^{-2t}$ be the time-change and let $(Z_t)_{t \geq 0}$ be the time-changed process given by $Z_t = B_{\sigma_t}/\sqrt{1 - \sigma_t}$. It can be shown by Itô formula that (Z_t) solves the stochastic differential equation

$$dZ_t = Z_t\,dt + \sqrt{2}\,d\beta_t$$

where $(\beta_t)_{t \geq 0}$ is a Brownian motion. Hence (Z_t) is a diffusion with the infinitesimal generator

$$\mathbf{L}_{\mathbf{Z}} = z\frac{\partial}{\partial z} + \frac{\partial^2}{\partial z^2}$$

for $z \in \mathbb{R}$. Substituting the time-change yields that

$$V_* = \inf_\sigma\ \mathbf{E}\left(\int_0^\sigma e^{-2u}\,c(|Z_u|)\,du \right) + 1 \,.$$

Hence the initial problem (6.11) reduces to solving

$$W_*(z) = \inf_\sigma\ \mathbf{E}_z\left(\int_0^\sigma e^{-2u}\,c(|Z_u|)\,du \right) \tag{6.12}$$

where the infimum is taken over all stopping times a for (Z_t) and $V_* = W_*(0) + 1$. This is a problem that can be solve with the recipe from Section 18.5.

*   **1.** *The domain of continued observation is a symmetric interval around zero, that is $C = \{ z \in \mathbb{R} \mid z \in (-z_*, z_*) \}$ and the value function is an even C^1-function or equivalent $W_*'(0) = 0$. From the observation in Example 3 one is led to formulate the corresponding free-boundary system of the problem (6.12)*

$$
\begin{aligned}
&\mathbf{L}_{\mathbf{Z}} W(z) - 2W(z) = -c(|z|) \text{ for } \ -z_* < z < z_* \\
&W(\pm z_*) = 0 \qquad\qquad \textit{(instantaneous stopping)} \\
&W'(\pm z_*) = 0 \qquad\qquad \textit{(smooth fit)} \\
&W'(0) = 0 \qquad\qquad\quad \textit{(normal reflection)} \,.
\end{aligned}
$$

2. *The solution of the free-boundary problem is given by*

$$W(z) = \Phi(z_*)\left(1 + z^2\right) - 2\Phi'(z) + \left(1 - z^2\right)\Phi(z) - 3/2$$

for $z \in [0, z_*]$ where z_* is the unique solution of the equation (6.13).

3. *By Itô formula it can be proved that $W(z)$ is the value function and $\sigma_* = \inf\{t > 0 : |Z_t| \geq z_*\}$ is an optimal stopping time. Transforming the value function and the optimal strategy back to the initial problem (6.11) the following result ensues (for more details see [Graversen, Peskir & Shiryaev, 2001]).*

THEOREM 6 *Consider the optimal stopping problem (6.11). Then the value function V_* is given by*

$$V_* = 2\Phi(z_*) - 1 \approx 0.73$$

where $z_* \approx 1.12$ is the unique root of the equation

$$4\Phi(z_*) - 2z_*\Phi'(z_*) - 3 = 0 . \tag{6.13}$$

The following stopping time is optimal (see Figure 3)

$$\tau_* = \inf\left\{t > 0 : \max_{0 \leq u \leq t} B_u - B_t \geq z_*\sqrt{1 - t}\right\} . \tag{6.14}$$
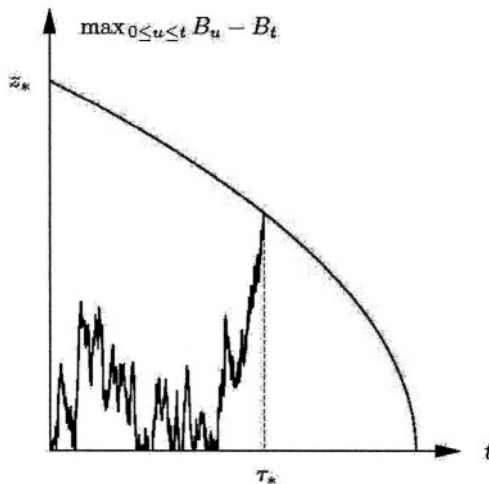
Figure 3. A computer drawing of the optimal stopping strategy (6.14).

Probability distance

   The problem was formulated and solved in [Pedersen, 2003]. Consider the optimal stopping problem with value function

$$W_*^{(\varepsilon)} = \sup_{\tau \in \mathcal{S}_1} \mathbf{P}(S_1 - B_\tau \leq \varepsilon) \qquad (6.15)$$

for $\varepsilon > 0$. Furthermore, in this case, the gain process is discontinuous. Using the stationary independents increments of (B_t) yields that

$$W_*^{(\varepsilon)} = \sup_{\tau \in \mathcal{S}_1} \mathbf{E}\Big(\mathbf{E}\big(\mathbf{1}_{[0,\varepsilon]}(S_1 - B_\tau)\,\big|\,\mathcal{F}_\tau\big)\Big)$$

$$= \sup_{\tau \in \mathcal{S}_1} \mathbf{E}\big(F_{1-\tau}(\varepsilon)\,;\, S_\tau - B_\tau \leq \varepsilon\big)$$

where $F_t(\varepsilon) = 2\Phi(\varepsilon/\sqrt{t}) - 1$ is the distribution function of S_t. By Lévy's theorem and the general optimal stopping theory the stopping problem (6.15) is equivalent to solving

$$W_*^{(\varepsilon)}(t, x) = \sup_{\tau \in \mathcal{S}_{1-t}} \mathbf{E}_x\big(F_{1-t-\tau}(\varepsilon)\,;\, |B_\tau| \leq \varepsilon\big)$$

for $t \leq 1$ and $x \in \mathbb{R}$. It can be shown that it is only optimal to stop if $|B_\tau| = \varepsilon$ on the set $\{\tau < 1 - t\}$. This observation – together with the
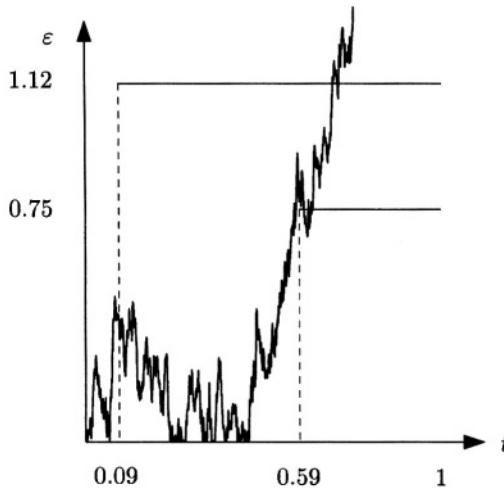
Figure 4.   A computer drawing of the optimal stopping strategy (6.16) when $\varepsilon = 0.75$ and $\varepsilon = 1.12$. Then $t_* = 0.75$ and $t_* = 0.09$ respectively.

Brownian scaling property – indicates that the optimal stopping time is of the form

$$\tau_\varepsilon^{t*} = \inf \{\, 0 < u \leq 1 - t \,:\, |B_u| = \varepsilon\, b_{t*}(t + u) \,\} \qquad (\inf \emptyset = 1 - t)$$

where the boundary function $b_{t*}(t) = \infty$ if $t < t_*$ and $b_{t*}(t) = 1$ elsewhere for some $0 \leq t_* \leq 1$ to be found. This shows that the principle of smooth fit is not satisfied in the sense that the value function $W_*^{(\varepsilon)}$ is not C^1 at all points of the boundary of the domain of continued observation. More precisely, the smooth fit breaks down in the state variable x because of the discontinuous gain function. However, due to the definition of the gain function the smooth fit should still hold in the time variable t and this implies – together with Itô formula and the shape of the domain of continued observation – that the principle of smooth fit at a single point should hold. This approach provides a method to determine t_*.
  Set

$$W^{(\varepsilon)}(t, x) = \mathbf{E}_x \left(F_{1-t-\tau_\varepsilon^0 \wedge (1-t)}(\varepsilon) \,;\, |B_{\tau_\varepsilon^0 \wedge (1-t)}| \leq \varepsilon \right)$$
$$= \mathbf{E}_x \left(F_{1-t-\tau_\varepsilon^0 \wedge (1-t)}(\varepsilon) \,;\, \tau_\varepsilon^0 \leq 1 - t \right) + \mathbf{P}_x(\tau_\varepsilon^0 > 1 - t)\, \mathbf{1}_{[0,\varepsilon]}(|x|).$$

For fixed $t < 1$, $x \mapsto W^{(\varepsilon)}(t, x)$ is in general only continuous at $|x| = \varepsilon$. Let $\varepsilon_* \approx 1.17$ be the point satisfying that $x \mapsto W^{(\varepsilon)}(0, x)$ is differentiable at $|x| = \varepsilon$. The result is the following theorem.

THEOREM 7 *Consider the optimal stopping problem* (6.15).   *Set $t_* = (1 - (\varepsilon/\varepsilon_*)^2) \vee 0$.*
  (i) If $t_* = 0$, then the value function is given by (see Figure 5)

$$W_*^{(\varepsilon)} = W^{(\varepsilon)}(0, 0)$$

$$= 1 + 2\varepsilon \int_0^1 \frac{F_{1-y}(-\varepsilon)}{y^{3/2}} \sum_{k=0}^\infty (-1)^k (2k+1)\, \varphi\left(\frac{(2k+1)\varepsilon}{\sqrt{y}} \right) dy$$

$$- 4 \sum_{k=0}^\infty (-1)^k \, \Phi\big(-(2k+1)\varepsilon \big)$$

(ii) If $t_* > 0$, then the value function is given by (see Figure 5)

$$W_*^{(\varepsilon)} = W_{t_*}^{(\varepsilon)}(0, 0) = \frac{2}{\sqrt{t_*}} \int_0^\infty W^{(\varepsilon)}(t_*, x)\, \varphi\left(\frac{x}{\sqrt{t_*}} \right) dx$$

where

$$W^{(\varepsilon)}(t,x) = 1 + \int_0^{1-t} F_{1-t-y}(\varepsilon)$$

$$\sum_{k=-\infty}^{\infty} (-1)^k \frac{x+(2k+1)\varepsilon}{y^{3/2}} \varphi\left(\frac{x+(2k+1)\varepsilon}{\sqrt{y}}\right) dy$$

$$- 2 \sum_{k=-\infty}^{\infty} (-1)^k \operatorname{sgn}\left(x+(2k+1)\varepsilon\right) \Phi\left(-\frac{|x+(2k+1)\varepsilon|}{\sqrt{1-t}}\right)$$

for $0 \le x < \varepsilon$ *and*

$$W^{(\varepsilon)}(t,x) = \int_0^{1-t} F_{1-t-y}(\varepsilon) \frac{x-\varepsilon}{y^{3/2}} \varphi\left(\frac{x-\varepsilon}{\sqrt{y}}\right) dy$$

for $x > \varepsilon$.

    In both cases, the optimal stopping time is given by (see Figure 4)

$$\tau_* = \inf\left\{ t_* < t \le 1 \ : \ \max_{0 \le u \le t} B_u - B_t = \varepsilon \right\} \qquad (\inf \emptyset = 1). \quad (6.16)$$
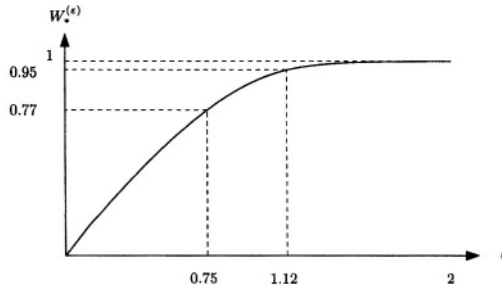
Figure 5.   A drawing of the value function $W_*^{(\varepsilon)}$ as a function of ε .

References

CHERNOFF, H. (1961). Sequential tests for the mean of a normal distribution. *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.* **1**, Univ. California Press (79-91).

CONZE, A. *and* VISWANATHAN, R. (1991). Path dependent options: The case of lookback options. *J. Finance* **46** (1893-1907).

DAVIS, B. (1976). On the L^p norms of stochastic integrals and other martingales. *Duke Math. J.* **43** (697-704).

DUBINS, L.E., SHEPP, L.A. *and* SHIRYAEV, A.N. (1993). Optimal stopping rules and maximal inequalities for Bessel processes. *Theory Probab. Appl.* **38** (226-261).

DYNKIN, E.B. (1963). Optimal choice of the stopping moment of a Markov process. *Dokl. Akad. Nauk SSSR* **150** (238-240). (In Russian).

EL KAROUI, N. (1981). Les aspects probabilites du contrôle stochastique. *Lecture Notes in Math.* **876**, Springer (73-238). (In French).

GRAVERSEN, S.E. *and* PESKIR, G. (1997). On Doob's maximal inequality for Brownian motion. *Stochastic Process. Appl.* **69** (111-125).

GRAVERSEN, S.E. *and* PESKIR, G. (1998). Optimal stopping and maximal inequalities for geometric Brownian motion. *J. Appl. Probab.* **42** (564-575).

GRAVERSEN, S.E. *and* PESKIR, G. (1998a). Optimal stopping in the *L* log *L*-inequality of Hardy and Littlewood. *Bull. London Math. Soc.* **30** (171-181).

GRAVERSEN, S.E., PESKIR, G. *and* SHIRYAEV, A.N. (2001). Stopping Brownian motion without anti-cipation as close as possible to its ultimate maximum. *Theory Probab. Appl.* **45** (41-50).

GRIGELIONIS, B.I. *and* SHIRYAEV, A.N. (1966). On the Stefan problem and optimal stopping rules for Markov processes. *Theory Probab. Appl.* **11** (541-558).

JACKA, S.D. (1991). Optimal stopping and bests constant for Doob-like inequalities I: The case $p = 1$. *Ann. Probab.* **19** (1798-1821).

LINDLEY, D.V. (1961). Dynamic programming and decision theory. *Appl. Statist.* **10** (39-51).

McKEAN, H.P. (1965). A free-boundary problem for the heat equation arising from the problem of mathematical economics. *Industrial Managem. Review* **6** (32-39).

MIKHALEVICH, V.S. (1958). Bayesian choice between two hypotheses for the mean value of a normal process. *Visnik Kiiv. Univ.* **1** (101-104). (in Ukrainian).

MYNENI, R. (1992). The pricing of the American option. *Ann. Appl. Probab.* **2** (1-23).

