This is an updated version of the preprint:

Data-adaptive histograms through statistical regular pavings RAAZESH SAINUDIIN† ∘, GLORIA TENG‡, JENNIFER HARLOW ∘ and WARWICK TUCKER? , ∘School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, ‡ NEXT Academy, Kuala Lumpur, Malaysia, and ? Department of Mathematics, Uppsala University, Uppsala, Sweden.

Due to the size of the original preprint above, the reviewers have advised to break it into two smaller papers.

Now the first smaller preprint:

Minimum distance estimation with universal performance guarantees over statistical regular pavings

RAAZESH SAINUDIIN† ∘ and GLORIA TENG‡ , ∘ Department of Mathematics, Uppsala University, Uppsala, Sweden, and ‡ NEXT Academy, Kuala Lumpur, Malaysia.

has been prepended on Thu Apr 5 23:16:46 CEST 2018 to the original preprint.

Minimum distance estimation with universal performance guarantees over statistical regular pavings

RAAZESH SAINUDIIN[†°] and GLORIA TENG[‡], [°]Department of Mathematics, Uppsala University, Uppsala, Sweden, and [‡]NEXT Academy, Kuala Lumpur, Malaysia.

We present a data-adaptive multivariate histogram estimator of an unknown density f based on n independent samples from it. Such data-dependent adaptive histograms are formalized as statistical regular pavings (SRPs). Regular pavings (RPs) are binary trees obtained by selectively bisecting a d-dimensional box in a recursive and regular manner. SRP augments RP by mutably caching the sufficient statistics of the data. Using a form of tree matrix collation with SRPs we obtain the minimum distance estimate (MDE) with universal performance guarantees over Yatracos classes and demonstrate the performance of the estimator with simulated data.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: *—Probabilistic algorithms (including Monte Carlo)*; *Statistical computing*; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Trees*; E.1 [**Data Structures**]: *—Trees*

General Terms: Algorithms, Design, Performance, Theory

Additional Key Words and Phrases: Rooted planar binary tree, Yatracos class, Tree matrix arithmetic

1. INTRODUCTION

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

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

Any density estimate $f_n(x) := f_n(x; X_1, X_2, \ldots, X_n) : \mathbb{R}^d \times \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ 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 $X_1, X_2, \ldots, X_n$ 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:

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

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* if it is valid no matter what the underlying f happens to be [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, for some $k \geq 0$, 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.

Author's addresses: [†] Corresponding Author: Raazesh Sainudiin, [°]Department of Mathematics, Uppsala University, Box 480, 751 06 Uppsala, Sweden. [‡]NEXT Academy, Kuala Lumpur 60000, Malaysia.

For a concrete class of unknown densities of pressing interest in current applications involving periodic bursts of sample size $n \approxeq 10^7$, consider an $f \notin C^0$, where $f \ll \lambda$, but its inverse image at 0 has finitely many distinct Lebesgue-measurable full compact sets (say, h distinct "clean holes"), within a box $\breve{\boldsymbol{x}} := [-M, M]^d$ for a large enough M, such that, $\mu(\breve{\boldsymbol{x}}^c) < \xi$, where $\breve{\boldsymbol{x}}^c := \mathbb{R}^d \setminus \breve{\boldsymbol{x}}$, for any given but fixed $\xi > 0$. Thus, $f^{[-1]}(0) := \{x : f(x) = 0\} = \cup_{i=1}^{h} H_i$, where each $H_i \subset \breve{\boldsymbol{x}}$ and $H_i \cap H_j = \emptyset$, for $i \neq j$. Such an f is in L_1, given that it is a density, and crucially, an estimator f_n with universal performance guarantees will estimate $f \in L_1$ and *not* merely assure asymptotic L_1 consistency while mathematically compromising to f actually *not* being in L_1 in order to solve the smoothing problem that is also required to obtain f_n for a given $n < \infty$.

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 [Vapnik and Chervonenkis 1971] on uniform convergence of empirical probability measures over classes of sets.

Tree based partitioning strategies are particularly suited to large sample sizes and we focus on tree based histograms here using statistical regular pavings. 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 to produce nonparametric data-adaptive density estimates in d dimensions with universal performance guarantees.

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. 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 implementations of the MDE using data-adaptive multivariate histograms for bursts of data (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. 2018] is the only publicly available implementation of such an MDE estimator.

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 $\mathrm{mid}(\boldsymbol{x}) := (\underline{x} + \overline{x})/2$. A box of dimension d with coordinates in $\Delta := \{1, 2, \ldots, d\}$ is an interval vector with ι as the first coordinate of maximum width:

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

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 ι gives the left and right child boxes of $\boldsymbol{x}$

$$\boldsymbol{x}_{\mathsf{L}} := [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_\iota, \mathrm{mid}(\boldsymbol{x}_\iota)) \times [\underline{x}_{\iota+1}, \overline{x}_{\iota+1}] \times \ldots \times [\underline{x}_d, \overline{x}_d] ,$$

$$\boldsymbol{x}_{\mathsf{R}} := [\underline{x}_1, \overline{x}_1] \times \ldots \times [\mathrm{mid}(\boldsymbol{x}_\iota), \overline{x}_\iota] \times [\underline{x}_{\iota+1}, \overline{x}_{\iota+1}] \times \ldots \times [\underline{x}_d, \overline{x}_d] .$$

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

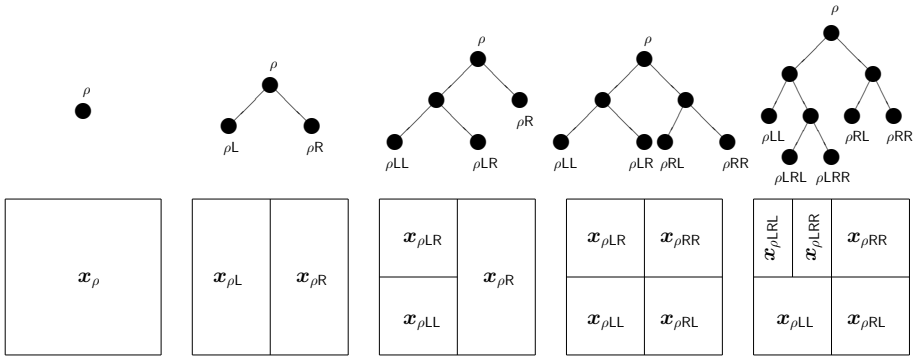


Fig. 1. A sequence of selective bisections of boxes (nodes) along the first widest coordinate, starting from the root box (root node) in two dimensions, produces an RP.

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 Figure 1 via a sequence of bisections of a square (2-dimensional) root box by always bisecting on the *first* widest coordinate.

Let $\mathbb{N} := \{1, 2, \ldots\}$ be the set of natural numbers. Let the j-th 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 $\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}}) = \prod_{j=1}^{d}(\overline{x}_{\rho\mathsf{v},j} - \underline{x}_{\rho\mathsf{v},j})$, the set of all nodes of an RP be $\mathbb{V} := \rho \cup \{\rho\{\mathsf{L},\mathsf{R}\}^j : j \in \mathbb{N}\}$, the set of all leaf nodes be $\mathbb{L}$ and the set of internal nodes or splits be $\breve{\mathbb{V}}(s) := \mathbb{V}(s) \setminus \mathbb{L}(s)$. 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.

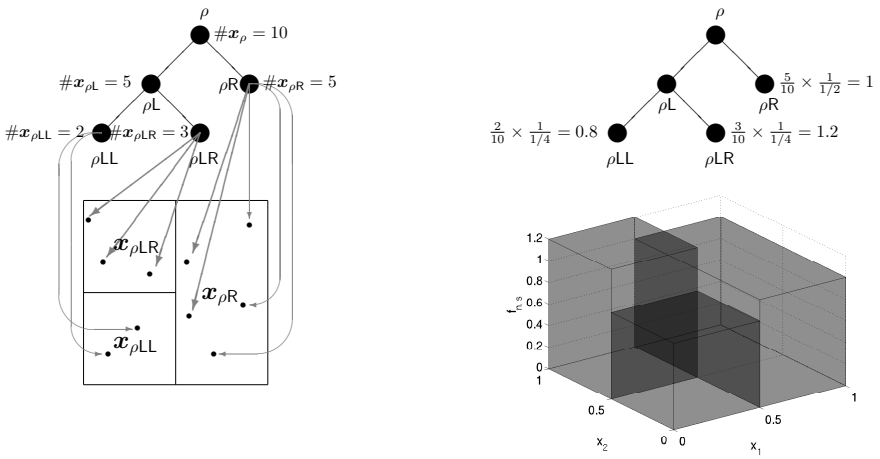
$$C_k = \frac{1}{k+1}\binom{2k}{k} = \frac{(2k)!}{(k+1)!(k!)} \ . \tag{1}$$

For $i, j \in \mathbb{Z}_+$, where $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$ 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 $k \in \{i, i+1, \ldots, j\}$. 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 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 $\boldsymbol{x}_\rho \in \mathbb{IR}^2$. The number of sample data points in the root box $\boldsymbol{x}_\rho$ is 10. Figure 2(a) 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 2(b)

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.



(a) An SRP tree and its constituents. (b) An SRP histogram $f_{n,s}$.

Fig. 2. An SRP and its corresponding histogram.

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

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

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 midpoint 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 Sec. 4).

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 $\texttt{SEBTreeMC}(s, \overline{\#}, \overline{m})$ of Algorithm 1 is $[s(0), s(1), \ldots, s(T)]$, 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

ALGORITHM 1: SEBTreeMC$(s, \overline{\#}, \overline{m})$

input : s, initial SRP with root node ρ,
$x = (x_1, x_2, \ldots, x_n)$, a data burst of size n,
$\# : \mathbb{L}^{\triangledown}(s) \to \mathbb{R}$, 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 $[s(0), s(1), \ldots, s(T)]$ 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: $\boldsymbol{x}_\rho \rightsquigarrow x$, make $\boldsymbol{x}_\rho$ such that $\cup_i^n x_i \subset \boldsymbol{x}_\rho$ if $\nexists$ domain knowledge or historical data,
$s \rightsquigarrow \boldsymbol{x}_\rho$, 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` $\left(\underset{\rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s)}{\operatorname{argmax}} \#(\rho\mathsf{v}) \right)$ `// sample uniformly from nodes with largest` $\#$

$s \leftarrow s$ with node $\rho\mathsf{v}$ split `// split the sampled node and update` s

$\mathbf{s}$.append(s) `// append the new SRP state with an additional split`

end

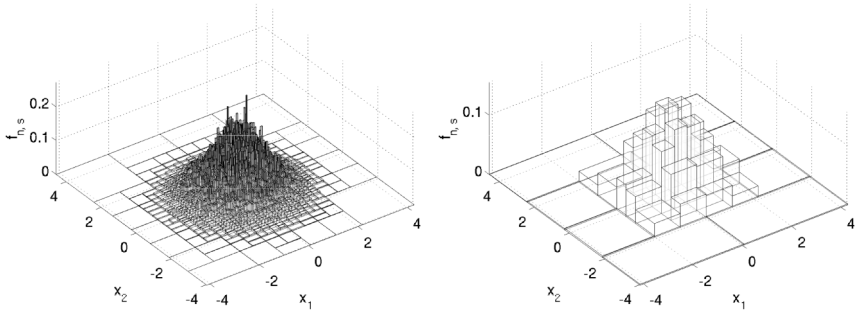


Fig. 3. Two histogram density estimates for the standard bivariate Gaussian density. The left figure shows a histogram with 1485 leaf nodes where $\overline{\#} = 50$ and the histogram on the right has $\overline{\#} = 1500$ resulting in 104 leaf nodes.

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 3.1, the L_1 consistency of SEBTreeMC requires that $\overline{m}$ must grow sublinearly (i.e. $\overline{m}/n \to 0$ as $n \to \infty$) 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 3.3.

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

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

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

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

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 $\boldsymbol{x} = [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_d, \overline{x}_d]$, $\mathrm{diam}(\boldsymbol{x}) = \sqrt{\sum_{i=1}^d (\overline{x}_i - \underline{x}_i)^2}$.

THEOREM 3.1 (L_1-CONSISTENCY). *Let $X_1, X_2, \ldots$ 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:

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

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 $\sqrt{d(\frac{2M}{2^i})^2}$, 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

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

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.,

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

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:

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

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

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

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 $\mathcal{T} := \{x_1, \ldots, x_{n-\varphi n}\}$ and the set of validation data by $\mathcal{V} := \{x_{n-\varphi n+1}, \ldots, x_n\} = \{y_1, \ldots, y_{\varphi n}\}$. For an ordered pair $(\theta, \vartheta) \in \Theta^2$, with $\theta \neq \vartheta$, the set:

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

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

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

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: $s_{\theta_1}, s_{\theta_2}, \ldots$ 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 is 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 Figure 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 $\check{\#}\boldsymbol{x}_{\rho v}$.

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

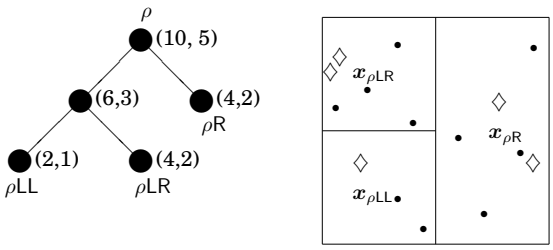


Fig. 4. An SRP s with training ($\bullet$) and validation data ($\diamond$) and their respective sample counts ($\#\boldsymbol{x}_{\rho v}, \check{\#}\boldsymbol{x}_{\rho v}$) that are updated recursively as data falls through the nodes of s.

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 Δ_θ 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. 2018] 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 3.2. *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 Δ_θ in* (3)*, we have*

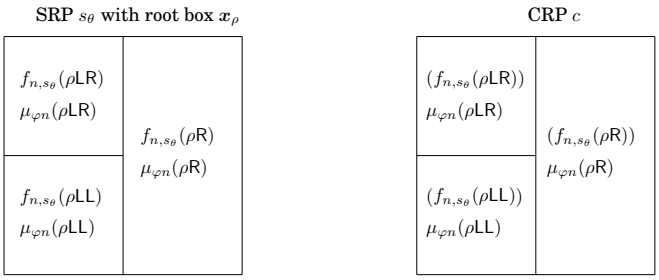
$$\int |f_{n-\varphi n,\theta^*} - f| \leq 3 \min_{\theta\in\Theta} \int |f_{n-\varphi n,\theta} - f| + 4\Delta \tag{4}$$

where

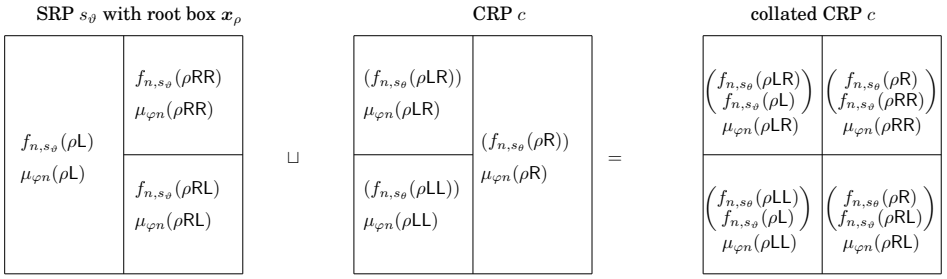
$$\Delta = \max_{A\in\mathcal{A}_\Theta} \left| \int_A f - \mu_{\varphi n}(A) \right| \quad . \tag{5}$$

Theorem 3.2 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 by using the cardinality bound [Devroye et al. 1996, Theorem 13.6, p. 219] for the shatter coefficient of $\mathcal{A}_\Theta$.



(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 $\boldsymbol{x}_\rho$.

Given $\{x_1, \ldots, x_n\}$ the n-th shatter coefficient of $\mathcal{A}_\Theta$ is defined as

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

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 n-th shatter coefficient is bounded as follows:

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

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 3.3. *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, φn, Θ and $f \in L_1$:*

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

PROOF. By Theorem 3.2,

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

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

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

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

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

□

In order 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 3.1, 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. PERFORMANCE EVALUATION

To evaluate the performance of our MDE we chose two multivariate densities: the spherically symmetric Gaussian and the highly structured Rosenbrock density (whose expression up to normalization is given in (7)) in d dimensions for various sample sizes.

$$\exp\left(-\sum_{i=2}^{d}(100(x_i - x_{i-1}^2)^2 + (1 - x_{i-1})^2)\right) \quad . \tag{7}$$

Table I. 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 $L_1(f_{n,\theta^*}, f)$, $L_1(f_{n,\theta^*}, f) - \min_{\theta\in\Theta} L_1(f_{n,\theta}, f)$				
10^4	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)

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 I), 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 I for each density and sample size. For comparison we used the posterior mean histograms based on the MCMC method [Sainudiin et al. 2013, see for details on this evaluation] (they are shown in the bottom panel of Table I along with their standard errors. Note how the L_1 errors decrease with the sample size and how the errors are comparable between the methods, albeit the MDE method is at least an order of magnitude faster than the MCMC method (for detailed CPU times of the MCMC method see [Sainudiin et al. 2013]).

Remark 4.1. The approximate integration methods based on quasi-random streams and their importance sampling extensions became unreliable and significantly slower for highly structured densities such as that of Rosenbrock (7) in dimensions as large as 5. Thus, we used *r*eal mapped regular paving or $\mathbb{R}$-MRP approximation of the true density that is within 0.01 in Hellinger distance of the true density (see [Sainudiin et al. 2013, Sec. 4.2] and [Harlow et al. 2012] for details) whose n samples were simulated exactly using interval enclosures of the range of the target density [Sainudiin and York 2013] over regularly paved partitions. The target density f can be any one with a locally Lipschitz arithmetical expression and not merely the two examples shown here (see `mrs2` [Sainudiin et al. 2018] `examples/MooreRejSam` module) and this

allows a skeptic to experiment for further evidence from simulations from this large class of densities on their own. By producing n samples from such piecewise constant $\mathbb{R}$-MRP densities, we can take advantage of $\mathbb{R}$-MRP arithmetic to obtain the exact L_1 error in Table I between the approximated $\mathbb{R}$-MRP representation of the density f and the $\mathbb{R}$-MRP representation of the estimate f_n produced by the MCMC or MDE methods. All experiments were performed on the same physical machine that is currently considered to be commodity hardware [Sainudiin et al. 2013, for machine specifications].

Thus, by 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. 2018]. 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 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 Apache Spark [Zaharia et al. 2016] as the sample size becomes too large for the memory of a single machine.

ACKNOWLEDGMENTS

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 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.

REFERENCES

DEVROYE, L., GYÖRFI, L., AND LUGOSI, G. 1996. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

DEVROYE, L. AND LUGOSI, G. 2001. *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York.

DEVROYE, L. AND 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.

GRAY, A. G. AND MOORE, A. W. 2003. Nonparametric Density Estimation: Towards Computational Tractability. In *SIAM International Conference on Data Mining*. SIAM, San Francisco, California, USA, 203–211.

HARLOW, J., SAINUDIIN, R., AND TUCKER, W. 2012. Mapped regular pavings. *Reliable Computing 16*, 252–282.

KIEFFER, M., JAULIN, L., BRAEMS, I., AND WALTER, E. 2001. Guaranteed set computation with subpavings. In *Scientific Computing, Validated Numerics, Interval Methods, Proceedings of SCAN 2000*, W. Kraemer and J. Gudenberg, Eds. Kluwer Academic Publishers, New York, 167–178.

KLEMELÄ, J. 2009. *Smoothing of Multivariate Data: Density Estimation and Visualization*. Wiley, Chichester, United Kingdom.

LU, L., JIANG, H., AND WONG, W. H. 2013. Multivariate density estimation by bayesian sequential partitioning. *Journal of the American Statistical Association 108,* 504, 1402–1410.

LUGOSI, G. AND NOBEL, A. 1996. Consistency of Data-Driven Histogram Methods for Density Estimation and Classification. *The Annals of Statistics 24,* 2, 687–706.

MAHALANABIS, S. AND STEFANKOVIC, D. 2008. Density estimation in linear time. In *21st Annual Conference on Learning Theory - COLT 2008*, R. A. Servedio and T. Zhang, Eds. Omnipress, Helsinki, Finland, 503–512.

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 University Press, Cambridge, United Kingdom.

SAINUDIIN, R., TENG, G., HARLOW, J., AND 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. AND YORK, T. 2013. An auto-validating, trans-dimensional, universal rejection sampler for locally Lipschitz arithmetical expressions. *Reliable Computing 18*, 15–54.

SAINUDIIN, R., YORK, T., HARLOW, J., TENG, G., TUCKER, W., AND GEORGE, D. 2008–2018. MRS 2.0, a C++ class library for statistical set processing and computer-aided proofs in statistics. `https://github.com/raazesh-sainudiin/mrs2`.

SAMET, H. 1990. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley Longman, Boston.

STANLEY, R. P. 1999. *Enumerative combinatorics. Vol. 2*. Cambridge Studies in Advanced Mathematics Series, vol. 62. Cambridge University Press, Cambridge.

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. AND 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, pp. 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., MENG, X., ROSEN, J., VENKATARAMAN, S., FRANKLIN, M. J., GHODSI, A., GONZALEZ, J., SHENKER, S., AND STOICA, I. 2016. Apache Spark: A unified engine for big data processing. *Commun. ACM 59,* 11, 56–65.

