# Novel Approaches to Uncertainty Quantification in Nonparametric Settings

by

**Shaun William McDonald**

M.A.St, University of Cambridge, 2017
B.Sc. (Hons.), University of Manitoba, 2016

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

**© Shaun William McDonald 2022**
**SIMON FRASER UNIVERSITY**
**Fall 2022**

Copyright in this work is held by the author. Please ensure that any reproduction
or re-use is done in accordance with the relevant national copyright legislation.

# Declaration of Committee

| | |
|---|---|
| **Name:** | **Shaun William McDonald** |
| **Degree:** | **Doctor of Philosophy** |
| **Thesis title:** | **Novel Approaches to Uncertainty Quantification in Nonparametric Settings** |
| **Committee:** | **Chair:** Joan Hu<br>Professor, Statistics and Actuarial Science |
| | **Dave Campbell**<br>Supervisor<br>Adjunct Professor, Statistics and Actuarial Science |
| | **Richard Lockhart**<br>Committee Member<br>Professor, Statistics and Actuarial Science |
| | **Jiguo Cao**<br>Examiner<br>Professor, Statistics and Actuarial Science |
| | **Martin Lysy**<br>External Examiner<br>Associate Professor<br>Department of Statistics and Actuarial Science<br>University of Waterloo |

# Abstract

This thesis contains explorations of uncertainty quantification in a variety of nonparametric statistical settings, focusing on novel uses of uncertainty to enhance inference in ways which may otherwise be overlooked.

The first two chapters concern the Laplace approximation for high-dimensional integrals. This approximation is commonly used in complex models — for instance, to obtain marginal likelihoods in hierarchical models for use in optimization. The quality of the approximation may depend intimately on the true shape of the integrand. To assess this, we use probabilistic numerics, recasting the approximation problem and its inherent uncertainty in the framework of probability theory. We develop a diagnostic tool for the Laplace approximation and its underlying shape assumptions with this framework. The tool is decidedly non-asymptotic and is not intended as a full substitute for other quadrature methods. Rather, it is simply meant to test the feasibility of the assumptions underpinning the Laplace approximation with as little computational burden as possible.

Next, we provide a comprehensive overview of uncertainty quantification methods for density estimation. There are many methods of estimating an unknown density and constructing "plausible" sets in which it may lie. Examples of the latter include pointwise intervals, simultaneous bands, or balls in a function space; and they may be frequentist or Bayesian in interpretation. Here, we thoroughly review literature on density inference, covering a broad spectrum of ideas ranging from theoretical to practical.

Finally, we propose a novel approach to modelling in a micro-macro situation, in which group-level outcomes are dependent on covariates measured at the level of individuals within groups. Although such models are perhaps underrepresented in the literature, they have applications in economics, epidemiology, and the social sciences. Our approach is an empirical Bayesian method which jointly infers group-specific covariate densities and uses them as predictors in a functional linear model. Unlike many similar methods, the assumptions made on the structure of the data are minimal, allowing for better inference and a fuller quantification of uncertainty in a wide variety of situations.

**Keywords:** Probabilistic numerics; multilevel modelling; nonparametric inference; uncertainty quantification; functional data analysis; Density estimation

# Dedication

To Donna, Jay, and Riley, who will likely be horrified to have this much math dedicated to them.

# Acknowledgements


> And if you need to know
> the measure of a man
> you simply count his friends.
>
> Paul Williams

A couple pages isn't enough for me to thank everyone who deserves it. If your name isn't mentioned here and you think it should have been, you're probably right.

First and foremost, I'd like to thank my family, especially my parents, Donna and Jay, and my brother Riley. Their unwavering love, encouragement, sympathy, and pride have kept me moving at every step of this journey. I'd also like to thank Colleen, who ensured I landed on my feet when I came to Vancouver.

I owe an enormous debt of gratitude to Theo Koulis, as well as CJ Mundy and Jens Ehn, who supervised the undergraduate research that started this entire journey.

Thanks to all the friends I made in Cambridge, who made my first year of grad school so fun that I signed up for another five.

