L_1-consistent adaptive multivariate histograms from a randomized queue prioritized for statistically equivalent blocks

UCDMS Research Report 2014/2 (Thu Oct 23 02:02:47 EDT 2014, Ithaca, NY)

Gloria Teng[1], Jennifer Harlow[2], and Raazesh Sainudiin[2]

[1] Universiti Tunku Abdul Rahman, Kuala Lumpur 53300, Malaysia
gloriateng@utar.edu.my
[2] University of Canterbury, Christchurch 8041, New Zealand
jenny.harlow@canterbury.ac.nz
raazesh.sainudiin@gmail.com

Abstract. An L_1-consistent data-adaptive histogram estimator driven by a randomized queue prioritized by a statistically equivalent blocks rule is obtained. Such data-dependent histograms are formalized as real mapped regular pavings ($\mathbb{R}$-MRP). A regular paving (RP) is a binary tree obtained by selectively bisecting boxes along their first widest side. A statistical regular paving (SRP) augments an RP by mutably caching the recursively computable sufficient statistics of the data. Mapping a real value to each element of the partition gives an $\mathbb{R}$-MRP that can be used to represent a piecewise-constant function density estimate on a multidimensional domain. $\mathbb{R}$-MRPs are closed under addition and allow for efficient averaging of histograms with different partitions in any dimension. A partitioning strategy driven by a randomized priority queue of the current leaf nodes of an SRP is formalized as a Markov chain over the space of SRPs and the conditions for its L_1-consistency are obtained.

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 simply 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 sample $X_1, X_2, \ldots, X_n$ drawn from f. This density estimate f_n 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, classification, regression and clustering.

There are two general approaches to density estimation: parametric density estimation and nonparametric density estimation. Here we are concerned with

nonparametric density estimation. Histograms and kernel density estimates are the two most common forms of nonparametric density estimate for data assumed to be drawn from a continuous distribution. Both can be used for univariate and multivariate data. Other density estimators include orthogonal series estimators and nearest neighbour estimators [16, chap. 2]. Adaptations and specializations of these density estimation methods may be used for particular types of data, high-dimensional data, and very large data sets [15, 8]. However it is formed, the density estimate is some smoothed representation of the observed data [18]. The density estimation method determines how this smoothing is performed. Data-adaptive density estimation methods adapt the amount of smoothing to the local density of the data [16, chap. 2].

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 visited by an MCMC algorithm with stationary samples from the posterior distribution was given in [12]. 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 a an *ad hoc* manner in [12]. This paper proposes the use of an L_1-consistent and data-dependent tree-based histogram, that is built using a data structure known as statistical regular paving (SRP), as an initializing strategy for the MCMC in [12]. SRP is an extension of a regular paving (RP) [13, 6, 5], a class of space-partitioning trees that can facilitate efficient arithmetical operations. A real mapped regular paving ($\mathbb{R}$-MRP) is an extension of an RP designed to represent a piecewise-constant function and allow efficient arithmetical operations with them, including the averaging of $\mathbb{R}$-MRP histogram states with *different* partitions that are visited by an MCMC algorithm as in [12]. An SRP augments an RP by mutably caching recursively computable sufficient statistics of the data. A histogram density estimate represented as an $\mathbb{R}$-MRP can then be created from an SRP. Moreover, such histogram density estimates allow for a wide range of subsequent statistical operations, such as, creating marginal and conditional density estimates or evaluating the density estimate at a large set of query points, to be performed efficiently [5].

The paper is laid out as follows. Section 2 reviews various tree-based histogram estimators in the literature. Section 3 introduces the arithmetic and algebra for RPs, $\mathbb{R}$-MRPs, and SRPs, and explains how a histogram can be built using these structures. Section 4 illustrates the use of a randomized priority queue to partition the histogram and a proof of the L_1-consistency of this adaptive partitioning scheme. Section 5 concludes.

2 Tree-based histograms

A histogram is based on a partition of the data space; the elements of the partition are commonly known as bins. The choice of bin width(s) is the smoothing problem: wider bins give more smoothing, narrower bins less smoothing. The bins of a *regular* histogram are all equally-sized; the bins of an *irregular* histogram can vary in size.