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 Δ_θ vector: $\Delta_{\Theta_{i-1}} \in \mathbb{R}^{(1\times(i))}$.

output : the updated Δ_θ vector: $\Delta_{\Theta_i} \in \mathbb{R}^{(1\times(i+1))}$.

if $i = 0$ **then**
  $\Delta_{\Theta_i} = \emptyset$
end
else
  `// Get` Δ_θ `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 Δ_θ into $\Delta_{\Theta_i}(\theta)$; `// insert into the` θ`-th entry of the vector` Δ_{Θ_i}
  end
  `// Get` Δ_θ `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 Δ_θ into $\Delta_{\Theta_i}(i+1)$
end
return Δ_{Θ_i}

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

input :
(1) The root node ρ 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` $\boldsymbol{f}_{n-\varphi n,c}(\rho v)$.
    $\mu_{\varphi n}(\rho v) \leftarrow \breve{\#}\boldsymbol{x}_{\rho v}/\varphi n$
  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 $\boldsymbol{f}_{n-\varphi n,c}(\rho)$
  $\mu_{\varphi n}(\rho^{(c)}) \leftarrow \breve{\#}\boldsymbol{x}_{\rho^{(c)}}/\varphi n$

  if (`IsLeaf`(ρ) & (!`IsLeaf`$(\rho^{(c)})$) **then**
    Make temporary nodes L', 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`(ρ) **then**
    Make temporary nodes L', 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`(ρ)) & (!`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: ρv^*;
(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 $\boldsymbol{x} \in \{\boldsymbol{x}_{\rho v^* \mathtt{L}}, \boldsymbol{x}_{\rho v^* \mathtt{R}}\}$ **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 $\boldsymbol{x}_\rho \in \{\boldsymbol{x}_{\rho v^* \mathtt{L}}, \boldsymbol{x}_{v^* \mathtt{R}}\}$ **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 $\mathcal{A}_{\Theta_i}$

Data-adaptive histograms through statistical regular pavings

RAAZESH SAINUDIIN[†°], GLORIA TENG[‡], JENNIFER HARLOW[°] and WARWICK TUCKER[⋆], [°]School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, [‡]NEXT Academy, Kuala Lumpur, Malaysia, and [⋆]Department of Mathematics, Uppsala University, Uppsala, Sweden.

We present data-adaptive multivariate histogram estimators of an uknown density f based on n independent samples from it. Such data-dependent adaptive histograms are formalized as statistical regular pavings (SRPs). Regular pavings (RPs) are binary trees obtained by selectively bisecting a d-dimensional box in a recursive and regular manner. Due to this regularity, RPs are algebraically closed under union operations and amenable to tree arithmetic. SRP augments RP by mutably caching the recursively computable sufficient statistics of the data.

Using tree arithmetic operations we are able to perform computationally efficient smoothing for a given sample size n to obtain three optimally smoothed data-adaptive multivariate histograms: (i) the L_2-risk minimizing posterior mean or Bayes estimate by averaging posterior sample histograms from a Monte Carlo Markov chain initialized about the *maximum a posterior* (MAP) estimate along the paths of asymptotically L_1 consistent priority-queued Markov chains, (ii) the optimal MAP (OPTMAP) estimate by minimizing the cross-validation estimator of L_2 risk as a function of the prior or penalty parameter, and (iii) the minimum L_1 distance estimate (MDE) with universal performance guarantees over *Yatracos classes*.

Finally, our histogram density estimates allow a wide range of subsequent statistical operations, such as, computing tail probabilities, or conditional predictive densities, to be performed efficiently. We demonstrate the utility of some such statistical operations with our optimally smoothed density estimates on simulated data as well as data from a measurable chaotic double pendulum.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: *—Probabilistic algorithms (including Monte Carlo)*; *Statistical computing*; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Trees*; E.1 [**Data Structures**]: *—Trees*

General Terms: Algorithms, Design, Performance, Theory

Additional Key Words and Phrases: multivariate histogram, rooted planar binary tree, shatter coefficient, Yatracos class, minimum distance estimate, universal performance guarantee

1. INTRODUCTION

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

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

Any density estimate $f_n(x) := f_n(x; X_1, X_2, \ldots, X_n) : \mathbb{R}^d \times \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ 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 $X_1, X_2, \ldots, X_n$ drawn from f. However it is obtained, the density estimate is some smoothed representation of the observed data [Whittle 1958]. A density estimator must not only be asymptotically L_1 consistent, i.e., $\int \mathrm{abs}(f_n(x) - f(x))dx \to 0$ as $n \to \infty$, but must also be optimally smoothed for any fixed sample of size n and generalizable to unobserved data. In other words, for a given n and a smoothing parameter s, our estimator should optimize a sensible expected objective function g using a hold-out method such as cross-validation, i.e., $f_{n,s^*} = \mathrm{argopt}_{f_{n,s}} E(g(f_{n,s}, f))$.

Author's addresses: [†] Corresponding Author: Raazesh Sainudiin, [°]School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8041, New Zealand. [‡]NEXT Academy, Kuala Lumpur 60000, Malaysia. [⋆]Department of Mathematics, Uppsala University, Box 480, 751 06 Uppsala, Sweden.

This asymptotically L_1 consistent and optimally smoothed density estimate f_{n,s^*} of the unknown f gives us a means of computing the probabilities of any Borel set $A \in \mathcal{B}^d$ or of computing the density at any point $x \in \mathbb{R}^d$. Density estimation is often the first step in many learning tasks, including, anomaly detection, classification, regression and clustering. Thus, such a density estimate f_{n,s^*} *is particulary useful if it is conducive to computationally efficient statistical operations* including point-wise image evaluation for a set of query points, highest density set computation, L_1 distance calculation, conditional and marginal density computations, etc. Another class of statistical operations include transformations of the density estimate f_{n,s^*} (through finitely many arithmetical operations) and integration, such as computing the integral of f^2_{n,s^*} in leave-one-out cross-validation. Such statistical operations directly on a density estimate can help with subsequent learning tasks.

There are two general approaches to density estimation: parametric and nonparametric. Although parametric estimates can achieve the optimal convergence rate of $n^{-1/2}$, they are generally not rich enough to represent the underlying structure of various real-world datasets, especially in moderate to high dimensions. Nonparametric density estimation methods avoid the bias inherent in the limited approximation power of parametric families by leveraging on the large sample size n. Among a large class of nonparametric multivariate density estimates, kernel density estimates (or KDEs) and histograms are the two most common forms. Data-adaptive density estimation methods adapt the amount of smoothing to the local density of the data [Silverman 1986, chap. 2] — a crucial and necessary strategy to capture the underlying inherent complexity in today's datasets with large samples sizes. Data-adaptive smoothing is necessary to control estimation errors as the number of dimensions increases [Scott 1992, chap. 7], but is also computationally expensive and difficult to achieve [Scott and Sain 2005].

1.1. Kernel density estimates

One of the most popular nonparametric estimators is the kernel density estimator (KDE), also known as the Parzen window estimator, which places a kernel on each of the n sample points [Rosenblatt 1956; Parzen 1962]. One of the theoretical properties of the KDE is its convergence rate to the true density, especially when the method is computationally feasible for small samples sizes in small dimensions. However, as shown in [Stone 1980], the optimal convergence rate for a density f with p bounded derivatives in d dimensions is of the order $n^{p/(2p+d)}$ — which can be very slow in high dimensions even if p is often unrealistically assumed to be relatively large to enforce smoothness on the *unknown* f for convenient theoretical arguments. Moreover, KDE methods suffer from a high computational complexity, since the computation time grows linearly with the product of the number of training samples n and the number of test samples $n_t < n$, i.e., $O(n \times n_t)$. The complexity becomes particularly acute, growing as $O(n \times n_t \times n_s)$, when one also needs to determine the optimal smoothing parameter from a finite set of size n_s.

Several approaches are available to reduce the computational burden of KDEs. For example, by binning the n training samples to b bins with $b < n$, one can reduce the time complexity to $O(b \times n_t \times n_s)$ [Scott and Sheather 1985], but at the expense of loosing any crucial structural information in the training sample within the bins. The fast multipole method [Rokhlin 1985], an approximation that combines co-located samples, is optimized in the fast Gauss transform [Greengard and Strain 1991] for Gaussian kernels. This method produces approximations in just $C_\epsilon(n + n_t) \times n_s$ steps, where the constant C_ϵ depends on the desired accuracy level ϵ, but at the expense of loss of information in the training set due to truncation of the underlying Hermite expan-

sions. Moreover, the time complexity of its improved variant grows polynomially in d after delicate Taylor expansions and domain partitions but is still limited in practice to fewer than 10 dimensions with n less than 10^4 or 10^5, even if one were to ignore the information lost in the truncations for a reasonable accuracy level [Yang et al. 2003]. Other improvements to KDE use fast nearest neighbor methods based on binary trees, such as kd-trees, ball trees and anchors hierarchy (eg. [Bentley 1975; Moore 2000; Gray and Moore 2003a; 2003b]). These methods, being tree-assisted approximations that exploit proximity relations among the samples, are faster than the naive KDE but do not scale well to large sample sizes in moderately large dimensions with reasonable accuracy. For instance, when the underlying density is highly unstructured these methods suffer from the curse of dimensionality as well as sample size [Moore 2000] as discussed in Sainudiin et al. [2013].

The faster KDE methods mentioned above typically consider spherical Gaussian kernels. Such isotropic kernels can significantly constrain the richness of the ensuing nonparameric class of densities and make it inadequate for datasets with complex local structures [López-Rubio and de Lazcano-Lobato 2008]. A typical solution to deal with this isotropic inadequacy is to extract the local principal directions of the dataset as done by manifold methods (eg. [Vincent and Bengio 2002; Ozertem and Erdogmus 2011]) at the steep cubic computational cost of learning the local principal directions from clusters of training samples.

Finally, when optimal smoothing parameters have to be chosen or tuned for the given dataset of finite sample size n, the KDE methods become prohibitively expensive. As the number of dimensions increases, both the choice of the kernel's shape and bandwidth needs to depend on the data. The kernel bandwidth is the only smoothing parameter in need of tuning for the simplest isotropic kernels, but n_s, the size of the typically discrete set obtained from an adaptive grid of smoothing parameters that need to be tuned via some hold-out procedure, can get very large for more flexible and data-adaptive kernels. For example, the MCMC-based method that automatically produces the KDE *along* with the optimal diagonal bandwidth matrix [Zhang et al. 2006] quickly becomes impractical for $n > 2000$ and $d > 2$. For these reasons many KDE methods, even after computational improvements, are only effective with data in less than five or six dimensions with sample sizes less than a few tens of thousands and generally reach computational bottle-necks when the sample size reaches 10^6, especially when the tuning of the smoothing parameter via a hold-out procedure is also necessary, even if we were to make the typically unrealistic assumption that the unknown underlying density is sufficiently smooth to ensure a fast enough theoretical convergence rate for the multivariate KDE [Stone 1980].

1.2. Histograms

A histogram [Pearson 1895] is based on a partition of the data space, with the density estimate over each element of the partition given by the product of the relative frequency of the data contained in it and its reciprocal Lebesgue measure. The elements of the partition are commonly known as bins and the choice of bin width(s) is the *smoothing problem*: wider bins give more smoothing, narrower bins less smoothing or more spiking. The bins of a *regular* histogram are all equally-sized; the bins of an *irregular* histogram can vary in size [Tukey 1947; Scott 1979].

Regular partitioning with small enough bins to suit the modes of the density will not only give too many bins in low or flat density areas [Rissanen et al. 1992] but also make the total number of bins grow exponentially with the number of dimensions. Regular partitioning with a bin width more suited to the overall variability of the data may compromise the potential of the histogram to show important local features in the highest density areas. Thus multivariate regular histograms with a single global

bin width are not able to adapt to spatially varying smoothing requirements [Klemelä 2009, chap. 17] akin to isotropic kernels with a single bandwidth parameter in KDEs.

A data-dependent partition, on the other hand, allows the bin width to vary in a way that is determined by the data and can adapt to reflect the complexity in the data. Data-dependent partitions can provide estimates which are theoretically superior to those using partitions based simply on the number of data points in the data set [Stone 1985]. A histogram density estimate with a data-dependent partitioning strategy can be asymptotically L_1-consistent, provided the following three conditions of Lugosi and Nobel [1996] are met: (i) a sub-exponential growth of a certain combinatorial geometric complexity measure of the class of allowed partitions (ii) a sub-linear growth of the number of cells and (iii) the shrinking of diameters of the cells in the partition with $f > 0$, as the sample size tends to infinity. See Devroye and Lugosi [2001] for a self-contained introduction to combinatorial methods in density estimation. On the other hand, an intuitively convincing data-adaptive partitioning strategy, say by isolating k points inside a cell along each coordinate axis and repeating such isolations on a rotational basis on all coordinate axes [Devroye et al. 1996, Fig. 21.3], need not be asymptotically L_1 consistent. We establish in Theorem 3.1 that the data-adaptive partitioning scheme we develop here for our multivariate histogram estimators satisfies the three conditions of Lugosi and Nobel [1996].

Once the asymptotic L_1 consistency of the data-adaptive partitioning strategy is established, there are several approaches to the problem of how to create optimally smoothed data-dependent partitions for a given dataset of finite sample size n — the smoothing problem. The common aim is to achieve some proper level of smoothing for each region of the data but to allow this to vary over the entire sample space. The differences are mainly about stopping and refining strategies: when is it 'good' to stop partitioning and how each refinement of the partition is determined. Tree based partitioning strategies are particularly suited to large sample sizes and we focus on tree based histograms here. We consider estimators from three schools of inference; two of them use the likelihood function while the third so-called L_1-school uses methods such as minimum distance estimation (MDE). 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].

1.2.1. Tree based partitions. Partitions of multi-dimensional space are usually represented using hierarchical data structures such as trees. A review of the use of trees to represent spatial data structures that discusses most of these points, and more, can be found in Samet [2006]. The main advantages are:

— Operations on the data structures are often well suited to spatial divide and conquer methods and hence to hierarchical data structures.
— A tree provides $\mathcal{O}(\log m)$ access time to any sub-box in a collection of m sub-boxes, regardless of the number of dimensions, without the need to impose a uniform grid partition on the space.
— Trees provide low-cost (constant time) insertion and deletion of sub-boxes, without the need to reallocate existing partitions in memory.
— Algorithms operating on trees can be expressed naturally and succinctly in a recursive form, allowing a simpler and more understandable programming implementation [Kruse 1987].

A tree-based structure is particularly suitable for *regular pavings*, the algebraically desirable binary tree based space partitioning structures in our approach, because an arithmetic operation on two pavings (for example, addition) requires 'matching' of pairs of sub-boxes, each pair having one sub-box from each operand paving. If a tree structure is used then the algorithm can easily be expressed and implemented very efficiently in a form that recurses simultaneously on both structures and automatically matches the appropriate boxes — this is what we mean by *tree arithmetic*.

A tree structure can be used in statistical and machine-learning algorithms for creating data-adaptive histograms. An introduction to such tree based methods is given by Devroye et al. [1996, Ch. 20]. Tree based methods have several advantages especially for density estimation in large dimensions involving massive samples sizes due to the ease with which they can be distributed over many computers using standard distributed graph algorithms [Malewicz et al. 2010; Xin et al. 2013] — a necessity when the sample dataset does not even fit in one computer. Trees are especially suitable where the algorithm uses some form of recursive partitioning strategy, often in association with a penalty function to control complexity and ensure regularity in a frequentist setting or a prior distribution in a Bayesian setting. A *greedy* partitioning algorithm makes locally optimal decisions (with respect to the chosen optimality criterion) based on the immediately available information in each step but is not guaranteed to find a globally optimal solution. Several greedy data-adaptive tree-structured histogram algorithms have been developed, including methods that grow the tree (partition) step-by-step or that grow the tree to represent the most complex allowable partition and then use a greedy algorithm to prune the tree (reduce the number of elements in the partition). For example, the tree-structured HIRED histogram of [Baltrunas et al. 2006] uses a similar recursive bisect-and-prune approach with an L_1 distance measure and a minimum bin volume to control complexity.

1.2.2. Bayesian methods. As remarked by Grenander [1981, p. 347], Bayesian inference, the oldest likelihood based method in statistics, assumes that (a) the *a priori* probabilities exist and (b) they are known at least approximately. If (a) and (b) hold for the underlying data-generating mechanism modelled via independent samples from f, say by elicitation from domain experts who are familiar with the mechanism underpinning f, then Bayesian methods that are based on the posterior distribution given by the product of the likelihood function and the prior distribution are naturally desirable.

An interesting tree-based class of priors called Pólya tree priors, proposed by Ferguson [1973] and elaborated further [Lavine 1992; 1994], are extensions of the Dirichlet process priors of Ferguson [1973] and special cases of the tailfree processes of Freedman [1963]. See Ghosh and Ramamoorthi [2003] for an introduction to such Bayesian nonparametric methods and their extensions. A Pólya tree prior starts from a bounded rectangular root box $\boldsymbol{x}_\rho \subset \mathbb{R}^d$ containing the sample points. It then defines a partition of $\boldsymbol{x}_\rho$ into sub-boxes that are seen as the new leaf nodes of the tree. The Pólya tree continues the partitioning process in a nested recursive manner by futher partitioning the current leaf boxes and assigning to each sub-box tree-path products from Beta-distributed random variables whose parameters give the priors. Pólya trees with Beta priors at each split node thus produce a nested sequence of partitions in all d dimensions and so experience exponential growth of 2^d sub-boxes when splitting each sub-box at each depth-level. Therefore these structures cannot computationally cope in dimensions larger than 2. Optional Pólya trees (OPT) of Wong and Ma [2010] partly ameliorate this by allowing optional stopping and random selection of sub-boxes to partition further. OPTs were made more efficient using a hierarchical maximum a posteriori estimate [Wong and Ma 2010] and further improved through an approximate

inference strategy via limited-lookahead optional Pólya trees coupled with piecewise linear interpolations [Jiang et al. 2016]. Another computationally efficient Bayesian sequential partitioning (BSP) method is presented by Lu et al. [2013] to obtain the marginal posterior distribution of the partition much faster than the OPT method. Our approach is closely related to the BSP method but without involving Dirichlet priors and emphasizing stricter partitioning rules to produce regularly paved partitions in order to take advantage of the ensuing algebra and arithmetic as explained in Sections 2.1.2 and 2.2.

1.2.3. Regularized maximum likelihood methods. Regularization methods for nonparmatric inference are given a self-contained introduction by [Grenander 1981] in the context of maximum likelihood estimators (MLEs) restricted over sieves whose complexity is allowed to grow at an appropriate rate with the sample size n (including the case of histograms). A common approach to create data-dependent partitions is to use a form of penalised likelihood estimator such as the Akaike information criterion (AIC) and the Bayesian information criteria (BIC). Penalised likelihood methods can be used for selecting the optimal bin width for regular histograms [Taylor 1987; Birgé and Rozenholc 2006] and irregular multivariate histograms [Castellan 1999; Rozenholc et al. 2009]. Both the AIC and the BIC may be unsuitable criteria for histogram density estimation [Massart 2007; Rozenholc et al. 2009]. As noted in [Birgé and Rozenholc 2006], in general optimality criteria that rely on an asymptotic estimate of risk (such as cross-validation) or any optimality criteria that depends on asymptotic performance (including many of the penalised likelihood approaches) do not necessarily perform well with small sample sizes (much less than 2000). Our optimal MAP estimator can also be seen as simple regularized/penalized MLE over the τ-parametric family of priors that penalize the partition based on the number of leaves. We did not find any significant improvements from simulation studies involving a wide range of densities (smooth, piece-wise constant and highly multi-modal) by using more sophisticated and computationally expensive functional forms for the penalty function. The optimal functional forms may depend on the assumptions about the unknown underlying density $f \in L_1$, as rightly remarked by Birgé and Rozenholc [2006].

1.2.4. Universal performace guarantees of minimum distance estimates. Recall that a density estimate $f_n = f_n(x; X_1, \ldots, X_n)$ is a mapping from $\left(\mathbb{R}^d\right)^{n+1}$ to $\mathbb{R}$ that is used to estimate probabilities of any Borel set $A \in \mathcal{B}^d$ or of computing the density at any point $x \in \mathbb{R}^d$. Its quality is naturally measured by how well it performs the assigned task of computing the probabilities of sets under the total variation criterion:

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

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* if it is valid no matter what the underlying f happens to be [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. But for a fixed n, however large but finite, classical studies of the rate of convergence of f_n to f, including the estimators in Sections 1.2.2 and 1.2.3 that rely on the likelihood function, require additional assumptions on the smoothness class, such as $f \in L_2$ as opposed to $f \in L_1$, and thereby violate the universality property.

Universal performance guarantee is provided by the *minimum distance estimate* 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 [Vapnik and Chervonenkis 1971] on uniform convergence of empirical proability measures over classes of sets.

1.3. Our approach and contributions

Statistical regular paving (SRP) is an extension of a regular paving (RP) [Samet 1990; Kieffer et al. 2001; Harlow et al. 2012], a class of space-partitioning trees that can facilitate efficient arithmetical operations. An SRP augments an RP by mutably caching recursively computable sufficient statistics of the data. A real mapped regular paving ($\mathbb{R}$-MRP) is an extension of an RP designed to represent a piecewise-constant function. A histogram density estimate represented as an $\mathbb{R}$-MRP can then be created from an SRP. SRPs and their $\mathbb{R}$-MRPs allow efficient arithmetical operations directly on their recursive tree structures that are closed under union operations. For density estimation purposes, such operations include (i) averaging of $\mathbb{R}$-MRP histogram states with *different* partitions (not necessarily refinements of one another) that are visited by a Markov chain Monte Carlo (MCMC) algorithm as in [Sainudiin et al. 2013], and (ii) evaluating Stone's leave-one-out cross validation score to quickly find the optimally smoothed *maximum a posteriori* (MAP) or regularized maximum likelihood estimate.

1.3.1. Optimal MAP/RML estimate. For a prior distribution parameterized by τ over SRPs, the posterior mode giving the *maximum a posterior probability* (MAP) estimate can also be thought of as a *regularized maximum likelihood* (RML) estimate, where the regularization (prior) parameter τ determines the extent of the penalty for the complexity encoded by an SRP, the adaptive histogram model of the underlying unknown density. Here we focus on the optimally smoothed RML estimate or prior-selected MAP estimate that is significantly faster to produce than the posterior mean estimate of [Sainudiin et al. 2013] with a fixed prior as described in Section 1.3.2. We view this paper's main contribution as one of fast prior selection using the leave-one-out rule during MAP/RML estimation over a set of priors parameters.