KSENDAL, B. (1998). *Stochastic differential equations. An introduction with applications.* (Fifth edition). Springer.

PEDERSEN, J.L. (2000). Best bounds in Doob's maximal inequality for Bessel processes. *J. Multivariate Anal.* **75** (36-46).

PEDERSEN, J.L. (2000a). Discounted optimal stopping problems for the maximum process. *J. Appl. Probab.* **37** (972-983).

PEDERSEN, J.L. (2003). Optimal prediction of the ultimate maximum of Brownian motion. *Stoch. Stoch Rep.* **75** (205-219).

PEDERSEN, J.L. *and* PESKIR, G. (1998). Computing the expectation of the Azéma-Yor stopping time. *Ann. Inst. H. Poincare Probab. Statist.* **34** (265-276).

PEDERSEN, J.L. *and* PESKIR, G. (2000). Solving non-linear optimal stopping problems by the method of time-change. *Stochastic Anal. Appl.* **18** (811-835).

PESKIR, G. (1998). Optimal stopping of the maximum process: The maximality principle. *Ann. Probab.* **26** (1614-1640).

PESKIR, G. *and* SHIRYAEV, A.N. (2000). Sequential testing problems for Poisson processes. *Ann. Statist.* **28** (837-859).

SHEPP, L.A. (1969). Explicit solutions to some problems of optimal stopping. *Ann. Math. Statist.* **40** (993-1010).

SHEPP, L.A. *and* SHIRYAEV, A.N. (1993). The Russian option: Reduced regret. *Ann. Appl. Probab.* **3** (631-640).

SHIRYAEV, A.N. (1961). The problem of quickest detection of a violation of stationary behavior. *Dokl. Akad. Nauk SSSR* **138** (1039-1042). (In Russian).

SHIRYAEV, A.N. (1969). Two problems of sequential analysis. *Cybernetics* **3** (63-69).

SHIRYAEV, A.N. (1978). *Optimal stopping rules.* Springer.

SNELL, J.L. (1952). Application of martingale system theorems. *Trans. Amer. Math. Soc.* **73** (293-312).

TAYLOR, H.M. (1968). Optimal stopping in a Markov process. *Ann. Math. Statist.* **39** (1333-1344).

VAN MOERBEKE, P. (1974). Optimal stopping and free boundary problems. *Rocky Mountain. J. Math.* **4** (539-578).

WALD, A. (1947). *Sequential Analysis.* John Wiley and Sons.

CRITICALITY IN EPIDEMICS:
THE MATHEMATICS OF SANDPILES EXPLAINS UNCERTAINTY IN EPIDEMIC OUTBREAKS

Nico Stollenwerk

School of Biological Sciences, Royal Holloway, University of London, Egham, Surrey TW20 OEX, UK

nks22@cam.ac.uk

19.1   Introduction

The universality of critical phenomena in phase transitions has attracted physicists for more than 25 years [Stanley, 1971]. Soon after also the relevance for epidemiological and in general birth-death processes was recognized ([Grassberger & de la Torre, 1979],[Grassberger, 1983]). For a recent popular account of universality and its applications in various scientific fields see [Warden, 2001].

Two case studies will be presented to demonstrate the various aspects of criticality in epidemiology. In our first case studies we will show how an epidemic system can display huge variability while crossing a critical threshold: Measles in decreasing vaccination levels caused by a loss of confidence in vaccines in an originally highly vaccinated population (e.g. due to ongoing discussions on vaccine side effects, especially the combined measles, mumps and rubella vaccine MMR claimed to cause autism, as discussed in Great Britain).

Not only criticality as such but development of a system towards this criticality has been postulated for physical systems ([Bak et al, 1987], [Bak et al, 1988]) with the paradigmatic system of a sand pile (see for an overview [Jensen, 1998]).

In our second case study we present a system consisting of host classes infected with different mutants of a pathogenic agent leading the epidemic system towards criticality: bacterial meningitis. This system is of much broader interest, since it potentially provides an explanation for uncertainties and huge fluctuations for more general models in evolutionary biology. This approach is more realistic than previous attempts in oversimplified evolutionary models ([Bak & Sneppen, 1993], [Flyvbjerg et al, 1993]). We show explicitly that a parameter is automatically driven towards its critical value. The pathogenicity evolves to small values near its critical value of zero. In the analysis it evolves

to zero, since for analytic treatability we use reasonable approximations that show correct qualitative behaviour. In the full system the pathogenicity will evolve to small values, in the order of magnitude of the mutation rate where competing strains can replace each other.

Epidemics with critical fluctuations have been described in the literature before ([Rhodes & Anderson, 1996], [Rhodes et al, 1997]) in forest fire like scenarios ([Jensen, 1998], p. 68). We present a non-spatial stochastic model, especially a master equation (time-continuous Markov process), leading in criticality to power laws with exponents of mean field type (essentially the branching process exponent 3/2), confirming that the system under investigation really establishes critical fluctuations with fat tail behaviour.

A spatial system analysis would require a renormalization approach to path integrals which are derived from the spatial master equation. This method is still under controversial debate, even in chemical systems' analysis ([Cardy, 1996], [Wijland, 2001], [Park et al, 2000]), and can only be sketched here.

19.2   Basic epidemiological model

In this section we describe the basic epidemiological model which will underlie in modifications the following sections. It describes a non-spatial homogeneous mixing population of hosts in different states of infection. A corresponding spatial model will be given and analyzed in the final sections.

Since we will describe fluctuations near critical states we have to consider stochastic models, Markov processes explicitly formulated in master equations, as used in physics and chemistry (see e.g. [van Kampen, 1992]).

19.2.1.   The SIR-model

The basic SIR-model for a host population of size N devided in subclasses of susceptible, infected and recovered hosts [Anderson & May, 1991] is construced as follows: With a rate α a resistent host becomes susceptible, or as a reaction scheme $R \xrightarrow{\alpha} S$. Then, susceptible meet infected with a transition rate β and proportional to the number of infected (devided by N to make the model scale invariant with population size, since we obtain a quadratic term in the variables, as opposed to the linear term in the previous transition). As a reaction scheme we have $S + I \xrightarrow{\beta} I + I$. Finally, infected hosts can recover and become temporally resistent with rate γ, hence $I \xrightarrow{\gamma} R$.

We could call this basic SIR-model also SIRS-model, since transitions from *R* to *S* are allowed, but stick to SIR, since later in an SIRYX-model, with additional classes of hosts to be introduced later, parallel transitions prohibit a simple way of labelling. Hence, here SIR just means that we have three classes of hosts, *S, I* and *R* to deal with, as opposed to 5 classes in the more complicated model.

The corresponding deterministic ordinary differential equation (ODE) system reads

$$\frac{dS}{dt} = \alpha \cdot R - \beta \frac{I}{N} \cdot S$$

$$\frac{dI}{dt} = \beta \frac{I}{N} \cdot S - \gamma \cdot I \qquad (2.1)$$

$$\frac{dR}{dt} = \gamma \cdot I - \alpha \cdot R$$

and describes merely the dynamic of the mean values for the total number of susceptibles, infected and recovered under the assumptions of mean field behaviour and homogeneous mixing, hence mean values of products can be replaced by products of means in the nonlinear contact term $(\beta/N)\, I \cdot S$.

19.2.2.   Stochastic modelling

We include demographic stochasticity into the description of the epidemic. As such, for the basic SIR-model we consider the dynamics of the probability $p(S, I, R, t)$ of the system to have S susceptibles, I infected and R recovered at time t, which is governed by a master equation ([van Kampen, 1992], [Gardiner, 1985], and in a recent application to a plant epidemic model [Stollenwerk & Briggs, 2000], [Stollenwerk, 2001]). For state vectors $\underline{n}$, here for the SIR-model $\underline{n} = (S, I, R)$, the master equation reads

$$\frac{dp(\underline{n})}{dt} = \sum_{\tilde{\underline{n}} \neq \underline{n}} w_{\underline{n},\tilde{\underline{n}}}\; p(\tilde{\underline{n}}) - \sum_{\tilde{\underline{n}} \neq \underline{n}} w_{\tilde{\underline{n}},\underline{n}}\; p(\underline{n}) \qquad (2.2)$$

with transition probabilities corresponding to the ones described above for the ODE-system. Here the rates $w_{\tilde{\underline{n}},\underline{n}}$ are

$$w_{(S+1,I,R-1),(S,I,R)} = \alpha \cdot R$$

$$w_{(S-1,I+1,R),(S,I,R)} = \beta \cdot \frac{I}{N} S \qquad (2.3)$$

$$w_{(S,I-1,R+1),(S,I,R)} = \gamma \cdot I$$

from which the rates $w_{\underline{n},\tilde{\underline{n}}}$ follow immediately as

$$w_{(S,I,R),(S-1,I,R+1)} = \alpha \cdot (R+1)$$

$$w_{(S,I,R),(S+1,I-1,R)} = \beta \cdot \frac{I-1}{N} (S+1) \qquad (2.4)$$

$$w_{(S,I,R),(S,I+1,R-1)} = \gamma \cdot (I+1) \quad .$$

This formulation defines the stochastic process completely and will be the basis for modified models, e.g. additional terms for vaccination in the next section.

19.3   Measles around criticality

Measles epidemics in human populations have been a subject of investigations for a long time ([London & Yorke, 1973], [London & Yorke, 1973a], [Dietz, 1976]), since rather good empirical time series are available, and various aspects of recent paradigmatic theories like deterministic chaos in pre-vaccination dynamics ([Schwartz & Smith, 1983], [Schenzle, 1984], [Aron & Schwartz, 1984], [Schaffer, 1985], [Schaffer & Kott, 1985], [Olsen & Schaffer, 1990], [May & Sugihara, 1990], [Rand & Wilson, 1991], [Grenfell, 1992], [Bolker & Grenfell, 1993], [Drepper et al, 1994]) and criticality in island populations have been investigated ([Rhodes & Anderson, 1996], [Rhodes et al, 1997]).

Here we investigate a vaccinated population, i.e. the only stable stationary state is the disease-free population and any invading disease cases lead to quickly extinct epidemics, in which the vaccination level drops below the critical threshold, where epidemics can take off. The consideration of dropping vaccination levels is motivated by the observation that in the United Kingdom of Great Britain a discussion on side-effects of vaccines led to a dramatic drop in vaccine uptake [Jansen et al, 2002].

19.3.1.   The ODE system for the SIR-model with vaccination

The ODE system for the SIR-model with vaccination reads

$$
\begin{aligned}
\dot{S} &= \mu(N-S) - \beta\frac{I}{N}\,S - v\cdot S \\
\dot{I} &= \beta\frac{S}{N}\,I - (\gamma+\mu)I \\
\dot{R} &= \gamma I - \mu R + vS
\end{aligned}
\tag{3.1}
$$

with $v := \alpha \cdot \rho$ the vaccination rate. Here α is a time rate for the vaccination and ρ the proportion of vaccinated susceptibles. Only the product of both has importance in the model.

19.3.2.   Stationary state and vaccination threshold

From Equ. (3.1), defining functions f, g and h as

$$
\begin{aligned}
\dot{S} &=: f(S,I,R) \\
\dot{I} &=: g(S,I,R) \\
\dot{R} &=: h(S,I,R)
\end{aligned}
\tag{3.2}
$$

we obtain the stationary state by the conditions $f(S^*, I^*, R^*) = 0$, $g(S^*, I^*, R^*) = 0$ and $h(S^*, I^*, R^*) = 0$. Since we have quadratic terms we find two equilibria.

In the stationary state $I^* = 0$ (no epidemics) we find

$$S_1^* = N\frac{\mu}{\mu + v} \quad , \quad I_1^* = 0 \quad , \quad R_1^* = N - S_1^* - I_1^* = N\frac{v}{\mu + v} \quad (3.3)$$

Stability analysis gives the condition for the vaccination threshold. The Jacobian matrix around the stationary state $\underline{x} := (S^*, I^*, R^*)$ is given by

$$\frac{d\underline{f}}{d\underline{x}} = \begin{pmatrix} \frac{\partial f}{\partial S} & \frac{\partial f}{\partial I} & \frac{\partial f}{\partial R} \\ \frac{\partial g}{\partial S} & \frac{\partial g}{\partial I} & \frac{\partial g}{\partial R} \\ \frac{\partial h}{\partial S} & \frac{\partial h}{\partial I} & \frac{\partial h}{\partial R} \end{pmatrix} \quad (3.4)$$

hence

$$\frac{d\underline{f}}{d\underline{x}} = \begin{pmatrix} -\mu - \beta\frac{I^*}{N} - v & -\beta\frac{S^*}{N} & 0 \\ \beta\frac{I^*}{N} & \beta\frac{S^*}{N} - (\gamma + \mu) & 0 \\ v & \gamma & -\mu \end{pmatrix} \quad (3.5)$$

The characteristic polynomial is given by

$$\left(-\mu - \beta\frac{I^*}{N} - v - \lambda\right)\left(\beta\frac{S^*}{N} - (\gamma + \mu) - \lambda\right)(-\mu - \lambda)$$

$$+ \beta\frac{I^*}{N} \cdot \beta\frac{S^*}{N} \cdot (-\mu - \lambda) = 0 \quad (3.6)$$

One eigenvalues is simply $\lambda_3 = -\mu$, and after some calculation two further eigenvalues are: $\lambda_2 = -(\mu + v)$ and

$$\lambda_1 = \beta\frac{S^*}{N} - (\gamma + \mu) \quad (3.7)$$

those two being interesting for the further considerations. The requirement $\lambda_1 = 0$ gives the threshold value v_c, or critical vaccination value,

$$v_c = \frac{\mu}{\gamma + \mu}\left(\beta - (\gamma + \mu)\right) \quad . \quad (3.8)$$

19.3.3.   Definition and expression for the reproduction number $\mathcal{R}$

In the endemic stationary state $I^* \neq 0$, where the disease is always present, we find

$$S_2^* = N\frac{\gamma+\mu}{\beta}, \quad I_2^* = N\left(\frac{\mu}{\beta}\left(\frac{\beta}{\gamma+\mu}-1\right) - \frac{v}{\beta}\right), \quad R_2^* = N - S_2^* - I_2^* \quad .$$
(3.9)

