Japanese Journal of Statistics and Data Science (2019) 2:507–527
https://doi.org/10.1007/s42081-019-00054-y

ORIGINAL PAPER

Information Theory and Statistics

Minimum distance histograms with universal performance guarantees

Raazesh Sainudiin[1] · Gloria Teng[2]

Received: 31 December 2018 / Accepted: 12 June 2019 / Published online: 13 July 2019

Abstract

We present a data-adaptive multivariate histogram estimator of an unknown density *f* based on *n* independent samples from it. Such histograms are based on binary trees called regular pavings (RPs). RPs represent a computationally convenient class of simple functions that remain closed under addition and scalar multiplication. Unlike other density estimation methods, including various regularization and Bayesian methods based on the likelihood, the minimum distance estimate (MDE) is guaranteed to be within an L_1 distance bound from f for a given n, no matter what the underlying f happens to be, and is thus said to have universal performance guarantees (Devroye and Lugosi, Combinatorial methods in density estimation. Springer, New York, 2001). Using a form of tree matrix arithmetic with RPs, we obtain the first generic constructions of an MDE, prove that it has universal performance guarantees and demonstrate its performance with simulated and real-world data. Our main contribution is a constructive implementation of an MDE histogram that can handle large multivariate data bursts using a tree-based partition that is computationally conducive to subsequent statistical operations.

Keywords Rooted planar binary tree · Yatracos class · Tree matrix arithmetic · Model selection · Regular paving · Density estimation

1 Introduction

Suppose our random variable X has an unknown density f on $\mathbb{R}^d$, then for all Borel sets $A \subseteq \mathbb{R}^d$,

$$\mu(A) := \Pr\{X \in A\} = \int_A f(x)\mathrm{d}x.$$

✉ Raazesh Sainudiin
raazesh.sainudiin@math.uu.se

[1] Department of Mathematics, and Combient Competence Centre for Data Engineering Sciences, Uppsala University, Box 480, 751 06 Uppsala, Sweden

[2] Department of Mathematics, Xiamen University Malaysia, Sepang, Selangor, Malaysia

Any density estimate $f_n(x) := f_n(x; X_1, X_2, \dots, X_n) : \mathbb{R}^d \times \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ is a map from $\left(\mathbb{R}^d\right)^{n+1}$ to $\mathbb{R}$. The objective in density estimation is to estimate the unknown f from an independent and identically distributed (IID) sample $X_1, X_2, \dots, X_n$ drawn from f. Density estimation is often the first step in many learning tasks, including anomaly detection, classification, regression and clustering.

The quality of f_n is naturally measured by how well it performs the assigned task of computing the probabilities of sets under the total variation criterion:

$$\mathrm{TV}(f_n, f) = \sup_{A \in \mathcal{B}^d} \left| \int_A f_n - \int_A f \right| = \frac{1}{2} \int |f_n - f|,$$

where $\mathcal{B}^d$ are Borel sets in $\mathbb{R}^d$. The last equality above is due to Scheffé's identity and this equates the L_1 distance between f_n and f, in the absolute scale of [0, 1], to the total variation distance between them.

A non-parametric density estimator is said to have universal performance guarantees when the underlying f is allowed to be any density in L_1 (Devroye and Lugosi 2001, p. 1). Histograms and kernel density estimators can approximate f in this universal sense in an asymptotic setting, i.e., as the number of data points n approaches infinity (the so-called asymptotic consistency of the estimator f_n). But for a fixed n, however large but finite, classical studies of the rate of convergence of f_n to f require additional assumptions on the smoothness class (to solve this so-called smoothing problem), such as $f \in L_2 \neq L_1$ or $f \in C^k$, the set of k times differentiable functions, as opposed to letting f simply belong to the set where densities exist, i.e., $f \in L_1$, and thereby violate the universality property.

Universal performance guarantee is provided by the minimum distance estimate (MDE) due to Devroye and Lugosi (2001, 2004). Their fundamentally combinatorial approach combined ideas from Yatracos (1985, 1988) on minimum distance methods and from Vapnik and Chervonenkis (1971) on uniform convergence of empirical probability measures over classes of sets. See Devroye and Lugosi (2001) for a self-contained introduction to combinatorial methods in density estimation. Unlike the likelihood based methods, MDE gives universal performance guarantees, i.e., MDE does not assume that f is in L_2 in order to address the smoothing problem for the given sample of size n, by directly minimizing the L_1 distance over the so-called Yatracos class—a certain class of subsets of the support set that are induced by the partitions of each ordered pair of histograms in the set of histograms from which one has to choose the optimally smoothed histogram (Devroye and Lugosi 2001).

The Yatracos class is not trivial to represent for the purposes of concretely obtaining the MDE in a nonparametric multivariate setting involving large sample sizes. The particular class of MDEs studied in Devroye and Lugosi (2001, 2004) were limited to kernel estimates and histograms under simpler partitioning rules. Inspired by this, here we develop an MDE over statistical regular pavings using tree-based partitioning strategies to produce a much more general nonparametric MDE that has (1) data-adaptive partitions (2) in d dimensions with (3) partitioning structures imbued with arithmetic for downstream statistical operations. Briefly, our approach exploits a recursive arithmetic using nodes imbued with recursively computable statistics and a specialized collator structure to compute the supremal deviation of the held-out empirical measure over the Yatracos class of the candidate densities.

Unlike other tree-based partitions, our regular paving structure restricts partitioning by only bisecting a box along its first widest coordinate to make the countable set of such trees closed under addition and scalar multiplication and thereby allowing for computationally efficient computer arithmetic over a dense set of simple functions. See Harlow et al. (2012) for statistical applications of this arithmetic, including conditional density regression and multivariate tail probability computations for anomaly detection. Although a more efficient algorithm (up to pre-processing the L_1 distances for each pair of densities) is characterized in Mahalanabis and Stefankovic (2008), we are not aware of any publicly available implementations of the MDE using data-adaptive multivariate histograms for bursts of data common in many industrial applications today, especially for downstream statistical operations with the density estimate, including anomaly detection (with $n \approxeq 10^7$ in dimensions up to 6 for instance in a non-distributed computational setting over one commodity machine).

To the best of our knowledge, the accompanying code of this paper in `mrs2` Sainudiin et al. (2008–2019) is the only publicly available implementation of such an MDE estimator. Our main contribution in this work is a rigorous implementation of the minimum distance estimate proposed by Devroye and Lugosi (2001) for the nonparametric multivariate setting that can handle large bursts of data. The estimator has been successfully used in industry-scale problems where one needs to construct a multivariate density estimate in a "batch" setting and use this estimate for producing anomaly scores.

In the next two sections, we give the definitions, algorithms, theorems and proofs needed for our minimum distance estimator. Three core algorithms are given in the Appendix for completeness. We finally conclude after evaluating the performance of the estimator on simulated and real-world datasets.

2 Regular pavings and histograms

Let $\boldsymbol{x} := [\underline{x}, \overline{x}]$ be a compact real interval with lower bound $\underline{x}$ and upper bound $\overline{x}$, where $\underline{x} \leq \overline{x}$. Let the space of such intervals be $\mathbb{IR}$. The width of an interval $\boldsymbol{x}$ is $\mathrm{wid}\,(\boldsymbol{x}) := \overline{x} - \underline{x}$. The midpoint is $\mathrm{mid}\,(\boldsymbol{x}) := \left(\underline{x} + \overline{x}\right)/2$. A box of dimension d with coordinates in $\Delta := \{1, 2, \ldots, d\}$ is an interval vector with $\imath$ as the first coordinate of maximum width:

$$\boldsymbol{x} := [\underline{x}_1, \overline{x}_1] \times \cdots \times [\underline{x}_d, \overline{x}_d] =: \bigotimes_{j \in \Delta} [\underline{x}_j, \overline{x}_j], \imath := \min\left(\underset{i}{\operatorname{argmax}} (\mathrm{wid}\,(\boldsymbol{x}_i))\right).$$

The set of all such boxes is $\mathbb{IR}^d$, i.e., the set of all interval real vectors in dimension d. A *bisection* or *split* of $\boldsymbol{x}$ perpendicularly at the mid-point along this first widest coordinate $\imath$ gives the left and right child boxes of $\boldsymbol{x}$:

$$\begin{aligned}
\boldsymbol{x}_{\mathsf{L}} &:= [\underline{x}_1, \overline{x}_1] \times \cdots \times [\underline{x}_\imath, \mathrm{mid}\,(\boldsymbol{x}_\imath)) \times [\underline{x}_{\imath+1}, \overline{x}_{\imath+1}] \times \cdots \times [\underline{x}_d, \overline{x}_d], \\
\boldsymbol{x}_{\mathsf{R}} &:= [\underline{x}_1, \overline{x}_1] \times \cdots \times [\mathrm{mid}\,(\boldsymbol{x}_\imath), \overline{x}_\imath] \times [\underline{x}_{\imath+1}, \overline{x}_{\imath+1}] \times \cdots \times [\underline{x}_d, \overline{x}_d].
\end{aligned}$$