In this paper, we use a novel pair of complementary *priority-queued Markov chains* (PQMCs) to find the MAP or RML estimate over the state space of binary tree-based adaptive histograms that are explored in an asymptotically L_1-consistent manner. The first PQMC carves in the support of the density by prioritizing large regions with low empirical measure for the next bisection, while multiple independent copies of the second PQMC branches tributary paths in the state space by starting from various states in the path taken by the first PQMC. The second PQMC is complementary to the first support-carving PQMC by prioritizing the regions of the support with the highest empirical measure for the next bisection in order to uphold the L_1 consistent statistically equivalent blocks rule. The search for the posterior mode is done dynamically along all states in the tributary paths visited collaboratively by the complementary pair of PQMCs. This is how we solve the smoothing problem of determining the MAP partition for a fixed prior. Finally, using a computationally efficient tree-based scheme, we are able to evaluate the leave-one-out cross-validation score in order to determine the optimal choice for the prior/penalty/regularization parameter τ.

1.3.2. MCMC over regular pavings. For a given prior distribution over SRPs, the posterior mean can be thought of as an L_2-loss minimizing Bayesian nonparametric density estimate. Such a Bayesian smoothing based on the sample mean of a sequence of histogram states over SRP histograms visited by an MCMC algorithm with stationary samples from the posterior distribution was given in [Sainudiin et al. 2013] and further refined and automated by [Harlow 2013, Ch. 6]. The averaging algorithm exploited the

arithmetic properties of regular paving trees in order to average arbitrary partitions in the space of SRP histograms to produce an average histogram in the same space but typically with many more leaves than that of each histogram being averaged — this is fundamentally different from the Pólya tree based methods which report the posterior density of a partition resulting from a sequential refinement strategy whereby one may only (optionally) split (selected) existing leaves typically along all (or a set of chosen) coordinate directions. The averaged regular paving histogram density estimate of Sainudiin et al. [2013] was tested with uniform data in up to $d = 1,000$ dimensions with $n = 10^6$ and found that the method coped well with a mean L_1 error of 0.0012 in the universal scale of $[0, 2]$ with this type of high-dimensional unstructured data. It was also shown that the method coped well in up to 5 or 6 dimensions when up to 10^7 sample points were drawn from concentrated Gaussian and highly structured Rosenbrock densities (represented as piecewise constant approximations within 0.01 in Hellinger distance from their continuous counterparts). A particular advantage of the method is its ability to computationally cope with large volumes of data, even n as high as 10^7, in stark contrast with other available methods (in a non-distributed setting).

However, the crucial strategy to initialize the MCMC chain from states with high posterior probability, in order to minimize the chance that the chain gets stuck in low posterior states, was done in an *ad hoc* manner in [Sainudiin et al. 2013]. Another significant disadvantage of the MCMC chain was the long time to convergence, under highly conservative and heuristic conditions of Harlow [2013] to diagnose convergence, in an attempt to nearly fully automate the MCMC inference in [Sainudiin et al. 2013]. The most significant limitation of the method was the reliance on a fixed prior distribution over SRP histogram states, the so-called *natural Catalan prior*. This prior was too penalizing and did not adapt to smaller sample sizes. Therefore, large samples sizes are required to ensure the likelihood could wage a reasonable 'war of evidence' against the hyper-penalizing force of this fixed prior to reach acceptably small mean L_1 errors.

This paper addresses two major deficiencies of this estimator: (i) asymptotic justification of the initialization strategy (Theorem 3.1) from states with high posterior probability to improve mixing of the MCMC and (ii) a fast cross-validation scheme for prior selection of the MAP estimate from a univariate family of priors. Interestingly, the optimally smoothed RML or prior-selected MAP estimate based on the fast cross-validation scheme is significantly faster to produce than the MCMC-based posterior mean estimate of [Sainudiin et al. 2013] with marginally better mean integrated absolute errors. Thus, the main emphasis in this paper is on the fast prior-selected MAP estimator.

1.3.3. Minimum distance estimate. The minimum distance estimate (MDE) minimizes the distance to the empirical measure in a metric that is reminiscent of the total variation distance. The particular class of estimators studied in [Devroye and Lugosi 2001; 2004] were limited to kernel estimates and histograms under simpler partitioning rules. Inspired by this, here we develop a minimum distance estimator (MDE) over statistical regular pavings to produce nonparametric data-adaptive density estimates in d dimensions with universal performance guarantees as described in Section 1.2.4.

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. Although a more efficient algorithm (upto pre-processing the L_1 distances for each pair of densities) is characterised in [Mahalanabis and Stefankovic 2008], we are not aware of any implementations of the MDE using data-adaptive multivariate histograms that are asymptotically consistent for large volumes of data ($n = 10^6$ in

dimensions up to 6 for instance). To the best of our knowledge, ours is the only publicly available implementation of such an MDE estimator.

1.3.4. Statistical operations with regular pavings. Our regular paving histogram density estimates allow for a wide range of subsequent statistical operations as described in Section 2.2. Briefly, such statistical operations include (i) evaluating of the density over a large set of query points for prediction or cross-validation, (ii) obtaining the highest density or coverage regions to hunt for bumps and density features, (iii) getting marginal densities as $\mathbb{R}$-MRPs by tree-based integration over a subset of the coordinates, or (iv) producing conditional densities as $\mathbb{R}$-MRPs for subsequent regression, if done according to the recursive tree-based algorithms in [Harlow et al. 2012], as highlighted in Section 2.2.

1.4. Plan of the paper

The rest of the paper is laid out as follows. Section 2 introduces the algebra for RPs, the arithmetic for $\mathbb{R}$-MRPs and SRPs, and explains how various multivariate data-adaptive histogram estimates can be built using these structures. Section 3 illustrates the use of a novel partitioning strategy using a complementary pair of PQMCs on the state space of SRPs and a proof of its asymptotic L_1-consistency. In Section 4 we evaluate the performance on simulated data and in Section 5 we briefly highlight some applications. Finally in Section 6 we summarize and conclude with a discussion.

2. REGULAR PAVINGS AND HISTOGRAMS

This section introduces the notions of regular pavings (RPs), statistical regular pavings (SRPs) and real mapped regular pavings ($\mathbb{R}$-MRPs), and explains how a histogram density estimate can be built using these tree based and arithmetically amenable data structures.

2.1. Regular pavings (RPs)

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 $\mathrm{mid}\,(\boldsymbol{x}) := (\underline{x} + \overline{x})/2$. A box of dimension d with coordinates in $\Delta := \{1, 2, \ldots, d\}$ is an interval vector:

$$\boldsymbol{x} := [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_d, \overline{x}_d] =: \bigboxtimes_{j \in \Delta} [\underline{x}_j, \overline{x}_j] \ .$$

The set of all such boxes is $\mathbb{IR}^d$, i.e., the set of all interval real vectors in dimension d. Consider a box $\boldsymbol{x}$ in $\mathbb{IR}^d$. Define the index ι to be the first coordinate of maximum width:

$$\iota := \min \left(\underset{i}{\mathrm{argmax}} (\mathrm{wid}\,(\boldsymbol{x}_i)) \right) \ .$$

A *bisection* or *split* of $\boldsymbol{x}$ perpendicularly at the mid-point along this first widest coordinate ι gives the left and right child boxes of $\boldsymbol{x}$

$$\boldsymbol{x}_{\mathsf{L}} := [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_\iota, \mathrm{mid}\,(\boldsymbol{x}_\iota)) \times [\underline{x}_{\iota+1}, \overline{x}_{\iota+1}] \times \ldots \times [\underline{x}_d, \overline{x}_d] \ ,$$

$$\boldsymbol{x}_{\mathsf{R}} := [\underline{x}_1, \overline{x}_1] \times \ldots \times [\mathrm{mid}\,(\boldsymbol{x}_\iota), \overline{x}_\iota] \times [\underline{x}_{\iota+1}, \overline{x}_{\iota+1}] \times \ldots \times [\underline{x}_d, \overline{x}_d] \ .$$

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 $\boldsymbol{x}$ in $\mathbb{IR}^d$ is

known as a *regular paving* [Kieffer et al. 2001] or n-tree [Samet 1990] of $\boldsymbol{x}$. A regular paving of $\boldsymbol{x}$ can also be seen as a binary tree formed by recursively bisecting the box $\boldsymbol{x}$ 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].

When the root box $\boldsymbol{x}_\rho$ is clear from the context we refer to an RP of $\boldsymbol{x}_\rho$ as merely an RP. Each node of an RP is associated with a sub-box of the root box that can be attained by a sequence of selective regular bisections. Each node in an RP can be distinctly labelled by L and R for the sequence of left and right child node selections, respectively, from the root node.

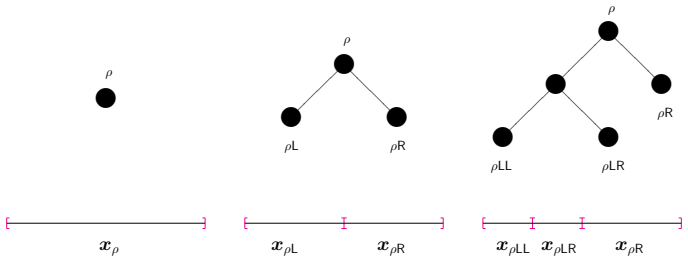


Fig. 1. A sequence of selective bisections of boxes (nodes) along the first widest coordinate, starting from the root box (root node) in one dimension, produces an RP.

The relationship of trees, labels and partitions is illustrated in Figure 1 using a simple one-dimensional example. The root node associated with root interval $\boldsymbol{x}_\rho \in \mathbb{IR}$ is labelled ρ. First, ρ is split into two child nodes, and the left child and right child nodes are labelled $\rho\mathsf{L}$ and $\rho\mathsf{R}$, respectively. The left half of $\boldsymbol{x}_\rho$ that is now associated with node $\rho\mathsf{L}$ is labelled $\boldsymbol{x}_{\rho\mathsf{L}}$. Similarly, the right half of $\boldsymbol{x}_\rho$ that is associated with the right child node $\rho\mathsf{R}$ is labelled $\boldsymbol{x}_{\rho\mathsf{R}}$. $\rho\mathsf{L}$ and $\rho\mathsf{R}$ are a pair of *sibling nodes* since they share the same parent node ρ. A node with no child nodes is called a *leaf node*. A *cherry node* is a sub-terminal node with a pair of child nodes that are both leaves. This pair of sibling nodes can be *reunited* or *merged* to its parent cherry node, thereby turning the cherry node into a leaf node. Next, the left node $\rho\mathsf{L}$ is split to get its left and right child nodes $\rho\mathsf{LL}$ and $\rho\mathsf{LR}$ with associated sub-intervals $\boldsymbol{x}_{\rho\mathsf{LL}}$ and $\boldsymbol{x}_{\rho\mathsf{LR}}$ respectively, formed by the bisection of interval $\boldsymbol{x}_{\rho\mathsf{L}}$ (because the root interval $\boldsymbol{x}_\rho$ is one-dimensional, each bisection is always on that single coordinate).

Figure 2 shows a sequence of bisections of a square (2-dimensional) root box. We start with the same sequence as in Figure 1, so the first three trees are identical to those in Figure 1 but in Figure 2 we can see the effect of always bisecting on the *first* widest coordinate. The first bisection, forming sub-boxes $\boldsymbol{x}_{\rho\mathsf{L}}$ and $\boldsymbol{x}_{\rho\mathsf{R}}$, takes place on the first widest coordinate of $\boldsymbol{x}_\rho$, which is the first coordinate because the box is a square. The next bisection, of box $\boldsymbol{x}_{\rho\mathsf{L}}$ to form $\boldsymbol{x}_{\rho\mathsf{LL}}$ and $\boldsymbol{x}_{\rho\mathsf{LR}}$, takes place on the second coordinate because this is the first widest coordinate of $\boldsymbol{x}_{\rho\mathsf{L}}$. We then extend the sequence with two further bisections. First we split the right child node $\rho\mathsf{R}$ into its child nodes $\rho\mathsf{RL}$ and $\rho\mathsf{RR}$, respectively (again bisecting on the second coordinate of $\boldsymbol{x}_{\rho\mathsf{R}}$). Then we select $\rho\mathsf{LR}$ to do a final split and obtain its child nodes $\rho\mathsf{LRL}$ and $\rho\mathsf{LRR}$. $\boldsymbol{x}_{\rho\mathsf{LR}}$ is square and is bisected on its first coordinate to form the sub-boxes $\boldsymbol{x}_{\rho\mathsf{LRL}}$ and $\boldsymbol{x}_{\rho\mathsf{LRR}}$.

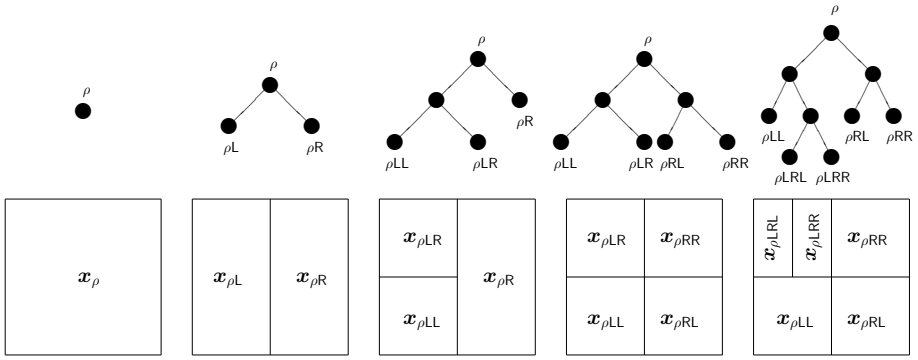


Fig. 2. A sequence of selective bisections of boxes (nodes) along the first widest coordinate, starting from the root box (root node) in two dimensions, produces an RP.

Figures 1 and 2 also illustrate an important point about these regular pavings. Because of our restricted bisection rule (splitting a box only at the mid-point along its first widest coordinate) a tree and its associated root box $\boldsymbol{x}_\rho$ will uniquely describe the partition of $\boldsymbol{x}_\rho$ into sub-boxes. In Figure 1 there is one and only one partition of the root box corresponding to each tree, and similarly in Figure 2. The same would apply for a root box of any dimension. For the same reason, if we have two RPs with the same root box and two nodes at exactly the same positions in their respective trees (i.e., nodes with the same label), then both of these nodes will have exactly the same box and can be considered to be 'equivalent'. It is this restriction that allows us to efficiently carry out operations on two regular pavings that will result in another regular paving and to extend this to arithmetic on mapped regular pavings.

We now return to a general description of regular pavings. Let the j-th interval of a box $\boldsymbol{x}_{\rho\mathsf{v}}$ be $[\underline{x}_{\rho\mathsf{v},j}, \overline{x}_{\rho\mathsf{v},j}]$. The volume or Lebesgue measure of a d-dimensional box $\boldsymbol{x}_{\rho\mathsf{v}}$ associated with the node $\rho\mathsf{v}$ of an RP of $\boldsymbol{x}_\rho$ is the product of the side-lengths of the box, i.e., $\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}}) = \prod_{j=1}^{d}(\overline{x}_{\rho\mathsf{v},j} - \underline{x}_{\rho\mathsf{v},j})$. The volume is associated with the depth of a node. The depth of a node $\rho\mathsf{v}$ in an RP is denoted by $\mathsf{d}_{\rho\mathsf{v}}$. A node has depth $\mathsf{d}_{\rho\mathsf{v}} = k$ in the tree if it can be reached by k splits from the root node. If an RP has root box $\boldsymbol{x}_\rho$ and a node $\rho\mathsf{v}$ in the regular paving has depth k, then the volume of the box $\boldsymbol{x}_{\rho\mathsf{v}}$ associated with that node is $\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}}) = 2^{-k}\mathrm{vol}\,(\boldsymbol{x}_\rho)$.

Any tree can be uniquely identified by the *sequence of its leaf node depths* if a consistent ordering of leaf nodes is used. For example the leaf nodes of the final regular paving in Figure 1, listed in left-to-right order, are $[\rho\mathsf{LL}, \rho\mathsf{LR}, \rho\mathsf{R}]$. The sequence $2, 2, 1$ uniquely identifies the tree and the tree (as discussed above) uniquely identifies the partition of the root box, and so the sequence of leaf node depths also uniquely describes the partitioning of the root box $\boldsymbol{x}_\rho$. In general, an RP is denoted by s. Where it is convenient, an RP may be labelled by its leaf node depth sequence. For example, the final regular paving in Figure 1 can be referred to as $s_{2,2,1}$ or simply s_{221} if the maximal depth is less than 10. The set of all nodes of an RP is denoted by $\mathbb{V} := \rho \cup \{\rho\{\mathsf{L},\mathsf{R}\}^j : j \in \mathbb{N}\}$. The set of all leaf nodes of an RP is denoted by $\mathbb{L}$ and the set of internal nodes by $\hat{\mathbb{V}}(s) := \mathbb{V}(s) \setminus \mathbb{L}(s)$. 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 $s \in \mathbb{S}_k$.

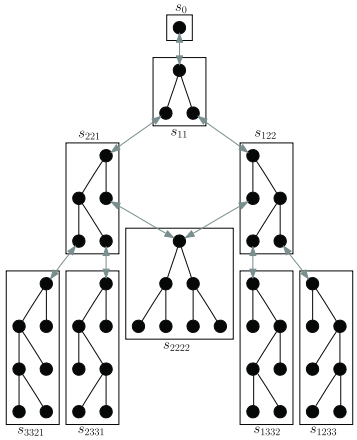


Fig. 3. Hasse (or transition) diagram over $\mathbb{S}_{0:3}$ with split/reunion transitions from one regular paving tree to another.

2.1.1. Combinatorics of regular paving trees. The number of distinct binary trees with k splits is equal to the Catalan number C_k.

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

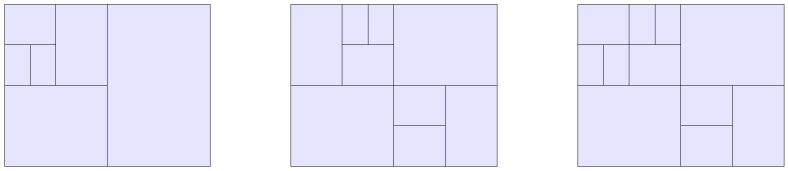
For $i, j \in \mathbb{Z}_+$, where $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$ and $i \leq j$, let $\mathbb{S}_{i:j} := \cup_{k=i}^{j} \mathbb{S}_k$ be the set of regular pavings with k splits where $k \in \{i, i+1, \ldots, j\}$. Let the set of all regular pavings be $\mathbb{S}_{0:\infty} := \lim_{j\to\infty}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

Figure 3 displays the Hasse or transition diagram over $\mathbb{S}_{0:3}$ where the gray arrows represent the immediate precedence relation from one regular paving tree to another through a split or reunion. There may be more than one path from the root node to a particular regular paving in $\mathbb{S}_k$, i.e., more than one distinct sequence of k splits may result in the same regular paving in $\mathbb{S}_k$. In Figure 3, for example, there are two paths to s_{2222}. More generally the number of distinct paths from the root node to the regular paving tree s, by splitting leaf nodes recursively, is given by:

$$\mathcal{C}(s) = \prod_{\rho\mathsf{v}\in\hat{\mathbb{V}}(s)} \binom{\wedge_{\rho\mathsf{v}}^L + \wedge_{\rho\mathsf{v}}^R}{\wedge_{\rho\mathsf{v}}^L} = (|\mathbb{L}(s)| - 1)! \prod_{\rho\mathsf{v}\in\hat{\mathbb{V}}(s)} \frac{1}{(\wedge_{\rho\mathsf{v}}^L + \wedge_{\rho\mathsf{v}}^R + 1)} \ , \tag{2}$$

where $\wedge_{\rho\mathsf{v}}^L$ and $\wedge_{\rho\mathsf{v}}^R$ are the number of split nodes in the left and right subtrees below the internal node $\rho\mathsf{v} \in \hat{\mathbb{V}}(s)$ in the regular paving tree s with $k = |\mathbb{L}(s)| - 1$ many splits. This product of binomial coefficients of splits, $\mathcal{C}(s)$, is called the *shape functional* by Dobrow and Fill [1995, Cor. 4.1] or the *Catalan coefficient* [Sainudiin 2012; Sainudiin and Véber 2016] and is the solution to an enumerative combinatorial exercise [Stanley 1997, Ch. 3, Ex. 1.b., p. 312]. The binomial terms in the product can be directly understood as the number of distinct ways in which the splits to the left and right of each internal node can be interleaved when counting the number of distinct paths through $\mathbb{S}_{0:k}$ from the root to a given $s \in \mathbb{S}_k$. These paths in $\mathbb{S}_{0:k}$ can be encoded by adding a rank from $\{1, 2, \ldots, k\}$ to each internal node to reflect the order in which it was split. This finer space of rooted planar *ranked* binary trees, is in bijective correspondence with permutations of $\{1, 2, \ldots, k\}$ through the *increasing binary tree-lifting* operation [Flajolet and Sedgewick 2009, Ex 17, p. 132] and thus satisfies $\sum_{s\in\mathbb{S}_k} \mathcal{C}(s) = k!$. Thus, regular paving trees are *rooted planar unranked binary trees*.

2.1.2. Algebra of regular paving trees. The union of two regular pavings $s^{(1)}$ and $s^{(2)}$ in $\mathbb{S}_{0:\infty}$ with the same root box $\boldsymbol{x}_\rho$ is denoted by $s^{(1)} \cup s^{(2)}$. Intuitively, the leaf boxes of the union of two regular pavings can be seen as being obtained from overlaying or superimposing the partitions of the operand regular pavings as shown in Figure 4.



(a) $s^{(1)}$. (b) $s^{(2)}$. (c) $s^{(1)} \cup s^{(2)}$.

Fig. 4. Union of two regular pavings of a root box in $\mathbb{R}^2$.

We can obtain union (or the joint partition) of two RPs through `RPUnion` [Harlow et al. 2012, Algorithm 1] by exploiting the tree structure to recurse on pairs of nodes, moving through both trees simultaneously. The tree thus provides a very simple way of matching up sub-boxes of the regular pavings 'layer by layer' in the tree hierarchies in order to find the finest partitioning of any part of the shared root box and copy it into the result of the union operation. The union operation $\cup$ is subject only to the restriction that the two operand regular pavings have the same root box $\boldsymbol{x}_\rho \in \mathbb{IR}^d$. The set of *regular pavings are closed under unions* (for a proof see [Meier 2008, Prop. 10.3]), i.e., if $s^{(1)}, s^{(2)} \in \mathbb{S}_{0:\infty}$ then $s^{(1)} \cup s^{(2)} =: s \in \mathbb{S}_{0:\infty}$.