Regular partitioning with small enough bins to suit the modes of the density will give too many bins in low or flat density areas [11]. 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. Multivariate histograms with a single bin width are not able to adapt to spatially varying smoothing requirements [7, chap. 17]. A data-dependent partition allows the bin width to vary in a way that is determined by 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 [17], and under certain conditions, a histogram density estimate with a data-dependent partition can be strongly L_1-consistent [9].

A tree structure can be used in algorithms for creating data-adaptive histograms. This is especially suitable where the algorithm uses some form of recursive partitioning strategy, often in association with a penalty function to control complexity. 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 to reduce the tree (reduce the number of elements in the partition. Partitioning trees can also be used in non-greedy complexity-penalized optimization algorithms that perform an exhaustive comparison of a limited set of possible partitions. For a discussion of such tree-based approaches see [7, chap. 17 & 18] and the references therein.

In general, computational efficiency of the methods described above suffer in two basic ways. First, they are not well-suited to very high-dimensional data because the computational complexity of most density estimation algorithms grows exponentially with the dimensions [14, chap. 7], irrespective of the complexity of the underlying density. Second, these methods cannot cope with large volumes of data, say with sample size n around 10^5 or 10^6, even in small dimensions, say dimension d up to 4 or 5. A Metropolis-Hastings Markov chain method was developed in [12] with stationary distribution given by a posterior distribution over regular paving histograms and used to estimate the posterior expectation by exploiting the arithmetic properties of regular pavings when averaging

histogram samples from the chain. The averaged regular paving histogram density estimate was tested with uniform data in up to $d = 1,000$ dimensions and found that the method coped well with this type of high-dimensional unstructured data. Results using data simulated from uniform mixture approximations to non-uniform structured densities such as multivariate Gaussian and Rosenbrock densities showed that the number of dimensions in which the method is computationally feasible with reasonably smaller mean integrated absolute errors is much lower, about $d = 5$ or $d = 6$. However, the method can computationally cope with large volumes of data, even n as high as 10^7, in stark contrast with other available methods.

Many conventional kernel density estimation methods are only effective with data in less than five or six dimensions [3, 19] with sample sizes around few thousands and generally reach computational bottle-necks when the sample size reaches 10^4. The posterior histogram estimate of [12] therefore has some attractions as a density estimation method in up to around five dimensions especially in situations where there is a large amount of sample data available and where it is advantageous to be able to carry out subsequent statistical operations efficiently, directly on the density estimate itself. Such statistical operations include (i) evaluating of the density over a large set of query points for cross-validation, (ii) obtaining the highest density or coverage regions, (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, according to the algorithms in [5].

3 Regular pavings and histograms

This section introduces the notions of RPs, SRPs, and $\mathbb{R}$-MRPs, and explains how a histogram density estimate can be built using these data structures.

3.1 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:

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

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 $\boldsymbol{x}$

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

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

Such a bisection is said to be *regular*. Note that this bisection gives the left child box a half-open interval $[\underline{x}_\iota, \operatorname{mid}(\boldsymbol{x}_\iota))$ 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* [6] or n-tree [13] 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. When the root box $\boldsymbol{x}$ is clear from the context we refer to an RP of $\boldsymbol{x}$ 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 the sequence of child node selections from the root node. We label these nodes and the associated boxes with strings composed of L and R for left and right, respectively. For example, in Figure 1, the root node associated with root box $\boldsymbol{x}_\rho$ is labeled ρ.

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 ρ, thereby turning the cherry node into a leaf node.

Returning to Figure 1, 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).

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 $\boldsymbol{x}_{\rho\mathsf{v}}$ associated with the node $\rho\mathsf{v}$ of an RP of $\boldsymbol{x}_\rho$ is the product of the side-lengths of the box, that is, $\operatorname{vol}(\boldsymbol{x}_{\rho\mathsf{v}}) = \prod_{j=1}^{d} (\overline{x}_{\rho\mathsf{v},j} - \underline{x}_{\rho\mathsf{v},j})$.