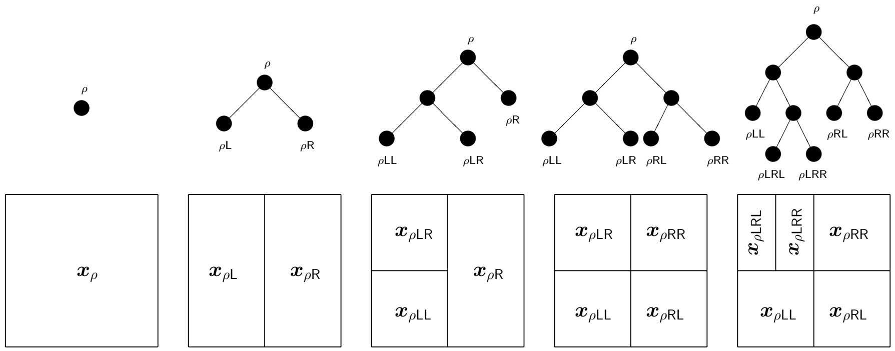

ρ
ρ
ρL
ρR
ρ
ρLL
ρLR
ρR
ρ
ρLL
ρLR
ρRL
ρRR
ρ
ρLL
ρLRL
ρLRR
ρRL
ρRR
$\boldsymbol{x}_\rho$
$\boldsymbol{x}_{\rho\mathsf{L}}$
$\boldsymbol{x}_{\rho\mathsf{R}}$
$\boldsymbol{x}_{\rho\mathsf{LR}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$
$\boldsymbol{x}_{\rho\mathsf{R}}$
$\boldsymbol{x}_{\rho\mathsf{LR}}$
$\boldsymbol{x}_{\rho\mathsf{RR}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$
$\boldsymbol{x}_{\rho\mathsf{RL}}$
$\boldsymbol{x}_{\rho\mathsf{LRL}}$
$\boldsymbol{x}_{\rho\mathsf{LRR}}$
$\boldsymbol{x}_{\rho\mathsf{RR}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$
$\boldsymbol{x}_{\rho\mathsf{RL}}$

Fig. 1 A sequence of selective bisections of boxes (nodes) along the first widest coordinate, starting from the root box (root node) in two dimensions, produces an RP

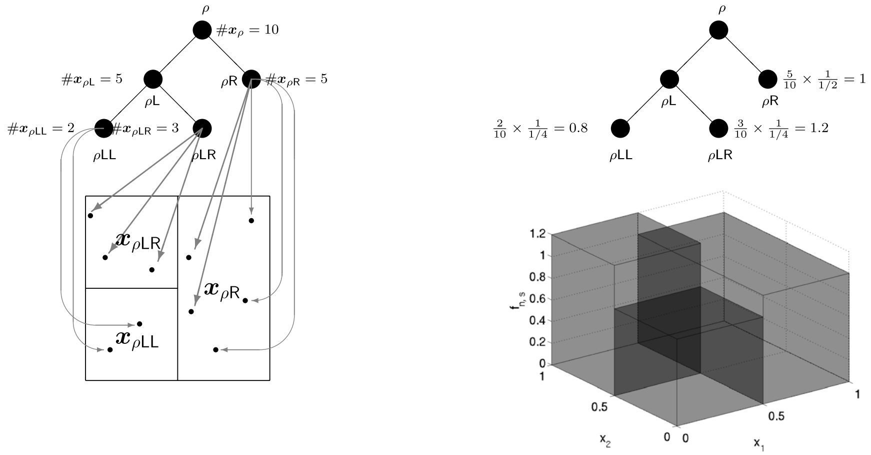
Such a bisection is said to be regular. Note that this bisection gives the left child box a half-open interval $[\underline{x}_\iota, \mathrm{mid}\,(\boldsymbol{x}_\iota))$ on coordinate ι so that the intersection of the left and right child boxes is empty. A recursive sequence of selective regular bisections of boxes, with possibly open boundaries, along the first widest coordinate, starting from the root box $\boldsymbol{x}_\rho$ in $\mathbb{IR}^d$ is known as a *regular paving* (Kieffer et al. 2001) or n-tree (Samet 1990) of $\boldsymbol{x}_\rho$. A regular paving of $\boldsymbol{x}_\rho$ can also be seen as a binary tree formed by recursively bisecting the box $\boldsymbol{x}_\rho$ at the root node. Each node in the binary tree has either no children or two children. These trees are known as plane binary trees in enumerative combinatorics (Stanley 1999, Ex. 6.19(d), p. 220) and as finite, rooted binary trees (frb-trees) in geometric group theory (Meier 2008, Chap. 10). The relationship of trees, labels and partitions is illustrated in Fig. 1 via a sequence of bisections of a square (2-dimensional) root box by always bisecting on the first widest coordinate.

Let $\mathbb{N} := \{1, 2, \dots\}$ be the set of natural numbers. Let the jth interval of a box $\boldsymbol{x}_{\rho\mathsf{v}}$ be $[\underline{x}_{\rho\mathsf{v},j}, \overline{x}_{\rho\mathsf{v},j}]$, the volume of a d-dimensional box $\boldsymbol{x}_{\rho\mathsf{v}}$ be $\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}}) = \prod_{j=1}^{d}(\overline{x}_{\rho\mathsf{v},j} - \underline{x}_{\rho\mathsf{v},j})$. Let the set of all nodes, leaf nodes and internal nodes (or splits) of a regular paving s be $\mathbb{V}(s) := \rho \cup \{\rho\{\mathsf{L},\mathsf{R}\}^j : j \in \mathbb{N}\}$, $\mathbb{L}(s)$ and $\breve{\mathbb{V}}(s) := \mathbb{V}(s) \setminus \mathbb{L}(s)$, respectively. The set of leaf boxes of a regular paving s with root box $\boldsymbol{x}_\rho$ is denoted by $\boldsymbol{x}_{\mathbb{L}(s)}$ and it specifies a partition of the root box $\boldsymbol{x}_\rho$. Let $\mathbb{S}_k$ be the set of all regular pavings with root box $\boldsymbol{x}_\rho$ made of k splits. Note that the number of leaf nodes $m = |\mathbb{L}(s)| = k + 1$ if $s \in \mathbb{S}_k$. The number of distinct binary trees with k splits is equal to the Catalan number C_k:

$$C_k = \frac{1}{k+1}\binom{2k}{k} = \frac{(2k)!}{(k+1)!(k!)}. \tag{1}$$

For $i,j \in \mathbb{Z}_+$, where $\mathbb{Z}_+ := \{0, 1, 2, \dots\}$ and $i \le j$, let $\mathbb{S}_{i:j} := \cup_{k=i}^{j}\mathbb{S}_k$ be the set of regular pavings with k splits where $k \in \{i, i+1, \dots, j\}$. Let the set of all regular pavings be $\mathbb{S}_{0:\infty} := \lim_{j\to\infty}\mathbb{S}_{0:j}$.

A statistical regular paving (SRP) denoted by s is an extension of the RP structure that is able to act as a partitioned 'container' and responsive summarizer for multivariate data. An SRP can be used to create a histogram of a data set. A recursively



(a) An SRP tree and its constituents.

(b) An SRP histogram $f_{n,s}$.

Fig. 2 An SRP and its corresponding histogram

computable statistic (Fisher 1925; Gray and Moore 2003) that an SRP node $\rho\mathsf{v}$ caches is $\#\boldsymbol{x}_{\rho\mathsf{v}}$, the count of the number of data points that fell into $\boldsymbol{x}_{\rho\mathsf{v}}$. A leaf node $\rho\mathsf{v}$ with $\#\boldsymbol{x}_{\rho\mathsf{v}} > 0$ is a non-empty leaf node. The set of non-empty leaves of an SRP s is $\mathbb{L}^{+}(s) := \{\rho\mathsf{v} \in \mathbb{L}(s) : \#\boldsymbol{x}_{\rho\mathsf{v}} > 0\} \subseteq \mathbb{L}(s)$.

Figure 2 depicts a small SRP s with root box $\boldsymbol{x}_{\rho} \in \mathbb{R}^2$. The number of sample data points in the root box $\boldsymbol{x}_{\rho}$ is 10. Figure 2a shows the tree, including the count associated with each node in the tree and the partition of the root box represented by the leaf boxes of this tree, with the sample data points superimposed on the boxes. Figure 2b shows how the density estimate is computed from the count and the volume of leaf boxes to obtain the density estimate $f_{n,s}$ as an SRP histogram.

An SRP histogram is obtained from n data points that fell into $\boldsymbol{x}_{\rho}$ of SRP s as follows:

$$f_{n,s}(x) = f_n(x) = \sum_{\rho\mathsf{v} \in \mathbb{L}(s)} \frac{\mathbb{1}_{\boldsymbol{x}_{\rho\mathsf{v}}}(x)}{n} \left(\frac{\#\boldsymbol{x}_{\rho\mathsf{v}}}{\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}})} \right). \tag{2}$$

It is the maximum likelihood estimator over the class of simple (piecewise-constant) functions given the partition $\boldsymbol{x}_{\mathbb{L}(s)}$ of the root box of s. We suppress subscripting the histogram by the SRP s for notational convenience. SRP histograms have some similarities to dyadic histograms [for eg. Klemelä (2009, chap. 18), Lu et al. (2013)]. Both are binary tree-based and partition so that a box may only be bisected at the mid-point of one of its coordinates, but the RP structure restricts partitioning further by only bisecting a box on its first widest coordinate in order to make $\mathbb{S}_{0:\infty}$ closed under addition and scalar multiplication and thereby allowing for computationally efficient computer arithmetic over a dense set of simple functions [see Harlow et al. (2012) for statistical applications of this arithmetic]. Crucially, when data bursts have large sample sizes, this restrictive partitioning does not affect the L_1 errors when compared to a computationally more expensive Bayes estimator (see Sect. 4).

```
ALGORITHM 1: SEBTreeMC($s, \overline{\#}, \overline{m}$)
input     : $s$, initial SRP with root node $\rho$,
            $x = (x_1, x_2, \dots, x_n)$, a data burst of size $n$,
            $\underline{\#} : \mathbb{L}^{\triangledown}(s) \rightarrow \mathbb{R}$, a priority function of counts,
            $\overline{\#}$, maximum value of $\#(\rho\mathsf{v}) \in \mathbb{L}^{\triangledown}(s)$ for any splittable leaf node in the final
SRP,
            $\overline{m}$, maximum number of leaves in the final SRP.
output    : a sequence of SRP states $[s(0), s(1), \dots, s(T)]$ such that $\mathbb{L}^{\triangledown}(s(T)) = \emptyset$ or
            $\#(\rho\mathsf{v}) \leq \overline{\#}\ \forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s(T))$ or $|\mathbb{L}(s(T))| = \overline{m}$ .
initialize: $\boldsymbol{x}_\rho \leftsquigarrow x$, make $\boldsymbol{x}_\rho$ such that $\cup_i^n x_i \subset \boldsymbol{x}_\rho$ if $\nexists$ domain knowledge or historical
            data,
            $s \leftsquigarrow \boldsymbol{x}_\rho$, specify the root box of $s$,
            $\mathbf{s} \leftarrow [s]$
while $\mathbb{L}^{\triangledown}(s) \neq \emptyset$ & $|\mathbb{L}(s)| < \overline{m}$ & $\max_{\rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s)} \#(\rho\mathsf{v}) > \overline{\#}$ do
    $\rho\mathsf{v} \leftarrow$ random_sample $\left( \operatorname{argmax}_{\rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s)} \#(\rho\mathsf{v}) \right)$   // sample uniformly from nodes with
     largest #
    $s \leftarrow s$ with node $\rho\mathsf{v}$ split                 // split the sampled node and update $s$
    s.append($s$)            // append the new SRP state with an additional split
end
```