2.2. Real mapped regular pavings ($\mathbb{R}$-MRPs)

Let $s \in \mathbb{S}_{0:\infty}$ be an RP with root node ρ and root box $\boldsymbol{x}_\rho \in \mathbb{IR}^d$. Let ${}^{\square}f : \mathbb{V}(s) \to \mathbb{R}$ map each node of s to an element in $\mathbb{R}$ as follows:

$$\{\rho\mathsf{v} \mapsto f_{\rho\mathsf{v}} : \rho\mathsf{v} \in \mathbb{V}(s), f_{\rho\mathsf{v}} \in \mathbb{R}\} \ .$$

Such a map ${}^{\square}f$, obtained by augmenting each node $\rho\mathsf{v}$ of the RP tree s with an additional data member $f_{\rho\mathsf{v}} \in \mathbb{R}$, is called an $\mathbb{R}$-*mapped regular paving* ($\mathbb{R}$-MRP). Next we show how $\mathbb{R}$-MRPs form a computationally amenable representation of piecewise constant real-valued functions (simple functions).

The sets of all nodes and leaf nodes of an $\mathbb{R}$-MRP ${}^{\square}f$ are denoted by $\mathbb{V}({}^{\square}f)$ and $\mathbb{L}({}^{\square}f)$, respectively. The set of all leaf node boxes is denoted by $\boldsymbol{x}_{\mathbb{L}({}^{\square}f)}$. The class of $\mathbb{R}$-MRPs over the leaf boxes of regular pavings of a root box $\boldsymbol{x}_\rho \in \mathbb{IR}^d$ is then

$${}^{\square}\mathcal{F} := \{\{\rho\mathsf{v} \mapsto f_{\rho\mathsf{v}} : \rho\mathsf{v} \in \mathbb{V}(s), f_{\rho\mathsf{v}} \in \mathbb{R}\} : s \in \mathbb{S}_{0:\infty}\} \ .$$

Sometimes we let ${}^{\square}f(\rho\mathsf{v}) = f_{\rho\mathsf{v}}$ for notational ease.

Arithmetic operations in $\mathbb{R}$ can be extended to $\mathbb{R}$-MRPs [Harlow et al. 2012]. For example, given any two $\mathbb{R}$-MRPs ${}^{\square}f^{(1)}$ and ${}^{\square}f^{(2)}$ with the same root box $\boldsymbol{x}_\rho$ and a binary operation $\star \in \{+, -, \cdot, /\}$, the $\mathbb{R}$-MRP ${}^{\square}f = {}^{\square}f^{(1)} \star {}^{\square}f^{(2)}$ can be obtained by `MRPOperate` [Harlow et al. 2012, Algorithm 4] — a simple extension of `RPUnion` for simple real function arithmetic. An $\mathbb{R}$-MRP ${}^{\square}f$ can also be easily transformed [Harlow et al. 2012, Algorithm `MRPTransform`] by any standard function $g \in \mathfrak{S} := \{\exp, \sin, \cos, \tan, \ldots\}$ to obtain the $\mathbb{R}$-MRP ${}^{\square}g = g({}^{\square}f)$. Finally, a binary operation of the form ${}^{\square}f \star x$ for an $\mathbb{R}$-MRP ${}^{\square}f$ and $x \in \mathbb{R}$ can also be carried out by encoding x as a constant $\mathbb{R}$-MRP on the common root box $\boldsymbol{x}_\rho$, and again the result ${}^{\square}g = {}^{\square}f \star x$ is an $\mathbb{R}$-MRP. All these properties are used to show that ${}^{\square}\mathcal{F}$ satisfies the conditions of a Stone-Weierstrass theorem and

is therefore dense in $\mathcal{C}(\boldsymbol{x}_\rho, \mathbb{R})$, the algebra of real-valued continuous functions over $\boldsymbol{x}_\rho$ [Harlow et al. 2012, Theorem 4.1]. This ensures that we can uniformly approximate any continuous density $f : \boldsymbol{x}_\rho \to \mathbb{R}$ using $\mathbb{R}$-MRPs in ${}^{\square}\mathcal{F}$. Note that $\mathbb{S}_{0:\infty}$, the set of RP tree partitions, contains the set of dyadic partitions of $\boldsymbol{x}_\rho$ represented by complete dyadic binary trees: $\{\mathbb{S}_0, \mathbb{S}_2, \mathbb{S}_4, \mathbb{S}_8, \ldots\} \subset \mathbb{S}_{0:\infty}$. Thus $\mathbb{S}_{0:\infty}$ can be used in principle to obtain any σ-additive measure using standard measure-theoretic constructions (but such constructions without efficiency considerations may be limited in practice due to finite machine memory and computing time).

Moreover, we can obtain an $\mathbb{R}$-*MRP arithmetical expression* that is specified by a directed acyclic graph made of finitely many sub-expressions involving constant $\mathbb{R}$-MRPs, binary arithmetic operations over two $\mathbb{R}$-MRPs, standard transformations of $\mathbb{R}$-MRPs by elements of $\mathfrak{S}$ and their compositions. Thus we can obtain $\mathbb{R}$-MRPs as arithmetical expressions of other $\mathbb{R}$-MRPs allowing us to work with recursively computable and machine-representable tree based set functions in the classical sense of Hahn and Rosenthal [1948]. The advantage of an $\mathbb{R}$-MRP representation is that all the arithmetic operations between real-valued simple functions in ${}^{\square}\mathcal{F}$ described above as well as several useful statistical operations as detailed in [Harlow et al. 2012] can be carried out efficiently and recursively using trees (see Section 5.1).

2.3. Statistical regular pavings (SRPs)

A *statistical regular paving* (SRP) 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. An SRP is effectively an association of a collection of data (the *data sample* or *data set*) with an RP-based structure where the nodes have additional properties:

— A node of an SRP tree can be associated with a subset of the sample data;
— A node of an SRP tree records recursively computable statistics relating to this sample subset.

An SRP is denoted by s. Denote $\mathbb{S}_{i:j}$ as the set of all statistical regular pavings with a given root box and k splits where $k \in \{i, i+1, \ldots, j\}$, where $i, j \in \mathbb{Z}_+$ and $i \leq j$. The space of all statistical regular pavings with a given root box is then $\mathbb{S}_{0:\infty} := \lim_{i \to \infty} \mathbb{S}_{0:i}$

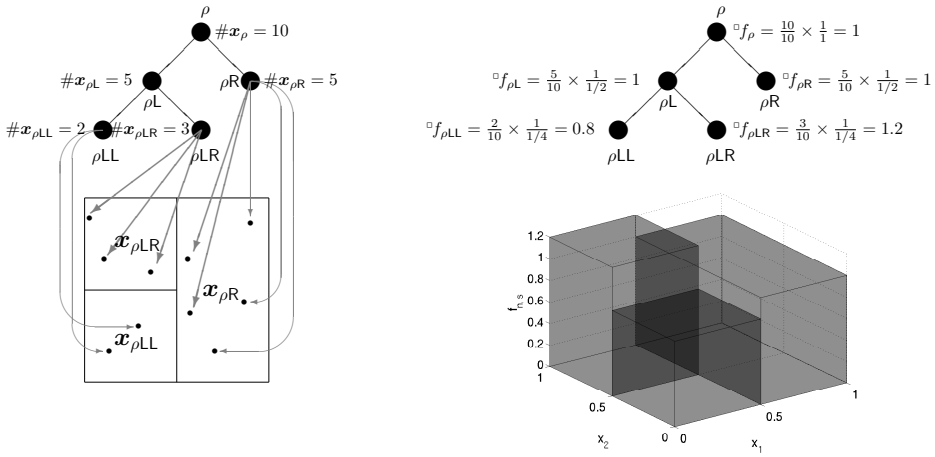
Take a data sample of size n, $X_1, X_2, \ldots, X_n$ and an SRP s. For convenience the sample will be referred to as ${}^{n}X$. Let ${}^{\subset n}X$ be a subset of ${}^{n}X$ and let ${}^{\subset n}X_{\rho\mathsf{v}}$ be the subset of ${}^{n}X$ contained in the box $\boldsymbol{x}_{\rho\mathsf{v}}$ associated with a node $\rho\mathsf{v}$ in s.

A recursively computable statistic of some data is a statistic whose value can be updated from the addition of new data using only the current value of the statistic and the new data (i.e., it is not necessary to know the individual data values from which the current value of the statistic is calculated). Formally, if $T({}^{\subset n}X)$ is some statistic of ${}^{\subset n}X$ and a new data point x is added to ${}^{\subset n}X$ so that $n' = n + 1$ and ${}^{\subset n'}X = {}^{\subset n}X \cup x$, then $T({}^{\subset n'}X)$ can be calculated using $u(T({}^{\subset n}X), x)$ where u is some updating function.

Recursively computable properties of sufficient statistics were discussed by Fisher [1925] and are used in the tree-based structures of [Gray and Moore 2003a] for example. For the purpose of this paper, the only such statistic that an SRP node $\rho\mathsf{v}$ is required to keep is the count of the number of data points in ${}^{\subset n}X_{\rho\mathsf{v}}$. This count is denoted by $\#\boldsymbol{x}_{\rho\mathsf{v}} = |{}^{\subset n}X_{\rho\mathsf{v}}|$. 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 $\mathbb{L}^+(s) := \{\rho\mathsf{v} \in \mathbb{L}(s) : \#\boldsymbol{x}_{\rho\mathsf{v}} > 0\} \subseteq \mathbb{L}(s)$. We use C-XSC's dot-precision accumulators [Hofschuster 2003] to recursively compute the mean vector and covariance matrix in each box, by rigorously and efficiently accounting for numerical errors. Although we do not show the results based on using these

first and second sample moment statistics in this work, they are readily available for exploring randomized algorithms that rely on them.

Figure 5 depicts a small SRP s with root box $\boldsymbol{x}_\rho \in \mathbb{IR}^2$. The number of sample data points in the root box $\boldsymbol{x}_\rho$ is 10. Figure 5(a) 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 5(b) shows how the density estimate is computed from the count and the volume of each box as ${}^{\square}f_{\rho\mathsf{v}}$ to obtain the $\mathbb{R}$-MRP ${}^{\square}f$ and its restriction to leaf boxes to obtain the density estimate $f_{n,s}$ as an SRP histogram.



(a) An SRP tree and its constituents. (b) An SRP histogram $f_{n,s}$ and its $\mathbb{R}$-MRP ${}^{\square}f$.

Fig. 5. An SRP and its corresponding histogram.

2.4. Statistical regular paving (SRP) histograms and their likelihoods

Given the count data recorded by each node, an SRP associated with data ${}^{n}X$ can be used to form a histogram. The bins are the elements in the partition, i.e., the boxes associated with the leaf nodes $\boldsymbol{x}_{\mathbb{L}(s)}$. If the total number of data points associated with the whole of an SRP s with root node ρ and root box $\boldsymbol{x}_\rho$ is $n = \#\boldsymbol{x}_\rho = \sum_{\rho\mathsf{v}\in\mathbb{L}(s)} \#\boldsymbol{x}_{\rho\mathsf{v}}$, then the corresponding histogram is:

$$f_{n,s}(x) = f_n(x) = \sum_{\rho\mathsf{v}\in\mathbb{L}(s)} \frac{\mathbb{1}_{\boldsymbol{x}_{\rho\mathsf{v}}}(x)}{n} \left(\frac{\#\boldsymbol{x}_{\rho\mathsf{v}}}{\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}})} \right) \quad . \tag{3}$$

A histogram obtained using Equation (3) is referred to as an SRP histogram. 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 ${}^{\square}\mathcal{F}$ closed under addition and scalar

multiplication and thereby allow for computationally efficient averaging of histograms with different partitions.

Recalling that the volume of a box is related to the depth of a node and the volume of the root box $\boldsymbol{x}_\rho$, the volume of the box $\boldsymbol{x}_{\rho\mathsf{v}}$ of a node $\rho\mathsf{v}$ with depth $\mathsf{d}_{\rho\mathsf{v}}$ is $\mathrm{vol}\left(\boldsymbol{x}_{\rho\mathsf{v}}\right) = \mathrm{vol}\left(\boldsymbol{x}_\rho\right)2^{-\mathsf{d}_{\rho\mathsf{v}}}$, we can recursively reexpress the log-likelihood in terms of leaf node depths as follows:

$$\begin{aligned}\ell\left(f_n\right) = \log\mathcal{L}\left(f_n\right) &= \sum_{\rho\mathsf{v}\in\mathbb{L}(s)} \#\boldsymbol{x}_{\rho\mathsf{v}}\log\left(\#\boldsymbol{x}_{\rho\mathsf{v}}\right) - \sum_{\rho\mathsf{v}\in\mathbb{L}(s)} \#\boldsymbol{x}_{\rho\mathsf{v}}\log\left(\mathrm{vol}\left(\boldsymbol{x}_{\rho\mathsf{v}}\right)\right) - n\log\left(n\right) \\ &= \sum_{\rho\mathsf{v}\in\mathbb{L}(s)} \#\boldsymbol{x}_{\rho\mathsf{v}}\log\left(\#\boldsymbol{x}_{\rho\mathsf{v}}\right) + \log\left(2\right)\sum_{\rho\mathsf{v}\in\mathbb{L}(s)} \mathsf{d}_{\rho\mathsf{v}}\#\boldsymbol{x}_{\rho\mathsf{v}} - n\log\left(\mathrm{vol}\left(\boldsymbol{x}_\rho\right)\right) - n\log\left(n\right) \quad .\end{aligned} \tag{4}$$

Using this recursive structure we can efficiently track the two elementary changes in the log-likelihoods when searching for an optimal SRP state along a path of a Markov chain over SRP states such that two consecutive states differ only by a split or a merge. These elementary changes in log-likelihood are (i) $\delta\ell_{\bigtriangledown_{\mathrm{SRP}}(\rho\mathsf{v})}$ or (ii) $\delta\ell_{\bigtriangleup_{\mathrm{SRP}}(\rho\mathsf{v})}$, due to (i) splitting a single leaf node $\rho\mathsf{v}$ to form two child nodes $\rho\mathsf{vL}$ and $\rho\mathsf{vR}$ or (ii) merging the children $\rho\mathsf{vL}$ and $\rho\mathsf{vR}$ of a single cherry or sub-terminal node $\rho\mathsf{v}$ back into $\rho\mathsf{v}$, respectively, and can be computed by:

$$\begin{aligned}\delta\ell_{\bigtriangledown_{\mathrm{SRP}}(\rho\mathsf{v})} &= \#\boldsymbol{x}_{\rho\mathsf{vL}}\log\left(\#\boldsymbol{x}_{\rho\mathsf{vL}}\right) + \#\boldsymbol{x}_{\rho\mathsf{vR}}\log\left(\#\boldsymbol{x}_{\rho\mathsf{vR}}\right) - \#\boldsymbol{x}_{\rho\mathsf{v}}\log\left(\#\boldsymbol{x}_{\rho\mathsf{v}}\right) + \#\boldsymbol{x}_{\rho\mathsf{v}}\log\left(2\right) \\ \delta\ell_{\bigtriangleup_{\mathrm{SRP}}(\rho\mathsf{v})} &= \#\boldsymbol{x}_{\rho\mathsf{v}}\log\left(\#\boldsymbol{x}_{\rho\mathsf{v}}\right) - \left(\#\boldsymbol{x}_{\rho\mathsf{vL}}\log\left(\#\boldsymbol{x}_{\rho\mathsf{vL}}\right) + \#\boldsymbol{x}_{\rho\mathsf{vR}}\log\left(\#\boldsymbol{x}_{\rho\mathsf{vR}}\right)\right) - \#\boldsymbol{x}_{\rho\mathsf{v}}\log\left(2\right) \quad .\end{aligned} \tag{5}$$

The optimal SRP state of interest in this work is the Bayesian maximum *a posteriori* or MAP estimate, which depends on the SRP likelihood and a prior distibution on SRPs.

Before we delve into the Bayesian setting, we emphasize the representation of an SRP histogram estimate as an $\mathbb{R}$-MRP for the arithmetic it allows. The SRP structure is designed for analysing and organising data and controlling Markov chains to drive data-adaptive paritioning, but it is not directly capable of arithmetic over pairs of distinct SRPs. However, since an SRP histogram f_n is a piecewise-constant function that can be represented as an $\mathbb{R}$-MRP ${}^{\square}f_n \in {}^{\square}\mathcal{F}$, all the $\mathbb{R}$-MRP operations described in Section 2.2 can be carried out with the $\mathbb{R}$-MRP histogram density estimate formed from the SRP. We take advantage of this recursive tree arithmetic with ${}^{\square}f_n$ and between ${}^{\square}f_n$ and f_n for fast prior selection via cross-validation in Section 2.5.5.

2.5. Bayesian posterior mean and fast MAP estimation with prior selection

Given realisations of the sample data $x_1, \ldots, x_n$, let π be the posterior distribution that is proportional to the product of the likelihood of the data given SRP s and the prior probability of s:

$$\pi(s) := \Pr\{s|x_1,\ldots,x_n\} \propto \Pr\{x_1,\ldots,x_n|s\}\Pr\{s\} \approx \mathcal{L}\left(f_n\right)\Pr\{s\} \quad .$$

Here, the likelihood of the data given SRP s is approximated by the maximum likelihood value from the histogram f_n with bins given by the partition $\boldsymbol{x}_{\mathbb{L}(s)}$ of the root box of s. We do not take Dirichlet priors as done in fully Bayesian settings primarily because we are also interested in datasets that are known to have full compact null sets within the root box (Section 5.2). See Section 6.2 for a discussion on natural Bayesian extensions of our approach.

We note that our MAP estimation can also be viewed from a frequentist perspective as regularized (penalized) maximum likelihood estimation. Our main focus here is on

constructive and recursively computable tree-based methods that exploit the arithmetic properties of regular pavings in order to efficiently find the prior-selected MAP estimate or, equivalently, find the optimally regularized maximum likelihood estimate. We defer the problem of optimizing the exact parametric forms for the penalty or prior for the future and focus here on the randomized search algorithm using a simple parametric family of priors that penalize partitions proportional to their size. We use the Bayesian language in this paper while minding the alternative regularized or penalized likelihood perspective.

2.5.1. A nonparametric family of priors. We want our prior distribution $\{\Pr(s)\}$ over $s \in \mathbb{S}_{0:\infty}$ to be proper and uninformative in some natural sense. Moreover, we also want our prior probabilities to decrease as the partition size increases in order to penalise large partitions. With these considerations, we proposed a nonparametric Catalan family of proper priors indexed by any convergent decreasing sequence [Sainudiin et al. 2013]. The prior probabilities should decrease as the partition size increases in order to penalise large partitions. An $\{a_k\}$-*penalised uninformative proper Catalan prior* that assigns states in $\mathbb{S}_k$ with probability $\frac{a_k}{a}$ and distributes this mass uniformly over $\mathbb{S}_k$ is given by:

$$\Pr(s) = \sum_{k=0}^{\infty} \mathbb{1}_{\mathbb{S}_k}(s) \frac{a_k}{a C_k} \ , \tag{6}$$

where $\{a_k\}$ for $k = 1, 2, \ldots$ is any decreasing sequence of positive real numbers such that $\sum_{k=1}^{\infty} a_k = a < \infty$.

2.5.2. Bayesian posterior mean estimate: smoothing by averaging. Using a fixed *natural Catalan prior* with $a_k = 1/C_k$ and $a = 2 + 4\pi/3^{5/2}$, Sainudiin et al. [2013] developed a Metropolis-Hastings Markov chain with a stationary distribution approximating the posterior distribution over $\mathbb{S}_{0:\infty}$ and estimated the expectation under this posterior distribution by exploiting the $\mathbb{R}$-MRP arithmetic properties to average samples from the chain over SRP histograms with different partitions. This method was further improved and automated by Harlow [2013]. Although the method performed extremely well for unstructured (uniform or mixtures of uniform) densities in high dimensions, it was limited by (1) a fixed prior and (2) by the slow mixing of the MCMC algorithm under its simplistic *stay-split-merge* base Markov chain and highly conservative (but heuristic) convergence diagnostics based on running multiple chains, especially for highly structured densities in dimensions more than 3 (and impractical for dimensions more than 6) with over 10^6 sample points. We focus here on removing the fixed prior assumption through a parametric family of priors and selecting the best prior using a computationally efficient and asymptotically L_1 consistent randomized algorithm for MAP estimation. This is to be contrasted with the slower but generally more accurate MCMC based posterior mean estimate on a fixed prior by Sainudiin et al. [2013] and further refined and automated by [Harlow 2013, Ch. 6] as briefly discussed in Section 1.3.2 and further elaborated below.

2.5.3. A parametric family of priors. We are interested in prior selection here over $\{\Pr(s;\tau) : \tau \in (0,\infty)\}$, a τ-parametric family of $\{a_k(\tau)\}$-penalized uninformative proper Catalan priors, given by:

$$\Pr(s;\tau) = \sum_{k=0}^{\infty} \mathbb{1}_{\mathbb{S}_k}(s) \Big(e^{\frac{1}{\tau}} - 1 \Big) e^{\left(-\frac{k+1}{\tau}\right)}, \ \text{ as specified in (6) by } \ a_k(\tau) = C_k e^{\left(-\frac{k+1}{\tau}\right)} \ .$$

This family has a physical interpretation in terms of the temperature parameter τ. As τ increases, the prior "heats up" becoming increasingly dissipative over $\mathbb{S}_{0:\infty}$ and

thereby penalizing models with more leaves less harshly in comparison to those with fewer leaves while maintaining uniformity in prior distribution over all trees with the same number of leaves at any given τ. And as τ decreases, the prior "freezes down" becoming increasingly concentrated on the root node in $\mathbb{S}_0$ and thereby penalizing models with more leaves more harshly.

