Computational Statistics and Data Analysis 220 (2026) 108367

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Random irregular histograms

Oskar Høgberg Simensen [a,b,*], Dennis Christensen [c], Nils Lid Hjort [b]

[a] Scientific Computing Center Karlsruhe Institute of Technology, Germany
[b] Department of Mathematics, University of Oslo, Norway
[c] Norwegian Defence Research Establishment (FFI), Norway

ARTICLE INFO

Keywords:
Bayesian model selection
Bayesian nonparametrics
Histograms
Nonparametric density estimation
Mode detection

ABSTRACT

A new method of histogram construction is proposed, providing a fully Bayesian approach to irregular histograms. The procedure applies Bayesian model selection to a piecewise constant model of the underlying distribution, resulting in a method that selects both the number of bins as well as their location based on the data in a fully automatic fashion. It is shown that the histogram estimate is consistent with respect to the Hellinger metric under mild regularity conditions, and that it attains a convergence rate equal to the minimax rate (up to a logarithmic factor) for Hölder continuous densities. Simulation studies indicate that the new method performs comparably to other histogram procedures in terms of minimizing estimation error, and that it often outperforms other statistical procedures for identifying modes. A software implementation of the proposed method is provided as part of a publicly available Julia package.

1. Introduction

The much celebrated histogram is the earliest example of a nonparametric density estimator, and remains widespread in use even to this day. Although more efficient density estimators have been devised since, histograms have retained their popularity due to their simple nature and interpretability. The key difficulty encountered when drawing a histogram is that the appearance of the density estimate is sensitive to the choice of partition used to draw the histogram. If the partition is not chosen well, the quality of the resulting histogram will be rather poor, failing to resemble the underlying distribution at all. As a result, the question of designing automatic histogram procedures, where the partition is chosen based on the given sample, has attracted considerable interest in the statistics community; see Scott (2010) for a review. However, this problem has turned out to be a difficult one, with no solution universally accepted as the best one.

Many of the histogram procedures proposed in the statistical literature simplify the problem of selecting the partition somewhat by exclusively considering regular partitions (Birgé and Rozenholc, 2006), where the bins are of equal width, so that one only needs to choose the number k of bins. However, even this simplified problem has not led to a canonical procedure to choose the number of intervals in the partition, as evidenced by the sheer number of different bin selection rules for regular histograms (Li et al., 2020).

Automatic irregular histogram methods provide a more flexible alternative to regular histograms by determining both the number of bins and their location based on the sample. The benefits of using optimally chosen cut points between the intervals to draw a histogram as opposed to a regular grid with the same number of bins are rather immediate, as the bin widths are in the irregular case able to adapt to the local behavior of the underlying density, offering more smoothing near modes and in the tails of the density, resulting in smaller estimation risks (Scott, 1992). Irregular histograms are also more robust to perturbations in local bandwidths;

* Corresponding author.
E-mail addresses: oskar.simensen@kit.edu (O.H. Simensen), dennis.christensen@ffi.no (D. Christensen), nils@math.uio.no (N.L. Hjort).

https://doi.org/10.1016/j.csda.2026.108367
Received 18 July 2025; Received in revised form 10 November 2025; Accepted 2 March 2026
Available online 4 March 2026

while small changes in bin widths can greatly influence the graphical display of a fixed-bandwidth histogram, adaptive methods allow for boundary adjustments at each point, reducing such sensitivity. While the adaptive nature of irregular histograms is an attractive feature from a theoretical perspective, Birgé and Rozenholc (2006) remark that the search for an appropriate set of cut points can lead to increases in the statistical risk of the procedure, which may result in worse density estimates in practice according to classical loss functions.

Although automatic irregular histogram procedures do not always yield smaller estimation risks than regular ones, they often have an advantage when it comes to detecting important features in a density such as modes. Automatic mode identification is a feat that cannot be achieved by regular histogram procedures designed to produce small risks with respect to classical loss functions; as shown by Scott (1992), the asymptotically optimal bin width in terms of $\mathbb{L}_2$ risk results in an undersmoothed histogram if the goal is to automatically infer the modes of a density. In contrast, our proposed irregular histogram procedure shows that there need not be a trade-off between low estimation error and automatic mode identification.

Traditional approaches to mode identification typically determine the number of modes via hypothesis testing (Ameijeiras-Alonso et al., 2019). The test statistics used to assess the modality of the sample are either based on kernel density estimates (Silverman, 1981; Fisher and Marron, 2001) or excess mass statistics (Müller and Sawitzki, 1991; Ameijeiras-Alonso et al., 2019). However, as all of these approaches rely on the use of a nonparametric bootstrap procedure to assess the significance level of the test, they are quite expensive to use in practice. An alternative testing-based approach based on estimating density derivatives was suggested in Genovese et al. (2015). Some recent proposals for identifying modes use model selection procedures instead of testing as a means of identifying the number of modes in a density, including the Gaussian mixture model approach of Chacón (2019) and the Bayesian taut spline method of Chacón and Fernández Serrano (2024). Previous work (Davies et al., 2009; Li et al., 2020) has shown that certain irregular histogram procedures often do well at mode identification, but we are unaware of any articles that directly compare irregular histograms to the more traditional approaches to mode hunting in terms of identifying both the number of modes and their location. We demonstrate through simulations that our histogram procedure often outperforms traditional methods for identifying modes, and does so at a considerably lower computational cost.

Despite the apparent advantages of irregular histogram procedures, they have not been as widely adopted as their regular counterparts (Davies et al., 2009). This is in no doubt in part due to the computational challenges encountered by most irregular histogram methods, which typically involves solving a more difficult optimization problem than that encountered in the regular case. Moreover, many previous proposals for irregular histograms depend on the selection of key tuning parameters, with no universal recommendation given for their default values, which has hindered their use by practitioners (Rozenholc et al., 2010). However, the past few decades has seen a renewed interest in irregular histogram procedures, with many different proposals appearing in the statistical literature that offer automatic histogram construction at speeds that make them an appealing alternative to regular histogram methods; see e.g. Davies and Kovac (2004), Rozenholc et al. (2010), Li et al. (2020), Mendizábal et al. (2023).

Parallel to the development of the irregular histogram literature, the topic of tree-based Bayesian nonparametric density estimators has received much attention in recent years. Examples include the optional Pólya tree of Wong and Ma (2010) and the binary partition-based estimator of Lu et al. (2013), which have both been proven to perform well in a large-sample setting (Castillo and Randrianarisoa, 2022; Liu et al., 2023). Further proposals in this spirit include ensemble estimators such as the Bayesian CART method of Jeong and Rockova (2023). Although both irregular histograms and tree-based estimators are based on a piecewise constant model for the data-generating process, they typically have a different focus. For irregular histogram procedures, the aim is typically to provide a parsimonious account of an observed sample using a moderate number of bins, while tree-based methods mostly attempt to minimize prediction error with little regard for model parsimony.

The irregular histogram model proposed in this paper provides a fully Bayesian approach to histogram estimation. In particular, our method is based on finding the partition which maximizes the posterior probability under a piecewise constant model of the data-generating density, resulting in an automatic data-based rule for choosing the histogram partition. Using a combination of search heuristics and dynamic programming, our approach results in a bin selection rule that is very quick to compute, even for large datasets. Unlike regular Bayesian histogram models, the random irregular histogram method is shown to excel at automatic mode detection while achieving similar performance in terms of classical loss functions, making it an attractive choice for exploratory data analysis. A software implementation of the irregular random histogram estimator proposed in this article can be found in the `AutoHist.jl` Julia package, available as part of the general package registry (Simensen, 2025). The source code used to create all the figures and tables presented in this article is available at the following GitHub repository: https://github.com/oskarhs/Random-Histograms---Paper.

The remainder of the paper is structured as follows. Section 2 gives an introduction to histogram construction, focusing on irregular methods. Section 3 introduces our irregular random histogram model and describes an algorithm to compute the approximate a posteriori most probable partition given an observed sample. Consistency and convergence rate results are given in Section 4. In Section 5, we describe two simulation studies in which we compare the performance of our method with other state-of-the-art histogram procedures from the statistical literature and other methods for identifying modes in a density. In Section 6 we illustrate the proposed density estimator by applying it to some real-world datasets. We conclude the article by discussing possible extensions of our method.

2. Histogram construction

The histogram approach to nonparametric density estimation is based on a piecewise constant model of the data-generating density f. For ease of exposition, we present all models for densities on the unit interval, but note that the methodology presented here can be extended to other compact intervals via a suitable affine transformation. For a given interval partition $\mathcal{I} = (I_1, I_2, \ldots, I_k)$ of $[0, 1]$,

the piecewise constant approximation to the density f is given by

$$f(x \mid \mathcal{I}, \boldsymbol{\theta}) = \sum_{j=1}^{k} \frac{\theta_j}{|\mathcal{I}_j|} \mathbb{1}_{\mathcal{I}_j}(x), \quad x \in [0,1], \tag{1}$$

where $\boldsymbol{\theta}$ belongs to the k-simplex $S_k = \{\boldsymbol{\theta} \in [0,1]^k : \sum_{j=1}^k \theta_j = 1\}$ and $\mathbb{1}_{\mathcal{A}}$ denotes the indicator function for a set $\mathcal{A}$, so $\mathbb{1}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$ and 0 otherwise. Although the dimension of $\boldsymbol{\theta}$ depends on the given partition $\mathcal{I}$, we generally omit this in the notation when the number of bins in the partition in question is clear from the context. For a given partition, estimation of the parameter $\boldsymbol{\theta}$ is usually done by maximizing the likelihood function based on f or minimizing an estimated $\mathbb{L}_2$ loss, yielding the familiar bin proportions $\widehat{\theta}_j = N_j/n$, where $N_j = \sum_{i=1}^n \mathbb{1}_{\mathcal{I}_j}(x_i)$ is the number of observations falling into bin $\mathcal{I}_j$. Although constructing a histogram for a given partition is a simple task, selecting a good partition based on the observed sample constitutes a much more difficult problem. The chosen partition controls the degree of smoothing, with larger intervals inducing a smoother density estimate at the cost of increased bias. Attempting to find a partition that strikes a good balance between smoothing and small bias is the basis of the vast majority of automatic histogram methods.

Most automatic histogram procedures are based on a decision-theoretic framework where the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ has been generated by a true density f_0 and seek to achieve good performance in terms of the frequentist risk,

$$R_n(f_0, \widehat{f}\,) = \mathbb{E}_{\boldsymbol{x} \sim f_0}\left\{\ell(f_0, \widehat{f}\,)\right\}, \tag{2}$$

where $\ell(f_0, \widehat{f}\,)$ is a loss function measuring the quality of using the estimate $\widehat{f}$ as an approximation of f_0. Typical loss functions include those based on powers of the $\mathbb{L}_r$ or Hellinger metrics,

$$\left\|f_0 - \widehat{f}\right\|_r = \left(\int_0^1 \left|f_0(x) - \widehat{f}(x)\right|^r \mathrm{d}x\right)^{1/r},$$

$$d_H(f_0, \widehat{f}\,) = \left(\int_0^1 \left\{\sqrt{f_0(x)} - \sqrt{\widehat{f}(x)}\right\}^2 \mathrm{d}x\right)^{1/2}.$$

On the other hand, approaches based on maximum likelihood attempt to minimize the Kullback–Leibler divergence,

$$K(f_0, \widehat{f}\,) = \int_0^1 f_0(x) \log \frac{f_0(x)}{\widehat{f}(x)}\, \mathrm{d}x.$$

In general, the risk of an estimator will depend on the true density, and there is no procedure that universally yields the smallest possible risk for all densities. As such, most histogram procedures aim for small risks over a large class of densities f_0. The vast majority of regular histogram methods take this approach to histogram estimation by minimizing an asymptotic expression for the risk, a penalized likelihood, or a direct estimate of the risk (Davies et al., 2009). Irregular histogram methods based on risk minimization include proposals based on $\mathbb{L}_2$ leave-one-out cross-validation (Rudemo, 1982; Celisse and Robin, 2008) and the penalized maximum likelihood approach of Rozenholc et al. (2010), which seeks to minimize an upper bound on the Hellinger risk. Not all histogram procedures follow the decision-theoretic paradigm; the approaches of Davies and Kovac (2004), Li et al. (2020) construct an irregular histogram based on approximating the empirical distribution function under a parsimony constraint.

The determination of the cut points of an irregular histogram constitutes a difficult problem from a computational perspective, and most irregular histogram procedures rely on various discretization schemes that specify a finite set of possible cut points between the intervals. The reduction in computational complexity brought about by discretization is in itself not enough to guarantee that the resulting optimization problem can be solved in a reasonable amount of time, even for datasets of modest size. As a result, many irregular histogram procedures are based on a selection criterion with a particular additive structure that enables the use of a dynamic programming algorithm (Kanazawa, 1988), efficient search heuristics (Mendizábal et al., 2023) or both (Rozenholc et al., 2010).

3. A Bayesian approach to irregular histograms

We now describe our proposed approach for constructing an irregular histogram. Assuming that we have observed an independent and identically distributed (i.i.d.) sample $x_1, \ldots, x_n$ on the unit interval, our Bayesian modeling approach proceeds by specifying a parameterized density f as in (1) and prior distributions on the model parameters, $\mathcal{I}, \boldsymbol{\theta}$. This choice of f leads to the following expression for the joint density of the sample $\boldsymbol{x}$,

$$f(\boldsymbol{x} \mid \mathcal{I}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left(\frac{\theta_j}{|\mathcal{I}_j|}\right)^{\mathbb{1}_{\mathcal{I}_j}(x_i)} = \prod_{j=1}^{k} \left(\frac{\theta_j}{|\mathcal{I}_j|}\right)^{N_j}, \tag{3}$$

which depends on the sample only through the bin counts $\boldsymbol{N} = (N_1, \ldots, N_k)$.

In our model, we restrict our attention to partitions with interval endpoints belonging to a given finite set $\mathcal{T}_n = \{\tau_{n,j} : 0 \le j \le k_n\}$, where $\tau_{n,0} = 0$, $\tau_{n,k_n} = 1$ and $\tau_{n,j-1} < \tau_{n,j}$ for all j, and k_n is a sequence of positive integers, growing with the sample size n. Let $\mathcal{P}_{\mathcal{T}_n,k}$ denote the set of interval partitions of $[0,1]$ consisting of k bins with endpoints in $\mathcal{T}_n$, and put $\mathcal{P}_{\mathcal{T}_n} = \cup_{k=1}^{k_n} \mathcal{P}_{\mathcal{T}_n,k}$. Since $\mathcal{P}_{\mathcal{T}_n}$ depends on n, the prior for $\mathcal{I}$ must also necessarily do so, and we write $p_n(\mathcal{I})$ to make this dependence explicit. The prior distribution on the partitions is most easily described by introducing the number k of bins as a further random variable, with prior distribution $k \sim p_n(k)$,

supported on $\{1, 2, \dots, k_n\}$. Conditional on k, the prior distribution on the partitions $p_n(\mathcal{I} \mid k)$ is the uniform distribution on $\mathcal{P}_{\mathcal{T}_n,k}$. Since $|\mathcal{P}_{\mathcal{T}_n,k}| = \binom{k_n-1}{k-1}$, it follows that the prior probability mass function of $\mathcal{I}$ is $p_n(\mathcal{I} \mid k) = \binom{k_n-1}{k-1}^{-1}$. We note that if $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$, then $p_n(\mathcal{I}, k')$ is nonzero only for $k' = k$ and as such, we have the equality $p_n(\mathcal{I}) = \sum_{k'=1}^{k_n} p(k')p_n(\mathcal{I} \mid k') = p_n(k)p_n(\mathcal{I} \mid k)$. Finally, as prior distribution for $\boldsymbol{\theta} \mid \mathcal{I}$ we take a k-dimensional $\mathrm{Dir}(\boldsymbol{a})$ distribution, which has density

$$p(\boldsymbol{\theta} \mid \mathcal{I}) = \frac{\Gamma(a)}{\prod_{j=1}^k \Gamma(a_j)} \prod_{j=1}^k \theta_j^{a_j-1}, \quad \boldsymbol{\theta} \in S_k,$$

where $\boldsymbol{a} = (a_1, \dots, a_k) \in (0, \infty)^k$ and $a = \sum_{j=1}^k a_j$. In general, the parameters a_j may depend on the partition $\mathcal{I}$, although we omit this in the notation. We defer the discussion on how to choose $p_n(k)$ and $\boldsymbol{a}$ to Section 3.2.

The prior-model specification can thus be summarized as

$$\begin{aligned} k &\sim p_n(k), \\ \mathcal{I} \mid k &\sim \mathrm{Unif}\left(\mathcal{I} \,\middle|\, \mathcal{P}_{\mathcal{T}_n,k}\right), \\ \boldsymbol{\theta} \mid \mathcal{I} &\sim \mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{a}), \\ f(x \mid \mathcal{I}, \boldsymbol{\theta}) &= \sum_{j=1}^k \frac{\theta_j}{|I_j|} \mathbb{1}_{I_j}(x), \\ x_i \mid f &\sim f. \end{aligned} \tag{4}$$

The model (4) is similar to some Bayesian approaches to regular histogram estimation, which have been investigated from a theoretical perspective (Hall and Hannan, 1988; Scricciolo, 2007) and from a computationally oriented one (Knuth, 2019). What separates our proposal from these prior approaches is that we consider more than one partition for each value of k, resulting in a more flexible procedure due to the variable bin widths. Our approach also shares similarities with the Bayesian sequential partitioning model of Lu et al. (2013), the main difference being that their model is based on binary partitions, while ours relies on a sample-size dependent grid.

3.1. Posterior distribution

We now derive an expression for the posterior distribution of $\mathcal{I}$. As a first step, we show that the Dirichlet distribution is a conjugate prior for f, conditional on the chosen partition. Indeed, as the likelihood function takes the form in (3), it follows from Bayes' rule that for $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$,

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}, \mathcal{I}) \propto p(\boldsymbol{\theta} \mid \mathcal{I})\, f(\boldsymbol{x} \mid \mathcal{I}, \boldsymbol{\theta}) \propto \prod_{j=1}^k \theta_j^{a_j-1} \prod_{j=1}^k \theta_j^{N_j} \propto \mathrm{Dir}\left(\boldsymbol{\theta} \mid \boldsymbol{a} + \boldsymbol{N}\right).$$

To find the posterior probability of $\mathcal{I}$ up to proportionality, we note that by Bayes' rule,

$$p_n(\mathcal{I} \mid \boldsymbol{x}) \propto p_n(\mathcal{I})\, p(\boldsymbol{x} \mid \mathcal{I}) = p_n(\mathcal{I})\, \frac{p(\boldsymbol{\theta} \mid \mathcal{I})\, f(\boldsymbol{x} \mid \mathcal{I}, \boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{x}, \mathcal{I})},$$

where the last equality follows from $p(\boldsymbol{\theta} \mid \mathcal{I})\, f(\boldsymbol{x} \mid \mathcal{I}, \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \mathcal{I})\, p(\boldsymbol{\theta} \mid \boldsymbol{x}, \mathcal{I})$. Plugging in the expressions for our chosen prior distributions along with the posterior for $\boldsymbol{\theta}$, we arrive at the following expression for the posterior probability of $\mathcal{I}$,

$$p_n(\mathcal{I} \mid \boldsymbol{x}) \propto \frac{p_n(k)}{\prod_{j=1}^k |I_j|^{N_j}} \frac{\prod_{j=1}^k \Gamma(a_j + N_j)}{\prod_{j=1}^k \Gamma(a_j)} \frac{\Gamma(a)}{\Gamma(a+n)} \binom{k_n - 1}{k-1}^{-1}, \quad \mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}. \tag{5}$$

Our proposed Bayesian histogram estimator is based on the partition which minimizes the Bayes risk,

$$\widehat{\mathcal{I}} = \operatorname*{arg\,min}_{\mathcal{J} \in \mathcal{P}_{\mathcal{T}_n}} \mathbb{E}_{\mathcal{I},\boldsymbol{\theta}}\left(\mathbb{E}_{\boldsymbol{x} \sim f}\left\{\ell(\mathcal{I}, \mathcal{J}) \,\middle|\, \mathcal{I}, \boldsymbol{\theta}\right\}\right),$$

where ℓ is a nonnegative loss function. To derive the optimal model selection procedure in terms of the Bayes risk, we work with the 0–1 loss function,

$$\ell(\mathcal{I}, \mathcal{J}) = \begin{cases} 1 & \text{if } \mathcal{J} = \mathcal{I}, \\ 0 & \text{else.} \end{cases}$$

To find the minimizer of the Bayes risk, it suffices to minimize the posterior expected loss, $\mathbb{E}_{\mathcal{I}}\left\{\ell(\mathcal{I}, \mathcal{J}) \,\middle|\, \boldsymbol{x}\right\}$ [cf. Lehmann and Casella (1998, p. 228)], which leads to the following rule for selecting the optimal partition

$$\widehat{\mathcal{I}} = \operatorname*{arg\,max}_{\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n}} p_n(\mathcal{I} \mid \boldsymbol{x}). \tag{6}$$

We refer to $\widehat{\mathcal{I}}$ as the maximum a posteriori (MAP) partition.

Since the histogram is first and foremost a graphical tool, providing point estimates of the density is of great interest. Having computed the maximizer of (6) we can construct a density estimate $\widehat{f}$ based on the conditional posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}, \widehat{\mathcal{I}})$. For a given partition $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$, the density f can be parameterized in terms of the k-dimensional vector $\boldsymbol{\theta}$, and we proceed by estimating $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$, the Bayes estimator under squared $\mathbb{L}_2$ loss conditional on $\mathcal{I}$. Writing $f_{\mathcal{I},\boldsymbol{\delta}}$ for the density given by $f(x \mid \mathcal{I}, \boldsymbol{\delta})$ as in (1) for $\boldsymbol{\delta} \in \mathcal{S}_k$, we can expand the loss as

$$\left\| f_{\mathcal{I},\boldsymbol{\theta}} - f_{\mathcal{I},\boldsymbol{\delta}} \right\|_2^2 = \sum_{j=1}^{k} \int_{I_j} \left\{ f_{\mathcal{I},\boldsymbol{\theta}}(x) - f_{\mathcal{I},\boldsymbol{\delta}}(x) \right\}^2 \mathrm{d}x = \sum_{j=1}^{k} |I_j|^{-1} \left\{ \theta_j - \delta_j \right\}^2.$$

Since the Bayes estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ minimizes the posterior expected loss, we need to solve the following optimization problem:

$$\min_{\boldsymbol{\delta} \in \mathcal{S}_k} \mathbb{E}_{\boldsymbol{\theta}} \left\{ \left\| f_{\mathcal{I},\boldsymbol{\theta}} - f_{\mathcal{I},\boldsymbol{\delta}} \right\|_2^2 \,\middle|\, \boldsymbol{x} \right\} = \min_{\boldsymbol{\delta} \in \mathcal{S}_k} \sum_{j=1}^{k} |I_j|^{-1} \mathbb{E}_{\boldsymbol{\theta}} \left\{ (\theta_j - \delta_j)^2 \,\middle|\, \boldsymbol{x} \right\}.$$

This problem is most easily solved by first removing the restriction $\boldsymbol{\delta} \in \mathcal{S}_k$, and showing that the unconstrained and constrained optima coincide. The unconstrained version of the above problem then reduces to minimizing $\mathbb{E}_{\boldsymbol{\theta}} \left\{ (\theta_j - \delta_j)^2 \,\middle|\, \boldsymbol{x} \right\}$ with respect to δ_j for each j, the solution of which is $\widehat{\theta}_j = \mathbb{E}_{\boldsymbol{\theta}} \{ \theta_j \mid \boldsymbol{x} \}$ (Lehmann and Casella, 1998, p. 228). Using the linearity and monotonicity properties of expectations, one quickly verifies that $\widehat{\boldsymbol{\theta}} \in \mathcal{S}_k$, so it must also be the solution to the constrained problem. As the posterior distribution of $\boldsymbol{\theta}$ given $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$ is a Dirichlet distribution, we can read off the posterior mean directly, yielding Bayes estimates

$$\widehat{\theta}_j = \frac{a_j + N_j}{a + n} = \frac{a}{a+n} \frac{a_j}{a} + \frac{n}{a+n} \frac{N_j}{n}, \quad j = 1, 2, \ldots, k. \tag{7}$$

Since the prior mean of θ_j is a_j/a, we see that the posterior mean is a convex combination of the prior mean and the maximum likelihood estimate N_j/n. If a is small relative to n, then the data-based part dominates the estimate, and the estimated bin probabilities end up being close to the corresponding maximum likelihood estimates. Our Bayesian approach thus differs from the penalized likelihood approach mainly in the selection of the optimal partition and not in the estimation of bin probabilities conditional on the selected model.

After computing the optimal partition and the corresponding Bayes estimates of the bin probabilities, a point estimate of the density f can be obtained via

$$\widehat{f}_{\widehat{\mathcal{I}}}(x) = \sum_{j=1}^{k} \frac{\widehat{\theta}_j}{\left| \widehat{I}_j \right|} \mathbb{1}_{\widehat{I}_j}(x). \tag{8}$$

We refer to the estimator (8) as the *Bayes histogram estimator*. The $\widehat{\theta}_j$ serve as estimates of the true bin probabilities $\theta_j = \int_{I_j} f(x)\,\mathrm{d}x$. Although the form of (8) is rather simple, the bin selection procedure itself is quite complex, as it requires finding the best partition according to (5) among a field of $2^{k_n - 1}$ candidate partitions. As argued by Wand (1997), the use of a complex bin selection procedure seems almost inevitable for a histogram method designed to perform well for a large class of densities. In fact, simulation studies show that most well working automatic regular and irregular histogram methods either solve an optimization problem over a set of possible partitions or estimate derivatives of the underlying density (Birgé and Rozenholc, 2006; Davies et al., 2009).

