Non-Product Data-Dependent Partitions for Mutual Information Estimation: Strong Consistency and Applications

Jorge Silva, *Member, IEEE,* and Shrikanth S. Narayanan, *Fellow, IEEE*

***Abstract*— A new framework for histogram-based mutual information estimation of probability distributions equipped with density functions in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is presented in this work. A general histogram-based estimate is proposed, considering non-product data-dependent partitions, and sufficient conditions are stipulated to guarantee a strongly consistent estimate for mutual information. Two emblematic families of density-free strongly consistent estimates are derived from this result, one based on statistically equivalent blocks (the Gessaman's partition) and the other, on a tree-structured vector quantization scheme.**

***Index Terms*— Mutual information, histogram-based estimation, data-dependent partitions, asymptotically sufficient partitions, *Vapnik-Chervonenkis* inequality, tree-structured vector quantization.**

I. Introduction

MUTUAL information (MI) specifies the level of statistical dependency between a pair of random variables [1], [2]. This quantity is fundamental to characterizing some of the most remarkable results in information theory: the performance limit for the rate of reliable communication through a noisy channel, and the achievable rate-distortion curve in lossy compression, among others [1], [2]. Mutual information has been also adopted in statistical learning-decision contexts. It has been used as a fidelity indicator, primarily because of Fano's inequality [2], finding important applications as a tool for statistical analysis [3], [4], in feature extraction [5]–[7], in detection problems [8], in image registration and segmentation [9]–[11], and recently in the characterization of performance limits on pattern recognition [12].

Typically these learning-decision applications rely on empirical data as the distributions are unknown. Hence, the problem of distribution-free MI estimation based on independent and identically distributed (i.i.d.) realizations of the involved probability measure becomes crucial, as pointed out in many of the mentioned works. The MI estimation scenario relates fundamentally with the well understood problem of distribution (density) estimation, as MI is a functional of a probability distribution. In this context strong consistency in classical L_1 sense is well known [13]. In particular for classical histogram-based estimates necessary and sufficient conditions are known for density-free estimation [13], [14]. Some extensions have been studied considering data-dependent partitions [15] and the family of histogram-based estimator proposed by Barron *et al.* [16], where sufficient conditions for L_1-consistency were stipulated. More recent work on the Barron-type of histogram-based estimator has considered consistency under topologically stronger notions, such as consistency in direct information divergence by Barron *et al.* [16] and Györfi *et al.* [17], χ^2-divergence and expected χ^2-divergence by Györfi *et al.* [18] and Vajda *et al.* [19] and the general family of Csiszár's ϕ-divergence by Beirlant *et al.* [20]. These results not only provide strong consistency for density estimation, with respect to the choice of dissimilarity measure between distributions (total variations, information divergence, Csiszár's ϕ-divergence), but also provide results characterizing rate of consistency [18]–[20] and the asymptotic distributions of normalized errors (asymptotic normality) [21].

In the context of estimating functionals of probability distributions, there is also an extensive literature dealing with the differential entropy estimation for distributions defined on a finite dimensional Euclidean space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, (see Beirlant *et al.* [22] and references therein for an excellent review). As the MI of continuous random variables can be expressed as the summation of differential entropies [2], the constructions and results derived from this estimation problem [22] extend to the MI estimation scenario. In particular, consistency results are well known for histogram-based and kernel plug-in estimates [22], [23]. Focusing on the important case of histogram-based estimation, the conventional approach requires the use of a product partition of the space, i.e., every coordinate of the space is partitioned independently to form the full partition of $\mathbb{R}^d$, and where in addition the partition is made only a function of the amount of data and not depending on how the data is distributed in the space. However, it is known that non-product data-driven partitions can approximate better the nature of the empirical data with few quantization bins and provide the flexibility to improve the approximation quality of histogram-based estimates [15], [24]. This has been shown theoretically in three emblematic non-parametric learning problems: density estimation, regression and classification [15], [25].

This fact was first observed by Darbellay and Vajda [24],

J. Silva is with the Department of Electrical Engineering, University of Chile, Av. Tupper 2007 Santiago, 412-3, Room 508, Chile, Tel: 56-2-9784090, Fax: 56-2-6953881, (email: josilva@ing.uchile.cl).

S. Narayanan is with the Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California. 3740 McClilntock Avenue, Room EEB430, Los Angeles, CA 90089 2564, USA, Tel: 213-740-6432, Fax: 213-740-4651 (email: shri@sipi.usc.edu)

The work of J. Silva was supported by funding from Fondecyt Grant 1090138, CONICYT-Chile.

The work of S. Narayanan was supported in part by grants from the Office of Naval Research (ONR) and the National Science Foundation (NSF).

who, consequently, proposed an MI estimate based on a non-product tree-structured data-dependent partition. This scheme partitions the space in statistically equivalent bins and uses a local stopping rule based on thresholding the conditional empirical MI gain obtained during an iterative bin-splitting process [24]. While this work showed promising empirical evidence of the goodness of non-product partition for MI estimation, as the authors mentioned in [24], strong consistency is a challenging and open problem for this type of construction. In fact, this lack of consistency reduces in practice to the fact that the stopping criterion has to be set empirically.

A. Contribution and Organization

The present paper provides an alternative approach for the problem of non-product data-dependent partition for MI estimation. In terms of methodology, we first study the problem of strong consistency in general terms, and we then apply these findings to the construction of specific histogram-based estimates based on non-product statistically equivalent partitions of the space. With regard to the first objective of this work, a non-product data-dependent construction is presented and a set of sufficient conditions are derived to make its induced histogram-based MI estimate strongly consistent. To achieve this, we adopt the celebrated *Vapnik-Chervonenkis* inequality [26], [27] and results concerning asymptotically sufficient partitions to control estimation and approximation errors, respectively, which are part of this learning problem [28]. For the second goal, we consider two important applications of the aforementioned consistency result, the first for the statistically equivalent block proposed by Gessaman [29], and the second for a tree-structured vector quantization (TSVQ) induced by binary statistically equivalent splits of the data [28]. In both contexts, our main result implies a range of design values where the induced estimates show the desired consistent behavior for the family of distributions in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ absolutely continuous with respect to the Lebesgue measure. Finally, some simulation scenarios are used to illustrate the advantage of proposed data-driven schemes when compared to conventional product histogram-based and kernel plug-in estimates.

This research continues our previous effort on Kullback-Leibler (KL) divergence estimation [30], [31]. However the setting and formulation of the problem here are different and address hitherto unexplored technical and practical challenges.

The paper is organized as follows. Section II provides preliminaries on some important statistical learning results that will be used in the rest of the paper. Section III introduces the problem and presents the non-product histogram-based estimator. Section IV formulates the main consistency result. Section V provides details about the two data-driven partition schemes used to estimate the MI. Finally, Section VI reports the experiments and Section VII, the final remarks.

II. Preliminaries

This work makes systematic use of the Vapnik and Chervonenkis theory [15], [27], [28], [32], which is briefly introduced in this section. We also introduce the notion of partition scheme and some standard notations for sequences.

A. Combinatorial Notions

Let us focus on the finite dimensional Euclidean space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ where $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel sigma field. Let $\mathcal{C} \subset \mathcal{B}(\mathbb{R}^d)$ be a collection of measurable events, and $x_1^n = (x_1, .., x_n)$ be a sequence of n points in $\mathbb{R}^d$. Then we define by $\mathcal{S}(\mathcal{C}, x_1^n)$ the number of different sets in

$$\{\{x_1, x_2, .., x_n\} \cap B : B \in \mathcal{C}\}, \tag{1}$$

and the *shatter coefficient* of $\mathcal{C}$ by $S_n(\mathcal{C}) = \sup_{x_1^n \in \mathbb{R}^{d\cdot n}} \mathcal{S}(\mathcal{C}, x_1^n)$ [27], [28]. The shatter coefficient is an indicator of the richness of $\mathcal{C}$ to dichotomize a finite sequence of points in the space, where by definition $S_n(\mathcal{C}) \leq 2^n$. The largest integer where $S_n(\mathcal{C})$ is strictly less than 2^n is called the Vapnik and Chervonenkis (VC) dimension of $\mathcal{C}$ [28]. If $S_n(\mathcal{C}) = 2^n$ for all n, then the class is set to have infinite VC-dimension. If $\mathcal{C}$ has a finite VC-dimension V, then the shatter coefficient is bounded by the following polynomial growth [28], [32], $\forall n > V$,

$$S_n(\mathcal{C}) \leq (1+n)^V. \tag{2}$$

Similarly, these notions can be extended to a collection of partitions [15]. Let us denote by π a finite measurable partition of $\mathbb{R}^d$ and by $|\pi|$, its cardinality. Let $\mathcal{A}$ be a collection of finite measurable partitions for $\mathbb{R}^d$, then the *maximum cell count* of $\mathcal{A}$ is given by [15]

$$\mathcal{M}(\mathcal{A}) = \sup_{\pi \in \mathcal{A}} |\pi|. \tag{3}$$

In addition, let us consider $x_1^n = (x_1, .., x_n) \in \mathbb{R}^{dn}$, then $\Delta(\mathcal{A}, x_1^n)$ denotes the number of different partitions of $\{x_1, x_2, .., x_n\}$ induced by the elements of $\mathcal{A}$ (partitions of the form $\{\{x_1, x_2, .., x_n\} \cap B : B \in \pi\}$ with $\pi \in \mathcal{A}$), where it is clear that $\Delta(\mathcal{A}, x_1^n) \leq \mathcal{M}(\mathcal{A})^n$ [15]. Analogous to the shatter coefficient, the *growth function* of $\mathcal{A}$ is given by

$$\Delta_n^*(\mathcal{A}) = \sup_{x_1^n \in \mathbb{R}^{d\cdot n}} \Delta(\mathcal{A}, x_1^n). \tag{4}$$

B. Vapnik-Chervonenkis Inequalities

Let $X_1, X_2, .., X_n$ be i.i.d. realizations of a random vector with values in $\mathbb{R}^d$, with $X \sim P$ and P a probability measure in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then for any measurable set $B \in \mathcal{B}(\mathbb{R}^d)$ the empirical distribution is given by,

$$P_n(B) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_B(X_i), \tag{5}$$

where $\mathbb{1}_B(x)$ is the indicator function of B[1].

A fundamental problem in statistical learning is being able to bound the deviation of the empirical distribution P_n with respect to P restricted to a collection of measurable events.

THEOREM 1: (Vapnik and Chervonenkis [32]) Let $X_1, X_2, ..$ be i.i.d realizations of a random variable in $\mathbb{R}^d$, with $X_i \sim P$ for all i and P a probability measure in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Let $\mathcal{C}$ be a collection of measurable events of $\mathbb{R}^d$, then $\forall n \in \mathbb{N}$, $\forall \epsilon > 0$,

$$\mathbb{P}\left(\sup_{B \in \mathcal{C}} |P_n(B) - P(B)| > \epsilon\right) \leq \mathcal{S}_n(\mathcal{C}) \cdot \exp^{-\frac{n\epsilon^2}{8}}, \tag{6}$$

[1] $\mathbb{1}_B(x)$ is one if $x \in B$ and zero otherwise.

where $\mathbb{P}$ denotes the distribution of the process $\{X_1, X_2, \cdots\}$. This is the celebrated *Vapnik and Chervonenkis inequality*, a distribution free inequality that uniformly bounds the deviation of P_n with respect to P in the events of $\mathcal{C}$. Notably the right hand side (RHS) of (6) is distribution-free and, furthermore, a function of the shatter coefficient of $\mathcal{C}$. Lugosi *et al.* [15] extended this concentration inequality for a collection of measurable partitions. In this case the empirical measure is restricted to a partition $\pi \subset \mathcal{B}(\mathbb{R}^d)$ where the total variational distance [28] is used to quantify the deviation between P and P_n restricted to the measurable space $(\mathbb{R}^d, \sigma(\pi))$ [2]. The following lemma states this result formally.

LEMMA 1: (Lugosi and Nobel [15]) Under the learning setting of Theorem 1, let $\mathcal{A}$ be a collection of finite measurable partitions for $\mathbb{R}^d$. Then $\forall n \in \mathbb{N}$, $\forall \epsilon > 0$,

$$\mathbb{P}\left(\sup_{\pi\in\mathcal{A}} \sum_{B\in\pi} |P_n(B) - P(B)| > \epsilon\right) \leq$$
$$4\Delta^*_{2n}(\mathcal{A})2^{\mathcal{M}(\mathcal{A})} \exp^{-\frac{n\epsilon^2}{32}}, \tag{7}$$

with $\mathbb{P}$ the process distribution of $\{X_1, X_2, \cdots\}$.