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

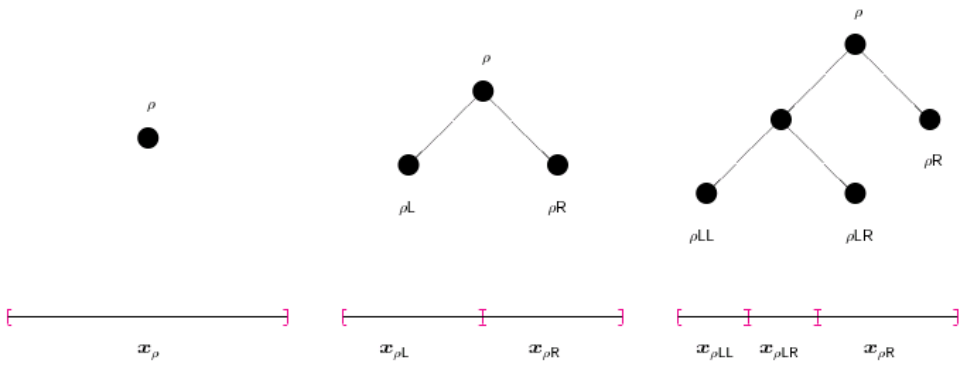


Fig. 1. A sequence of selective bisections, starting from the root, produces an RP.

the regular paving has depth k, then the volume of the box $\boldsymbol{x}_{\rho\mathsf{v}}$ associated with that node is $\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}}) = 2^{-k}\mathrm{vol}\,(\boldsymbol{x}_{\rho})$. This is because any split always results in the child node's box having half the volume of the parent node's box.

In general, an RP is denoted by s. 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 $\mathbb{L}$. The boxes associated with the leaf nodes of an RP are the partition of the root box $\boldsymbol{x}_{\rho}$. The set of leaf boxes of a regular paving s with root box $\boldsymbol{x}_{\rho}$ is denoted by $\boldsymbol{x}_{\mathbb{L}(s)}$. 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.

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

For $i, j \in \mathbb{Z}_+$, where $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$ and $i \leq j$, let $\mathbb{S}_{i:j}$ be the set of regular pavings with k splits where $k \in \{i, i+1, \ldots, j\}$. The space of all regular pavings is then $\mathbb{S}_{0:\infty} := \lim_{j\to\infty} \mathbb{S}_{0:j}$. The size of the space of all regular pavings with between i and j splits, $|\mathbb{S}_{i:j}|$, is given by the sum of Catalan numbers:

$$|\mathbb{S}_{i:j}| = \sum_{k=i}^{j} C_k \quad . \tag{2}$$

The size of the space of all regular pavings with up to k splits is $|\mathbb{S}_{0:k}|$.

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

A real mapped regular paving ($\mathbb{R}$-MRP) is an extension of an RP. 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:

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

Such a map ${}^{\square}f$ is called an $\mathbb{R}$-*mapped regular paving* ($\mathbb{R}$-MRP). Thus, an $\mathbb{R}$-MRP ${}^{\square}f$ is obtained by augmenting each node $\rho\mathsf{v}$ of the RP tree s with an additional data member $f_{\rho\mathsf{v}} \in \mathbb{R}$.

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

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

Arithmetic operations in $\mathbb{R}$ can be extended to $\mathbb{R}$-MRPs [5]. 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. An $\mathbb{R}$-MRP ${}^{\square}f$ can also be transformed using any standard function $\tau \in \mathfrak{S} := \{\exp, \sin, \cos, \tan, \ldots\}$ to obtain the $\mathbb{R}$-MRP $\tau({}^{\square}f)$. 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, 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 therefore dense in $\mathcal{C}(\boldsymbol{x}_\rho, \mathbb{R})$, the algebra of real-valued continuous functions over $\boldsymbol{x}_\rho$ [5]. This ensures that we can uniformly approximate any continuous density $f : \boldsymbol{x}_\rho \to \mathbb{R}$ using $\mathbb{R}$-MRPs in ${}^{\square}\mathcal{F}$.