With the heuristic definition of the reproduction level, called $\mathcal{R}$, measured in stationarity

$$\mathcal{R} \cdot \frac{S_2^*}{N} := 1$$
(3.10)

we obtain

$$\frac{S_2^*}{N} = \frac{1}{\mathcal{R}} = \frac{\gamma+\mu}{\beta} \quad .$$
(3.11)

Then the critical vaccination threshold can be expressed as function of $\mathcal{R}$

$$v_c = \frac{\mu}{\gamma+\mu}\left(\beta - (\gamma+\mu)\right) = \mu\mathcal{R} - \mu \quad .$$
(3.12)

19.3.4.   Vaccination level at criticality v_c

At the criticality threshold v_c we obtain the classical results for the vaccination threshold [Anderson & May, 1991], namely $c_c = 1 - 1/\mathcal{R}$, where c_c is the critical value of the vaccination level c when writing the ODE for S in the form

$$\dot{S} = \mu((1-c)N - S) - \beta\frac{I}{N}S$$
(3.13)

as opposed to

$$\dot{S} = \mu(N - S) - \beta\frac{I}{N}S - v \cdot S$$
(3.14)

Explicitly the argument goes as follows: At criticality $v_c = \mu(\mathcal{R}-1)$, and from the definition $\mathcal{R} = \beta/(\gamma+\mu)$, we obtain

$$\frac{S_c^*}{N} = \frac{\mu}{\mu+v_c} = \frac{1}{\mathcal{R}} \quad ,$$
(3.15)

hence

$$v_c \cdot S_c^* = \mu(\mathcal{R}-1) \cdot \frac{N}{\mathcal{R}} = \mu\left(1 - \frac{1}{\mathcal{R}}\right)N \quad .$$
(3.16)

>From

$$\dot{S} = \mu(N - S) - \beta\frac{I}{N}S - v \cdot S$$
(3.17)

we therefore have in stationarity

$$\dot{S} = \mu(N - S) - \beta\frac{I}{N}\,S - v_c \cdot S_c^* \cdot .$$ (3.18)

With Equ. (3.16) we finally get the analogous form of Equ. (3.13)

$$\dot{S} = \mu\left(\left(1 - \left(1 - \frac{1}{\mathcal{R}}\right)\right)N - S\right) - \beta\frac{I}{N}\,S$$ (3.19)

from which it follows directly that $c_c = 1 - 1/\mathcal{R}$.

19.3.5.   Parameters for measles epidemics

Rough estimates for measles parameters are average life time $\mu^{-1} = 75$ years, average infection period $\gamma^{-1} = 0.02$ years from an estimate of around 1 week.

Mean age of infection $(\beta \cdot I^*/N)^{-1} = 5$ years, with I^*/N in endemic equilibrium without vaccination, gives

$$\beta = (\gamma + \mu)\left(\frac{\mu^{-1}}{5\text{years}} + 1\right) \approx \gamma\frac{\mu^{-1}}{5\text{years}} = 750\text{years}^{-1} \quad .$$ (3.20)

The average age of vaccination can be $\alpha^{-1} \approx 1$ year to 3 years. Since it only varies the percentage of to-be-vaccinated sucseptibles ρ, we do not have to specify this parameter very accurately, taking $\alpha^{-1} = 3$ years.

19.3.6.   Stochastic simulations

Simulations are done in the frame work of master equations to capture the population noise, using Gillespie's algorithm [Gillespie, 1976]. The Gillespie algorithm, often also called minimal process algorithm, is a Monte Carlo method, in which after an event, i.e. a transition from state $\underline{n}$ to another state $\underline{\tilde{n}}$, the exponential waiting time is calculated as a random variable from the sum of all transition rates, after which the next transition is chosen randomly from all now possible transitions, according to their relative transition rates.

In analogy to the SIR-model described previously, using Equ. (2.2), the rates $w_{\tilde{n},n}$ for our model with vaccination are

$$
\begin{aligned}
w_{(S-1,I+1,R),(S,I,R)} &= \beta \cdot \frac{I}{N}\,S \\
w_{(S,I-1,R+1),(S,I,R)} &= \gamma \cdot I \\
w_{(S+1,I,R-1),(S,I,R)} &= \mu \cdot R \\
w_{(S+1,I-1,R),(S,I,R)} &= \mu \cdot I \\
w_{(S-1+1,I,R),(S,I,R)} &= \mu \cdot S \\
w_{(S-1,I,R+1),(S,I,R)} &= v \cdot S
\end{aligned}
$$ (3.21)

from which the rates $w_{\underline{n},\tilde{\underline{n}}}$ follow immediately as

$$
\begin{aligned}
w_{(S,I,R),(S+1,I-1,R)} &= \beta \cdot \frac{I-1}{N}(S+1) \\
w_{(S,I,R),(S,I+1,R-1)} &= \gamma \cdot (I+1) \\
w_{(S,I,R),(S-1,I,R+1)} &= \mu \cdot (R+1) \\
w_{(S,I,R),(S-1,I+1,R)} &= \mu \cdot (I+1) \\
w_{(S,I,R),(S+1-1,I,R)} &= \mu \cdot S \\
w_{(S,I,R),(S+1,I,R-1)} &= v \cdot (S+1) \quad .
\end{aligned}
\tag{3.22}
$$

19.3.7.   Bifurcation diagram for vaccine uptake

We plot for each value for the vaccine uptake c the size of several epidemics after 3 years, when starting with one infected at the starting time. This shows that for high uptake rates only small epidmics are found, but for low values either the epidemic takes off with high epidemic levels or still dies out quickly (bifurcation diagram). Large fluctuations are visible around the deterministic threshold value for c.
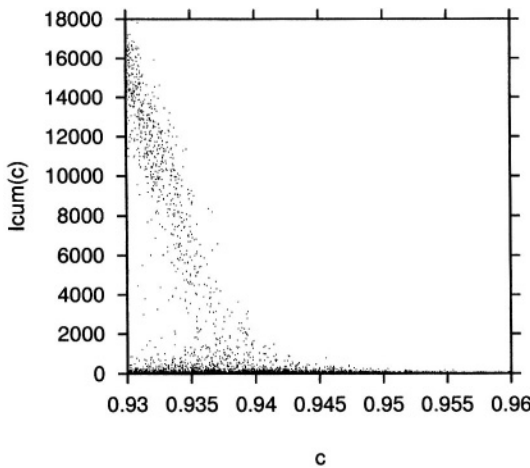
Figure 1. Bifurcation diagram for vaccine uptake c.

At the equilibrium without infected (see above), we have

$$
S_1^* = N \cdot \frac{\mu}{\mu + v}
\tag{3.23}
$$

or in terms of c instead of v

$$S_1^* = N \cdot (1 - c) \tag{3.24}$$

hence

$$v(c) = \frac{\mu c}{1 - c} \tag{3.25}$$

or

$$c(v) = \frac{v}{\mu + v} \quad . \tag{3.26}$$

In Fig. 1 we show stochastic simulations for various values of c, recalculating $v(c)$ for the simulations and starting each in the stationary values for S, R and one infected $I = 1$. The simulations are done for 3 years of epidemics. This summarizes the previous plots.

19.3.8. Epidemics when dropping the vaccine uptake

We consider the size of epidemics when lowering the uptake from 96% to 80%, introducing one infected at time t_i.
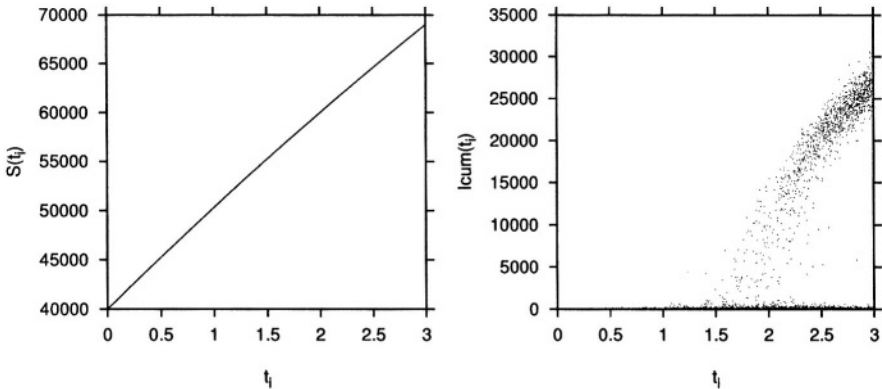
Figure 2. a) $S(t)$ with $c_1 = 96\%$ and $c_2 = 80\%$ starting at $S(t_0) := S_1^*(c_1) = N(1 - c_1)$, but with c_2 (respectively $v(c_2)$) all the time, b) Size of epidemics when dropping the uptake from 96% to 80%, introducing one infected at t

From

$$\dot{S} = \mu((1 - c)N - S) - \beta \frac{I}{N} S \tag{3.27}$$

with $c := c_2$ and $I(t) = 0,$ no infected around in the system, we obtain with $S(t_0) := S_1^*(c_1) = N(1 - c_1)$

$$S(t) = N(1 - c_1) + N(c_2 - c_1) \cdot (1 - e^{-\mu(t-t_0)}) \tag{3.28}$$

i.e. $S(t_\infty) = N(1 - c_2) = S_1^*(c_2)$ and $N(c_2 - c_1) = S(t_\infty) - S(t_0)$.

For the Fig. 2 b) we take $S(t) = S(t_i)$ as starting conditions for a stochastic simulation for 1 year of epidemics introducing exactly one infected at time t_i into the system. For the stochastic simulations and $S(t)$ we have to consider the dynamics of the fast vaccination time scale with $v(c)$ instead of c itself. So we start with

$$\dot{S} = \mu(N - S) - vS \qquad (3.29)$$

giving

$$S(t) = S(t_0) + (S(t_\infty) - S(t_0)) \cdot (1 - e^{-(\mu + v_2)(t - t_0)}) \qquad (3.30)$$

with $S(t_0) = N(1 - c_1) = N\mu/(\mu + v_1)$ and $S(t_\infty) = N(1 - c_2) = N\mu/(\mu + v_2)$ equally if expressed in v or c. This results in the faster time scale for $S(t)$ with $(\mu + v_2)$ in the exponential instead of the slow μ only.

In summary this shows that the decrease in vaccine uptake to low levels shows only after some time, during which the number of susceptibles is built up, large epidemics are becoming more and more likely. Translated into the situation in the UK, large outbreaks of measles are to be expected soon, since the vaccination level, varying regionally, has dropped from around 96% to as low as 85% and in some parts of London even below 80%.

19.4   Meningitis around criticality

This section is based on previous work [Stollenwerk & Jansen, 2002], but also includes later results. Though meningitis and septicaemia are only rarely observed diseases, and often in linked smaller or larger epidemics, the bacteria causing the disease can be detected in as many as 30 or 40 % of the host population as harmless comensals. Rarely, mutations in these bacteria occur and from time to time they make the severe mistake to harm their hosts heavily, in former times almost always fatally.

We model the host dynamics for meningitis and septicaemia as a simple SIR-model for the harmless strain of bacteria, and additional classes for the infection with mutant bacteria, called Y hosts, and heavily diseased cases X. With this model we can show that huge fluctuations appear when the chance of a mutant causing a diseased case, called pathogenicity, is small. Furthermore, we can show that in systems with mutations of various values of pathogenicity only those with small pathogenicity are present for significant periods of time. For such small values of the pathogenicity we can furthermore show power law behaviour of the size distribution of epidemics (see [Stollenwerk & Jansen, 2002] for details), hence demonstrate that the system is in criticality. The aspect of evolution towards criticality is first described here.

19.4.1.   The meningitis model

In order to describe the behaviour of pathogenic strains added to the basic SIR-system we include a new class Y of individuals infected with a potentially pathogenic strain. We will assume that such strains arise by e.g point mutations or recombination through a mutation process with a rate μ in the "reaction scheme" $S + I \xrightarrow{\mu} Y + I$. (For symmetry, we also allow the mutants to backmutate with rate ν, hence $S + Y \xrightarrow{\nu} I + Y$.)

The major point here in introducing the mutant is that the mutant has the same basic epidemiological parameters α, β and γ as the original strain and only differs in its additional transition to pathogenicity with rate ε.

These mutants cause disease with rate ε, which will turn out to be small later on, hence the reaction scheme is $S + Y \xrightarrow{\varepsilon} X + Y$. This sends susceptible hosts into an X class, which contains all hosts who develop the symptomatic disease. These are the cases wich are detectable as opposed to hosts in classes Y and I who are asymptomatic carriers who cannot be detected easily.

The state vector in the extended model is now $\underline{n} = (S, I, R, Y, X)$. The mutation transition $S + I \xrightarrow{\mu} Y + I$ fixes the master equation transition rate $w_{(S-1,I,R,Y+1,X),(S,I,R,Y,X)} = \mu \cdot (I/N) \cdot S$. In order to denote the total contact rate still with the parameter β, we keep the balancing relation

$$
\begin{aligned}
w_{(S-1,I+1,R,Y,X),(S,I,R,Y,X)} & \\
+ w_{(S-1,I,R,Y+1,X),(S,I,R,Y,X)} &= \beta \cdot \frac{I}{N} \cdot S
\end{aligned}
\tag{4.1}
$$

and obtain for the ordinary infection of normal carriage the transition rate $w_{(S-1,I+1,R,Y,X),(S,I,R,Y,X)} = (\beta - \mu) \cdot (I/N) \cdot S$. Respectively, to denote the total rate of contacts a susceptible host can make with any infected, either normal carriage I or mutant carriage Y, by β, we obey the balancing equation

$$
\sum_{\tilde{\underline{m}} \neq \underline{m}} w_{(S-1,\tilde{\underline{m}}),(S,\underline{m})} = \beta \, \frac{I+Y}{N} \cdot S
\tag{4.2}
$$

for $\underline{m} = (I, R, Y, X)$. With the above mentioned transitions this fixes the master equation rate $w_{(S-1,I,R,Y+1,X),(S,I,R,Y,X)} = (\beta - \nu - \varepsilon) \cdot (Y/N) \cdot S$.

For completeness, we introduce a recovery from the severe meningitis respectively septicaemia with rate φ, hence $X \xrightarrow{\varphi} S$. With regard to meningitis and septicaemia in many cases the disease is fatal, hence $\varphi = 0$. With medication the sufferers often survive, but are hospitalized for a long time and then suffer from resulting impairments. So for the theoretical analysis we will still keep $\varphi = 0$, which might be changed when analysing more realistic situations or recent data.