Remark 1. Although the primary focus of our article is the study of the Bayes histogram estimator, we note that full posterior inference for f is available through the posterior distribution for $\mathcal{I}$ given by (5). In particular, Markov chain Monte Carlo (MCMC) methods or importance sampling estimators as in Lu et al. (2013) can be used to generate approximate posterior samples, which in turn can be used to estimate posterior quantities of interest, such as the posterior probability distribution of the number of modes in the density.

Remark 2. The estimator in (8) is not the overall Bayes estimator of the density f under the standard $\mathbb{L}_2$ loss as it does not minimize the Bayes risk. The minimizer of the posterior expected loss is rather the posterior mean of f,

$$\mathbb{E}_f \left\{ f(x) \,\middle|\, \boldsymbol{x} \right\} = \sum_{\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n}} p_n(\mathcal{I} \mid \boldsymbol{x})\, f\left(x \mid \mathcal{I}, \widehat{\boldsymbol{\theta}}\right), \tag{9}$$

which is a model averaging estimator, weighting each conditional posterior mean $\mathbb{E}_f \left\{ f(x) \,\middle|\, \boldsymbol{x}, \mathcal{I} \right\}$ according to the posterior probability of the corresponding partition $\mathcal{I}$. Since the partition $\mathcal{J}$ which includes every grid point in $\mathcal{T}_n$ is a refinement of every $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n}$, the estimator (9) is also a piecewise constant density with cut points in $\mathcal{T}_n$, one could also consider using it as a histogram-type estimator of the density. We do not pursue this approach here for two reasons. The first reason for favoring (8) over the posterior mean is that it is quick to compute; the posterior mean is a sum of $2^{k_n - 1}$ terms, making the evaluation of (9) infeasible even for moderate values of k_n. Although simulation from the posterior $\mathcal{I} \mid \boldsymbol{x}$ is possible through MCMC, such a procedure may suffer from slowly mixing chains and requires performing convergence diagnostics, which makes the model fitting procedure rather involved compared to alternative automatic histogram methods. On the other hand, an approximation to the Bayes histogram estimator can be computed quickly using the methods of the next subsection without requiring any tuning from the user. Another reason for preferring the Bayes histogram estimator is that of model parsimony, as it depends only on the chosen partition and the corresponding interval probability estimates (7). Our simulations show that (8) often yields quite good visual summaries of the underlying density with a relatively small number of bins. In contrast, the computation of $\mathbb{E}_f \left\{ f(x) \,\middle|\, \boldsymbol{x} \right\}$ involves computing a weighted average of the estimated bin probabilities from many different models, which leads to a complicated expression for the resulting bin heights.

3.2. Computing the Bayes histogram estimator

The key observation that allows us to find the maximizer of $p_n(\mathcal{I} \mid \boldsymbol{x})$ efficiently is that we can write the logarithm of the right hand side in (5) as

$$\Phi(\mathcal{I}) + \Psi_n(k), \quad \mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}, \tag{10}$$

where Φ and Ψ_n are given by

$$\Phi(\mathcal{I}) = \sum_{j=1}^{k} \left\{ \log\Gamma(a_j + N_j) - \log\Gamma(a_j) - N_j \log|I_j| \right\},$$

$$\Psi_n(k) = \log p_n(k) + \log\Gamma(a) - \log\Gamma(a+n) - \log\binom{k_n - 1}{k-1}.$$

Φ is additive with respect to the intervals of the partition in the sense that we can write $\Phi = \sum_{j=1}^{k} \Phi_0(I_j)$, where $\Phi_0(I_j)$ depends only on I_j. This allows for the use of the dynamic programming algorithm due to Kanazawa (1988) which determines the maximizer of (10) using $\mathcal{O}(k_n^3)$ floating-point operations.

Although the dynamic programming algorithm succeeds in identifying the MAP partition within a reasonable time budget for smaller datasets, the $\mathcal{O}(k_n^3)$ runtime is prohibitive for large values of k_n. It is generally desirable to have k_n increase at a rate close to n, as otherwise $\mathcal{T}_n$ can end up being too coarse, and the resulting histogram may miss key details in the data that would have been shown if a high-resolution grid was used. To reduce the computational burden, we have implemented a greedy search heuristic similar to that of Rozenholc et al. (2010) to construct a reduced grid $\mathcal{Q}_n \subseteq \mathcal{T}_n$ whose cardinality q_n grows more slowly with n. After computing $\mathcal{Q}_n$, we run the dynamic programming algorithm to maximize $p_n(\mathcal{I} \mid \boldsymbol{x})$ over $\mathcal{P}_{\mathcal{Q}_n}$, and take $\widehat{\mathcal{I}}$ as the resulting maximizer. Combining these two heuristics allows us to let k_n increase at a rate close to n, while retaining a fast implementation. Although this approach is not guaranteed to find the exact maximizer of $p_n(\mathcal{I} \mid \boldsymbol{x})$ over $\mathcal{T}_n$, we have found that in practice the chosen partition typically yields a histogram close to the one based on the true MAP in terms of Hellinger loss in cases where exact evaluation is feasible.

3.3. Prior elicitation and practical implementation

The model specified by (4) leaves the statistician with some flexibility through the choice of the prior for k and the hyperparameter $\boldsymbol{a}$. To have a fully automatic procedure, however, default choices for $p_n(k)$ and $\boldsymbol{a}$ are needed, which we provide in this section.

As prior for k, we have found that in practice using the uniform prior on $\{1, 2, \ldots, k_n\}$ with $k_n = \lceil 4n/\log^2(n) \rceil$ appears to work well compared to other alternatives in simulations, as documented in Appendix A.1. As such, we have used this uniform prior for k as part of our default procedure.

The hyperparameters $\boldsymbol{a}$ of the Dirichlet prior for $\boldsymbol{\theta} \mid \mathcal{I}$ can be chosen to center the prior distribution of f on a particular reference density. If g_0 is a initial guess for the density f, we let

$$a_j = a \int_{I_j} g_0(x)\mathrm{d}x, \quad \mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}, \quad j = 1, 2, \ldots, k,$$

where $a > 0$ is a constant. Conditional on the partition $\mathcal{I}$, we see that the prior mean of f is

$$\mathbb{E}_{\boldsymbol{\theta}}\{ f(x) \mid \mathcal{I} \} = \sum_{j=1}^{k} \frac{\mathbb{E}_{\boldsymbol{\theta}}\{\theta_j\}}{|I_j|} \mathbb{1}_{I_j}(x) = \sum_{j=1}^{k} |I_j|^{-1} \int_{I_j} g_0(y)\mathrm{d}y \, \mathbb{1}_{I_j}(x),$$

which is the best piecewise constant approximation of the density g_0 based on the partition $\mathcal{I}$, in the sense that it minimizes the $\mathbb{L}_2$ distance to g_0 among all piecewise constant densities on $\mathcal{I}$. To guarantee that this choice of a_j always yields a proper posterior for $\boldsymbol{\theta} \mid \boldsymbol{x}, \mathcal{I}$, we must ensure that $a_j > 0$, which is true if the density g_0 is chosen to be positive almost everywhere in $[0, 1]$. If prior information is not available, we can take g_0 equal to the uniform distribution in the unit interval, leading to the non-informative choice $a_j = a|I_j|$ for all j. We have made the choice $g_0 = \mathbb{1}_{[0,1]}$ the default in the available software implementation of the random irregular histogram, and we have adopted this choice in all subsequent simulations and applications of the histogram method in this paper.

The choice of the parameter a is more delicate, as it is difficult to directly assess its impact on the chosen partition $\widehat{\mathcal{I}}$. Through (7), smaller values of a yield Bayes estimates $\hat{\theta}_j$ which are closer to the maximum likelihood estimate of θ_j, depending more on the data and less on the prior mean. However, if a is small relative to n its effect on the chosen partition is typically of much greater importance than its influence on the estimated bin probabilities, since the data-based part of the estimate tends to dominate for moderate values of n, which is confirmed by the simulations in Appendix A.2. Based on the results from these preliminary simulations, we suggest $a = 5$ as a reasonable default choice for a.

As presented, the random irregular histogram method applies to densities supported on a known closed interval. To apply the methodology to data with unknown support, the support of the data has to be estimated. Our preferred method in this regard is to rescale the data to the unit interval via the affine transformation $z_i = (x_i - x_{(1)})/(x_{(n)} - x_{(1)})$, apply the Bayesian histogram method to the z_i and rescale the density estimates to the original scale, which is equivalent to the approach taken by most histogram procedures, see e.g. Birgé and Rozenholc (2006). For a discussion of other approaches to estimating the support of a probability distribution for the purposes of density estimation, we refer the reader to Section 1.4 of Guan (2016) and the references therein.

Another issue to consider is the construction of the grid $\mathcal{T}_n$. One possibility is to use a fine, regular mesh. Data-based grids are also possible; one such approach is to construct $\mathcal{T}_n$ from the sample quantiles, so that $\tau_{n,j} = \widehat{Q}_n(j/k_n)$ for $j = 1, 2, \ldots, k_n - 1$ where $\widehat{Q}_n(q)$ is the q-quantile of the sample $\boldsymbol{x}$. A further data-based option is to use the order statistics of the data as possible cut points. All of these approaches have their advantages and disadvantages, and the best choice of grid is ultimately dependent on the density we are trying to estimate; see e.g. Rozenholc et al. (2010). As part of our default procedure, we propose using a data-based mesh for $\mathcal{T}_n$, but note that in practical applications, one may also want to test other grids to check the sensitivity of the procedure to the choice of mesh.

4. Asymptotic results

To study the large-sample behavior of the Bayes histogram estimator (8), we use the theory developed by Ghosal et al. (1999, 2000) to study posterior consistency and posterior convergence rates, but with some modifications to account for the fact that our estimator is based on the MAP partition $\widehat{\mathcal{I}}$. In the following, we study the case where the data $\boldsymbol{x} = (x_1, \ldots, x_n)$ is an i.i.d. sample from a true density f_0 on the unit interval, although the statements of the following Theorems remain valid more generally for densities with known compact support.

4.1. Consistency

Our first result shows that the Bayes histogram estimator is Hellinger consistent under very general conditions on the prior and the true density f_0. The proof of Theorem 1 is given in Appendix B.2.

Theorem 1. *Suppose that f_0 is a probability density on the unit interval satisfying $\int_0^1 \{f_0(x)\}^r \mathrm{d}x < \infty$ for some $r \in (1, 2]$, and suppose that the random histogram prior sequence $\{P_n\}_{n\in\mathbb{N}}$ satisfies the following conditions:*

(*i*) *The sequence of prior probability mass functions for k, $p_n(k)$, is fully supported on $\{1, 2, \ldots, k_n\}$, where $k_n \to \infty$ and $k_n = o(n)$. Moreover, we have uniformly for all $k = 1, 2, \ldots, k_n$ and n,*

$$\log p_n(k) = o(n).$$

(*ii*) *Conditional on the partition $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$, the prior for $\theta \mid \mathcal{I}$ is $\mathrm{Dir}(\boldsymbol{a})$ where $0 < a_j \le \Sigma$ for a constant $\Sigma > 0$.*

(*iii*) *The mesh sequence $\mathcal{T}_n$ satisfies*

$$\max_{1\le j\le k_n} \{\tau_{n,j} - \tau_{n,j-1}\} = o(1).$$

Then the Bayes histogram estimator $\widehat{f}_{\widehat{\mathcal{I}}}$ satisfies

$$\mathbb{E}_{\boldsymbol{x}\sim f_0}\left\{ d_H^2\left(f_0, \widehat{f}_{\widehat{\mathcal{I}}}\right)\right\} = o(1).$$

The first and third condition ensure that the prior eventually assigns positive mass to an arbitrarily fine grid, which is needed to approximate f_0 in the Hellinger sense. The second condition is fulfilled if, for instance, $a_j = a\int_{I_j} g_0(x)\mathrm{d}x$ for a constant $a > 0$ and a strictly positive density g_0, or if the a_j are all equal to some fixed, positive number. Theorem 1 mirrors the result by Barron et al. (1999, Section 3.1), who studied posterior consistency for a regular random histogram prior, showing that the posterior distribution is consistent under the slightly stronger condition that f_0 is bounded.

4.2. Convergence rate

To establish a convergence rate result, we need to make some additional smoothness assumptions on the true density f_0. A function $f_0 : [0, 1] \to \mathbb{R}$ is said to be α-Hölder continuous for $\alpha \in (0, 1]$ if there exists a constant $L_0 > 0$ such that $|f_0(x) - f_0(y)| \le L_0|x - y|^\alpha$ for all $x, y \in [0, 1]$. The parameter α determines the smoothness of the function f_0, with larger values of α corresponding to smoother functions. The Hölder exponent relates directly to the bias in the approximation of f_0 by a piecewise constant function on the partition $\mathcal{I}$ (cf. Lemma 4 in the appendix), with larger exponents yielding better approximations for a given number of bins, leading to a faster convergence rates as a result.

Theorem 2. *Suppose that $f_0 : [0, 1] \to \mathbb{R}$ is an α-Hölder continuous density which is lower bounded away from 0. Suppose further that the random irregular histogram prior sequence $\{P_n\}_{n\in\mathbb{N}}$ satisfies the following.*

(*i*) *The sequence of prior probability mass functions for k, $p_n(k)$, is fully supported on $\{1, 2, \ldots, k_n\}$, where $n^\gamma = \mathcal{O}(k_n)$ for any $\gamma \in (0, 1)$ and $k_n = \mathcal{O}(n)$. Moreover, there are positive constants D_1, D_2, d_1, d_2 such that for all $k \le k_n$,*

$$D_1 \exp\big(-d_1 k \log(k)\big) \le p_n(k) \le D_2 \exp\big(-d_2 k \log(k)\big).$$

(*ii*) *Conditional on the partition $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$, the prior for $\theta \mid \mathcal{I}$ is $\mathrm{Dir}(\boldsymbol{a})$ where $\sigma k^{-1} \le a_j \le \Sigma$ for all j and two constants $\sigma, \Sigma > 0$.*

(*iii*) *The mesh sequence $\mathcal{T}_n$ satisfies*

$$\max_{1\le j\le k_n} \{\tau_{n,j} - \tau_{n,j-1}\} = \mathcal{O}\big(k_n^{-1}\big),$$

$$\min_{1\le j\le k_n} \{\tau_{n,j} - \tau_{n,j-1}\} \ge B k_n^{-1}, \text{ for some } B > 0.$$

Then for $\epsilon_n = \{n/\log(n)\}^{-\alpha/(2\alpha+1)}$, *the Bayes histogram estimator satisfies*

$$\mathbb{E}_{\boldsymbol{x}\sim f_0}\left\{ d_H^2\left(f_0, \widehat{f}_{\widehat{I}}\right)\right\} = \mathcal{O}(\epsilon_n^2).$$

The proof of Theorem 2 can be found in Appendix B.1. The first condition of the theorem is satisfied if, for instance, $p_n(k)$ is taken to be a Poisson distribution truncated to the interval $\{1, 2, \dots, k_n\}$. The third condition ensures that the mesh $\mathcal{T}_n$ does not deviate too much from a regular one, which guarantees that the deterministic component of the risk is at most of order $k^{-2\alpha}$ with respect to the squared Hellinger metric.

The rate obtained by the Bayes histogram estimator is equivalent to the minimax rate for α-Hölder continuous densities with respect to the squared Hellinger metric of $n^{-2\alpha/(2\alpha+1)}$ up to a logarithmic factor (Massart, 2007, Proposition 7.16). Moreover, the prior is rate-adaptive; it attains a near-optimal convergence rate without knowing the smoothness of the true density f_0. The convergence rate of the Bayes histogram estimator mirrors the posterior convergence rate obtained by Scricciolo (2007) for the corresponding regular histogram model and that of Liu et al. (2023) for Bayesian trees, with the minor difference that our rate is slightly better in terms of the logarithmic factor. The additional logarithmic factors that appear in the convergence rate are rather typical of Bayesian density estimators based on mixtures with Dirichlet-distributed weights; see, for example, the polygonally smoothed prior of Scricciolo (2007) and the extended Bernstein prior of Kruijer and van der Vaart (2008). Hall and Hannan (1988) consider a regular random histogram model with a uniform prior on the interval probabilities. Their main result implies that the histogram estimator based on maximization of $p_n(k)$ converges at rate $\{n/\log(n)\}^{-1/3}$ if f_0 is bounded away from 0 and has a uniformly continuous derivative f_0', which is equal to the convergence rate of Theorem 2 under the weaker assumption that $\alpha = 1$. In light of this result, we conjecture that the rate obtained in Theorem 2 is sharp for any $\alpha \in (0, 1]$.

Remark 3. The consistency and rate results presented here can also be extended to similar models for regular histograms, with minor modifications to the proofs. In particular, this means that the results of Theorems 1 and 2 are also valid for the histogram estimator proposed by Knuth (2019).

5. Simulation study

To showcase the performance of our proposed irregular histogram method, we have conducted two simulation studies. In the first simulation study, we compare our approach to multiple state-of-the-art automatic histogram procedures on the set of test densities D given in Fig. 1. In the second study, we further compare the automatic mode hunting capabilities of the random irregular histogram to other mode detection procedures from the statistics literature.

The set of test densities was chosen to reflect a variety of features with regard to skewness, tail behavior, modality and sharpness of individual peaks. Densities 10, 12 and 14–16 are taken from Rozenholc et al. (2010), densities 7–9 are from Marron and Wand (1992), while density 13 was suggested by Li et al. (2020).

5.1. Loss functions

Although classical loss functions such as the $\mathbb{L}_2$ or Hellinger metrics provide a natural way of measuring the discrepancy between a density estimate and the true density in a simulation study, they do not adequately assess the quality of a histogram procedure in terms of identifying key features of a density, such as modes (Denby and Mallows, 2009). To also account for this important aspect of histogram estimation, we have included the peak identification loss of Davies et al. (2009) as a performance metric.

Definition 1 (Davies et al., 2009). We say that a density f has a *finite mode* or *peak* at t if there is an interval[1] C such that t is the midpoint of C and $f(s) = c$ for all $s \in C$, and there exists a $\gamma > 0$ such that $f(s) < c$ for all $s \in C_\gamma \setminus C$, where $C_\gamma = \cup_{s\in C}\{x : |s - x| \le \gamma\}$. The density f is said to have an *infinite peak* or *mode* at t if either one-sided limit at t equals ∞.

Suppose that f_0 has l_0 modes $t_1, t_2, \dots, t_{l_0}$ and that the histogram estimate $\widehat{f}$ has modes at $s_1, \dots, s_l$. Given a tolerance vector $\boldsymbol{\delta} \in (0, \infty)^{l_0}$, we say that a peak of the histogram $\widehat{f}$ *matches the peak at* t_j *of* f_0 *at tolerance* δ_j if $\min_{1\le i\le l} |t_j - s_i| < \delta_j$. The peak s_i is said to be *spurious* if it does not match any peak of f_0. Letting $C = \sum_{j=1}^{l_0} \mathbb{1}_{[0,\delta_j)}(\min_i |s_i - t_j|)$ denote the number of correctly identified modes, we define the *peak identification* (PID) *loss* as the sum of the number of spurious and unidentified peaks, $\ell\left(f_0, \widehat{f}\right) = (l_0 - C) + (l - C)$.

The behavior of the PID loss depends on the choice of the tolerance parameter $\boldsymbol{\delta}$. To determine the tolerance δ_j of a finite peak at t_j of a density f_0, we have used the following approach: find the smallest $\gamma > 0$ such that

$$\frac{\int_{C_{j,\gamma}} \left|f_0(x) - \bar{f}_{0,j,\gamma}\right| \mathrm{d}x}{\int_{C_{j,\gamma}} f_0(x)\mathrm{d}x} > 0.2, \quad C_{j,\gamma} = [t_j - \gamma, t_j + \gamma],$$

where we have defined $\bar{f}_{0,j,\gamma} = |C_{j,\gamma}|^{-1} \int_{C_{j,\gamma}} f_0(x)\mathrm{d}x$. We then take $\delta_j = \gamma/2$ as our tolerance for the peak at t_j. This criterion essentially measures how well f_0 is approximated by a constant function around any given mode, reflecting the fact that we should expect higher precision when trying to determine the location of a sharp peak. We visually confirmed that this criterion yielded sensible and non-overlapping tolerance regions for each of the test densities in Fig. 1. For the infinitely peaked Chisq(1) we set the tolerance

[1] We include the case where C is a singleton set, which corresponds to the usual definition of a local maximum.

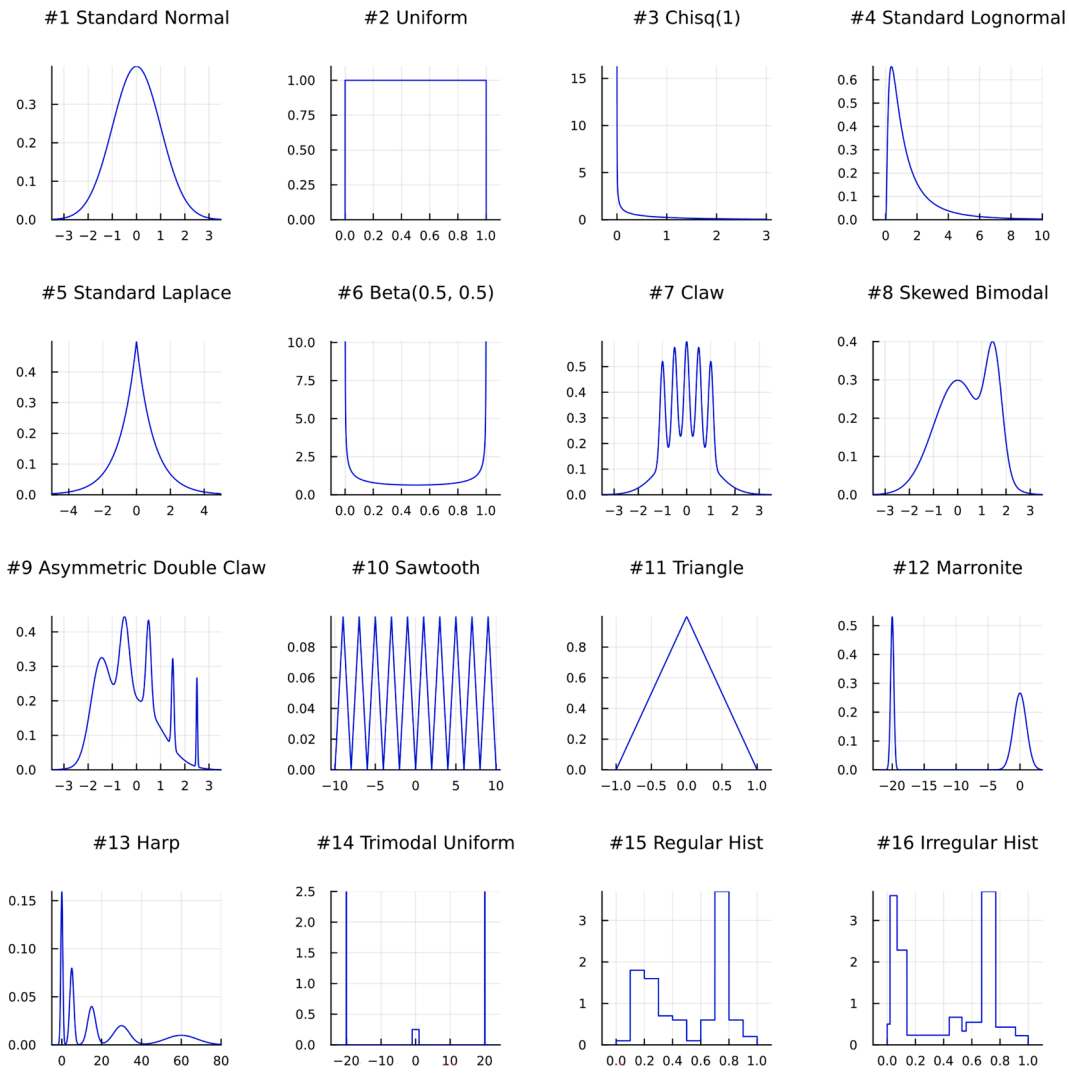

Fig. 1. Test densities included in the simulation study.

region of the peak at 0 to $[0, q_{0.1}]$, where q_α denotes the α-quantile of this distribution, and for the $\mathrm{Beta}(0.5, 0.5)$, we used $[0, q_{0.1}]$ and $[q_{0.9}, 1]$ for the peaks at 0 and 1, respectively.