2.5.4. MAP estimation. Let $m = m(s) = |\mathbb{L}(s)|$ denote the number of cells or leaf boxes in the partition given by the SRP s, i.e., $m = k + 1$ if $s \in \mathbb{S}_k$. Thus, the τ-parametric posterior probability of $s \in \mathbb{S}_{0:\infty}$, up to proportionality, is:

$$\pi(s;\tau) \propto \mathcal{L}(f_n) \left(e^{\frac{1}{\tau}} - 1\right) e^{\left(-\frac{m(s)}{\tau}\right)} \ . \tag{7}$$

For a given τ, finding the MAP estimate $f_{n,\tau}$, which may also be interpreted as a regularized (penalized) maximum likelihood estimate, amounts to the following maximization problem over $\mathbb{S}_{0:\infty}$:

$$f_{n,\tau} := \underset{s \in \mathbb{S}_{0:\infty}}{\operatorname{argmax}} \log(\pi(s;\tau)) = \underset{s \in \mathbb{S}_{0:\infty}}{\operatorname{argmax}} \ \log \mathcal{L}(f_n) - \left(\frac{m(s)}{\tau}\right) \ .$$

Here we develop a novel randomized algorithm that efficiently finds $f_{n,\tau}$ along a sequence of data-adaptively partitioned SRP states that is proved in Section 3 to be asymptotically L_1-consistent using the result of Lugosi and Nobel [1996] as detailed by Devroye et al. [1996, chap. 21]. Finally, since $\mathbb{S}_{0:\infty}$ is discrete and our prior is unimodal, the sequence of Bayes estimators under the shrinking 0-1 loss approaches the MAP estimator. Moreover, the MAP estimator can be used to inform the initial conditions and minimize burn-in time for the multiple independent MCMC-based posterior mean estimate, i.e., the Bayes estimator under the L_2-loss, as already done by Harlow [2013] (see Section 2.5.2). However, in this paper our main emphasis is on MAP estimation and prior selection.

2.5.5. Prior selection via cross-validation. Note that for any fixed sample size n, as $\tau \to 0$, the MAP estimate $f_{n,\tau}$ approaches the SRP with the unsplit root node, i.e., the uniform density over the root box containing the sample data, and as $\tau \to \infty$, the proper prior becomes increasingly uniform over $\mathbb{S}_{0:\infty}$ and $f_{n,\tau}$ approaches the degenerate maximum likelihood estimator. Thus, for any fixed n, we have a smoothing or prior-selection problem of finding the optimal parameter $\check{\tau}$ between 0 and ∞.

We will solve this prior-selection problem by minimizing Stone's leave-one-out cross-validation score $\widehat{J}(\tau)$, a nearly unbiased estimator of the expected L_2 loss $\int (f_{n,\tau}(x) - f(x))^2 dx$ (eg. [Wasserman 2003]), given by:

$$\widehat{J}(\tau) = \int \left(f_{n,\tau}(x)\right)^2 dx - \frac{2}{n} \sum_{i=1}^{n} f_{n,\tau}^{(-i)}(x_i) \ ,$$

where $f_{n,\tau}^{(-i)}$ is the MAP estimate obtained from the data set by leaving out x_i. To compute the first term for $\widehat{J}(\tau)$ we simply square the $\mathbb{R}$-MRP representation ${}^{\square}f_{n,\tau}$ of $f_{n,\tau}$, the MAP (SRP histogram) estimate associated with the underlying SRP s for the fixed prior parameter τ, and compute the integral. Both these operations are possible efficiently and exactly for $\mathbb{R}$-MRPs as discussed in Section 2.2. The second term for $\widehat{J}(\tau)$ simplifies to the following sum involving ${}^{\square}f_{n,\tau}$ and $f_{n,\tau}$ over the non-empty leaf nodes of s:

$$-2/(n-1) \sum_{\rho \mathsf{v} \in \mathbb{L}^+(s)} (\#\boldsymbol{x}_{\rho \mathsf{v}} - 1) \, {}^{\square}f_{n,\tau}(\rho \mathsf{v}) \ .$$

This sum is accumulated efficiently using tree look-ups by recursively descending into ${}^{\square}f_{n,\tau}$ and $f_{n,\tau}$ simultaneously until the non-empty leaf nodes of $f_{n,\tau}$ are reached (for details see `getLeave1OutCVScore` in MRS 2.0 [Sainudiin et al. 2016]).

Thus the prior-selected MAP estimate $f_{n,\check{\tau}}$ is optimally L_2-smoothed corresponding to the optimal prior parameter:

$$\check{\tau} = \operatorname*{argmin}_{\tau \in (0,\infty)} \widehat{J}(\tau) \ .$$

The actual minimization over the prior parameter $\tau \in (0,\infty)$ is performed by an adaptive grid-based heuristic search method within a wide-enough interval $[\underline{\tau}, \overline{\tau}]$.

Remark 2.1. Note that using L_2 smoothing does indeed restrict the unknown f to be in L_2 as opposed to L_1. We make this restrictive assumption due to the computational convenience of $\widehat{J}(\tau)$ and do address the problem of computationally amenable represention of the Yatracos class over $\mathbb{S}_{0:\infty}$, needed for minimum distance estimation in the vein of the L_1 school [Devroye and Lugosi 2001, chap. 6] with universal performance guarantees in Section 3.4. Moreover, $f_{n,\check{\tau}}$ can not only be used to speed-up the computation of the posterior mean estimate under the optimal prior $\check{\tau}$, but also to obtain a good set of candidate histograms to select from during minimum distance estimation.

3. PARTITIONING WITH A COMPLEMENTARY PAIR OF PRIORITY-QUEUED MARKOV CHAINS

Here we describe the priority-queued Markov chain (PQMC), our fundamental data-adaptive domain partitioning strategy, by starting from the motivating primitive pair of "yin and yang" chains with explicit path and state probabilities. Then we describe how two complementary PQMCs collaborate to explore adaptive histograms over the state space of SRPs to find a τ-specific MAP estimate $f_{n,\tau}$. We establish the asymptotic L_1 consistency of this partitioning strategy.

3.1. A complementary primitive pair of yin and yang Markov chains on regular pavings

Without loss of generality, let f be a probability density supported on $\boldsymbol{x}_\rho = [0,1]^d$. Consider the Markov chain $\{S(t)\}_{t\in\mathbb{Z}_+}$ on $\mathbb{S}_{0:\infty}$ initialized at the root regular paving $s(0)$ with root box $\boldsymbol{x}_\rho$. For the current RP state s of the chain, associate the *split transition probability* $p(\rho\mathsf{v})$ for each leaf node $\rho\mathsf{v} \in \mathbb{L}(s)$, such that $\sum_{\rho\mathsf{v}\in\mathbb{L}(s)} p(\rho\mathsf{v}) = 1$. The Hasse diagram on $\mathbb{S}_{0:\infty}$ limited to the immediate precendence relation based on a single split when weighted by the corresponding split transition probability gives the transition diagram of our primitive Markov chain $\{S(t)\}_{t\in\mathbb{Z}_+}$ with $S(0) = s(0)$.

If $p(\rho\mathsf{v})$ only depends on invariant features of $\rho\mathsf{v}$, such as $\boldsymbol{x}_{\rho\mathsf{v}}$, $\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}})$ and $\int_{\boldsymbol{x}_{\rho\mathsf{v}}} f(x)dx$, then all the path probabilities of reaching an RP $s = s(k) \in \mathbb{S}_k$, are identically equal to:

$$\begin{aligned}&\Pr\left(S(0) = s(0), S(1) = s(1), \ldots, S(k) = s(k) = s\right)\\ &\qquad = \prod_{t=1}^{k} \Pr\left(S(t) = s(t) | S(t-1) = s(t-1)\right) \Pr\left(S(0) = s(0)\right) = \prod_{\rho\mathsf{v}\in\hat{\mathbb{V}}(s)} p(\rho\mathsf{v}) \ ,\end{aligned}$$

because the order in which one splits the internal nodes of s in $\hat{\mathbb{V}}(s)$, corresponding to each distinct path to reach s from the root, does not affect the total product of probabilites. Therefore, by (2), the probability that the chain visits a particular $s \in \mathbb{S}_k$ is

simply:

$$\Pr\left(S(k)=s|S(0)=s(0)\right)=\mathcal{C}(s)\prod_{\rho\mathsf{v}\in\hat{\mathbb{V}}(s)}p(\rho\mathsf{v})=k!\prod_{\rho\mathsf{v}\in\hat{\mathbb{V}}(s)}\frac{p(\rho\mathsf{v})}{(\wedge^{L}_{\rho\mathsf{v}}+\wedge^{R}_{\rho\mathsf{v}}+1)} \tag{8}$$

Now, we get our *yang chain* that splits leaves proportional to the probability of the corresponding leaf boxes with p in (8) given by:

$$p(\rho\mathsf{v})=\int_{\boldsymbol{x}_{\rho\mathsf{v}}}f(x)dx\ , \tag{9}$$

and our *yin chain* that proportionally splits the leaf nodes with the least probability but with the most Lebesgue measure (or vol) with p in (8) given by:

$$p(\rho\mathsf{v})=\frac{\tilde{p}(\rho\mathsf{v})}{N_s},\quad \tilde{p}(\rho\mathsf{v}):=\left(1-\int_{\boldsymbol{x}_{\rho\mathsf{v}}}f(x)dx\right)\mathrm{vol}\left(\boldsymbol{x}_{\rho\mathsf{v}}\right),\quad N_s:=\sum_{\rho\mathsf{v}\in\mathbb{L}(s)}\tilde{p}(\rho\mathsf{v})\ . \tag{10}$$

The exact probabilities for these "yin and yang" Markov chains can be obtained from the above three equations: (8), (9) and (10). The yin and yang chains have their natural empirical extensions over the space of statistical regular pavings (SRPs) such that recursively computable statistics at each node $\rho\mathsf{v}$ determine its transition probability $p(\rho\mathsf{v})$ and thereby uphold (8). For instance, $\int_{\boldsymbol{x}_{\rho\mathsf{v}}}f(x)dx$ is replaced by the empirical measure $\#\boldsymbol{x}_{\rho\mathsf{v}}/n$ in (9) and (10) to obtain the empirical variants over SRPs for the yang and yin chains, respectively. These chains and their empirical variants were chosen as our primitive complementary Markov chains due to their simplicity, computational ease and interpretability. This choice was based on extensive simulation experiments with 16 other natural splitting probabilities that also depended on higher-order recursively computable statistics of $\rho\mathsf{v}$, including sample mean vector and variance-covariance matrix (not reported here, but see [Sainudiin et al. 2016]).

Finally, our primitive yin and yang pair of chains can be made to collaborate in complementary ways in order to stochastically search for optimal data-adaptive partitions by simply starting a set of yang chains from states along the path of a yin chain (reminiscent of a tributary system of paths taken by independent yang chains starting from states along a path taken by the yin chain). They are complementary because the yin chain focusses on carving away large regions of the root box with hardly any density while the yang chain initialized from various support-carved states of the yin chain can focus on the splitting of regions with the most density in order to achieve the classical statistical rule of nearly *statistically equivalent blocks* [Tukey 1947].

There is a disadvantage with this yin and yang partitioning strategy if one is primarily interested in finding the best partitions (say in the MAP sense or the regularized-ML sense) as efficiently as possible. The empirical versions of these chains require the intensive updating of simulable large discrete probability vectors $(p(\rho\mathsf{v}):\rho\mathsf{v}\in\mathbb{L}(s))$ after every split. More crucially, the state probabilities under the primitive Markov chain on $\mathbb{S}_{0:\infty}$ allows transitions by splitting *any* leaf node of the current state and this generally leads to unnecessarily dissipative state probabilities over $\mathbb{S}_k$ — in the sense of reaching states in $\mathbb{S}_k$ that do not have the highest state probability often enough. Our simple *randomized greedy* solution to make these yin and yang chains more concentrated along paths in $\mathbb{S}_{0:\infty}$ with the highest state probabilities is by introducing a randomized priority queue for the transition probabilities so that only nodes with the highest p in (8) are split. As explained in the next section, the priority-queued empirical variants of the yin and yang chain yield data-adaptive RP partitions that are not only computationally efficient but also satisfy the three conditions of Lugosi and Nobel [1996] for their asymptotic L_1 consistency.

3.2. Priority-queued Markov chains for data-adaptive partitioning

The pseudo-code to generate sample paths from a generic priority-queued Markov chain (PQMC) over the state space of statistical regular pavings is given in Algorithm 1. A leaf node of a statistical regular paving (SRP) is splittable if it contains data and the child nodes from the split can be represented in the computer (as described in the next paragraph). A PQMC is a Markov chain on SRPs whose transition probabilities are given by ordering the elements of $\mathbb{L}^{\triangledown}(s)$, the splittable leaf nodes of the current SRP state s, by a randomized queue that is prioritized according to a given priority function $\psi : \mathbb{L}^{\triangledown}(s) \to \mathbb{R}$. This priority-queued collection of splittable leaf nodes is used to select the next node to be split from $\operatorname{argmax}_{\rho\mathsf{v}\in\mathbb{L}^{\triangledown}(s)} \psi(\rho\mathsf{v})$, the set of splittable leaf nodes of s which are equally 'large' when measured using ψ. If there is more than one such 'largest' node the choice is made uniformly at random from this set; this is the 'randomized' aspect of the process.

Three criteria can be specified to stop the PQMC. A straightforward stopping condition is to stop partitioning when the number of leaves in the SRP reaches a specified maximum $\overline{m}$. The other stopping condition relates to the priority function so that partitioning stops when the value of the largest node under the priority function ψ is less than or equal to a specified value $\overline{\psi}$. A PQMC will also stop partitioning if there are no splittable leaf nodes in the SRP, a constraint that can be naturally reached due to finite precision and thus making the Markov chains necessarily have a finite state space in its machine implementable version (see [Sainudiin et al. 2013, Sec. 3] or [Harlow 2013, App. C]). Such a constraint can also be algorithmically imposed by requiring for instance that a leaf node is splittable only if the box associated with it has volume larger than some specified $\epsilon > 0$. The latter consideration is helpful (i) to account for the *limits on empirical resolution* [Sainudiin 2005; Ferson et al. 2007], i.e., when the machine sensor generating the data has finite empirical resolution and the root box is aligned with the sensor's measurement range as in the case of the chaotic double pendulum in Section 5.2 or (ii) when the sample size is so large that we need to account for finite machine memory to represent the SRPs, even when distributed over multiple machines or (iii) to simply prevent the PQMC from visiting states with too many leaves at each level in a hierarchial partitioning strategy.

ALGORITHM 1: $\texttt{PQMC}(s, \psi, \overline{\psi}, \overline{m})$

input : s, initial SRP with root box $\boldsymbol{x}_\rho$,
$\psi : \mathbb{L}^{\triangledown}(s) \to \mathbb{R}$, a priority function,
$\overline{\psi}$, maximum value of $\psi(\rho\mathsf{v}) \in \mathbb{L}^{\triangledown}(s)$ 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 $[s(0), s(1), \ldots, s(T)]$ such that $\mathbb{L}^{\triangledown}(s(T)) = \emptyset$ or $\psi(\rho\mathsf{v}) \leq \overline{\psi}$ $\forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s(T))$ or $|\mathbb{L}(s(T))| \leq \overline{m}$.

initialize: s $\leftarrow [s]$

while $\mathbb{L}^{\triangledown}(s) \neq \emptyset \;\;\&\;\; |\mathbb{L}(s)| < \overline{m} \;\;\&\;\; \psi\left(\operatorname{argmax}_{\rho\mathsf{v}\in\mathbb{L}^{\triangledown}(s)} \psi(\rho\mathsf{v})\right) > \overline{\psi}$ **do**

$\rho\mathsf{v} \leftarrow \texttt{random\_sample}\left(\underset{\rho\mathsf{v}\in\mathbb{L}^{\triangledown}(s)}{\operatorname{argmax}}\, \psi(\rho\mathsf{v})\right)$ `// sample uniformly from nodes with largest` ψ

$s \leftarrow s$ with node $\rho\mathsf{v}$ split `// split the sampled node and update` s

$\texttt{s.append}(s)$ `// append the new SRP state with an additional split`

end

The output of $\texttt{PQMC}(s, \psi, \overline{\psi}, \overline{m})$ algorithm is $[s(0), s(1), \ldots, s(T)]$, a sequence of SRP states giving a sample path from the PQMC $\{S(t)\}_{t\in\mathbb{Z}_+}$ on $\mathbb{S}_{0:\overline{m}-1}$, such that $\mathbb{L}^{\triangledown}(s(T)) =$

$\emptyset$ or $\psi(\rho\mathsf{v}) \leq \overline{\psi}\ \forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s(T))$ or $|\mathbb{L}(s(T))| \leq \overline{m}$ and T is a corresponding random stopping time. Care must be taken in defining splittable nodes as well as the values of the stopping parameters $\overline{\psi}$ and $\overline{m}$, to ensure that $\psi(\rho\mathsf{v}) \leq \overline{\psi}\ \forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s(T))$ ***and*** $|\mathbb{L}(s(T))| \leq \overline{m}$. In Section 3.3, we specify the conditions that need to be met by these stopping parameters, as functions of the sample size n, in order to ensure that the data-adaptive partitioning strategy of the PQMC is asymptotically L_1-consistent, as $n \to \infty$.

If the initial state $S(t=0)$ is the root $s \in \mathbb{S}_0$ then PQMC $\{S(t)\}_{t \in \mathbb{Z}_+}$ 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. Note that the initial state can be specified by a sample from any distribution on $\mathbb{S}_{0:\overline{m}-1}$. In fact, we will use a distribution defined from sample paths of one PQMC with a specific priority function to initialize several independent PQMCs with a different and complementary priority function to find our MAP density estimate. These complementary PQMCs which will be used collaboratively in order to search the space of SRPs for the MAP estimate, are described next.

3.2.1. Statistically Equivalent Blocks PQMC or SEB-PQMC. 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] (except possibly in blocks on the boundary of the partitioned space). A statistically equivalent blocks (SEB)-based SRP partitioning scheme specified by the PQMC with (i) priority function $\psi(\rho\mathsf{v}) = \#\boldsymbol{x}_{\rho\mathsf{v}}$, i.e., the number of sample points associated with a node $\rho\mathsf{v}$, (ii) ψ-related stopping condition $\overline{\psi} = \overline{\#}$ and (iii) $\overline{m}$, is denoted by SEB-PQMC. SEB-PQMC, as motivated in Section 3.1, is our solution for the problem of finding states with high posterior density more efficiently than with the empirical variant of the yang chain on SRPs.

Thus, at stopping time T, the SRP s realized by the SEB-PQMC will be such that either $\mathbb{L}^{\triangledown}(s) = \emptyset$ or $|\mathbb{L}(s)| \leq \overline{m}$ or $\#\boldsymbol{x}_{\rho\mathsf{v}} \leq \overline{\#}\ \forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s)$. The operation may only be considered to be successful if $|\mathbb{L}(s)| \leq \overline{m}$ and $\#\boldsymbol{x}_{\rho\mathsf{v}} \leq \overline{\#}\ \forall \rho\mathsf{v} \in \mathbb{L}^{\triangledown}(s)$. Care must be taken to ensure that the operation is successful.

Therefore, an SEB-PQMC can be used to create a final SRP at stopping time T such that each leaf node has at most $\overline{\#}$ of the sample data points associated with it and the total number of leaves is at most $\overline{m}$. As we will see in Section 3.3, the L_1 consistency of SEB-PQMC requires that $\overline{m}$ must grow sublinearly (i.e. $\overline{m}/n \to 0$ as $n \to \infty$) while the volume of leaf boxes shrink such that a combinatorial complexity measure of the partitions in the support of the SEB-PQMC grows sub-exponentially.

Intuitively, SEB-PQMC prioritises the splitting of leaf nodes with the largest numbers of data points associated with them. Figure 6 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 (left) while the other histogram with a larger $\overline{\#}$ used as the SEB stopping criterion is over-smoothed (right). We will use the leave-one-out cross-validation to choose the optimally smoothed MAP estimate in the sequel.

Unfortunately, under an SEB-PQMC partitioning strategy over SRPs, the nodes with least data associated with them will remain unsplit for longer (and will possibly never be split). This tends to result in relatively large regions of very low density in the tails of the density estimate f_n formed from the SRP. Figure 7 shows two partitions of an SRP associated with a highly correlated dataset super-imposed on it. As the number of leaves in the partition increases from 20 (Figure 7(a)) to 40 (Figure 7(b)), large sub-boxes containing very little sample data remain unsplit. The effect can be es-

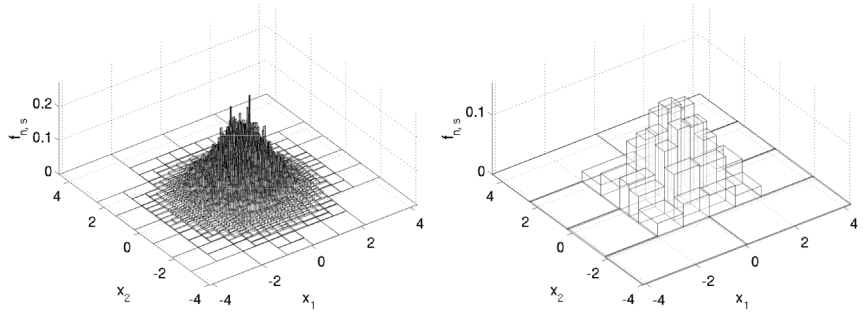
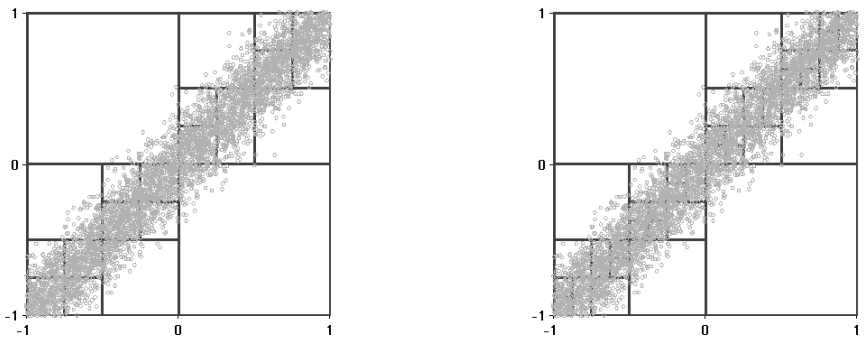


Fig. 6. Two histogram density estimates for the standard bivariate Gaussian density. The left figure shows a histogram with 1485 leaf nodes where $\overline{\#} = 50$ and the histogram on the right has $\overline{\#} = 1500$ resulting in 104 leaf nodes.

pecially distorting to the resulting histogram when the axis-aligned root box required by the SRP is a poor fit to the data (when large areas of the root box contain no data points). A general transformation of the data, beyond translations and coordinate-wise rescaling, to mitigate the problem may not be desirable because such a transformation could preclude the subsequent creation of marginal or conditional densities on the coordinates of the untransformed data [Harlow 2013] and thereby prevent us from exploiting the underpinning algebra of planar binary trees and the ensuing tree arithmetic as described in Section 2.2. Thus, we would like to carve out empty space within the root box while remaining in $\mathbb{S}_{0:\infty}$. The next PQMC ameliorates this undesirable feature of SEB-PQMC and can be used collaboratively with SEB-PQMC to still ensure asymptotic L_1 consistency of the joint partitioning strategy.