$\mathbb{R}$-MRPs are important structures in this paper because an $\mathbb{R}$-MRP can be used to represent a piecewise-constant function. [5] describes function approximation using $\mathbb{R}$-MRPs in general. 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 can be carried out efficiently and recursively using trees. A box in any type of RP has real volume ($\mathrm{vol}\,(\boldsymbol{x}_{\rho\mathsf{v}}) \in \mathbb{R}$). This allows operations using both node volume and node value, such as integrating, normalizing and marginalizing, to be carried out on $\mathbb{R}$-MRPs. A *non-negative* $\mathbb{R}$-MRP ${}^{\square}f$ can be used to represent a (possibly non-normalized) probability density function. An $\mathbb{R}$-MRP ${}^{\square}f$ is non-negative if $f_{\rho\mathsf{v}} \geq 0 \ \forall \rho\mathsf{v} \in \mathbb{L}({}^{\square}f)$. Figure 2 shows an $\mathbb{R}$-MRP density estimate of an example density which is a mixture of two bivariate Normal densities for $x \in \mathbb{R}^2$.

3.3 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;

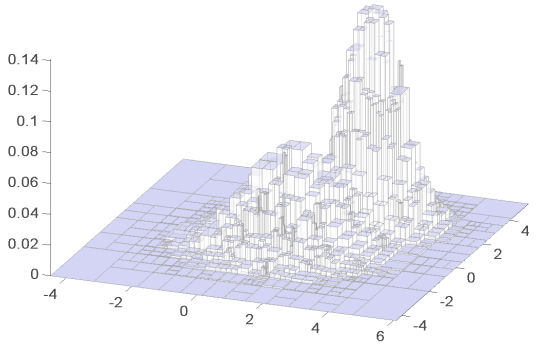


Fig. 2. $\mathbb{R}$-MRP density estimate of a bivariate Gaussian mixture.

– A node of an SRP tree records recursively computable statistics relating to this sample subset.

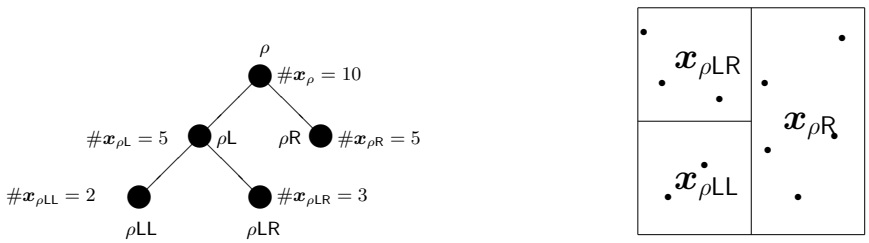
An SRP is denoted by s. Denote $\mathbb{S}_{i:j}$ as the set of all statistical regular pavings with a given root box and k splits where $k \in \{i, i+1, \ldots, j\}$, where $i, j \in \mathbb{Z}_+$ and $i \leq j$. The space of all statistical regular pavings with a given root box is then $\mathbb{S}_{0:\infty}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

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.

For the purpose of this paper, the only 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 $\mathbb{L}^+(s) := \{\rho\mathsf{v} \in \mathbb{L}(s) : \#\boldsymbol{x}_{\rho\mathsf{v}} > 0\} \subseteq \mathbb{L}(s)$.

Figure 3 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 3(a) shows the tree, including the count associated with each node in the tree. Figure 3(b) shows the partition of the root box represented by this tree, with the sample data points superimposed on the box.



(a) SRP tree.

(b) SRP paving.

Fig. 3. A small SRP.

3.4 Statistical regular paving (SRP) histograms as $\mathbb{R}$-MRPs

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:

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

A histogram obtained using Equation (3) is referred to as an SRP histogram. SRP histograms have some similarities to dyadic histograms [7, chap. 18]. 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.

This SRP histogram is a piecewise-constant function that can be represented as an $\mathbb{R}$-MRP in ${}^{\square}\mathcal{F}$. Thus all the $\mathbb{R}$-MRP operations described above can be carried out with the $\mathbb{R}$-MRP histogram density estimate formed from the SRP. Section 4 discusses how the partitioning of the root box of an SRP can be carried out in a data-adaptive and asymptotically L_1-consistent manner.