C. Data-Dependent Partitions

A *n-sample partition rule* $\pi_n(\cdot)$ is a mapping from $\mathbb{R}^{dn}$ to the space of finite-measurable partitions for $\mathbb{R}^d$, that we denote by $\mathcal{A}(\mathbb{R}^d)$, where a *partition scheme* for $\mathbb{R}^d$ is a countable collection of n-sample partition rules $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), ...\}$. Let Π be an arbitrary partition scheme for $\mathbb{R}^d$, then for every partition rule $\pi_n(\cdot) \in \Pi$ with $n \in \mathbb{N}$, we can define its associated collection of measurable partitions by [15]

$$\mathcal{A}_n = \left\{\pi_n(x_1, .., x_n) : (x_1, .., x_n) \in \mathbb{R}^{dn}\right\} \subset \mathcal{A}(\mathbb{R}^d). \tag{8}$$

Here, for a given n-sample partition rule $\pi_n(\cdot)$ and a sequence $(x_1, .., x_n) \in \mathbb{R}^{dn}$, $\pi_n(x|x_1, .., x_n)$ denotes the mapping from any point x in $\mathbb{R}^d$ to its unique cell in $\pi_n(x_1, .., x_n)$, such that $x \in \pi_n(x|x_1, .., x_n)$, $\forall x \in \mathbb{R}^d$.

D. Asymptotic Relationships for Sequences

Let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be two sequences of non-negative real numbers, we say that $(a_n)_{n\in\mathbb{N}}$ dominates $(b_n)_{n\in\mathbb{N}}$, denoted by $(b_n) \preceq (a_n)$ (or alternatively, (b_n) is $O(a_n)$), if there exists $C > 0$ and $k \in \mathbb{N}$ such that $b_n \leq C \cdot a_n$ for all $n \geq k$. We say that $(b_n)_{n\in\mathbb{N}}$ and $(a_n)_{n\in\mathbb{N}}$ are asymptotically equivalent, denoted by $(b_n) \approx (a_n)$, if there exists $C > 0$ such that $\lim_{n\to\infty} \frac{a_n}{b_n} = C$. A nonnegative sequence $(a_n)_{n\in\mathbb{N}}$ is $o(f(n))$, for some non-negative increasing function $f(\cdot) : \mathbb{N} \to \mathbb{R}$, if $\lim_{n\to\infty} \frac{a_n}{f(n)} = 0$.

III. The Mutual Information Estimate

Let X and Y be two random variables taking values on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with a joint distribution denoted by $P_{X,Y}$. We focus on the finite dimensional Euclidean space, i.e., $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$ and consequently $P_{X,Y}$ is defined on the Borel sigma field $\mathcal{B}(\mathbb{R}^d)$ for $d = p+q$. In this case the MI between X and Y can be expressed by [2]

$$I(X;Y) = D(P_{X,Y}||P_X \times P_Y), \tag{9}$$

where $P_X \times P_Y$ is the probability distribution on $\mathbb{R}^d$ induced by multiplication of the marginals of X and Y (joint probability where X and Y are independent) and $D(P||Q)$ denotes the *Kullback-Leibler (KL) divergence* given by [1], [33]

$$D(P||Q) = \sup_{\pi\in\mathcal{A}(\mathbb{R}^d)} \sum_{B\in\pi} P(B) \cdot \log \frac{P(B)}{Q(B)}, \tag{10}$$

with $\mathcal{A}(\mathbb{R}^d)$ being the collection of finite measurable partitions of $\mathbb{R}^d$.

We are interested in the problem of estimating $I(X;Y)$ based on $(X_1, Y_1), \cdots (X_n, Y_n)$, i.i.d. realizations of the joint distribution $P_{X,Y}$. To simplify the notation we denote by Z_i the joint vector (X_i, Y_i) on $\mathbb{R}^d$ and by Z_1^k the sequence of realizations $(Z_1, .., Z_k)$ on $\mathbb{R}^{dk}$. In particular, in this work we focus on the histogram-based approach [15], [28] based on a partition scheme, Section II-C.

A. The Histogram-Based Estimator

Let $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), \cdots\}$ be a partition scheme, where we impose the condition that every bin induced by this family of partition rules has a *product form*, i.e., $\forall z_1^n = (z_1, .., z_n) \in \mathbb{R}^{dn}$, every measurable set $B \in \pi_n(z_1^n)$ can be expressed in the following Cartesian product form,

$$B = B_1 \times B_2, \tag{11}$$

where $B_1 \in \mathcal{B}(\mathbb{R}^p)$ and $B_2 \in \mathcal{B}(\mathbb{R}^q)$.

Let $Z_1, .., Z_n$ be i.i.d. realizations with probability distribution $P_{X,Y}$. To simplify notation, we denote by P, the joint distribution and by P_n, its empirical version given in (5). Then, the proposed MI estimate is given by

$$\hat{I}_n(X;Y) =$$
$$\sum_{B\in\pi_n(Z_1^n)} P_n(B) \cdot \log \frac{P_n(B)}{P_n(B_1 \times \mathbb{R}^q) \cdot P_n(\mathbb{R}^p \times B_2)}, \tag{12}$$

where $B_1 \times B_2$ denotes the product form of the event $B \in \pi_n(Z_1^n)$. Note that this construction involves three steps: first, obtain a random partition from the data $\pi_n(Z_1^n)$, second, estimate the empirical distributions restricted to the events in $\sigma(\pi_n(Z_1^n))$, and third, plug the empirical distributions in the finite alphabet version of the KL divergence in (12).

Remark 1: The condition in (11) does not imply that the partition $\pi_n(z_1^n)$ has a product structure, i.e., it can be written by $Q_1 \times Q_2$, with Q_1 and Q_2 being individual measurable partitions of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively.

Remark 2: The product bin condition in (11) is strictly necessary for being able to estimate $P_{X,Y}$ as well as the reference measure $P_X \times P_Y$ only based on the i.i.d. realizations of $P_{X,Y}$.

As pointed out in [24], this kind of estimate is not formally a MI quantity, in other words it is not the MI between quantized versions of the two random variables X and Y (this requires a product structure of the partition $\pi_n(Z_1^n)$). Instead,

[2] $\sigma(\pi)$ denotes the smallest sigma-field that contains π, which for the case of a partition π is the collection of sets that can be written as unions of events in π.

as considered by Darbellay *et al.* [24], the proposed empirical construction $\hat{I}_n(X;Y)$ is the KL divergence between the empirical joint distribution and its empirical product counterpart (multiplication of marginals empirical distributions), restricted to the sub-sigma field $\sigma(\pi_n(Z_1^n))$ [1]. This is motivated by (9) and (10).

The main challenge is to find distribution-free conditions on the partition scheme Π that guarantee the MI estimate in (12) to be strongly consistent with respect to $\mathbb{P}$, the distribution of the empirical process $\{Z_1, Z_2, \ldots\}$. The answer to this question is formally addressed in the next section.

IV. Strong Consistency

The difference between our estimator in (12) and $I(X;Y)$ can be bounded by the following two terms,

$$\left|\hat{I}_n(X;Y) - I(X;Y)\right| \leq \left|\sum_{B\in\pi_n(Z_1^n)} P_{X,Y}(B)\cdot\log\frac{P_{X,Y}(B)}{P_X\times P_Y(B)} - I(X;Y)\right| + \left|\hat{I}_n(X;Y) - \sum_{B\in\pi_n(Z_1^n)} P_{X,Y}(B)\cdot\log\frac{P_{X,Y}(B)}{P_X\times P_Y(B)}\right|. \quad (13)$$

The first term in the upper bound is *the approximation error* (or bias of the estimate), which only considers the effect of quantizing the space — it is well known that quantization reduces the magnitude of information theoretic quantities [1], [24], [34]. The second term in (13) is the *estimation error* (or the variance term) that quantifies the deviation between the empirical and true distribution in the finite alphabet MI functional. A natural direction is to find a good compromise between these sources of error as a function of the structural properties of the data-dependent partition scheme. Specifically, the objective is to make the two errors vanish asymptotically with probability one with respect to $\mathbb{P}$. We first deal with the approximation error to later integrate this analysis with the estimation error in the main theorem of this work.

A. Controlling the Approximation Error

For a measurable event $B \in \mathcal{B}(\mathbb{R}^d)$, the diameter of the set is given by,

$$\text{diam}(B) = \sup_{x,y\in B} ||x-y||, \quad (14)$$

where $||\cdot||$ denotes the Euclidian norm in $\mathbb{R}^d$.

THEOREM 2: Let $P_{X,Y}$ be a probability measure absolutely continuous with respect to the Lebesgue measure λ in $\mathbb{R}^d$ and let $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), ...\}$ be a partition scheme driven by $Z_1, Z_2, \cdots$, i.i.d. realizations with $Z_i \sim P_{X,Y}$ for all i. If $\forall\delta > 0$,

$$\lim_{n\to\infty} P_{X,Y}\left(\left\{z\in\mathbb{R}^d : \text{diam}(\pi_n(z|Z_1^n)) > \delta\right\}\right) \to 0, \quad (15)$$

$\mathbb{P}$ (the process distribution of $\{Z_1, Z_2, \cdots\}$) almost surely (a.s.), then,

$$\lim_{n\to\infty} \sum_{B\in\pi_n(Z_1^n)} P_{X,Y}(B)\cdot\log\frac{P_{X,Y}(B)}{P_X\times P_Y(B)} = I(X;Y), \quad (16)$$

$\mathbb{P}$-a.s. (The proof of this result is presented in Appendix I.)

The result says that if the diameter of the random partition $\pi_n(Z_1^n)$ vanishes in a probabilistic sense as the number of samples tends to infinity, as given by (15), we can approximate with arbitrary precision the distributions for the purpose of estimating the MI (or equivalently to say that Π is $\mathbb{P}$-almost surely asymptotically sufficient for $I(X;Y)$). This approximation property in (15) is called a *shrinking cell condition*. Different flavors of this notion have been introduced for controlling approximation error in histogram-based regression, classification and density estimation problems [15], [25], [28], [35]. A similar shrinking cell condition was presented in [30] for the problem of KL divergence approximation. In fact as part of the proof of Theorem 2, a stronger result for the KL divergence scenario is presented (we refer the reader to Theorem 7 in Appendix I for details).

B. The Result

Before stating the result, we introduce some definitions. For any partition rule $\pi_n(\cdot) \in \Pi$ and $z_1^n \in \mathbb{R}^{dn}$, we consider its product bin structure in (11) to define the following collections of coordinated-projected sets, $\mathcal{C}_{[1,p]}(z_1^n) \equiv \left\{\xi_{[1,p]}(B) : B \in \pi_n(z_1^n)\right\}$, $\mathcal{C}_{[p+1,d]}(z_1^n) \equiv \left\{\xi_{[p+1,d]}(B) : B\in\pi_n(z_1^n)\right\}$, with $1 \leq p < d$ and $\xi_{[1,p]}(B)$ denoting the set operator that returns the collection of projected elements of B in the range of coordinate dimensions $\{1,..,p\}$ [3]. Then in addition to $\mathcal{A}_n = \left\{\pi_n(z_1^n) : z_1^n \in \mathbb{R}^{d\cdot n}\right\} \subset \mathcal{A}(\mathbb{R}^d)$, the following collections of measurable sets will be associated with the partition rule $\pi_n(\cdot)$:

$$\mathcal{C}_{[1,p],n} \equiv \bigcup_{z_1^n\in\mathbb{R}^{d\cdot n}} \mathcal{C}_{[1,p]}(z_1^n) \subset \mathcal{A}(\mathbb{R}^p) \quad (17)$$

$$\mathcal{C}_{[p+1,d],n} \equiv \bigcup_{z_1^n\in\mathbb{R}^{d\cdot n}} \mathcal{C}_{[p+1,d]}(z_1^n) \subset \mathcal{A}(\mathbb{R}^q). \quad (18)$$

Finally we have all the elements to state the main consistency result.

THEOREM 3: Let X and Y be random variables in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, with joint distribution $P_{X,Y}$ absolutely continuous with respect to the Lebesgue measure λ in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. In addition, let us consider a partition scheme $\Pi = \{\pi_1(\cdot), \ldots\}$ with the product bin structure and driven by i.i.d. realizations $Z_1, Z_2, \ldots$ with $Z_i \sim P_{X,Y}$ for all i. If there exists $\tau \in (0,1)$ for which the following set of conditions are satisfied:

- **c.1:** $\lim_{n\to\infty}\frac{1}{n^\tau}\log\mathcal{S}_n(\mathcal{C}_{[1,p],n}) = 0$, $\lim_{n\to\infty}\frac{1}{n^\tau}\log\mathcal{S}_n(\mathcal{C}_{[p+1,d],n}) = 0$,
- **c.2:** $\lim_{n\to\infty}\frac{1}{n^\tau}\log\Delta_n^*(\mathcal{A}_n) = 0$,
- **c.3:** $\lim_{n\to\infty}\frac{1}{n^\tau}\mathcal{M}(\mathcal{A}_n) = 0$,
- **c.4:** $\exists\ (k_n)_{n\in\mathbb{N}}$ a sequence of non-negative numbers, with $(k_n) \approx (n^{0.5+\tau/2})$, such that, $\forall n > 0$ and $\forall(z_1,..,z_n)\in\mathbb{R}^{dn}$,

$$\inf_{B\in\pi(z_1^n)} P_n(B) \geq \frac{k_n}{n},$$

[3] By construction any set $B \in \pi_n(z_1^n)$ can be expressed by $B = B_1 \times B_2$, with $B_1 \in \mathbb{R}^p$ and $B_2 \in \mathbb{R}^q$, and consequently $\xi_{[1,p]}(B) = B_1$ and $\xi_{[p+1,d]}(B) = B_2$.

c.5: and $\forall \delta > 0$,

$$\lim_{n\to\infty} P_{X,Y}\left(\left\{z \in \mathbb{R}^d : \mathrm{diam}(\pi_n(z|Z_1^n)) > \delta\right\}\right) \to 0,$$

$\mathbb{P}$-a.s.,

then,

$$\lim_{n\to\infty} \hat{I}_n(X;Y) = I(X;Y), \quad \mathbb{P}-a.s. \tag{19}$$

(The proof is presented in Appendix II.)