One objection to the use of the PID loss as a measure of mode detection capabilities is that it only captures a histogram procedure's ability to detect modes automatically and that no statistician would interpret any mode of a histogram as one corresponding to a mode in the true density. Scott (1979) remarks that many data analysts prefer the use of relatively small bin widths when drawing a regular histogram, leaving the final smoothing to be done by eye. Although we agree that this is a valid objection to the use of the peak identification loss in the regular case, the subjective manner in which the number and the location of modes have to be located is somewhat unsatisfactory. In contrast, a procedure that does well with respect to the PID loss can identify modes automatically, which is a major reason for the introduction of irregular histograms to begin with.

In addition to the PID loss, we have also included results for the Hellinger and $\mathbb{L}_2$ metrics in our comparison with other histogram methods. Since closed-form expressions for the latter two loss functions are generally not available, Gauss quadrature was used to approximate the loss, performing the numeric integration piecewise over each interval where both the histogram estimate and the true density are continuous.

5.2. A comparison with other histogram methods

In our first simulation study we have chosen to focus on histogram methods that have performed well in past simulation studies, see Birgé and Rozenholc (2006), Davies et al. (2009). In addition, the method of Knuth (2019) and the stochastic complexity approach of Hall and Hannan (1988) were also included, as these effectively correspond to a regular equivalent of the random irregular histogram.

Table 1
Overview of the methods included in the simulation study.

Type	Method	Abbreviation	Description
Regular	Wand's 2-step rule	Wand	Wand (1997)
	AIC	AIC	Taylor (1987)
	BIC	BIC	Davies et al. (2009)
	Birgé and Rozenholc's rule	BR	Birgé and Rozenholc (2006)
	Knuth's rule	Knuth	Knuth (2019)
	Stochastic Complexity	SC	Hall and Hannan (1988)
Irregular	Random irregular histogram	RIH	Section 3
	Penalty B in Rozenholc et al.	RMG-B	Rozenholc et al. (2010)
	Penalty R in Rozenholc et al.	RMG-R	Rozenholc et al. (2010)
	Taut string	TS	Davies and Kovac (2004)
	$\mathbb{L}_2$ cross-validation	L2CV	Rudemo (1982)
	Kullback–Leibler cross-validation	KLCV	Section 5.2

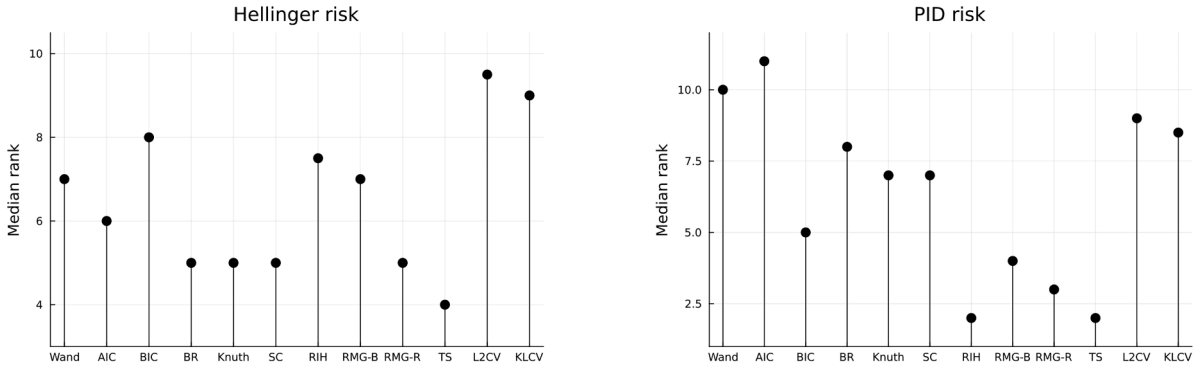

Fig. 2. Median ranks from the simulation study in terms of Hellinger risk (left) and PID risk (right).

The Essential Histogram of Li et al. (2020) was excluded due to the high computational cost of its available software implementation, which made it impractical for the largest sample sizes considered in this study. We have also included an irregular histogram method based on Kullback–Leibler cross-validation, an approach that, to our knowledge, has so far not been explored for irregular histograms. This approach chooses the partition which maximizes $\mathrm{CV}_n(\mathcal{I}) = \sum_{j=1}^{k} N_j \log(N_j - 1) - \sum_{j=1}^{k} N_j \log|I_j|$ for all partitions with $N_j \geq 2$ for all j, and is an irregular counterpart to the cross-validation criterion proposed by Hall (1990). A detailed list of the included methods and their corresponding abbreviations can be found in Table 1. C.1 provides further details on the implementations used for each procedure.

To judge the quality of the different histogram procedures, we used Monte Carlo simulations to estimate the risks of each method for a set of test densities $f_0 \in D$ with respect to the Hellinger, $\mathbb{L}_2$ and PID losses, using the sample averages

$$\widehat{R}_n(f_0, \widehat{f}_m) = \frac{1}{B} \sum_{b=1}^{B} \ell(f_0, \widehat{f}_m), \quad (n, m, f_0) \in \mathcal{N} \times \mathcal{M} \times D, \tag{11}$$

where $\mathcal{N} = \{50, 200, 1000, 5000, 25000\}$ denotes the set of sample sizes used in the study and $\mathcal{M}$ is the set of model indices. The Monte Carlo average in (11) was computed using $B = 500$ replications.

5.2.1. Results

Tables showing the estimated risks $\widehat{R}_n(f_0, \widehat{f}_m)$ for each of the three loss functions can be found in Appendix C.2, along with boxplots comparing the performances of the different methods included in the study. To provide a more compact summary of the results, we ranked each method according to their risk for each combination of loss function, density and sample size, and computed the median rank by method over D and $\mathcal{N}$. The resulting medians for the Hellinger and PID losses are shown in Fig. 2.

Although the relative performance of the different procedures depends on the true density at hand, our results indicate some broader trends among the regular and irregular histogram procedures.

- Whether or not the Hellinger risk favors regular histogram methods varies based on the true density. For spatially homogeneous densities such as the standard normal or skewed bimodal densities, there is little benefit to using an irregular grid to begin with, resulting in generally smaller risks for regular histograms. For the heavy-tailed $\mathrm{Chisq}(1)$ and standard lognormal densities and the infinitely peaked $\mathrm{Beta}(1/2, 1/2)$, the irregular methods tend to perform better.
- With the exception of the L2CV and KLCV criteria, all the irregular histogram procedures included in the study outperformed their regular counterparts in terms of automatic mode identification for almost every combination of f_0 and n. This is apparent from

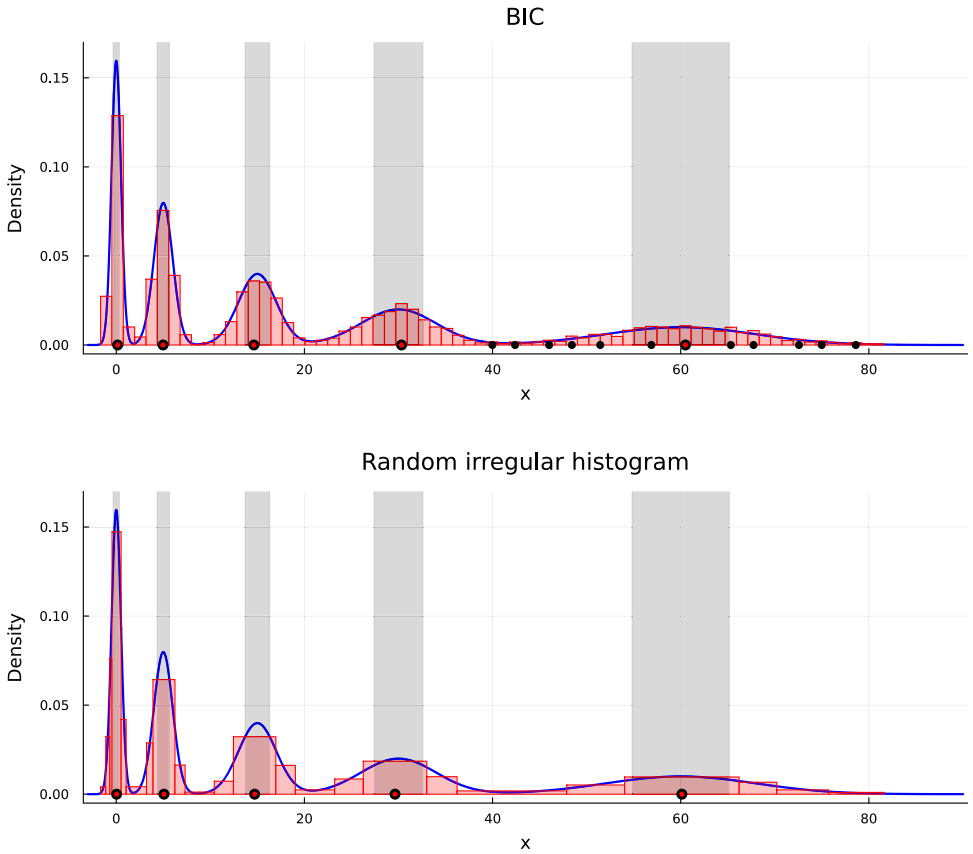

Fig. 3. Illustration of automatic mode detection and the PID loss for a $n = 5000$ sample from the Harp density.

the right panel of Fig. 2; four of the irregular procedures produce much smaller median ranks than any of the other methods. In addition, we note that these four irregular methods exhibit lower PID risks as the sample size increases, a pattern that is often reversed for regular histograms.

Regarding the relative performance of the individual methods, our main findings were as follows:

- The two cross-validation based procedures perform much worse in terms of mode identification than the other methods, and yielded the highest $\mathbb{L}_2$ and Hellinger losses for many densities. Plots of the histograms resulting from these procedures revealed that they still have a tendency to choose partitions with rather narrow bins even for smooth densities, despite the fact that a minimum bin width was enforced.
- Among the regular histogram procedures, the Bayesian criteria Knuth and SC are the overall best performers with respect to the Hellinger metric for the smallest sample sizes, while the criterion of Birgé and Rozenholc (2006) does better for larger samples. The BIC-based histogram is the clear winner in terms of PID risk, but does relatively poorly with respect to the two classical loss functions.
- The Taut String method of Davies and Kovac (2004) is overall the best performer in terms of Hellinger and $\mathbb{L}_2$ risk out of the irregular histogram methods. The two criteria proposed by Rozenholc et al. (2010) perform quite similarly with respect to the Hellinger loss, while penalty RMG-R tends to produce lower $\mathbb{L}_2$ risks, especially for smaller sample sizes. The random irregular histogram method performs comparably to the other non-cross-validation irregular procedures with respect to classical loss functions, and does particularly well at mode detection, as evidenced by its low PID risks across the board.

To illustrate the advantage of irregular histograms over their regular counterparts in automatic mode detection, we generated a sample of size $n = 5000$ from the Harp density in Fig. 1. The Harp density has multiple modes at different scales, making it difficult for regular histogram procedures to correctly identify the number of modes and the location of all individual peaks simultaneously, even for larger samples. Fig. 3 shows the resulting histograms from fitting the random irregular histogram and the regular BIC histogram to the simulated sample, which was chosen to yield a PID loss close to the corresponding estimated risks in Table C.7. The tolerance region for each individual peak is shown in gray, with the red dots indicating the positions of the histogram modes that match the peaks of the true density, whereas the spurious modes are drawn in black.

The disadvantages of using a fixed global bin width in this case are immediately apparent from Fig. 3; the bin width chosen by the BIC histogram results in too much smoothing in the region surrounding the leftmost mode, while the density estimate is far too

Table 2
Estimated PID risks and median inference times in seconds for each mode detection method.

		Estimated PID risk					Median inference time (seconds)				
Density	n	RIH	SI	FM	WJ	GMM	RIH	SI	FM	WJ	GMM
1	200	**0.005**	0.01	0.085	0.295	0.015	**0.005**	1.11	1.64	0.19	0.12
	600	**0.0**	0.005	0.035	0.08	**0.0**	**0.081**	1.17	6.16	2.31	0.27
	2000	**0.0**	0.005	0.02	1.37	**0.0**	**0.066**	1.08	55.92	26.32	0.68
2	200	**0.01**	0.285	0.97	2.735	2.1	**0.005**	1.06	1.69	0.26	0.27
	600	**0.0**	0.34	1.295	9.055	3.39	**0.059**	1.08	11.445	2.955	1.58
	2000	**0.0**	0.435	1.765	45.105	5.255	**0.061**	1.19	116.01	47.285	8.555
4	200	**1.5**	2.0	*	6.03	2.785	**0.005**	1.135	10.07	0.1	0.08
	600	**1.17**	2.0	*	8.93	2.54	**0.06**	1.215	47.845	0.96	0.28
	2000	**0.43**	2.005	*	18.66	1.925	**0.06**	1.35	540.47	15.625	1.08
7	200	**4.195**	5.155	*	7.42	4.29	**0.005**	1.06	1.64	0.275	0.13
	600	4.085	5.0	4.86	10.33	**2.19**	**0.058**	1.09	52.815	3.135	0.35
	2000	0.205	4.655	4.81	6.04	**0.0**	**0.065**	1.14	544.895	15.515	0.9
8	200	1.135	1.285	0.82	1.985	**0.49**	**0.005**	1.08	3.055	0.19	0.12
	600	1.165	0.9	0.55	2.255	**0.07**	**0.058**	1.105	11.42	1.95	0.33
	2000	1.03	0.415	0.42	12.005	**0.0**	**0.063**	2.22	113.79	32.12	0.97
11	200	0.01	**0.0**	0.065	0.09	0.135	**0.005**	1.06	1.56	0.19	0.18
	600	**0.0**	0.01	0.07	0.25	0.29	**0.059**	1.08	5.67	1.745	0.55
	2000	**0.0**	0.005	0.03	**0.0**	0.135	**0.062**	1.09	53.945	26.645	1.5
13	200	1.005	4.54	5.14	4.285	**0.0**	**0.004**	2.75	15.17	0.09	0.05
	600	0.025	3.08	5.03	3.075	**0.005**	**0.055**	5.02	50.47	0.7	0.12
	2000	**0.0**	2.705	5.0	3.365	**0.0**	**0.06**	4.905	481.525	13.53	0.39
14	200	**0.0**	*	*	**0.0**	0.895	**0.004**	9.76	13.74	0.03	0.08
	600	**0.0**	*	*	**0.0**	4.315	**0.056**	9.84	46.42	0.22	0.29
	2000	**0.0**	*	*	**0.0**	5.54	**0.082**	15.4	695.645	4.205	1.19

coarse near the rightmost mode, resulting in a high number of spurious peaks near this mode and in the right tail of the density. In contrast, the irregular histogram is able to adapt the bin width locally, providing more smoothing in flatter regions of the density, which results in high-precision estimates of each of the five true modes. On the other hand, the Hellinger loss for each of the two procedures are 0.176 for the BIC and 0.272 for the random irregular histogram, favoring the regular histogram. Ultimately, whether one prefers the regular or irregular estimate in this case comes down to the analysts preference for low estimation risk or automatic mode detection.

5.3. A comparison with other mode identification methods

To further investigate how the random irregular histogram fares in terms of its mode hunting capabilities, we conducted a second simulation study where we included several non-histogram mode detection methods. In our simulations, we compared against the testing-based procedures of Silverman (1981) (SI) and Fisher and Marron (2001) (FM) and the Gaussian mixture model (GMM) method of Chacón (2019). In addition to the three aforementioned procedures, we included a third method, termed WJ, based on computing a kernel estimate of the first derivative of the density using the plug-in approach of Wand and Jones (1994) to select the bandwidth and taking the resulting maxima of the kernel density as estimates of the true mode locations. Further details on the software used to implement each method can be found in Appendix D.1. To assess the quality of each procedure, we computed Monte Carlo estimates of the PID risk on densities #1, #2, #4, #7, #8, #11, #13 and #14 in Fig. 1 for samples of size $n \in \{200, 600, 2000\}$. In addition, the median time used to estimate the mode locations was recorded for each method. All simulations were performed on a computer equipped with an Intel Core i7-1255U (12th Gen) processor running at 1.7 GHz and 16 GB of RAM under Windows 11. The results of the simulation study are shown in Table 2, where the best method for each combination of sample size and density is outlined in bold.[2]

The results can be summarized as follows:

- In terms of PID risk, the GMM and RIH procedures were the best performers and have the lowest risk for almost all combinations of densities and sample sizes. The GMM method does very well at identifying the modes of the normal mixture densities, but the performance is worse than the RIH method for the other densities, particularly for the uniform (#2) and the trimodal uniform (#14). The RIH approach performs well for both continuous and discontinuous densities, although it tends to lose out to the GMM method for the multimodal normal mixtures in smaller samples.

[2] The procedure for finding the locations of the modes occasionally failed for the SI and FM methods for certain combinations of densities and sample size, and the PID risk has been marked with an asterisk in those case.

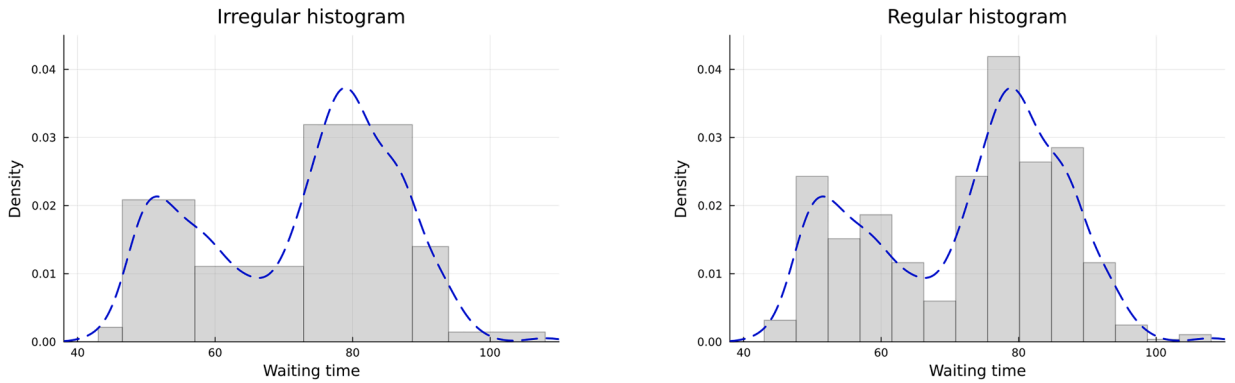

Fig. 4. Irregular (left) and regular (right) histograms fitted to the Old Faithful data. The dashed blue line shows the kernel estimate. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- The other methods generally performed worse than the RIH and GMM approaches across the board, although the testing-based SI and FM procedures show decent performance for continuous densities with few modes.
- The random irregular histogram method is by far the best performer in terms of inference time, and frequently outperforms the other methods by an order of magnitude or more. The FM and WJ methods both scale very poorly with increasing sample sizes. This is especially true for the FM method, for which single iterations could frequently take several minutes to complete.

6. Applications

In this section we apply the irregular histogram method proposed in this paper to two real-world datasets. For comparison, we also include results for the histogram method of Knuth (2019), as this method is a regular equivalent of the random irregular histogram.

6.1. Old Faithful geyser data

In our first application we consider the Old Faithful dataset from Härdle (1991), which consists of measured waiting times in minutes between the consecutive eruptions of the Old Faithful geyser in Yellowstone National Park. Previous analyses of this dataset have typically shown a clear bimodal structure; see e.g. Venables and Ripley (1999) and Kwasniok (2021) for analyses using several different density estimation procedures.

We fitted both the irregular random histogram and the regular histogram of Knuth (2019) to the dataset, estimating the support of the histogram for both procedures. For comparison, we also computed a kernel density estimate using the `density()` function from base R with the bandwidth selector of Sheather and Jones (1991). The resulting density estimates are shown in Fig. 4.

The irregular histogram method yields a smooth density estimate, using a relatively small number of bins, and displays a clear bimodal structure. In contrast, the regular histogram procedure chooses a fairly large number of bins in this case, resulting in a much rougher appearance. The bimodal structure is also less apparent in this case, as there are two peaks near each of the two modes in the kernel estimate. Overall, the irregular histogram displays a greater degree of agreement with the kernel estimate than Knuth's method.

6.2. Multiple hypothesis testing

In a multiple hypothesis testing scenario, statistical procedures often aim at controlling the false discovery rate (FDR), the expected number of true null hypotheses among the rejected ones (Benjamini and Hochberg, 1995). It is therefore of great interest to estimate the FDR based on the observed test statistics. This task is often carried out by using the p-values of the individual tests to construct an estimate of the proportion of true null hypotheses, from which the FDR can be estimated, for instance via the approach of Storey and Tibshirani (2003). A model-based approach to estimating the true null proportion assumes that the p-values are distributed according to a mixture density of the form

$$f(p) = \pi_0 \mathbb{1}_{[0,1]}(p) + (1 - \pi_0)h(p), \tag{12}$$

where π_0 is the theoretical proportion of true null hypotheses and h is the density of the false null hypotheses, assumed to have most of its mass near 0 and $h(1) = 0$. A nonparametric approach to estimating π_0 is to use the observed p-values to compute a nonparametric density estimate $\widehat{f}$ and take $\widehat{\pi}_0 = \widehat{f}(1)$ as the estimate of π_0, which is reasonable in light of our assumptions that imply $\pi_0 = f(1)$. For some previous proposed nonparametric approaches to estimating the density f specifically for this purpose, see e.g. Langaas et al. (2005), Guan et al. (2008) and the references therein. Our irregular histogram procedure seems like an attractive method for estimating f in this case, as the support the distribution of the p-values is known. This avoids having to use boundary corrections,

Table 3
Average RMSEs for $\beta = 4$ and $\beta = 10$.

		$\beta = 4$		$\beta = 10$	
π_0	n	Irregular	Regular	Irregular	Regular
0.5	200	0.118	**0.078**	**0.065**	0.117
	1000	0.059	**0.043**	**0.031**	0.063
	5000	0.033	**0.025**	**0.015**	0.034
0.8	200	0.147	**0.109**	**0.059**	0.099
	1000	0.06	**0.038**	**0.038**	0.062
	5000	0.033	**0.022**	**0.017**	0.034
0.95	200	**0.052**	0.057	**0.053**	0.057
	1000	0.051	**0.05**	**0.044**	0.051
	5000	0.04	**0.027**	**0.015**	0.023

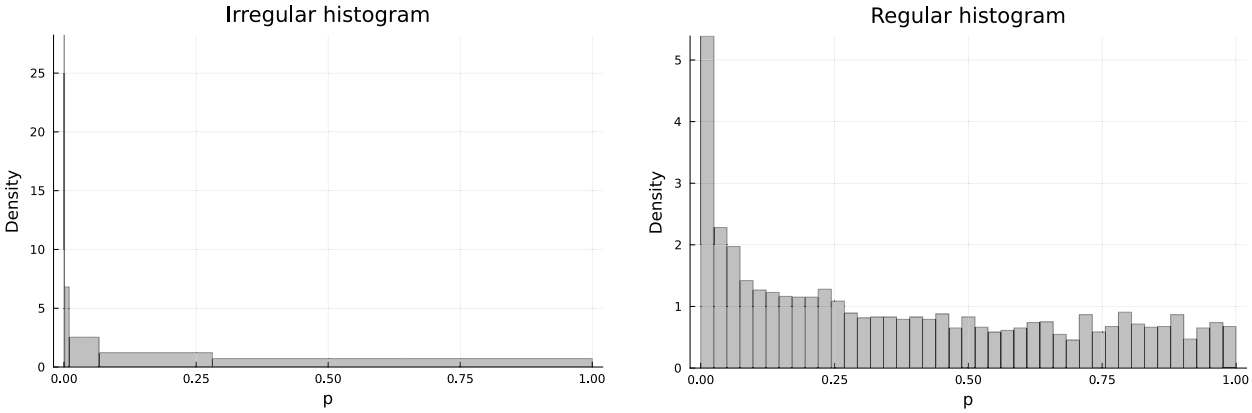

Fig. 5. Irregular (left) and regular (right) histograms fitted to the Hedenfalk et al. (2001) data.

which are necessary for some classical density estimators such as kernel-based ones, since we are primarily concerned with estimating a value at the boundary of the support of the density.

To investigate the performance of our irregular histogram procedure in this setting, we conducted a small simulation study based on the model in (12), before applying it to a real dataset. In our simulations, we took the probability density h of the alternative hypotheses equal to the $\text{Beta}(1, \beta)$ density for values of β in the range $[2, 10]$. We varied the parameter $\pi_0 \in \{0.5, 0.8, 0.95\}$, and considered samples of size $n \in \{200, 1000, 5000\}$. To measure the quality of the procedure, the root mean squared error (RMSE) of $\widehat{\pi}_0$ was estimated by $B = 500$ Monte Carlo samples. For comparison, we included estimates of π_0 based on using the automatic histogram procedure of Knuth (2019). The average RMSEs from the Monte Carlo simulations are shown in Tables 3, with the lowest risk in each scenario outlined in boldface.

The relative performance of the two procedures depend heavily on the value of β in the simulations, with Knuth's method performing better for $\beta = 4$ while the irregular procedure wins out for $\beta = 10$. This is related to the amount of smoothing introduced by the two procedures in the tail of the density. The tendency of the irregular procedure to select a large bin in regions where the density is very flat worsens the performance for the $\text{Beta}(1, 4)$, as this density still assigns a considerable amount of mass to $p > .5$, resulting in greater bias. In contrast, for $\beta = 10$ the right tail of h decays very quickly, and the bias of the irregular histogram procedure becomes negligible. In this case, the variance reduction brought about by the greater amount of smoothing leads to better performance in terms of RMSE.

