

Random Histograms

Automatic Histogram Construction via Bayesian Nonparametrics

Oskar Høgberg Simensen

SMR: Statistics
60 ECTS study points

Department of Mathematics
Faculty of Mathematics and Natural Sciences

Oskar Høgberg Simensen

Random Histograms

Automatic Histogram Construction via Bayesian
Nonparametrics

Supervisors:

Nils Lid Hjort, Dennis Christensen & Ida Scheel

Abstract

This thesis approaches density estimation from a nonparametric Bayesian viewpoint. In particular, we consider two estimators based on a piecewise constant approximation of the true density, which can be used to construct a histogram using Bayesian decision theory. We also propose a third model based on linear spline densities. We give details on the asymptotic behavior of the posterior distribution in all the models under study by proving consistency- and convergence rate results. Additionally, algorithms are provided for posterior inference in all three models. We also conduct a simulation study to investigate the finite-sample performance of the two histogram procedures.

Contents

1	Introduction	1
2	Background	5
	2.1 Background and notation	5
	2.1.1 Mathematical background	5
	2.1.2 Notation	6
	2.1.3 A note on estimated support	7
	2.2 Likelihoods and frequentist estimation	7
	2.2.1 Likelihood-based inference	7
	2.2.2 Numerical optimization of log-likelihoods: the EM algorithm	9
	2.2.3 Model selection	11
	2.2.4 Parametric and nonparametric models	12
	2.3 Loss functions and statistical risk	13
	2.3.1 Metric and normed spaces	14
	2.4 Asymptotics	18
	2.4.1 Convergence of random variables	18
	2.4.2 Stochastic order symbols	19
	2.4.3 Consistency and convergence rates	20
3	Bayesian inference	23
	3.1 Basics of Bayesian inference	23
	3.1.1 Posterior inference in nonconjugate models	25
	3.1.2 Bayesian model selection	29
	3.1.3 Bayesian decision theory	31
	3.2 Nonparametric Bayes	32
	3.2.1 The Kullback-Leibler support condition	34
	3.2.2 Various notions of consistency	35
	3.2.3 Covering and bracketing numbers	36
	3.2.4 Establishing strong convergence	38
	3.2.5 Posterior convergence rates	39
4	Regular Histograms	41
	4.1 Automatic regular histogram procedures	42
	4.1.1 Rules for selecting the number of bins	42
	4.1.2 Choice of loss function for histogram estimation	44
	4.2 A Bayesian approach to histograms	46
	4.2.1 Computation in the random histogram model	47
	4.3 Prior support under random histogram priors	51
	4.3.1 Assumptions and notation	51
	4.3.2 The master theorem	52
	4.3.3 Proving the master Theorem	53

4.4	Posterior consistency	58
4.4.1	Posterior consistency of regular histogram priors	58
4.5	Consistency of the histogram point estimator	61
4.6	Convergence rate under regular random histogram priors	66
4.6.1	Approximation properties of piecewise constant densities	66
4.6.2	Convergence rates of Bayesian histogram methods	68
4.7	Dirichlet-based Histograms	73
4.7.1	Choice of prior for the bin probabilities	73
4.7.2	Choice of prior for the number of bins	79
4.8	Simulation study: regular histograms	80
5	Irregular Histograms	87
5.1	An introduction to irregular histograms	87
5.1.1	Methods of irregular histogram construction	89
5.2	Bayesian histograms with irregular grids	90
5.2.1	Our irregular histogram prior	90
5.2.2	Computing the optimal partition by dynamical programming	94
5.2.3	Computing the posterior mean and uncertainty quantification	98
5.3	Posterior Consistency of Bayesian irregular histograms	101
5.4	Consistency of irregular histogram point estimation	109
5.5	Posterior rate of convergence of irregular random histograms	114
5.6	Simulation study: Irregular histograms	119
5.6.1	Software implementations used in the simulation study	119
5.6.2	Choice of discretization schemes	121
5.6.3	Irregular histograms: results	122
5.6.4	A comparison between regular and irregular histograms	126
6	Linear Spline Priors	129
6.1	Beyond histogram models	129
6.2	Bayesian inference for the linear spline model	131
6.2.1	Metropolis–Hastings for posterior inference	132
6.2.2	Validating the Metropolis–Hastings output	135
6.3	Penalized Maximum Likelihood Estimation	136
6.4	Big support and posterior consistency	138
6.4.1	Big support	138
6.4.2	Posterior consistency	140
6.5	Posterior convergence rates	141
6.5.1	An approximation result	142
6.5.2	Posterior convergence rate	143
6.6	Applications to real world data sets	146
7	Conclusion	149
A	Appendix A.	152
A.1	Technical proofs	152
B	Appendix B.	160
C	Appendix C	162
C.1	Test densities	162
C.2	Results from simulations: Regular Histograms	163
C.3	Results from simulations: Irregular Histograms	165
D	Appendix D	167
D.1	Diagnostics for MCMC	167
D.2	Model diagnostics for examples	168

Preface

This thesis started out with the rather open problem of analyzing a particular Bayesian model that can be used to construct a regular histogram. The main motivation for studying this statistical model was its computational simplicity, but it also proved to be a fruitful starting point for asymptotic analyses. Along the way, I developed two other models that in many ways can be seen as modifications of the framework introduced by the regular histogram model, but where model fitting is not as straightforward as in the original model, requiring the use of more sophisticated computational tools. The mathematical results developed for the regular histogram model turned out to be rather useful when studying these new models, as the proofs proceed in a similar fashion with minor modifications.

I very much enjoyed the process of writing this thesis, as I was able to pursue not only computational details in the models under study, but also mathematical limit results in a large-sample setting. This mix of mathematics and programming is a big part of the reason why I decided to pursue statistics as a field of study to begin with, and I am overall very happy that I got to explore both of these aspects in my thesis. I can only hope that the reader gets as much enjoyment out of reading this text as I had writing it.

A big thank you goes out to my supervisors, Nils Lid Hjort, Dennis Christensen and Ida Scheel. In particular, I am much indebted to Nils and Dennis for providing me with an interesting thesis topic, giving me useful feedback during the writing process and for bringing related work to my attention along the way.

Finally, I would like to thank my family for supporting me during my time writing the thesis. In particular, I thank my wife, Dorothea Simensen, for listening to my rants about topics she is not particularly familiar with and for providing me with much-needed emotional support. This thesis would not have turned out as well as it did without you by my side.

Chapter 1

Introduction

Statistics is at heart the science of making inferences from data, providing useful summaries that simplify a given data set at hand, often in an interpretable manner. Statistical methods are routinely employed to solve problems from a variety of different fields, and can be used to tackle diverse tasks such as estimating the effect of an additional year of schooling on a country's GDP or classifying whether or not an e-mail is likely to be spam. An often encountered task is to provide an informative visual display of the distribution of a given data set. Such visual displays are often used in exploratory data analyses, which, due to their exploratory nature, require the use of procedures that work well with minimal model assumptions. When faced with such a problem, most non-statisticians would likely resort to drawing a histogram. Histograms are notable for their mathematical simplicity; their construction is so simple that they are routinely taught to students with limited mathematical backgrounds.¹ Due to this fact, histograms have remained popular despite the fact that more sophisticated methods of density estimation have appeared in recent decades, and their continued use seems to be a virtual guarantee.

The histogram has a long and rich history, dating back to the work of John Graunt in the 17th century (Scott, 1992). Even though it is the oldest density estimator, histograms are often still taught in courses that cover nonparametric density estimation, as their simple construction makes them particularly suited for mathematical analysis. Given a sample $x_1, x_2, \dots, x_n$ of data points, the first step to constructing a standard histogram is to partition the interval $[\min_i x_i, \max_i x_i]$ into k disjoint intervals or bins $I_1, I_2, \dots, I_k$. Once the partition has been chosen, constructing a histogram over a given interval partition is a simple task; all one has to do is to merely count the number of observations in each bin. In practice so-called *regular partitions*, one where each interval is of equal length, are commonly employed to simplify the issue of having to choose a partition based on the data to one of selecting the number of bins. Letting N_j denote the number of observations in cell j , the histogram estimate of the density of the sample is simply

$$\hat{f}_n(x) = \frac{N_j}{n|I_j|}, \quad \text{for } x \in I_j, \quad (1.1)$$

where $|I_j|$ denotes the length of interval j .

While the estimator in (1.1) is easy to compute and simple to understand, we have refrained from addressing one key question until now: how do we select the partition upon which the histogram should be drawn? By restricting our attention to regular partitions, the problem simplifies somewhat. However, this still leaves us with the problem of selecting the number k of bins used for the regular partition. It is easy to see

¹The main author has personal experience in teaching histograms to Norwegian high school students.

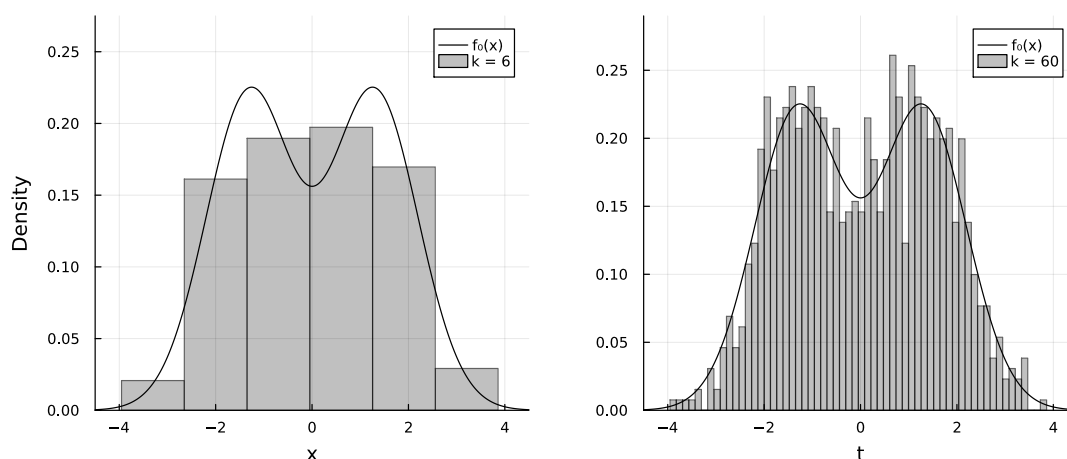


Figure 1.1: An oversmoothed and an undersmoothed histogram drawn based on the same simulated sample.

why the question of choosing an appropriate number of bins is of great importance; too coarse a partition results in an estimate which is too smooth, causing one to potentially miss key features of the given data set. On the other hand, choosing too many bins will lead to many spurious bumps in the drawn histogram, highlighting features where there are none. To illustrate these tradeoffs in practice, we simulated a sample of size $n = 1000$ from a symmetric mixture of normal distributions, and plotted the true density and two regular histograms with $k = 6$ and $k = 60$ bins for the generated sample. The results are shown in Figure 1.1. The choice of $k = 6$ is seen to produce an oversmoothed histogram, failing to identify the bimodal nature of the density, while the $k = 60$ histogram is too bumpy, with many spurious modes. An appropriate bin width for this sample will lie somewhere in between these two extremes.

Practitioners are likely to rely on default procedures in software when first drawing a histogram. This creates a need for automatic procedures that choose the number of bins from the sample with no input from the user. However as it turns out, there is no universally accepted answer as to how one should select the number of bins even for regular histograms, with many different proposals being provided in the statistical literature, see for instance Birgé and Rozenholc (2006) for a brief review. It is upon reading this literature that one realizes that the construction of histograms is a lot more complicated than what one might have thought at first. The proposed rules for selecting the number of bins range from simple rules-of-thumb to more sophisticated methods, such as those based on cross-validation or information-theoretic model selection criteria. Even to trained statisticians, it is not clear which procedure one should choose when drawing a histogram.

To motivate the need for good default histogram procedures, we will consider a real-life dataset from the stock market ². The S&P 500 index tracks the value of the stocks of the 500 largest companies listed on stock exchanges in the United States. For our present analyses we consider the daily log-returns to the stock index, spanning the 5-year period from 25/02/2020 to 24/02/2025 for a total of $n = 1256$ observations. To analyze the data, we use the Python library `matplotlib` (Hunter, 2007). The package provides functionality for plotting a histogram with a default choice for the number

²Data retrieved from Nasdaq (2025).

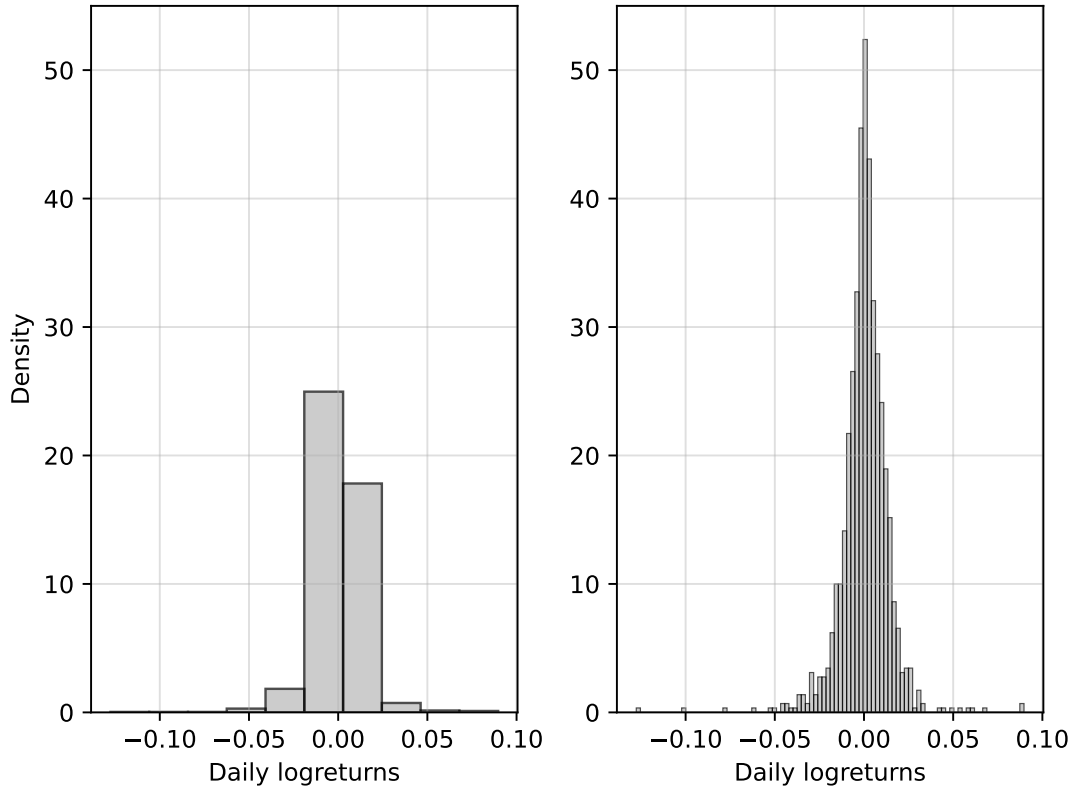


Figure 1.2: Two histograms of daily logreturns of the S&P 500 index using the default procedure in matplotlib (left) and the 'auto' procedure (right)

of bins. The resulting histogram can be seen on the left of Figure 1.2. Upon reading the documentation (Harris et al., 2020), one learns that the default procedure draws a histogram with $k = 10$ bins, with no reference to the sample at all! This procedure seems rather odd, as one would generally expect the number of bins used to increase with the sample size, showing finer details in the data as more samples make themselves available. Further inspection of the documentation reveals that matplotlib also provides an automatic procedure named 'auto', which in this case corresponds to the data-based rule proposed by Freedman and Diaconis (1981). By specifying that the number of bins should be chosen according to this procedure we obtain the histogram on the right in Figure 1.2. The 'auto' procedure produces a histogram consisting of $k = 95$ bins instead of the default $k = 10$, an almost ten-fold difference! The difference in the two histograms are rather immediate as the histogram on the right shows more of the detail in the data and has a much more pronounced central peak than the one on the left. Again, this example stresses the importance of having a well-working rule for selecting the number of bins.

Scope of the thesis

The first major objective of this thesis is to analyze different automatic histogram procedures, that is, procedures that can be used to draw a histogram without requiring the user to specify the values of any hyperparameters. Our main contribution is to propose modifications of a previously studied Bayesian method for regular histograms and to provide a novel Bayesian irregular histogram construction. We compare different

automatic histogram procedures through the use of simulations, where the focus is both to produce a good estimate of the density as measured by classical loss functions, but also to identify key features such as the number and locations of modes. Some of our work is also of a more theoretical nature, as we analyze the performance of the proposed Bayesian methods with respect to various metrics in the large-sample setting. In particular, we derive consistency results and find convergence rates for our estimators under suitable smoothness assumptions on the data-generating density.

Various mathematical and statistical concepts are used quite heavily throughout the text. Chapter 2 provides an introduction to the mathematical and statistical theory needed in later chapters. Since our own work is mostly from a Bayesian perspective, we made Chapter 3 to cover the most important parts of Bayesian inference, with the second section dedicated to the somewhat more exotic field of Bayesian nonparametrics. The theory is illustrated with examples from the world of histogram estimation. These two chapters combined serve as background material to be drawn on throughout the subsequent chapters. In each of the remaining chapters, we will mostly study one particular class of methods, making comparisons to the methods of previous chapters when appropriate. In Chapter 4, we focus on regular histograms. We introduce many of the most popular methods for constructing regular histograms, and study a Bayesian method, with particular focus on computations. The second part of the chapter is of a more theoretical nature, as we establish consistency results and rates of convergence for the Bayesian procedure. In the final part of the chapter, we study the performance of the method through simulations and compare it to other estimators from the regular histogram literature. Chapter 5 will deal with irregular histogram constructions. In addition to the theoretical analyses, we also provide a simulation study to compare the performance of regular and irregular histogram procedures. In Chapter 6, we propose a new Bayesian continuous density estimator based on a piecewise linear model of the true density, sharing a similar structure to the regular histograms of Chapter 4, but which yields continuous density estimates. We prove asymptotic results for the given estimator and show how inferences for the data-generating density can be carried out via simulations. Our thesis ends with a brief discussion of the work presented here, with a mention of possible extensions and improvements of our work that could serve as pointers for future research into the topic.

The algorithms for fitting the statistical models of Chapters 4 to 6 have all been implemented by the main author in the **Julia** language (Bezanson et al., 2017). The code used to produce the figures and tables included in this thesis is publicly available as part of a github repository, available at <https://github.com/oskarhs/Density-Estimation-and-Regression-by-Piecewise-Constant-Procedures>. The core functionality used to fit the regular and irregular random histogram models of Chapters 4 and 5 are also available as part of the **AutoHist.jl** package (Simensen, 2025).

Chapter 2

Background

This chapter provides a brief introduction to basic definitions and statistical concepts used throughout the text. In the first section we spend some time discussing recommended prior knowledge and briefly introduce the notation used in the thesis. A short introduction to some important mathematical concepts are also given. The second section deals with issues pertaining to statistical estimation. We start by introducing likelihood-based methods for statistical inference and cover related model selection procedures. Other issues covered include parametric and nonparametric procedures, statistical risk and asymptotics. We illustrate the theory with a running example of estimating a density based on a histogram.

2.1 Background and notation

We introduce some necessary concepts and notation to be used later in the text. In particular we give a rather general description of our notation, discuss statistical estimation problems and introduce tools for studying the asymptotic behavior of random variables. The theory is illustrated by applying it in the context of histogram estimation, briefly explaining the theoretical justifications of different histogram procedures found in the statistical literature.

2.1.1 Mathematical background

As with almost any statistical text, some prior mathematical and statistical knowledge is required to get the most out of reading this thesis. To fully appreciate the material covered in this thesis, some familiarity with real analysis is helpful. In particular, one should be comfortable with the concepts of metrics, norms and convergence. We do not assume any detailed knowledge of measure theory, although Lebesgue's Dominated Convergence Theorem will come in handy at several points in the text. We have gathered those measure-theoretic results we will take for granted throughout the text in Appendix B. Prior exposure to mathematical statistics will in no doubt also be an advantage. In particular, the reader should be familiar with the different notions of convergence of random variables and the Law of Large Numbers. Some familiarity with Bayesian statistics is advantageous, although we do cover the basics of Bayesian inference in Section 3.1. Knowledge of nonparametric statistics, both frequentist and Bayesian, might also come in handy, but we shall cover the necessary concepts here in sufficient detail so that no prior knowledge is required.

2.1.2 Notation

We will generally write vectors in boldface, so that $\mathbf{x} = (x_1, x_2, \dots, x_k)$ is a vector in $\mathbb{R}^k$ for a positive integer k . This text is primarily concerned with density estimation problems and we will generally write $\mathbf{x}_{1:n}$ for the observed sample $\mathbf{x}_1, \dots, \mathbf{x}_n$. We denote functions $h: \mathcal{X} \rightarrow \mathbb{R}$, by either h or $h(\cdot)$. We generally use $\mathcal{X}$ to denote a subset of $\mathbb{R}$, unless otherwise specified. In particular, if h is a function depending on a parameter vector $\boldsymbol{\lambda}$, the notation $h(\cdot | \boldsymbol{\lambda})$ refers to the function when the parameter vector takes the value $\boldsymbol{\lambda}$. We shall use $\mathcal{I}(\cdot)$ to denote the indicator function, defined by

$$\mathcal{I}(x \in \mathcal{A}) = \begin{cases} 1 & \text{if } x \in \mathcal{A}, \\ 0 & \text{else,} \end{cases}$$

where $\mathcal{A}$ is a set. If $\mathcal{A}$ is a discrete set, we let $|\mathcal{A}|$ denote the cardinality of $\mathcal{A}$, that is, the number of elements in $\mathcal{A}$, while for an interval I , we write $|I|$ for the Lebesgue measure of the set I . For a positive-valued function $h: \mathcal{X} \rightarrow \mathbb{R}$ and a real number γ , we write h^γ to mean the function obtained by raising h to the power γ , e.g. $h^2(x) = \{h(x)\}^2$ for all $x \in \mathcal{X}$, and it does not refer to the composition of h with itself. Throughout the text we adopt the convention that $0 \times \log 0 = 0$, in line with the well-known limit result $\lim_{x \downarrow 0} x \log(x) = 0$. As far as abbreviations are concerned, we adopt the standard moniker iid. to mean independent and identically distributed, and use pdf. and pmf. for probability density function and probability mass function, respectively. Statements that are said to hold almost everywhere are always made with respect to the Lebesgue measure on a suitable subset of $\mathbb{R}^k$, and are abbreviated a.e. Likewise, a statement of measurability is always made with respect to the Lebesgue measure, unless otherwise specified. Since we shall work with densities with respect to the Lebesgue measure throughout the entire text, the notation $f = g$ is taken to mean that the two densities f and g are equal a.e. For a random vector $\mathbf{z}$, we write $\mathbb{E}_{\mathbf{z}}[\cdot]$ for the expectation taken with respect to the distribution of $\mathbf{z}$. For a given parametric family of distributions $\mathcal{D}$ indexed by the parameter $\boldsymbol{\theta}$, we will write $\mathcal{D}(x | \boldsymbol{\theta})$ for the corresponding density when the parameter is equal to $\boldsymbol{\theta}$. As an example, we will write $\mathcal{N}(x | \mu, \sigma^2)$ for the density of the normal distribution with mean μ and variance σ^2 . For sequences of elements belonging to a set $a_i \in \mathcal{A}$ we write $\{a_i\}_{i=1}^k$ for $k \in \mathbb{N} \cup \{\infty\}$. By convention, we will often merely write $\{a_i\}$ for countably infinite sequences, corresponding to $k = \infty$. For most of the text we shall be working with probability measures defined on the Borel sigma algebra on $[0, 1]$. As such, statements like ‘Let P_0 be a probability measure on the unit interval’ should be interpreted as P_0 being defined on the measure space $([0, 1], \mathcal{B}([0, 1]))$. Throughout the thesis, we will frequently work with parameter vectors that represent probabilities. In a random experiment with k different outcomes, we write p_j for the probability of outcome j . We can then collect all these probabilities into a vector $\mathbf{p} = (p_1, \dots, p_k)$. Since the components of $\mathbf{p}$ are probabilities of disjoint events, it follows from the axioms of probability that $p_j \geq 0$ for all j and that $\sum_{j=1}^k p_j = 1$. The set on which such probability vectors $\mathbf{p}$ live is called the $k - 1$ -dimensional *simplex*, which we denote by $\mathcal{S}_{k-1} \subseteq \mathbb{R}^k$, where

$$\mathcal{S}_{k-1} = \left\{ \boldsymbol{\pi} \in [0, 1]^k : \sum_{j=1}^k \pi_j = 1 \right\}. \quad (2.1)$$

2.1.3 A note on estimated support

Throughout this thesis, we will present our models for random samples $x_1, \dots, x_n$ on $[0, 1]$. These models can be easily generalized to samples in another known compact interval. Although our methods are presented in this context, they can easily be applied to samples with unknown support by estimating the support of the data-generating density. Our preferred approach is to transform the sample into the unit interval via $z_i = (x_i - x_{(1)}) / (x_{(n)} - x_{(1)})$, calculate the density estimator on $z_1, \dots, z_n$, and then scale the density estimator back to the original interval, producing an estimated density supported on $[x_{(1)}, x_{(n)}]$. This is also the approach most commonly used to draw a histogram of a sample from a distribution whose support is unknown. For an overview of alternative approaches to estimating the support of an unknown density, we refer the reader to Zhong (2016, Section 1.4).

We note that other approaches could in principle also be applied to obtain a sample on the unit interval, such as the transformation $z_i = F(x_i)$, where $F(\cdot)$ is a strictly increasing cumulative distribution function, although we do not explore this approach further in this thesis.

2.2 Likelihoods and frequentist estimation

2.2.1 Likelihood-based inference

An important task in statistics is that of making inferences about unobserved parameters from data. Given an observed sample $\mathbf{x}_{1:n} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, we would like to make inferences about a related parameter, denoted $\boldsymbol{\theta}$, belonging to some suitable parameter space Θ . Typical problems the statistician deals with is designing good point estimates of $\boldsymbol{\theta}$, and quantifying the uncertainty in the knowledge of the parameter. Many statistical procedures are based on formulating a probabilistic model for the data, where $\mathbf{x}_{1:n} \sim p(\mathbf{x}_{1:n} | \boldsymbol{\theta})$ for some pdf or pmf $p(\cdot | \boldsymbol{\theta})$ parametrized by the parameter $\boldsymbol{\theta}$. In the likelihood-based paradigm, inferences are based on the *likelihood function*

$$L_n(\boldsymbol{\theta}) = p(\mathbf{x}_{1:n} | \boldsymbol{\theta}). \quad (2.2)$$

In particular, if the sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ is independent the likelihood function (2.2) becomes

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{x}_i | \boldsymbol{\theta}), \quad (2.3)$$

where $p(\mathbf{x}_i | \boldsymbol{\theta})$ is the pdf./pmf. of $\mathbf{x}_i$. It is often more convenient to work with the *log-likelihood function*

$$l_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}). \quad (2.4)$$

This is especially the case for independent data, where the log-likelihood can be written as $l_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(\mathbf{x}_i | \boldsymbol{\theta})$. A popular approach to estimating the parameter $\boldsymbol{\theta}$ is the *method of maximum likelihood*, which is based on maximizing the likelihood function for the obtained sample $\mathbf{x}_{1:n}$ with respect to the parameter $\boldsymbol{\theta}$, yielding the *maximum likelihood estimator* (MLE),

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} L_n(\boldsymbol{\theta}). \quad (2.5)$$

Since the logarithm is a monotone transformation, this is equivalent to the problem of maximizing the log-likelihood (2.4). Popularized by Ronald Fisher (Aldrich, 1997), the method of maximum likelihood is based on the principle that values of $\boldsymbol{\theta}$ for which the likelihood $L_n(\boldsymbol{\theta})$ is large are more plausible.

Example 2.1 (Maximum likelihood histogram). Consider the problem of estimating the density of an iid sample $\mathbf{x}_{1:n} = (x_1, \dots, x_n) \sim f$ on the unit interval. A possible model for the distribution f of the given dataset is a piecewise constant one. We start by partitioning $[0, 1]$ into k subintervals $I_{1:k} = (I_1, I_2, \dots, I_k)$, not necessarily of equal length. Since the density $f(\cdot)$ is piecewise constant over each of the k intervals, we can parametrize it by the values it takes on each interval $f_1, f_2, \dots, f_k$, so that $f(x) = f_j$ if $x \in I_j$. This means that we can write

$$f(x) = \sum_{j=1}^k f_j \mathcal{I}(x \in I_j) = \prod_{j=1}^k f_j^{\mathcal{I}(x \in I_j)}.$$

Alternatively, we can parametrize the given $f(\cdot)$ by the *bin probabilities* $p_j = \int_{I_j} f(x) dx$ instead. A brief computation shows that

$$p_j = \int_{I_j} f(x) dx = \int_{I_j} f_j dx = f_j |I_j|.$$

Rearranging, one obtains $f_j = p_j / |I_j|$ so we may represent $f(\cdot)$ as

$$f(x) = \prod_{j=1}^k \left(\frac{p_j}{|I_j|} \right)^{\mathcal{I}(x \in I_j)}. \quad (2.6)$$

For $f(\cdot)$ to be a probability density, we also require that $f(x) \geq 0$ for all $x \in [0, 1]$, which is equivalent to $p_j \geq 0$ for any j . Moreover, as f must integrate to one, we find that

$$\int_0^1 f(x) dx = \sum_{j=1}^k \int_{I_j} f(x) dx = \sum_{j=1}^k \int_{I_j} \frac{p_j}{|I_j|} dx = \sum_{j=1}^k p_j = 1,$$

so the probability vector $\mathbf{p}$ must sum to one. Thus, we see that $\mathbf{p}$ belongs to the simplex $\mathcal{S}_{k-1}$ (2.1). Since f can be parametrized in terms of the partition $I_{1:k}$ and the corresponding simplex vector $\mathbf{p}$, we introduce the notation $f(\cdot) = f(\cdot | k, I_{1:k}, \mathbf{p})$ for the piecewise constant density given in (2.6). If the partition $I_{1:k}$ is uniquely determined by the value of k , we omit it in our notation, writing $f(\cdot | k, \mathbf{p})$ instead for the histogram density.

From (2.3) and (2.6), we derive the likelihood function:

$$L_n(\mathbf{p} | k, I_{1:k}) = \prod_{i=1}^n f(x_i | k, I_{1:k}, \mathbf{p}) = \prod_{i=1}^n \prod_{j=1}^k \left(\frac{p_j}{|I_j|} \right)^{\mathcal{I}(x_i \in I_j)} = \prod_{j=1}^k \prod_{i: x_i \in I_j} \frac{p_j}{|I_j|} \quad (2.7)$$

From (2.7), we learn that the likelihood contribution from any $x_i \in I_j$ is $p_j / |I_j|$. Denoting the number of observations falling into interval I_j as $N_j = \sum_{i=1}^n \mathcal{I}(x_i \in I_j)$ for $j = 1, \dots, k$, we can write the likelihood more compactly as

$$L_n(\mathbf{p} | k, I_{1:k}) = \prod_{j=1}^k \left(\frac{p_j}{|I_j|} \right)^{N_j}. \quad (2.8)$$

Maximizing the corresponding log-likelihood under the constraint that $\mathbf{p} \in \mathcal{S}_{k-1}$, we find that the maximum likelihood estimate is given as a solution to the problem

$$\hat{\mathbf{p}} = \arg \max_{\mathbf{p} \in \mathcal{S}_k} \sum_{i=1}^n N_j \log(p_j / |I_j|). \quad (2.9)$$

Solving (2.9) may be done using the method of Lagrange multipliers, yielding $\hat{p}_j = N_j/n$ for all j . The resulting estimated density is

$$f(x | k, I_{1:k}, \hat{\mathbf{p}}) = \prod_{j=1}^k \left(\frac{\hat{p}_j}{|I_j|} \right)^{\mathcal{I}(x \in I_j)} \quad (2.10)$$

This proves that the standard histogram is in fact a maximum likelihood estimator among the class of piecewise constant densities. $\blacktriangle$

From (2.10), it is clear that in the setting where the data is generated according to a true density f_0 , the histogram estimator based on the partition $I_{1:k}$ is aiming for a piecewise constant variant of f_0 , as the observed bin frequencies $\hat{p}_j$ are natural estimators of the bin probabilities π_j . We formalize the concept of a piecewise constant approximation in Definition 2.1.

Definition 2.1 (Piecewise constant version). For an integrable function $h: \mathcal{X} \rightarrow \mathbb{R}$, we write $h_{I_{1:k}}$ for the *piecewise constant version of h* based on the interval partition $I_{1:k} = (I_1, \dots, I_k)$ of $\mathcal{X}$, that is,

$$h_{I_{1:k}}(x) = \sum_{j=1}^k h_{k,j} \mathcal{I}(x \in I_j), \quad h_{k,j} = |I_j|^{-1} \int_{I_j} h(s) \, ds. \quad (2.11)$$

2.2.2 Numerical optimization of log-likelihoods: the EM algorithm

The setup of (2.1) is somewhat special in that the optimization problem was analytically tractable, and we were able to find an explicit expression for the maximum likelihood estimator. In general, this will not be the case in maximum likelihood problems, such as the density estimator of Section 6.3, and one has to resort to numerical optimization techniques in order to compute the MLE. Although there are numerous optimization techniques that can be used to maximize the log-likelihood $\log L_n(\boldsymbol{\theta})$, we will here focus our attention on the *Expectation Maximization algorithm* (EM algorithm). Our presentation of the EM algorithm focuses mostly on the problem of estimating a mixture density as in Example 2.2, but the method is applicable in a more general setting. We refer the reader to Robert and Casella (2004) for a more detailed account of the EM algorithm, its properties and other use cases.

The EM algorithm is an iterative method, forming a sequence $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(T)}$ of approximations to the maximum likelihood estimator through successive solution of related optimization problems. Given an observed iid sample $\mathbf{x}_{1:n}$ from $f(\cdot | \boldsymbol{\theta})$, the EM algorithm proceeds by considering an augmented density function $f(\mathbf{x}_{1:n}, \mathbf{c}_{1:n} | \boldsymbol{\theta})$, where $\mathbf{c}_{1:n}$ can be thought of as missing data, and the augmented density $f(\mathbf{x}_{1:n}, \mathbf{c}_{1:n})$ satisfies $\mathbb{E}_{\mathbf{c}_{1:n}} \{f(\mathbf{x}_{1:n}, \mathbf{c}_{1:n} | \boldsymbol{\theta})\} = f(\mathbf{x}_{1:n} | \boldsymbol{\theta})$. Based on the augmented density, we form the complete log-likelihood function of the observed and ‘missing’ data, $\log L_n^c(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(x_i, c_i | \boldsymbol{\theta})$. Given an initial guess on the maximum likelihood estimator $\boldsymbol{\theta}^{(0)}$, the EM algorithm proceeds by alternating between an expectation (E) and maximization (M) step as follows:

$$\text{Compute } Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\mathbf{c}_{1:n}} \{ \log L_n^c(\boldsymbol{\theta}) | \mathbf{x}_{1:n}, \boldsymbol{\theta}^{(t)} \}, \quad (\text{E-step})$$

$$\text{Maximize } \boldsymbol{\theta}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}). \quad (\text{M-step})$$

Despite not working with the likelihood function directly, the sequences of iterations $\{\boldsymbol{\theta}^{(t)}\}$ from the EM algorithm can be shown to satisfy

$$\log L_n(\boldsymbol{\theta}^{(t+1)}) \geq \log L_n(\boldsymbol{\theta}^{(t)}),$$

see Robert and Casella (2004, Theorem 5.15) for a proof. Thus, the EM iterations are guaranteed to never yield a decrease in the log-likelihood at every step.

It should be noted that although the EM algorithm requires the solution of an optimization problem at each iteration, this maximization is often much simpler to solve than the original problem, as shown in Example 2.2. We will return to an example where the EM algorithm provides an efficient way of computing the MLE in the linear spline model covered in Chapter 6, as it shares the structure of the class of models considered in the following example.

Example 2.2 (EM algorithm for mixture models). Suppose that we observe a random sample $x_1, \dots, x_n \sim f$ where f is a density of the form

$$f(x | \boldsymbol{\theta}) = \sum_{j=1}^k \theta_j \omega_j(x), \quad (2.12)$$

where $\{\omega_j\}_{j=1}^k$ are known probability densities and $\boldsymbol{\theta} \in \mathcal{S}_{k-1}$ is unknown. We note that the histogram model of (2.6) is a special case of the formulation in (2.12), as one obtains the histogram model by using the basis densities $\omega_j(x) = |I_j|^{-1} \mathcal{I}(x \in I_j)$. Thus, (2.12) can be seen as a generalization of the histogram.

The log-likelihood function for the sample is then

$$\log L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log \left(\sum_{j=1}^k \theta_j \omega_j(x_i) \right). \quad (2.13)$$

Maximizing the log-likelihood analytically is in general infeasible due to the presence of the sum within the logarithm.

Densities of the form (2.12) are known as *mixture models*, and it is instructive to think of them in terms of how we can generate samples from such a model. We start by randomly drawing a component c with probability $\mathbb{P}(c = j) = \theta_j$, and then subsequently take x to be a sample from $\omega_c(\cdot)$. The joint density of x, c is in this case given by $f(x, c | \boldsymbol{\theta}) = \theta_c \omega_c(x)$, since the probability of drawing component c is θ_c . By marginalizing over the distribution of c , we see that x has density

$$\sum_{c=1}^k f(x, c | \boldsymbol{\theta}) = \sum_{c=1}^k \theta_c \omega_c(x) = f(x), \quad (2.14)$$

so $x \sim f$ as claimed. The main feature that makes the EM algorithm particularly attractive in this case is that by introducing the class labels $\mathbf{c}_{1:n}$ as missing data, the resulting optimization problem in the M-step can be solved analytically.

To derive the EM updates for this problem, we first need to form the complete log-likelihood L_n^c . When c_i is known to equal j , the sample x_i only contributes to the log-likelihood through the component $\omega_j(x_i)$. Hence, we can write the complete log-likelihood on the form

$$\log L_n^c(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^k \mathcal{I}(c_i = j) \left\{ \log(\theta_j) + \log \omega_j(x_i) \right\}$$

The Q function in the subsequent E-step is rather straightforward to compute, since conditional on $\mathbf{x}_{1:n}, \boldsymbol{\theta}^{(t)}$ the only random quantity is $\mathbf{c}_{1:n}$. Hence,

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^n \sum_{j=1}^k \mathbb{E}_{\mathbf{c}_{1:n}} \{ \mathcal{I}(c_i = j) | x_i, \boldsymbol{\theta}^{(t)} \} \{ \log(\theta_j) + \log \omega_j(x_i) \} \\ &= \sum_{i=1}^n \sum_{j=1}^k \mathbb{P}(c_i = j | x_i, \boldsymbol{\theta}^{(t)}) \{ \log(\theta_j) + \log \omega_j(x_i) \} \end{aligned} \quad (2.15)$$

The conditional probabilities here can be computed using Bayes' rule

$$\mathbb{P}(c_i = j | x_i, \boldsymbol{\theta}^{(t)}) = \frac{\mathbb{P}(c_i = j | \boldsymbol{\theta}^{(t)}) f(x_i | c_i = j, \boldsymbol{\theta}^{(t)})}{\sum_{j'=1}^k \mathbb{P}(c_i = j' | \boldsymbol{\theta}^{(t)}) f(x_i | c_i = j', \boldsymbol{\theta}^{(t)})} = \frac{\theta_j^{(t)} \omega_j(x_i)}{\sum_{j'=1}^k \theta_{j'}^{(t)} \omega_{j'}(x_i)}. \quad (2.16)$$

The maximum of (2.15) under the constraint that $\boldsymbol{\theta} \in \mathcal{S}_{k-1}$ can be found analytically, yielding a closed-form solution for update in the M-step:

$$\theta_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \mathbb{P}(c_i = j | x_i, \boldsymbol{\theta}^{(t)}), \quad \text{for } j = 1, 2, \dots, k. \quad (2.17)$$

We can then numerically approximate the maximum of (2.13) by iterating (2.17) until some convergence criterion has been met. ▲

2.2.3 Model selection

Most statisticians recognize that any given probability model $p(\mathbf{x}_{1:n} | \boldsymbol{\theta})$ for the data $\mathbf{x}_{1:n}$ is at best a good approximation of the truth. In any given situation, a statistician may posit many different probabilistic models for the data, with no good reason to strongly favor any particular model on an a priori basis. This introduces a need for data-based procedures that can choose the best model among the field of candidates, a process we refer to as *model selection*. A key problem pertaining to the task of model selection is that of model complexity. More complex models provide greater flexibility, but may overfit to the data, identifying patterns where there are none. On the other hand, the use of too simple a model may cause the analyst to miss some key underlying structure present in the dataset. Statistical model selection provides a framework for discriminating between different models based on the given sample in an attempt to find a model that strikes the right balance between good fit and an appropriate level of complexity.

In a model selection problem, we have a set of candidate models $\mathcal{M}$, each with their own parameters $\boldsymbol{\theta}_m$ belonging to different parameter spaces Θ_m . The goal is to find the model m which gives the best account of the observed sample in some sense. For likelihood-based models, *information criteria* are a popular tool used to perform model selection. The information criteria considered here are in general of the form of a penalized maximal log-likelihood:

$$\text{IC}(m) = \max_{\boldsymbol{\theta}_m \in \Theta_m} \{ \log L_n(\boldsymbol{\theta}_m | m) - \text{pen}(d_m) \}, \quad (2.18)$$

where $d_m = \dim(\boldsymbol{\theta}_m)$, $\log L_n(\boldsymbol{\theta}_m | m)$ is the log-likelihood under model m and $\text{pen}(\cdot)$ is a penalty function, penalizing models with larger values of d_m more. The best model according to the information criterion is the one that maximizes¹ (2.18). The information

¹Traditionally, information criteria are defined using the opposite sign to (2.18) and are minimized.

criterion defined above attempts to find a good trade-off between good fit, rewarding higher values of the log-likelihood, and model complexity, assigning a greater penalty to more complex models.

The Akaike's Information Criterion (AIC) (Akaike, 1998), is the original information criterion, and corresponds to using a penalty function of $\text{pen}(d_m) = d_m$. Another popular information criterion is the Bayesian Information Criterion (BIC) (Schwarz, 1978), which corresponds to the penalty $\text{pen}(d_m) = 0.5 \log(n)d_m$, penalizing complexity to a greater extent than AIC for moderate to large samples.

2.2.4 Parametric and nonparametric models

Statistical procedures are typically categorized as either *parametric* or *nonparametric*. The term parametric is reserved for those models that can be described in terms of a parameter of fixed-, or low dimension. Those procedures which are infinite-dimensional or of high dimension, possibly growing with the number of samples, are termed nonparametric. The parametric statistical paradigm typically relies on more rigid assumptions on the data-generating mechanism; e.g. assuming that the data-generating density belongs to a given parametric family of distributions such as the normal. Nonparametric methods typically make less restrictive assumptions about the data, making them a robust alternative to parametric ones.

One particular class of nonparametric estimators we shall make frequent use of in this thesis are *sieve estimators*. Introduced by Grenander (1981), sieve estimators are finite-dimensional models that use gradually more and more parameters as the sample size increases to estimate a given quantity. Typically, the true parameter will not belong to any of the finite-dimensional models under consideration. However, by constructing a sieve with good approximation properties, the sieve estimator can come close to this parameter.

Example 2.3 (Histograms with an unknown number of bins). Consider the setting of estimating a density based on an iid. sample $x_1, \dots, x_n \sim f_0$ where f_0 is a density supported on $[0, 1]$. As in Example 2.1 we model the data as coming from a piecewise constant density based on a regular partition. In this section we focus on histograms based on regular partitions, where $[0, 1]$ has been partitioned into the intervals

$$\begin{aligned} I_{k,1} &= [0, 1/k], \\ I_{k,j} &= ((j-1)/k, j/k], \quad 2 \leq j \leq k. \end{aligned} \tag{2.19}$$

When the value of k is clear from the context, we will omit the subscript k from the notation, so that the above intervals are denoted $I_1, \dots, I_k$, where k is initially treated as a fixed parameter value. Since $|I_j| = 1/k$ for all j we can simplify the likelihood from (2.8) slightly, leading to the expression

$$L_n(\mathbf{p} | k) = k^n \prod_{j=1}^k p_j^{N_j}, \tag{2.20}$$

with maximum likelihood estimates $\hat{p}_j = N_j/n$ as before.

This model is somewhat unsatisfactory unless we have a strong reason a priori to believe that the data is well approximated by a piecewise constant density over the given partition. To illustrate why this can be problematic, we assume that the true density f_0 is that of the Beta(2, 2) distribution, and let $k = 5$. Even when the best possible

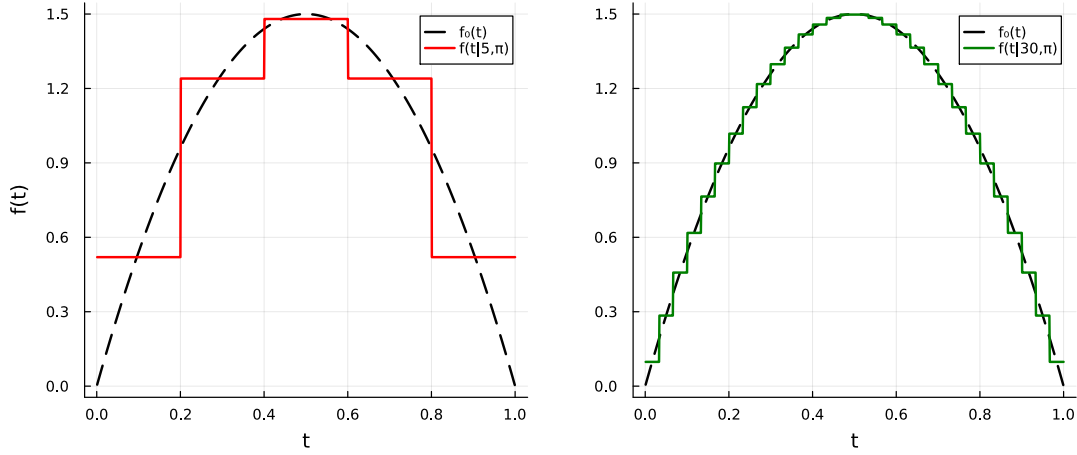


Figure 2.1: Approximation of a Beta(2,2)-density by piecewise constant densities with $k = 5$ and $k = 30$ bins

estimates of the bin probabilities $\pi_j = \int_{I_j} f_0(x) dx$ are known, the resulting piecewise constant density $f(\cdot | k, \pi)$ is a rather crude approximation of $f_0(\cdot)$, as can be seen in Figure 2.1 on the left.

As a way to overcome the rigid assumptions of the histogram with fixed number of bins is to rather treat the number k of bins as a parameter itself to be chosen based on the observed sample. The increased flexibility of this nonparametric approach is rather immediate: by choosing a suitably large value of k , we observe that the precision of the approximation $f(\cdot | k, \pi)$ improves substantially.

However, the use of a more flexible procedure does not come without its drawbacks, as it introduces the difficult problem of how to choose the number k of bins. One option is to use a sieve estimator. For histograms, we can construct such an estimator by searching for the optimal number $\hat{k}$ of bins among a set of candidates $\{1, 2, \dots, k_n\}$, where $\{k_n\}$ is a nondecreasing sequence of natural numbers. The resulting histogram is then drawn using $\hat{k}$ bins. However, it remains to find a suitable criterion for identifying the best $\hat{k}$. As mentioned in Section 2.2.3, likelihood alone will not be a good choice for choosing the best value of k as it will tend to favor overly complex models. In fact, it can be proven mathematically that the likelihood function (2.20) grows to positive infinity as we let $k \rightarrow \infty$, see Proposition A.1. To have a data-dependent selection rule which appropriately rewards goodness-of-fit while penalizing complexity, one could then use an information criterion (2.18) to find the proper number of bins. Using AIC as an example, a possible rule for selecting the number $\hat{k}$ of bins for a given sample is to find the maximizer of $\log L_n(\hat{\mathbf{p}} | k) - k$. ▲

2.3 Loss functions and statistical risk

To determine the performance of any given statistical procedure, we need a way of quantifying how large the deviation is between our estimate and the quantity which we are trying to estimate. If we are trying to estimate f_0 with an estimator $\hat{f}_n$ based on a sample $\mathbf{x}_{1:n} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, a good estimator should produce results as close to f_0 as possible. To achieve this, we specify a nonnegative *loss function* $\ell(f_0, \hat{f}_n)$, chosen such that lower values of ℓ corresponds to greater similarity between our estimate $\hat{f}_n$ and the

true density f_0 . For a given loss function, we define the *frequentist risk* by

$$R_n(f_0, \hat{f}_n) = \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ \ell(f_0, \hat{f}_n) \right\}. \quad (2.21)$$

By integrating the loss ℓ over the sampling distribution we get a measure of performance of using the estimator $\hat{f}_n$ in repeated experiments, where samples $\mathbf{x}_{1:n}$ are drawn according to the same data-generating mechanism. Ideally, we want the risk to be as small as possible.

In the parametric setting we assume that the data-generating density belongs to a family of distributions indexed by a parameter vector $\boldsymbol{\theta}$, which we would like to estimate. In such problems it is quite common to specify a loss function directly on the parameter of interest itself, so that the loss function becomes $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ with corresponding risk $R_n(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$.

We finally note that the risk $R_n(f_0, \hat{f}_n)$ will in general depend on the unknown f_0 , and as such, good performance at any particular f_0 will not guarantee good performance for other densities. To ensure that a given procedure is robust, the estimator should rather be designed to produce small risks for a general class of densities f_0 . In parametric problems, this means that we aim for small risk in most of the parameter space Θ , while nonparametric procedures should try to produce small risks for general classes of densities $\mathcal{F}$, such as those with bounded derivative. In either case, the loss function is typically taken to be equal to a metric on a suitable set. Since most of the loss functions we will be working with are metrics, a proper introduction to metric spaces is in order.

2.3.1 Metric and normed spaces

To measure the distance between two real numbers a and b , one merely has to compute the absolute difference $|a - b|$ between the two numbers. However, it is less obvious how exactly one should quantify the distance between two functions f and g . *Metrics* generalize the notion of distance from $\mathbb{R}$ to more abstract spaces, such as spaces of Euclidian vectors and functions. In the following, we let $\mathcal{X}$ be a set.

Definition 2.2 (Metric space). We say that the function $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a *metric* on $\mathcal{X}$ if for any $x, y, z \in \mathcal{X}$, the following properties hold:

- (i) $d(x, y) \geq 0$ and $d(x, y) = 0 \iff x = y$ (positivity)
- (ii) $d(x, y) = d(y, x)$ (symmetry)
- (iii) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

We refer to the tuple $(\mathcal{X}, d)$ as a *metric space*.

Example 2.4 (Metrics on Euclidian spaces). An example of a metric space is the Euclidian space $(\mathbb{R}^k, d_2)$, where the metric d_2 is the usual Euclidian distance,

$$d_2(\mathbf{x}, \mathbf{y}) = \left\{ \sum_{j=1}^k (x_j - y_j)^2 \right\}^{1/2}. \quad (2.22)$$

A generalization of the above metric is that of $\mathbb{L}_r$ metrics,

$$d_r(\mathbf{x}, \mathbf{y}) = \left\{ \sum_{j=1}^k |x_j - y_j|^r \right\}^{1/r}, \quad r \geq 1, \quad (2.23)$$

which can also be shown to satisfy the three properties of Definition 2.2. Another important metric in Euclidian spaces is the *supremum metric*,

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_{1 \leq j \leq k} |x_j - y_j|, \quad (2.24)$$

which is the pointwise limit of the $\mathbb{L}_r$ metrics defined above as $r \rightarrow \infty$. Finally, a metric that features prominently in this work is the *Hellinger metric*, defined on $\mathbb{R}_+^k$ via

$$d_H(\mathbf{x}, \mathbf{y}) = \left\{ \sum_{j=1}^k (\sqrt{x_j} - \sqrt{y_j})^2 \right\}^{1/2}. \quad (2.25)$$

Letting $\sqrt{\mathbf{x}}$ denote the vector with element j equal to $\sqrt{x_j}$, we can denote the Hellinger metric more compactly via $d_H(\mathbf{x}, \mathbf{y}) = d_2(\sqrt{\mathbf{x}}, \sqrt{\mathbf{y}})$, i.e. as the $\mathbb{L}_2$ distance between the root-vectors. $\blacktriangle$

The metrics of Example 2.4 on $\mathbb{R}^k$ serve as natural loss functions in parametric estimation problems where the goal is to estimate a k -dimensional parameter $\boldsymbol{\theta}$. If our estimator is $\hat{\boldsymbol{\theta}}$, then the loss is simply $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = d(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, where d is one of the metrics defined above.

We now give a few examples of norms on spaces of functions and the metrics they induce. These metrics are natural to use as loss functions in the context of density estimation, and we shall make use of them both in more theoretical work later, as well as in simulations to measure the performance of different histogram procedures.

Example 2.5 (Metrics on function spaces). All of the metrics considered in (2.4) have analogues defined on spaces of functions. Letting $\mathcal{X} \subseteq \mathbb{R}$ denote a (measurable) set, we define the space $\mathbb{L}_r(\mathcal{X})$ for $r \in [1, \infty]$ to be the set of measurable functions $f: \mathcal{X} \rightarrow \mathbb{R}$, satisfying

$$\begin{cases} \int_{\mathcal{X}} |f(x)|^r dx < \infty & \text{if } r \in [1, \infty), \\ \sup_{x \in \mathcal{X}} |f(x)| < \infty & \text{if } r = \infty. \end{cases} \quad (2.26)$$

For two functions $f, g \in \mathbb{L}_r(\mathcal{X})$, we define the so-called $\mathbb{L}_r$ metric by

$$d_r(f, g) = \begin{cases} (\int_{\mathcal{X}} |f(x) - g(x)|^r dx)^{1/r} & \text{if } r \in [1, \infty), \\ \sup_{x \in \mathcal{X}} |f(x) - g(x)| & \text{if } r = \infty. \end{cases} \quad (2.27)$$

The requirement that $f, g \in \mathbb{L}_r(\mathcal{X})$ ensures that the metrics in (2.27) are finite. The *Hellinger metric* is defined on the subset of nonnegative-valued functions in $\mathbb{L}_1(\mathcal{X})$ by

$$d_H(f, g) = \left\{ \int_{\mathcal{X}} \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx \right\}^{1/2}, \quad (2.28)$$

or equivalently $d_H(f, g) = d_2(\sqrt{f}, \sqrt{g})$. $\blacktriangle$

Importantly, metrics are used to define the notions of open balls, generalizing the notion of a ball to non-Euclidian spaces.

Definition 2.3 (Open ball). Let $(\mathcal{X}, d)$ be a metric space and let $x \in \mathcal{X}$. For $\epsilon > 0$, we define the *open ball* at x of radius ϵ by

$$B_\epsilon(x) = \{y \in \mathcal{X} : d(x, y) < \epsilon\}. \quad (2.29)$$

A related notion that can be used to define distances is that of a norm. Since almost all of the metrics we will be considering can be defined in terms of a norm, we give a short introduction to norms and normed spaces. Norms generalize the concept of size from $\mathbb{R}$ to general vector spaces, much in the same way that metrics generalize distances.

Definition 2.4 (Normed space). Let $\mathcal{V}$ be a vector space over the real numbers and $\|\cdot\|: \mathcal{V} \rightarrow \mathbb{R}$ a function. We say that $\|\cdot\|$ is a *norm* on $\mathcal{V}$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ and $\lambda \in \mathbb{R}$, we have

- (i) $\|\mathbf{x}\| \geq 0$ and $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$ (positivity)
- (ii) $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$ (homogeneity)
- (iii) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)

We refer to the tuple $(\mathcal{V}, \|\cdot\|)$ as a *normed space*.

Importantly, any norm $\|\cdot\|$ on a vector space $\mathcal{V}$ can be used to define a metric on any subset $\mathcal{A} \subseteq \mathcal{V}$ by

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|, \quad \mathbf{x}, \mathbf{y} \in \mathcal{A}. \quad (2.30)$$

We say that the norm $\|\cdot\|$ *induces* the metric d on $\mathcal{A}$.

Example 2.6 (Metrics induced by norms). Almost all of the metrics encountered so far are induced by norms. In particular, the Euclidian $\mathbb{L}_r$ metric are induced by the norm $\|\cdot\|_r$, given by

$$\|\mathbf{x}\|_r = \begin{cases} \left(\sum_{j=1}^k |x_j|^r\right)^{1/r} & \text{if } r \in [1, \infty) \\ \sup_{1 \leq j \leq k} |x_j| & \text{if } r = \infty \end{cases} \quad (2.31)$$

Similarly, the metric d_r defined on $\mathbb{L}_r(\mathcal{X})$ is induced by

$$\|f\|_r = \begin{cases} \left(\int_{\mathcal{X}} |f(x)|^r dx\right)^{1/r} & \text{if } r \in [1, \infty) \\ \sup_{x \in \mathcal{X}} |f(x)| & \text{if } r = \infty \end{cases} \quad (2.32)$$

On the other hand, the Hellinger metrics are not induced by a norm, but they are intrinsically related to the $\mathbb{L}_2$ norm via $d_H(f, g) = \|\sqrt{f} - \sqrt{g}\|_2$. ▲

Example 2.7 (Loss functions for histogram estimation). We consider the problem of estimating a bounded density f_0 on $[0, 1]$, using a regular histogram consisting of k bins. Since k is fixed, we can parameterize the model $f(\cdot) = f(\cdot | k, \mathbf{p})$ by the simplex-vector $\mathbf{p}$. It thus becomes possible to use a loss function on the parameter $\mathbf{p}$ directly, where the goal is to estimate the true bin probabilities $\pi_j = \int_{I_j} f_0(x) dx$ accurately. For instance, using the loss function induced by the squared $\mathbb{L}_2$ norms on $\mathbb{R}^k$, we obtain

$$\ell(\boldsymbol{\pi}, \hat{\mathbf{p}}) = \|\boldsymbol{\pi} - \hat{\mathbf{p}}\|_2^2 = \sum_{j=1}^k |\pi_j - \hat{p}_j|^2, \quad (2.33)$$

Since we are trying to estimate a density, it is arguably more natural to use a loss function placed on the space of densities instead. Using the squared $\mathbb{L}_2$ norm on functions, we get the loss function

$$\tilde{\ell}(f_0, \hat{f}_n) = \|f_0 - \hat{f}_n\|_2^2 = \int_0^1 \{f_0(x) - \hat{f}_n(x)\}^2 dx. \quad (2.34)$$

It turns out that these two loss functions are equivalent in this case. A computation shown in the Appendix (see Proposition A.2) reveals that

$$\tilde{\ell}(f_0, \hat{f}_n) = \|f_0 - f_{0,k}\|_2^2 + \|f_{0,k} - \hat{f}_n\|_2^2 \quad (2.35)$$

where $f_{0,k} = f_{0,I_{1:k}}$ is the piecewise constant version of f_0 based on using k bins, see Definition 2.1. The first term in (2.35) is a constant for given values of k as it does not depend on $\hat{f}_n$. Hence, we can drop the first term from consideration for now and focus on the second term. For two piecewise constant densities we have that

$$\begin{aligned} \|f(\cdot | k, \boldsymbol{\pi}) - f(\cdot | k, \hat{\mathbf{p}})\|_2^2 &= \int_0^1 \left\{ f(x | k, \boldsymbol{\pi}) - f(x | k, \hat{\mathbf{p}}) \right\}^2 dx \\ &= \sum_{j=1}^k \int_{I_j} \left\{ f(x | k, \boldsymbol{\pi}) - f(x | k, \hat{\mathbf{p}}) \right\}^2 dx \\ &= \sum_{j=1}^k \int_{I_j} \{k\pi_j - k\hat{p}_j\}^2 dx \\ &= k \sum_{j=1}^k \{\pi_j - \hat{p}_j\}^2 \\ &= k \|\boldsymbol{\pi} - \hat{\mathbf{p}}\|_2^2 \\ &= k \ell(\boldsymbol{\pi}, \hat{\mathbf{p}}). \end{aligned} \quad (2.36)$$

From (2.36) and (2.35), we obtain the relation

$$\|f_0 - \hat{f}_n\|_2^2 = \|f_0 - f_{0,k}\|_2^2 + k \|\boldsymbol{\pi} - \hat{\mathbf{p}}\|_2^2. \quad (2.37)$$

Since k is constant, we conclude that there is a one-to-one correspondence between the two loss functions (2.33) and (2.34). $\blacktriangle$

2.3.1a Constructing estimators based on risk minimization

Performance with respect to the risk R_n can also be used to construct estimators. In particular, consider a class of estimators $\hat{f}_n = \hat{f}_n(\cdot; \lambda)$ depending on some tuning parameter λ . The tuning parameter can for instance be the number k of bins chosen when constructing a histogram. Ideally, the tuning parameter should be chosen such that (2.21) is minimized among all estimators belonging to the given class of densities. However, as the risk will generally depend on the true density of the data f_0 , one cannot use this criterion directly to design the risk-optimal estimator. To circumvent this issue, one can estimate the risk of $\hat{f}_n(\cdot; \lambda)$ based on the data instead. The resulting estimator is then taken to be the one based on the value of λ which minimizes the estimated risk. Since more complex models will generally fit the data more closely, naively estimating the risk on the sample used to construct the estimator will lead to overly optimistic estimates, where the degree of optimism generally increases with model complexity. As such, out-of-sample risk estimates are often used to select the tuning parameter λ .

Example 2.8 (Bin selection via estimated risk minimization). We consider the problem of estimating a density f_0 on $[0, 1]$ with a regular histogram. As noted in Example 2.3, the number k of bins is a key tuning parameter that may greatly influence the resulting estimator. To select k we use the approach detailed in this section; we estimate the risk based on squared $\mathbb{L}_2$ loss for different values of k and choose the one that minimizes our

estimates to construct the final histogram. Let $\hat{f}_n(\cdot | k) = f(\cdot | k, \hat{\mathbf{p}})$ denote the histogram estimate of f_0 using k bins. For squared $\mathbb{L}_2$ loss, we have

$$\ell(f_0, \hat{f}_n(\cdot | k)) = \int_0^1 \hat{f}_n^2(x | k) dx - 2 \int_0^1 \hat{f}_n(x | k) f_0(x) dx + \int_0^1 f_0^2(x) dx,$$

leading to the following expression for the risk:

$$R_n(f_0, \hat{f}_n(\cdot | k)) = \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ \int_0^1 \hat{f}_n^2(x | k) dx \right\} - 2 \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ \int_0^1 \hat{f}_n(x | k) f_0(x) dx \right\} + \int_0^1 f_0^2(x) dx. \quad (2.38)$$

The last term is constant with respect to $\hat{f}_n$ as it only depends on f_0 , we may ignore it when minimizing the risk estimate, focusing only on the first two terms. Dealing with the second term is not straightforward, as it depends on the unknown density f_0 which we are trying to estimate. However, if we let x_{n+1} be a new sample from f_0 independent of $\mathbf{x}_{1:n}$, then we can write this term in the form of an expectation,

$$\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ \int_0^1 \hat{f}_n(x | k) f_0(x) dx \right\} = \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ \mathbb{E}_{x_{n+1}} \left[\hat{f}_n(x_{n+1} | k) \right] \right\} = \mathbb{E}_{\mathbf{x}_{1:n}, x_{n+1}} \left\{ \hat{f}_n(x_{n+1} | k) \right\}. \quad (2.39)$$

The approach taken by Rudemo (1982) is to estimate (2.39) through the use of *leave-one-out cross-validation* (LOOCV). The LOOCV approach proceeds by constructing the estimator $\hat{f}_{n-1}^{-i}(\cdot; k)$ based on the sample $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$, that is, we exclude x_i from the sample used to construct the estimator. The LOOCV estimate of (2.39) is then $n^{-1} \sum_{i=1}^n \hat{f}_{n-1}^{-i}(x_i; k)$. The first term in (2.38) can be estimated² simply by $\int_0^1 \hat{f}_n^2(x | k) dx$. This approach leads to a criterion which is equivalent to maximizing the following (Rudemo, 1982, p. 68)

$$\text{CV}_n(k) = \frac{k(n+1)}{n^2} \sum_{j=1}^k N_j^2 - 2k. \quad (2.40)$$

The criterion (2.40) can be quickly computed for each value of k belonging to a suitable finite set $\mathcal{K} \subseteq \mathbb{N}$. The chosen number $\hat{k}$ of bins resulting from this procedure is then $\hat{k} = \arg \max_{k \in \mathcal{K}} \text{CV}_n(k)$. ▲

2.4 Asymptotics

Asymptotics informally refers to the study of what happens when the sample size grows to infinity. In particular, we are often interested in the *large-sample* properties of estimators. If the sample $\mathbf{x}_{1:n} = (x_1, \dots, x_n)$ has been generated by the density f_0 , then we would like for an estimator of the density $\hat{f}_n$ based on the observed sample to be able to recover f_0 as the sample size approaches infinity. The notion of convergence of random variables provides a framework under which one can prove that $\hat{f}_n$ possesses such a property.

2.4.1 Convergence of random variables

We now give a brief introduction to convergence of random variables and stochastic order notation, concepts that we will later use to study the large-sample behavior

²The first term in (2.38) can in principle also be estimated using LOOCV. The two procedures can be shown to yield very similar optimality criteria, see Rudemo (1982).

of histogram estimators. Specifically, we cover stochastic convergence for real-valued random variables, as these are the building blocks we use to extend these concepts to more general spaces later.

Definition 2.5 (Convergence of random variables). Let X be real-valued random variable and $\{X_n\}$ a sequence of random variables taking values in $\mathbb{R}$. We say that X_n converges to X in probability, denoted $X_n \xrightarrow{\mathbb{P}} X$ if

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0, \quad \forall \epsilon > 0. \quad (2.41)$$

We say that X_n converges to X almost surely, denoted $X_n \xrightarrow{\text{a.s.}} X$ if

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1. \quad (2.42)$$

Remark 2.1. Although we shall mostly work with convergence in probability, the notion of almost sure convergence remains useful, as it can be used to establish the former form of convergence. In particular, if $X_n \xrightarrow{\text{a.s.}} X$, then $X_n \xrightarrow{\mathbb{P}} X$ as well (van der Vaart, 2000, Theorem 2.7). The converse statement does not hold in general, as there are sequences of random variables X_n that converge to X in probability, but where a.s. convergence does not take place.

Another useful tool for establishing convergence in probability is that of Markov's inequality, as it relates probabilities to expectations. Often, the latter is easier to compute or bound, allowing us to control the former.

Proposition 2.1 (Markov's Inequality). *Let X be a random variable taking values in $\mathcal{X} \subseteq \mathbb{R}$ and $\varphi: \mathcal{X} \rightarrow \mathbb{R}$ a nonnegative, nondecreasing function. Then for any $\epsilon > 0$,*

$$\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}_X\{\varphi(X)\}}{\varphi(\epsilon)}. \quad (2.43)$$

Proof. Since $\varphi(\cdot)$ is nondecreasing and nonnegative, we have that

$$\mathbb{I}(X \geq \epsilon) \leq \mathbb{I}(X \geq \epsilon)\varphi(X)/\varphi(\epsilon) \leq \varphi(X)/\varphi(\epsilon).$$

Using the monotonicity property of expectations, one obtains

$$\mathbb{P}(X \geq \epsilon) = \mathbb{E}_X\{\mathbb{I}(X \geq \epsilon)\} \leq \mathbb{E}_X\left\{\frac{\varphi(X)}{\varphi(\epsilon)}\right\} = \frac{\mathbb{E}_X\{\varphi(X)\}}{\varphi(\epsilon)},$$

which concludes the proof. $\square$

2.4.2 Stochastic order symbols

A concept strongly related to that of convergence in probability is stochastic order symbols, serving as a useful tool for studying the asymptotic behavior of random variables. The stochastic order symbols $\mathcal{O}_{\mathbb{P}}(\cdot)$ and $o_{\mathbb{P}}(\cdot)$ form a counterpart to the more familiar deterministic order symbols $\mathcal{O}(\cdot)$ and $o(\cdot)$. Recall that if $\{a_n\}$ and $\{b_n\}$ are sequences of positive real numbers then $b_n = \mathcal{O}(a_n)$ means that there exists some $M > 0$ such that $a_n/b_n \leq M$ for all sufficiently large n . We write $b_n = o(a_n)$ if $\lim_{n \rightarrow \infty} b_n/a_n = 0$. The definitions for the corresponding stochastic order symbols are rather similar.

Definition 2.6 (Stochastic order symbols). Let $\{X_n\}$ be a sequence of random variables and $\{a_n\}$ a sequence of positive real numbers. We write $X_n = \mathcal{O}_{\mathbb{P}}(a_n)$ if for any $\epsilon > 0$ there is an $M > 0$ and an integer $m \in \mathbb{N}$ such that

$$\mathbb{P}(|X_n| \geq Ma_n) < \epsilon, \quad \forall n \geq m. \quad (2.44)$$

We write $X_n = o_{\mathbb{P}}(a_n)$ if for any $\epsilon > 0$,

$$\mathbb{P}(|X_n| \geq \epsilon a_n) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (2.45)$$

Equivalently, (2.45) can be written as $X_n/a_n \xrightarrow{\mathbb{P}} 0$.

From Definition 2.6, it follows rather immediately that if a_n and b_n are deterministic sequences, then $a_n = \mathcal{O}(b_n)$ implies $a_n = \mathcal{O}_{\mathbb{P}}(b_n)$ and likewise $a_n = o(b_n)$ is sufficient to ensure that $a_n = o_{\mathbb{P}}(b_n)$. Proposition 2.2 reviews some important properties of stochastic order symbols.

Proposition 2.2 (Properties of stochastic order symbols).

- (i) Let $\{a_n\}$, $\{b_n\}$ and $\{c_n\}$ be sequences of positive real numbers and let $\{X_n\}$ be a sequence of random variables. If $a_n/b_n \rightarrow d$ for some $d \geq 0$ as $n \rightarrow \infty$, then the following holds:

$$\mathcal{O}_{\mathbb{P}}(a_n) + \mathcal{O}_{\mathbb{P}}(b_n) = \mathcal{O}_{\mathbb{P}}(b_n), \quad (2.46)$$

$$o_{\mathbb{P}}(a_n) + o_{\mathbb{P}}(b_n) = o_{\mathbb{P}}(b_n). \quad (2.47)$$

$$c_n \mathcal{O}_{\mathbb{P}}(a_n) = \mathcal{O}_{\mathbb{P}}(c_n a_n) \quad (2.48)$$

$$c_n o_{\mathbb{P}}(a_n) = o_{\mathbb{P}}(c_n a_n) \quad (2.49)$$

$$X_n = o_{\mathbb{P}}(a_n) \implies X_n = \mathcal{O}_{\mathbb{P}}(a_n), \quad (2.50)$$

$$X_n = \mathcal{O}_{\mathbb{P}}(a_n) \text{ and } a_n = o(1) \implies X_n = o_{\mathbb{P}}(1). \quad (2.51)$$

- (ii) Suppose that $\{X_n\}$ is a sequence of random variables converging to X in probability. Then $X_n = \mathcal{O}_{\mathbb{P}}(1)$. Moreover, $X_n = X + o_{\mathbb{P}}(1)$.

- (iii) Fix $k \in \mathbb{N}$ and suppose that $\{X_{n,1}, \dots, X_{n,k}\}_{n=1}^{\infty}$ is a sequence of random variables satisfying $X_{n,j} = o_{\mathbb{P}}(a_n)$ for all j . Then $\max_{1 \leq j \leq k} |X_{n,j}| = o_{\mathbb{P}}(a_n)$.

The proof of Proposition 2.2 can be found in Appendix A.1.

2.4.3 Consistency and convergence rates

We now return to the key concepts of consistency and convergence rates. Informally, consistency ensures that our estimator $\hat{f}_n$ will be close to the true f_0 with high probability as the sample size approaches infinity.

Definition 2.7 (Consistency of point estimators). Given a loss function ℓ , we say that an estimator $\hat{f}_n$ is *consistent* at f_0 with respect to ℓ if for any $\epsilon > 0$,

$$\mathbb{P}(\ell(f_0, \hat{f}_n) \geq \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (2.52)$$

The formulation in (2.52) is equivalent to saying that the real-valued random variable $\ell(f_0, \hat{f}_n)$ satisfies $\ell(f_0, \hat{f}_n) \xrightarrow{\mathbb{P}} 0$.

We may also be interested in quantifying how quickly our estimator $\hat{f}_n$ approaches f_0 . This leads to the concept of convergence rates.

Definition 2.8 (Convergence rate of point estimators). Let $\{\epsilon_n\}$ be a sequence of positive real numbers such that $\epsilon_n = o(1)$. We say that $\hat{f}_n$ converges to f_0 at rate ϵ_n , with respect to the loss function ℓ if for any sequence $\{M_n\}$ such that $M_n \rightarrow \infty$, we have that

$$\mathbb{P}(\ell(f_0, \hat{f}_n) \geq M_n \epsilon_n) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (2.53)$$

Remark 2.2. A related, but somewhat stronger notion of convergence rates is as follows: we say that $\hat{f}_n$ converges at rate ϵ_n if

$$\ell(f_0, \hat{f}_n) = \mathcal{O}_{\mathbb{P}}(\epsilon_n). \quad (2.54)$$

This is seen to imply (2.53), so this other notion of convergence is somewhat stronger than that of Definition 2.8. Our reason for adapting the former definition is that it corresponds more closely to its Bayesian counterpart to be defined later.

Since $\epsilon_n = o(1)$, Definition 2.8 implies that if $\hat{f}_n$ converges at a given rate, then it is also consistent. To establish rates of convergence one generally needs to make more restrictive smoothness assumptions on the true density f_0 than what one needs to merely establish consistency. To show that a given estimator possesses a certain rate, it is often easiest to work with the risk R_n of the estimator. If we take ℓ to be the loss induced by the $\mathbb{L}_r$ -norms for $r \geq 1$, an application of Markov's inequality (2.43) yields

$$\mathbb{P}(\|f_0 - \hat{f}_n\|_r \geq M \epsilon_n) \leq \frac{\mathbb{E}_{\mathbf{x}_{1:n}} \|f_0 - \hat{f}_n\|_r}{M \epsilon_n} = \frac{R_n}{M \epsilon_n}. \quad (2.55)$$

If $R_n = \mathcal{O}(\epsilon_n)$, then $R_n/\epsilon_n = \mathcal{O}(1)$, and it follows that we can make the left hand side of (2.55) arbitrarily small for sufficiently large M , so the rate is ϵ_n in the stronger sense of (2.54). Of primary interest is finding the sequence ϵ_n decaying to 0 as quickly as possible while satisfying (2.53). Aside from being of independent interest, studies of the rate at which the risk R_n goes to zero can be used to construct estimators which are optimal in the asymptotic sense. By designing a procedure that minimizes the terms that dominate the risk asymptotically, the resulting estimator will perform well with respect to the risk in a large-sample setting.

Example 2.9 (Scott's rule). We return to the example of estimating a density on $[0, 1]$ using regular histograms as in Example 2.3. One of the more widely used rules for selecting the number of bins of a histogram is Scott's rule (Scott, 1979), which is derived by minimizing an asymptotic expansion of the risk based on squared $\mathbb{L}_2$ loss, $R_n(f_0, \hat{f}_n) = \mathbb{E}_{\mathbf{x}_{1:n}} \|f_0 - \hat{f}_n\|_2^2$, where $\hat{f}_n(\cdot) = f(\cdot; k, \hat{\mathbf{p}})$ is the estimated density using a histogram with k bins (2.10). Let $h = k^{-1}$ denote the interval width. The calculation of the asymptotic risk is somewhat lengthy and we don't pursue it here, but Scott (1979) obtains an expression for the risk of $\hat{f}_n$ under suitable regularity conditions on f_0 :

$$R_n(f_0, \hat{f}_n) = \frac{1}{nh} + \frac{1}{12} h^2 \int_0^1 \{f_0'(x)\}^2 dx + \mathcal{O}\left(\frac{1}{n} + h^3\right). \quad (2.56)$$

In the case where $h \rightarrow 0$ and $nh^{-1} \rightarrow \infty$ as $n \rightarrow \infty$, the first two terms in (2.56) will dominate the expression of the risk. Minimizing these two terms with respect to h , one finds that the asymptotically optimal interval width is

$$h^* = \left(\frac{6}{n \int_0^1 \{f_0'(x)\}^2 dx} \right)^{1/3}. \quad (2.57)$$

One cannot really use (2.57) in practical problems as it depends on the true density f_0 which we are trying to estimate to begin with. A way of dealing with this problem is to replace $\int_0^1 \{f'_0(t)\}^2 dt$ with a data-dependent estimate by assuming that f_0 belongs to some given parametric family of distributions, such as the Beta distribution. $\blacktriangle$

Before we end our discussion of asymptotics and risk for now, we introduce the concept of minimax rates. Recall from the previous section that the risk of a given estimator will in general depend on the unknown density f_0 . It turns out that if we want to estimate any such arbitrary density belonging to a general class of functions $\mathcal{F}$, this may put a lower limit on the convergence rate, no matter which estimator is being used, depending on the class under consideration.

Definition 2.9 (Minimax risk). Let $\mathcal{F}$ be a class of probability distributions and ℓ a loss function. The *minimax risk* over the class $\mathcal{F}$ with respect to ℓ is defined as

$$R_n^M = \inf_{\hat{f}_n} \sup_{f_0 \in \mathcal{F}} \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ \ell(f_0, \hat{f}_n) \right\}, \quad (2.58)$$

where $\mathbf{x}_{1:n}$ is a random sample from f_0 and the infimum is taken over all estimators $\hat{f}_n$. If there exists a constant $m > 0$ and a sequence of positive reals $\{\epsilon_n\}$ such that $R_n^M \geq m\epsilon_n$ for all large n , we say that ϵ_n is a *minimax rate* for $\mathcal{F}$ under loss function ℓ .

Minimax rates are a very important concept in asymptotic statistics, as they provide lower bounds on the convergence rate for any procedure. If the minimax rate for $\mathcal{F}$ is ϵ_n , and one can construct an estimator whose risk contracts at the same rate, then this estimator is asymptotically optimal in the sense that no other procedure can attain a faster convergence rate uniformly over the class $\mathcal{F}$. Typically, the minimax risk is related to the smoothness of the functions in the class under consideration.

Definition 2.10 (α -smooth functions). We say that a function $h: \mathcal{X} \rightarrow \mathbb{R}$ is α -smooth³ for $0 < \alpha \leq 1$ if there exists some $L > 0$ such that for all $s, t \in \mathcal{X}$,

$$|h(s) - h(t)| \leq L|s - t|^\alpha. \quad (2.59)$$

For $\alpha \in (1, 2]$, h is said to be α -smooth if it is differentiable, and there exists $L > 0$ such that the derivative satisfies

$$|h'(s) - h'(t)| \leq L|s - t|^{\alpha-1}, \quad (2.60)$$

for every $s, t \in \mathcal{X}$.

Remark 2.3. The property (2.59) is also commonly known under the name α -Hölder continuity. The parameter α controls the smoothness of the function h , with larger values corresponding to smoother functions.

Example 2.10. Let $\mathcal{F}_\alpha$ denote the class of probability density functions in $[0, 1]$ whose square roots are α -smooth. Then the minimax rate with respect to the squared Hellinger metric is $n^{-2\alpha/(2\alpha+1)}$ (Massart, 2007, Proposition 7.16). We note that α increases, i.e. as the class under consideration becomes smoother, the minimax rate improves, with the rate $n^{-4/5}$ for $\alpha = 2$. Later, in Proposition 4.6, we will show that this rate is actually achieved for certain histogram constructions when $\alpha \in (1/2, 1]$, showing that they achieve the optimal rate. $\blacktriangle$

³Not to be confused with smoothness in the sense of being infinitely many times continuously differentiable.

Chapter 3

Bayesian inference

3.1 Basics of Bayesian inference

Similarly to the likelihood-based methods of the previous section, the *Bayesian* approach starts with formulating a probability-based model for the data $\mathbf{x}_{1:n}$ indexed by a parameter $\boldsymbol{\theta}$, leading to a corresponding likelihood function $L_n(\boldsymbol{\theta}) = p(\mathbf{x}_{1:n} | \boldsymbol{\theta})$. However, the two approaches differ in the treatment of the unknown parameter $\boldsymbol{\theta}$. In the Bayesian paradigm, the parameter $\boldsymbol{\theta}$ is considered a random variable, and we form a *prior distribution* $p(\boldsymbol{\theta})$ to represent our *prior* beliefs about this parameter. The Bayesians belief about the parameter $\boldsymbol{\theta}$ upon observing the sample $\mathbf{x}_{1:n}$ is captured in the posterior distribution $p(\boldsymbol{\theta} | \mathbf{x}_{1:n})$. The posterior density is, in principle, readily computed using Bayes formula,

$$p(\boldsymbol{\theta} | \mathbf{x}_{1:n}) = \frac{p(\boldsymbol{\theta}) p(\mathbf{x}_{1:n} | \boldsymbol{\theta})}{p(\mathbf{x}_{1:n})} = \frac{p(\boldsymbol{\theta}) p(\mathbf{x}_{1:n} | \boldsymbol{\theta})}{\int_{\Theta} p(\boldsymbol{\theta}') p(\mathbf{x}_{1:n} | \boldsymbol{\theta}') d\boldsymbol{\theta}'} \quad (3.1)$$

The normalization constant $p(\mathbf{x}_{1:n})$ is known as the *marginal likelihood*. The marginal likelihood is in general difficult to compute in a Bayesian model, owing to the difficulty of computing the integral in the denominator of (3.1). However, since the marginal likelihood does not depend on $\boldsymbol{\theta}$, one often works with an unnormalized posterior instead; this is primarily due to the fact that many of the methods used for posterior inference in Bayesian models only require knowledge of the posterior density up to proportionality. We then obtain the following relation

$$p(\boldsymbol{\theta} | \mathbf{x}_{1:n}) \propto p(\boldsymbol{\theta}) p(\mathbf{x}_{1:n} | \boldsymbol{\theta}). \quad (3.2)$$

Since the posterior distribution is a properly normalized density, one can additionally ignore any factor in the prior and the likelihood not depending on $\boldsymbol{\theta}$, so long as one only wants to know the posterior distribution up to proportionality.

Our first example of Bayesian inference will deal with inference for multinomial parameters. Since a Bayesian approach to the problem requires the specification of a prior distribution, we will briefly introduce the Dirichlet distribution as a possible prior distribution for this problem.

The Dirichlet distribution

In the present thesis, we shall encounter the Dirichlet distribution quite frequently, due to its use as a probabilistic model for probabilities of disjoint outcomes. In a problem where we model the probability vector $\mathbf{p}$ as a random quantity, it is often

useful to work with a parametric family of distributions for $\mathbf{p}$, the Dirichlet being one such family particularly suited for probabilities. We say that a random vector $\mathbf{p}$ follows a k -dimensional Dirichlet($\mathbf{a}$)-distribution, where $\mathbf{a} \in \mathbb{R}_+^k$ and $a = \sum_{j=1}^k a_j$, if $\mathbf{p}$ has a density given by

$$p(\mathbf{p}) = \frac{\Gamma(a)}{\prod_{j=1}^k \Gamma(a_j)} \prod_{j=1}^k p_j^{a_j-1} \text{ for } \mathbf{p} \in S_{k-1}, \quad (3.3)$$

and 0 elsewhere, where S_{k-1} is the simplex (2.1). We will frequently use of the parametrization $\mathbf{a} = a\mathbf{p}_0$, where $\mathbf{p}_0 \in S_{k-1}$. The following Proposition gives some important properties of the Dirichlet distribution. For a proof, the reader may consult Ng et al. (2011).