Interpreting the result, the first four conditions are stipulated to asymptotically control the estimation error quantity in (13). They impose asymptotic bounds on the combinatorial complexity of the family of sets induced by $\pi_n(\cdot)$ (**c.1**, **c.2** and **c.3**), as well as in the number of sample points that every bin of the resulting data-driven partition $\pi_n(Z_1^n)$ should have (**c.4**). The argument used to make this error vanish is based on the *Vapnik-Chervonekis* (VC) inequalities and the *Borel-Cantelli* lemma [36]. Concerning the approximation error, **c.5** just invokes the sufficient condition presented in Theorem 2.

Remark 3: From the domain of values stipulated for τ, these conditions are stronger than the one obtained for the problem of density estimation consistent in the L_1 sense [15]. These stronger conditions are necessary to handle the unbounded behavior of the $\log(\cdot)$ function in the neighborhood of zero — the function is not absolutely continuous in $(0, \infty)$, which is the most critical part to guarantee a strongly consistent result for the MI estimation problem.

Considering the condition **c.1** of Theorem 3, if a collection of measurable events $\mathcal{C}$ has a finite VC dimension, let us say $V > 0$, then $\forall n > V$, $\mathcal{S}_n(\mathcal{C}) \leq (n+1)^V$ [28], [37], and consequently $\forall \tau \in (0,1)$, $\lim_{\to\infty} \frac{1}{n^\tau} \log \mathcal{S}_n(\mathcal{C}) = 0$. It is interesting, however, to extend this idea to a sequence of measurable events. The following proposition guarantees **c.1** of Theorem 3 when a collections of sets have finite VC-dimensions.

PROPOSITION 1: Let $\{\mathcal{C}_n : n \in \mathbb{N}\}$ be a collection of measurable events with finite VC dimension sequence $(V_n)_{n\in\mathbb{N}}$. If $(V_n)_{n\in\mathbb{N}}$ is $o\left(\frac{n^{\tau_o}}{\log(n+1)}\right)$ for some $\tau_o \in (0,1)$, then

$$\lim_{n\to\infty} \frac{1}{n^{\tau_o}} \log \mathcal{S}_n(\mathcal{C}_n) = 0. \tag{20}$$

Proof: From the fact that $(V_n)_{n\in\mathbb{N}}$ is $o\left(\frac{n^{\tau_o}}{\log(n+1)}\right)$, there exists N such that $\forall n > N$, $n > V_n$. Then from the definition of the VC dimension [28], $\forall n > N$, $\mathcal{S}_n(\mathcal{C}_n) \leq (n+1)^{V_n}$. Then $\limsup_{n\to\infty} \frac{1}{n^{\tau_o}} \log \mathcal{S}_n(\mathcal{C}_n) \leq \lim_{n\to\infty} \frac{V_n \log(n+1)}{n^{\tau_o}} = 0$ by the hypothesis. □

In conclusion, Theorem 3 provides a set of sufficient conditions for strong consistency of the histogram-based estimator in (12). However at this point a valid question to ask is how this result translates into specific design conditions when working with some specific data-dependent partition schemes. In other words, is it possible to find a class of strongly consistent MI estimates based on Theorem 3? This will be the focus for the rest of the paper, where we show how these general conditions provide specific design setting in the implementation of two widely adopted non-product partition schemes.

V. Applications

A. Statistical Equivalent Data-Dependent Partitions

Here we consider a data-dependent partition scheme based on the notion of statistically equivalent blocks [28], and more precisely the axis-parallel scheme proposed by Gessaman [29]. The idea is to use the data $Z_1, .., Z_n$ to partition the space in such a way to create cells with equal empirical mass. In Gessaman's approach, this is done by sequentially splitting every coordinate of $\mathbb{R}^d$ using axis-parallel hyperplanes. More precisely, let $l_n > 0$ denote the number of samples points that we ideally want to have in every bin of $\pi_n(Z_1^n)$, and let us choose a particular sequential order for the axis-coordinates, such as the standard $(1, .., d)$. With that, $T_n = \lfloor (n/l_n)^{1/d} \rfloor$ is the number of partitions to create in every coordinate. Then the inductive construction goes as follows: first, project the i.i.d. samples $Z_1, .., Z_n$ into the first coordinate, which for simplicity we denote by $Y_1, .., Y_n$. Compute the order statistics $Y^{(1)}, Y^{(2)}, .., Y^{(n)}$ or the permutation of $Y_1, .., Y_n$ such that $Y^{(1)} < Y^{(2)} < \cdots < Y^{(n)}$ — this permutation exists with probability one if $P_{X,Y}$ is absolutely continuous with respect to the Lebesgue measure [28]. Based on this, the following set of intervals to partition the real line is induced,

$$\{I_i : i = 1, .., T_n\} = \left\{(-\infty, Y^{(s_n)}], (Y^{(s_n)}, Y^{(2\cdot s_n)}], .., (Y^{((T_n-1)\cdot s_m)}, \infty)\right\}, \tag{21}$$

where $s_n = \lfloor n/T_n \rfloor$. Then assigning the samples of $Z_1, .., Z_n$ to the different resulting bins, i.e., $\left\{I_i \times \mathbb{R}^{d-1} : i = 1, .., T_n\right\}$, we can conduct the same process in each of those bins by projecting its data into the second coordinate. Iterating this approach until the last coordinate we get the Gessaman data-dependent partition $\pi_n(Z_1^n)$. Note that by construction if $n = (l_n)^d$, then we are in the ideal scenario where every bin has been assigned with l_n empirical points of $Z_1, .., Z_n$. Importantly for our consistency result, $P_n(A) \geq \frac{l_n}{n}$, $\forall A \in \pi_n(Z_1^n)$. Consequently, we can use this product-bin construction for estimating the MI based on (12). The following result applies Theorem 3 to this scenario.

THEOREM 4: Under the general hypothesis presented in Theorem 3, the Gessaman's partition scheme provides a strongly consistent estimate for $I(X,Y)$, if $(l_n) \approx (n^{0.5+\tau/2})$ for some $\tau \in (1/3, 1)$.

The proof of this result reduces to checking the sufficient condition stated in Theorem 3. Note that the result does not impose any condition on the joint distribution $P_{X,Y}$ and provides a family of density-free strongly consistent MI estimates.

Proof: Let us consider an arbitrary $\tau \in (1/3, 1)$. The trivial case to check is **c.4)**, because by construction we can consider $k_n = l_n$, $\forall n \in \mathbb{N}$, where the hypothesis of the theorem gives the result. For **c.1)**, from the construction of $\pi_n(\cdot)$, it is noted that $\mathcal{C}_{[1,p],n}$ and $\mathcal{C}_{[p+1,d],n}$ are contained in the collection of rectangles of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, which are well known to have finite VC dimensions [37]. Hence from Proposition 1 we get the result. Concerning **c.3)**, again by construction we have that $\mathcal{M}(\mathcal{A}_n) \leq n/l_n + 1$, then

$n^{-l}\mathcal{M}(\mathcal{A}_n) \leq n^{1-\tau}/l_n + n^{-\tau}$. Given that $(l_n) \approx (n^{0.5+\tau/2})$ and $\tau \in (1/3, 1)$ it follows that, $\lim_{n\to\infty} n^{-\tau}\mathcal{M}(\mathcal{A}_n) = 0$. For **c.2)**, Lugosi *et al.* [15] showed that $\Delta^*_n(\mathcal{A}_n) \leq \binom{T_n+n}{n}^d$, where using that $\log\binom{s}{t} \leq s \cdot h(t/s)$ [28], with $h(x) = -x\log(x) - (1-x)\log(1-x)$ for $x \in [0,1]$ the binary entropy function [2], and defining $\bar{T}_n \equiv \lfloor n/l_n \rfloor \geq T_n$, it follows that,

$$\begin{aligned} n^{-\tau}\log\left(\Delta^*_n(\mathcal{A}_n)\right) &\leq n^{-\tau} d \cdot \log\tbinom{\bar{T}_n+n}{n} \\ &\leq 2dn^{1-\tau} \cdot h\left(\frac{1}{n/\bar{T}_n + 1}\right) \leq 2dn^{1-\tau} \cdot h\left(\frac{1}{l_n}\right). \end{aligned}$$

Consequently we have that, $\forall n \in \mathbb{N}$,

$$\begin{aligned} n^{-\tau}\log(\Delta^*_n(\mathcal{A}_n)) \leq -\frac{2dn^{1-\tau}}{l_n}\log(1/l_n) \\ - 2dn^{1-\tau}(1-1/l_n)\log(1-1/l_n). \quad (22) \end{aligned}$$

The first term on the right hand side (RHS) of (22) behaves like $n^{0.5-3/2\cdot\tau} \cdot \log(l_n)$, where as long as the exponent of the first term is negative (equivalent to $\tau > 1/3$) this sequence tends to zero as n tends to infinity — considering that by construction $(l_n) \preceq (n)$. The second term on the RHS of (22) behaves asymptotically like $-n^{1-\tau} \cdot \log(1 - 1/l_n)$ which is upper bounded by the sequence $\frac{n^{1-\tau}}{l_n} \cdot \frac{1}{1-1/l_n}$ — using that $\log(x) \leq x - 1$, $\forall x > 0$. This upper bound tends to zero because $(l_n) \approx (n^{0.5+\tau/2})$ and $\tau > 1/3$. Consequently from (22), $\lim_{n\to\infty} n^{-\tau}\log(\Delta^*_n(\mathcal{A}_n)) = 0$. Finally concerning **c.5)**, Lugosi *et al.* [15] (*Theorem 4*) proved that to get this shrinking cell condition is sufficient to show that (l_n) is o(n), which is the case considering that $\tau < 1$. □

Remark 4: The condition of Theorem 4 implies that $\lim_{n\to\infty} l_n = \infty$ and (l_n) is $o(n)$, which are the necessary and sufficient conditions for the Gessaman histogram based density estimate to be strongly consistent in $\mathbf{L}_1$, by Abou-Jaoude [38]. The fact that stronger conditions are needed to get consistency in the MI functional agrees with findings on density-free estimation of differential entropy [23] (using classical product partition), and with the new results on the convergence analysis of the differential entropy by Piera and Parada [39].

B. Tree-Structured Partition Schemes

In this section we consider a version of what is known as *balanced search tree* [28](*Chapter 20.3*), in particular the binary case. More precisely, given $Z_1, Z_2, .., Z_n$ i.i.d. realizations of the joint distribution, this data-dependent partition chooses a dimension of the space in a sequential order, say the dimension i for the first step, and then the i axis-parallel halfspace by

$$H_i(Z_1^n) = \left\{x \in \mathbb{R}^d : x(i) \leq Z^{(\lceil n/2 \rceil)}(i)\right\}, \quad (23)$$

where $Z^{(1)}(i) < Z^{(2)}(i) <,..,< Z^{(n)}(i)$ denotes the order statistics of the sampling points $\{Z_1, .., Z_n\}$ projected in the target dimension i. Using this hyper-plane, $\mathbb{R}^d$ is divided into two statistically equivalent rectangles with respect to the coordinate dimension i, denoted by $U_{(1,0)}$ and $U_{(1,1)}$. Reallocating the sampling points in the respective intermediate cells, $U_{(1,0)}$ and $U_{(1,1)}$, we can choose a new dimension in the mentioned sequential order and continue in an inductive fashion with this splitting process. In particular in the iteration k of the algorithm (assuming that the stopping rule is not violated) the intermediate rectangles $U_{(k-1,l)}$ for $l \in \{0, .., 2^{k-1}-1\}$ are partitioned in terms of their respective statistically equivalent k-axis parallel hyper-planes to create $\{U_{(k,2l)}, U_{(k,2l+1)} : l = 0, .., 2^{k-1}-1\}$. The termination criterion is based on a stopping rule that guarantees a minimum number of sample points per cell, denoted by $k_n > 0$. This stopping rule is fundamental to obtaining our consistency result. After the first iteration, the resulting cells have at most $n/2 + 1$ and at least $n/2 - 1$ sampling points. The second iteration implies the creation of 4 cells with at most $n/4+2$ and at least $n/4 - 2$ sampled points, and consequently inductively the k-th iteration — if the stopping criterion is not violated — creates a balanced tree of 2^k cells with at least $n/2^k - k$ and at most $n/2^k + k$ sampling points. Note that at the end of the process it is not guaranteed that $\pi_n(Z_1^n)$ has either perfect statistically equivalent cells (rectangles with equal empirical mass) or a balanced tree structure.