A statistically equivalent block (SEB) partition of a sample space is some partitioning scheme that results in equal numbers of data points in each element (block) of the partition (Tukey 1947). The output of SEBTreeMC($s, \overline{\#}, \overline{m}$) of Algorithm 1 is $[s(0), s(1), \dots, s(T)]$, a sequence of SRP states visited by a sample path of the Markov chain $\{S(t)\}_{t \in \mathbb{Z}_+}$ on $\mathbb{S}_{0:\overline{m}-1}$, such that, $\mathbb{L}^{\triangledown}(s(T)) = \emptyset$, or $\#(\rho\mathsf{v}) \leq \overline{\#}\ \forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s(T))$, or $|\mathbb{L}(s(T))| = \overline{m}$ and T is a corresponding random stopping time. As the initial state $S(t = 0)$ is the root $s \in \mathbb{S}_0$, the Markov chain $\{S(t)\}_{t \in \mathbb{Z}_+}$ on $\mathbb{S}_{0:\overline{m}-1}$ satisfies $S(t) \in \mathbb{S}_t$ for each $t \in \mathbb{Z}_+$, i.e., the state at time t has $t + 1$ leaves or t splits. The operation may only be considered to be successful if $|\mathbb{L}(s)| \leq \overline{m}$ and $\#\boldsymbol{x}_{\rho\mathsf{v}} \leq \overline{\#}\ \forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s)$. Therefore, the sequence of SRP histogram states visited by SEBTreeMC that successfully terminates at stopping time T will have the terminal histogram with at most $\overline{\#}$ many of the n data points in each of its leaf nodes and with at most $\overline{m}$ many leaf nodes.

Intuitively, SEBTreeMC($s, \overline{\#}, \overline{m}$) prioritizes the splitting of leaf nodes with the largest numbers of data points associated with them. As we will see in Theorem 1, the L_1 consistency of SEBTreeMC requires that $\overline{m}$ must grow sublinearly (i.e., $\overline{m}/n \rightarrow 0$ as $n \rightarrow \infty$) while the volume of leaf boxes shrink such that a combinatorial complexity measure of the partitions in the support of the SEBTreeMC grows sub-exponentially. Figure 3 shows two different SRP histograms constructed using two different values of $\overline{\#}$ for the same dataset of $n = 10^5$ points simulated under the standard bivariate Gaussian density. A small $\overline{\#}$ produces a histogram that is under-smoothed with unnecessary spikes (Fig. 3 left), while the other histogram with a larger $\overline{\#}$ is over-smoothed (Fig. 3 right). We will obtain the minimum distance estimate from the SRP histograms visited by the SEBTreeMC in Theorem 3.

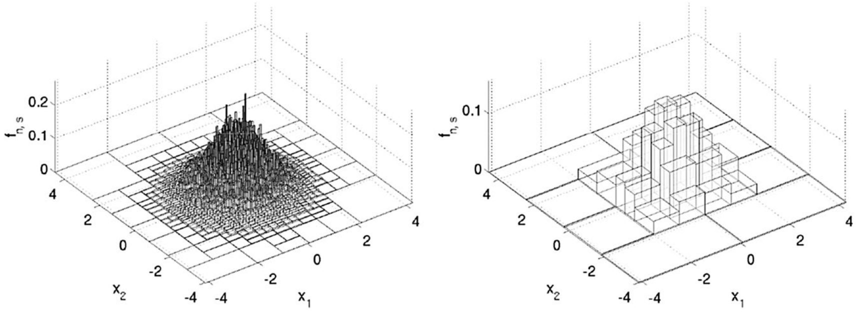


Fig. 3 Two histogram density estimates for the standard bivariate Gaussian density. The left figure shows a histogram with 1485 leaf nodes where $\overline{\#} = 50$ and the histogram on the right has $\overline{\#} = 1500$ resulting in 104 leaf nodes

3 Minimum distance estimation using statistical regular pavings

We show that the SRP density estimate from the SEBTreeMC-based partitioning scheme is asymptotically L_1-consistent as $n \to \infty$ provided that $\overline{\#}$, the maximum sample size in any leaf box in the partition, and $\overline{m}$, the maximum number of leaf boxes in the partition grow with the sample size n at appropriate rates. This is done by proving the three conditions in Theorem 1 of Lugosi and Nobel (1996). We will need to show that as the number of sample points increases linearly, the following conditions are met:

1. the number of leaf boxes grows sub-linearly;
2. the partition grows sub-exponentially in terms of a combinatorial complexity measure;
3. and the volume of the leaf boxes in the partition is shrinking.

Let $\{S_n(i)\}_{i=0}^{\dot{I}}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain of algorithm SEBTreeMC. The Markov chain terminates at some state $\dot{s}$ with partition $\mathbb{L}(\dot{s})$. Associated with the Markov chain is a fixed collection of partitions:

$$\mathcal{L}_n := \left\{\mathbb{L}(\dot{s}) : \dot{s} \in \mathbb{S}_{0:\infty}, \Pr\{S(\dot{I}) = \dot{s}\} > 0\right\},$$

and the size of the largest partition $\mathbb{L}(\dot{s})$ in $\mathcal{L}_n$ is given by

$$m(\mathcal{L}_n) := \sup_{\mathbb{L}(\dot{s}) \in \mathcal{L}_n} |\mathbb{L}(\dot{s})| \leq \overline{m},$$

such that $\mathcal{L}_n \subseteq \{\mathbb{L}(s) : s \in \mathbb{S}_{0:\overline{m}-1}\}$.

Given n fixed points $\{x_1, \dots, x_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Pi\left(\mathcal{L}_n, \{x_1, \dots, x_n\}\right)$ be the number of distinct partitions of the finite set $\{x_1, \dots, x_n\}$ that are induced by partitions $\mathbb{L}(\dot{s}) \in \mathcal{L}_n$:

$$\Pi(\mathcal{L}_n, \{x_1, \dots, x_n\}) := \left|\{\{\boldsymbol{x}_{\rho v} \cap \{x_1, \dots, x_n\} : \boldsymbol{x}_{\rho v} \in \mathbb{L}(\dot{s})\} : \mathbb{L}(\dot{s}) \in \mathcal{L}_n\}\right|.$$

For any fixed set of n points, the growth function of $\mathcal{L}_n$ is then

$$\Pi^*(\mathcal{L}_n, \{x_1, \dots, x_n\}) = \max_{\{x_1,\dots,x_n\} \in \left(\mathbb{R}^d\right)^n} \Pi(\mathcal{L}_n, \{x_1, \dots, x_n\}).$$

Let $A \subseteq \mathbb{R}^d$. Then, the diameter of A is the maximum Euclidean distance between any two points of A, i.e., $\mathrm{diam}(A) := \sup_{x,y \in A} \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$. Thus, for a box $\boldsymbol{x} = [\underline{x}_1, \overline{x}_1] \times \cdots \times [\underline{x}_d, \overline{x}_d]$, $\mathrm{diam}(\boldsymbol{x}) = \sqrt{\sum_{i=1}^{d} (\overline{x}_i - \underline{x}_i)^2}$.

Theorem 1 (L_1-Consistency) *Let $X_1, X_2, \dots$ be independent and identical random vectors in $\mathbb{R}^d$ whose common distribution μ has a non-atomic density f, i.e., $\mu \ll \lambda$. Let $\{S_n(i)\}_{i=0}^{\dot{I}}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using* `SEBTreeMC` *(Algorithm 1) with terminal state $\dot{s}$ and histogram estimate $f_{n,\dot{s}}$ over the collection of partitions $\mathcal{L}_n$. As $n \to \infty$, if $\overline{\#} \to \infty$, $\overline{\#}/n \to 0$, $\overline{m} \geq n/\overline{\#}$, and $\overline{m}/n \to 0$, then the density estimate $f_{n,\dot{s}}$ is asymptotically consistent in L_1, i.e.,*

$$\int |f(x) - f_{n,\dot{s}}(x)| dx \to 0 \text{ with probability 1.}$$

Proof We will assume that $\overline{\#} \to \infty$, $\overline{\#}/n \to 0$, $\overline{m} \geq n/\overline{\#}$, and $\overline{m}/n \to 0$, as $n \to \infty$, and show that the three conditions:

(a) $n^{-1} m(\mathcal{L}_n) \to 0$,
(b) $n^{-1} \log \Pi_n^*(\mathcal{L}_n) \to 0$, and
(c) $\mu(x : \mathrm{diam}(\boldsymbol{x}(x)) > \gamma) \to 0$ with probability 1 for every $\gamma > 0$,

are satisfied. Then, by Theorem 1 of Lugosi and Nobel (1996) our density estimate $f_{n,\dot{s}}$ is asymptotically consistent in L_1.

Condition (a) is satisfied by the assumption that $\overline{m}/n \to 0$ since $m(\mathcal{L}_n) \leq \overline{m}$.

The largest number of distinct partitions of any n point subset of $\mathbb{R}^d$ that are induced by the partitions in $\mathcal{L}_n$ is upper bounded by the size of the collection of partitions $\mathcal{L}_n \subseteq \mathbb{S}_{0:\overline{m}-1}$, i.e.,

$$\Pi_n^*(\mathcal{L}_n) \leq |\mathcal{L}_n| \leq \sum_{k=0}^{\overline{m}-1} C_k,$$

where k is the number of splits.