Thank you to Joel, Neil, Sam, Tory, Dan, and Chelsey, who made Vancouver so much easier to live in and so much harder to leave. To Will and Trevor for all the coffee runs. To Tom Loughin for giving me the *really* important education.

Thank you to all my friends and fellow musicians in Winnipeg, who continue to provide me with the all-important second half of "work-life balance".

A collective thanks to all of the professors, support staff, and students in SFU's Department of Statistics and Actuarial Science. I am honoured to have been part of it.

Thank you to Eric Cator, Zhong Guan, Maria Lomeli, Omiros Papaspiliopoulos, Yushi Shi, Richard Nickl, Bin Wang, and Anders Nielsen, who clarified many of the ideas in Chapters 3–4 through e-mail correspondence. Thanks are also owed to the anonymous reviewers who gave excellent feedback for the original publication of Chapter 4. I also extend my gratitude to the collective online Stan community, without whom Chapter 5 would have taken far longer to finish.

Thank you to Richard Lockhart for his supervisory support, and for providing some characteristically brilliant insight which helped to turn the ideas in Chapters 2–3 into a

reality. I'd also like to thank Saman Muthukumarana and Alex Leblanc at the U of M for their contributions to the original ideas underpinning Chapter 5.

Finally, I thank my supervisor, Dave Campbell. None of this would've been possible without his patience, advice, encouragement, and support. I am immeasurably grateful for the opportunities he has provided me, and for the values and principles I have been fortunate enough to share with him. It has been a pleasure and an honour to work with him, and I truly hope for more opportunities to do so.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Overview

In the broadest sense of the term, uncertainty quantification (UQ) is one of the most fundamental components of statistics. The very concept of statistical inference means assessing the plausibility of an estimate or a hypothesis, establishing in some concrete sense what our expectations of "the truth" may be, and describing the degree and nature of the error that may exist in these expectations. Although they allow for an enormous variety of philosophies, interpretations, and methods, these principles are central to almost any type of statistics.

However, as advances in data and technology have allowed for the development of more complex methodology, the uncertainty inherent in various modelling tasks is not always fully accounted for. Depending on the context, it may be taken for granted entirely. Complicated models often rely on either simplifying assumptions or approximations of various types in order to be computationally viable. Although these are usually necessary concessions in models which typically provide otherwise acceptable estimation and inference in practical applications, there are certainly cases in which a better understanding of uncertainty — in either the data or the modelling strategy applied to it — would be beneficial. More comprehensive UQ can provide better insights about the data-generating process and any approximations thereof. In turn, such insights can aid assessment of our modelling strategies and their shortcomings, possibly motivating the development of improved strategies.

This thesis comprises a collection of studies on uncertainty quantification in various nonparametric contexts. Each chapter either discusses UQ in a context where it may be underused, or proposes new methods of using uncertainty to better understand a given statistical problem.

## 1.2 Organization of the thesis

In Chapter 2, we detail a diagnostic tool which uses probabilistic numerics to assess the appropriateness of the Laplace approximation to high-dimensional integrals, based on the work of Zhou [301]. The tool is based on a non-asymptotic philosophy which is designed to use uncertainty about the shapes of high-dimensional integrands in a practical way. Chapter 3 details the diagnostic's use in high dimensions, with discussions of the associated challenges and an application to real state-space data. Chapter 4 is a comprehensive literature review of nearly all known methods of uncertainty quantification for various types of probability density estimators, originally published as a standalone paper [197]. In Chapter 5, we propose a new Bayesian method of multilevel regression which combines density inference and functional data analysis to flexibly model group-level responses based on individual-level predictors. The method is shown to accommodate data structures of a very general type, fully accommodating for uncertainty in both levels of the data. We apply the method to a variety of simulated data, showing that its theoretical generality and inferential power are realized in practice.