For the SIRYX-system the transition probabilities $w_{\tilde{\underline{n}},\underline{n}}$ are then given (omitting unchanged indices in $\tilde{\underline{n}}$, with respect to $\underline{n}$) by

$$
\begin{aligned}
w_{(R-1,S+1),(R,S)} &= \alpha \cdot R &,& & R &\xrightarrow{\alpha} S \\
w_{(S-1,I+1),(S,I)} &= (\beta - \mu) \cdot \tfrac{I}{N} S &,& & S+I &\xrightarrow{\beta-\mu} I+I \\
w_{(S-1,Y+1),(S,Y)} &= \mu \cdot \tfrac{I}{N} S &,& & &\xrightarrow{\mu} Y+I \\
w_{(I-1,R+1),(I,R)} &= \gamma \cdot I &,& & I &\xrightarrow{\gamma} R \\
w_{(S-1,Y+1),(S,Y)} &= (\beta - \nu - \varepsilon) \cdot \tfrac{Y}{N} S &,& & S+Y &\xrightarrow{\beta-\nu-\varepsilon} Y+Y \\
w_{(S-1,I+1),(S,I)} &= \nu \cdot \tfrac{Y}{N} S &,& & &\xrightarrow{\nu} I+Y \\
w_{(S-1,X+1),(S,X)} &= \varepsilon \cdot \tfrac{Y}{N} S &,& & &\xrightarrow{\varepsilon} X+Y \\
w_{(Y-1,R+1),(Y,R)} &= \gamma \cdot Y &,& & Y &\xrightarrow{\gamma} R \\
w_{(X-1,S+1),(X,S)} &= \varphi \cdot X &,& & X &\xrightarrow{\varphi} S
\end{aligned}
$$
$$(4.3)$$

along with the respective reaction schemes. Again from $w_{\tilde{n},n}$ the rates $w_{n,\tilde{n}}$ follow immediately. This defines the master equation for the full SIRYX-system.

19.4.2.   The invasion dynamics of mutant strains

Before we proceed with further theoretical analysis of the model we now demonstrate basic properties of our SIRYX-model in simulations of the master equation, using the Gillespie algorithm, also known as minimal process algorithm [Gillespie, 1976]. This is a Monte Carlo method, in which after an event, i.e. a transition from state $\underline{n}$ to another state $\underline{\tilde{n}}$, the exponential waiting time is calculated as a random variable from the sum of all transition rates, after which the next transition is chosen randomly from all now possible transitions, according to their relative transition rates.

To investigate the dynamics of the infection with mutants, class Y, in relation to the normal carriage I with harmless strains, we first fix the basic SIR-subsystem's parameters to the values $\alpha := 0.1$, $\beta := 0.2$ and $\gamma := 0.1$.

The endemic equilibrium of the SIR-system is given by

$$
S^* = N\frac{\gamma}{\beta} \quad , \qquad I^* = N\frac{\alpha}{\beta}\left(\frac{\beta-\gamma}{\alpha+\gamma}\right) \quad , \qquad R^* = N - S^* - I^* \quad (4.4)
$$

as can be seen from Equs. (2.1) setting the left hand side of each subequation to zero and $\gamma := 0.1$. This equilibrium would correspond to labelling 2, hence S_2^* etc., in previous chapters. As for the parameters used, we find in equilibrium a normal level of carriage of harmless infection of about 25% in our total population of size N. This is in agreement with reported levels of carriage for *Neisseria meningitidis*. Average duration of carriage is in the order of 10 months, hence we choose $\gamma = 0.1$. We assume the duration of immunity to be the same as the duration of carriage. In equilibrium this results in the ratio of

$S^* : I^* : R^* = 2 : 1 : 1$. However, the qualitative results are not affected by these first guesses of parameter values, but rather the order of magnitude.

Interesting behaviour is observed when the pathogenicity ε is too large for the hyperinvasive strain to take over but small enough to create large outbreaks of mutant infecteds Y before becoming extinct again. In Fig. 3 we show two simulations in this ε-region, first $\varepsilon = 0.05$, Fig. 3 a), b), then a ten times smaller ε, Fig. 3 c), d). For high pathogenicity ε we find relatively low levels of mutants Y, in Fig. 3 a) less than 20 cases, and at the end of the simulation roughly between 15 and 80 hospital cases X, Fig. 3 b). For smaller pathogenicity ε, Fig. 3 c), we find much larger fluctuations in the number of mutants Y with peaks of more than 80 mutant infected hosts. Though the probability rate to cause disease ε is ten times smaller than in the previous simulation we find at the end of this simulation similar numbers of disease cases X, Fig. 3 d). We observed larger fluctuations and sometimes much more outbreaks of diseased cases though the probability to create disease is smaller.

This counter-intuitive result can be understood by considering the dynamics of the hyperinvasive lineage in detail. We will do so by analyzing a simplified version of our SIRYX-model analytically.

19.4.3.   Divergent fluctuations for vanishing pathogenicity

For pathogenicity ε larger than the mutation rate μ the hyperinvasive lineage normally does not attain very high densities compared to the total population size. Therefore, we can consider the full system as composed of a dominating SIR-system which is not really affected by the rare Y and X cases, calling it the SIR-heat bath, and our system of interest, namely the Y cases and their resulting pathogenic cases X, considered to live in the SIR-heat bath.

Taking into account Equs. (4.4) for the stationary values of the SIR-system we obtain for the transition rates (compare Equs. (4.3)) of the remaining YX-system

$$
\begin{aligned}
w_{(S^*,Y+1),(S^*,Y)} &= \mu \cdot \tfrac{S^*}{N} I^* &&=: \; c \\
w_{(S^*,Y+1),(S^*,Y)} &= (\beta - \nu - \varepsilon) \cdot \tfrac{S^*}{N} Y &&=: \; b \cdot Y \\
w_{(S^*,X+1),(S^*,X)} &= \varepsilon \cdot \tfrac{S^*}{N} Y &&=: \; g \cdot Y \qquad (4.5) \\
w_{(Y-1,R^*),(Y,R^*)} &= \gamma \cdot Y &&=: \; a \cdot Y \\
w_{(X-1,S^*),(X,S^*)} &= \varphi \cdot X &&\;.
\end{aligned}
$$

All terms not involving Y or X vanish from the master equation, since the gain and loss terms cancel each other out for such transitions. If we neglect the recovery of the disease cases to susceptibility, $\varphi = 0$, as is reasonable for meningitis, we are only left with Y-dependent transition rates. Hence for the
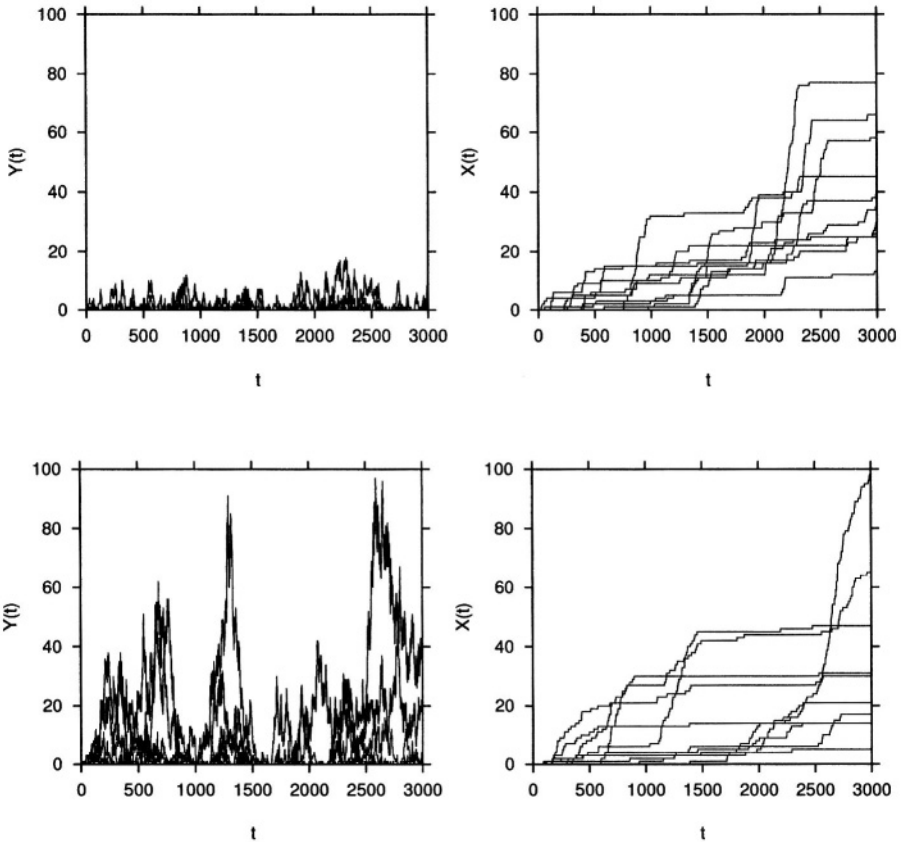
Figure 3.   a) Time series of ten runs showing the mutant carriage Y for pathogenicity $\varepsilon = 0.05$. b) Number of seriously diseased cases X for pathogenicity $\varepsilon = 0.05$. c) and d) as a) and b) with pathogenicity ten times smaller, hence $\varepsilon = 0.005$. Although the pathogenicity ε is of the factor ten smaller, the damage in the number of seriously diseased cases X remains high and even varies more than for larger ε.

YX-system we obtain the master equation

$$\frac{d}{dt}p(Y, X, t) = (b \cdot (Y - 1) + c)\, p(Y - 1, X, t) + a \cdot (Y + 1)\, p(Y + 1, X, t)$$
$$+ g \cdot Y\, p(Y, X - 1, t) - (bY + aY + gY + c)\, p(Y, X, t)$$
$$\tag{4.6}$$

This gives for the marginal distribution $p(Y, t) := \sum_{X=0}^{\infty} p(Y, X, t)$ the master equation for a simple birth-death process with birth rate $b := (\beta - \nu - \varepsilon) \cdot \frac{S^*}{N}$,

death rate $a := \gamma$ and a migration rate $c := \mu \cdot \frac{S^*}{N} \, I^*$. In the definition of the marginal distribution we take the upper limit of the summation to infinity, since we assume numbers of X and Y cases to be well below the stationary values of the SIR-system, i.e. they will not be affected by any finite upper boundary. We will check the validity of this assumption later with simulations of the full SIRYX-system.

Hence we have for $Y \in \mathbb{N}$

$$\frac{d}{dt} p(Y, t) \;=\; (b \cdot (Y - 1) + c) \, p(Y - 1, t) + a \cdot (Y + 1) \, p(Y + 1, t)$$

(4.7)

$$-(bY + aY + c) \, p(Y, t)$$

and for $Y = 0$ as boundary equation

$$\frac{d}{dt} p(Y = 0, t) = a \cdot p(Y = 1, t) - c \cdot p(Y = 0, t) \quad . \tag{4.8}$$

For the ensemble mean $\langle Y \rangle := \sum_{Y=0}^{\infty} Y \cdot p(Y, t)$ we obtain, using the above master equation,

$$\frac{d}{dt} \langle Y \rangle = (b - a) \cdot \langle Y \rangle + c \quad . \tag{4.9}$$

And for the variance, $Var(t) := \langle Y^2 \rangle - \langle Y \rangle^2$, we obtain

$$\frac{d}{dt} Var(t) = 2(b - a) Var(t) + (b + a) \cdot \langle Y \rangle + c \quad . \tag{4.10}$$

We neglect the mutation and backmutation terms, setting $c = 0$, and $\nu = 0$ in the definition for b. In this case

$$b - a = (\beta - \varepsilon) \cdot \frac{S^*}{N} - \gamma \quad = -\varepsilon \cdot \frac{S^*}{N} \tag{4.11}$$

is proportional to ε. We set $g := \varepsilon \cdot \frac{S^*}{N}$, and the ODEs for mean $Y(t) := \langle Y \rangle$ and variance $Var(t)$ then read

$$\dot{Y}(t) \;=\; -g \cdot Y(t)$$

(4.12)

$$\dot{Var}(t) \;=\; -2g \cdot Var(t) + (2\gamma - g) Y(t)$$

with initial conditions $Y(t = 0) = 1$, $Var(t = 0) = 0$. The solutions are

$$Y(t) \;=\; e^{-g(t-t_0)} \quad ,$$

(4.13)

$$Var(t) \;=\; \frac{(2\gamma - g)}{g} e^{-g(t-t_0)} \left(1 - e^{-g(t-t_0)} \right) \quad .$$

19.4.4.   Evolution towards criticality

We show now that in a population of equally distributed pathogenicity ε after some time the hosts with mutants of low pathogenicity remain in the system.   We assume initially one infected with a mutant of pathogenicity for all possible pathogenicities, and then consider the relative frequency of infected with certain pathogenicity.

$$\hat{p}(\varepsilon, t) := \frac{\langle Y \rangle(\varepsilon, t)}{\int_0^{\varepsilon_m} \langle Y \rangle(\varepsilon, t) \ d\varepsilon} \qquad (4.14)$$

with

$$\langle Y \rangle(\varepsilon, t) = e^{-gt} \qquad (4.15)$$

and $g := \varepsilon \cdot \gamma/\beta$. This is derived from the ODE $\frac{d}{dt}\langle Y \rangle = (b - a)\langle Y \rangle$ with $b - a = -g$. The result is

$$\hat{p}(\varepsilon, t) = \frac{\frac{\gamma}{\beta} t \cdot e^{-\varepsilon \frac{\gamma}{\beta} t}}{1 - e^{-\varepsilon_m \frac{\gamma}{\beta} t}} \qquad (4.16)$$

with initial distribution $\hat{p}(\varepsilon, t_0) = 1/\varepsilon_m$ for $\varepsilon \in [0, \varepsilon_m]$. For time going towards infinity $\hat{p}(\varepsilon, t \to \infty) = \delta(\varepsilon)$, hence all mass at $\varepsilon = 0$.

19.4.5.   Simulation for the full SIRYX model

In simulations of the full SIRYX-system we consider a variety of pathogenicities ε_i and for each of those we perform a large number of runs j, recording the number of mutant infected $Y_j(\varepsilon_i, t)$ over time. Hence the distribution of pathogenicities in an ensemble of hosts infected with different mutant strains is given by

$$\hat{p}(\varepsilon_i, t) := \frac{\sum_j Y_j(\varepsilon_i, t)}{\sum_i \sum_j Y_j(\varepsilon_i, t) \cdot \Delta\varepsilon} \qquad (4.17)$$

with $\Delta\varepsilon$ the length of the considered ε-interval times the number of ε-values. We compare the simulation results with the previous theoretical results in Fig. 5.