THEOREM 5: Let us consider the tree-structure partition (TSP) with binary axis-parallel statistically equivalent splits and a stopping rule governed by a sequence of non-negative numbers $(k_n)_{n\in\mathbb{N}}$. Under the general hypothesis of Theorem 3, if $(k_n) \approx (n^{0.5+\tau/2})$ for some $\tau \in (1/3, 1)$ the MI estimator $\hat{I}_n(X;Y)$ induced by (12) is strongly consistent.

Proof: We check the sufficient conditions of Theorem 3. First **c.1)** is guaranteed by the same reasons stated for the Gessaman's partition scheme in Section V-A, where **c.4)** is obtained by the hypotheses of the theorem. Considering **c.3)**, $|\pi_n(Z_1^n)|$ is uniformly upper bounded by n/k_n. Then it follows that $\mathcal{M}(\mathcal{A}_n) \leq n/k_n$ and consequently $n^{-\tau}\mathcal{M}(\mathcal{A}_n) \leq \frac{n^{1-\tau}}{k_n} \approx n^{0.5-\frac{3}{2}\tau}$. This upper bound tends to zero as $n \to \infty$ given that $\tau > 1/3$. For **c.2)**, we use the upper bound proposed by Lugosi *et al.* [15], specifying that every polytope (or cell) of $\pi_n(Z_1^n)$ is induced by at most $\mathcal{M}(\mathcal{A}_n)$ hyperplane splits. Each binary split can dichotomize $n \geq 2$ points in $\mathbb{R}^d$ in at most n^d ways [40]. Consequently we have that $\Delta^*_n(\mathcal{A}_n) \leq (n^d)^{n/k_n}$, and then $n^{-\tau}\log\Delta^*_n(\mathcal{A}_n) \leq \frac{n^{1-\tau}}{k_n} d\log n$. Again this bound tends to zero as $n \to \infty$ because $\tau > 1/3$. The final condition **c.5)** is the most technically challenging. To prove this shrinking cell condition we need to introduce some notations, definitions and a preliminary result.

We represent $\pi_n(Z_1^n)$ by the collection of pairs (k, l), or nodes, obtained during the iterative construction of $\pi_n(Z_1^n)$ (the hyper-plane splitting process). Adopting Breiman *et al.* conventions [35], this collection of nodes represents a rooted binary-tree[4], denoted by T_n, where the direct decedents of a non-terminal node (k, l) are $(k+1, 2l)$ and $(k+1, 2l+1)$. The set of terminal nodes (the pairs (k, l) with no direct decedents), denoted by $\mathcal{L}(T_n)$, indexes the partition by the following relationship, $\pi_n(Z_1^n) = \{U_{(k,l)} : (k,l) \in \mathcal{L}(T_n)\}$. Then, the TSP scheme Π can be indexed and represented by $\{T_1, T_2, \cdots\}$. In general, $(0,0)$ denotes the root of any of our

[4] A *binary tree* T is a collection of nodes with only one with degree 2 (the *root* node), and the remaining nodes with degree 3 (*internal* nodes) or degree 1 (*leaf* or *terminal* nodes) [35]. Note that the arcs are implicit in this convention.

binary trees $\{T_1, T_2, \cdots\}$.

If $\bar{T} \subset T$ and $\bar{T}$ is a binary tree by itself, we say that $\bar{T}$ is a *subtree* of T and moreover if both have the same root we say that $\bar{T}$ is a *pruned* version of T, denoted by $\bar{T} \ll T$. In particular in our construction, if we consider $\bar{T}_n \ll T_n$, then it is simple to show that $\pi_n(Z_1^n)$ is a refinement of $\pi_{\bar{T}_n}(Z_1^n) \equiv \left\{U_t : t \in \mathcal{L}(\bar{T}_n)\right\}$. Let T be a binary tree. For all $t \in T$ let $depth(t)$ denote the *depth* of t — the number of arcs that connect t with the root of T. In this context, let T^r denote the truncated version of T, formally given by $T^r = \{t \in T : depth(t) \leq r\}$, where by construction $T^r \ll T$.

Definition 1: Let T be a binary tree, we say that T is a *balanced tree* of height r if $\forall t \in \mathcal{L}(T)$, $depth(t) = r$.

Definition 2: A TSP scheme $\Pi = \{T_1, T_2, \cdots\}$ is a *uniform balanced tree-structured scheme* (UBTSS), if each partition rule, represented by T_n, forms a balanced tree of height d_n (only function of n).

In the context of UBTSS we have the following result.

THEOREM 6: Let $\Pi = \{T_1, T_2, \cdots\}$ be a UBTSS induced by the statistically equivalent splitting process presented in Section V-B. Let $(d_n)_{n \in \mathbb{N}}$ denote its height sequence, then Π satisfies the shrinking cell condition of Theorem 2, if there exists a non-negative real sequence $(q_n) \approx (n^\theta)$, for some $\theta > 0$, such that

$$\frac{n}{d_n 2^{d_n}} - \frac{q_n}{d_n} \to \infty \; and \; d_n \to \infty, \text{ as } n \text{ tends to infinity.}$$

Theorem 6 derives from the ideas presented by Devroye *et al.* (Theorem 20.2) [28], where a weak version of our shrinking cell condition was proved for a similar balanced tree-structured partition. The proof was first derived in the context of KLD estimation in [41] (Chapter 4, Lemma 4.3). For sake of completeness the argument is presented in Appendix III.

Returning to the proof of Theorem 5, by the binary tree structure of Π and the stopping rule, it is simple to show that, $\forall z_1^n \in \mathbb{R}^{d \cdot n}$,

$$r(n) \equiv \lfloor \log_2(n) \rfloor - \lceil \log_2(k_n) \rceil \leq \min_{t \in \mathcal{L}(T_n(z_1^n))} depth(t), \quad (24)$$

and consequently $T_n^{r(n)}$ is a balanced tree. Defining $\bar{\Pi} = \left\{T_1^{r(1)}, T_2^{r(2)}, \cdots\right\}$, it suffices to check the shrinking cell condition on $\bar{\Pi}$ [5]. Given that $\bar{\Pi}$ is a UBTSS, we can check the sufficient condition of Theorem 6. Let $\bar{d}_n (= r(n))$ denote the height of $T_n^{r(n)}$. By construction $\bar{d}_n \geq \log_2(n/k_n) - 2$ and consequently tends to infinity ($(k_n) \approx (n^{0.5+\tau/2})$ with $\tau < 1$). On the other hand, if we consider an arbitrary non-negative sequence $(q_n) \approx (n^\theta)$ with $\theta \in \left(0, \frac{2}{3}\right]$, then

$$\frac{n}{\bar{d}_n 2^{\bar{d}_n}} - \frac{q_n}{\bar{d}_n} \geq \frac{n}{d_n \cdot 2^{\log_2(n/k_n)}} - \frac{q_n}{d_n} = \frac{k_n - q_n}{d_n} \to \infty \quad (25)$$

as $n \to \infty$, because $(d_n) \preceq (\log_2(n))$, $(k_n) \approx (n^{0.5+\tau/2})$ and by hypothesis $\tau > 1/3$, which proves the result.

□

VI. Simulations

A classical product histogram-based estimate [23] and a kernel plug-in estimate are evaluated and contrasted with the histogram-based constructions presented in Section V. These two techniques are strongly consistent for the differential entropy estimation [22], [23] and, consequently, for the MI under the absolutely continuous assumption studied in this work. Following the experimental setting in [24], we consider X and Y to be scalar Gaussian random variables with correlation coefficient denoted by r. We simulate 1000 i.i.d. realizations of the joint Gaussian distribution with different correlation coefficients, $\{0, 0.3, 0.8\}$. Before performing comparisons, we evaluate all the methods with respect to their respective design variables (τ for the asymptotic sub-linear rate of (l_n) and (k_n), the window of the kernel and the length of the rectangle intervals generated by the product histogram, respectively), restricted to the range of those variables that makes these techniques strongly consistent. Then, for each technique we have chosen a design value that demonstrates good empirical performance among all the correlation scenarios explored in our experiments. In particular for our data-driven techniques we chose τ in $(0.4, 0.6)$.

Tables I provides the performance comparison among the different techniques by evaluating performances across the sampling lengths, $n \in \{33, 179, 564, 3164, 5626\}$ (uniformly spaced in log domain) for all the correlation scenarios. These results show that under different levels of statistical dependency between X and Y our non-product histogram-based constructions (Gessaman and TSVQ) present performance improvements compared to the two classical techniques. In particular, there is a clear bias difference in the sample regime $[1 - 10^4]$, with respect to classical product histogram approaches, supporting our conjecture that data-dependent partitions can improve the performance of classical product histogram based constructions in the small sample regime. These techniques also perform better than the kernel plug-in estimate, particularly clear in the sampling range $[1 - 200]$ and across all correlation scenarios.

Finally, similar performance trends were observed simulating Gaussian vectors in higher dimensions in a number of settings (in terms of the structure of the covariance matrix and the parameter of the techniques). Just to illustrate, two specific cases are presented in Table II. For these, the covariance matrixes were chosen with a pairwise coordinate dependency of the form [1.04 0 0.4 0; 0 1.04 0 0.4; 0.4 0 1.04 0; 0 0.4 0 1.04] for the dimension $d = 4$ and similar pairwise dependency for $d = 6$. Performances are reported under the parameter choice of the four estimation techniques adopted previously. These results show the advantage of the two data-driven techniques with respect to the competitive methods, which is congruent with what has been observed in the scalar scenario, in Table I.

A. Computational Complexity

We conclude this study with a comparison of the computational complexity of the four aforementioned methods. We start with the tree-structured partition scheme of Section V-B,

[5] Π is a refinement of $\bar{\Pi}$ in the sense that $\forall n \in \mathbb{N}$, $\forall z_1^n \in \mathbb{R}^{d \cdot n}$, $T_n^{r(n)}(z_1^n) \ll T_n(z_1^n)$, then by definition (15) the shrinking cell condition of $\bar{\Pi}$ implies the property for Π.

	33	179	564	3164	5626
TSVQ:	1.3e-02 (1.7e-03)	3.0e-03 (2.1e-04)	1.7e-03 (6.0e-05)	2.7e-04 (4.2e-06)	8.5e-05 (1.5e-06)
GESS:	2.5e-02 (2.5e-03)	7.7e-03 (3.8e-04)	2.8e-03 (6.0e-05)	2.4e-04 (4.3e-06)	1.6e-04 (2.0e-06)
KERN:	3.5e-02 (1.2e-02)	9.7e-03 (2.3e-03)	3.2e-03 (6.0e-05)	3.1e-04 (9.4e-05)	9.0e-05 (4.9e-05)
PROD:	4.5e-01 (2.4e-02)	3.9e-02 (2.0e-03)	1.1e-02 (6.0e-05)	5.3e-03 (8.6e-05)	5.7e-03 (6.2e-05)
TSVQ:	8.6e-03 (2.6e-03)	1.8e-03 (5.3e-04)	1.2e-03 (1.9e-04)	1.4e-04 (3.0e-05)	2.4e-05 (1.4e-05)
GESS:	1.9e-02 (3.4e-03)	5.5e-03 (6.6e-04)	2.0e-03 (1.9e-04)	1.3e-04 (3.0e-05)	8.2e-05 (1.6e-05)
KERN:	4.5e-02 (1.2e-02)	1.4e-02 (2.5e-03)	5.3e-03 (1.9e-04)	9.1e-04 (1.1e-04)	4.4e-04 (6.4e-05)
PROD:	4.8e-01 (2.3e-02)	4.4e-02 (2.5e-03)	1.2e-02 (1.9e-04)	6.2e-03 (1.2e-04)	6.7e-03 (7.7e-05)
TSVQ:	3.2e-02 (3.6e-03)	2.1e-02 (1.3e-03)	7.3e-03 (6.6e-04)	5.8e-03 (1.4e-04)	6.7e-03 (7.7e-05)
GESS:	1.8e-02 (4.9e-03)	1.4e-02 (1.5e-03)	1.0e-02 (6.6e-0 4)	5.9e-03 (1.4e-04)	4.4e-03 (8.4e-05)
KERN:	2.0e-01 (1.5e-02)	8.6e-02 (3.4e-03)	4.6e-02 (6.6e-04)	1.8e-02 (2.3e-04)	1.4e-02 (1.2e-04)
PROD:	8.0e-01 (2.8e-02)	1.3e-01 (4.8e-03)	4.6e-02 (6.6e-04)	3.0e-02 (4.6e-04)	3.3e-02 (3.3e-04)