To end the section, we consider an example from breast cancer research analyzed in Storey and Tibshirani (2003). The study of Hedenfalk et al. (2001) attempted to identify differences in gene expressivity in individuals with BRCA1- and BRCA2-mutation-positive tumors based on microarray measurements. Storey and Tibshirani (2003) use permutation tests to compute p-values for the measured differences in gene expressivity, leaving us with a sample of $n = 3170$ p-values. The resulting density estimates from applying the irregular histogram procedure and the regular histogram of Knuth are shown in Fig. 5. The irregular histogram displays a very pronounced peak near 0, owing to the fact that a non-negligible proportion of the p-values in this dataset concentrate in a narrow interval around the origin. Despite having a very narrow width, the leftmost interval contains a total of 73 observations, accounting for more than 2 % of the total sample. Although the regular histogram also shows a distinct mode near 0, it is much less pronounced than that of the irregular histogram. The regular histogram also appears to undersmooth significantly in the right tail, as there are many sporadic bumps here that are unlikely to reflect any real structure in the data. The resulting estimates of π_0 are $\widehat{\pi}_{0,\text{irr}} = 0.714$ and $\widehat{\pi}_{0,\text{reg}} = 0.675$ for the irregular and regular histogram models, respectively. For comparison, we also estimated π_0 using the convex, decreasing density estimator in Langaas et al. (2005), and the local false discovery rate method of Phipson (2013).

This yielded estimates of $\hat{\pi}_{0,\mathrm{cd}} = 0.671$ for the former method and $\hat{\pi}_{0,\mathrm{lfdr}} = 0.744$ for the latter, indicating that the results from both of the histogram methods are quite reasonable in this case.

7. Discussion

We conclude the article with a brief discussion of possible extensions of the random irregular histogram, and how the ideas presented here can be used to construct computationally tractable Bayesian models for regression and hazard rate estimation.

The results of our simulation study show that the random irregular histogram model excels at mode detection for larger samples, a feat that cannot be achieved by regular histogram procedures designed for optimal performance with respect to classical estimation error criteria (Scott, 1992). However, the improved performance of regular histograms in terms of classical loss functions for spatially homogeneous densities has led some authors to argue for their continued use, cf. Birgé and Rozenholc (2006). Rozenholc et al. (2010) propose constructing both a regular and an irregular histogram based on a penalized likelihood, and the resulting histogram is chosen to be the one with the lowest penalized likelihood of the two. Although this approach often succeeds in reducing the estimation error compared to an irregular procedure, the resulting density estimates still suffer from poor automatic mode detection whenever the regular histogram is chosen. Another possible approach in this regard would be to use a prior distribution (or penalty in the penalized likelihood case) such that, for a given value of k, the partitions which have more regular bin widths are assigned more prior probability or a smaller penalty. This could lead to an improved estimation error for spatially homogeneous densities, while possibly keeping the advantage of an irregular procedure in terms of mode identification.

Although we have at present focused on density estimation in the univariate case, the random irregular histogram model can be extended to multivariate data as well. For practical applications, the two-dimensional case is of particular interest, as one can graphically display the computed histograms with relative ease. A possible extension of our method in this case is based on the following model for the density f:

$$f(x \mid \boldsymbol{\theta}, \mathcal{I}, \mathcal{J}) = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\theta_{i,j}}{|\mathcal{I}_i| \times |\mathcal{J}_j|} \mathbb{1}_{\mathcal{I}_j \times \mathcal{J}_j}(\boldsymbol{x}), \quad \boldsymbol{x} \in [0,1]^2, \tag{13}$$

where $\mathcal{I} = (\mathcal{I}_1, \ldots, \mathcal{I}_k)$ and $\mathcal{J} = (\mathcal{J}_1, \ldots, \mathcal{J}_m)$ are two (random) interval partitions of $[0,1]$ and $\boldsymbol{\theta} \in \mathcal{S}_{km}$. With a km-dimensional Dirichlet prior on $\boldsymbol{\theta}$, we have conjugacy conditional on the chosen partition, and we obtain a closed-form expression for the posterior probability of $\mathcal{I}, \mathcal{J}$, similar to the one-dimensional case. However, in this setting, the problem of computing the MAP partition is more difficult due to the increased dimension of the solution space. We leave it as an interesting future research question to come up with a clever way of approximating the maximum a posteriori partition in this case.

It would also be of interest to explore the asymptotic properties of the multivariate random irregular histogram. Liu et al. (2023) obtained a posterior convergence rate independent of the dimension of the data under a spatial sparsity constraint for a Bayesian tree method based on binary partitions. In light of these findings, an interesting avenue for further research is to investigate whether similar theoretical results hold for the random irregular histogram in the multivariate case.

In this article, we have concerned ourselves with the task of finding the best piecewise constant density estimate, but many of the ideas presented here can be extended to other areas of statistics such as semiparametric regression and hazard rate estimation. In the semiparametric regression setup, we could consider the model $y_i = m(x_i) + \epsilon_i$ where $m : [0,1] \to \mathbb{R}$ is an unknown regression function and the ϵ_i are independent and $\mathrm{Normal}(0, \tau^{-2})$. A possible choice of nonparametric prior for m is a piecewise constant model, where m takes the fixed value m_j on $\mathcal{I}_j$. In this case, letting $m_j \mid \mathcal{I} \sim \mathrm{Normal}(\mu, \beta^{-2})$ leads to a conjugate model for given $\mathcal{I}$, which results in an analytical expression for the marginal likelihood. The computational techniques outlined in this paper can then be used to find the MAP partition $\hat{\mathcal{I}}$ in this model, and the corresponding estimated regression function is a Bayesian variant of the regressogram; the regression equivalent of the histogram (Klemelä, 2012). In the hazard rate case, Gamma priors can be used to yield a similarly computationally tractable expression for the posterior of $\mathcal{I}$.

Acknowledgements

We thank the anonymous reviewers for insightful comments that greatly helped improve the presentation of this article.

Appendix A. Sensitivity to choices of prior parameters

To investigate the sensitivity of the irregular random histogram procedure to the choice of prior distribution, we conducted some small numerical experiments. In each experiment, we estimated the Hellinger risk for a given configuration of the prior parameters by generating random samples $\boldsymbol{x}^{(b)}$ of size $n \in \mathcal{N} = \{100, 1000, 10000\}$ from three test densities for $b = 1, 2, \ldots, B$. The risks were estimated by a simple Monte Carlo average,

$$\hat{R}_n(f_0, j) = \frac{1}{B} \sum_{i=1}^{B} d_H\left(f_0, \hat{f}_j\right), \quad (n, m, f_0) \in \mathcal{N} \times \mathcal{M} \times \mathcal{D}', \tag{A.1}$$

where $\mathcal{D}'$ denotes a set of test densities, $\mathcal{M}$ is the model index set for the different prior configurations and $\hat{f}_j$ is the Bayes histogram estimate when prior j is used. In our simulations, we took $B = 500$. The chosen test densities were the $\mathrm{Gamma}(3,3)$, the $\mathrm{Beta}(3,3)$ and the t-distribution with 3 degrees of freedom. For the $\mathrm{Beta}(3,3)$ we took $[0,1]$ to be the support of the density estimate, while for the

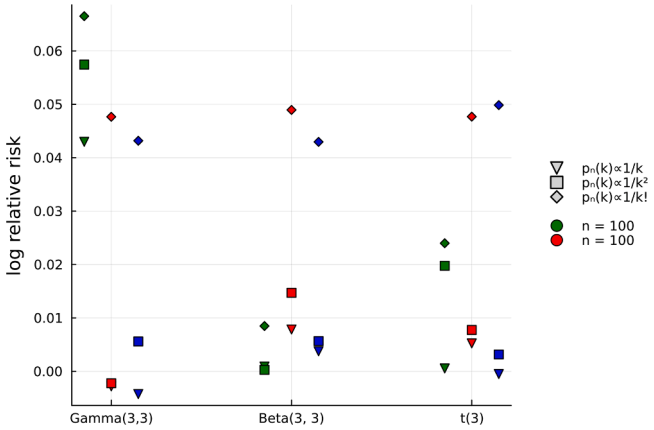

Fig. A.6. Logarithm of risk relative to the uniform prior $p_n(k) \propto 1$.

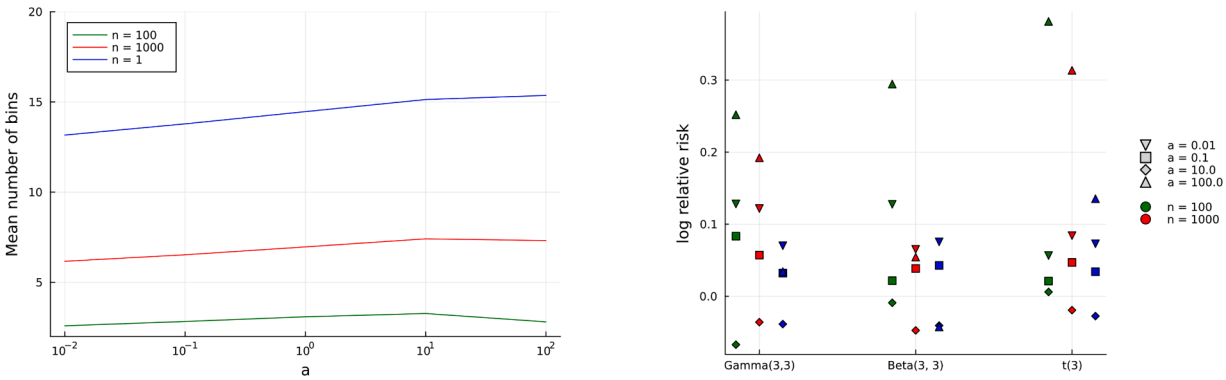

Fig. A.7. Left: Mean value of k aggregated over all test densities by sample size. Right: Logarithm of risk relative to $a = 1$.

Gamma$(3, 3)$ we computed the histogram estimate on $[0, x_{(n)}]$. For the t-distribution the left- and rightmost endpoints of the histogram were set equal to the minimum and maximum of the sample, respectively.

A.1. Prior for the number of bins

To start, we investigated the effect of different prior distributions for the number k of bins. In the experiment, the value of the Dirichlet concentration parameter a was kept fixed at 1. We tested the prior distributions $p_n(k) \propto 1/k^m$ for $m \in \{0, 1, 2\}$ and the Poiss(1) distribution with $p_n(k) \propto 1/k!$ as this prior fulfills the conditions of Theorem 2. In order to present the results, we computed the logarithm of the risk in (A.1) relative to that of the uniform prior. The resulting log-relative risks are shown in Fig. A.6.

The uniform prior appears to yield the best performance for most combinations of the sample size and true density, with the exception of the Gamma$(3, 3)$ for $n = 10^4$. Both $p_n(k) \propto 1/k$ and $p_n(k) \propto 1/k^2$ generally perform similarly and not much worse than the uniform. The Poisson prior does the worst in terms of Hellinger risk, indicating that it puts too high a penalty on larger values of k compared to what is optimal in this case.

A.2. Choice of the Dirichlet concentration parameter

To study the effect of the concentration parameter a on the quality of the resulting density estimates, a similar numerical experiment to that in the previous subsection was conducted. We tested the values $a \in \{0.01, 0.1, 1, 10, 100\}$. Across all values of a, we kept the prior for k fixed to a uniform one. To investigate the effect of the concentration parameter on the number of bins in the MAP partition, we computed the mean number of bins chosen in the simulations for each choice of a. Furthermore, the relative Hellinger risk of the procedure was estimated for different values of a, using $a = 1$ as a reference. The results are shown in Fig. A.7.

The mean number of bins chosen by the procedure tends to increase along with the concentration parameter a. This illustrates the point made in Section 3.3; although smaller values of a yield estimates of the bin probabilities that assign more weight to the data through (7), its effect on the posterior of $\mathcal{I}$ cannot be neglected, and it can thus be beneficial to set the concentration parameter to a larger value in practice. We also plotted a against the mean number of bins chosen separately for each density, but this showed a very similar picture to the aggregated results and is therefore omitted.

In terms of risk, the procedure with $a = 10$ performs the best in all cases. Its lower risk compared to $a = 1$ can likely be attributed to its tendency of choosing a greater number of bins. The smallest values of the concentration parameter $a = 0.01$ performs quite poorly, and the choice $a = 0.1$ is only slightly better. The choice $a = 100$ does very poorly for the smallest sample sizes. This is rather unsurprising in light of (7), as for $n = 100$ the prior mean ends up contributing significantly to the interval probability estimates. As the sample size increases, the influence of the prior mean weakens, and the relative risk of this procedure improves considerably. Overall, it seems that a choice of $a \in [1, 10]$ should perform reasonably well for the sample sizes considered here.

Appendix B. Proofs

Our proofs are mainly based on the theory developed in Ghosal et al. (1999, 2000). These are based on the use of covering numbers, which provide a measure of the complexity of a given model. For a metric space $(\mathcal{X}, d)$ the covering number $N(\epsilon, \mathcal{Y}, d)$ of a set $\mathcal{Y} \subseteq \mathcal{X}$ is defined as the minimal number of open balls in $\mathcal{X}$ needed to cover $\mathcal{Y}$. Covering numbers can be controlled using the slightly stronger concept of bracketing numbers. For a metric space $(\mathcal{F}, d)$ consisting of functions from $\mathcal{A} \to \mathcal{B}$, we define the bracket $[f, g]$ for two functions $f, g \in \mathcal{F}$ to be the set of functions $h\colon \mathcal{A} \to \mathcal{B}$ satisfying $f(x) \le h(x) \le g(x)$ for all $x \in \mathcal{A}$. For two vectors in a Euclidian metric space, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$, we define the bracket $[\boldsymbol{a}, \boldsymbol{b}]$ as the set of those $\boldsymbol{c}$ satisfying $a_j \le c_j \le b_j$ for all j. We define the bracketing number of a set $\mathcal{H} \subseteq \mathcal{F}$ to be the minimal number of brackets in $\mathcal{F}$ needed to cover $\mathcal{H}$.

Before proceeding with the proofs, we review some notation. Given an observed i.i.d. sample $\boldsymbol{x} = (x_1, \dots, x_n)$, a likelihood function f and a prior P_n on $\mathcal{F}$, the space of probability densities on $[0, 1]$, the posterior probability of a measurable set $\mathcal{A}$ is

$$P_n(f \in \mathcal{A} \mid \boldsymbol{x}) = \frac{\int_{\mathcal{A}} \prod_{i=1}^n f(x_i) \mathrm{d}P_n(f)}{\int_{\mathcal{F}} \prod_{i=1}^n f(x_i) \mathrm{d}P_n(f)}.$$

From the form of the random histogram prior given in (4), this probability can be decomposed as

$$P_n(f \in \mathcal{A} \mid \boldsymbol{x}) = \frac{\sum_{\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n}} p_n(\mathcal{I}) \int_{\mathcal{A}_{\mathcal{I}}} p(\boldsymbol{\theta} \mid \mathcal{I}) \prod_{i=1}^n f_{\mathcal{I},\boldsymbol{\theta}}(x_i) \mathrm{d}\boldsymbol{\theta}}{\sum_{\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n}} p_n(\mathcal{I}) \int_{S_k} p(\boldsymbol{\theta} \mid \mathcal{I}) \prod_{i=1}^n f_{\mathcal{I},\boldsymbol{\theta}}(x_i) \mathrm{d}\boldsymbol{\theta}},$$

where $\mathcal{A}_{\mathcal{I}} = \{\boldsymbol{\theta} \in S_k\colon f_{\mathcal{I},\boldsymbol{\theta}} \in \mathcal{A}\}$ and $f_{\mathcal{I},\boldsymbol{\theta}}$ is the density given in (1). Since our primary goal is to study the posterior mean of f conditional on $\widehat{\mathcal{I}}$, the posterior distribution $f \mid \boldsymbol{x}, \mathcal{I}$ is of interest. Conditional on the chosen partition $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$, we have

$$P_n(f_{\mathcal{I},\boldsymbol{\theta}} \in \mathcal{A} \mid \boldsymbol{x}, \mathcal{I}) = \frac{\int_{\mathcal{A}_{\mathcal{I}}} p(\boldsymbol{\theta} \mid \mathcal{I}) \prod_{i=1}^n f_{\mathcal{I},\boldsymbol{\theta}}(x_i) \mathrm{d}\boldsymbol{\theta}}{\int_{S_k} p(\boldsymbol{\theta} \mid \mathcal{I}) \prod_{i=1}^n f_{\mathcal{I},\boldsymbol{\theta}}(x_i) \mathrm{d}\boldsymbol{\theta}}.$$

Moreover, the sample $x_1, \dots, x_n$ is assumed to be i.i.d. from a true density f_0 and we denote by P_0^n the n-fold product measure induced by f_0.

B.1. Convergence rate

For $M > 0$ and a sequence $\epsilon_n = o(1)$, we define $\mathcal{A}_n = \{f\colon d_H(f_0, f) \ge M\epsilon_n\}$. In line with the notation introduced in the beginning of the section, for $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$ we denote by $\mathcal{A}_{n,\mathcal{I}}$ the set given by $\{\boldsymbol{\theta} \in S_k\colon d_H(f_0, f_{\mathcal{I},\boldsymbol{\theta}}) \ge M\epsilon_n\}$.

Our first result concerns the convergence rate of the conditional posterior mean based on a maximum a posteriori model selection procedure.

Lemma 1. *Let $\epsilon_n, \bar{\epsilon}_n$ be sequences satisfying $\bar{\epsilon}_n \le \epsilon_n$ and $n\bar{\epsilon}_n^2 \to \infty$. Suppose that for some $b > 0$ and a constant $c > 0$, we can find a sequence of interval partitions $\{\mathcal{I}^{(n)}\}_{n \in \mathbb{N}}$ of $[0, 1]$, where $\mathcal{I}^{(n)} \in \mathcal{P}_{\mathcal{T}_n}$, such that the event*

$$e^{(2+c)n\bar{\epsilon}_n^2} p(\boldsymbol{x} \mid \mathcal{I}^{(n)})\, p_n(\mathcal{I}^{(n)}) / \prod_{i=1}^n f_0(x_i) > b, \tag{B.1}$$

denoted $\mathcal{B}_n$, satisfies $P_0^n(\boldsymbol{x} \in B_n^c) = \mathcal{O}(\epsilon_n^2)$. Suppose further that we can find sets $\{\mathcal{F}_n\}$ in $\mathcal{F}$ such that

$$N(\epsilon_n/2, \mathcal{F}_n, d_H) < \exp(n\epsilon_n^2), \tag{B.2}$$

$$P_n(\mathcal{F}_n^c) < \exp(-\{c + 4\} n\bar{\epsilon}_n^2), \tag{B.3}$$

Then $\mathbb{E}_{\boldsymbol{x} \sim f_0}\{d_H^2(f_0, f_{\widehat{\mathcal{I}}})\} = \mathcal{O}(\epsilon_n^2)$.

Proof. Take $\{\mathcal{I}^{(n)}\}$ to satisfy (B.1). For a partition $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n}$, we write $D_n(\mathcal{I}) = p(\boldsymbol{x} \mid \mathcal{I})\, p_n(\mathcal{I}) / \prod_{i=1}^n f_0(x_i)$. We work with the following posterior probability:

$$\begin{aligned} P_n\left(f_{\widehat{\mathcal{I}},\boldsymbol{\theta}} \in \mathcal{A}_n \,\middle|\, \boldsymbol{x}, \widehat{\mathcal{I}}\right) &= \frac{\int_{\mathcal{A}_{n,\widehat{\mathcal{I}}}} p(\boldsymbol{\theta} \mid \widehat{\mathcal{I}}) \prod_{i=1}^n f_{\widehat{\mathcal{I}},\boldsymbol{\theta}}(x_i) \mathrm{d}\boldsymbol{\theta}}{\int_{S_k} p(\boldsymbol{\theta} \mid \widehat{\mathcal{I}}) \prod_{i=1}^n f_{\widehat{\mathcal{I}},\boldsymbol{\theta}}(x_i) \mathrm{d}\boldsymbol{\theta}} \\ &\le \frac{\int_{\mathcal{A}_{n,\widehat{\mathcal{I}}}} p(\boldsymbol{\theta} \mid \widehat{\mathcal{I}}) \prod_{i=1}^n f_{\widehat{\mathcal{I}},\boldsymbol{\theta}}(x_i) \mathrm{d}\boldsymbol{\theta}}{\int_{S_k} p(\boldsymbol{\theta} \mid \widehat{\mathcal{I}}) \prod_{i=1}^n f_{\widehat{\mathcal{I}},\boldsymbol{\theta}}(x_i) \mathrm{d}\boldsymbol{\theta}} \mathbb{1}_{B_n}(\boldsymbol{x}) + \mathbb{1}_{B_n^c}(\boldsymbol{x}). \end{aligned}$$

We now turn our attention toward bounding the first term in the above inequality. To this end, we note that multiplying by $p_n(\hat{I})/\prod_{i=1}^n f_0(x_i)$ in the numerator results in the following inequality:

$$\begin{aligned}
&D_n(\hat{I})^{-1} p_n(\hat{I}) \int_{\mathcal{A}_{n,\hat{I}}} p(\boldsymbol{\theta} \,|\, \hat{I}) \prod_{i=1}^n \frac{f_{\hat{I},\boldsymbol{\theta}}(x_i)}{f_0(x_i)} \,\mathrm{d}\boldsymbol{\theta}\, \mathbb{1}_{\mathcal{B}_n}(\boldsymbol{x}) \\
&\leq D_n(\mathcal{I}^{(n)})^{-1} p_n(\hat{I}) \int_{\mathcal{A}_{n,\hat{I}}} p(\boldsymbol{\theta} \,|\, \hat{I}) \prod_{i=1}^n \frac{f_{\hat{I},\boldsymbol{\theta}}(x_i)}{f_0(x_i)} \,\mathrm{d}\boldsymbol{\theta}\, \mathbb{1}_{\mathcal{B}_n}(\boldsymbol{x}),
\end{aligned}$$

where we used that $p_n(\hat{I})\, p(\boldsymbol{x} \,|\, \hat{I}) \geq p_n(\mathcal{I}^{(n)})\, p(\boldsymbol{x} \,|\, \mathcal{I}^{(n)})$ and hence $D_n(\hat{I}) \geq D_n(\mathcal{I}^{(n)})$ for the second inequality, courtesy of the argmax property of $\hat{I}$. The above expression can be further bounded by

$$\begin{aligned}
&b^{-1} e^{(2+c)n\bar{\epsilon}_n^2} p_n(\hat{I}) \int_{\mathcal{A}_{n,\hat{I}}} p(\boldsymbol{\theta} \,|\, \hat{I}) \prod_{i=1}^n \frac{f_{\hat{I},\boldsymbol{\theta}}(x_i)}{f_0(x_i)} \,\mathrm{d}\boldsymbol{\theta} \\
&\leq b^{-1} e^{(2+c)n\bar{\epsilon}_n^2} \sum_{\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n}} p_n(\mathcal{I}) \int_{\mathcal{A}_{n,\mathcal{I}}} p(\boldsymbol{\theta} \,|\, \mathcal{I}) \prod_{i=1}^n \frac{f_{\mathcal{I},\boldsymbol{\theta}}(x_i)}{f_0(x_i)} \,\mathrm{d}\boldsymbol{\theta} \\
&= b^{-1} e^{(2+c)n\bar{\epsilon}_n^2} \int_{\mathcal{A}_n} \prod_{i=1}^n \frac{f(x_i)}{f_0(x_i)} \mathrm{d}P_n(f).
\end{aligned}$$

Arguing as in the proof of Theorem 8.9 in Ghosal and van der Vaart (2017), under the conditions (B.2) and (B.3) the expectation of the right hand side of the above under P_0^n is $\mathcal{O}(e^{-\rho n\epsilon_n^2})$ when it is multiplied by $e^{(2+c)n\bar{\epsilon}_n^2}$ for sufficiently large $M > 0$ and some $\rho > 0$ not depending on n. From this we deduce that

$$\mathbb{E}_{\boldsymbol{x} \sim f_0} \left\{ P_n\left(f_{\hat{I},\boldsymbol{\theta}} \in \mathcal{A}_n \,\middle|\, \boldsymbol{x}, \hat{I} \right) \right\} \leq \mathcal{O}(e^{-\rho n\epsilon_n^2}) + P_0^n(\boldsymbol{x} \in \mathcal{B}_n^c) = \mathcal{O}(\epsilon_n^2), \tag{B.4}$$

where the last statement follows from the assumption that $P_0^n(\boldsymbol{x} \in \mathcal{B}_n^c) = \mathcal{O}(\epsilon_n^2)$. To conclude the proof, we note that by arguing as on page 507 of Ghosal et al. (2000), we have

$$d_H^2(f_0, \hat{f}_{\hat{I}}) \leq M^2\epsilon_n^2 + 2P_n\left(f_{\hat{I},\boldsymbol{\theta}} \in \mathcal{A}_n \,\middle|\, \boldsymbol{x}, \hat{I} \right),$$

which follows from the convexity of the map $g \mapsto d_H^2(f_0, g)$, the conditional version of Jensen's inequality and the fact that the Hellinger metric is bounded by $\sqrt{2}$. Since both terms on the right hand side of (B.4) are $\mathcal{O}(\epsilon_n^2)$, we obtain

$$\mathbb{E}_{\boldsymbol{x} \sim f_0} \{ d_H^2(f_0, \hat{f}_{\hat{I}}) \} = \mathcal{O}(\epsilon_n^2),$$

which proves the claim. □

We now turn our attention to finding useful bound on the covering number. Define

$$\mathcal{H}_{\mathcal{T}_n,k} = \left\{ f : [0,1] \to \mathbb{R} : f(x) = \sum_{j=1}^k |I_j|^{-1} \theta_j \mathbb{1}_{I_j}(x) \text{ for } \mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k} \text{ and } \boldsymbol{\theta} \in \mathcal{S}_k \right\}. \tag{B.5}$$

To show that the conditions of Lemma 1 hold, we need to bound the bracketing number of $\mathcal{H}_{\mathcal{T}_n,k}$. This is achieved by relating it to the bracketing number of the simplex $\mathcal{S}_k$. Variants of the following lemmas appear in the proofs of Ghosal (2001), Scricciolo (2007). We include the argument in its full length here for completeness.