19.4.6.   Power law at criticality

We have shown previously [Stollenwerk & Jansen, 2002] that the size of the epidemics, once the epidemics have died out, follows a power law as observed in branching processes. These power laws are a characteristic sign for criticality.

In a simplified model, where the SIR-subsystem is assumed to be stationary (due to its fast dynamics), we can show analytically divergence of variance and power law behaviour for the size of the epidemics $p(X)$ as soon as the
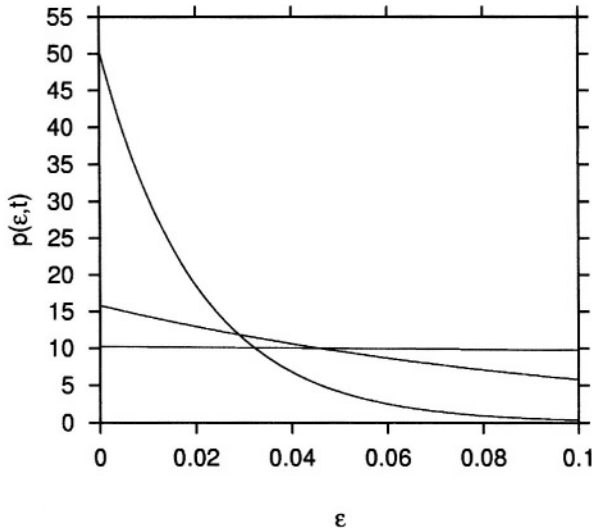
Figure 4. Times $t = 1$, horizontal line, $t = 20$, slightly tilted line, and $t = 100$, where all the probability is going towards small pathogenicity values.

pathogenicity is going towards zero. Hence the counter-intuitively large number of disease cases in some realizations of the process can be understood as large scale fluctuations in a critical system with order parameter ε towards zero.

The master equation for YX in stationary SIR results in a birth-death process

$$\frac{d}{dt} p(Y, X, t) = (b \cdot (Y - 1) + c) \, p(Y - 1, X, t)$$
$$+ a \cdot (Y + 1) \, p(Y + 1, X, t) + g \cdot Y \, p(Y, X - 1, t) \quad (4.18)$$
$$- (bY + aY + gY + c) \, p(Y, X, t) \quad .$$

Considering $\varepsilon \to 0$ and large X, we obtain power law behaviour for the size distribution of the epidemic

$$p_\varepsilon(X) := \lim_{t \to \infty} p(Y = 0, X, t) \sim \frac{1}{2\sqrt{\pi \beta}} \cdot \varepsilon^{\frac{1}{2}} \cdot X^{-\frac{3}{2}} \quad . \quad (4.19)$$

This was obtained by approximations to a solution with the hypergeometric function

$$p_\varepsilon(X) = \sqrt{\varepsilon} \cdot \frac{2^{-(X+1)}}{\sqrt{\beta}} \cdot {}_2F_1 \left(\frac{3 - X}{2}, \frac{2 - X}{2}; 2; 1 - \frac{\varepsilon}{\beta} \right). \quad (4.20)$$
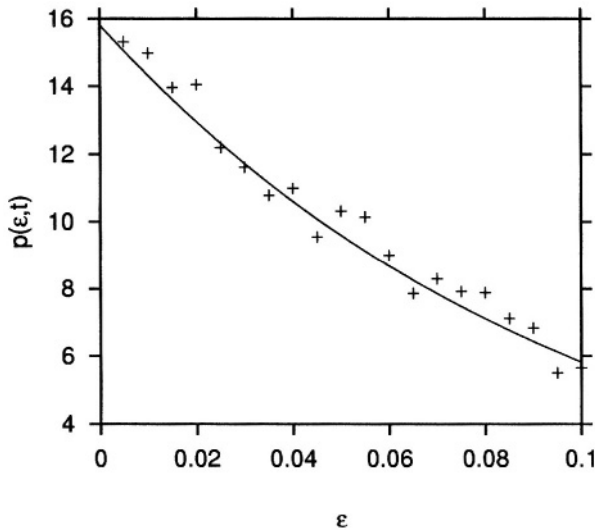
Figure 5.   Comparison of simulations of the complete SIRYX-system with the theoretical curve from the YX-subsystem and assumption of SIR in stationarity. Here time $t = 20$ is shown.

Such behaviour near criticality is also observed in the ful SIRYX-system in simulations where the pathogenicity ε is small, i.e. in the range of the mutation rate μ.

   In spatial versions of this model it is expected that the critical exponents are those of directed percolation (private communication, H.K. Jansen, Duessel-dorf, see also [Janssen, 1981]). We will discuss the directed percolation and its relation to birth-death processes in a subsequent section.

19.5   Spatial stochastic epidemics

   Non-spatial stochastic processes, as described e.g. in [van Kampen, 1992] for chemical and physical processes, have been applied to biology for a long time [Goel & Richter-Dyn, 1974], whereas spatial aspects have more recently enjoyed considerable attention among biologists, especially ecologists and epi-demiologists (e.g. [Keeling et al, 1997], for an overview of the development during the 1990s see [Rand, 1999], and recently [Dieckmann et al, 2000]).

   As a starting point we use the master equation approach for a spatial system as for example used in [Glauber, 1963] and derive from it equations for the dynamics of moments, which under additional assumptions give closed ODE-

systems (moment closure methods). Such ODE-systems have very recently been used to manage real world epidemics [Ferguson et al, 2001]. In the easiest moment closure, the mean field assumption, the usual ODEs are found back which classically were used as starting points for deterministic models. We will show this explicitly for the easiest SIS-model. The approach can be applied easily to more complicated models with some more writing effort.

The spatial master equation as used here will also be applied to investigate the fluctuations around critical points, a situation in which the simple moment closure assumptions do not hold any more. For detailed analysis see [Cardy & Täuber, 1998], [Brunel et al, 2000]) and related [Grassberger, 1983], [Grassberger & Scheunert, 1980], [Peliti, 1985]. The basic procedure will be described in the following section.

19.5.1.   Spatial master equation

One of the simplest and best studied spatial processes is the birth-death process with birth rate b and death rate a on N sites, of which each can be either inhabited $I := 1$, or empty or solo $S := 1,$ hence $I = 0$ (in general $S := I - 1$).

Translated into epidemiology, I is the infected, S the susceptible class, b the infection rate, a the recovery. We refer to it as SIS-system. (In this section we use letters a and b etc. as is conventional for spatial birth-death processes with no reference to notations used in previous sections.) The master equation for the spatial SIS-system is for N lattice points

$$\frac{d}{dt}p(I_1, ..., I_N, t) = \sum_{i=1}^{N} w_{I_i, 1-I_i} \; p(I_1, ..., 1 - I_i, ..., I_N, t)$$

(5.1)

$$- \sum_{i=1}^{N} w_{1-I_i, I_i} \; p(I_1, ..., I_i, ..., I_N, t) \quad ,$$

where $I_i \in \{0, 1\}$ and transition rates

$$w_{I_i, 1-I_i} = b \left(\sum_{j=1}^{N} J_{ij} I_j \right) \cdot I_i + a \cdot (1 - I_i) \quad , \tag{5.2}$$

and

$$w_{1-I_i, I_i} = b \left(\sum_{j=1}^{N} J_{ij} I_j \right) \cdot (1 - I_i) + a \cdot I_i \quad , \tag{5.3}$$

with b birth or infection rate and a death or recovery rate. Here (J_{ij}) is the adjacency matrix containing 0 for no connection and 1 for a connection between sites i and j, hence $J_{ij} = J_{ji} \in \{0, 1\}$ for $i \neq j$ and $J_{ii} = 0$.

Define the number of clusters with certain shapes, for total number

$$[I] := \sum_{i=1}^{N} I_i \tag{5.4}$$

and respectively

$$[S] := \sum_{i=1}^{N} (1 - I_i) \tag{5.5}$$

and for pairs

$$[II] := \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} \, I_i \cdot I_j \tag{5.6}$$

and triples

$$[III] := \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} J_{ij} J_{jk} \cdot I_i I_j I_k \tag{5.7}$$

or triangles

$$[\Delta] := \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} J_{ij} J_{jk} J_{ki} \cdot I_i I_j I_k \tag{5.8}$$

and so on.

These space averages, e.g $[I] := \sum_{i=1}^{N} I_i$, depend on the ensemble $(I_1, ..., I_N)$ which changes with time. Hence we define the ensemble average, e.g.

$$\langle I \rangle(t) := \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1} [I] \, p(I_1, ..., I_N, t)$$

or more generally for any function $f = f(I_1, ..., I_N)$ of the state variables we define the ensemble average as

$$\langle f \rangle(t) := \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1} f(I_1, ..., I_N) \, p(I_1, ..., I_N, t) \quad . \tag{5.9}$$

We will consider mainly functions like $f = [I]$, $f = [II]$ etc.. Then the time evolution is determined by

$$\frac{d}{dt} \langle f \rangle(t) := \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1} f(I_1, ..., I_N) \frac{d}{dt} p(I_1, ..., I_N, t) \tag{5.10}$$

where the master equation is to be inserted again giving terms of the form $\langle f \rangle$ and other expressions $\langle g(I_1, ..., I_N) \rangle$.

By defining marginal distributions

$$p(I_i, t) := \sum_{I_1=0}^{1} \cdots \not\!\!\sum_{I_i=0}^{1} \cdots \sum_{I_N=0}^{1} p(I_1, ..., I_N, t) \qquad (5.11)$$

and respectively

$$p(I_i, I_j, t) := \sum_{I_1=0}^{1} \cdots \not\!\!\sum_{I_i=0}^{1} \cdots \not\!\!\sum_{I_j=0}^{1} \cdots \sum_{I_N=0}^{1} p(I_1, ..., I_N, t) \qquad (5.12)$$

one obtains for its realizations useful expressions like

$$p(I_i = 1, I_j = 0, t) := \sum_{I_1=0}^{1} \cdots \not\!\!\sum_{I_i=0}^{1} \cdots \not\!\!\sum_{I_j=0}^{1} \cdots \sum_{I_N=0}^{1} p(I_1, ..., I_i = 1,$$

$$I_j = 0, ..., I_N, t)$$

$$= \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1} I_i(1 - I_j)\, p(I_1, ..., I_N, t)$$

$$=: \langle I_i(1 - I_j) \rangle = \langle I_i \rangle - \langle I_i I_j \rangle \qquad (5.13)$$

which we will consider extensively in the subsequent text. The crossed out summation signs in $\sum_{I_1=0}^{1} \cdots \not\!\!\sum_{I_i=0}^{1} \cdots \sum_{I_N=0}^{1}$ indicate summation with respect to all sites I_1 to I_N, only excluding summation over I_i.

Hence it follows

$$\langle II \rangle(t) = \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} \langle I_i I_j \rangle \qquad (5.14)$$

and with $\langle S_i I_j \rangle = \langle (1 - I_i) I_j \rangle$

$$\langle SI \rangle(t) = \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} \langle S_i I_j \rangle = \sum_{i=1}^{N} \langle I_i \rangle \left(\sum_{j=1}^{N} J_{ij} \right) - \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} \langle I_i I_j \rangle \qquad (5.15)$$

with

$$\sum_{i=1}^{N} \langle I_i \rangle \left(\sum_{j=1}^{N} J_{ij} \right) = Q \cdot \sum_{i=1}^{N} \langle I_i \rangle = Q \cdot \langle I \rangle \qquad (5.16)$$

for $Q_i := \left(\sum_{j=1}^{N} J_{ij} \right) = Q$ the number of neighbours to site i, here assumed to be constant Q.

In more general, terms of the form

$$\langle II \rangle_\nu := \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij}^{\nu} \cdot I_i I_j \tag{5.17}$$

will appear with any ν^{th} power of the adjacency matrix, e.g. $J_{ij}^2 = \sum_{k=1}^{N} J_{ik} J_{kj}$, and respectively

$$\langle III \rangle_{\mu,\nu} := \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} J_{ij}^{\mu} J_{jk}^{\nu} \cdot I_i I_j I_k \tag{5.18}$$

and so on.

19.5.2.   Time evolution of marginals and local expectations

For the marginals we can put forward some rules which are rigorous but also intuitively obtained from the master equation.

The birth-death process (or equivalently the SIS-epidemics, and for a more general class of processes specified below) presents the following expressions for the dynamics of local quantities (like $\langle I_i \rangle$ etc.)

$$\frac{d}{dt} \langle I_i \rangle := \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1} \frac{d}{dt} p(I_1, ..., I_N, t) \tag{5.19}$$

$$= \sum_{\{I\}} I_i \left(\sum_{k=1}^{N} w_{I_k, 1-I_k} \; p(I_1, ..., 1 - I_k, ..., I_N, t) \right.$$

$$\left. - \sum_{k=1}^{N} w_{1-I_k, I_k} \; p(I_1, ..., I_k, ..., I_N, t) \right)$$

using the definition $\sum_{\{I\}} := \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1}$ for the ensemble average and by inserting the master equation for the time derivative of the probability.