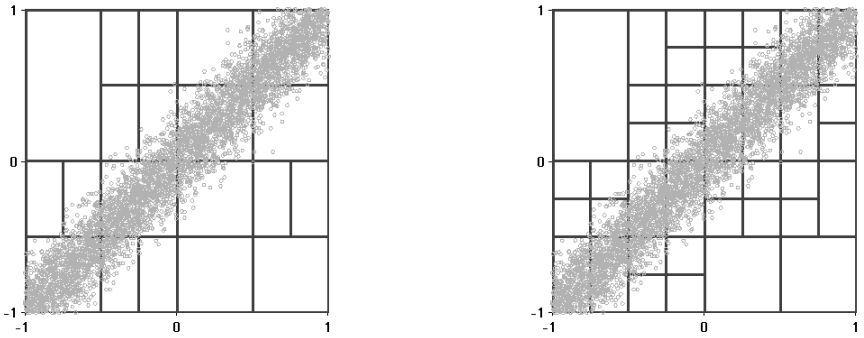
(a) 20 leaves. (b) 40 leaves.

Fig. 7. Partition using an SEB-PQMC.

3.2.2. Support Carving PQMC or SPC-PQMC. A support-carving (SPC) SRP partitioning scheme specified by the PQMC with (i) $\psi(\rho\mathsf{v}) = \Xi(\rho\mathsf{v}) = (1 - \#\boldsymbol{x}_{\rho\mathsf{v}}/n)\mathrm{vol}\,(\rho\mathsf{v})$, the SPC priority function which prioritizes node $\rho\mathsf{v}$ according to the relative lack of its empirical measure when further scaled by its volume, (ii) ψ-related stopping condition $\overline{\Xi}$ and (iii) maximum number of leaves $\overline{m}^{\Xi}$, is denoted by SPC-PQMC.

Using SPC-PQMC to carve out "empty space" in the complement of the empirical support can be thought of as an 'inversion' of the SEB-PQMC, as motivated in Section 3.1 to speed up search relative to the empirical variant of the yin chain on SRPs: instead of prioritising splitting of nodes with the largest number of data points associated with them as done by SEB-PQMC, the SPC-PQMC prioritises splitting of non-empty leaf nodes with large boxes but few data points and thus is likely to result in one of the child nodes being devoid of any sample data. Hence, the two PQMCs are thought to be complementary.

The carving effect is illustrated in Figure 8. The partitions are created for the same set of correlated data shown in Figure 7 using SPC-PQMC. Figures 8(a) and 8(b) show the partition when the SRP has 20 and 40 leaves, respectively. Partitioning is concentrated in the regions of sparse sample data and the effect is to reduce the size of the sub-boxes of the partition into which these sparse data points fall, in effect more tightly enclosing the support of the data as desired while still remaining in $\mathbb{S}_{0:\infty}$.



(a) 20 leaves. (b) 40 leaves.

Fig. 8. Partition using an SPC-PQMC.

Algorithm 1 with the SPC priority function Ξ can be used to obtain a *core support-carved path* in $\mathbb{S}_{0:\overline{m}^{\Xi}}$ by initializing from the root SRP $s \in \mathbb{S}_0$ through the procedure $\texttt{PQMC}(s, \Xi, \overline{\Xi}, \overline{m}^{\Xi})$ with $\overline{\Xi} = 0.0$. Thus, the partitioning process of this core SPC-PQMC only stops when the SRP has $\overline{m}^{\Xi}$ leaves or aborts if there a no splittable nodes. More general support-carved paths can be generated by specifying $\overline{\Xi} > 0.0$ or imposing constraints on the minimum volume of splittable nodes if deemed appropriate for the underlying data generation process.

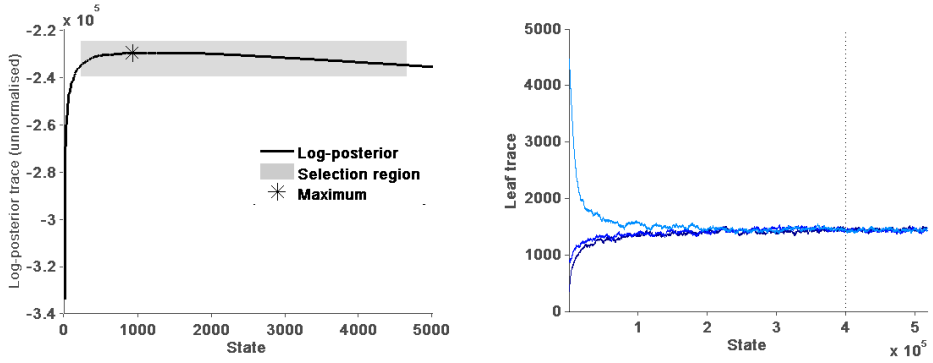
3.2.3. Joint exploration. An SPC-PQMC of Section 3.2.2 alone will not give an effective data-driven partitioning strategy, but used in conjunction with an SEB-PQMC it can improve the SRP histogram. An initial SPC-PQMC can be run for a short time (specifying $\overline{\Xi} = 0.0$ and a relatively low value of $\overline{m}^{\Xi}$, say a small fraction of the total numer of leaves $\overline{m}$), followed by an SEB-PQMC. The empty elements of the partition 'carved out' will be ignored by the SEB-PQMC, under which partitioning will be concentrated on the areas where most of the sample data has fallen.

The core support-carved path from SPC-PQMC over $\mathbb{S}_{0:\overline{m}^{\Xi}}$ through $\texttt{PQMC}(s, \Xi, \overline{\Xi}, \overline{m}^{\Xi})$ with $\overline{\Xi} = 0.0$ will be used to determine c^{Ξ} many spread-out initial conditions for launching multiple independent SEB-PQMCs. It is along these c^{Ξ} many joint SPC/SEB-PQMC paths, which may be viewed as a system of c^{Ξ} many SEB-PQMC tributaries spread along the core support-carved path of the SPC-PQMC, that we conduct MAP

estimation for a given τ-specific prior by simply identifying the SRP state s with the highest log-posterior value. The search for the MAP estimate is conducted sequentially along each of the SPC/SEB-PQMC paths, i.e., along a random sequence of states in $\mathbb{S}_{0:\overline{m}}$ with each state having one additional leaf than its preceding state. This sequential search for the MAP estimate along each joint SPC/SEB-PQMC path is relatively straightforward due to the recursive updates in (5) and the unimodal nature of the log-posterior: thanks to the counteracting forces of the spiking likelihood and the penalizing prior.

Finally, to gaurd against over-carving with too large a value of $\overline{m}^{\Xi}$ relative to $\overline{m}$, the c^{Ξ} many states are dispersed over the core support-carving path from the root node. The only requirement is the following *over-carving constraint*, whereby, $\overline{m}^{\Xi}$, the maximum number of leaves allowed in the SPC-PQMC is smaller than $\overline{m}$, the maximum number of leaves allowed in the subsequent SEB-PQMC (usually much smaller than $\overline{m}$ to ensure that the SEB-PQMC in the second phase is able to reach $\overline{m}$ leaves).

The c^{Ξ} many initial conditions from the core support-carved path for launching the SEB-PQMCs, should be well spread out in order to facilitate a better search strategy over $\mathbb{S}_{0:\overline{m}}$ for the MAP estimate, and in particular, contain the SEB-PQMC launched from the root node as one of its joint SPC/SEB-PQMC paths. In essence, we can view the joint exploration as one that will necessarily improve upon the MAP estimate found by the SEB-PQMC initialized from the root node (which we prove to be asymptotically L_1-consistent in the next Section). However, this is not a rigorous or exhaustive global optimization for the MAP estimate over every state in $\mathbb{S}_{0:\overline{m}}$ — an impractical task over the exponentially large state space. The implementational details of this joint exploratory search for optimal MAP estimation can be found in `examples/StatsSubPav/CVOptMAP` [Sainudiin et al. 2016]. See Harlow [2013, App. F] for detailed descriptions of a more general form of this joint explorative process that finds a collection of high posterior states about the joint SPC/SEB-PQMC paths for initializing multiple independent MCMC chains to improve and automate the posterior mean histogram estimate of Sainudiin et al. [2013].



(a) histogram states in the selection region. (b) Leaf trace for three chains.

Fig. 9. Selecting multiple initial states with extended selection region.

Figure 9(a) shows the log-posterior as a function of the number of leaves in each of the SRP states in the selection region — the sequence of SEB-PQMC tributary paths along the core support-carving path of SPC-PQMC resulting from the joint exploration described above. Figure 9(b) shows the leaf traces from three Markov chains started at the first state, maximum log-posterior or MAP state, and last state in the sequence. It

takes longer for the more widely dispersed chains to converge but the leaf traces eventually settle down as in Figure 9(b). The time to convergence of the MCMC method to obtain posterior mean estimate when initilized from multiple high posterior states in the selection region can be impractical when $d > 6$ and $n > 10^6$ and the sensity is concentrated or structured [Sainudiin et al. 2013]. However, MAP estimation which amounts to finding the state with the highest log-posterior in the selection region is more feasible for large d and n, when compared to hours or days for a fully automated MCMC with convergence diagnostics from multiple chains [Harlow 2013, Ch. 6]. We will focus here on MAP estimation with optimal prior selection to obtain optimal MAP estimates (or optimally regularized ML estimates) that have nearly the same L_1 error as a well-smoothed but time-consuming MCMC posterior mean estimate with significantly more leaves due to the averaging over SRP histogram states.

3.3. L_1-consistency

We now show that an RMRP density estimate based on an SRP created using the SEB-PQMC 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 are shrinking.

Let $\{S_n(i)\}_{i=0}^{\dot{I}}$ on $\mathbb{S}_{0:\infty}$ be the Markov chain formed using SEB-PQMC. 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 $\{X_1, \ldots, X_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\Pi\left(\mathcal{L}_n, \{X_1, \ldots, X_n\}\right)$ be the number of distinct partitions of the finite set $\{X_1, \ldots, X_n\}$ that are induced by partitions $\mathbb{L}(\dot{s}) \in \mathcal{L}_n$:

$$\Pi(\mathcal{L}_n, \{X_1, \ldots, X_n\}) := \left| \{ \{ \boldsymbol{x}_{\rho v} \cap \{X_1, \ldots, X_n\} : \boldsymbol{x}_{\rho v} \in \mathbb{L}(\dot{s}) \} \, : \mathbb{L}(\dot{s}) \in \mathcal{L}_n \} \right| \ .$$

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

$$\Pi^*(\mathcal{L}_n, \{X_1, \ldots, X_n\}) = \max_{\{X_1, \ldots, X_n\} \in (\mathbb{R}^d)^n} \Pi(\mathcal{L}_n, \{X_1, \ldots, X_n\}) \ .$$

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 $\boldsymbol{x} = [\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_d, \overline{x}_d]$, $\mathrm{diam}(\boldsymbol{x}) = \sqrt{\sum_{i=1}^d (\overline{x}_i - \underline{x}_i)^2}$.

We now check the three conditions for L_1-consistency of the histogram estimate constructed using SEB-PQMC.

THEOREM 3.1 (L_1-CONSISTENCY). *Let $X_1, X_2, \ldots$ 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 SEB-PQMC 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) $\mu(x : \text{diam}(\boldsymbol{x}(x)) > \gamma) \to 0$ 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 approximated as ([Mattarei 2010]):

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

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:

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

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

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

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 $\sqrt{d(\frac{2M}{2^i})^2}$, we can have at most 2^{di} boxes in the interior of $\hat{\boldsymbol{x}}$ and δ boxes at the lower dimensional boundaries of $\hat{\boldsymbol{x}}$, i.e., there are at most m_γ disjoint boxes in $\hat{\boldsymbol{x}}$ that have diameter greater than γ, where

$$m_\gamma < 2^{di} + \delta, \quad \delta = \left(2^d + \sum_{j=1}^{d-1} 2^{d-j}\binom{d}{j}2^{ij}\right) \ . \tag{11}$$

By choosing i large enough we can upper bound m_γ by $(2M\sqrt{d}/\gamma)^d + 2^d + \sum_{j=1}^{d-1} 2^{d-j}\binom{d}{j}(2M\sqrt{d}/\gamma)^j$, a quantity that is independent of n, such that

$$\begin{aligned}
\mu(x : \mathrm{diam}(\boldsymbol{x}(x)) > \gamma) &\leq \xi + \mu\left(\{x : \mathrm{diam}(\boldsymbol{x}(x)) > \gamma\} \cap \hat{\boldsymbol{x}}\right) \\
&\leq \xi + m_\gamma\left(\max_{\boldsymbol{x}\in\mathbb{L}(\dot{s})} \mu(\boldsymbol{x})\right) \\
&\leq \xi + m_\gamma\left(\max_{\boldsymbol{x}\in\mathbb{L}(\dot{s})} \mu_n(\boldsymbol{x}) + \max_{\boldsymbol{x}\in\mathbb{L}(\dot{s})} |\mu(\boldsymbol{x}) - \mu_n(\boldsymbol{x})|\right) \\
&\leq \xi + m_\gamma\left(\frac{\overline{\#}}{n} + \sup_{\boldsymbol{x}\in\mathbb{R}^d} |\mu(\boldsymbol{x}) - \mu_n(\boldsymbol{x})|\right).
\end{aligned}$$

The first term in the parenthesis converges to zero since $\overline{\#}/n \to 0$ by assumption. For $\epsilon > 0$, the second term goes to zero by applying the Vapnik-Chervonenkis theorem to boxes in $\mathbb{R}^d$ with shatter coefficient $s(\mathbb{R}^d, n) = 2^{2d}$ [Devroye et al. 1996, p. 220], i.e.,

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

By the Borel-Cantelli lemma,

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

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

$$\limsup_{n\to\infty} \mu(\{x : \mathrm{diam}(\boldsymbol{x}(x)) > \gamma\}) \leq \xi.$$

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

Remark 3.2. To see that the joint partitioning strategy from the SPC/SEB-PQMC of Section 3.2.3 is also asymptotically L_1 consistent first note that the SEB-PQMC initialized at the root node is one of the $c^{\overline{\Xi}}$ many paths explored and that all these paths being SEB-PQMCs in their second phase terminate at a state with no more than $\overline{m}$ many leaves with at most $\overline{\#}$ many points in each one due to the over-carving constraint. And, finally since the union of these paths are still a subset of $\mathbb{S}_{0:\overline{m}}$, the combinatorial complexity bound in the theorem still holds and was not too much of an upper bound after all.

3.4. Minimum distance estimation using SRP

The minimum distance estimate (MDE) minimizes the distance to the empirical measure in a metric that is reminiscent of the total variation distance. Unlike Bayes or MAP estimates that are based on the likelihood function, MDEs come with universal performance guarantees, i.e., the unknown f is allowed to remain in L_1. 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 SEB-PQMC. For a fixed sample size n, let $\{s_\theta : \theta \in \Theta\}$ be an ordered subset of states visited by the PQMC, with $s_\theta \prec s_\vartheta$ if s_ϑ is a refinement of s_θ, i.e. if s_θ is visited by the PQMC 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 $\mathcal{T} := \{x_1, \ldots, x_{n-\varphi n}\}$ and the set of validation data by $\mathcal{V} := \{x_{n-\varphi n+1}, \ldots, x_n\} = \{y_1, \ldots, y_{\varphi n}\}$. For an ordered pair $(\theta, \vartheta) \in \Theta^2$, with $\theta \neq \vartheta$, the set:

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

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

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

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

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-\varphi n,\theta}(A) - \mu_{\varphi n}(A) \right| \ . \tag{12}$$

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 $\mathcal{A}_\Theta$.

3.4.1. Recursively Computable Statistics for Validation Data. The SRP is adapted for MDE to mutably cache recursively computable statistics such as counts, mean, etc. for traning and validation data separately. Thus, 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 priority queued Markov chain $\texttt{PQMC}(s_0, \psi, \overline{\psi}, \overline{m})$ to produce a sequence of SRP states: $s_{\theta_1}, s_{\theta_2}, \ldots$ 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 is 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 Figure 10. Computing the MDE objective Δ_{θ_i} in (12) 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 $\check{\#}\boldsymbol{x}_{\rho v}$.

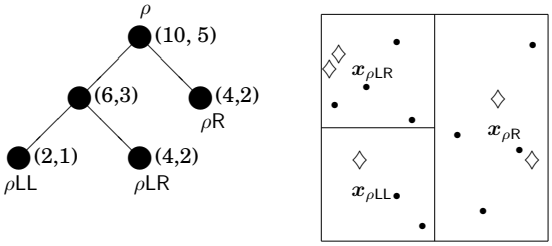
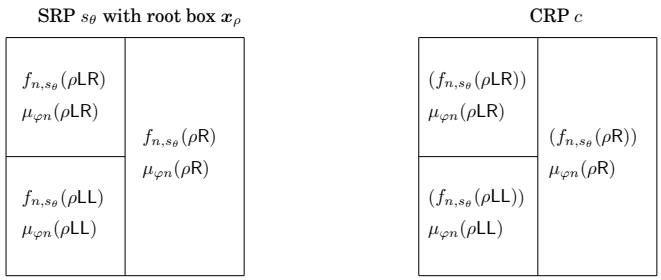


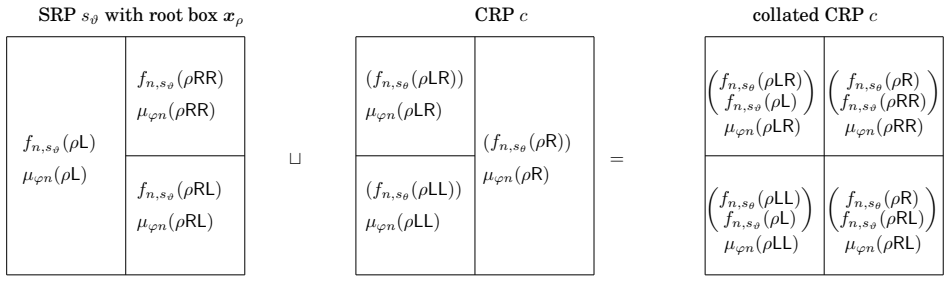
Fig. 10. An SRP s with training ($\bullet$) and validation data ($\diamond$) and their respective sample counts $(\#\boldsymbol{x}_{\rho v}, \check{\#}\boldsymbol{x}_{\rho v})$ that are updated recursively as data falls through the nodes of s.

3.4.2. Regular Paving as a Collator for MDE. 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 PQMC. Using nodes imbued with recursively computable statistics for both training and validation data, and a specialized collation according to Algorithm 2 over SRPs, we compute the objective Δ_θ in (12) using Algorithm 4 via a dynamically grown Yatracos Matrix with pointers to all Scheffé sets constituting the Yatracos class according to Algorithm 3. We briefly outline these core ideas next.

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, 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 11 shows how CRP c can collate two SRPs s_θ and s_ϑ using Algorithm 2.



(a) Make the SRP s_θ into a CRP c.



(b) Collate another SRP s_ϑ onto CRP c.

Fig. 11. Collating two SRPs s_θ and s_ϑ with the same root box $\boldsymbol{x}_\rho$.

3.4.3. Error Bounds for MDE. 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 3.3. *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 Δ_θ in* (12)*, we have*

$$\int |f_{n-\varphi n,\theta^*} - f| \leq 3 \min_{\theta\in\Theta} \int |f_{n-\varphi n,\theta} - f| + 4\Delta \tag{13}$$

where

$$\Delta = \max_{A\in\mathcal{A}_\Theta} \left| \int_A f - \mu_{\varphi n}(A) \right| \ . \tag{14}$$

Theorem 3.3 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 by using the cardinality bound [Devroye et al. 1996, Theorem 13.6, p. 219] for the shatter coefficient of $\mathcal{A}_\Theta$. Given $\{x_1, \ldots, x_n\}$ the n-th shatter coefficient of $\mathcal{A}_\Theta$ is defined as

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

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` of Algorithm 3, by Theorem 13.6 of Devroye et al. [1996, p. 219] its n-th shatter coefficient is bounded as follows:

$$S(\mathcal{A}_\Theta, n) \leq |\mathcal{A}_\Theta| \leq (|\Theta| + 1)^2 - (|\Theta| + 1) = |\Theta|(|\Theta| + 1) \ . \tag{15}$$

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 3.4. *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, φn, Θ and $f \in L_1$:*

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

PROOF. By Theorem 3.3,

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

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

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

Finally by Theorem 3.1 in [Devroye and Lugosi 2001, p. 18] and (15),

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

□

In order 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 SEB-PQMC 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 SEB-PQMC 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 preceeding states. Then we simply zoom-in around the states with the lowest Δ_θ values and add another k states along the same SEB-PQMC 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 3.1, we know that the SEB-PQMC is asymptotically consistent. Thus, the adaptive search set Θ that is selected iteratively from the set of histogram states along the path of the SEB-PQMC 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].