Lemma 2. *Let $\mathcal{I}$ be a partition of $[0,1]$. Let $l = \sum_{j=1}^k a_j\omega_j$ and $u = \sum_{j=1}^k b_j\omega_j$ with $\omega_j(x) = |I_j|^{-1}\mathbb{1}_{I_j}(x)$. Suppose that $h(x) = \sum_{j=1}^k \theta_j\omega_j(x)$, where $\boldsymbol{\theta} \in \mathcal{S}_k$. If $[\mathbf{a}, \mathbf{b}]$ is an ϵ-bracket containing $\boldsymbol{\theta}$, then $[l,u]$ is an ϵ-bracket containing h.*

Proof. The claim that $h \in [l,u]$, follows from the fact that $|I_j|^{-1}a_j \leq |I_j|^{-1}\theta_j \leq |I_j|^{-1}b_j$ for all j by assumption.

It remains to show that $d_H(l,u) < \epsilon$. We start off by noting that

$$d_H^2(a_j\omega_j, b_j\omega_j) = \int_{I_j} \left(\sqrt{|I_j|^{-1}a_j} - \sqrt{|I_j|^{-1}b_j} \right)^2 \mathrm{d}x = \left(\sqrt{a_j} - \sqrt{b_j} \right)^2.$$

From Lemma 4 in Genovese and Wasserman (2000), we have that $d_H^2(l,u) \leq \sum_{j=1}^k d_H^2(a_j\omega_j, b_j\omega_j)$ and hence,

$$d_H^2(l,u) \leq \sum_{j=1}^k \left(\sqrt{a_j} - \sqrt{b_j} \right)^2 = d_H^2(\mathbf{a}, \mathbf{b}) < \epsilon^2,$$

which was to be shown. □

Define $\mathcal{H}_{\mathcal{T}_n,\mathcal{I}}$ to be the set of piecewise constant densities based on the partition $\mathcal{I}$. We now give a bound on the covering number of $\mathcal{H}_{\mathcal{T}_n,\mathcal{I}}$.

Lemma 3. *Let $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$. Then the ϵ-bracketing number of $\mathcal{H}_{\mathcal{T}_n,\mathcal{I}}$ for $\epsilon \leq 1$ satisfies*

$$N_{[]}\big(\epsilon, \mathcal{H}_{\mathcal{T}_n,\mathcal{I}}, d_H\big) \leq \frac{k(2\pi e)^{k/2}}{\epsilon^{k-1}}, \tag{B.6}$$

Moreover, the ϵ-convering number of $\mathcal{F}_n = \cup_{k=1}^{s_n} \mathcal{H}_{\mathcal{T}_n,k}$ satisfies

$$N_{[]}\big(\epsilon, \mathcal{F}_n, d_H\big) \leq \left(\frac{Lk_n}{\epsilon s_n}\right)^{s_n}, \tag{B.7}$$

for a universal constant $L > 0$.

Proof. To show that the first claim holds, let $[\mathbf{a}_1, \mathbf{b}_1], \ldots, [\mathbf{a}_m, \mathbf{b}_m]$ be an ϵ-bracketing of $\mathcal{S}_k$, and assume that m is chosen so that the number of brackets is minimal. Define the associated functions $l_1, u_1, \ldots, l_m, u_m$ as in Lemma 2,

$$l_i(x) = \sum_{j=1}^{k} a_{i,j}|\mathcal{I}_j|^{-1}\mathbb{1}_{\mathcal{I}_j}(x), \quad u_i(x) = \sum_{j=1}^{k} b_{i,j}|\mathcal{I}_j|^{-1}\mathbb{1}_{\mathcal{I}_j}(x).$$

Let $h \in \mathcal{H}_{\mathcal{T}_n,\mathcal{I}}$ be given by $h(x) = \sum_{j=1}^{k} |\mathcal{I}_j|^{-1}\theta_j \mathbb{1}_{\mathcal{I}_j}(x)$, for some $\theta \in \mathcal{S}_k$. Since $\mathcal{S}_k \subseteq \cup_{l=1}^{m}[\mathbf{a}_l, \mathbf{b}_l]$, we have $\theta \in [\mathbf{a}_i, \mathbf{b}_i]$ for some i. By Lemma 2, $[l_i, u_i]$ is an ϵ-bracket containing h, and since h was arbitrary, we conclude that $\mathcal{H}_{\mathcal{T}_n,\mathcal{I}} \subseteq \cup_{i=1}^{m}[l_i, u_i]$. By the minimality of m, this shows that the Hellinger bracketing number of $\mathcal{H}_{\mathcal{T}_n,\mathcal{I}}$ can be no greater than that of $\mathcal{S}_k$. By Lemma 2 in Genovese and Wasserman (2000), the bracketing number of the k-simplex $\mathcal{S}_k$ satisfies

$$N_{[]}\big(\epsilon, \mathcal{S}_k, d_H\big) \leq \frac{k(2\pi e)^{k/2}}{\epsilon^{k-1}}.$$

Using this fact we find that

$$N_{[]}\big(\epsilon, \mathcal{H}_{\mathcal{T}_n,\mathcal{I}}, d_H\big) \leq N_{[]}\big(\epsilon, \mathcal{S}_k, d_H\big) \leq \frac{k(2\pi e)^{k/2}}{\epsilon^{k-1}},$$

which was to be shown.

To prove the second part of the lemma, we note that for any $k \leq s_n$, a piecewise constant density based on the partition $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$ can be written as

$$f_{\theta,\mathcal{I}}(x) = \sum_{j=1}^{k} |\mathcal{I}_j|^{-1}\theta_j \mathbb{1}_{\mathcal{I}_j}(x), \text{ for } x \in [0,1],$$

for some $\theta \in \mathcal{S}_k$. Now by choosing $\mathcal{J} \in \mathcal{P}_{\mathcal{T}_n,s_n}$ such that $\mathcal{J}$ is a refinement of $\mathcal{I}$ and $\tilde{\theta} \in \mathcal{S}_{s_n}$ appropriately, we have that

$$f_{\mathcal{J},\tilde{\theta}}(x) = f_{\mathcal{I},\theta}(x), \text{ for each } x \in [0,1].$$

We conclude that $\mathcal{P}_{\mathcal{T}_n,k} \subseteq \mathcal{P}_{\mathcal{T}_n,s_n}$, and as such, $\mathcal{F}_n = \cup_{k=1}^{s_n}\mathcal{H}_{\mathcal{T}_n,k} \subseteq \mathcal{H}_{\mathcal{T}_n,s_n}$. Thus, it suffices to bound the covering number of $\mathcal{P}_{\mathcal{T}_n,s_n}$. Let $\mathcal{H}_{\mathcal{T}_n,\mathcal{I}}$ be the set of piecewise constant densities based on the partition $\mathcal{I}$. By construction, we have that

$$\mathcal{H}_{\mathcal{T}_n,s_n} = \bigcup_{\mathcal{I}\in\mathcal{P}_{\mathcal{T}_n,s_n}} \mathcal{H}_{\mathcal{T}_n,\mathcal{I}}.$$

By (B.6), the bracketing number of $\mathcal{H}_{\mathcal{T}_n,\mathcal{I}}$ for $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,s_n}$ is bounded by the bracketing number of $\mathcal{S}_{s_n}$. We note that there are $\binom{k_n-1}{s_n-1}$ possible partitions of $[0,1]$ of size s_n. By a union bound, we have for any $\epsilon \leq 1$ that

$$\begin{aligned} N_{[]}\Big(\epsilon, \mathcal{H}_{\mathcal{T}_n,s_n}, d_H\Big) &\leq \sum_{\mathcal{I}\in\mathcal{P}_{\mathcal{T}_n,s_n}} N_{[]}\Big(\epsilon, \mathcal{H}_{\mathcal{T}_n,\mathcal{I}}, d_H\Big) \\ &\leq \sum_{\mathcal{I}\in\mathcal{P}_{\mathcal{T}_n,s_n}} N_{[]}\Big(\epsilon, \mathcal{S}_{s_n}, d_H\Big) \\ &= \binom{k_n-1}{s_n-1} N_{[]}\Big(\epsilon, \mathcal{S}_{s_n}, d_H\Big). \end{aligned}$$

Using the fact that the bracketing number bounds the covering number for the same ϵ together with previous estimates, we deduce the following chain of inequalities:

$$\begin{aligned} N\big(\epsilon, \mathcal{H}_{\mathcal{T}_n,s_n}, d_H\big) &\leq \sum_{\mathcal{I}\in\mathcal{P}_{\mathcal{T}_n,s_n}} N_{[]}\big(\epsilon, \mathcal{H}_{\mathcal{T}_n,\mathcal{I}}, d_H\big) \\ &\leq \binom{k_n-1}{s_n-1} N_{[]}\big(\epsilon, \mathcal{S}_{s_n}, d_H\big) \\ &\leq \big(ek_n/s_n\big)^{s_n} \frac{s_n(2\pi e)^{s_n/2}}{\epsilon^{s_n-1}}, \end{aligned}$$

where we used that $\binom{k_n-1}{s_n-1} \le \binom{k_n}{s_n} \le \left(ek_n/s_n\right)^{s_n}$. To conclude the proof, we note that as $s_n \le 2^{s_n}$ for $s_n \ge 1$,

$$N\left(\epsilon/2, \mathcal{H}_{\mathcal{T}_n,s_n}, d_H\right) \le \left(\frac{ek_n}{s_n}\right)^{s_n} \frac{s_n(2\pi e)^{s_n/2}}{(\epsilon/2)^{s_n-1}} \le \left(\frac{Lk_n}{\epsilon s_n}\right)^{s_n},$$

for some constant $L > 0$, which was to be shown. □

In the sequel, for a given function $h \in \mathbb{L}_1\left([0,1]\right)$ we write $h_{\mathcal{I}}$ for the piecewise constant version of h based on the partition $\mathcal{I} \in \mathcal{P}_{\mathcal{T}_n,k}$, defined by

$$h_{\mathcal{I}}(x) = \sum_{j=1}^{k} h_j \mathbb{1}_{I_j}(x), \quad \text{for } x \in [0,1],$$

where $h_j = |I_j|^{-1} \int_{I_j} h(x)\,\mathrm{d}x$. The following lemma establishes an approximation rate result for Hölder continuous functions.

Lemma 4. *Let $h : [0,1] \to \mathbb{R}$ be α-Hölder continuous. Let $\left\{\mathcal{I}^{(k)}\right\}_{k=1}^{\infty}$ be a sequence of partitions of $[0,1]$ such that $\max_j \left|I_j^{(k)}\right| = \mathcal{O}\left(k^{-1}\right)$. Then the sequence of piecewise constant approximations $\{h_{\mathcal{I}^{(k)}}\}_{k=1}^{\infty}$ of h satisfies*

$$\left\|h - h_{\mathcal{I}^{(k)}}\right\|_\infty = \mathcal{O}\left(k^{-\alpha}\right).$$

Moreover, if h is strictly positive we have $d_H^2\left(h, h_{\mathcal{I}^{(k)}}\right) = \mathcal{O}\left(k^{-2\alpha}\right)$

Proof. Fix $k \in \mathbb{N}$ and let $x \in [0,1]$. Let j be such that $x \in I_j^{(k)}$. Then,

$$\begin{aligned}
|h(x) - h_{\mathcal{I}^{(k)}}(x)| &= \left|h(x) - \left|I_j^{(k)}\right|^{-1} \int_{I_j^{(k)}} h(s)\,\mathrm{d}s\right| \\
&\le \left|I_j^{(k)}\right|^{-1} \int_{I_j^{(k)}} |h(x) - h(s)|\,\mathrm{d}s \\
&\le \left|I_j^{(k)}\right|^{-1} \int_{I_j^{(k)}} L_0|x-s|^{\alpha}\,\mathrm{d}s \\
&\le |I_j^{(k)}|^{-1} L_0 \left|I_j^{(k)}\right|^{1+\alpha} \\
&= L_0 \left|I_j^{(k)}\right|^{\alpha},
\end{aligned}$$

where we used that $|x-s| \le |I_j^{(k)}|$ for $x, s \in I_j^{(k)}$. Since $\max_j \left|I_j^{(k)}\right| = \mathcal{O}(k^{-1})$, we find

$$\left\|h - h_{\mathcal{I}^{(k)}}\right\|_\infty = \sup_{x\in[0,1]} |h(x) - h_{\mathcal{I}^{(k)}}(x)| \le L_0 \max_j \left|I_j^{(k)}\right|^{\alpha} = \mathcal{O}(k^{-\alpha}),$$

For the second part, we note that

$$d_H^2\left(h, h_{\mathcal{I}^{(k)}}\right) \le \int_0^1 \frac{\left\{h(x) - h_{\mathcal{I}^{(k)}}(x)\right\}^2}{h(x)}\,\mathrm{d}x,$$

cf. Lemmas 2.5 and 2.7 in Tsybakov (2009). Since h is continuous and strictly positive, the last expression is upper bounded by a constant multiple of $\|h - h_{\mathcal{I}^{(k)}}\|_\infty^2$ as a consequence of the extreme value theorem. □

To show that the condition (B.1) holds, it suffices to show that the prior distribution assigns sufficient probability to the event that $f \in \mathcal{N}'_{\epsilon_n}(f_0)$, where

$$\mathcal{N}'_{\epsilon}(f_0) = \left\{f : d_H^2\left(f_0, f\right)\left\|f_0/f\right\|_\infty \le \epsilon^2\right\}, \quad \epsilon > 0.$$

The following lemma gives a lower bound on the probability of $\mathcal{N}'_{\epsilon}(f_0)$.

Lemma 5. *Suppose the hypotheses of Theorem 2 are met. Let C_1, C_2 be positive constants and let $k = k(\epsilon) \in \mathbb{N}$ satisfy*

$$C_1\epsilon^{-1/\alpha} \le k \le C_2\epsilon^{-1/\alpha}. \tag{B.8}$$

Then we can find a partition $\mathcal{I}^{(n)} \in \mathcal{P}_{\mathcal{T}_n,k}$ and two constants c_1, c_2 not depending on ϵ, k and $\mathcal{I}^{(n)}$ such that for all sufficiently small $\epsilon > 0$,

$$P_n\left(\mathcal{N}'_{\epsilon}(f_0)\right) \ge p_n\left(\mathcal{I}^{(n)}\right) c_1 \exp\left(-c_2 k \log(1/\epsilon)\right), \tag{B.9}$$

provided n is sufficiently large.

Proof. We argue as in the beginning of the proof of Theorem 2 in Scricciolo (2007). Let $\epsilon > 0$ to be determined later. If n is such that $k(\epsilon) > k_n$, then the inequality (B.9) is trivially true as the right hand side is zero. Our strategy will be to relate the neighborhood $\mathcal{N}'_{\epsilon}(f_0)$ to neighborhoods in terms of the squared Hellinger distance, as the latter is more convenient to work with. To accomplish this, we first need an estimate of the squared Hellinger distance between the piecewise constant approximation and the true density

f_0. We start by constructing the partition $\mathcal{I}^{(n)}$ by including every $\lceil k_n/k \rceil$ index in the partition. The maximal size of a bin in this partition is then bounded by

$$\max_j \left|\mathcal{I}_j^{(n)}\right| \leq \lceil k_n/k \rceil \max_j \{\tau_{n,j} - \tau_{n,j-1}\} \leq A\lceil k_n/k \rceil k_n^{-1} \leq 2Ak^{-1},$$

provided n is sufficiently large. This shows that $\max_j \left|\mathcal{I}_j^{(n)}\right| = \mathcal{O}(k^{-1})$. By a similar argument, we arrive at $\min_j \left|\mathcal{I}_j^{(n)}\right| \geq Bk^{-1}/2$. Let $\gamma = c_0\epsilon$ for a positive constant c_0 to be determined later. To bound the squared Hellinger distance between f_0 and $f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}$ we apply the triangle inequality together with the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, yielding

$$d_H^2\left(f_0, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right) \leq 2d_H^2\left(f_0, f_{0,\mathcal{I}^{(n)}}\right) + 2d_H^2\left(f_{0,\mathcal{I}^{(n)}}, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right). \tag{B.10}$$

To bound the first term of (B.10), we leverage Lemma 4 to find that

$$d_H^2\left(f_0, f_{0,\mathcal{I}^{(n)}}\right) \leq \tilde{A}k^{-2\alpha} \leq \tilde{A}C_1^{-2\alpha}\epsilon^2 = \tilde{A}C_1^{-2\alpha}c_0^2\gamma^2,$$

for some constant $\tilde{A} > 0$. We now bound the last term on the right hand side of (B.10). From $d_H^2\left(f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}, f_{0,\mathcal{I}^{(n)}}\right) \leq \left\|f_{\mathcal{I}^{(n)},\boldsymbol{\theta}} - f_{0,\mathcal{I}^{(n)}}\right\|_1$ we obtain

$$\begin{aligned} d_H^2\left(f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}, f_{0,\mathcal{I}^{(n)}}\right) &\leq \left\|f_{\mathcal{I}^{(n)},\boldsymbol{\theta}} - f_{0,\mathcal{I}^{(n)}}\right\|_1 \\ &= \sum_{j=1}^k \int_{\mathcal{I}_j^{(n)}} \left|\mathcal{I}_j^{(n)}\right|^{-1} \left|\theta_j - \pi_j^{(n)}\right| \mathrm{d}x \\ &= \sum_{j=1}^k \left|\theta_j - \pi_j^{(n)}\right| \\ &= \left\|\boldsymbol{\theta} - \boldsymbol{\pi}^{(n)}\right\|_1, \end{aligned}$$

for $\pi_j^{(n)} = \int_{\mathcal{I}_j^{(n)}} f_0(x)\,\mathrm{d}x$. Now for any simplex vector $\boldsymbol{\theta} \in S_k$ satisfying the inequality $\left\|\boldsymbol{\theta} - \boldsymbol{\pi}^{(n)}\right\|_1 \leq C_1^{-2\alpha}c_0^2\gamma^{1+1/\alpha}$ it follows from (B.10) that

$$\begin{aligned} d_H^2\left(f_0, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right) &\leq 2\tilde{A}k^{-2\alpha} + 2\left\|\boldsymbol{\theta} - \boldsymbol{\pi}^{(n)}\right\|_1 \\ &\leq 2\tilde{A}k^{-2\alpha} + 2C_1^{-2\alpha}c_0^2\gamma^{1+1/\alpha} \\ &\leq 2(\tilde{A}C_1^{-2\alpha}c_0^2 + C_1^{-2\alpha}c_0^2)\gamma^2 \\ &= c_3\gamma^2, \end{aligned}$$

where we used that $\gamma^{1+1/\alpha} \leq \gamma^2$ for $\gamma \in (0,1)$ and $c_3 = 2c_0^2C_1^{-2\alpha}(\tilde{A}+1)$ is a constant. Note that $c_3 \downarrow 0$ as $c_0 \downarrow 0$, so c_3 can be made arbitrarily small by choosing c_0 to be suitably small.

Denote $\mathcal{A}_\gamma = \left\{\boldsymbol{\theta} \in S_k : \left\|\boldsymbol{\theta} - \boldsymbol{\pi}^{(n)}\right\|_1 \leq C_1^{-2\alpha}c_0^2\gamma^{1+1/\alpha}\right\}$. To bound the likelihood ratio $\|f_0/f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\|_\infty$, we note that for $m = \inf_{x\in[0,1]} f_0(x)$, we have $\pi_j^{(n)} \geq m|\mathcal{I}_j^{(n)}|$ for $j = 1, 2, \ldots, k$. For sufficiently small γ, we thus have on $\mathcal{A}_\gamma$ that

$$\theta_j \geq \pi_j^{(n)} - \left\|\boldsymbol{\theta} - \boldsymbol{\pi}^{(n)}\right\|_1 \geq m|\mathcal{I}_j^{(n)}| - C_1^{-2\alpha}c_0^2\gamma^{1+1/\alpha}.$$

To ensure that $\theta_j \geq m\left|\mathcal{I}_j^{(n)}\right|/2$, we note that $\gamma^{1+1/\alpha} \leq \mathcal{O}(k^{-1-\alpha}) = o(k^{-1})$ and hence, any constant multiple of $\gamma^{1+1/\alpha}$ is upper bounded by $mBk^{-1}/4$ for large values of k. Consequently $C_1^{-2\alpha}c_0^2\gamma^{1+1/\alpha} \leq m\min_j|\mathcal{I}_j^{(n)}|/2$ which further implies $\theta_j \geq m|\mathcal{I}_j^{(n)}|/2$ by the above display. From the above arguments we arrive at $f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}(x) \geq m/2$ and hence

$$\left\|f_0/f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right\|_\infty \leq \frac{2\|f_0\|_\infty}{m}.$$

It follows that $d_H^2\left(f_0, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right)\left\|f_0/f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right\|_\infty \leq c_4\gamma^2$, for a constant $c_4 = c_4(c_0) \downarrow 0$ as $c_0 \downarrow 0$. Applying the above estimates, we thus have that $\mathcal{A}_\gamma \subseteq \mathcal{N}_\epsilon'(f_0)$. As $\gamma^{1+1/\alpha} = o(k^{-1})$, any constant multiple of $\gamma^{1+1/\alpha}$ is less than $1/(\Sigma k)$ for all sufficiently large k. In addition, the lower bound $\sigma k^{-1} \leq a_{k,j}$ implies that the Dirichlet parameters are lower bounded by a constant multiple of $\gamma^{-b(1+1/\alpha)}$ for some $b > 0$, so the conditions of Lemma G.13 in Ghosal and van der Vaart (2017) are met, and we obtain the lower bound

$$P_n\left(\mathcal{A}_\gamma \mid \mathcal{I}^{(n)}\right) \geq c_1 \exp\left(-c_5 k \log(2C_1^{2\alpha}c_0^{-2}/\gamma^{1+1/\alpha})\right),$$

for some $c_5 > 0$. It remains to show that we can write the probability on the right hand side in terms of ϵ only. Observe that for $\gamma = c_0\epsilon$ the expression $\log(2C_1^{2\alpha}c_0^{-2}/\gamma^{1+1/\alpha})/\log(1/\epsilon)$ converges to a positive, finite limit as $\epsilon \downarrow 0$. From this we conclude that for a suitable constant $c_2 > 0$,

$$P_n\left(\mathcal{N}_\epsilon'(f_0)\right) \geq p_n\left(\mathcal{I}^{(n)}\right) c_1 \exp\left(-c_2 k \log(1/\epsilon)\right),$$

which was to be shown. □

We are finally in a position to prove Theorem 2.

Proof of Theorem 2. We verify the conditions of Lemma 1, tending to the first condition with $\bar{\epsilon}_n = \epsilon_n = \{n/\log(n)\}^{-\alpha/(2\alpha+1)}$.