The growth function is thus bounded by the total number of partitions with 0 to $\overline{m} - 1$ splits, i.e., the $(\overline{m} - 1)$th partial sum of the Catalan numbers. The partial sum can be asymptotically equivalent to (Mattarei 2010):

$$\sum_{k=0}^{\overline{m}-1} C_k \sim \frac{4^{\overline{m}}}{\left(3(\overline{m} - 1)\sqrt{\pi(\overline{m} - 1)}\right)} \quad \text{as } \overline{m} \to \infty.$$

Taking logs and dividing by n on both sides of the above two equations, and using the assumption that $\overline{m}/n \to 0$ as $n \to \infty$, we can see that condition (b) is satisfied

$$\log \mathit{\Pi}_n^*(L_n)/n \leq \log(|\mathcal{L}_n|)/n \to \tfrac{1}{n}\left(\overline{m}\log 4 - \tfrac{3}{2}\log(\overline{m}-1) - \log 3\sqrt{\pi}\right) \to 0.$$

We now prove the final condition. Fix $\gamma, \xi > 0$. There exists a box $\check{\boldsymbol{x}} = [-M, M]^d$ for a large enough M, such that $\mu(\check{\boldsymbol{x}}^c) < \xi$, where $\check{\boldsymbol{x}}^c := \mathbb{R}^d \setminus [-M, M]^d$. Consequently

$$\mu(\{x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma\}) \leq \xi + \mu(\{x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma\} \cap \check{\boldsymbol{x}}).$$

Using 2^{di} hypercubes of equal volume $(2M)^d/2^{di}, i = \left\lceil \log_2\left(2M\sqrt{d}/\gamma\right)\right\rceil$ with side length $2M/2^i$ and diameter $\sqrt{d\left(\frac{2M}{2^i}\right)^2}$, we can have at most $m_\gamma < 2^{di}$ boxes in $\check{\boldsymbol{x}}$ that have diameter greater than γ. By choosing i large enough, we can upper bound m_γ by $(2M\sqrt{d}/\gamma)^d$, a quantity that is independent of n, such that

$$\begin{aligned}
\mu(x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma) \leq & \xi + \mu(\{x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma\} \cap \check{\boldsymbol{x}}) \\
\leq & \xi + m_\gamma\left(\max_{\boldsymbol{x}\in\mathbb{L}(\dot{s})} \mu(\boldsymbol{x})\right) \\
\leq & \xi + m_\gamma\left(\max_{\boldsymbol{x}\in\mathbb{L}(\dot{s})} \mu_n(\boldsymbol{x}) + \max_{\boldsymbol{x}\in\mathbb{L}(\dot{s})} \left|\mu(\boldsymbol{x}) - \mu_n(\boldsymbol{x})\right|\right), \\
& \qquad\qquad \text{where, } \mu_n(\boldsymbol{x}) := \frac{\#(\boldsymbol{x})}{n} \\
\leq & \xi + m_\gamma\left(\frac{\overline{\#}}{n} + \sup_{\boldsymbol{x}\in\mathbb{IR}^d} |\mu(\boldsymbol{x}) - \mu_n(\boldsymbol{x})|\right).
\end{aligned}$$

The first term in the parenthesis converges to zero since $\overline{\#}/n \to 0$ by assumption. For $\epsilon > 0$ and $n > 4d$, the second term goes to zero by applying the Vapnik–Chervonenkis (VC) theorem to boxes in $\mathbb{IR}^d$ with VC dimension $2d$ and shatter coefficient $S(\mathbb{IR}^d, n) \leq (en/2d)^{2d}$ (Devroye et al. 1996, Thms. 12.5, 13.3 and p. 220), i.e.,

$$\Pr\left\{\sup_{\boldsymbol{x}\in\mathbb{IR}^d} \left|\mu_n(\boldsymbol{x}) - \mu(\boldsymbol{x})\right| > \epsilon\right\} \leq 8 \cdot (en/2d)^{2d} \cdot e^{-n\epsilon^2/32}.$$

For any $\epsilon > 0$ and finite d, the right-hand side of the above inequality can be made arbitrarily small for n large enough. This convergence in probability is equivalent to the following almost sure convergence by the bounded difference inequality:

$$\lim_{n\to\infty} \sup_{\boldsymbol{x}\in\mathbb{IR}^d} \left|\mu_n(\boldsymbol{x}) - \mu(\boldsymbol{x})\right| = 0 \quad \text{w.p. } 1.$$

Thus, for any $\gamma, \xi > 0$,

$$\lim_{n\to\infty} \mu(\{x : \operatorname{diam}(\boldsymbol{x}(x)) > \gamma\}) \leq \xi \quad \text{w.p. } 1.$$

Therefore, condition (c) is satisfied and this completes the proof. □

Let Θ index a set of finitely many density estimates: $\{f_{n,\theta} : \theta \in \Theta\}$, such that $\int f_{n,\theta} = 1$ for each $\theta \in \Theta$. We can index the SRP trees by $\{s_\theta : \theta \in \Theta\}$, where θ is the sequence of leaf node depths that uniquely identifies the SRP tree, and denote the density estimate corresponding to s_θ by f_{n,s_θ} or simply by $f_{n,\theta}$. Now, consider the asymptotically consistent path taken by the Markov chain of SEBTreeMC. For a fixed sample size n, let $\{s_\theta : \theta \in \Theta\}$ be an ordered subset of states visited by the Markov chain, with $s_\theta \prec s_\vartheta$ if s_ϑ is a refinement of s_θ, i.e., if s_θ is visited before s_ϑ. The goal is to select the optimal estimate from $|\Theta|$ many candidates.

When our candidate set of densities are additive like the histograms, we can use the hold-out method proposed by Devroye and Lugosi (2001, Sec. 10.1) for minimum distance estimation as follows. Let $0 < \varphi < 1/2$. Given n data points, use $n - \varphi n$ points as the training set and the remaining φn points as the validation set (by φn we mean $\lfloor \varphi n \rfloor$). Denote the set of training data by $\mathcal{T} := \{x_1, \dots, x_{n-\varphi n}\}$ and the set of validation data by $\mathcal{V} := \{x_{n-\varphi n+1}, \dots, x_n\} = \{y_1, \dots, y_{\varphi n}\}$. For an ordered pair $(\theta, \vartheta) \in \Theta^2$, with $\theta \neq \vartheta$, the set

$$A_{\theta,\vartheta} := A\big(f_{n-\varphi n,\theta}, f_{n-\varphi n,\vartheta}\big) := \big\{x : f_{n-\varphi n,\theta}(x) > f_{n-\varphi n,\vartheta}(x)\big\},$$

is known as a *Scheffé set*. The *Yatracos class* (Yatracos 1985) is the collection of all such Scheffé sets over Θ:

$$\mathcal{A}_\Theta = \big\{\big\{x : f_{n-\varphi n,\theta}(x) > f_{n-\varphi n,\vartheta}(x)\big\} : (\theta, \vartheta) \in \Theta^2, \theta \neq \vartheta\big\}.$$

Let $\mu_{\varphi n}$ be the empirical measure of the validation set $\mathcal{V}$. Then, the *minimum distance estimate* or MDE $f_{n-\varphi n,\theta^*}$ is the density estimate $f_{n-\varphi n,\theta}$ constructed from the training set $\mathcal{T}$ with the smallest index θ^* that minimizes:

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-\varphi n,\theta}(A) - \mu_{\varphi n}(A) \right|. \tag{3}$$

Thus, the MDE $f_{n-\varphi n,\theta^*}$ minimizes the supremal absolute deviation from the held-out empirical measure $\mu_{\varphi n}$ over the Yatracos class $\mathcal{A}_\Theta$.

The SRP is adapted for MDE to mutably cache the counts for training and validation data separately and the $n - \varphi n$ training data points in $\mathcal{T}$ and the φn validation data points in $\mathcal{V}$ are accessible from any leaf node ρv of the SRP via pointers to $x_i \in \mathcal{T}$ and $y_i \in \mathcal{V}$, respectively. The training data drive the Markov chain SEBTreeMC$(s, \overline{\#}, \overline{m})$ to produce a sequence of SRP states: $s_{\theta_1}, s_{\theta_2}, \dots$ that are further selected to build the candidate set of adaptive histogram density estimates given by $\{f_{n-\varphi n,\theta_i} : \theta_i \in \Theta\}$. For each $\theta_i \in \Theta$, the validation data are allowed to flow through s_{θ_i} and drop into the leaf boxes of s_{θ_i}. A graphical representation of an SRP with training counter $\#\boldsymbol{x}_{\rho v}$ and validation counter $\check{\#}\boldsymbol{x}_{\rho v}$ is shown in Fig. 4. Computing the MDE objective Δ_{θ_i} in (3) requires the histogram estimate $f_{n-\varphi n}(\rho v) = \#\boldsymbol{x}_{\rho v}/n\lambda(\boldsymbol{x}_{\rho v})$ and the empirical measure of the validation data $\mu_{\varphi n}(\boldsymbol{x}_{\rho v}) = \check{\#}\boldsymbol{x}_{\rho v}/\varphi n$ at any node ρv. These can be readily obtained from $\#\boldsymbol{x}_{\rho v}$ and $\check{\#}\boldsymbol{x}_{\rho v}$.