TABLE I

BIAS (AND VARIANCE) FOR THE MUTUAL INFORMATION ESTIMATES (HISTOGRAM-BASED USING GESSAMAN PARTITION SCHEME (GESS), TREE-STRUCTURED VECTOR QUANTIZATION (TSVQ), CLASSICAL PRODUCT PARTITION (PROD) AND A KERNEL PLUG-IN ESTIMATE (KERN)) OBTAINED FROM 1000 I.I.D. REALIZATIONS OF THE EMPIRICAL PROCESS. SIMULATED DATA CONSIDERS X AND Y TO BE SCALAR GAUSSIAN RANDOM VARIABLES, RESPECTIVELY. PERFORMANCE VALUES ARE REPORTED WITH RESPECT TO SAMPLING LENGTHS $\{33, 179, 564, 3164, 5626\}$ (COLUMNS) AND FOR THE CROSS CORRELATION COEFFICIENTS $r = 0, 0.3, 0.8$ (ROWS)

	33	179	564	3164	5626
TSVQ:	4.4e-03 (2.2e-03)	1.2e-03 (5.5e-04)	1.1e-03 (2.1e-04)	9.5e-04 (4.9e-05)	6.0e-04 (2.6e-05)
GESS:	4.4e-03 (2.2e-03)	2.1e-05 (4.1e-04)	3.3e-03 (2.1e-04)	4.5e-04 (4.8e-05)	3.0e-04 (2.7e-05)
KERN:	2.1e-02 (1.2e-02)	3.2e-04 (2.8e-03)	2.6e-03 (2.1e-04)	1.5e-02 (1.6e-04)	1.9e-02 (9.2e-05)
PROD:	6.9e-01 (4.0e-02)	1.1e+00 (2.1e-02)	2.3e-03 (2.1e-04)	4.1e+00 (1.2e-02)	4.8e+00 (9.8e-03)
TSVQ:	4.4e-04 (1.6e-03)	3.5e-04 (4.9e-04)	1.6e-06 (2.2e-04)	6.3e-05 (4.3e-05)	2.0e-04 (2.5e-05)
GESS:	1.8e-02 (2.7e-03)	4.0e-03 (7.3e-04)	1.6e-06 (2.2e-04)	7.6e-04 (3.7e-05)	9.2e-04 (2.1e-05)
KERN:	1.3e-04 (1.2e-02)	1.2e-02 (2.8e-03)	3.4e-02 (2.2e-04)	7.4e-02 (2.0e-04)	8.7e-02 (1.2e-04)
PROD:	1.1e-01 (4.5e-02)	1.0e+00 (3.0e-02)	2.4e-00 (2.2e-04)	6.3e+00 (1.2e-02)	7.2e+00 (9.5e-03)

TABLE II

BIAS (AND VARIANCE) FOR THE NON-PARAMETRIC MUTUAL INFORMATION ESTIMATES (GESSAMAN PARTITION SCHEME (GESS), TREE-STRUCTURED VECTOR QUANTIZATION (TSVQ), CLASSICAL PRODUCT PARTITION (PROD) AND A KERNEL PLUG-IN ESTIMATE (KERN)) OBTAINED FROM EMPIRICAL DATA OF A MULTIVARIATE GAUSSIAN VECTOR (X, Y) OF DIMENSIONS 4 AND 6, RESPECTIVELY. PERFORMANCE VALUES ARE REPORTED FOR THE SAMPLING LENGTHS $\{33, 179, 564, 3164, 5626\}$ AND FROM 1000 I.I.D. REALIZATIONS OF THE EMPIRICAL PROCESS.

which is a function of the number of sample points n and the design variable $k_n = O(n^{0.5+\tau/2})$ with $\tau \in (1/3, 1)$. We recognize three phases: the constructions of the partition, the estimation of the empirical distributions and the computation of the MI functional. For the first, the number of operations is proportional to the number of binary splits on the space, which is $O(n/k_n)$. For the second, the estimation of $P_{X,Y}$ on the elements of $\pi_n(z_1^n)$ is $O(n)$, however the estimation of $P_X \times P_Y$ is proportional to $n \cdot |\pi_n(z_1^n)|$ that is $O(n^2/k_n)$. The last phase is a linear proportion of $|\pi_n(z_1^n)|$ by (12). At the end, $O(n^2/k_n)$ is the complexity of the algorithm, and considering $\tau > \frac{1}{3}$ the worst case scenario is $O(n^{5/4})$. For the Gessaman's partition scheme, we obtain the same results, i.e., it is $O(n^2/l_n)$ with $l_n = O(n^{0.5+\tau/2})$ and $\tau \in (1/3, 1)$.

The classical product partition scheme (under the same number of quantization bins n/k_n with $k_n = O(n^{0.5+\tau/2})$ and $\tau \in (1/3, 1)$) offers the same complexity $O(n^2/k_n)$. Consequently, there is no penalization in the algorithmic cost of histogram-based methods by incorporating data-driven partitions. On the other hand, kernel-based methods have higher complexity, $O(n^2)$, as they require the computation of pairwise differences between the element of the data [22].

VII. DISCUSSION AND FUTURE WORK

This paper provides theoretical results to show how non-product data-dependent partitions can be incorporated as inference tools for mutual information estimation in the finite dimensional continuous setting. We stipulated a strong consistency result that was applied in two emblematic non-product constructions — statistically equivalent blocks and tree-structured vector quantization. In both scenarios, specific ranges of design values were obtained where density-free strong consistency is guaranteed.

One inherent limitation of this work is that we only stipulate sufficient conditions for consistency. It is a topic of further research to find either necessary and sufficient conditions for our histogram-based estimate to be strongly consistent (resulting in an optimal range for (k_n) and (l_n)) or, alternatively, tighter inequalities for the estimation and approximation errors that could result in a refined consistent range for (k_n) and (l_n).

In particular for the Gessaman partition, it is unknown if one could match the weaker conditions obtained for L_1-strongly consistent density estimate presented in [15], [38], which are optimal for this problem. More generally, the conjecture that always extra conditions are needed to make L_1-consistent histogram-based density estimates consistent for the estimation of information theoretic quantities remains open, a question originally underscored by Györfi and van der Meulen in [23].

Another line of future research is to choose the design variable of the family of consistent estimates optimally (the parameter τ in our problem) with respect to some criterion in order to guarantee not only consistency, but, for instance, a nearly optimal rate of convergence. Results along this direction have been presented for histogram-based density estimation and classification [19], [20], [42]. The idea would be to shrink the gap with respect to the ideal oracle result where, in the domain of consistent design values ($\tau \in (1/3, 1)$), we choose the one with the best small sample performance (the one with the optimal tradeoff between the estimation and the approximation error) for a given joint distribution and sampling length. Improvement can be obtained from the inductive nature of tree-structured partitions, as explored by Darbellay *et al.* [24], and in theory motivated from results in the context of complexity regularized regression and classification trees [35], [42], [43]. We are currently working on this problem, and we hope to present result on it in the near future.

VIII. Acknowledgment

We want to express our deepest gratitude to the anonymous reviewers and Patricio Parada for their rigorous and constructive comments. These contributed enormously to improve the technical content of the manuscript, as well as the clarity of the exposition.

Appendix I
Proof of Theorem 2

A. Preliminary Result for the KL Divergence

The next result addressees the approximation error for the KL divergence estimation problem. As the MI is a particular instance of the KL divergence, this result proves Theorem 2.

THEOREM 7: Let P and Q be two probability measures in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, absolutely continuous with respect to the Lebesgue measure, such that the divergence $D(P||Q)$ is finite. Let us consider a partition scheme $\Pi = \{\pi_1(\cdot), \pi_2(\cdot), \cdots\}$ driven by a random sequence $Z_1, Z_2, \cdots$ with $Z_i \sim P$ for all i, and process distribution denoted by $\mathbb{P}$. If $\forall \delta > 0$,

$$\lim_{n\to\infty} P\left(\left\{z \in \mathbb{R}^d : \text{diam}(\pi_n(z|Z_1^n)) > \delta\right\}\right) \to 0, \quad \mathbb{P}\text{-a.s.,} \tag{26}$$

then,

$$\lim_{n\to\infty} \sum_{B\in\pi_n(Z_1^n)} P(B)\cdot \log\frac{P(B)}{Q(B)} = D(P||Q), \quad \mathbb{P}\text{-a.s..} \tag{27}$$

Proof of Theorem 7: First, we use that

$$D(P||Q) = \sup_{\pi\in\mathcal{A}(\mathbb{R}^d)} D_\pi(P||Q) \tag{28}$$

with $\mathcal{A}(\mathbb{R}^d)$ being the collection of finite measurable partitions of $\mathbb{R}^d$ and $D_\pi(P||Q) \equiv \sum_{B\in\pi}\log\frac{P(B)}{Q(B)}\cdot P(B)$ the KL divergence restricted to $\sigma(\pi) \subset \mathcal{B}(\mathbb{R}^d)$ [1]. Then, for any sequence $z_1^n \in \mathbb{R}^{d\cdot n}$, $D_{\pi_n(z_1^n)}(P||Q) \leq D(P||Q)$, and then,

$$\limsup_{n\to\infty} D_{\pi_n(Z_1^n)}(P||Q) \leq D(P||Q), \ \mathbb{P}\text{-a.s.} \tag{29}$$

Consequently, it is sufficient to show that, $\forall\epsilon > 0$,

$$D(P||Q) < \liminf_{n\to\infty} D_{\pi_n(Z_1^n)}(P||Q) + \epsilon, \ \mathbb{P}\text{-a.s.} \tag{30}$$

We first introduce some definitions.

- **Definition 3:** We define the set $B_n(\delta) \equiv \bigcup_{\substack{A\in\pi_n(z_1^n)\\ diam(A)>\delta}} A$, as the support of the partition $\pi_n(z_1^n)$ with bins of diameter strictly greater than δ.
- **Definition 4:** $\forall B \in \mathcal{B}(\mathbb{R}^d)$, we denote by $\pi_n[B|z_1^n] \equiv \bigcup_{\substack{A\in\pi_n(z_1^n)\\ s.t. A\cap B\neq\emptyset}} A$, the smallest measurable support of $\pi_n(z_1^n)$ that fully contains B.
- **Definition 5:** For any partition π, we define the following measurable function: $f_\pi(P||Q)(x) \equiv \sum_{B\in\pi}\log\frac{P(B)}{Q(B)}\cdot \mathbb{I}_B(x)$, $\forall x \in \mathbb{R}^d$, with $\mathbb{I}_B(\cdot)$ denoting the indicator function[6].
- **Definition 6:** For a partition $\pi = \{A_1, \cdots, A_L\}$ and an event B, let $\bar{\pi}/B \equiv \{A_1\cap B, .., A_L\cap B\}$ denote the partition of B induced by π.

Continuing with the proof, let us consider an arbitrary $\epsilon > 0$. Then from (28) there exists a partition $\bar{\pi} = \{A_1, .., A_L\}$ such that,

$$D(P||Q) < D_{\bar{\pi}}(P||Q) + \epsilon/2. \tag{31}$$

By the continuity of $x\log x$ function, the fact that $P \ll Q$ and that $|\pi| < \infty$, it is simple to show that there exists $\delta(\epsilon/2) > 0$ such that for any partition $\hat{\pi} = \left\{\hat{A}_1, .., \hat{A}_L\right\}$ that satisfies both $\sup_{i=1,..,L}\left|P(A_i) - P(\hat{A}_i)\right| < \delta(\epsilon/2)$ and $\sup_{i=1,..,L}\left|Q(A_i) - Q(\hat{A}_i)\right| < \delta(\epsilon/2)$, then,

$$|D_{\bar{\pi}}(P||Q) - D_{\hat{\pi}}(P||Q)| < \frac{\epsilon}{2}. \tag{32}$$

We will use this result to approximate the events of $\bar{\pi}$ by set operations of our data-dependent collection $\left\{\pi_1(z_1), \pi_2(z_1^2), \cdots\right\}$. More precisely, note that $P \ll \lambda$ and $Q \ll \lambda$, with λ representing the *Lebesgue measure* in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Consequently there exists a bounded set B_o such that $P(B_o^c) < 0.5\cdot\delta(\epsilon/2)$ and $Q(B_o^c) < 0.5\cdot\delta(\epsilon/2)$ [36], with B_o^c the complement of B_o. We will find a good approximation for the bounded set of events $\bar{\pi}/B_o = \{A_1\cap B_o, .., A_L\cap B_o\}$. For that we need to introduce the following *oracle* data dependent partition.