Letting $\{s_n\}$ be a sequence of integers satisfying $C_1\epsilon_n^{-1/\alpha} \leq s_n \leq C_2\epsilon_n^{-1/\alpha}$ for positive constants C_1, C_2, the result of Lemma 5 yields

$$P_n\Big(f : d_H^2(f_0, f)\Big\|f_0/f\Big\|_\infty \leq \epsilon^2\Big) \geq p_n\big(\mathcal{I}^{(n)}\big)\, c_1 \exp\big(-c_2 s_n \log(1/\epsilon_n)\big), \tag{B.11}$$

for some $\mathcal{I}^{(n)} \in \mathcal{H}_{\mathcal{T}_n, s_n}$. To continue our argument, we need to lower bound the prior probability of the partition $\mathcal{I}^{(n)}$. To this end, note that

$$p_n\big(\mathcal{I}^{(n)} \,\big|\, s_n\big) = \binom{k_n - 1}{s_n - 1}^{-1} \geq \binom{k_n}{s_n}^{-1} \geq \left(\frac{ek_n}{s_n}\right)^{-s_n} \geq \exp\big(-s_n \log(k_n)\big), \tag{B.12}$$

where the last inequality holds for all sufficiently large n. Since $k_n = \mathcal{O}(n)$ by assumption, we have, for all sufficiently large n,

$$p_n\big(\mathcal{I}^{(n)} \mid s_n\big) \geq \exp\big(-s_n \log(A_1 n)\big) \geq \exp\big(-A_2 s_n \log(n)\big),$$

for two constants $A_1, A_2 > 0$.

$$p_n(s_n) \geq D_1 \exp\big(-d_1 s_n \log(s_n)\big) \geq D_1 \exp\big(-d_1 s_n \log(n)\big).$$

It is easily verified that $\log(1/\epsilon_n)/\log(n)$ converges to a positive, finite limit as $n \to \infty$. Moreover we have $n\epsilon_n^2 = n^{1/(2\alpha+1)}\{\log(n)\}^{2\alpha/(2\alpha+1)}$ and hence,

$$s_n \log(n) \leq C_2 n^{1/(2\alpha+1)}\{\log(n)\}^{(2\alpha+1-1)/(2\alpha+1)} = C_2 n\epsilon_n^2.$$

To proceed, note that since $\mathcal{I}^{(n)} \in \mathcal{P}_{\mathcal{T}_n, s_n}$ and it follows that $p_n\big(\mathcal{I}^{(n)}\big) = p_n(s_n)\, p_n\big(\mathcal{I}^{(n)} \mid s_n\big)$. We deduce that the right hand side of (B.11) is lower bounded by

$$D_1 \exp\big(-d_1 s_n \log(n)\big) \exp\big(-A_2 s_n \log(n)\big)\, c_1 \exp\big(-c_2 C_2 n\epsilon_n^2\big),$$

where we used the assumption on $p_n(k)$. Since $\log(s_n)/\log(n)$ converges to a positive, finite limit as $n \to \infty$, we conclude that we can find constants $c_3, c_4 > 0$ such that

$$P_n\Big(f : d_H^2\big(f_0, f\big)\Big\|f_0/f\Big\|_\infty \leq \epsilon_n^2\Big) \geq c_3 \exp\big(-c_4 n\epsilon_n^2\big).$$

We now turn our attention to verifying that $P_0(\boldsymbol{x} \in \mathcal{B}_n) = \mathcal{O}\big(\epsilon_n^2\big)$ for $\mathcal{B}_n$ the event that $e^{(2+c)n\bar{\epsilon}_n^2} p(\boldsymbol{x} \mid \mathcal{I}^{(n)})\, p_n(\mathcal{I}^{(n)})/\prod_{i=1}^n f_0(x_i) > b$, we appeal to Lemma 8.10 in Ghosal and van der Vaart (2017), which together with equation (8.8) in the same work implies that $\mathcal{B}_n$ has probability at least $d_6 D^2(\sqrt{n}/\epsilon_n)^{-6} = \mathcal{O}\big(\epsilon_n^2\big)$ under f_0. The rate condition of Lemma 1 is thus seen to hold under the stated assumptions.

To verify the prior condition (B.3), we let $\{t_n\}$ be a sequence of integers taken to satisfy

$$E_1 n^{1/(2\alpha+1)}\{\log(n)\}^{-1/(2\alpha+1)} \leq t_n \leq E_2 n^{1/(2\alpha+1)}\{\log(n)\}^{-1/(2\alpha+1)},$$

for two positive constants E_1, E_2 to be chosen later. From the definition of t_n and previous calculations it is clear that $t_n \log(n) \geq E_1 n\epsilon_n^2$. For $\mathcal{F}_n = \cup_{k=1}^{t_n} \mathcal{H}_{\mathcal{T}_n, k}$, where $\mathcal{H}_{\mathcal{T}_n, k}$ is as in (B.5), we then have that $P_n\big(\mathcal{F}_n^c\big) = \sum_{k=k_n+1}^\infty p_n(k)$. Using the assumption on $p_n(k)$ this is bounded above by $c_5 \exp\big(-d_2 t_n \log(t_n)\big)$ for $c_5 = D_2 \sum_{k=0}^\infty \exp\big(-d_2 k \log(k)\big) < \infty$. Furthermore, as $\log(t_n)/\log(n)$ converges to a positive limit as $n \to \infty$, we can find a new positive constant d_2' such that

$$\sum_{k=k_n+1}^{\infty} p_n(k) < c_5 \exp\big(-d_2' t_n \log(n)\big).$$

Taking $E_1 = (c_4 + 4)/d_2'$, we have $d_2' t_n \geq (c_4 + 4) n\epsilon_n^2$ which combined with previous estimates implies that $P_n\big(\mathcal{F}_n^c\big) \geq \exp\{-(c_4 + 4)n\epsilon_n^2\}$, verifying (B.3).

Finally, we show that the entropy condition (B.5) holds. By Lemma 3, $N(\epsilon_n/2, \mathcal{F}_n, d_H)$ is upper bounded by $\exp\big(t_n \log(Lk_n/\{\epsilon_n s_n\})\big)$. Since $\log(k_n) + \log(1/\epsilon_n) = \mathcal{O}\big(\log(n)\big)$, the covering number can be further bounded by $\exp\big(c_5 t_n \log(n)\big)$ for a suitable constant c_5. Since $t_n \log(n) = \mathcal{O}(n\epsilon_n^2)$ this implies that $N(\epsilon_n/2, \mathcal{F}_n, d_H) \leq \exp\big(c_6 n\epsilon_n^2\big)$ for some $c_6 > 0$ as desired. □

B.2. Consistency

The proof of Theorem 1 is in parts very similar to that of Theorem 2, so we only fill in the details where needed.

To prove Theorem 1, we need the following approximation result.

Lemma 6. *Let $f_0 : [0,1] \to \mathbb{R}$ be a probability density. Then for any $\epsilon > 0$ we can find a $\delta > 0$ such that if $\mathcal{I}$ is any interval partition of $[0,1]$ with $\max_j |I_j| < \delta$, then there is a density h, possibly depending on ϵ, that is piecewise constant on the partition $\mathcal{I}$ satisfying $\big\|f_0 - h\big\|_1 < \epsilon$.*

Proof. We will first show that for any $\epsilon > 0$ and an arbitrary sequence of partitions $\mathcal{J}^{(n)}$ with $\max_j \big|J_j^{(n)}\big| \to 0$, for all sufficiently large $N \in \mathbb{N}$ we can find a density h_N that is piecewise constant on $\mathcal{J}^{(N)}$ and which satisfies $\big\|f_0 - h_N\big\|_1 < \epsilon$. To this end, we make use of the fact that step functions are dense in $\mathbb{L}_r\big([0,1]\big)$ for $r \geq 1$ (Schilling, 2005, Corollary 12.11). Now fix $\epsilon \in (0, 1/2)$ and let g_ϵ be

a strictly positive step function such that $\|f_0 - g_\epsilon\|_1 < \epsilon/5$. From the inverse triangle inequality, we have $1 - \epsilon/5 < \|g_\epsilon\|_1 < 1 + \epsilon/5$. Letting $m_\epsilon = g_\epsilon/\|g_\epsilon\|_1$, it follows by the triangle inequality that

$$\left\|f_0 - m_\epsilon\right\|_1 \leq \left\|f_0 - g_\epsilon\right\|_1 + \left\|g_\epsilon - m_\epsilon\right\|_1 < \frac{\epsilon}{5} + \left\|g_\epsilon\right\|_1 \left(1 - \frac{1}{\|g_\epsilon\|_1}\right).$$

Using the fact that $|1 - 1/x| < 2x$ for $x > 1/2$, we further obtain $\left\|f_0 - m_\epsilon\right\|_1 < \frac{\epsilon}{5} + \frac{2\epsilon}{5}(1 + \epsilon/5) < 4\epsilon/5$. Since m_ϵ is a step function, it is almost everywhere continuous on $[0,1]$. Let $\{\mathcal{J}^{(n)}\}$ be a sequence of interval partitions of $[0,1]$ such that $\max_j \left|\mathcal{J}_j^{(n)}\right| \to 0$. Define $h_n = m_{\epsilon,\mathcal{J}^{(n)}}$ to be the piecewise constant version of m_ϵ based on the partition $\mathcal{J}^{(n)}$. To show convergence in $\mathbb{L}_1$, we first show that h_n converges to m_ϵ almost everywhere. To this end, fix $\mu > 0$ and let $x \in [0,1]$ be a continuity point of m_ϵ so that $|m_\epsilon(x) - m_\epsilon(y)| < \mu$ whenever $|x - y| < \nu$ for all sufficiently small ν. For all large n we have $\max_j \left|\mathcal{J}_j^{(n)}\right| \leq \nu$, so for $x, y \in \mathcal{J}_j^{(n)}$, integrating the inequalitites $-\mu < m_\epsilon(x) - m_\epsilon(y) < \mu$ over this interval and multiplying by $1/\left|\mathcal{J}_j^{(n)}\right|$ shows that

$$-\mu < m_\epsilon(x) - m_{\epsilon,\mathcal{J}^{(n)}}(x) < \mu.$$

Since μ was arbitrary we conclude that $h_n(x) \to m_\epsilon(x)$. As $h_n \to m_\epsilon$ almost everywhere and h_n, m_ϵ are densities for all n, an application of Scheffé's Lemma yields $\left\|m_\epsilon - h_n\right\|_1 \to 0$. In particular, this result implies that $\|m_\epsilon - h_N\|_1 < \epsilon/5$ for all sufficiently large $N \in \mathbb{N}$, so that

$$\left\|f_0 - h_N\right\|_1 \leq \left\|f_0 - m_\epsilon\right\|_1 + \left\|m_\epsilon - h_N\right\|_1 < \epsilon,$$

which proves our claim.

Suppose now that the hypothesis of the Lemma is false for the sake of contradiction. Then there exists an $\epsilon > 0$ such that for every $n \in \mathbb{N}$, we can find a partition $\mathcal{I}^{(n)}$ with $\max_j \left|\mathcal{I}_j^{(n)}\right| < 1/n$ such that $\left\|f_0 - h_n\right\|_1 \geq \epsilon$ for all densities h_n which are piecewise constant on $\mathcal{I}^{(n)}$. Since $\max_j \left|\mathcal{I}_j^{(n)}\right| \to 0$, this contradicts our previous result. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. By an argument similar to that of Lemma 1, it can be established the Bayes histogram estimator is consistent if for a constant $c > 0$ and a sequence of sets $\mathcal{F}_n \subseteq \mathcal{F}$,

$$N\left(\epsilon, \mathcal{F}_n, d_H\right) < n\epsilon^2,$$
$$P_n(\mathcal{F}_n^c) < \exp(-cn),$$

in addition to the condition that for all $\delta > 0$, we can find a sequence of partitions $\mathcal{I}^{(n)} \in \mathcal{P}_{\mathcal{T}_n}$ such that the event $e^{n\delta} p\left(\mathcal{I}^{(n)}\right) p\left(\boldsymbol{x} \mid \mathcal{I}^{(n)}\right) / \prod_{i=1}^n f_0(x_i) > b$ for some $b > 0$, denoted $\mathcal{B}_n$, has probability tending to 1. The first two conditions can be handled as in the proof of Theorem 2 by taking $\mathcal{F}_n = \cup_{k=1}^{k_n} \mathcal{H}_{\mathcal{T}_n,k}$, as $N\left(\epsilon, \mathcal{F}_n, d_H\right) < n\epsilon^2$ for all sufficiently large n and $P_n\left(\mathcal{F}_n^c\right) = 0$ in this case.

We first show that f_0 can be approximated to arbitrary precision by a piecewise constant density in the squared Hellinger metric. Since $d_H^2(f,g) \leq \|f - g\|_1$ for all densities f, g, we work with the $\mathbb{L}_1$ metric instead as it is more convenient in this case. Let now $\mu \in (0,1)$ to be decided later. We will now show that we can always construct a partition $\mathcal{I}^{(n)} \in \mathcal{P}_{\mathcal{T}_n,k}$ and find a density h_n piecewise constant on $\mathcal{I}^{(n)}$ with $k \in \mathbb{N}$ fixed such that $\left\|f_0 - h_n\right\|_1 < \mu$. By Lemma 6, we can find such an h_n whenever $\max_j \left|\mathcal{I}_j^{(n)}\right| < \nu$ for some $\nu > 0$. Let $k > 2/\nu$, and let $N \in \mathbb{N}$ be such that $\max_{1 \leq j \leq k} \{\tau_{n,j} - \tau_{n,j-1}\} < \nu/3$ for all $n \geq N$. For each $l \in \{1, 2, \ldots, k\}$, we can then find at least one $\tau_{n,j} \in ((l-1)/k, l/k]$. We now construct the partition $\mathcal{I}^{(n)}$ by including exactly one $\tau_{n,j}$ for each regular interval $\left((l-1)/k, l/k\right]$. Since the length of every such regular interval is less than $\nu/2$, it follows that the adjacent endpoints in the interval partition $\mathcal{I}^{(n)}$ constructed this way are at most ν apart and consequently we must have $\left\|f_0 - h_n\right\|_1 < \mu$.

Denote $V(f_0, f) = \int_0^1 f_0(x) \log^2 \left\{f_0(x)/f(x)\right\} \mathrm{d}x$ and define $\mathcal{N}_\epsilon''(f_0) = \left\{f : K(f_0, f) \leq \epsilon^2, V(f_0, f) \leq \epsilon^2\right\}$. To relate Hellinger neighbourhoods to $\mathcal{N}_\epsilon''(f_0)$, we appeal to Theorem 5 in Wong and Shen (1995), which implies that

$$\max\left\{K(f_0, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}), V(f_0, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}})\right\} < \epsilon^2 \text{ whenever } d_H^2\left(f_0, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right) < 3\mu$$

for all sufficiently small $\mu > 0$ provided that for some $\gamma \in (0,1]$ and the set $\mathcal{D}_\gamma = \left\{f_0/f_{\mathcal{I}^{(n)},\boldsymbol{\theta}} \geq e^{1/\gamma}\right\}$

$$L_\gamma^2 = \int_{\mathcal{D}_\gamma} f_0(x) \left\{\frac{f_0(x)}{f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}(x)}\right\}^\gamma \mathrm{d}x < \infty.$$

This holds true for $\gamma = r - 1$ if $\min_j \theta_j > \mu^2/2$ due to the fact that $f_0 \in \mathbb{L}_r\left([0,1]\right)$ for some $r \in (1,2]$ by assumption. To show that the prior assigns sufficient mass to Hellinger neighbourhoods of f_0, note that by the triangle inequality,

$$d_H^2\left(f_0, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right) \leq \left\|f_0 - h_n\right\|_1 + \left\|h_n - f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}\right\|_1 = \left\|f_0 - h_n\right\|_1 + \left\|\boldsymbol{\eta}^{(n)} - \boldsymbol{\theta}\right\|_1,$$

where $\eta_j^{(n)} = \int_{\mathcal{I}_j^{(n)}} h_n(x) \mathrm{d}x$. The first term in the above expression is less than μ for all sufficiently large n by the previous calculation, and we conclude that $\{f_{\mathcal{I}^{(n)},\boldsymbol{\theta}} : d_H^2(f_0, f_{\mathcal{I}^{(n)},\boldsymbol{\theta}}) < 3\mu\} \subseteq \mathcal{N}_\epsilon''(f_0)$ for $\boldsymbol{\theta}$ belonging to the set

$$\mathcal{A}_\mu = \left\{\boldsymbol{\theta} \in \mathcal{S}_{s_n} : \left\|\boldsymbol{\theta} - \boldsymbol{\eta}^{(n)}\right\|_1 \leq 2\mu, \min_j \theta_j > \mu^2/2\right\}.$$

Since the prior for $\boldsymbol{\theta} \mid \mathcal{I}^{(n)}$ is a k-dimensional Dir($\boldsymbol{a}$) distribution and $\boldsymbol{a} \in (0, \Sigma)^k$ by assumption, we can find $z > 0$ such that a_j is lower bounded by a constant multiple of μ^z for all j. Appealing to Lemma G.13 in Ghosal and van der Vaart (2017), we find that $P\big(\mathcal{A}_\mu \mid \mathcal{I}^{(n)}\big) > c_1 \exp\big(-c_2 k \log(1/\mu)\big)$ for two positive constants c_1, c_2. From (B.12) we have $p_n\big(\mathcal{I}^{(n)} \mid k\big) \geq \exp\big(-k \log(k_n)\big)$ and hence,

$$\begin{aligned} P_n\big(f \in \mathcal{N}_\epsilon''(f_0)\big) &\geq p_n(k)\, p_n\big(\mathcal{I}^{(n)} \mid k\big)\, P\big(\mathcal{A}_\mu \mid \mathcal{I}^{(n)}\big) \\ &\geq c_3 \exp\big(\log p_n(k) - c_4 k \log(n)\big), \end{aligned}$$

for constants $c_3, c_4 > 0$. By Lemma 8.10 in Ghosal and van der Vaart (2017), we obtain the lower bound

$$\frac{e^{n\delta} p\big(\mathcal{I}^{(n)}\big) p\big(\boldsymbol{x} \mid \mathcal{I}^{(n)}\big)}{\prod_{i=1}^n f_0(x_i)} \geq c_3 \exp\big(n\delta + \log p_n(k) - c_4 k \log(n) - (1+D) n\epsilon^2\big),$$

where $D > 0$, except for on a set of probability less than $(\sqrt{n}\epsilon^2)^{-2}$. By our assumptions the terms in the exponent on the right hand side converge to 0 when divided by n, expect for $n\delta - (1+D)n\epsilon^2$. Hence, the right hand side of the expression diverges to ∞ provided ϵ is taken to be sufficiently small, so that B_n^c is contained in a set with probability tending to 0, which concludes the proof. □

Appendix C. First simulation study

The following section includes some further details on the setup of the simulation study in Section 5.2 and presents the results in more detail.

C.1. Methods used and implementation

An overview of all the methods included in the simulation study is given in Table 1. The software implementations used for each method are as follows: For the Wand method we used the `dpih()` function in the `KernSmooth` library (Wand, 2021). The `ftnonpar` package (Davies and Kovac, 2012) was used to compute the Taut String histogram. Finally, for the two approaches of Rozenholc et al. (2010) we used the `histogram` R library (Mildenberger et al., 2019). For all the other methods we have used our own software implementation.

For all the densities in Fig. 1, we estimated the support in the way described in the beginning of Section 3.3. This was done even for those densities that have known compact supports to ensure that all density estimates were comparable, as some of the software implementations used in the simulation study did not allow us to manually specify the support of the histogram estimate.

With the exception of the Wand method, we computed the optimal value of k for each regular histogram procedure by enumerating the optimization criterion over all regular histograms consisting of fewer than $\lceil 4n/\log^2(n) \rceil + 1$ bins. Although Birgé and Rozenholc (2006) specifically recommend maximizing the penalized likelihood up to $\lceil n/\log(n) \rceil$ for their method, we have found that for the test densities under consideration here, the maximum always occurs at a much lower value of k for the largest sample sizes, so we can save a considerable amount of computation by restricting our search to a smaller set.

For the RIH, L2CV and KLCV methods we used the greedy search heuristic of Rozenholc et al. (2010) to reduce the computational burden of computing the optimal partition according to these criteria. Based on some preliminary simulations, we decided to use a data-driven grid for all three methods. For cross-validation-based approaches, we restricted our search to only consider partitions with a minimum bin width greater than $\log^{1.5}(n)/n$, similar to the approach used by Rozenholc et al. (2010) for the L2CV method. This restriction was put in place because these criteria would frequently yield density estimates with sharp and narrow spikes even for smooth densities in the unrestricted case.

C.2. Results

This subsection contains tables showing the complete output from the simulation study for the Hellinger, PID and $\mathbb{L}_2$ losses, respectively. Complete tables of the results with respect to each loss function are shown in Tables C.4–C.9. The best performing method for each combination of sample size and density has been outlined in boldface.

To provide a more compact visual summary of the results with respect to the Hellinger and $\mathbb{L}_2$ losses, we computed the logarithm of the risk relative to the best-performing method for each density $f_0 \in \mathcal{D}$ and sample size $n \in \mathcal{N}$,

$$\mathrm{LRR}_n(f_0, m) = \log \widehat{R}_n\big(f_0, \widehat{f}_m\big) - \log \min_{m' \in \mathcal{M}} \widehat{R}_n\big(f_0, \widehat{f}_{m'}\big).$$

Boxplots of $\sqrt{\mathrm{LRR}_n(f_0, m)}$ for sample sizes of $\{50, 200, 5000, 25000\}$ are shown in Figs. C.8 and C.11 for the Hellinger and $\mathbb{L}_2$ losses, respectively. Figs. C.9 and C.10 shows boxplots of the square roots of the estimated PID risks for the four irregular methods RIH, RMG-B, RMG-R and TS. The irregular L2CV and KLCV histograms were excluded from this comparison, as they tended to perform much worse than the other irregular procedures in this regard.

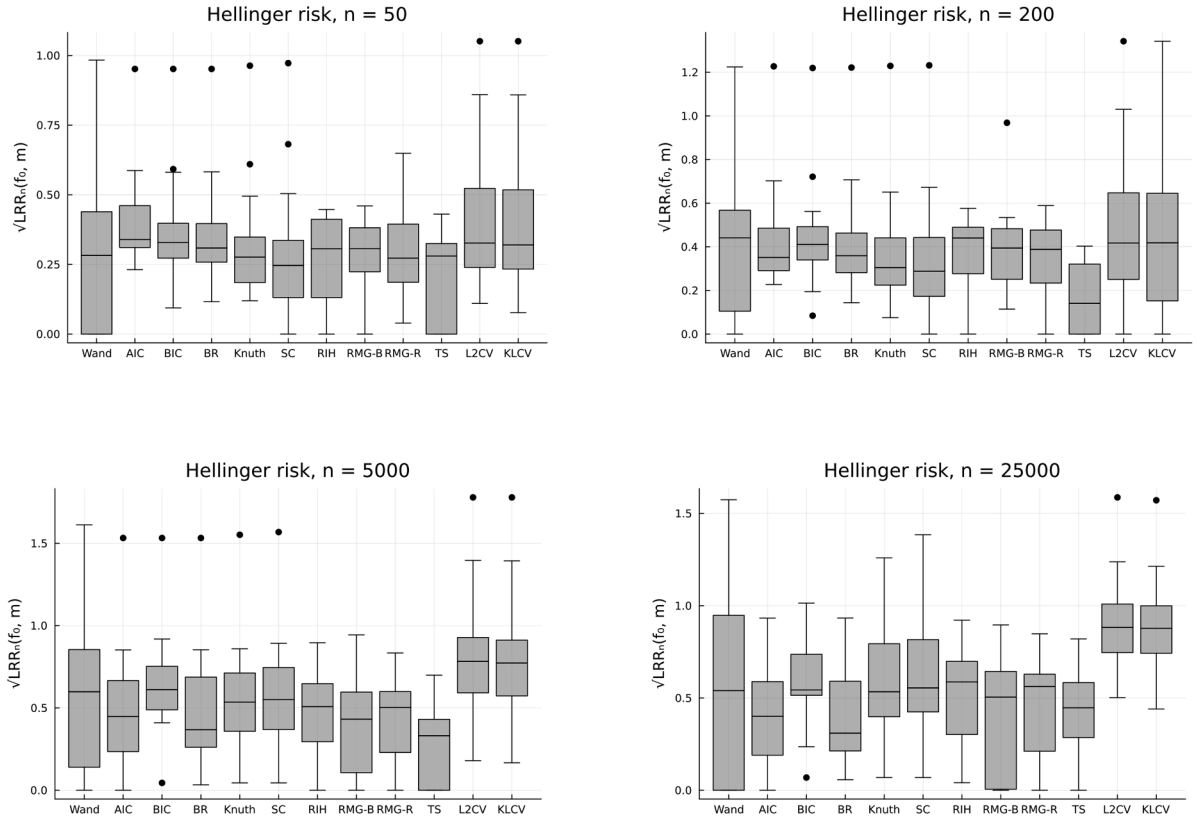

Fig. C.8. Boxplots of $\sqrt{\mathrm{LRR}_n(f_0, m)}$ for d_H.

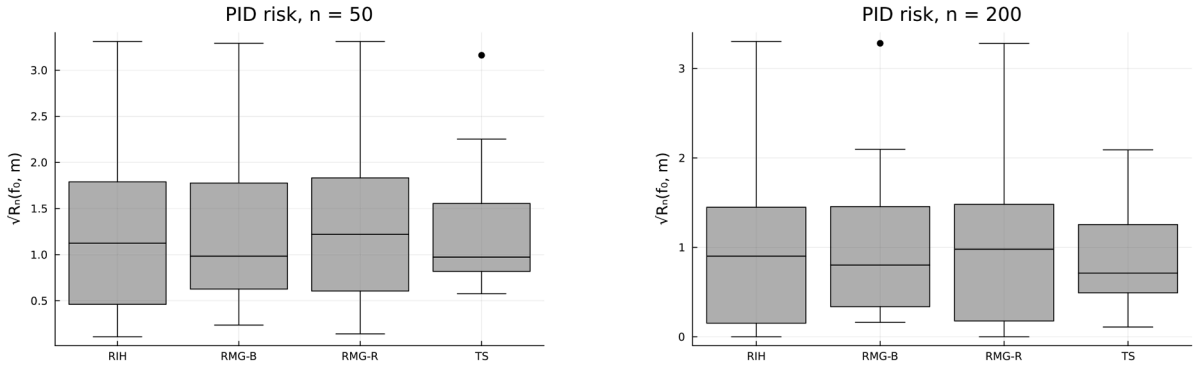

Fig. C.9. Boxplots of root-PID risks for four of the irregular histogram methods for sample sizes $n = 50$ and $n = 200$.