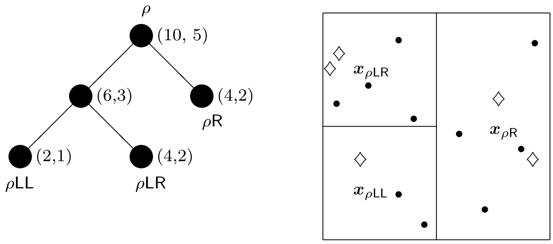

ρ
(10, 5)
(6,3)
(4,2)
ρR
(2,1)
(4,2)
ρLL
ρLR
$\boldsymbol{x}_{\rho\mathsf{LR}}$
$\boldsymbol{x}_{\rho\mathsf{R}}$
$\boldsymbol{x}_{\rho\mathsf{LL}}$

Fig. 4 An SRP s with training (•) and validation data (◇) and their respective sample counts $(\#\boldsymbol{x}_{\rho v}, \check{\#}\boldsymbol{x}_{\rho v})$ that are updated recursively as data fall through the nodes of s

Our approach to obtaining the MDE $f_{n-\varphi n,\theta^*}$ with optimal SRP s_{θ^*} exploits the partition refinement order in $\{s_\theta : \theta \in \Theta\}$, a subset of states along the path taken by the SEBTreeMC. Using nodes imbued with recursively computable statistics for both training and validation data, and a specialized collation according to `SRPCollate` (Algorithm 3) over SRPs, we compute the objective $\varDelta_\theta$ in (3) using `GetDelta` (Algorithm 2) via a dynamically grown Yatracos Matrix with pointers to all Scheffé sets constituting the Yatracos class according to `GetYatracos` (Algorithm 4). We briefly outline the core ideas in these three algorithms next [see Appendix for their pseudocode and `mrs2` Sainudiin et al. (2008–2019) for details].

In the MDE procedure, pairwise comparisons of the heights of the candidate density estimates $f_{n-\varphi n,\theta}$ and $f_{n-\varphi n,\vartheta}$ are needed to get the Scheffé sets that make up the Yatracos class. An efficient way to approach this is to collate the SRPs corresponding to the density estimates onto a *collator regular paving* (CRP) where the space of CRP trees is also $\mathbb{S}_{0:\infty}$. Consider now two SRPs s_θ and s_ϑ for which the corresponding histogram estimates $f_{n,\theta}$ and $f_{n,\vartheta}$ are computed. Both SRPs s_θ and s_ϑ have the same root box $\boldsymbol{x}_\rho$. By collating the two SRPs, we get a CRP c with the same root box and the tree obtained from a union of s_θ and s_ϑ. Unlike the union operation over RPs (Harlow et al. 2012, Algorithm 1), each node ρv of the SRP collator c stores $f_{n,\theta}$ and $f_{n,\vartheta}$ as a vector $\boldsymbol{f}_{n,c}(\rho v) := (f_{n,\theta}(\rho v), f_{n,\vartheta}(\rho v))$. The empirical measure of the validation data $\mu_{\varphi n}(\boldsymbol{x}_{\rho v})$ will also be stored at each node ρv and can be easily accessed via pointers. Figure 5 shows how CRP c can collate two SRPs s_θ and s_ϑ using `SRPCollate`.

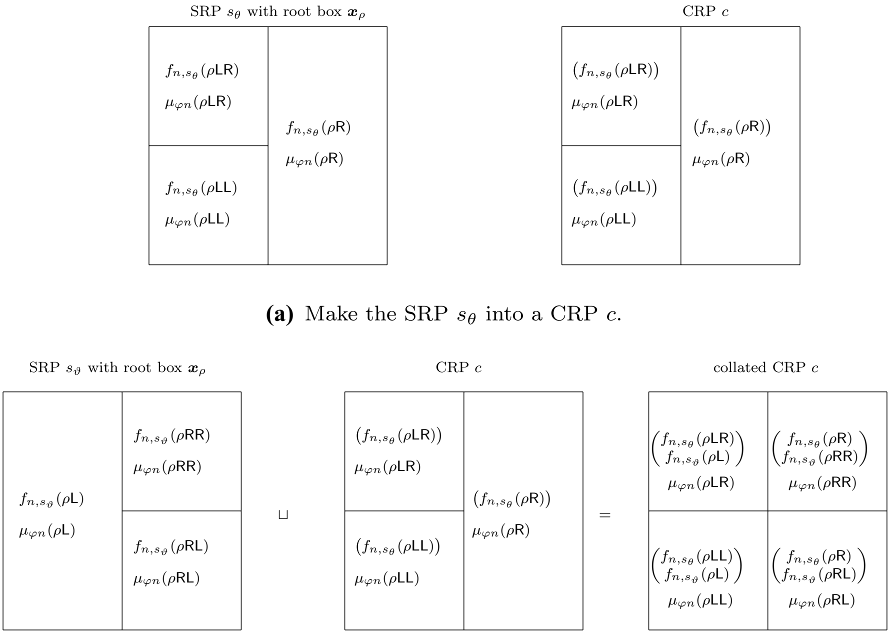
We now use Theorem 10.1 of Devroye and Lugosi (2001, p. 99) and Theorem 6.6 of Devroye and Lugosi (2001, p. 54) to obtain the L_1-error bound of the minimum distance estimate $f_{n-\varphi n,\theta^*}$, with $\theta^* \in \Theta$ and $|\Theta| < \infty$.

Theorem 2 *If $\int f_{n-\varphi n,\theta} = 1$ for all $\theta \in \Theta$, then for the minimum distance estimate $f_{n-\varphi n,\theta^*}$ obtained by minimizing $\varDelta_\theta$ in (3),we have*

$$\int \left| f_{n-\varphi n,\theta^*} - f \right| \leq 3 \min_{\theta \in \Theta} \int \left| f_{n-\varphi n,\theta} - f \right| + 4\varDelta, \tag{4}$$

where

$$\varDelta = \max_{A \in \mathcal{A}_\Theta} \left| \int_A f - \mu_{\varphi n}(A) \right|. \tag{5}$$


SRP s_θ with root box $\boldsymbol{x}_\rho$
$f_{n,s_\theta}(\rho\mathsf{LR})$
$\mu_{\varphi n}(\rho\mathsf{LR})$
$f_{n,s_\theta}(\rho\mathsf{R})$
$\mu_{\varphi n}(\rho\mathsf{R})$
$f_{n,s_\theta}(\rho\mathsf{LL})$
$\mu_{\varphi n}(\rho\mathsf{LL})$
CRP c
$(f_{n,s_\theta}(\rho\mathsf{LR}))$
$\mu_{\varphi n}(\rho\mathsf{LR})$
$(f_{n,s_\theta}(\rho\mathsf{R}))$
$\mu_{\varphi n}(\rho\mathsf{R})$
$(f_{n,s_\theta}(\rho\mathsf{LL}))$
$\mu_{\varphi n}(\rho\mathsf{LL})$
SRP s_ϑ with root box $\boldsymbol{x}_\rho$
$f_{n,s_\vartheta}(\rho\mathsf{RR})$
$\mu_{\varphi n}(\rho\mathsf{RR})$
$f_{n,s_\vartheta}(\rho\mathsf{L})$
$\mu_{\varphi n}(\rho\mathsf{L})$
$f_{n,s_\vartheta}(\rho\mathsf{RL})$
$\mu_{\varphi n}(\rho\mathsf{RL})$
$\sqcup$
CRP c
$(f_{n,s_\theta}(\rho\mathsf{LR}))$
$\mu_{\varphi n}(\rho\mathsf{LR})$
$(f_{n,s_\theta}(\rho\mathsf{R}))$
$\mu_{\varphi n}(\rho\mathsf{R})$
$(f_{n,s_\theta}(\rho\mathsf{LL}))$
$\mu_{\varphi n}(\rho\mathsf{LL})$
$=$
collated CRP c
$\begin{pmatrix} f_{n,s_\theta}(\rho\mathsf{LR}) \\ f_{n,s_\vartheta}(\rho\mathsf{L}) \end{pmatrix}$
$\mu_{\varphi n}(\rho\mathsf{LR})$
$\begin{pmatrix} f_{n,s_\theta}(\rho\mathsf{R}) \\ f_{n,s_\vartheta}(\rho\mathsf{RR}) \end{pmatrix}$
$\mu_{\varphi n}(\rho\mathsf{RR})$
$\begin{pmatrix} f_{n,s_\theta}(\rho\mathsf{LL}) \\ f_{n,s_\vartheta}(\rho\mathsf{L}) \end{pmatrix}$
$\mu_{\varphi n}(\rho\mathsf{LL})$
$\begin{pmatrix} f_{n,s_\theta}(\rho\mathsf{R}) \\ f_{n,s_\vartheta}(\rho\mathsf{RL}) \end{pmatrix}$
$\mu_{\varphi n}(\rho\mathsf{RL})$

(a) Make the SRP s_θ into a CRP c.

(b) Collate another SRP s_ϑ onto CRP c.

Fig. 5 Collating two SRPs s_θ and s_ϑ with the same root box $\boldsymbol{x}_\rho$

Theorem 2 can be proved directly by a conditional application of Theorem 6.3 of Devroye and Lugosi (2001, p. 54) and is nothing but the finite Θ version of their Theorem 10.1 (Devroye and Lugosi 2001, p. 99) without the additional 3 / n term due to $|\Theta| < \infty$.

When f is unknown and $2^n > |\mathcal{A}_\Theta|$, Δ may be approximated using the cardinality bound (Devroye et al. 1996, Theorem 13.6, p. 219) for the shatter coefficient of $\mathcal{A}_\Theta$. Given $\{x_1, \dots, x_n\}$ the nth shatter coefficient of $\mathcal{A}_\Theta$ is defined as

$$S\big(\mathcal{A}_\Theta, n\big) = \max_{x_1,\dots,x_n \in \mathbb{R}^d} \Big| \big\{ \{x_1, \dots, x_n\} \cap A \,:\, A \in \mathcal{A}_\Theta \big\} \Big|.$$

Since $\mathcal{A}_\Theta$ is finite, containing at most quadratically many Scheffé sets $A_{\theta,\vartheta}$ with distinct ordered pairs $(\theta, \vartheta) \in \Theta^2$ given by the non-diagonal elements of the Yatracos matrix returned by `GetYatracos` , by Theorem 13.6 of Devroye et al. (1996, p. 219) its nth shatter coefficient is bounded as follows:

$$S\big(\mathcal{A}_\Theta, n\big) \leq |\mathcal{A}_\Theta| \leq (|\Theta| + 1)^2 - (|\Theta| + 1) = |\Theta|(|\Theta| + 1). \tag{6}$$

Finally, given that adaptive multivariate histograms based on statistical regular pavings in $\mathbb{S}_{0:\infty}$ form a class of regular additive density estimates, we can slightly modify Theorem 10.3 of Devroye and Lugosi (2001, p. 103) for the case with finite Θ to get the following error bound that further accounts for splitting the data.