Proposition 3.1. *Let $\mathbf{p}$ be a random vector in S_{k-1} following a $\text{Dir}(a\mathbf{p}_0)$ -distribution.*

(i) *The marginal distributions of each p_j is given by*

$$p_j \sim \text{Beta}(ap_{0,j}, a(1 - p_{0,j})). \quad (3.4)$$

(ii) *The mean of p_j is given by*

$$\mathbb{E}_{p_j}\{p_j\} = p_{0,j}. \quad (3.5)$$

(iii) *The variances and covariances of $\mathbf{p}$ are*

$$\begin{aligned} \text{Var}_{p_j}(p_j) &= \frac{p_{0,j}(1 - p_{0,j})}{a + 1}, \\ \text{Cov}_{p_j, p_l}(p_j, p_l) &= -\frac{p_{0,j}p_{0,l}}{a + 1}, \quad j \neq l \end{aligned} \quad (3.6)$$

(iv) *The collapsability property. Let $A_1, A_2, \dots, A_M$ be a partition of $\{1, 2, \dots, k\}$. Define the new vectors $\bar{\mathbf{p}}$ and $\bar{\mathbf{p}}_0$ by*

$$\bar{p}_m = \sum_{j \in A_m} p_j, \quad \bar{p}_{0,m} = \sum_{j \in A_m} p_{0,j},$$

for $m = 1, 2, \dots, M$. Then $\bar{\mathbf{p}} \sim \text{Dir}(a\bar{\mathbf{p}}_0)$. For instance, if $(p_1, p_2, p_3, p_4, p_5) \sim \text{Dir}(a_1, a_2, a_3, a_4, a_5)$, then $(p_1 + p_2, p_3 + p_5, p_4) \sim \text{Dir}(a_1 + a_2, a_3 + a_5, a_4)$.

When $\mathbf{p} \sim \text{Dir}(a\mathbf{p}_0)$, we simply refer to $\mathbf{p}_0$ as the prior mean, a naming convention justified by (ii) in Proposition 3.1. We shall refer to a as the *Dirichlet concentration parameter*, since a governs the degree to which the prior concentrates around its mean through (3.6).

The following example shows how (3.2) can be used to quickly obtain the posterior distribution in a Bayesian model.

Example 3.1 (Dirichlet–Multinomial). We consider the problem of estimating the parameters of a multinomial distribution. Assume that $\mathbf{N} \sim \text{Multinom}(n, \mathbf{p})$ for some $\mathbf{p} \in S_{k-1}$. As a prior for the unknown parameter $\mathbf{p}$, we shall use a proper Dirichlet distribution, $\mathbf{p} \sim \text{Dir}(a\mathbf{p}_0)$. Inspecting the likelihood function and the prior as functions of $\mathbf{p}$, we learn that

$$p(\mathbf{p} | \mathbf{N}) \propto \prod_{j=1}^k p_j^{N_j}, \quad p(\mathbf{p}) \propto \prod_{j=1}^k p_j^{ap_{0,j}-1}.$$

To compute the posterior distribution of $\mathbf{p} \mid \mathbf{N}$, we use the unnormalized version of Bayes' Theorem (3.2):

$$\begin{aligned}
 p(\mathbf{p} \mid \mathbf{N}) &\propto p(\mathbf{p}) p(\mathbf{N} \mid \mathbf{p}) \\
 &\propto \prod_{j=1}^k p_j^{ap_{0,j}-1} \prod_{j=1}^k p_j^{N_j} \\
 &\propto \prod_{j=1}^k p_j^{ap_{0,j}+N_j-1} \\
 &\propto \text{Dir}(\mathbf{p} \mid a\mathbf{p}_0 + \mathbf{N}).
 \end{aligned}$$

Since the posterior distribution must be a properly normalized density, we conclude that it is in fact equal to the $\text{Dir}(a\mathbf{p}_0 + \mathbf{N})$ -distribution. $\blacktriangle$

The setting of Example 3.1 is rather special in that both the prior and the posterior distributions are Dirichlet distributions with different parameters. In general, we say that the prior $p(\boldsymbol{\theta})$ is *conjugate* for the likelihood $p(\mathbf{x}_{1:n} \mid \boldsymbol{\theta})$ if the prior and posterior densities belong to the same parametric family of distributions. In conjugate models, it is often possible to infer the normalization constant directly by inspection as in Example 3.1. Other choices of prior distributions for $\boldsymbol{\theta}$ are of course possible, but in general it will no longer be possible to infer the normalization constant by inspection, so the best one can typically hope for is to know the posterior distribution up to proportionality.

The posterior predictive density

In many practical applications, the main objective is to make predictions from a statistical model. Assuming that we have observed a sample $\mathbf{x}_{1:n}$ from f_0 , we want to predict and/or quantify uncertainty in a new observation $\mathbf{x}_{n+1}$ from the same density. In the purely likelihood-based framework of Section 2.2.1, one can, for instance, obtain point predictions by computing the MLE $\hat{\boldsymbol{\theta}}$ in the model $p(\mathbf{x}_{1:n} \mid \boldsymbol{\theta})$, and then make inferences about $\mathbf{x}_{n+1}$ based on the pdf/pmf corresponding to the MLE, $p(\mathbf{x}_{n+1} \mid \hat{\boldsymbol{\theta}})$. The Bayesian approach differs quite markedly from this, as the primary object of interest in predictive modeling is the *posterior predictive distribution*,

$$p(\mathbf{x}_{n+1} \mid \mathbf{x}_{1:n}) = \int_{\Theta} p(\mathbf{x}_{n+1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}_{1:n}) d\boldsymbol{\theta}. \quad (3.7)$$

From (3.7) it is immediate how the Bayesian approach to prediction differs from the frequentist one. Whereas in the frequentist paradigm the prediction for $\mathbf{x}_{n+1}$ is computed using only a single value for the parameter $\boldsymbol{\theta}$, the predictive distribution of (3.7) weights the likelihood for a given $\boldsymbol{\theta}$ according to how probable that value of $\boldsymbol{\theta}$ is, as measured by the posterior distribution. Point estimates and uncertainty measures for $\mathbf{x}_{n+1}$ are then obtained through the posterior predictive distribution (3.7).

3.1.1 Posterior inference in nonconjugate models

The inability to infer the normalization constant in the nonconjugate setting has led to the development of methods that allow us to perform approximate inference in the posterior model. We focus our attention on *Markov chain Monte Carlo* (MCMC) methods for posterior inference, but note that there are other alternative procedures to approximating the posterior distribution, such as variational inference (Blei et al., 2017).

Our review of MCMC methods is by no means intended to be exhaustive, and the reader may consult Robert and Casella (2004) for a more thorough introduction to the theory of MCMC. Our presentation mostly follows that of Robert and Casella (2004, Chapter 6.1).

The main goal of MCMC in Bayesian inference is to generate approximate samples from the posterior distribution $p(\boldsymbol{\theta} | \mathbf{x}_{1:n})$. For an index set $\mathcal{T} \subseteq \mathbb{N} \cup \{0\}$, we say that a sequence of random variables $\{\boldsymbol{\theta}^{(t)}\}_{t \in \mathcal{T}}$ is a *Markov chain* if the distribution of $\boldsymbol{\theta}^{(t)}$ given the entire past only depends on the previous state $\boldsymbol{\theta}^{(t-1)}$, that is, $p(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t-1)}, \dots, \boldsymbol{\theta}^{(0)}) = p(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(t-1)})$ for all $t \in \mathcal{T}$. At present, we shall only concern ourselves with *time homogeneous* chains, referring to those Markov chains where the distribution of $\boldsymbol{\theta}^{(t)}$ given $\boldsymbol{\theta}^{(t-1)}$ is independent of t . In this case, the Markov chain can be characterized in terms of the initial distribution of $\boldsymbol{\theta}^{(0)}$ and a *transition kernel* $K(\cdot | \cdot)$, where $K(\cdot | \boldsymbol{\theta}^{(t-1)})$ is the probability distribution of $\boldsymbol{\theta}^{(t)}$ given $\boldsymbol{\theta}^{(t-1)}$.

If the Markov chain is well-behaved enough, then it admits a *stationary distribution*, say $\pi(\cdot)$, with the property that if $\boldsymbol{\theta}^{(t)} \sim \pi$ then $\boldsymbol{\theta}^{(t+1)} \sim \pi$ as well. This stationary distribution also acts as a limiting distribution of the chain in the sense that the distribution of $\boldsymbol{\theta}^{(t)}$ converges to π as $t \rightarrow \infty$. Thus, if the chain is run for long enough, then the distribution of $\boldsymbol{\theta}^{(t)}$ approximates that of π . The two conditions that ensure that this property holds is that the chain is *aperiodic* and *irreducible*. The *periodicity* of a chain simply means that the chain moves in cycles, i.e. we can partition the state space Θ such that the chain visits the sets of the partition in a periodic manner with probability 1. The chain is then said to be aperiodic if it is not periodic. Irreducibility refers to the fact that the chain can move from any single state $\boldsymbol{\theta}$ to all sets $\mathcal{A}$ with positive probability under π in a finite number of steps. More formally, under some conditions on the chain, the distribution of $\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^{(0)}$, denoted $K^{(t)}(\cdot | \boldsymbol{\theta}^{(0)})$, converges to π in the $\mathbb{L}_1$ sense,

$$d_1(\pi, K^{(t)}(\cdot | \boldsymbol{\theta}^{(0)})) \rightarrow 0, \quad \text{as } t \rightarrow \infty. \quad (3.8)$$

Another key result from the theory of Markov chains is that a variant of the Law of Large Numbers holds in the limit as $t \rightarrow \infty$. Specifically, if h is a function satisfying $\mathbb{E}_{\boldsymbol{\theta}}\{|h(\boldsymbol{\theta})|\} < \infty$ where $\boldsymbol{\theta} \sim \pi$, then

$$\frac{1}{T} \sum_{t=0}^{T-1} h(\boldsymbol{\theta}^{(t)}) \xrightarrow{\text{a.s.}} \mathbb{E}_{\boldsymbol{\theta}}\{h(\boldsymbol{\theta})\}. \quad (3.9)$$

The result of Equation (3.9) suggests that we can approximate $\mathbb{E}_{\boldsymbol{\theta}}\{h(\boldsymbol{\theta})\}$ by a simple Monte Carlo average, if we generate enough samples from a Markov chain with stationary distribution π .

The modus operandi of MCMC is to construct a Markov chain which has the posterior as the stationary distribution. A popular way of constructing such a Markov chain with the desired property is via the *Metropolis–Hastings* (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). The MH algorithm proceeds iteratively by proposing a new point $\boldsymbol{\theta}^*$ from a proposal distribution $q(\cdot | \boldsymbol{\theta}^{(t)})$. One then either accepts or rejects the proposed sample,

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}^*, & \text{with probability } \min\{1, A(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(t)})\} \\ \boldsymbol{\theta}^{(t)}, & \text{else} \end{cases}$$

where $A(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(t)})$ is the so-called *Metropolis–Hastings ratio*,

$$A(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(t)}) = \frac{p(\boldsymbol{\theta}^* | \mathbf{x}_{1:n}) q(\boldsymbol{\theta}^{(t)} | \boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t)} | \mathbf{x}_{1:n}) q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(t)})}. \quad (3.10)$$

Importantly, this sampling step does not require knowledge of the normalization constant in (3.2) since the MH ratio only depends on the posterior density through the ratio $p(\boldsymbol{\theta}^* | \mathbf{x}_{1:n})/p(\boldsymbol{\theta}^{(t)} | \mathbf{x}_{1:n})$. This fact makes the MH algorithm an attractive choice for posterior inference in nonconjugate Bayesian models. The proposal $q(\cdot | \cdot)$ need not be any prespecified distribution, and can be adapted to the problem at hand, leaving the statistician with a great deal of flexibility when designing a Metropolis–Hastings algorithm for a given problem.

Another important feature of MCMC algorithms in multivariate problems is that all components of the vector $\boldsymbol{\theta}$ do not have to be updated in each iteration in order for the generated samples to be valid approximations to samples from the target π . For example, if the vector $\boldsymbol{\theta}$ has been divided into blocks $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_p)$, where $\boldsymbol{\theta}_l$ is a k_l -dimensional vector, it can be shown that using an MH-step to update one of the blocks of the vector in a cyclical or random fashion yields a Markov chain with the desired stationary distribution, and the same convergence guarantees apply as previously, provided the chain is aperiodic and irreducible. We refer to the cyclical algorithm described above as *systematic scan Metropolis–Hastings*, and the random variant is termed *random scan Metropolis–Hastings*.

Although the convergence result of (3.8) justify the use of an MCMC sample as an approximate sample from the posterior provided that T is large enough, they do not tell us what value of T is appropriate in practice. Depending on the target density and the chosen proposal distribution, the chain may be very slow to properly explore the part of the parameter space with high probability for a given finite iteration budget. In such a setting, the MCMC sample can be a rather poor approximation to the target distribution. Another issue is that the variance of the estimate in (3.9) can be quite large due to the dependence of the samples, which can lead to imprecise point estimates of key posterior quantities of interest. To assess whether the Markov chain has reached the stationary distribution and that we have generated a sufficient amount of samples to estimate a quantity of interest to the desired precision, various convergence diagnostics and effective sample size measures are employed in practice. As we shall rely on some of these convergence checks for our MCMC examples later in the text, a brief introduction is given in Appendix D.1.

Example 3.2 (Gibbs sampler for Bayesian mixture density estimation). We revisit the mixture model of Example 2.2. For a Bayesian analysis of such a mixture model, we elicit a prior distribution on the parameters $\boldsymbol{\theta}$, one convenient choice being $\boldsymbol{\theta} \sim \text{Dir}(a\boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \mathcal{S}_{k-1}$ and $a > 0$. This yields the following model for the data and the parameters:

$$\begin{aligned}\boldsymbol{\theta} &\sim \text{Dir}(a\boldsymbol{\theta}_0), \\ f(x | \boldsymbol{\theta}) &= \sum_{j=1}^k \theta_j \omega_j(x), \\ x_i | f &\sim f(x_i | \boldsymbol{\theta}).\end{aligned}$$

The posterior distribution in this model then becomes

$$p(\boldsymbol{\theta} | \mathbf{x}_{1:n}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^n f(x_i | \boldsymbol{\theta}) \propto \prod_{j=1}^k \theta_j^{a\theta_{0,j}-1} \prod_{i=1}^n \sum_{j=1}^k \theta_j \omega_j(x_i). \quad (3.11)$$

The expression for the posterior distribution (3.11) is difficult to work with directly, since it does not belong to any known family of distributions for which we can easily generate

samples or compute simple posterior summaries such as posterior means and variances. However, by introducing the hidden class label $\mathbf{c}_{1:n}$ the posterior becomes amenable to posterior inference via a special type of systematic scan MH algorithm, allowing us to generate approximate samples from (3.11). Since we can write the likelihood contribution of sample x_i as $\theta_{c_i} \omega_{c_i}(x_i) = \prod_{j=1}^k \{\theta_j \omega_j(x_i)\}^{\mathcal{I}(c_i=j)}$ in this case, we derive that the joint posterior of $\boldsymbol{\theta}, \mathbf{c}_{1:n}$ becomes

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{c}_{1:n} | \mathbf{x}_{1:n}) &\propto p(\mathbf{x}_{1:n}, \mathbf{c}_{1:n}, \boldsymbol{\theta}) \\ &\propto \prod_{j=1}^k \theta_j^{a\theta_{0,j}-1} \prod_{i=1}^n \prod_{j=1}^k \{\theta_j \omega_j(x_i)\}^{\mathcal{I}(c_i=j)} \\ &= \prod_{j=1}^k \theta_j^{a\theta_{0,j}+N_j-1} \prod_{i=1}^n \omega_{c_i}(x_i), \end{aligned} \quad (3.12)$$

where $N_j = \sum_{i=1}^n \mathcal{I}(c_i = j)$ is the number of observations with class label equal to j . This setup is especially amenable to the class of MCMC methods known as *Gibbs sampling* (Geman and Geman, 1984). The Gibbs sampler proceeds by constructing a Markov chain by sampling from the conditional distributions $p(\boldsymbol{\theta} | \mathbf{x}_{1:n}, \mathbf{c}_{1:n})$ and $p(\mathbf{c}_{1:n} | \mathbf{x}_{1:n}, \boldsymbol{\theta})$ in an alternating fashion. Given initial values for $\mathbf{c}_{1:n}^{(0)}$, the Gibbs sampler proceeds as follows:

- 1) Set $t = 0$ and initialize $\mathbf{c}_{1:n}^{(0)}$.
- 2) Sample $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta} | \mathbf{x}_{1:n}, \mathbf{c}_{1:n}^{(t)})$
- 3) Sample $\mathbf{c}_{1:n}^{(t+1)} \sim p(\mathbf{c}_{1:n} | \mathbf{x}_{1:n}, \boldsymbol{\theta}^{(t+1)})$
- 4) Increment t by 1 and return to 2) until the desired number of samples has been generated.

The above scheme in fact corresponds to a systematic scan MH algorithm, where we use the conditional distributions $p(\boldsymbol{\theta} | \mathbf{x}_{1:n}, \mathbf{c}_{1:n})$ and $p(\mathbf{c}_{1:n} | \mathbf{x}_{1:n}, \boldsymbol{\theta})$ as proposal distributions, see for instance Theorem 10.13 in (Robert and Casella, 2004). It can be shown in this case that the MH ratio (3.10) always equals 1 in this case, so we will always accept the proposed sample.

The main benefit of running a Gibbs sampler in this case is that both the required conditional distributions are easy to sample from. For the first conditional distribution, direct inspection of (3.12) reveals that

$$p(\boldsymbol{\theta} | \mathbf{x}_{1:n}, \mathbf{c}_{1:n}) \propto p(\mathbf{x}_{1:n}, \mathbf{c}_{1:n}, \boldsymbol{\theta}) \propto \prod_{j=1}^k \theta_j^{a\theta_{0,j}+N_j-1} \propto \text{Dir}(\boldsymbol{\theta} | a\boldsymbol{\theta}_0 + \mathbf{N}),$$

which is easily sampled from via standard software. For the conditional distribution of the class labels, we find that

$$p(\mathbf{c}_{1:n} | \mathbf{x}_{1:n}, \boldsymbol{\theta}) \propto p(\mathbf{x}_{1:n}, \mathbf{c}_{1:n}, \boldsymbol{\theta}) \propto \prod_{i=1}^n \prod_{j=1}^k \{\theta_j \omega_j(x_i)\}^{\mathcal{I}(c_i=j)}$$

From the above expression we see that $\mathbf{c}_{1:n}$ are independent conditional on $\boldsymbol{\theta}$, and hence we can sample from this distribution by generating $c_1, \dots, c_n$ independently from categorical distributions with

$$p(c_i = j | \mathbf{x}_{1:n}, \boldsymbol{\theta}) \propto \theta_j \omega_j(x_i), \quad \text{for } j = 1, 2, \dots, k.$$

Having generated MCMC samples $\{\boldsymbol{\theta}^{(t)}, \mathbf{c}_{1:n}^{(t)}\}$ from the Gibbs sampler, inference for the posterior distribution of the density of the sample can be carried out through the implied samples for f ,

$$f^{(t)}(x) = \sum_{j=1}^k \theta_j^{(t)} \omega_j(x).$$

From the above samples, we can compute point estimates and quantify uncertainty. Point estimates for the density f are easily obtained by a simple Monte Carlo average,

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^T f^{(t)}(x),$$

which is an approximation of the posterior mean $\mathbb{E}_f\{f(x) \mid \mathbf{x}_{1:n}\}$. A common uncertainty measure for the value of the density f at a point x is a *credible interval* at level $1 - \alpha$, which is an interval $[a(x), b(x)]$ satisfying

$$P(f(x) \in [a(x), b(x)] \mid \mathbf{x}_{1:n}) = 1 - \alpha. \quad (3.13)$$

Given access to approximate samples from the posterior distribution of $f(x)$, an approximate credibility interval can then be constructed via the sample $\alpha/2$ - and $1 - \alpha/2$ -quantiles. $\blacktriangle$

3.1.2 Bayesian model selection

In the Bayesian setting there is also a need for model selection procedures when multiple models are posited for a given dataset. As it turns out, once a full Bayesian model has been specified, the task of model selection is in principle quite straightforward, as there is no need to consider which information criterion to use; Bayes' theorem automatically identifies the best model for us. The key difference to the frequentist model selection setting of Section 2.2.3 is that the model $m \in \mathcal{M}$, where $\mathcal{M}$ is a discrete set, is now considered a random variable, with its own prior distribution $p(m)$. For a given model m , we proceed more or less as before, specifying a prior distribution $p(\boldsymbol{\theta}_m \mid m)$ for the parameter vector $\boldsymbol{\theta}_m$ and a likelihood function $p(\mathbf{x}_{1:n} \mid m, \boldsymbol{\theta}_m)$. This yields the corresponding *marginal model likelihood* for model m :

$$p(\mathbf{x}_{1:n} \mid m) = \int_{\Theta_m} p(\mathbf{x}_{1:n} \mid m, \boldsymbol{\theta}_m) p(\boldsymbol{\theta}_m \mid m) d\boldsymbol{\theta}_m. \quad (3.14)$$

Having specified the required priors and likelihoods, we can compute the posterior probability of each model $p(m \mid \mathbf{x})$ by applying the discrete version of Bayes' formula:

$$p(m \mid \mathbf{x}_{1:n}) = \frac{p(m) p(\mathbf{x}_{1:n} \mid m)}{p(\mathbf{x}_{1:n})} = \frac{p(m) p(\mathbf{x}_{1:n} \mid m)}{\sum_{m' \in \mathcal{M}} p(m') p(\mathbf{x}_{1:n} \mid m')}. \quad (3.15)$$

The best candidate model is the one for which the posterior probability (3.15) is maximized,

$$\hat{m} = \arg \max_{m \in \mathcal{M}} p(m \mid \mathbf{x}_{1:n}). \quad (3.16)$$

Since the likelihood $p(\mathbf{x}_{1:n})$ is the same for each model, $\hat{m}$ is also the maximizer of the unnormalized posterior,

$$p(m \mid \mathbf{x}_{1:n}) \propto p(m) p(\mathbf{x}_{1:n} \mid m). \quad (3.17)$$

We refer to $\hat{m}$ as the *Maximum A Posteriori* (MAP) model, since it maximizes the posterior probability.

Although the Bayesian approach to model selection is conceptually simple, one often runs into computational challenges in practice. For nonconjugate models, the marginal model likelihoods (3.14) cannot be computed in closed form. This in itself does not preclude the use of such a model, as we can resort to various approximation techniques for marginal likelihood computation.

3.1.2a The BIC approximation

Although it is routinely used to discriminate between purely frequentist models, the BIC of Section 2.2.3 was originally motivated on entirely Bayesian grounds, as a scaled variant of the BIC of model m can be shown to be a valid large-sample approximation to the log-posterior probability of model m (Claeskens and Hjort, 2008, p. 80). To be specific, under some regularity conditions, we can write

$$\log p(m | \mathbf{x}_{1:n}) = \log L_n(\hat{\boldsymbol{\theta}}_m | m) - \frac{1}{2}d_m \log(n) + \mathcal{O}_{\mathbb{P}}(1), \quad (3.18)$$

where $\hat{\boldsymbol{\theta}}_m$ is the MLE in model m . Since the first term is $\mathcal{O}_{\mathbb{P}}(n)$ and the second grows as $\log(n)$, the two first terms will tend to dominate the expression for the log-posterior when n is large. Thus, when selecting the model m among a fixed set of models $\mathcal{M}$, BIC and standard Bayesian model selection procedures will tend to select the same model when the sample size is sufficiently large. The BIC is often very quick to compute, as evaluating the maximal log-likelihood of a model is typically not a very computationally demanding task, allowing the statistician to discriminate quickly between different models.

3.1.2b Model averaging

In density estimation problems, we are usually interested in obtaining a good estimate of the density, possibly together with some form of uncertainty quantification. In such cases, when multiple models have been posited for the density, *model averaging* constitutes an attractive alternative to model selection as a means of providing a point estimate. Given a quantity h with a common interpretation across models, the full posterior distribution for h is given by

$$p(h | \mathbf{x}_{1:n}) = \sum_{m \in \mathcal{M}} p(h | m, \mathbf{x}_{1:n}) p(m | \mathbf{x}_{1:n}). \quad (3.19)$$

A possible point estimate of h is the posterior mean, which then becomes

$$\hat{h} = \sum_{m \in \mathcal{M}} \mathbb{E}_h\{h | m\} p(m | \mathbf{x}_{1:n}), \quad (3.20)$$

which is simply a convex combination of the posterior means in each of the candidate models. The estimator (3.20) reflects that the model choice itself is uncertain, and weights each of the candidate models according to their posterior probability. This is in stark contrast to model selection-based estimators of h , where inferences are based on $p(h | \mathbf{x}_{1:n}, \hat{m})$, where $\hat{m}$ is chosen according to some model selection criterion, and not the full posterior distribution (3.19).

Frequentist analogies to (3.20) have also been developed to take into account uncertainty in the model choice in the estimation procedure (Hjort and Claeskens, 2003),

yielding estimators of the form

$$\hat{h}_{\text{MA}} = \sum_{m \in \mathcal{M}} w_m \hat{h}_m, \quad (3.21)$$

where $\hat{h}_m$ is the maximum likelihood estimator of h_m under model m and w_m are weights determined by some information criterion (IC), such as the BIC or the AIC, through

$$w_m = \frac{\exp\{\text{IC}(m)\}}{\sum_{m' \in \mathcal{M}} \exp\{\text{IC}(m')\}}. \quad (3.22)$$

In the parametric Bayesian setting, the frequentist model averaging estimator using BIC weights can be seen as an approximation to the true model averaging estimator, due to the fact that the BIC of a model is large-sample equivalent to its posterior probability, see (3.18).

3.1.3 Bayesian decision theory

When performing Bayesian analysis, inference for parameters of interest is available through their respective posterior distributions. However, the statistician will often be interested in providing point estimates of the parameters upon seeing the data. From the perspective of Bayesian inference, the optimal point estimate is given by the choice of prior distribution for the parameter as well as the chosen loss function ℓ . The Bayesian approach seeks to find the action $\delta = \delta(\mathbf{x}_{1:n})$ minimizing the so-called *Bayes risk*, which is the expected frequentist risk (2.21),

$$\text{BR}(\delta) = \int_{\Theta} R_n(\theta, \delta) p(\theta) d\theta = \int_{\Theta} \mathbb{E}_{\mathbf{x}_{1:n}} \{\ell(\theta, \delta)\} p(\theta) d\theta. \quad (3.23)$$

It can be shown that the minimizer of the Bayes risk (3.23) is the minimizer of the posterior expected loss (Lehmann and Casella, 2006, p. 228),

$$\hat{\theta}_B = \arg \min_{\delta} \int \ell(\theta, \delta) p(\theta | \mathbf{x}_{1:n}) d\theta. \quad (3.24)$$

The resulting estimator $\hat{\theta}_B$ is then termed the *Bayes estimator*. Throughout the text, if a particular loss function has not been specified, the term Bayes estimator refers to the Bayes estimator under $\mathbb{L}_2$ loss.

Example 3.3 (Model selection). The model selection procedure suggested in Equation (3.16) can be motivated on decision-theoretic grounds. If we take our loss function to be the so-called 0-1 loss, $\ell(m, \mu) = \mathcal{I}(m \neq \mu)$, then the posterior expected loss becomes

$$\begin{aligned} \mathbb{E}_m \{\mathcal{I}(m \neq \mu(\mathbf{x}_{1:n})) | \mathbf{x}_{1:n}\} &= \sum_{m \in \mathcal{M}} P(m \neq \mu | \mathbf{x}_{1:n}) \mathcal{I}(\mu(\mathbf{x}_{1:n}) = m) \\ &= \sum_{m \in \mathcal{M}} \{1 - p(m | \mathbf{x}_{1:n})\} \mathcal{I}(\mu(\mathbf{x}_{1:n}) = m). \end{aligned} \quad (3.25)$$

The expression (3.25) is then minimized by taking μ equal to the a posteriori most probable model as in Equation (3.16). $\blacktriangle$

Example 3.4 (Estimation of multinomial parameters). Consider the setup of Example 3.1 where we have observed multinomial counts $\mathbf{N} | p \sim \text{Multinom}(n, \mathbf{p})$ and use the prior $\mathbf{p} \sim \text{Dir}(a\mathbf{p}_0)$. We want to obtain a point estimate of the probability vector $\mathbf{p}$. A reasonable loss function is that induced by the $\mathbb{L}_r$ norms on $\mathbb{R}^k$,

$$\ell(\mathbf{p}, \boldsymbol{\delta}) = \|\mathbf{p} - \boldsymbol{\delta}\|_r^r = \sum_{j=1}^k |p_j - \delta_j|^r.$$

Since $\mathbf{p} \in \mathcal{S}_{k-1}$, we would like for the estimator $\boldsymbol{\delta}$ to be a k -simplex vector as well, so that we may interpret the resulting estimate as probabilities. Following the recipe of (3.24), the Bayes estimator is given by

$$\arg \min_{\boldsymbol{\delta} \in \mathcal{S}_{k-1}} \int \sum_{j=1}^k |p_j - \delta_j|^r p(\mathbf{p} | \mathbf{N}) d\mathbf{p} = \arg \min_{\boldsymbol{\delta} \in \mathcal{S}_{k-1}} \sum_{j=1}^k \int |p_j - \delta_j|^r p(p_j | \mathbf{N}) dp_j. \quad (3.26)$$

The constraint in (3.26) makes it hard to simplify the expression for the Bayes estimator in general. However, for $r = 2$ we can write down the resulting estimator in closed form by considering the corresponding unconstrained problem to (3.26). It is well known that the minimizer of

$$\min_{\delta_j} \int (p_j - \delta_j)^2 p(p_j | \mathbf{N}) dp_j$$

are the marginal posterior means, $\hat{p}_{B,j} = \mathbb{E}_{p_j}\{p_j | \mathbf{N}\}$, see for instance Lehmann and Casella (2006, pp. 228). Since the unconstrained minimum of (3.26) for $r = 2$ is given by $\hat{\mathbf{p}}_B$, the solution to the constrained problem cannot attain a smaller value of the criterion since adding a constraint means that we must search for the optimum over a smaller set of possible solutions. Now, by linearity of expectation,

$$\sum_{j=1}^k \hat{p}_{B,j} = \sum_{j=1}^k \mathbb{E}_{p_j}\{p_j | \mathbf{N}\} = \mathbb{E}_{\mathbf{p}} \left\{ \sum_{j=1}^k p_j \mid \mathbf{N} \right\} = 1.$$

Moreover, as $P(p_j \geq 0) = 1$ it follows that $\mathbb{E}_{p_j}\{p_j | \mathbf{N}\} \geq 0$ and hence we conclude that $\hat{\mathbf{p}}_B \in \mathcal{S}_{k-1}$. Thus, $\hat{\mathbf{p}}_B$ is also a solution to the constrained problem. Using Proposition 3.1 and the fact that $\mathbf{p} | \mathbf{N} \sim \text{Dir}(a\mathbf{p}_0 + \mathbf{N})$, the Bayes estimator becomes

$$\hat{p}_{B,j} = \mathbb{E}_{p_j}\{p_j | \mathbf{N}\} = \frac{ap_{0,j} + N_j}{a + n} = \frac{a}{a + n}p_{0,j} + \frac{n}{a + n}N_j/n. \quad (3.27)$$

The Bayes estimator in this case is a convex combination of the prior mean $p_{0,j}$ and the observed proportion N_j/n , which is the maximum likelihood estimator for p_j in this case. As n grows large, we see that the weight assigned to the prior mean goes to zero, so the Bayes estimator and MLE will tend to be very close for larger sample sizes. $\blacktriangle$

3.2 Nonparametric Bayes

As in the frequentist case, nonparametric Bayesian procedures offer a more flexible alternative to its parametric counterparts. By using a prior distribution which spreads its mass over a larger space of functions, a nonparametric Bayesian model provides additional flexibility not offered by parametric models, as the models are able to use an infinite or growing number of parameters along with the sample size. Bayesian models can, in general, be described through the specification of a likelihood $\mathbf{x}_{1:n} | f \sim f$ for

some probability density f , and a prior distribution on f , $f \sim P(f)$, inducing a posterior distribution $P(\cdot | \mathbf{x}_{1:n})$. In the parametric setting, the prior for f could be indexed by a finite-dimensional parameter θ , so that the prior for f could be induced by specifying the prior density $p(\theta)$. In a Bayesian nonparametric model, the prior for f will in general take a more complicated form and often cannot be described by a single density, as in Example 3.5.

Example 3.5 (Random Histograms). We once again consider the regular histogram model of Example 2.3. We first consider a parametric Bayesian procedure, where we fix $k = 5$. In this case, we can induce a prior on f by using a simplex-prior for the probability vector $\mathbf{p}$, with density $p(\mathbf{p})$. As mentioned in Example 2.3 we can get a more flexible model by treating the number k of bins as a parameter. To have a fully Bayesian model, this entails specifying a prior distribution for k , so that $k \sim p(k)$. For each value of k , we denote the associated $\mathcal{S}_{k-1}$ -valued probability vector by $\mathbf{p}_k$. In the Bayesian setting, each $\mathbf{p}_k$ is equipped with its own prior distribution $p(\mathbf{p}_k | k)$. The overall prior distribution can thus be written as an infinite mixture of simplex priors,

$$P(\cdot) = \sum_{k=1}^{\infty} p(k)P(\cdot | k),$$

where $P(\cdot | k)$ is the probability measure induced by the density $p(\mathbf{p}_k | k)$. For given values of k , and a probability vector $\mathbf{p}_k \in \mathcal{S}_{k-1}$, the likelihood function for a sample $x_i | k, \mathbf{p}_k \sim f(x_i | k, \mathbf{p}_k)$ is still given by

$$f(x_i | k, \mathbf{p}_k) = k \prod_{j=1}^k p_{k,j}^{\mathcal{I}(x_i \in I_{k,j})}.$$

We can write the full model for the prior and likelihood as a hierarchical scheme,

$$\begin{aligned} k &\sim p(k), \\ \mathbf{p}_k | k &\sim p(\mathbf{p}_k | k), \\ f(x | k, \mathbf{p}_k) &= k \prod_{j=1}^k p_{k,j}^{\mathcal{I}(x \in I_{k,j})}, \\ x_i | k, \mathbf{p}_k &\sim f(x_i | k, \mathbf{p}_k). \end{aligned} \tag{3.28}$$

By using a prior distribution that puts positive mass on any $k \in \mathbb{N}$, we have a similarly flexible model to that of Example 2.3, provided the priors $p(\mathbf{p}_k | k)$ assign probability to any set in the interior of the simplex of nonzero Lebesgue measure. $\blacktriangle$

Example 3.5 illustrates well the main motivation behind using a nonparametric Bayesian procedure over a parametric one, in that the nonparametric prior assigns positive probability to probability distributions closely resembling a quite general class of densities, a condition we informally refer to as a *big support* condition. The mention of ‘closeness’ and ‘resemblance’ begs the question of how these term should be measured. The goal of the next section is to formalize the notion of big support mathematically, and in what sense we would like for big support conditions to hold.

Remark 3.1. Before delving into the specifics of big support, we make a brief detour to highlight some issues pertaining to the choice of prior distributions. In some Bayesian problems we may wish to use a sequence of prior distributions $\{P_n\}$ for f instead of a fixed prior P . If we observe a sample $\mathbf{x}_{1:n} | f \sim f$, of size n , then the prior distribution for

f is given by $f \sim P_n(f)$. In particular, the prior distribution may be chosen to depend on the sample size. This is at odds with purely Bayesian philosophy, where the prior should represent our beliefs about the parameter with no reference to the observed sample at all. However, sample-size dependent priors turn out to be quite useful in certain nonparametric problems. A strong argument in favor of using sample-size dependent priors is that highly flexible models will tend to overfit for small datasets. This means that using a large amount of parameters relative to the number of samples will in many situations lead to models that track the data too closely, resulting in rather poor fits. By assigning zero prior probability to models that are very complex, one can circumvent this issue, as the posterior probability of any such model would also be zero. To ensure that the model remains flexible, we can choose the prior sequence such that the maximal number of potential parameters used grows with the sample size. This approach is also quite beneficial from a purely pragmatic perspective, as we only have to deal with a finite number of parameters for any given sample of size n , which can significantly simplify posterior computations.

The idea of using a prior distribution whose support increases with the sample size is similar to the sieve estimators introduced in Section 2.2.4. Due to this similarity, priors of this type are generally referred to as *sieve priors*.

3.2.1 The Kullback-Leibler support condition

The matter of establishing large prior support has been at the heart of nonparametric Bayesian methods since its inception with the Dirichlet Process (Ferguson, 1973). When establishing that a given nonparametric procedure has so-called large prior support, one is typically interested in showing that the prior puts positive probability on densities that are in some sense close to an arbitrary $f_0(\cdot)$ belonging to some ‘large’ class of densities $\mathcal{F}$. In nonparametric Bayes, the class $\mathcal{F}$ should be rather general, for instance the class of all bounded density functions on $[0, 1]$. It is rather immediate why establishing such conditions is important; if the prior places zero mass on neighborhoods around f_0 , one cannot hope that the posterior will concentrate on such sets as the sample size tends to infinity since such sets will also be assigned zero posterior mass. Often, it will be the case that the prior puts zero mass on a subset $\mathcal{G} \subseteq \mathcal{F}$ to which the true f_0 belongs. However, the prior may still put positive mass on arbitrary neighborhoods of any such f_0 . As an example, if we take $\mathcal{G}$ to be the set of all continuous densities for which $f(0) = 0$ and $f(1) = 1$, then the nonparametric random histogram prior of Example 3.5 assigns zero mass to $\mathcal{G}$, since all densities in $\mathcal{G}$ are not piecewise constant. However, it is the case, as we shall prove in later sections, that the prior assigns positive mass to elements closely resembling those in $\mathcal{G}$.

Let P_f, P_g be probability measures on $\mathcal{X} \subseteq \mathbb{R}$ with densities f and g , respectively. We denote the *Kullback–Leibler divergence* between f and g by $K(f, g)$, defined as

$$K(f, g) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx, \quad (3.29)$$

if $P_g(\mathcal{A}) = 0 \implies P_f(\mathcal{A}) = 0$ for all measurable sets $\mathcal{A}$. If there exists a measurable set $\mathcal{A}$ such that $P_f(\mathcal{A}) > 0$ and $P_g(\mathcal{A}) = 0$, we put $K(f, g) = \infty$. We note that the expression in (3.29) is not symmetric, so the KL-divergence is not a metric. However, the KL-divergence satisfies $K(f, g) \geq 0$ with equality if and only if $f = g$, see Proposition A.5 for a proof. A KL-ball of radius ϵ around a density f_0 is defined¹ as $\{f \in \mathcal{F} : K(f_0, f) < \epsilon\}$.

¹Due to the aforementioned asymmetry, the order of the arguments is important here.

When establishing large prior support one is usually interested in showing that the prior puts positive probability on Kullback–Leibler (KL) neighborhoods of the true density.

Definition 3.1 (Kullback–Leibler support condition). Let P be a prior distribution on $\mathcal{F}$. We say that f_0 is in the *KL-support* of P if, for all $\epsilon > 0$ we have that

$$P(f \in \mathcal{F} : K(f_0, f) < \epsilon) > 0. \quad (3.30)$$

In many nonparametric problems one may want to use a sample-size dependent sequence of prior distributions $\{P_n\}$ instead. In this case, the *KL-support condition* becomes

$$\liminf_{n \rightarrow \infty} P_n(f \in \mathcal{F} : K(f_0, f) < \epsilon) > 0, \quad \text{for every } \epsilon > 0, \quad (3.31)$$

and we say that f_0 is in the *KL-support* of the sequence $\{P_n\}$.

Showing that f_0 lies in the KL-support of the prior is of particular importance as it can be used to show that the posterior distribution concentrates around f_0 under some additional conditions on the prior.

3.2.2 Various notions of consistency

This subsection provides a brief introduction to the issue of consistency of posterior distribution, with particular attention paid to those issues most relevant to the field of Bayesian nonparametrics. For a more complete review of issues pertaining to posterior consistency in nonparametric problems, we refer the reader to Hjort et al. (2010, Section 2.4) or Ghosal and van der Vaart (2017, Section 6.1).

The tools used to study point estimators in the nonparametric Bayesian setting are somewhat different than the ones typically used to study consistency of frequentist procedures. While frequentists are most often concerned with studying the convergence of a point estimator $\hat{f}_n$ to the true f_0 , Bayesian nonparametric studies of consistency are mostly based on studying the behavior of the posterior distribution. It should be noted that the concept of consistency is a purely frequentist one, as it is built on the assumption that the data is generated according to a fixed unknown density f_0 . When the Bayesian model has been formulated, the frequentist framework treats the posterior distribution itself as a random quantity, depending only on the observed sample $\mathbf{x}_{1:n}$. For Bayesian approaches to density estimation problems, consistency informally means that we would like the (random) posterior distribution to increasingly concentrate on densities arbitrarily close to the true data-generating density with probability tending to one.

While being of interest in its own right, Bayesian consistency also ensures that the choice of prior distribution does not impact the resulting posterior too much when the sample sizes grow larger, ensuring a kind of robustness to the choice of prior distribution. In fact, posterior consistency is both a sufficient and necessary condition for the posterior distributions based on different priors to become arbitrarily close with respect to some metric as the sample size approaches infinity, see Ghosal and van der Vaart (2017, Theorem 6.6).

The issue of consistency introduces the need for a way to measure the distance between probability densities belonging to a set of probability densities $\mathcal{F}$. One way is to define neighborhoods based on a metric d on $\mathcal{F}$. In nonparametric Bayes, it is often most convenient to work with the $\mathbb{L}_1$ metric (2.27) or the Hellinger distance (2.28), although in some problems, it can be advantageous to work with the metric induced by

the $\mathbb{L}_2$ norm instead. Once a suitable metric has been chosen, open balls around any given f_0 are given by (2.29).

Another possibility is to use the topology of weak convergence, where we define the ball of radius $\epsilon > 0$ around any $f_0 \in \mathcal{F}$ by

$$B_\epsilon(f_0) = \left\{ f \in \mathcal{F} : \left| \int g_i(x)f(x) dx - \int g_i(x)f_0(x) dx \right| < \epsilon, i = 1, 2, \dots, K \right\},$$

where the g_i are continuous and bounded probability density functions on $\mathbb{R}$.

Once a suitable notion of open sets has been chosen, one can define what it means for the posterior distribution to be consistent at f_0 in the given topology:

Definition 3.2 (Posterior consistency). Let $\{P_n\}$ be a sequence of prior distributions and f_0 a probability density function. We say that the posterior distribution is *consistent* at f_0 if for every $\epsilon > 0$,

$$P_n(B_\epsilon(f_0) \mid \mathbf{x}_{1:n}) \xrightarrow{\mathbb{P}} 1 \text{ as } n \rightarrow \infty. \quad (3.32)$$

If the posterior is consistent with respect to the weak topology on $\mathcal{F}$, we say that it is *weakly consistent* at f_0 , while consistency in the metric topology is termed *strong consistency*. It turns out that for nonparametric problems, showing the former is quite a lot easier than showing the latter. This is due to a result from Schwartz (1965), stated below. For a proof, the reader may consult Walker (2003, Corollary 2).

Theorem 3.1. *If f_0 lies in the Kullback-Leibler support of the prior, then the posterior distribution is weakly consistent at f_0 .*

While Theorem 3.1 gives one a criterion that is fairly easy to check, establishing posterior consistency in the strong sense is in general more difficult. Barron et al. (1999) argue that a weak neighborhood around f_0 contains many densities that fail to resemble f_0 , so that there is a need to verify that strong consistency holds. In a few nonparametric problems where the prior is conjugate, one can establish posterior consistency with respect to the metric d_2 by working with the posterior distribution directly, see for instance Theorem 4.4. To establish posterior consistency in a nonconjugate model, one typically has to resort to more general results. Papers providing sufficient conditions for strong consistency typically require additional properties to hold in addition to the KL-support condition in order to establish consistency with respect to either the $\mathbb{L}_1$ or Hellinger metric, see for instance Barron et al. (1999); Ghosal et al. (1999); Walker (2003, 2004) for examples of such theorems. The approach we pursue here towards establish strong posterior consistency makes use of covering numbers. Before introducing our main tool for establishing convergence, we therefore make a brief detour, discussing both covering numbers and the related notion of bracketing numbers.

3.2.3 Covering and bracketing numbers

In brief, covering and bracketing numbers allow us to measure the size of a given set. These play an important role in the convergence proofs later in the text, as controlling the size of sets consisting of functions of a given type becomes necessary. We will keep our discussion of these quantities rather brief, focusing only on the properties we will need in our proofs later down the line. For a more thorough introduction to covering and bracketing numbers, one may consult van de Geer (2000). In the following, we will study these quantities in the context of a metric space $(\mathcal{X}, d)$. We start with covering numbers.

Definition 3.3 (Covering number). An ϵ -cover of a set $\mathcal{Y} \subseteq \mathcal{X}$ is a finite collection of open balls of radius ϵ centered at points $x_1, \dots, x_n \in \mathcal{X}$ such that

$$\mathcal{Y} \subseteq \cup_{i=1}^n B_\epsilon(x_i). \quad (3.33)$$

The ϵ -covering number of the set $\mathcal{Y}$, denoted $N(\epsilon, \mathcal{Y}, d)$, is defined to be the minimal number of balls of size ϵ needed to cover $\mathcal{Y}$, that is,

$$N(\epsilon, \mathcal{Y}, d) = \min \{n \in \mathbb{N}: \mathcal{Y} \subseteq \cup_{i=1}^n B_\epsilon(x_i) \text{ for some } x_1, \dots, x_n \in \mathcal{X}\}. \quad (3.34)$$

Since the metrics we will be using are often those induced by a norm, we will occasionally write $N(\epsilon, \mathcal{Y}, \|\cdot\|)$ to mean the ϵ -covering number using the metric induced by $\|\cdot\|$.

Another notion of the size of a given set are bracketing numbers. To define bracketing numbers, we also require an *order relation* $\leq$ on the given metric space $\mathcal{X}$. That $\leq$ is an order relation means that it is *transitive, reflexive and antisymmetric*. This means that the following holds for all $x, y, z \in \mathcal{X}$:

$$x \leq y \text{ and } y \leq z \implies x \leq z, \quad (\text{Transitivity})$$

$$x \leq x, \quad (\text{Reflexivity})$$

$$x \leq y \text{ and } y \leq x \implies x = y. \quad (\text{Antisymmetry})$$

The order relation does not have to be a *total order*, meaning that we need not have $x \leq y$ or $y \leq x$ for any two $x, y \in \mathcal{X}$. We refer to such order relations as *partial order relations*. Since we will need bracketing numbers for both spaces of functions and Euclidian vectors, we define them separately, as we need to consider different order relations $\leq$ on each space. We start by considering Euclidian spaces:

Definition 3.4 (Bracketing number of Euclidian space). Let $\mathcal{Y}$ be a subset of the Euclidian metric space $(\mathcal{X}, d)$, where $\mathcal{X} \subseteq \mathbb{R}^k$. We define the order relation $\leq$ on $\mathcal{X}$ by $\mathbf{a} \leq \mathbf{b}$ if the two vectors $\mathbf{a}, \mathbf{b} \in \mathcal{X}$ satisfy

$$a_j \leq b_j, \text{ for } j = 1, 2, \dots, k.$$

The *bracket* $[\mathbf{a}, \mathbf{b}]$ is the set consisting of those $\mathbf{x} \in \mathcal{X}$ satisfying $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$. If $d(\mathbf{a}, \mathbf{b}) < \epsilon$, then we say that $[\mathbf{a}, \mathbf{b}]$ is an ϵ -bracket. The ϵ -bracketing number of $\mathcal{Y}$ is then defined as the smallest number of ϵ -brackets needed to cover $\mathcal{Y}$, that is

$$N_{[]}(\epsilon, \mathcal{Y}, d) = \min \{n \in \mathbb{N}: \mathcal{Y} \subseteq \cup_{i=1}^n [\mathbf{a}_i, \mathbf{b}_i] \text{ for some } \mathbf{a}_1, \mathbf{b}_1, \dots, \mathbf{a}_n, \mathbf{b}_n \in \mathcal{X}\}. \quad (3.35)$$

We note that if $\mathcal{X} \subseteq \mathbb{R}$, then the bracket $[a, b]$ for $a < b$ is simply the closed interval from a to b , so our notation is in line with that normally used in mathematics texts in the one-dimensional case.

For functions, bracketing numbers are defined as follows:

Definition 3.5 (Bracketing number of function space). Let $\mathcal{F}$ be a set of functions $f: \mathcal{X} \rightarrow \mathbb{R}$ and let d be a metric on $\mathcal{F}$. We define the order relation $\leq$ on $\mathcal{F}$ by $l \leq u$ if $l(x) \leq u(x)$ for all $x \in \mathcal{X}$. Given two functions $l, u \in \mathcal{F}$ satisfying $l \leq u$, we define the *bracket* $[l, u]$ to be the set of functions $f \in \mathcal{F}$ satisfying $l \leq f \leq u$. We say that $[l, u]$ is an ϵ -bracket if $d(l, u) < \epsilon$. The ϵ -bracketing number of a set $\mathcal{G} \subseteq \mathcal{F}$, denoted $N_{[]}(\epsilon, \mathcal{G}, d)$, is the smallest number of ϵ -brackets needed to cover $\mathcal{G}$,

We now review some important properties of covering and bracketing numbers.

Proposition 3.2 (Union bound). *Let $(\mathcal{X}, d)$ be a metric space and let $\mathcal{Y} \subseteq \mathcal{X}$. Let $\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_k$ be subsets of $\mathcal{X}$ such that $\mathcal{Y} = \cup_{j=1}^k \mathcal{Y}_j$. For any $\epsilon > 0$, the ϵ -covering number of $\mathcal{Y}$ satisfies*

$$N\left(\epsilon, \cup_{j=1}^k \mathcal{Y}_j, d\right) \leq \sum_{j=1}^k N(\epsilon, \mathcal{Y}_j, d). \quad (3.36)$$

If we assume in addition that $\leq$ is an order relation on $\mathcal{X}$, then the following inequality holds for the ϵ -bracketing number:

$$N_{[]}(\epsilon, \cup_{j=1}^k \mathcal{Y}_j, d) \leq \sum_{j=1}^k N_{[]}(\epsilon, \mathcal{Y}_j, d). \quad (3.37)$$

Proof. We only prove (3.36), noting that a similar argument establishes (3.37). Let $C_1, C_2, \dots, C_k$ be ϵ -covers of $\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_k$, respectively, and let n_j denote the number of balls in C_j . We may assume that each cover has been chosen such that the number of balls in each individual cover is minimal. Since $\mathcal{Y}_j \subseteq C_j$ and $\mathcal{Y} = \cup_{j=1}^k \mathcal{Y}_j$, it follows that $\cup_{j=1}^k C_j$ is an ϵ -cover of $\mathcal{Y}$. Thus, we have constructed an ϵ -cover of $\mathcal{Y}$ using $\sum_{j=1}^k n_j$ balls, and by definition $N(\epsilon, \mathcal{Y}, d)$ can be no greater than this number, establishing (3.36). $\square$

The next proposition shows how one can use the bracketing number to bound the covering number for sets of functions $\mathcal{G}$. This result is useful in situations where obtaining an upper bound on the bracketing number is simpler than working with the covering number directly.

Proposition 3.3. *Let $\mathcal{F}$ be a set of functions equipped with a metric d and $\mathcal{G} \subseteq \mathcal{F}$. Suppose that the metric d satisfies $d(f, g) \leq d(l, u)$ whenever $l \leq f \leq u$ and $l \leq g \leq u$. Then for any $\epsilon > 0$,*

$$N(\epsilon, \mathcal{G}, d) \leq N_{[]}(\epsilon, \mathcal{G}, d). \quad (3.38)$$

Remark 3.2. We note that the condition on the metric in Proposition 3.3 can be verified for all the metrics we have covered thus far. To show that it holds for the $\mathbb{L}_r$ metrics, one simply has to note that for functions l, u, f, g the conditions $l \leq f \leq u$ and $l \leq g \leq u$ imply that $|f - g| \leq |l - u|$. Using the monotonicity property of integrals, and the fact that the map $x \mapsto x^r$ for $r \geq 1$ is increasing for nonnegative x , we obtain the result. One may prove that the property holds for the supremum metric and the Hellinger distance in a similar fashion.

Proof of Proposition 3.3. Let $[l_1, u_1], \dots, [l_n, u_n]$ be an ϵ -bracketing of $\mathcal{G}$, where n is chosen to be minimal. We claim that $\mathcal{G} \subseteq \cup_{i=1}^n B_\epsilon(g_i)$ where $g_i = (l_i + u_i)/2$. Let $g \in \mathcal{G}$. Then there is an index i such that $g \in [l_i, u_i]$. Since $l_i \leq g \leq u_i$ and $l_i \leq g_i \leq u_i$, we have that

$$d(g, g_i) \leq d(l_i, u_i) < \epsilon.$$

Thus, $\mathcal{G} \subseteq \cup_{i=1}^n B_\epsilon(g_i)$ as claimed. Hence, the ϵ -covering number of $\mathcal{G}$ can be no greater than n . $\square$

3.2.4 Establishing strong convergence

We now return to our discussion of posterior convergence. In establishing strong convergence, we will make use of a result due to Ghosal and van der Vaart (2017). Theorem 3.2 gives sufficient conditions for strong convergence in terms of ϵ -covering numbers. The additional conditions involve finding a sequence $\{\mathcal{F}_n\}$ of subsets of $\mathcal{F}$

which ‘almost’ makes up the entire support of the prior, in the sense that $P_n(\mathcal{F}_n^c)$ is assigned exponentially small probability. If this sequence in addition is not ‘too large’ as measured by ϵ -covering numbers, then posterior consistency holds. The following result is a variation on Ghosal and van der Vaart (2017, Theorem 6.32).

Theorem 3.2. *Let $\{P_n\}$ be a sequence of prior distributions on $\mathcal{F}$ and let d be either the $\mathbb{L}_1$ or Hellinger metric. Suppose that f_0 is in the KL-support of the prior distribution (3.31). If we can find a sequence of subsets $\{\mathcal{F}_n\}$ of $\mathcal{F}$ and a constant c_1 such that for all sufficiently large n :*

$$N(\epsilon, \mathcal{F}_n, d) < \exp(n\epsilon^2), \quad (3.39)$$

$$P_n(\mathcal{F}_n^c) < \exp(-c_1 n), \quad (3.40)$$

then the posterior based on P_n satisfies $P_n(d(f_0, f) \geq \epsilon \mid \mathbf{x}_{1:n}) \xrightarrow{a.s.} 0$ for all $\epsilon > 0$.

We note that while the above theorem yields consistency with respect to either the Hellinger or the $\mathbb{L}_1$ metric, this is in fact also sufficient to conclude that posterior consistency also holds with respect to the other metric. This is due to the fact that $d_H^2(f, g) \leq d_1(f, g) \leq 2d_H(f, g)$, (cf. Proposition A.3), so one can choose to work with the distance which is most mathematically convenient for the given problem at hand when establishing consistency results via Theorem 3.2.

Although results about consistency of the posterior distribution is of independent interest, we may be more concerned with studying the large-sample behavior of point estimators derived from the Bayesian model. Luckily, posterior consistency can be used to prove consistency results for certain point estimators. For the posterior mean, the following result, which is a simplified version of Theorem 6.8 in Ghosal and van der Vaart (2017), shows that strong consistency of the posterior is sufficient for the posterior mean to converge:

Theorem 3.3. *Suppose the sequence of priors $\{P_n\}$ is consistent at f_0 with respect to either the $\mathbb{L}_1$ or Hellinger metric. Then the posterior mean is a consistent estimator of f_0 with respect to the same metric.*

3.2.5 Posterior convergence rates

While consistency results may yield an affirmative answer to the question ‘Does the posterior distribution concentrate around the true density’, they do not tell us how fast the posterior convergence takes place. The study of *posterior convergence rates* offer more refined results concerning the convergence of the posterior distribution, also yielding a rate of convergence. To measure the rate at which the posterior contracts, one needs a metric d , typically the $\mathbb{L}_1$ or Hellinger to quantify the distance between the true density f_0 and any f in the support of the prior.

Definition 3.6 (Posterior convergence rate). Let $\{P_n\}$ be a sequence of prior distributions. We say that the posterior distribution *converges at rate* $\epsilon_n = o(1)$ at f_0 with respect to the metric d if for every sequence $\{M_n\}$ satisfying $M_n \rightarrow \infty$ as $n \rightarrow \infty$,

$$P_n(f: d(f, f_0) \geq M_n \epsilon_n \mid \mathbf{x}_{1:n}) \xrightarrow{\mathbb{P}} 0. \quad (3.41)$$

In general, we want to find a sequence ϵ_n decaying to zero as quickly as possible, while (3.41) still holds. Since a posterior convergence rate of ϵ_n with respect to a given metric d implies posterior convergence in the sense of Definition 3.2, it will in general

be harder to deduce a posterior convergence rate than to show posterior consistency. To properly describe how posterior convergence rates can be established, we need some new notation. Put $V(f_0, f) = \int_{\mathcal{X}} f_0(x) \log^2\{f_0(x)/f(x)\} dx$. Our main tool for deducing convergence rates is the following, which is a variation on Ghosal and van der Vaart (2001, Theorem 2.1):

Theorem 3.4. *Let $\{P_n\}$ be a sequence of prior distributions on $\mathcal{F}$ and suppose that α_n, β_n are sequences such that $n \min\{\alpha_n^2, \beta_n^2\} \rightarrow 0$ as $n \rightarrow \infty$. Let d denote the Hellinger or $\mathbb{L}_1$ -metric and suppose further that there are constants $c_1, c_2, c_3, c_4 > 0$ and subsets $\{\mathcal{F}_n\}$ of $\mathcal{F}$ such that the following holds:*

$$N(\alpha_n/2, \mathcal{F}_n, d) \leq \exp\left(c_1 n \alpha_n^2\right), \quad (3.42)$$

$$P_n(\mathcal{F}_n^c) \leq c_3 \exp\left(-\{c_2 + 4\}\beta_n^2\right), \quad (3.43)$$

$$P_n\left(f: K(f_0, f) \leq \beta_n^2, V(f_0, f) \leq \beta_n^2\right) \geq c_4 \exp\left(-c_2 n \beta_n^2\right). \quad (3.44)$$

Then there is a constant $M > 0$ such that $P_n(d(f_0, f) \geq M\epsilon_n \mid \mathbf{x}_{1:n}) \xrightarrow{\mathbb{P}} 0$ holds for the sequence $\epsilon_n = \max\{\alpha_n, \beta_n\}$.

If the sequence β_n satisfies $\sum_{n=1}^{\infty} \exp(-En\beta_n^2) < \infty$ for any $E > 0$, the conclusion still holds if condition (3.44) is replaced by

$$P_n\left(f: d_H^2(f_0, f) \|f_0/f\|_{\infty} \leq \beta_n^2\right) \geq c_4 \exp\left(-c_2 n \beta_n^2\right). \quad (3.45)$$

We note that the sense of convergence rate obtained from Theorem 3.4 is stronger than that of Definition 3.6, since we can take $M_n = M$ for all M , and it is not required that this sequence diverges to positive infinity. The conditions of Theorem 3.4 are quite similar to those of Theorem 3.2, with the new entropy condition (3.42) and the prior mass conditions (3.43) replacing (3.39) and (3.40), respectively. The KL-support condition is replaced by the stronger condition of (3.44), ensuring that the prior mass around the true density does not go to zero too quickly.

Posterior convergence rates also have important consequences for the convergence rates of point estimators derived from the posterior distribution. Theorem 8.7 of Ghosal and van der Vaart (2017) provides a method of constructing a point estimator that converges at the same rate as the posterior distribution. An important consequence of this is that the posterior convergence rate over a class of densities $\mathcal{F}$ cannot be lower than the minimax rate for $\mathcal{F}$, making the two directly comparable.

Chapter 4

Regular Histograms

Regular histogram construction is routine when doing exploratory data analysis, partly due to the fact that virtually any default histogram procedure in statistical software builds a histogram based on a regular grid. In this chapter we focus on methods of regular histogram construction that choose the number of bins automatically. An overview of many popular methods for regular histograms is given. We then turn our attention to a Bayesian method for constructing histograms, showing how to carry out posterior inference. We derive big support conditions for the procedure and derive posterior consistency under some conditions on the prior distribution. Posterior convergence rates are also derived under suitable smoothness assumptions on the true density. The chapter concludes with a large simulation study, comparing the risks of several methods for constructing regular histogram.

Should one prefer regular or irregular histograms?

Before we get into the meat of the chapter, some comments on the use of regular as opposed to irregular histogram constructions are in order. Recall that a *regular histogram* procedure is one in which the partitioning intervals are of equal length; see (2.19). *Irregular histogram* procedures, in contrast, refer to any histogram procedure where it is not given in advance that the chosen partition is a regular one. Although the histogram construction in theory easily allows for the specification of irregular grids, much of the literature on histograms, theoretical and applied, has focused on regular histogram construction (Davies et al., 2009). At first glance, this may seem a bit odd, as the advantages of using an irregular grid are rather immediate; adaptive meshes allow for the bin width to adopt locally to the density, ideally selecting smaller bin widths in areas where the density is rapidly changing and choosing larger bin widths in flatter regions. For a given value of k , an optimally chosen partition can thus potentially have much lower bias. This is not without its drawbacks; the search for a good partition itself will incur additional variance in the estimation procedure (Birgé and Rozenholc, 2006). Thus, it is not necessarily clear that an irregular histogram estimator will produce lower risks. Scott (1992, p. 68) calculated the relative reduction in asymptotic risk based on L_2 loss, (2.56), when using the optimal grid compared to a regular one. For the standard normal distribution, the gains are rather moderate, with the asymptotic risk only being reduced by a factor of 0.107, while for the standard lognormal the reduction is much larger at 0.449. We conclude that it will not necessarily be advantageous to use an irregular histogram procedure in all cases. This is somewhat comforting, as regular histogram construction is likely to remain routine in the future.

Although this chapter focuses solely on regular histograms, we will return to the topic of irregular histograms in Chapter 5, offering a comparison of the two approaches.

4.1 Automatic regular histogram procedures

Due to the popularity and computational simplicity of constructing regular histograms, there are many procedures available which select the number k of bins automatically. The justifications for such procedures differ quite a lot, however, the methods considered in this chapter can be broadly categorized into five categories, and we will cover four of these in this section. The last category consists of those procedures that can be motivated on Bayesian grounds. However, as these all fall under the umbrella of mixtures of Dirichlet distributions, we delay the exposition of these procedures to the next section.

4.1.1 Rules for selecting the number of bins

The following section offers a brief summary of the various rules that have been proposed in the statistical literature. We have by no means attempted to create an exhaustive list over all regular histogram methods, but see for instance Birgé and Rozenholc (2006) for a description of some procedures that were not included here. We have decided to focus on procedures that are popular in use or that have performed well in previous simulation studies, such as those of Birgé and Rozenholc (2006) and Davies et al. (2009).

4.1.1a Rules of thumb

The first category of bin selection rules is one that we term ‘rules of thumb’. These rules often lack theoretical justification, and are rather chosen on the basis that they are particularly simple to understand. Popular examples include Sturges’ rule $k = \lceil \log_2(n) \rceil + 1$, (Sturges, 1926) or the aptly named square root rule $k = \sqrt{n}$ (Nuzzo, 2019). As these rules do not aim at any particular optimality property and were not chosen based on particularly good performance in simulations, one should not come to expect good behavior with respect to any particular loss function when using these methods. Despite this, the methods remain popular, with both of the aforementioned procedures being available in most statistical software. Sturges rule is notably the default procedure used by R’s `hist` function.

4.1.1b Plug-in rules

Plug-in rules are derived based on minimizing an asymptotic expansion of the risk under suitable smoothness assumptions on the true underlying density f_0 ; see Example 2.9 for such an approach when using integrated square error as the loss function. This approach generally yields an asymptotically optimal bin width depending on a functional $C(\cdot)$ of the unknown density f_0 , so it cannot be applied directly in practice. To circumvent this issue, a commonly used approach is to rather estimate the unknown functional by assuming that the density belongs to a given parametric family of distributions. For densities supported on the real line, Scott (1979) suggests using the normal density as a reference, leading to the following bin width selection rule $\hat{h} = 3.49sn^{-1/3}$, where s denotes the sample standard deviation. This plug-in approach is commonly known as Scott’s rule in the literature. Another popular plug-in rule is that of Freedman and Diaconis (1981), which suggests to take $\hat{h} = 2\text{IQR}(\mathbf{x}_{1:n})n^{-1/3}$, where $\text{IQR}(\mathbf{x}_{1:n})$ is the interquartile range of the sample, also yielding the asymptotically optimal rate of

convergence. Such plug-in rules can also be derived with respect to other loss functions, but these are far less commonly used than Scott's rule. If ℓ is taken to be the Hellinger or $\mathbb{L}_r$ loss for $r \geq 1$, the resulting asymptotic minimizer of the risk R_n is of the form $h = C(f_0)n^{-1/3}$ under suitable smoothness assumptions on f_0 , where $C(\cdot)$ is a functional depending on the chosen loss. We note that this implies that histograms based on Scott's rule also converges at the optimal rate for these loss functions as well.

A somewhat more sophisticated plug-in style method of estimating the functional $C(f_0)$ under $\mathbb{L}_2$ loss is that of Wand (1997). Wand's procedure is based on using a kernel-type estimator to estimate the unknown functional $C(f_0)$ by $\hat{C}(h)$, where h is a bandwidth parameter. The kernel-estimate also requires the specification of the pilot bandwidth $\hat{h}_0$, which is chosen using a simple rule like that of Scott. This procedure can be repeated multiple times to produce a sequence of bandwidths $\hat{h}_0, \dots, \hat{h}_m$, where the final chosen bin width $\hat{h}_m$ is then taken to be the corresponding asymptotically optimal bin width when $\hat{C}(\hat{h}_{m-1})$ is used in place of $C(f_0)$.

4.1.1c Methods based on cross-validation

Instead of using asymptotic expansions of the risk to choose the bin width, cross-validation-based procedures estimate the risk or loss function directly, and the resulting bin width is taken to be the minimizer of this estimate. The case of $\mathbb{L}_2$ loss was covered in Example 2.8, leading to maximization of (2.40). Another cross-validation approach is one based on minimizing an estimated Kullback–Leibler loss (Hall, 1990), which corresponds to maximizing

$$CV_n(k) = \sum_{j=1}^k N_j \log(N_j - 1) + n \log(k). \quad (4.1)$$

We note that this expression is only defined in the cases where $N_j \geq 2$, and thus we define $CV_n(k) = -\infty$ whenever $N_j \leq 1$ for at least one j . It should be noted that while this approach is not based on any asymptotic expansions of the loss or risk, cross-validation can be shown to have good asymptotic properties under suitable regularity conditions; see Hall (1990) for Kullback–Leibler loss and Stone (1984) for $\mathbb{L}_2$ loss.

4.1.1d Penalized likelihood methods

The penalized likelihood approaches to regular histogram estimation are all based on the fact that the histogram is a maximum likelihood estimator; cf. Example 2.1. These estimators seek to maximize the penalized log-likelihood,

$$\log L_n(\hat{\mathbf{p}} | k) - \text{pen}_n(k), \quad (4.2)$$

for a suitable increasing penalty function $\text{pen}_n(\cdot)$. In particular, the popular model selection criteria AIC (Hall, 1990) and BIC (Davies et al., 2009) have been used for this purpose. Birgé and Rozenholc (2006) suggest a penalized likelihood procedure based on minimizing an upper bound on the Hellinger risk. This results in a penalty of the form

$$\text{pen}_n(k) = k - 1 + \{\log(n)\}^{2.5}. \quad (4.3)$$

Other criteria that either fit the penalized likelihood formulation (4.2) or a criterion close to it are those based on the information-theoretic criteria of Rissanen (1983). In summary, these criteria seek to find a simpler encoding of the sample $\mathbf{x}_{1:n}$ using a string of the shortest length. Based on these ideas, Kontkanen and Myllymäki (2007)

proposes the selection of the histogram based on the Normalized Maximum Likelihood (NML). However, their criterion includes a penalty term whose computation has a time complexity that is linear in n , making the method computationally expensive compared to other penalized likelihood methods. To reduce the computational complexity of the NML criterion, we will instead consider an asymptotic expansion of the penalty which can be computed in constant time. The following approximation is due to Szpankowski (1998), and has an error of order $\mathcal{O}(n^{-3/2})$ (Kontkanen, 2009):

$$\begin{aligned} \text{pen}_n(k) = & \frac{k-1}{2} \log(n/2) + \log \frac{\sqrt{\pi}}{\Gamma(k/2)} + n^{-1/2} \frac{\sqrt{2k}\Gamma(k/2)}{3\Gamma(k/2 - 1/2)} \\ & + n^{-1} \left(\frac{3 + k(k-2)(2k+1)}{36} - \frac{\Gamma(k/2)^2 k^2}{9\Gamma(k/2 - 1/2)^2} \right). \end{aligned} \quad (4.4)$$

The penalized regular histogram procedure using the penalty (4.4) was considered in Davies et al. (2009). Hall and Hannan (1988) formulate a criterion based on the Minimum Description Length principle, which leads to maximization of

$$\text{MDL}_n(k) = \sum_{j=1}^k (N_j - 1/2) \log(N_j - 1/2) - n(n - k/2) \log(n - k/2) \log(k) - k/2 \log(n), \quad (4.5)$$

where $\text{MDL}_n(k) = -\infty$ if $N_j = 0$ for at least one j .

4.1.2 Choice of loss function for histogram estimation

The choice of loss function when estimating a histogram is an important one. Not only is the performance of a given method of histogram procedure typically judged based on estimates of the risk on an appropriate set of test densities, but the loss function may itself be used to motivate the histogram estimator. Much of the early work on histograms was based on $\mathbb{L}_2$ loss (Scott, 1979; Freedman and Diaconis, 1981). The reason is in no doubt of a partly mathematical nature, as it is particularly amenable to mathematical analyses, and it can be nicely decomposed into integrated bias and integrated variance (A.7).

Since the histogram is a maximum likelihood estimator, some authors have studied the use Kullback–Leibler loss (Hall, 1990). As argued in Birgé and Rozenholc (2006), the use of the KL divergence as a loss function in a simulation study is somewhat problematic when working with histogram estimators, as $K(f_0, \hat{f}_n) = \infty$ whenever a maximum likelihood histogram has an empty bin that overlaps with a set of positive probability under f_0 .

More recent work has focused on the $\mathbb{L}_1$ and Hellinger metrics as loss functions for histogram estimation (Devroye and Györfi, 1985; Kanazawa, 1993; Birgé and Rozenholc, 2006). These two losses are arguably more natural than the $\mathbb{L}_2$ loss in the context of density estimation, due to the fact that they are defined for any pair of densities, while the $\mathbb{L}_2$ distance between the true density and a histogram is only defined if f_0 is square integrable. See also Devroye and Györfi (1985) and Kanazawa (1993) for further arguments in favor of $\mathbb{L}_1$ and Hellinger loss, respectively. Ultimately, there is no clear best choice in terms of loss function, and as such the chosen loss function will in some respects be the result of the modelers preference. As for our study, we will mostly focus on performance with respect to the Hellinger loss. We also include results for the $\mathbb{L}_2$ distance when applicable.

Although integral-based loss functions have been widely used in past work on histograms, their use is not without criticism. Compared to more modern methods of density estimation, like those based on kernels or cubic splines, the histogram is asymptotically less efficient if the true density is sufficiently smooth. Thus, some statisticians argue that histograms serve mostly as an exploratory visualization tool useful for identifying key features in a dataset, and as such, classical loss functions are inappropriate for judging the performance of a given histogram procedure, as one would rather choose another class of estimators if the primary concern is to estimate a density well with respect to a classical loss function (Denby and Mallows, 2009). This attitude can also be found in the works of Davies and Kovac (2004) and Li et al. (2020), where the primary concern is to correctly infer key features of a random sample rather than attaining low risk with respect to classical loss functions. To empirically judge the performance of histograms, it is thus desirable to also have a loss function that rewards a procedure for capturing important features of a density based on a random sample. The primary concern of the peak identification loss (pid loss) of Davies et al. (2009) is to locate modes in the true density. The pid loss rewards those histogram procedures that are able to correctly identify modes of the true density, while putting a penalty on the number of false modes. In order to tell whether a histogram has been able to correctly identify a mode of the true density f_0 , we define a mode/peak of a density in the following manner:

Definition 4.1 (Mode). We say that a density f has a *finite mode* or *peak* at t if there is an interval I such that t is the midpoint of I and $f(s) = c$ for all $s \in I$, and there exists a $\delta > 0$ such that $f(s) < c$ for all $s \in I_\delta \setminus I$, where $I_\delta = \cup_{s \in I} \{x: |s - x| \leq \delta\}$. The density f is said to have an *infinite peak* or *mode* at t if either the left- or right limit at t equals ∞ .

Remark 4.1. Definition 4.1 is a bit more cumbersome to work with than the more commonly adopted definition of a mode, which states that t is a mode of f if it is a local maximum. This standard definition is not well suited for histograms due to their piecewise constant nature; if $t \in I_{k,j}$ is a mode according to this definition, then so is any $s \in I_{k,j}$ since $f(s) = f(t)$ for a density f which is piecewise constant on $I_{k,j}$. Definition 4.1 on the other hand, says that a histogram has a mode if the density in a given bin is greater than the density in the two neighboring bins. In this case, the mode of the density is placed at the midpoint of the modal bin. In cases where the local maximum t satisfies $f(s) < f(t)$ for all $s \neq t$ in some neighborhood of t , our definition coincides with the classical one. This case is rather common among distributions used in statistical models, such as the Normal or Laplace densities.

An illustration of definition Definition 4.1 is shown in Figure 4.1.

Definition 4.2 (Peak identification loss). Suppose that f_0 has l_0 modes $t_1, t_2, \dots, t_{l_0}$ and that the histogram estimate $\hat{f}_n$ has modes at $s_1, \dots, s_l$. Given a tolerance vector $\boldsymbol{\delta} \in \mathbb{R}_+^{l_0}$, we say that a peak of the histogram $\hat{f}_n$ *matches* the peak at t_i of f_0 at tolerance δ_j if $\min_{1 \leq i \leq l} |t_j - s_i| < \delta_j$. The peak s_i is said to be *spurious* if it does not match any peak of f_0 . Letting $C = \sum_{j=1}^{l_0} \mathcal{I}(\min_i |s_i - t_j| < \delta_j)$ denote the number of correctly identified modes, we define the *peak identification loss* as

$$\ell(f_0, \hat{f}_n) = (l_0 - C) + (l - C). \quad (4.6)$$

The choice of the tolerance vector δ_j specifies how close the mode of the histogram estimator should be to t_j before they are said to be matching. In particular, the tolerance

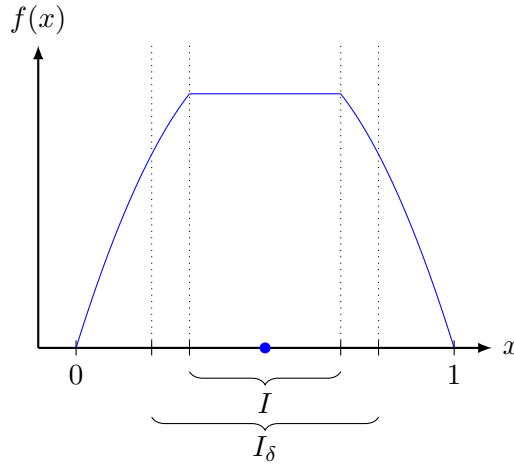


Figure 4.1: Illustration of Definition 4.1. The location of the single mode of f is shown with a blue dot. Note that $f(x)$ is constant on I , while $f(s) < f(x)$ for $s \in I_\delta \setminus I$.

should be chosen so that a given mode in the estimator $\hat{f}_n$ is at most able to match one mode of the true density at a time.

We think both the approach based on classical loss functions as well as the one based on the PID loss are valuable, and good performance with respect to both are as such desired. More classical loss functions measure whether the procedure is able to capture the shape of the true density well, but fails to take into account that certain features of the true density may be of greater interest when doing exploratory data analysis. As such, we think that the two loss functions nicely complement each other, and are both yield good measures of the quality of a given histogram approximation. Example 4.1 illustrates the use of the two loss functions on simulated data.

Example 4.1. We simulated one sample from a bimodal symmetric mixture of normal distributions of size $n = 5000$, and used both the AIC and the BIC to choose the number k of bins. For both methods, we computed the hellinger (2.28) and peak identification losses (4.6) for both of the methods. For the tolerance vector δ , we used a tolerance of 0.02 times the sample range for both peaks. The resulting histograms are shown in Figure 4.2. The locations of the modes of the histogram densities are shown with dots, with the correctly identified modes marked in green/red and the spurious modes marked in black. If a mode of the histograms is located within either of the shaded gray regions, then the corresponding peak is counted as being correctly identified according to the PID loss. From the figure, we see that the AIC-based histogram manages to correctly identify the both peaks, but it also produces 2 spurious modes, leading to a PID loss of 2. The BIC on the other hand performs better and is able to identify both peaks of the density, while producing no spurious modes, leading to a perfect PID loss of 0. In terms of Hellinger loss the AIC outperforms the BIC by a significant margin, with the BIC based histogram producing a loss 17.2% greater. ▲

4.2 A Bayesian approach to histograms

In the present Section we expand on the Bayesian histogram procedure introduced in Example 3.5, focusing on computational aspects. In particular, we show how to use the posterior distribution to compute a histogram density estimator and provide

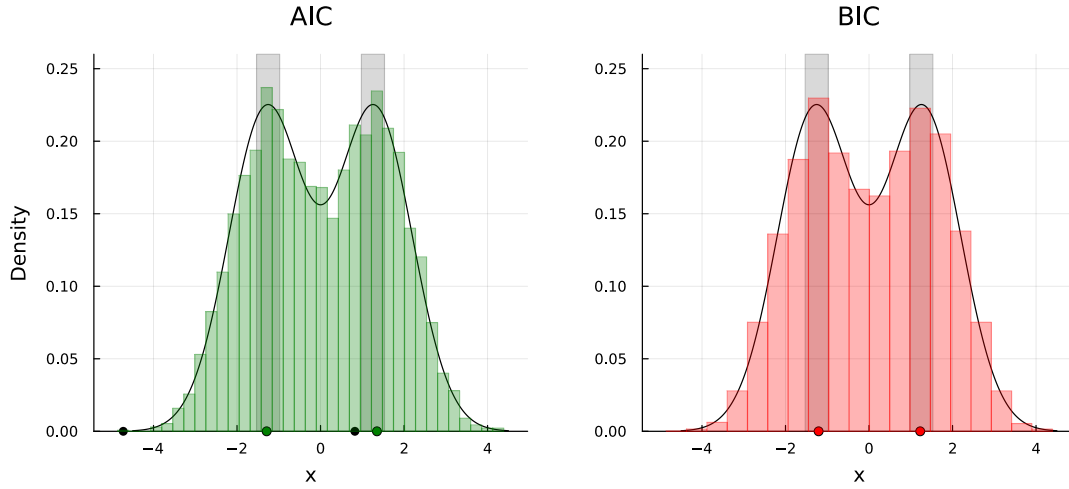


Figure 4.2: Illustration of the PID loss: Histograms with number of bins chosen by AIC (left) and BIC (right).

a simulation-based approach to computing credible intervals. We note that Bayesian approaches to histogram estimation are not new, dating back to Leonard (1973) who, to our knowledge, was the first to suggest a Bayesian histogram procedure. However, Leonard’s procedure requires the number k of bins to be selected in advance, while our goal is to offer a procedure where k is selected automatically based on the sample. Prior work on Bayesian histogram models with automatic bin selection has for the most part focused on prior distributions which are mixtures of Dirichlet distributions of varying dimension, the details of which are given in Example 4.2. The first paper to present a Bayesian argument for selecting the number of bins in a histogram was that of Hall and Hannan (1988), where the prior distributions were chosen to be uniform; a special case of the Dirichlet mixture approach. Other work in this area has attacked the problem of Bayesian histogram estimation from a more computationally oriented perspective (Knuth, 2019), but also through a theoretical lens (Barron et al., 1999; Scricciolo, 2007). Although all the aforementioned works have studied the case where the true density is supported on a compact interval, we note that there are some Bayesian histogram-like approaches that can be used as priors for densities supported on $\mathbb{R}$. Such an approach is that of Gasparini (1996), who constructs a prior on piecewise constant functions on the full real line, also based on the Dirichlet distribution. This approach differs somewhat from the others, as a continuous prior is placed on the interval width h rather than a discrete prior on k .

4.2.1 Computation in the random histogram model

In the regular histogram model the choice of the partition reduces to choosing the number k of bins, and as such, any automatic histogram procedure must provide a rule for selecting k . In the Bayesian setting, we treat the problem of finding k as a Bayesian model selection problem, as in Section 3.1.2. Specifically, we will construct the histogram based on maximizing the posterior probability of the number k of bins given the observed sample $\mathbf{x}_{1:n}$,

$$\hat{k} = \arg \max_k p_n(k | \mathbf{x}_{1:n}). \quad (4.7)$$

A histogram can then be constructed based on the MAP model $\hat{k}$ by using $\hat{k}$ bins and estimating the cell probabilities p_j through an appropriate procedure. We note that the construction offered below can be readily extended to cover non-regular histograms, provided that the bins are constructed in a given way for each k .