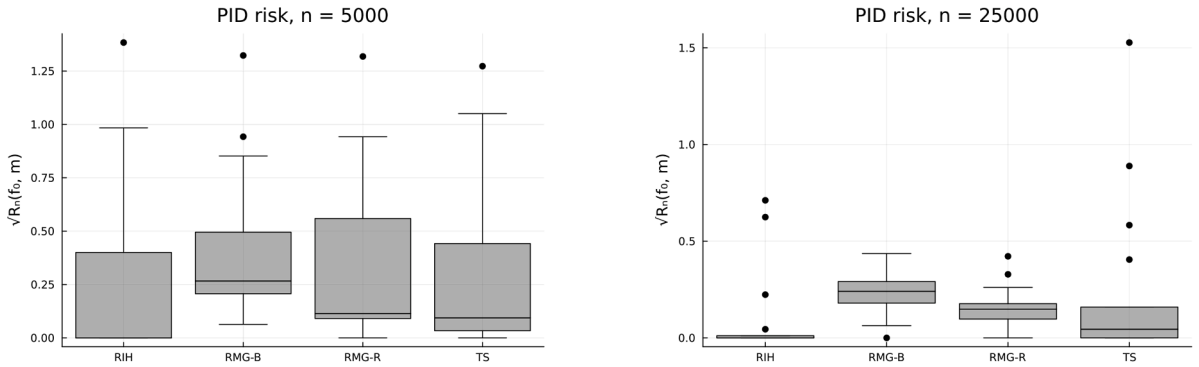

Fig. C.10. Boxplots of root-PID risks for four of the irregular histogram methods for samples of size $n = 5000$ and $n = 25000$.

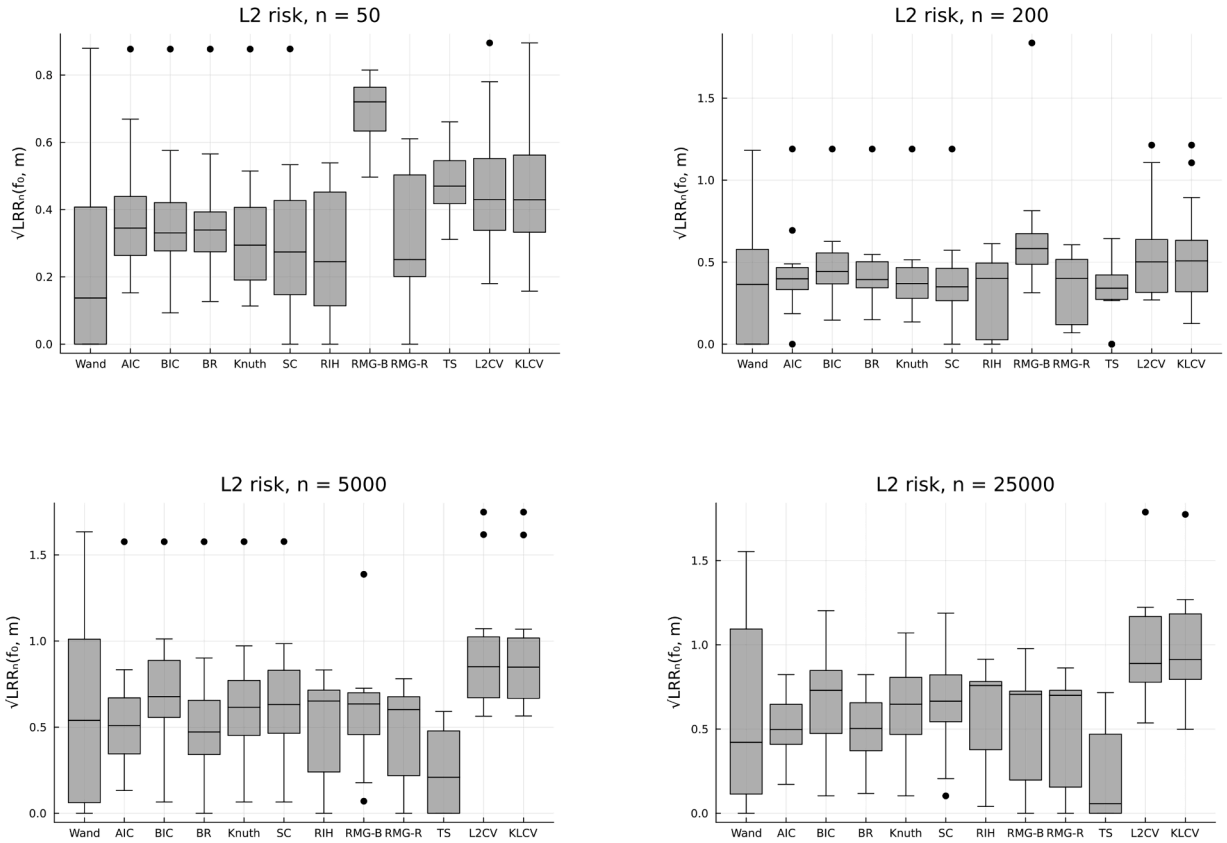

Fig. C.11. Boxplots of $\sqrt{\mathrm{LRR}_n(f_0, m)}$ for the $\mathbb{L}_2$ metric.

Table C.4
Estimated Hellinger risks for densities $1-8$.

Density	n	Wand	AIC	BIC	BR	Knuth	SC	RIH	RMG-B	RMG-R	TS	L2CV	KLCV
1	50	**0.27**	0.298	0.3	0.297	0.298	0.293	0.325	0.333	0.321	0.325	0.292	0.294
	200	**0.167**	0.185	0.194	0.187	0.185	0.183	0.22	0.222	0.218	0.189	0.194	0.195
	1000	**0.096**	0.104	0.117	0.106	0.109	0.11	0.14	0.139	0.136	0.127	0.139	0.137
	5000	**0.057**	0.06	0.071	0.061	0.066	0.067	0.089	0.085	0.084	0.092	0.112	0.11
	25,000	**0.034**	0.035	0.044	0.035	0.04	0.041	0.055	0.052	0.051	0.066	0.088	0.086
2	50	0.226	0.239	0.194	0.197	0.197	0.207	**0.189**	0.194	**0.189**	0.204	0.249	0.246
	200	0.13	0.117	**0.095**	0.099	0.096	0.097	**0.095**	0.096	**0.095**	0.102	0.188	0.187
	1000	0.07	0.052	**0.043**	0.045	**0.043**	**0.043**	**0.043**	**0.043**	**0.043**	0.047	0.152	0.151
	5000	0.04	0.022	0.019	0.019	0.019	0.019	**0.018**	0.019	0.018	0.02	0.13	0.129
	25,000	0.023	0.01	**0.008**	0.009	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**	0.009	0.104	0.099
3	50	0.433	0.446	0.443	0.444	0.404	0.408	0.384	0.391	0.393	**0.316**	0.417	0.421
	200	0.336	0.339	0.348	0.341	0.316	0.326	0.268	0.261	0.286	**0.207**	0.32	0.321
	1000	0.253	0.254	0.274	0.255	0.247	0.259	0.173	0.161	0.195	**0.136**	0.239	0.239
	5000	0.192	0.194	0.218	0.194	0.196	0.208	0.109	0.1	0.133	**0.094**	0.183	0.181
	25,000	0.146	0.148	0.173	0.148	0.156	0.167	0.068	**0.062**	0.09	0.065	0.141	0.136
4	50	0.361	0.349	0.345	0.346	0.323	0.334	0.347	0.346	0.338	**0.284**	0.33	0.34
	200	0.266	0.231	0.24	0.232	0.222	0.236	0.237	0.232	0.232	**0.183**	0.211	0.213
	1000	0.188	0.166	0.178	0.167	0.166	0.178	0.153	0.144	0.15	**0.13**	0.148	0.147
	5000	0.129	0.136	0.156	0.142	0.151	0.157	0.097	0.09	0.095	**0.089**	0.11	0.109
	25,000	0.095	0.091	0.151	0.093	0.122	0.148	0.063	**0.058**	0.063	0.076	0.075	0.07
5	50	0.32	0.334	0.327	0.327	0.322	0.321	0.357	0.363	0.356	**0.316**	0.325	0.327
	200	0.209	0.218	0.222	0.214	0.208	0.212	0.249	0.249	0.245	**0.205**	0.205	0.206
	1000	0.129	0.129	0.142	0.128	0.129	0.134	0.16	0.156	0.154	0.139	0.128	**0.127**
	5000	0.08	**0.077**	0.091	0.078	0.081	0.086	0.101	0.096	0.096	0.098	0.098	0.096
	25,000	0.049	**0.046**	0.059	0.047	0.051	0.056	0.063	0.058	0.059	0.069	0.079	0.074
6	50	0.322	0.384	0.345	0.347	0.346	0.331	**0.32**	0.332	0.325	0.317	0.334	0.331
	200	0.248	0.271	0.267	0.263	0.26	0.253	**0.228**	0.229	0.229	0.211	0.259	0.258
	1000	0.194	0.197	0.204	0.193	0.193	0.189	0.151	0.144	0.153	**0.14**	0.202	0.202
	5000	0.154	0.147	0.159	0.144	0.147	0.144	0.098	0.089	0.1	**0.093**	0.165	0.162
	25,000	0.122	0.11	0.124	0.109	0.113	0.111	0.061	**0.056**	0.065	0.062	0.13	0.123
7	50	0.337	0.356	0.355	0.353	0.349	0.343	0.357	0.363	0.35	0.37	**0.341**	0.339
	200	0.25	0.255	0.251	0.248	0.243	**0.242**	0.246	0.249	0.246	0.245	0.244	0.245
	1000	0.169	0.157	0.172	0.151	0.147	0.15	0.19	0.179	0.177	**0.131**	0.152	0.15
	5000	0.091	0.094	0.113	0.094	0.099	0.103	0.111	0.109	0.108	**0.073**	0.108	0.106
	25,000	0.055	0.055	0.071	0.055	0.061	0.064	0.072	0.066	0.066	**0.043**	0.087	0.083
8	50	**0.285**	0.312	0.309	0.307	0.306	0.302	0.312	0.314	0.305	0.318	0.302	0.301
	200	**0.174**	0.191	0.199	0.191	0.189	0.189	0.206	0.208	0.204	0.193	0.2	0.2
	1000	**0.102**	0.111	0.127	0.113	0.116	0.117	0.138	0.136	0.133	0.125	0.141	0.14
	5000	**0.061**	0.064	0.077	0.065	0.07	0.071	0.086	0.083	0.082	0.068	0.112	0.11
	25,000	**0.036**	0.038	0.048	0.038	0.043	0.044	0.055	0.051	0.051	0.049	0.088	0.086

Table C.5
Estimated Hellinger risks for densities 9 – 16.

Density	n	Wand	AIC	BIC	BR	Knuth	SC	RIH	RMG-B	RMG-R	TS	L2CV	KLCV
9	50	**0.309**	0.34	0.336	0.334	0.335	0.329	0.338	0.343	0.334	0.347	0.32	0.322
	200	**0.229**	0.242	0.24	0.234	0.232	0.231	0.251	0.252	0.247	0.241	0.235	0.235
	1000	0.159	0.152	0.17	0.152	0.154	0.154	0.165	0.16	0.157	**0.148**	0.16	0.16
	5000	0.104	0.09	0.111	0.09	0.095	0.098	0.103	0.1	0.099	**0.073**	0.117	0.115
	25,000	0.062	0.059	0.069	0.059	0.061	0.063	0.067	0.062	0.061	**0.043**	0.09	0.088
10	50	0.404	0.417	0.389	0.39	0.391	0.397	**0.385**	0.39	0.386	0.417	0.416	0.414
	200	0.349	0.305	0.345	0.33	0.31	**0.285**	0.345	0.347	0.346	0.335	0.355	0.355
	1000	0.322	0.183	0.216	0.185	0.183	**0.177**	0.23	0.23	0.226	0.191	0.201	0.2
	5000	0.189	**0.107**	0.134	**0.107**	0.114	0.112	0.162	0.151	0.147	0.112	0.133	0.133
	25,000	0.079	**0.063**	0.083	**0.063**	0.071	0.07	0.103	0.095	0.093	0.07	0.105	0.104
11	50	**0.256**	0.291	0.285	0.282	0.288	0.282	0.295	0.301	0.297	0.301	0.277	0.277
	200	**0.155**	0.171	0.178	0.172	0.171	0.168	0.193	0.195	0.193	0.178	0.194	0.193
	1000	**0.087**	0.094	0.108	0.097	0.101	0.098	0.127	0.126	0.123	0.11	0.148	0.147
	5000	**0.052**	0.054	0.066	0.056	0.061	0.059	0.08	0.078	0.076	0.075	0.126	0.124
	25,000	**0.031**	**0.031**	0.041	0.032	0.037	0.036	0.05	0.047	0.046	0.053	0.101	0.099
12	50	0.586	0.491	0.54	0.531	0.552	0.606	0.46	0.416	0.408	**0.381**	0.797	0.796
	200	0.294	0.282	0.299	0.285	0.331	0.373	0.308	0.279	0.282	**0.244**	0.705	0.7
	1000	0.183	0.163	0.241	0.174	0.221	0.263	0.187	0.171	0.172	**0.152**	0.214	0.19
	5000	**0.09**	0.093	0.135	0.098	0.132	0.156	0.115	0.105	0.107	0.104	0.093	0.093
	25,000	**0.048**	0.054	0.081	0.056	0.08	0.093	0.071	0.064	0.065	0.072	0.065	0.064
13	50	0.586	0.563	0.589	0.588	0.523	**0.508**	0.518	0.542	0.537	0.559	0.586	0.575
	200	0.489	0.395	0.477	0.417	0.377	0.371	0.485	0.889	0.492	**0.348**	0.526	0.524
	1000	0.275	0.231	0.273	0.235	0.229	0.229	0.42	0.588	0.391	**0.2**	0.323	0.32
	5000	0.155	0.13	0.171	0.131	0.134	0.136	0.271	0.295	0.243	**0.121**	0.183	0.182
	25,000	**0.074**	0.075	0.101	0.075	0.084	0.085	0.173	0.165	0.152	**0.074**	0.13	0.13
14	50	1.079	1.015	1.015	1.015	1.038	1.056	0.484	**0.41**	0.625	0.428	1.239	1.239
	200	0.913	0.919	0.903	0.907	0.924	0.93	0.25	0.209	**0.204**	0.227	1.236	1.236
	1000	0.699	0.747	0.747	0.747	0.764	0.78	0.113	0.092	**0.091**	0.105	0.965	0.964
	5000	0.548	0.426	0.426	0.426	0.452	0.476	0.051	**0.041**	**0.041**	0.048	0.964	0.964
	25,000	0.217	0.019	0.019	0.019	0.089	0.124	0.023	**0.018**	**0.018**	0.023	0.036	0.039
15	50	0.398	0.39	0.398	0.397	0.363	**0.352**	0.388	0.405	0.39	0.391	0.404	0.407
	200	0.273	0.249	0.251	0.244	0.236	0.234	0.238	0.242	0.238	**0.224**	0.228	0.221
	1000	0.184	0.141	0.14	0.134	0.134	0.133	**0.103**	0.104	**0.103**	0.129	0.152	0.146
	5000	0.128	0.07	0.068	0.068	0.068	0.068	**0.049**	0.05	0.05	0.063	0.123	0.12
	25,000	0.086	0.03	0.03	0.03	0.03	0.03	0.025	0.024	**0.023**	0.027	0.098	0.094
16	50	0.415	0.416	0.41	0.41	0.368	0.359	**0.355**	0.374	0.358	0.356	0.467	0.468
	200	0.297	0.29	0.281	0.278	0.268	0.264	**0.224**	0.228	0.225	0.23	0.242	0.229
	1000	0.2	0.171	0.226	0.172	0.17	0.166	**0.111**	**0.111**	**0.111**	0.116	0.155	0.148
	5000	0.141	0.089	0.119	0.09	0.107	0.103	0.055	0.054	**0.053**	0.063	0.124	0.12
	25,000	0.084	0.04	0.04	0.04	0.04	0.039	**0.022**	**0.022**	**0.022**	0.03	0.1	0.094

Table C.6
Estimated PID risks for densities 1 − 8.

Density	n	Wand	AIC	BIC	BR	Knuth	SC	RIH	RMG-B	RMG-R	TS	L2CV	KLCV
1	50	0.492	0.696	0.25	0.262	0.428	0.63	0.046	0.108	0.052	0.338	0.112	0.126
	200	0.514	1.09	0.14	0.19	0.246	0.28	0.0	0.052	0.002	0.012	0.426	0.388
	1000	1.124	1.842	0.066	0.356	0.19	0.176	0.0	0.044	0.006	0.0	4.45	4.136
	5000	3.166	3.582	0.216	1.538	0.486	0.426	0.0	0.034	0.012	0.002	24.022	22.716
	25,000	7.394	6.938	0.8	4.326	1.34	1.112	0.0	0.034	0.028	0.004	80.834	75.788
2	50	1.368	0.654	0.052	0.072	0.108	0.34	0.024	0.168	0.02	0.368	0.794	0.864
	200	1.814	0.616	0.012	0.108	0.014	0.05	0.0	0.062	0.004	0.252	2.232	2.178
	1000	3.648	0.626	0.0	0.06	0.0	0.0	0.0	0.036	0.012	0.27	9.586	9.356
	5000	7.708	0.558	0.0	0.038	0.0	0.0	0.0	0.024	0.0	0.47	38.31	37.486
	25,000	15.572	0.662	0.0	0.026	0.0	0.0	0.0	0.0	0.0	0.34	118.566	104.676
3	50	4.932	4.516	3.452	3.474	4.308	4.194	1.37	0.738	2.0	0.84	2.008	2.004
	200	10.596	7.696	4.824	6.352	6.968	6.36	0.6	0.142	1.988	0.308	2.242	2.152
	1000	27.918	19.048	8.706	16.974	14.674	11.84	0.0	0.044	1.132	0.0	4.298	3.908
	5000	75.49	54.402	19.5	50.702	33.506	26.298	0.0	0.068	0.004	0.0	15.022	13.738
	25,000	196.452	154.122	48.028	147.712	83.588	63.848	0.0	0.064	0.008	0.0	52.39	40.55
4	50	3.082	3.256	2.9	2.894	3.132	3.022	1.592	1.554	1.8	1.062	1.94	1.938
	200	8.214	4.148	3.702	3.918	3.958	3.846	1.454	1.39	1.644	0.432	1.578	1.54
	1000	20.592	6.088	5.116	5.474	5.424	5.2	0.904	0.74	1.106	0.092	1.53	1.344
	5000	52.7	19.38	7.302	14.198	7.708	7.294	0.066	0.062	0.276	0.028	6.004	5.14
	25,000	168.888	59.098	13.02	56.224	30.608	15.322	0.0	0.054	0.01	0.0	26.102	19.264
5	50	1.578	1.38	0.874	0.874	1.202	1.212	0.384	0.442	0.402	0.514	0.388	0.452
	200	2.978	2.382	0.704	0.936	1.156	1.018	0.03	0.114	0.046	0.026	0.114	0.088
	1000	7.386	4.472	1.162	2.536	1.978	1.676	0.0	0.034	0.002	0.0	1.566	1.298
	5000	17.622	9.266	2.696	6.878	4.092	3.446	0.0	0.05	0.014	0.0	14.264	12.85
	25,000	38.554	18.642	5.962	15.826	8.36	6.824	0.0	0.076	0.024	0.0	60.984	52.118
6	50	2.072	3.29	1.658	1.754	2.676	3.314	0.942	0.752	1.138	0.83	1.926	1.892
	200	3.206	6.048	2.872	3.854	4.702	5.812	0.418	0.178	0.592	0.284	3.222	3.132
	1000	5.602	17.178	5.074	10.554	8.73	12.602	0.002	0.034	0.004	0.0	7.648	7.402
	5000	11.966	52.07	10.422	40.59	21.156	27.398	0.0	0.046	0.01	0.0	34.92	33.774
	25,000	31.296	153.122	30.11	138.168	57.666	69.232	0.0	0.06	0.026	0.0	118.792	98.536
7	50	5.832	5.664	5.486	5.488	5.568	5.56	5.218	5.24	5.216	5.076	5.388	5.364
	200	5.996	5.32	5.516	5.516	5.306	5.254	4.26	4.26	4.28	3.73	4.818	4.862
	1000	5.548	7.412	4.338	4.776	3.988	3.858	3.276	1.864	2.0	0.05	1.202	0.946
	5000	12.228	16.064	3.204	12.736	6.936	5.874	0.0	0.074	0.01	0.0	11.762	10.34
	25,000	38.996	30.568	9.886	26.936	15.29	13.61	0.0	0.066	0.028	0.002	66.768	57.108
8	50	1.376	1.408	1.356	1.38	1.37	1.422	1.162	1.21	1.206	1.342	1.31	1.33
	200	0.826	1.408	0.854	0.834	0.912	0.958	1.112	1.138	1.13	1.502	0.638	0.644
	1000	1.698	2.658	0.37	0.818	0.586	0.598	1.136	1.102	1.09	0.988	3.934	3.676
	5000	5.262	5.212	0.58	2.94	1.184	1.07	0.968	0.888	0.888	0.104	22.972	21.732
	25,000	12.188	10.952	1.422	7.308	2.392	2.084	0.39	0.114	0.108	0.002	79.588	74.554

Table C.7
Estimated PID risks for densities 9 − 16.

Density	n	Wand	AIC	BIC	BR	Knuth	SC	RIH	RMG-B	RMG-R	TS	L2CV	KLCV
9	50	5.182	5.034	5.186	5.174	5.106	5.026	4.99	4.968	**4.886**	4.948	5.042	4.962
	200	4.574	5.034	4.83	4.682	4.742	4.692	4.476	4.394	4.4	**4.368**	3.6	3.52
	1000	4.326	7.814	3.864	4.642	4.006	4.066	2.928	2.482	2.454	**1.698**	3.42	3.448
	5000	7.61	14.008	3.572	10.462	5.726	5.338	0.906	0.726	0.648	**0.13**	18.086	17.208
	25,000	20.534	36.982	6.222	25.948	9.582	8.848	0.05	0.056	0.016	**0.0**	73.482	67.678
10	50	10.182	10.856	10.892	10.846	10.93	10.864	10.966	10.838	10.968	**10.0**	10.476	10.522
	200	12.484	5.586	10.912	7.986	7.216	5.838	10.904	10.768	10.762	**4.056**	10.074	9.948
	1000	6.906	3.606	3.426	2.18	1.974	2.434	0.628	0.472	**0.16**	0.532	0.38	0.43
	5000	0.964	7.982	0.104	3.924	0.49	0.874	**0.0**	0.172	0.05	0.012	10.444	9.786
	25,000	0.012	17.92	0.018	11.796	0.624	1.012	**0.0**	0.138	0.068	0.002	84.522	80.274
11	50	0.484	0.74	0.132	0.218	0.318	0.654	**0.02**	0.056	0.044	0.332	0.162	0.152
	200	0.378	0.88	0.082	0.144	0.17	0.382	0.002	0.026	**0.0**	0.036	0.976	0.9
	1000	0.462	1.594	0.026	0.294	0.148	0.228	**0.0**	0.034	0.008	0.014	7.152	6.78
	5000	0.812	3.098	0.04	1.058	0.238	0.31	**0.0**	0.034	0.01	0.002	33.902	32.454
	25,000	1.498	6.306	0.04	2.596	0.258	0.35	**0.002**	0.028	0.02	**0.002**	108.578	102.754
12	50	2.846	2.534	2.582	2.586	2.53	2.49	**0.296**	0.396	0.272	0.82	1.508	1.444
	200	0.526	0.384	0.392	0.36	0.36	0.362	**0.044**	0.112	0.06	0.212	0.978	1.024
	1000	0.524	1.72	0.198	1.08	0.684	0.268	**0.0**	0.044	0.002	0.002	0.034	0.19
	5000	3.652	3.784	0.866	2.86	1.35	0.98	**0.0**	0.106	0.002	0.002	3.544	3.274
	25,000	16.588	8.876	1.502	7.492	2.96	1.934	**0.0**	0.054	0.014	0.002	29.402	26.248
13	50	4.708	5.412	5.526	5.524	5.428	5.378	4.738	4.492	5.43	**3.356**	5.464	5.608
	200	5.01	5.584	4.762	5.136	5.414	5.38	1.058	1.232	0.802	**0.588**	3.264	3.322
	1000	4.33	15.188	3.942	12.916	11.226	10.878	**0.0**	0.138	0.008	0.044	2.706	2.2
	5000	16.696	42.412	12.32	36.552	25.872	25.232	**0.0**	0.12	0.016	0.006	17.222	15.536
	25,000	64.008	77.988	26.232	71.332	41.472	40.33	**0.0**	0.116	0.042	0.006	83.18	67.84
14	50	4.886	4.0	4.0	4.0	4.0	4.0	0.012	**0.386**	0.786	0.728	4.0	4.0
	200	4.962	4.568	4.0	4.148	4.26	4.032	0.008	0.288	**0.0**	1.062	4.0	4.0
	1000	4.884	5.194	5.182	5.194	5.194	5.194	**0.0**	0.034	**0.0**	1.378	2.384	2.372
	5000	7.358	3.388	3.388	3.388	3.388	3.388	**0.0**	0.004	**0.0**	1.62	3.61	3.602
	25,000	11.128	5.914	5.914	5.914	5.914	5.914	**0.0**	0.004	**0.0**	2.332	8.258	7.356
15	50	2.142	2.07	2.114	2.124	2.108	2.084	**1.804**	1.85	1.844	1.824	2.794	2.818
	200	2.372	3.648	**1.37**	1.86	2.528	2.756	1.858	1.89	1.842	1.484	1.532	1.482
	1000	7.056	7.612	**0.896**	3.216	2.34	2.616	1.98	2.008	1.96	1.324	6.514	6.224
	5000	17.35	7.138	**0.052**	1.78	0.38	0.384	1.914	1.75	1.738	0.482	30.226	28.682
	25,000	38.954	1.576	**0.0**	0.222	**0.0**	**0.0**	0.506	0.19	0.178	0.164	102.66	91.492
16	50	4.218	3.294	3.502	3.52	3.314	3.29	2.758	2.762	2.912	**2.138**	4.448	4.472
	200	2.68	4.886	2.436	2.84	3.61	3.56	2.896	2.882	2.874	**1.798**	2.114	2.044
	1000	7.692	14.472	3.706	12.306	11.216	11.548	1.738	1.522	1.524	**1.192**	5.562	5.292
	5000	19.014	28.098	12.53	25.878	17.404	18.646	0.686	0.542	**0.436**	1.104	29.646	27.908
	25,000	44.898	28.788	28.602	28.694	28.602	28.602	**0.0**	0.014	0.008	0.79	106.614	89.904

Table C.8
Estimated $\mathbb{L}_2$ risks for densities $1 - 8$ (where applicable).