Theorem 3 *Let $0 < \varphi < 1/2$ and $n < \infty$. Let the finite set Θ determine a class of adaptive multivariate histograms based on statistical regular pavings with $\int f_{n-\varphi n,\theta} = 1$ for all $\theta \in \Theta$. Let f_{n,θ^*} be the minimum distance estimate. Then for all $n, \varphi n, \Theta$ and $f \in L_1$:*

$$\begin{aligned} E\left\{ \int \left| f_{n-\varphi n,\theta^*} - f \right| \right\} &\leq 3 \min_\theta E\left\{ \int \left| f_{n,\theta} - f \right| \right\} \left(1 + \frac{2\varphi}{1-\varphi} + 8\sqrt{\varphi} \right) \\ &+ 8\sqrt{\frac{\log 2|\Theta|(|\Theta|+1)}{\varphi n}}. \end{aligned} \tag{7}$$

Proof By Theorem 2,

$$\int \left| f_{n-\varphi n,\theta^*} - f \right| \leq 3 \min_\theta \int \left| f_{n-\varphi n,\theta} - f \right| + 4\Delta$$

Taking expectations on both sides and using Theorem 10.2 in Devroye and Lugosi (2001, p. 99)

$$\begin{aligned} E\left\{ \int \left| f_{n-\varphi n,\theta^*} - f \right| \right\} &\leq 3 \min_\theta E\left\{ \int \left| f_{n-\varphi n,\theta} - f \right| \right\} + 4E\Delta \\ &\leq 3 \min_\theta E\left\{ \int \left| f_{n,\theta} - f \right| \right\} \left(1 + \frac{2\varphi n}{(1-\varphi)n} + 8\sqrt{\frac{\varphi n}{n}} \right) + 4E\Delta. \end{aligned}$$

Finally, by Theorem 3.1 in Devroye and Lugosi (2001, p. 18) and (6),

$$\begin{aligned} 4E\Delta = 4E\left\{ \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f - \mu_{\varphi n}(A) \right| \right\} &\leq 4 \cdot 2 \cdot \sqrt{\frac{\log 2S(\mathcal{A}_\Theta, \varphi n)}{\varphi n}} \\ &\leq 4 \cdot 2 \cdot \sqrt{\frac{\log 2|\Theta|(|\Theta|+1)}{\varphi n}}. \end{aligned}$$

□

4 Performance evaluation

4.1 Practical minimum distance estimation

To effectively use the error bound, we need to ensure that $|\Theta|$ is not too large and the densities in Θ are close to the true density f. Next, we highlight the effectiveness and limitations of our MDE.

The size of Θ is kept small (typically less than 100) and independent of n by an adaptive search. Note that $|\Theta|$ is upper-bounded by $\overline{m}$ if we were to exhaustively consider each SRP state along the entire path of the `SEBTreeMC` in Θ, our set of candidate SRP partitions. Such an exhaustive approach is computationally inefficient as the Yatracos matrix that updates the Scheffé sets grows quadratically with $|\Theta|$. We take a simple adaptive search approach by considering only k (typically $10 \leq k \leq 20$) SRP states in each iteration. In the initial iteration, we add k states to Θ by picking uniformly spaced states from a long-enough `SEBTreeMC` path that starts from the root node and ends at a state with a large number of leaves and a significantly higher Δ_θ score than its preceding states. Then, we simply zoom-in around the states with the lowest Δ_θ values and add another k states along the same `SEBTreeMC` path close to such optimal states from the first iteration. We repeat this adaptive search process until we are unable to zoom-in further. Typically, we are able to find nearly optimal states within 5 or fewer iterations. By Theorem 1, we know that the histogram partitioning strategy of `SEBTreeMC` is asymptotically consistent. Thus, the adaptive search set Θ that is selected iteratively from the set of histogram states along the path of `SEBTreeMC` with optimal Δ_θ values will naturally contain densities that approach f as n increases. However, the rate at which the L_1 distance between the best density in Θ and f approach 0 will depend on the complexity of f in terms of the number of leaves needed to uniformly approximate f using simple functions with SRP partitions, a class that is dense in $\mathcal{C}(\boldsymbol{x}_\rho, \mathbb{R})$, the algebra of real-valued continuous functions over the root box $\boldsymbol{x}_\rho$ by the Stone–Weierstrass Theorem (Harlow et al. 2012, Theorem 4.1). This dependence on the structural complexity of f is evaluated next.

4.2 Simulations

To evaluate the performance of our MDE we first choose the unstructured multivariate uniform density. Although the dimension d of the uniform density on $[0,1]^d$ ranges in $\{1, 10, 100, 1000\}$, the true density is given by the root box, the first candidate density indexed by Θ. Based on the mean integrated absolute errors (MIAE) shown in Table 1 for each d and n in $\{10^2, 10^3, 10^4, 10^5, 10^6, 10^7\}$, there is a dimension-free performance by the MDE for such a target density that immediately belongs to the set of candidate densities indexed by Θ. The sample mean of the integrated absolute errors was taken over five replicate simulations with standard error less than half of the MIAE values. When the sample size is 10^7 and dimension is 1000, the data cannot be represented in a machine with 32GB of memory (as indicated by the '–' entry in Table 1).

We independently verified the inequalities in Eqs. 4 and 7 of Theorems 2 and 3, respectively, by explicitly computing the left and right hand-sides of the inequalities for MDEs and checking that they are indeed satisfied for the simulated datasets in the previous sub-section. This verification is in `examples/StatsSubPav/MinimumDistanceEstimation/MDETest` of the `mrs2` module `companions/mrs-1.0-YatracosThis/`. These results are shown for multivariate uniform densities in Fig. 6. The bounds are not sharp although they do decrease

Table 1 The MIAE for MDE with different sample sizes n for the 1D-, 10D-, 100D- and 1000D-Uniform densities

Dimension	Sample size n					
d	10^7	10^6	10^5	10^4	10^3	10^2
1	3.439e – 04	1.981e – 03	2.866e – 03	1.405e – 02	3.237e – 02	1.000e – 01
10	3.606e – 04	1.803e – 03	2.689e – 03	1.156e – 02	3.191e – 02	1.470ee – 01
100	3.446e – 04	1.908e – 03	2.953e – 03	1.540e – 02	2.898e – 02	1.520e – 01
1000	–	1.720e – 03	2.576e – 03	1.619e – 02	2.998e – 02	1.125e – 01

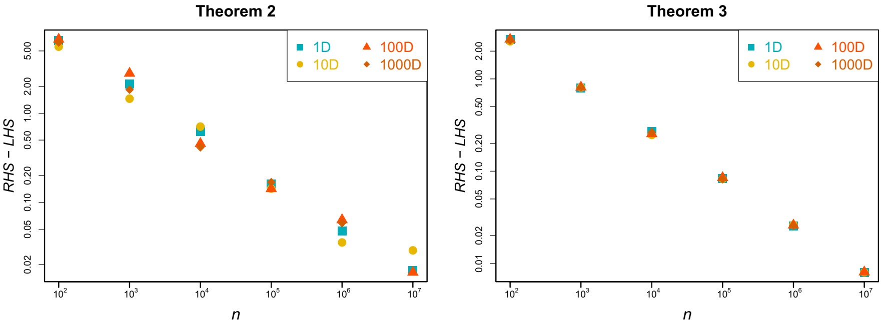


Fig. 6 Computational verification of the inequalities of Theorems 2 and 3 for the multivariate uniform density in various dimensions over several sample sizes

with the sample size. This is because they are extremely general by construction and based merely on the cardinality of the set of candidate densities.

To evaluate the performance of our MDE, we also chose two structured multivariate densities: the spherically symmetric Gaussian with a simple concentrated structure and the highly structured Rosenbrock density [whose expression up to normalization is given in (8)] in d dimensions for various sample sizes:

$$\exp\left(-\sum_{i=2}^{d}(100(x_i - x_{i-1}^2)^2 + (1 - x_{i-1})^2)\right). \tag{8}$$

The sample standard deviations about the mean integrated absolute errors or MIAEs for the MDE method, i.e., $L_1(f_{n,\theta^*}, f)$ (shown in the top panel of Table 2), based on ten trials, are below 10^{-3} and 10^{-4} for values of n in $\{10^4, 10^5\}$ and $\{10^6, 10^7\}$, respectively. Thus, these standard errors are not shown. However, the L_1 distance between the MDE and the best estimate in the candidate set Θ, $L_1(f_{n,\theta^*}, f) - \min_{\theta\in\Theta} L_1(f_{n,\theta}, f)$, is shown in Table 2 for each density and sample size. For comparison, as shown in the bottom panel of Table 2, we used the Bayes estimator from the posterior mean histograms (Sainudiin et al. 2013, see for details on this evaluation). Note how the L_1 errors decrease with the sample size and how the errors

Table 2 The MIAE for MDE and posterior mean estimates with different sample sizes for the 1D-, 2D-, and 5D-Gaussian densities, as well as the 2D- and 5D-Rosenbrock densities

n	Standard Gaussian densities			Rosenbrock densities	
	1D	2D	5D	2D	5D
Minimum distance estimate's mean $L_1(f_{n,\theta^*},f)$, $L_1(f_{n,\theta^*},f) - \min_{\theta\in\Theta} L_1(f_{n,\theta},f)$					
10^2	0.4154, 0.0348	0.6018, 0.0325	1.4944, 0.1093	1.1843, 0.0208	1.6853, 0.0424
10^3	0.2643, 0.0091	0.3515, 0.0144	0.8521, 0.0053	0.7533, 0.0119	1.3323, 0.0061
10^4	0.0888, 0.0058	0.2038, 0.0044	0.6764, 0.0020	0.4502, 0.0050	1.0154, 0.0018
10^5	0.0504, 0.0046	0.1140, 0.0014	0.4744, 0.0006	0.2476, 0.0024	0.7278, 0.0060
10^6	0.0204, 0.0014	0.0656, 0.0014	0.3310, 0.0006	0.1430, 0.0006	0.4772, 0.0034
10^7	0.0100, 0.0004	0.0376, 0.0002	0.2548, 0.0014	0.0828, 0.0012	0.2661, 0.0016
MCMC posterior mean estimate's MIAE (standard error)					
10^4	0.0565 (0.0053)	0.1673 (0.0046)	0.6467 (0.0051)	0.3717 (0.0103)	1.0190 (0.0059)
10^5	0.0274 (0.0011)	0.0932 (0.0002)	0.4655 (0.0020)	0.1982 (0.0067)	0.7250 (0.0011)
10^6	0.0129 (0.0006)	0.0533 (0.0005)	0.3274 (0.0009)	0.1102 (0.0006)	0.4812 (0.0012)
10^7	0.0060 (0.0001)	0.0304 (0.0002)	0.2292 (0.0034)	0.0608 (0.0049)	0.3302 (0.0004)

are comparable between the methods, albeit the MDE method is at least an order of magnitude faster than the posterior mean histogram that does not provide universal performance guarantees like most density estimators.