For any function $f(I_i, I_j)$ we have

$$\sum_{\{I\}} f(I_i, I_j) \, p(I_1, ..., 1 - I_i, ..., I_N, t) = \sum_{\{I\}} f(1 - I_i, I_j) \, p(I_1, ..., I_i, ..., I_N, t). \tag{5.20}$$

This is obtained from the elementary consideration

$$\sum_{I_i=0}^{1} f(I_i)p(1-I_i,t)$$

$$= f(0)p(1,t) + f(1)p(0,t)$$
$$= f(1)p(0,t) + f(0)p(1,t)$$
$$= \sum_{I_i=0}^{1} f(1-I_i)p(I_i,t)$$

and results in

$$\frac{d}{dt}\langle I_i\rangle = \sum_{\{I\}} I_i \left(\sum_{k=1,k\neq i}^{N} w_{1-I_k,I_k} \; p(I_1,...,I_k,...,I_N,t) \right.$$

$$\left. - \sum_{k=1,k\neq i}^{N} w_{1-I_k,I_k} \; p(I_1,...,I_k,...,I_N,t) \right)$$

$$+ \sum_{\{I\}} (1-I_i) \; w_{1-I_i,I_i} \; p(I_1,...,I_i,...,I_N,t)$$

$$- I_i \; w_{1-I_i,I_i} \; p(I_1,...,I_i,...,I_N,t)$$

$$(5.21)$$

$$= \sum_{\{I\}} w_{1-I_i,I_i} \left((1-I_i) - I_i \right) \; p(I_1,...,I_i,...,I_N,t) \qquad .$$

For the variable $I_i \in \{0,1\}$ we obtain the equations $I_i^2 = I_i$ and $(1-I_i)^2 = (1-I_i)$ and hence $I_i(1-I_i) = 0$ so that for the birth-death process

$$w_{1-I_i,I_i} \cdot \left((1-I_i) - I_i \right) = b \left(\sum_{j=1}^{N} J_{ij}I_j \right) \cdot \underbrace{\left((1-I_i)^2 - (1-I_i)I_i \right)}_{=(1-I_i)}$$

$$+ a \underbrace{\left(I_i(1-I_i) - I_i^2 \right)}_{=-I_i}$$

$$(5.22)$$

$$=: \; \tilde{w}_{1-I_i,I_i}$$

with a function $\tilde{w}$ with additive birth and subtractive death term.

The equation

$$w_{1-I_i,I_i} \cdot \left((1 - I_i) - I_i \right) = \tilde{w}_{1-I_i,I_i} \tag{5.23}$$

holds for general transition probabilities of the functional form

$$w_{1-I_i,I_i} = f(\{I_j\}_{j \neq i}) \cdot (1 - I_i) + g(\{I_j\}_{j \neq i}) \cdot I_i \tag{5.24}$$

with arbitrary functions f for birth terms and g for death terms and $\tilde{w}$ defined as

$$\tilde{w}_{1-I_i,I_i} := f(\{I_j\}_{j \neq i}) \cdot (1 - I_i) - g(\{I_j\}_{j \neq i}) \cdot I_i \quad . \tag{5.25}$$

Hence we obtain

$$
\begin{aligned}
\frac{d}{dt} \langle I_i \rangle &= \sum_{\{I\}} \tilde{w}_{1-I_i,I_i} \; p(I_1, ..., I_i, ..., I_N, t) \\
\\
&= b \sum_{j=1}^{N} J_{ij} \langle I_j (1 - I_i) \rangle - a \langle I_i \rangle \\
\\
&= b \sum_{j=1}^{N} J_{ij} \langle S_i I_j \rangle - a \langle I_i \rangle
\end{aligned}
\tag{5.26}
$$

where in the last line we used again $S_i := 1 - I_i$. This provides an easy and intuitive way to calculate generally such dynamics of local expectation values.

19.5.3.   Moment equations

For the total number $\langle I \rangle := \sum_{i=1}^{N} \langle I_i \rangle$ we obtain the dynamics

$$
\begin{aligned}
\frac{d}{dt} \langle I \rangle &= \sum_{i=1}^{N} \frac{d}{dt} \langle I_i \rangle \\
\\
&= \sum_{i=1}^{N} \left(-a \langle I_i \rangle + b \sum_{j=1}^{N} J_{ij} (\langle I_j \rangle - \langle I_i I_j \rangle) \right) \\
\\
&= -a \underbrace{\sum_{i=1}^{N} \langle I_i \rangle}_{=\langle I \rangle} + b \underbrace{\sum_{j=1}^{N} \langle I_j \rangle \underbrace{\sum_{i=1}^{N} J_{ij}}_{=Q_j = Q}}_{=Q\langle I \rangle} - b \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} \langle I_i I_j \rangle}_{=\langle II \rangle_1} \\
\\
&= -a \langle I \rangle + bQ \langle I \rangle - b \langle II \rangle_1 \quad .
\end{aligned}
\tag{5.27}
$$

Hence

$$\frac{d}{dt}\langle I\rangle = b\left(Q\langle I\rangle - \langle II\rangle_1\right) - a\langle I\rangle$$

$$= b\langle SI\rangle_1 - a\langle I\rangle \tag{5.28}$$

with $\langle SI\rangle_1 := \sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}\langle S_iI_j\rangle = Q\langle I\rangle - \langle II\rangle_1$. To obtain the dynamics for the total number of pairs

$$\frac{d}{dt}\langle II\rangle_1 = \sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}\frac{d}{dt}\langle I_iI_j\rangle \tag{5.29}$$

we have to calculate first $\frac{d}{dt}\langle I_iI_j\rangle$ from the rules given above and using the master equation. We thereby have

$$\frac{d}{dt}\langle I_iI_j\rangle = \sum_{\{I\}} I_iI_j\, p(I_1,...,I_N,t)$$

$$= \sum_{\{I\}} \underbrace{\left(I_j((1-I_i)-I_i)w_{1-I_i,I_i} + I_i((1-I_j)-I_j)w_{1-I_j,I_j}\right)}_{\text{see equation (5.21)}}$$

$$\cdot\, p(I_1,...,I_N,t)$$

$$= \underbrace{\langle I_j\tilde{w}_{1-I_i,I_i} + I_i\tilde{w}_{1-I_j,I_j}\rangle}_{\text{see equation (5.22)}}$$

$$= \langle I_j(b\sum_{k=1}^{N} J_{ik}I_k(1-I_i) - aI_i)\rangle + \langle I_i(b\sum_{k=1}^{N} J_{jk}I_k(1-I_j) - aI_j)\rangle$$

$$= b\sum_{k=1}^{N} J_{ik}\langle I_jI_k(1-I_i)\rangle - a\langle I_jI_i\rangle + b\sum_{k=1}^{N} J_{jk}\langle I_iI_k(1-I_j)\rangle - a\langle I_iI_j\rangle \tag{5.30}$$

Hence for the dynamics of nearest neighbour pairs we obtain

$$
\begin{aligned}
\frac{d}{dt}\langle II\rangle_1 &= \sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}\frac{d}{dt}\langle I_iI_j\rangle \\
&= -2a\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}\langle I_iI_j\rangle \\
&\quad - 2b\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} J_{ij}\,J_{jk}\langle I_iI_jI_k\rangle \\
&\quad + 2b\sum_{i=1}^{N}\sum_{k=1}^{N}\underbrace{\left(\sum_{j=1}^{N} J_{ij}\,J_{jk}\right)}_{=:(J^2)_{ik}}\langle I_iI_k\rangle \quad .
\end{aligned}
\tag{5.31}
$$

Here $(J^2)_{ij}$ is the matrix J squared and then taken the $ij^{\,th}$ element of that matrix J^2. This last term gives a contribution of the form $\langle II\rangle_2$, see equation (5.17).

In total we obtain for the pair dynamics

$$
\begin{aligned}
\frac{d}{dt}\langle II\rangle_1 &= 2b\left(\langle II\rangle_2 - \langle III\rangle_{1,1}\right) - 2a\langle II\rangle_1 \\
&= b\langle ISI\rangle_{1,1} - 2a\langle II\rangle_1
\end{aligned}
\tag{5.32}
$$

with $\langle ISI\rangle_{1,1} := \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} J_{ij}J_{jk}\langle I_i(1-I_j)I-k\rangle$. Again the ODE for the nearest neighbours pair $\langle II\rangle_1$ involves higher moment terms like $\langle II\rangle_2$ and $\langle III\rangle_{1,1}$.

We now try to approximate the higher moments in terms of lower in order to close the ODE system. The quality of the approximation will depend on the actual parameters of the birth-death process, i.e. a and b. We first investigate the mean field approximation, expressing $\langle II\rangle_1$ in terms of $\langle I\rangle$. Then other schemes to approximate higher moments are shown, like the BBGKY-approximation (after Bogolyubov, Born, Green, Kirkwood, Yvon).

19.5.4.   Mean field behaviour

In mean field approximation, in the interaction term the exact number of inhabited neighbours is replaced by the average number of inhabitants in the full system, acting like a mean field on the actually considered site. Hence we

set

$$\sum_{j=1}^{N} J_{kj} I_j \approx \sum_{j=1}^{N} J_{kj} \frac{\langle I \rangle}{N}$$

$$= \frac{Q}{N} \cdot \langle I \rangle \quad ,$$

(5.33)

and get for $\langle II \rangle_1$ in equation (5.35)

$$\langle II \rangle_1 \quad = \quad \langle \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij} I_i I_j \rangle$$

$$= \quad \langle \sum_{i=1}^{N} I_i \sum_{j=1}^{N} J_{ij} I_j \rangle$$

$$\approx \quad \langle \sum_{i=1}^{N} I_i \frac{Q}{N} \cdot \langle I \rangle \rangle \qquad = \frac{Q}{N} \cdot \langle I \rangle \cdot \langle \sum_{i=1}^{N} I_i \rangle$$

$$= \quad \frac{Q}{N} \cdot \langle I \rangle^2 \quad ,$$

(5.34)

hence

$$\frac{d}{dt} \langle I \rangle \quad = \quad b \left(Q \langle I \rangle - \frac{Q}{N} \langle I \rangle^2 \right) - a \langle I \rangle$$

$$= \quad b \frac{Q}{N} \ (N - \langle I \rangle) \langle I \rangle - a \langle I \rangle \quad .$$

(5.35)

For homogeneous mixing, i.e. the number of neighbours equals the total population size $Q = N$, we obtain the logistic equation for the total number of inhabited sites

$$\frac{d}{dt} \langle I \rangle = b \ (N - \langle I \rangle) \langle I \rangle - a \langle I \rangle$$

(5.36)

or for the proportion $\frac{\langle I \rangle}{N} =: x \in [0, 1]$

$$\frac{d}{dt} \frac{\langle I \rangle}{N} = Nb \ \left(1 - \frac{\langle I \rangle}{N} \right) \frac{\langle I \rangle}{N} - a \frac{\langle I \rangle}{N}$$

(5.37)

and hence

$$\frac{dx}{dt} = Nb \ (1 - x) \cdot x - a \cdot x \quad .$$

(5.38)

19.5.5.   Pair approximation

For the simplest pair approximation scheme we obtain the closed ODE system

$$\frac{d}{dt}\langle I\rangle \;=\; b\left(Q\langle I\rangle - \langle II\rangle_1\right) - a\langle I\rangle$$

$$\frac{d}{dt}\langle II\rangle_1 \;\approx\; 2b\cdot\frac{2(Q\langle I\rangle - \langle II\rangle_1)}{N - \langle I\rangle} - 2a\langle II\rangle_1$$

(5.39)

where the tripple appearing originally in the second ODE is approximated by pairs and singles. For further details on approximation schemes and simulation evaluations (see e.g. [Rand, 1999]) and references there.

19.6   Directed percolation and path integrals

For a long time it has been numerically established that simple birth-death processes for mutually excluding particles on a lattice belong in criticality to the universality class of directed percolation [Grassberger & de la Torre, 1979]. But only recently, attempts have started to describe such hard-core particles in a field theory [Park et al, 2000] and even more recently in a formalism easily treated analytically to obtain such field theories, i.e. bosonic theories [Wijland, 2001]. Van Wijland uses δ-functions built from bose operators.

We show that the δ-bosons used by [Wijland, 2001] can mimic the spin 1/2 operators used in [Grassberger & de la Torre, 1979] and derive a path integral which can be compared to those analysed for directed percolation [Janssen, 1981]. To make the link between such hard-core processes and directed percolation precise is especially important for modelling epidemics, which naturally happen in entities of uninfected or single infected individuals, e.g. in plant epidemics plants on regular lattice points (see e.g. [Stollenwerk & Briggs, 2000]), or in animal and human epidemics on social network lattices (e.g. [Rand, 1999]).

19.6.1.   Master equation of the birth-death-process

One of the simplest and best studied spatial processes is the birth-death process with birth rate β and death rate α on N sites, of which each can be either inhabited $I := 1$, or empty or solo $S := 1$, hence $I = 0$ (in general $S := I$–1). In this section, α and β will stand for death respectively birth rate, since a will be used for annihilators, as is convention in particle and stochastical physics.

Translated into epidemiology, I is the infected, S the susceptible class, β the infection rate, α the recovery. We refer to it as SIS-system. The master equation for the spatial SIS-system is for N lattice points using the master equation approach for a spatial system in a form as for example used in [Glauber, 1963]

for a spin dynamics,

$$\frac{d}{dt}p(I_1, ..., I_N, t) = \sum_{i=1}^{N} w_{I_i, 1-I_i}(t) \; p(I_1, ..., 1-I_i, ..., I_N, t)$$

(6.1)

$$- \sum_{i=1}^{N} w_{1-I_i, I_i}(t) \; p(I_1, ..., I_i, ..., I_N, t)$$

for $I_i \in \{0, 1\}$ and transition rate

$$w_{I_i, 1-I_i} = \beta \left(\sum_{j=1}^{N} J_{ij} I_j \right) \cdot I_i + \alpha \cdot (1 - I_i) \quad ,$$

(6.2)

and

$$w_{1-I_i, I_i} = \beta \left(\sum_{j=1}^{N} J_{ij} I_j \right) \cdot (1 - I_i) + \alpha \cdot I_i \quad ,$$

(6.3)

with β birth or infection rate and α death or recovery rate. Here (J_{ij}) is the adjacency matrix containing 0 for no connection and 1 for a connection between sites i and j, hence $J_{ij} = J_{ji} \in \{0, 1\}$ for $i \neq j$ and $J_{ii} = 0$.

The master equation can be transformed into a Schrodinger-like equation using operators common in quantum theory ([Grassberger & Scheunert, 1980], [Peliti, 1985]), from which a path integral can be derived for the renormalization analysis.