# Chapter 2

# Proof of concept for a probabilistic diagnostic tool to assess Laplace approximations

## 2.1 Introduction

Many statistical models assume the existence of "unseen" variables which influence the actual observed data, but are distinct from the model parameters that are of interest for inference. One such model is the *state-space model* (SSM), which has become a staple of ecological modelling [e.g. 2, and references therein] and will serve as a motivating example throughout this chapter and the next. Briefly, the SSM assumes that (possibly vector-valued) data $y_t$ are observed at discrete time steps $t = 1, \ldots, T$. At a given time $t$, the distribution of $y_t$ depends on an unobserved or "hidden" state $x_t \in \mathbb{R}^q$ (typically the dimensionality of $x_t$ is the same for all $t$, but it may differ from the dimensionality of the $y_t$'s). In turn, the distribution of $x_t$ depends on the previous hidden state, $x_{t-1}$. The reader may recognize this as the structure of a *hidden Markov model (HMM)*, although that term is typically used when the domain of the hidden states is discrete [e.g 51]. Here, they are assumed to be continuous and possibly multivariate.

In mathematical terms, the SSM is characterized by the joint likelihood[1]

$$p_{x,y}\left(\boldsymbol{x}, \boldsymbol{y} \mid \theta\right) = p\left(x_1 \mid \theta\right) \left[\prod_{t=2}^{T} p\left(x_t \mid x_{t-1}, \theta\right)\right] \left[\prod_{t=1}^{T} p\left(y_t \mid x_t, \theta\right)\right], \tag{2.1}$$

[1]There are several possible formulations for the distribution of the first hidden state (the $p\left(x_1 \mid \theta\right)$ term in (2.1)). Some literature assumes it to depend on an "initial state" $x_0$ which is given its own prior in turn [e.g. 205] or simply point estimated [e.g. 268]. The latter is essentially equivalent to specifying an "unconditional" distribution for $x_1$, another common approach [e.g. 51, 164]. Some authors omit the $p\left(x_1 \mid \theta\right)$ term entirely, thereby implicitly assigning $x_1$ an "improper uniform prior" [e.g. 208, which is the formulation used in Section 3.2]. The general model form given in (2.1) will suffice for the purposes of this chapter.

where $\boldsymbol{x} = (x_1, \ldots, x_T)$ is a vector of dimension $d = qT$ concatenating the hidden states, $\boldsymbol{y}$ is defined analogously, and $\theta$ is a vector of model parameters. These parameters are conceptually different from the hidden states even though both are unobserved: $\theta$ represents the *fixed effects* of the model, whereas $\boldsymbol{x}$ represents *random effects*[2].

There are a variety of methods for both frequentist and Bayesian inference with SSM's [e.g. 65, 268, and references therein]. In the frequentist framework, one typically wishes to estimate $\theta$ by maximizing the marginal likelihood of the data,

$$p_y(\boldsymbol{y} \mid \theta) = \int_{\mathbb{R}^d} p_{x,y}(\boldsymbol{x}, \boldsymbol{y} \mid \theta) \, \mathrm{d}\boldsymbol{x}. \tag{2.2}$$

Unfortunately, the necessary integral over the hidden states is $d$-dimensional, and as such the marginal likelihood cannot realistically be computed — much less optimized — in most cases. Instead, frequentist inference methods for SSM's typically rely on approximations of various types to obtain a suitable estimate of $\theta$. Examples include methods based on particle filtering, as described by Kantas et al. [149]. Another common — and less computationally demanding [e.g. 2] — approach is use of the *Laplace approximation* (LA). The Laplace approximation of the marginal likelihood is reasonably easy to compute and optimize as a function of $\theta$, but it is based on certain assumptions about the shape of the joint likelihood as a function of $\boldsymbol{x}$: namely, that it is well approximated by a $d$-dimensional Gaussian density. If this assumption is not satisfied, the LA may not be suitable, and different methods for SSM inference may need to be invoked.