Recall that the random histogram model of Example 3.5 could be described by the hierarchical scheme

$$\begin{aligned} k &\sim p_n(k), \\ \mathbf{p} | k &\sim p(\mathbf{p} | k), \\ f(x | k, \mathbf{p}) &= k \prod_{j=1}^k p_j^{\mathcal{I}(x \in I_j)}, \\ x_i | k, \mathbf{p} &\sim f(x_i | k, \mathbf{p}). \end{aligned} \tag{4.8}$$

The probability vector $\mathbf{p}$ is understood to belong to $\mathcal{S}_{k-1}$ for given k . This formulation is slightly more general than that of (3.28), as we now consider the possibility that the prior for k , $p_n(k)$ depends on the sample size n . To carry out posterior inference in the model (3.28), let us first consider the likelihood function conditional on k . Given the sample $\mathbf{x}_{1:n}$, the likelihood function for $\mathbf{p} | k$ is as in (4.8):

$$L_n(\mathbf{p} | k) = p(\mathbf{x}_{1:n} | k, \mathbf{p}) = k^n \prod_{j=1}^k p_j^{N_j}, \tag{4.9}$$

where $N_j = \sum_{i=1}^n \mathcal{I}(x_i \in I_j)$ is the bin count for bin j . By integrating out $\mathbf{p}$ we can compute the marginal likelihood for k ,

$$p(\mathbf{x}_{1:n} | k) = \int_{\mathcal{S}_{k-1}} p(\mathbf{x}_{1:n}, \mathbf{p} | k) d\mathbf{p} = \int_{\mathcal{S}_{k-1}} p(\mathbf{p} | k, \mathbf{x}_{1:n}) p(\mathbf{x}_{1:n} | k, \mathbf{p}) d\mathbf{p}. \tag{4.10}$$

Once the marginal likelihood for k has been worked out, we can compute the posterior probability mass function for k up to proportionality by Bayes' rule (3.17):

$$p_n(k | \mathbf{x}_{1:n}) \propto p_n(k) p(\mathbf{x}_{1:n} | k). \tag{4.11}$$

By computing the expression (4.11) for k in the support of $p_n(k)$, we can find the maximizer of the above expression and the resulting MAP model $\hat{k}$. After computing the MAP for k we can characterize the posterior distribution of the resulting histogram $f(\cdot | \hat{k}, \mathbf{p})$ through the posterior distribution of the probability vector $\mathbf{p}$. Working with (4.9) it becomes possible to compute this posterior of $\mathbf{p}$ for given values of k through Bayes rule (3.2):

$$p(\mathbf{p} | k, \mathbf{x}_{1:n}) \propto p(\mathbf{p} | k) p(\mathbf{x}_{1:n} | k, \mathbf{p}). \tag{4.12}$$

A key use case for histograms is to visualize a given dataset, and as such, it is of great interest to provide point estimates of $f(x)$ given the observed data $\mathbf{x}_{1:n}$. Let $\boldsymbol{\delta} = \boldsymbol{\delta}(\mathbf{x}_{1:n})$ denote an estimator of $\mathbf{p}$. Using the decision-theoretic framework of Section 3.1.3, the histogram point estimator $f(x | k, \hat{\mathbf{p}}_B)$ emerges so soon as we have specified the prior distribution, likelihood and the loss function. We will here motivate our histogram estimator based on squared $\mathbb{L}_2$ loss. Recall from (2.37) that

$$\|f - f(\cdot | k, \boldsymbol{\delta})\|_2^2 = \|f - f(\cdot | k, \mathbf{p})\|_2^2 + k \|\mathbf{p} - \boldsymbol{\delta}\|_2^2, \tag{4.13}$$

where $\mathbf{p}$ are the bin probabilities. As such, it suffices to minimize the second term of (4.13). From Example 3.4, we recall that the minimizer of $\mathbb{E}_{p_j}\{\|\mathbf{p} - \boldsymbol{\delta}\|_2^2 | \mathbf{N}\}$ is

the posterior mean, leading to $\hat{p}_{B,j} = \mathbb{E}_{p_j}\{p_j | \mathbf{x}_{1:n}, k\}$. Plugging in the estimated probabilities $\hat{p}_{B,j}$ and the MAP $\hat{k}$, we arrive at our point estimate

$$\hat{f}_{B,n}(x) = f(x | \hat{k}, \hat{\mathbf{p}}_B) = \prod_{j=1}^{\hat{k}} \left(\hat{k} \hat{p}_{B,j} \right)^{\mathcal{I}(x \in I_j)}. \quad (4.14)$$

We will refer to the estimator in (4.14) as the *Bayes histogram estimator*.

Remark 4.2. Although we used the squared $\mathbb{L}_2$ distance as our loss function to motivate our estimator in this case, it is possible to construct point estimators based on other loss functions as well. Our main motivation for using the squared $\mathbb{L}_2$ loss here is that the resulting point estimates $\hat{\mathbf{p}}_B$ are aiming for the cell probabilities $\boldsymbol{\pi}$. For example, the piecewise constant density that minimizes the Hellinger distance to f_0 is generally not equal to $f_{0,k}$ (Birgé and Rozenholc, 2006). Since the area under a given bin in a histogram is more or less always understood to be an estimate of the probability of an observation landing in that bin, we feel that using a point estimator of f_0 whose corresponding bin probability estimates $\hat{\mathbf{p}}_B$ are not aiming for $\boldsymbol{\pi}$ would be somewhat misleading, and as such, we prefer to motivate our estimator using $\mathbb{L}_2$ distance rather than other loss functions.

Example 4.2. A computationally tractable random histogram model emerges by putting Dirichlet prior distributions on the cell probabilities. For each k , we take $\mathbf{p} | k \sim \text{Dir}(a\mathbf{p}_0)$. It turns out that this choice of prior distribution satisfies a certain conjugacy property. From (4.9), the likelihood function for $\mathbf{p}$ is given by $p(\mathbf{x}_{1:n} | k, \mathbf{p}) \propto \prod_{j=1}^k p_j^{N_j}$. Using (4.12), we find that the conditional posterior $\mathbf{p} | k, \mathbf{x}_{1:n}$ is given by

$$\begin{aligned} p(\mathbf{p} | k, \mathbf{x}_{1:n}) &\propto p(\mathbf{p} | k) p(\mathbf{x}_{1:n} | k, \mathbf{p}) \\ &\propto \prod_{j=1}^k p_j^{ap_{0,j}-1} \prod_{j=1}^k p_j^{N_j} \\ &= \prod_{j=1}^k p_j^{ap_{0,j}+N_j-1} \\ &\propto \text{Dir}(\mathbf{p} | a\mathbf{p}_0 + \mathbf{N}), \end{aligned} \quad (4.15)$$

mirroring the computations of the Dirichlet–Multinomial situation in Example 3.1. Thus, we conclude that the posterior distribution of $\mathbf{p}$ for given values of k is also Dirichlet. Due to this conjugacy property, we can bypass (4.10) when computing the marginal likelihood. By rearranging Bayes rule (3.1), we can write

$$p(\mathbf{x}_{1:n} | k) p(\mathbf{p} | \mathbf{x}_{1:n}, k) = p(\mathbf{p} | k) p(\mathbf{x}_{1:n} | k, \mathbf{p}). \quad (4.16)$$

This can then be used to compute the posterior for k up to proportionality:

$$p_n(k | \mathbf{x}_{1:n}) \propto p_n(k) p(\mathbf{x}_{1:n} | k) = p_n(k) \frac{p(\mathbf{p} | k) p(\mathbf{x}_{1:n} | k, \mathbf{p})}{p(\mathbf{p} | \mathbf{x}_{1:n}, k)}. \quad (4.17)$$

Since the prior and posterior for $\mathbf{p} | k$ is Dirichlet, we have

$$\begin{aligned} p(\mathbf{p} | k) p(\mathbf{x}_{1:n} | k, \mathbf{p}) &= k^n \frac{\Gamma(a)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \prod_{j=1}^k p_j^{ap_{0,j}+N_j-1}, \\ p(\mathbf{p} | \mathbf{x}_{1:n}, k) &= \frac{\Gamma(a+n)}{\prod_{j=1}^k \Gamma(ap_{0,j}+N_j)} \prod_{j=1}^k p_j^{ap_{0,j}+N_j-1}. \end{aligned}$$

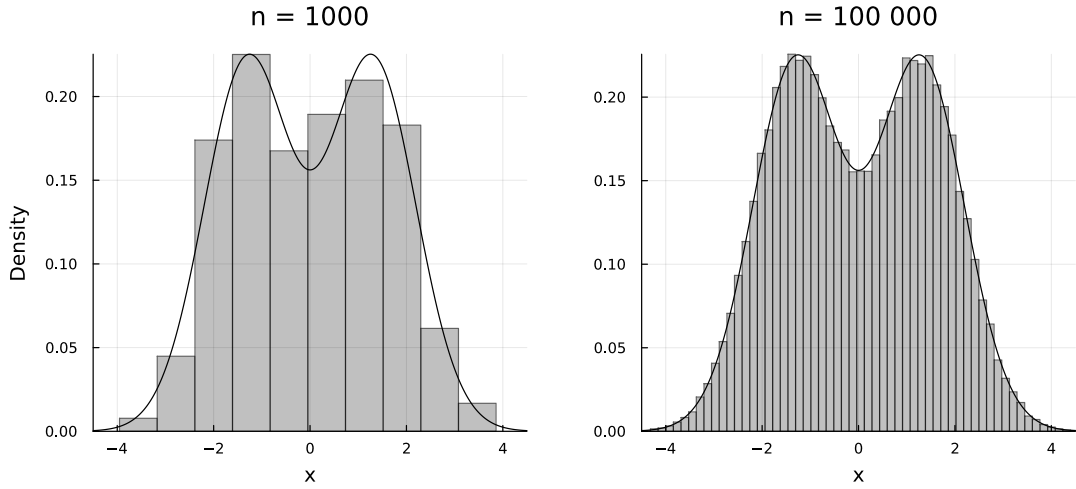


Figure 4.3: Random histogram model fit to two random samples of size $n = 1000$ (left) and $n = 100\,000$ (right) from a symmetric mixture of normals.

Plugging this fact into (4.17) we obtain

$$p_n(k | \mathbf{x}_{1:n}) \propto p_n(k) k^n \frac{\prod_{j=1}^k \Gamma(ap_{0,j} + N_j)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \frac{\Gamma(a)}{\Gamma(a+n)}. \quad (4.18)$$

Using (4.18) we can easily compute the MAP $\hat{k}$ by computing the expression on the right hand side for each value of $1 \leq k \leq k_n$, and then taking $\hat{k}$ to be the k for which this quantity was maximized. Under the $\mathbb{L}_2$ loss the Bayes estimator of the bin probabilities are given by the posterior means, so that

$$\hat{p}_{B,j} = \mathbb{E}_{p_j}\{p_j | \mathbf{x}_{1:n}, k\} = \frac{a}{a+n} p_{0,j} + \frac{n}{a+n} N_j/n, \quad (4.19)$$

which is the familiar convex combination of prior mean and the MLE of Example 3.4.

To illustrate the above inferential pipeline in practice, we simulated two iid samples $x_1, \dots, x_{n_j}$ from the normal mixture of Figure 1.1, where $n_1 = 1000$ and $n_2 = 100\,000$. For each of the samples, we used a random histogram prior with the prior for k being the uniform distribution on $\{1, 2, \dots, \lceil 4n_m/\log^2(n_m) \rceil\}$. The resulting histograms can be found in Figure 4.3.

For the $n_2 = 100\,000$ sample, the histogram approximation of the true density is better. As one might expect, the posterior distribution selects a larger number of bins for the second sample. This seems quite reasonable, as successful nonparametric procedures typically will tend to use a larger number of parameters as the sample size increases, leading to more accurate approximations of the true density. $\blacktriangle$

Remark 4.3. From a purely Bayesian point of view, the estimator (4.14) is in some sense not the natural point estimate of the density as it does not minimize the $\mathbb{L}_2$ posterior expected loss. Instead, the optimal estimator from this decision-theoretic point of view is the model-averaged estimator

$$\mathbb{E}_f\{f(x) | \mathbf{x}_{1:n}\} = \sum_{k=1}^{\infty} p_n(k | \mathbf{x}_{1:n}) f(x | k, \hat{\mathbf{p}}_B). \quad (4.20)$$

In particular, when the posterior pmf for k assigns high mass to multiple values of k , the estimator (4.20) can look quite different to the Bayes histogram estimator (4.14).

We note that this estimator is not a regular histogram, but rather a convex combination thereof, and thus its form is quite a bit more complicated than that of (4.14). In the machine learning community, model averaging estimators such as (4.20) are known as *ensemble estimators*, which often have better predictive performance than estimators based on model selection, but these methods are generally more difficult to interpret. Indeed, the main appeal of the Bayes histogram estimator is precisely its interpretability and parsimony. The probability mass content of a single bin I_j provides a direct estimate of the probability that $x \in I_j$, and, while the true Bayes estimator (4.20) is infinite- or high-dimensional, the Bayes histogram estimator only uses $\hat{k}$ parameters. Breiman (2001) discusses this issue at length, and termed the conflict between interpretability and predictive accuracy ‘The Occam dilemma’ in reference to *Occam’s razor*, a philosophical device often invoked in favor of choosing simpler models over more complex ones (Domingos, 1999).

When the prior distribution of $\mathbf{p} \mid k$ is Dirichlet and the prior for k has finite support, it becomes possible to compute (4.20) since we have an explicit expression for $\hat{\mathbf{p}}_B$. Since the distributions of $k \mid \mathbf{x}_{1:n}$ and $\mathbf{p} \mid \mathbf{x}_{1:n}, k$ are easy to sample from in this case, we are able to easily estimate posterior quantities of interest via simulations by first generating $k^{(t)} \sim p(k \mid \mathbf{x}_{1:n})$ and then sampling $\mathbf{p}^{(t)} \sim p(\mathbf{p} \mid \mathbf{x}_{1:n}, k^{(t)})$. The tuple $(k^{(t)}, \mathbf{p}^{(t)})$ defines a sample $f^{(t)}$ from the posterior distribution of f , which can for instance be used to compute credible intervals (3.13) for $f(x)$ via sample quantiles.

While the remaining parts of this section focus mostly on the histogram estimator based on $\hat{k}$, we return to the Bayes estimator in Section 6.6.

4.3 Prior support under random histogram priors

In this section we show that the regular random histogram prior distribution has the KL-property (3.31) at any bounded, a.e. continuous f_0 .

4.3.1 Assumptions and notation

Before we proceed with the proof, we introduce some new notation. Recall that for a function $h: [0, 1] \rightarrow \mathbb{R}$, we denote the piecewise constant version of h based on a given partition $I_{1:k} = (I_{k,1}, \dots, I_{k,k})$ of the unit interval by $h_{I_{1:k}}$ (2.11). In the following, we will mostly work with a slightly different parametrization of $h_{I_{1:k}}$,

$$h_{I_{1:k}}(x) = \prod_{j=1}^k (\rho_j / |I_{k,j}|)^{\mathcal{I}(x \in I_{k,j})}, \quad (4.21)$$

where $\rho_j = \int_{I_{k,j}} h(s) ds$. Recall that the Kullback–Leibler divergence between two densities f, g on $[0, 1]$ is given by

$$K(f, g) = \int_0^1 f(x) \log \frac{f(x)}{g(x)} dx. \quad (4.22)$$

We note that the Kullback–Leibler divergence is well-defined in this setting (although it may be infinite), provided that both f and g are bounded. In particular, the boundedness of f implies that

$$\int_0^1 f(x) |\log f(x)| dx < \infty,$$

see Proposition A.4 for a proof. In the following, we will assume that the sample $\mathbf{x}_{1:n} = x_1, \dots, x_n$ is iid from a given probability density f_0 on $[0, 1]$. Recall that the regular histogram prior is given by the hierarchical scheme in (4.8), and hence it is determined by the sequence of prior distributions $k \sim p_n(k)$ and the k -simplex distribution $\mathbf{p} | k \sim p(\mathbf{p} | k)$. Given a sample of size $n \in \mathbb{N}$, the prior for f , $P_n(\cdot)$, has the representation

$$P_n(f \in \mathcal{A}) = \sum_{k=1}^{\infty} p_n(k) P(f_k \in \mathcal{A} | k), \quad (4.23)$$

where $\mathcal{A}$ is a measurable set and $P(f \in \mathcal{A} | k) = \int_{\mathcal{S}_{k-1}} \mathcal{I}(f(\cdot | k, \mathbf{p}) \in \mathcal{A}) p(\mathbf{p} | k) d\mathbf{p}$ for $f(\cdot | k, \mathbf{p})$ given as in (4.8). We denote the sequence of regular random histogram priors by $\{P_n\}$, where P_n is the probability measure on the space of densities induced by $p_n(k)$ and $p(\mathbf{p} | k)$ by (4.23).

In the interest of being as general as possible, in the following we shall also allow the prior for k , $p_n(k)$ to have full support in $k \in \mathbb{N}$ for all $n \in \mathbb{N}$, although this is not the method we have implemented in practice in Section 4.2. To establish that $f_0(\cdot)$ lies in the support of the prior, we require the following conditions to be met by the sequence of priors $\{P_n\}$:

Assumption 4.1. We assume the following:

- (i) The sequence of prior probability mass functions for k , $p_n(k)$, satisfies

$$\liminf_{n \rightarrow \infty} p_n(k) > 0, \quad \forall k \in \mathbb{N}. \quad (4.24)$$

- (ii) For given k , $\mathbf{p}$ has the prior probability density $p(\mathbf{p} | k) = \text{Dir}(\mathbf{p} | \mathbf{a}_k)$, where $\mathbf{a}_k$ satisfies $a_{k,j} > 0$ for all k and j .

Unless the true density is itself a regular histogram density, there will be some error in the piecewise constant approximation $f_{0,I_{1:k}}$ of f_0 for any finite $k \in \mathbb{N}$. Assumption (i) ensures that large values of k are eventually given sufficiently large prior mass. The second assumption secures that positive mass is assigned to those g close to $f_{0,I_{1:k}}$ by ensuring that the event $\{\mathbf{p} \in \mathcal{S}_{k-1} : \|\mathbf{p} - \boldsymbol{\pi}\|_1 < \delta\}$ has positive probability for all k and arbitrary simplex vectors $\boldsymbol{\pi}$.

4.3.2 The master theorem

Below we state the main theorem of the section, showing that conditions (i) – (ii) are sufficient for the KL-support condition (3.31). Theorem 4.1 is a generalization of the result of section 3.1 in Barron et al. (1999), who derived big support for a random regular histogram model, using symmetric Dirichlet priors. Our result allows the prior for k , $p_n(k)$ to depend on the sample size, providing a generalization of the result of Barron et al. (1999).

Theorem 4.1. *Let $f_0: [0, 1] \rightarrow \mathbb{R}$ be a bounded and a.e. continuous probability density. Let $\{P_n\}$ be a sequence of prior distributions satisfying Assumption 4.1. Then for all $\epsilon > 0$,*

$$\liminf_{n \rightarrow \infty} P_n(f : K(f_0, f) < \epsilon) > 0. \quad (4.25)$$

Theorem 4.1 can be used to immediately verify big support conditions in two important cases.

Corollary 4.1. *Let $p(\cdot)$ be a probability distribution on $\mathbb{N}$ satisfying $p(k) > 0$ for all k , and let P denote the histogram prior based on $p(\cdot)$, required to additionally satisfy assumption (ii). Then for any bounded, a.e. continuous probability density f_0 on $[0, 1]$ and $\epsilon > 0$,*

$$P(f : K(f_0, f) < \epsilon) > 0. \quad (4.26)$$

Moreover, if $\{k_n\}$ is any sequence in $\mathbb{N}$ satisfying $k_n \rightarrow \infty$, the histogram prior based on $p_n(\cdot)$, the truncation of $p(\cdot)$ to the set $\{1, \dots, k_n\}$, satisfies equation (4.25).

Proof of Corollary 4.1. We verify that assumption (i) holds. By taking $p_n = p$ for all n , we have that

$$\liminf_{n \rightarrow \infty} p_n(k) = p(k) > 0,$$

proving (4.24). Since $P = P_n$, the left hand side is independent of n , and we conclude that the classical big support condition (4.26) also holds.

To prove the second part, put $S_n = \sum_{k=1}^{k_n} p(k)$. Then for any k and sufficiently large n , the truncation of p to $\{1, \dots, k_n\}$ satisfies $p_n(k) = p(k)/S_n$. Since $S_n \leq 1$ for all n we then have that

$$\liminf_{n \rightarrow \infty} p_n(k) = \liminf_{n \rightarrow \infty} p(k)/S_n \geq p(k) > 0,$$

showing that assumption (i) holds for this prior sequence. $\square$

4.3.3 Proving the master Theorem

To prove Theorem 4.1 we will make use of a few lemmas. Proposition 4.1 establishes that the piecewise constant version of f_0 yields an arbitrarily good approximation to the density at every continuity point as the chosen grid is made arbitrarily fine. Although the ultimate goal of this section is to show a big support condition for the regular random histogram prior, we give many of the following results for more general partitions since the necessary approximation properties hold for more general grids, of which regular partitions are a special case. We will make use of the results for irregular grids later when we shift our focus toward irregular histograms.

Proposition 4.1. *Let $h(\cdot)$ be a bounded, almost everywhere continuous function supported on $[0, 1]$. Let $\{I_{k,1}, \dots, I_{k,k}\}_{k=1}^\infty$ be a sequence of interval partitions of the unit interval satisfying $\max_j |I_{k,j}| \rightarrow 0$ as $k \rightarrow \infty$. Then $h_{I_{1:k}} \rightarrow h$ almost everywhere as $k \rightarrow \infty$.*

Proof. To show a.e. convergence, fix $\epsilon > 0$ and let $x \in [0, 1]$ be a continuity point of h . By the continuity of h at x , we can find $\delta > 0$ such that $|h(x) - h(s)| < \epsilon$ whenever $|x - s| < \delta$. Note that as $I_{k,1}, \dots, I_{k,k}$ forms a partition of the unit interval, we always have $x \in I_{k,j_k}$ for some $j_k \in \{1, 2, \dots, k\}$ no matter the value of k . Denote $\delta_k = \max_j |I_{k,j}|$. Since $\delta_k \rightarrow 0$, as $k \rightarrow \infty$, we can find $K \in \mathbb{N}$ such that $\delta_k < \delta$ for all $k \geq K$. For any $k \geq K$, we consequently have $|I_{k,j_k}| < \delta$ from which it is clear that every $s \in I_{k,j_k}$ satisfies $|h(x) - h(s)| < \epsilon$. Recalling that $h_{I_{1:k}}(x) = |I_{k,j_k}|^{-1} \int_{I_{k,j_k}} h(s) ds$

for $x \in I_{k,j_k}$, we deduce from this fact that

$$\begin{aligned}
 |h(x) - h_k(x)| &= |h(x) - h_{k,j_k}| \\
 &= \left| |I_{k,j_k}|^{-1} \int_{I_{k,j_k}} \{h(x) - h(s)\} ds \right| \\
 &\leq |I_{k,j_k}|^{-1} \int_{I_{k,j_k}} |h(x) - h(s)| ds \\
 &< |I_{k,j_k}|^{-1} |I_{k,j_k}| \epsilon \\
 &= \epsilon,
 \end{aligned}$$

which proves the claim. $\square$

The following lemma shows that we can approximate f_0 by another density g which is a.e. continuous and bounded from below.

Lemma 4.1. *Let f_0 be a bounded, a.e. continuous probability density on $[0, 1]$. Then for all $\epsilon > 0$, there exists an a.e. continuous, bounded probability density g_ϵ on $[0, 1]$ such that g_ϵ is bounded away from 0 which satisfies $K(f_0, g_\epsilon) < \epsilon$. Moreover, if f_0 is continuous then g_ϵ can be chosen to be continuous as well.*

Proof. Fix $\epsilon > 0$. We first note that it suffices to show that we can find a sequence of probability density functions $\{g_n\}$ where each $g_n(\cdot)$ is bounded away from 0, and satisfying

$$\lim_{n \rightarrow \infty} \int_0^1 f_0(x) \log g_n(x) dx = \int_0^1 f_0(x) \log f_0(x) dx. \quad (4.27)$$

To see why, note that if (4.27) holds, then $K(f_0, g_n) \rightarrow 0$, which implies that $K(f_0, g_n) < \epsilon$ for all $n \geq N_\epsilon$. But then g_{N_ϵ} is bounded away from 0 and satisfies $K(f_0, g_{N_\epsilon}) < \epsilon$, and the statement of the lemma thus holds for $g_\epsilon = g_{N_\epsilon}$.

Denote $M = \sup_{x \in [0,1]} f_0(x)$. Define the family of functions $\{h_\delta\}_{\delta \in \mathbb{R}_+}$ by

$$h_\delta(x) = \begin{cases} f_0(x) & \text{if } f_0(x) \geq \delta \\ \delta & \text{if } f_0(x) < \delta \end{cases}$$

By construction, it is clear that h_δ is a.e. continuous as well. Moreover, for any $x \in [0, 1]$.

$$\lim_{\delta \downarrow 0} h_\delta(x) = f_0(x).$$

Finally, we note that for all $\delta \leq M$ we have that $f_0(x) \leq h_\delta(x) \leq M$ for all x . Integrating these inequalities over $[0, 1]$, we discover that

$$1 \leq \int_0^1 h_\delta(x) dx \leq M.$$

Moreover, an application of the Dominated Convergence Theorem yields

$$\int_0^1 h_\delta(x) dx \rightarrow \int_0^1 f_0(x) dx = 1 \text{ as } \delta \downarrow 0.$$

Let $\{\delta_n\}$ be a sequence of positive real numbers satisfying $\delta_n \downarrow 0$ as $n \rightarrow \infty$. We define a family of probability density functions $\{g_n\}$ by

$$g_n(x) = \frac{h_{\delta_n}(x)}{\int_0^1 h_{\delta_n}(s) ds} \text{ for } x \in [0, 1].$$

The a.e. continuity of g_n is then an immediate consequence of the a.e. continuity of h_{δ_n} .

We will now turn our attention to showing that g_n approximates f_0 in the Kullback–Leibler sense. By taking n sufficiently large, we have that $\delta_n \leq M$. From the above results it follows that

$$\lim_{n \rightarrow \infty} g_n(x) = \frac{f_0(x)}{1} = f_0(x),$$

establishing pointwise convergence on $[0, 1]$. Moreover, we obtain the following bounds on g_n :

$$\frac{1}{M} h_{\delta_n}(x) \leq g_n(x) \leq M.$$

Furthermore, as $f_0(x) \leq h_\delta(x)$ for any $\delta > 0$ by construction, the lower bound further implies that $\frac{1}{M} f_0(x) \leq g_n(x)$. Since $x \mapsto \log x$ is a strictly increasing function, we obtain the following bounds:

$$\begin{aligned} f_0(x) |\log g_n(x)| &\leq \max \left\{ f_0(x) \left| \log \left(\frac{1}{M} f_0(x) \right) \right|, f_0(x) |\log M| \right\} \\ &\leq f_0(x) \left| \log \left(\frac{1}{M} f_0(x) \right) \right| + f_0(x) |\log M|. \end{aligned}$$

Since f_0 is bounded, the integral of the right hand side of this inequality is finite, and its value independent of n , provided n is sufficiently large. Moreover, as $g_n \rightarrow f_0$ pointwise on $[0, 1]$, another application of the Dominated Convergence Theorem establishes (4.27)

For the final claim of the lemma, we observe that if f_0 is continuous, then h_{δ_n} is also continuous for any n . Since g_n is just a rescaled version of h_{δ_n} , we conclude that g_n is continuous for all $n \in \mathbb{N}$, establishing the claim. $\square$

The next lemma shows that if f_0 and g are close in the KL-sense, then so are f_0 and $g_{I_{1:k}}$ provided the partition $I_{1:k}$ is sufficiently fine.

Lemma 4.2. *Let g be a bounded and a.e. continuous probability density on the unit interval, that is bounded away from 0. Let $\{I_{k,1}, \dots, I_{k,k}\}_{k=1}^\infty$ be a sequence of interval partitions such that $\max_j |I_{k,j}| \rightarrow 0$ as $k \rightarrow \infty$. Then the sequence $\{g_{I_{1:k}}\}_{k=1}^\infty$ satisfies*

$$\lim_{k \rightarrow \infty} \int_0^1 f_0(x) \log \frac{g(x)}{g_{I_{1:k}}(x)} dx = 0. \quad (4.28)$$

Moreover, for each $\epsilon > 0$ we can find a $\kappa \in \mathbb{N}$ and a piecewise constant function g_κ on $I_{1:\kappa}$ such that

$$K(f_0, g_\kappa) < \epsilon. \quad (4.29)$$

Proof. To prove that (4.28) holds it suffices to show that the sequence $\{g_{I_{1:k}}\}_{k=1}^\infty$ satisfies

$$\lim_{k \rightarrow \infty} \int_0^1 f_0(x) \log g_{I_{1:k}}(x) dx = \int_0^1 f_0(x) \log g(x) dx. \quad (4.30)$$

Denote $m = \inf_{x \in [0,1]} g(x)$ and $M = \sup_{x \in [0,1]} g(x)$ and fix $k \in \mathbb{N}, x \in [0, 1]$. Let j be such that $x \in I_{k,j}$. Integrating the inequalities $m \leq g(x) \leq M$ over $I_{k,j}$, we find that

$$m |I_{k,j}| \leq \int_{I_{k,j}} g(x) dx \leq M |I_{k,j}|. \quad (4.31)$$

Using the fact that $g_{I_{1:k}}(x) = |I_{k,j}|^{-1} \int_{I_{k,j}} g(s) ds$ for $x \in I_{k,j}$ and multiplying (4.31) by $|I_{k,j}|^{-1}$, we obtain $m \leq g_k(x) \leq M$, from which it follows that

$$\begin{aligned} f_0(x) |\log g_k(x)| &\leq \max \{f_0(x) |\log m|, f_0(x) |\log M|\} \\ &\leq f_0(x) |\log m| + f_0(x) |\log M|, \end{aligned}$$

which is integrable over $[0, 1]$, and does not depend on k . By Proposition 4.1, the sequence $g_{I_{1:k}}$ converges to g almost everywhere on $[0, 1]$, so the conditions for the Dominated Convergence Theorem are met, establishing (4.30).

To prove the second part of the Lemma, we note that by Lemma 4.1, we can find a lower bounded, a.e. continuous density g_ϵ such that $K(f_0, g_\epsilon) < \epsilon/2$. By (4.28), we can find a $\kappa \in \mathbb{N}$ such that $\int_0^1 f_0(x) \log \{g_\epsilon(x)/g_{\epsilon, I_{1:\kappa}}(x)\} dx < \epsilon/2$. Hence, we have that

$$K(f_0, g_{\epsilon, I_{1:\kappa}}) = K(f_0, g_\epsilon) + \int_0^1 f_0(x) \log \frac{g_\epsilon(x)}{g_{\epsilon, I_{1:\kappa}}(x)} dx < \epsilon.$$

Taking $g_\kappa = g_{\epsilon, I_{1:\kappa}}$ established the claim. $\square$

We will now show a useful decomposition of the Kullback–Leibler divergence between any bounded probability density f on $[0, 1]$ and a piecewise constant function on a given partition of the interval. We will reuse this result later on when we investigate consistency properties of the procedure.

Recall from Example 2.1 that for a partition $I_{1:k}$ and $\boldsymbol{\rho} \in \mathcal{S}_{k-1}$, we write $f(\cdot | k, I_{1:k}, \boldsymbol{\rho})$ for the probability density function given by

$$f(x | k, I_{1:k}, \boldsymbol{\rho}) = \sum_{j=1}^k \frac{\rho_j}{|I_j|} \mathcal{I}(x \in I_j), \quad x \in [0, 1]. \quad (4.32)$$

Lemma 4.3. *Let $f_0: [0, 1] \rightarrow \mathbb{R}$ be a bounded density and let $I_{1:k}$ be an interval partition of $[0, 1]$. Then for any $\boldsymbol{\rho} \in \mathcal{S}_{k-1}$, the piecewise constant density $g_k(\cdot) = f(\cdot | k, I_{1:k}, \boldsymbol{\rho})$ satisfies*

$$K(f_0, g_k) = K(f_0, f_{0, I_{1:k}}) + K(f_{0, I_{1:k}}, g_k), \quad (4.33)$$

Proof. Put $\pi_j = \int_{I_j} f_0(x) dx$. By elementary properties of the logarithm, we have that

$$\begin{aligned} K(f_0, g_k) &= \int_0^1 f_0(x) \log \frac{f_0(x)}{f_{0, I_{1:k}}(x)} dx + \int_0^1 f_0(x) \log \frac{f_{0, I_{1:k}}(x)}{g_k(x)} dx \\ &= K(f_0, f_{0, I_{1:k}}) + \sum_{j=1}^k \int_{I_j} f_0(x) \log \frac{f_{0, I_{1:k}}(x)}{g_k(x)} dx \\ &= K(f_0, f_{0, I_{1:k}}) + \sum_{j=1}^k \int_{I_j} f_0(x) \log \frac{\pi_j / |I_j|}{\rho_j / |I_j|} dx \\ &= K(f_0, f_{0, I_{1:k}}) + \sum_{j=1}^k \int_{I_j} \pi_j \log \frac{\pi_j / |I_j|}{\rho_j / |I_j|} dx, \end{aligned}$$

where we used that $g_k(x) = \rho_j / |I_j|$ and $f_{0, I_{1:k}}(x) = \pi_j / |I_j|$ on I_j . Since $\int_{I_j} f_{0, I_{1:k}}(x) dx = \pi_j$, we further have that

$$\begin{aligned} K(f_0, g_k) &= K(f_0, f_{0, I_{1:k}}) + \sum_{j=1}^k \int_{I_j} f_{0, I_{1:k}}(x) \log \frac{f_{0, I_{1:k}}(x)}{g_k(x)} dx \\ &= K(f_0, f_{0, I_{1:k}}) + K(f_{0, I_{1:k}}, g_k), \end{aligned}$$

which proves the claim. $\square$

It follows rather immediately from Lemma 4.3 that the approximation $f_{0,I_{1:k}}$ is closer to f_0 than $g_{I_{1:k}}$ in terms of KL-divergence, since the KL-divergence is nonnegative. Using the above fact, we deduce that $f_{0,I_{1:k}}$ can be used to approximate f_0 in the Kullback–Leibler sense to arbitrary precision, provided k is sufficiently large.

Corollary 4.2. *For any $\epsilon > 0$, if $\{I_{k,1}, \dots, I_{k,k}\}_{k=1}^\infty$ is a sequence of interval partitions such that $\max_j |I_{k,j}| \rightarrow 0$ as $k \rightarrow \infty$ the piecewise constant approximation of f_0 , $f_{0,I_{1:k}}$ satisfies $K(f_0, f_{0,I_{1:k}}) < \epsilon$, for all sufficiently large k .*

We now return to the proof of Theorem 4.1, which relies on Lemma 4.1 and Lemma 4.2. Since Theorem 4.1 concerns the regular random histogram prior, we will in the following proof let $I_{1:k}$ denote the regular partition of the unit interval using k bins.

Proof of Theorem 4.1. Let $\epsilon > 0$ be arbitrary. Since $\max_j |I_{k,j}| = 1/k \rightarrow 0$ as $k \rightarrow \infty$, Lemma 4.2 guarantees that we can find $\kappa \in \mathbb{N}$ and a lower bounded piecewise constant density $g_{I_{1:\kappa}}$, such that

$$\int_0^1 f_0(x) \log \frac{f_0(x)}{g_{I_{1:\kappa}}(x)} dx < \epsilon/2. \quad (4.34)$$

Put $f_\kappa = f(\cdot | \kappa, I_{1:\kappa}, \mathbf{p})$, and let $\rho_j = \int_{I_j} g_\kappa(x) dx$. Since $x \mapsto \log(x)$ is continuous at any $x > 0$, we can find $\delta > 0$, possibly depending on κ , such that $|p_j - \rho_j| < \delta$ for all $j = 1, 2, \dots, \kappa$ implies $|\log p_j - \log \rho_j| < \epsilon/2$ for all such j . An application of the triangle inequality shows that

$$\begin{aligned} & \left| \int_0^1 f_0(x) \log f_\kappa(x) dx - \int_0^1 f_0(x) \log g_{I_{1:\kappa}}(x) dx \right| \\ &= \left| \sum_{j=1}^\kappa \int_{I_{\kappa,j}} f_0(x) (\log p_j - \log \rho_j) dx \right| \\ &\leq \sum_{j=1}^\kappa |\log p_j - \log \rho_j| \int_{I_{\kappa,j}} f_0(x) dx \\ &< \frac{\epsilon}{2} \sum_{j=1}^\kappa \int_{I_{\kappa,j}} f_0(x) dx \\ &= \frac{\epsilon}{2}. \end{aligned}$$

Since $\|\mathbf{p} - \boldsymbol{\rho}\|_\infty \leq \|\mathbf{p} - \boldsymbol{\rho}\|_1$, we have the inclusion

$$\{\mathbf{p} \in \mathcal{S}_{\kappa-1} : \|\mathbf{p} - \boldsymbol{\rho}\|_1 < \delta\} \subseteq \left\{ \mathbf{p} \in \mathcal{S}_{\kappa-1} : \int_0^1 f_0(x) \log \frac{g_{I_{1:\kappa}}(x)}{f_\kappa(x)} dx < \frac{\epsilon}{2} \right\}. \quad (4.35)$$

Combining (4.34) and (4.35), we see that any $\mathbf{p} \in \mathcal{S}_{\kappa-1}$ satisfying $\|\mathbf{p} - \boldsymbol{\rho}\|_1 < \delta$ yields

$$K(f_0, f_\kappa) = \int_0^1 f_0(x) \log \frac{f_0(x)}{f_\kappa(x)} dx < \epsilon.$$

Hence, we conclude that

$$\{k = \kappa, \mathbf{p} \in \mathcal{S}_{\kappa-1} : \|\mathbf{p} - \boldsymbol{\rho}\|_1 < \delta\} \subseteq \{k = \kappa, f_\kappa : K(f_0, f_\kappa) < \epsilon\}.$$

Finally, we lower bound the probability of $\{f: K(f_0, f) < \epsilon\}$,

$$\begin{aligned}
P_n(f: K(f_0, f) < \epsilon) &\geq P_n(k = \kappa, f_\kappa: K(f_0, f_\kappa) < \epsilon) \\
&\geq P_n(k = \kappa, \mathbf{p} \in \mathcal{S}_{\kappa-1}: \|\mathbf{p} - \boldsymbol{\rho}\|_1 < \delta) \\
&= P_n\left(\|\mathbf{p} - \boldsymbol{\rho}\|_1 < \delta \mid \kappa\right) p_n(\kappa) \\
&= P\left(\|\mathbf{p} - \boldsymbol{\rho}\|_1 < \delta \mid \kappa\right) p_n(\kappa),
\end{aligned}$$

where we used that $p(\mathbf{p} \mid \kappa)$ does not depend on n . By assumption, we have that $\liminf_{n \rightarrow \infty} p_n(k) > 0$, and the assumption that $\mathbf{p} \mid k$ is Dirichlet together with Lemma A.1 ensures that the first factor is strictly positive, establishing the claim. $\square$

4.4 Posterior consistency

The objective of the present section is to establish conditions under which posterior consistency can be verified for the regular random histogram prior. As it turns out, posterior consistency holds for a large class of densities; the regularity assumptions on the true density function f_0 are rather minimal, all we require is boundedness and a.e. continuity.

We note that there are some previously published results concerning posterior consistency of histogram priors. Barron et al. (1999) gave a consistency result for regular histogram priors with a symmetric Dirichlet prior distribution on the cell probabilities. However, their result does not cover the case of using a sequence of sample size dependent priors, which we will use in our study, so the results presented here are more general than those of Barron et al. (1999). Scricciolo (2007) studied posterior convergence rates for regular histogram priors. Her results also imply posterior consistency, but involve making more regularity assumptions on the true density f_0 , so they do not apply quite as generally as the results we derive in this section. We return to the results of Scricciolo in Section 4.6.

4.4.1 Posterior consistency of regular histogram priors

In the following, we let $I_{1:k}$ be a potentially irregular interval partition of $[0, 1]$. We define

$$\mathcal{H}(I_{1:k}) = \left\{ h: [0, 1] \rightarrow \mathbb{R}: h(x) = \sum_{j=1}^k |I_j|^{-1} p_j \mathcal{I}(x \in I_j) \text{ for some } \mathbf{p} \in \mathcal{S}_{k-1} \right\} \quad (4.36)$$

to be the set of piecewise constant densities based on the partition $I_{1:k}$. Since the Kullback–Leibler property was already established in Theorem 4.1, we need to find a sieve $\{\mathcal{F}_n\}$ satisfying the additional conditions of Theorem 4.2 in order to show that the posterior is consistent. In particular, we will study a sequence of sets of the form $\mathcal{F}_n = \cup_{k=1}^{k_n} \mathcal{H}(I_{1:k})$, where k_n diverges to infinity, but $n^{-1}k_n \rightarrow 0$. To verify the covering number condition (3.42), we derive a bound on the ϵ -covering number of $\mathcal{F}_n$, valid for arbitrarily small $\epsilon > 0$. To this end, our strategy is to bound the covering number of $\mathcal{H}(I_{1:k})$ in terms of the bracketing number of the k -simplex, using the Hellinger metric on $\mathbb{R}_+^k$,

$$d_H(\mathbf{x}, \mathbf{y}) = \left\{ \sum_{j=1}^k (\sqrt{x_j} - \sqrt{y_j})^2 \right\}^{1/2}.$$

As we shall make use of Lemma 4.4 for irregular histograms as well, we allow for irregular partitions in the following two results. Before we state our first result, it is important to recall Definitions 3.4 and 3.5, as we work with bracketing numbers in the following.

Lemma 4.4. *Let $I_{1:k}$ 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 $\omega_j(x) = |I_j|^{-1} \mathcal{I}(x \in I_j)$. Suppose that $h(x) = \sum_{j=1}^k p_j \omega_j(x)$, where $\mathbf{p} \in \mathcal{S}_{k-1}$. If $[\mathbf{a}, \mathbf{b}]$ is an ϵ -bracket containing $\mathbf{p}$, then $[l, u]$ is an ϵ -bracket containing h .*

Proof. To establish $h \in [l, u]$, we simply note that for any $x \in [0, 1]$, $h(x) = |I_j|^{-1} p_j$ for some j . The claim that $h \in [l, u]$ now follows immediately from the fact that $|I_j|^{-1} a_j \leq |I_j|^{-1} p_j \leq |I_j|^{-1} b_j$ for all j by construction.

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

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

From Lemma B.1, 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. $\square$

We now give a bound on the covering number of $\mathcal{H}(I_{1:k})$ and use it to bound the covering number of $\mathcal{F}_n$.

Lemma 4.5. *Let $I_{1:k}$ be an interval partition of the unit interval. Then the ϵ -bracketing number of $\mathcal{H}(I_{1:k})$ for $\epsilon \leq 1$ satisfies*

$$N_{[]}(\epsilon, \mathcal{H}(I_{1:k}), d_H) \leq N_{[]}(\epsilon, \mathcal{S}_{k-1}, d_H) \leq \frac{k(2\pi e)^{k/2}}{\epsilon^{k-1}} \quad (4.37)$$

Moreover, the ϵ -bracketing number of $\mathcal{F}_n = \cup_{k=1}^{k_n} \mathcal{H}(I_{1:k})$ satisfies

$$N_{[]}(\epsilon, \mathcal{F}_n, d_H) \leq \sum_{k=1}^{k_n} \frac{k(2\pi e)^{k/2}}{\epsilon^{k-1}}. \quad (4.38)$$

Proof. To show that the first claim holds, let $[\mathbf{a}_1, \mathbf{b}_1], \dots, [\mathbf{a}_m, \mathbf{b}_m]$ be an ϵ -bracketing of $\mathcal{S}_{k-1}$, and assume that m is chosen so that the number of brackets is minimal. Define the associated functions $l_1, u_1, \dots, l_m, u_m$ as in Lemma 4.4, that is, define

$$l_i(x) = \sum_{j=1}^k a_{i,j} |I_j|^{-1} \mathcal{I}(x \in I_j), \quad (4.39)$$

$$u_i(x) = \sum_{j=1}^k b_{i,j} |I_j|^{-1} \mathcal{I}(x \in I_j). \quad (4.40)$$

Let $h \in \mathcal{H}(I_{1:k})$ be given by $h(x) = \sum_{j=1}^k |I_j|^{-1} p_j \mathcal{I}(x \in I_j)$, for some $\mathbf{p} \in \mathcal{S}_{k-1}$. Since $\mathcal{S}_{k-1} \subseteq \cup_{l=1}^m [\mathbf{a}_l, \mathbf{b}_l]$, we have $\mathbf{p} \in [\mathbf{a}_i, \mathbf{b}_i]$ for some i . By Lemma 4.4, $[l_i, u_i]$ is an ϵ -bracket containing h , and since h was arbitrary, we conclude that $\mathcal{H}(I_{1:k}) \subseteq \cup_{i=1}^m [l_i, u_i]$. By the minimality of m , this shows that the Hellinger bracketing number of $\mathcal{H}(I_{1:k})$ can be no

greater than that of $\mathcal{S}_{k-1}$. By Lemma B.2, the bracketing number of the k -simplex $\mathcal{S}_{k-1}$ satisfies

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

Using this fact we find that

$$N_{[]}(\epsilon, \mathcal{H}(I_{1:k}), d_H) \leq N_{[]}(\epsilon, \mathcal{S}_{k-1}, d_H) \leq \frac{k(2\pi e)^{k/2}}{\epsilon^{k-1}}, \quad (4.41)$$

which was to be shown.

To prove the second part of the lemma, we use the union bound of Proposition 3.2, obtaining

$$N_{[]}(\epsilon, \mathcal{F}_n, d_H) = N_{[]}(\epsilon, \cup_{k=1}^{k_n} \mathcal{H}(I_{1:k}), d_H) \leq \sum_{k=1}^{k_n} N_{[]}(\epsilon, \mathcal{H}(I_{1:k}), d_H) \leq \sum_{k=1}^{k_n} \frac{k(2\pi e)^{k/2}}{\epsilon^{k-1}},$$

as desired. $\square$

We are now ready to prove the main result of this section:

Theorem 4.2. *Let $\{P_n\}$ denote a sequence of regular random histogram priors satisfying Assumption 4.1. Let $\{k_n\}$ be a sequence of natural numbers satisfying*

$$k_n \rightarrow \infty \quad \text{and} \quad k_n/n \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (4.42)$$

If $\sum_{k=k_n+1}^{\infty} p_n(k) < \exp(-c_1 n)$ for some constant $c_1 > 0$, then the posterior distribution is Hellinger consistent at any bounded, a.e. continuous density f_0 on the unit interval.

Proof. Since the Kullback–Leibler property holds by Theorem 4.1, it suffices to verify conditions (3.40) and (3.39). To this end, we define $\mathcal{F}_n = \cup_{k=1}^{k_n} \mathcal{H}(I_{1:k})$, where $I_{1:k}$ now denotes the regular interval partition of $[0, 1]$ using k bins. To show that the prior mass condition (3.40) of Theorem 3.2 holds, we note that the prior only assigns mass to the set $\cup_{k=1}^{\infty} \mathcal{H}(I_{1:k})$ and hence,

$$P_n(\mathcal{F}_n^c) = P_n\left(\cup_{k=k_n+1}^{\infty} \mathcal{H}(I_{1:k})\right) \leq \sum_{k=k_n+1}^{\infty} P_n(\mathcal{H}(I_{1:k})) = \sum_{k=k_n+1}^{\infty} p_n(k) < \exp(-c_1 n),$$

establishing (3.40).

To prove that the covering number condition (3.39) is satisfied, we note that it suffices to consider $\epsilon \leq 1$. We argue in a similar way to the proof of Theorem 2.2 in Ghosal (2001). From equation (4.41), we have that

$$N_{[]}(\epsilon, \mathcal{F}_n, d_H) \leq \sum_{k=1}^{k_n} \frac{k(2\pi e)^{k/2}}{(2\epsilon)^{k-1}}.$$

We claim that $k(2\pi e)^{k/2}/\epsilon^{k-1} \leq (L/\epsilon)^k$ for a universal constant $L > 1$. Indeed, as the following calculation shows,

$$\frac{k(2\pi e)^{k/2}}{\epsilon^{k-1}} = \frac{k(\sqrt{2\pi e})^k}{\epsilon^{k-1}} \leq \frac{k(\sqrt{2\pi e})^k}{\epsilon^k} \leq \frac{2^k(\sqrt{2\pi e})^k}{\epsilon^k} = \left(\frac{2\sqrt{2\pi e}}{\epsilon}\right)^k,$$

where we used that $\epsilon \leq 1$ for the first inequality and the fact that $k \leq 2^k$ for any $k \in \mathbb{N}$ for the second. The claim thus holds for $L = 2\sqrt{2\pi e}$.

Since the function $k \mapsto x^k$ is increasing for any $x > 1$, we have $(L/\epsilon)^k \leq (L/\epsilon)^{k_n}$ for all $k \leq k_n$ and hence,

$$N(\epsilon, \mathcal{F}_n, d_H) \leq N_{[]}(\epsilon, \mathcal{F}_n, d_H) \leq \sum_{k=1}^{k_n} (L/\epsilon)^k \leq k_n (L/\epsilon)^{k_n} = e^{\{k_n \log(L/\epsilon) + \log k_n\}},$$

where we used Proposition 3.3 for the first inequality. Since $k_n/n \rightarrow 0$ by assumption, it holds for all sufficiently large n ,

$$n^{-1}[k_n \log(L/\epsilon) + \log k_n] < \epsilon^2.$$

We deduce that for such n ,

$$N(\epsilon, \mathcal{F}_n, d_H) \leq \exp\left(n^{-1}\{k_n \log(L/\epsilon) + \log k_n\}n\right) < \exp(n\epsilon^2).$$

Hellinger consistency now follows from Theorem 3.2. $\square$

4.5 Consistency of the histogram point estimator

While the consistency result of Section 4.4 yields a consistent posterior, it does not necessarily imply that the Bayes histogram estimator (4.14) will be consistent. The two notions of consistency are of a somewhat different nature; posterior consistency concerns the behavior of the random variable $P_n(d(f_0, f) > \epsilon | \mathbf{x}_{1:n})$, while the consistency of the estimator $\hat{f}_{B,n}$ depends on the behavior of $d(f_0, \hat{f}_{B,n})$, and statements regarding the former do not necessarily translate directly into a statement about the latter. In the present section, we show that, indeed, the resulting histogram point estimator is consistent whenever the true f_0 is bounded and a.e. continuous, under a new set of slightly different conditions on the prior distribution sequence. As before, we state necessary assumptions on both the true density f_0 and the prior in order for consistency to hold.

Assumption 4.2. The regular random histogram prior sequence $\{P_n\}$ satisfies the following:

- (i) The prior for k , $p_n(k)$, is supported on the set $\{1, 2, \dots, k_n\}$ for all n , where $\{k_n\}$ is a sequence diverging to infinity that satisfies $k_n = o(n/\log(n))$. Moreover, p_n satisfies, for all $k \leq k_n$,

$$\log p_n(k) = o(n). \quad (4.43)$$

- (ii) The prior for $\mathbf{p} | k$ is $\text{Dir}(\mathbf{a}_k)$, where $\mathbf{a}_k$ satisfies $0 < a_{k,j} \leq A$ for a positive constant A and all k, j .

We note that assumption (i) holds in some cases where Assumption 4.1 (i) does not. For instance, if $p_n(k)$ is taken to be the uniform distribution on $\{1, 2, \dots, k_n\}$, then $\log p_n(k) = -\log(k_n)$, and hence (4.43) holds, while (4.24) does not. Thus, the result of Theorem 4.3 can be used to verify the posterior consistency of $\hat{f}_{B,n}$ for some choices of $p_n(k)$ where Theorem 4.2 cannot be applied to prove a consistent posterior.

The main result of this section shows that the histogram estimator based on the MAP $\hat{k}$ is consistent. Past studies on the consistency of random histograms have mainly been interested in showing that the posterior distribution is consistent, and have not investigated the question of consistency for the Bayes histogram estimator. Theorem 4.3 is as such, to the best of our knowledge, a completely new result in this regard.

Theorem 4.3. *Suppose that Assumption 4.2 holds. Then the Bayes histogram estimator $\hat{f}_{B,n} = f(\cdot | \hat{k}, \hat{\mathbf{p}}_B)$ is $\mathbb{L}_1$ consistent at f_0 . Moreover,*

$$\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_0, \hat{f}_{B,n}) \right\} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (4.44)$$

For notational simplicity, we shall in the following let $f_{0,k}$ denote for the piecewise constant version of f_0 based on using $k \in \mathbb{N}$ regular bins, and write $\pi_j = \int_{I_j} f_0(x) dx$ for the bin probabilities, since this subsection is only concerned with regular histograms. The key to the proof of Theorem 4.3 is the result of Lemma 4.7, which implies that for any two numbers of bins $\kappa_1, \kappa_2 \in \mathbb{N}$, the posterior will eventually favor the value that minimizes the KL divergence between f_0 and its piecewise constant approximation based on κ_1 or κ_2 bins. Some care must be taken to handle the case where the true density f_0 is a regular step function, so that $f_0 = f_{0,k_0}$ for some $k_0 \in \mathbb{N}$, as in this case the Kullback–Leibler divergence satisfies $K(f_0, f_{0,mk_0}) = 0$ for any $m \in \mathbb{N}$. For regular step functions, the distribution of $\hat{k}$ may concentrate on bounded sets as the sample size approaches infinity. If f_0 is not a regular step function we will instead have $K(f_0, f_{0,k}) > 0$ for all $k \in \mathbb{N}$ with the KL divergence approaching 0 as $k \rightarrow \infty$. As we shall see in Lemma 4.7, this leads to $\hat{k}$ diverging to positive infinity in probability.

To establish the desired result, we first need to work with the log-posterior for k in the large-sample limit.

Lemma 4.6. *Suppose that Assumption 4.2 holds. Then for any $k \in \mathbb{N}$ in the support of the prior, the log-posterior $\log p_n(k | \mathbf{x}_{1:n})$ in the Dirichlet histogram model satisfies*

$$\log p_n(k | \mathbf{x}_{1:n}) = \log L_n(\hat{\mathbf{p}} | k) - \log(Z_n) + o(n), \quad (4.45)$$

where $\hat{\mathbf{p}}$ is the MLE, $L_n(\hat{\mathbf{p}} | k) = \sum_{j=1}^n N_j \log(k \hat{p}_j)$, and Z_n is a random variable not depending on k .

Proof (sketch). By Bayes' rule (4.17), we can write

$$\log p_n(k | \mathbf{x}_{1:n}) = \log p(\mathbf{x}_{1:n} | k) + \log p_n(k) - \log(Z_n),$$

where $p(\mathbf{x}_{1:n} | k)$ is the marginal likelihood given k and Z_n is the normalization constant in the posterior for k . By arguing as in Lemma 5.3, one can deduce that the log-marginal likelihood satisfies

$$\log p(\mathbf{x}_{1:n} | k) = \log L_n(\hat{\mathbf{p}} | k) + o(n). \quad (4.46)$$

Since $\log p_n(k) = o(n)$ by assumption, (4.45) follows. $\square$

The proof of (4.46) is omitted, as it is established using techniques identical to those used in the proof of Lemma 5.3, relying on Stirling's approximation to the loggamma function. The result of Lemma 4.6 ensures that in the large-sample limit, the value of the log-likelihood eventually dominates the expression for the log-posterior. This is the key fact that allows us to show that the MAP $\hat{k}$ will eventually concentrate on values that come arbitrarily close f_0 in the Kullback–Leibler sense. This is the approach taken in the proof of Lemma 4.7. Before we state the lemma, define for $K > 0$,

$$\mathcal{A}_K = \{k \in \mathbb{N} : k > K \text{ or } f_0 = f_{0,k}\}. \quad (4.47)$$

Lemma 4.7. *For any bounded, a.e. continuous density f_0 and $K > 0$, we have that*

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{k} \in \mathcal{A}_K^c) = 0. \quad (4.48)$$

Proof. Fix $K > 0$. Our strategy is to show that the posterior probability of any $k \notin \mathcal{A}_K$ converges to zero as $n \rightarrow \infty$ via the large-sample approximation in (4.45). By the positivity of the KL-divergence we have that $K(f_0, f_{0,k}) > 0$ for all $k \in \mathcal{A}_K^c$, since $f_0 \neq f_{0,k}$. We conclude that $\min_{k \notin \mathcal{A}_K} K(f_0, f_{0,k}) > 2\epsilon$ for some $\epsilon > 0$ since there are only finitely many $k \in \mathcal{A}_K^c$. On the other hand, by Corollary 4.2 we can find $\kappa \in \mathbb{N}$ such that $K(f_0, f_{0,\kappa}) < \epsilon$, and this κ can be chosen to be larger than K . From Lemma 4.6, we have that

$$n^{-1} \log p_n(k | \mathbf{x}_{1:n}) = \sum_{j \in J_k} \hat{p}_j \log(k \hat{p}_j) - \log(Z_n) + o(1), \quad (4.49)$$

where $J_k = \{j: \hat{p}_j > 0\}$ are the nonempty bins. Subtracting and adding $C_n = n^{-1} \sum_{i=1}^n \log f_0(x_i) - \log(Z_n)$ to (4.49) results in

$$n^{-1} \log p_n(k | \mathbf{x}_{1:n}) = \sum_{j \in J_k} \hat{p}_j \log(k \hat{p}_j) - \sum_{i=1}^n \log f_0(x_i) + C_n + o(1). \quad (4.50)$$

Note that

$$\sum_{i=1}^n \log f_0(x_i) = \sum_{i=1}^n \log \frac{f_0(x_i)}{f_{0,k}(x_i)} + \sum_{i=1}^n \log f_{0,k}(x_i), \quad (4.51)$$

where $\log f_{0,k}(x) = \sum_{j=1}^k \mathcal{I}(x \in I_j) \log(k \pi_j)$, for $\pi_j = \int_{I_j} f_0(s) ds$. The last term of (4.51) is thus seen to be equal to $\sum_{j \in J_k} N_j \log(k \pi_j)$. To handle the first term, we note that by the Law of Large Numbers, it holds for any $k \in \{1, 2, \dots, K\} \cup \{\kappa\}$,

$$n^{-1} \sum_{i=1}^n \log \frac{f_0(x_i)}{f_{0,k}(x_i)} = K(f_0, f_{0,k}) + o_{\mathbb{P}}(1).$$

Combining this fact with (4.50), we obtain

$$\begin{aligned} n^{-1} \log p_n(k | \mathbf{x}_{1:n}) &= \sum_{j \in J_k} \hat{p}_j \log(k \hat{p}_j) - \sum_{i=1}^n \log f_{0,k}(x_i) - \sum_{i=1}^n \log \frac{f_0(x_i)}{f_{0,k}(x_i)} + C_n + o(1). \\ &= \sum_{j \in J_k} \hat{p}_j \log(k \hat{p}_j) - \sum_{j \in J_k} \hat{p}_j \log(k \pi_j) - K(f_0, f_{0,k}) + C_n + o_{\mathbb{P}}(1) \end{aligned}$$

We will now show that the difference of the two first terms on the right hand side converge to zero in probability. By the inequality $\log(x) \leq x - 1$ (A.14), we have

$$\hat{p}_j \log \frac{\hat{p}_j}{\pi_j} \leq \hat{p}_j \left(\frac{\hat{p}_j}{\pi_j} - 1 \right).$$

Summing over $j \in J_k$, we obtain

$$\sum_{j \in J_k} \hat{p}_j \log \frac{\hat{p}_j}{\pi_j} \leq \sum_{j \in J_k} \left\{ \frac{\hat{p}_j^2}{\pi_j} - \hat{p}_j \right\} = \sum_{j \in J_k} \frac{(\hat{p}_j - \pi_j)^2}{\pi_j}. \quad (4.52)$$

Moreover, since $\sum_{j \in J_k} \hat{p}_j \log\{\hat{p}_j/\pi_j\}$ can be written as the Kullback–Leibler divergence between $\hat{f}_n$ and $f_{0,k}$, where $\hat{f}_n$ is the maximum likelihood histogram estimator based on k regular bins (2.10), the left hand side of (4.52) is nonnegative. Taking expectations on the right hand side of (4.52), we observe that

$$\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ \sum_{j \in J_k} \frac{(\hat{p}_j - \pi_j)^2}{\pi_j} \middle| k \right\} \leq \sum_{j: \pi_j > 0} \frac{\text{Var}_{\mathbf{x}_{1:n}}(\hat{p}_j | k)}{\pi_j} = \sum_{j: \pi_j > 0} \frac{\pi_j(1 - \pi_j)}{n\pi_j} = \frac{k-1}{n} \leq n^{-1}k_n,$$

where the first inequality is due to the fact that $\pi_j = 0$ implies $\hat{p}_j = 0$ almost surely together with the monotonicity of expectations. The last expression in the above converges to zero by assumption. We conclude via Markov's inequality that $\sum_{j \in J_k} \hat{p}_j \log(\hat{p}_j/\pi_j) = o_{\mathbb{P}}(1)$ and thus for any $k \in \mathcal{A}_K^c \cup \{\kappa\}$,

$$n^{-1} \log p_n(k | \mathbf{x}_{1:n}) = -K(f_0, f_{0,k}) + C_n + o_{\mathbb{P}}(1). \quad (4.53)$$

For any $k \notin \mathcal{A}_K$, we thus get the following upper bound on the posterior pmf:

$$n^{-1} \log p_n(k | \mathbf{x}_{1:n}) \leq -\min_{k \notin \mathcal{A}_K} K(f_0, f_{0,k}) + C_n + o_{\mathbb{P}}(1) < -2\epsilon + C_n + o_{\mathbb{P}}(1).$$

On the other hand, since $K(f_0, f_{0,\kappa}) < \epsilon$, (4.53) also implies that

$$n^{-1} \log p_n(\kappa | \mathbf{x}_{1:n}) \geq -\epsilon + C_n + o_{\mathbb{P}}(1).$$

Recalling that the maximum of a fixed, finite number of $o_{\mathbb{P}}(1)$ random variables is $o_{\mathbb{P}}(1)$ (see Proposition 2.2), we conclude that

$$\max_{k \in \mathcal{A}_K^c} \log p_n(k | \mathbf{x}_{1:n}) - \log p_n(\kappa | \mathbf{x}_{1:n}) \leq n \{-\epsilon + o_{\mathbb{P}}(1)\} \xrightarrow{\mathbb{P}} -\infty,$$

which implies that $\max_{k \in \mathcal{A}_K^c} p_n(k | \mathbf{x}_{1:n})/p_n(\kappa | \mathbf{x}_{1:n}) = o_{\mathbb{P}}(1)$. In particular, this means that

$$\mathbb{P}(\hat{k} \in \mathcal{A}_K^c) \leq \mathbb{P}\left(\max_{k \in \mathcal{A}_K^c} p_n(k | \mathbf{x}_{1:n}) \geq p_n(\kappa | \mathbf{x}_{1:n})\right) \rightarrow 0,$$

which was to be shown. $\square$

We are finally in a position to prove the main result of this section:

Proof of Theorem 4.3. It suffices to show that (4.44) holds, as $\mathbb{L}_1$ consistency is then immediate from Markov's inequality. Now fix $\epsilon > 0$ and let $K > 0$, to be chosen later. Then the $\mathbb{L}_1$ risk R_n of $\hat{f}_{B,n}$ can be decomposed as

$$\begin{aligned} R_n &= \mathbb{E}_{\mathbf{x}_{1:n}} \{d_1(f_0, \hat{f}_{B,n})\} \\ &= \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_0, \hat{f}_{B,n}) \mathcal{I}(\hat{k} \in \mathcal{A}_K^c) \right\} + \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_0, \hat{f}_{B,n}) \mathcal{I}(\hat{k} \in \mathcal{A}_K) \right\} \\ &\leq 2 \mathbb{P}(\hat{k} \in \mathcal{A}_K^c) + \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_0, \hat{f}_{B,n}) \mathcal{I}(\hat{k} \in \mathcal{A}_K) \right\} \end{aligned}$$

where we used that $d_1(f, g) \leq 2$ for all densities f, g . By the triangle inequality, this last expression is upper bounded by

$$\underbrace{2 \mathbb{P}(\hat{k} \in \mathcal{A}_K^c)}_{(i)} + \underbrace{\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_0, f_{0,\hat{k}}) \mathcal{I}(\hat{k} \in \mathcal{A}_K) \right\}}_{(ii)} + \underbrace{\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_{0,\hat{k}}, \hat{f}_{B,n}) \mathcal{I}(\hat{k} \in \mathcal{A}_K) \right\}}_{(iii)} \quad (4.54)$$

To show that we can make (4.54) arbitrarily small, we deal with (ii) first. Note that the integrand of $\int_0^1 |f_0(x) - f_{0,k}(x)| dx$ is bounded from above by $2 \sup_{x \in [0,1]} f_0(x)$. It follows from Proposition 4.1 and the Dominated Convergence Theorem that $d_1(f_0, f_{0,k}) \rightarrow 0$ as $k \rightarrow \infty$. Hence, by making K sufficiently large, we have that $d_1(f_0, f_{0,k}) < \epsilon$ for all $k > K$. Consequently, as $d_1(f_0, f_{0,\hat{k}})$ only depends on the data through $\hat{k}$,

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_0, f_{0,\hat{k}}) \mathcal{I}(\hat{k} \in \mathcal{A}_K) \right\} \\ &\leq \mathbb{E}_{\hat{k}} \left\{ d_1(f_0, f_{0,\hat{k}}) \mathcal{I}(\hat{k} > K) \right\} + \mathbb{E}_{\hat{k}} \left\{ d_1(f_0, f_{0,\hat{k}}) \mathcal{I}(\hat{k} \in \{k : f_0 = f_{0,k}\}) \right\} < \epsilon, \end{aligned}$$

where we used that for given $\hat{k} > K$ we have $d_1(f_0, f_{0,\hat{k}}) \mathcal{I}(\hat{k} > K) < \epsilon$ and $d_1(f_0, f_{0,\hat{k}}) \mathcal{I}(\hat{k} \in \{k: f_0 = f_{0,k}\}) = 0$.

To handle (iii) in (4.54), we develop an upper bound on bound the expected $\mathbb{L}_1$ distance between $f_{0,\hat{k}}$ and $\hat{f}_{B,n}$ independent of the value of $\hat{k}$. By the triangle inequality,

$$\begin{aligned} d_1(f_{0,\hat{k}}, \hat{f}_{B,n}) &= \sum_{j=1}^{\hat{k}} |\pi_j - \hat{p}_{B,j}| \\ &\leq \sum_{j=1}^{\hat{k}} \left| \pi_j - \mathbb{E}_{\mathbf{N}_{\hat{k}}} \left\{ \hat{p}_{B,j} \mid \hat{k} \right\} \right| + \sum_{j=1}^{\hat{k}} \left| \mathbb{E}_{\mathbf{N}_{\hat{k}}} \left\{ \hat{p}_{B,j} \mid \hat{k} \right\} - \hat{p}_{B,j} \right|. \end{aligned} \quad (4.55)$$

Recall that $\hat{p}_{B,j} = p_{0,j}a/(a+n) + \hat{p}_j n/(a+n)$ where $\hat{p}_j = N_j/n$ and hence $\mathbb{E}_{\mathbf{N}_{\hat{k}}} \left\{ \hat{p}_{B,j} \mid \hat{k} \right\} = p_{0,j}a/(a+n) + \pi_j n/(a+n)$. From this we deduce that the first term on the right hand side of (4.55) becomes

$$\sum_{j=1}^{\hat{k}} \left| \pi_j - \mathbb{E}_{\mathbf{N}_{\hat{k}}} \left\{ \hat{p}_{B,j} \mid \hat{k} \right\} \right| = \frac{a}{a+n} \sum_{j=1}^{\hat{k}} |\pi_j - p_{0,j}| \leq 2an^{-1}.$$

Taking the expected value of the second term in (4.55) conditional on $\hat{k}$, we end up with

$$\frac{n}{a+n} \sum_{j=1}^{\hat{k}} \mathbb{E}_{\mathbf{N}_{\hat{k}}} \left\{ |\pi_j - \hat{p}_j| \mid \hat{k} \right\}.$$

An application of the Cauchy–Schwarz inequality yields the following upper bound on the last expression:

$$\mathbb{E}_{\mathbf{N}_{1:\hat{k}}} \left\{ |\pi_j - \hat{p}_j| \mid \hat{k} \right\} \leq \left[\mathbb{E}_{\mathbf{N}_{1:\hat{k}}} \left\{ (\pi_j - \hat{p}_j)^2 \mid \hat{k} \right\} \right]^{1/2} \leq \sqrt{\pi_j(1-\pi_j)/n} \leq \sqrt{\pi_j/n}.$$

It follows that the conditional expectation of $\sum_{j=1}^{\hat{k}} |\mathbb{E}_{\mathbf{N}_{\hat{k}}} \left\{ \hat{p}_{B,j} \mid \hat{k} \right\} - \hat{p}_{B,j}|$ is upper bounded by

$$\frac{\sqrt{n}}{a+n} \sum_{j=1}^{\hat{k}} \sqrt{\pi_j} \leq n^{-1/2} \sum_{j=1}^{\hat{k}} \sqrt{\pi_j}.$$

Since f_0 is bounded from above by M , we have that $\pi_j \leq M\hat{k}^{-1}$ for all j , and thus, we can further bound the right hand side of the above by $\sqrt{M\hat{k}/n}$. Putting together our estimates, we arrive at the following upper bound

$$\mathbb{E}_{\mathbf{N}_{\hat{k}}} \left\{ d_1(f_{0,\hat{k}}, \hat{f}_{B,n}) \mid \hat{k} \right\} \leq 2a/n + \sqrt{M\hat{k}/n} \leq 2a/n + \sqrt{Mk_n/n}.$$

Due to the boundedness assumption on the $a_{k,j}$, we have $a = \mathcal{O}(k_n)$. Moreover, since $k_n = o(n)$, it follows that the right hand side of the above tends to zero as $n \rightarrow \infty$. Using the tower property, it follows that

$$\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_{0,\hat{k}}, \hat{f}_{B,n}) \mathcal{I}(\hat{k} \in \mathcal{A}_K) \right\} \leq \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_{0,\hat{k}}, \hat{f}_{B,n}) \right\} < 2a/n + \sqrt{Mk_n/n} = o(1).$$

Lastly, we deal with (i) of (4.54). By Lemma 4.7, we have $\mathbb{P}(\hat{k} \in \mathcal{A}_K^c) \rightarrow 0$ as $n \rightarrow \infty$. Putting together previous estimates with the upper bound from (4.54) then yields the following inequality,

$$\limsup_{n \rightarrow \infty} R_n \leq \lim_{n \rightarrow \infty} 2\mathbb{P}(\hat{k} \in \mathcal{A}_K^c) + \epsilon + \lim_{n \rightarrow \infty} \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_{0,\hat{k}}, \hat{f}_{B,n}) \mathcal{I}(\hat{k} \in \mathcal{A}_K) \right\} = \epsilon.$$

Since $\epsilon > 0$ was arbitrary we conclude that $\lim_{n \rightarrow \infty} R_n = 0$ as claimed. $\square$

4.6 Convergence rate under regular random histogram priors

In the present section we describe convergence results for the regular random histogram priors under appropriate regularity assumptions on the true density f_0 . We start by considering the approximation properties of histograms for functions of varying degrees of smoothness. These results are then used to study the convergence rate of Bayesian histogram procedures.

4.6.1 Approximation properties of piecewise constant densities

The rate at which the posterior distribution converges is strongly linked to the rate at which the distance between the piecewise constant version $f_{0,I_{1:k}}$ and the true density f_0 goes to zero as a function of the number of bins. In the following, we will study the case where either $\sqrt{f_0}$ or f_0 is α -Hölder continuous (Definition 2.10), which yields an approximation rate with respect to a suitable metric as a function of k depending on the exponent α . Since this approximation rate will come in handy in later proofs, we state this approximation property more formally in the form of a proposition.

Proposition 4.2. *Let $h: [0, 1] \rightarrow \mathbb{R}$ be an α -Hölder continuous function for $\alpha \in (0, 1]$. Let $\{I_{k,1}, \dots, I_{k,k}\}_{k=1}^\infty$ be a sequence of partitions of $[0, 1]$ such that $\max_j |I_{k,j}| = \mathcal{O}(k^{-1})$. Then the sequence of piecewise constant approximations $\{h_{I_{1:k}}\}_{k=1}^\infty$ of h (2.11) satisfies*

$$d_\infty(h, h_{I_{1:k}}) = \mathcal{O}(k^{-\alpha}). \quad (4.56)$$

From (4.56) we see that the rate at which the piecewise constant functions $\{h_{I_{1:k}}\}$ approximate h depends on its smoothness α . Smoother functions are approximated at a faster rate, with the best approximation rate obtained for Lipschitz continuous functions at $\mathcal{O}(k^{-1})$.

Proof of Proposition 4.2. Fix $k \in \mathbb{N}$ and let $x \in [0, 1]$. Let j be such that $x \in I_{k,j}$. Then,

$$\begin{aligned} |h(x) - h_{I_{1:k}}(x)| &= \left| h(x) - |I_{k,j}|^{-1} \int_{I_{k,j}} h(s) \, ds \right| \\ &\leq |I_{k,j}|^{-1} \int_{I_{k,j}} |h(x) - h(s)| \, ds \\ &\leq |I_{k,j}|^{-1} \int_{I_{k,j}} L_0 |x - s|^\alpha \, ds \\ &\leq |I_{k,j}|^{-1} L_0 |I_{k,j}|^{1+\alpha} \\ &= L_0 |I_{k,j}|^\alpha \end{aligned}$$

where we used that $|x - s| \leq |I_{k,j}|$ for $x, s \in I_{k,j}$. Since $\max_j |I_{k,j}| = \mathcal{O}(k^{-1})$, we find

$$d_\infty(h, h_{I_{1:k}}) = \sup_{x \in [0,1]} |h(x) - h_{I_{1:k}}(x)| \leq L_0 \max_j |I_{k,j}|^\alpha = \mathcal{O}(k^{-\alpha}),$$

which was to be shown. $\square$

Since most of our convergence rate results concern to the Hellinger metric, approximation rates with respect to this distance is needed. From the above proposition, a trivial bound on the approximation rate for an α -Hölder continuous density f_0 on

$[0, 1]$ in the squared Hellinger distance can be obtained by utilizing the following chain of inequalities from Proposition A.3:

$$d_H^2(f_0, f_{0,I_{1:k}}) \leq d_1(f_0, f_{0,I_{1:k}}) \leq d_\infty(f_0, f_{0,I_{1:k}}),$$

with the implied rate being $\mathcal{O}(k^{-\alpha})$. However, the above approximation bound is too loose if the goal is to establish a convergence rate close to the minimax one. One way of strengthening the approximation rate is to additionally assume that f_0 is strictly positive, as is done in the work of Scricciolo (2007), as this allows for the use of the inequalities $d_H^2(f_0, f_{0,I_{1:k}}) \leq m^{-1}d_2^2(f_0, f_{0,I_{1:k}}) \leq m^{-1}d_\infty^2(f_0, f_{0,I_{1:k}})$ of Proposition A.3, where m is a positive lower bound on f_0 . However, it turns out that this assumption is stronger than what is strictly necessary for our purposes, as it can be shown that the approximation rate can be bettered under a set of weaker assumptions. In particular, if we assume that the root-density is Hölder continuous, we obtain the desired approximation rate without any additional assumptions.

Proposition 4.3. *Let $f_0: [0, 1] \rightarrow \mathbb{R}$ be a probability density function such that $\sqrt{f_0}$ is α -Hölder continuous. Then for any sequence of partitions $\{I_{k,1}, \dots, I_{k,k}\}_{k=1}^\infty$ satisfying $\max_j |I_{k,j}| = \mathcal{O}(k^{-1})$,*

$$d_H^2(f_0, f_{0,I_{1:k}}) = \mathcal{O}(k^{-2\alpha}).$$

Proof. Letting $\pi_{k,j} = \int_{I_{k,j}} f_0(x) dx$ as previously, we note that by the mean value theorem for integrals, we can find an $s_{k,j} \in I_{k,j}$ such that $f_{0,I_{1:k}}(x) = \pi_{k,j}/|I_{k,j}| = f_0(s_{k,j})$ for all $x \in I_{k,j}$. Hence,

$$\begin{aligned} d_H^2(f_0, f_{0,I_{1:k}}) &= \sum_{j=1}^k \int_{I_{k,j}} \left(\sqrt{f_0(x)} - \sqrt{f_0(s_{k,j})} \right)^2 dx \\ &\leq \sum_{j=1}^k \int_{I_{k,j}} L_0^2 |x - s_{k,j}|^{2\alpha} dx \\ &\leq AL_0^2 \sum_{j=1}^k |I_{k,j}| k^{-2\alpha} \\ &\leq AL_0^2 k^{-2\alpha}, \end{aligned}$$

where we used that $\sqrt{f_0}$ is α -Hölder continuous for the first inequality and that $|x - s_{k,j}| \leq |I_{k,j}|$ for $x \in I_{k,j}$. $\square$

The following proposition gives a condition under which the Hölder continuity of f_0 is equivalent to the Hölder continuity of $\sqrt{f_0}$.

Proposition 4.4. *Suppose $h: \mathcal{X} \rightarrow \mathbb{R}$ satisfies $m \leq h \leq M$ for some $m > 0$. Then h is α -Hölder continuous if and only if $\sqrt{h}$ is α -Hölder continuous.*

Proof. Assume that $a, b \in [m, M]$ and without loss of generality that $a < b$. Applying the mean value theorem to the map $x \mapsto \sqrt{x}$ we find for some $c \in (a, b)$,

$$|\sqrt{a} - \sqrt{b}| = \frac{1}{2\sqrt{c}}|a - b|.$$

Taking the supremum and infimum over the larger set $c \in [m, M]$, we obtain the following chain of inequalities

$$\frac{1}{2\sqrt{M}}|a - b| \leq |\sqrt{a} - \sqrt{b}| \leq \frac{1}{2\sqrt{m}}|a - b|.$$

In particular, inserting $a = h(x)$ and $b = h(y)$ for $x, y \in \mathcal{X}$ yields

$$\begin{aligned} |h(x) - h(y)| &\leq 2\sqrt{M} \left| \sqrt{h(x)} - \sqrt{h(y)} \right|, \\ \left| \sqrt{h(x)} - \sqrt{h(y)} \right| &\leq 2^{-1} m^{-1/2} |h(x) - h(y)|. \end{aligned}$$

Hence, the Hölder continuity of h and $\sqrt{h}$, are equivalent. $\square$

As a consequence of Proposition 4.4, the assumption that $\sqrt{f_0}$ is α -Hölder continuous is in fact strictly weaker than the assumption that f_0 is α -Hölder continuous and bounded away from 0. As an example of a family of distributions which satisfies the former criterion but not the latter, take $\mathcal{F}$ to be the set of all Beta(a, b) densities with $a, b \geq 3$. It is easily verified that the square root of any $f_0 \in \mathcal{F}$ is continuously differentiable on $[0, 1]$, implying Lipschitz continuity of $\sqrt{f_0}$. On the other hand, $f_0(0) = 0$ for any such f_0 , so no element in $\mathcal{F}$ is bounded away from 0.

An interesting question is whether a better approximation rate than $\mathcal{O}(k^{-2})$ is attainable with respect to the squared Hellinger metric under stronger smoothness assumptions than Lipschitz continuity on the root-density $\sqrt{f_0}$. However, the following result from Freedman and Diaconis (1981) shows that k^{-2} in fact serves as an upper limit on how well we can approximate smooth functions by histograms.

Proposition 4.5 (Proposition 2.7, (Freedman and Diaconis, 1981), slightly rewritten). *Let $\{I_{1:k}\}_{k=1}^\infty$ be the sequence of regular partitions of the unit interval and let $h: [0, 1] \rightarrow \mathbb{R}$ be such that $\sqrt{h}$ has a square integrable a.e. derivative g . Then*

$$\int_0^1 \left\{ \sqrt{h(x)} - \sqrt{h_{I_{1:k}}(x)} \right\}^2 dx = \frac{1}{12} k^{-2} \int_0^1 g^2(x) dx + o(k^{-2}) \text{ as } k \rightarrow \infty. \quad (4.57)$$

The result (4.57) shows that in the case where $(\sqrt{f_0})' \neq 0$, then the squared Hellinger distance between f_0 and its piecewise constant approximation using k bins is of order k^{-2} . Thus, the approximation rate is $\mathcal{O}(k^{-2})$ even for densities which are smoother than Lipschitz continuous ones. We note that a similar result holds for the squared $\mathbb{L}_2$ metric, as a consequence of the above proposition. Due to this fact, we restrict our subsequent analyses to the class of α -Hölder continuous functions, as there is nothing to be gained in terms of approximation rates by assuming additional smoothness beyond Lipschitz continuity.

4.6.2 Convergence rates of Bayesian histogram methods

In the following subsection, we present convergence results for regular Bayesian histogram procedures.

4.6.2a Convergence rates when the number of bins increase deterministically

Before discussing the convergence results of Scricciolo (2007), we derive the posterior convergence rate for a related Bayesian model, where we have the same Dirichlet prior distribution for $\mathbf{p} | k$, and where the number k of bins is given. In this model, we take $k = k_n$ with $\{k_n\}$ being a sequence of positive integers diverging to infinity at a given rate depending on the smoothness of the true density f_0 . This model is easier to analyze mathematically; for given $n \in \mathbb{N}$ the posterior distribution of f only depends on $k = k_n$, and as such it is determined through the posterior distribution of the k_n -simplex vector

p. From the conjugacy argument of (4.15), we have $\mathbf{p} \mid \mathbf{x}_{1:n} \sim \text{Dir}(\mathbf{p} \mid a\mathbf{p}_0 + \mathbf{N})$, so the posterior of f can be characterized in terms of a single Dirichlet distribution. This allows us to develop explicit expressions for the $\mathbb{L}_2$ risk of the Bayes estimator $\hat{f}_{B,n}$ in this model, which is given by

$$\hat{f}_{B,n}(x) = k_n \sum_{j=1}^{k_n} \hat{p}_{B,j} \mathcal{I}(x \in I_j), \quad x \in [0, 1], \quad (4.58)$$

where $\hat{p}_{B,j} = \mathbb{E}_{p_j}\{p_j \mid \mathbf{x}_{1:n}\} = a_n/(a_n + n)p_{0,j} + n/(a_n + n)\hat{p}_j$ and $\hat{p}_j = N_j/n$.

Our main motivation for studying the above setting is to see if treating k as a random variable to be estimated worsens the convergence rate compared to the case where we choose it deterministically. In the interest of generality, as the prior distribution for the $\mathcal{S}_{k_n-1}$ -valued random variable $\mathbf{p}$ we shall take the $\text{Dir}(a_n\mathbf{p}_0)$ -distribution where $\{a_n\}$ is a sequence of positive real numbers and $\mathbf{p}_0$ belongs to the interior of the simplex for any k_n . In this case, the number of bins will be chosen deterministically as a function of the sample size. A variant of the model considered here was studied by Castillo (2014), based on dyadic regular partitions of the unit interval. Castillo derived a convergence rate of $\{\log(n)/n\}^{\alpha/(2\alpha+1)}$ with respect to the supremum norm, which is the minimax rate for this class of densities. His results rely on the assumptions that f_0 is bounded away from 0 and that $\alpha > 1/2$. Additionally, Castillo (2014, Lemma 10) derives the same convergence rate with respect to the Hellinger metric for all $\alpha \in (0, 1]$, using the tools of Ghosal et al. (2000). Under a similar set of assumptions, we show that the posterior converges at the minimax rate with respect to the Hellinger metric for $\alpha > 1/2$, although we require k_n to grow at a faster rate than in Castillo's proof. For mathematical convenience, we derive our first convergence result in terms of the $\mathbb{L}_2$ metric $d_2^2(f, g) = \int_0^1 \{f(x) - g(x)\}^2 dx$, as it enables a rather direct convergence rate proof due to the conjugacy property of the Dirichlet distribution. We then use this result to also obtain the same convergence rate in the Hellinger metric d_H under some additional assumptions in Proposition 4.6.