19.6.2. Schrödinger-like equation

The master equation (6.1) can be written in the following form of a linear operator equation

$$\frac{d}{dt}|\Psi(t)\rangle = L|\Psi(t)\rangle$$

(6.4)

for a Liouville operator L to be calculated from the master equation (Equ. (5.1)) and with state vector $|\Psi(t)\rangle$ defined by

$$|\Psi(t)\rangle := \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1} p(I_1, ..., I_N, t) \left(c_1^+\right)^{I_1} \cdots \left(c_N^+\right)^{I_N} |0\rangle$$

(6.5)

$$=: \sum_{\{I\}} p(\{I\}, t) \left(\prod_{i=1}^{N} (c_i^+)^{I_i} \right) |0\rangle$$

and vacuum state $|0\rangle$. The creation and annihilation operators are defined by $c_i^+|0\rangle = |1\rangle$ and $c_i|1\rangle = |0\rangle$, and $\left(c_i^+\right)^2|0\rangle = 0$ and $c_i|0\rangle = 0$, hence

$$c_i\ |I_i\rangle\ =\ I_i \cdot |1 - I_i\rangle \tag{6.6}$$
$$c_i^+|I_i\rangle\ =\ (1 - I_i) \cdot |1 - I_i\rangle \tag{6.7}$$

and $\left(c_i^+\right)^2|I_i\rangle = c_i^2|I_i\rangle = 0$. We have anti-commutator rules on single lattice sites

$$[c_i, c_i^+]_+ := c_i c_i^+ + c_i^+ c_i = 1 \tag{6.8}$$

and ordinary commutators for different lattice sites $i \neq j$

$$[c_i, c_j^+]_- := c_i c_j^+ - c_j^+ c_i = 0 \tag{6.9}$$

respectively

$$[c_i, c_j]_- = 0 \quad , \quad [c_i^+, c_j^+]_- = 0 \quad . \tag{6.10}$$

These are exactly the raising and lowering operators in [Brunel et al, 2000] with

$$c^+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad , \quad c = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \quad , \tag{6.11}$$

for vectors

$$|1\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad , \quad |0\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad , \tag{6.12}$$

respectively product spaces of it for many particle systems as considered here. [Brunel et al, 2000] then use the Jordan-Wigner transformation to change to pure Fermi operators with anti-commutation on single sites and on different sites to obtain their path integrals. We use a different way.

The dynamics is expressed by

$$\frac{d}{dt}|\Psi(t)\rangle = \sum_{\{I\}} \left(\frac{d}{dt} p(\{I\}, t)\right) \prod_{i=1}^{N} \left(c_i^+\right)^{I_i} |0\rangle = \ldots = L|\Psi(t)\rangle \tag{6.13}$$

where the master equation has to be used to obtain the specific form of the operator L. The explicit calculations, here only denoted by ..., will be shown below.

For the birth-death process (Equ. (5.1)) the Liouville operator is after some calculation

$$L = \sum_{i=1}^{N}(1 - c_i)\beta \left(\sum_{j=1}^{N} J_{ij} c_j^+ c_j\right) c_i^+ + \sum_{i=1}^{N}(1 - c_i^+)\alpha\ c_i \quad . \tag{6.14}$$

The term $(1 - c_i)$ guarantees the normalization of the master equation solution and $\beta J_{ij} c_j^+ c_j c_i^+$ creates one infected at site i from a neighbour j which is

itself not altered. $c_j^+ c_j$ is simply the number operator on site j. Furtheron, αc_i removes a particle from site i, again ensuring normalization with $(1 - c_i^+)$.

This form Equ. (6.4) and Equ. (6.14) is exactly the form given as well in ([Grassberger & de la Torre, 1979], there pp. 392–394, Appendix A), hence using the raising and lowering operators.

However, it seems not an easy task to construct from such a Liouville operator the path integral since no coherent states are constructed for the raising and lowering operators. Therefore, [Brunel et al, 2000] proceed from these spin 1/2 operators to ferminon operators, using Grassman variables for the coherent states, whereas [Cardy & Täuber, 1998], use bose operators from the start for which coherent states are easily available (e.g. [Le Bellac, 1991], [Zinn-Justin, 1989]) hoping that rarely more than one particle will appear at a single site. But [Park et al, 2000] have emphasized once again the need for a rigorous fomulation in terms of hard core particles for which the exclusion principle on a single site is guaranteed and commutation on different sites as well.

This aim can be achieved by constructing δ-**operators** for bosons [Wijland, 2001], as will be demonstrated for our birth-death process now.

19.6.3. δ-bosons for hard-core particles

Defining Bose operators a^+ and a for states $|n\rangle$ with $n \in \mathbf{N_0}$ particles on one site by

$$a^+ |n\rangle := |n + 1\rangle \tag{6.15}$$

and

$$a|n\rangle := n \cdot |n - 1\rangle \tag{6.16}$$

and the number operator $\hat{n} := a^+ a$ with

$$a^+ a |n\rangle = n |n\rangle \tag{6.17}$$

we can use δ-functions

$$\delta_{\hat{n},k} |m\rangle = \delta_{m,k} |m\rangle \tag{6.18}$$

with a suitable representation, e.g.

$$\delta_{\hat{n},k} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{iu(\hat{n}-k)} \, du \tag{6.19}$$

[Wijland, 2001]. $\delta_{m,k}$ is the ordinary Kroneker delta whereas $\delta_{\hat{n},k}$ is an operator defined by Equation (6.18).

Then we obtain for the birth-death process the following Liouville operator

$$L = \sum_{i=1}^{N} (a_i^+ - 1)\beta \left(\sum_{j=1}^{N} J_{ij} \delta_{\hat{n}_j,1} \right) \delta_{\hat{n}_i,0} + \sum_{i=1}^{N} (a_i - 1)\alpha \, \delta_{\hat{n}_i,1} \quad . \tag{6.20}$$

which can be understood easily when replacing $\delta_{\hat{n}_i,1}$ in the bosonic theory by $c_i^+ c_i$ in the spin 1/2 theory, and $\delta_{\hat{n}_i,0}$ by $(1 - c_i^+ c_i)$ and simply replacing a_i by c_i and a_i^+ by c_i^+. Then evaluating the resulting Liouville operator in terms of the spin 1/2 commutation rules results exactly in Equ. (6.14) again.

19.6.4. Path integral for hard-core particles in a birth-death process

The path integral follows from integrating (6.4)

$$|\Psi(t)\rangle = \prod_{\nu=1}^{M} (1 + \Delta t \cdot L(t - \nu \cdot \Delta t)) |\Psi(t_M)\rangle$$

Hence with $\Delta t \to 0$, $M \to \infty$ and the finite time interval $\Delta t \cdot M = t - t_0$ we obtain for any expectation value $\langle f \rangle$ defined as

$$\langle f \rangle(t) := \sum_{\{n\}} f(\{n\}) p(\{n\}, t) = \langle P | f | \Psi(t) \rangle$$

with a Felderhof projection state $\langle P | := \langle 0 | e^{\sum_{i=1}^{N} a_i}$ [Felderhof, 1971] the path integral

$$\langle P | f | \Psi(t) \rangle = \int \cdots \int \mathcal{D}\Phi_j^*(\tilde{t}) \mathcal{D}\Phi_j(\tilde{t}) f(t) \cdot e^{- \int_{t_0}^{t} d\tilde{t} \sum_{j=1}^{N} \left(\Phi_j^*(\tilde{t}) \frac{\partial \Phi_j(\tilde{t})}{\partial t} - \mathcal{L} \right)}$$

with

$$\mathcal{D}\Phi_j(\tilde{t}) := \prod_{j=1}^{N} \prod_{\nu=0}^{M} \frac{\Phi_j(t_0 + \nu \cdot \Delta t)}{(2\pi i)^{N \cdot M}} \qquad (6.21)$$

again in the limit $M \to \infty$ and $\Delta t \to 0$. The field variables $\Phi_j^*(\tilde{t})$ and $\Phi_j(\tilde{t})$ are introduced by coherent state integrals and replace the creation and annihilation operators by complex scalar variables. Here the Lagrange function is

$$\mathcal{L}(\Phi^*, \Phi) = \left(\beta \sum_{k=1}^{N} J_{jk} |\Phi_k|^2 e^{|\Phi_k|^2} - \alpha \Phi_j \right) \cdot \left(\Phi_j^* - 1 \right) e^{|\Phi_k|^2} \quad . \qquad (6.22)$$

This compares well with the path integrals used as a starting point for further analysis of directed percolation [Janssen, 1981] when we only use the lowest order of Φ in Taylor's expansion. Higher orders are expected to give irrelevant renormalization fields.

The path integral is now ready for a further renormalization analysis (see [Cardy & Täuber, 1998]). On the numerical side the real space renormalization as initially described by [Ma, 1976] is promising for further progress in understanding the spatial birth-death process near criticality. In the following we give the derivations in more detail:

19.6.5. Product space for spin 1/2 many particle systems

With single particle creation and annihilation operators

$$c^+ := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad , \quad c := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \quad , \tag{6.23}$$

and single particle state vectors

$$|1\rangle := \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad , \quad |0\rangle := \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad , \tag{6.24}$$

the corresponding two-particle system would be constructed as a product space with 4-dimensional state vectors and 4×4- matrices. Hence the vacuum state is

$$|0\rangle := \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \quad , \tag{6.25}$$

and a state containing one particle at site 1 and no particle at site 2, hence the state $|1, 0\rangle$ is

$$|1, 0\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad , \tag{6.26}$$

being created from $c_1^+|0\rangle = c_1^+|0, 0\rangle = |1, 0\rangle$. Hence the creation operators for the two particles are the 4×4-matrices built from 2×2-matrices

$$c_1^+ = \begin{pmatrix} c^+ & 0 \\ 0 & 1| \end{pmatrix} \quad , \quad c_2^+ = \begin{pmatrix} 1| & 0 \\ 0 & c^+ \end{pmatrix} \quad , \tag{6.27}$$

with 2×2-unit matrix $1|$, or written out e.g.

$$c_1^+ = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad . \tag{6.28}$$

The other operators c_1 and c_2 follow directly from this, and commutation rules e.g. $[c_1, c_2^+]_- = 0$ can be shown easily.

19.6.6. Path integral using coherent states for hard-core bosons

The Schrödinger-like equation

$$\frac{d}{dt}|\Psi(t)\rangle = L|\Psi(t)\rangle \tag{6.29}$$

with the Liouville operator

$$L = \sum_{i=1}^{N}(a_i^+ - 1)\beta \left(\sum_{j=1}^{N} J_{ij}\delta_{\hat{n}_j,1}\right)\delta_{\hat{n}_i,0} + \sum_{i=1}^{N}(a_i - 1)\alpha\,\delta_{\hat{n}_i,1} \quad . \quad (6.30)$$

can be integrated formally using $\Delta t \to \infty$ from the quotient of differences

$$\frac{d}{dt}|\Psi(t)\rangle \approx \frac{1}{\Delta t}\left(|\Psi(t)\rangle - |\Psi(t - \Delta t)\rangle\right) = L(t - \Delta t)\,|\Psi(t - \Delta t)\rangle \quad (6.31)$$

showing

$$|\Psi(t)\rangle = (1 + \Delta t \cdot L(t - \Delta t))\,|\Psi(t - \Delta t)\rangle$$

and for several subsequent time steps

$$|\Psi(t)\rangle = \prod_{\mu=1}^{M}(1 + \Delta t \cdot L(t - \mu \cdot \Delta t))\,|\Psi(t - M \cdot \Delta t)\rangle$$

where $t - M \cdot \Delta t =: t_s$ is the starting time of the stochastic process.
    With the Felderhof projection operator [Felderhof, 1971]

$$\langle P| := \langle 0|e^{\sum_{i=1}^{N} a_i} \quad\quad (6.32)$$

and the definition for the state vector

$$|\Psi(t)\rangle = \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1} p(I_1, ..., I_N, t)\left(\prod_{i=1}^{N}(a_i^+)^{I_i}\right)|0\rangle \quad\quad (6.33)$$

any measurable quantity A as a function of the state variables I_i in the master equation formulation, respectively number operator $a_i^+ a_i$

$$A := A(I_1, ..., I_N) = A(\{I_i\}) = A(\{a_i^+ a_i\}) \quad\quad (6.34)$$

using he notation $\{I_i\} := \{I_1, ..., I_N\}$ has for its expectation value

$$\langle A\rangle(t) := \sum_{\{I_i\}} A(\{I_i\})p(\{I_i\}, t) \quad\quad (6.35)$$

the following expressions

$$\langle A\rangle(t) = \langle P|A(\{a_i^+, a_i\})|\Psi(t)\rangle \quad . \quad\quad (6.36)$$

Again we use

$$\sum_{\{I_i\}} := \sum_{I_1=0}^{1} \cdots \sum_{I_N=0}^{1} \quad . \quad\quad (6.37)$$

The path integral for an expectation value is then expressed by

$$\langle P|A|\Psi(t_f)\rangle = \langle P|A \prod_{\nu=1}^{M}(1 + \Delta t \cdot L_\nu)|\Psi(t_s)\rangle \tag{6.38}$$

with final time t_f and starting time t_s and times t_ν such that $t_0 = t_f$ and $t_M = t_s$, $L_\nu := L(t_\nu)$.

With coherent states $|\Phi\rangle := e^{\Phi \cdot a^+}|0\rangle$ and its completeness relation

$$1| = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{d\Phi^* \, d\Phi}{2\pi i} \ e^{-\Phi^*\Phi} \ |\Phi\rangle \, \langle\Phi| \tag{6.39}$$

and abbreviation $\int d^2\Phi := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{d\Phi^* \, d\Phi}{2\pi i}$ we have for N site with operators a_j, a_j^+ the completeness relation

$$1| = \int \left(\prod_{j=1}^{N} d^2\Phi(t_\nu) \right) e^{-\sum_{j=1}^{N}|\Phi_j(t_\nu)|^2} \ |\{\Phi_j(t_\nu)\}_{j=1}^{N}\rangle \, \langle\{\Phi_j(t_\nu)\}_{j=1}^{N}| \tag{6.40}$$

with $|\{\Phi_j(t_\nu)\}_{j=1}^{N}\rangle := |\Phi_1(t_\nu), ..., \Phi_N(t_\nu)\rangle$.