4. PERFORMANCE EVALUATION

We are mainly interested in cases where the sample size is two orders of magnitude larger than the feasible sample size of $n_k = 2000$ that can be handled in a reasonable amount of time (say, $\leq 10^4$ seconds) by a KDE Method with smoothing [Zhang et al. 2006] in up to five dimensions for a multivariate Gaussian mixture (Density I in Appendix 7). The MCMC method with the fixed Catalan prior has lower estimated L_1 error for larger sample sizes, but it is prohibitively slow due to the stringent convergence diagnostic procedures under the slowly mixing *stay-split-merge* base chain.

The optimal MAP estimate (OPTMAP) is comparable in mean L_1 error estimate with the posterior mean estimate from the MCMC method, but the OPTMAP is one to two orders of magnitude faster.

Table I shows a summary of results for $\hat{L}_1$ with true density Density I (see Appendix 7) for $d = 2, 3, 4$ and 5. The estimates of the L_1 error (from 10^7 quasi-random samples) are shown for the KDE (using $n_K = 2,000$ sample points from the true density) and also for an $\mathbb{R}$-MRP posterior mean estimate with the fixed natural Catalan prior, specified by setting $a_k = 1/C_k$ and $a = 2 + 4\pi/3^{5/2}$ in (6). The $\mathbb{R}$-MRP posterior mean estimate was formed by averaging 100 SRP histogram samples taken after burn-in from a Markov chain (thin-out 100), using the refined and automated extension [Harlow 2013, Ch. 6] of the method of Sainudiin et al. [2013]. The effect of increasing the sample size n for each dimension d on the estimated error between the three optimally smoothed density estimates and the true Density A, in terms of (i) the total variation distance given by $\hat{L}_1/2$ and (ii) its upper bound given by the square-root of $\hat{d}_{\mathrm{KL}}/2$ (due to Pinsker's inequality), is shown in Table I. The values shown for $\hat{L}_1$ for the averaged posterior mean histogram estimate in Table I are averages over 10 replications of the process with different sample data and different pseudo-random number sequences used for each replication, except for the results of the posterior mean estimate with $d = 5$ and $n = 100,000$, where time constraints meant that only three replications could be completed. The variability between replications is low (for example, for $d = 2$ and $n = 10,000$, minimum and maximum $\hat{L}_1$ are 0.217 and 0.228, respectively). There is wider variability in the time taken for each replication and the number of leaves in the final averaged $\mathbb{R}$-MRP esimate of the posterior mean and so the minimum and maximum over the replications are also shown. Note how the timings depend much more on the sample size than on the dimension and scale for this multivariate mixture (Density I in Appendix 7). As discussed in Section 1.1, on the basis of the optimal convergence rate of KDEs by Stone [1980], when $n = 2000$ the KDE method becomes computationally feasible and gives smaller L_1 errors when compared to the OPTMAP method limited to histogram partitions for the smooth Density I. However, this optimal convergence rate for a density f with p bounded derivatives in d dimensions is of the order $n^{p/(2p+d)}$ — which can be very slow in high dimensions especially when the sample sizes are large and the unknown f is not sufficiently smooth. Observe that the OPTMAP method can hande sample sizes that are two orders of magnitude larger that $n_K = 2000$ and produce histogram estimates with smaller L_1 errors in time that is about one order of magnitude faster in up to five dimensions. More crucially, unlike the KDE methods that are computationally cursed by the sample size, the data-adaptive SRP histogram methods such as the minimum distance estimate (MDE) will give universal performance guarantees for less smooth or more wildly varying densities in L_1 by taking advantage of the large sample sizes in a computationally efficient manner.

The approximate integration methods based on quasi-random streams and importance sampling used to obtain $\hat{L}_1$ and $\hat{d}_{\mathrm{KL}}$ in Table I became unreliable and significantly slower for highly structured densities such as the Rosenbrock (Sec. 7.3) in dimensions as large as 5. Thus, we used $\mathbb{R}$-MRP approximation of the true density that is within 0.01 in Hellinger distance of the true density (see [Sainudiin et al. 2013, Sec. 4.2]). By producing n samples from such piecewise constant $\mathbb{R}$-MRP densities, we can take advantage of $\mathbb{R}$-MRP arithmetic to obtain the exact L_1 error in Table II between the approximated $\mathbb{R}$-MRP density and its $\mathbb{R}$-MRP estimate produced by the MCMC or MDE methods.

The Posterior mean histograms were obtained from 10,000 samples collected by the MCMC method under the Catalan prior [Sainudiin et al. 2013, Table II with $\Lambda = 10^6$] along with their standard errors, and are shown in Table II. The standard deviations

Table I. Estimated errors for KDE, posterior mean and optimal MAP $\mathbb{R}$-MRP histograms.

	estimated errors		Time (s)		Leaves	
	$\frac{1}{2}\hat{L}_1$	$\sqrt{\frac{1}{2}\hat{d}_{\mathrm{KL}}}$	min.	max.	min.	max.
2-d						
KDE ($n_K = 2,000$)	0.10	0.14	5,000	7,200	n/a	
$\mathbb{R}$-MRP posterior mean with natural Catalan prior						
$n = 10,000$	0.11	0.17	2	13	811	902
$n = 50,000$	0.08	0.12	15	2,168	1,546	1,719
$\mathbb{R}$-MRP MAP with optimal τ-prior or OPTMAP						
$n = 2,000$	0.17	0.41	5	6	178	483
$n = 10,000$	0.11	0.27	23	26	552	857
$n = 50,000$	0.08	0.22	400	661	2,000	2,075
$n = 100,000$	0.06	0.16	1,429	2,590	2,055	2,756
3-d						
KDE ($n_K = 2,000$)	0.18	0.25	5,600	7,200	n/a	
$\mathbb{R}$-MRP posterior mean with natural Catalan prior						
$n = 10,000$	0.21	0.35	21	451	1,573	1,718
$n = 50,000$	0.15	0.24	295	27,832	3,507	3,783
$\mathbb{R}$-MRP MAP with optimal τ-prior or OPTMAP						
$n = 2,000$	0.30	0.58	6	7	260	833
$n = 10,000$	0.22	0.50	30	35	743	4,050
$n = 50,000$	0.16	0.33	341	445	2,338	6,394
$n = 100,000$	0.14	0.30	1,407	1,827	3,818	7,740
4-d						
KDE ($n_K = 2,000$)	0.26	0.35	7,200	8,050	n/a	
$\mathbb{R}$-MRP posterior mean with natural Catalan prior						
$n = 50,000$	0.24	0.40	2,524	53,190	6,241	6,570
$n = 100,000$	0.21	0.35	10,382	82,684	9,431	9,775
$\mathbb{R}$-MRP MAP with optimal τ-prior or OPTMAP						
$n = 2,000$	0.42	0.78	6	8	337	1,252
$n = 10,000$	0.33	0.53	34	36	1,218	1,548
$n = 50,000$	0.26	0.40	420	453	3,682	4,703
$n = 100,000$	0.23	0.37	1,555	1,667	5,068	10,374
5-d						
KDE ($n_K = 2,000$)	0.33	0.45	7,350	8,880	n/a	
$\mathbb{R}$-MRP posterior mean with natural Catalan prior						
$n = 50,000$	0.34	0.57	28,841	277,071	9,342	9,803
$n = 100,000$	0.30	0.51	24,244	399,016	15,160	15,563
$\mathbb{R}$-MRP MAP with optimal τ-prior or OPTMAP						
$n = 2,000$	0.53	1.02	8	9	514	1,329
$n = 10,000$	0.44	0.78	35	41	1,827	5,954
$n = 50,000$	0.35	0.56	492	511	5,631	8,120
$n = 100,000$	0.31	0.50	1,453	1,660	9,289	12,847

Table II. 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 $L_1(f_{n,\theta^*}, f)$, $L_1(f_{n,\theta^*}, f) - \min_{\theta\in\Theta} L_1(f_{n,\theta}, f)$				
10^4	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)

about the MIAEs for the MDE method, i.e., $L_1(f_{n,\theta^*}, f)$ in Table II, 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 II for each density and sample size. Note how the L_1 errors decrease with the sample size and how the errors are comparable between the methods, albeit the MDE method is at least an order of magnitude faster than the MCMC method.

5. APPLICATIONS

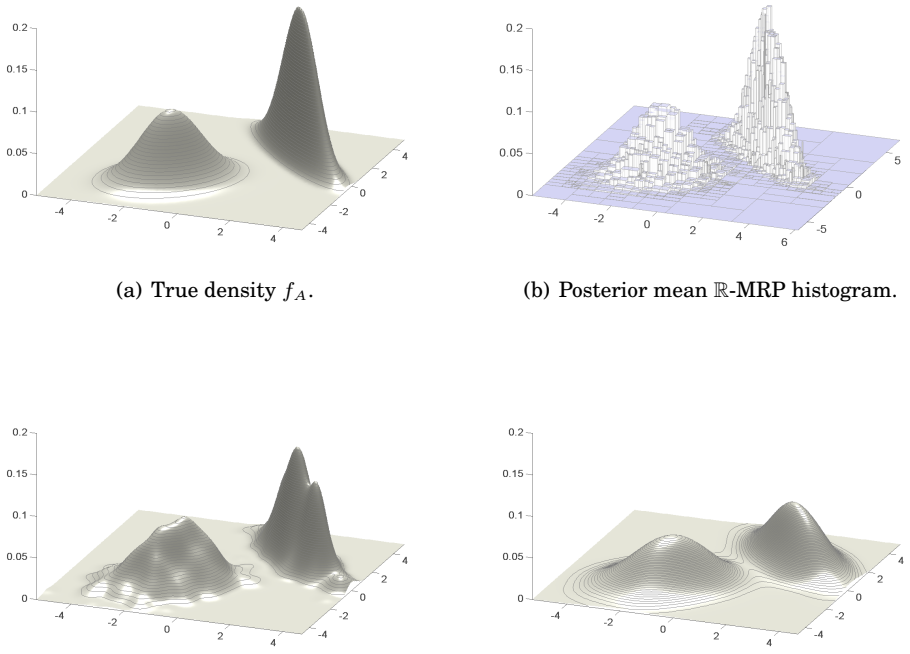
5.1. Statistical Operations using $\mathbb{R}$-MRPs

Using Density A studied by Zhang et al. [2006], the 2-dimensional version of Density I in Appendix 7, we illustrate the most useful tree arithmetic operations with $\mathbb{R}$-MRPs for statistical purposes in the next three Sections.

5.1.1. Comparing some estimates of density A. Figure 12 shows estimates of Density A using $n_{\kappa} = 1,000$ sample points simulated from the true density. Figure 12(a) is a visualisation of the true density. Figure 12(c) is a visualisation of the KDE created using the optimal diagonal bandwidth matrix as described in [Zhang et al. 2006] (with the same burn-in of 5,000 iterations and the same 250,000 recorded iterations as were used in that study). Figure 12(d) is a visualisation of the KDE created using the 'Normal reference rule'. Figure 12 illustrates the smoothness of the KDE method over a histogram-based method but also demonstrates the oversmoothing that can occur if the KDE bandwidth matrix is not suitable, such as bandwidths chosen using the Normal reference rule used with non-Normal data (Figure 12(d)). Figure 12(b) is the $\mathbb{R}$-MRP with 1551 leaf boxes formed using $n = 50,000$ data points drawn from Density A and averaging 100 SRP histogram samples taken after burn-in from a Markov chain (with thin-out 100), using the method described in Chapter [Harlow 2013, Ch. 6], a refinement of the MCMC method that produces the posterior mean in [Sainudiin et al. 2013] by taking advantage of the addition operations over $\mathbb{R}$-MRP histograms visited by a Markov chain whose stationary distribution is the posterior distribution over $\mathbb{S}_{0:\infty}$ with a fixed prior (see Section 2.5.2).

5.1.2. Uniformly approximating a KDE *by an $\mathbb{R}$-MRP.* Recall that the KDE of Figure 12(c) is a procedure that takes as input an $x \in \mathbb{R}^2$ and returns the density estimate at x, typically by evaluating the optimally smoothed kernels at all n_{κ} sample points used to obtain the KDE itself. This procedure can be made more efficient by using tree-based structures as in [Gray and Moore 2003b; 2003a], but is not conducive to arithmetic operations for subsequent statistical purposes. An interesting application of $\mathbb{R}$-MRPs, especially for small sample sizes where KDE methods are more optimal, is to uniformly approximate a KDE density estimate by an $\mathbb{R}$-MRP and use this arithmetically amenable structure for various statistical computations of interest. We illustrate this next. Figure 13 shows the results of using `RPQApproximate` [Harlow 2013, Algorithm 7.1], an adaptation of `RPQEnclose`$^{\triangledown}$ [Harlow et al. 2012, Algorithm 6], in order to uniformly approximate the KDE $\hat{f}_{\kappa}$ shown in Figure 12(c) for various values of $\overline{\psi}$, which specifies the uniform error bound between the approximating $\mathbb{R}$-MRP ${}^{\square}f$ and the KDE $\hat{f}_{\kappa}$. Note that the approximation error $\overline{\psi}$ is inversely proportional to the number of leaf boxes in ${}^{\square}f$ and the time taken to make these approximations increases as $\overline{\psi}$ decreases but is of the order of seconds (compared with minutes, or hours, or days, depending on the sample size n_{κ}, for the KDE itself) [Harlow 2013, Table 7.1].

5.1.3. Point-wise, Marginal, Conditional and Coverage Operations as $\mathbb{R}$-MRP tree arithmetic. Figure 14 shows some of the various operations that can then be carried out on the $\mathbb{R}$-MRP approximation of the KDE or an $\mathbb{R}$-MRP histogram produced directly as the posterior mean in [Sainudiin et al. 2013; Harlow 2013] or as the optimally smoothed *maximum a*



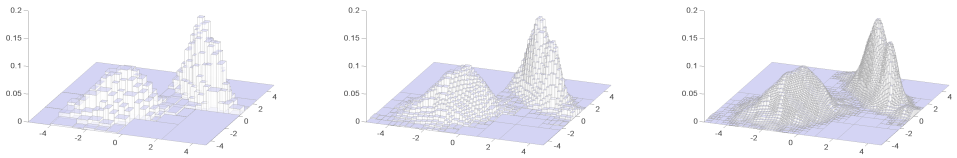
(a) True density f_A.

(b) Posterior mean $\mathbb{R}$-MRP histogram.

(c) MCMC bandwidth KDE $\hat{f}_K$.

(d) Normal reference rule KDE.

Fig. 12. Estimating Density A using three methods.



(a) $\overline{\psi} = 0.001$ (187 leaves).

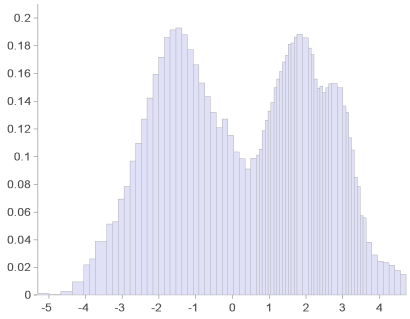
(b) $\overline{\psi} = 0.0001$ (919 leaves).

(c) $\overline{\psi} = 0.00001$ (4420 leaves).

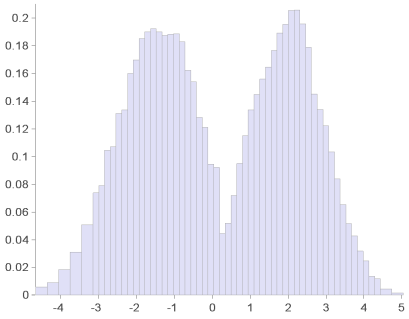
Fig. 13. Approximating the KDE $\hat{f}_K$ using $\mathbb{R}$-MRP ${}^{\square}f$ with 187, 919 and 4420 leaves, respectively.

posteriori (OPTMAP) estimate or as the *minimum distance estimate* (MDE) developed here. The $\mathbb{R}$-MRP used here to illustrate tree-based operations that output marginal and conditional densities as well as coverage regions is that shown in Figure 13(a) with uniform error bound $\overline{\psi} = 0.001$ and 187 leaf boxes.

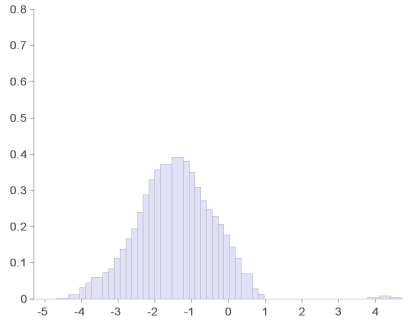
Obtaining point-wise image of n_q query points from an $\mathbb{R}$-MRP representation of the density can be carried out by `PointWiseImage` [Harlow et al. 2012, Algorithm 2] by simply dropping the n_q points in the RP tree until each point reaches a leaf node by recursively descending from the root through inequality checks and looking-up the density at the leaf boxes containing the n_q points. Contrast this with the $O(n \times n_t \boldsymbol{n}_s)$



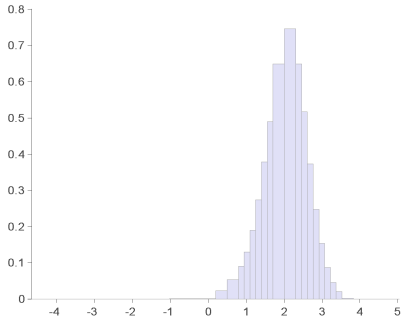
(a) Marginal on coordinate 1, ${}^{\square}f^{\{1\}}(x_1)$.



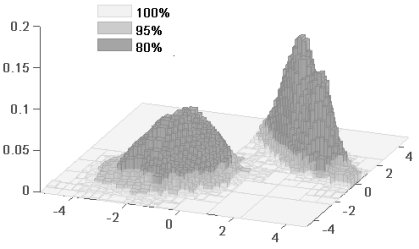
(b) Marginal on coordinate 2, ${}^{\square}f^{\{2\}}(x_2)$.



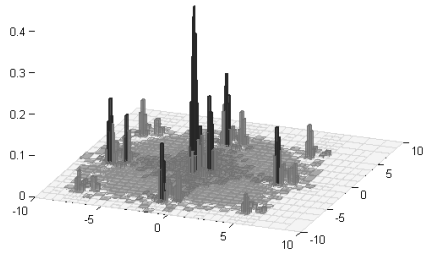
(c) Conditional density ${}^{\square}f(x_1 \mid x_2 = -1.5)$.



(d) Conditional density ${}^{\square}f(x_2 \mid x_1 = 2.0)$.



(e) Coverage regions of ${}^{\square}f(x_1, x_2)$.



(f) Coverage regions of ${}^{\square}f(x_1, x_2) \approxeq \dot{l}$ of Levy.

Fig. 14. Density operations using RMRP approximations to Density I (App. 7) and the highly multimodal Levy Density App. 7.2.

operations needed to obtain the smoothed KDE as a procedure with $O(n \times n_q)$ query cost.

Obtaining marginal densities over a subset of coordinates of a joint density is a fundamental operation with multivariate densities in statistical tasks. Figures 14(a) and 14(b) show the marginalised approximations on coordinates 1 (${}^{\square}f^{\{1\}}(x_1)$) and 2 (${}^{\square}f^{\{2\}}(x_2)$), respectively, through `Marginalise` [Harlow et al. 2012, Algorithm 10]. Figure 14(c) shows ${}^{\square}f^{|x_2=-1.5}(x_1)$, the normalised slice on $x_2 = -1.5$, an estimate of

$f_A(x_1 \mid x_2 = -1.5)$ the univariate conditional density of x_1 given $x_2 = -1.5$. Figure 14(d) shows ${}^{\square}f^{|x_1=2.0}(x_2)$, the normalised slice on $x_1 = 2.0$, an estimate of $f_A(x_2 \mid x_1 = 2.0)$ the univariate conditional density of x_2 given $x_1 = 2.0$. These conditional densities are obtained through `Slice` [Harlow et al. 2012, Algorithm 11] and can be used along with `Marginalise` for nonparametric regressions, i.e., predicting the entire distribution of one or more of the variates by conditioning on others in the joint nonparametric density estimate of all the variates available as an $\mathbb{R}$-MRP ${}^{\square}f$. Figure 14(e) shows the 95% (dark-gray) and 80%(mid-gray) coverage regions of the $\mathbb{R}$-MRP against the 100% coverage (light-gray). Again, the coverage region computations are done efficiently using the $\mathbb{R}$-MRP tree structures coupled with a sort according to `CoverageRegion` [Harlow et al. 2012, Algorithm 9]. Such coverage regions can be used for "bump-hunting" and data exploration. In Figure 14(f), the coverage of the $\mathbb{R}$-MRP approximation to Levy density $\dot{l}$ in App. 7.2 with 700 modes is shown for the highest 90%, 50% and 10% corresponding to α values of 0.9, 0.5 and 0.1 in shades of dark gray, light gray, and black. Moreover they can be used to obtain the highest Bayesian posterior sets as $\mathbb{B}$-MRPs or *Boolean mapped regular pavings*, where elements of the Boolean set $\mathbb{B} = \{\texttt{true}, \texttt{false}\}$ are mapped to leaf boxes of the regular paving in order to represent subsets of the root box $\boldsymbol{x}_\rho \in \mathbb{IR}^d$ (practical typically for $d \leq 6$ in a non-distributed single-machine setting considered here). Such $\mathbb{B}$-MRPs are closed under tree arithmetic operations of unions, intersections, set-difference, etc. akin to $\mathbb{R}$-MRPs [Harlow et al. 2012, Example 2 and Fig. 9], and can be be processed further to take set-valued decisions (eg. [Jaulin et al. 2001; Tucker 2011]).

5.2. Phase density estimation for a chaotic double pendulum

We obtain data from a mechatronically measurable double pendulum [Lawrence et al. 2010] with sufficient energy in the initial condition to produce chaotic trajectories in the first 25 seconds of observation. Data from three such trajectories with nearly the same initial conditions are used to estimate the density of the angular position of each arm in Figure 15. The number of sample points for the angular position of the two arms was $348,713$ and the number of leaves in the OPTMAP estimate was $71,720$.