The example of the SSM provides motivation for the broader goal of this chapter, which is to develop a diagnostic tool to check the assumptions underpinning the LA. In particular, our interest is in assessing whether or not a given function is "close enough" to the Gaussian shape to justify using the Laplace approximation of its integral. In making this assessment, we strive for a "middle ground" of computational effort: the diagnostic will naturally be more complex than the LA itself, but much less expensive than a full-fledged numerical estimate of the integral. Expanding on the work of Zhou [301], here we describe such a diagnostic tool based on the machinery of *probabilistic numerics*, a burgeoning field which exploits probability theory to tackle numerical problems. The tool is an application of the probabilistic numerical technique of *Bayesian quadrature* (BQ), which allows for both estimation and inference of unknown integrals. Unlike "conventional" BQ, however, the actual integral value is of secondary importance, as the tool is primarily intended to capture as much information as possible about the *shape of the integrand*. In keeping with the aforementioned objective of "medium effort", the tool is also decidedly non-asymptotic: it is meant to deliver as much information as possible with a modest amount of computation,

[2]Of course, in a Bayesian setting, both model components are given priors and essentially treated in the same way. In that case, the difference between them is more of a "philosphical" matter.

without consideration of any type of limiting behaviour. The goal is a fast, informal method that can be readily deployed to determine if additional modelling efforts are needed beyond the LA.

The remainder of the chapter proceeds as follows. Section 2.2 defines the LA and establishes the notation used throughout Chapters 2–3, while Section 2.3 provides more detail about the workings of probabilistic numerics and BQ in particular. Sections 2.4–2.5 provide technical details about the design of our diagnostic tool, and Section 2.6 shows a low-dimensional application. Chapter 3 of the thesis is focused on challenges, applications, and discussion of the diagnostic in high-dimensional settings.

## 2.2 Framework and notation

Consider a positive function $f : \mathbb{R}^d \to \mathbb{R}_{>0}$ and its integral $F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x$. More rigorous treatments of the Laplace approximation are available in, for instance, de Bruijn [62] and Barndorff-Nielsen et al. [12], but for this exposition it suffices to assume that all second-order partial derivatives of $f$ exist and are continuous, and that $f$ attains a maximum at some point $\hat{x} \in \mathbb{R}^d$. To reflect the common use case where $f$ is a density or likelihood, $\hat{x}$ is called a *mode*. Let $H$ be the Hessian of $\log f$ at $\hat{x}$ and suppose that it is negative definite. Taking a second-order Taylor expansion of $\log f$ about $\hat{x}$ gives the approximation

$$\log f(x) \approx \log f\left(\hat{x}\right) + \frac{1}{2}\left(x - \hat{x}\right)^\top H \left(x - \hat{x}\right), \tag{2.3}$$

since all first-order partial derivatives of $\log f$ are equal to zero at the mode. Exponentiating (2.3) gives an approximation for $f$ in the form of (up to normalizing constants) a Gaussian density centered at $\hat{x}$ with covariance matrix $-H^{-1}$. In turn, integrating this exponentiated function (hereafter called the *Gaussian approximation to* $f$) produces the *Laplace approximation* to $F$:

$$\begin{aligned} F \approx L\left(f\right) &:= f\left(\hat{x}\right) \int_{\mathbb{R}^d} \exp\left[\frac{1}{2}\left(x - \hat{x}\right)^\top H \left(x - \hat{x}\right)\right] \mathrm{d}x \\ &= f\left(\hat{x}\right) \sqrt{(2\pi)^d \det\left(-H^{-1}\right)}. \end{aligned} \tag{2.4}$$

The LA has a long history of use in statistics [e.g. 178, 278]. It is exact (or "true") if the integrand $f$ is itself proportional to a Gaussian density. There are other function shapes for which this may be the case, but such instances may be thought of as "coincidence". Certainly, the derivation of the LA via (2.3) is based on an assumption of approximately Gaussian shape (insofar as it assumes that the second-order Taylor series is a reasonable approximation to $\log f$), and as noted in Section 2.1, this assumption is our main interest.