Due to the use of space-partitioning regularly paved trees, our MDE histograms cannot provide small L_1 errors for highly structured densities beyond 10 or so dimensions on the basis of sample sizes in the order of millions. The reason is simply due to the large L_1 distance between the best candidate density in Θ based on a reasonable maximal number of splits. However, using modern dimensionality reduction techniques including auto-encoders we can often project high dimensional data non-linearly to a lower dimensional space and use the MDE histograms to construct a density estimate and do further statistical processing as we show below.

All experiments were performed on the same physical machine that is currently considered to be commodity hardware (Sainudiin et al. 2013, for machine specifications and CPU times).

4.3 Detecting bot flows using MDE tail probabilities

We apply the MDE histogram on the real-world scenario 11 of the CTU-13 dataset of botnet traffic on a computer network (Garcia et al. 2014). The dataset captured 8164 real botnet traffic flows mixed with 99087 normal and background traffic flows. These flows were augmented into 80 dimensions using Word2Vec embeddings of the flows (datapoints) and reduced to 8 dimensions by training a deep auto-encoder with a bottleneck layer of eight nodes by Ramström (2019), who gives domain-specific details on how the raw data was augmented and fitted to an auto-encoder. For the purpose of our application, it suffices to note that a deep auto-encoder was trained on appropriately augmented normal flows in order to non-linearly reduce the dimensions from 80 to 8. We then used the $n = 99087$ samples in 8 dimensions to obtain the MDE histogram. For each data point x from the normal as well as the botnet flow, we computed its multivariate tail probability (Harlow et al. 2012, Algorithm 9). Briefly, this is given by 1 minus the sum of the probability mass of all leaf nodes in the MDE histogram whose density ("height") is larger than that of the leaf node whose box contains x. The tail probability can be directly used as a score of how unlikely an event is under the density estimate constructed with the normal flows. We obtain these tail probabilities for all 107251 flows (mixed with 7.6% botnet flows) and sort them by their tail probabilities. Our histogram estimate was able to identify 87% and 99.1% of the botnet flows, i.e., 7115 and 8090 out of 8164 botnet flows were within the lowest 7.6% and 10% of the tail probabilities, respectively. Thus, using the tail probabilities of the MDE histogram estimated from the normal flows was extremely effective in identifying the botnet flows.

5 Conclusion and future directions

Thus, using the collator regular paving (CRP), we obtain the minimum distance estimate (MDE) with universal performance guarantees. All the methods are implemented and available in `mrs2` Sainudiin et al. (2008–2019), including the downstream statistical operations for anomaly detection and conditional density regression (Harlow et al. 2012). We limited our minimum distance estimate (MDE) to the candidate set given by the SRP histograms visited along the path of the Markov chain `SEBTreeMC`. This was done to take advantage of the structure of consecutive refinements of the tree partitions along a single path of `SEBTreeMC`.

However, obtaining the MDE from an arbitrary set of SRP histograms taken from $\mathbb{S}_{0:\infty}$ will need more sophisticated collators. Initial experiments using the Scheffé tournament approach (as opposed to the MDE) to find the best estimate in a candidate set of arbitrary SRP histograms (not just those along a path in $\mathbb{S}_{0:\infty}$) look feasible. Such a Scheffé tournament will allow us to compare estimates from entirely different methodological schools (Bayesian, penalized likelihood, etc.). Finally, the pure tree structure allows one to possibly extend this MDE to a distributed fault-tolerant computational setting such as Zaharia et al. (2016) as the sample size becomes too large for the memory of a single machine.

Acknowledgements RS and GT proved the theorems and GT implemented the three MDE algorithms based on codes by Jennifer Harlow and RS in `mrs2`. This research began from a conversation RS had with Luc Devroye at the World Congress in Probability and Statistics in 2008 and was partly supported by RS's external consulting revenues from the New Zealand Ministry of Tourism, UC College of Engineering Sabbatical Grant and Visiting Scholarship at Department of Mathematics, Cornell University, Ithaca NY, USA and completed through the project CORCON: Correctness by Construction, Seventh Framework Programme of the European Union, Marie Curie Actions-People, International Research Staff Exchange Scheme (IRSES) with counter-part funding from the Royal Society of New Zealand. The application to botnet detection was partially supported by Combient Mix AB and the Department of Mathematics, Uppsala University.

Funding This study was funded by FP7 People: Marie-Curie Actions (project CORCON) (Grant number: 612638).

Appendix: MDE algorithms

ALGORITHM 2: `GetDelta`

```
input    :
 1. the current number of splits: $i$;
 2. the collated regular paving CRP: $c$ with pointers to the vector $\boldsymbol{f}_{n-\varphi n,c}(\rho)$
    and $\mu_{\varphi n}(\rho)$ of each node in $c$
 3. the Yatracos matrix: $\mathcal{A}_{\Theta_i}$;
 4. the current $\Delta_\theta$ vector: $\Delta_{\Theta_{i-1}} \in \mathbb{R}^{(1\times(i))}$.
output   : the updated $\Delta_\theta$ vector: $\Delta_{\Theta_i} \in \mathbb{R}^{(1\times(i+1))}$.
if $i = 0$ then
    $\Delta_{\Theta_i} = \emptyset$
end
else
    // Get $\Delta_\theta$ for all $\theta \in \Theta_{i-1}$ for the sets in the
       $(i+1)$-column and the $(i+1) - th$ row of $\mathcal{A}_{\Theta_i}$.
    foreach $\theta \in \Theta_{i-1}$ do
        foreach $A \in \{\mathcal{A}_{\Theta_i}(\cdot, i+1), \mathcal{A}_{\Theta_i}(i+1, \cdot)\}$ do
            $\Delta \leftarrow 0$
            foreach $\boldsymbol{x} \in A$ do
                $\Delta \leftarrow \Delta + \left[\left(\boldsymbol{f}^{(\theta)}_{n-\varphi n,c}(\boldsymbol{x}) * \mathrm{vol}\,(\boldsymbol{x})\right) - \mu_{\varphi n}(\boldsymbol{x})\right]$
            end
            $\Delta \leftarrow |\Delta|$
            $\Delta_\theta \leftarrow \max\{\Delta, \Delta_\theta\}$
        end
        insert $\Delta_\theta$ into $\Delta_{\Theta_i}(\theta)$ ; // insert into the $\theta$-th entry of the
         vector $\Delta_{\Theta_i}$
    end
    // Get $\Delta_\theta$ for $\theta = i$
    foreach $A \in \{\mathcal{A}_{\Theta_i}$ do
        $\Delta \leftarrow 0$
        foreach $\boldsymbol{x} \in A$ do
            $\Delta \leftarrow \Delta + \left[\left(\boldsymbol{f}^{(\theta)}_{n-\varphi n,c}(\boldsymbol{x}) * \mathrm{vol}\,(\boldsymbol{x})\right) - \mu_{\varphi n}(\boldsymbol{x})\right]$
        end
        $\Delta \leftarrow |\Delta|$
        $\Delta_\theta \leftarrow \max\{\Delta, \Delta_\theta\}$
    end
    insert $\Delta_\theta$ into $\Delta_{\Theta_i}(i+1)$
end
return $\Delta_{\Theta_i}$
```