Definition 7: For the bounded set B_o, $\forall\delta > 0$ and for any $n \in \mathbb{N}$, we denote by $\pi_n^\delta(z_1^n)$ a refined version of $\pi_n(z_1^n)$ constructed from, $\forall A \in \pi_n(z_1^n)$:

if $A\cap B_o = \emptyset$, then $A \in \pi_n^\delta(z_1^n)$.
if $A\cap B_o \neq \emptyset$, and $\text{diam}(A) < \delta$, then $A \in \pi_n^\delta(z_1^n)$.

[6] Note that $f_\pi(P||Q)(\cdot)$ is P-integrable ($\in \mathcal{L}_1(P)$), in fact from its definition, $\int f_\pi(P||Q)\partial P(x) = D_\pi(P||Q) \leq D(P||Q) < \infty$.

else, Partition A into a finite collection of events, with the condition that every resulting event intersecting B_o has diameter strictly smaller than δ, and assign those sets to $\pi_n^{\delta}(z_1^n)$. [7]

By definition $\pi_n^{\delta}(z_1^n)$ is a refinement of $\pi_n(z_1^n)$, constructed in such a way that all the bins on $\pi_n^{\delta}[B_o|z_1^n]$ have a diameter strictly less than δ. We use the following construction to approximate $\bar{\pi}/B_o$ from $\pi_n^{\delta}(z_1^n)$,

$$C_{1,n}^{\delta} = \bigcup_{\substack{A\in\pi_n^{\delta}(z_1^n)\\ A\cap(B_o\cap A_1)\neq\emptyset}} A,$$

$$C_{2,n}^{\delta} = \bigcup_{\substack{A\in\pi_n^{\delta}(z_1^n)\\ A\cap(B_o\cap A_2)\neq\emptyset}} A\setminus C_{1,n}^{\delta}, \cdots,$$

$$C_{L,n}^{\delta} = \bigcup_{\substack{A\in\pi_n^{\delta}(z_1^n)\\ A\cap(B_o\cap A_L)\neq\emptyset}} A\setminus \bigcup_{k=1}^{L-1} C_{k,n}^{\delta},$$

and we denote the approximation by $\bar{\pi}_n^{\delta} \equiv \left\{C_{1,n}^{\delta}, .., C_{L,n}^{\delta}\right\}$ (without loss of generality we assume that $\forall i \in \{1, .., L\}$, $B_o \cap A_i \neq \emptyset$). By the continuity of a measure under monotone set sequence [36], [44], it can be shown that $\forall \bar{\epsilon} > 0$ $\exists \bar{\delta} > 0$ sufficiently small such that $\sup_{i=1,..,L} \lambda((A_i \cap B_o)\triangle C_{i,n}^{\bar{\delta}}) < \bar{\epsilon}$, uniformly $\forall n \in \mathbb{N}$. Then in particular, using that $P \ll \lambda$ and $Q \ll \lambda$, we can choose $\bar{\delta}$ such that $\sup_{i=1,..,L} \left|P(A_i \cap B_o) - P(C_{i,n}^{\bar{\delta}})\right| < 0.5 \cdot \delta(\epsilon/2)$ and $\sup_{i=1,..,L} \left|Q(A_i \cap B_o) - Q(C_{i,n}^{\bar{\delta}})\right| < 0.5 \cdot \delta(\epsilon/2)$, $\forall n \in \mathbb{N}$. Using this result and (32), then for this $\bar{\delta}(\epsilon/2)$ we have that,

$$\left|D_{\bar{\pi}}(P||Q) - \sum_{i=1}^{L} \log \frac{P(C_{i,n}^{\bar{\delta}})}{Q(C_{i,n}^{\bar{\delta}})} \cdot P(C_{i,n}^{\bar{\delta}})\right| < \epsilon/2,\ \forall n \in \mathbb{N}. \tag{33}$$

On the other hand, by the hypothesis in (26),

$$\lim_{n\to\infty} P(B_n(\bar{\delta})) = 0 \tag{34}$$

for $\mathbb{P}$-almost every sequence $z_1, z_2, \cdots \in \mathbb{R}^{d\cdot\mathbb{N}}$. Let us concentrate on one of those typical sequences and show that for any of them, (30) is satisfied. Let $z_1, z_2, \ldots$ be a *typical sequence* with respect to $\bar{\delta}$ (i.e., a realization of the process where (34) is satisfied). Then,

$$\begin{aligned}
&D_{\bar{\pi}}(P||Q) - D_{\pi_n(z_1^n)}(P||Q)\\
&< \frac{\epsilon}{2} + \sum_{i=1}^{L} \log \frac{P(C_{i,n}^{\bar{\delta}})}{Q(C_{i,n}^{\bar{\delta}})} \cdot P(C_{i,n}^{\bar{\delta}}) - D_{\pi_n(z_1^n)}(P||Q)\\
&\leq \frac{\epsilon}{2} + D_{\pi_n^{\bar{\delta}}(z_1^n)}(P||Q) - D_{\pi_n(z_1^n)}(P||Q) = \frac{\epsilon}{2} +\\
&\int_{B_n(\bar{\delta})} f_{\pi_n^{\bar{\delta}}(z_1^n)}(P||Q)\partial P(x) - \int_{B_n(\bar{\delta})} f_{\pi_n(z_1^n)}(P||Q)\partial P(x),
\end{aligned} \tag{35}$$

the first inequality because of (33), the second due to the monotonic behavior of $D_{\pi}(P||Q)$ under refined partitions [1], [34] and the fact that by construction $\bar{\pi}_n^{\bar{\delta}} \ll \pi_n^{\bar{\delta}}(z_1^n)$, and the last equality because by construction $\pi_n^{\bar{\delta}}(z_1^n)$ and $\pi_n(z_1^n)$ are equivalent in the support $B_n^c(\bar{\delta})$. Again using the monotonicity of the KL divergence under sequence of refined partitions, we have that,

$$\begin{aligned}
\log \frac{P(B_n(\bar{\delta}))}{Q(B_n(\bar{\delta}))} \cdot P(B_n(\bar{\delta})) &\leq \int_{B_n(\bar{\delta})} f_{\pi_n^{\bar{\delta}}(z_1^n)}(P||Q)\partial P(x)\\
&\leq \int_{B_n(\bar{\delta})} f(P||Q)\partial P(x).
\end{aligned} \tag{36}$$

Given that $D(P||Q) < \infty$ and that $\lim_n P(B_n(\delta)) = 0$, by the *dominated convergence theorem* [36]

$$\lim_{n\to\infty} \int_{B_n(\delta)} f(P||Q)\partial P(x) = 0,$$

then from (36) $\limsup_{n\to\infty} \int_{B_n(\bar{\delta})} f_{\pi_n^{\bar{\delta}}(z_1^n)}(P||Q)\partial P(x) \leq 0$. On the other hand,

$$P(B_n(\bar{\delta})) \cdot \log \frac{P(B_n(\bar{\delta}))}{Q(B_n(\bar{\delta}))} \geq P(B_n(\bar{\delta})) \cdot \log P(B_n(\bar{\delta}))$$

and the fact that $\lim_{x\to 0} x \cdot \log x = 0$ implies that $\liminf_{n\to\infty} \int_{B_n(\bar{\delta})} f_{\pi_n^{\bar{\delta}}(z_1^n)}(P||Q)\partial P(x) \geq 0$. Consequently,

$$\lim_{n\to\infty} \int_{B_n(\bar{\delta})} f_{\pi_n^{\bar{\delta}}(z_1^n)}(P||Q)\partial P(x) = 0.$$

By the same argument,

$$\lim_{n\to\infty} \int_{B_n(\bar{\delta})} f_{\pi_n(z_1^n)}(P||Q)\partial P(x) = 0.$$

Finally taking limits in the previous set of inequalities (35), $D_{\bar{\pi}}(P||Q) < \liminf_{n\to\infty} D_{\pi_n(z_1^n)}(P||Q) + \frac{\epsilon}{2}$, and from (31),

$$D(P||Q) < \liminf_{n\to\infty} D_{\pi_n(z_1^n)}(D||P) + \epsilon,$$

for any typical sequence (or $\mathbb{P}-$a.s.). □

B. *Proof of Theorem 2:*

It is just a direct consequence of Theorem 7, considering the measures P=$P_{X,Y}$ and $Q = P_X \times P_Y$. □

Appendix II
Proof of Theorem 3

We know that $I(X;Y) = D(P||Q)$ with P denoting the joint distribution $P_{X,Y}$ and $Q = P_X \times P_Y$. We denote by P_n and Q_n the empirical versions of P and Q induced by $Z_1, .., Z_n$ and the product bin structure of $\pi_n(\cdot)$. Then the empirical MI estimate in (12) can be expressed by $D_{\pi_n(Z_1^n)}(P_n||Q_n)$. To prove the result we use the following inequality,

$$\begin{aligned}
&\left|D_{\pi_n(Z_1^n)}(P_n||Q_n) - D(P||Q)\right| \leq\\
&\left|D_{\pi_n(Z_1^n)}(P_n||Q_n) - D_{\pi_n(Z_1^n)}(P||Q)\right|\\
&+ \left|D_{\pi_n(Z_1^n)}(P||Q) - D(P||Q)\right|.
\end{aligned} \tag{37}$$

The last term in the right hand side of (37) is the approximation error, which from Theorem 2 converges to zero $\mathbb{P}$-a.s. as n

[7] Note that this oracle partition is possible (not unique) as B_o is a bounded set (referring to the refinement step on the bins intersecting B_o with diameter greater or equal to δ).

tends to infinity. Then we just need to focus on the estimation error term. From triangular inequality

$$\left|D_{\pi_n(Z_1^n)}(P_n||Q_n) - D_{\pi_n(Z_1^n)}(P||Q)\right|$$
$$\leq \left|\sum_{A\in\pi_n(Z_1^n)} \left[P_n(A)\log P_n(A) - P(A)\log P(A)\right]\right| \quad (38)$$
$$+\left|\sum_{A\in\pi_n(Z_1^n)} \left[P_n(A)\log Q_n(A) - P(A)\log Q(A)\right]\right|. \quad (39)$$

Concerning the term in (38), it is upper bounded by

$$\left|\sum_{A\in\pi_n(Z_1^n)} \left[P_n(A) - P(A)\right]\log P_n(A)\right|$$
$$+\left|\sum_{A\in\pi_n(Z_1^n)} \left[\log P_n(A) - \log P(A)\right]P(A)\right| \leq$$
$$\sum_{A\in\pi_n(Z_1^n)} \left|P_n(A) - P(A)\right| \log\frac{n}{k_n} +$$
$$\sup_{A\in\pi_n(Z_1^n)} \left|\log P(A) - \log P_n(A)\right|, \quad (40)$$

where we use that $P_n(A) \geq \frac{k_n}{n}$ $\forall A \in \pi_n(Z_1^n)$. The first term in the RHS of (40), from left to right, is bounded by,

$$\mathbb{P}\left(\sum_{A\in\pi_n(Z_1^n)} \left|P_n(A) - P(A)\right| \cdot \log\frac{n}{k_n} > \epsilon\right)$$
$$\leq 4\Delta^*_{2n}(\mathcal{A}_n)2^{\mathcal{M}(\mathcal{A}_n)} \cdot \exp\left\{-\frac{n\epsilon^2}{(\log n/k_n)^2\cdot 32}\right\}, \quad (41)$$

where the inequality follows from $\pi_n(Z_1^n) \subset \mathcal{A}_n$ and Lemma 1. Note that the exponential term $\exp\left\{-\frac{n\epsilon^2}{(\log n/k_n)^2\cdot 32}\right\} \leq \exp\left\{-\frac{n\epsilon^2}{(\log n)^2\cdot 32}\right\}$, where this last sequence is (uniformly in ϵ) dominated by the sequence $(\exp\{-m^{\bar{\tau}}\})_{n\in\mathbb{N}}$, $\forall\bar{\tau} \in (0,1)$. Consequently from **c.2**, **c.3** and **c.4**, it is simple to show that $\forall\epsilon > 0$, $\limsup_{n\to\infty}\frac{1}{m^\tau}\cdot \log\mathbb{P}\left(\sum_{A\in\pi_n(Z_1^n)}\left|P_n(A) - P(A)\right| > \frac{\epsilon}{\log n/k_n}\right)\leq C_o$, being $C_o < 0$. Finally from the fact that $\sum_{n\geq 0}\exp\{C_o\cdot m^\tau\} < \infty$ and the *Borel-Cantelli* lemma [36], $\lim_{n\to\infty}\sum_{A\in\pi_n(Z_1^n)}\left|P_n(A) - P(A)\right|\log\frac{n}{k_n} = 0$, $\mathbb{P}$-a.s.

Concerning the second term in the RHS of (40), we use the following result.