Before proceeding to further details about the construction of the diagnostic tool, it is worthwhile to connect these concepts to the SSM example described in Section 2.1. For

given observations $\boldsymbol{y}$ and parameter values $\theta$, the joint likelihood $p_{xy}(\cdot, \boldsymbol{y} \mid \theta)$ takes the role of the integrand, viewed as a function of the hidden states $\boldsymbol{x} \in \mathbb{R}^d$. In turn, one can see from (2.2) that the marginal likelihood $p_y(\boldsymbol{y} \mid \theta)$ takes the role of the integral over $\mathbb{R}^d$ to be approximated by $L(p_{xy})$. Note, however, that this approximation is itself a function of $\boldsymbol{y}$ and $\theta$, as both

$$\hat{x} = \underset{\boldsymbol{x}}{\operatorname{argmax}}\ p_{xy}(\boldsymbol{y}, \boldsymbol{x} \mid \theta) \quad \text{and} \quad H = \left.\frac{\partial^2 \log p_{xy}}{\partial \boldsymbol{x}^2}\right|_{(\boldsymbol{y}, \hat{x}, \theta)}$$

may depend on these quantities. Indeed, one of the most common ways to "fit an SSM" in the frequentist sense is to maximize $L(p_{xy})$ with respect to $\theta$ (given observed $\boldsymbol{y}$), typically using standard numerical algorithms. Fitting the model in this way becomes a matter of *nested* optimization, since in each iteration $\hat{x} = \hat{x}(\theta, \boldsymbol{y})$ must be (numerically) calculated for the current $\theta$-value [see 165, for instance].

Implicit in the use of such methods for SSM's is the assumption that the LA is reasonably accurate given $\boldsymbol{y}$ and for each $\theta$-value calculated during the optimization steps. If the shape of $p_{xy}$ with respect to $\boldsymbol{x}$ is not "sufficiently Gaussian" at a given iteration, then the ultimate estimate of $\theta$ may not be close to the actual MLE for the marginal likelihood. Therefore, it would be desirable to check the validity of the LA at each step, using the diagnostic tool detailed below.

## 2.3 Probabilistic numerics and Bayesian quadrature

Broadly speaking, probabilistic numerics is the use of probability theory, from a somewhat Bayesian perspective, to simultaneously perform estimation and uncertainty quantification in standard numerical problems [131]. For instance, Chkrebtii et al. [50] developed a probabilistic solver for differential equations. For a given equation, they jointly modelled the function and its derivatives with a Gaussian process prior, then sequentially conditioned on evaluations of the true derivative to conduct posterior inference on the entire solution.

The approach briefly described above — using Gaussian process priors and finitely many function evaluations to obtain posteriors for the functions and quantities of interest — is at the core of many probabilistic numerical methods. In particular, it is the standard framework with which *Bayesian quadrature* (BQ) is usually conducted [see 26, 54, and references therein]. As the name suggests, BQ is a probabilistic analogue to standard numerical integration that uses a combination of prior belief and gathered information about a function. The remainder of this section, in which the diagnostic for the LA is developed, will also serve as an explanation of the mathematical machinery underpinning BQ.

Literature on BQ commonly assumes that the integral of interest is with respect to a probability (i.e. finite) measure $G$ on the domain [e.g 26], and a standard choice for $\mathbb{R}^d$ is a $d$-dimensional Gaussian measure [215, 150]. Accordingly, we use an "importance weighting

trick" [151, 96, 216] to re-express the integral of interest. Recalling the notation of Section 2.2, the integral of $f$ over $\mathbb{R}^d$ is

$$F = \int_{\mathbb{R}^d} f(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)g(x)\mathrm{d}x = \int_{\mathbb{R}^d} r(x)\mathrm{d}G(x), \tag{2.5}$$

where $r := f/g$ and $g$ is the density of the aforementioned Gaussian measure $G$, the parameters of which will be discussed later. It is this "re-weighted" function $r$ that is modelled with a Gaussian process prior [151]. The mean function of the GP prior, $m_0^x$, is taken to be the (similarly re-weighted) Gaussian approximation of $f$ underpinning (2.3) and (2.4):

$$m_0^x(x) := \frac{f(\hat{x}) \exp\left[\frac{1}{2}(x-\hat{x})^\top H (x-\hat{x})\right]}{g(x)}, x \in \mathbb{R}^d. \tag{2.6}$$

The covariance operator for the GP is a (positive-definite) kernel $C_0^x$ on $\mathbb{R}^d \times \mathbb{R}^d$, defined in Section 2.4.2. Because integration is a linear projection, such a prior on $g$ induces a univariate normal prior on $F$ with mean $m_0 := \int_{\mathbb{R}^d} m_0^x(x)\mathrm{d}G(x) = L(f)$ and variance $C_0 := \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} C_0^x(x,z)\mathrm{d}G(x)\mathrm{d}G(z)$ [e.g. 96, 131].