Theorem 4.4. *Let $\epsilon_n = n^{-\alpha/(2\alpha+1)}$ and suppose that there exists two constants $C_1, C_2 > 0$ such that the sequence $\{k_n\}$ satisfies*

$$C_1 \epsilon_n^{-1/\alpha} \leq k_n \leq C_2 \epsilon_n^{-1/\alpha}. \quad (4.59)$$

Suppose further that for each n , $\mathbf{p}$ follows the distribution $\text{Dir}(a\mathbf{p}_0)$ on $\mathcal{S}_{k_n-1}$ where $\{a_n\}$ is a sequence of positive numbers satisfying $a_n = \mathcal{O}(n^{1/2})$. If f_0 is α -Hölder continuous, then the posterior distribution converges to f_0 at the rate ϵ_n with respect to d_2 in the sense of (3.41).

The result in Theorem 4.4 relates the smoothness of f_0 to the convergence rate directly, where a greater degree of smoothness corresponds to faster convergence. We also note that to obtain the optimal convergence rate, k_n has to depend on the level of smoothness α , with less smooth functions requiring a greater number of bins. This is directly related to the approximation property of (4.56), since more bins will in general be needed to reach a given approximation level for less smooth functions. In the limit as $\alpha \rightarrow 0$, we require that k_n grows at a rate approaching n , while for Lipschitz continuous f_0 the optimal rate for k_n is the familiar $n^{1/3}$ of Example 2.9. We note that unlike the convergence rate result we derive for the Hellinger metric later, Theorem 4.4 does not need the hypothesis that f_0 is bounded from below.

Proof of Theorem 4.4. Let f_{0,k_n} be the piecewise constant version of f_0 using k_n regular bins, and write $\pi_j = \int_{I_j} f_0(x) dx$. Since the number k of bins is fixed to $k = k_n$

in this case, the posterior distribution of $f(\cdot)$ can be parametrized by the k_n -simplex vector $\mathbf{p}$ so that $f(x) = f(x|k_n, \mathbf{p}) = k_n \sum_{j=1}^{k_n} p_j \mathcal{I}(x \in I_j)$. Our strategy will be to bound the non-deterministic components of the risk in expectation, and leverage Markov's inequality to deduce convergence in probability. By the triangle inequality together with the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$ (Proposition A.7), we have

$$d_2^2(f_0, f) \leq 2d_2^2(f_0, f_{0,k_n}) + 2d_2^2(f_{0,k_n}, f). \quad (4.60)$$

The first term in (4.60) can be bounded by Proposition 4.2 since $\max_j |I_{1:k_n}| = k_n^{-1}$ for regular histograms. We thus have

$$d_2^2(f_0, f_{0,k_n}) \leq d_\infty^2(f_0, f_{0,k_n}) \leq L_0^2 k_n^{-2\alpha} \leq L_0^2 C_1^{-2\alpha} \epsilon_n^2,$$

where we used that $k_n^{-2\alpha} \leq C_1^{-2\alpha} \epsilon_n^2$ due to (4.59) for the last inequality. To control the second term in (4.60), we apply the triangle inequality again to obtain

$$d_2^2(f_{0,k_n}, f) \leq 2d_2^2(f_{0,k_n}, \hat{f}_{B,n}) + 2d_2^2(\hat{f}_{B,n}, f), \quad (4.61)$$

where $\hat{f}_{B,n}$ is the Bayes estimator (4.58). To bound the two terms in (4.61), we use Lemma A.3 to deduce that

$$\begin{aligned} \mathbb{E}_{\mathbf{p}} \left\{ d_2^2(\hat{f}_{B,n}, f) \mid \mathbf{x}_{1:n} \right\} &\leq C_2 \epsilon_n^2, \\ \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_2^2(f_{0,k_n}, \hat{f}_{B,n}) \right\} &\leq (4a_n^2 n^{-1} + 1) \epsilon_n^2. \end{aligned}$$

We will now show that the posterior converges at the desired rate in the sense of Definition 3.6. Let $\{M_n\}$ be a sequence of numbers diverging to positive infinity. Applying Markov's inequality we have that

$$\begin{aligned} A_n &= P_n(d_2(f_0, f) \geq M_n \epsilon_n \mid \mathbf{x}_{1:n}) \\ &\leq M_n^{-2} \epsilon_n^{-2} \mathbb{E}_{\mathbf{p}} \left\{ d_2^2(f_0, f) \mid \mathbf{x}_{1:n} \right\} \\ &\leq 2M_n^{-2} \epsilon_n^{-2} \left(d_2^2(f_0, f_{0,k_n}) + 2d_2^2(f_{0,k_n}, \hat{f}_{B,n}) + 2\mathbb{E}_{\mathbf{p}} \left\{ d_2^2(\hat{f}_{B,n}, f) \mid \mathbf{x}_{1:n} \right\} \right) \\ &\leq 2M_n^{-2} \left\{ L_0 C_1^{-2\alpha} + 2C_2 + 2\epsilon_n^{-2} d_2^2(f_{0,k_n}, \hat{f}_{B,n}) \right\}. \end{aligned}$$

Note that $a_n^2 n^{-1}$ is bonded by a positive constant for large n due to the assumption on the sequence $\{a_n\}$. Taking expectation over $\mathbf{x}_{1:n}$, and using Markov's inequality again, we have for any $\delta > 0$,

$$\begin{aligned} \mathbb{P}(A_n \geq \delta) &\leq \delta^{-1} \mathbb{E}_{\mathbf{x}_{1:n}} \{A_n\} \\ &\leq 2\delta^{-1} M_n^{-2} \left(L_0 C_1^{-2\alpha} + 2C_2 + 2\epsilon_n^{-2} \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_2^2(f_{0,k_n}, \hat{f}_{B,n}) \right\} \right) \\ &\leq 2\delta^{-1} M_n^{-2} \left(L_0 C_1^{-2\alpha} + 2C_2 + 2(4a_n^2 n^{-1} + 1) \right). \end{aligned}$$

The right hand side converges to zero by assumption, showing that $A_n \xrightarrow{\mathbb{P}} 0$. □

Remark 4.4. The proof of Theorem 4.4 also shows that the Bayes estimator $\hat{f}_{B,n}$ attains the convergence rate ϵ_n in the $\mathbb{L}_2$ metric, since both $\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_2^2(f_{0,k_n}, \hat{f}_{B,n}) \right\}$ and $d_2^2(f_0, f_{0,k_n})$ are $\mathcal{O}(\epsilon_n^2)$. This fact, together with the triangle inequality and $(a+b)^2 \leq 2a^2 + 2b^2$, establishes $\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_2^2(f_0, \hat{f}_{B,n}) \right\} = \mathcal{O}(\epsilon_n^2)$.

The result of Theorem 4.4 gives the convergence rate in terms of the $\mathbb{L}_2$ distance, making it difficult to compare it directly to the results of Scricciolo (2007), as these concern the Hellinger metric. However, it turns out that we can leverage the result of Theorem 4.4 to obtain the same convergence rate for the Hellinger distance as well in a more restrictive but similar setting:

Proposition 4.6. *Suppose that the assumptions of Theorem 4.4 are met. If in addition f_0 is bounded away from 0 and $\alpha > 1/2$, then the posterior distribution converges to f_0 at the rate $\epsilon_n = n^{-\alpha/(2\alpha+1)}$ in the sense of (3.41) in the Hellinger metric as well.*

The rate obtained here avoids the logarithmic factors in Lemma 10 of Castillo (2014), and offers an improved convergence rate in the case where $\alpha > 1/2$. We note that the result of Proposition 4.6 is similar to that of Ghosal et al. (2000, Theorem 4.5), who obtain the same rate of convergence in the Hellinger metric for log-spline models, containing regular histograms as a special case. However, they use a prior constructed on parameters on the log-scale, and not on the p_j directly as we do.

Proof of Proposition 4.6. We will use the fact that we can bound the Hellinger metric d_H in terms of d_2 when the densities involved are lower bounded. To start, let $\{M_n\}$ be a sequence of real numbers diverging to positive infinity and denote $m = \inf_{x \in [0,1]} f_0(x)$. We note that by the second part of Proposition A.3, if $f(x) \geq m/2$ for all x , then $d_H^2(f_0, f) \leq m^{-1}d_2^2(f_0, f)$. From the definition of f , we have that $f(x) \geq m/2$ if and only if $\min_j p_j \geq k_n^{-1}m/2$. It follows that

$$\begin{aligned} P_n(d_H(f_0, f) \geq M_n \epsilon_n \mid \mathbf{x}_{1:n}) \\ \leq P_n(d_2(f_0, f) \geq m^{-1/2} M_n \epsilon_n \mid \mathbf{x}_{1:n}) + P_n(\min_j p_j < k_n^{-1}m/2 \mid \mathbf{x}_{1:n}). \end{aligned} \quad (4.62)$$

The first term on the right hand side can be handled using Theorem 4.4, so we focus on bounding the second term. Note that as $f_0(x) \geq m$ for all $x \in [0, 1]$, we have that $\pi_j = \int_{I_j} f_0(x) dx \geq m k_n^{-1}$. Hence, if $p_j < m k_n^{-1}/2$, then $\pi_j - p_j > m k_n^{-1}/2$, so the following inclusion holds:

$$\{p_j : p_j < k_n^{-1}m/2\} \subseteq \{p_j : |p_j - \pi_j| \geq k_n^{-1}m/2\}.$$

A simple union bound now yields

$$\begin{aligned} P_n(\min_j p_j < k_n^{-1}m/2 \mid \mathbf{x}_{1:n}) &= P_n\left(\bigcup_{j=1}^{k_n} \{p_j < k_n^{-1}m/2\} \mid \mathbf{x}_{1:n}\right) \\ &\leq \sum_{j=1}^{k_n} P_n(p_j < k_n^{-1}m/2 \mid \mathbf{x}_{1:n}) \\ &\leq \sum_{j=1}^{k_n} P_n(|p_j - \pi_j| \geq k_n^{-1}m/2 \mid \mathbf{x}_{1:n}) \\ &\leq 4k_n^2 m^{-2} \sum_{j=1}^{k_n} \mathbb{E}_{p_j}\{(p_j - \pi_j)^2 \mid \mathbf{x}_{1:n}\}, \end{aligned} \quad (4.63)$$

where we used Markov's inequality for the last line. To further bound (4.63) we note that by Proposition A.7,

$$\mathbb{E}_{p_j}\{(p_j - \pi_j)^2 \mid \mathbf{x}_{1:n}\} \leq 2\mathbb{E}_{p_j}\{(p_j - \hat{p}_{B,j})^2 \mid \mathbf{x}_{1:n}\} + 2(\hat{p}_{B,j} - \pi_j)^2.$$

The first term on the right hand side is the posterior variance, which by (A.24) satisfies

$$\text{Var}_{p_j}(p_j | \mathbf{x}_{1:n}) \leq \hat{p}_{B,j} n^{-1}.$$

The second term can be controlled in expectation. Using a similar bias-variance decomposition to (A.25), we have that

$$\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ (\hat{p}_{B,j} - \pi_j)^2 \right\} \leq 2a^2 n^{-2} (p_{0,j} + \pi_j) + n^{-1} \pi_j.$$

Plugging these estimates into (4.63), we have

$$\begin{aligned} P_n(\min_j p_j < k_n^{-1} m/2 | \mathbf{x}_{1:n}) &\leq 8k_n^2 m^{-2} \sum_{j=1}^{k_n} \left\{ \hat{p}_{B,j} n^{-1} + (\hat{p}_{B,j} - \pi_j)^2 \right\} \\ &\leq 8k_n^2 m^{-2} \left(n^{-1} + \sum_{j=1}^{k_n} (\hat{p}_{B,j} - \pi_j)^2 \right). \end{aligned}$$

Taking the expectation with respect to $\mathbf{x}_{1:n}$ on both sides then yields

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ P_n(\min_j p_j < k_n^{-1} m/2 | \mathbf{x}_{1:n}) \right\} \\ &\leq 8k_n^2 m^{-2} \left(n^{-1} + \sum_{j=1}^{k_n} \left\{ 2a_n^2 n^{-2} (p_{0,j} + \pi_j) + n^{-1} \pi_j \right\} \right) \\ &\leq 16k_n^2 m^{-2} n^{-1} (2 + 2a_n^{-2} n^{-1}). \end{aligned} \tag{4.64}$$

Since $k_n^2 \leq C_2^2 n^{2/(2\alpha+1)}$, we have that $k_n^2 n^{-1} = C_2^2 n^{(1-2\alpha)/(2\alpha+1)}$, which goes to zero as $n \rightarrow \infty$ provided $\alpha > 1/2$. Hence, the right hand side of (4.64) also converges to zero. By Markov's inequality this implies that $P_n(\min_j p_j < k_n^{-1} m/2 | \mathbf{x}_{1:n}) = o_{\mathbb{P}}(1)$ as well. By Theorem 4.4, the first term on the right hand side of (4.62) also converges to 0 in probability. Thus, $P_n(d_H(f_0, f) \geq M_n \epsilon_n | \mathbf{x}_{1:n}) \xrightarrow{\mathbb{P}} 0$, and we conclude that the convergence rate with respect to the Hellinger metric is ϵ_n . $\square$

4.6.2b Posterior convergence rate of fully Bayesian regular histograms

We now give convergence rate results for a fully Bayesian regular histogram prior given by the hierarchical scheme (4.8), with Dirichlet prior distributions on the bin probabilities $\mathbf{p}$. Theorem 4.5 gives two convergence rate depending on the choice of some hyperparameters in the prior distributions. Before we state the theorem, we state the necessary assumptions on the prior distribution.

Assumption 4.3. For the irregular histogram prior sequence $\{P_n\}$, the following holds:

- (i) The sequence of prior probability mass functions for k , $p_n(k)$, assigns positive mass to all $k \in \{1, 2, \dots, k_n\}$, where k_n is either a sequence diverging to positive infinity such that $k_n = \mathcal{O}(n^\gamma)$ for all $\gamma < 1$, or $k_n = \infty$ for all sufficiently large n . Moreover, there are positive constants D_1, d_1, D_2, d_2 such that for all $k \leq k_n$,

$$D_1 \exp(-d_1 k \log(k)) \leq p_n(k) \leq D_2 \exp(-d_2 k \log(k)). \tag{4.65}$$

- (ii) Given k and a partition $I_{1:k}$, the prior for $\mathbf{p} | k, I_{1:k}$ is $\text{Dir}(\mathbf{a}_k)$, where $\sigma k^{-1} \leq a_{k,j} \leq \Sigma$ for two positive constants σ, Σ .

The conditions here are somewhat stronger than those of Assumption 4.1. The requirement (ii) enables the use of Lemma B.3, which is needed to establish that the prior assigns sufficient probability to shrinking neighborhoods of the true density. Note that this requirement is covered in two important cases: if $a_{k,j} = a \int_{I_{k,j}} g_0(x) dx$ for some fixed $a > 0$ and g_0 is a strictly positive probability density, then (ii) holds. On the other hand, if $a_{k,j} = b$ for all k, j and some fixed $b > 0$ the requirement is also satisfied. The condition (5.42) on the prior for k is satisfied if k follows a $\text{Poisson}(\lambda)$ distribution. In this case we have $p(k) \propto \lambda^k/k!$, and Stirling's approximation to the factorial then shows that the logarithm of this expression is dominated asymptotically by the term $-k \log(k)$. This argument also holds true if we take $p_n(k)$ to be a Poisson pmf truncated to the set $\{1, 2, \dots, k_n\}$.

The main result of the section is given in Theorem 4.5, establishing convergence rates for the regular random histogram prior. The proof of the theorem follows the basic strategy outlined in the proof of Theorem 2 in Scricciolo (2007). Since we will be exposed to most of these arguments in the proof of Theorem 5.4 later, we omit the proof of Theorem 4.5.

Theorem 4.5. *Assume that Assumption 4.3 holds and that $f_0: [0, 1] \rightarrow \mathbb{R}$ is a probability density whose square root is α -Hölder continuous. Then the posterior distribution converges at the rate*

$$\epsilon_n = n^{-\alpha/(2\alpha+1)} \log(n)^{(\alpha+1)/(2\alpha+1)}. \quad (4.66)$$

Since the minimax rate of estimation for the class of Hölder continuous densities in the Hellinger metric is $n^{-\alpha/(2\alpha+1)}$, the regular random histogram prior attains the optimal convergence rate up to logarithmic factors. Interestingly, the rate (4.66) is attained without using a prior distribution depending on α , so the histogram prior is able to adapt to f_0 without knowledge of the level of smoothness of the true density. We note that the rate is a bit slower than the $n^{-\alpha/(2\alpha+1)} \log(n)^{(2\alpha+1/2)/(2\alpha+1)}$ rate obtained by Scricciolo (2007) for small $\alpha > 0$, which is mainly due to the fact that we do not make the assumption that f_0 is bounded away from 0. Compared to the case where k_n is chosen to grow at the optimal rate in a deterministic fashion as in Theorem 4.4, the results of Theorem 4.5 suggest that the loss in convergence rate is rather small.

4.7 Dirichlet-based Histograms

In this section we further explore the choice of prior distributions in the random histogram model. For the bin probabilities $\mathbf{p}|k$, we will only consider Dirichlet distributions for all values of $k \in \mathbb{N}$, so that $\mathbf{p}|k \sim \text{Dir}(a_k \mathbf{p}_0)$ for a parameter $a_k > 0$ possibly depending on k and $\mathbf{p}_0 \in \mathcal{S}_{k-1}$. Although one may in principle use another family of simplex priors for the bin probabilities, computing the Bayes histogram estimator (4.14) in such models becomes difficult due to the marginal likelihood (4.10) not being available in closed form. One can resort to various approximate techniques, but these will generally be slower and/or require significant tuning in the fitting procedure, as is the case for MCMC. This is completely at odds with our stated goal of designing a histogram procedure with quick inference and minimal input required from the user.

4.7.1 Choice of prior for the bin probabilities

One appealing property of the random histogram model with $\mathbf{p}|k \sim \text{Dir}(a_k \mathbf{p}_0)$ is the ease at which the prior distribution can be centered at a given prior guess of the true

density g_0 . Centering the prior at g_0 can be accomplished by setting the prior mean $\mathbf{p}_0$ of $\mathbf{p} | k$ to, for each $k \in \mathbb{N}$,

$$p_{0,j} = \int_{I_j} g_0(x) dx, \text{ for } j = 1, 2, \dots, k. \quad (4.67)$$

This choice results in the density f being centered on g_0 in the sense that for given k , the prior mean of f is $g_{0,I_{1:k}}$, the piecewise constant version of g_0 based on the regular partition $I_{1:k}$. Indeed, since the prior mean of p_j is $p_{0,j}$, we have

$$\begin{aligned} \mathbb{E}_f\{f(x) | k\} &= \mathbb{E}_{\mathbf{p}} \left\{ k \sum_{j=1}^k p_j \mathcal{I}(x \in I_{k,j}) \mid k \right\} \\ &= k \sum_{j=1}^k \mathbb{E}_{\mathbf{p}}\{p_j | k\} \mathcal{I}(x \in I_{k,j}) \\ &= k \sum_{j=1}^k p_{0,j} \mathcal{I}(x \in I_{k,j}) \\ &= \sum_{j=1}^k k \int_{I_{k,j}} g_0(s) ds \mathcal{I}(x \in I_{k,j}) \\ &= g_{0,I_{1:k}}(x). \end{aligned}$$

In practice, g_0 should be chosen such that the set $\{x \in [0, 1] : g_0(x) > 0\}$ is of full Lebesgue measure, as otherwise the prior is not a proper probability distribution, with no guarantee of a proper posterior.

In situations where no prior information is available, it is desirable to have a non-informative ‘default’ choice for $\mathbf{p}_0$. A symmetric Dirichlet distribution presents itself as a natural choice as prior for the bin probabilities in this setting. Indeed, when prior information is absent, it would seem natural to require that the prior distribution for $\mathbf{p} | k$ is invariant to permutations of the cell probabilities, as we have no reason to think that the probability of a particular observation landing in, say, the interval I_j is different from the probability of landing in I_l , since both intervals are of equal length. This permutation invariance is satisfied by the $\text{Dir}(a_k/k, a_k/k, \dots, a_k/k)$ density for $a_k > 0$, since it is proportional to $\prod_{j=1}^k p_j^{a_k/k-1}$, which is invariant to permutations of the indices $j = 1, 2, \dots, k$. Another interpretation of the symmetric Dirichlet prior is that we center the prior $\mathbf{p} | k$ on the uniform distribution in the sense of (4.67), since the prior means of $p_j | k$ are all $1/k$ in this model.

4.7.1a Noninformative choices of the concentration parameter

While it is easy to find a noninformative choice for the prior mean $\mathbf{p}_0$, noninformative choices for the Dirichlet concentration parameter a are less obvious. At first, it might seem natural to choose a to be a very small value. From Proposition 3.1, the prior variance of p_j is given by $\text{Var}_{p_j}(p_j) = p_{0,j}(1 - p_{0,j})/(a + 1)$. Letting a tend to zero yields greater prior variance, expressing greater prior uncertainty about the parameter. Moreover, choosing a small value of a yields point estimates that give more weight to the data, since the posterior means of the cell probabilities from (4.19) are given by $\hat{p}_{B,j} = \lambda_a p_{0,j} + (1 - \lambda_a) N_j/n$, where $\lambda_a = a/(a + n)$. When a approaches 0, so does λ_a , and the resulting estimate assigns more weight to the term based on the data. However, we note that although greater prior variance is generally desirable, this may

lead to a prior distribution which has some rather undesirable properties when it comes to the behavior of the posterior distribution for k . As a decreases, it is evident that the prior distribution for $\mathbf{p}$ becomes increasingly concentrated near the edges of the simplex, assigning nearly all of its mass to those $\mathbf{p}$ for which one component is close to 1 and the rest are near zero. We will now formalize this notion mathematically. Let $\mathcal{A} \subseteq \mathcal{S}_{k-1}$ be a set bounded away from the boundary of $\mathcal{S}_{k-1}$, that is, there exists a $\delta > 0$ such that for any $\mathbf{p} \in \mathcal{A}$, we have $p_j > \delta$ for all j . Proposition 4.7 shows that the prior probability of $\mathbf{p} \in \mathcal{A}$ tends to zero as $a \rightarrow 0$. The proof can be found in Appendix A.1.

Proposition 4.7. *Let $\mathbf{p} \sim \text{Dir}(a\mathbf{p}_0)$ be k -dimensional Dirichlet distribution, and let $\mathcal{A} \subseteq \mathcal{S}_{k-1}$ be a set bounded away from the boundary of $\mathcal{S}_{k-1}$. Then*

$$\lim_{a \rightarrow 0} P(\mathbf{p} \in \mathcal{A} | k) = 0.$$

The Dirichlet concentration parameter a has a significant impact on the posterior distribution of k . We recall from Example 4.2 that the posterior for k under this prior is given by

$$p_n(k | \mathbf{x}_{1:n}) \propto p_n(k) k^n \frac{\prod_{j=1}^k \Gamma(ap_{0,j} + N_j)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \frac{\Gamma(a)}{\Gamma(a+n)}. \quad (4.68)$$

For simplicity, we will analyze the behavior of the posterior distribution when a is the same across different values of k . The following Proposition shows that as $a \rightarrow 0$, the posterior $p_n(k | \mathbf{x}_{1:n})$ will concentrate on those k for which there is only one nonempty bin.

Proposition 4.8. *Suppose that $a \rightarrow 0$. Then $p_n(k | \mathbf{x}_{1:n}) \rightarrow 0$ at any k for which more than one bin is nonempty.*

Proof. Write $f(x) \asymp g(x)$ if $f(x)/g(x) \rightarrow 1$ as $x \rightarrow 0$, and note that $\Gamma(x) \asymp 1/x$. Fix $k \in \mathbb{N}$. Since the posterior pmf $p_n(k | \mathbf{x}_{1:n})$ of (4.68) is invariant to joint permutations of the components of $\mathbf{N}$ and $\mathbf{p}_0$, we may assume that intervals $I_1, \dots, I_{m(k)}$ are the nonempty ones. This implies that as $a \rightarrow 0$,

$$\frac{\prod_{j=1}^k \Gamma(ap_{0,j} + N_j)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \asymp \frac{\prod_{j=1}^{m(k)} \Gamma(N_j)}{\prod_{j=1}^{m(k)} 1/(ap_{0,j})} = a^m \prod_{j=1}^{m(k)} \{p_{0,j} \Gamma(N_j)\}, \quad (4.69)$$

where we also used the continuity of $\Gamma(\cdot)$. Since a is the same for all values of k , we ignore the factor $\Gamma(a)/\Gamma(a+n)$ in (4.68). Write

$$h(k, a) = p_n(k) k^n \frac{\prod_{j=1}^k \Gamma(ap_{0,j} + N_j)}{\prod_{j=1}^k \Gamma(ap_{0,j})}.$$

From this it follows that $p_n(k | \mathbf{x}_{1:n}) \propto h(k, a)$ and by (4.69),

$$h(k, a) \asymp p_n(k) k^n a^{m(k)} \prod_{j=1}^{m(k)} \{p_{0,j} \Gamma(N_j)\}, \text{ as } a \rightarrow 0.$$

Now let k be such that at least two bins are nonempty, i.e. $m(k) \geq 2$. Now, for $k = 1$ we have $m(1) = 1$ since all the observations will necessarily fall into the one bin in this case. From this, we observe that

$$\begin{aligned}
p_n(k | \mathbf{x}_{1:n}) &= \frac{h(k, a)}{\sum_{\kappa=1}^{k_n} h(\kappa, a)} \leq \frac{h(k, a)}{h(1, a)} \asymp \frac{p_n(k) k^n a^{m(k)} \prod_{j=1}^{m(k)} \{p_{0,j} \Gamma(N_j)\}}{p_n(1) a \Gamma(n)} \\
&= a^{m(k)-1} \frac{p_n(k) \prod_{j=1}^{m(k)} p_{0,j} \Gamma(N_j)}{p_n(1) \Gamma(n)}.
\end{aligned} \tag{4.70}$$

Since $m(k) \geq 2$ by assumption, $a^{m(k)-1} \rightarrow 0$ as $a \rightarrow 0$, and hence $\lim_{a \rightarrow 0} p_n(k | \mathbf{x}_{1:n}) = 0$ by squeezing. $\square$

The result of Proposition 4.8 is similar to that obtained by Petrone (1999). Petrone studied a hierarchical Dirichlet mixture of Beta distributions, and found that as the concentration parameter of the Dirichlet distribution approaches zero, then the posterior will concentrate on a single Beta density. In both our case and Petrone's case, the posterior concentrates on those k for which a single mixture component dominates.

It is also interesting to study the limiting behavior of the posterior distribution at the other extreme, namely as $a \rightarrow \infty$. It turns out that in the case where the prior mean of the Dirichlet distribution is uniform, we get a particularly simple limit result showing that the posterior for k approaches the prior distribution as $a \rightarrow \infty$.

Proposition 4.9. *The posterior pmf of k satisfies*

$$\log p_n(k | \mathbf{x}_{1:n}) = n \log(k) + \sum_{j=1}^k N_j \log(p_{0,j}) - \log(Z_a) + \mathcal{O}(a^{-1}) \text{ as } a \rightarrow \infty,$$

where $Z_a = \sum_{k=1}^{k_n} p_n(k | \mathbf{x}_{1:n})$ is the normalization constant. Additionally, if the prior mean $\mathbf{p}_0 \in \mathcal{S}_{k-1}$ of the Dirichlet distributions of $\mathbf{p} | k$ are all uniform, then $p_n(k | \mathbf{x}_{1:n}) \rightarrow p_n(k)$ for each k .

Proof. For simplicity, we opt to work on the log-scale. Stirling's approximation to the loggamma function then yields

$$\begin{aligned}
\log \Gamma(ap_{0,j} + N_j) &= (ap_{0,j} + N_j - \frac{1}{2}) \log(ap_{0,j} + N_j) - ap_{0,j} - N_j + \frac{1}{2} \log(2\pi) + \mathcal{O}(a^{-1}), \\
\log \Gamma(ap_{0,j}) &= (ap_{0,j} - \frac{1}{2}) \log(ap_{0,j}) - ap_{0,j} + \frac{1}{2} \log(2\pi) + \mathcal{O}(a^{-1}), \\
\log \Gamma(a + n) &= (a + n - \frac{1}{2}) \log(a + n) - a - n + \frac{1}{2} \log(2\pi) + \mathcal{O}(a^{-1}), \\
\log \Gamma(a) &= (a - \frac{1}{2}) \log(a) - a + \frac{1}{2} \log(2\pi) + \mathcal{O}(a^{-1}).
\end{aligned}$$

To get a tractable asymptotic expansion for the log-posterior of k , apply the above approximations to $\sum_{j=1}^k \{\log \Gamma(ap_{0,j} + N_j) - \log \Gamma(ap_{0,j})\} + \log \Gamma(a) - \log \Gamma(a + n)$, yielding

$$\sum_{j=1}^k \left\{ (ap_{0,j} - \frac{1}{2}) \log(1 + N_j/[ap_{0,j}]) + N_j \log(ap_{0,j} + N_j) \right\} - (a - \frac{1}{2}) \log(1 + n/a) + \mathcal{O}(a^{-1}).$$

Using the fact that $\log(1 + x) = x + \mathcal{O}(x^2)$ as $x \rightarrow 0$, the above expression is equal to the following:

$$\begin{aligned}
&\sum_{j=1}^k \left\{ ap_{0,j} \frac{N_j}{ap_{0,j}} + N_j (\log(ap_{0,j}) + \log(1 + N_j/[ap_{0,j}])) \right\} - a \frac{n}{a} + \mathcal{O}(a^{-1}) \\
&= \sum_{j=1}^k N_j \log(ap_{0,j}) + \mathcal{O}(a^{-1}) = n \log(a) + \sum_{j=1}^k N_j \log(p_{0,j}) + \mathcal{O}(a^{-1}).
\end{aligned}$$

Since the term $n \log(a)$ is common to all values of k , the log-posterior for k (4.68) permits the asymptotic expansion

$$\log p_n(k | \mathbf{x}_{1:n}) = \log p_n(k) + n \log(k) + \sum_{j=1}^k N_j \log(p_{0,j}) - \log(Z_a) + \mathcal{O}(a^{-1}), \quad (4.71)$$

proving the first part of the claim. To prove the second part, assume that $p_{0,j} = k^{-1}$ for all k and j . In this case, $\sum_{j=1}^k N_j \log(p_{0,j}) = -n \log(k)$, which together with (4.71) shows that the limit of $p_n(k | \mathbf{x}_{1:n})$ is $p_n(k)$ as $a \rightarrow \infty$, where we used that

$$\lim_{a \rightarrow \infty} Z_a = \lim_{a \rightarrow \infty} \sum_{k=1}^{k_n} \exp \left\{ \log p_n(k) + \mathcal{O}(a^{-1}) \right\} = \sum_{k=1}^{k_n} p_n(k) = 1,$$

concluding the proof. $\square$

In the uniform case, the result of (4.9) shows that as $a \rightarrow \infty$, the posterior for k becomes independent of the data. The results shown here show directly that choosing a value of a too large yields a posterior distribution with undesirable properties.

We will now turn our attention to Dirichlet priors that have previously been proposed in the literature as noninformative prior distributions in the context of Bayesian histogram density estimation. The first is the uniform prior, $p(\mathbf{p} | k) \propto 1$, which corresponds to a symmetric $\text{Dir}(a\mathbf{p}_0)$ distribution with $a = k$. Although the choice of the uniform prior has been criticized as actually being informative and leading to poor performance in certain estimation problems (Berger et al., 2015), it has been defended as being an objective choice for multinomial estimation problems based on arguments pertaining to the symmetry of the resulting posterior predictive distribution (Gerlach et al., 2009). Uniform simplex priors have been used to estimate histograms by Hall and Hannan (1988), who choose the number of bins by maximizing the posterior probability (4.18). It should be noted that the approach here is not fully Bayesian, as they use maximum likelihood estimates $\hat{p}_j = N_j/n$ for the bin probabilities, which differs from the Bayes estimates under a uniform prior.

The second noninformative prior distribution is Jeffreys' prior (Jeffreys, 1946), which is defined by the relation

$$p(\mathbf{p} | k) \propto \left| \det \mathbf{I}(\mathbf{p} | k) \right|^{1/2},$$

where $\mathbf{I}(\mathbf{p} | k) \in \mathbb{R}^{k \times k}$ is the *Fisher information matrix*, which has elements given in terms of the log-likelihood of a single sample $x \sim f$,

$$[\mathbf{I}(\mathbf{p} | k)]_{lj} = -\mathbb{E}_x \left[\frac{\partial^2}{\partial p_l \partial p_j} \log f(x | k, \mathbf{p}) \right].$$

The Jeffreys prior has the desirable property of being invariant to one-to-one transformations of the model parameters $\boldsymbol{\gamma} = h(\mathbf{p})$, in that for any measurable set $A \subseteq \mathcal{S}_{k-1}$,

$$\int_A p(\mathbf{p} | \mathbf{x}) d\mathbf{p} = \int_{h(A)} p(\boldsymbol{\gamma} | \mathbf{x}) d\boldsymbol{\gamma}.$$

This invariance property means that under Jeffreys' prior, the choice of parametrization of the likelihood function does not affect the resulting posterior inference.

With the likelihood given by (2.20), straightforward calculations (see Proposition A.8) show that the Jeffreys' prior takes the form $p(\mathbf{p} | k) \propto \prod_{j=1}^k p_j^{-1/2}$, which we recognize as an unnormalized symmetric Dirichlet distribution with $a = k/2$. We note that Jeffreys' prior was the choice of prior for the random histograms of Knuth (2019).

4.7.1b A basic coherence property

While both prior distributions covered so far has a increasing linearly with k , we argue that keeping a as a fixed number for different values of k , possibly depending on n , yields a prior better suited for the goal of estimating a histogram, as this approach satisfies a basic coherence property which does not hold for the two aforementioned priors. Let k and $m \geq 2$ be positive integers. In this case, the regular partition of $[0, 1]$ using mk bins, $I_{1:mk}$, is a refinement of the partition $\bar{I}_{1:k}$ based on k bins. In particular, we have $\bar{I}_j = \cup_{l=1}^m I_{m(j-1)+l}$ for each j . Let $\mathbf{p}_{mk} \in \mathcal{S}_{mk-1}$ denote the interval probabilities for the partition $I_{1:mk}$ and write $\bar{\mathbf{p}}_k \in \mathcal{S}_{k-1}$ for the bin probabilities based on $\bar{I}_{1:k}$. From the above definitions, it follows that

$$\bar{p}_j = \int_{\bar{I}_j} f(x) dx = \sum_{l=1}^m \int_{I_{m(j-1)+l}} f(x) dx = \sum_{l=1}^m p_{m(j-1)+l}, \quad \text{for } j = 1, 2, \dots, k. \quad (4.72)$$

Now, due to the above equality, it would be somewhat strange if $\bar{p}_j$ and $\sum_{l=1}^m p_{m(j-1)+l}$ had different prior distributions since they both represent the probability that a given sample x_i falls into the interval $\bar{I}_j$. Let now $\bar{\mathbf{p}}_k \sim \text{Dir}(\mathbf{a})$ and $\mathbf{p}_{km} \sim \text{Dir}(\mathbf{b})$, where $\mathbf{a} \in [0, \infty)^k$ and $\mathbf{b} \in [0, \infty)^{km}$. Write $a_+ = \sum_{j=1}^k a_j$, $b_{j,+} = \sum_{l=1}^m b_{m(j-1)+l}$ and $b_+ = \sum_{j=1}^k b_{j,+}$ for the sums. Using property (iv) of Proposition 3.1 together with (3.4), we observe that under the model based on the partition $I_{1:km}$, we have that

$$\sum_{l=1}^m p_{m(j-1)+l} \sim \text{Beta}(b_{j,+}, b_+ - b_{j,+}).$$

On the other hand, in the model based on the partition $I_{1:k}$ we have

$$\bar{p}_j \sim \text{Beta}(a_j, a_+ - a_j).$$

In order for these two distributions to be equal for each j , we need $b_{j,+} = a_j$ to hold. Let us now take $\mathbf{p} | k \sim \text{Dir}(a\mathbf{p}_0)$ where $a > 0$ does not depend on k and $\mathbf{p}_0 \in \mathcal{S}_{k-1}$ where $p_{0,j} = \int_{I_j} g(s) ds$ for a fixed density $g(\cdot)$ on $[0, 1]$. Then the aforementioned coherence property holds. Indeed,

$$b_{j,+} = \sum_{l=1}^m a \int_{I_{m(j-1)+l}} g(x) dx = a \int_{I_j} g(x) dx = ap_{0,j} = a_j.$$

If we use the Jeffreys prior instead, $\mathbf{p} | k \sim \text{Dir}(1/2, \dots, 1/2)$ for all k , then

$$b_{j,+} = \sum_{l=1}^m \frac{1}{2} = \frac{m}{2} = ma_j \neq a_j.$$

A similar result can be shown for the uniform prior. While both the uniform and Jeffreys prior have a claim to being noninformative, the models induced by choosing these two priors result in incoherent prior distributions.

The discussion above yields no clear choice as to what the best choice for a should be. Smaller values of a leads to Bayes estimates $\hat{p}_{B,j}$ closer to the maximum likelihood estimates $\hat{p}_j = N_j/n$, but as we saw in Proposition 4.8, choosing a too small can have detrimental effects on the number of bins chosen. Berger et al. (2015) recommend the choice $a = 1$ as being uninformative in a Dirichlet–Multinomial estimation problem. However, it is not necessarily clear that this choice of prior will work well in the case of histogram estimation due to the influence of a on the posterior distribution of k .

Another possibility is to let a depend on the sample size n , so that $\mathbf{p} | k \sim \text{Dir}(a_n \mathbf{p}_0)$ for a sequence of positive real numbers $\{a_n\}$. It might seem a bit odd at first to introduce sample-size dependence in the prior for $\mathbf{p} | k$ here, but it turns out that there are theoretical reasons as to why one might prefer to use such a prior. In the case of estimating multinomial parameters under squared $\mathbb{L}_2$ loss (see Example 3.4), the choice $a_n = \sqrt{n}$ and $\mathbf{p}_0 = (k^{-1}, \dots, k^{-1})$ yields the so-called *minimax* estimator (Lehmann and Casella, 2006, Example 4.5). The name derives from the fact that the estimator (3.27) for $a_n = \sqrt{n}$ minimizes the worst-case risk, namely

$$\sup_{\mathbf{p} \in \mathcal{S}_{k-1}} R_n(\mathbf{p}, \hat{\mathbf{p}}_B) = \inf_{\delta} \sup_{\mathbf{p} \in \mathcal{S}_{k-1}} R_n(\mathbf{p}, \delta).$$

Although the limit result of Proposition 4.9 showed that $a \rightarrow \infty$ yielded a posterior with undesirable properties, this result was obtained under the assumption that the sample size stays fixed. As such, the implied divergence of a to infinity under the minimax prior need not be problematic in itself, as a grows at a slower rate than n . We note that the case where a_n diverges to ∞ with the sample size is not covered by our consistency and rate theorems, so this choice does not come with the same theoretical guarantees as the Jeffreys and uniform priors. Unlike the other choices of prior distributions covered in this subsection, the minimax rule does not claim to be uninformative but can be seen as a conservative choice, as it attempts to minimize the worst-case risk for given k . As k is variable, the choice $a_n = \sqrt{n}$ will not necessarily yield the overall minimax Bayes histogram, since the choice of a also has an effect of the chosen value of k , and the chosen partition typically has a larger effect on the loss of the resulting histogram estimator, than small changes to the bin probability estimates $\hat{p}_{B,j}$.

4.7.2 Choice of prior for the number of bins

For the number k of bins we argue that using a sample-size dependent sequence of prior distributions $p_n(k)$, is well-suited for practical estimation problems, for reasons partially laid out in Remark 3.1. In particular, we advocate that the support of the prior $p_n(k)$ should be truncated to the set $\{1, 2, \dots, k_n\}$ for each finite n , where k_n is a sequence of integers satisfying $n^{-1}k_n \rightarrow 0$ and $k_n \rightarrow \infty$. The reason for this is two-fold. Most importantly, if k is large relative to n then there will be k probability parameters to estimate with relatively little data. In this case, the model will be prone to overfitting, producing histograms that fail to resemble the true density at all. The other reason for using finitely supported priors is computational in nature, since maximization of the posterior (4.18) now becomes feasible as we only have a finite number of possible models to consider.

Among the possible priors supported on $\{1, \dots, k_n\}$, the discrete uniform prior may seem like a reasonable, noninformative choice. However, when the models under consideration are of varying dimension, with some models being nested, one could argue in favor of using a prior which penalizes more complex models (Claeskens and Hjort, 2008, pp. 218), with one possibility being the proper prior $p_n(k) \propto 1/k$ on $\{1, \dots, k_n\}$. Both the papers of Hall and Hannan (1988) and Knuth (2019) suggest to use uniform priors on k , on the basis that this choice is noninformative. The work of Hall & Hannan is of a more theoretical nature, and no exact advice is given for choosing the support of the uniform prior. However, their theoretical results suggest that k_n should increase at a rate slower than n . Knuth suggests to take k_n equal to the integer closest to the sample range $x_{(n)} - x_{(1)}$ divided by the smallest distance between two non-identical data points. This choice inevitably leads to $k_n \geq n$. This choice does not seem desirable

from a frequentist point of view, as Theorem 4.3 only guarantees consistency under the condition that k_n grows slower than n when using a uniform prior. In such a case, the uniform prior may not place a large enough penalty on large values of k , which can result in a large value of $\hat{k}$ relative to the sample size. This will almost inevitably lead to an undersmoothed density estimate, as k is the main regularization parameter of our procedure. Overall, it seems reasonable to require that k_n grows at a rate slower than n , especially when using a uniform prior.

4.8 Simulation study: regular histograms

To compare the performance of the Bayesian histogram method with the automatic histogram procedures of Section 4.1, we conducted a large simulation study. The performance of each procedure is judged by both its performance with respect to classical loss functions such as the $\mathbb{L}_2$ (2.27) and Hellinger distances (2.28), but also with respect to its ability to correctly identify modes as measured by the PID loss (4.6). The test densities used in the simulation study, denoted $\mathcal{T}$, are shown in Figure 4.4. Densities 7, 8, 10, 12 and 14-16 are taken from Rozenholc et al. (2010), density 9 is from Marron and Wand (1992) while density 13 was suggested by Li et al. (2020). The test suite was chosen to include densities that exhibit different behavior in terms of modality, skewness and tail behavior. Since the PID loss of Definition 4.2 allows for some error in the estimation of the location of true modes, we only included densities where the peaks have enough separation to yield non-overlapping tolerance regions. In particular, we set the tolerance regions δ_j to 0.02 times the sample range for all modes and all densities throughout the simulation study. We assessed the risk for different sample sizes, generating samples of size $n \in \mathcal{J} = \{50, 200, 1000, 5000, 25\,000\}$ for each density. To estimate the risk for a given method m , we generated iid random samples $\mathbf{x}_{1:n}^{(b)} \sim f_0$ for $b = 1, 2, \dots, B$, computed the estimator $\hat{f}_n^{(b)}(\cdot | m)$ on the given sample and recorded the value of the loss function. The risk of a given procedure was estimated by the simple Monte Carlo average,

$$\hat{R}_n(m, f_0) = \frac{1}{B} \sum_{b=1}^B \ell(f_0, \hat{f}_n^{(b)}(\cdot | m)), \quad (f_0, m, n) \in \mathcal{T} \times \mathcal{M} \times \mathcal{J}, \quad (4.73)$$

where $\mathcal{M}$ denotes the histogram methods included in the simulation study. For our simulations here we used $B = 500$. Since densities #3 and #6 are not in $\mathbb{L}_2(\mathbb{R})$, we did not attempt to estimate the $\mathbb{L}_2$ risk for these densities. For all the test densities in Figure 4.4, the support of the density was estimated by using the minimum and maximum of the obtained sample as detailed in Section 2.1.3. We also followed this procedure when dealing with distributions with known support, such as the Beta(0.5, 0.5) and the LogNormal(0, 1) densities. Our reason for taking this approach here, was to ensure that all the methods tested here yield density estimates with the same support, as some of the histogram software used in the irregular simulations in Section 5.6 did not allow us to specify the support of the density manually, which would have made comparisons between the different methods more difficult.

An overview of the histogram procedures included in the simulation study is provided in Table 4.1. Note that all the random histogram procedures listed in Table 4.1 are based on the noninformative symmetric choice for the prior mean, $\mathbf{p}_0 = (k^{-1}, \dots, k^{-1})$ for given k . For the Wand procedure, we used the two-step plug-in rule from the `dpin` function implemented in the R package `KernSmooth` (Wand, 2021). For the Bayesian and

4.8. Simulation study: regular histograms

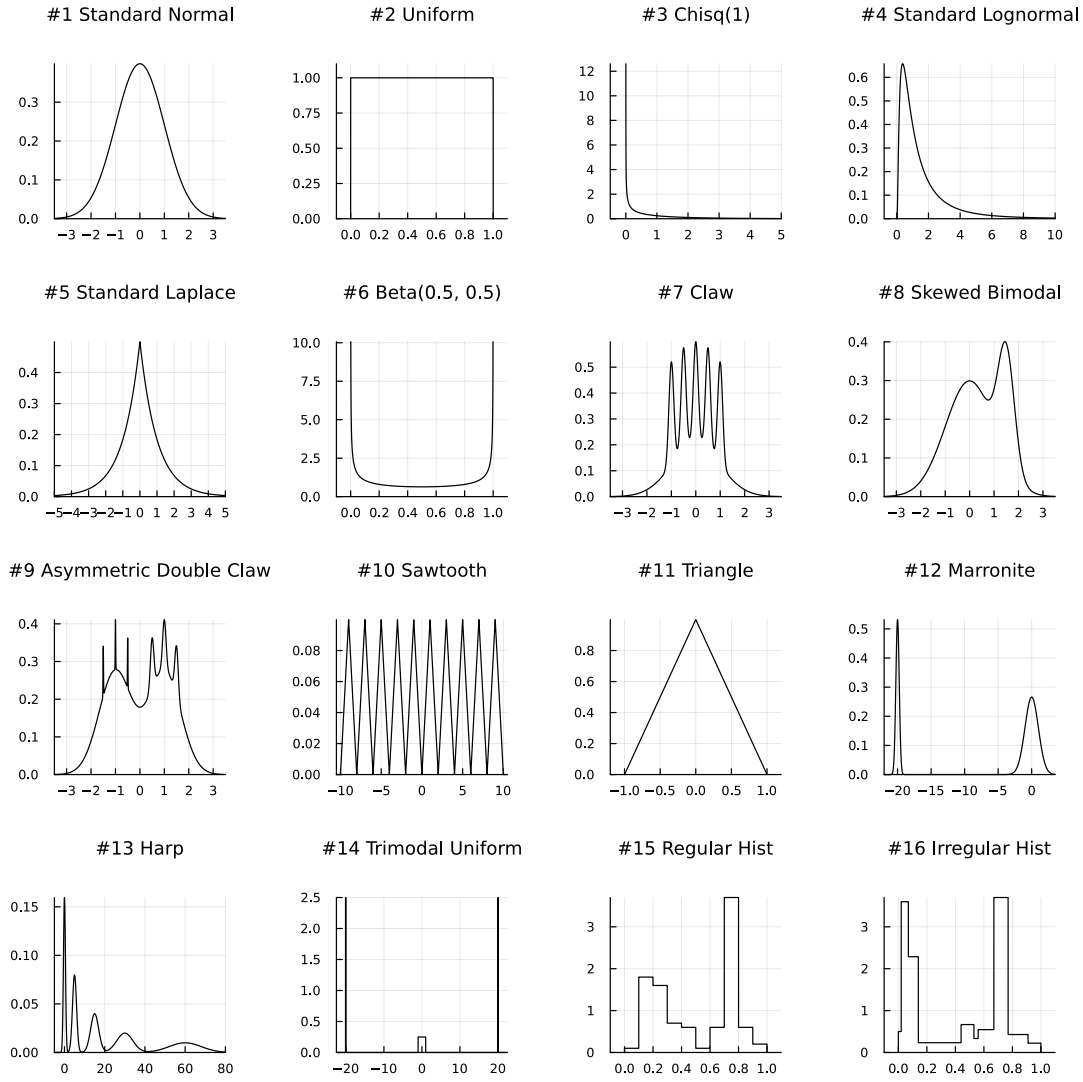


Figure 4.4: Test densities used in the simulation study

penalized maximum likelihood methods, we performed the maximization for k over the set $\{1, 2, \dots, k_n\}$ where $k_n = \lceil 4n / \log^2(n) \rceil$. We note that some of the methods tested here, such as that of Birgé and Rozenholc (2006), came with a suggestion of maximizing the log-likelihood over a larger set of k -values, but we found in our simulations that the maxima typically occurs for values of k much smaller than k_n , so our choice can save a considerable amount of computation for larger sample sizes.

To compare the relative performance of the different methods, we computed the ranks $\{r_n(m, f_0)\}_{m \in \mathcal{M}}$ of the estimated risks (4.73) for each tuple $(f_0, n) \in \mathcal{T} \times \mathcal{J}$. As a measure of the relative performance of the method m for a given sample size n , we computed the rank averaged over all densities for that sample size,

$$\bar{r}_n(m) = \frac{1}{|\mathcal{T}|} \sum_{f_0 \in \mathcal{T}} r_n(m, f_0). \quad (4.74)$$

Similarly, as a measure of the overall performance of a method m we consider the average

Table 4.1: Overview of the methods included in the simulation study

Method	Abbreviation	Description
Sturges' rule	St	Section 4.1.1a
Square root rule	Sqrt	
Scott's rule	Scott	Section 4.1.1b
Freedman and Diaconis' rule	FD	
Wand's 2-step rule	Wand	
Kullback–Leibler cross-validation	KLCV	Section 4.1.1c
$\mathbb{L}_2$ cross-validation	L2CV	
AIC	AIC	Section 4.1.1d
BIC	BIC	
Birgé and Rozenholc	BR	
Normalized Maximum Likelihood	NML	
Minimum Description Length	MDL	
Random histogram, $a = 1$	B-B	Section 4.7.1a
Random histogram, $a = k$	B-SC	
Random histogram, $a = k/2$	B-K	
Random histogram, $a = \sqrt{n}$	B-MM	

$\bar{r}(m)$ of $\bar{r}_n(m)$ over all sample sizes,

$$\bar{r}(m) = \frac{1}{|\mathcal{J}|} \sum_{n \in \mathcal{J}} \bar{r}_n(m) = \frac{1}{|\mathcal{T}||\mathcal{J}|} \sum_{(f_0, n) \in \mathcal{T} \times \mathcal{J}} r_n(m, f_0). \quad (4.75)$$

The overall average ranks $\bar{r}(m)$ (4.75) for the Hellinger and PID losses are shown in Figure 4.5, and the results for each sample size are given in Tables 4.2 and 4.3, with the best performing method for each sample size outlined in bold. Figure 4.6 shows a box plot of m against the logarithm of the Hellinger risk relative to the best method,

$$\text{LRR}_n(m, f_0) = \log \left\{ \hat{R}_n(m, f_0) / \min_{m' \in \mathcal{M}} \hat{R}_n(m', f_0) \right\} \quad (4.76)$$

for samples of size $n = 200$ (left) and $n = 25\,000$ (right). Note that some of the outliers for the worse performing methods are not shown here in order to give more detail to the differences between the better performing methods. Finally, Tables 4.4 and 4.5 shows the estimated risks $\hat{R}_n(m, f_0)$ for all $(f_0, m, n) \in \mathcal{T} \times \mathcal{M} \times \mathcal{J}$ for the Hellinger and PID losses, respectively. The results for the $\mathbb{L}_2$ loss are deferred to Appendix C.2.

The Bayesian methods MM, SC and K all perform very well with respect to the Hellinger metric, as indicated by their low average ranks in Figure 4.5, with MM yielding the lowest overall rank out of all the methods. The left boxplot in Figure 4.6 shows that these three methods are the overall best performers for $n = 200$. Inspection of Table 4.2 shows that the relative performance of the Bayesian methods are particularly good for smaller sample sizes; for example, the Hellinger risk MM procedure is never much worse than that of the best method. The Bayesian model B with $a = 1$ performs considerably worse with respect to Hellinger risk. Some additional simulations showed that this method tended to select smaller values of k than the other Bayesian approaches, leading to histograms that are oversmoothed with respect to the Hellinger metric. The BIC-based histograms also tended to perform poorly, having the highest average rank out of all the procedures based on penalized maximum likelihood. Out of the non-Bayesian approaches, the criteria BR, AIC and L2CV yield good performance overall

4.8. Simulation study: regular histograms

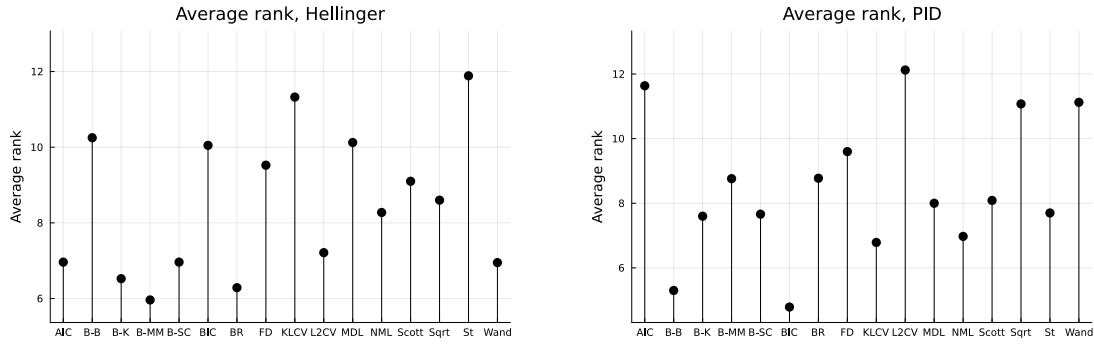


Figure 4.5: Left: overall average ranks by Hellinger distance. Right: overall average ranks by PID loss.

Table 4.2: Average ranks, Hellinger risk

n	50	200	1000	5000	25000
AIC	10.9	9.44	6.06	4.44	3.94
B-B	8.44	11.1	10.9	10.6	10.2
B-K	6.88	5.75	5.69	7.19	7.12
B-MM	5.19	4.81	6.25	6.88	6.69
B-SC	6.0	5.25	6.81	8.5	8.25
BIC	9.0	10.4	10.9	10.2	9.69
BR	8.69	8.06	6.12	4.12	4.44
FD	9.06	9.75	9.62	9.62	9.56
KLCV	10.8	11.1	11.6	11.3	11.8
L2CV	9.56	8.69	6.62	5.44	5.75
MDL	11.5	9.81	9.0	9.56	10.8
NML	10.0	8.0	7.94	7.88	7.56
Scott	7.0	8.94	9.56	10.1	9.94
Sqrt	8.06	8.06	8.56	9.06	9.25
St	7.44	8.94	13.2	14.8	15.1
Wand	7.44	7.81	7.12	6.38	6.0

Table 4.3: Average ranks, PID risk

n	50	200	1000	5000	25000
AIC	10.2	10.4	12.1	13.2	12.2
B-B	5.62	4.88	5.25	4.88	5.88
B-K	8.5	7.31	7.69	7.0	7.5
B-MM	8.5	7.69	8.75	8.88	10.0
B-SC	8.62	8.31	7.5	6.81	7.06
BIC	5.31	6.12	4.31	3.94	4.25
BR	5.38	7.19	9.44	10.8	11.1
FD	10.2	10.9	9.94	8.75	8.19
KLCV	6.38	6.5	7.19	7.38	6.5
L2CV	9.69	9.62	13.1	14.1	14.2
MDL	9.75	9.69	8.44	6.75	5.38
NML	7.19	6.5	7.31	6.81	7.06
Scott	7.81	8.88	9.12	7.75	6.88
Sqrt	12.8	11.7	9.0	10.2	11.7
St	10.6	8.94	6.69	7.12	5.19
Wand	9.44	11.3	10.2	11.6	13.0

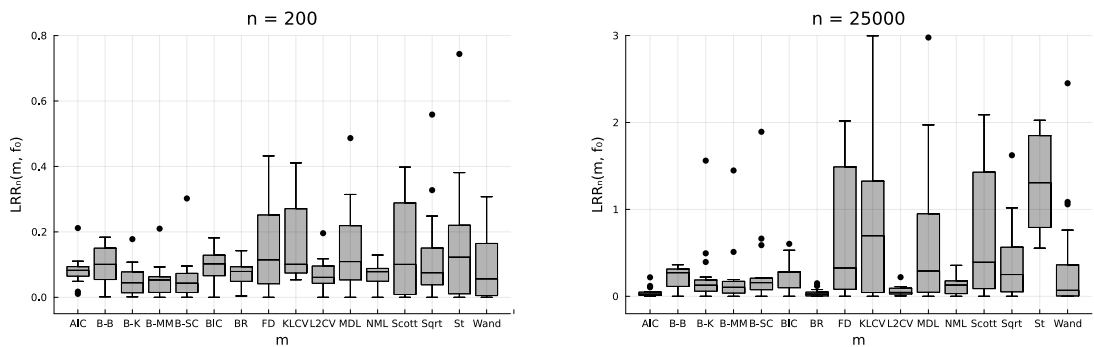


Figure 4.6: Boxplots of log-relative Hellinger risk by method for sample sizes $n = 200$ (left) and $n = 25000$ (right).

and especially for larger sample sizes as can be seen from the right boxplot Figure 4.6; their relative risk is never much larger than that of the best method for any given

density. The good performance of AIC and the similar BR criterion for larger sample sizes is expected in light of the results of Kanazawa (1993), who showed that under suitable regularity assumptions, the AIC asymptotically aims for the Hellinger-optimal bin width. L2CV performs well for large sample sizes, but the Hellinger risk is relatively high for small samples. On the other hand, Kullback–Leibler cross-validation performed quite poorly across the board as it was among the worst performers for all samples sizes. Out of the coding-length based procedures MDL and NML, the NML procedure does a bit better. Finally, the simple plug-in rules Sqrt, St, Scott and FD performed poorly compared to most of the other methods. To our surprise, the square root rule yielded the best performance out of these in our simulations, although it is not derived from some asymptotic criterion, and it does not even increase at the rate $k = \mathcal{O}(n^{-1/3})$ which is optimal for most of the densities considered here. The poor performance of Sturges’ rule for large sample sizes is unsurprising, as the number of bins grows much slower than what is optimal according to asymptotic criteria. We note that the $\mathbb{L}_2$ risks showed similar results to the Hellinger risks, with the Bayesian method MM obtaining the overall lowest rank.

For the PID loss, the BIC is the clear winner, producing the lowest average rank overall and for every sample size but $n = 200$. We note that the BIC was also found to be the best performer in terms of mode identification among the regular histogram procedures considered in the simulation study of Davies et al. (2009). The Bayesian method B also performs well with respect to mode identification, and ranks second only to the BIC in this regard. Both methods tend to choose smaller values of k when compared to the penalized likelihood methods AIC and BR or the Bayesian methods K, SC and MM. This indicates that for our test density suite (Figure 4.4) and regular histogram procedures, there exists a trade-off between the appropriate number of bins for mode identification and for good performance with respect to Hellinger loss. In particular, methods such as L2CV, BR or AIC tend to produce histograms with a large number of spurious modes, leading to larger PID loss. Although the BIC performs well in mode identification relative to the other regular histogram methods considered here, the results are overall not very impressive, as the PID risks in Table 4.5 are large for many densities. This is particularly noticeable for the heavy-tailed Chisq(1) and LogNormal(0, 1) densities (#3 and #4), where the use of a global bin width causes undersmoothing in the tails of the density, leading to a large amount of spurious modes. Moreover, the PID risks are often increasing with the sample size, so the BIC histogram does not appear to get better at locating modes with increasing sample size.

Our results here suggest that the optimal regular histogram procedure is very much dependent on the overall goal of the analysis. Methods that perform well with respect to classical loss functions are not among the best performers when it comes to identifying modes of the true density and vice versa. If the goal is to locate the true modes of a density, then the BIC criterion, or the B-B criterion if a Bayesian analysis is preferred, are the most appealing choices among the regular procedures. On the other hand, if good performance with respect to the Hellinger metric is desired, then AIC, BR or the Bayesian MM seem more appropriate.

4.8. Simulation study: regular histograms

Table 4.4: Hellinger risks from the simulation study.

Density	n	AIC	B-B	B-K	B-MM	B-SC	BIC	BR	FD	KLCV	L2CV	MDL	NML	Scott	Sqrt	St	Wand
1	50	0.298	0.308	0.3	0.288	0.293	0.303	0.298	0.277	0.287	0.299	0.307	0.306	0.269	0.282	0.275	0.272
	200	0.186	0.195	0.184	0.183	0.183	0.192	0.185	0.172	0.183	0.187	0.185	0.185	0.166	0.179	0.166	0.167
	1000	0.104	0.12	0.108	0.109	0.109	0.117	0.106	0.1	0.115	0.105	0.103	0.109	0.096	0.108	0.108	0.096
	5000	0.06	0.073	0.066	0.066	0.067	0.072	0.061	0.059	0.078	0.06	0.063	0.066	0.056	0.068	0.083	0.056
	25000	0.035	0.045	0.04	0.041	0.041	0.044	0.035	0.035	0.063	0.035	0.042	0.04	0.033	0.044	0.075	0.033
2	50	0.232	0.188	0.191	0.201	0.201	0.19	0.192	0.224	0.205	0.22	0.294	0.191	0.222	0.267	0.255	0.224
	200	0.117	0.095	0.095	0.1	0.096	0.095	0.099	0.126	0.113	0.115	0.11	0.094	0.125	0.165	0.138	0.128
	1000	0.051	0.042	0.042	0.043	0.042	0.042	0.044	0.065	0.05	0.051	0.042	0.042	0.065	0.098	0.066	0.07
	5000	0.023	0.019	0.019	0.019	0.019	0.019	0.019	0.034	0.023	0.023	0.019	0.019	0.034	0.063	0.032	0.039
	25000	0.01	0.008	0.008	0.008	0.008	0.008	0.008	0.019	0.01	0.01	0.008	0.008	0.019	0.041	0.015	0.024
3	50	0.448	0.427	0.405	0.399	0.409	0.446	0.446	0.423	0.505	0.437	0.433	0.448	0.428	0.419	0.422	0.435
	200	0.336	0.335	0.315	0.309	0.325	0.347	0.339	0.323	0.467	0.326	0.385	0.337	0.348	0.332	0.363	0.332
	1000	0.252	0.261	0.245	0.238	0.258	0.272	0.253	0.25	0.474	0.246	0.38	0.255	0.282	0.269	0.355	0.251
	5000	0.193	0.209	0.195	0.185	0.207	0.217	0.193	0.204	0.458	0.19	0.368	0.198	0.236	0.22	0.347	0.192
	25000	0.148	0.166	0.155	0.145	0.166	0.173	0.148	0.171	0.448	0.146	0.362	0.155	0.201	0.182	0.351	0.146
4	50	0.352	0.329	0.325	0.324	0.335	0.349	0.349	0.354	0.499	0.343	0.383	0.35	0.332	0.335	0.333	0.365
	200	0.23	0.224	0.22	0.223	0.233	0.239	0.233	0.248	0.498	0.228	0.357	0.231	0.219	0.228	0.254	0.266
	1000	0.166	0.166	0.166	0.17	0.178	0.178	0.167	0.175	0.597	0.172	0.425	0.169	0.161	0.166	0.273	0.189
	5000	0.137	0.15	0.151	0.132	0.158	0.156	0.143	0.123	0.682	0.126	0.512	0.151	0.14	0.144	0.316	0.129
	25000	0.091	0.112	0.122	0.089	0.149	0.151	0.093	0.082	0.765	0.085	0.593	0.118	0.114	0.13	0.383	0.095
5	50	0.336	0.327	0.321	0.319	0.32	0.327	0.326	0.336	0.341	0.329	0.315	0.333	0.303	0.309	0.305	0.319
	200	0.214	0.221	0.208	0.207	0.211	0.221	0.214	0.217	0.253	0.213	0.207	0.213	0.195	0.198	0.207	0.207
	1000	0.129	0.142	0.13	0.129	0.135	0.143	0.13	0.136	0.222	0.134	0.154	0.133	0.121	0.123	0.174	0.129
	5000	0.077	0.09	0.082	0.081	0.086	0.092	0.078	0.083	0.198	0.082	0.126	0.083	0.074	0.075	0.158	0.08
	25000	0.046	0.057	0.051	0.051	0.055	0.059	0.047	0.051	0.191	0.049	0.118	0.052	0.045	0.047	0.161	0.049
6	50	0.384	0.336	0.344	0.333	0.329	0.343	0.344	0.321	0.332	0.374	0.351	0.37	0.32	0.328	0.324	0.319
	200	0.269	0.269	0.259	0.251	0.251	0.265	0.261	0.264	0.257	0.267	0.269	0.265	0.263	0.242	0.245	0.246
	1000	0.193	0.206	0.189	0.187	0.186	0.2	0.189	0.227	0.191	0.195	0.197	0.19	0.221	0.181	0.211	0.19
	5000	0.145	0.164	0.146	0.146	0.142	0.158	0.143	0.201	0.144	0.153	0.144	0.146	0.196	0.143	0.196	0.152
	25000	0.11	0.129	0.113	0.114	0.111	0.124	0.109	0.178	0.109	0.122	0.109	0.113	0.172	0.114	0.189	0.122
7	50	0.358	0.355	0.348	0.343	0.342	0.35	0.349	0.338	0.344	0.36	0.341	0.357	0.333	0.344	0.336	0.335
	200	0.254	0.251	0.243	0.239	0.241	0.25	0.248	0.247	0.256	0.25	0.236	0.249	0.249	0.242	0.249	0.249
	1000	0.157	0.178	0.148	0.149	0.15	0.171	0.152	0.166	0.202	0.159	0.169	0.152	0.181	0.155	0.217	0.168
	5000	0.094	0.115	0.099	0.095	0.103	0.113	0.094	0.104	0.176	0.096	0.137	0.101	0.118	0.091	0.205	0.09
	25000	0.055	0.073	0.061	0.058	0.064	0.071	0.055	0.063	0.163	0.057	0.12	0.062	0.072	0.054	0.203	0.054
8	50	0.309	0.309	0.305	0.295	0.3	0.306	0.303	0.28	0.294	0.307	0.318	0.309	0.28	0.291	0.284	0.281
	200	0.194	0.205	0.191	0.19	0.19	0.2	0.193	0.177	0.195	0.194	0.19	0.193	0.179	0.183	0.177	0.177
	1000	0.112	0.13	0.117	0.115	0.117	0.127	0.113	0.103	0.126	0.112	0.111	0.118	0.105	0.111	0.125	0.103
	5000	0.064	0.078	0.07	0.07	0.071	0.077	0.065	0.061	0.093	0.064	0.07	0.07	0.062	0.069	0.1	0.061
	25000	0.038	0.048	0.043	0.043	0.044	0.047	0.038	0.036	0.077	0.038	0.051	0.043	0.037	0.044	0.093	0.036
9	50	0.314	0.312	0.309	0.299	0.305	0.311	0.31	0.287	0.301	0.313	0.332	0.312	0.287	0.295	0.288	0.287
	200	0.196	0.206	0.194	0.191	0.192	0.201	0.195	0.181	0.193	0.197	0.201	0.195	0.182	0.187	0.18	0.18
	1000	0.118	0.133	0.122	0.121	0.122	0.13	0.119	0.114	0.12	0.12	0.117	0.123	0.115	0.114	0.128	0.112
	5000	0.072	0.088	0.076	0.076	0.077	0.085	0.072	0.074	0.086	0.073	0.075	0.077	0.076	0.073	0.104	0.07
	25000	0.046	0.058	0.052	0.051	0.052	0.056	0.046	0.05	0.07	0.047	0.054	0.052	0.051	0.048	0.096	0.045
10	50	0.424	0.388	0.393	0.399	0.401	0.39	0.392	0.404	0.401	0.415	0.439	0.392	0.403	0.428	0.416	0.405
	200	0.306	0.344	0.31	0.304	0.286	0.344	0.33	0.347	0.302	0.303	0.295	0.326	0.347	0.339	0.357	0.349
	1000	0.183	0.235	0.182	0.186	0.176	0.216	0.184	0.337	0.183	0.183	0.177	0.187	0.337	0.214	0.335	0.321
	5000	0.107	0.145	0.114	0.115	0.111	0.134	0.107	0.282	0.106	0.108	0.107	0.115	0.285	0.12	0.314	0.189
	25000	0.063	0.09	0.071	0.071	0.069	0.083	0.063	0.227	0.063	0.065	0.067	0.071	0.217	0.064	0.299	0.079
11	50	0.288	0.288	0.287	0.273	0.281	0.285	0.282	0.26	0.272	0.286	0.317	0.291	0.255	0.274	0.268	0.256
	200	0.172	0.183	0.172	0.166	0.168	0.179	0.173	0.156	0.167	0.174	0.194	0.174	0.155	0.171	0.155	0.155
	1000	0.095	0.112	0.101	0.097	0.099	0.109	0.098	0.088	0.094	0.098	0.097	0.102	0.088	0.102	0.092	0.089
	5000	0.054	0.069	0.061	0.058	0.06	0.066	0.056	0.051	0.054	0.056	0.058	0.061	0.052	0.064	0.064	0.052
	25000	0.031	0.042	0.037	0.035	0.037	0.04	0.032	0.03	0.031	0.033	0.036	0.037	0.031	0.041	0.054	0.031
12	50	0.49	0.496	0.552	0.547	0.605	0.543	0.536	0.834	0.997	0.479	0.997	0.49	0.871	0.621	0.64	0.584
	200	0.277	0.283	0.327	0.338	0.371	0.294	0.281	0.73	0.975	0.274	0.975	0.279	0.685	0.38	0.576	0.29
	1000	0.163	0.181	0.22	0.204	0.263	0.243	0.175	0.69	0.969	0.151	0.969	0.181	0.624	0.286	0.48	0.183
	5000	0.094	0.111	0.132	0.128	0.155	0.135	0.098	0.48	0.969	0.084	0.968	0.111	0.419	0.128	0.39	0.09
	25000	0.054	0.068	0.079	0.081	0.094	0.082	0.056	0.364	0.969	0.05	0.95	0.068	0.391	0.062	0.362	0.048
13	50	0.512	0.515	0.476	0.471	0.468	0.522	0.521	0.541	0.52	0.504	0.501	0.513	0.516	0.513	0.532	0.538
	200	0.377	0.415	0.356	0.357	0.352	0.41	0.388	0.446	0.419	0.367	0.396	0.379	0.491	0.451	0.438	0.438
	1000	0.226	0.255	0.219	0.2242												

Chapter 4. Regular Histograms

Table 4.5: Peak identification risks.

Density	n	AIC	B-B	B-K	B-MM	B-SC	BIC	BR	FD	KLCV	L2CV	MDL	NML	Scott	Sqrt	St	Wand
1	50	2.154	1.586	1.8	1.89	2.032	1.596	1.638	2.418	1.764	2.158	3.304	1.728	1.776	2.784	2.282	1.906
	200	2.432	1.598	1.654	1.774	1.716	1.592	1.62	2.72	1.73	2.57	3.024	1.632	1.876	3.6	1.74	1.934
	1000	2.838	1.268	1.42	1.584	1.398	1.248	1.598	4.046	1.55	3.15	1.93	1.408	2.29	6.846	1.204	2.396
	5000	4.41	1.006	1.342	1.868	1.23	0.996	2.264	7.048	1.258	4.57	1.578	1.336	3.898	16.168	1.002	3.904
	25000	7.156	1.13	1.676	2.62	1.446	1.118	4.376	13.596	1.036	8.092	1.384	1.676	7.392	41.034	0.8	7.366
2	50	1.302	0.304	0.372	0.736	0.704	0.362	0.476	2.64	0.756	1.132	3.49	0.348	2.64	3.864	3.09	2.656
	200	1.158	0.046	0.046	0.332	0.116	0.046	0.276	3.244	0.97	1.108	0.796	0.038	3.316	5.662	3.72	3.134
	1000	1.042	0.008	0.008	0.124	0.012	0.008	0.192	4.534	1.032	1.052	0.03	0.008	4.564	10.54	4.414	5.176
	5000	1.12	0.004	0.004	0.076	0.018	0.004	0.208	6.872	1.1	1.126	0.026	0.004	6.628	22.846	5.95	8.814
	25000	1.09	0.008	0.008	0.056	0.008	0.008	0.234	9.926	1.09	1.124	0.012	0.008	10.014	51.67	6.588	16.722
3	50	4.608	3.842	4.448	4.588	4.274	3.544	3.546	4.046	2.206	4.798	2.748	4.46	2.78	3.56	3.316	4.892
	200	6.8	5.302	6.434	6.698	5.994	4.756	6.024	5.96	2.306	6.452	3.052	6.464	3.786	4.962	3.308	8.44
	1000	16.602	8.582	12.7	15.818	9.95	6.884	14.668	10.438	2.29	17.808	3.038	13.256	5.56	6.364	3.232	26.144
	5000	52.408	21.138	31.474	46.312	23.986	17.324	48.808	22.474	2.372	57.502	3.064	33.222	9.93	14.69	3.336	73.294
	25000	155.402	56.21	81.524	121.616	62.508	45.672	147.412	46.188	2.33	191.312	3.01	86.058	20.966	33.792	3.22	196.586
4	50	3.29	3.042	3.176	3.21	3.076	2.91	2.908	3.052	2.088	3.126	2.318	3.17	2.846	3.122	3.052	3.088
	200	3.22	3.328	3.196	3.204	3.268	3.41	3.254	5.836	2.162	3.504	2.408	3.2	2.556	2.994	3.27	7.496
	1000	5.304	4.03	3.978	5.278	3.574	3.42	4.376	14.032	2.144	11.518	2.398	4.066	5.684	5.43	3.572	20.558
	5000	19.134	7.5	7.286	23.274	6.288	6.162	13.922	33.488	2.162	37.42	2.452	7.454	14.094	11.524	3.922	52.908
	25000	58.202	35.368	29.59	64.55	14.612	12.452	55.734	76.654	2.176	109.14	2.502	32.578	32.852	24.298	4.098	167.292
5	50	2.302	1.676	2.086	2.058	2.076	1.664	1.658	2.94	1.646	2.262	2.232	2.032	1.85	2.298	2.052	2.438
	200	2.642	1.284	1.602	1.87	1.498	1.278	1.48	4.526	1.258	3.264	1.844	1.588	2.116	2.656	1.666	3.4
	1000	4.5	1.396	2.12	3.032	1.824	1.332	2.77	9.71	1.026	7.698	1.388	2.158	3.686	4.796	1.478	7.474
	5000	9.26	2.774	3.984	6.222	3.336	2.596	6.836	20.916	0.73	18.29	1.226	4.052	8.754	12.25	1.3	17.69
	25000	18.982	6.156	8.104	12.752	6.876	5.776	15.448	44.138	0.598	36.928	1.348	8.206	19.866	30.376	1.06	38.872
6	50	5.372	3.454	4.69	4.81	5.312	3.64	3.718	3.96	4.198	5.428	6.088	4.438	3.982	4.966	4.706	4.092
	200	7.998	4.458	6.708	6.76	7.68	4.85	5.9	4.142	7.37	8.05	8.052	6.456	4.182	7.068	5.254	5.17
	1000	17.07	6.518	9.612	8.932	12.536	7.434	11.02	4.558	15.282	22.864	21.08	9.198	4.832	8.024	5.468	8.246
	5000	52.824	7.852	21.384	20.262	28.064	10.306	41.534	5.206	49.44	83.592	42.468	20.69	5.52	19.61	5.526	11.698
	25000	153.954	22.878	57.026	53.41	69.17	29.868	138.66	6.226	148.528	307.734	82.716	56.134	7.08	46.874	5.162	31.54
7	50	5.606	5.428	5.47	5.584	5.574	5.452	5.468	5.882	5.522	5.64	5.57	5.518	5.702	5.998	5.832	5.778
	200	4.61	5.408	5.054	4.85	4.938	5.428	5.316	5.916	5.324	4.428	4.93	5.11	5.706	5.364	5.58	5.844
	1000	7.198	4.434	3.674	4.258	3.44	4.008	4.516	3.216	4.638	9.334	3.58	3.656	3.406	4.152	5.32	3.23
	5000	15.976	2.53	6.972	9.984	5.776	2.806	12.748	6.322	3.654	21.638	3.292	7.012	4.096	11.25	4.97	12.316
	25000	30.398	9.162	15.106	19.314	13.34	9.848	26.806	13.386	3.12	45.2	2.84	15.222	8.792	28.588	4.842	38.718
8	50	3.134	2.592	2.736	2.868	3.04	2.602	2.63	3.148	2.714	3.062	4.12	2.696	2.996	3.714	3.508	3.088
	200	3.65	2.486	2.774	2.964	2.94	2.554	2.706	3.074	2.676	3.558	3.878	2.784	2.842	4.268	2.944	3.128
	1000	4.56	2.17	2.546	2.98	2.58	2.252	2.956	3.196	2.526	4.91	3.178	2.532	2.924	7.064	2.354	3.688
	5000	6.332	1.848	2.328	3.054	2.286	1.852	3.918	4.324	2.05	6.582	2.518	2.338	3.56	15.31	1.976	6.008
	25000	11.246	1.822	2.858	4.576	2.614	1.952	7.278	7.984	1.398	12.474	2.156	2.86	6.476	39.074	1.542	12.248
9	50	6.776	6.794	6.834	6.754	6.818	6.8	6.75	6.576	6.702	6.818	7.162	6.832	6.466	6.846	6.87	6.482
	200	6.658	6.344	6.482	6.512	6.508	6.402	6.506	6.398	6.446	6.48	6.762	6.468	6.348	6.734	6.264	6.354
	1000	5.412	6.54	5.976	5.568	5.854	6.412	5.61	5.742	5.428	5.688	5.262	5.996	5.726	5.878	5.83	5.54
	5000	5.986	4.51	3.324	3.17	3.322	4.284	3.72	3.09	4.118	7.922	3.346	3.342	3.236	11.568	5.778	3.444
	25000	16.354	2.418	2.912	4.046	2.842	2.522	9.496	3.476	3.618	28.22	3.168	2.908	3.248	33.878	5.332	7.138
10	50	10.738	10.92	10.87	10.85	10.694	10.84	10.758	9.97	10.844	10.768	9.986	10.91	9.402	11.442	11.806	9.844
	200	2.36	10.968	5.146	4.03	2.78	10.788	6.968	12.466	2.744	2.216	2.366	5.978	12.516	6.244	11.636	12.284
	1000	2.37	1.012	0.708	0.428	1.146	0.546	1.0	7.246	0.492	1.602	2.078	0.606	6.852	0.364	9.594	6.372
	5000	7.644	0.004	0.434	0.268	0.776	0.008	3.774	3.76	3.202	3.414	3.064	0.35	2.926	0.044	5.016	0.012
	25000	18.248	0.002	0.604	0.63	1.112	0.018	12.266	0.0	14.396	7.536	2.47	0.526	0.0	2.086	3.012	0.026
11	50	2.038	1.348	1.614	1.778	1.858	1.408	1.43	2.342	1.654	2.034	3.932	1.522	1.916	3.062	2.596	1.996
	200	2.28	1.098	1.34	1.468	1.538	1.148	1.284	2.328	1.9	2.254	4.25	1.308	2.176	4.172	1.684	1.83
	1000	2.712	0.902	1.174	1.36	1.266	0.932	1.352	2.668	2.51	2.4	1.508	1.158	2.42	8.322	0.504	1.638
	5000	4.074	0.968	1.202	1.496	1.29	1.016	2.1	3.19	4.012	3.164	1.526	1.188	2.234	19.466	1.988	1.66
	25000	6.354	0.734	0.624	0.886	0.676	0.648	2.984	4.582	6.21	2.886	0.808	0.624	2.344	47.28	1.996	1.788
12	50	1.978	1.946	2.016	1.784	2.132	2.51	2.438	3.638	3.0	1.736	3.0	1.966	3.984	3.276	2.944	3.124
	200	0.492	0.476	0.502	0.486	0.504	0.51	0.516	2.764	3.0	0.458	3.0	0.482	2.796	0.5	3.884	0.558
	1000	1.382	0.724	0.648	1.856	0.412	0.308	0.874	3.992	3.0	2.034	3.0	0.732	3.96	0.114	0.452	0.422
	5000	3.706	1.548	1.342	4.518	0.838	0.6	2.826	0.352	3.0	14.284	2.994	1.498	1.6	0.588	1.668	3.53
	25000	8.738	3.362	3.042	8.696	2.004	1.714	7.398	0.738	2.99	30.928	2.844	3.236	0.204	4.16	0.13	16.53
13	50	6.684	6.248	6.592	6.612	6.63	6.358	6.348	5.534	6.27	6.702	6.206	6.58	6.466	5.27	5.084	5.086
	200	5.522	5.838	5.51	5.486	5.494	5.772	5.55	5.216	5.904	5.368	5.458	5.524	4.4	6.224	5.548	5.958
	1000	1															

Chapter 5

Irregular Histograms

Although automatic regular histogram constructions have proven valuable in many settings, even outperforming irregular histogram constructions in terms of estimated risks in many density estimation problems (Davies et al., 2009), irregular histograms remain an attractive alternative as they are able to adapt the degree of smoothing to the local behavior of a given density by varying the bin widths. This chapter focuses on irregular histogram constructions and introduces the topic through a brief overview of various automatic methods. We propose a new irregular Bayesian histogram construction and study its asymptotic properties in detail. The chapter concludes with a simulation study that compares the different irregular histogram methods with respect to classical loss functions and in terms of mode identification. Finally, we compare the irregular histograms with their regular cousins, evaluating their relative strengths and weaknesses.

5.1 An introduction to irregular histograms

The first section offers a brief introduction to irregular histogram construction, describing some of the procedures that have been proposed in the literature.

The regular histogram procedures of Chapter 4 reduced the problem of selecting the best partition to one of choosing the number of bins. Irregular histogram construction on the other hand constitutes a much more difficult problem, as one has to choose both the number of bins that are used in addition to their location. Any irregular partition $I_{1:k}$ of $[0, 1]$ can be described in terms of the endpoints $\xi_0, \xi_1, \dots, \xi_k$ of the constituent intervals, where $I_1 = [\xi_0, \xi_1]$, $I_j = (\xi_{j-1}, \xi_j]$ for $2 \leq j \leq k$ and the endpoints satisfy $\xi_0 = 0$, $\xi_k = 1$ and $\xi_{j-1} < \xi_j$. The problem of constructing the best irregular histogram for a given value of k then reduces to finding the vector ξ_k which produces the best histogram for the given dataset according to a suitable optimality criterion. However, since we are both interested in finding the number k of bins as well as the locations of the bin edges ξ_k , a naive approach where the bin edges are allowed to vary continuously would quickly become computationally prohibitive, as it would involve the solution of many k -dimensional optimization problems, where k would typically be allowed to be quite large for larger sample sizes. In practice some form of discretization is usually employed to ease this computational burden, so that the bin edges ξ_k are only allowed to take values in a finite set $\{\tau_{n,j} : 0 \leq j \leq k_n\}$, where $\tau_{n,0} = 0$, $\tau_{n,k_n} = 1$, $\tau_{n,j-1} < \tau_{n,j}$ and k_n is the maximal number of grid points. In this discretized approach, it is customary to have the ‘resolution’ of the grid increase along with the sample size, allowing finer details in the data to be revealed for larger samples. Letting the grid become finer with the sample size is also necessary if one is interested in having a consistent procedure;

it is not hard to see that $\max_{1 \leq j \leq k_n} \{\tau_{n,j} - \tau_{n,j-1}\} \rightarrow 0$ is a necessary condition in order to have consistency at an arbitrary continuous density f_0 . We note that in the discretization scheme described above not much generality is lost, since one can allow the grid to depend on the values of the sample $\mathbf{x}_{1:n}$.