4 Randomized priority queue for adaptive partitioning

A randomized priority queue (RPQ) partitioning method orders the elements of $\mathbb{L}^{\triangledown}(s)$, the splittable leaf nodes of an SRP s, according to some priority function $\psi : \mathbb{L}^{\triangledown}(s) \to \mathbb{R}$, in order to select the next node to be split from

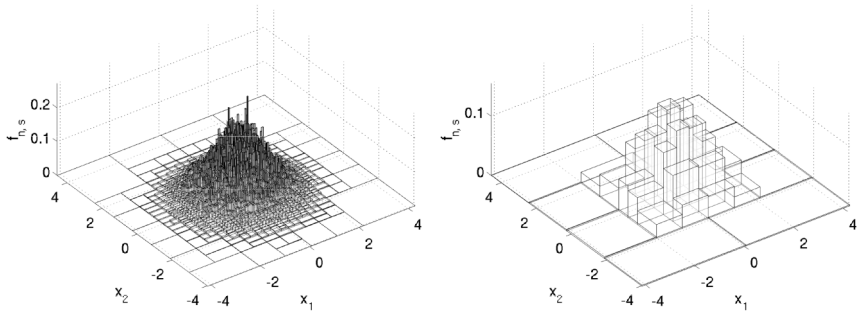


Fig. 4. Two histogram density estimates for the standard bivariate Gaussian density. The left figure shows a histogram with 1485 leaf nodes where $\overline{\#} = 50$ and the histogram on the right has $\overline{\#} = 1500$ resulting in 104 leaf nodes.

$\text{argmax}_{\rho\mathsf{v}\in\mathbb{L}^{\triangledown}(s)}\,\psi(\rho\mathsf{v})$, 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. Two criteria can be specified to stop the RPQ partitioning. 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 $\overline{\psi}$. An RPQ will also stop partitioning if there are no splittable leaf nodes in the SRP.

The RPQ process generates a sequence of states $\{S(t)\}_{t\in\mathbb{Z}_+}$ on $\mathbb{S}_{1:\overline{m}-1}$. If the initial state $S(t=0)$ is the root $s \in \mathbb{S}_0$ then this can be seen as a sequence $\{S(k)\}_{k\in\mathbb{Z}_+}$ on $\mathbb{S}_{0:\overline{m}-1}$ such that $S(k) \in \mathbb{S}_k$, i.e. the $(k+1)^{\text{th}}$ state has $k+1$ leaves or k splits.

A statistically equivalent block (SEB)-based SRP partitioning scheme driven by an RPQ can be used to create a final SRP where each leaf node has at most $\overline{\#}$ of the sample data points associated with it and the total number of leaves is at most $\overline{m}$. The SEB priority function is given by $\psi(\rho\mathsf{v}) = \#\boldsymbol{x}_{\rho\mathsf{v}}$.

Figure 4 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). An approach to solve the smoothing problem by a Bayesian MCMC that exploits the efficient averaging of $\mathbb{R}$-MRP histograms in ${}^{\square}\mathcal{F}$ was proposed in [12] (as discussed in Section 2). The crucial initialization strategy for the MCMC was not justified in [12]. It is proved here to be L_1-consistent.

We now show that an RMRP density estimate based on an SRP created using the SEB RPQ partitioning scheme is asymptotically L_1-consistent provided that $\overline{\#}$ and $\overline{m}$ grow with the sample size n at appropriate rates. This is done by proving the three conditions in Theorem 1 of [9]. 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 RPQ. The Markov chain terminates at some state $\dot{s}$ with partition $\ell(\dot{s})$. Associated with the Markov chain is a fixed collection of partitions

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

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

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

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

Given n fixed points $\{X_1, \ldots, X_n\} \in \left(\mathbb{R}^d\right)^n$. Let $\mathit{\Pi}\left(\mathcal{L}_n, \{X_1, \ldots, X_n\}\right)$ be the number of distinct partitions of the finite set $\{X_1, \ldots, X_n\}$ that are induced by partitions $\ell(\dot{s}) \in \mathcal{L}_n$:

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

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

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

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 RPQ.