Density	n	Wand	AIC	BIC	BR	Knuth	SC	RIH	RMG-B	RMG-R	TS	L2CV	KLCV
1	50	**0.165**	0.194	0.195	0.19	0.195	0.189	0.22	0.32	0.217	0.234	0.197	0.197
	200	**0.106**	0.122	0.128	0.123	0.121	0.119	0.15	0.172	0.15	0.119	0.145	0.145
	1000	**0.062**	0.07	0.081	0.071	0.074	0.074	0.099	0.103	0.096	0.068	0.115	0.115
	5000	**0.038**	0.04	0.051	0.042	0.046	0.047	0.063	0.063	0.06	0.039	0.097	0.097
	25,000	**0.022**	0.024	0.032	0.024	0.028	0.029	0.04	0.037	0.037	**0.022**	0.071	0.076
2	50	0.309	0.308	0.209	0.218	0.214	0.243	**0.197**	0.33	0.198	0.259	0.362	0.356
	200	0.197	0.154	0.097	0.108	0.098	0.102	**0.095**	0.134	0.096	0.117	0.325	0.323
	1000	0.117	0.069	**0.043**	0.05	**0.043**	0.044	**0.043**	0.052	0.044	0.053	0.288	0.287
	5000	0.071	0.029	0.019	0.021	0.019	0.019	**0.018**	0.024	**0.018**	0.024	0.253	0.251
	25,000	0.044	0.014	0.009	0.01	0.009	0.009	**0.008**	**0.008**	**0.008**	0.01	0.205	0.196
4	50	0.195	0.224	0.253	0.254	0.227	0.233	0.248	0.363	0.252	**0.216**	0.255	0.262
	200	0.147	0.165	0.202	0.184	0.178	0.189	0.179	0.196	0.184	**0.136**	0.174	0.177
	1000	**0.101**	0.145	0.173	0.152	0.158	0.168	0.125	0.122	0.128	0.11	0.137	0.137
	5000	**0.06**	0.12	0.16	0.136	0.155	0.159	0.08	0.073	0.084	0.063	0.095	0.094
	25,000	**0.037**	0.064	0.157	0.067	0.116	0.151	0.051	0.045	0.054	**0.037**	0.049	0.047
5	50	**0.176**	0.196	0.2	0.2	0.195	0.19	0.225	0.313	0.229	0.218	0.198	0.201
	200	**0.114**	0.127	0.152	0.142	0.136	0.138	0.166	0.222	0.165	0.129	0.134	0.134
	1000	**0.069**	0.08	0.108	0.089	0.094	0.098	0.111	0.11	0.107	0.075	0.084	0.084
	5000	**0.041**	0.049	0.075	0.054	0.064	0.069	0.072	0.069	0.069	0.044	0.064	0.064
	25,000	**0.024**	0.03	0.051	0.032	0.043	0.047	0.045	0.041	0.043	0.025	0.048	0.051
7	50	**0.277**	0.291	0.289	0.287	0.287	0.281	0.286	0.384	0.28	0.329	0.286	0.284
	200	0.237	**0.216**	0.233	0.229	0.224	0.221	0.219	0.239	0.22	0.243	0.235	0.236
	1000	0.171	**0.134**	0.174	0.139	0.142	0.143	0.19	0.188	0.175	0.135	0.146	0.146
	5000	0.082	0.08	0.119	0.085	0.102	0.106	0.109	0.109	0.107	**0.071**	0.099	0.099
	25,000	0.043	0.047	0.076	0.049	0.063	0.067	0.071	0.066	0.066	**0.039**	0.076	0.079
8	50	**0.184**	0.209	0.199	0.197	0.2	0.198	0.2	0.311	0.194	0.219	0.207	0.206
	200	**0.117**	0.132	0.133	0.129	0.128	0.127	0.138	0.168	0.137	0.13	0.152	0.151
	1000	**0.071**	0.077	0.09	0.08	0.082	0.082	0.099	0.101	0.095	0.082	0.117	0.117
	5000	0.042	0.045	0.058	0.046	0.051	0.052	0.06	0.059	0.058	**0.039**	0.097	0.097
	25,000	0.025	0.026	0.036	0.027	0.032	0.033	0.04	0.037	0.037	**0.022**	0.073	0.076

Table C.9
Estimated $\mathbb{L}_2$ risks for densities $9 - 16$.

Density	n	Wand	AIC	BIC	BR	Knuth	SC	RIH	RMG-B	RMG-R	TS	L2CV	KLCV
9	50	**0.215**	0.241	0.233	0.231	0.235	0.231	0.234	0.32	0.231	0.256	0.23	0.231
	200	**0.169**	0.174	0.177	0.172	0.172	0.169	0.188	0.233	0.185	0.19	0.183	0.182
	1000	0.117	**0.11**	0.127	**0.11**	0.114	0.112	0.134	0.139	0.127	0.114	0.129	0.129
	5000	0.074	0.065	0.081	0.066	0.071	0.072	0.081	0.081	0.079	**0.057**	0.101	0.1
	25,000	0.044	0.041	0.053	0.041	0.048	0.049	0.054	0.051	0.05	**0.031**	0.076	0.078
10	50	0.145	0.149	0.136	0.137	0.137	0.142	0.136	0.173	**0.135**	0.209	0.152	0.151
	200	0.131	0.115	0.128	0.123	0.118	**0.11**	0.129	0.136	0.129	0.166	0.136	0.136
	1000	0.123	0.071	0.082	0.071	**0.07**	0.069	0.09	0.105	0.089	0.082	0.08	0.08
	5000	0.069	**0.042**	0.049	0.041	0.043	**0.042**	0.064	0.061	0.058	0.043	0.058	0.058
	25,000	0.027	0.025	0.029	**0.024**	0.025	0.025	0.041	0.038	0.037	**0.024**	0.047	0.047
11	50	**0.256**	0.313	0.306	0.299	0.312	0.307	0.323	0.364	0.326	0.347	0.309	0.308
	200	**0.164**	0.191	0.19	0.185	0.184	0.183	0.229	0.296	0.226	0.2	0.247	0.246
	1000	**0.097**	0.111	0.116	0.108	0.11	0.108	0.156	0.16	0.151	0.117	0.208	0.207
	5000	**0.058**	0.064	0.069	0.062	0.065	0.064	0.1	0.098	0.095	0.071	0.182	0.181
	25,000	**0.034**	0.038	0.041	0.036	0.038	0.038	0.063	0.06	0.058	0.043	0.141	0.146
12	50	0.335	0.302	0.32	0.317	0.3	0.306	**0.23**	0.348	0.239	0.253	0.393	0.393
	200	0.216	0.203	0.219	0.207	0.205	0.209	0.188	0.216	0.188	**0.16**	0.355	0.355
	1000	0.129	0.119	0.178	0.125	0.131	0.156	0.122	0.119	0.12	**0.084**	0.128	0.132
	5000	0.067	0.07	0.107	0.074	0.088	0.099	0.079	0.074	0.076	**0.048**	0.066	0.066
	25,000	0.034	0.041	0.066	0.043	0.055	0.062	0.049	0.044	0.047	**0.027**	0.046	0.047
13	50	0.162	0.15	0.16	0.16	0.148	**0.146**	0.15	0.224	0.155	0.205	0.162	0.162
	200	0.145	0.121	0.138	0.126	0.121	0.12	0.115	2.822	0.116	**0.097**	0.143	0.142
	1000	0.097	0.071	0.096	0.074	0.076	0.077	0.085	1.185	0.08	**0.049**	0.094	0.095
	5000	0.058	0.038	0.057	0.04	0.047	0.048	0.054	0.184	0.049	**0.027**	0.044	0.044
	25,000	0.024	0.022	0.039	0.023	0.031	0.031	0.034	0.038	0.031	**0.015**	0.027	0.027
14	50	1.129	1.124	1.124	1.124	1.124	1.125	**0.521**	0.88	0.756	0.676	1.161	1.161
	200	1.074	1.096	1.095	1.095	1.095	1.095	**0.266**	0.371	0.269	0.358	1.16	1.16
	1000	0.904	1.003	1.003	1.003	1.003	1.004	0.119	**0.122**	0.119	0.162	1.116	1.116
	5000	0.758	0.63	0.63	0.63	0.63	0.632	0.053	0.053	**0.052**	0.074	1.116	1.116
	25,000	0.255	0.027	0.027	0.027	0.028	0.032	0.024	**0.023**	**0.023**	0.038	0.04	0.04
15	50	0.783	0.704	0.766	0.765	0.69	**0.662**	0.753	1.199	0.75	0.829	0.809	0.816
	200	0.556	0.465	0.484	0.465	0.455	0.451	0.395	0.487	**0.401**	0.424	0.425	0.402
	1000	0.393	0.271	0.278	0.263	0.264	0.263	**0.172**	0.189	0.175	0.224	0.314	0.304
	5000	0.274	0.139	0.135	0.134	0.134	0.134	**0.09**	0.093	0.091	0.114	0.263	0.26
	25,000	0.189	0.061	0.06	0.06	0.06	0.06	0.046	**0.042**	**0.042**	0.055	0.187	0.209
16	50	0.857	0.771	0.795	0.795	0.745	0.729	**0.761**	1.309	0.763	0.887	1.001	1.007
	200	0.644	0.579	0.606	0.588	0.576	0.573	**0.47**	0.558	0.472	0.513	0.51	0.477
	1000	0.462	0.348	0.53	0.361	0.373	0.367	0.236	0.249	**0.232**	0.257	0.328	0.315
	5000	0.331	0.187	0.279	0.192	0.248	0.239	0.102	0.104	**0.1**	0.142	0.267	0.262
	25,000	0.197	0.084	0.084	0.084	0.084	0.084	**0.043**	**0.043**	**0.043**	0.071	0.182	0.212

Appendix D. Second simulation study

The following section includes additional notes on the simulation study in Section 5.3.

D.1. Methods used and implementation

To perform the required hypothesis tests for the SI and FM methods, we used the `multimode` R package (Ameijeiras-Alonso et al., 2021). The number of modes for each method was identified by finding the smallest $m \in \mathbb{N}$ at which the null hypothesis $H_0 :$ f has m modes was not rejected against the alternative that f has more than m modes, using a significance level of $\alpha = 0.1$. We used 100 bootstrap samples to compute the p-values for each of the tests. Once the number of modes had been determined for each procedure, the `locmodes` function from the same package was used to estimate their location.

To identify the modes using the GMM method, we followed the proposal of Chacón (2019) by first estimating a Gaussian mixture model using the `mclust` R package (Scrucca et al., 2023), with the number of mixture components and the variance structure chosen by BIC. After estimating the optimal number of components and the corresponding mixture parameters, we used the modal expectation maximization algorithm of Li et al. (2007) to estimate the mode locations.

For the WJ proposal, we used the `ks` R package (Duong, 2007) to compute the plug-in bandwidths for the kernel estimator. The modes of the resulting kernel estimates were calculated by the mean shift algorithm (Fukunaga and Hostetler, 1975), using the implementation in the `LPCM` R package (Einbeck and Evers, 2024).

In addition to the four mode detection methods that we included in our simulation study, our literature search revealed three additional methods that can be used to identify modes, which we ended up excluding from the simulation study for various reasons. The test-based approach of Ameijeiras-Alonso et al. (2019) was excluded on the grounds that the available software implementation (Ameijeiras-Alonso et al., 2021) was too slow to be included in the simulation study, due to the high computation times for testing $H_0 :$ $m = l$ against the alternative for $l \geq 2$. Similarly, we did not include the Bayesian taut spline method of Chacón and Fernández Serrano (2024) because of the high cost of the inference procedure. We also considered including the Hessian-based mode testing method of Genovese et al. (2015). The authors suggest choosing a bandwidth to maximize the number of significant modes found by their testing procedure over a grid of bandwidths, but since the theory supporting their approach is only valid for bandwidths that decay to 0 with the sample size at a sufficiently slow rate, the procedure is prone to fail if the lower bound on the bandwidth grid is chosen too small. In our preliminary test runs we found that the heuristic suggested by the authors to choose the lower bound on the bandwidths works poorly in practice, resulting in the procedure selecting a bandwidth too small and overestimating the number of modes as a result. Due to the lack of a well-calibrated rule for choosing the lower bound, we decided against including this method.

References

Ameijeiras-Alonso, J., Crujeiras, R.M., Rodriguez-Casal, A., 2021. multimode: an R package for mode assessment. J. Stat. Softw. 97 (9), 1–32. https://doi.org/10.18637/jss.v097.i09

Ameijeiras-Alonso, J., Crujeiras, R.M., Rodríguez-Casal, A., 2019. Mode testing, critical bandwidth and excess mass. Test 28, 900–919. https://doi.org/10.1007/s11749-018-0611-5

Barron, A., Schervish, M.J., Wasserman, L., 1999. The consistency of posterior distributions in nonparametric problems. Ann. Stat. 27, 536–561. https://doi.org/10.1214/aos/1018031206

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodol.) 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Birgé, L., Rozenholc, Y., 2006. How many bins should be put in a regular histogram. ESAIM Probab. Stat. 10, 24–45. https://doi.org/10.1051/ps:2006001

Castillo, I., Randrianarisoa, T., 2022. Optional Pólya trees: posterior rates and uncertainty quantification. Electron. J. Stat. 16, 6267 – 6312. https://doi.org/10.1214/22-EJS2086

Celisse, A., Robin, S., 2008. Nonparametric density estimation by exact leave-p-out cross-validation. Comput. Stat. Data Anal. 52, 2350–2368. https://doi.org/10.1016/j.csda.2007.10.002

Chacón, J.E., 2019. Mixture model modal clustering. Adv. Data Anal. Classif. 13, 379–404. https://doi.org/10.1007/s11634-018-0308-3

Chacón, J.E., Fernández Serrano, J., 2024. Bayesian taut splines for estimating the number of modes. Comput. Stat. Data Anal. 196, 107961. https://doi.org/10.1016/j.csda.2024.107961

Davies, P.L., Gather, U., Nordman, D., Weinert, H., 2009. A comparison of automatic histogram constructions. ESAIM Probab. Stat. 13, 181–196. https://doi.org/10.1051/ps:2008005

Davies, P.L., Kovac, A., 2004. Densities, spectral densities and modality. Ann. Stat. 32, 1093–1136. https://doi.org/10.1214/009053604000000364

Davies, P.L., Kovac, A., 2012. ftnonpar: Features and Strings for Nonparametric Regression. R package version 0.1–88. https://cran.r-project.org/src/contrib/Archive/ftnonpar/.

Denby, L., Mallows, C., 2009. Variations on the histogram. J. Comput. Graph. Stat. 18, 21–31. https://doi.org/10.1198/jcgs.2009.0002

Duong, T., 2007. ks: kernel density estimation and kernel discriminant analysis for multivariate data in R. J. Stat. Softw. 21 (7), 1–16. https://doi.org/10.18637/jss.v021.i07

Einbeck, J., Evers, L., 2024. LPCM: Local Principal Curve Methods. R package version 0.47-6. https://doi.org/10.32614/CRAN.package.LPCM

Fisher, N.I., Marron, J.S., 2001. Mode testing via the excess mass estimate. Biometrika 88 (2), 499–517. https://doi.org/10.1093/biomet/88.2.499

Fukunaga, K., Hostetler, L., 1975. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans. Inf. Theory 21 (1), 32–40.

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L., 2015. Non-parametric inference for density modes. J. R. Stat. Soc. Ser. B Stat. Methodol. 78 (1), 99–126. https://doi.org/10.1111/rssb.12111

Genovese, C.R., Wasserman, L., 2000. Rates of convergence for the Gaussian mixture sieve. Ann. Stat. 28, 1105–1127. https://doi.org/10.1214/aos/1015956709

Ghosal, S., 2001. Convergence rates for density estimation with Bernstein polynomials. Ann. Stat. 29, 1264–1280. https://doi.org/10.1214/aos/1013203453

Ghosal, S., Ghosh, J.K., Ramamoorthi, R.V., 1999. Posterior consistency of Dirichlet mixtures in density estimation. Ann. Stat. 27, 143–158. https://doi.org/10.1214/aos/1018031105

Ghosal, S., Ghosh, J.K., van der Vaart, A.W., 2000. Convergence rates of posterior distributions. Ann. Stat. 28, 500–531. https://doi.org/10.1214/aos/1016218228

Ghosal, S., van der Vaart, A.W., 2017. Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press. https://doi.org/10.1017/9781139029834

Guan, Z., 2016. Efficient and robust density estimation using Bernstein type polynomials. J. Nonparametr. Stat. 28, 250–271. https://doi.org/10.1080/10485252.2016.1163349

Guan, Z., Wu, B., Zhao, H., 2008. Nonparametric estimator of false discovery rate based on Bernstein polynomials. Stat. Sin. 18, 905–923.
Hall, P., 1990. Akaike's information criterion and Kullback–Leibler loss for histogram density estimation. Probab. Theory Relat. Fields 85, 449–467. https://doi.org/10.1007/BF01203164
Hall, P., Hannan, E.J., 1988. On stochastic complexity and nonparametric density estimation. Biometrika 75, 705–714. https://doi.org/10.1093/biomet/75.4.705
Härdle, W., 1991. Smoothing Techniques: With Implementation in S. Springer New York, NY. https://doi.org/10.1007/978-1-4612-4432-5
Hedenfalk, I., Duggan, D., Chen, Y., et al., 2001. Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med. 344, 539–548. https://doi.org/10.1056/NEJM200102223440801
Jeong, S., Rockova, V., 2023. The art of BART: minimax optimality over nonhomogeneous smoothness in high dimension. J. Mach. Learn. Res. 24, 1–65.
Kanazawa, Y., 1988. An optimal variable cell histogram. Commun. Stat. Theory Methods 17, 1401–1422. https://doi.org/10.1080/03610928808829688
Klemelä, J., 2012. Bin smoother. Wiley Interdiscip. Rev. Comput. Stat. 4, 384–393. https://doi.org/10.1002/wics.1214
Knuth, K.H., 2019. Optimal data-based binning for histograms and histogram-based probability density models. Digit. Signal Process. 95. https://doi.org/10.1016/j.dsp.2019.102581
Kruijer, W., van der Vaart, A.W., 2008. Posterior convergence rates for Dirichlet mixtures of beta densities. J. Stat. Plan. Inference 138, 1981–1992. https://doi.org/10.1016/j.jspi.2007.07.012
Kwasniok, F., 2021. Semiparametric maximum likelihood probability density estimation. PLoS ONE 16. https://doi.org/10.1371/journal.pone.0259111
Langaas, M., Lindqvist, B.H., Ferkingstad, E., 2005. Estimating the proportion of true null hypotheses, with application to DNA microarray data. J. R. Stat. Soc. Ser. B (Methodol.) 67, 555–572. https://doi.org/10.1111/j.1467-9868.2005.00515.x
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation. Springer New York, NY. 2nd ed. https://doi.org/10.1007/b98854
Li, H., Munk, A., Sieling, H., Walther, G., 2020. The essential histogram. Biometrika 107, 347–364. https://doi.org/10.1093/biomet/asz081
Li, J., Ray, S., Lindsay, B.G., 2007. A nonparametric statistical approach to clustering via mode identification. J. Mach. Learn. Res. 8, 1687–1723.
Liu, L., Li, D., Wong, W.H., 2023. Convergence rates of a class of multivariate density estimation methods based on adaptive partitioning. J. Mach. Learn. Res. 24, 1–64.
Lu, L., Jiang, H., Wong, W.H., 2013. Multivariate density estimation by Bayesian sequential partitioning. J. Am. Stat. Assoc. 108, 1402–1410. https://doi.org/10.1080/01621459.2013.813389
Marron, J.S., Wand, M.P., 1992. Exact mean integrated squared error. Ann. Stat. 20, 712–736. https://doi.org/10.1214/aos/1176348653
Massart, P., 2007. Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003. Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-48503-2
Mendizábal, V.Z., Boullé, M., Rossi, F., 2023. Fast and fully-automated histograms for large-scale data sets. Comput. Stat. Data Anal. 180. https://doi.org/10.1016/j.csda.2022.107668
Mildenberger, T., Rozenholc, Y., Zasada, D., 2019. Histogram: Construction of Regular and Irregular Histograms with Different Options for Automatic Choice of Bins. R package version 0.0–25. https://doi.org/10.32614/CRAN.package.histogram
Müller, D.W., Sawitzki, G., 1991. Excess mass estimates and tests for multimodality. J. Am. Stat. Assoc. 86, 738–746. https://doi.org/10.1080/01621459.1991.10475103
Phipson, B., 2013. Empirical Bayes Modelling of Expression Profiles and their Associations.]bibinfotypePh.D. thesis. Department of Mathematics and Statistics, University of Melbourne.
Rozenholc, Y., Mildenberger, T., Gather, U., 2010. Combining regular and irregular histograms by penalized likelihood. Comput. Stat. Data Anal. 54, 3313–3323. https://doi.org/10.1016/j.csda.2010.04.021
Rudemo, M., 1982. Empirical choice of histograms and kernel density estimators. Scand. J. Stat. 9, 65–78.
Schilling, R.L., 2005. Measures, Integrals and Martingales. Cambridge University Press. https://doi.org/10.1017/CBO9780511810886
Scott, D.W., 1979. On optimal and data-based histograms. Biometrika 66, 605–610. https://doi.org/10.1093/biomet/66.3.605
Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley. https://doi.org/10.1002/9780470316849
Scott, D.W., 2010. Histogram. Wiley Interdiscip. Rev. Comput. Stat. 2, 44–48. https://doi.org/10.1002/wics.59
Scricciolo, C., 2007. On rates of convergence for Bayesian density estimation. Scand. J. Stat. 34, 626–642. https://doi.org/10.1111/j.1467-9469.2006.00540.x
Scrucca, L., Fraley, C., Murphy, T.B., Raftery, A.E., 2023. Model-Based Clustering, Classification, and Density Estimation Using mclust in R. Chapman and Hall/CRC. https://doi.org/10.1201/9781003277965
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B (Methodol.) 53, 683–690. https://doi.org/10.1111/j.2517-6161.1991.tb01857.x
Silverman, B.W., 1981. Using kernel density estimates to investigate multimodality. J. R. Stat. Soc. Ser. B (Methodol.) 43 (1), 97–99. https://doi.org/10.1111/j.2517-6161.1981.tb01155.x
Simensen, O.H., 2025. AutoHist.jl: a Julia package for fast and automatic histogram construction. J. Open Source Softw. 10, 8850. https://doi.org/10.21105/joss.08850
Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. 100, 9440–9445. https://doi.org/10.1073/pnas.1530509100
Taylor, C.C., 1987. Akaike's information criterion and the histogram. Biometrika 74, 636–639. https://doi.org/10.1093/biomet/74.3.636
Tsybakov, A.B., 2009. Introduction to Nonparametric Estimation. Springer New York, NY. https://doi.org/10.1007/b13794
Venables, W.N., Ripley, B.D., 1999. Modern Applied Statistics with S-PLUS. Springer New York, NY. 3rd ed. https://doi.org/10.1007/978-1-4757-3121-7
Wand, M.P., 1997. Data-based choice of histogram bin width. Am. Stat. 51, 59–64. https://doi.org/10.2307/2684697
Wand, M.P., 2021. KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995). R package version 2.23-20. https://doi.org/10.32614/CRAN.package.KernSmooth
Wand, M.P., Jones, M.C., 1994. Kernel Smoothing. CRC press. https://doi.org/10.1201/b14876
Wong, W.H., Ma, L., 2010. Optional Pólya tree and Bayesian inference. Ann. Stat. 38, 1433 – 1459. https://doi.org/10.1214/09-AOS755
Wong, W.H., Shen, X., 1995. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. Ann. Stat. 23, 339–362. https://doi.org/10.1214/aos/1176324524