Although discretizing the grid manages to reduce the computational complexity of the problem somewhat, irregular histogram constructions based on such grids still require the use of more sophisticated optimization methods than a brute-force approach. For regular histograms, it is possible to compute the optimality criterion for each value of k belonging to a finite subset of $\mathbb{N}$, provided the set is not too large. In the irregular case, however, such a brute-force approach does not work well due to the fact that the number of possible partitions of $[0, 1]$ with endpoints in $\{\tau_{n,j} : 0 \leq j \leq k_n\}$ is 2^{k_n-1} . For methods that are based on maximizing a goodness-of-fit criterion, such as cross-validation or penalized likelihood, enumerating the criterion for every possible partition quickly becomes computationally infeasible even for moderate values of k_n . Due to this difficulty, many of the optimality criteria considered for irregular histograms have a particular additive structure to them that allows for the use of a polynomial time dynamical programming algorithm, efficient search heuristics (Mendizábal et al., 2023) or a combination of both (Rozenholc et al., 2010). More details on the required structure as well as the optimization procedures are given in Section 5.2.2.

Before delving into some of the automatic irregular histogram procedures that have been proposed in the literature, we highlight some of the difficulties encountered when developing optimality criteria for irregular histograms.

Example 5.1. Since the histogram is a maximum likelihood estimator as seen in Example 2.1, one could consider criteria based on the maximized log-likelihood. However, this approach fails miserably unless some form of regularization is used in addition. For regular histograms, it was shown in Proposition A.1 that the log-likelihood tends to infinity when the number of bins is allowed to become arbitrarily large, which leads us to consider an approach based on maximizing a penalized likelihood (4.2). In the irregular case one also encounters a similar problem when maximizing the log-likelihood, suggesting a need for a penalty on the number k of bins. However, the irregular histogram procedure also introduces the additional problem of having to account for the fact that there is more than one possible partition for each value of k . This is one of the key points made by Rozenholc et al. (2010), who offer both theoretical and empirical support to the idea that not taking into account the number of candidate models through the use of a stronger penalty leads to a histogram procedure which will perform poorly in practice. In irregular histogram methods based on penalized likelihoods, we often consider criteria of the form

$$\log L_n(\hat{\mathbf{p}} \mid k, I_{1:k}) - \text{pen}_n(k, I_{1:k}),$$

where the penalty is now allowed to depend on the chosen partition $I_{1:k}$, and will in general penalize more those values of k for which there are many possible partitions. Due to the heavier penalty used by irregular histogram methods, the MAP choice for the number of bins tends to be smaller than the corresponding MAP in a regular histogram procedure, counteracting some of the reduction in bias from using an irregular partition. We shall see later that the need to use a heavier penalty in the irregular case is accounted for in a natural way by the Bayesian procedure described later by the choice of appropriate proper prior distributions. $\blacktriangle$

5.1.1 Methods of irregular histogram construction

A brief introduction to various automatic irregular histogram procedures is given below. Many of these have clear regular histogram analogues, while others are motivated on entirely different grounds.

5.1.1a Penalized likelihood methods

As discussed in Example 5.1, penalized likelihood methods can also be used for irregular histograms. Rozenholc et al. (2010) propose what is essentially a modified variant of the criterion of Birgé and Rozenholc (4.3), where the number of different possible models for each value of k in the penalty is taken into account. Based on performance in simulations, they suggest the two penalties

$$\text{pen}_n^B(k, I_{1:k}) = k - 1 + \log^{2.5}(k) + \log \binom{k_n - 1}{k - 1}, \quad (5.1)$$

$$\text{pen}_n^R(k, I_{1:k}) = \frac{1}{2n} \sum_{j=1}^k \frac{N_j}{|I_j|} + \log^{2.5}(k) + \log \binom{k_n - 1}{k - 1}. \quad (5.2)$$

Note that the latter penalty depends on the data, unlike the other penalties considered so far. The Stochastic Complexity and Minimal Description Length criteria of Rissanen (1983) have also been applied to irregular histograms. Rissanen et al. (1992) extends the work of Hall and Hannan (1988) by proposing a stochastic complexity criterion which can be used to select both the number of bins and their location. The approximation of the Normalized Maximum Likelihood (NML) histogram of Kontkanen and Myllymäki (2007) mirrors the penalty of the regular case (4.4), but has an additional term to penalize the number of models under consideration for each given value of k ,

$$\begin{aligned} \text{pen}_n(k, I_{1:k}) = & \frac{k-1}{2} \log(n/2) + \log \frac{\sqrt{\pi}}{\Gamma(k/2)} + n^{-1/2} \frac{\sqrt{2k}\Gamma(k/2)}{3\Gamma(k/2 - 1/2)} \\ & + n^{-1} \left(\frac{3 + k(k-2)(2k+1)}{36} - \frac{\Gamma(k/2)^2 k^2}{9\Gamma(k/2 - 1/2)^2} \right) + \log \binom{k_n - 1}{k - 1}. \end{aligned} \quad (5.3)$$

5.1.1b Bayesian methods

A few of the irregular histogram methods appearing in the statistics literature also have a Bayesian rationale. The G-Enum histogram of Mendizábal et al. (2023) yields a similar criterion to the NML, based on a Bayesian generative model for the data. The Bayesian Blocks method of Scargle et al. (2013) is based on a piecewise constant model for the density and uses Bayesian change point analysis to determine both the number of change points or bins, as well as their respective break points.

5.1.1c Methods based on cross-validation

The LOOCV approach based on the $\mathbb{L}_2$ loss of Rudemo (1982) can also be used to select an irregular histogram, with a criterion similar to (2.40). The LOOCV error of a histogram based on the partition $I_{1:k}$ can be computed easily in closed form, and leads to the maximization of

$$\text{CV}_n(k, I_{1:k}) = \frac{n+1}{n^2} \sum_{j=1}^k N_j^2 / |I_j| - 2 \sum_{j=1}^k N_j / |I_j|.$$

Similarly, the CV-estimate of the Kullback–Leibler loss of Hall (1990) is easily extended to irregular partitions, and the resulting optimal partition is the maximizer of the expression

$$\text{CV}_n(k, I_{1:k}) = \sum_{j=1}^k N_j \log(N_j - 1) - \sum_{j=1}^k N_j \log |I_j|.$$

As before, we define $\text{CV}_n(k, I_{1:k}) = -\infty$ whenever $N_j \leq 1$ for at least one j .

5.1.1d Methods based on approximating the empirical distribution

The Essential Histogram of Li et al. (2020) is based on finding a simple piecewise constant density whose cumulative distribution function is a good approximation of the empirical cdf. The basis of their procedure is a confidence set of distribution functions based on likelihood ratio tests of interval probabilities. The resulting irregular histogram is chosen to be the piecewise constant density whose distribution belongs to this confidence set with the lowest number of bins. The procedure crucially depends on the level α of the confidence set, with higher values of α leading to a histogram consisting of more bins. The authors give some guidelines for choosing the parameter, with $\alpha \in \{0.1, 0.5, 0.9\}$ serving as default choices.

Another procedure that can be used to draw an irregular histogram is the Taut String method of Davies and Kovac (2004). The Taut String (TS) procedure constructs an irregular histogram with the minimal number of modes whose distribution function is at most a certain distance from the empirical distribution function of the sample. This approach differs quite a lot from methods based on decision theory, as the main concern is to find an approximation of the empirical distribution with control over the modality of the estimate rather than to estimate a hypothetical density. The Taut String histogram is somewhat unique in that it explicitly controls the modality and not the number of bins of the histogram. Due to this fact, the Taut String method typically results in a histogram with a large number of bins compared to the other approaches presented here, since these methods explicitly penalize larger numbers of bins.

We note that some of the proposed irregular schemes in the literature, such as that of Kanazawa (1988), are used to find an optimal partition of the data given the number k of bins, and are as such not fully automatic as they do not provide a default choice for how this parameter should be chosen. Since our focus is on fully automatic procedures, we have not included such methods in the above summary.

5.2 Bayesian histograms with irregular grids

In this section, we introduce an irregular histogram procedure which in many ways extends the regular random histogram of the last chapter to irregular partitions, allowing us to leverage previously gained insights in the new situation, in particular when it comes to computing posterior quantities for a given partition. Our irregular histogram method relies on conjugacy to find the optimal partition of $[0, 1]$ quickly. However, new difficulties arise from the problem of having to find the best partition based on the sample, which introduces the need for new optimization techniques.

5.2.1 Our irregular histogram prior

To keep the problem computationally tractable, we partition $[0, 1]$ into a finest possible mesh, consisting of $k_n + 1$ grid points, $\{\tau_{n,j} : 0 \leq j \leq k_n\}$, and we restrict the possible

partitions of $[0, 1]$ to use the $\tau_{n,j}$ as grid points. The maximal permitted number of intervals in a partition of $[0, 1]$ is thus k_n , which is chosen to be a nondecreasing function of the sample size. Due to the requirement that the endpoints of the intervals are chosen among the $\tau_{n,j}$, the support of the number k of bins must necessarily be a subset of $\{1, 2, \dots, k_n\}$. Thus, to have a flexible nonparametric procedure, the prior for k cannot be fixed and has to depend on the sample size, so that $k \sim p_n(k)$. Unlike the regular histogram model, the partition $I_{1:k}$ for given k is now a parameter, and we shall treat it as its own random variable, with a corresponding prior distribution. We note that since the sequence $\{\tau_{n,j}\}$ and also the number of possible partitions of size k does not stay fixed as n increases, the prior for $I_{1:k} | k$ will also depend on the sample size, so we write $I_{1:k} | k \sim p_n(I_{1:k} | k)$. Once the number k of bins and the partition $I_{1:k}$ have been fixed, we can specify a prior distribution on the bin probabilities $p_j = \int_{I_j} f(x) dx$ of the piecewise constant density f , so that $\mathbf{p} | k, I_{1:k} \sim p(\mathbf{p} | k, I_{1:k})$. The fully Bayesian model for the histogram can thus be written down in the form

$$\begin{aligned} k &\sim p_n(k), \\ I_{1:k} | k &\sim p_n(I_{1:k} | k), \\ \mathbf{p} | k, I_{1:k} &\sim p(\mathbf{p} | k, I_{1:k}), \\ f(x | k, I_{1:k}, \mathbf{p}) &= \prod_{j=1}^k \left(\frac{p_j}{|I_j|} \right)^{\mathcal{I}(x \in I_j)}, \\ x_i | f &\sim f(x_i | k, I_{1:k}, \mathbf{p}). \end{aligned} \tag{5.4}$$

This prior-model specification of (5.4) is similar to its regular counterpart (4.8), but includes an additional level of the hierarchy due to the fact that the partition $I_{1:k}$ is not given in advance once k is known. The problem of finding the best histogram now consists of finding k and a corresponding interval partition $I_{1:k} = (I_1, I_2, \dots, I_k)$ of $[0, 1]$ with $\tau_0, \dots, \tau_{k_n}$ as endpoints that maximizes the joint posterior distribution, namely we are interested in finding

$$(\hat{k}, \hat{I}_{1:\hat{k}}) = \arg \max_{k, I_{1:k}} p_n(k, I_{1:k} | \mathbf{x}_{1:n}). \tag{5.5}$$

The pair $(k, I_{1:k})$ maximizing the above is also the maximizer of the (unnormalized) log-posterior, so we may choose to work with the logarithm of the posterior distribution instead out of computational convenience. To compute the posterior probability $p_n(k, I_{1:k} | \mathbf{x}_{1:n})$ of a given partition, we can proceed as in Section 4.2.1 by first computing the marginal likelihood given the partition as in (4.10),

$$p(\mathbf{x}_{1:n} | k, I_{1:k}) = \int_{\mathcal{S}_{k-1}} p(\mathbf{x}_{1:n} | k, I_{1:k}, \mathbf{p}) p(\mathbf{p} | k, I_{1:k}) d\mathbf{p}.$$

The posterior probability of the given partition can now be computed using Bayes rule (3.17),

$$p_n(k, I_{1:k} | \mathbf{x}_{1:n}) \propto p_n(k, I_{1:k}) p(\mathbf{x}_{1:n} | k, I_{1:k}) = p_n(k) p_n(I_{1:k} | k) p(\mathbf{x}_{1:n} | k, I_{1:k}). \tag{5.6}$$

Maximizing the latter expression then yields the a posteriori most probable partition. Once the MAP $(\hat{k}, \hat{I}_{1:\hat{k}})$ has been computed, a histogram can be obtained as a point estimate of the density, based on the chosen partition. For this purpose, we will once again use squared $\mathbb{L}_2$ distance as our loss function, for reasons laid out in Remark 4.2. Using Proposition A.2, we have that

$$\|f - f(\cdot | k, I_{1:k}, \boldsymbol{\rho})\|_2^2 = \|f - f(\cdot | k, I_{1:k}, \mathbf{p})\|_2^2 + \|f(\cdot | k, I_{1:k}, \mathbf{p}) - f(\cdot | k, I_{1:k}, \boldsymbol{\rho})\|_2^2.$$

Thus, in order to minimize the posterior expected loss for given $k, I_{1:k}$, it suffices to minimize the posterior expectation of the second term. Some simplifications reveal that

$$\begin{aligned} \|f(\cdot | k, I_{1:k}, \mathbf{p}) - f(\cdot | k, I_{1:k}, \boldsymbol{\rho})\|_2^2 &= \sum_{j=1}^k \int_{I_j} \left\{ |I_j|^{-1} p_j - |I_j|^{-1} \rho_j \right\}^2 dx \\ &= \sum_{j=1}^k |I_j|^{-1} \{p_j - \rho_j\}^2. \end{aligned} \quad (5.7)$$

The Bayes estimator $\hat{\mathbf{p}}_B$ is thus the solution to

$$\arg \min_{\boldsymbol{\delta} \in \mathcal{S}_{k-1}} \sum_{j=1}^k |I_j|^{-1} \mathbb{E}_{p_j} \left\{ (p_j - \rho_j)^2 \mid \mathbf{x}_{1:n}, k, I_{1:k} \right\}.$$

Arguing as in Example 3.4, we find that the Bayes estimator is given by the posterior means, $\hat{p}_{B,j} = \mathbb{E}_{p_j} \{p_j \mid \mathbf{x}_{1:n}, k, I_{1:k}\}$ for all j . The corresponding Bayes histogram estimator is then given by

$$f(x \mid \hat{k}, \hat{I}_{1:\hat{k}}, \hat{\mathbf{p}}_B) = \sum_{j=1}^{\hat{k}} \frac{\hat{p}_{B,j}}{|\hat{I}_j|} \mathcal{I}(x \in \hat{I}_j). \quad (5.8)$$

We will now show how to compute the histogram estimator (5.8) when using Dirichlet prior distributions of the bin probabilities.

Example 5.2 (Irregular Dirichlet Histograms). The Dirichlet distribution is also a particularly attractive choice as a prior for the cell probabilities for irregular histograms. Unlike in the regular case, the prior on the cell probabilities can now also depend on the partition $I_{1:k}$. As before, we take a $p(\mathbf{p} \mid k, I_{1:k}) = \text{Dir}(\mathbf{p} \mid a\mathbf{p}_0)$ -prior for the interval probabilities for a given partition and number of cells. This is mainly to ensure that the prior distribution is conjugate, so that the MAP estimate (5.5) can be computed efficiently. To choose the prior mean $\mathbf{p}_0$, one can center the prior on a particular density function $g_0(\cdot)$ by choosing $p_{0,j} = \int_{I_j} g_0(x) dx$. As a noninformative choice, one may take $g_0(\cdot)$ to be a uniform distribution so that the prior mean is given by $p_{0,j} = |I_j|$ for all j .

As a prior distribution on the partitions for a given value of k , we choose a uniform distribution. For each value of $k \in \{1, 2, \dots, k_n\}$, there are $\binom{k_n-1}{k-1}$ possible histograms with k bins, so that the pmf of the uniform prior on the partitions becomes

$$p_n(I_{1:k} \mid k) = \binom{k_n-1}{k-1}^{-1}.$$

For a given value of k and partition $I_{1:k}$, the likelihood function for $\mathbf{p}$ of an iid sample $x_1, \dots, x_n$ is still as in (2.8):

$$L_n(\mathbf{p} \mid k, I_{1:k}) = p(\mathbf{x}_{1:n} \mid k, I_{1:k}, \mathbf{p}) = \prod_{i=1}^n \prod_{j=1}^k \left(\frac{p_j}{|I_j|} \right)^{\mathcal{I}(x_i \in I_j)} = \prod_{j=1}^k \left(\frac{p_j}{|I_j|} \right)^{N_j}.$$

The desired posterior distribution $p_n(k, I_{1:k} \mid \mathbf{x}_{1:n})$ results from a similar calculation as that for $p_n(k \mid \mathbf{x}_{1:n})$ for regular histograms. Indeed, a computation similar to (4.15)

shows that we have conjugacy for $\mathbf{p}$ for given $k, I_{1:k}$:

$$\begin{aligned} p(\mathbf{p} | k, I_{1:k}, \mathbf{x}_{1:n}) &\propto p(\mathbf{p} | k, I_{1:k}) p(\mathbf{x}_{1:n} | k, I_{1:k}, \mathbf{p}) \\ &\propto \prod_{j=1}^k p_j^{ap_{0,j}-1} \prod_{j=1}^k p_j^{N_j} \\ &\propto \text{Dir}(\mathbf{p} | a\mathbf{p}_0 + \mathbf{N}). \end{aligned} \quad (5.9)$$

Due to the conjugacy property of (5.9), we may obtain the posterior distribution of $k, I_{1:k}$ as in (4.17):

$$p_n(k, I_{1:k} | \mathbf{x}_{1:n}) \propto p_n(k) p_n(I_{1:k} | k) p(\mathbf{x}_{1:n} | k, I_{1:k}), \quad (5.10)$$

where the marginal likelihood for given $k, I_{1:k}$ can be computed via

$$\begin{aligned} p(\mathbf{x}_{1:n} | k, I_{1:k}) &= \frac{p(\mathbf{p} | k, I_{1:k}) p(\mathbf{x}_{1:n} | k, I_{1:k}, \mathbf{p})}{p(\mathbf{p} | k, I_{1:k}, \mathbf{x}_{1:n})} \\ &= \frac{\prod_{j=1}^k \Gamma(ap_{0,j} + N_j)}{\prod_{j=1}^k |I_j|^{N_j} \Gamma(ap_{0,j})} \frac{\Gamma(a)}{\Gamma(a+n)}. \end{aligned} \quad (5.11)$$

Plugging in this result in (5.10), we arrive at the following expression for the posterior:

$$p_n(k, I_{1:k} | \mathbf{x}_{1:n}) \propto \frac{p_n(k)}{\prod_{j=1}^k |I_j|^{N_j}} \frac{\prod_{j=1}^k \Gamma(ap_{0,j} + N_j)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \frac{\Gamma(a)}{\Gamma(a+n)} \binom{k_n-1}{k-1}^{-1}. \quad (5.12)$$

To estimate the cell probabilities, we use the fact that the posterior for $\mathbf{p} | k, I_{1:k}$ is Dirichlet, and hence by Proposition 3.1,

$$\hat{p}_{B,j} = \mathbb{E}_{p_j} \{p_j | \mathbf{x}_{1:n}, k, I_{1:k}\} = \frac{a}{a+n} p_{0,j} + \frac{n}{a+n} N_j/n, \quad (5.13)$$

where $N_j = \sum_{i=1}^n \mathcal{I}(x_i \in I_j)$ is the number of samples falling into interval I_j . The computation of the estimated bin probabilities is thus more or less the same as in the regular histogram model; all that is required is to count the number of observations landing in each interval. Once the Bayes estimator has been computed from (5.13), we can plug them into (5.8) to obtain the a histogram estimate of the density. $\blacktriangle$

The main difference between the regular and irregular histogram models lies in the computation of the MAP partition. The maximization of (5.12) is not as straightforward as finding the maximum of (4.18). However, by exploiting the additive structure of the log-posterior (5.10) we can compute the best partition in polynomial time by the use of a dynamical programming algorithm.

Remark 5.1. We note that until now, we have justified the use of a Dirichlet prior out of computational concerns, being able to handle the posterior analytically. However, if we let a be a fixed positive number, then it can be shown that the prior satisfies the coherence property of Section 4.7.1b. Namely, if $I_{1:k}$ is an interval partition of $[0, 1]$ and $\bar{I}_{1:k}$ a refinement of $I_{1:k}$, then the prior distribution of the bin probability $p_j = \int_{I_j} f(t) dt$ is the same under the two models. Thus, the two models express the same belief in the interval probability p_j , but there is uncertainty connected to how fine a grid one should use to model the data-generating density. As before this will not be the case if a is taken to be an increasing function of k .

5.2.2 Computing the optimal partition by dynamical programming

To compute the optimal partition with respect to (5.12), we will use a dynamical programming algorithm due to Kanazawa (1988), which we will refer to as the *KDP algorithm* from now on. This algorithm exploits the fact that the objective we are trying to maximize has a certain additive structure to it, which allows us to break the problem up into smaller subproblems, whose solutions can be combined in an efficient manner to find the global optimum. To show that our optimization problem has this particular structure, we introduce some new notation. Denote

$$\begin{aligned}\Phi_k(I_{1:k}) &= \sum_{j=1}^k \{ \log \Gamma(ap_{0,j} + N_j) - \log \Gamma(ap_{0,j}) - N_j \log |I_j| \}, \\ \Psi_n(k) &= \log p_n(k) + \log \Gamma(a) - \log \Gamma(a + n) - \log \binom{k_n - 1}{k - 1}.\end{aligned}$$

Taking the logarithm of (5.12), we find that the log-posterior for $(k, I_{1:k})$ can be decomposed as

$$\log p_n(k, I_{1:k} | \mathbf{x}_{1:n}) = \Phi_k(I_{1:k}) + \Psi_n(k) + C_n, \quad (5.14)$$

where $C_n \in \mathbb{R}$ is a constant depending on the sample $\mathbf{x}_{1:n}$, but not on $k, I_{1:k}$. Denoting $\Phi_0(I_j) = \log \Gamma(ap_{0,j} + N_j) - \log \Gamma(ap_{0,j}) - N_j \log |I_j|$, we see that $\Phi_k(\cdot)$ is additive with respect to the intervals $I_{1:k}$ in the sense that

$$\Phi_k(I_{1:k}) = \sum_{j=1}^k \Phi_0(I_j). \quad (5.15)$$

The above equality is additivity property that allows for the use of the KDP algorithm as a method of finding the MAP partition. We note that this property also holds for the criteria from the penalized likelihood and cross-validation based approaches to irregular histogram construction of Section 5.1.1, for suitable choices of the functions Φ_k, Φ_0 and Ψ_n . As such, the dynamical programming algorithm detailed below can also be used to find the best partition with respect to these criteria as well. The algorithm is used to solve the following problem: for each $k \in \{1, 2, \dots, k_n\}$, find the optimal interval partition $I_{1:k}$ of $[0, 1]$ consisting of k intervals. The dynamical programming algorithm solves this problem in polynomial time, with a complexity of $\mathcal{O}(k_n^3)$ (Rozenholc et al., 2010). We will now give a detailed description of the KDP algorithm, along with a formal proof of its correctness. A detailed description of the algorithm is also given in section 4 of Gather et al. (2009).

5.2.2a Details of the dynamical programming algorithm

As the penalty term $\Psi_n(k)$ only depends on k and not the chosen partition given k , if we are able to compute the partition maximizing $\Phi_k(I_{1:k})$ for each value of k , then the maximum of (5.14) by tabulating this criterion for each value of k and choosing the value of k and the corresponding partition maximizing this criterion to be the one we base our histogram estimate on. Thus, we turn our focus to maximizing $\Phi_k(I_{1:k})$ for $k = 1, 2, \dots, k_n$ first. To keep notation simple, we shall reformulate this problem slightly. Given the grid points $\{\tau_{n,j} \in [0, 1] : 0 \leq j \leq k_n\}$, the problem of finding the optimal partition consisting of k intervals is seen to be equivalent to choosing which $k - 1$ grid points from $\{\tau_{n,j}\}$ to use in building the partition. We are thus interested in finding the

Algorithm 1 The KDP algorithm

Require: A table of values $U(i, j)$ for $0 \leq i < j \leq k_n$.

```

for  $i = 1, 2, \dots, k_n$  do
     $V_1(i) = U(0, i)$ 
end for
for  $k = 2, 3, \dots, k_n$  do
    for  $j = k, k + 1, \dots, k_n$  do
         $i_k(j) = \arg \max_{k-1 \leq i \leq j-1} \{V_{k-1}(i) + U(i, j)\}$ 
         $V_k(j) = V_{k-1}(i_k(j)) + U(i_k(j), j)$ 
    end for
end for
return  $\{V_k(k_n)\}, \{i_k(j)\}$  for each  $k, j$ 

```

subset $\{a_1, a_2, \dots, a_{k-1}\} \subseteq \{1, 2, \dots, k_n - 1\}$ (where $a_0 = 0$, $a_k = k_n$ and $a_j < a_{j+1}$ for all j) such that

$$\sum_{l=0}^{k-1} \Phi_0((\tau_{a_l}, \tau_{a_{l+1}}])$$

is maximized. For any pair $0 \leq i < j \leq k_n$, we define $U(i, j) = \Phi_0((\tau_i, \tau_j])$. This allows us to restate the problem of finding the indices corresponding to the optimal partition as

$$\max_{A \in \mathcal{A}_{k_n-1}^{k-1}} \sum_{l=0}^{k-1} U(a_l, a_{l+1}), \quad (5.16)$$

where $\mathcal{A}_{j-1}^{k-1}$ is the set consisting of all subsets of $\{1, 2, \dots, j-1\}$ of cardinality $k-1$, and $a_1, \dots, a_{k-1}$ is the increasing enumeration of A . We define $a_0 = 0$ and $a_{k_n} = j$ for any $A \in \mathcal{A}_{j-1}^{k-1}$, and put $V_1(i) = U(0, i)$.

The KDP algorithm is given in full detail as Algorithm 1. We also offer our own proof of its correctness in Theorem 5.1.

Theorem 5.1. *The output of Algorithm 1 satisfies, for $k = 1, 2, \dots, k_n$,*

$$V_k(k_n) = \max_{I_{1:k}} \Phi_k(I_{1:k}).$$

Moreover for each k the maximizer of (5.16), $A = \{a_1, a_2, \dots, a_{k-1}\}$, is given by the recursion

$$\begin{aligned} a_k &= k_n \\ a_j &= i_{j+1}(a_{j+1}), \quad j = k-1, k-2, \dots, 2 \end{aligned}$$

Proof of Theorem 5.1. We claim that

$$V_k(j) = \max_{A \in \mathcal{A}_{j-1}^{k-1}} \sum_{l=0}^{k-1} U(a_l, a_{l+1}) \quad (5.17)$$

for all $k = 1, 2, \dots, k_n$ and $k \leq j \leq k_n$. If we can prove that equation (5.17) holds, then the first part of the theorem follows by taking $j = k_n$ and noting the one-to-one correspondence between the indices in A and the partition $I_{1:k}$. To show that this is indeed the case, we proceed by induction.

For the base case $k = 1$, the sum in (5.17) reduces to $V_1(j) = U(0, j) = U(a_0, a_1)$, which is the only possibility for this value of k , agreeing with the induction hypothesis.

Now assume that (5.17) holds for some $1 \leq k < k_n$. Let j satisfy $k + 1 \leq j \leq k_n$, and let $A \in \mathcal{A}_{j-1}^k$, with enumeration $a_1, a_2, \dots, a_k$. Note that this implies that $a_k \geq k$. Put $a_{k+1} = j$. We can thus write

$$\begin{aligned} \sum_{l=0}^k U(a_l, a_{l+1}) &= \sum_{l=0}^{k-1} U(a_l, a_{l+1}) + U(a_k, a_{k+1}) \\ &\leq \max_{B \in \mathcal{A}_{a_k-1}^{k-1}} \sum_{l=0}^{k-1} U(b_l, b_{l+1}) + U(a_k, a_{k+1}) \\ &= V_k(a_k) + U(a_k, a_{k+1}), \end{aligned}$$

where the inequality follows from the fact that $\{a_1, a_2, \dots, a_{k-1}\} \in \mathcal{A}_{a_k-1}^{k-1}$ as $a_k \geq k$, and the last equality is due to the induction hypothesis. Now, using the definition of V_{k+1} , we see that

$$\sum_{l=0}^k U(a_l, a_{l+1}) \leq V_k(a_k) + U(a_k, a_{k+1}) \leq \max_{k \leq i \leq j-1} \{V_k(i) + U(i, j)\} = V_{k+1}(j).$$

Since $A \in \mathcal{A}_{j-1}^k$ was arbitrary, we conclude that

$$\max_{A \in \mathcal{A}_{j-1}^k} \sum_{l=0}^k U(a_l, a_{l+1}) \leq V_{k+1}(j)$$

for $j \geq k + 1$, so that $V_{k+1}(j)$ is an upper bound on the maximum. Since this upper bound is clearly attainable, we conclude that

$$V_{k+1}(j) = \max_{A \in \mathcal{A}_{j-1}^k} \sum_{l=0}^k U(a_l, a_{l+1}). \quad (5.18)$$

The second part of the claim is immediate from the fact that $a_k = k_n$ and the definition of $i_k(j)$. $\square$

Since Algorithm 1 finds the partition maximizing $\Phi_k(\cdot)$ for each k , the solutions can be used to compute the MAP partition in a straightforward manner. This is accomplished by utilizing the fact that due to Theorem 5.1,

$$\max_{k, I_{1:k}} \log p_n(k, I_{1:k} | \mathbf{x}) = \max_{k, I_{1:k}} \{\Phi_k(I_{1:k}) + \Psi_n(k) + C_n\} = \max_k \{V_k(k_n) + \Psi_n(k) + C_n\},$$

Performing the maximization on the right hand side for k is a simple task, all that needs to be done is to compute $V_k(k_n) + \Psi_n(k)$ for $k = 1, 2, \dots, k_n$, and take $\hat{k}$ to be the value maximizing this expression. To find the MAP partition $\hat{I}_{1:\hat{k}}$ corresponding to $\hat{k}$, we utilize the recursive scheme of Theorem 5.1, starting at $a_k = \hat{k}$.

5.2.2b A greedy search heuristic

Although the KDP algorithm finds the best possible partition in polynomial time, the algorithm can still be quite computationally demanding, especially for larger sample sizes. For instance, if one wants to use each unique value in the sample $\mathbf{x}_{1:n}$ as a possible

endpoint in the grid $\{\tau_{n,i}: 0 \leq i \leq n\}$, the computational complexity is of order $\mathcal{O}(n^3)$, which is problematic for larger sample sizes.

To reduce the computational burden, we follow the proposal of Rozenholc et al. (2010), selecting a subset consisting of $q_n \ll k_n$ grid points $\{t_{n,j}: 0 \leq j \leq q_n\}$ of the full grid $\{\tau_{n,j}: 0 \leq j \leq k_n\}$ in a preprocessing step, and then running the KDP algorithm on this subset. The runtime complexity of the KDP algorithm will now be $\mathcal{O}(q_n^3)$, so by choosing q_n to be of order much lower than n , we get a computationally tractable optimization problem. For our code, we settled on an order of $q_n \sim (n \log(n))^{1/3}$. This leads to a runtime complexity of order $\mathcal{O}(n \log(n))$ for the dynamic programming algorithm. We felt like this choice of q_n is appropriate, since most irregular histogram methods require sorting the input data, and this operation has a well-known runtime complexity of order $n \log(n)$.

To construct the reduced grid¹ $\{t_{n,j}: 0 \leq j \leq q_n\}$, we use the greedy search heuristic of Rozenholc et al. (2010), which we will now briefly describe. The algorithm proceeds by starting with the grid points $\{0, 1\}$, corresponding to a histogram consisting of one bin. Suppose that the grid points included in the partition at iteration l are $\{t_{n,i}: 0 \leq i \leq l\}$. The iteration proceeds by identifying the grid point $\tau_{n,j}$ maximizing the log-likelihood of the histogram based on the partition including every grid point in $\{\tau_{n,j}\} \cup \{t_{n,i}: 0 \leq i \leq l\}$. The maximizer $\tau_{n,j}$ is then subsequently added to the grid, and we proceed with the next iteration. We repeat this procedure until the partition consists of $q_n + 1$ points, including the endpoints 0 and 1. In constructing the smaller partition, we perform a greedy maximization of the log-likelihood function $L_n(\hat{\mathbf{p}} | k, I_{1:k}) = \sum_{j=1}^k N_j \log(\hat{p}_j / |I_j|)$. Since we are maximizing a penalized variant of (5.15) rather than a penalized maximum likelihood, it might have been more natural to apply the greedy search heuristic to $\Phi(\cdot)$ directly instead. However, by we discovered through simulations that maximizing the log-likelihood to build the finest grid typically yielded solutions that were closer to the exact optimal solution of (5.5) obtained by running the dynamic programming algorithm on the finest grid altogether.

Example 5.3 (Irregular histogram estimate). To illustrate the irregular histogram procedure in practice, we repeated the simulation experiment of Example 4.2 by generating two iid samples from the normal mixture shown in Figure 1.1 of size $n = 1000$ and $n = 100000$. For our Dirichlet prior distributions we used $a = 1$ and took g_0 to be the uniform distribution. As potential grid points, we used the sample values $\mathbf{x}_{1:n}$. The results are shown in Figure 5.1.

For the sample of size $n = 1000$, the MAP histogram selected a partition based on $\hat{k} = 5$ intervals. The resulting density estimate fails to show the two peaks of the density; both modes are covered by a single large bin. This compares unfavorably with the corresponding regular histogram of Figure 4.3, which was able to capture the bimodal structure of the true density. The Hellinger loss is also 14.7% larger for the irregular histogram than for its regular counterpart. In the larger sample, the number of bins chosen was $\hat{k} = 29$. For the large sample, the resulting density estimate is better able to capture the shape of the true distribution. In this case, the procedure is able to leverage some of the advantages of using an irregular grid. The bins are narrower in regions where the true density is changing rapidly, while larger bins are preferred in regions where the density is more flat, such as in the tails of the density and near the three critical points. The Hellinger loss still favors the regular histogram but the difference is smaller than was the case for $n = 1000$, with the irregular histogram producing a loss approximately 6.5% larger.

¹For notational purposes, we do not assume that the $t_{n,j}$ are ordered increasingly here.

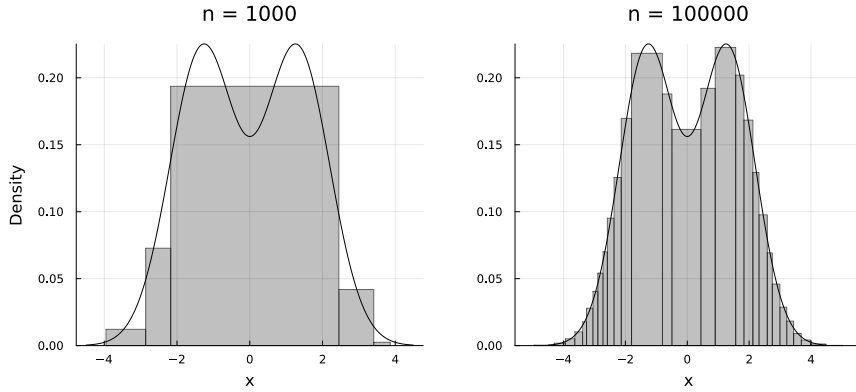


Figure 5.1: Irregular random histogram model fit to two random samples from a symmetric mixture of normals of size $n = 1000$ (left) and $n = 100\,000$ (right).

We note that the number of bins selected in each example was smaller than that chosen in the regular histogram model. This is due to the presence of the additional factor $\binom{k_n-1}{k-1}^{-1}$ in (5.12), which places an additional penalty on larger values of k compared to the regular histogram criterion (4.18). Although the bias of a well-constructed irregular histogram is lower than a regular histogram for a given value of k , the additional penalty introduced by having to choose the partition based on the data often results in choosing a lower value of k . Thus, the bias of the irregular histogram may end up being larger than for the corresponding histogram in practice. This example illustrates well the fact that even though irregular histograms based on an optimally chosen irregular grid have lower bias in theory than a regular histogram using the same number of bins, the search for a good data-based partition can lead to worse density estimates in practice. ▲

5.2.3 Computing the posterior mean and uncertainty quantification

As noted in Remark 4.3, the posterior mean of f provides a more natural estimate of the density in the Bayesian framework than the one based on the MAP for $(\hat{k}, \hat{I}_{1:\hat{k}})$ from a decision-theoretic perspective. When using a Dirichlet prior on $\mathbf{p} | k, I_{1:k}$ and a prior for k supported on $\{1, 2, \dots, k_n\}$, the Bayes estimator may in principle be computed exactly via the formula

$$\mathbb{E}_f\{f(x) | \mathbf{x}_{1:n}\} = \sum_{k=1}^{k_n} \sum_{I_{1:k} \in \mathcal{P}_{n,k}} f(x | k, I_{1:k}, \hat{\mathbf{p}}_B) p_n(k, I_{1:k} | \mathbf{x}_{1:n}), \quad (5.19)$$

where $\mathcal{P}_{n,k}$ denotes the set of interval partitions of $[0, 1]$ with endpoints in $\{\tau_{n,j} : 0 \leq j \leq k_n\}$ consisting of k intervals. However, the expression (5.19) is intractable in practice even for moderate values of k_n , as it is a sum consisting of 2^{k_n-1} terms. Direct simulation from the posterior is also infeasible from a computational standpoint, since $p_n(k, I_{1:k} | \mathbf{x}_{1:n})$ is only known up to proportionality, and hence drawing samples $(k, I_{1:n})$ from directly from their posterior distribution would involve tabulating the expression for $p_n(k, I_{1:k} | \mathbf{x}_{1:n})$ at each of the 2^{k_n-1} possible values, incurring an initial computational burden equal to that of evaluating (5.19). This latter fact also means that computing credible intervals is more difficult than in the regular case, since implementing a simple Monte Carlo scheme is too computationally demanding.

In order to compute the Bayes estimator and credible intervals in a reasonable amount of time, we propose an MCMC algorithm that allows us to draw approximate

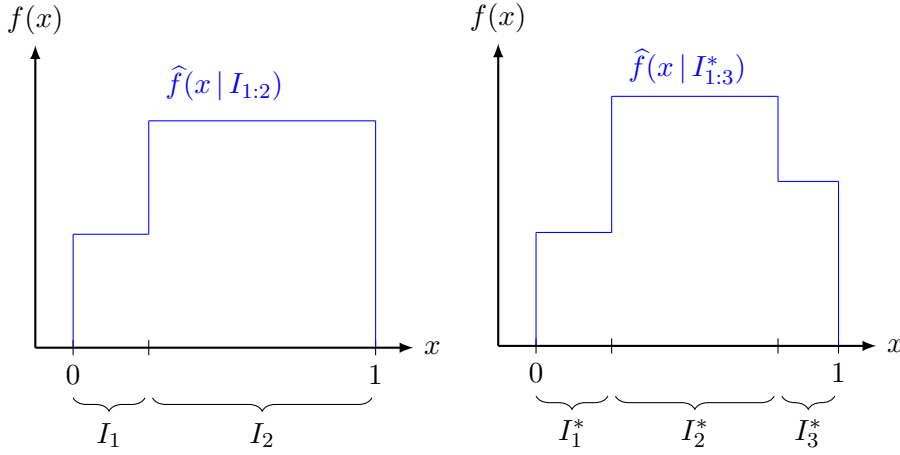


Figure 5.2: Illustration of a proposed split move. The current state is shown on the left and the proposed state on the right. The posterior mean conditional on the chosen partition is shown in blue.

samples of $(k, I_{1:k})$ from the posterior distribution (5.12). Our MCMC algorithm is a Metropolis–Hastings scheme consisting of three types of proposed moves: splits, merges and relocations. The two first types of moves are dimension-changing, incrementing the number of bins by 1, while the last move type only relocates the position of an endpoint of a bin included in the partition.

5.2.3a Split and merge moves

Given an interval partition $I_{1:k}$ of $[0, 1]$, the split move chooses one of the constituent intervals, say I_j , and splits it into two disjoint subintervals I_j^* and I_{j+1}^* , resulting in a new interval partition $I_{1:k+1}^*$ where $I_l^* = I_l$ for $l < j$ and $I_{l+1}^* = I_l$ for $l > j$. The merge move effectively reverses the action of the corresponding split move, by choosing an interval I_j for some $j \leq k - 1$ from the partition $I_{1:k}$ to be merged with the interval neighboring it to the right, yielding a proposed interval partition $I_{1:k-1}^*$ consisting of $I_l^* = I_l$ for $l < j$ and $I_{l-1}^* = I_l$ for $l > j$. Figure 5.2 illustrates a split move in a hypothetical example.

In Figure 5.2 the current state is the interval partition $I_{1:2}$, and the interval I_2 has been chosen for a proposed split, forming the two new subintervals I_2^* and I_3^* . We note that the interval $I_1^* = I_1$ is left unchanged by the proposed move, so that the Bayes estimator of the bin probability p_1 remains the same, assuming that a does not change as a function of k .

We now describe the split move in greater detail. Given the current partition $I_{1:k}$ the split move starts by sampling j uniformly from $\{1, 2, \dots, k\}$. Once interval j has been chosen, we proceed by sampling a partition of I_j into two new subintervals I_j^* and I_{j+1}^* according to proposal distribution

$$q(I_j^*, I_{j+1}^* | I_j) \propto \frac{\Gamma(ap_{0,j}^* + N_j^*)}{|I_j^*|^{N_j^*} \Gamma(ap_{0,j}^*)} \frac{\Gamma(ap_{0,j+1}^* + N_{j+1}^*)}{|I_{j+1}^*|^{N_{j+1}^*} \Gamma(ap_{0,j+1}^*)}.$$

Note that the requirement that I_j^*, I_{j+1}^* form a partition of I_j means that we only need to determine the right endpoint of I_j^* as this is also the left endpoint of I_{j+1}^* .

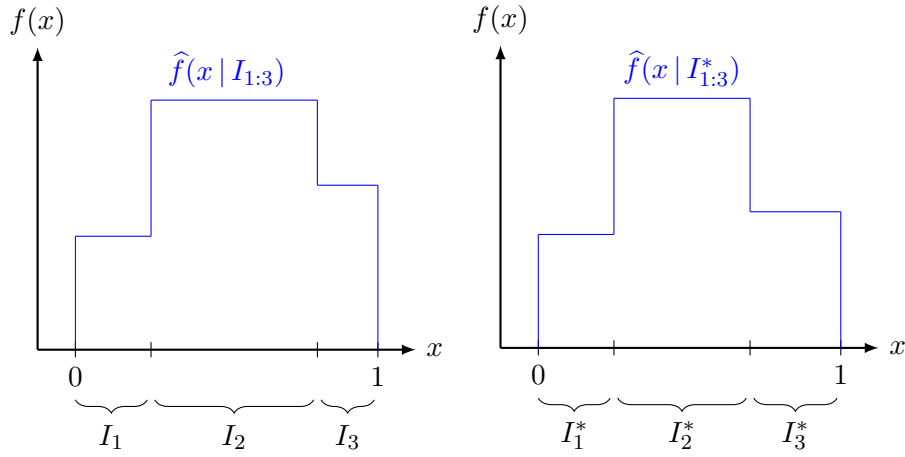


Figure 5.3: Illustration of a proposed relocation move. The current state is shown on the left and the proposed state on the right. The posterior mean conditional on the chosen partition is shown in blue.

5.2.3b Relocation moves

The relocation move chooses a right endpoint of one of the intervals included in the partition (except for the rightmost interval) and proposes a new value for the endpoint that lies to the right of the left endpoint in the same interval, and to the left of the right endpoint of the interval to the right of this one. This move type is included to ensure better mixing for partitions when the value of k is fixed, as moving the endpoints of an interval in a partition through merges and splits only can be a rather slow process. Given the current partition $I_{1:k}$, the relocation move proceeds by selecting an index j in $\{1, 2, \dots, k-1\}$ uniformly at random. We then propose a new location for the right endpoint uniformly among the possible endpoints $\tau_{k_n, m}$ that lie between the left endpoint of I_j and the right endpoint of I_{j+1} , leading to the proposed intervals I_j^* and I_{j+1}^* . All the other intervals in the current partition remain unchanged in the proposed partition, so that $I_l^* = I_l$ for all $l \notin \{j, j+1\}$. The Metropolis–Hastings ratio for this move then becomes

$$\frac{\Gamma(ap_{0,j}^* + N_j^*)}{|I_j^*|^{N_j^*} \Gamma(ap_{0,j}^*)} \frac{\Gamma(ap_{0,j+1}^* + N_{j+1}^*)}{|I_{j+1}^*|^{N_{j+1}^*} \Gamma(ap_{0,j+1}^*)} \frac{|I_j|^{N_j} \Gamma(ap_{0,j})}{\Gamma(ap_{0,j} + N_j)} \frac{|I_{j+1}|^{N_{j+1}} \Gamma(ap_{0,j+1})}{\Gamma(ap_{0,j+1} + N_{j+1})},$$

due to the fact that the factors involving the other intervals cancel in the numerator and the denominator. An illustration of a proposed relocation move is shown in Figure 5.3.

We implemented the MCMC algorithm in a random scan form, by selecting a relocation move with probability 0.5, and otherwise selecting between a split or a merge move with equal probability.

Example 5.4. We return to the simulated data of Example 5.3, with a focus on constructing credible bands for the density estimate and validating the corresponding output from our Metropolis–Hastings algorithm. For each sample size $n \in \{1000, 100\,000\}$, we ran four chains with different starting values for the number k of bins, with a random initialization of the partition $I_{1:k}$ in each case. Each chain had a length of $T = 15\,000$, and we used the first $D = 5000$ samples as burn-in. To decorrelate the output from the chains, we used thinning, keeping the output of every 5th iteration of the algorithm. The resulting histogram density estimate along with credible bands

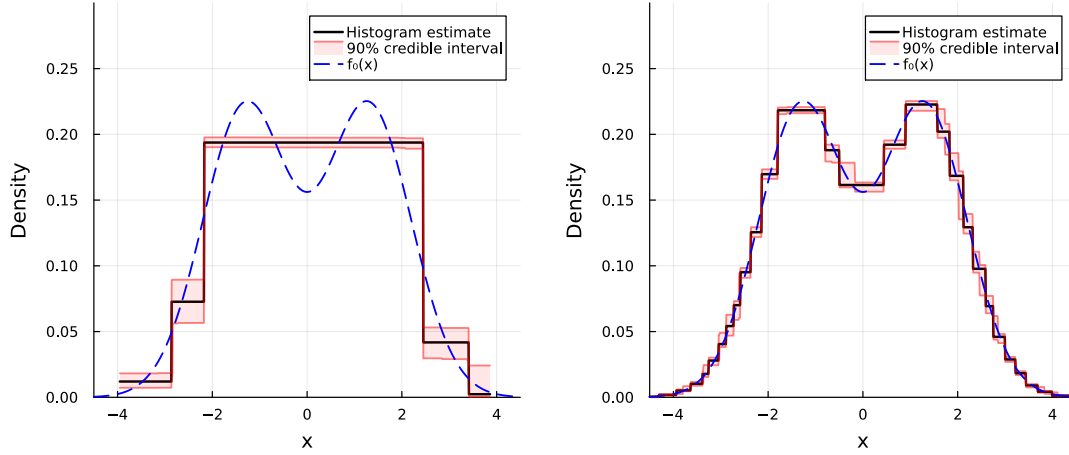


Figure 5.4: Histogram estimates along with pointwise 90% credibility intervals for f in a simulated example. The left plot corresponds to the sample of size $n = 1000$ and the right to the sample of size $n = 100\,000$.

are shown in Figure 5.4, and the corresponding MCMC diagnostics can be found in Appendix D.2.

The resulting credible intervals exhibit sharp discontinuities, due to the discontinuous nature of the model used for the data. The discontinuities are for the most part aligned with those of the histogram estimate, especially for the smaller sample size. This reflects the fact that the posterior mass appears to be highly concentrated on a few very similar partitions in this case. This was also reflected in the estimated posterior mean, which was very similar to the histogram estimate. For the larger sample size, the contributions from the non-MAP partitions are more noticeable, with significant shifts in the credible bands also occurring at points where the histogram estimate is continuous, reflecting that the posterior distribution on the partitions is more spread out in this case. ▲

5.3 Posterior Consistency of Bayesian irregular histograms

The proposed random irregular histogram method of the previous section is a novel Bayesian nonparametric construction, and as such it is of interest to derive consistency and convergence rate results for this model. The proposed method of Section 5.2 shares a lot of similarities with the regular random histogram model of Chapter 4, allowing us to reuse some of the results we previously developed here. However, the fact that there is now more than one possible model for each value of k complicates matters in the irregular setting, leading us to take a slightly different route towards proving posterior consistency.

Recall that the posterior probability of a set $\mathcal{A}$ is given by

$$P_n(\mathcal{A} | \mathbf{x}_{1:n}) = \frac{\int_{\mathcal{A}} \prod_{i=1}^n f(x_i) dP_n(f)}{\int_{\mathcal{F}_n} \prod_{i=1}^n f(x_i) dP_n(f)} = \frac{\int_{\mathcal{A}} \prod_{i=1}^n \frac{f(x_i)}{f_0(x_i)} dP_n(f)}{\int_{\mathcal{F}_n} \prod_{i=1}^n \frac{f(x_i)}{f_0(x_i)} dP_n(f)}. \quad (5.20)$$

To show posterior consistency, i.e. that the above expression converges to 0 for sets of the form $\mathcal{A} = \{f: d(f_0, f) \geq 2\epsilon\}$ for a suitable metric d , the techniques typically employed handle the numerator and denominator separately. The conditions (3.39) and (3.40) take care of the numerator, while the Kullback–Leibler property (3.31) is usually used

to take care of the denominator. However, it turns out that this last condition is a little stronger than what is really necessary. In this section, we pursue a strategy similar to that employed by Ghosh and Ramamoorthi (2003) to the Dirichlet density process of Gasparini (1996), showing posterior consistency in this model without the use of the KL support condition. Establishing the KL property becomes difficult in the setting of the irregular histogram prior, as the grid used is allowed to change with n . Due to the fact that k_n is in general growing with n , the prior probability of any given partition goes to zero, which makes it more difficult to show (3.31) directly than in the regular histogram setting. However, if we can establish that for any $\delta > 0$, the sequence of prior distributions $\{P_n\}$ supported on $\mathcal{F}$ satisfy

$$e^{n\delta} \int_{\mathcal{F}} \prod_{i=1}^n \frac{f(x_i)}{f_0(x_i)} dP_n(f) \xrightarrow{\mathbb{P}} \infty, \quad (5.21)$$

then posterior consistency follows provided the additional conditions (3.39) and (3.40) hold. Since we are working with a conjugate prior, this becomes possible under suitable assumptions on the true density and the hyperparameters of the prior. The proof of this claim follows by suitable modifications of the proof of Ghosal and van der Vaart (2017, Theorem 6.17). Since the proof is not explicitly given there, we fill in the details in the proof of Theorem 5.2.

To characterize the sequence of irregular random histogram priors $\{P_n\}$ considered in this section, we appeal to the hierarchical scheme of (5.4). For a given sample size n , the prior distribution P_n is determined by the prior distributions $p_n(k)$, $p_n(I_{1:k} | k)$ and $p(\mathbf{p} | k, I_{1:k})$ by the relation

$$P_n(f \in \mathcal{A}) = \sum_{k=1}^{k_n} \sum_{I_{1:k} \in \mathcal{P}_{n,k}} p_n(k) p_n(I_{1:k} | k) P(f \in \mathcal{A} | k, I_{1:k}) \quad (5.22)$$

where $\mathcal{A}$ is a measurable set, $\mathcal{P}_{n,k}$ is the set of all interval partitions of the unit interval with endpoints in $\{\tau_{n,j} : 1 \leq j \leq k_n\}$ and

$$P(f \in \mathcal{A} | k, I_{1:k}) = \int_{\mathcal{S}_{k-1}} \mathcal{I}(f(\cdot | k, I_{1:k}, \mathbf{p}) \in \mathcal{A}) p(\mathbf{p} | k, I_{1:k}) d\mathbf{p}.$$

The characterization of the prior in (5.22) is similar to that in (4.23), except that there is an extra component to the prior due to the fact that there can be more than one partition for a given value of k .

Before we state and prove the main result of the section, we provide the assumptions needed to establish the theorem.

Assumption 5.1. For the irregular histogram prior sequence $\{P_n\}$ the following holds:

- (i) The sequence of prior probability mass functions for k , $p_n(k)$, is fully supported on $\{1, 2, \dots, k_n\}$, where the sequence $\{k_n\}$ is nondecreasing, $k_n = o(n/\log(n))$ and $k_n \rightarrow \infty$. Moreover, the prior satisfies

$$\max_{1 \leq k \leq k_n} n^{-1} \log p_n(k) \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (5.23)$$

- (ii) Given k and a partition $I_{1:k}$, the prior for $\mathbf{p} | k, I_{1:k}$ is $\text{Dir}(\mathbf{a}_k)$ where $0 < a_{k,j} \leq A$ for a constant $A > 0$ and all k, j .

(iii) The sequence of partitions satisfy

$$\max_{1 \leq j \leq k_n} \{\tau_{n,j} - \tau_{n,j-1}\} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (5.24)$$

(iv) The prior for $I_{1:k} | k$ is uniform.

The key difference from Assumption 4.1 is that we now need to ensure that the bins become arbitrarily small as $n \rightarrow \infty$. In addition, the first assumption (5.23) is somewhat weaker than the corresponding assumption in the regular case, which is due to the fact that we are not using the KL property to establish consistency. For example, if $p_n(k)$ is the uniform distribution of $\{1, 2, \dots, k_n\}$, then (5.23) holds, but (4.24) does not.

Our consistency result is the following:

Theorem 5.2. *Suppose Assumption 5.1 holds. Then the posterior distribution is Hellinger consistent at any bounded and a.e. continuous density $f_0: [0, 1] \rightarrow \mathbb{R}$.*

To prove the theorem, we pursue the strategy outlined in the beginning of the section. We first need to handle the numerator in (5.20) which requires estimating the covering numbers of a sequence of sets $\mathcal{F}_n$. Before we proceed with proofs, we establish some new notation. For any $k \in \{1, \dots, k_n\}$, let $I_{1:k}$ denote an interval partition of $[0, 1]$ with endpoints in $\{\tau_{n,j}: 0 \leq j \leq k_n\}$. Define

$$\mathcal{H}_k = \left\{ f: [0, 1] \rightarrow \mathbb{R}: f(x) = \sum_{j=1}^k |I_j|^{-1} p_j \mathcal{I}(x \in I_j) \text{ for some } I_{1:k} \in \mathcal{P}_{n,k} \text{ and } \mathbf{p} \in \mathcal{S}_{k-1} \right\} \quad (5.25)$$

to be the set of possible histogram densities using k bins based on partitions with endpoints in $\{\tau_{n,j}\}$.

To prove that the posterior converges at a given rate, the entropy condition (3.42) requires that the covering number of the model does not grow too quickly. Lemma 5.1 estimates the covering number of $\mathcal{H}_k$ by bounding it in terms of the covering number of the simplex, in a similar fashion to Lemma 4.5.

Lemma 5.1. *Let $\{s_n\}$ be a sequence of natural numbers such that $s_n \leq k_n$ for all n and let $\mathcal{F}_n = \cup_{k=1}^{s_n} \mathcal{H}_k$. Then the Hellinger ϵ -covering number of $\mathcal{F}_n$ satisfies*

$$N(\epsilon/2, \mathcal{F}_n, d_H) \leq \left(\frac{Lk_n}{\epsilon s_n} \right)^{s_n}, \text{ for every } \epsilon \leq 2 \quad (5.26)$$

for some constant $L > 0$.

Proof. Note that for any $k \leq s_n$, a histogram density based on the partition $I_{1:k}$ can be written as

$$f(x | k, I_{1:k}, \mathbf{p}_k) = \sum_{j=1}^k |I_j|^{-1} p_j \mathcal{I}(x \in I_j) \text{ for } x \in [0, 1]$$

Now by choosing $I_{1:s_n}$ and $\tilde{\mathbf{p}}_{s_n} \in \mathcal{S}_{s_n-1}$ appropriately, we have that

$$f(x | k_n, I_{1:s_n}, \tilde{\mathbf{p}}_{s_n}) = f(x | k, I_{1:k}, \mathbf{p}_k)$$

for each $x \in [0, 1]$ due to the fact that for any partition $I_{1:k}$, we can find a refinement using s_n bins, $I_{1:s_n}$. We conclude that $\mathcal{H}_k \subseteq \mathcal{H}_{s_n}$, and as such, $\mathcal{F}_n = \cup_{k=1}^{s_n} \mathcal{H}_k \subseteq \mathcal{H}_{s_n}$. Thus, it suffices to bound the covering number of $\mathcal{H}_{s_n}$. To estimate the bracketing number of $\mathcal{H}_{s_n}$ we argue as in the proof of Lemma 4.5 by bounding it in terms of the

Hellinger bracketing number of the simplex $\mathcal{S}_{s_n-1}$. Let $\mathcal{H}(I_{1:k})$ be the set of piecewise constant densities based on the partition $I_{1:k}$ as in (4.36), and let $\mathcal{P}_{n,k}$ denote the set of partitions of $[0, 1]$ by k intervals with endpoints in $\{\tau_{n,j} : 0 \leq j \leq k_n\}$. By construction, we have that

$$\mathcal{H}_{s_n} = \bigcup_{I_{1:s_n} \in \mathcal{P}_{n,s_n}} \mathcal{H}(I_{1:s_n}).$$

By (4.37), the bracketing number of $\mathcal{H}(I_{1:k})$ is bounded by the bracketing number of $\mathcal{S}_{k-1}$. We note that there are $\binom{k_n-1}{s_n-1}$ possible partitions of $[0, 1]$ of size s_n . Using the union bound of Proposition 3.2, we have for any $\epsilon \leq 1$ that

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

Using the bound of Lemma B.2 together with Proposition 3.3, we find that

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

where we used that $\binom{k_n-1}{s_n-1} \leq \binom{k_n}{s_n} \leq (ek_n/s_n)^{s_n}$. Now an argument among the lines of that in the proof of Theorem 4.2 yields

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

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

Lemma 5.1 together with our assumption on the prior $p_n(k)$ ensures that the two conditions (3.39) and (3.40) are met, so it remains to verify that (5.21) holds, which is shown in Lemma 5.4. To show that this lemma holds, we need to establish a few helpful results first.

Lemma 5.2. *Let $f_{0,I_{1:k}}$ be the piecewise constant version of f_0 based on the partition $I_{1:k}$, and denote $\pi_j = \int_{I_j} f_0(x) dx$. Then provided $k_n = o(n)$, we have, for all $k \leq k_n$,*

$$n^{-1} \sum_{i=1}^n \log \frac{f_0(x_i)}{f_{0,I_{1:k}}(x_i)} - K(f_0, f_{0,I_{1:k}}) = o_{\mathbb{P}}(1). \quad (5.27)$$

Proof. Since $\mathbb{E}_{x_i} \{\log f_0(x_i)/f_{0,I_{1:k}}(x_i)\} = K(f_0, f_{0,I_{1:k}})$, it suffices to show that $\text{Var}_{x_i} \{\log f_0(x_i)/f_{0,I_{1:k}}(x_i)\} = \mathcal{O}(k_n)$, and Markov's inequality will yield the result. Since $\text{Var}_y(y) \leq \mathbb{E}_y\{y^2\}$ for any random variable y , this is implied if we can show that

$$\int_0^1 f_0(x) \log^2 \left\{ \frac{f_0(x)}{f_{0,I_{1:k}}(x)} \right\} dx = \mathcal{O}(k_n). \quad (5.28)$$

To estimate the integral in (5.28), we note that this is bounded from above by

$$2 \int_0^1 f_0(x) \log^2 f_0(x) dx + 2 \int_0^1 f_0(x) \log^2 f_{0,I_{1:k}}(x) dx, \quad (5.29)$$

where we used that $(a+b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$ (Proposition A.7). The first term in the above can be bounded independently of k . To see why, note that the integral is taken over a finite interval, and it thus suffices to show that the integrand is bounded. The boundedness of $f_0(x) \log^2 f_0(x)$ is ensured by the fact that for $M = \sup_{x \in [0,1]} f_0(x)$, the map $h: [0, M] \rightarrow \mathbb{R}$ defined by $h(x) = x \log^2(x) \mathcal{I}(x > 0)$ is continuous (the limit at 0 is 0). By the Extreme Value Theorem, h is bounded on its domain, ensuring that the integrand is bounded. To bound the second integral in (5.29), we note that since $f_{0,I_{1:k}}$ is piecewise constant,

$$\int_0^1 f_0(x) \log^2 f_{0,I_{1:k}}(x) dx = \sum_{j=1}^k \pi_j \log^2 (\pi_j / |I_j|)$$

To finish the proof, it suffices to show that this display is $\mathcal{O}(k_n)$. Applying the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ once again, we find that

$$\sum_{j=1}^k \pi_j \log^2 (\pi_j / |I_j|) \leq 2 \sum_{j=1}^k \pi_j \log^2 (\pi_j) + 2 \sum_{j=1}^k \pi_j \log^2 (|I_j|)$$

Defining $h: [0, 1] \rightarrow \mathbb{R}$ as before, a continuity argument reveals that the summands of the first term are bounded by a constant not depending on $k, I_{1:k}$. Thus the sum itself is of order $\mathcal{O}(k)$. To handle the second sum, it suffices to note that $\pi_j = \int_{I_j} f_0(x) dx \leq M |I_j|$ by the boundedness assumption on f_0 . Hence, we obtain the bound

$$\sum_{j=1}^k \pi_j \log^2 (|I_j|) \leq M \sum_{j=1}^k |I_j| \log^2 (|I_j|).$$

Since $|I_j| \leq 1$, the same continuity argument as before yields summands of order $\mathcal{O}(1)$, and the sum is thus also $\mathcal{O}(k)$. Putting together all estimates, we have that

$$\int_0^1 f_0(x) \log^2 \left\{ \frac{f_0(x)}{f_{0,I_{1:k}}(x)} \right\} dx \leq \mathcal{O}(k) \leq \mathcal{O}(k_n).$$

Since $k_n/n \rightarrow 0$ by assumption, the claim follows. $\square$

The following Lemma shows that in the large-sample limit, the log marginal likelihood eventually becomes dominated by the maxmial log-likelihood only.

Lemma 5.3. *The marginal likelihood satisfies for all $k \leq k_n$,*

$$\log \int_{S_{k-1}} \prod_{i=1}^n f(x_i | k, I_{1:k}, \mathbf{p}) p(\mathbf{p} | \mathbf{x}_{1:n}) d\mathbf{p} = \log L_n(\hat{\mathbf{p}} | k, I_{1:k}) + o(n). \quad (5.30)$$

Proof. We now study the large-sample behavior of the marginal likelihood. We note that the left hand side of (5.30) is $p(\mathbf{x}_{1:n} | k, I_{1:k})$, and from (5.11) we thus have that

$$\int_{S_{k-1}} \prod_{i=1}^n f(x_i | k, I_{1:k}, \mathbf{p}) p(\mathbf{p} | k, I_{1:k}) d\mathbf{p} = \prod_{j=1}^k \left\{ |I_j|^{-N_j} \frac{\Gamma(a_{k,j} + N_j)}{\Gamma(a_{k,j})} \right\} \frac{\Gamma(a)}{\Gamma(a+n)}. \quad (5.31)$$

Consider now the logarithm of the above expression:

$$\sum_{j \in J(I_{1:k})} \left\{ \log \Gamma(a_{k,j} + N_j) - \log \Gamma(a_{k,j}) - N_j \log |I_j| \right\} + \log \Gamma(a) - \log \Gamma(a + n), \quad (5.32)$$

where $J(I_{1:k}) = \{j: N_j \geq 1\}$ are the nonempty cell indices. A Stirling approximation to the terms of the form $\Gamma(a_{k,j} + N_j)$ yields

$$\begin{aligned} & \sum_{j \in J(I_{1:k})} \left\{ (a_{k,j} + N_j - 1/2) \log(a_{k,j} + N_j) - a_{k,j} - N_j + \frac{1}{2} \log(2\pi) + \mathcal{O}(N_j^{-1}) \right\} \\ &= \sum_{j \in J(I_{1:k})} \left\{ (a_{k,j} + N_j - 1/2) \left[\log(N_j) + \mathcal{O}(N_j^{-1}) \right] - a_{k,j} - N_j + \frac{1}{2} \log(2\pi) + \mathcal{O}(N_j^{-1}) \right\}. \end{aligned}$$

We will now show that most of the terms in the sum converge to 0 when multiplied with n^{-1} . By assumption, we can find a constant $\tilde{A}$ such that $|a_{k,j} - 1/2| \leq \tilde{A}$ for all k, j . From this fact it follows that

$$|a_{k,j} - 1/2| \log(N_j) \leq \tilde{A} \log(n),$$

and this bound does not depend on the partition $I_{1:k}$. Since the remainder of the summand is $\mathcal{O}(1)$ except for $N_j \log(N_j) - N_j$, we obtain

$$n^{-1} \sum_{j \in J(I_{1:k})} \{N_j \log(N_j) - N_j + \mathcal{O}(\log(n))\} = \sum_{j \in J(I_{1:k})} \frac{N_j}{n} \log(N_j) - 1 + \mathcal{O}(n^{-1} k \log(n)).$$

Since $k \leq k_n$ and $n^{-1} \log(n) k_n = o(1)$ it follows that the last term is $o(1)$. Furthermore, by using some elementary rules for the logarithm we observe that

$$\sum_{j \in J(I_{1:k})} \frac{N_j}{n} \log N_j = \sum_{j \in J(I_{1:k})} \hat{p}_j \log(\hat{p}_j) + \sum_{j \in J(I_{1:k})} \hat{p}_j \log(n),$$

where $\hat{p}_j = N_j/n$. The last sum works out to $\log(n)$. We now handle the terms of the form $\log \Gamma(a_{k,j})$. Note that $\log \Gamma(x) = -\log(x) + o(1)$ as $x \rightarrow 0$ and that $\Gamma(\cdot)$ is bounded away from 0 on $(0, \Sigma]$. Since $\Gamma(\cdot)$ is continuous for positive arguments, the Extreme Value Theorem guarantees that $\log \Gamma(\cdot)$ is bounded on sets of the form $[x, \Sigma]$ for all $x > 0$. As such, for all sufficiently small x , the maximum of $|\log \Gamma(y)|$ for $y \in [x, \Sigma]$ occurs at x . This, together with the lower bound $a_{k,j} > \sigma k^{-1}$ yields $\log \Gamma(a_{k,j}) = \mathcal{O}(\log k)$. We conclude that

$$n^{-1} \sum_{j \in J(I_{1:k})} \log \Gamma(a_{k,j}) = \mathcal{O}(n^{-1} k \log(k)) = o(1).$$

To estimate the remaining terms in (5.32), we again resort to a Stirling approximation. After some algebra, one arrives at

$$\log \Gamma(a) - \log \Gamma(a + n) = (a - 1/2) \log(a) - (a + n - 1/2) \log(a + n) + n + \mathcal{O}(a^{-1}).$$

Working with the right hand side, we have that

$$\begin{aligned} (a + n - 1/2) \log(a + n) &= (a + n - 1/2) \log(1 + a/n) + (a + n - 1/2) \log(n) \\ &= (a + n - 1/2) \{a/n + o(a/n)\} + (a + n - 1/2) \log(n). \end{aligned}$$

Due to the boundedness assumption on $a_{k,j}$ we have $a = \mathcal{O}(k)$. Multiplying the above by n^{-1} , we have

$$(an^{-1} + 1 - 0.5n^{-1})\{a/n + o(a/n)\} + (an^{-1} + 1 - 0.5n^{-1})\log(n).$$

By assumption, all of the terms in the above tend to zero with the exception of $\log(n)$. Similarly, $n^{-1}(a + 1/2)\log(a) = \mathcal{O}(n^{-1}k\log(k))$ which also tends to 0. We conclude that

$$n^{-1}\{\log \Gamma(a) - \log \Gamma(a + n)\} = 1 - \log(n) + o(1).$$

Combining all previous estimates, we see that (5.32) times n^{-1} works out to

$$\begin{aligned} & \sum_{j \in J(I_{1:k})} \left\{ \hat{p}_j \log(\hat{p}_j) - \hat{p}_j \log |I_j| \right\} - 1 + \log(n) + 1 - \log(n) + o(1) \\ &= \sum_{j \in J(I_{1:k})} \hat{p}_j \log(\hat{p}_j / |I_j|) + o(1). \end{aligned}$$

We recognize the first term in the above expression as $n^{-1} \log L_n(\mathbf{p} | k, I_{1:k})$. $\square$

Finally, we combine the two preceding Lemmas to establish Lemma 5.4.

Lemma 5.4. *For every $\delta > 0$, we have that*

$$e^{n\delta} \int_{\mathcal{F}} \prod_{i=1}^n \frac{f(x_i)}{f_0(x_i)} dP_n(f) \xrightarrow{\mathbb{P}} \infty \quad (5.33)$$

Proof. We study the histogram based on using the partition including all k_n grid points. We note that by restricting to the set where the chosen partition is $k, I_{1:k_n}$, the integral in (5.33) becomes smaller and hence,

$$\begin{aligned} & \int_{\mathcal{F}} \prod_{i=1}^n \frac{f(x_i)}{f_0(x_i)} dP_n(f) \\ & \geq p_n(k_n) p_n(I_{1:k_n} | k_n) \int_{\mathcal{S}_{k_n-1}} \prod_{i=1}^n f(x_i | k_n, I_{1:k_n}, \mathbf{p}) p(\mathbf{p} | k_n, I_{1:k_n}) d\mathbf{p} \left\{ \prod_{i=1}^n f_0(x_i) \right\}^{-1} \end{aligned}$$

By the result of Lemma 5.3, we see that the term on the right hand side times n^{-1} is equal to

$$A_n = n^{-1} \log p_n(k_n) + n^{-1} \log p_n(I_{1:k_n} | k_n) + \sum_{j \in J(I_{1:k_n})} \hat{p}_j \log \frac{\hat{p}_j}{|I_j|} - n^{-1} \sum_{i=1}^n \log f_0(x_i) + o_{\mathbb{P}}(1). \quad (5.34)$$

By Lemma 5.2, we note that

$$\begin{aligned} n^{-1} \sum_{i=1}^n \log f_0(x_i) &= n^{-1} \sum_{i=1}^n \log f_{0,I_{1:k_n}}(x_i) + n^{-1} \sum_{i=1}^n \log \frac{f_0(x_i)}{f_{0,I_{1:k_n}}(x_i)} \\ &= n^{-1} \sum_{i=1}^n \log f_{0,I_{1:k_n}}(x_i) + K(f_0, f_{0,I_{1:k_n}}) + o_{\mathbb{P}}(1). \end{aligned}$$

Since the density $f_{0,I_{1:k_n}}$ is piecewise constant on the partition $I_{1:k_n}$ we can write $f_{0,I_{1:k_n}}(x) = \prod_{j=1}^{k_n} (\pi_j / |I_j|)^{\mathcal{I}(x \in I_j)}$, and we deduce the equality $n^{-1} \sum_{i=1}^n \log f_{0,I_{1:k_n}}(x_i) = \sum_{j \in J_{k_n}} \hat{p}_j \log \frac{\pi_j}{|I_j|}$. Returning to the expression (5.34), we observe that

$$A_n = \log p_n(k) + \log p_n(I_{1:k_n} | k) + \sum_{j \in J(I_{1:k_n})} \hat{p}_j \log \frac{\hat{p}_j}{|I_j|} - \sum_{j \in J(I_{1:k_n})} \hat{p}_j \log \frac{\pi_j}{|I_j|} - K(f_0, f_{0,I_{1:k_n}}) + o_{\mathbb{P}}(1).$$

Noting that $K(\hat{f}_n, f_{0,I_{1:k_n}}) = \sum_{j \in J(I_{1:k_n})} \hat{p}_j \log \frac{\hat{p}_j}{|I_j|} - \sum_{j \in J(I_{1:k_n})} \hat{p}_j \log \frac{\pi_j}{|I_j|}$, where $\hat{f}_n$ is the maximum likelihood estimate based on the partition $I_{1:k}$ (2.10), we conclude that

$$A_n = \log p_n(k_n) + \log p_n(I_{1:k_n} | k_n) - K(f_0, f_{0,I_{1:k_n}}) + K(\hat{f}_n, f_{0,I_{1:k_n}}) + o_{\mathbb{P}}(1).$$

Since $p_n(I_{1:k} | k)$ is uniform, $p_n(I_{1:k_n} | k_n) = 1$, as there is only one possible partition which includes all grid points. Due to the fact that $K(\hat{f}_n, f_{0,I_{1:k_n}}) \geq 0$ and $n^{-1}\{\log p_n(k_n) + \log p_n(I_{1:k_n} | k_n)\} \rightarrow 0$ by the assumption on $p_n(k)$, we find the inequality

$$A_n \geq -K(f_0, f_{0,I_{1:k_n}}) + o_{\mathbb{P}}(1).$$

To finish the proof, let $\delta > 0$. Since $\max_{1 \leq j \leq k_n} \{\tau_{n,j} - \tau_{n,j-1}\} \rightarrow 0$, the intervals of the partition $I_{1:k_n}$ have lengths tending to 0 and it follows from Corollary 4.2 that there exists an N such that $K(f_0, f_{0,I_{1:k_n}}) < \delta/2$ for all $n \geq N$. This implies that $A_n \geq -\delta/2 + o_{\mathbb{P}}(1)$. Exponentiating this expression and multiplying by $e^{n\delta}$ we obtain the inequality

$$e^{n\delta} \int_{\mathcal{F}} \prod_{i=1}^n \frac{f(x_i)}{f_0(x_i)} dP_n(f) \geq \exp\{n\delta/2 + o_{\mathbb{P}}(n)\}. \quad (5.35)$$

The right hand side diverges to positive infinity in probability, establishing the claim. $\square$

Having finally established Lemma 5.4, we conclude the section with the proof of Theorem 5.2.

Proof of Theorem 5.2. Let $\mathcal{F}_n = \cup_{k=1}^{k_n} \mathcal{H}_k$ where $\mathcal{H}_k$ is as in (5.25). By Lemma 5.1 the ϵ covering number of $\mathcal{F}_n$ satisfies

$$N(\epsilon, \mathcal{F}_n, d_H) \leq \left(\frac{2L}{\epsilon}\right)^{k_n} = \exp(L'k_n)$$

for all small ϵ and a suitable constant $L' > 0$ possibly depending on ϵ . Since $k_n = o(n)$ by assumption, it follows that for all sufficiently large n , $L'k_n < n\epsilon^2$ and hence

$$N(\epsilon, \mathcal{F}_n, d_H) < \exp(n\epsilon^2).$$

We also note that $\mathcal{F}_n$ as defined above makes up the entire support of the prior, meaning that $P_n(f \in \mathcal{F}_n^c) = 0$ for all n .

Due to the entropy and prior mass condition deduced above, it follows from Ghosal and van der Vaart (2017, Theorems 6.17 and 6.23) that we can write

$$P_n\left(f \in \mathcal{F}_n : d_H(f_0, f) \geq \epsilon \mid \mathbf{x}_{1:n}\right) \leq A_n + \frac{B_n}{\int_{\mathcal{F}} \prod_{i=1}^n f(x_i)/f_0(x_i) dP_n(f)}, \quad (5.36)$$

where $A_n \xrightarrow{\mathbb{P}} 0$ and $B_n = \mathcal{O}_{\mathbb{P}}(e^{-n\delta})$ for some $\delta > 0$. Writing C_n for the denominator of the last term, we see that it suffices to show that B_n/C_n tends to zero in probability. To this end, note that for $\gamma > 0$ the positivity of C_n means that

$$\begin{aligned} \mathbb{P}\left(\left|\frac{B_n}{C_n}\right| \geq \gamma\right) &= \mathbb{P}(B_n \geq \gamma C_n) \\ &= \mathbb{P}\left(B_n \geq \gamma C_n, C_n \geq \gamma^{-1}e^{-n\delta/2}\right) + \mathbb{P}\left(B_n \geq \gamma C_n, C_n < \gamma^{-1}e^{-n\delta/2}\right) \\ &\leq \mathbb{P}\left(B_n \geq e^{-n\delta/2}\right) + \mathbb{P}\left(e^{n\delta/2}C_n < \gamma^{-1}\right). \end{aligned}$$

The last term in the above converges to 0 by Lemma 5.4. To prove that the second term is also $o(1)$, we appeal to Proposition 2.2. Note that $e^{n\delta/2}B_n = \mathcal{O}_{\mathbb{P}}(e^{-n\delta/2})$ which is $o_{\mathbb{P}}(1)$ by (2.51). We conclude that $\mathbb{P}(e^{n\delta/2}B_n \geq 1) = o(1)$ and hence $B_n/C_n = o_{\mathbb{P}}(1)$, establishing the claim. $\square$

5.4 Consistency of irregular histogram point estimation

We now study the consistency of the point estimator $f(\cdot | \hat{k}, \hat{I}_{1:\hat{k}}, \hat{\mathbf{p}}_B)$. The proof follows a similar strategy to that employed in Section 4.5, relying on the large-sample expansion Lemma 5.3 to show that the MAP $\hat{I}_{1:\hat{k}}$ will eventually select a partition $I_{1:k}$ which yields a piecewise constant density close to f_0 . Matters are a bit more complicated here than in Section 4.5, since the argument used there relies on the fact that for sufficiently large H , the regular piecewise constant version $f_{0,\kappa}$ of f_0 satisfies $K(f_0, f_{0,\kappa}) < \epsilon$ for every $\kappa \geq H$. In the irregular histogram setting, the piecewise constant approximation based on the grid $I_{1:\kappa}$, denoted $f_{0,I_{1:\kappa}}$, can still fail to get close to f_0 in the KL sense even as $\kappa \rightarrow \infty$ if $\max_j |I_{\kappa,j}|$ does not approach 0. Due to this fact, we cannot directly adapt the argument from Lemma 4.7 to the irregular histogram setting. Hence, there is a need to develop new tools to prove consistency in the irregular histogram model.

We establish consistency of the histogram point estimator under the following conditions:

Assumption 5.2. For the irregular histogram prior sequence, (ii) and (iii) of Assumption 5.1 hold. In addition, we require that the prior for k , $p_n(k)$ is supported on $\{1, 2, \dots, k_n\}$ for each n , where $k_n = \mathcal{O}(n^\gamma)$ for some $\gamma \in (0, 1)$.

This assumption is identical to Assumption 5.1 except for the requirement that k_n grows at a slower rate, which allows us to bound the maximum in the remainder term of (5.37), which is crucial for the argument we develop later to establish convergence. Before we begin with proofs, recall that for a partition $I_{1:k}$ and a simplex vector $\boldsymbol{\rho} \in \mathcal{S}_{k-1}$, we write $f(\cdot | k, I_{1:k}, \boldsymbol{\rho})$ for the function given in (4.32).

Lemma 5.5. Let $\{I_{1:k}\}_{k=1}^\infty$ be a sequence of partitions of the unit interval such that $\max_j |I_{k,j}| \rightarrow 0$ and suppose that f_0 is a bounded, a.e. continuous density on the unit interval. Then provided $k_n = \mathcal{O}(n^\gamma)$ for some $\gamma \in (0, 1)$, the piecewise constant version $f_{0,I_{1:k}}$ of f_0 satisfies, for all $k \leq k_n$,

$$n^{-1} \sum_{i=1}^n \log \frac{f_0(x_i)}{f_{0,I_{1:k}}(x_i)} = K(f_0, f_{0,I_{1:k}}) + r_n(I_{1:k}), \quad (5.37)$$

where $\max_{I_{1:k}} |r_n(I_{1:k})| = o_{\mathbb{P}}(1)$.