PROPOSITION 2: (Silva *et al.* [30]) Under the conditions **c.2**, **c.3** and **c.4** of the Theorem 3,

$$\lim_{n\to\infty}\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A)}{P_n(A)} - 1\right| = 0, \quad \mathbb{P}-a.s. \quad (42)$$

(Proof presented at the end of this section).

Then from (42), it is simple to show that $\lim_{n\to\infty}\sup_{A\in\pi_n(Z_1^n)}\frac{P(A)}{P_n(A)} = 1$ and $\lim_{n\to\infty}\sup_{A\in\pi_n(Z_1^n)}\frac{P_n(A)}{P(A)} = 1$ $\mathbb{P}$-a.s. On the other hand, $\forall A \in \pi_n(Z_1^n)$,

$$\left|\frac{P_n(A)}{P(A)} - 1\right| \leq \frac{\mid P(A) - P_n(A)\mid}{P_n(A)}\cdot\frac{P_n(A)}{P(A)},$$

then $\lim_{n\to\infty}\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P_n(A)}{P(A)} - 1\right| = 0$ $\mathbb{P}$-a.s. Finally noting that $\forall n$,

$$\sup_{A\in\pi_n(Z_1^n)}\left|\log\frac{P(A)}{P_n(A)}\right| \leq$$
$$\max\left\{\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A)}{P_n(A)} - 1\right|, \sup_{A\in\pi_n(Z_1^n)}\left|\frac{P_n(A)}{P(A)} - 1\right|\right\},$$

this last inequality shows the result. Concerning the term in (39), we bounded it by the expression in (43) (see figure at the top of the next page), where considering the product bin structure of $\pi_n(\cdot)$, we have that $\forall A \in \pi_n(Z_1^n)$, $Q_n(A) = P_n(A_{[1,p]}\times\mathbb{R}^q)P_n(\mathbb{R}^p\times A_{[p+1,d]})$, with $A_{[1,p]}$ and $A_{[p+1,d]}$ a short-hand notation for $\xi_{[1,p]}(A)$ and $\xi_{[p+1,d]}(A)$, respectively. We focus attention on just one of the terms in (43), since by symmetry the derivation for the other is equivalent. We have that $\left|\sum_{A\in\pi_n(Z_1^n)}\left[P(A)\log P(A_{[1,p]}\times\mathbb{R}^q) - P_n(A)\log P_n(A_{[1,p]}\times\mathbb{R}^q)\right]\right| \leq$

$$\sum_{A\in\pi_n(Z_1^n)} \left|P_n(A) - P(A)\right| \log\frac{n}{k_n} +$$
$$\sup_{A\in\pi_n(Z_1^n)} \left|\log P(A_{[1,p]}\times\mathbb{R}^q) - \log P_n(A_{[1,p]}\times\mathbb{R}^q)\right|, \quad (44)$$

where it has been proved that the first term of the bound tends to zero $\mathbb{P}$-a.s as n tends to infinity. Concerning the second term in (44), from one of the previous arguments it is sufficient to show that $\lim_{n\to\infty}\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A_{[1,p]}\times\mathbb{R}^q)}{P_n(A_{[1,p]}\times\mathbb{R}^q)} - 1\right| = 0$ $\mathbb{P}$-a.s. Analyzing this expression, we have that, $\forall\epsilon > 0$,

$$\mathbb{P}\left(\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A_{[1,p]}\times\mathbb{R}^q)}{P_n(A_{[1,p]}\times\mathbb{R}^q)} - 1\right| > \epsilon\right) \leq$$
$$\mathbb{P}\left(\sup_{A\in\pi_n(Z_1^n)}\left|P(A_{[1,p]}\times\mathbb{R}^q) - P_n(A_{[1,p]}\times\mathbb{R}^q)\right| > \frac{k_n\cdot\epsilon}{n}\right)$$
$$\leq \mathcal{S}_n(\mathcal{C}_{[1,p],n})\cdot\exp\left\{-\frac{k_n^2\cdot\epsilon^2}{n\cdot 8}\right\}, \quad (45)$$

the first inequality results from the fact that $P_n(A_{[1,p]}\times\mathbb{R}^d) \geq P_n(A) \geq \frac{k_n}{n}$, $\forall A \in \pi_n(Z_1^n)$, and the second from $\mathcal{C}_{[1,p]}(Z_1^n) \subset \mathcal{C}_{[1,p],n}$ and the Vapnik-Chervonenkis inequality in Theorem 1. Finally considering that $(k_n) \approx (n^{0.5+\tau/2})$ and **c.1**,

$$\limsup_{n\to\infty}\frac{1}{n^\tau}\log\mathbb{P}\left(\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A_{[1,p]}\times\mathbb{R}^q)}{P_n(A_{[1,p]}\times\mathbb{R}^q)} - 1\right| > \epsilon\right)$$

$< C(\epsilon)$ a constant function of ϵ that is strictly negative. Then again from the *Borel-Cantelli* lemma, $\lim_{n\to\infty}\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A_{[1,p]}\times\mathbb{R}^q)}{P_n(A_{[1,p]}\times\mathbb{R}^q)} - 1\right| = 0$ $\mathbb{P}$-a.s, which is the last piece of result needed to prove the theorem.

□

$$\left|\sum_{A\in\pi_n(Z_1^n)}\left[P_n(A)\log Q_n(A)-P(A)\log Q(A)\right]\right|\leq$$
$$\left|\sum_{A\in\pi_n(Z_1^n)}\left[P(A)\log P(A_{[1,p]}\times\mathbb{R}^q)-P_n(A)\log P_n(A_{[1,p]}\times\mathbb{R}^q)\right]\right|$$
$$+\left|\sum_{A\in\pi_n(Z_1^n)}\left[P(A)\log P(\mathbb{R}^p\times A_{[p+1,d]})-P_n(A)\log P_n(\mathbb{R}^p\times A_{[p+1,d]})\right]\right| \quad (43)$$

A. Proof of Proposition 2

We have that

$$\mathbb{P}\left(\sup_{A\in\pi_n(Z_1^n)}\left|\frac{P(A)}{P_n(A)}-1\right|>\epsilon\right)$$
$$\leq\mathbb{P}\left(\sup_{A\in\pi_n(Z_1^n)}|P(A)-P_n(A)|>\epsilon\cdot k_n/n\right)$$
$$\leq\mathbb{P}\left(\sup_{\pi\in\mathcal{A}_n}\sup_{A\in\pi}|P(A)-P_n(A)|>\epsilon\cdot k_n/n\right)$$
$$\leq 4\Delta^*_{2n}(\mathcal{A}_n)2^{\mathcal{M}(\mathcal{A}_n)}\exp^{-\frac{(\epsilon\cdot k_n)^2}{n\cdot 32}},$$

where the last inequality is from Lemma 1. Using **c.2**, **c.3** and **c.4** from Theorem 3, there exits a $\tau\in(0,1)$ such that $\lim_{m\to\infty}\frac{1}{n^\tau}\cdot\log\mathbb{P}\left(\sup_{A\in\pi_m(Z_1^m)}|P_n(A)-P(A)|>\epsilon\cdot k_n/n\right)$ $\leq-\lim_{n\to\infty}\epsilon^2\cdot\frac{k_n^2}{n^{1+l}}=-\epsilon\cdot C$, for some $C>0$. Then *Borel-Cantelli* proves the result. □

Appendix III
Proof of Theorem 6

Proof: As considered in Appendix I and II, we address this result for the more general scenario of the KLD. For the rest, let $\pi_n(Z_1^n)=\pi_{T_n}(Z_1^n)$ denote the n-sample partition rule of the UBTSS Π.

A. Reducing the Problem to the Bounded Measurable Space $([0,1]^d,\mathcal{B}([0,1]^d))$

Note that Π is *monotone transformation invariant* [28], in the sense that $\forall\pi_n\in\Pi$, $\forall z\in\mathbb{R}^d$, $\forall z_1^n\in\mathbb{R}^{d\cdot n}$,

$$\pi_n(z|z_1,..z_n)=\pi_n(F(z)|F(z_1),..F(z_n)),$$

where $F:\mathbb{R}^d\to\mathbb{R}^d$ is an arbitrary function that can be expressed by $F(x)=(f_1(x(1)),\cdots,f_d(x(d)))$, for some collection of strictly increasing real functions $\{f_i(\cdot):i=1,..,d\}$. In particular, we can consider $f_i(\cdot)$ to be the distribution function of the probability Q restricted to events on the i-coordinate $\forall i\in\{1,..,d\}$. Without loss of generality we can restrict to the case when $\{f_i(\cdot):i=1,..,d\}$ are strictly increasing. Consequently, the induced distributions of the transform space, denoted by $\bar{Q}$ and $\bar{P}$ respectively, have support on $[0,1]^d$ and satisfies that [1]

$$D(P||Q)=D(\bar{P}||\bar{Q}), \quad (46)$$

because $F(\cdot)$ is one-to-one measurable mapping from $\mathbb{R}^d$ to $[0,1]^d$ (more precisely $\left\{F^{-1}(A):A\in\mathcal{B}([0,1]^d)\right\}=\mathcal{B}(\mathbb{R}^d)$). Moreover, if we apply Π in the transform domain, it is simple to check that

$$D_{\pi(Z_1^n)}(P||Q)=D_{\pi(F(Z_1),..,F(Z_n)))}(\bar{P}||\bar{Q}), \quad (47)$$

and from (46) and (47) without loss of generality we can assume that Q and P are defined on $([0,1]^d,\mathcal{B}([0,1]^d)$.

B. Formulation of a Sufficient Condition

Given that $\pi_n(Z_1^n)$ is induced by axis-parallel hyperplanes, every cell $U\in\pi_n(Z_1^n)$ can be represented by a finite dimensional rectangle of the form $\otimes_{i=1}^d[l_i,u_i)$ (with the possible open and closed interval variations). In this scenario, $\forall U\in\pi_n(Z_1^n)$,

$$diam(U)\leq\sum_{i=1}^d length_i(U), \quad (48)$$

with $length_i(U)$ denoting the Lebesgue measure of the projection of U on the i-coordinate. Then from *Markov's inequality*, for proving the shrinking cell condition it suffices to show that [28],

$$\lim_{n\to\infty}E_P\left(\sum_{i=1}^d length_i(\pi_n(Z|Z_1^n))\right)=$$
$$\lim_{n\to\infty}\int_{[0,1]^d}\sum_{i=1}^d length_i(\pi_n(z|Z_1^n))\partial P(z)=0, \quad (49)$$

almost surely with respect to the process distribution of $Z_1,Z_2\cdots$.

C. A Preliminary Definition

Let $U=\otimes_{i=1}^d[l_i,u_i]$ be a rectangle in $\mathcal{B}([0,1]^d)$ and let $\left\{H_0^0,H_0^1,H_1^1,\cdots,H_0^{d-1},..,H_{2^{d-1}-1}^{d-1}\right\}$ be a sequence of axis-parallel hyperplanes used to recursively split U in every coordinate. This partitions U in 2^d cells. More precisely, H_0^0 parallel to the 1-coordinate splits $U_0^0=U$ into two rectangles U_0^1, U_1^1, then H_0^1 and H_1^1 parallel to the 2-coordinate split U_0^1 and U_1^1 into U_0^2, U_1^2, and U_2^2, U_3^2 respectively, and inductively at the end of the process a TSP for U is created $\left\{U_j^d:j=0,..,2^d-1\right\}$.

Definition 8: Let P be a probability measure in $([0,1]^d,\mathcal{B}([0,1]^d))$, and for the aforementioned construction let $p_j^l=P(U_j^l)$ denote the probability of every induced rectangle. We say that $\left\{H_0^0, H_0^1, H_1^1, \cdots, H_0^{d-1}, .., H_{2^{d-1}-1}^{d-1}\right\}$ is a *sequence of ϵ-good median cuts* for U with respect to P if: $\forall l \in \{0,..,d-1\}$ and $j \in \left\{0,..,2^l-1\right\}$,

$$\max(p_{2j}^{l+1}, p_{2j+1}^{l+1}) \leq \frac{1}{2}(1+\epsilon)^{1/d} \cdot p_j^l. \tag{50}$$

PROPOSITION 3: Let U be a finite dimensional rectangle in $\mathcal{B}([0,1]^d)$ with probability $P(U) = p > 0$, and $\left\{U_j^d : j = 0,..,2^d-1\right\}$ a partition of U induced by sequence of ϵ-good median cuts. Then,

$$\sum_{j=0}^{2^d-1} p_j^d \cdot \sum_{i=1}^{d} length_i(U_j^d) \leq \frac{1+\epsilon}{2} \cdot p \cdot \sum_{i=1}^{d} length_i(U). \tag{51}$$

The proof is a simple consequence of (50).