We now can introduce unit operators $1|$ in between every time slice of the path integral and then insert the completeness relations for the coherent states

$$\begin{aligned} \langle P|A|\Psi(t_f)\rangle \ &= \ \langle P|A \, 1| \left(\prod_{\nu=1}^{M}(1 + \Delta t \cdot L_\nu) \ 1| \right) |\Psi(t_s)\rangle \\[2mm] &= \ \int \left(\prod_{j=1}^{N} d^2\Phi(t_f) \right) \langle P|A|\{\Phi_j(t_f)\}_{j=1}^{N}\rangle \\[2mm] &\quad \cdot \ \int \left(\prod_{j=1}^{N}\prod_{\nu=1}^{M} d^2\Phi(t_\nu) \right) \left(\prod_{\nu=1}^{M} e^{-\sum_{j=1}^{N}|\Phi_j(t_{\nu-1})|^2} \right. \\[2mm] &\qquad \left. \langle\{\Phi_j(t_{\nu-1})\}_{j=1}^{N}| (1 + \Delta t \cdot L_\nu)|\{\Phi_j(t_\nu)\}_{j=1}^{N}\rangle \right) \\[2mm] &\quad \cdot \ e^{-\sum_{j=1}^{N}|\Phi_j(t_s)|^2} \ \langle\{\Phi_j(t_f)\}_{j=1}^{N}|\Psi(t_s)\rangle \end{aligned} \tag{6.41}$$

considering the non-boundary terms

$$(*) := \int \left(\prod_{j=1}^{N} \prod_{\nu=1}^{M} d^2\Phi(t_\nu) \right) \left(\prod_{\nu=1}^{M} e^{-\sum_{j=1}^{N} |\Phi_j(t_{\nu-1})|^2} \right.$$

$$\left. \langle \{\Phi_j(t_{\nu-1})\}_{j=1}^{N} | (1 + \Delta t \cdot L_\nu) | \{\Phi_j(t_\nu)\}_{j=1}^{N} \rangle \right) \tag{6.42}$$

further in the following.

It is

$$\langle \{\Phi_j(t_{\nu-1})\}_{j=1}^{N} | (1 + \Delta t \cdot L_\nu) | \{\Phi_j(t_\nu)\}_{j=1}^{N} \rangle$$

$$= e^{-\sum_{j=1}^{N} \Phi_j^*(t_{\nu-1}) \cdot \Phi_j(t_\nu)} + \Delta t \cdot \tilde{L}_\nu \tag{6.43}$$

with

$$\langle \{\Phi_j(t_{\nu-1})\}_{j=1}^{N} | \{\Phi_j(t_\nu)\}_{j=1}^{N} \rangle = e^{-\sum_{j=1}^{N} \Phi_j^*(t_{\nu-1}) \cdot \Phi_j(t_\nu)} \tag{6.44}$$

and

$$\tilde{L}_\nu := \langle \{\Phi_j(t_{\nu-1})\}_{j=1}^{N} | L_\nu | \{\Phi_j(t_\nu)\}_{j=1}^{N} \rangle$$

$$= \sum_{k=1}^{N} \alpha \cdot \langle \{\Phi_j(t_{\nu-1})\}_{j=1}^{N} | a_k \, \delta_{\hat{n}_k,1} | \{\Phi_j(t_\nu)\}_{j=1}^{N} \rangle - \alpha \cdot \Phi_k(t_{\nu-1}) \, \Phi_k(t_\nu)$$

$$+ \sum_{k=1}^{N} \beta \cdot \left(\sum_{\ell=1}^{N} J_{k,\ell} \Phi_\ell(t_{\nu-1}) \, \Phi_\ell(t_\nu) e^{\sum_{j=1, j \neq k, \ell}^{N} \Phi_j(t_{\nu-1}) \, \Phi_j(t_\nu)} \right)$$

$$\cdot \left(\Phi_k(t_{\nu-1}) - 1 \right) e^{\sum_{j=1, j \neq k}^{N} \Phi_j(t_{\nu-1}) \, \Phi_j(t_\nu)} \tag{6.45}$$

with

$$\langle \{\Phi_j(t_{\nu-1})\}_{j=1}^{N} | a_k \, \delta_{\hat{n}_k,1} | \{\Phi_j(t_\nu)\}_{j=1}^{N} \rangle = \Phi_k(t_\nu) \cdot e^{\sum_{j=1, j \neq k}^{N} \Phi_j(t_{\nu-1}) \, \Phi_j(t_\nu)} \tag{6.46}$$

et cetera using the coherent state definition

$$|\{\Phi_j(t_\nu)\}_{j=1}^{N} \rangle := e^{\sum_{j=1}^{N} \Phi_j(t_\nu) \, a_j^+} |0\rangle \quad . \tag{6.47}$$

In this way we obtain completely the path integral as given above.

19.7   Summary

We have described epidemic processes near criticality, and have given analysis for mean field models under homogeneous mixing conditions. In one case

we found that an epidemiological system evolves on its own towards criticality, hence self-organizes itself towards the critical state. For spatial systems we have presented the basic description of the master equation and have shown the connection with the previous sections under the explicit analysis of mean field assumptions. A complete analysis of the spatial system would reveal qualitatively the same behaviour, in particular again power laws for the distributions of epidemics, but with different exponents. The detailed analysis via renormalization is still under debate. criticality,self organized

Acknowledgments

References

Stanley, H.E. (1971) *An Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, Oxford).

Grassberger, P., & de la Torre, A. (1979) Reggeon Field Theory (Schlögel's First Model) on a Lattice: Monte Carlo Calculations of Critical Behaviour. *Annals of Physics* **122,** 373–396.

Grassberger, P. (1983) On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences* **63,** 157–172.

Warden, M. (2001) *Universality: the underlying theory behind life, the university and everything* (Macmillan, London).

Bak, P., Tang, C., & Wiesenfeld, K. (1987) Self-Organized Criticality: An explanation of 1/f Noise. *Phys. Rev. Lett.* **59,** 381–384.

Bak, P., Tang, C., & Wiesenfeld, K. (1988) Self-organized criticality. *Phys. Rev.* **A 38**, 364–374.

Jensen, H.J. (1998) *Self-organized criticality, emergent complex behaviour in physical and biological systems* (Cambridge University Press, Cambridge).

Bak, P., & Sneppen, K. (1993) Punctuated equilibrium and criticality in a simple model of evolution, *Phys. Rev. Lett.* **71,** 4083–4086.

Flyvbjerg, H., Sneppen, K., & Bak, P. (1993) Mean field theory for a simple model of evolution. *Phys. Rev. Lett.* **71,** 4087–4090.

Rhodes, C.J., & Anderson, R.M. (1996) *Nature* **381,** 600–604.

Rhodes, C.J., Jensen, H.J., & Anderson, R.M. (1997) *Proc. R. Soc. London* **B 264,** 1639–1649.

Cardy, J. (1996) *Scaling and Renormalization in Statistical Physics.* (Cambridge University Press).

Wijland, F. van (2001) Field theory for reaction-diffusion processes with hard-core particles, *Physical Review* **E 63,** 022101-1-4.

Park, S.C., Kim, D., & Park, J.M. (2000) Path-integral formulation of stochastic processes for exclusive particle systems, *Physical Review* **E 62,** 7642–7645.

van Kampen, N. G. (1992). *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam).

Anderson, R.M., & May, R. (1991). *Infectious diseases in humans* (Oxford University Press, Oxford).

Gardiner, C.W. (1985) *Handbook of stochastic methods* (Springer, New York).

Stollenwerk, N., & Briggs, K.M. (2000) Master equation solution of a plant disease model, *Physics Letters* **A 274,** 84–91.

Stollenwerk, N. (2001) Parameter estimation in nonlinear systems with dynamic noise, in *Integrative Systems Approaches to Natural and Social Sciences - System Science 2000* , eds. M. Matthies, H. Malchow & J. Kriz, (Springer-Verlag, Berlin).

Abramowitz, M., & Stegun, I.A. (1972) *Handbook of mathematical functions* (Dover Publications, New York).

Feistel, R. (1977) Betrachtung der Realisierung stochastischer Prozesse aus automatentheoretischer Sicht. *Wss. Z. WPU Rostock* **26,** 663–670.

Landau, D.P., & Binder, K. (2000) *Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge).

Dietz, K. (1976) The incident of infectious diseases under the influence of seasonal fluctuations. *Lecture Notes Biomath.* **11,** 1–15.

London, W.P. & Yorke, J.A. (1973) Recurrent outbreaks of measles, chickenpocks and mumps I. *Am J. Epidemiology* **98,** 453–468.

Yorke, J.A. & London, W.P. (1973) Recurrent outbreaks of measles, chickenpocks and mumps II. *Am J. Epidemiology* **98,** 469–482.

Olsen, L.F. & Schaffer W.M. (1990) Chaos versus noisy periodicity: Alternative hypotheses for childhood epidemics. *Science* **249,** 499–504.

May, R.M. & Sugihara, G. (1990) Nonlinear forecasting as a way of distinguishing chaos measurement errors in time series. *Nature* **344,** 734–741.

Grenfell, B.T. (1992) Chances and chaos in measles dynamics. *J. Royal Statist. Soc.* **B 54,** 383–398.

Drepper, F.R., Engbert, R., & Stollenwerk, N. (1994) Nonlinear time series analysis of empirical population dynamics, *Ecological Modelling* **75/76,** 171–181.

Aron, J.L. & Schwartz, I.B. (1984) Seasonality and period–doubling bifurcations in an epidemic model. *J. Theor. Biology* **110,** 665–679.

Schaffer, W.M. (1985) Order and chaos in ecological systems. *Ecology* **66,** 93–106.

Schaffer, W.M., & Kott, M. (1985) Nearly one dimensional dynamics in an epidemic. *J. Theor. Biology* **112,** 403–427.

Rand, D.A., & Wilson, H.B. (1991) Chaotic stochasticity: A ubiquitous source of unpredictability in epidemics. *Proc. of the Royal Society* **B 246,** 179–184.

Schwartz, I.B., & Smith, H.L. (1983) Infinite subharmonic bifurcation in an SEIR epidemic model. *J. Math. Biology* **B 18,** 233–253.

Bolker, B.M., & Grenfell, B.T. (1993) Chaos and biological complexity in measles dynamics. *Proc. R. Soc. Lond.* **B 251,** 75–81.

Schenzle, D. (1984) An age-structured model of pre- and post-vaccination measles transmission. *IMA J. Math. appl. Med. Biology* **1,** 169–191.

Jansen, V.A.A., Stollenwerk, N., Jensen, H.J., Edmunds, W.J., & Rhodes, C.J. (2002) Measles outbreaks in populations with declining vaccine uptake, *manuscript in preparation*.

Le Bellac, M. (1991). *Quantum and Statistical Field Theory* (Oxford University Press, Oxford).

Binney, J.J., Dowrick, N.J., Fisher, A.J., & Newman, M.E.J. (1992). *The Theory of Critical Phenomena, An Introduction to the Renormalization Group* (Oxford University Press, Oxford).

Yeomans, J.M. (1992). *Statistical Mechanics of Phase Transitions* (Oxford University Press, Oxford).

Zinn-Justin, J. (1989). *Quantum Field Theory and critical phenomena* (Oxford University Press, Oxford).

Privman, V. (1997). *Nonequilibrium Statistical Mechanics in One Dimension* (Cambridge University Press, Cambridge).

Rand, D.A. (1999) Correlation equations and pair approximations for spatial ecologies, in: *Advanced Ecological Theory,* ed. J. McGlade, (Blackwell Science, Oxford, London, Edinburgh, Paris), 100–142.

Keeling, M.J., Rand, D.A., & Morris, A.J. (1997) Correlation models for childhood epidemics, *Proc. Royal Soc. London* **B 264,** 1149–1156.

Stollenwerk, N., & Jansen, V. A.A. (2001) Meningitis, pathogenicity near criticality, *presented as talk at the conference "Scaling Concepts and Complex Systems", Merida, Mexico, 9.–14. July, 2001.*

Stollenwerk, N., & Jansen, V.A.A. (2002) Meningitis, pathogenicity near criticality: The epidemiology of meningococcal disease as a model for accidental pathogens. *Journal of Theoretical Biology, accepted for publication, 17.12.2002.*

Goel, N.S., & Richter-Dyn, N. (1974) *Stochastic models in biology* (Academic Press, New York).

Dieckmann, U., Law, R. & Metz, J.A.J. (2000) *The Geometry of Ecological Interactions* (Cambridge University Press, Cambridge).

vanBaalen, M. (2000) Pair Approximations for Different Spatial Geometries, in: *The Geometry of Ecological Interactions,* eds. Dieckmann, U., Law, R. & Metz, J.A.J. (Cambridge University Press, Cambridge), 359–387.

Kleczkowski, A., Bailey, D. J., & Gilligan, C. A. (1996) Dynamically generated variability in plant-pathogen systems with biological control. *Proc. Royal Soc. London* **B 263,** 777–783.

Gillespie, D.T. (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics* **22,** 403–434.

Gillespie, D.T. (1978) Monte Carlo simulation of random walks with residence time dependent transition probability rates. *Journal of Computational Physics* **28,** 395–407.

Glauber, R.J. (1963) Time-dependent statistics of the Ising model. *J. Math. Phys.* **4,** 294–307.

Stollenwerk, N. (1999) A method of numerical likelihood estimates tested on plant disease curves, *presented as talk at Dynamics Days, Como, June 1999.*

Cardy, J., & Täuber, U.C. (1998) Field theory of branching and annihilating random walks. *J. Stat. Phys.* **90,** 1–56.

Grassberger, P., & Scheunert, M. (1980) Fock-space methods for identical classical objects. *Fortschritte der Physik* **28,** 547–578.

Abarbanel, H.D.I., & Bronzan, J.B. (1974) Structure of the Pomeranchuk singularity in Reggeon field theory. *Physical Review* **D 9,** 2397–2410.

Peliti, L. (1985) Path integral approach to birth-death processes on a lattice. *J. Physique* **46,** 1469–1483.

Cardy, J. and Grassberger, P. (1985) Epidemic models and percolation. *J. Phys. A: Math. Gen.* **18,** L267–L271.

Brunel, V., Oerding, K., & Wijland, F. (2000) Fermionic field theory for directed percolation in (1+1)-dimension, *J. Phys.* **A 33,** 1085–1097.

Ma, Sh.-K. (1976) Renormalization group by Monte Carlo methods, *Phys. Rev. Lett.* **37,** 461–464.

Menéndez de la Prida, L., Stollenwerk, N., & Sánchez-Andrés, J.V. (1997) Bursting as a source for predictability in biological neural network activity, *Physica* **D 110,** 323–331.

Ferguson, N.M., Donnelly, C.A., & Anderson, R.M. (2001) The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact pf intervention, *Science* **292,** 1155–1160.

Müller, J., Schönfisch, B., & Kirkilionis, M. (2000) Ring vaccination, *Journal of Mathematical Biology* **41,** 143–171.

Janssen, H.K. (1981) On the nonequilibrium phase transition in reaction-diffusion systems with an absorbing stationary state, *Z. Phys.* **B 42,** 151–154.

Felderhof, B.U. (1971) Spin relaxation of the Ising chain. *Rep. math. Phys.* **1,** 215–234.

Index