Proof. Using basic properties of the logarithm, we can write

$$\begin{aligned} r_n(I_{1:k}) &= n^{-1} \sum_{i=1}^n \log \frac{f_0(x_i)}{f_{0,I_{1:k}}(x_i)} - K(f_0, f_{0,I_{1:k}}) \\ &= n^{-1} \sum_{i=1}^n \log f_0(x_i) - \int_0^1 f_0(x) \log f_0(x) dx \\ &\quad - n^{-1} \sum_{i=1}^n \log f_{0,I_{1:k}}(x_i) + \int_0^1 f_{0,I_{1:k}}(x) \log f_{0,I_{1:k}}(x) dx. \end{aligned}$$

We note that by the Law of Large Numbers, $n^{-1} \sum_{i=1}^n \log f_0(x_i) - \int_0^1 f_0(x) \log f_0(x) dx$ converges to 0 in probability. Moreover, as this expression does not depend on the chosen partition, the convergence is uniform over $\cup_{k=1}^{k_n} \mathcal{P}_{n,k}$. Thus, it suffices to show that the

remaining terms also converge to 0 uniformly in probability. Since $f_{0,I_{1:k}}$ is piecewise constant, we can write

$$\begin{aligned} n^{-1} \sum_{i=1}^n \log f_{0,I_{1:k}}(x_i) &= \sum_{j=1}^k \hat{p}_j \log(\pi_j/|I_j|), \\ \int_0^1 f_{0,I_{1:k}}(x) \log f_{0,I_{1:k}}(x) dx &= \sum_{j=1}^k \pi_j \log(\pi_j/|I_j|). \end{aligned}$$

where $\hat{p}_j = N_j/n$ are the maximum likelihood estimates. Plugging this into the above, we deduce the simplified expression

$$n^{-1} \sum_{i=1}^n \log f_{0,I_{1:k}}(x_i) - \int_0^1 f_{0,I_{1:k}}(x) \log f_{0,I_{1:k}}(x) dx = \sum_{j=1}^k (\hat{p}_j - \pi_j) \log(\pi_j/|I_j|).$$

Using the triangle inequality, the absolute value of the right hand side of the above is bounded by

$$\sum_{j=1}^k |\hat{p}_j - \pi_j| \left| \log \pi_j - \log |I_j| \right| \leq - \sum_{j=1}^k |\hat{p}_j - \pi_j| \log \pi_j - \sum_{j=1}^k |\hat{p}_j - \pi_j| \log |I_j|,$$

where we used that $\pi_j, |I_j| \leq 1$ for the last inequality. Since the partition $I_{1:k_n}$ is a refinement of $I_{1:k}$, we deduce from the triangle inequality that the last display is upper bounded by

$$- \sum_{j=1}^{k_n} |\hat{p}_j - \pi_j| \log \pi_j - \sum_{j=1}^{k_n} |\hat{p}_j - \pi_j| \log |I_j|. \quad (5.38)$$

If we can show that (5.38) tends to 0, then it follows that $|r_n(I_{1:k})| \xrightarrow{\mathbb{P}} 0$ uniformly as claimed. We treat the first term first. Let $\mathcal{A}_n = \{j: \pi_j \geq n^{-1+\epsilon}\}$ and define $\mathcal{E}_n = \cup_{j \in \mathcal{A}_n} \{|\hat{p}_j - \pi_j| \geq \pi_j n^{-\epsilon/3}\}$, for some $\epsilon \in (0, 1)$ to be determined later. Our strategy is to split the first sum in (5.38) into two parts,

$$- \sum_{j \in \mathcal{A}_n} |\hat{p}_j - \pi_j| \log \pi_j - \sum_{j \in \mathcal{A}_n^c} |\hat{p}_j - \pi_j| \log \pi_j, \quad (5.39)$$

and show that each of these converge to 0 in probability. We handle the first term first. By a Chernoff bound we have for a universal constant d [see Vershynin (2018, Exercise 2.3.5)]

$$\begin{aligned} \mathbb{P}(\mathcal{E}_n) &= \mathbb{P}\left(\cup_{j \in \mathcal{A}_n} |N_j - n\pi_j| \geq n\pi_j/n^{\epsilon/3}\right) \\ &\leq 2 \sum_{j \in \mathcal{A}_n} \exp\left\{-dn\pi_j n^{-2\epsilon/3}\right\} \\ &\leq 2 \sum_{j \in \mathcal{A}_n} \exp\left\{-dn^{\epsilon/3}\right\} \\ &\leq 2k_n \exp\{-dn^{\epsilon/3}\}. \end{aligned}$$

The last display is $o(1)$, since $k_n = o(n)$. On the set $\mathcal{E}_n^c$, a simple calculation reveals that

$$- \sum_{j \in \mathcal{A}_n} |\hat{p}_j - \pi_j| \log(\pi_j) \leq - \sum_{j \in \mathcal{A}_n} \frac{\pi_j}{n^{\epsilon/3}} \log(\pi_j) \leq \sum_{j \in \mathcal{A}_n} \frac{(1-\epsilon)\pi_j}{n^{\epsilon/3}} \log(n) \leq \frac{1}{n^{\epsilon/3}} \log(n),$$

which is $o(1)$. It follows that

$$\begin{aligned} & \mathbb{P} \left(- \sum_{j \in \mathcal{A}_n} |\hat{p}_j - \pi_j| \log(\pi_j) \geq \gamma \right) \\ & \leq \mathbb{P} \left(- \sum_{j \in \mathcal{A}_n} |\hat{p}_j - \pi_j| \log(\pi_j) \geq \gamma, \mathcal{E}_n^c \right) + \mathbb{P}(\mathcal{E}_n) \end{aligned}$$

which converges to 0 by previous estimates. For the second term in (5.39) we have

$$-\mathbb{E}_{\mathbf{x}_{1:n}} \sum_{j \in \mathcal{A}_n^c} |\hat{p}_j - \pi_j| \log \pi_j = - \sum_{j \in \mathcal{A}_n^c} \mathbb{E}_{\mathbf{x}_{1:n}} |\hat{p}_j - \pi_j| \log \pi_j \leq - \sum_{j \in \mathcal{A}_n^c} \sqrt{\pi_j(1 - \pi_j)/n} \log(\pi_j),$$

where we used Cauchy–Schwarz for the last inequality. Thus, the last display is upper bounded by

$$n^{-1/2} \sum_{j \in \mathcal{A}_n^c} -\pi_j^{1/2} \log(\pi_j) = n^{-1/2} \sum_{j \in \mathcal{A}_n^c} g(\pi_j),$$

where $g: [0, 1] \rightarrow \mathbb{R}$ is defined by $g(t) = -\sqrt{t} \log(t) \mathcal{I}(t > 0)$. Since $\lim_{t \downarrow 0} g(t) = 0$, g is continuous. Moreover, its derivative on $(0, 1)$ is given by $g'(t) = -\frac{1}{2\sqrt{t}} \log(t) - \frac{1}{\sqrt{t}}$ which is positive for all t in all sufficiently small strictly positive neighbourhoods of 0. We conclude that if n is large enough, then the supremum of $|g(t)|$ over the set $[0, n^{-1+\epsilon}]$ is attained at $t = n^{-1+\epsilon}$, as a direct consequence of the mean value theorem. For $j \in \mathcal{A}_n^c$ we have that $\pi_j < n^{-1+\epsilon}$, from which it follows that

$$\begin{aligned} n^{-1/2} \sum_{j \in \mathcal{A}_n^c} g(\pi_j) & \leq n^{-1/2} \sum_{j \in \mathcal{A}_n^c} (1 - \epsilon) n^{-1/2+\epsilon/2} \log(n) \\ & \leq (1 - \epsilon) k_n n^{-1+\epsilon/2} \log(n) \\ & = \mathcal{O}(n^{\gamma-1+\epsilon/2} \log(n)). \end{aligned}$$

By taking ϵ to be sufficiently small, this tends to 0, and Markov's inequality consequently yields

$$- \sum_{j \in \mathcal{A}_n} |\hat{p}_j - \pi_j| \log \pi_j - \sum_{j \in \mathcal{A}_n^c} |\hat{p}_j - \pi_j| \log \pi_j \xrightarrow{\mathbb{P}} 0.$$

To handle the second term in (5.38) we can rely on previous estimates. Since f_0 is bounded by assumption we have $\pi_j \leq M|I_j|$ and hence as $|I_j| \leq 1$,

$$- \sum_{j=1}^{k_n} |\hat{p}_j - \pi_j| \log |I_j| \leq \sum_{j=1}^{k_n} |\hat{p}_j - \pi_j| \{ \log M - \log \pi_j \}.$$

Reexamining the proof, we see that the argument laid out previously goes through when $\log \pi_j$ is replaced with $\log M$, and as such,

$$- \sum_{j=1}^{k_n} |\hat{p}_j - \pi_j| \log |I_j| \xrightarrow{\mathbb{P}} 0,$$

which establishes the claim. $\square$

The following lemma serves the same purpose as Lemma 4.7, but has been adapted for the irregular setting.

Lemma 5.6. *Let $\mathcal{A}_\epsilon = \{I_{1:k} : K(f_0, f_{0,I_{1:k}}) < \epsilon\}$ for some $\epsilon > 0$. Then the MAP partition $\hat{I}_{1:k}$, satisfies*

$$\mathbb{P}(\hat{I}_{1:k} \in \mathcal{A}_\epsilon^c) \rightarrow 0.$$

Proof. We start by arguing as in Lemma 4.6. By Bayes' rule (5.6), we can write the log-posterior probability of a partition as

$$\log p_n(k, I_{1:k} | \mathbf{x}_{1:n}) = \log p_n(k) + \log p_n(I_{1:k} | k) + \log p_n(\mathbf{x}_{1:n} | k, I_{1:k}) - \log(Z_n),$$

where Z_n is the normalization constant, which does not depend on $k, I_{1:k}$. Combining this with the result of Lemma 5.3, the log-posterior can thus be written as

$$\log p_n(k, I_{1:k} | \mathbf{x}_{1:n}) = \log L_n(\hat{\mathbf{p}} | k, I_{1:k}) - \log(Z_n) + o(n).$$

Dividing by n , subtracting and adding $C_n = n^{-1} \sum_{i=1}^k \log f_0(x_i) - \log(Z_n)$ as in Lemma 4.7 to the right hand side of the above yields

$$\sum_{j \in J(I_{1:k})} \hat{p}_j \log(\hat{p}_j / |I_j|) - \sum_{j \in J(I_{1:k})} \hat{p}_j \log(\pi_j / |I_j|) - n^{-1} \sum_{i=1}^n \log \frac{f_0(x_i)}{f_{0,I_{1:k}}(x_i)} + C_n + o(n), \quad (5.40)$$

where $J(I_{1:k}) = \{j : N_j > 0\}$ and we used that $\log f_{0,I_{1:k}}(x) = \log(\pi_j / |I_j|)$ for $x \in I_j$ and hence

$$\sum_{i=1}^n \log f_{0,I_{1:k}}(x_i) = \sum_{j \in J(I_{1:k})}^k N_j \log(\pi_j / |I_j|).$$

Combining the result of Lemma 5.5 with (5.40), we see that $n^{-1} \log p_n(k, I_{1:k} | \mathbf{x}_{1:n})$ equals

$$\sum_{j=1}^k \hat{p}_j \log(\hat{p}_j / |I_j|) - \sum_{j=1}^k \hat{p}_j \log(\pi_j / |I_j|) - K(f_0, f_{0,I_{1:k}}) + C_n + r_n(I_{1:k}),$$

where $\max_{I_{1:k}} |r_n(I_{1:k})| = o_{\mathbb{P}}(1)$. To further simplify this expression, an argument similar to that used to establish (4.52) yields

$$0 \leq \sum_{j \in J(I_{1:k})} \hat{p}_j \log(\hat{p}_j / \pi_j) \leq \sum_{j \in J(I_{1:k_n})} \hat{p}_j \log(\hat{p}_j / \pi_j) \leq \sum_{j \in J(I_{1:k_n})} \frac{(\hat{p}_j - \pi_j)^2}{\pi_j}.$$

where we used that $I_{1:k_n}$ is a refinement of $I_{1:k}$ together with the log-sum inequality (Lemma B.5) to deduce the middle inequality. The expectation of this latter expression is bounded by

$$\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ \sum_{j \in J(I_{1:k_n})} \frac{(\hat{p}_j - \pi_j)^2}{\pi_j} \right\} \leq \sum_{j: \pi_j > 0} \frac{\text{Var}_{\mathbf{x}_{1:n}}(\hat{p}_j)}{\pi_j} = \sum_{j: \pi_j > 0} \frac{\pi_j(1 - \pi_j)}{n\pi_j} \leq n^{-1} k_n,$$

which is $o(1)$. We conclude by Markov's inequality that

$$\max_{I_{1:k}} \sum_{j \in J(I_{1:k})} \hat{p}_j \log(\hat{p}_j / \pi_j) \leq \sum_{j \in J(I_{1:k_n})} \{\hat{p}_j - \pi_j\}^2 / \pi_j \leq o_{\mathbb{P}}(1).$$

It follows that we can write

$$n^{-1} \log p_n(k, I_{1:k} | \mathbf{x}_{1:n}) = -K(f_0, f_{0,I_{1:k}}) + \tilde{r}_n(I_{1:k}) + C_n, \quad (5.41)$$

where $\max_{I_{1:k}} |\tilde{r}_n(I_{1:k})| \xrightarrow{\mathbb{P}} 0$. To conclude the proof, we note that since $\max_{1 \leq j \leq k_n} |I_{k_n, j}| \rightarrow 0$ by assumption, Corollary 4.2 guarantees that $K(f_0, f_{0, I_{1:k_n}}) < \epsilon/2$ for all sufficiently large n . This, combined with (5.41) yields the lower bound

$$n^{-1} \log p_n(k_n, I_{1:k_n} | \mathbf{x}_{1:n}) \geq -\epsilon/2 + o_{\mathbb{P}}(1).$$

On the other hand, (5.41) also implies the following upper bound,

$$n^{-1} \max_{I_{1:k} \in \mathcal{A}_{\epsilon}^c} \log p_n(k, I_{1:k} | \mathbf{x}_{1:n}) \leq -\epsilon + o_{\mathbb{P}}(1),$$

since any $I_{1:k} \in \mathcal{A}_{\epsilon}^c$ satisfies $K(f_0, f_{0, I_{1:k}}) \geq \epsilon$. Combining the two bounds on the log-posterior pmf results in.

$$\max_{I_{1:k} \in \mathcal{A}_{\epsilon}^c} \log p_n(k, I_{1:k} | \mathbf{x}_{1:n}) - \log p_n(k_n, I_{1:k_n} | \mathbf{x}_{1:n}) \leq -n\epsilon/2 + o_{\mathbb{P}}(n),$$

where we used that the sum of two $o_{\mathbb{P}}(1)$ variables in $o_{\mathbb{P}}(1)$ together with $n o_{\mathbb{P}}(1) = o_{\mathbb{P}}(n)$. The right hand of the above side diverges to $-\infty$ in probability as $n \rightarrow \infty$. It follows that the random variable $\max_{I_{1:k} \in \mathcal{A}_{\epsilon}^c} p_n(k, I_{1:k} | \mathbf{x}_{1:n}) / p_n(k_n, I_{1:k_n} | \mathbf{x}_{1:n})$ is $o_{\mathbb{P}}(1)$, and we deduce that

$$\mathbb{P}(\hat{I}_{1:\hat{k}} \in \mathcal{A}_{\epsilon}^c) \leq \mathbb{P}\left(\max_{I_{1:k} \in \mathcal{A}_{\epsilon}^c} p_n(k, I_{1:k} | \mathbf{x}_{1:n}) \geq p_n(k_n, I_{1:k_n} | \mathbf{x}_{1:n})\right) \rightarrow 0,$$

which was to be shown. $\square$

We now in a position to prove the main result of the section.

Theorem 5.3. *Suppose that Assumption 5.2 holds and that $f_0: [0, 1] \rightarrow \mathbb{R}$ is a bounded, a.e. continuous density function. Then the irregular histogram estimator is $\mathbb{L}_1$ consistent for f_0 .*

Proof (sketch). The proof follows the strategy outlined in the proof of Theorem 4.3. Fix $\epsilon > 0$. Arguing as in (4.54), we can bound the $\mathbb{L}_1$ risk R_n of $\hat{f}_{B,n}$ by

$$\underbrace{2\mathbb{P}(\hat{I}_{1:\hat{k}} \in \mathcal{A}_{\epsilon}^c)}_{(i)} + \underbrace{\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_0, f_{0, \hat{I}_{1:\hat{k}}}) \mathcal{I}(\hat{I}_{1:\hat{k}} \in \mathcal{A}_{\epsilon}^c) \right\}}_{(ii)} + \underbrace{\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_1(f_{0, \hat{I}_{1:\hat{k}}}, \hat{f}_{B,n}) \mathcal{I}(\hat{I}_{1:\hat{k}} \in \mathcal{A}_{\epsilon}) \right\}}_{(iii)}.$$

By the preceding Lemma, (i) is $o(1)$. To get an upper bound on (ii), we use Pinsker's inequality which states that $d_1(f, g) \leq \sqrt{2K(f, g)}$ for two densities f, g (see Tsybakov (2009, Lemma 2.5)). Consequently we have that

$$d_1(f_0, f_{0, \hat{I}_{1:\hat{k}}}) \mathcal{I}(\hat{I}_{1:\hat{k}} \in \mathcal{A}_{\epsilon}^c) \leq \sqrt{2K(f_0, f_{0, \hat{I}_{1:\hat{k}}})} \mathcal{I}(\hat{I}_{1:\hat{k}} \in \mathcal{A}_{\epsilon}^c) \leq \sqrt{2\epsilon} \mathcal{I}(\hat{I}_{1:\hat{k}} \in \mathcal{A}_{\epsilon}^c) \leq \sqrt{2\epsilon},$$

so (ii) is bounded by $\sqrt{2\epsilon}$ from above. We can upper bound (iii) as in the proof of Theorem 4.3, since the arguments used there are also valid for given irregular partitions provided $k_n = o(n)$. Thus, (iii) is also $o(1)$, and we conclude that $R_n \leq \sqrt{2\epsilon} + o(1)$. Since ϵ was arbitrary, we obtain $\lim_{n \rightarrow \infty} R_n = 0$, which was to be shown. $\square$

5.5 Posterior rate of convergence of irregular random histograms

As for regular histogram priors, we establish that the posterior distribution based on the irregular histogram prior possesses a convergence rate close to the minimax one under suitable regularity conditions. The basic workhorses are the lemmas presented in Section 4.6, due to the fact that irregular histograms also share the approximation properties of their regular cousins, provided the size of the largest interval goes to zero sufficiently fast. This enables a proof along similar lines to that of Scricciolo (2007, Theorem 2).

Assumption 5.3. The irregular histogram prior sequence $\{P_n\}$ satisfies

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

$$D_1 \exp(-d_1 k \log(k)) \leq p_n(k) \leq D_2 \exp(-d_2 k \log(k)). \quad (5.42)$$

- (ii) Given k and a partition $I_{1:k}$, the prior for $\mathbf{p} \mid k, I_{1:k}$ is $\text{Dir}(\mathbf{a}_k)$ where $\sigma k^{-1} \leq a_{k,j} \leq \Sigma$ for constants $\sigma, \Sigma > 0$.
- (iii) The sequence of partitions satisfy

$$\max_{1 \leq j \leq k_n} \{\tau_{n,j} - \tau_{n,j-1}\} = \mathcal{O}(k_n^{-1}). \quad (5.43)$$

Overall, we note that the stated assumptions here are mostly similar to those of Assumption 4.3. The only difference is that since the grid is not given in advance, we must ensure that the largest bin in the finest possible partition is of the correct order. The condition (5.43) restricts the grid so be sufficiently regular to leverage the approximation rate of Proposition 4.2. The condition (5.42) on the prior for k can be fulfilled by letting $p_n(k)$ be the Poisson distribution truncated to the set $\{1, 2, \dots, k_n\}$.

Theorem 5.4. *Suppose that Assumption 5.3 holds and that the sequence k_n satisfies $n^\gamma/k_n \rightarrow 0$ for any $\gamma \in [0, 1)$ and $k_n = \mathcal{O}(n)$. Suppose further that $\sqrt{f_0}$ is α -Hölder continuous for some $\alpha \in (0, 1]$. Then the posterior distribution converges at the rate*

$$\epsilon_n = n^{-\alpha/(2\alpha+1)} \log(n)^{(\alpha+1)/(2\alpha+1)}. \quad (5.44)$$

Before proceeding with the proof of Theorem 5.4, we make some brief comments on the result. The rate differs from the minimax one, $n^{-\alpha/(2\alpha+1)}$, by a logarithmic factor. Again, the posterior convergence rate obtained is slightly worse than that of Theorem 2 in Scricciolo (2007) for the corresponding regular histogram model, but only by a logarithmic factor for small values of α . This is due to the fact that Scricciolo makes the assumption that f_0 is bounded away from 0, and as such it becomes possible to get a tighter estimate of the probability of the ball $\mathcal{B}_\epsilon(f_0)$ defined below, since the Hellinger distance and the KL divergence are equivalent in this setting (Ghosal and van der Vaart, 2017, p. 199). With suitable modifications of our proof, it is possible to show that the irregular histogram prior attains the same slightly better rate if f_0 is lower bounded by 0. We note that the irregular histogram prior does not depend on the unknown smoothness parameter α . The procedure thus adapts to the smoothness of f_0 without knowledge of the unknown α .

Since we rely on Theorem 3.4 to establish the convergence result, we need an estimate of the prior probability of a certain type of a Kullback–Leibler type neighborhood. Let $V(f_0, f) = \int_0^1 f_0(x) \log^2 \{f_0(x)/f(x)\} dx$, and define the pseudo-ball $\mathcal{B}'_\epsilon(f_0) = \{f \in \mathcal{F} : \max\{K(f_0, f), V(f_0, f)\} \leq \epsilon^2\}$. The following Lemma provides a lower bound on the probability of $\mathcal{B}'_\epsilon(f_0)$.

Lemma 5.7. *Suppose that Assumption 5.3 holds. Then there are constants $C_1, C_2 > 0$ not depending on $\epsilon > 0$ such that for any $k = k(\epsilon) \in \mathbb{N}$ satisfying*

$$C_1 \{\epsilon / \log(1/\epsilon)\}^{-1/\alpha} \leq k \leq C_2 \{\epsilon / \log(1/\epsilon)\}^{-1/\alpha}.$$

there exists a partition $I_{1:k}$ and two constants c_1, c_2 not depending on n, ϵ, k and $I_{1:k}$ that satisfy

$$P_n(\mathcal{B}'_\epsilon(f_0)) \geq p_n(k) p_n(I_{1:k} | k) c_1 \exp\{-c_2 k \log(1/\epsilon)\}, \quad (5.45)$$

for all sufficiently small $\epsilon > 0$.

Proof. We argue as in the beginning of the proof of Theorem 2 in Scricciolo (2007). Let $\epsilon \in (0, (1 - e^{-1})/\sqrt{2}]$. If n is such that $k(\epsilon) > k_n$, then the inequality (5.45) is trivially true as the right hand side is zero. Thus, we may assume that n is such that $k \leq k_n$ throughout the proof. In this case we define $\gamma = \epsilon / \log(1/\epsilon)$. Let $k = k(\gamma)$ be such that $C_1 \gamma^{-1/\alpha} \leq k \leq C_2 \gamma^{-1/\alpha}$, where C_1, C_2 are positive constants to be determined later. Our strategy will be to relate the neighborhood $\mathcal{B}'_\epsilon(f_0)$ 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 $I_{1:k}$ by including every $\lceil k_n/k \rceil$ cut point in the partition. The maximal size of a bin in this partition is then bounded by, for a constant $\tilde{A} > 0$,

$$\max_j |I_j| \leq \lceil k_n/k \rceil \max_j \{\tau_{n,j} - \tau_{n,j-1}\} \leq \tilde{A} \lceil k_n/k \rceil k_n^{-1} \leq 2\tilde{A}k^{-1},$$

where we used (5.43), provided n is sufficiently large. This shows that $\max_j |I_j| = \mathcal{O}(k^{-1})$.

Let $f_{0,I_{1:k}}$ be the piecewise constant version of f_0 based on the partition $I_{1:k}$, and let $f_{I_{1:k}}(\cdot) = f(\cdot | k, I_{1:k}, \mathbf{p})$. To bound the squared Hellinger distance between f_0 and $f_{I_{1:k}}$ we apply the triangle inequality together with Proposition A.7, yielding

$$d_H^2(f_0, f_{I_{1:k}}) \leq 2d_H^2(f_0, f_{0,I_{1:k}}) + 2d_H^2(f_{0,I_{1:k}}, f_{I_{1:k}}). \quad (5.46)$$

To bound the first term of (5.46), we leverage Proposition 4.3 to find that

$$d_H^2(f_0, f_{0,I_{1:k}}) \leq Ak^{-2\alpha} \leq AC_1^{-2\alpha} \gamma^2, \quad (5.47)$$

for some constant $A > 0$. We now bound the last term on the right hand side of (5.46). Denoting $\pi_j = \int_{I_{k,j}} f_0(x) dx$ and using the fact that $d_H^2(f_{I_{1:k}}, f_{0,I_{1:k}}) \leq d_1(f_{I_{1:k}}, f_{0,I_{1:k}})$, we obtain

$$d_H^2(f_{I_{1:k}}, f_{0,I_{1:k}}) \leq d_1(f_{I_{1:k}}, f_{0,I_{1:k}}) = \sum_{j=1}^k \int_{I_{k,j}} |I_{k,j}|^{-1} |p_j - \pi_j| dx = \sum_{j=1}^k |p_j - \pi_j| = d_1(\mathbf{p}, \boldsymbol{\pi}).$$

Now for any simplex vector $\mathbf{p} \in \mathcal{S}_{k-1}$ satisfying $d_1(\mathbf{p}, \boldsymbol{\pi}) \leq C_1^{-2\alpha} \gamma^{1+1/\alpha}$ it follows from (5.46) that

$$\begin{aligned} d_H^2(f_0, f_{I_{1:k}}) &\leq 2Ak^{-2\alpha} + 2d_1(\mathbf{p}, \boldsymbol{\pi}) \\ &\leq 2Ak^{-2\alpha} + 2C_1^{-2\alpha} \gamma^{1+1/\alpha} \\ &\leq 2(AC_1^{-2\alpha} + C_1^{-2\alpha})\gamma^2 \\ &= c_3\gamma^2, \end{aligned} \tag{5.48}$$

where we used that $\gamma^{1+1/\alpha} \leq \gamma^2$ for $\gamma \in (0, 1)$ and $c_3 = 2C_1^{-2\alpha}(A + 1)$ is a constant. Note that $c_3 \downarrow 0$ as $C_1 \rightarrow \infty$, so c_3 can be made arbitrarily small by choosing C_1 to be suitably large. To relate the Hellinger distance to the KL divergence, we verify that the conditions of Lemma B.4 are met for $\delta = 1$. Assume that $p_j > (C_1^{-2\alpha} \gamma^{1+1/\alpha}/2)^4$ for all j which implies that the corresponding density $f_{I_{1:k}}$ is lower bounded by $(2A)^{-1}k(C_1^{-2\alpha} \gamma^{1+1/\alpha}/2)^4$, since $\max_j |I_{k,j}| \leq 2Ak^{-1}$ by a previous calculation. It follows that

$$L_1^2 = \int_{f_0^2/f_k \geq e^{1/\delta}} f_0^2(x)/f_{I_{1:k}}(x) dx \leq \int_0^1 \frac{2^5 AM^2}{k(C_1^{-2\alpha} \gamma^{1+1/\alpha})^4} dx \leq \frac{2^5 AM^2}{kC_1^{-8\alpha} \gamma^{4+4/\alpha}} \leq \frac{2^5 AM^2}{C_1^{-8\alpha} \gamma^{4+4/\alpha}},$$

where we used the boundedness of f_0 . Thus, as $\gamma \downarrow 0$,

$$\gamma^2 \log^r(L_1/\gamma) = \gamma^2 \log^r \left(\frac{2^{5/2} C_1^{4\alpha} A^{1/2} M}{\gamma^{1+2+2\alpha}} \right) \leq G_r \gamma^2 \log^r(1/\gamma),$$

for $r = 1, 2$ and positive constants G_1, G_2 . As $\gamma \downarrow 0$, the expression with $r = 2$ is larger and we conclude from the result of Lemma B.4 that

$$\max \{K(f_0, f_k), V(f_0, f_k)\} \leq G_3 c_3^2 \gamma^2 \log^2(1/\gamma), \tag{5.49}$$

for some $G_3 > 0$. Recalling that $\gamma = \epsilon/\log(1/\epsilon)$, we have that

$$\begin{aligned} G_3 c_3^2 \gamma^2 \log^2(1/\gamma) &= G_3 c_3^2 \gamma^2 \log^2(\gamma) \\ &= G_3 \frac{c_3^2 \epsilon^2}{\log^2(\epsilon)} \log^2 \{\epsilon/\log(1/\epsilon)\} \\ &= G_3 c_3^2 \epsilon^2 \left\{ \frac{\log(\epsilon/\log[1/\epsilon])}{\log(\epsilon)} \right\}^2 \\ &= G_3 c_3^2 \epsilon^2 \left\{ \frac{\log(\epsilon) + \log(-1/\log[\epsilon])}{\log(\epsilon)} \right\}^2, \end{aligned}$$

The limit of the factor inside the curly brackets is 1 as ϵ tends to 0. By continuity, this expression can as such be bounded uniformly by a universal constant for $\epsilon \leq (1 - e^{-1})/\sqrt{2}$ as it does not depend on any other variable than ϵ . It follows that by taking C_1 to be sufficiently large and hence c_3 sufficiently small, we have $G_2 c_3^2 \gamma^2 \log^2(1/\gamma) \leq \epsilon^2$, and combining this with the estimate (5.49), we see that

$$\mathcal{A}_\gamma = \left\{ \mathbf{p} \in \mathcal{S}_{k-1} : d_1(\mathbf{p}, \boldsymbol{\pi}) \leq C_1^{-2\alpha} \gamma^{1+1/\alpha}, \min_j p_j > (C_1^{-2\alpha} \gamma^{1+1/\alpha}/2)^4 \right\} \subseteq \mathcal{B}'_\epsilon(f_0).$$

Note that $\gamma^{1+1/\alpha} \leq \mathcal{O}(k^{-1-\alpha}) = o(k^{-1})$, which means that 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 B.3 are met, and we obtain the lower bound

$$P_n(\mathcal{A}_\gamma | k, I_{1:k}) \geq c_1 \exp \left(-c_4 k \log(2C_1^{2\alpha}/\gamma^{1+1/\alpha}) \right), \quad (5.50)$$

for some $c_4 > 0$. Note that as C_1 is a constant, we have that

$$c_4 k \log(2C_1^{2\alpha}/\gamma^{1+1/\alpha}) = c_4 \{ \log(2C_1^{2\alpha}) + (1 + 1/\alpha) \log(1/\gamma) \} = c_5 \log(1/\gamma),$$

for a suitable constant $c_5 > 0$ as $\gamma \downarrow 0$. In particular, this estimate together with (5.50) implies that

$$P_n(\mathcal{B}'_\epsilon(f_0)) \geq p_n(k) p_n(I_{1:k} | k) c_1 \exp \{ -c_5 k \log(1/\gamma) \} \quad (5.51)$$

It remains to show that we can write (5.51) in terms of ϵ only. To do this, we note that

$$\log \left(\frac{1}{\gamma} \right) = \log \left(\frac{1}{\epsilon / \log(1/\epsilon)} \right) = \log \left(\frac{1}{\epsilon} \right) + \log \{ \log(1/\epsilon) \}.$$

Since $\log \{ \log(1/\epsilon) \} / \log(1/\epsilon) \rightarrow 0$ as $\epsilon \downarrow 0$, the above display is bounded by $\log(1/\epsilon)$ times a constant for all sufficiently small $\epsilon > 0$. Thus, for a suitable constant $c_2 > 0$, we have $c_5 k \log(1/\gamma) \leq c_2 k \log(1/\epsilon)$. From this we conclude that

$$P_n(\mathcal{B}'_\epsilon(f_0)) \geq p_n(k) p_n(I_{1:k} | k) c_1 \exp \{ -c_2 k \log(1/\epsilon) \},$$

which was to be shown. $\square$

We now proceed with the proof of the main result of the section.

Proof of Theorem 5.4. We will show that the conditions of Theorem 3.4 hold for the two sequences defined by $\alpha_n = \beta_n = n^{-\alpha/(2\alpha+1)} \log(n)^{(\alpha+1)/(2\alpha+1)}$. We first focus our attention on condition (3.44). Let s_n be a sequence of natural numbers such that $C_1 \{ \beta_n / \log(\beta_n) \}^{-1/\alpha} \leq s_n \leq C_2 \{ \beta_n / \log(\beta_n) \}^{-1/\alpha}$, where C_1 is the constant from Lemma 5.7 and $C_2 > C_1$. An application of Lemma 5.7 with $k = s_n$ and $\epsilon = \beta_n$ then yields

$$P_n(\mathcal{B}'_{\beta_n}(f_0)) \geq p_n(s_n) p_n(I_{1:s_n} | s_n) c_1 \exp \{ -c_2 s_n \log(1/\beta_n) \} \quad (5.52)$$

for a partition $I_{1:s_n}$ and constants $c_1, c_2 > 0$. To continue our argument, we need to lower bound the probability of the partition $I_{1:s_n}$ given $k = s_n$. To this end, note that

$$p_n(I_{1:s_n} | s_n) = \binom{k_n - 1}{s_n - 1}^{-1} \geq \binom{k_n}{s_n}^{-1} \geq \left(\frac{ek_n}{s_n} \right)^{-s_n} \geq \exp \{ -s_n \log(k_n) \},$$

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 ,

$$p_n(I_{1:s_n} | s_n) \geq \exp \{ -s_n \log(A_1 n) \} \geq \exp \{ -A_2 s_n \log(n) \}, \quad (5.53)$$

for two constants $A_1, A_2 > 0$, since $\log(A_1 n) \leq A_2 \log(n)$ for all n provided A_2 is sufficiently large. A lower bound on $p_n(s_n)$ follows from the fact that $s_n \leq n$ and hence,

$$p_n(s_n) \geq D_1 \exp \{ -d_1 s_n \log(s_n) \} \geq D_1 \exp \{ -d_1 s_n \log(n) \}$$

Now note that

$$\begin{aligned} s_n &\leq C_2 \left\{ n^{-\alpha/(2\alpha+1)} \log(n)^{(\alpha+1)/(2\alpha+1)} / \log \left(n^{\alpha/(2\alpha+1)} \log(n)^{-(\alpha+1)/(2\alpha+1)} \right) \right\}^{-1/\alpha} \\ &= C_2 n^{1/(2\alpha+1)} \left\{ \log(n)^{(\alpha+1)/(2\alpha+1)} / \left[\alpha(2\alpha+1)^{-1} \log(n) + o(\log(n)) \right] \right\}^{-1/\alpha}, \end{aligned}$$

where we used the following to establish the last equality:

$$\lim_{n \rightarrow \infty} \log \left(\log(n)^{-(\alpha+1)/(2\alpha+1)} \right) / \log(n) = 0.$$

Thus, as $n \rightarrow \infty$,

$$s_n = \mathcal{O} \left(n^{1/(2\alpha+1)} \log(n)^{1/(2\alpha+1)} \right).$$

Which implies that

$$s_n \log(n) = \mathcal{O} \left(n^{1/(2\alpha+1)} \log(n)^{(2\alpha+2)/(2\alpha+1)} \right). \quad (5.54)$$

On the other hand,

$$n\beta_n^2 = n^{1-2\alpha/(2\alpha+1)} \log(n)^{(2\alpha+2)/(2\alpha+1)} = n^{1/(2\alpha+1)} \log(n)^{(2\alpha+2)/(2\alpha+1)}, \quad (5.55)$$

so $s_n = \mathcal{O}(n\beta_n^2)$, and we conclude by (5.53) that $p_n(I_{1:s_n} | s_n) \geq \exp(-A_3 n\beta_n^2)$ for a constant $A_3 > 0$. Observe that $\log(1/\beta_n)/\log(n)$ converges to a positive finite limit as $n \rightarrow \infty$, and an application of (5.54) yields

$$s_n \log(1/\beta_n) = \mathcal{O}(n\beta_n^2),$$

Using (5.52), we have

$$\begin{aligned} P_n(\mathcal{B}'_{\beta_n}(f_0)) &\geq p_n(s_n) p_n(I_{1:s_n} | s_n) c_1 \exp \{ -c_2 s_n \log(1/\beta_n) \} \\ &\geq D_1 c_1 \exp \{ -d_1 n\beta_n^2 \} \exp \{ -A_3 n\beta_n^2 \} \exp \{ -c_2 s_n \log(1/\beta_n) \} \\ &\geq A_4 \exp \{ -A_5 n\beta_n^2 \}, \end{aligned}$$

where A_4 and A_5 are positive constants and we used that

$$d_1 n\beta_n^2 + A_3 n\beta_n^2 + c_2 s_n \log(1/\beta_n) = \mathcal{O}(n\beta_n^2).$$

We conclude that (3.44) holds.

To verify the covering number condition (3.42) and corresponding prior mass condition (3.43), we consider the sieve $\mathcal{F}_n = \cup_{k=1}^{t_n} \mathcal{H}_k$, where $\mathcal{H}_k$ is as defined in (5.25) and t_n is a sequence of positive integers satisfying $E_1 n^{1/(2\alpha+1)} \log(n)^{1/(2\alpha+1)} \leq t_n \leq E_2 n^{1/(2\alpha+1)} \log(n)^{1/(2\alpha+1)}$ for two positive constants $E_1 < E_2$ to be chosen later.

We first show the prior condition (3.43). First, we deduce the upper bound

$$P_n(\mathcal{F}_n^c) \leq \sum_{k=t_n+1}^{k_n} p_n(k) \leq D_2 \sum_{k=t_n+1}^{\infty} \exp(-d_2 k \log(k)) \leq D_3 \exp(-d_2 t_n \log(t_n)).$$

Since $\log(t_n)/\log(n)$ converges to a positive finite limit as $n \rightarrow \infty$, we can find a constant d'_2 such that the last expression is upper bounded by $D_3 \exp(-d'_2 t_n \log(n)) = D_3 \exp(-d'_2 n\beta_n^2)$, where the equality follows from (5.55). Choosing E_1, E_2 such that $E_1 \geq (A_5 + 4)/d'_2$, we obtain

$$P_n(\mathcal{F}_n^c) \leq D_3 \exp(-\{A_5 + 4\}n\beta_n^2),$$

establishing (3.43).

Finally, we need to establish the entropy condition (3.42). We define the sequence $\alpha_n = n^{-\alpha/(2\alpha+1)} \log(n)^{(\alpha+1)/(2\alpha+1)}$, from which it follows that

$$n\alpha_n^2 = n^{1-2\alpha/(2\alpha+1)} \log(n)^{(\alpha+1)/(2\alpha+1)} = n^{1/(2\alpha+1)} \log(n)^{(\alpha+1)/(2\alpha+1)}.$$

Using Lemma 5.1, we have

$$N(\alpha_n/2, \mathcal{F}_n, d_H) \leq \left(\frac{Lk_n}{\alpha_n t_n} \right)^{t_n} \leq \exp(t_n \{\log(1/\alpha_n) + \log(Lk_n)\}).$$

Note that $\log(1/\alpha_n) + \log(Lk_n) = \mathcal{O}(\log(n))$. It follows that

$$\begin{aligned} t_n \{\log(1/\alpha_n) + \log(Lk_n)\} &\leq E_2 n^{1/(2\alpha+1)} \log(n)^{1/(2\alpha+1)} \mathcal{O}(\log(n)) \\ &= \mathcal{O}\left(n^{1/(2\alpha+1)} \log(n)^{(2\alpha+2)/(2\alpha+1)}\right) \\ &= \mathcal{O}(n\alpha_n^2). \end{aligned}$$

Taken together, the above inequality implies that for some constant $c_5 > 0$,

$$N(\alpha_n/2, \mathcal{F}_n, d_H) \leq \exp(c_5 n\alpha_n^2),$$

establishing (3.42).

Since all the conditions of Theorem 3.4 are met, we conclude that the posterior distribution converges at rate $\epsilon_n = \max\{\alpha_n, \beta_n\} = \alpha_n$. $\square$

5.6 Simulation study: Irregular histograms

To end the chapter, we investigate the performance of our irregular histogram procedure in simulations, comparing it to the irregular histograms procedures of Section 5.1.1. We conclude the section with a discussion of the relative strengths of regular and irregular histograms.

5.6.1 Software implementations used in the simulation study

An overview of the irregular histogram methods included in the study are shown in Table 5.1. For both the random histogram priors in Table 5.1, note that we used the $\text{Dir}(a\mathbf{p}_0)$ distribution with $p_{0,j} = |I_{k,j}|$ as the prior for the bin probabilities $\mathbf{p} \mid k, I_{1:k}$. For the choices of the Dirichlet concentration parameter a , we have considered the choices $a = 1$ and $a = \sqrt{n}$. The last choice $a = \sqrt{n}$ is motivated by the good performance of the corresponding choice in the regular histogram model.

The choice of software implementation is critical in a simulation study. First, if a given implementation runs too slowly, it may not be feasible to include it in the simulation study to begin with, as the histogram has to be computed repeatedly for different samples. Another important point is that some of the methods proposed in Section 5.1.1 are based on maximizing criteria where computing the exact maximizer is slow and as such various heuristics must be employed to compute an approximation the optimum in a reasonable amount of time. In such cases, the quality of the heuristic used becomes part of the procedure in practice, and it should as such be accounted for when estimating the risk of the procedure.

For our simulation study, we either used our own **Julia** code to find the optimal histogram, or relied on prewritten libraries by the authors of the respective histogram methods. Specifically, the implementations used for each method are as follows:

Table 5.1: Overview of the methods included in the simulation study.

Method	Abbreviation	Description
Rozenholc et al. pen^B	RMG-B	Section 5.1.1a
Rozenholc et al. pen^R	RMG-R	
Normalized Maximum Likelihood	NML	
G-Enum	G-Enum	Section 5.1.1b
Bayesian Blocks	BB	
Kullback–Leibler cross-validation	KLCV	Section 5.1.1c
$\mathbb{L}_2$ cross-validation	L2CV	
Essential Histogram	EH	Section 5.1.1d
Taut String	TS	
Irregular Random histogram, $a = 1$	B-B	Section 5.6.1
Irregular Random histogram, $a = \sqrt{n}$	B-MM	

- For most of the methods, we used the approach of Rozenholc et al. (2010) by first using their greedy algorithm to construct a grid consisting of at most $q_n = \{n \log(n)\}^{1/3}$ points and then subsequently running the KDP algorithm to find the optimal partition based on this reduced grid, see Section 5.2.2 for the computational details. We used this approach when maximizing the criteria L2CV, KLCV, RMG and NML as well as the Bayesian criteria B and MM. To compute the optimal histograms for each method, we used the **Julia** implementation of `AutoHist.jl` (Simensen, 2025), although we note that many of these criteria have also been implemented in the **R** package `histogram` (Mildenberger et al., 2019).
- For the Taut String (TS) method we used the `ftnonpar` **R** package (Davies and Kovac, 2012).
- For the Essential Histogram (EH) we used the **R** package `essHist` (Li and Sieling, 2019).
- For the Bayesian Blocks (BB) histogram, we relied on the implementation in **Python**’s `astropy` library (Price-Whelan et al., 2022).
- For the G-Enum histogram we used the binary standalone implementation of Mendizábal et al. (2022).

Due to the intense computational demands of Monte Carlo simulations, we had to limit our simulation study to those methods that were able to compute the optimal histogram for hundreds of random samples in a reasonable amount of time. This meant that we had to exclude some of the methods covered in Section 5.1.1, as their slow runtimes made them unsuitable for our simulation study.

- The Stochastic Complexity approach of Rissanen et al. (1992) crucially depends on the choice of certain hyperparameters, where no default choice is given by the authors. Rissanen et al. (1992) suggests performing an exhaustive hyperparameter search, making it too computationally demanding for our study.
- The Essential Histogram and Bayesian Blocks implementations showed adequate performance for smaller datasets ($n \leq 10^3$), but turned out to be too slow for larger ones ($n \geq 5 \times 10^3$). As such, we decided to exclude them from the subsequent simulation study.

5.6.2 Choice of discretization schemes

All of the histogram approaches included in our simulation study are based on discretizing $[x_{(1)}, x_{(n)}]$ into a finest possible grid, defined by the set $\{\tau_{n,j} \in [0, 1] : 0 \leq j \leq k_n\}$ of permitted interval endpoints. As such, the choice of discretization scheme will have an impact on the resulting histogram estimators. For many of the methods considered here, such as RMG-B, RMG-R, TS, NML and G-Enum, the authors give a suggestion as to how the grid should be constructed. However, for the L2CV, NML and KLCV approaches as well as our Bayesian methods from Example 5.2 there are no universal recommendations given in the literature, and we therefore give a short description of some of the issues pertaining to the choice of the finest possible grid.

One popular approach to discretization is to simply use the data points themselves as the finest permitted grid; this is the discretization scheme used by RMG-B, RMG-R and TS. Another commonly employed discretization scheme is based on using a fine regular grid, constructed so that the ratio k_n/n tends to zero rather slowly; such an approach is used by the G-Enum method. Rozenholc et al. (2010) compared the use of regular and data-based grid for their proposed penalized likelihood criteria, as well as L2CV. Their results show that the regular grid approach often yields small improvements in terms of Hellinger risk for some densities, but that the performance worsens significantly for heavy-tailed densities and densities with pronounced and narrow peaks. To further investigate this issue, we ran some simulations on the test suite in Figure C.1, recording the estimated Hellinger risk for both data-based and regular grids. In most of the simulations, the L2CV and KLCV histograms performed better when using a data-based grid. For the Bayesian methods, the regular grid tended to produce slightly lower risks, although the differences were generally small in relative terms. The suggestion of Kontkanen and Myllymäki (2007) for the NML histogram is to put two grid points on either side of each data point, leading to a total of $k_n = 2n - 1$ maximal number of bins. In the interest of reducing the computational complexity of the method, we instead used the data-dependent grid with a grid point at each observation when using the NML criterion (5.3) to build a histogram.

Another important matter to consider when drawing an irregular histogram is whether or not one should impose restrictions on the minimal allowed bin width. If such restrictions are not in place, then the irregular histogram procedure may produce very small intervals with very high estimated density, but with few observations, even when the true density has no peak in the vicinity of this interval. Not only can these narrow intervals negatively impact the quality of the resulting density estimates in terms of risk, but they often lead to visually displeasing density estimates, where the graphical display is dominated by a few narrow spikes that are very large relative to the rest of the density estimate. Typical restrictions imposed on the optimal partition are of the form $\min_j |I_{k,j}| \geq b_n$, where $b_n > 0$ is the minimum allowed bin width, which in general decreases as the sample size increases. The use of such restrictions are not without their drawbacks, however; if the true density has narrow, pronounced peaks then the estimated density may be unable to recover them due to the restriction on the smallest allowed interval.

Rozenholc et al. (2010) found that the irregular L2CV histogram performed poorly for most densities when no restrictions on the minimal allowed bin width are in place, but that the imposition of such a restriction at best had a small positive effect on penalized likelihood-based histograms. To decide on whether to impose minimum bin width restrictions or not, we ran some exploratory simulations on the 6 densities showed in Figure C.1 with data-driven grids. We found that the L2CV and KLCV histograms

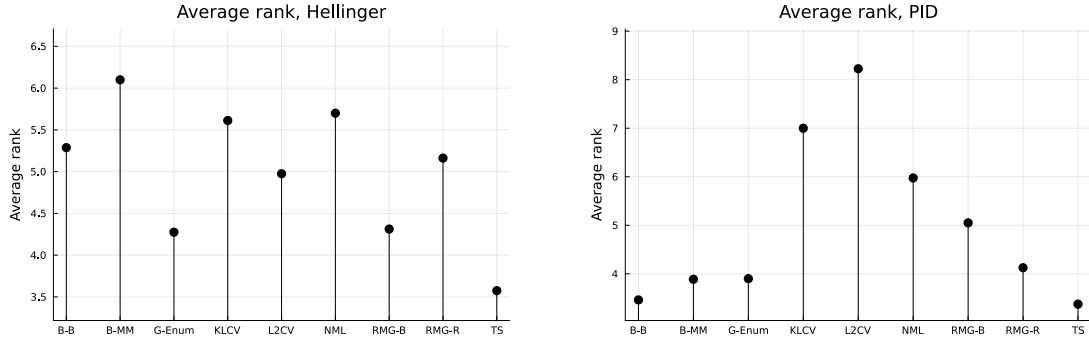


Figure 5.5: Left: overall average ranks by Hellinger distance. Right: overall average ranks by PID loss.

Table 5.2: Average ranks, Hellinger risk

n	50	200	1000	5000	25000
B-B	3.44	4.75	5.44	6.31	6.5
B-MM	4.31	6.12	6.88	6.62	6.56
G-Enum	5.81	5.06	3.69	3.5	3.31
KLCV	5.25	5.56	6.12	5.75	5.38
L2CV	4.62	5.31	5.38	4.94	4.62
NML	7.75	5.31	5.19	5.31	4.94
RMG-B	5.06	5.12	3.94	3.88	3.56
RMG-R	4.94	5.19	5.62	5.12	4.94
TS	3.81	2.56	2.75	3.56	5.19

Table 5.3: Average ranks, PID risk

n	50	200	1000	5000	25000
B-B	2.88	3.31	4.0	3.56	3.56
B-MM	3.62	5.44	4.5	3.12	2.75
G-Enum	4.88	4.19	3.12	4.06	3.25
KLCV	6.44	5.75	6.81	8.0	8.0
L2CV	7.69	7.31	8.12	9.0	9.0
NML	8.06	5.44	5.75	5.31	5.31
RMG-B	3.38	4.88	5.75	5.69	5.56
RMG-R	4.94	5.25	4.06	2.62	3.75
TS	3.12	3.44	2.88	3.62	3.81

frequently selected very narrow intervals, producing sharp, spurious peaks in cases where the true density has none. This was not a problem for the Bayesian methods of Example 5.2, suggesting that using a prior that disallows the inclusion of small intervals would at best have a modest positive effect on the risk of the estimator. Imposing the restriction $\min_j |I_{k,j}| \geq (x_{(n)} - x_{(1)}) \log^{1.5}(n)/n$ on the L2CV and KLCV criteria as in Rozenholc et al. (2010) removed these narrow spikes, and resulted in density estimates with better performance in terms of risk. As such, we decided to enforce this minimal interval width restriction on the L2CV and KLCV methods in our subsequent simulation study.

5.6.3 Irregular histograms: results

Figure 5.5 shows the overall average rank for each method $\bar{r}(m)$ (4.75) in terms of Hellinger and PID risk, while Tables 5.2 and 5.3 show the average rank $\bar{r}_n(m)$ (4.74) for the Hellinger and PID losses respectively, separated by sample size. Finally, a boxplot of $\text{LRR}_n(m, f_0)$ (4.76) for $n = 200$ and $n = 25000$ can be found in Figure 5.6. The results for the $\mathbb{L}_2$ loss are deferred to Appendix C.3.

The Taut String method is the overall winner in terms of both Hellinger and PID risk. It is never far from the best method in terms of Hellinger risk, and produces the overall lowest risk for many of the different densities. The results of the Taut String method are especially impressive for the Claw (#7), Asymmetric Double Claw (#9) and Sawtooth (#10) densities, where none of the other methods come close to the Taut String in terms of Hellinger risk for larger sample sizes. Both the G-enum

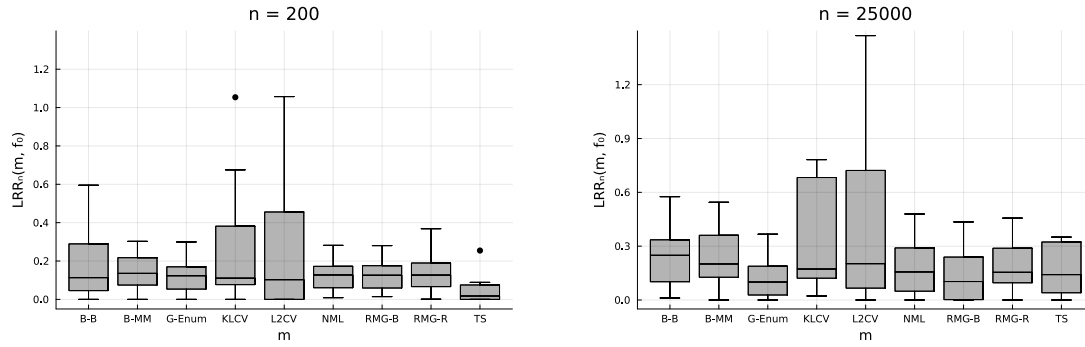


Figure 5.6: Boxplots of log-relative Hellinger risk by method for sample sizes $n = 200$ (left) and $n = 25000$ (right).

and RMG-B methods perform well with respect to the Hellinger metrics, and are never much worse than the Taut String. The RMG-R method performs quite similarly to the RMG-B method for most densities, but does considerably worse for the Chisq(1) (#3). Curiously, the Bayesian MM method ($a = \sqrt{n}$) was the worst performer, losing out to the B-B method ($a = 1$). This is a reversal of the situation in Section 4.8, where the MM choice was among the best performing estimators in terms of Hellinger risk, and the B method among the worst. The B-B method performs better for most densities, with the exception of the Trimodal Uniform (#14), where B-MM does much better. The cross-validation based estimators show good performance for densities #1, #5 and #8, but have much higher risks at other densities. The performance of the Bayesian methods are better with respect to the $\mathbb{L}_2$ loss, with the B-B method doing particularly well for smaller sample sizes. The Taut String method also does very well in terms of $\mathbb{L}_2$ risk, especially for larger sample sizes. In terms of PID risk, both of the Bayesian methods are among the top performers, with the B-B method achieving an overall rank similar to the Taut String method. The G-enum and RMG-R histograms perform about equally well. L2CV and KLCV do very poorly in terms of PID risk, mainly due to the fact that these both tend to produce a large number of spurious modes.

Overall, the results shown here are in agreement with the simulation studies of Davies et al. (2009) and Rozenholc et al. (2010), who both found that the Taut String method was among the best performers in terms of Hellinger risk. Davies et al. (2009) also found that the Taut String performed well in terms of PID loss compared to the other methods in their simulation study, but all the competing procedures in this study were based on regular histograms. For the PID loss, our results indicate that the Taut String procedure also does comparatively well to the other procedures included in our study, but the advantage is smaller than the one reported in Davies et al. (2009).

Table 5.5: PID risks for the densities in Figure 4.4.

Density	n	B-B	B-MM	G-Enum	KLCV	L2CV	NML	RMG-B	RMG-R	TS
1	50	1.43	1.436	1.48	1.6	1.878	1.892	1.462	1.448	1.604
	200	1.05	1.088	1.14	1.784	3.02	1.084	1.08	1.048	1.18
	1000	0.608	0.576	0.668	3.542	5.586	0.646	0.66	0.588	0.588
	5000	0.146	0.14	0.218	7.336	8.566	0.16	0.226	0.148	0.288
	25000	0.012	0.004	0.024	9.854	10.676	0.014	0.094	0.014	0.05
2	50	0.29	0.272	0.276	1.882	3.034	0.492	0.43	0.272	0.818
	200	0.008	0.004	0.0	3.796	5.62	0.044	0.138	0.008	0.428
	1000	0.008	0.0	0.0	7.358	8.998	0.008	0.04	0.008	0.43
	5000	0.0	0.0	0.0	9.872	11.026	0.0	0.012	0.0	0.856
	25000	0.0	0.0	0.0	11.77	12.52	0.0	0.004	0.004	0.666
3	50	0.356	0.62	1.264	2.01	2.348	2.958	0.174	1.48	0.24
	200	0.002	0.0	0.02	2.144	2.962	0.044	0.04	0.108	0.024
	1000	0.002	0.0	0.002	1.06	2.004	0.05	0.072	0.0	0.0
	5000	0.002	0.0	0.0	2.842	3.48	0.028	0.086	0.002	0.0
	25000	0.0	0.0	0.002	3.64	4.016	0.016	0.1	0.002	0.004
4	50	1.42	1.576	1.708	1.998	2.176	2.488	1.386	1.72	0.866
	200	0.74	0.92	0.768	0.902	0.828	0.742	0.722	1.008	0.118
	1000	0.004	0.024	0.004	0.442	1.072	0.034	0.05	0.024	0.002
	5000	0.002	0.0	0.002	1.458	2.012	0.008	0.054	0.0	0.0
	25000	0.002	0.0	0.0	2.486	2.832	0.004	0.05	0.002	0.002
5	50	1.244	1.316	1.396	1.322	1.686	2.494	1.266	1.324	1.256
	200	0.53	0.624	0.55	0.5	1.39	0.552	0.574	0.592	0.448
	1000	0.008	0.004	0.008	0.764	1.996	0.056	0.084	0.008	0.004
	5000	0.0	0.0	0.002	3.454	4.49	0.028	0.09	0.0	0.0
	25000	0.002	0.0	0.0	5.45	6.124	0.012	0.066	0.006	0.0
6	50	2.06	2.336	2.958	3.934	4.306	2.426	1.49	2.728	1.472
	200	0.942	1.318	1.85	4.98	6.712	0.676	0.652	1.662	0.578
	1000	0.006	0.004	0.002	4.534	5.936	0.03	0.046	0.004	0.0
	5000	0.008	0.0	0.002	6.222	7.066	0.022	0.068	0.0	0.004
	25000	0.004	0.0	0.002	7.3	7.742	0.004	0.066	0.004	0.002
7	50	5.248	5.262	5.34	5.412	5.856	5.934	5.258	5.288	5.136
	200	4.218	4.22	4.354	5.0	4.262	4.242	4.242	4.216	3.524
	1000	2.104	3.882	1.046	0.49	1.08	2.212	1.71	2.96	0.022
	5000	0.004	0.0	0.0	1.892	2.458	0.008	0.05	0.002	0.008
	25000	0.0	0.0	0.0	3.17	3.622	0.012	0.088	0.008	0.0
8	50	2.582	2.566	2.562	2.874	3.34	2.78	2.608	2.574	2.66
	200	2.904	2.948	2.884	2.642	3.73	2.922	2.92	2.946	2.878
	1000	2.086	2.458	1.702	3.716	5.486	2.22	2.068	2.252	2.12
	5000	0.964	0.992	0.776	6.33	7.638	0.986	1.0	0.972	0.538
	25000	0.22	0.536	0.054	9.092	9.832	0.276	0.182	0.196	0.068
9	50	6.832	6.84	6.828	6.506	6.714	6.962	6.814	6.848	6.62
	200	6.884	6.932	6.694	6.144	6.466	6.89	6.868	6.934	6.494
	1000	5.588	6.06	5.15	3.94	4.784	5.692	5.502	5.704	5.65
	5000	4.24	4.232	4.016	5.316	6.24	4.232	4.154	4.24	3.516
	25000	1.978	2.886	1.408	6.688	7.3	2.17	1.534	1.868	1.656
10	50	10.894	10.93	10.948	10.27	11.176	10.822	10.708	10.94	9.404
	200	10.792	10.936	10.988	10.304	7.618	10.868	10.706	10.88	2.162
	1000	0.204	2.924	0.022	0.076	0.51	0.404	0.352	0.266	0.08
	5000	0.0	0.0	0.008	0.804	1.194	0.072	0.16	0.0	0.008
	25000	0.008	0.0	0.02	0.442	0.468	0.04	0.17	0.026	0.004
11	50	1.352	1.35	1.172	1.674	2.1	1.698	1.38	1.352	1.506
	200	1.31	1.32	0.796	2.306	4.008	1.354	1.368	1.322	1.404
	1000	1.02	1.076	1.008	5.27	7.456	1.052	1.052	1.036	1.012
	5000	0.578	0.556	0.612	9.408	10.604	0.594	0.652	0.554	0.552
	25000	0.144	0.164	0.112	12.946	13.654	0.152	0.196	0.146	0.136
12	50	0.248	0.252	0.288	1.696	1.86	3.172	0.294	0.26	0.596
	200	0.032	0.052	0.066	1.274	1.324	0.168	0.14	0.052	0.086
	1000	0.006	0.0	0.004	0.042	0.18	0.034	0.084	0.004	0.0
	5000	0.0	0.0	0.0	0.692	1.224	0.036	0.076	0.0	0.004
	25000	0.002	0.0	0.002	2.098	2.48	0.012	0.092	0.0	0.002
13	50	4.548	5.006	5.444	6.078	4.866	8.808	4.314	5.506	3.512
	200	1.668	2.14	1.668	4.66	4.938	1.748	1.752	1.868	1.822
	1000	0.518	0.508	0.598	1.812	2.696	0.562	0.638	0.512	0.78
	5000	0.156	0.188	0.26	2.724	3.132	0.216	0.3	0.16	0.23
	25000	0.028	0.016	0.044	2.454	2.662	0.032	0.16	0.042	0.044
14	50	0.006	0.002	0.008	4.0	4.0	8.17	0.056	0.888	0.316
	200	0.006	0.0	0.0	4.0	4.0	0.03	0.03	0.0	0.512
	1000	0.0	0.0	0.0	2.0	2.0	0.004	0.016	0.0	0.616
	5000	0.0	0.0	0.004	2.694	2.926	0.0	0.0	0.0	0.958
	25000	0.0	0.0	0.0	1.29	1.482	0.0	0.0	0.0	1.266
15	50	2.282	2.428	2.788	3.476	3.59	5.018	2.336	2.418	2.264
	200	1.954	1.97	1.94	1.712	2.502	2.03	2.022	1.964	1.94
	1000	2.01	1.996	1.986	5.542	6.816	2.038	2.05	2.008	1.708
	5000	1.864	1.96	1.744	10.088	10.876	1.92	1.828	1.88	0.982
	25000	0.348	0.772	0.198	13.536	14.098	0.506	0.298	0.348	0.35
16	50	2.812	2.992	3.278	4.642	4.706	5.536	2.904	3.04	2.264
	200	2.804	2.844	2.784	2.244	1.946	2.824	2.826	2.83	1.854
	1000	1.518	1.994	1.416	4.28	5.706	1.608	1.508	1.718	1.348
	5000	0.504	0.778	0.566	9.188	9.91	0.666	0.52	0.52	1.188
	25000	0.002	0.0	0.0	12.656	13.094	0.002	0.01	0.002	1.068

Table 5.4: Estimated Hellinger risks for the densities in Figure 4.4.

Density	n	B-B	B-MM	G-Enum	KLCV	L2CV	NML	RMG-B	RMG-R	TS
1	50	0.331	0.338	0.358	0.29	0.276	0.351	0.338	0.338	0.333
	200	0.221	0.226	0.217	0.196	0.177	0.224	0.223	0.224	0.191
	1000	0.14	0.143	0.129	0.128	0.117	0.14	0.138	0.139	0.128
	5000	0.09	0.092	0.079	0.083	0.076	0.088	0.085	0.086	0.093
	25000	0.057	0.058	0.048	0.049	0.047	0.055	0.052	0.053	0.065
2	50	0.186	0.185	0.189	0.244	0.235	0.194	0.193	0.185	0.202
	200	0.094	0.093	0.096	0.184	0.165	0.094	0.097	0.094	0.101
	1000	0.042	0.042	0.043	0.129	0.11	0.042	0.043	0.042	0.046
	5000	0.019	0.019	0.019	0.074	0.063	0.019	0.019	0.019	0.021
	25000	0.008	0.008	0.008	0.038	0.033	0.008	0.008	0.008	0.009
3	50	0.385	0.392	0.397	0.424	0.408	0.414	0.388	0.402	0.316
	200	0.271	0.273	0.273	0.322	0.312	0.259	0.26	0.292	0.202
	1000	0.176	0.175	0.174	0.245	0.241	0.161	0.159	0.197	0.131
	5000	0.113	0.112	0.112	0.182	0.181	0.101	0.098	0.134	0.092
	25000	0.072	0.07	0.071	0.132	0.132	0.063	0.06	0.091	0.064
4	50	0.358	0.35	0.342	0.343	0.32	0.366	0.349	0.354	0.286
	200	0.25	0.237	0.226	0.214	0.2	0.232	0.231	0.237	0.184
	1000	0.168	0.158	0.146	0.148	0.146	0.147	0.145	0.155	0.13
	5000	0.111	0.1	0.089	0.113	0.117	0.093	0.089	0.098	0.088
	25000	0.073	0.062	0.054	0.063	0.071	0.058	0.055	0.061	0.064
5	50	0.358	0.362	0.37	0.325	0.305	0.383	0.363	0.363	0.315
	200	0.251	0.254	0.237	0.204	0.188	0.249	0.249	0.252	0.205
	1000	0.164	0.166	0.145	0.124	0.119	0.159	0.156	0.16	0.14
	5000	0.105	0.105	0.089	0.08	0.077	0.1	0.096	0.099	0.098
	25000	0.066	0.065	0.054	0.049	0.048	0.062	0.058	0.06	0.068
6	50	0.316	0.327	0.33	0.328	0.31	0.346	0.33	0.325	0.315
	200	0.223	0.233	0.229	0.257	0.24	0.228	0.227	0.23	0.212
	1000	0.141	0.151	0.144	0.193	0.181	0.142	0.14	0.151	0.135
	5000	0.091	0.098	0.094	0.133	0.129	0.09	0.087	0.1	0.09
	25000	0.058	0.063	0.06	0.09	0.089	0.057	0.054	0.065	0.061
7	50	0.358	0.361	0.389	0.338	0.346	0.387	0.362	0.36	0.37
	200	0.25	0.246	0.25	0.245	0.241	0.248	0.248	0.246	0.245
	1000	0.183	0.199	0.163	0.151	0.147	0.183	0.177	0.187	0.131
	5000	0.114	0.115	0.103	0.092	0.09	0.111	0.108	0.11	0.073
	25000	0.072	0.074	0.062	0.056	0.055	0.07	0.067	0.068	0.043
8	50	0.311	0.315	0.327	0.297	0.287	0.323	0.313	0.315	0.312
	200	0.209	0.211	0.211	0.202	0.186	0.211	0.209	0.209	0.195
	1000	0.139	0.142	0.128	0.13	0.12	0.139	0.136	0.138	0.126
	5000	0.088	0.089	0.078	0.082	0.074	0.086	0.083	0.084	0.068
	25000	0.056	0.057	0.048	0.049	0.046	0.054	0.051	0.052	0.049
9	50	0.307	0.309	0.313	0.301	0.292	0.319	0.311	0.309	0.319
	200	0.209	0.213	0.217	0.209	0.192	0.213	0.211	0.21	0.207
	1000	0.144	0.15	0.135	0.139	0.126	0.146	0.143	0.145	0.116
	5000	0.095	0.097	0.086	0.09	0.083	0.095	0.092	0.093	0.074
	25000	0.062	0.064	0.054	0.054	0.051	0.061	0.058	0.059	0.045
10	50	0.388	0.387	0.384	0.416	0.408	0.408	0.393	0.387	0.42
	200	0.345	0.344	0.344	0.357	0.348	0.345	0.346	0.345	0.337
	1000	0.225	0.254	0.221	0.209	0.211	0.23	0.229	0.229	0.191
	5000	0.152	0.173	0.14	0.125	0.125	0.156	0.15	0.152	0.111
	25000	0.098	0.107	0.087	0.083	0.083	0.1	0.095	0.096	0.07
11	50	0.295	0.298	0.301	0.275	0.258	0.308	0.301	0.299	0.299
	200	0.191	0.198	0.191	0.193	0.172	0.199	0.197	0.196	0.18
	1000	0.125	0.132	0.117	0.134	0.119	0.129	0.127	0.127	0.111
	5000	0.078	0.083	0.071	0.086	0.076	0.081	0.078	0.078	0.075
	25000	0.048	0.052	0.043	0.051	0.047	0.05	0.047	0.048	0.053
12	50	0.492	0.426	0.409	0.799	0.81	0.449	0.415	0.413	0.378
	200	0.347	0.302	0.277	0.703	0.705	0.279	0.279	0.291	0.245
	1000	0.222	0.188	0.166	0.211	0.233	0.174	0.172	0.179	0.152
	5000	0.142	0.117	0.102	0.096	0.093	0.109	0.105	0.109	0.103
	25000	0.091	0.073	0.061	0.053	0.051	0.068	0.064	0.067	0.072
13	50	0.485	0.502	0.521	0.529	0.547	0.526	0.496	0.511	0.458
	200	0.306	0.331	0.318	0.387	0.456	0.316	0.318	0.321	0.271
	1000	0.205	0.222	0.199	0.248	0.282	0.207	0.204	0.21	0.152
	5000	0.129	0.141	0.12	0.124	0.128	0.129	0.125	0.129	0.093
	25000	0.081	0.089	0.073	0.066	0.066	0.081	0.077	0.079	0.057
14	50	0.519	0.42	0.36	1.239	1.264	0.532	0.411	0.656	0.432
	200	0.32	0.213	0.176	1.237	1.262	0.206	0.205	0.202	0.228
	1000	0.192	0.094	0.086	0.954	1.027	0.089	0.09	0.089	0.103
	5000	0.12	0.038	0.032	0.953	0.968	0.036	0.036	0.036	0.045
	25000	0.078	0.015	0.015	0.095	0.406	0.014	0.014	0.014	0.019
15	50	0.387	0.412	0.447	0.409	0.418	0.443	0.406	0.415	0.388
	200	0.239	0.243	0.253	0.221	0.281	0.241	0.242	0.241	0.223
	1000	0.103	0.105	0.112	0.135	0.172	0.104	0.103	0.103	0.127
	5000	0.049	0.049	0.052	0.084	0.092	0.049	0.049	0.049	0.062
	25000	0.024	0.026	0.025	0.046	0.048	0.025	0.024	0.024	0.027
16	50	0.348	0.363	0.395	0.468	0.472	0.406	0.367	0.365	0.349
	200	0.225	0.228	0.239	0.231	0.298	0.229	0.229	0.228	0.229
	1000	0.111	0.116	0.116	0.143	0.193	0.112	0.112	0.113	0.117
	5000	0.053	0.056	0.057	0.084	0.099	0.055	0.053	0.053	0.063
	25000	0.022	0.022	0.024	0.046	0.05	0.022	0.022	0.022	0.03

5.6.4 A comparison between regular and irregular histograms

To end the section, we comment on some of the differences between irregular and regular histograms.

The most striking difference between the regular and irregular histogram procedures considered in our simulation study is that the latter perform much better with respect to mode identification than the former. The PID risks of regular histograms (Table 4.5) are consistently much larger than the corresponding PID risks of irregular histograms (Table 5.5). In particular, irregular histogram procedures typically become better at mode identification as the sample size increases. This is not true of regular histograms, where the ability to find modes in the true density frequently worsens for larger samples. As an example of a density where regular histograms struggle with peak identification, consider the Harp density of Figure 4.4. This density has multiple modes at different scales; the leftmost peak is very narrow, and the other modes become flatter as one moves from left to right. To showcase the superior performance of irregular histograms in this case, we generated a sample of size $n = 5000$ from the Harp density, and fitted the Bayesian regular B histogram as well as the Bayesian irregular B histogram. This particular sample was chosen so that the PID loss was close to the value of the PID risk for both methods. The resulting histograms, along with their respective modes, are shown in Figure 5.7, with the tolerance regions for the PID loss shaded in gray. The first coordinate of correctly identified modes are marked by colored dots, while spurious modes are marked with black dots. The regular histogram can identify the locations of the first four modes, but misses the fifth. Although the regular histogram can detect most of the true modes in this case, this comes at the cost of a large number of spurious peaks in the histogram density, especially near the rightmost mode. The use of a global bin width $h = 1/k$ results in a somewhat unsatisfactory compromise between smoothing and small bias, as the bins are too wide near the leftmost peak and far too narrow in the right tail of the density, resulting in a high number of spurious modes near the flattest peak. In contrast, the irregular histogram is able to correctly identify all peaks in the true density and does not produce any spurious modes in the process. The advantage of the irregular histogram is rather immediate, as the bin widths adapt locally to the shape of the density, choosing smaller bins in regions where the density is rapidly changing and a larger bin width in the flatter regions. Although it may be argued that a statistician using a regular histogram will not necessarily conclude that a histogram mode at a given location indicates that the true density has a mode close to this point, most of the irregular histogram methods studied here appear to reliably detect well-separated modes automatically without the need for the user's judgment for larger sample sizes.

To compare the Hellinger risks of regular and irregular histograms, we limited our attention to a few of the best performing methods. We included the regular histogram methods B-MM and AIC, as well as the TS and G-Enum irregular histograms. For these four methods, we computed the logarithm of the risk relative to the best method at each density for sample sizes $n = 200$ and $n = 25\,000$. The results are shown in Figure 5.8. The regular histogram methods yielded large relative risks $\text{LRR} \geq 1.6$ for the trimodal uniform density (#14), and are not shown in the above plots, with the exception of AIC and $n = 25\,000$, whose log-relative risk of 0.24 is shown.

The irregular methods produce much lower risks for the heavy-tailed Chisq(1) (#3) and LogNormal(0, 1) (#4) densities. This difference is particularly pronounced for the larger sample size. Regular histograms struggle somewhat with heavy-tailed densities, since the use of a global bin width results in oversmoothing in high-probability regions and undersmoothing in the tails. The irregular histograms also have an advantage for

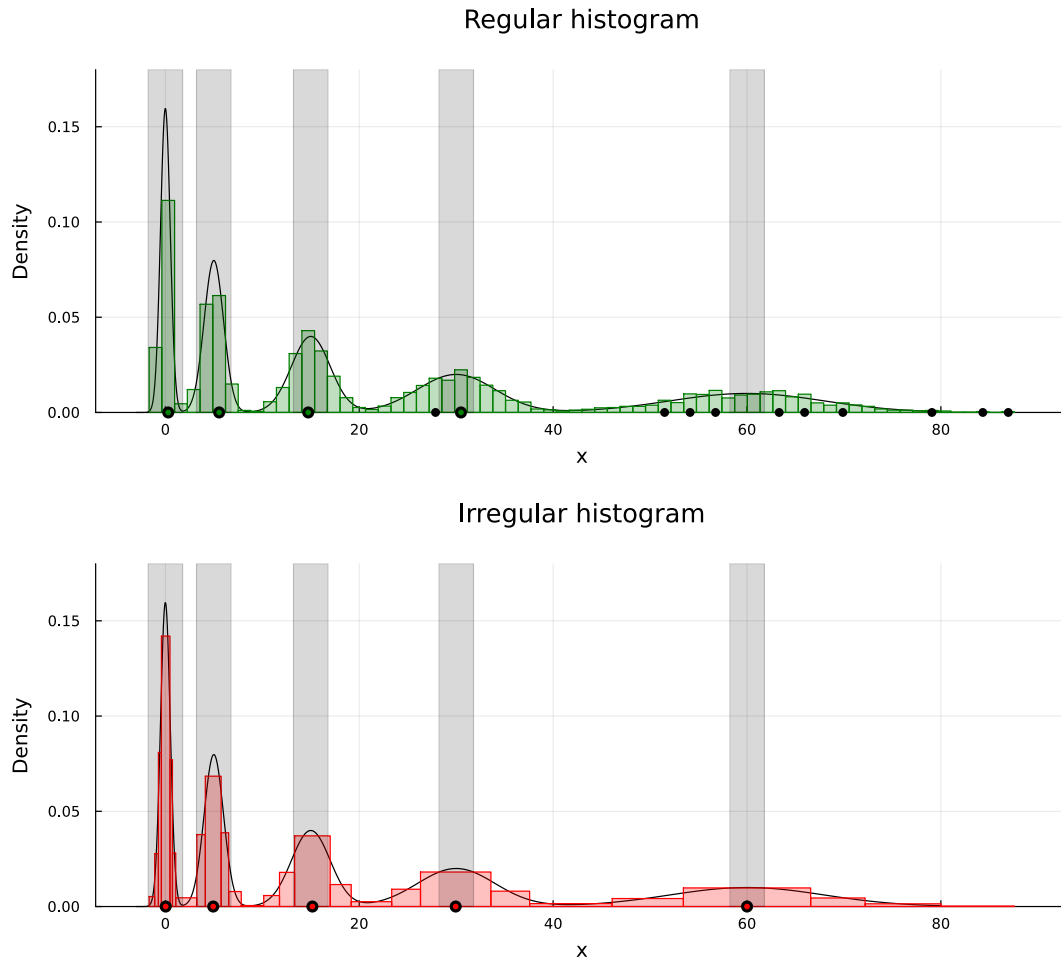


Figure 5.7: Regular and irregular random histogram models fit to an $n = 5000$ sample from the Harp density.

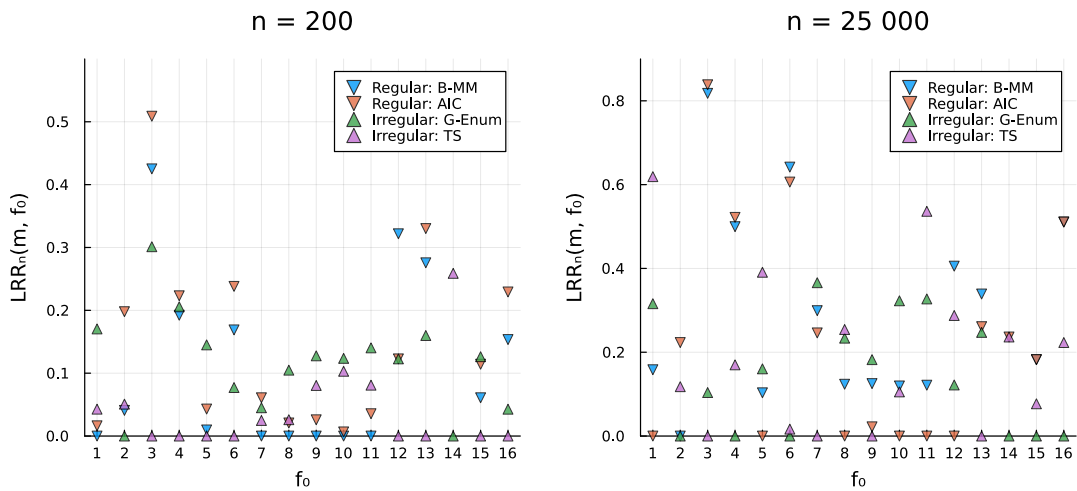


Figure 5.8: Logarithm of Hellinger risk relative to the best method by density.

the spatially inhomogeneous densities #13 and #14, as well as for the infinitely peaked Beta(0.5, 0.5) (#6). In cases where the benefits of a locally varying bin width are less immediate, the regular histograms are often the best performers. This is especially true of classical densities such as those of the standard Normal and Laplace distributions. In this case, the benefit of using an adaptive mesh is outweighed by the need to search for a good partition, resulting in higher overall risks. The Taut String method performs particularly well for $n = 200$, achieving the smallest risk for half of the test densities, and it never produces a Hellinger risk greater than 1.1 times that of the best regular histogram method. For the larger sample size, the Hellinger risk favors the AIC by a fair margin for some densities, although the risk of the G-Enum method is never greater than 1.5 times that of the AIC.

The comparisons undertaken here indicate that the performance of regular histograms relative to irregular histograms in terms of Hellinger risk depends heavily on the true density. Although regular histograms perform better with respect to classical loss functions for some spatially homogeneous densities, this comes at the cost of poor performance with respect to mode identification, a task that the irregular procedures excel at. Overall, we find the arguments in favor of using an irregular histogram procedure over a regular one convincing in light of our findings here.

Chapter 6

Linear Spline Priors

6.1 Beyond histogram models

Although the histogram estimators studied in the two preceding sections provided a flexible framework for nonparametric density estimation, the discontinuity of the estimator may be somewhat undesirable. One reason is that the statistician may prefer a continuous estimate for visual reasons. The discontinuity of the histogram estimator can also be shown to be suboptimal from an asymptotic point of view in certain settings. Indeed, if the true density f_0 is assumed to satisfy additional smoothness properties beyond Hölder continuity then the convergence rate of the histogram estimator is provably suboptimal. The goal of this section is to develop a continuous estimator using similar ideas to those used for the random histogram.

One of the simplest ways of approximating a function in a continuous fashion is to use *piecewise linear interpolation*. Given a probability density $f: [0, 1] \rightarrow \mathbb{R}$ and a set of interpolation points $\{\xi_j: 0 \leq j \leq k\}$ satisfying $\xi_0 = 0$, $\xi_k = 1$ and $\xi_{j-1} < \xi_j$, we form the piecewise linear approximation $\tilde{f}_k$ of f by requiring $\tilde{f}_k(\xi_j) = f(\xi_j)$ for all j , and setting $\tilde{f}_k$ to be a linear function between the interpolation points. We refer to $\tilde{f}_k$ as the *linear interpolating spline* of f with *knots* at $\xi_0, \dots, \xi_k$. To ensure that the piecewise linear model for f is continuous, we make use of a convenient parameterization which we now describe. We shall particularly focus on the case where the knots ξ_j form a regular grid, so that $\xi_j = j/k$ for all j . Let us define the corresponding intervals $I_1 = [\xi_0, \xi_1]$ and $I_j = (\xi_{j-1}, \xi_j]$ for $j = 2, \dots, k$. One quickly verifies that one can write down an expression for the linear interpolating spline of the following form:

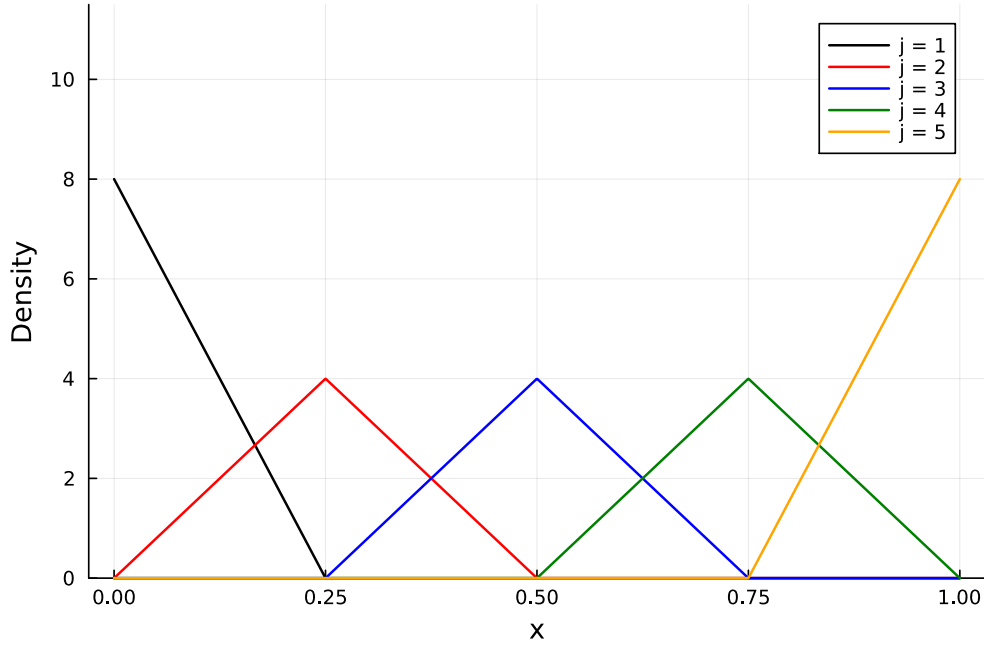
$$\tilde{f}_k(x) = \frac{1}{2k} f(0) \omega(x|k, 0) + \frac{1}{k} \sum_{j=1}^{k-1} f(j/k) \omega(x|k, j) + \frac{1}{2k} f(1) \omega(x|k, k) \quad (6.1)$$

where we introduced the basis functions $\{\omega(\cdot|k, j)\}_{j=0}^k$, defined by

$$\begin{aligned} \omega(x|k, j) &= k(1 - |kx - j|)_+, \quad \text{for } j = 1, 2, \dots, k-1 \\ \omega(x|k, j) &= 2k(1 - |kx - j|)_+, \quad \text{for } j = 0, k, \end{aligned} \quad (6.2)$$

with $(x)_+ = \max\{x, 0\}$. The factor additional factor 2 appearing in the definition of $\omega(\cdot|k, 0)$ and $\omega(\cdot|k, k)$ ensures that $\omega(\cdot|k, j)$ is a probability density for all k, j , since $\omega(\cdot|k, j) \geq 0$ and $\int_0^1 \omega(x|k, j) dx = 1$. The basis functions for $k = 4$ are shown in Figure 6.1.

We note that the supports of the basis functions are overlapping, with $\omega(\cdot|k, j-1)$ and $\omega(\cdot|k, j)$ both being nonzero on I_j . This is unlike the piecewise constant model

Figure 6.1: Basis functions $\{\omega(\cdot | k, j)\}_j$ for $k = 4$.

for f_k used in the regular random histogram model (4.8), where the corresponding basis functions, $k\mathcal{I}(x \in I_{k,j})$, do not have overlapping supports. It is not hard to see that (6.1) leads to a piecewise linear function which interpolates f at $\{j/k: j = 0, \dots, k\}$; only $\omega(\cdot | k, j)$ is nonzero at j/k and each basis functions is linear on each interval I_j .

Although (6.1) offers a reasonable approximation of the density f it will not in general be a probability density function, due to the fact that it may fail to integrate to unity. While it is certainly possible to work with density estimates that are not proper densities, and there are techniques for turning such estimates into densities (see for instance Glad et al. (2003)), we prefer working with an estimator which produces proper probability density functions to begin with. As such, we will introduce an integrability constraint of the parameters of the model. To create a piecewise linear model for f that is a proper probability density, we use a parameterization in terms of a vector $\theta \in \mathcal{S}_k$. Following (6.1), a reasonable model for f is the piecewise linear density

$$f_k(x) = f(x | k, \theta) = \sum_{j=0}^k \theta_j \omega(x | k, j), \quad x \in [0, 1], \quad (6.3)$$

with the θ_j being unknown parameters to be estimated from a random sample $\mathbf{x}_{1:n}$. For this particular parameterization it is easily verified that f_k is a probability density. Since $\omega(\cdot | k, j) \geq 0$ and $\theta_j \geq 0$ it is clear that $f_k \geq 0$. Moreover, the integral of f_k over its domain can be computed as

$$\int_0^1 f_k(x) dx = \int_0^1 \sum_{j=0}^k \theta_j \omega(x | k, j) dx = \sum_{j=0}^k \theta_j \int_0^1 \omega(x | k, j) dx = \sum_{j=0}^k \theta_j = 1,$$

where we used that $\int_0^1 \omega(x | k, j) dx = 1$ for all j . Since each $\omega(\cdot | k, j)$ is a probability density function, the model (6.3) can be seen as a mixture density, with θ_j being the probability that a sample $x \sim f_k$ is generated according to component j in the mixture.