ALGORITHM 3: `SRPCollate`$(\rho, \rho^{(c)})$

```
input    :
 1. The root node $\rho$ of an SRP $s$ with root box $\boldsymbol{x}_\rho$.
 2. The root node $\rho^{(c)}$ of an CRP $c$.
output   : The updated root node $\rho^{(c)}$ of the CRP $c$.
if $\rho^{(c)} = \emptyset$ // Nothing has been collated yet.
 then
    Make a new node $\rho^{(c)}$ with box $\boldsymbol{x}_\rho$
    foreach $\rho v \in s$ do
        $f_{n-\varphi n,s}(\rho v) \leftarrow \#\boldsymbol{x}_{\rho v}/((n-\varphi n) * \rho v)$
        Insert $f_{n-\varphi n,s}(\rho v)$ into $\boldsymbol{f}_{n-\varphi n,c}(\rho v)$ ;  // This is a ‘‘pushback’’
         operation, i.e keep $f_{n-\varphi n,s}(\rho v)$ in a vector $\boldsymbol{f}_{n-\varphi n,c}(\rho v)$.
        $\mu_{\varphi n}(\rho v) \leftarrow \check{\#}\boldsymbol{x}_{\rho v}/\varphi n$
    end
    return $c$
end
else
    Make a new node $\rho^{(c)}$ with box $\boldsymbol{x}_\rho$
    $f_{n-\varphi n,s}(\rho^{(c)}) \leftarrow \#\boldsymbol{x}_{\rho^{(c)}}/(n * \rho^{(c)})$
    Insert $f_{n-\varphi n,s}(\rho^{(c)})$ into $\boldsymbol{f}_{n-\varphi n,c}(\rho)$
    $\mu_{\varphi n}(\rho^{(c)}) \leftarrow \check{\#}\boldsymbol{x}_{\rho^{(c)}}/\varphi n$
    if (IsLeaf($\rho$) & (!IsLeaf($\rho^{(c)}$)) then
        Make temporary nodes $\mathsf{L}'$, $\mathsf{R}'$
        $\boldsymbol{x}_{\mathsf{L}'} \leftarrow \boldsymbol{x}_{\rho\mathsf{L}}$, $\boldsymbol{x}_{\mathsf{R}'} \leftarrow \boldsymbol{x}_{\rho\mathsf{R}}$
        $f_{n-\varphi n,s}(\mathsf{L}') \leftarrow f_{n-\varphi n,s}(\rho)$, $f_{n-\varphi n,s}(\mathsf{R}') \leftarrow f_{n-\varphi n,s}(\rho)$
        Graft onto $\rho^{(c)}$ as left child the node SRPCollate($\mathsf{L}', \rho^{(c)}\mathsf{L}$)
        Graft onto $\rho^{(c)}$ as right child the node SRPCollate($\mathsf{R}', \rho^{(c)}\mathsf{R}$)
    end
    if (IsLeaf($\rho^{(c)}$) & (!IsLeaf($\rho$) then
        Make temporary nodes $\mathsf{L}'$, $\mathsf{R}'$
        $\boldsymbol{x}_{\rho\mathsf{L}'} \leftarrow \boldsymbol{x}_{\rho^{(c)}\mathsf{L}}$, $\boldsymbol{x}_{\mathsf{R}'} \leftarrow \boldsymbol{x}_{\rho^{(c)}\mathsf{R}}$
        $f_{n-\varphi n,s}(\mathsf{L}') \leftarrow f_{n-\varphi n,s}(\rho^{(c)})$, $f_{n-\varphi n,s}(\mathsf{R}') \leftarrow f_{n-\varphi n,s}(\rho^{(c)})$
        Graft onto $\rho^{(c)}$ as left child the node SRPCollate($\rho\mathsf{L}, \mathsf{L}'$)
        Graft onto $\rho^{(c)}$ as right child the node SRPCollate($\rho\mathsf{R}, \mathsf{R}'$)
    end
    if (!IsLeaf($\rho$)) & (!IsLeaf($\rho^{(c)}$) then
        Graft onto $\rho^{(c)}$ as left child the node SRPCollate($\rho\mathsf{L}, \rho^{(c)}\mathsf{L}$)
        Graft onto $\rho^{(c)}$ as right child the node SRPCollate($\rho\mathsf{R}, \rho^{(c)}\mathsf{R}$)
    end
    return $\rho^{(c)}$
end
```

ALGORITHM 4: `GetYatracos`

```
input    :
 1. the node that was split: $\rho v^*$;
 2. the vector of histogram estimates: $\boldsymbol{f}_{n-\varphi n,c}$;
 3. the current number of splits: $i$;
 4. the current Yatracos matrix: $\mathcal{A}_{\Theta_{i-1}}$.
output   : the updated Yatracos matrix: $\mathcal{A}_{\Theta_i}$.
if $\boldsymbol{x}_{\rho v^*} = \boldsymbol{x}_\rho$ then
  | $A_{0,0} \leftarrow \emptyset$
end
for $j = 0 : (i-1)$ do
    check the i-th column // Iterating through the entries of the
        $(i-1)$-th column to check if the entry $A_{j,i-1}$ contains
        $\boldsymbol{x}_{\rho v^*}$
    if $(A_{j,i-1} \neq \emptyset)$ & $(\boldsymbol{x}_{\rho v^*} \in A_{j,i-1})$ then
        // The entry $A_{j,i}$ takes all the elements of $A_{j,i-1}$ except
            $\boldsymbol{x}_{\rho v^*}$
        $A_{j,i} \leftarrow A_{j,i-1} \setminus \boldsymbol{x}_{\rho v^*}$
    end
    else
        $A_{j,i} \leftarrow A_{j,i-1}$
    end
    // Compare the estimates at each child node
    foreach $\boldsymbol{x} \in \{\boldsymbol{x}_{\rho v^* \mathsf{L}}, \boldsymbol{x}_{\rho v^* \mathsf{R}}\}$ do
        if $\boldsymbol{f}^{(j)}_{n-\varphi n,c}(\boldsymbol{x}_\rho) > \boldsymbol{f}^{(i)}_{n-\varphi n,c}(\boldsymbol{x}_\rho)$ then
            // Take the union of the elements in entry $A_{j,i}$ with
                $\boldsymbol{x}_\rho$
            $A_{j,i} \leftarrow \left\{ \bigcup_{\boldsymbol{x}_v \in A_{j,i}} \boldsymbol{x}_{\rho v} \cup \boldsymbol{x}_\rho \right\}$
        end
    end
    check the i-th row // Iterating through the entries of the
        $(i-1)$-th row to check if the entry $A_{i-1,j}$ contains $\boldsymbol{x}_{\rho v^*}$
    if $(A_{i-1,j} \neq \emptyset)$ & $(\boldsymbol{x}_{\rho v^*} \in A_{i-1,j})$ then
        // The entry $A_{i,j}$ takes all the elements of $A_{i-1,j}$ except
            $\boldsymbol{x}_{\rho v^*}$
        $A_{i,j} \leftarrow A_{i-1,j} \setminus \boldsymbol{x}_{\rho v^*}$
    end
    else
        $A_{i,j} \leftarrow A_{i-1,j}$
    end
    // Compare the estimates at each child node
    foreach $\boldsymbol{x}_\rho \in \{\boldsymbol{x}_{\rho v^* \mathsf{L}}, \boldsymbol{x}_{v^* \mathsf{R}}\}$ do
        if $\boldsymbol{f}^{(i)}_{n-\varphi n,i}(\boldsymbol{x}_\rho) > \boldsymbol{f}^{(j)}_{n-\varphi n,j}(\boldsymbol{x}_\rho)$ then
            // Take the union of the elements in entry $A_{i,j}$ with
                $\boldsymbol{x}_\rho$
            $A_{i,j} \leftarrow \left\{ \bigcup_{\boldsymbol{x}_{\rho v} \in A_{i,j}} \boldsymbol{x}_{\rho v} \cup \boldsymbol{x}_\rho \right\}$
        end
    end
end
$A_{i,i} \leftarrow \emptyset$ // The diagonal entry is always an empty set
return $\mathcal{A}_{\Theta_i}$
```

References

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer-Verlag.

Devroye, L., & Lugosi, G. (2001). *Combinatorial methods in density estimation*. New York: Springer-Verlag.

Devroye, L., & Lugosi, G. (2004). Bin width selection in multivariate histograms by the combinatorial method. *TEST*, *13*(1), 129–145.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

Garcia, S., Grill, M., Stiborek, H., & Zunino, A. (2014). An empirical comparison of botnet detection methods. *Computers and Security Journal*, *45*, 100–123.

Gray, A. G., & Moore, A. W. (2003). Nonparametric density estimation: Towards computational tractability. In SIAM international conference on data mining (pp. 203–211). San Francisco, California, USA: SIAM.

Harlow, J., Sainudiin, R., & Tucker, W. (2012). Mapped regular pavings. *Reliable Computing*, *16*, 252–282.

Kieffer, M., Jaulin, L., Braems, I., & Walter, E. (2001). Guaranteed set computation with subpavings. In W. Kraemer & J. Gudenberg (Eds.), *Scientific computing, validated numerics, interval methods, proceedings of SCAN 2000* (pp. 167–178). New York: Kluwer Academic Publishers.

Klemelä, J. (2009). *Smoothing of multivariate data: density estimation and visualization*. Chichester: Wiley.

Lu, L., Jiang, H., & Wong, W. H. (2013). Multivariate density estimation by bayesian sequential partitioning. *Journal of the American Statistical Association*, *108*(504), 1402–1410. https://doi.org/10.1080/01621459.2013.813389.

Lugosi, G., & Nobel, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, *24*(2), 687–706.

Mahalanabis, S., & Stefankovic, D. (2008). Density estimation in linear time. In R. A. Servedio & T. Zhang (Eds.), *21st annual conference on learning theory—COLT 2008* (pp. 503–512). Finland: Omnipress, Helsinki.

Mattarei, S. (2010). Asymptotics of partial sums of central binomial coefficients and Catalan numbers. arXiv.0906.4290v3

Meier, J. (2008). *Groups, graphs and trees: an introduction to the geometry of infinite groups*. Cambridge: Cambridge University Press.

Ramström, K. (2019). Botnet detection on flow data using the reconstruction error from Autoencoders trained on Word2Vec network embeddings. Msc thesis, Uppsala University

Sainudiin, R., Teng, G., Harlow, J., & Lee, D. S. (2013). Posterior expectation of regularly paved random histograms. *ACM Transactions on Modeling and Computer Simulation*, *23*(26), 6:1–6:20.

Sainudiin, R., York, T., Harlow, J., Teng, G., Tucker, W., & George, D. (2008–2019). MRS 2.0, a C++ class library for statistical set processing and computer-aided proofs in statistics. https://github.com/lamastex/mrs2

Samet, H. (1990). *The design and analysis of spatial data structures*. Boston: Addison-Wesley Longman.

Stanley, R.P. (1999). Enumerative combinatorics. Vol. 2, Cambridge Studies in Advanced Mathematics, vol 62. Cambridge University Press, Cambridge. https://books.google.fr/books?id=zg5wDqT6T-UC&hl=fr&source=gbs_book_other_versions

Tukey, J. W. (1947). Non-parametric estimation II. Statistically equivalent blocks and tolerance regions—The continuous case. *The Annals of Mathematical Statistics*, *18*(4), 529–539.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab Appl*, *16*, 264–280.

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, *13*(2), 768–774.

Yatracos, Y. G. (1988). A note on l1 consistent estimation. *The Canadian Journal of Statistics*, *16*(3), 283–292.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., et al. (2016). Apache spark: A unified engine for big data processing. *Commun ACM*, *59*(11), 56–65. https://doi.org/10.1145/2934664.