Theorem 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*

RPQ 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 strongly 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 strongly consistent in L_1.

Condition (a) is satisfied by the assumption that $\overline{m}/n \to 0$ since $m(\mathcal{L}_n) \leq \overline{m}$ (see Remark 1).

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.

$$\Pi^*_n(\mathcal{L}_n) \leq |\mathcal{L}_n| \leq \sum_{i=0}^{\overline{m}-1} C_i$$

where i 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 ([10]):

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

Taking logs and dividing by n on both sides we get

$$\begin{aligned} \log \Delta^*_n(L_n)/n &\leq \log \left(\frac{4^{(m(\mathcal{L}_n)+1)}}{3m(\mathcal{L}_n)\sqrt{\pi m(\mathcal{L}_n))}} \right) /n \\ &\leq \tfrac{1}{n}(m(\mathcal{L}_n)+1)\log 4 - \tfrac{1}{n}\log 3\sqrt{(\pi)} - \tfrac{3}{2n}\log m(\mathcal{L}_n). \end{aligned}$$

The first and third term goes to 0 by an application of condition (a). The second term which is just a constant divided by n also vanishes as $n \to \infty$. Therefore, condition (b) is satisfied.

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, $\mu(\hat{\boldsymbol{x}}^c) < \xi$. 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 $\hat{\mathbb{X}}$ and δ boxes at the lower dimensional boundaries of $\hat{\mathbb{X}}$, i.e. there are at most m_γ disjoint boxes in $\hat{\boldsymbol{x}}$ that have diameter greater than γ, where

$$m_\gamma < 2^{di} + \delta, \quad \delta = \left(2^d + \sum_{j=1}^{d-1} 2^{d-j} \binom{d}{j} 2^{ij} \right) \ . \tag{4}$$

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

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

The first term in the parenthesis converges to zero since $\overline{\#}/n \to 0$ by assumption. For $\epsilon > 0$, the second term goes to zero by applying the Vapnik-Chervonenkis theorem to boxes in $\mathbb{IR}^d$ with shatter coefficient $s(\mathbb{IR}^d, n) = 2^{2d}$ [1, p. 220], i.e.

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

By the Borel-Cantelli lemma,

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

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

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

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

Remark 1. We can choose $\overline{\#}$ to be some sub-linear function of n, say n^α. Then $\alpha > 0$ so that $\overline{\#} \to \infty$ and $\alpha < 1$ so that $\overline{\#}/n \to 0$. Now let $\overline{m} = n^\beta$, then $\beta > 0$ so that $\overline{m} \geq n/\overline{\#}$. The above constraints imply that $\alpha + \beta \geq 1$. Finally $\beta < 1$ such that $\overline{m}/n \to 0$.

5 Conclusions

In this paper we formalized the RP data structure and its extensions, SRP and $\mathbb{R}$-MRP, and showed that by using an SEB RPQ partitioning scheme, an L_1-consistent adaptive histogram can be obtained. This can be used to initialize MCMC as in [12] to obtain Bayesian smoothed density estimates.

Note that the regular paving structure places some restrictions on the density estimate due to the way the bisections are selected, but has the advantage of allowing a wide range of statistical operations to be carried out efficiently on piecewise-constant density estimate from the dense class of ${}^{\square}\mathcal{F}$, real mapped regular pavings [5]. In addition, a collection of histogram density estimates from ${}^{\square}\mathcal{F}$ with different partitions in any dimension can be efficiently averaged even with large sample sizes. Up to a given prior distribution, the resulting nonparametric Bayesian density estimate is a smoothed $\mathbb{R}$-MRP representation of the posterior sample mean with lower mean integrated absolute error than any particular $\mathbb{R}$-MRP histogram state visited by the MCMC algorithm [12]. A major advantage of the SEB-based RPQ algorithm is that, run for a limited number of states, it can be an effective way to create an over-smoothed histogram that reflects the major features of the density of the sample data. This gives an SRP that provides a much better starting point for further data-adaptive partitioning than the original, unpartitioned, SRP. By partitioning deeper into the state space with small $\overline{\#}$ we can find histograms with higher posterior densities along the asymptotically consistent path taken by the SEB-based RPQ Markov chain. Such high posterior states can be used to initialize the MCMC algorithm of [12] and thereby minimize the mixing time as done in [4, chap. 6].