This fact will be used in Section 6.3, where we discuss maximum likelihood estimation of the parameter vector $\boldsymbol{\theta}$. Like the regular random histogram procedure of Chapter 4, the procedure described here crucially depends on the choice of the smoothing parameter k , which will be a major focus in the two coming sections. We consider both approaches based on penalized likelihood and a Bayesian approach, with a prior distribution on k and Dirichlet priors for $\boldsymbol{\theta} | k$.

The use of splines in density estimation is not novel, and we refer the reader to Kirkby et al. (2023) for a comprehensive review of such proposals. Edwards et al. (2019) consider Bayesian estimation of spectral densities using a spline-based prior with a variable number of knots, although they mainly focus on cubic splines. Some Bayesian approaches to density estimation based on splines include the log-spline models of Ghosal et al. (2000) and the polygonally smoothed model of Scricciolo (2007). In the former model, the selection of the number $k+1$ of spline terms is not automatic, and it therefore differs from our approach. The polygonally smoothed model of Scricciolo (2007) is much closer to ours, as it results in a piecewise linear and continuous density, with a prior on the number of spline terms used. The main difference to our spline model is that our estimator is linear on the entire support while Scricciolo's estimator is piecewise constant near the edges. As we shall see in Section 6.5, our choice leads to an improvement in the obtained posterior convergence rate. Notable frequentist proposals that yield continuous, piecewise linear density estimators are the frequency polygon (Scott, 1985) and the edge frequency polygon (Jones et al., 1998). The frequency polygon approach differs from ours as the estimators are computed from binned data; the resulting density estimates are formed by computing linear combinations of the bin heights in a histogram.

6.2 Bayesian inference for the linear spline model

Similarly to the regular histograms of Chapter 4, our regular spline model can be represented via a hierarchical scheme,

$$\begin{aligned} k &\sim p_n(k), \\ \boldsymbol{\theta} | k &\sim \text{Dir}(a\boldsymbol{\theta}_0), \\ f(x | k, \boldsymbol{\theta}) &= \sum_{j=0}^k \theta_j \omega(x | k, j), \\ x_i | k, \boldsymbol{\theta} &\sim f(x_i | k, \boldsymbol{\theta}). \end{aligned} \tag{6.4}$$

The structure of the model and the prior distribution is very similar to that of (4.8), with a Dirichlet prior on the mixture probabilities and a prior on the smoothness parameter k . The only difference is the likelihood function, which is piecewise linear and continuous instead of piecewise constant. The form of the likelihood function in the linear spline model makes posterior computation less straightforward than in the regular histogram model, as the supports of the basis functions $\{\omega(\cdot | k, j)\}$ overlap in this case. Borrowing terminology from the mixture model in Example 3.2, this means that the class labels are unknown, and the model is no longer conjugate for $\boldsymbol{\theta}$ given k . This fact makes it difficult to compute the marginal likelihood and hence the posterior distribution for k without resorting to approximate techniques. The fact that k is random in our model also complicates matters further. If k had been a fixed design parameter set by the user, then the Gibbs sampler of Example 3.2 could be applied to yield samples from the posterior density. In principle, it is possible to estimate the posterior for k by running separate MCMC chains for different values of k , as efficient methods for estimating the

marginal likelihood based on posterior MCMC samples exist (Chib, 1995). However, this approach is computationally costly if the maximal value for k , k_n , is taken to be of even moderate size. As a solution to these problems, we develop a Metropolis–Hastings algorithm that allows us to generate approximate samples from the posterior distribution for variable k .

6.2.1 Metropolis–Hastings for posterior inference

The basis of our proposed MH algorithm uses ideas from Gibbs sampling by introducing auxiliary random variables $\mathbf{u}_{1:n} \in \mathbb{R}^n$ that allow us to identify the class label of each sample once the values of the auxiliary variables are known. This in turn allows us to integrate out $\boldsymbol{\theta}$ and further enables the use of an MH algorithm on k and $\mathbf{u}_{1:n}$ where the dimension stays fixed throughout the iterations.

Suppose that $x \in I_{k,j}$. We start by considering the augmented density $f(x, u | k, \boldsymbol{\theta})$ given by

$$k \sum_{j=1}^k \left\{ \gamma_j \theta_j \mathcal{I}(0 \leq u \leq \rho(x | k, j)) + \gamma_{j-1} \theta_{j-1} \mathcal{I}(\rho(x | k, j) < u \leq 1) \right\} \mathcal{I}(x \in I_{k,j}), \quad (6.5)$$

where $\rho(x | k, j) = k^{-1} \gamma_j^{-1} \omega(x | k, j)$, $\gamma_j = 2$ for $j = 0, k$ and $\gamma_j = 1$ otherwise. Integrating out u , we see that this implies that x has marginal distribution

$$\begin{aligned} f(x | k, \boldsymbol{\theta}) &= \int_0^1 f(x, u | k, \boldsymbol{\theta}) du \\ &= k \sum_{j=1}^k \left\{ \gamma_j \theta_j \rho(x | k, j) + \gamma_{j-1} \theta_{j-1} (1 - \rho(x | k, j)) \right\} \mathcal{I}(x \in I_{k,j}) \\ &= 2\theta_0 k (1 - \rho(x | k, 1)) \mathcal{I}(x \in I_{k,1}) + 2\theta_k k \rho(x | k, k) \mathcal{I}(x \in I_{k,k}) \\ &\quad + k \sum_{j=2}^{k-1} \theta_j \left\{ \rho(x | k, j) \mathcal{I}(x \in I_{k,j}) + (1 - \rho(x | k, j+1)) \mathcal{I}(x \in I_{k,j+1}) \right\}. \end{aligned}$$

From the definition of (6.2), it follows that for $x \in I_{k,j}$ we have $\rho(x | k, j) + \rho(x | k, j-1) = 1$ for all $j = 1, 2, \dots, k$. Hence, the above expression reduces to

$$\begin{aligned} f(x | k, \boldsymbol{\theta}) &= 2k\theta_0 \rho(x | k, 0) \mathcal{I}(x \in I_{k,1}) + 2k\theta_k \rho(x | k, k) \mathcal{I}(x \in I_{k,k}) \\ &\quad + k \sum_{j=1}^{k-1} \theta_j \rho(x | k, j) \mathcal{I}(x \in I_{k,j} \cup I_{k,j+1}) \\ &= \sum_{j=0}^k \theta_j \omega_j(x | k, j). \end{aligned}$$

Since the augmented density (6.5) implies the same model for f , we can work with this density instead when setting up our MCMC algorithm. To derive an expression for the likelihood function of the sample $\mathbf{x}_{1:n}$ it is helpful to rewrite (6.5) slightly. Let $C_{k,j}(x, u) = \mathcal{I}(0 \leq u \leq \rho(x | k, j))$, which implies that $1 - C_{k,j}(x, u) = \mathcal{I}(\rho(x | k, j) < u \leq 1)$. From (6.5) we observe that if $x \in I_{k,j}$ and $C_{k,j}(x, u) = 1$, then the value of the density is equal to $\gamma_j k \theta_j$. On the other hand, if $x \in I_{k,j}$ and $1 - C_{k,j}(x, u) = 1$ then

$f(x, u | k, \boldsymbol{\theta}) = \gamma_j k \theta_{j-1}$. This means that we can rewrite the augmented density as

$$\begin{aligned} f(x, u | k, \boldsymbol{\theta}) &= k \prod_{j=1}^k \left\{ (\gamma_{j-1} \theta_{j-1})^{1-C_{k,j}(x,u)} (\gamma_j \theta_j)^{C_{k,j}(x,u)} \right\}^{\mathcal{I}(x \in I_{k,j})} \\ &= k \prod_{j=0}^k (\gamma_j \theta_j)^{\mathcal{I}(x \in I_{k,j}) C_{k,j}(x,u) + \mathcal{I}(x \in I_{k,j+1}) (1 - C_{k,j}(x,u))}, \end{aligned} \quad (6.6)$$

where $C_{k,0} = C_{k,k+1} = 0$. To write the likelihood function of the sample in a convenient form, define $\mathbf{N}_k(\mathbf{u}_{1:n}) \in \mathbb{R}^{k+1}$ by

$$N_{k,j}(\mathbf{u}_{1:n}) = \sum_{i=1}^n \left\{ \mathcal{I}(x_i \in I_{k,j}) C_{k,j}(x_i, u_i) + \mathcal{I}(x_i \in I_{k,j+1}) (1 - C_{k,j+1}(x_i, u_i)) \right\}. \quad (6.7)$$

From (6.6), we can then write the augmented likelihood as

$$f(\mathbf{x}_{1:n}, \mathbf{u}_{1:n} | k, \boldsymbol{\theta}) = k^n \prod_{j=0}^k (\gamma_j \theta_j)^{N_{k,j}(\mathbf{u}_{1:n})}.$$

From this form of the augmented likelihood, we derive the marginal posterior distribution of $k, \mathbf{u}_{1:n}$ via a conjugacy argument. To this end, note that the full conditional distribution of $\boldsymbol{\theta}$ is given by

$$\begin{aligned} p_n(\boldsymbol{\theta} | \mathbf{x}_{1:n}, k, \mathbf{u}_{1:n}) &\propto p_n(k) p(\boldsymbol{\theta} | k) f(\mathbf{x}_{1:n}, \mathbf{u}_{1:n} | k, \boldsymbol{\theta}) \\ &\propto \prod_{j=0}^k \theta_j^{a\theta_{0,j}-1} \prod_{j=0}^k \theta_j^{N_{k,j}(\mathbf{u}_{1:n})} \\ &\propto \text{Dir}(\boldsymbol{\theta} | a\boldsymbol{\theta}_0 + \mathbf{N}_k(\mathbf{u}_{1:n})) \end{aligned}$$

Arguing as in (4.17), we derive that the joint posterior of $k, \mathbf{u}_{1:n}$ becomes

$$\begin{aligned} p(k, \mathbf{u}_{1:n} | \mathbf{x}_{1:n}) &\propto p_n(k) \frac{p(\boldsymbol{\theta} | k) f(\mathbf{x}_{1:n}, \mathbf{u}_{1:n} | k, \boldsymbol{\theta})}{p(\boldsymbol{\theta} | \mathbf{x}_{1:n}, k, \mathbf{u}_{1:n})} \\ &\propto p_n(k) k^n \frac{\prod_{j=0}^k \Gamma(a\theta_{0,j} + N_{k,j}(\mathbf{u}_{1:n}))}{\prod_{j=0}^k \Gamma(a\theta_{0,j})} \frac{\Gamma(a)}{\Gamma(a+n)}. \end{aligned} \quad (6.8)$$

The posterior distribution in (6.8) is of a similar form as the posterior for k in the histogram model (4.18), the difference being that the former expression relies on auxiliary random variables to determine the class labels for each observation. Note that if we change the value of k and keep $\mathbf{u}_{1:n}$ fixed, the value of the joint posterior will change, as the implied class labels from $\mathbf{u}_{1:n}$ will be different in this case.

Our proposed MCMC algorithm alternates between sampling $(k, \mathbf{u}_{1:n})$ via a Metropolis–Hastings step and sampling $\mathbf{u}_{1:n} | k$ via a Gibbs step. Specifically, the MH step is based on proposing k^* uniformly from $\{k-3, k-2, k-1, k+1, k+2, k+3\}$ with adjustments made at the edges, i.e. we don't propose $k \leq 0$ or $k > k_n$ as these are not in the support of the prior. Based on k^* , we propose the new value $\mathbf{u}_{1:n}^*$ via a Gibbs-like proposal density. This is accomplished by first uniformly drawing a random permutation $i_1, i_2, \dots, i_n$ of $\{1, 2, \dots, n\}$, and initializing $\mathbf{u}_{1:n}^* = \mathbf{u}_{1:n}$. The components of $\mathbf{u}_{1:n}^*$ are then updated according to a Gibbs step in the order prescribed by the permutation by sampling from $p(u_{i_l}^* | \mathbf{x}_{1:n}, k, \mathbf{u}_{-i_l}^*)$ for $l = 1, 2, \dots, n$, i.e. we update i_1 first, then

i_2 and so on. The resulting proposal is then either accepted or rejected based on the Metropolis–Hasting ratio (3.10),

$$A((k^*, \mathbf{u}_{1:n}^*) | (k, \mathbf{u}_{1:n})) = \frac{p(k^*, \mathbf{u}_{1:n}^* | \mathbf{x}_{1:n}) q_k(k | k^*) q_{\mathbf{u}}(\mathbf{u}_{1:n} | k, \mathbf{u}_{1:n}^*)}{p(k, \mathbf{u}_{1:n} | \mathbf{x}_{1:n}) q_k(k^* | k) q_{\mathbf{u}}(\mathbf{u}_{1:n}^* | k^*, \mathbf{u}_{1:n})}, \quad (6.9)$$

where q_k and $q_{\mathbf{u}}$ denote the proposal distributions for k and $\mathbf{u}$ given k , respectively. We note that the probability from drawing the permutation used in the Gibbs step cancels in the MH ratio (6.9), since drawing the reverse permutation has the same probability.

In order to perform the Gibbs step, we need the conditional distributions $p(u_i | \mathbf{x}_{1:n}, k, \mathbf{u}_{-i})$, which we now derive. Suppose that $x \in I_{k,j}$. Let $\mathbf{N}_k(\mathbf{u}_{-l})$ denote the version of $\mathbf{N}_k(\mathbf{u}_{1:n})$ with sample l removed, i.e. where the sum in (6.7) is taken over all $i \neq l$. Then we have the following relationship:

$$N_{k,j}(\mathbf{u}_{1:n}) = \begin{cases} N_{k,j}(\mathbf{u}_{-i}) + 1 & \text{if } u_i \in [0, p_{k,j}(x_i)] \\ N_{k,j}(\mathbf{u}_{-i}) & \text{if } u_i \in (p_{k,j}(x_i), 1]. \end{cases}$$

On the other hand,

$$N_{k,j-1}(\mathbf{u}_{1:n}) = \begin{cases} N_{k,j}(\mathbf{u}_{-i}) & \text{if } u_i \in [0, p_{k,j}(x_i)] \\ N_{k,j}(\mathbf{u}_{-i}) + 1 & \text{if } u_i \in (p_{k,j}(x_i), 1]. \end{cases}$$

Using the fact that $\Gamma(z+1) = z\Gamma(z)$ for all $z > 0$, we derive from (6.8) that

$$\begin{aligned} & p_n(u_i | \mathbf{x}_{1:n}, k, \mathbf{u}_{-i}) \\ & \propto \gamma_j \mathcal{I}(0 \leq u_i \leq \rho(x_i | k, j)) \prod_{m=1}^k \Gamma(a\theta_{0,m} + N_{k,m}(\mathbf{u}_{1:n})) \\ & \quad + \gamma_{j-1} \mathcal{I}(\rho(x_i | k, j) < u_i \leq 1) \prod_{m=1}^k \Gamma(a\theta_{0,m} + N_{k,m}(\mathbf{u}_{1:n})) \\ & \propto \gamma_j \mathcal{I}(0 \leq u_i \leq \rho(x_i | k, j)) \Gamma(a\theta_{0,j} + N_{k,j}(\mathbf{u}_{-i}) + 1) \Gamma(a\theta_{0,j} + N_{k,j-1}(\mathbf{u}_{-i})) \\ & \quad + \gamma_{j-1} \mathcal{I}(\rho(x_i | k, j) < u_i \leq 1) \Gamma(a\theta_{0,j} + N_{k,j-1}(\mathbf{u}_{-i}) + 1) \Gamma(a\theta_{0,j} + N_{k,j}(\mathbf{u}_{-i})) \\ & \propto \gamma_j \mathcal{I}(0 \leq u_i \leq \rho(x_i | k, j)) \{a\theta_{0,j} + N_{k,j}(\mathbf{u}_{-i})\} \\ & \quad + \gamma_{j-1} \mathcal{I}(\rho(x_i | k, j) < u_i \leq 1) \{a\theta_{0,j} + N_{k,j-1}(\mathbf{u}_{-i})\}, \end{aligned}$$

which is a mixture of two uniform distributions. For our MCMC algorithm we thus use $q_{u_i}(u_i | k, \mathbf{u}_{-i}) = p(u_i | \mathbf{x}_{1:n}, k, \mathbf{u}_{-i})$ as proposal distribution when sampling $\mathbf{u}_{1:n}^*$. Our MCMC algorithm may then be summarized as follows

- 1) Sample $\mathbf{u}_{1:n}$ using the Gibbs step outlined in steps 3) and 4).
- 2) Sample $k^* \sim q_k(k^* | k)$
- 3) Draw a random permutation $i_1, \dots, i_n$ of $\{1, \dots, n\}$ and set $\mathbf{u}_{1:n}^* = \mathbf{u}_{1:n}$.
- 4) For $l = 1, 2, \dots, n$
 - a) Sample $u_{i_l}^* \sim q_u(u_{i_l}^* | k, \mathbf{u}_{-i_l}^*)$

- 5) Accept $(k^*, \mathbf{u}_{1:n}^*)$ with probability $\min\{1, A((k^*, \mathbf{u}_{1:n}^*) | (k, \mathbf{u}_{1:n}))\}$, else keep the old values $(k, \mathbf{u}_{1:n})$.

This process is repeated until the desired number of samples has been generated.

Having generated samples $\{(k^{(t)}, \mathbf{u}_{1:n}^{(t)})\}_{t=1}^T$, we can estimate the posterior mean of f by

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\boldsymbol{\theta}} \{f(x | k^{(t)}, \mathbf{u}_{1:n}^{(t)}, \boldsymbol{\theta}) | \mathbf{x}_{1:n}\} = \frac{1}{T} \sum_{t=1}^T \sum_{j=0}^k \mathbb{E}_{\theta_j} \{\theta_j | k^{(t)}, \mathbf{u}_{1:n}^{(t)}, \mathbf{x}_{1:n}\} \omega(x | k^{(t)}, j). \quad (6.10)$$

This is a natural Monte Carlo estimator of the posterior expectation since $k^{(t)}, \mathbf{u}_{1:n}^{(t)}$ are approximate samples from the posterior distribution and

$$\mathbb{E}_f \{f(x) | \mathbf{x}_{1:n}\} = \mathbb{E}_{k, \mathbf{u}_{1:n}} \mathbb{E}_{\boldsymbol{\theta}} \{f(x | k, \mathbf{u}_{1:n}, \boldsymbol{\theta}) | k, \mathbf{u}_{1:n}, \mathbf{x}_{1:n}\},$$

which the sample average in (6.10) approximates. Moreover, the estimate (6.10) avoids the need to simulate $\boldsymbol{\theta} | k^{(t)}, \mathbf{u}_{1:n}^{(t)}$ which would introduce additional variance in the estimate. The conditional expectation in (6.10) is tractable, since $\boldsymbol{\theta}$ has a Dirichlet distribution conditional on $k, \mathbf{u}_{1:n}$.

Posterior uncertainty measures are easily obtained by simulation, since

$$\boldsymbol{\theta} | k^{(t)}, \mathbf{u}_{1:n}^{(t)} \sim \text{Dir}(a\boldsymbol{\theta}_0 + \mathbf{N}_{k^{(t)}}(\mathbf{u}_{1:n}^{(t)})).$$

Thus, by drawing samples $\boldsymbol{\theta}^{(t)}$ from the Dirichlet distribution in the above display, the tuple $(k^{(t)}, \boldsymbol{\theta}^{(t)})$ defines a sample $f^{(t)}$ from the posterior of f . This can be used to construct credibility intervals for $f(x)$ at a given point or other posterior summaries of interest.

We now make some comments about design choices we made when developing our proposed MCMC algorithm. Since it is in principle easy to sample from the conditional distributions of (6.8), one might consider running a Gibbs sampler to draw random variates from the joint posterior. We found that in practice, the chains resulting from running such a Gibbs sampler exhibited poor mixing properties for k , since the value of k would not change at all or stay the same for many iterations. After running some numerical experiments with our MH sampler, we found out the the posterior distribution for k occasionally produced modes that were relatively far apart, with little probability mass in between the two modes. In this case, the output from different chains differed significantly, as moving between the two modes became difficult. The issue of moving between distant modes could possibly be solved by using a technique such as simulated tempering (Marinari and Parisi, 1992), which can allow for more rapid moves throughout the parameter space, although we do not explore this approach further here.

6.2.2 Validating the Metropolis–Hastings output

As an illustration of the proposed method, we consider a simulated example where $n = 1000$ samples have been drawn from the symmetric normal mixture of Figure 1.1. We transformed the obtained sample to $[0, 1]$ and ran the Metropolis-Hastings sampler of Section 6.2.1 on the transformed sample to estimate the density. We ran 4 separate chains of length $T = 11\,000$ with a burn-in period of $D = 1000$. As our prior distribution for $\boldsymbol{\theta} | k$ we took the Dirichlet concentration parameter to be $a = 1.0$, with a prior mean centered on the uniform distribution on the unit interval, $\boldsymbol{\theta}_0 = (0.5k^{-1}, k^{-1}, \dots, k^{-1}, 0.5k^{-1})$. As our prior for k we used a uniform distribution on $\{1, 2, \dots, 15\}$, and used $k = 1, 5, 10$ and 15 as starting points for the four chains. The resulting density estimates, together with the true density of the data for the 4 chains, are shown separately in Figure 6.2, together with the true density.

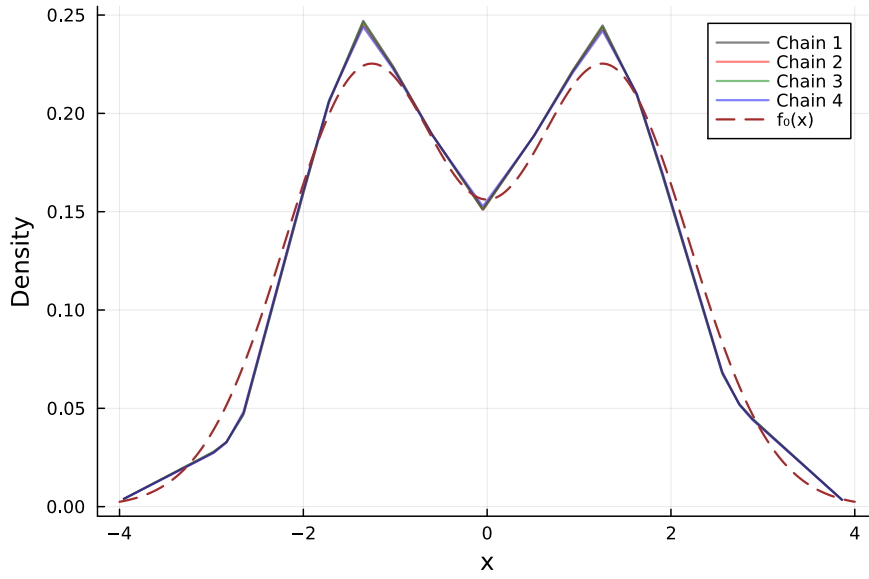


Figure 6.2: True and estimated density from four different chains in the simulated example

The density estimates from the different chains are visually indistinguishable, suggesting that the burn-in period was sufficient. The resulting density estimates track the shape of the true density well. In addition, the posterior mean was able to correctly identify the location and the number of modes of the density.

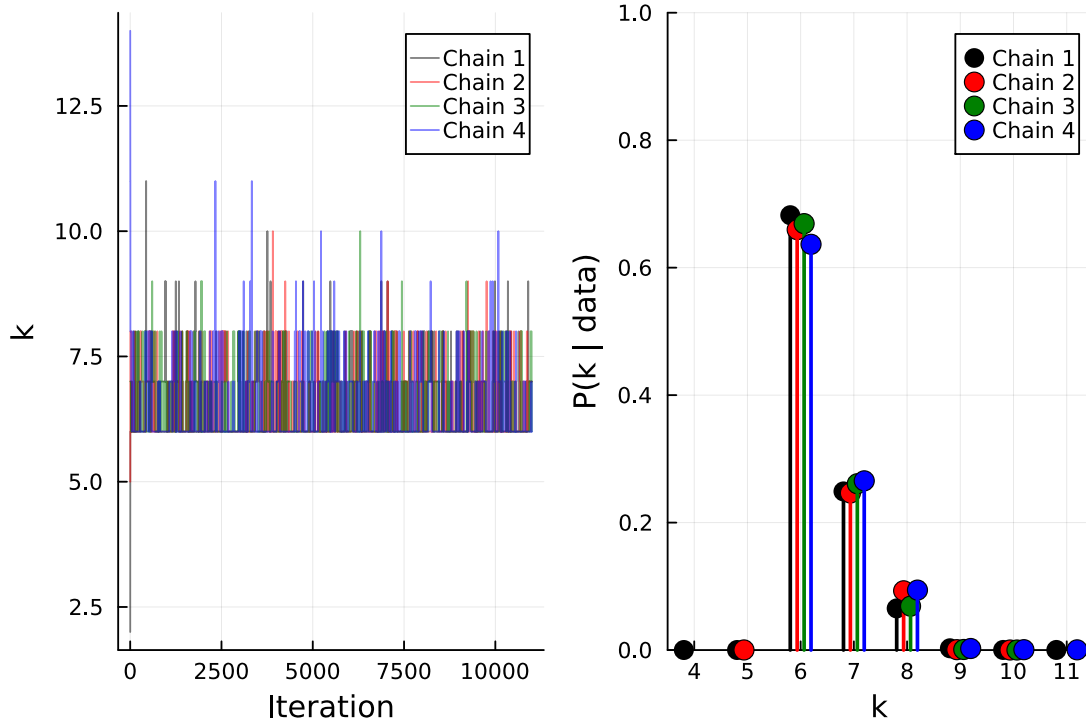
To further investigate the mixing properties of the chain, we inspected the trace plots for the smoothing parameter k for each of the chains. Additionally, we estimated posterior probabilities separately for each of the four chains, via the empirical probabilities,

$$\hat{p}_n(k | \mathbf{x}_{1:n}) = \frac{1}{T} \sum_{t=1}^T \mathcal{I}(k^{(t)} = k), \quad \text{for } k = 1, 2, \dots, k_n. \quad (6.11)$$

The trace plots and the estimated posterior for k are shown in Figure 6.3. All four chains exhibit good mixing for k , as they all rapidly explore the part of the parameter space that has a high posterior probability. This suggests that the burn-in period of $D = 1000$ was more than adequate. All four chains yielded similar estimates of the posterior probabilities $\hat{p}_n(k | \mathbf{x}_{1:n})$. As a final measure of the quality of the chains, we computed the sample autocorrelation function of $\{k^{(t)}\}$ for each of the four chains. The results are shown in Figure D.1. The autocorrelation decays rapidly to 0 in each of the four chains, indicating good mixing. We conclude that the chains all show rapid mixing and yield very similar density estimates for this example.

6.3 Penalized Maximum Likelihood Estimation

The parameters of the $k+1$ -knot linear spline density (6.3) can also be estimated using a maximum likelihood approach. Since the complexity of the model is increasing with the number k of intervals used to construct the estimator, we consider an approach based on penalized maximum likelihood estimation in order to determine the value of smoothing parameter k . Instead of using a single value of k as chosen by AIC or BIC, we shall

Figure 6.3: Trace plots and estimated posterior probability for k for each chain.

focus on model-averaged density estimators of the type

$$\hat{f}_n(x) = \sum_{k=1}^{k_n} w_k f(x | k, \hat{\theta}), \quad (6.12)$$

where the w_k are as in (3.22) and $\hat{\theta} | k$ is the maximum likelihood estimator for θ given k . Since the expression in (6.12) relies on the maximized log-likelihood and the MLE for given k , the computation of the estimator essentially reduces to solving k_n different maximum likelihood problems, and thus a method of computing the (approximate) MLE is needed.

If $x_i | k, \theta \sim f$, then the likelihood function becomes

$$L_n(\theta | k) = \prod_{i=1}^n \sum_{j=0}^k \theta_j \omega(x_i | k, j). \quad (6.13)$$

If $x_i \in I_{k,j}$ then the form of the likelihood can be simplified somewhat by noting that of all the basis functions, only $\omega(\cdot | k, j-1)$ and $\omega(\cdot | k, j)$ are supported on this interval. Thus, the likelihood contribution of x_i becomes $\theta_{j-1} \omega(x_i | k, j-1) + \theta_j \omega(x_i | k, j)$.

Since the model (6.3) is a special case of a mixture model, the EM algorithm can be used to compute a numerical approximation to the maximum likelihood estimator as in Example 2.2. Recall that the updates for the parameter vector θ are given by

$$\theta_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \frac{\theta_j^{(t)} \omega(x_i | k, j)}{\sum_{j'=1}^k \theta_{j'}^{(t)} \omega(x_i | k, j')}, \quad \text{for } j = 1, 2, \dots, k. \quad (6.14)$$

The updates in (6.14) can be particularly efficiently computed in this case, since there are only two mixture components contributing to the likelihood of the sample x_i .

As with any numerical optimization routine, a suitable stopping criterion should be employed to determine when the algorithm has reached a point close to the optimum. Since we are trying to estimate a density, it seems appropriate to using a stopping criterion based on a norm on the density function induced by $\boldsymbol{\theta}^{(t)}$. Note that we have the inequality $d_H^2(f(\cdot | k, \boldsymbol{\theta}^{(t+1)}), f(\cdot | k, \boldsymbol{\theta}^{(t)})) \leq d_H^2(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)})$ from the proof of Lemma 6.3, so controlling the Hellinger distance between subsequent parameter vectors provides an inexpensive way of ensuring that the Hellinger distance of the iterates of densities has converged.

6.4 Big support and posterior consistency

In this section we show that the linear spline prior has big support and yields a consistent posterior under rather mild assumptions on the prior distribution.

6.4.1 Big support

The first post on the program of the section is to show that the prior distribution possesses the Kullback–Leibler property. As before, for a given sample size n we denote the prior distribution for f by $P_n(\cdot)$. Since the prior for f is given by the hierarchical scheme of (6.4), we have for a given measurable set $\mathcal{A}$ that

$$P_n(f \in \mathcal{A}) = \sum_{k=1}^{\infty} p_n(k) P(f_k \in \mathcal{A} | k), \quad (6.15)$$

where $P(f_k \in \mathcal{A} | k) = \int_{\mathcal{S}_k} \mathcal{I}(f(\cdot | k, \boldsymbol{\theta}) \in \mathcal{A}) p(\boldsymbol{\theta} | k) d\boldsymbol{\theta}$. To show that the prior indeed has the KL property, we need some mild assumptions on the prior:

Assumption 6.1. The sequence of prior distributions $\{P_n\}$ satisfies the following:

- (i) The prior for k satisfies $\liminf_{n \rightarrow \infty} p_n(k) > 0$ for any $k \in \mathbb{N}$.
- (ii) The prior for $\boldsymbol{\theta} | k$ is $\text{Dir}(\mathbf{a}_k)$, where $a_{k,j} > 0$ for all k, j .

These assumptions are more or less the same as those used to establish that the regular histogram prior has big support. The strategy used to prove consistency of the linear spline prior follows similar steps as those used for the histogram prior.

Let $f_0: [0, 1] \rightarrow \mathbb{R}$ be a bounded probability density. Let $\mathcal{G}$ denote the set of strictly positive, bounded functions on $[0, 1]$. We define the functional $T_{f_0}: \mathcal{G} \rightarrow \mathbb{R}$ by

$$T_{f_0}(g) = \int_0^1 f_0(x) \log g(x) dx. \quad (6.16)$$

Before we discuss the properties of T_{f_0} , we note that for any $g \in \mathcal{G}$ satisfying $m \leq g \leq M$ with $m > 0$, we have $|\log g(x)| \leq \max\{|\log m|, |\log M|\}$ and hence

$$\int_0^1 f_0(x) |\log g(x)| dx \leq \int_0^1 f_0(x) \max\{|\log m|, |\log M|\} dx < \infty,$$

so T_{f_0} is a well-defined operator.

We also make use of the following fact in the sequel: if $g: [0, 1] \rightarrow \mathbb{R}$ is a continuous and strictly positive function then g we can find a positive number m such that $g \geq m$. This is a consequence of the extreme value theorem, which states that g attains its infimum on $[0, 1]$. Hence, for continuous g strict positivity is equivalent to being bounded away from 0.

Lemma 6.1. T_{f_0} is a continuous functional when $\mathcal{G}$ is equipped with the supremum metric.

Proof. Let $g \in \mathcal{G}$ and fix $\epsilon \in (0, 1)$. Then g is lower bounded by, say $m > 0$. Let $h \in \mathcal{G}$ satisfy $d_\infty(g, h) < m\epsilon/2$, from which we deduce that $h \geq m/2$. By the mean value theorem we have that

$$|\log g(x) - \log h(x)| = \frac{1}{\eta(x)} |g(x) - h(x)|,$$

where $\eta(x)$ is a number between $g(x)$ and $h(x)$. Since $g, h \geq m/2$ it follows that $\eta(x) \geq m/2$. We conclude that $|\log g(x) - \log h(x)| \leq 2m^{-1}d_\infty(g, h)$.

$$\begin{aligned} |T_{f_0}(g) - T_{f_0}(h)| &= \left| \int_0^1 f_0(x) \{ \log g(x) - \log h(x) \} dx \right| \\ &\leq \int_0^1 f_0(x) |\log g(x) - \log h(x)| dx \\ &\leq 2m^{-1}d_\infty(g, h) \int_0^1 f_0(x) dx \\ &= 2m^{-1}d_\infty(g, h). \end{aligned}$$

The last display is less than ϵ by assumption, so we conclude that T_{f_0} is continuous at g . $\square$

From Lemma 4.1, we know that any f_0 can be approximated by a continuous density g bounded away from zero in the Kullback–Leibler sense. To show that we can find a linear spline density g_k such that $K(f_0, g_k)$ can be made arbitrarily small, it thus suffices to show that any such g can be approximated by a linear spline density in the supremum metric by the preceding lemma.

Lemma 6.2. For any continuous probability density $f_0: [0, 1] \rightarrow \mathbb{R}$ and any $\epsilon > 0$ we can find a strictly positive linear spline density g_k satisfying $K(f_0, g_k) < \epsilon$.

Proof. We proceed in steps. Our first result is to show that we can find a sequence of strictly positive linear splines converging uniformly to g , where g is continuous, strictly positive and satisfies $K(f_0, g) < \epsilon/2$. The existence of such a g is ensured by Lemma 4.1. To this end, define $\tilde{g}_k: [0, 1] \rightarrow \mathbb{R}$ by

$$\tilde{g}_k(t) = \frac{1}{2k} g(0) \omega(t | k, 0) + \frac{1}{k} \sum_{j=1}^k g(j/k) \omega(t | k, j) + \frac{1}{2k} g(1) \omega(t | k, k). \quad (6.17)$$

We claim that $\tilde{g}_k \rightarrow g$ uniformly. By construction, $\tilde{g}_k$ interpolates g at $\{j/k, j = 0, \dots, k\}$ and is piecewise linear in between these points. Note that for fixed $x \in \mathbb{R}$ and $y \in [a, b]$, the function $z_x(y) = |x - (c_0 + c_1 y)|$ attains its maximum at one of the endpoints of the interval. As such, $|g(t) - \tilde{g}_k(t)|$ is upper bounded by $\max\{|g(t) - \tilde{g}_k((j-1)/k)|, |g(t) - \tilde{g}_k(j/k)|\}$ for $t \in I_{k,j} = ((j-1)/k, j/k]$ by the piecewise linearity of $\tilde{g}_k$. This implies that

$$\sup_{t \in [0, 1]} |g(t) - \tilde{g}_k(t)| \leq \max_j \left\{ \sup_{t \in I_{k,j}} \{|g(t) - g((j-1)/k)| + \sup_{t \in I_{k,j}} |g(t) - g(j/k)|\} \right\}, \quad (6.18)$$

where we used the interpolation property of $\tilde{g}_k$ to obtain the last inequality. Since $[0, 1]$ is compact, g is uniformly continuous and as such since $|I_{k,j}| \rightarrow 0$ uniformly as $k \rightarrow \infty$,

we have $\max_j \sup_{t,s \in I_{k,j}} |g(t) - g(s)| \rightarrow 0$ as well. We conclude $\sup_{t \in [0,1]} |g(t) - \tilde{g}_k(t)| \rightarrow 0$ by squeezing.

To show that convergence also holds for spline densities, let $g_k(t) = \tilde{g}_k(t) / \int_0^1 \tilde{g}_k(s) ds$. Clearly, g_k integrates to 1, and it is strictly positive as a consequence of the strict positivity of $\tilde{g}_k$. Since $\tilde{g}_k$ converges uniformly to g it follows that $\int_0^1 \tilde{g}_k(s) ds \rightarrow 1$. Applying Lemma A.4 to the sequences $h_k = \tilde{g}_k$ and $a_k = \int_0^1 \tilde{g}_k(s) ds$, it follows that g_k converges to g in d_∞ . By the continuity of T_{f_0} at g , it holds that for all sufficiently large k , $|T_{f_0}(g) - T_{f_0}(g_k)| < \epsilon/2$. Finally, we note that

$$K(f_0, g_k) = K(f_0, g) + T_{f_0}(g) - T_{f_0}(g_k) < \epsilon/2 + \epsilon/2 = \epsilon,$$

which concludes the proof. $\square$

Using the result of the above lemma, we can show that the linear interpolation prior possesses the Kullback–Leibler property under conditions mirroring the condition for histograms.

Theorem 6.1. *If Assumption 6.1 holds, then any continuous $f_0: [0, 1] \rightarrow \mathbb{R}$ is in the Kullback–Leibler support of the linear interpolation prior.*

Proof. By Lemma 6.2, we can find a linear spline density g_κ with $\kappa + 1$ knots bounded away from 0 such that $K(f_0, g_\kappa) < \epsilon/2$. We will show that the prior distribution assigns positive probabilities to KL-neighborhoods of this density. Letting $\boldsymbol{\theta} \in \mathcal{S}_k$, we have that for $f_\kappa = \sum_{j=0}^\kappa \theta_j \omega(\cdot | \kappa, j)$ and $g_\kappa = \sum_{j=0}^\kappa \rho_j \omega(\cdot | \kappa, j)$,

$$d_\infty(g_\kappa, f_\kappa) \leq \max_{0 \leq j \leq \kappa} |\rho_j - \theta_j| \sup_{x \in [0,1]} |\omega(x | \kappa, j)| = 2\kappa d_\infty(\boldsymbol{\rho}, \boldsymbol{\theta}) \leq 2\kappa d_1(\boldsymbol{\rho}, \boldsymbol{\theta}), \quad (6.19)$$

since $|\omega(x | \kappa, j)| \leq 2\kappa$ for all j and $x \in [0, 1]$. Observe that $K(f_0, f_\kappa) = K(f_0, g_\kappa) + T_{f_0}(g_\kappa) - T_{f_0}(f_\kappa)$. The continuity of T_{f_0} at g_κ implies that we can find a $\delta > 0$ such that $|T_{f_0}(g_\kappa) - T_{f_0}(f_\kappa)| < \epsilon/2$ whenever $d_\infty(g_\kappa, f_\kappa) < \delta$. Hence, if the prior probability of the event $\{k = \kappa, d_1(\boldsymbol{\rho}, \boldsymbol{\theta}) < \gamma\}$ does not tend to 0 for arbitrary $\gamma > 0$, $k \in \mathbb{N}$ and $\boldsymbol{\rho} \in \mathcal{S}_k$, the prior has the Kullback–Leibler property. This can be verified by arguing as in the proof of Theorem 4.1. $\square$

6.4.2 Posterior consistency

To verify the conditions of Theorem 4.2 we need to estimate the covering numbers of a sieve consisting of linear interpolation densities. As for histograms, this is achieved by relating the bracketing numbers of the model to the bracketing numbers of the simplex. Before we start with proofs we establish some new notation. Let $\mathcal{H}_k$ denote the set of linear spline densities with knots at $\{j/k: 0 \leq j \leq k\}$, that is,

$$\mathcal{H}_k = \left\{ f: [0, 1] \rightarrow \mathbb{R}: f(x) = \sum_{j=0}^k \theta_j \omega(x | k, j) \text{ for some } \boldsymbol{\theta} \in \mathcal{S}_k \right\} \quad (6.20)$$

The following result is an analogue Lemma 4.4 for linear spline densities.

Lemma 6.3. *Let $l(\cdot) = \sum_{j=1}^k a_j \omega(\cdot | k, j)$ and $u(\cdot) = \sum_{j=1}^k b_j \omega(\cdot | k, j)$. Suppose that $f_k = \sum_{j=0}^k \theta_j \omega(\cdot | k, j)$ for some $\boldsymbol{\theta} \in \mathcal{S}_k$. Then if $[\mathbf{a}, \mathbf{b}]$ is an ϵ -bracket containing $\boldsymbol{\theta}$, then $[l, u]$ is an ϵ -bracket containing f_k . Moreover, the covering number of $\mathcal{F}_n = \cup_{k=1}^{k_n} \mathcal{H}_k$ is upper bounded by $k_n(L/\epsilon)^{k_n+1}$ for a universal constant $L > 0$.*

Proof. Let $[\mathbf{a}, \mathbf{b}]$ be an ϵ Hellinger bracket. If $\boldsymbol{\theta} \in [\mathbf{a}, \mathbf{b}]$ then $a_j \leq \theta_j \leq b_j$ for all j , which implies that $f_k \in [l, u]$. Note that as $\omega(\cdot | k, j)$ integrates to one,

$$d_H^2(a_j \omega(\cdot | k, j), b_j \omega(\cdot | k, j)) = \int_0^1 (\sqrt{a_j} - \sqrt{b_j})^2 \omega(t | k, j) dt = (\sqrt{a_j} - \sqrt{b_j})^2.$$

Hence, using Lemma B.1, we obtain

$$d_H^2(l, u) \leq \sum_{j=1}^k d_H^2(a_j \omega(\cdot | k, j), b_j \omega(\cdot | k, j)) \leq \sum_{j=1}^k (\sqrt{a_j} - \sqrt{b_j})^2 < \epsilon^2,$$

so $[l, u]$ is an ϵ Hellinger bracket as claimed. The last part of the claim follows by arguing as in the proof of Lemma 4.5 and Theorem 4.2, since

$$N(\epsilon, \mathcal{F}_n, d_H) \leq \sum_{k=1}^{k_n} N_{[]}(\epsilon, \mathcal{H}_k, d_H) \leq \sum_{k=1}^{k_n} N_{[]}(\epsilon, \mathcal{S}_k, d_H) \leq \sum_{k=1}^{k_n} \frac{(k+1)(2\pi e)^{(k+1)/2}}{\epsilon^k},$$

which was to be shown. $\square$

Having established the Kullback–Leibler property for our prior distribution and an estimate of the covering number, consistency follows under a mild condition on the prior distribution. The proof of Theorem 6.2 is similar to that of Theorem 4.2 and is therefore omitted.

Theorem 6.2. *Suppose that Assumption 6.1 holds. Suppose further that for a sequence $\{k_n\}$ diverging to ∞ satisfying $k_n = o(n)$, we have that*

$$\sum_{k=k_n+1}^{\infty} p_n(k) \leq \exp(-c_1 n),$$

where $c_1 > 0$ is a constant. Then the linear spline prior is consistent at any continuous f_0 .

6.5 Posterior convergence rates

The goal of this section is to derive convergence rates for our linear spline estimator. Our main result shows that for sufficiently smooth functions we are able to obtain a rate equivalent to that achieved by standard kernel-type estimators¹, up to a logarithmic factor. The key to this result is the stronger approximation properties of linear splines compared to piecewise constant functions, allowing us to attain a better approximation rate.

Assumption 6.2. For the regular spline prior sequence $\{P_n\}$, the following holds:

- (i) For given k , the prior distribution of $\boldsymbol{\theta} \in \mathcal{S}_k$ is $\text{Dir}(\mathbf{a}_k)$, and there are two constants $\sigma, \Sigma > 0$ such that $\sigma k^{-1} \leq a_{k,j} \leq \Sigma$ for all $j = 0, 1, \dots, k$.
- (ii) There are constants $d_1, d_2, D_1, D_2 > 0$ such that for all $k \in \mathbb{N}$ we have

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

¹By standard kernel-type estimators we refer to those based on positive, second order kernels.

6.5.1 An approximation result

The main advantage of the linear spline model is that it can exploit additional smoothness beyond mere Hölder continuity, obtaining better approximation rates for α -smooth functions when $\alpha \in (1, 2]$, see Definition 2.10. Our first result concerns the rate of approximation when using a linear spline to approximate an α -smooth function.

Proposition 6.1. *If $h: [0, 1] \rightarrow \mathbb{R}$ is α -smooth for $\alpha \in (0, 2]$, then the linear spline interpolant of h at $\{j/k: 0 \leq j \leq k\}$, denoted h_k , satisfies $d_\infty(h, h_k) \leq \mathcal{O}(k^{-\alpha})$.*

Proof. Fix $k \in \mathbb{N}$ and $x \in [0, 1]$. Write $\xi_j = j/k$ for $j = 0, 1, \dots, k$ and suppose that $x \in I_j = (\xi_{j-1}, \xi_j]$. Since h_k interpolates h at ξ_{j-1} and ξ_j and is linear on I_j , we can write $h_k(x) = h(\xi_{j-1}) + k(x - \xi_{j-1})\{h(\xi_j) - h(\xi_{j-1})\}$ on I_j . We consider the two cases $\alpha \in (0, 1]$ and $\alpha \in (1, 2]$ separately. For $\alpha \in (0, 1]$ we note that

$$\begin{aligned} |h(x) - h_k(x)| &= |h(x) - h(\xi_{j-1}) - k\{h(\xi_j) - h(\xi_{j-1})\}(x - \xi_{j-1})| \\ &\leq |h(x) - h(\xi_{j-1})| + |h(\xi_j) - h(\xi_{j-1})| \\ &\leq L|x - \xi_{j-1}|^\alpha + Lk^{-\alpha} \\ &\leq 2Lk^{-\alpha}, \end{aligned}$$

where we used $|x - \xi_{j-1}| \leq k^{-1}$.

We now turn our attention to $\alpha \in (1, 2]$. By the Mean Value Theorem, we can find $\zeta_1, \zeta_2 \in I_j$ such that $h(\xi_j) - h(\xi_{j-1}) = h'(\zeta_1)k^{-1}$ and $h(x) - h(\xi_{j-1}) = h'(\zeta_2)\{x - \xi_{j-1}\}$. Hence,

$$\begin{aligned} |h(x) - h_k(x)| &= |h(x) - h(\xi_{j-1}) - k\{h(\xi_j) - h(\xi_{j-1})\}(x - \xi_{j-1})| \\ &= |h'(\zeta_2) - h'(\zeta_1)|(x - \xi_{j-1}) \\ &\leq L|\zeta_2 - \zeta_1|^{\alpha-1}(x - \xi_{j-1}) \\ &\leq Lk^{-\alpha+1}k^{-1} \\ &= Lk^{-\alpha}, \end{aligned}$$

where we used the Hölder continuity of h' for the second to last inequality. Since the two bounds do not depend on x , we conclude that $d_\infty(h, h_k) = \mathcal{O}(k^{-\alpha})$. $\square$

In the sequel, we write

$$\tilde{f}_{0,k}(x) = \frac{1}{2k}f_0(0)\omega(x|k, 0) + \frac{1}{k}\sum_{j=1}^{k-1}f_0(j/k)\omega(x|k, j) + \frac{1}{2k}f_0(1)\omega(x|k, k) \quad (6.21)$$

The above result shows that if we use (6.21) to approximate f_0 , then the approximation error is of order $\mathcal{O}(k^{-\alpha})$ in the supremum metric which implies that the $\mathbb{L}_2$ distance between the two root-functions are of order $\mathcal{O}(k^{-2\alpha})$, provided the densities involved are strictly positive for all k . The following lemma establishes that we can find a set of spline densities using the same set of knots which attains the same approximation rate as $\tilde{f}_{0,k}$.

Lemma 6.4. *Let $f_0: [0, 1] \rightarrow \mathbb{R}$ be a strictly positive, α -smooth probability density. Then we can find a linear spline density $f_{0,k}$ using a regular set of knots such that*

$$d_H^2(f_0, f_{0,k}) = \mathcal{O}(k^{-2\alpha}).$$

Proof. Let $\tilde{f}_{0,k}$ be the linear spline interpolating f_0 at $\{j/k: 0 \leq j \leq k\}$. To start we will show that $d_H^2(\tilde{f}_{0,k}, f_{0,k}) = \mathcal{O}(k^{-2\alpha})$, where $f_{0,k}(x) = \tilde{f}_{0,k}(x) / \int_0^1 \tilde{f}_{0,k}(s) ds$ is the normalized version of $\tilde{f}_{0,k}$. A straightforward calculation shows that

$$\begin{aligned} d_H^2(\tilde{f}_{0,k}, f_{0,k}) &= \left\| \sqrt{\tilde{f}_{0,k}} \left(1 - 1/\sqrt{\|\tilde{f}_{0,k}\|_1} \right) \right\|_2^2 \\ &= \int_0^1 \tilde{f}_{0,k}(x) \left(1 - 1/\sqrt{\|\tilde{f}_{0,k}\|_1} \right)^2 dx \\ &= \|\tilde{f}_{0,k}\|_1 \left(1 - 1/\sqrt{\|\tilde{f}_{0,k}\|_1} \right)^2. \end{aligned} \quad (6.22)$$

Thus, we turn our attention to showing that $\|\tilde{f}_{0,k}\|_1^{-1/2}$ approaches 1 sufficiently quickly. We note that

$$\left| \|\tilde{f}_{0,k}\|_1 - 1 \right| = \left| \int_0^1 \tilde{f}_{0,k}(x) - f_0(x) dx \right| \leq \int_0^1 |\tilde{f}_{0,k}(x) - f_0(x)| dx \leq \|\tilde{f}_{0,k} - f_0\|_\infty. \quad (6.23)$$

By the α -smoothness of f_0 and Proposition 6.1, the right hand side of (6.23) is $\mathcal{O}(k^{-\alpha})$. It follows that for all sufficiently large k , we have that $\|\tilde{f}_{0,k}\|_1 \in (1/2, 2)$, so to show that the expression on the right hand side of (6.22) is of the claimed order, it suffices to show that the second factor in this expression tends to zero at rate $\mathcal{O}(k^{-2\alpha})$. By a first order Taylor expansion with remainder we find that

$$\frac{1}{\sqrt{\|\tilde{f}_{0,k}\|_1}} = 1 + \frac{1}{2}(\|\tilde{f}_{0,k}\|_1 - 1) + o(\|\tilde{f}_{0,k}\|_1 - 1) = 1 + \mathcal{O}(k^{-\alpha}),$$

from which it follows that $\left(1 - 1/\sqrt{\|\tilde{f}_{0,k}\|_1} \right)^2 = \mathcal{O}(k^{-2\alpha})$. By (6.22) we conclude that $d_H^2(\tilde{f}_{0,k}, f_{0,k}) = \mathcal{O}(k^{-2\alpha})$ as claimed. We now turn our attention to $d_H^2(\tilde{f}_{0,k}, f)$. Note that from the proof of Proposition 4.4 we have that

$$\left| \sqrt{f_0(x)} - \sqrt{\tilde{f}_{0,k}(x)} \right| \leq 2^{-1} m^{-1/2} |f_0(x) - \tilde{f}_{0,k}(x)| \leq 2^{-1} m^{-1/2} \|f_0 - \tilde{f}_{0,k}\|_\infty. \quad (6.24)$$

Furthermore, we note that the approximation property of Proposition 6.1 together with the above upper bound implies

$$d_H^2(f_0, \tilde{f}_{0,k}) = \int_0^1 \left(\sqrt{f_0(x)} - \sqrt{\tilde{f}_{0,k}(x)} \right)^2 dx \leq \left\| \sqrt{f_0} - \sqrt{\tilde{f}_{0,k}} \right\|_\infty^2 = \mathcal{O}(k^{-2\alpha}),$$

where the last inequality follows from (6.24) and $\|f_0 - \tilde{f}_{0,k}\|_\infty^2 = \mathcal{O}(k^{-2\alpha})$. Using the fact that $d_H^2(f_0, f_{0,k}) \leq 2d_H^2(f_0, \tilde{f}_{0,k}) + 2d_H^2(\tilde{f}_{0,k}, f_{0,k})$, the claim now follows. $\square$

Remark 6.1. For $\alpha \in (0, 1]$ the approximation rate of Lemma 6.4 is the same as for the piecewise constant model (Proposition 4.3). For $\alpha \in (1, 2]$ the approximation rate of the linear spline model is better, because it is able to utilize the regularity of the derivative f'_0 .

6.5.2 Posterior convergence rate

With an approximation result and an estimate of the covering number in hand, we are ready to establish a convergence result for the linear spline prior. To start, we estimate the prior probability of a suitable ‘neighbourhood’ of the true density f_0 . Since we here assume that f_0 is strictly positive, we give an estimate that is useful in showing that (3.45), rather than (3.44) which was used to establish the convergence rate for random histograms.

Lemma 6.5. *Let $f_0: [0, 1] \rightarrow \mathbb{R}$ be an α -smooth, strictly positive density for some $\alpha \in (0, 2]$, and suppose that Assumption 6.2 holds. Then if $k = k(\epsilon) \in \mathbb{N}$ satisfies $C_1\epsilon^{-1/\alpha} \leq k \leq C_2\epsilon^{-1/\alpha}$ for two positive constants C_1, C_2 then we can find constants $c_1, c_2 > 0$ such that for all sufficiently small $\epsilon > 0$,*

$$P_n\left(f: d_H^2(f_0, f) \|f_0/f\|_\infty \leq \epsilon^2\right) \geq p_n(k) c_1 \exp\{-c_2 k \log(1/\epsilon)\}.$$

Proof. Denote $f_{0,k}(\cdot) = \sum_{j=0}^k \rho_j \omega(\cdot | k, j)$ as in the preceding lemma for some $\rho \in \mathcal{S}_k$ to be specified later, and take $\gamma = c_0 \epsilon$ for a constant c_0 to be determined later. Since $d_H^2(f_0, f_{0,k}) = \mathcal{O}(k^{-2\alpha})$ by Lemma 6.4 and $k = \mathcal{O}(\epsilon^{-1/\alpha})$ by assumption, it follows that $d_H^2(f_0, f_{0,k}) = \mathcal{O}(\gamma^2)$. Let $\mathcal{A}_\gamma = \{\theta \in \mathcal{S}_k: d_1(\theta, \rho) \leq \gamma^{1+2/\alpha}\}$ and define $\rho \in \mathcal{S}_k$ by

$$\begin{aligned} \rho_j &= \frac{f(\xi_j)}{k \|\tilde{f}_{0,k}\|_1}, \quad \text{for } j = 1, 2, \dots, k-1, \\ \rho_j &= \frac{f(\xi_j)}{2k \|\tilde{f}_{0,k}\|_1}, \quad \text{for } j = 0, k, \end{aligned}$$

where $\tilde{f}_{0,k}$ is as in (6.21). On the event $\mathcal{A}_\gamma$ we have that

$$d_H^2(f_k, f_{0,k}) \leq d_1(f_k, f_{0,k}) = d_1(\theta, \rho) \leq \gamma^{1+2/\alpha} \leq \gamma^2.$$

By the triangle inequality, it follows that $d_H^2(f_0, f_k) \leq 2d_H^2(f_0, f_{0,k}) + 2d_H^2(f_{0,k}, f_k) \leq \mathcal{O}(\gamma^2)$. We now turn our attention to bounding $\|f_0/f_k\|_\infty$ on $\mathcal{A}$, tending to the denominator first. To this end, we note that by construction, $\rho_j \geq f(\xi_j)/(2k \|\tilde{f}_{0,k}\|_1)$. Appealing to (6.23), we have for small ϵ and hence large k that $\|\tilde{f}_{0,k}\|_1 \leq 2$, which implies that $\rho_j \geq m/[4k]$. Since $\sum_{j=0}^k \omega(x | k, j) \geq k$ for all $x \in [0, 1]$, we conclude that $f_{0,k} = \sum_{j=0}^k \rho_j \omega(\cdot | k, j) \geq m/4$. We further observe that

$$d_\infty(f_{0,k}, f_k) = \sup_{x \in [0, 1]} \sum_{j=0}^k |\rho_j - \theta_j| \omega(x | k, j) \leq \sum_{j=0}^k |\rho_j - \theta_j| \sup_{x \in [0, 1]} \omega(x | k, j) \leq 2k \sum_{j=0}^k |\rho_j - \theta_j|$$

where we used that $\omega(x | k, j) \leq 2k$ for all x and j . Since $d_1(f_{0,k}, f_k) = \mathcal{O}(\epsilon^{1+1/\alpha})$ and $k = \mathcal{O}(\epsilon^{-1/\alpha})$ it follows that $d_\infty(f_{0,k}, f_k) = \mathcal{O}(\epsilon)$. In particular, this implies that for small enough ϵ , we have that $d_\infty(f_{0,k}, f_k) \leq m/8$ which further yields $f_k = f_{0,k} - (f_{0,k} - f_k) \geq m/4 - m/8 = m/8$. As such, we find that

$$\|f_0/f_k\|_\infty \leq \frac{\sup_{x \in [0, 1]} f_0(x)}{\inf_{x \in [0, 1]} f_k(x)} \leq 8M/m.$$

As an immediate consequence of the above upper bound and previous estimates, $d_H^2(f_0, f_k) \|f_0/f_k\|_\infty = \mathcal{O}(\gamma^2)$. By taking the constant c_0 to be small enough, we thus ensure that $d_H^2(f_0, f_k) \|f_0/f_k\|_\infty \leq \epsilon^2$.

We now verify the conditions of Lemma B.3 in order to lower bound $P_n(\mathcal{A}_\gamma)$. Note that $\gamma^{1+2/\alpha} \leq 1/(\Sigma k)$ for all sufficiently large k , due to the fact that $\gamma^{1+2/\alpha} = \mathcal{O}(\epsilon^{1+2/\alpha}) = \mathcal{O}(k^{-2-\alpha}) = o(k^{-1})$ by assumption. Furthermore, $a_{k,j}$ is lower bounded by a multiple of γ^b since $a_{k,j} > \sigma k^{-1}$ by assumption. Applying Lemma B.3 to a constant multiple of $\gamma^{1+2/\alpha}$, we find that

$$P_n\left(f_k: d_H^2(f_0, f_k) \|f_0/f_k\|_\infty \leq \epsilon^2 | k\right) \geq P_n(\mathcal{A}_\gamma) \geq c_1 \exp\{-c_3(k+1) \log(1/\gamma)\},$$

where $c_1, c_3 > 0$ are constants. Noting that $(k+1)/k$ and $\log(1/\gamma)/\log(1/\epsilon)$ both converge to positive constants as $\epsilon \downarrow 0$, we can find a constant $c_2 > 0$ such that the above expression is lower bounded by $c_1 \exp\{-c_2 k \log(1/\epsilon)\}$, which further implies that

$$P_n(f: d_H^2(f_0, f) \|f_0/f_k\|_\infty \leq \epsilon^2) \geq c_1 p_n(k) \exp\{-c_2 k \log(1/\epsilon)\},$$

which was to be shown. $\square$

Following the estimate Lemma 6.5, we are now in a position to prove the main result of the section. The proof follows along similar lines as that of Theorem 5.4, but we verify the condition (3.45) instead of (3.44). In this case we are able to utilize that f_0 is strictly positive, in which case the Hellinger distance and KL-divergence become comparable, and we don't incur any additional logarithmic factors from using Lemma B.4 as we did in Theorem 5.4.

Theorem 6.3. *Let $f_0: [0, 1] \rightarrow \mathbb{R}$ be a strictly positive α -smooth probability density with $\alpha \in (0, 2]$. If Assumption 6.2 holds, then the posterior distribution converges at the rate $\epsilon_n = \{n/\log(n)\}^{-\alpha/(2\alpha+1)}$.*

Proof. We verify the conditions of Theorem 3.4, tending to (3.45) first for $\beta_n = \{n/\log(n)\}^{-\alpha/(2\alpha+1)}$. As our first step, note that $\sum_{n=1}^\infty \exp(-En\beta_n^2) < \infty$ for any $E > 0$, satisfying the requirement for this version of Theorem 3.4.

Letting $\{s_n\}$ be a sequence of integers satisfying $C_1\beta_n^{-1/\alpha} \leq s_n \leq C_2\beta_n^{-1/\alpha}$ for positive constants C_1, C_2 , the result of Lemma 6.5 yields

$$P_n(f: d_H^2(f_0, f) \|f_0/f\|_\infty \leq \epsilon^2) \geq p_n(s_n) c_1 \exp\{-c_2 s_n \log(1/\beta_n)\}.$$

It can be shown that $\log(1/\beta_n)/\log(n)$ converges to a positive, finite limit as $n \rightarrow \infty$. Moreover, a quick computation shows that $n\beta_n^2 = n^{1/(2\alpha+1)}\{\log(n)\}^{2\alpha/(2\alpha+1)}$ and hence $s_n \log(n) \leq C_2 n^{1/(2\alpha+1)}\{\log(n)\}^{(2\alpha+1-1)/(2\alpha+1)} = C_2 n\beta_n^2$. We deduce that the above display is further bounded from below by

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

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

$$P_n(f: d_H^2(f_0, f) \|f_0/f\|_\infty \leq \epsilon^2) \geq c_3 \exp(-c_4 n\beta_n^2),$$

which establishes (3.45).

To verify the prior condition (3.43), 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 $t_n \log(n) \geq E_1 n\beta_n^2$. For $\mathcal{F}_n = \cup_{k=1}^{t_n} \mathcal{H}_k$, where $\mathcal{H}_k$ is as in (6.20), we then have that $P_n(\mathcal{F}_n^c) = \sum_{k=k_n+1}^\infty p_n(k)$. Using the assumption on $p_n(k)$ this is bounded above by $c_5 \exp(-d_2 t_n \log(t_n))$ for $c_5 = D_2 \sum_{k=0}^\infty \exp(-d_2 k \log(k)) < \infty$. Furthermore, as $\log(t_n)/\log(n)$ converges to a positive limit as $n \rightarrow \infty$, we can find a new constant d'_2 such that

$$\sum_{k=k_n+1}^\infty p_n(k) < c_5 \exp(-d'_2 t_n \log(n)).$$

Taking $E_1 = (c_4 + 4)/d'_2$, we have $d'_2 t_n \geq (c_4 + 4)n\beta_n^2$ which combined with previous estimates implies that $\mathcal{P}_n(\mathcal{F}_n^c) \geq \exp\{-(c_4 + 4)n\beta_n^2\}$, verifying (3.43).

Finally, we show that the entropy condition (3.42) holds. Let $\alpha_n = \{n/\log(n)\}^{-\alpha/(2\alpha+1)}$. By Lemma 6.3 the covering number $N(\alpha_n/2, \mathcal{F}_n, d_H)$ is upper bounded by $\exp\{\log(t_n) + (t_n + 1)\log(L/\alpha_n)\}$. Again, $\log(1/\alpha_n)/\log(n)$ converges to a positive finite limit, and as such the covering number can be further bounded by $\exp\{c_5 t_n \log(n)\}$ for a suitable constant c_5 . Since $t_n \log(n) = \mathcal{O}(n^{1/(2\alpha+1)}\{\log(n)\}^{2\alpha/(2\alpha+1)}) = \mathcal{O}(n\alpha_n^2)$ this implies that $N(\alpha_n/2, \mathcal{F}_n, d_H) \leq \exp(c_6 n\alpha_n^2)$ for some $c_6 > 0$ as desired.

By Theorem 3.4 we conclude that the convergence rate is $\epsilon_n = \max\{\alpha_n, \beta_n\} = \alpha_n$ which concludes the proof. $\square$

The rate implied by Theorem 6.3 for $\alpha \in (0, 1]$ is equal to the minimax rate up to logarithmic factors. The rate is also similar to the rate obtained by the random histogram procedure, so we do not really gain anything in terms of asymptotic rate by using the linear spline model over the histogram model if f_0 does not possess any additional smoothness. For $\alpha \in (1, 2]$ we also obtain the minimax rate up to logarithmic factors. In this case, the rate is also better than that obtained by the random histogram prior, since unlike piecewise constant functions, linear splines are able to exploit regularity of the derivative of the true density f_0 . The obtained convergence rate is faster than that obtained by the polygonally smoothed prior of Scricciolo (2007), owing to the fact that our model is piecewise linear over the entire unit interval, while Scricciolo's model is piecewise constant near the edges. In the case where f_0'' is absolutely continuous, a stronger assumption than α -smoothness with $\alpha = 2$, the $\{n/\log(n)\}^{-2/5}$ convergence rate is also comparable to the $n^{-2/5}$ rate of frequency polygons of Scott (1985); Jones et al. (1998).

6.6 Applications to real world data sets

Finally, we apply the models developed in this chapter and previous ones to some real-world data examples.

Example 6.1. For our first example, we consider a data set analyzed in Cunen et al. (2018), consisting of $n = 109$ observations of the yearly number of days where a snow depth greater than 25 cm was recorded at Bjørnholt, Nordmarka, or so-called 'skiing days', recorded during the years 1897 – 1937 and 1955 – 2022. Since this is a fairly small dataset, we decided to use a uniform prior for k on $\{1, 2, \dots, 30\}$, and we centered the Dirichlet prior distributions on the uniform distribution. To investigate the effect of the choice of the prior concentration parameter, we computed the density estimates and the estimated posterior probabilities (6.11) for different values of a . The results are shown in Figure 6.4.

The chosen value of a is seen to have a moderate effect on the resulting density estimates, with the choices $a = 5.5$ and $a = 10.0$ yielding rather similar density estimates. For the posterior probabilities of k , the effect of the concentration parameter is more drastic, as the posterior mass shifts towards larger values of k as a increases. In particular, the difference between $a = 1.0$ and $a \in \{5.5, 10.0\}$ is large, as the former choice yields a posterior very concentrated on $k = 3$, while the posterior mass for $a \in \{5.5, 10.0\}$ is more evenly spread out. The results here mirrors those of the random histogram model, where smaller values of a lead to greater concentration of posterior mass at smaller values

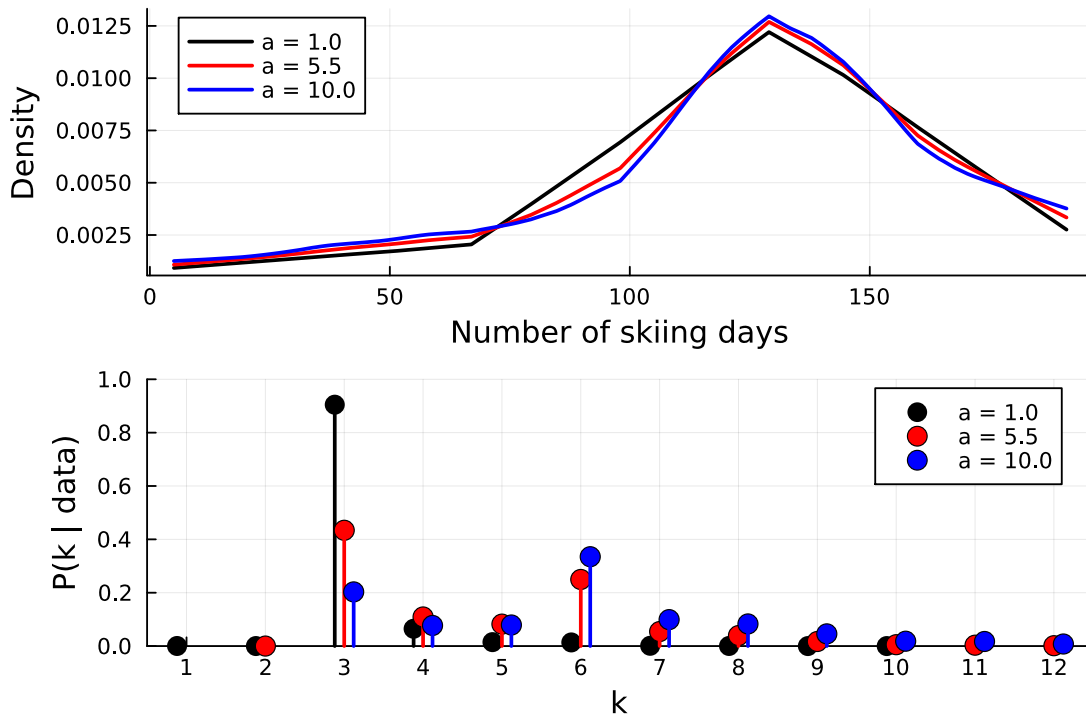


Figure 6.4: Density estimate and estimated posterior for k for different choices of the prior concentration parameter.

of k . Although the parameter a controls the degree of belief in the prior one also has to account for the fact that it greatly influences the resulting posterior for k , which will typically have a large effect on the density estimate. In the regular random histogram model, we discovered that the risk performance of the procedure improved when the concentration parameter increased slowly with n . As such, we conjecture that it will similarly be beneficial to increase the prior concentration parameter with the sample size in this case, but further investigation is needed before any strong conclusions can be drawn. Another possibility that might be investigated in the future is to let the concentration parameter increase linearly with k as in the Jeffreys' and uniform priors in the random histogram case.

For our further analyses, we took $a = 5.5$ reflecting a moderate degree of belief in the prior mean for $\theta | k$. Additionally, we fitted the regular random histogram model of Example 4.2 to the same dataset with the same value of the concentration parameter, and the same uniform distribution as our prior for k . The resulting posterior mean, along with 90% credible bands for the posterior density in the two models are shown in Figure 6.5. The resulting density estimates differ quite a bit reflecting the different models used for the density. Both estimates are unimodal, showing that a typical year during the sampling period will have between 100 and 150 skiing days. The density estimate and the credible bands based on the regular histogram model has many large jumps, reflecting the discontinuity of the prior. In particular, the most notable jumps correspond to the value of $k = 5$, reflecting that most approximately 90% of the posterior mass is located here. The linear spline model on the other hand yields a smoother and more visually appealing estimate in this case, reflecting the continuous model used for the density. The credible bands are quite wide in this case, reflecting large posterior uncertainty due to the small sample size. ▲

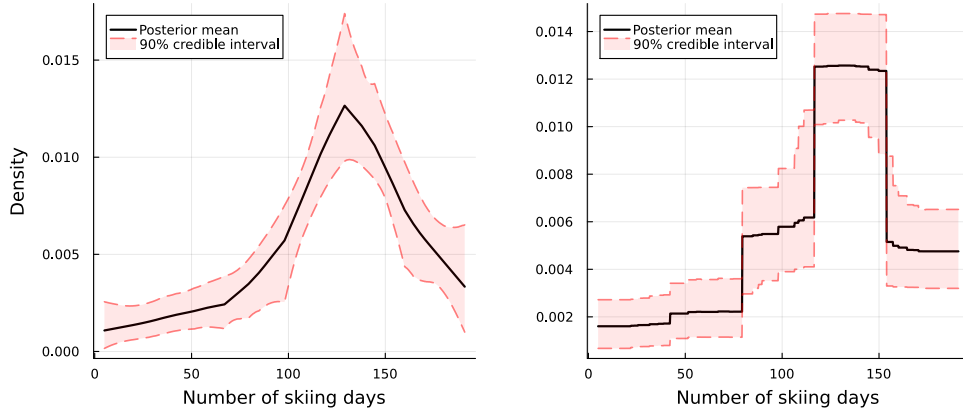


Figure 6.5: Density estimates with 90% credible bands using a linear spline (left) and random histogram prior. (right).

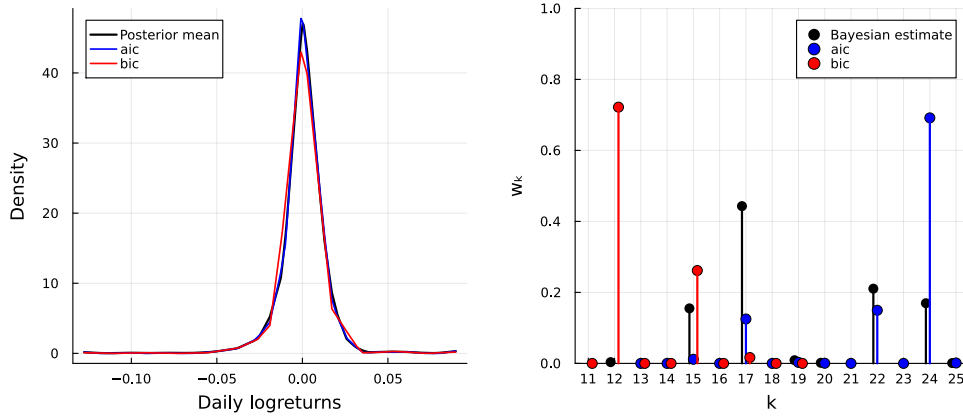


Figure 6.6: Left: Bayesian, AIC- and BIC-based density estimates based on daily log-returns. Right: Model weights w_k for the three different models. Only $w_k > 10^{-3}$ are shown.

Example 6.2. We return to the S&P 500 stock data from the introduction, analyzing the daily logreturns using the fully Bayesian linear spline model as well as the maximum likelihood-based estimator (6.12), using model weights derived from AIC and BIC respectively. For the Bayesian model, we chose a value of $a = 8$ and used a uniform prior for k on $\{1, 2, \dots, 40\}$. The three density estimators are shown in figure Figure 6.6, along with the model weights w_k given by (3.22) for the information criterion-based models and $w_k = p(k | \mathbf{x}_{1:n})$ for the fully Bayesian model.

All three density estimates show a dominant mode around 0. The Bayesian model and the AIC-based estimator yielded remarkably similar density estimates; the two densities are hardly distinguishable bar a minor difference near the peak. The density estimate based on the BIC exhibits a less pronounced peak than the other two. The plot of the model weights reveal a stark difference between the three models, with the BIC showing a clear preference for smaller values of k than the other two, reflecting the fact that the BIC penalizes model complexity to a greater extent than AIC. Although the model weights for the AIC-based and Bayesian models differ quite a bit, the resulting density estimates are very similar as noted above. ▲

Chapter 7

Conclusion

The main goal of this thesis has been to study nonparametric density estimation from a Bayesian point of view, focusing on models with automatic selection of smoothing parameters. Our approach is based on mixture densities of the form

$$f(x) = \sum_{j=1}^k \theta_j \omega_j(x),$$

where the number k of components in the mixture is treated as a random variable. We considered both the case where the ω_j are piecewise constant and the case where they are piecewise linear and continuous. In particular, the two piecewise constant models can be used to construct a histogram in an automatic fashion, offering a Bayesian alternative to the numerous existing frequentist histogram procedures.

In Chapter 4, we focused our attention on regular histograms. In particular, we studied a computationally convenient method of histogram construction based on Bayesian model selection. We undertook analyses of the asymptotic behavior of the posterior distribution as well as the Bayes histogram estimator using the toolbox from Ghosal et al. (1999, 2000), providing results on posterior consistency and posterior convergence rates. We concluded the chapter with a large simulation study, comparing the Bayesian method to previous proposals from the statistical literature, the results of which show that the Bayesian model is highly competitive with existing methods in terms of Hellinger risk, especially for smaller sample sizes.

Chapter 5 explored the world of irregular histograms. Much of the chapter is dedicated to the study of a novel Bayesian irregular histogram model, which extends the Bayesian model from the regular case. We show how to efficiently compute the Bayes histogram estimator using a dynamical programming algorithm, and discuss an approximate method based on a greedy search heuristic that can speed up computations for larger sample sizes. Additionally, we proposed a Metropolis–Hastings algorithm to generate approximate samples from the posterior distribution. Posterior asymptotics in the irregular histogram model were investigated, as we show posterior consistency and derive the posterior convergence rate in addition to providing a consistency result for the Bayes histogram estimator. We concluded the chapter with a large simulation study that compared the performance of our Bayesian method with other irregular histogram approaches with respect to classical loss functions and the ability to correctly determine the number and location of modes. The Bayesian method is shown to perform well in terms of both mode identification and L_2 loss. Finally, we compared the performance of some of regular and irregular histogram procedures, finding that neither class of methods

has a particularly large and consistent advantage with respect to classical loss functions, but that irregular histograms are much better at automatic mode detection.

The last chapter broadened the horizon by extending the piecewise constant model of Chapter 4 to a continuous one, with Chapter 6 focusing on the case of a piecewise linear model. To carry out posterior inference, we proposed an MCMC algorithm to generate approximate samples from the posterior distribution and apply the method to a few real-world datasets. The posterior distribution in the model was shown to be consistent under general conditions, and we gave explicit posterior convergence rates under suitable smoothness assumptions.

Many avenues for future research presented themselves during the writing of this thesis, and it is regrettable that we did not have the time to explore these further. Although our asymptotic results shine some light on the behavior of the posterior distribution with respect to classical loss functions, it would also be interesting to prove statements on the asymptotic behavior of the methods presented here in terms of mode identification. Results regarding mode identification are provided in Davies and Kovac (2004) and Li et al. (2020) for the Taut String and the Essential histogram, respectively, and the techniques therein could possibly be adapted to study the ability of other irregular histogram procedures to identify modes. Additionally, deducing precise convergence rate results with respect to the Hellinger metric for the Bayes histogram estimators in the regular and irregular random histogram models would be of great interest. We have pursued some results in this direction, using the techniques developed in Ghosal et al. (2000), but were unable to finish them in time for publication. Since histograms are primarily a graphical tool, it would also have been interesting to explore the behavior of our estimator with respect to the visual error criteria of Marron and Tsybakov (1995).

In our applied work, we have exclusively relied on the Dirichlet distribution as a prior for the mixture parameters, in part due to its attractive conjugacy properties. One disadvantage of this approach is that the Dirichlet distribution cannot incorporate smoothness assumptions into the prior distributions, since the covariances between different mixture probabilities are always negative. A possible approach to induce positive correlations in the simplex-valued parameter θ is to use the proposal of Leonard (1973), by starting with a k -dimensional multivariate normal parameter γ with a non-diagonal covariance matrix, and transforming it to the simplex via $\theta_j = \exp\{\gamma_j\} / \sum_l \exp\{\gamma_l\}$. In the irregular random histogram model, it would also be interesting to explore the use of different prior distributions on the partitions $I_{1:k} | k$ than the uniform. In particular, if we have prior information that the density we are trying to estimate is spatially homogeneous, then a piecewise constant function based on a highly irregular partition is unlikely to result in a reasonable approximation to the true density. This calls for the use of a prior distribution that assigns higher prior mass to more regular partitions, which could possibly better the performance of the irregular random histogram model with respect to classical loss functions for such densities.

In cases where prior information is not available, an alternative to using a noninformative prior is to use an empirical Bayes approach, where the hyperparameters of the prior distribution are estimated from the data. As we discussed in Section 4.7.1, it is possible to center $\mathbf{p} | k \sim \text{Dir}(a\mathbf{p}_0)$ on a reference density g_0 by setting $p_{0,j} = \int_{I_{k,j}} g_0(s) ds$ for all k and j , where we now estimate g_0 from the data. One possible approach is to set g_0 equal to the density arising from maximum likelihood or moment matching estimates in a parametric family of distributions on $[0, 1]$, such as the Beta family or a low-dimensional mixture thereof. Even in the case where the true density is

not close to any member of the parametric family in question, the estimated prior mean would likely be a better fit to the data than the one based on the uniform, leading to improved density estimates. This approach shares similarities with that of Hjort and Glad (1995), where a parametric density estimate is used as an ‘initial guess’ of the density which is subsequently corrected by a nonparametric kernel estimate.

More work could also be done to improve the mixing properties of the MCMC algorithm presented in Sections 5.2.3 and 6.2.1. In Example 6.2, the posterior pmf $p_n(k | \mathbf{x}_{1:n})$ exhibited modes at different values of k that were separated by a region with low posterior mass, that made it difficult for the MCMC algorithm to properly explore the parameter space. One possible approach would be to use simulated tempering (Marinari and Parisi, 1992), which introduces an auxiliary variable that allows for more rapid exploration of the parameter spaces, making it suited for multimodal problems. The fact that we have to resort to MCMC in the linear spline model is a weakness from a computational perspective, as more efficient continuous and piecewise linear estimators exist in the form of frequency polygons (Scott, 1985; Jones et al., 1998). Future work could also investigate systematically whether there is a gain in performance from using the Bayesian or penalized likelihood linear spline model over frequency polygons through the use of simulations.

Appendix A

Appendix A

A.1 Technical proofs

Proof of Proposition 2.2

Proof. To show (2.50), take $\epsilon = 1$ in (2.45), and (2.44) follows immediately from the definition of convergence for sequences of real numbers.

To establish (2.46), suppose $U_n = \mathcal{O}_{\mathbb{P}}(a_n)$ and $V_n = \mathcal{O}_{\mathbb{P}}(b_n)$. We note that

$$\mathbb{P}(|U_n + V_n| \geq Mb_n) \leq \mathbb{P}(|U_n| \geq Mb_n/2) + \mathbb{P}(|V_n| \geq Mb_n/2).$$

This follows by considering the complementary events of the inclusion

$$\{|U_n| < Mb_n/2\} \cup \{|V_n| < Mb_n/2\} \subseteq \{|U_n + V_n| < Mb_n\} \quad (\text{A.1})$$

Since $a_n/b_n \rightarrow c$ for $c \geq 0$, we eventually have $a_n \leq (c+1)b_n$. Hence, for all large n ,

$$\mathbb{P}(|U_n + V_n| \geq Mb_n) \leq \mathbb{P}(|U_n| \geq M(c+1)^{-1}a_n/2) + \mathbb{P}(|V_n| \geq Mb_n/2).$$

By making M large enough, both terms become arbitrarily small, establishing (2.46). (2.47) is established using a similar argument.

Let $U_n = \mathcal{O}_{\mathbb{P}}(a_n)$. We prove (2.48) by noting that

$$\mathbb{P}(c_n|U_n| \geq Mc_na_n) = \mathbb{P}(|U_n| \geq Ma_n). \quad (\text{A.2})$$

The right hand side can be made arbitrarily small for sufficiently large n and M by assumption, proving the claim. One proves (2.49) by an analogous argument.

For the proof of (2.51), fix $\epsilon > 0$. We note that by assumption for any $\gamma > 0$ there is an $M > 0$ such that for all sufficiently large n ,

$$\mathbb{P}(|X_n| \geq Ma_n) < \gamma. \quad (\text{A.3})$$

Moreover as $a_n \rightarrow 0$ we have $Ma_n < \epsilon$ for all large n . Hence, for any such n ,

$$\mathbb{P}(|X_n| \geq \epsilon) \leq \mathbb{P}(|X_n| \geq Ma_n) < \gamma.$$

Taking the limit superior of the above chain of inequalities, we conclude that $\limsup_{n \rightarrow \infty} \mathbb{P}(|X_n| \geq \epsilon) < \gamma$. Since $\gamma > 0$ was arbitrary, we conclude that $X_n \xrightarrow{\mathbb{P}} 0$ as claimed.