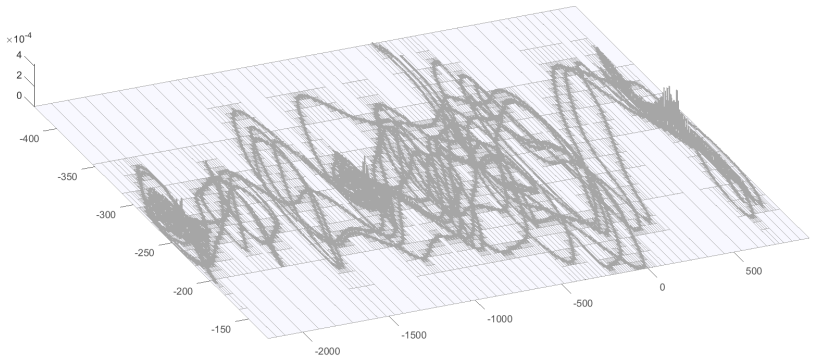


Fig. 15. The OPTMAP histogram density estimate of the positions taken by the two arms (in degrees) from the first 25 seconds of three independently sampled trajectories from the mechatronically measurable double pendulum.

6. DISCUSSION AND FUTURE DIRECTIONS

6.1. Summary

In this paper we formalized the regular paving (RP) data structure and its extensions such as real mapped regular paving or $\mathbb{R}$-MRP and statistical regular paving or SRP, and showed that by using a statistically-equivalent-blocks based priority-queued Markov chain (SEB-PQMC) to partition in a data-adaptive manner, an L_1-consistent adaptive histogram can be obtained. Furthermore, by a complementary partitioning scheme based on a support-carving priority-queued Markov chain (SPC-PQMC), we were able to produce better partitions collaboratively with the SEB-PQMC. The high posterior states along these PQMC paths can be used to initialize MCMC as in [Sainudiin et al. 2013; Harlow 2013] in order to obtain Bayes posterior mean estimates or to obtain the maximum *a posteriori* (MAP) estimates for a given prior. Futhermore, by minimizing Stone's leave-one-out score we can obtain the optimal prior for the MAP estimate (OPTMAP). Finally, by using the collator regular paving (CRP), we can obtain the minimum distance estimate (MDE) with universal performance guarantees. All the methods are implemented and available in MRS 2.0 [Sainudiin et al. 2016].

6.2. Fully Bayesian Extensions

A current intentional limitation of our posterior mean density estimate over SRP histograms is the approximation of the likelihood of the data given SRP s is approximated by the maximum likelihood value from the histogram f_n with bins given by the partition $\boldsymbol{x}_{\mathbb{L}(s)}$ of the root box of s. We do not take Dirichlet priors as done in fully Bayesian settings primarily because we are interested in datasets that are known to have null sets that are full compact subsets of the root box as in the case of the chaotic double pendulum experiment of Section 5.2, where priors that assign positive probability over $\boldsymbol{x}_\rho \in \mathbb{IR}^3$ are unacceptable on experimental grounds. However, in many situations it may be desirable to allow for such priors. A fully Bayesian approach involving a prior distribution on a class of simple functions over leaf boxes of s that are non-negative and integrate to 1 by an adaptation of Dirichlet distributions used in the constructions of classical Pólya trees [Lavine 1992; 1994] would be a natural extension of our estimator. The posterior form in [Lu et al. 2013, Eq. (1)] is one of the easiest Dirichlet extensions one can incorporate into our posterior density over SRPs since their partition prior parameter β is nothing but our $1/\tau$ up to proportionality. Such Dirichlet process priors over SRP histogram trees would not only have the arithmetical efficiency of averaging histograms with different partitions that are not mere sequential refinements in the strict Pólya sense (for eg. [Lavine 1992; Wong and Ma 2010; Lu et al. 2013; Jiang et al. 2016]), but also benefit from the fully Bayesian setting with SRP-adaptations of various Pólya tree and Dirichlet priors. We hope that research in this integrative direction will continue.

6.3. More flexible PQMCs from higher-order sample moments

SEB-based PQMC partitioning will tend to result in inefficient partitioning and undersmoothing in areas where the data is relatively high but has a nearly flat density. Using priority functions that use the deviation of the recursively computable variance covariance matrix from that of the uniform density on the box can be used to ameliorate this issue. A similar priority function can also be defined from the deviation of the mean vector in each box from the center of the box. As long as the three conditions in the theorem of Lugosi and Nobel [1996] are satisfied, such refined priority functions from higher-order recursively computable statistics may be of help for highly structured densities.

6.4. Towards minimal distance estimation across arbitrary SRP histograms

We limited our minimum distance estimate (MDE) to the candidate set given by the SRP histograms visited along the path of an SEB-PQMC. This was done to take advantage of the the structure of consecutive refinements of the tree partitions along a single path of the SEB-PQMC. It will be straighforward to apply this procedure to the optimal path from the joint or collaborative SPC/SEB-PQMC that is identified from the OPTMAP method. However, obtaing 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.) as long as they can be represented by $\mathbb{R}$-MRPs. We plan to implement efficient versions of the Scheffé tournament in the future.

7. APPENDICES: EXAMPLE DENSITIES

7.1. Density I

Density I is a mixture of two multivariate Normal densities for $x \in \mathbb{R}^d$. Density I has no correlation between data coordinates and moderate bimodality:

$$f_I(x \,|\, \mu_a, \Sigma_a, \mu_b, \Sigma_b) = \frac{1}{2}\varphi(x \,|\, \mu_a, \Sigma_a) + \frac{1}{2}\varphi(x \,|\, \mu_b, \Sigma_b),$$

where $\varphi(x \,|\, \mu, \Sigma)$ is the multivariate Normal density with mean $\mu \in \mathbb{R}^d$ and $d \times d$ variance-covariance matrix Σ, and

$$\mu_a = \begin{pmatrix} 1.0 \\ 0.0 \\ \vdots \\ 0.0 \end{pmatrix}, \quad \Sigma_a = \begin{pmatrix} \sigma_a(x_1, x_1) & 0 & \cdots & 0 \\ 0 & \sigma_a(x_2, x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_a(x_d, x_d) \end{pmatrix},$$

$$\mu_b = \begin{pmatrix} 2.5 \\ \vdots \\ 2.5 \end{pmatrix}, \quad \Sigma_b = \begin{pmatrix} \sigma_b(x_1, x_1) & 0 & \cdots & 0 \\ 0 & \sigma_b(x_2, x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_b(x_d, x_d) \end{pmatrix},$$

and

$$\sigma_a(x_i, x_i) = \frac{1.5}{1 + \left(\frac{i-1}{2}\right)}, \sigma_b(x_i, x_i) = \frac{0.625}{1 + \left(\frac{i-1}{4}\right)} \; i = 1, \ldots, d, .$$

When $d = 2$ Density I is a mixture of two bivariate Normal densities with

$$\mu_a = \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}, \quad \Sigma_a = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.0 \end{pmatrix}, \quad \mu_b = \begin{pmatrix} 2.5 \\ 2.5 \end{pmatrix}, \quad \Sigma_b = \begin{pmatrix} 0.625 & 0 \\ 0 & 0.5 \end{pmatrix}.$$

7.2. Levy Density

The bivariate Levy density $\dot{l}(t_1, t_2)$ over $\boldsymbol{x}_\rho = [-10, 10]^2$ with normalising constant $N_l := \int_{[-10,10]^2} l(t_1, t_2) dt_1 dt_2$ has 700 modes and is given by:

$$\begin{aligned}\dot{l}(t_1, t_2) =& \frac{1}{N_l} l(t_1, t_2), \quad l(t_1, t_2) = \exp\left(-\Upsilon(t_1, t_2)\right), \\ \Upsilon(t_1, t_2) =& \sum_{i=1}^{5} i \cos\left((i-1)t_1 + i\right) \sum_{j=1}^{5} j \cos\left((j+1)t_2 + j\right) \\ &+ (t_1 + 1.42513)^2 + (t_2 + 0.80032)^2.\end{aligned}$$

7.3. Multivariate Rosenbrock

We obtain $r_d(t)$, the Rosenbrock density in d dimensions over some box $\boldsymbol{x}_\rho \in \mathbb{IR}^d$, by appropriately normalising the Rosenbrock shape given by:

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

AUTHORS' CONTRIBUTIONS

RS and GT formalized the Markov chains over SRPs and proved the main Theorems. JH, RS and WT formalized arithmetics over RPs, JH and RS extended them over SRPs, and RS and GT extended them over CRPs for MDE. JH designed and implemented most of the data-structures and algorithms. RS wrote the first draft. All authors revised the final draft.

ACKNOWLEDGMENTS

RS thanks Hosam Mahmoud for pointer to the shape functional of Dobrow and Fill [1995], Robert C. Griffiths for combinatorial guidance on planar binary trees, and Jan-Erik Björk for pointers in analysis of set functions. This research was partly supported by RS's external consulting revenues from the New Zealand Ministry of Tourism, University of Canterbury (UC) MSc Scholarship to JH, UC College of Engineering Sabbatical Grant and Visiting Scholarship at Department of Mathematics, Cornell University, Ithaca NY, USA and completed through the 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.

REFERENCES

BALTRUNAS, L., MAZEIKA, A., AND BOHLEN, M. 2006. Multi-Dimensional Histograms with Tight Bounds for the Error. In *Proceedings of the 10th International Database Engineering and Applications Symposium*. IEEE Computer Society, Washington, D.C., 105–112.

BENTLEY, J. L. 1975. Multidimensional binary search trees used for associative searching. *Commun. ACM 18,* 9, 509–517.

BIRGÉ, L. AND ROZENHOLC, Y. 2006. How Many Bins Should be put in a Regular Histogram. *ESAIM: Probability and Statistics 10*, 24–25.

CASTELLAN, G. 1999. Modified Akaike's Criterion for Histogram Density Estimation. Tech. rep., Université Paris-Sud, Orsay.

DEVROYE, L., GYÖRFI, L., AND LUGOSI, G. 1996. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

DEVROYE, L. AND LUGOSI, G. 2001. *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York.

DEVROYE, L. AND LUGOSI, G. 2004. Bin Width Selection in Multivariate Histograms by the Combinatorial Method. *TEST 13,* 1, 129–145.

DOBROW, R. P. AND FILL, J. A. 1995. On the markov chain for the move-to-root rule for binary search trees. *The Annals of Applied Probability 5,* 1, 1–19.

FERGUSON, T. S. 1973. A bayesian analysis of some nonparametric problems. *The Annals of Statistics 1,* 2, 209–230.

FERSON, S., KREINOVICH, V., HAJAGOS, J., OBERKAMPF, W., AND GINZBURG, L. 2007. *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*. SANDIA REPORT SAND2007-0939. Sandia National Laboratories, Albuquerque, N.M.

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

FLAJOLET, P. AND SEDGEWICK, R. 2009. *Analytic Combinatorics* 1 Ed. Cambridge University Press, New York, NY, USA.

FREEDMAN, D. A. 1963. On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics 34,* 4, 1386–1403.

GHOSH, J. K. AND RAMAMOORTHI, R. 2003. *Bayesian nonparametrics*. Springer series in statistics. Springer, New York, Berlin, Paris.

GRAY, A. G. AND MOORE, A. W. 2003a. Nonparametric Density Estimation: Towards Computational Tractability. In *SIAM International Conference on Data Mining*. SIAM, San Francisco, California, USA, 203–211.

GRAY, A. G. AND MOORE, A. W. 2003b. Rapid Evaluation of Multiple Density Models. In *Proceedings of the Ninth Conference on Artificial Intelligence and Statistics (AISTATS)*, C. Bishop and F. B.J., Eds. Society for Artificial Intelligence and Statistics, Key West, Florida, USA.

GREENGARD, L. AND STRAIN, J. 1991. The fast gauss transform. *SIAM Journal on Scientific and Statistical Computing 12,* 1, 79–94.

GRENANDER, U. 1981. *Abstract inference*. Wiley series in probability and mathematical statistics. Wiley, New York. A Wiley-Interscience publication.

HAHN, H. AND ROSENTHAL, A. 1948. *Set Functions*. University of New Mexico Press, Albuquerque, N.M.

HARLOW, J. 2013. Data-adaptive multivariate density estimation using regular pavings, with applications to simulation-intensive inference. M.S. thesis, University of Canterbury.

HARLOW, J., SAINUDIIN, R., AND TUCKER, W. 2012. Mapped regular pavings. *Reliable Computing 16*, 252–282.

HOFSCHUSTER, K. 2003. C-XSC 2.0 - a C++ Class Library for Extended Scientific Computing. Tech. rep., Universität Wuppertal, BUW-WRSWT 2003/5.

JAULIN, L., KIEFFER, M., DIDRIT, O., AND WALTER, E. 2001. *Applied Interval Analysis with Examples in Parameter and State Estimation, Robust Control and Robotics*. Springer-Verlag, London.

JIANG, H., MU, J. C., YANG, K., DU, C., LU, L., AND WONG, W. H. 2016. Computational aspects of optional pÃşlya tree. *Journal of Computational and Graphical Statistics 25,* 1, 301–320.

KIEFFER, M., JAULIN, L., BRAEMS, I., AND WALTER, E. 2001. Guaranteed set computation with subpavings. In *Scientific Computing, Validated Numerics, Interval Methods, Proceedings of SCAN 2000*, W. Kraemer and J. Gudenberg, Eds. Kluwer Academic Publishers, New York, 167–178.

KLEMELÄ, J. 2009. *Smoothing of Multivariate Data: Density Estimation and Visualization*. Wiley, Chichester, United Kingdom.

KRUSE, R. L. 1987. *Data Structures and Program Design* Second Ed. Prentice-Hall, Englewood Cliffs, NJ, Chapter 8, 273–317.

LAVINE, M. 1992. Some aspects of polya tree distributions for statistical modelling. *The Annals of Statistics 20*, 1222–1235.

LAVINE, M. 1994. More aspects of polya tree distributions for statistical modelling. *The Annals of Statistics 22*, 1161–1176.

LAWRENCE, P., STUART, M., BROWN, R., TUCKER, W., AND SAINUDIIN, R. 2010. A mechatronically measurable double pendulum for machine interval experiments. *Indian Statistical Institute Technical Report, isibang/ms/2010/11*, 1–40.

LÓPEZ-RUBIO, E. AND DE LAZCANO-LOBATO, J. M. O. 2008. Soft clustering for nonparametric probability density function estimation. *Pattern Recognition Letters 29,* 16, 2085–2091.

LU, L., JIANG, H., AND WONG, W. H. 2013. Multivariate density estimation by bayesian sequential partitioning. *Journal of the American Statistical Association 108,* 504, 1402–1410.

LUGOSI, G. AND NOBEL, A. 1996. Consistency of Data-Driven Histogram Methods for Density Estimation and Classification. *The Annals of Statistics 24,* 2, 687–706.

MAHALANABIS, S. AND STEFANKOVIC, D. 2008. Density estimation in linear time. In *21st Annual Conference on Learning Theory - COLT 2008*, R. A. Servedio and T. Zhang, Eds. Omnipress, Helsinki, Finland, 503–512.

MALEWICZ, G., AUSTERN, M. H., BIK, A. J., DEHNERT, J. C., HORN, I., LEISER, N., AND CZAJKOWSKI, G. 2010. Pregel: A system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. SIGMOD '10. ACM, New York, NY, USA, 135–146.

MASSART, P. 2007. *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilitiés de Saint-Flour XXXIII — 2003*. Springer-Verlag, Berlin, Germany.

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 University Press, Cambridge, United Kingdom.

MOORE, A. W. 2000. The anchors hierarchy: Using the triangle inequality to survive high dimensional data. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*. UAI'00. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 397–405.

OZERTEM, U. AND ERDOGMUS, D. 2011. Locally defined principal curves and surfaces. *J. Mach. Learn. Res. 12*, 1249–1286.

PARZEN, E. 1962. On estimation of a probability density function and mode. *Ann. Math. Statist. 33,* 3, 1065–1076.

PEARSON, K. 1895. Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 186*, 343–414.

RISSANEN, J., SPEED, T., AND YU, B. 1992. Density Estimation by Stochastic Complexity. *IEEE Transactions on Information Theory 38,* 2, 315–323.

ROKHLIN, V. 1985. Rapid solution of integral equations of classical potential theory. *Journal of Computational Physics 60,* 2, 187–207.

ROSENBLATT, M. 1956. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist. 27,* 3, 832–837.

ROZENHOLC, Y., MILDENBERGER, T., AND GATHER, U. 2009. Constructing Irregular Histograms by Penalized Likelihood. Tech. rep., Technische Universität Dortmund, Sonderforschungsbereich 475.

SAINUDIIN, R. 2005. *Machine interval experiments*. Cornell University, Ithaca, New York, USA.

SAINUDIIN, R. 2012. Sequence A185155, The On-line Encyclopedia of Integer Sequences. published electronically.

SAINUDIIN, R., TENG, G., HARLOW, J., AND 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. AND VÉBER, A. 2016. A beta-splitting model for evolutionary trees. *Royal Society Open Science 3,* 5, 1–12.

SAINUDIIN, R., YORK, T., HARLOW, J., TENG, G., TUCKER, W., AND GEORGE, D. 2008–2016. MRS 2.0, a C++ class library for statistical set processing and computer-aided proofs in statistics. `https://github.com/raazesh-sainudiin/mrs2`.

SAMET, H. 1990. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley Longman, Boston.

SAMET, H. 2006. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufman, San Francisco.

SCOTT, D. W. 1979. On optimal and data-based histograms. *Biometrika 66,* 3, 605–610.

SCOTT, D. W. 1992. *Multivariate Density Estimation*. Wiley, New York.

SCOTT, D. W. AND SAIN, S. R. 2005. Multidimensional Density Estimation. In *Handbook of Statistics*, C. R. Rao, E. J. Wegman, and J. L. Solka, Eds. Vol. 24. Elsevier, Amsterdam, The Netherlands, Chapter 9, 229–262.

SCOTT, D. W. AND SHEATHER, S. J. 1985. Kernel density estimation with binned data. *Communications in Statistics - Theory and Methods 14,* 6, 1353–1359.

SILVERMAN, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

STANLEY, R. P. 1997. *Enumerative combinatorics. Vol. 1*. Cambridge Studies in Advanced Mathematics Series, vol. 49. Cambridge University Press, Cambridge. With a foreword by Gian-Carlo Rota, Corrected reprint of the 1986 original.

STANLEY, R. P. 1999. *Enumerative combinatorics. Vol. 2*. Cambridge Studies in Advanced Mathematics Series, vol. 62. Cambridge University Press, Cambridge.

STONE, C. J. 1980. Optimal rates of convergence for nonparametric estimators. *Ann. Statist. 8,* 6, 1348–1360.

STONE, C. J. 1985. An Asymptotically Optimal Histogram Selection Rule. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II*. Wadsworth, Belmont, CA, 513–520.

TAYLOR, C. C. 1987. Akaike's Information Criterion and the Histogram. *Biometrika 74,* 3, 636–639.

TUCKER, W. 2011. *Validated Numerics: A Short Introduction to Rigorous Computations*. Princeton University Press, Princeton, New Jersey.

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. AND CHERVONENKIS, A. Y. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl. 16*, 264–280.

VINCENT, P. AND BENGIO, Y. 2002. Manifold parzen windows. In *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA, USA, 825–832.

WASSERMAN, L. 2003. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York.

WHITTLE, P. 1958. On the Smoothing of Probability Density Functions. *Journal of the Royal Statistical Society . Series B (Methodological) 20,* 2, 334–343.

WONG, W. H. AND MA, L. 2010. Optional pÃŞlya tree and bayesian inference. *The Annals of Statistics 38,* 3, 1433–1459.

XIN, R. S., GONZALEZ, J. E., FRANKLIN, M. J., AND STOICA, I. 2013. Graphx: A resilient distributed graph system on spark. In *First International Workshop on Graph Data Management Experiences and Systems*. GRADES '13. ACM, New York, NY, USA, 2:1–2:6.

YANG, C., DURAISWAMI, R., GUMEROV, N. A., AND DAVIS, L. 2003. Improved fast gauss transform and efficient kernel density estimation. In *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, Nice, France, 664–671.

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

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

ZHANG, X., KING, M. L., AND HYNDMAN, R. J. 2006. A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Density Estimation. *Computational Statistics & Data Analysis 50,* 11, 3009–3031.

8. APPENDIX: ALGORITHMS FOR MINIMUM DISTANCE ESTIMATE

ALGORITHM 2: `SRPCollate`$(\rho, \rho^{(c)})$

input :
(1) The root node ρ 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`(ρ) & (!`IsLeaf`$(\rho^{(c)})$) **then**
    Make temporary nodes L', 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`(ρ) **then**
    Make temporary nodes L', 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`(ρ)) & (!`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 3: `GetYatracos`

input :
(1) the node that was split: ρv^*;
(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 $\boldsymbol{x} \in \{\boldsymbol{x}_{\rho v^* \mathtt{L}}, \boldsymbol{x}_{\rho v^* \mathtt{R}}\}$ **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 $\boldsymbol{x}_\rho \in \{\boldsymbol{x}_{\rho v^* \mathtt{L}}, \boldsymbol{x}_{v^* \mathtt{R}}\}$ **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 $\mathcal{A}_{\Theta_i}$

ALGORITHM 4: `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 Δ_θ vector: $\Delta_{\Theta_{i-1}} \in \mathbb{R}^{(1\times(i))}$.

output : the updated Δ_θ vector: $\Delta_{\Theta_i} \in \mathbb{R}^{(1\times(i+1))}$.

if $i = 0$ **then**
  $\Delta_{\Theta_i} = \emptyset$
end
else
  `// Get` Δ_θ `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 Δ_θ into $\Delta_{\Theta_i}(\theta)$; `// insert into the` θ`-th entry of the vector` Δ_{Θ_i}
  end
  `// Get` Δ_θ `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 Δ_θ into $\Delta_{\Theta_i}(i+1)$
end
return Δ_{Θ_i}