D. Final Argument

Let $\Pi = \{T_1, T_2, \cdots\}$ be the UBTSS of height (d_n), i.e. $|\pi_n(z_1^n)| = 2^{d_n}$, $\forall n > 0$ and $\forall z_1^n \in \mathbb{R}^{dn}$. In addition, let us consider the pruned UBTSS $\bar{\Pi} = \left\{\bar{T}_1, \bar{T}_2, \cdots\right\}$, where $\bar{T}_n \equiv T_n^{\bar{d}_n}$ and $\bar{d}_n = d \cdot \lfloor d_n/d \rfloor \leq d_n$. It is sufficient to prove the shrinking cell condition for $\bar{\Pi}$. The reason for this reduction is that by construction the height of $\bar{T}_n$ is a power of d, and then we can recursively apply *Proposition* 3 to bound $E_P\left(\sum_{i=1}^d length_i(\bar{\pi}_n(Z|Z_1^n))\right)$, where $\bar{\pi}_n(Z_1^n) \equiv \pi_{\bar{T}_n}(Z_1^n)$. More precisely, let $B_n(\epsilon) \in \mathcal{B}(\mathbb{R}^{d \cdot n})$ be the collection of sequences where all the axis-parallel hyperplanes that induce $\bar{\pi}_n(z_1^n)$ are ϵ-good median cuts with respect to P. Then from (51), for all $z_1^n \in B_n(\epsilon)$ we have that,

$$E_P\left(\sum_{i=1}^{d} length_i(\bar{\pi}_n(Z|z_1^n))\right) \leq \left[\frac{1+\epsilon}{2}\right]^{r_n} \cdot d, \tag{52}$$

with $r_n = \lfloor d_n/d \rfloor$. Let us choose $\epsilon_0 > 0$ sufficiently small in order that $1 + \epsilon_0 < 2$. Then from (52) as $r_n \to \infty$ (when $n \to \infty$), the event $A_n(\epsilon) = \left\{z_1^n \in \mathbb{R}^{d \cdot n} : E_P\left(\sum_{i=1}^d length_i(\bar{\pi}_n(Z|z_1^n))\right) > \epsilon\right\}$ $\in \mathcal{B}(\mathbb{R}^{d \cdot n})$ is eventually contained in $B_n(\epsilon_0)^c$, $\forall \epsilon > 0$. Consequently, let us focus on the analysis of $\mathbb{P}(B_n(\epsilon_0)^c)$. By definition $B_n(\epsilon_0)^c$ is the event that one of the cuts of $\bar{T}_n$ is not ϵ_0-median good. By construction the number of hyperplanes splitting $\bar{T}_n$ is given by $(1 + 2 + \cdots + 2^{\bar{d}_n - 1})$, then

$$\mathbb{P}(B_n(\epsilon_0)^c) \leq 2^{\bar{d}_n} \cdot \mathbb{P}(B_n^o(\epsilon_0)) \tag{53}$$

with $B_n^o(\epsilon_0)$ denoting the event that a cut is not ϵ_0-median good. Devroye *et al.* [28] (*Theorem 20.2*) showed for this case of balanced trees that,

$$\mathbb{P}(B_n^o(\epsilon_0)) \leq 2 \cdot \exp\left(-\frac{n}{2^{\bar{d}_n+2}} \cdot ((1+\epsilon_0)^{1/d} - 1)^2\right), \tag{54}$$

for n sufficiently large. Consequently, from (53) and (54), there exists $K > 0$ such, $\mathbb{P}(B_n(\epsilon_0)^c) \leq$

$$K \cdot \exp\left(\log(2) \cdot \bar{d}_n - \frac{n}{2^{\bar{d}_n+2}} \cdot ((1+\epsilon_0)^{1/d} - 1)^2\right), \tag{55}$$

$\forall n \in \mathbb{N}$. From the definition of $\bar{d}_n$, we have that $d_n - d < \bar{d}_n \leq d_n$, and consequently from the hypothesis, there exists $(a_n) \approx (n^\theta)$ for some $\theta > 0$, such that

$$\frac{n}{\bar{d}_n 2^{\bar{d}_n}} - \frac{a_n}{\bar{d}_n} \to \infty, \tag{56}$$

which from (55) is sufficient to show that,

$$\lim_{n\to\infty} \frac{\mathbb{P}(B_n(\epsilon_0)^c)}{\exp\left(-n^\theta\right)} = 0. \tag{57}$$

Finally, $\limsup_n A_n(\epsilon) \subset \limsup_n B_n(\epsilon_0)^c$, $\forall \epsilon > 0$, then given that $\sum_n \mathbb{P}(B_n(\epsilon_0)^c) < \infty$ from (57), and the *Borel-Cantelli lemma*, $E_P\left(\sum_{i=1}^d length_i(\bar{\pi}_n(Z|Z_1^n))\right)$ tends to zero with probability one with respect to $\mathbb{P}$, which concludes the proof. □

References

[1] R. M. Gray, *Entropy and Information Theory*. Springer - Verlag, New York, 1990.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Interscience, New York, 1991.

[3] J. W. Fisher III, M. Wainwright, E. Sudderth, and A. S. Willsky, "Statistical and information-theoretic methods for self-organization and fusion of multimodal, networked sensors," *International Journal of High Performance Computing Applications*, vol. 16, no. 3, pp. 337–353, 2002.

[4] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Transactions on Image Processing*, vol. 10, no. 11, pp. 1647–1658, November 2001.

[5] M. Padmanabhan and S. Dharanipragada, "Maximizing information content in feature extraction," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 512–519, July 2005.

[6] J. Silva and S. Narayanan, "Minimum probability of error signal representation," in *IEEE Workshop Machine Learning for Signal Processing*, August 2007.

[7] ——, "Discriminative wavelet packet filter bank selection for pattern recognition," *IEEE Transactions on Signal Processing*, vol. 57, no. 5, pp. 1796–1810, May 2009.

[8] M. L. Cooper and M. I. Miller, "Information measures for object recognition accommodating signature variability," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1896–1907, August 2000.

[9] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2083–2099, December 2000.

[10] T. Butz and J.-P. Thiran, "From error probability to information theoretic (multi-modal) signal processing," *Elsevier Signal Processing*, vol. 85, pp. 875–902, 2005.

[11] J. Kim, J. W. Fisher III, A. Yezzi, M. Cetin, and A. S. Willsky, "A nonparametric statistical method for image segmentation using information theory and curve evolution," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1486–1502, October 2005.

[12] M. B. Westover and J. A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 299–320, January 2008.

[13] L. Devroye and L. Györfi, *Nonparametric density estimation: The L_1 view*. Wiley Interscience, New York, 1895.

[14] S. Abou-Jaoude, "Condition nécessaires et sufficantes de convergence L_1 en probabilité de l'hitograme pour une densité," *Ann. Inst. H. Poincaré*, vol. 12, pp. 213–231, 1976.

[15] G. Lugosi and A. B. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *The Annals of Statistics*, vol. 24, no. 2, pp. 687–706, 1996.

[16] A. Barron, L. Györfi, and E. C. van der Meulen, "Distribution estimation consistent in total variation and in two types of information divergence," *IEEE Transactions on Information Theory*, vol. 38, no. 5, pp. 1437–1454, September 1992.

[17] L. Györfi and E. C. van der Meulen, "Density estimation consistent in information divergence," in *IEEE International Symposium on Information Theory*, 1994, pp. 35–35.

[18] L. Györfi, F. Liese, I. Vajda, and E. C. van der Meulen, "Distribution estimates consistent in χ^2- divergence," *Statistics*, vol. 32, no. 1, pp. 31–57, 1998.

[19] I. Vajda and E. C. van der Meulen, "Optimization of barron density estimates," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1867–1883, July 2001.

[20] A. Berlinet, I. Vajda, and E. C. van der Meulen, "About the asymptotic accuracy of Barron density estimate," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 999–1009, May 1998.

[21] A. Berlinet, L. Devroye, and L. Györfi, "Asymtotic normality of L_1-error in density estimation," *Statistics*, vol. 26, no. 329-343, 1995.

[22] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Non-parametric entropy estimation: An overview," *Int. J. of Math. and Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[23] L. Györfi and E. van der Meulen, "Density-free convergence properties of various estimators of entropy," *Computational Statistics and Data Analysis*, vol. 5, pp. 425–436, 1987.

[24] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partition of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[25] A. B. Nobel, "Histogram regression estimation using data-dependent partitions," *The Annals of Statistics*, vol. 24, no. 3, pp. 1084–1105, 1996.

[26] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer - Verlag, New York, 1999.

[27] ——, *Statistical Learning Theory*. John Wiley, 1998.

[28] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[29] M. P. Gessaman, "A consistent nonparametric multivariate density estimator based on statistically equivalent blocks," *Ann. Math. Statist.*, vol. 41, pp. 1344–1346, 1970.

[30] J. Silva and S. Narayanan, "Universal consistency of data-driven partitions for divergence estimation," in *IEEE International Symposium on Information Theory*, June 2007.

[31] ——, "Histogram-based estimation for the divergence revisited," in *IEEE International Symposium on Information Theory*. IEEE, June 2009.

[32] V. Vapnik and A. J. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability Apl.*, vol. 16, pp. 264–280, 1971.

[33] S. Kullback, *Information theory and Statistics*. New York: Wiley, 1958.

[34] F. Liese, D. Morales, and I. Vajda, "Asymptotically sufficient partition and quantization," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5599–5606, 2006.

[35] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[36] P. R. Halmos, *Measure Theory*. Van Nostrand, New York, 1950.

[37] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. Springer - Verlag, New York, 2001.

[38] S. Abou-Jaoude, "La convergence L_1 e L_∞ de l'estimateur de la partition aléatoire pour une densité," *Ann. Inst. H. Poincaré*, vol. 12, pp. 299–317, 1976.

[39] F. Piera and P. Parada, "On convergence properties of Shannon entropy," *Problems of Information Transmission*, vol. 45, no. 2, pp. 75–94, June 2009.

[40] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. on Electronic Computers*, vol. 14, pp. 326–334, 1965.

[41] J. Silva, "On optimal signal representation for statistical learning and pattern recognition," Ph.D. dissertation, University of Southern California, http://digitallibrary.usc.edu/assetserver/controller/item/etd-Silva-2450.pdf, December 2008.

[42] C. Scott and R. D. Nowak, "Minimax-optimal classification with dyadic decision trees," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1335–1353, April 2006.

[43] A. B. Nobel, "Analysis of a complexity-based pruning scheme for classification tree," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2362–2368, 2002.

[44] L. Breiman, *Probability*. Addison-Wesley, 1968.

Jorge Silva is Assistant Professor at the Electrical Engineering Department, University of Chile, Santiago, Chile. He received the Master of Science (2005) and Ph.D (2008) in Electrical Engineering from the University of Southern California (USC). He is IEEE member of the Signal Processing and Information Theory Societies and he has participated as a reviewer in various IEEE journals on Signal Processing. Jorge Silva was research assistant at the Signal Analysis and Interpretation Laboratory (SAIL) at USC (2003-2008) and was also research intern at the Speech Research Group, Microsoft Corporation, Redmond (Summer 2005).

Jorge Silva is recipient of the Outstanding Thesis Award 2009 for Theoretical Research of the Viterbi School of Engineering, the Viterbi Doctoral Fellowship 2007-2008 and Simon Ramo Scholarship 2007-2008 at USC. His research interests include: non-parametric learning; optimal signal representation for pattern recognition; speech processing; tree-structured vector quantization for lossy compression and statistical learning; universal quantization; sequential detection; distributive learning and sensor networks.

Shrikanth (Shri) Narayanan (Ph.D'95, UCLA) is the Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics and Psychology. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC he directs the Signal Analysis and Interpretation Laboratory. His research focuses on human-centered information processing and communication technologies.

Shri Narayanan is an Editor for the Computer Speech and Language Journal and an Associate Editor for the IEEE Transactions on Multimedia, IEEE Transactions on Affective Computing, and for the Journal of the Acoustical Society of America. He was also previously an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000-04) and the IEEE Signal Processing Magazine (2005-2008). He served on the Speech Processing technical committee (2005-2008) and Multimedia Signal Processing technical committees (2004-2008) of the IEEE Signal Processing Society and presently serves on the Speech Communication committee of the Acoustical Society of America and the Advisory Council of the International Speech Communication Association.

Shri Narayanan is a Fellow of the Acoustical Society of America, IEEE, and the American Association for the Advancement of Science and a member of Tau-Beta-Pi, Phi Kappa Phi and Eta-Kappa-Nu. He is a recipient of a number of honors including Best Paper awards from the IEEE Signal Processing society in 2005 (with Alex Potamianos) and in 2009 (with Chul Min Lee) and appointment as a Signal Processing Society Distinguished Lecturer for 2010-11. Papers with his students have won awards at ICSLP'02, ICASSP'05, MMSP06, MMSP'07 and DCOSS09 and InterSpeech2009-Emotion Challenge. He has published over 350 papers and has seven granted U.S. patents.