To show (ii) let X_n converge to X in probability. Note that for any $M > 0$,

$$\mathbb{P}(|X_n| \geq M) \leq \mathbb{P}(|X_n - X| \geq M/2) + \mathbb{P}(|X| \geq M/2). \quad (\text{A.4})$$

By taking M sufficiently large, the second term in (A.4) can be made arbitrarily small. Since $X_n \xrightarrow{\mathbb{P}} X$, the first term can also be made arbitrarily small provided n is sufficiently large, showing that $X_n = \mathcal{O}_{\mathbb{P}}(1)$. The last claim follows from writing $X_n = X + (X_n - X)$, where the last term is seen to be $o_{\mathbb{P}}(1)$.

Finally, (iii) can be shown by noting that from the definition, it is clear that $X_{n,j} = o_{\mathbb{P}}(a_n)$ if and only if $|X_{n,j}| = o_{\mathbb{P}}(a_n)$. Note that $\max_{1 \leq j \leq k} |X_{n,j}| \leq \sum_{j=1}^k |X_{n,j}|$, which is $o_{\mathbb{P}}(a_n)$ by (2.47), along with a standard inductive argument. $\square$

Proposition A.1. *Let $\mathbf{x}_{1:n} = (x_1, \dots, x_n)$ be numbers in the interval $[0, 1]$. Let $\log L_n(\mathbf{p} | k)$ denote the log-likelihood function of a regular histogram based on the sample $\mathbf{x}_{1:n}$. Then,*

$$\log L_n(\hat{\mathbf{p}} | k) \rightarrow \infty \text{ as } k \rightarrow \infty. \quad (\text{A.5})$$

Proof. Let $x_1^*, \dots, x_M^*$ denote the unique values in the sample, and let N_m^* denote the number of x_i that take on the value x_m^* , that is, $N_m^* = \sum_{i=1}^n \mathcal{I}(x_i = x_m^*)$. By choosing k sufficiently large, the bin width k^{-1} will be smaller than the smallest distance between two unique values in the sample. As such, the interval $I_{k,j}$ satisfying $x_m^* \in I_{k,j}$ contains N_m^* values for such k . Plugging in the maximum likelihood estimates for $\hat{\mathbf{p}}$, the likelihood function (2.20) becomes

$$L_n(\hat{\mathbf{p}} | k) = k^n \prod_{j=1}^k \hat{p}_j^{N_j} = k^n \prod_{m=1}^M \left(\frac{N_m^*}{n} \right)^{N_m^*}.$$

Taking the logarithm of the above expression yields the log-likelihood

$$\log L_n(\hat{\mathbf{p}} | k) = n \log(k) + \sum_{m=1}^M N_m^* \log(N_m^*/n). \quad (\text{A.6})$$

By the above argument, the second term in (A.6) is a constant for all sufficiently large k , while the first term diverges to ∞ as $k \rightarrow \infty$, establishing (A.5). $\square$

Proposition A.2. *Let f_0 be a bounded probability density on $[0, 1]$. Let $f_{0,I_{1:k}}$ denote the piecewise constant version of f_0 based on the partition $I_{1:k}$ of $[0, 1]$. Then for any function $g_{I_{1:k}}$ which is piecewise constant on $I_1, \dots, I_k$, we have that*

$$\|f_0 - g_{I_{1:k}}\|_2^2 = \|f_0 - f_{0,I_{1:k}}\|_2^2 + \|f_{0,I_{1:k}} - g_{I_{1:k}}\|_2^2. \quad (\text{A.7})$$

Proof. Denote $\pi_j = \int_{I_j} f_0(x) dx$, and write $g_{I_{1:k}}(x) = \prod_{j=1}^k (\rho_j / |I_j|)^{\mathcal{I}(x \in I_j)}$. Using the fact that $g_{I_{1:k}}$ is constant on I_j we find that the left hand side of (A.7) becomes

$$\begin{aligned} \|f_0 - g_{I_{1:k}}\|_2^2 &= \sum_{j=1}^k \int_{I_j} \{f_0(x) - g(x)\}^2 dx \\ &= \sum_{j=1}^k \left\{ \int_{I_j} f_0(x)^2 dx - 2|I_j|^{-1} \rho_j \int_{I_j} f_0(x) dx + |I_j|^{-1} \rho_j^2 \right\} \\ &= \sum_{j=1}^k \left\{ \int_{I_j} f_0(x)^2 dx - 2|I_j|^{-1} \rho_j \pi_j + |I_j|^{-1} \rho_j^2 \right\}. \end{aligned} \quad (\text{A.8})$$

Now plugging in $f_{0,I_{1:k}}$ for $g_{I_{1:k}}$ in this expression, we deduce that

$$\|f_0 - f_{0,I_{1:k}}\|_2^2 = \sum_{j=1}^k \left\{ \int_{I_j} f_0(x)^2 dx - 2|I_j|^{-1} \pi_j \pi_j + |I_j|^{-1} \pi_j^2 \right\} = \sum_{j=1}^k \left\{ \int_{I_j} f_0(x)^2 dx - |I_j|^{-1} \pi_j^2 \right\}.$$

We proceed by, plugging in $f_{0,I_{1:k}}$ for f_0 in (A.8). This, together with the fact that $\int_{I_j} f_{0,I_{1:k}}(x)^2 dx = \pi_j^2$ yields

$$\|f_{0,I_{1:k}} - g_{I_{1:k}}\|_2^2 = \sum_{j=1}^k \left\{ \pi_j^2 - 2|I_j|^{-1} \rho_j \pi_j + |I_j|^{-1} \rho_j^2 \right\}.$$

Collecting terms, we have

$$\|f_0 - f_{0,I_{1:k}}\|_2^2 + \|f_{0,I_{1:k}} - g_{I_{1:k}}\|_2^2 = \sum_{j=1}^k \left\{ \int_{I_j} f_0(x)^2 dx - 2|I_j|^{-1} \rho_j \pi_j + |I_j|^{-1} \rho_j^2 \right\} = \|f_0 - g_{I_{1:k}}\|_2^2,$$

which proves (A.7). $\square$

Lemma A.1. *Let $\boldsymbol{\pi} \in \mathcal{S}_{k-1}$ and $\mathbf{p}$ follow a k -dimensional Dirichlet distribution with parameter vector $\mathbf{a}_k \in \mathbb{R}_+^k$. Then*

$$\mathbb{P}(\|\mathbf{p} - \boldsymbol{\pi}\|_1 < \delta) > 0, \quad \forall \delta > 0.$$

Proof. This can be shown by arguing as in Ghosal and van der Vaart (2017, Lemma G.13). $\square$

Lemma A.2. *Let $f, g: \mathcal{X} \rightarrow \mathbb{R}$ be probability densities. Then $d_H^2(f, g) \leq \chi^2(f, g)$, where $\chi^2(f, g)$ is the chi-square divergence between f and g , given by*

$$\chi^2(f, g) = \int_{\mathcal{X}} \frac{\{f(x) - g(x)\}^2}{g(x)} dx. \quad (\text{A.9})$$

Proof. If $\chi^2(f, g)$ is infinite, then the inequality is trivially strict, since $d_H^2(f, g) \leq 2$ for any pair of densities f, g . To show that the inequality holds for finite $\chi^2(f, g)$, we claim that $|\sqrt{x} - 1| \leq |x - 1|$ for all $x \geq 0$. This inequality follows from the fact that $\sqrt{x} - 1 \leq x - 1$ for $x > 1$ and $\sqrt{x} - 1 > x - 1$ for $x < 1$, since we have equality for $x = 1$. From this we deduce that

$$\begin{aligned} d_H^2(f, g) &= \int_{\mathcal{X}} g(x) \left(\sqrt{\frac{f(x)}{g(x)}} - 1 \right)^2 dx \leq \int_{\mathcal{X}} g(x) \left(\frac{f(x)}{g(x)} - 1 \right)^2 dx \\ &= \int_{\mathcal{X}} \frac{\{f(x) - g(x)\}^2}{g(x)} dx, \end{aligned}$$

which was to be shown. $\square$

Proposition A.3 (Relationship between different distances). *Let $\mathcal{F}$ be the space of nonnegative bounded functions from $\mathcal{X} \subseteq \mathbb{R}$ to $\mathbb{R}$. Let d_1 and d_H be the $\mathbb{L}_1$ and Hellinger metrics on $\mathcal{F}$ respectively. Then for any $f, g \in \mathcal{F}$,*

$$d_H^2(f, g) \leq d_1(f, g) \leq 2d_H(f, g). \quad (\text{A.10})$$

If in addition f, g are lower bounded away from zero then

$$d_H^2(f, g) \leq m^{-1} d_2^2(f, g), \quad (\text{A.11})$$

where d_2 is the $\mathbb{L}_2$ metric on $\mathcal{F}$ and $m > 0$ is any number satisfying $g \geq m$.

Proof. To show the first inequality in (A.10), note that

$$\begin{aligned} d_H^2(f, g) &= \int_{\mathcal{X}} \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx \\ &\leq \int_{\mathcal{X}} \left| \sqrt{f(x)} - \sqrt{g(x)} \right| \left(\sqrt{f(x)} + \sqrt{g(x)} \right) dx \\ &= \int_{\mathcal{X}} |f(x) - g(x)| dx \\ &= d_1(f, g). \end{aligned}$$

The second inequality in (A.10) is proven using a similar trick:

$$\begin{aligned} d_1(f, g) &= \int_{\mathcal{X}} |f(x) - g(x)| dx \\ &= \int_{\mathcal{X}} \left| \sqrt{f(x)} - \sqrt{g(x)} \right| \left(\sqrt{f(x)} + \sqrt{g(x)} \right) dx \\ &\leq \left\{ \int_{\mathcal{X}} \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx \right\}^{1/2} \left\{ \int_{\mathcal{X}} \left(\sqrt{f(x)} + \sqrt{g(x)} \right)^2 dx \right\}^{1/2}, \end{aligned} \quad (\text{A.12})$$

where we used Cauchy–Schwarz for the last inequality. The first factor on the right hand side is $d_H(f, g)$, while the second can be upper bounded as follows:

$$\begin{aligned} \int_{\mathcal{X}} \left(\sqrt{f(x)} + \sqrt{g(x)} \right)^2 dx &= \int_{\mathcal{X}} \left\{ f(x) + g(x) + 2\sqrt{f(x)g(x)} \right\} dx \\ &= 2 + 2 \int_{\mathcal{X}} \sqrt{f(x)g(x)} dx \\ &\leq 4, \end{aligned}$$

where we used Cauchy–Schwarz again for the last inequality and the fact that f, g are densities. Combining this estimate together with (A.12), we have that $d_1(f, g) \leq 2d_H(f, g)$ as desired.

To prove the second part of the proposition, we note that by Lemma A.2 and the fact that $g \geq m > 0$,

$$d_H^2(f, g) \leq \chi^2(f, g) = \int_{\mathcal{X}} \frac{\{f(x) - g(x)\}^2}{g(x)} dx \leq m^{-1} d_2^2(f, g),$$

which concludes the proof. $\square$

Proposition A.4 (Finite entropy). *Let $f_0: [0, 1] \rightarrow \mathbb{R}$ be a bounded probability density. Then*

$$\int_0^1 f_0(t) |\log f_0(t)| dt < \infty. \quad (\text{A.13})$$

Proof. We will make use of an elementary inequality for the logarithm in order to bound the integrand in A.13 from above. We claim that

$$1 - \frac{1}{x} \leq \log(x) \leq x - 1, \quad \forall x > 0, \quad (\text{A.14})$$

and the inequalities are strict for $x \neq 1$. We prove the lower bound first. For $x = 1$ we have equality. For the case $x > 1$, we note that since $1/u < 1$ for $u > 1$,

$$\log(x) = \int_1^x \frac{du}{u} < \int_1^x du = x - 1. \quad (\text{A.15})$$

To prove the claim for $0 < x < 1$ we employ a similar strategy:

$$\log(x) = - \int_x^1 \frac{du}{u} > \int_x^1 -\frac{du}{u^2} = \left[\frac{1}{u} \right]_x^1 = 1 - \frac{1}{x},$$

where we used that $1/u^2 > 1/u$ and hence $-1/u^2 < -1/u$ for $0 < u < 1$.

To show the upper bound, we note that for $x < 1$,

$$\log(x) = - \int_x^1 \frac{du}{u} > - \int_1^x \frac{du}{x} = 1 - 1/x.$$

For $0 < x < 1$ we can use a similar integral bound:

$$\log(x) = - \int_x^1 \frac{du}{u} < \int_x^1 -du = [-u]_x^1 = x - 1,$$

where we used that for $0 < u < 1$, $1/u > 1$ and hence $-1/u < -1$. Finally, for $x = 1$ we have equality. Hence all the claimed inequalities hold for any $x > 0$.

We now prove the statement (A.13). Using equation (A.14), we have that for any x such that $f_0(x) > 0$,

$$1 - \frac{1}{f_0(x)} \leq \log f_0(x) \leq f_0(x) - 1$$

Multiplying by $f_0(x)$ on both sides, we derive the inequality

$$f_0(x) - 1 \leq f_0(x) \log f_0(x) \leq f_0^2(x) - f_0(x). \quad (\text{A.16})$$

Denote $M = \sup_{x \in [0,1]} |f_0(x)|$. Then it follows from (A.16) that

$$\begin{aligned} f_0(x) |\log f_0(x)| &\leq \max \left\{ |f_0(x) - 1|, |f_0^2(x) - f_0(x)| \right\} \\ &\leq |f_0(x) - 1| + |f_0^2(x) - f_0(x)| \\ &\leq f_0^2(x) + 2f_0(x) + 1 \\ &\leq M^2 + 2M + 1 \end{aligned}$$

As the right hand side is constant and hence clearly integrable over $[0, 1]$, we conclude that (A.13) holds, as the set where $f_0(x) = 0$ makes no contribution to the integral. $\square$

Proposition A.5. *Let P_f, P_g be probability measures on $\mathcal{X} \subseteq \mathbb{R}$ with densities f and g such that $P_g(\mathcal{A}) = 0 \implies P_f(\mathcal{A}) = 0$ for all measurable sets $\mathcal{A}$. Then the KL-divergence $K(f, g) = \int_{\mathcal{X}} f(x) \log \{f(x)/g(x)\} dx$ satisfies $K(f, g) \geq 0$ with equality if and only if $f = g$.*

Proof. Using the inequality $1 - 1/x \leq \log(x)$ for $x > 0$ from (A.14), we have

$$K(f, g) = \int_{\{g>0\}} f(x) \log \frac{f(x)}{g(x)} dx \geq \int_{\{g>0\}} f(x) \left\{ 1 - \frac{g(x)}{f(x)} \right\} dx = 1 - 1 = 0. \quad (\text{A.17})$$

If $f \neq g$ on a set of positive measure, then the above inequality becomes strict due to the strict inequality $1 - 1/x > \log(x)$ for $x \neq 1$ together with the strict monotonicity property of integrals. Finally, if $f = g$ a.e. then $\log\{f(x)/g(x)\} = 0$ a.e. which implies $K(f, g) = 0$. $\square$

Proof of Proposition 4.7

Proof. Recall that the pdf of a $\text{Dir}(a\mathbf{p}_0)$ -distribution is

$$p(\mathbf{p} | k) = \frac{\Gamma(a)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \prod_{j=1}^k p_j^{ap_{0,j}-1}. \quad (\text{A.18})$$

To show that $\mathcal{A}$ has prior probability tending to zero, note that for $a \leq 1$,

$$P(\mathbf{p} \in \mathcal{A} | k) = \frac{\Gamma(a)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \int_{\mathcal{A}} \prod_{j=1}^k p_j^{ap_{0,j}-1} d\mathbf{p} \leq \frac{\Gamma(a)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \delta^{-k} \int_{\mathcal{A}} d\mathbf{p}, \quad (\text{A.19})$$

since $p_j > \delta$ for all j on $\mathcal{A}$. The last two factors in (A.19) do not depend on a , so it suffices to show that the fraction approaches 0. To show establish this fact, we note that

$$\frac{\Gamma(a)}{\prod_{j=1}^k \Gamma(ap_{0,j})} \asymp \frac{\prod_{j=1}^k ap_{0,j}}{a} = a^{k-1} \prod_{j=1}^k p_{0,j} \rightarrow 0 \quad \text{as } a \rightarrow 0, \quad (\text{A.20})$$

and hence,

$$P(\mathbf{p} \in \mathcal{A} | k) = \int_{\mathcal{A}} p(\mathbf{p} | k) d\mathbf{p} \rightarrow 0, \quad \text{as } a \rightarrow 0, \quad (\text{A.21})$$

which was to be shown. $\square$

Lemma A.3. *Suppose the assumptions of Theorem 4.4 hold. Then the Bayes estimator,*

$$\hat{f}_{B,n}(x) = k_n \sum_{j=1}^{k_n} \hat{p}_{B,j} \mathcal{I}(x \in I_j), \quad x \in [0, 1],$$

satisfies

$$\mathbb{E}_{\mathbf{p}} \left\{ d_2^2(\hat{f}_{B,n}, f) \mid \mathbf{x}_{1:n} \right\} \leq C_2 \epsilon_n^2, \quad (\text{A.22})$$

$$\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_2^2(f_{0,k_n}, \hat{f}_{B,n}) \right\} \leq (4a_n^2 n^{-1} + 1) \epsilon_n^2. \quad (\text{A.23})$$

Proof. We start by showing (A.22). To this end, we note that

$$d_2^2(\hat{f}_{B,n}, f) = \sum_{j=1}^{k_n} \int_{I_j} (k_n \hat{p}_{B,j} - k_n p_j)^2 dx = k_n \sum_{j=1}^{k_n} (p_j - \hat{p}_{B,j})^2.$$

Taking posterior expectations on both sides, we obtain

$$\begin{aligned} \mathbb{E}_{\mathbf{p}} \left\{ d_2^2(\hat{f}_{B,n}, f) \mid \mathbf{x}_{1:n} \right\} &= k_n \sum_{j=1}^{k_n} \mathbb{E}_{p_j} \left\{ (p_j - \hat{p}_{B,j})^2 \mid \mathbf{x}_{1:n} \right\} \\ &= k_n \sum_{j=1}^{k_n} \text{Var}_{p_j} \{ p_j \mid \mathbf{x}_{1:n} \}. \end{aligned}$$

Since the posterior distribution of $\mathbf{p}$ is $\text{Dir}(a_n \mathbf{p}_0 + \mathbf{N})$, we can read off the variance using (3.6), yielding

$$\text{Var}_{p_j} (p_j \mid \mathbf{x}_{1:n}) = \frac{\hat{p}_{B,j}(1 - \hat{p}_{B,j})}{a_n + n + 1} \leq n^{-1} \hat{p}_{B,j}. \quad (\text{A.24})$$

Thus,

$$\mathbb{E}_{\mathbf{P}} \left\{ d_2^2(\widehat{f}_{B,n}, f) \mid \mathbf{x}_{1:n} \right\} \leq k_n \sum_{j=1}^{k_n} n^{-1} \widehat{p}_{B,j} = n^{-1} k_n.$$

Plugging in the rate of growth of k_n , we have that

$$n^{-1} k_n = C_2 n^{-1} \epsilon_n^{-1/\alpha} = C_2 n^{-1} n^{1/(2\alpha+1)} = C_2 n^{-2\alpha/(2\alpha+1)} = C_2 \epsilon_n^2,$$

so we conclude that $\mathbb{E}_{\mathbf{P}} \left\{ d_2^2(\widehat{f}_{B,n}, f) \mid \mathbf{x}_{1:n} \right\} = C_2 \epsilon_n^2$, which is (A.22) precisely. To control (A.23), we note that

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{1:n}}(\widehat{p}_{B,j}) &= \frac{a_n}{a_n + n} p_{0,j} + \frac{n}{a_n + n} \pi_j, \\ \text{Var}_{\mathbf{x}_{1:n}}(\widehat{p}_{B,j}) &= \text{Var}_{\mathbf{x}_{1:n}} \left(\frac{n}{a_n + n} \widehat{p}_j \right) = \frac{n}{(a_n + n)^2} \pi_j (1 - \pi_j) \leq \frac{\pi_j}{n}. \end{aligned}$$

Thus, the squared bias can be bounded by

$$\begin{aligned} \text{bias}^2(\pi_j, \widehat{p}_{B,j}) &= \left(\pi_j - \frac{a_n}{a_n + n} p_{0,j} - \frac{n}{a_n + n} \pi_j \right)^2 = \frac{a_n^2}{(a_n + n)^2} (\pi_j - p_{0,j})^2 \\ &\leq \frac{2a_n^2}{n^2} \{p_{0,j} + \pi_j\}. \end{aligned}$$

A risk calculation then reveals

$$\mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_2^2(f_{0,k_n}, \widehat{f}_{B,n}) \right\} = \sum_{j=1}^{k_n} \int_{I_j} \mathbb{E}_{\mathbf{x}_{1:n}} (k_n \pi_j - k_n \widehat{p}_{B,j})^2 ds = k_n \sum_{j=1}^{k_n} \mathbb{E}_{\mathbf{x}_{1:n}} (\pi_j - \widehat{p}_{B,j})^2.$$

A bias-variance decomposition of the last term yields

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{1:n}} \left\{ d_2^2(f_{0,k_n}, \widehat{f}_{B,n}) \right\} &= k_n \sum_{j=1}^{k_n} \left\{ \text{bias}^2(\pi_j, \widehat{p}_{B,j}) + \text{Var}_{\mathbf{x}_{1:n}}(\widehat{p}_{B,j}) \right\} \\ &\leq k_n \sum_{j=1}^{k_n} \left\{ \frac{2a_n^2}{n^2} (p_{0,j} + \pi_j) + \frac{\pi_j}{n} \right\} \\ &= 4a_n^2 n^{-2} k_n + n^{-1} k_n \\ &= 4a_n^2 n^{-1} \epsilon_n^2 + \epsilon_n^2, \end{aligned} \tag{A.25}$$

establishing (A.23). □

Proposition A.6. *Let $\{a_n\}_{n=1}^\infty, \{b_n\}_{n=1}^\infty$ be two non-negative sequences of real numbers. Then*

$$\liminf_{n \rightarrow \infty} a_n b_n \geq (\liminf_{n \rightarrow \infty} a_n) (\liminf_{n \rightarrow \infty} b_n). \tag{A.26}$$

Proof. Since the numbers are nonnegative we have that

$$a_n b_n \geq a_n \inf_{m \geq n} b_m \geq \left(\inf_{m \geq n} a_m \right) \left(\inf_{m \geq n} b_m \right).$$

Hence,

$$\inf_{m \geq n} a_m b_m \geq \left(\inf_{m \geq n} a_m \right) \left(\inf_{m \geq n} b_m \right).$$

Using continuity of the product, (A.26) follows by taking the limit as $n \rightarrow \infty$. □

Proposition A.7. *Let $a, b \in \mathbb{R}$ and $\gamma > 0$. Then $(a + b)^2 \leq (1 + \gamma)a^2 + (1 + \gamma^{-1})b^2$. In particular, $(a + b)^2 \leq 2a^2 + 2b^2$*

Proof. We start with $0 \leq (\gamma^{1/2}a - \gamma^{-1/2}b)^2 = \gamma a^2 - 2ab + \gamma^{-1}b^2$. Rearranging yields $2ab \leq \gamma a^2 + \gamma^{-1}b^2$, and hence

$$(a + b)^2 = a^2 + 2ab + b^2 \leq (1 + \gamma)a^2 + (1 + \gamma^{-1})b^2$$

Plugging in $\gamma = 1$ proves the second part. $\square$

Lemma A.4. *Let $\{h_k\}$ be a sequence of functions from $\mathcal{X}$ to $\mathbb{R}$ converging in the supremum metric to some h satisfying $\|h\|_\infty < \infty$ and $\{a_k\}$ a sequence of real numbers converging to $a \neq 0$. Then h_k/a_k converges to h/a in the supremum metric as well.*

Proof. A simple calculation shows that

$$|h_k(t)/a_k - h(t)/a| \leq |h_k(t)| |1/a_k - 1/a| + |a^{-1}| |h_k(t) - h(t)|. \quad (\text{A.27})$$

To show that the first term goes to 0, we note that since $h_k \rightarrow h$ uniformly, the reverse triangle inequality implies that $\|h_k\|_\infty \leq \|h\|_\infty + 1$ for all sufficiently large k and we obtain

$$|h_k(t)/a_k - h(t)/a| \leq (\|h\|_\infty + 1) |1/a_k - 1/a| + |a^{-1}| \|h_k - h\|_\infty. \quad (\text{A.28})$$

Both of these terms converge to zero uniformly in k by assumption, which proves the claim. $\square$

Proposition A.8. *Let $f(x | k, \mathbf{p}) \propto \prod_{j=1}^k p_j^{\mathcal{I}(x \in I_j)}$. Then the Jeffreys prior for $\mathbf{p} | k$ is proportional to $\prod_{j=1}^k p_j^{-1/2}$.*

Proof. From the definition of f we see that $\log f(x | k, \mathbf{p})$ equals $\sum_{j=1}^k \mathcal{I}(x \in I_j) \log p_j$ up to an additive constant. Differentiating this expression twice, we obtain

$$\frac{\partial^2}{\partial p_j^2} \log f(x | k, \mathbf{p}) = -\frac{\mathcal{I}(x \in I_j)}{p_j^2},$$

while the mixed partial derivatives are 0. Since $\mathbb{E}_x[\mathcal{I}(x \in I_j)] = \mathbb{P}(x \in I_j) = p_j$, it follows

$$-\mathbb{E}_x \left[\frac{\partial^2}{\partial p_j^2} \log f(x | k, \mathbf{p}) \right] = \frac{1}{p_j}.$$

Thus, the Fisher information matrix is given by $\mathbf{I}(\mathbf{p} | k) = \text{diag}\{1/p_1, 1/p_2, \dots, 1/p_k\}$, from which we derive that the Jeffreys prior for $\mathbf{p} | k$ is

$$p(\mathbf{p} | k) \propto \sqrt{\det[\mathbf{I}(\mathbf{p} | k)]} = \sqrt{\prod_{j=1}^k p_j^{-1}} = \prod_{j=1}^k p_j^{-1/2}, \quad (\text{A.29})$$

which was to be shown. $\square$

Appendix B

Appendix B

This Appendix contains results from other works that are used in the proofs. Some of the results stated here are important theorems, others are estimates of probabilities and covering numbers

Prerequisites

Theorem B.1 (Dominated Convergence Theorem, slight rewrite of Theorem 7.6.5 in Lindstrøm (2017)). *Let $g: \mathcal{X} \rightarrow \mathbb{R}$ be a Lebesgue integrable function and let $\{f_k\}_{k=1}^\infty$ be a sequence of measurable functions converging to f almost everywhere. Then if $|f_k(x)| \leq g(x)$ for all $x \in \mathcal{X}$ and all $k \in \mathbb{N}$,*

$$\lim_{k \rightarrow \infty} \int_{\mathcal{X}} f_k(x) dx = \int_{\mathcal{X}} f(x) dx.$$

Theorem B.2 (Law of Large Numbers, Theorem 2.4.1 in Durrett (2019)). *Let $\{X_n\}_{n=1}^\infty$ be a sequence of iid random variables satisfying $\mathbb{E}_{X_1}|X_1| < \infty$. Then*

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} \mathbb{E}_{X_1}\{X_1\}.$$

Some useful estimates

Lemma B.1 (Lemma 4 in Genovese and Wasserman (2000)). *Suppose that $l_1, \dots, l_k$ and $u_1, \dots, u_k$ are nonnegative integrable functions and let $a_1, \dots, a_k$ and $b_1, \dots, b_k$ be nonnegative real numbers. Then for $l = \sum_{j=1}^k a_j l_j$ and $u = \sum_{j=1}^k b_j u_j$, we have that*

$$d_H^2(l, u) \leq \sum_{j=1}^k d_H^2(a_j l_j, b_j u_j). \quad (\text{B.1})$$

Lemma B.2 (Lemma 2 in Genovese and Wasserman (2000)). *For any $\epsilon \leq 1$ the covering number of the simplex $\mathcal{S}_{k-1}$ satisfies*

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

Lemma B.3 (Lemma A.1 in Ghosal (2001)). *Let $\mathbf{p}$ follow a k -dimensional $\text{Dir}(a\mathbf{p}_0)$ -distribution. Suppose that $A\epsilon^b \leq ap_{0,j} \leq M$ for some constants $A, M, b > 0$. Let*

$\pi \in \mathcal{S}_{k-1}$, and suppose that $\epsilon \leq 1/(Mk)$. Then there exists constants c_1, c_2 depending only on A, M and a such that¹

$$\mathbb{P}\left(\mathbf{p} \in \mathcal{S}_{k-1} : \sum_{j=1}^k |p_j - \pi_j| \leq 2\epsilon, \min_j p_j > \epsilon^4\right) \geq c_1 \exp(-c_2 k \log(1/\epsilon)). \quad (\text{B.3})$$

Lemma B.4 (Theorem 5 in Wong and Shen (1995)). *Let $f, g: \mathcal{X} \rightarrow \mathbb{R}$ be two densities satisfying $d_H^2(f, g) \leq \epsilon^2$ and suppose that $L_\delta^2 = \int_{f/g \geq e^{1/\delta}} f^{1+\delta}(t)/g^\delta(t) dt$ is finite for some $\delta \in (0, 1]$. Then for every $\epsilon^2 \leq \frac{1}{2}(1 - e^{-1})^2$ it holds that*

$$\begin{aligned} \int_{\mathcal{X}} f(t) \log\left(\frac{f(t)}{g(t)}\right) dt &= \mathcal{O}\left(\epsilon^2 \max\{1, \log(L_\delta/\epsilon)\}\right) \\ \int_{\mathcal{X}} f(t) \log^2\left(\frac{f(t)}{g(t)}\right) dt &= \mathcal{O}\left(\epsilon^2 \max\{1, \log^2(L_\delta/\epsilon)\}\right) \end{aligned} \quad (\text{B.4})$$

Lemma B.5 (Log-sum inequality, Theorem 2.7.1 in Cover and Thomas (2006)). *For nonnegative numbers $a_1, \dots, a_k, b_1, \dots, b_k$ we have that*

$$\sum_{j=1}^k a_j \log \frac{a_j}{b_j} \geq a \log \frac{a}{b}, \quad (\text{B.5})$$

where $a = \sum_{j=1}^k a_j$ and $b = \sum_{j=1}^k b_j$.

¹The original statement of the Lemma states that the lower bound depends on m . However, inspection of the proof (see e.g. the proof of Lemma G.13 in Ghosal and van der Vaart (2017)) implies that we can find a bound independent of m since $\Gamma(m)$ is bounded away from 0 for all positive m .

Appendix C

Appendix C

C.1 Test densities

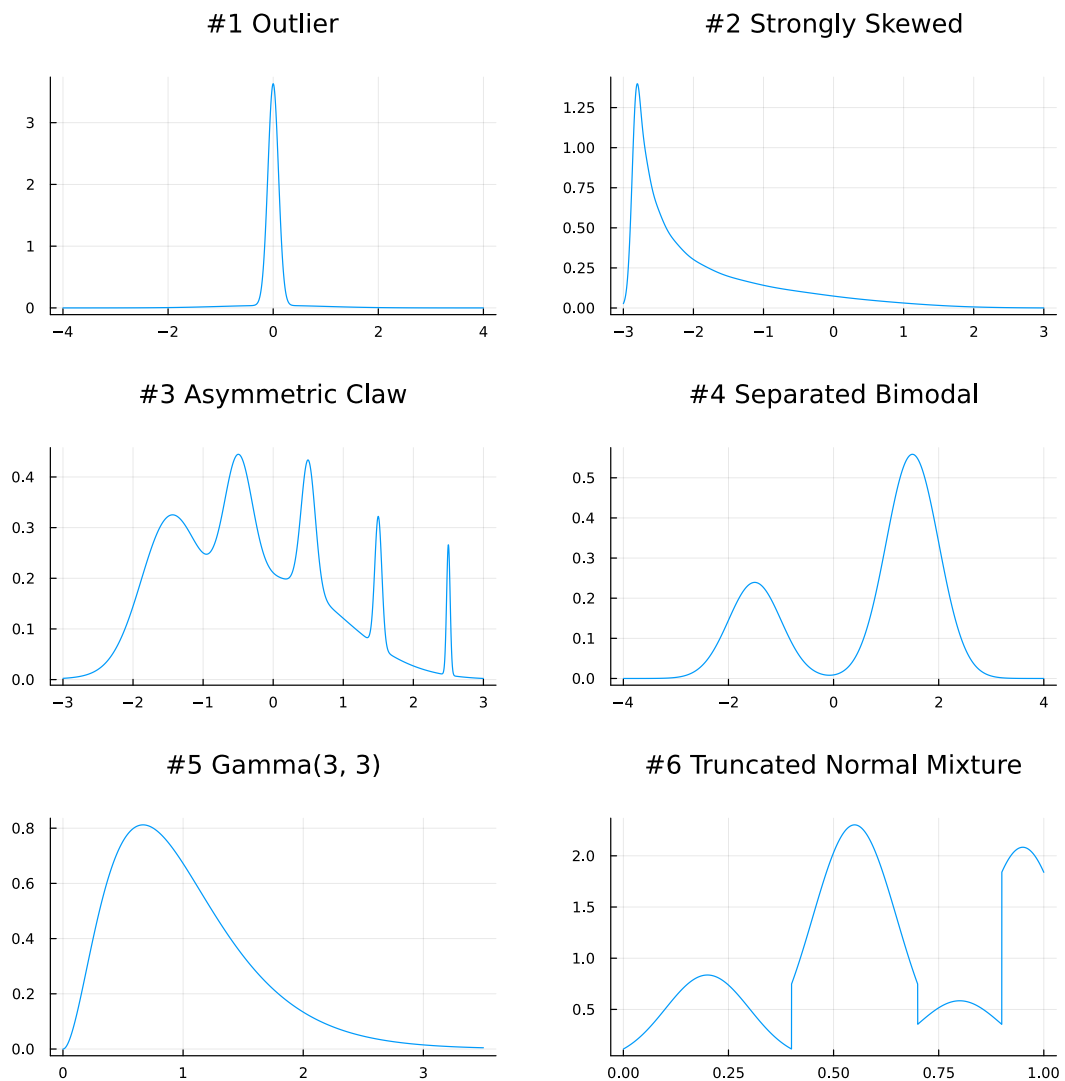


Figure C.1: Distributions used for some preliminary simulations.

C.2 Results from simulations: Regular Histograms

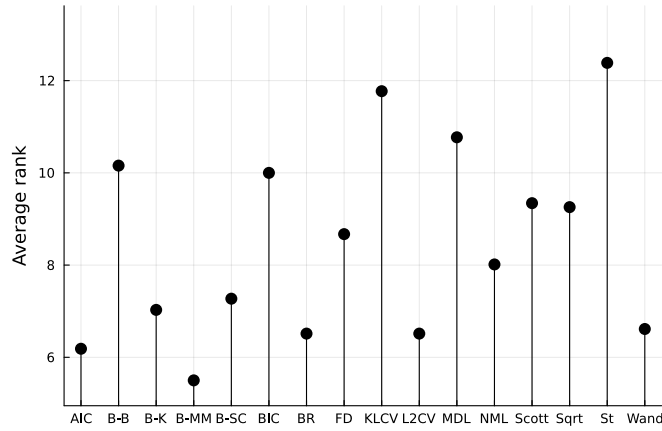


Figure C.2: Overall average ranks, $\mathbb{L}_2$ loss.

Table C.1: Average ranks by sample size, $\mathbb{L}_2$ loss. Lowest average rank for each sample size is shown in boldface.

n	50	250	1000	5000	25000
AIC	8.93	7.0	5.14	4.93	4.93
B-B	8.64	11.1	11.3	9.93	9.79
B-K	7.0	5.71	7.36	7.43	7.64
B-MM	4.43	4.79	5.29	6.5	6.5
B-SC	6.86	5.36	7.29	8.36	8.5
BIC	8.64	10.5	11.1	10.2	9.5
BR	8.43	8.21	5.86	5.0	5.07
FD	8.07	8.79	8.57	9.07	8.86
KLCV	10.6	11.7	12.1	11.7	12.6
L2CV	9.29	7.0	6.14	5.0	5.14
MDL	13.9	11.1	8.93	9.36	10.6
NML	8.57	7.29	8.5	7.93	7.79
Scott	8.0	9.29	9.43	10.2	9.79
Sqrt	9.14	10.1	9.29	9.36	8.43
St	8.43	10.4	13.3	14.7	15.1
Wand	7.07	7.64	6.36	6.29	5.71

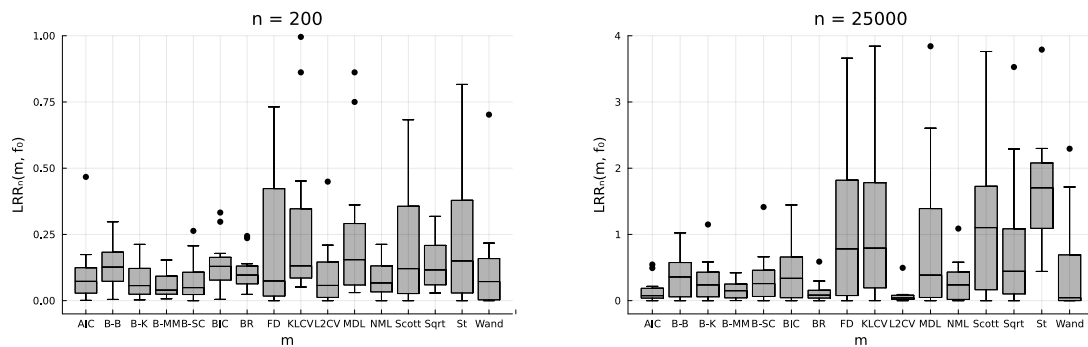


Figure C.3: Boxplots of log-relative $\mathbb{L}_2$ risk by method for sample sizes $n = 200$ (left) and $n = 25000$ (right).

Appendix C. Appendix C

Table C.2: Estimated $\mathbb{L}_2$ risks. Note that no estimate is provided for densities 3 and 6 of Figure 4.4 as they are not in $\mathbb{L}_2(\mathbb{R})$.

Density	n	AIC	B-B	B-K	B-MM	B-SC	BIC	BR	FD	KLCV	L2CV	MDL	NML	Scott	Sqrt	St	Wand
1	50	0.193	0.2	0.195	0.179	0.187	0.195	0.191	0.172	0.181	0.2	0.223	0.199	0.163	0.181	0.175	0.166
	200	0.125	0.13	0.122	0.118	0.12	0.128	0.123	0.112	0.122	0.131	0.129	0.123	0.107	0.119	0.107	0.107
	1000	0.07	0.083	0.074	0.071	0.074	0.081	0.072	0.065	0.079	0.074	0.071	0.075	0.063	0.071	0.074	0.063
	5000	0.04	0.052	0.046	0.043	0.047	0.051	0.042	0.039	0.056	0.042	0.044	0.046	0.037	0.044	0.06	0.037
	25000	0.023	0.032	0.028	0.026	0.029	0.032	0.024	0.023	0.046	0.024	0.031	0.028	0.022	0.028	0.055	0.022
2	50	0.299	0.196	0.205	0.234	0.232	0.203	0.211	0.309	0.246	0.282	0.466	0.202	0.303	0.417	0.389	0.309
	200	0.154	0.097	0.097	0.112	0.101	0.097	0.109	0.189	0.145	0.151	0.132	0.096	0.184	0.28	0.218	0.195
	1000	0.065	0.042	0.042	0.045	0.043	0.042	0.047	0.106	0.065	0.065	0.043	0.042	0.104	0.18	0.107	0.117
	5000	0.03	0.019	0.019	0.019	0.019	0.019	0.021	0.06	0.03	0.03	0.019	0.019	0.06	0.12	0.053	0.071
	25000	0.013	0.008	0.008	0.008	0.008	0.008	0.009	0.035	0.013	0.013	0.008	0.008	0.035	0.08	0.026	0.045
4	50	0.227	0.24	0.229	0.214	0.236	0.257	0.257	0.195	0.374	0.216	0.299	0.235	0.238	0.217	0.228	0.199
	200	0.162	0.183	0.177	0.16	0.187	0.2	0.183	0.143	0.388	0.152	0.303	0.175	0.163	0.163	0.22	0.148
	1000	0.146	0.158	0.157	0.144	0.167	0.172	0.152	0.115	0.448	0.126	0.364	0.156	0.14	0.147	0.277	0.101
	5000	0.121	0.154	0.154	0.109	0.159	0.16	0.137	0.078	0.482	0.078	0.413	0.154	0.138	0.145	0.318	0.061
	25000	0.064	0.103	0.117	0.056	0.152	0.157	0.067	0.046	0.512	0.04	0.453	0.11	0.106	0.131	0.368	0.037
5	50	0.194	0.203	0.193	0.187	0.19	0.202	0.202	0.179	0.215	0.196	0.197	0.199	0.179	0.176	0.177	0.177
	200	0.127	0.152	0.138	0.13	0.139	0.152	0.143	0.114	0.177	0.127	0.139	0.14	0.121	0.116	0.143	0.113
	1000	0.08	0.108	0.096	0.086	0.1	0.108	0.09	0.069	0.168	0.078	0.118	0.096	0.077	0.073	0.143	0.069
	5000	0.05	0.074	0.065	0.055	0.069	0.076	0.054	0.041	0.156	0.046	0.104	0.065	0.047	0.044	0.138	0.041
	25000	0.03	0.049	0.043	0.035	0.047	0.051	0.032	0.024	0.153	0.026	0.1	0.043	0.028	0.025	0.142	0.024
7	50	0.297	0.29	0.289	0.282	0.283	0.287	0.287	0.28	0.282	0.306	0.298	0.294	0.276	0.291	0.284	0.279
	200	0.214	0.233	0.224	0.217	0.22	0.233	0.229	0.233	0.238	0.212	0.222	0.227	0.237	0.224	0.238	0.237
	1000	0.133	0.181	0.142	0.139	0.143	0.173	0.139	0.167	0.206	0.132	0.173	0.143	0.188	0.148	0.226	0.169
	5000	0.08	0.121	0.101	0.092	0.105	0.119	0.085	0.108	0.189	0.077	0.149	0.101	0.127	0.085	0.221	0.082
	25000	0.047	0.079	0.063	0.056	0.067	0.077	0.049	0.065	0.178	0.045	0.134	0.063	0.079	0.047	0.221	0.043
8	50	0.205	0.197	0.197	0.188	0.196	0.196	0.194	0.178	0.19	0.208	0.233	0.2	0.178	0.193	0.186	0.18
	200	0.133	0.138	0.129	0.125	0.128	0.134	0.131	0.117	0.132	0.137	0.136	0.131	0.118	0.124	0.118	0.118
	1000	0.078	0.092	0.083	0.079	0.083	0.09	0.08	0.071	0.089	0.082	0.08	0.084	0.074	0.076	0.09	0.071
	5000	0.045	0.059	0.052	0.048	0.052	0.057	0.046	0.043	0.068	0.046	0.052	0.052	0.044	0.046	0.076	0.042
	25000	0.026	0.037	0.032	0.029	0.033	0.036	0.027	0.025	0.058	0.027	0.039	0.032	0.026	0.029	0.072	0.025
9	50	0.211	0.198	0.203	0.195	0.203	0.2	0.201	0.186	0.199	0.216	0.248	0.202	0.185	0.201	0.195	0.188
	200	0.141	0.143	0.136	0.132	0.135	0.14	0.136	0.129	0.137	0.145	0.151	0.137	0.129	0.136	0.128	0.129
	1000	0.092	0.1	0.094	0.092	0.093	0.099	0.093	0.09	0.094	0.094	0.092	0.095	0.091	0.087	0.099	0.089
	5000	0.059	0.074	0.065	0.062	0.065	0.072	0.06	0.065	0.072	0.059	0.064	0.065	0.066	0.057	0.088	0.06
	25000	0.039	0.052	0.047	0.044	0.047	0.051	0.04	0.046	0.061	0.038	0.049	0.047	0.047	0.037	0.083	0.04
10	50	0.152	0.135	0.138	0.142	0.144	0.136	0.137	0.145	0.143	0.15	0.167	0.137	0.145	0.158	0.153	0.146
	200	0.116	0.128	0.118	0.116	0.11	0.127	0.124	0.13	0.116	0.116	0.115	0.122	0.13	0.129	0.135	0.131
	1000	0.07	0.09	0.07	0.071	0.068	0.082	0.07	0.129	0.07	0.071	0.07	0.072	0.129	0.079	0.128	0.123
	5000	0.042	0.053	0.042	0.043	0.042	0.049	0.041	0.107	0.041	0.042	0.041	0.043	0.108	0.043	0.12	0.069
	25000	0.025	0.031	0.025	0.025	0.025	0.029	0.024	0.075	0.025	0.025	0.024	0.025	0.075	0.024	0.113	0.027
11	50	0.309	0.311	0.31	0.285	0.304	0.308	0.301	0.265	0.287	0.315	0.389	0.315	0.256	0.298	0.285	0.256
	200	0.193	0.196	0.185	0.176	0.182	0.192	0.185	0.168	0.185	0.2	0.233	0.187	0.164	0.2	0.165	0.162
	1000	0.11	0.121	0.111	0.106	0.108	0.117	0.108	0.101	0.109	0.113	0.107	0.111	0.098	0.129	0.098	0.098
	5000	0.065	0.072	0.065	0.062	0.064	0.069	0.062	0.059	0.065	0.065	0.063	0.065	0.058	0.085	0.067	0.057
	25000	0.038	0.042	0.038	0.037	0.038	0.041	0.036	0.035	0.038	0.037	0.037	0.038	0.034	0.056	0.053	0.034
12	50	0.303	0.301	0.301	0.296	0.307	0.323	0.32	0.414	0.454	0.298	0.454	0.302	0.427	0.349	0.351	0.336
	200	0.199	0.203	0.202	0.195	0.206	0.216	0.205	0.396	0.452	0.191	0.452	0.202	0.378	0.262	0.348	0.214
	1000	0.119	0.128	0.131	0.113	0.155	0.178	0.126	0.379	0.452	0.109	0.452	0.129	0.368	0.222	0.298	0.129
	5000	0.071	0.086	0.088	0.068	0.098	0.106	0.074	0.3	0.453	0.057	0.453	0.086	0.29	0.1	0.278	0.067
	25000	0.041	0.054	0.055	0.041	0.063	0.067	0.043	0.248	0.453	0.033	0.45	0.054	0.262	0.049	0.257	0.034
13	50	0.145	0.153	0.143	0.143	0.142	0.152	0.152	0.16	0.157	0.144	0.15	0.146	0.156	0.149	0.155	0.157
	200	0.119	0.134	0.119	0.12	0.119	0.133	0.124	0.142	0.136	0.116	0.131	0.12	0.154	0.144	0.139	0.141
	1000	0.07	0.095	0.076	0.078	0.076	0.095	0.073	0.145	0.116	0.066	0.099	0.077	0.133	0.099	0.133	0.096
	5000	0.038	0.057	0.048	0.048	0.048	0.056	0.04	0.127	0.106	0.034	0.083	0.048	0.131	0.068	0.143	0.058
	25000	0.022	0.042	0.031	0.03	0.031	0.039	0.023	0.113	0.097	0.02	0.072	0.031	0.106	0.037	0.147	0.024
14	50	1.123	1.123	1.123	1.123	1.124	1.123	1.123	1.146	1.161	1.123	1.161	1.123	1.158	1.153	1.148	1.131
	200	1.094	1.096	1.094	1.095	1.094	1.095	1.094	1.128	1.151	1.092	1.151	1.094	1.149	1.116	1.141	1.069
	1000	0.999	0.999	0.999	0.999	1.0	0.999	0.999	1.108	1.183	0.996	1.183	0.999	1.149	1.079	1.12	0.913
	5000	0.631	0.631	0.631	0.631	0.632	0.631	0.631	1.07	1.196	0.631	1.196	0.631	1.144	1.023	1.146	0.764
	25000	0.026	0.026	0.027	0.027	0.031	0.026	0.026	0.998	1.196	0.026	1.196	0.026	1.104	0.873	1.136	0.254
15	50	0.697	0.793	0.688	0.674	0.662	0.777	0.777	0.922	0.817	0.685	0.727	0.718	0.914	0.765	0.768	0.778
	200	0.467	0.498	0.457	0.453	0.448	0.483	0.466	0.737	0.516	0.462	0.473	0.463	0.722	0.55	0.686	0.557
	1000	0.275	0.286	0.267	0.269	0.267	0.282	0.267	0.299								

C.3 Results from simulations: Irregular Histograms

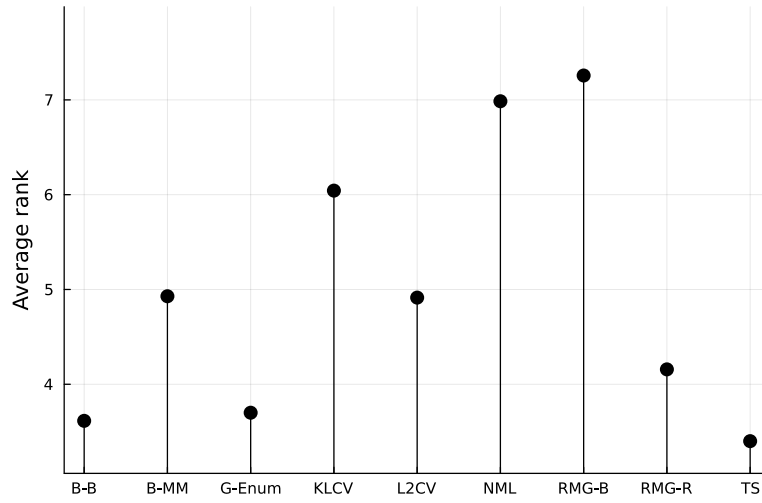


Figure C.4: Overall average ranks, $\mathbb{L}_2$ loss.

Table C.3: Average ranks by sample size, $\mathbb{L}_2$ loss. Lowest average rank for each sample size is shown in boldface.

n	50	200	1000	5000	25000
B-B	2.93	2.79	3.29	4.0	5.07
B-MM	3.21	4.14	5.21	5.43	6.64
G-Enum	4.64	4.0	3.0	3.36	3.5
KLCV	5.0	6.0	7.07	6.5	5.64
L2CV	3.93	4.93	5.57	5.21	4.93
NML	8.79	7.57	6.5	6.14	5.93
RMG-B	7.79	7.79	7.64	7.36	5.71
RMG-R	3.57	3.86	4.14	4.36	4.86
TS	5.14	3.93	2.57	2.64	2.71

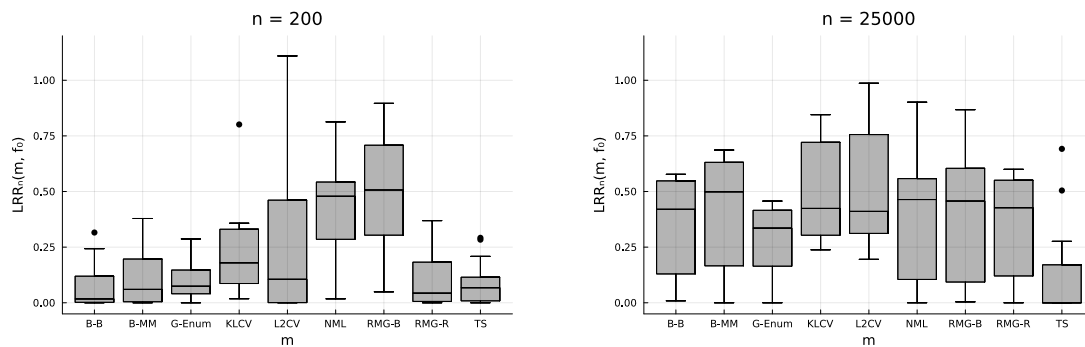


Figure C.5: Boxplots of log-relative $\mathbb{L}_2$ risk by method for sample sizes $n = 200$ (left) and $n = 25000$ (right).

Table C.4: Estimated $\mathbb{L}_2$ risks. Note that no estimate is provided for densities 3 and 6 of Figure 4.4 as they are not in $\mathbb{L}_2(\mathbb{R})$.

Density	n	B-B	B-MM	G-Enum	KLCV	L2CV	NML	RMG-B	RMG-R	TS
1	50	0.23	0.233	0.247	0.192	0.167	0.481	0.421	0.232	0.244
	200	0.149	0.156	0.147	0.147	0.12	0.208	0.208	0.155	0.123
	1000	0.095	0.102	0.089	0.103	0.091	0.104	0.122	0.098	0.069
	5000	0.06	0.065	0.054	0.066	0.059	0.062	0.063	0.061	0.039
	25000	0.038	0.041	0.033	0.038	0.035	0.039	0.039	0.037	0.022
2	50	0.195	0.19	0.193	0.356	0.346	0.337	0.424	0.19	0.254
	200	0.094	0.094	0.096	0.318	0.285	0.151	0.23	0.094	0.116
	1000	0.043	0.042	0.043	0.242	0.209	0.045	0.053	0.043	0.051
	5000	0.019	0.019	0.019	0.141	0.123	0.019	0.02	0.019	0.025
	25000	0.008	0.008	0.008	0.074	0.065	0.008	0.008	0.008	0.01
4	50	0.249	0.262	0.251	0.265	0.243	0.582	0.408	0.267	0.218
	200	0.175	0.185	0.175	0.176	0.162	0.229	0.242	0.187	0.137
	1000	0.12	0.132	0.12	0.139	0.139	0.14	0.142	0.132	0.111
	5000	0.076	0.084	0.071	0.107	0.115	0.075	0.074	0.086	0.063
	25000	0.048	0.052	0.044	0.051	0.058	0.047	0.046	0.055	0.037
5	50	0.229	0.235	0.239	0.202	0.181	0.537	0.417	0.236	0.225
	200	0.161	0.171	0.156	0.135	0.117	0.199	0.19	0.17	0.131
	1000	0.107	0.116	0.098	0.084	0.078	0.168	0.187	0.112	0.076
	5000	0.069	0.074	0.061	0.054	0.053	0.07	0.07	0.071	0.044
	25000	0.043	0.047	0.037	0.033	0.033	0.043	0.043	0.044	0.025
7	50	0.292	0.292	0.313	0.286	0.286	0.542	0.406	0.29	0.338
	200	0.221	0.219	0.226	0.236	0.229	0.24	0.241	0.219	0.245
	1000	0.175	0.201	0.156	0.146	0.142	0.194	0.217	0.186	0.133
	5000	0.108	0.112	0.1	0.089	0.087	0.115	0.12	0.109	0.071
	25000	0.068	0.074	0.061	0.054	0.054	0.068	0.073	0.068	0.039
8	50	0.195	0.2	0.207	0.2	0.183	0.283	0.243	0.199	0.207
	200	0.137	0.142	0.14	0.152	0.129	0.168	0.175	0.14	0.132
	1000	0.095	0.102	0.089	0.106	0.094	0.1	0.1	0.099	0.083
	5000	0.059	0.062	0.054	0.066	0.058	0.063	0.064	0.059	0.039
	25000	0.038	0.042	0.033	0.038	0.035	0.054	0.052	0.038	0.022
9	50	0.2	0.195	0.197	0.206	0.192	0.318	0.264	0.195	0.218
	200	0.144	0.148	0.152	0.159	0.141	0.228	0.324	0.146	0.153
	1000	0.107	0.114	0.1	0.115	0.103	0.114	0.12	0.11	0.088
	5000	0.074	0.077	0.068	0.076	0.07	0.08	0.08	0.074	0.058
	25000	0.05	0.053	0.044	0.045	0.044	0.053	0.052	0.049	0.036
10	50	0.144	0.138	0.134	0.152	0.148	0.319	0.271	0.135	0.194
	200	0.129	0.128	0.127	0.137	0.134	0.13	0.134	0.129	0.169
	1000	0.089	0.099	0.086	0.083	0.082	0.134	0.142	0.09	0.082
	5000	0.06	0.068	0.055	0.051	0.05	0.073	0.077	0.06	0.043
	25000	0.039	0.043	0.034	0.033	0.033	0.041	0.044	0.038	0.024
11	50	0.333	0.331	0.326	0.303	0.262	0.573	0.529	0.328	0.342
	200	0.222	0.232	0.216	0.245	0.203	0.441	0.487	0.229	0.203
	1000	0.152	0.161	0.138	0.184	0.161	0.181	0.184	0.155	0.118
	5000	0.097	0.104	0.086	0.117	0.104	0.106	0.11	0.098	0.071
	25000	0.061	0.065	0.052	0.068	0.063	0.062	0.06	0.06	0.043
12	50	0.231	0.238	0.25	0.394	0.394	0.636	0.291	0.239	0.248
	200	0.182	0.198	0.185	0.356	0.356	0.293	0.272	0.196	0.16
	1000	0.117	0.128	0.112	0.146	0.144	0.125	0.136	0.124	0.085
	5000	0.074	0.081	0.069	0.066	0.065	0.079	0.093	0.077	0.048
	25000	0.047	0.051	0.041	0.035	0.035	0.047	0.046	0.048	0.027
13	50	0.143	0.151	0.157	0.157	0.161	0.474	0.273	0.152	0.16
	200	0.09	0.097	0.095	0.129	0.146	0.203	0.192	0.095	0.091
	1000	0.062	0.067	0.06	0.092	0.101	0.073	0.087	0.064	0.046
	5000	0.039	0.044	0.036	0.042	0.042	0.057	0.058	0.04	0.025
	25000	0.025	0.028	0.022	0.02	0.02	0.03	0.03	0.025	0.014
14	50	0.519	0.542	0.542	1.164	1.17	2.479	1.084	0.788	0.713
	200	0.261	0.261	0.271	1.165	1.17	0.385	0.357	0.262	0.349
	1000	0.117	0.114	0.132	1.089	1.105	0.153	0.159	0.114	0.158
	5000	0.044	0.041	0.048	1.087	1.09	0.047	0.05	0.041	0.071
	25000	0.018	0.016	0.023	0.064	0.454	0.016	0.017	0.016	0.032
15	50	0.773	0.81	0.904	0.821	0.857	1.93	1.17	0.815	0.819
	200	0.392	0.405	0.451	0.402	0.579	0.551	0.53	0.403	0.419
	1000	0.173	0.174	0.199	0.275	0.369	0.201	0.203	0.172	0.221
	5000	0.09	0.089	0.101	0.173	0.195	0.091	0.115	0.089	0.114
	25000	0.044	0.048	0.046	0.094	0.101	0.046	0.043	0.043	0.054
16	50	0.762	0.799	0.848	1.013	1.031	2.302	1.204	0.795	0.825
	200	0.477	0.479	0.514	0.485	0.668	0.542	0.54	0.479	0.511
	1000	0.233	0.253	0.246	0.309	0.445	0.259	0.259	0.241	0.261
	5000	0.108	0.101	0.113	0.18	0.225	0.12	0.122	0.099	0.14
	25000	0.044	0.043	0.049	0.1	0.115	0.043	0.044	0.043	0.071

Appendix D

Appendix D

D.1 Diagnostics for MCMC

The material in this section draws on Givens and Hoeting (2012, Chapter 7.3), to which we refer readers looking for a more detailed account of convergence diagnostics of MCMC output.

A key issue in MCMC that greatly influences the quality of the resulting inference that can be drawn from such a chain is that of *autocorrelation*. Crucially, the variance of the estimate $\hat{h} = \frac{1}{T} \sum_{t=1}^T h(X_t)$ depends on the correlations between $h(X_t)$ and $h(X_{t+s})$ for positive s . When $\hat{h}$ is estimated from MCMC output, these correlations will typically be positive, leading to higher variance compared to a corresponding estimate based on an iid sample of the same size, with smaller correlations leading to less of a difference between the variances of the two estimates. Given a sequence of random variables $\{X_t\}$ such that X_t has the same marginal distribution for each t , we define the *autocorrelation function of X_t at lag $s \in \mathbb{N} \cup \{0\}$* to be

$$\rho(s) = \frac{\text{Cov}(X_t, X_{t+s})}{\text{Var}(X_t)}. \quad (\text{D.1})$$

For a key quantity of interest, it is thus desirable to have an autocorrelation function decaying quickly to 0 as s increases. To check whether this is indeed the case or not, so-called *autocorrelation plots* are commonly employed. These plot the estimated autocorrelation for different values of the lag s .

A particularly useful and simple tool to diagnose MCMC output are that of *traceplots*. A traceplot plots the iteration number against the value of a given parameter. Ideally, the parameter should move away from its starting value quickly (if it is not already in a region of high probability under the stationary distribution) and explore the support of the posterior quickly. In particular, if the parameter value changes little from iteration to iteration, then the exploration of the stationary distribution will likely be too slow, leading to highly autocorrelated output.

It is in general difficult to verify that a Markov chain has been able to explore the stationary distribution to an adequate extent. This is especially difficult in settings where the target distribution has more than one mode that are well-separated, as jumping between two such modes can be difficult in practice. Therefore, it is common to run multiple chains with different starting values to improve confidence that all chains have explored the high-probability regions of interest. If the output from the resulting chains do not differ much between the chains, one can be more confident that the chains have been able to adequately explore the stationary distribution.

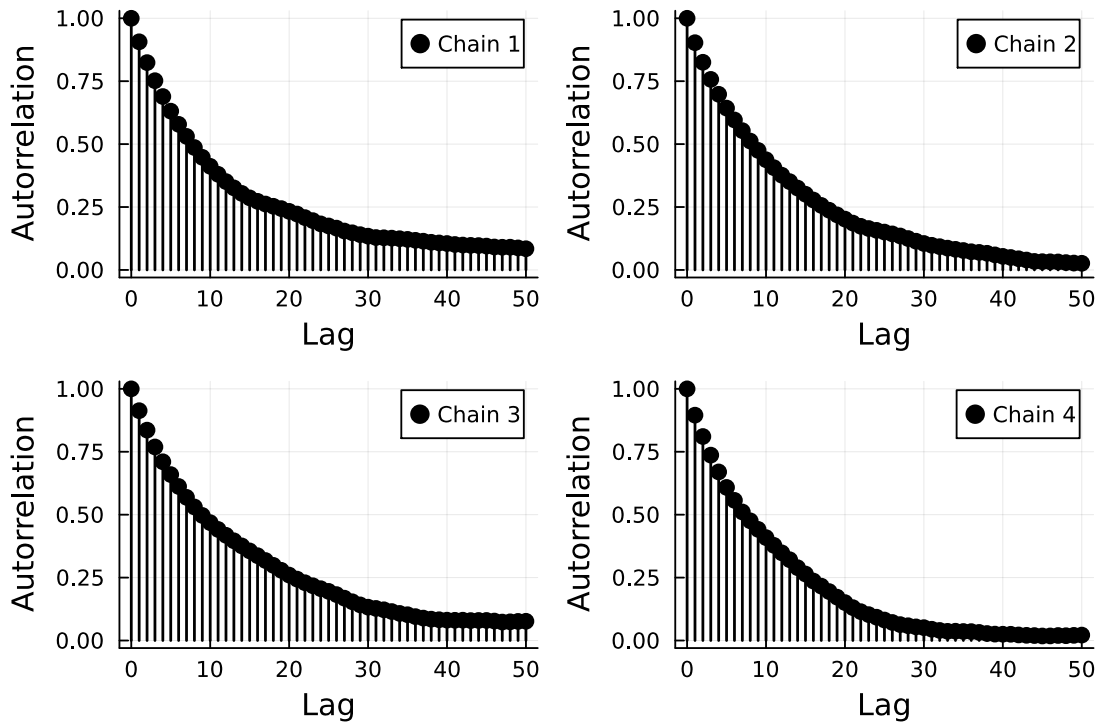


Figure D.1: Autocorrelation functions for each of the 4 chains from the example in Section 6.2.2

The theory for Markov chains only guarantee that the Markov Chain has the posterior distributions as a limiting distribution. In particular, for smaller values of t , the distribution of X_t can be quite different from the stationary distribution depending on the starting value X_0 . To circumvent this issue in practice, a *burn-in period* is commonly employed, where one runs the chain from a given starting point X_0 , discards the first D samples, and uses the remaining samples $X_D, X_{D+1}, \dots, X_T$ for inferences about the stationary distribution. To determine what burn-in length is appropriate, traceplots can be used to assess when a key parameter of each chain has apparently converged to the stationary distribution.

D.2 Model diagnostics for examples

Convergence diagnostics for Example 5.4.

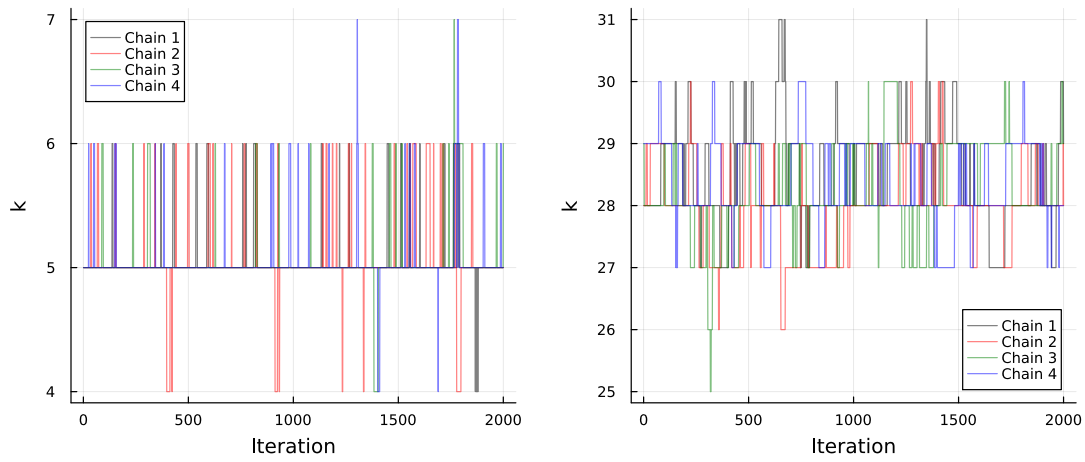


Figure D.2: Traceplots for k for the samples of size $n = 10^3$ (left) and $n = 10^5$ (right).

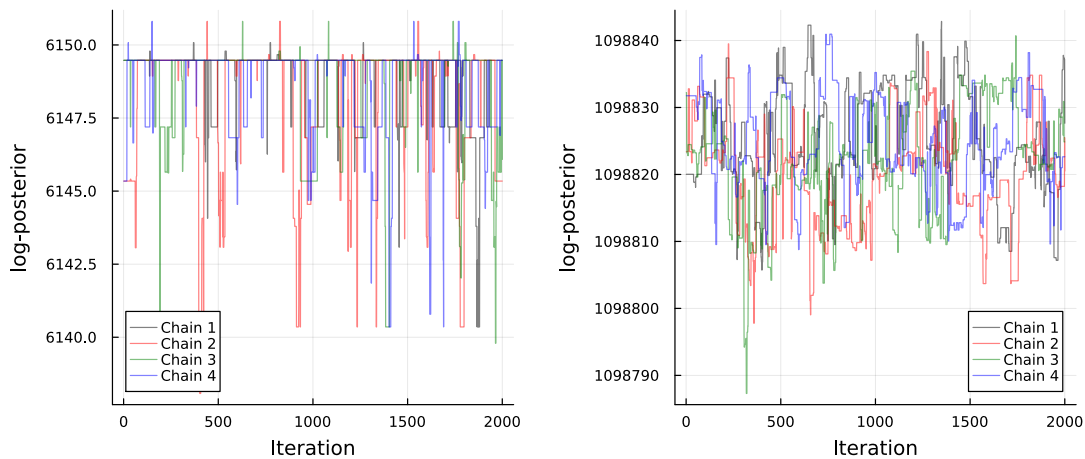


Figure D.3: Traceplots of the unnormalized log-posterior for the sample of size $n = 10^3$ (left) and $n = 10^5$ (right).

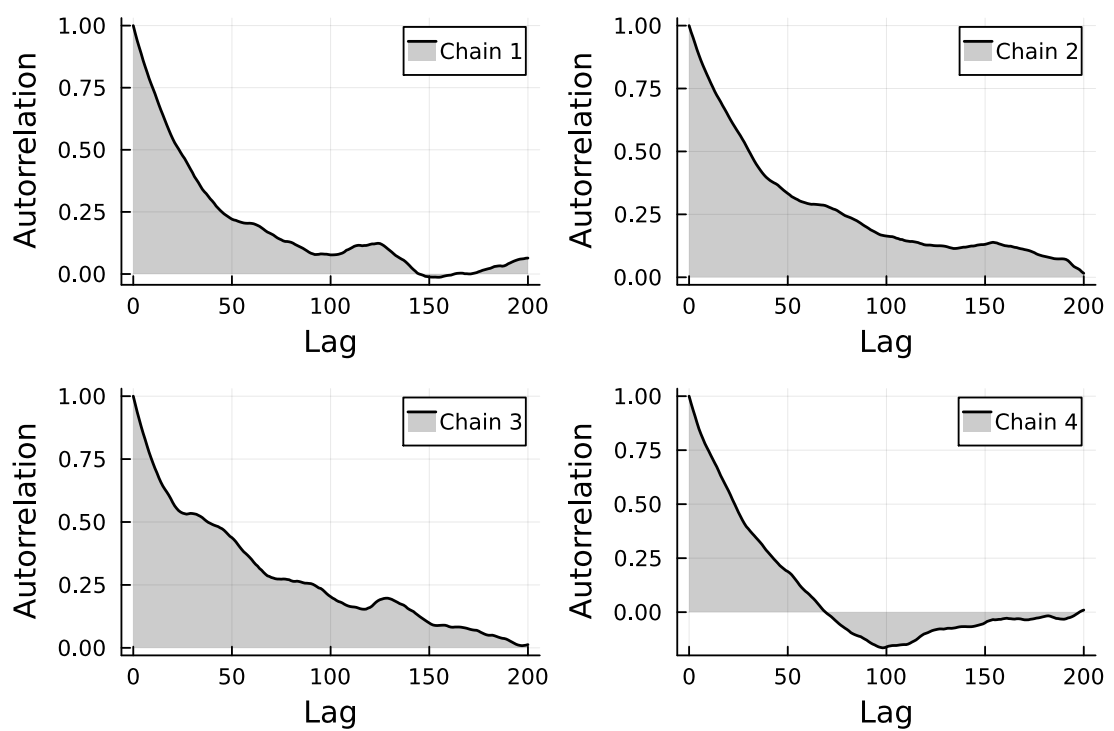


Figure D.4: Sample autocorrelation function for the log-posterior of the $n = 10^5$ sample.

Bibliography

- H. Akaike. Information theory and an extension of the maximum likelihood principle. pages 199–213, 1998. doi:10.1007/978-1-4612-1694-0_15.
- J. Aldrich. R. A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12:162–176, 1997. doi:10.1214/ss/1030037906.
- A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27:536–561, 1999. doi:10.1214/aos/1018031206.
- J. O. Berger, J. M. Bernardo, and D. Sun. Overall objective priors. *Bayesian Analysis*, 10:189–221, 2015. doi:10.1214/14-BA915.
- J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59:65–98, 2017. doi:10.1137/141000671.
- L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM: Probability and Statistics*, 10:24–45, 2006. doi:10.1051/ps:2006001.
- D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112:859–877, 2017. doi:10.1080/01621459.2017.1285773.
- L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16:199 – 231, 2001. doi:10.1214/ss/1009213726.
- I. Castillo. On Bayesian supremum norm contraction rates. *The Annals of Statistics*, 42:2058 – 2091, 2014. doi:10.1214/14-AOS1253.
- S. Chib. Marginal likelihood from the Gibbs output. *Journal of the american statistical association*, 90:1313–1321, 1995. doi:10.2307/2291521.
- G. Claeskens and N. L. Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008. doi:10.1017/CBO9780511790485.
- T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2 edition, 2006. doi:10.1002/047174882X.
- C. Cunen, G. H. Hermansen, and N. L. Hjort. Confidence distributions for change-points and regime shifts. *Journal of Statistical Planning and Inference*, 195:14–34, 2018. doi:10.1016/j.jspi.2017.09.009.
- P. L. Davies and A. Kovac. Densities, spectral densities and modality. *The Annals of Statistics*, 32:1093–1136, 2004. doi:10.1214/009053604000000364.

- P. L. Davies and A. Kovac. *ftnonpar: Features and Strings for Nonparametric Regression*, 2012. URL <https://CRAN.R-project.org/package=ftnonpar>. R package version 0.1-88.
- P. L. Davies, U. Gather, D. Nordman, and H. Weinert. A comparison of automatic histogram constructions. *ESAIM: Probability and Statistics*, 13:181–196, 2009. doi:10.1051/ps:2008005.
- L. Denby and C. Mallows. Variations on the histogram. *Journal of Computational and Graphical Statistics*, 18:21–31, 2009. doi:10.1198/jcgs.2009.0002.
- L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L1 view*. John Wiley & sons, Inc., 1985.
- P. Domingos. The role of Occam’s razor in knowledge discovery. *Data mining and knowledge discovery*, 3:409–425, 1999. doi:10.1023/A:1009868929893.
- R. Durrett. *Probability: theory and examples*. Cambridge university press, 5 edition, 2019. doi:10.1017/9781108591034.
- M. C. Edwards, R. Meyer, and N. Christensen. Bayesian nonparametric spectral density estimation using B-spline priors. *Statistics and Computing*, 29:67–78, 2019. doi:10.1007/s11222-017-9796-9.
- T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973. doi:10.1214/aos/1176342360.
- D. Freedman and P. Diaconis. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57:453–476, 1981. doi:10.1007/BF01025868.
- M. Gasparini. Bayesian density estimation via Dirichlet density processes. *Journal of Nonparametric Statistics*, 6:355–366, 1996. doi:10.1080/10485259608832681.
- U. Gather, T. Mildenberger, and Y. Rozenholc. Combining regular and irregular histograms by penalized likelihood. Technical report, Technische Universität Dortmund, 2009. doi: 10.17877/de290r-15901.
- S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984. doi:10.1109/TPAMI.1984.4767596.
- C. R. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28:1105–1127, 2000. doi:10.1214/aos/1015956709.
- R. Gerlach, K. Mengersen, and F. Tuyl. Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Analysis*, 4:151–158, 2009. doi:10.1214/09-BA405.
- S. Ghosal. Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics*, 29:1264–1280, 2001. doi:10.1214/aos/1013203453.
- S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics*, 29:1233–1263, 2001. doi:10.1214/aos/1013203452.

- S. Ghosal and A. W. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press, 2017. doi:10.1017/9781139029834.
- S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27:143–158, 1999. doi:10.1214/aos/1018031105.
- S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000. doi:10.1214/aos/1016218228.
- J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. Springer New York, NY, 1 edition, 2003. doi:10.1007/b97842.
- G. H. Givens and J. A. Hoeting. *Computational statistics*, volume 703. John Wiley & Sons, 2 edition, 2012. doi:10.1002/9781118555552.
- I. K. Glad, N. L. Hjort, and N. G. Ushakov. Correction of density estimators that are not densities. *Scandinavian Journal of Statistics*, 30:415–427, 2003. doi:10.1111/1467-9469.00339.
- U. Grenander. *Abstract inference*. Wiley, New York, 1981.
- P. Hall. Akaike’s information criterion and Kullback–Leibler loss for histogram density estimation. *Probability Theory and Related Fields*, 85:449–467, 1990. doi:10.1007/BF01203164.
- P. Hall and E. J. Hannan. On stochastic complexity and nonparametric density estimation. *Biometrika*, 75:705–714, 1988. doi:10.1093/biomet/75.4.705.
- C. R. Harris, K. J. Millman, S. J. van der Walt, et al. numpy.histogram_bin_edges. https://numpy.org/doc/stable/reference/generated/numpy.histogram_bin_edges.html, 2020. Accessed: 2024-04-09.
- W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970. doi:10.1093/biomet/57.1.97.
- N. H. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98:879–899, 2003. doi:10.1198/016214503000000828.
- N. L. Hjort and I. K. Glad. Nonparametric density estimation with a parametric start. *The Annals of Statistics*, 23:882–904, 1995. doi:10.1214/aos/1176324627.
- N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. *Bayesian Nonparametrics*, volume 28. Cambridge University Press, 2010. doi:10.1017/CBO9780511802478.
- J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9:90–95, 2007. doi:10.1109/MCSE.2007.55.
- H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186:453–461, 1946. doi:10.1098/rspa.1946.0056.
- M. C. Jones, M. Samiuddin, A. H. Al-Harbey, and T. A. H. Maatouk. The edge frequency polygon. *Biometrika*, 85:235–239, 1998. doi:10.1093/biomet/85.1.235.

- Y. Kanazawa. An optimal variable cell histogram. *Communications in Statistics-Theory and Methods*, 17:1401–1422, 1988. doi:10.1080/03610928808829688.
- Y. Kanazawa. Hellinger distance and Akaike’s information criterion for the histogram. *Statistics & probability letters*, 17:293–298, 1993. doi:10.1016/0167-7152(93)90205-W.
- J. L. Kirkby, Á. Leitaó, and D. Nguyen. Spline local basis methods for nonparametric density estimation. *Statistic Surveys*, 17:75–118, 2023. doi:10.1214/23-SS142.
- K. H. Knuth. Optimal data-based binning for histograms and histogram-based probability density models. *Digital Signal Processing*, 95:102581, 2019. doi:10.1016/j.dsp.2019.102581.
- P. Kontkanen. *Computationally efficient methods for MDL-optimal density estimation and data clustering*. PhD thesis, University of Helsinki, 2009.
- P. Kontkanen and P. Myllymäki. Mdl histogram density estimation. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 219–226, 2007.
- E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Science & Business Media, 2 edition, 2006. doi:10.1007/b98854.
- T. Leonard. A Bayesian method for histograms. *Biometrika*, 60:297–308, 1973. doi:10.1093/biomet/60.2.297.
- H. Li and S. Sieling. *essHist: The Essential Histogram*, 2019. URL <https://CRAN.R-project.org/package=essHist>. R package version 1.2.2.
- H. Li, A. Munk, H. Sieling, and G. Walther. The essential histogram. *Biometrika*, 107:347–364, 2020. doi:10.1093/biomet/asz081.
- T. L. Lindstrøm. *Spaces: An Introduction to Real Analysis*, volume 29. American Mathematical Soc., 2017.
- E. Marinari and G. Parisi. Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters*, 19:451, 1992. doi:10.1209/0295-5075/19/6/002.
- J. S. Marron and A. B. Tsybakov. Visual error criteria for qualitative smoothing. *Journal of the American Statistical Association*, 90:499–507, 1995. doi:10.2307/2291060.
- J. S. Marron and M. P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20:712–736, 1992. doi:10.1214/aos/1176348653.
- P. Massart. *Concentration inequalities and model selection: Ecole d’Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007. doi:10.1007/978-3-540-48503-2.
- V. Z. Mendizábal, M. Boullé, and F. Rossi. *genum histogram tool*, 2022. URL <http://www.marc-bouille.fr/genum/index.html>. Retrieved 07-03-2025.
- V. Z. Mendizábal, M. Boullé, and F. Rossi. Fast and fully-automated histograms for large-scale data sets. *Computational Statistics & Data Analysis*, 180:107668, 2023. doi:10.1016/j.csda.2022.107668.

- N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087–1092, 1953. doi:10.1063/1.1699114.
- T. Mildenberger, Y. Rozenholc, and D. Zasada. *histogram: Construction of Regular and Irregular Histograms with Different Options for Automatic Choice of Bins*, 2019. URL <https://CRAN.R-project.org/package=histogram>. R package version 0.0-25.
- Nasdaq. S&p 500 (spx) historical data, 2025. URL https://www.nasdaq.com/market-activity/index/spx/historical?page=1&rows_per_page=10&timeline=y5. Last visited: 2025-25-02.
- K. W. Ng, G. Tian, and M. Tang. *Dirichlet and Related Distributions: Theory, Methods and Applications*. John Wiley & Sons, 1 edition, 2011. doi:10.1002/9781119995784.
- R. L. Nuzzo. Histograms: A useful data analysis visualization. *PM&R*, 11:309–312, 2019. doi:10.1002/pmrj.12145.
- S. Petrone. Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, 27:105–126, 1999. doi:10.2307/3315494.
- A. M. Price-Whelan, P. L. Lim, N. Earl, N. Starkman, L. Bradley, et al. The astropy project: Sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package*. *The Astrophysical Journal*, 935:167, 2022. doi:10.3847/1538-4357/ac7c74.
- J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11:416–431, 1983. doi:10.1214/aos/1176346150.
- J. Rissanen, T. P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38:315–323, 1992. doi:10.1109/18.119689.
- C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2 edition, 2004. doi:10.1007/978-1-4757-4145-2.
- Y. Rozenholc, T. Mildenberger, and U. Gather. Combining regular and irregular histograms by penalized likelihood. *Computational statistics & data analysis*, 54: 3313–3323, 2010. doi:10.1016/j.csda.2010.04.021.
- M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9:65–78, 1982.
- J. D. Scargle, J. P. Norris, B. Jackson, and J. Chiang. Studies in astronomical time series analysis. vi. Bayesian block representations. *The Astrophysical Journal*, 764: 167, 2013. doi:10.1088/0004-637X/764/2/167.
- L. Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4:10–26, 1965. doi:10.1007/BF00535479.
- G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6:461–464, 1978. doi:10.1214/aos/1176344136.
- D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979. doi:10.1093/biomet/66.3.605.

Bibliography

- D. W. Scott. Frequency polygons: theory and application. *Journal of the American Statistical Association*, 80:348–354, 1985. doi:10.1080/01621459.1985.10478121.
- D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992. doi:10.1002/9780470316849.
- C. Scricciolo. On rates of convergence for Bayesian density estimation. *Scandinavian Journal of Statistics*, 34:626–642, 2007. doi:10.1111/j.1467-9469.2006.00540.x.
- O. H. Simensen. *AutoHist.jl: A pure Julia library for fast automatic histogram construction*, 2025. URL <https://github.com/oskarhs/AutoHist.jl.git>.
- C. J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 513–520, 1984.
- H. A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21:65–66, 1926. doi:10.1080/01621459.1926.10502161.
- W. Szpankowski. On asymptotics of certain recurrences arising in universal coding. *Problems of Information Transmission*, 34:142–146, 1998.
- A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 1 edition, 2009. doi:10.1007/b13794.
- S. A. van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.
- A. W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000. doi:10.1017/CBO9780511802256.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018. doi:10.1017/9781108231596.
- S. G. Walker. On sufficient conditions for Bayesian consistency. *Biometrika*, 90:482–488, 2003. doi:10.1093/biomet/90.2.482.
- S. G. Walker. New approaches to Bayesian consistency. *The Annals of Statistics*, 32: 2028–2043, 2004. doi:10.1214/009053604000000409.
- M. P. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51: 59–64, 1997. doi:10.2307/2684697.
- M. P. Wand. *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*, 2021. URL <https://CRAN.R-project.org/package=KernSmooth>. R package version 2.23-20.
- W. H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23:339–362, 1995. doi:10.1214/aos/1176324524.
- G. Zhong. Efficient and robust density estimation using Bernstein type polynomials. *Journal of Nonparametric Statistics*, 28:250–271, 2016. doi:10.1080/10485252.2016.1163349.