However, a drawback of the SRP RPQ algorithm as a density estimator is that, without some idea of the characteristics of the density to be estimated, it is extremely hard to determine suitable values for the parameters controlling the partitioning process. As with the other greedy algorithms discussed in Section 2, the locally optimal choices made by an RPQ algorithm may be globally suboptimal. An SEB-based RPQ has nevertheless been shown to be able to produce an asymptotically consistent density estimate. Cross-validation or minimum distance estimation [2, chap. 6], or other smoothing techniques, could potentially be used with SRP RPQs to produce $\mathbb{R}$-MRP density estimates. These possibilities are currently being explored.

Acknowledgements

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.

References

1. Luc Devroye, László Györfi, and G'aabor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.
2. Luc Devroye and G'aabor Lugosi. *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York, 2001.
3. Alexander G Gray and Andrew W Moore. Nonparametric Density Estimation: Towards Computational Tractability. In *SIAM International Conference on Data Mining*, pages 203–211. SIAM, 2003.
4. J. Harlow. Data-adaptive multivariate density estimation using regular pavings, with applications to simulation-intensive inference. Master's thesis, University of Canterbury, 2013.
5. J. Harlow, R. Sainudiin, and W. Tucker. Mapped regular pavings. *Reliable Computing*, 16:252–282, 2012.
6. M. Kieffer, L. Jaulin, I. Braems, and E. Walter. Guaranteed set computation with subpavings. In W. Kraemer and J.W. Gudenberg, editors, *Scientific Computing, Validated Numerics, Interval Methods, Proceedings of SCAN 2000*, pages 167–178. Kluwer Academic Publishers, New York, 2001.
7. Jussi Klemelä. *Smoothing of Multivariate Data: Density Estimation and Visualization.* Wiley, Chichester, United Kingdom, 2009.
8. Dongryeol Lee and Alexander Gray. Fast High-Dimensional Kernel Summations Using the Monte Carlo Multipole Method. In *Advances in Neural Information Processing Systems (NIPS), 21 (2008)*, pages 929–936. MIT Press, 2009.
9. Gábor Lugosi and Andrew Nobel. Consistency of Data-Driven Histogram Methods for Density Estimation and Classification. *The Annals of Statistics*, 24(2):687–706, 1996.
10. S. Mattarei. Asymptotics of partial sums of central binomial coefficients and Catalan numbers. arXiv.0906.4290v3, January 2010.
11. Jorma Rissanen, TP Speed, and Bin Yu. Density Estimation by Stochastic Complexity. *IEEE Transactions on Information Theory*, 38(2):315–323, 1992.
12. R. Sainudiin, G. Teng, J. Harlow, and D. S. Lee. Posterior expectation of regularly paved random histograms. *ACM Transactions on Modeling and Computer Simulation*, 23(26), 2013.
13. H. Samet. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley Longman, Boston, 1990.
14. David W. Scott. *Multivariate Density Estimation.* Wiley, New York, 1992.
15. David W Scott and Stephan R Sain. Multidimensional Density Estimation. In C. R. Rao, E. J. Wegman, and J. L. Solka, editors, *Handbook of Statistics*, volume 24, chapter 9, pages 229–262. Elsevier, Amsterdam, The Netherlands, 2005.
16. B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.
17. Charles J Stone. An Asymptotically Optimal Histogram Selection Rule. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II*, pages 513–520, Belmont, CA, 1985. Wadsworth.
18. P. Whittle. On the Smoothing of Probability Density Functions. *Journal of the Royal Statistical Society . Series B (Methodological)*, 20(2):334–343, 1958.
19. Xibin Zhang, Maxwell L. King, and Rob J. Hyndman. A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Density Estimation. *Computational Statistics & Data Analysis*, 50(11):3009–3031, July 2006.