In what follows, let $\boldsymbol{s} = (s_1, \ldots, s_n)^\top \in \mathbb{R}^{n\times d}$ be a row-wise concatenation of $n$ (transposed) vectors in $\mathbb{R}^d$ (we will sometimes call it a "*grid*" of $n$ "points" in $\mathbb{R}^d$). Then, for instance, the notation $r(\boldsymbol{s})$ will refer to the column vector $(r(s_1), \ldots, r(s_n))^\top \in \mathbb{R}^n$, and $C_0^x(\boldsymbol{s}, \boldsymbol{s})$ will denote the $n \times n$ matrix with $(i,j)^{\text{th}}$ entry $C_0^x(s_i, s_j)$. Using standard GP identities [e.g. 234], one may use true function values at the *interrogation points* $\boldsymbol{s}$ to obtain a posterior distribution for $g$ (with another slight abuse of notation):

$$r \mid r(\boldsymbol{s}) \sim \mathcal{GP}(m_1^x, C_1^x), \tag{2.7}$$

$$m_1^x(x) = m_0^x(x) + C_0^x(x,\boldsymbol{s})^\top [C_0^x(\boldsymbol{s},\boldsymbol{s})]^{-1}(r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})), \tag{2.8}$$

$$C_1^x(x,z) = C_0^x(x,z) - C_0^x(x,\boldsymbol{s})^\top [C_0^x(\boldsymbol{s},\boldsymbol{s})]^{-1} C_0^x(z,\boldsymbol{s}). \tag{2.9}$$

In turn, the posterior distribution on the integral $F$ is [e.g. 26, or, indeed, virtually any BQ paper]

$$F \mid r(\boldsymbol{s}) \sim \mathcal{N}(m_1, C_1), \tag{2.10}$$

$$m_1 = L(f) + \left[\int_{\mathbb{R}^d} C_0^x(z,\boldsymbol{s})\mathrm{d}G(z)\right]^\top [C_0^x(\boldsymbol{s},\boldsymbol{s})]^{-1}(r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})), \tag{2.11}$$

$$C_1 = C_0 - \left[\int_{\mathbb{R}^d} C_0^x(x,\boldsymbol{s})\mathrm{d}G(x)\right]^\top [C_0^x(\boldsymbol{s},\boldsymbol{s})]^{-1}\left[\int_{\mathbb{R}^d} C_0^x(x,\boldsymbol{s})\mathrm{d}G(x)\right]; \tag{2.12}$$

where the integrals are row-wise over $\boldsymbol{s}$:

$$\int_{\mathbb{R}^d} C_0^x(x,\boldsymbol{s})\mathrm{d}G(x) = \left(\int_{\mathbb{R}^d} C_0^x(x,s_1)\,\mathrm{d}G(x), \ldots, \int_{\mathbb{R}^d} C_0^x(x,s_n)\,\mathrm{d}G(x)\right)^\top.$$

It is useful to think of the posterior means and variances as their prior counterparts modified by the addition or subtraction of some "correction term".

The posterior (2.10) will serve as the diagnostic for the Laplace approximation. Borrowing from the traditional notion of hypothesis testing, one may deem the Laplace approximation (or, perhaps more accurately, the shape assumptions motivating it) acceptable or valid if $L(f)$ falls within the range spanned by the $(0.025, 0.975)$ quantiles of (2.10): the 95% "confidence interval" centered at the posterior mean. Conversely, if $L(f)$ is outside of this interval, the Laplace approximation would be deemed inappropriate ("rejection"), and one could proceed to use a more involved method of estimating $F$.

## 2.4 Design decisions

In general terms, there are three major categories of "design" choices one must make in order to conduct BQ, each of which will be explored in the following subsections. First, we must decide where to place interrogation points $\boldsymbol{s}$; second, a covariance kernel $C_0^x$ must be chosen for the GP prior; and finally, we must specify the measure $G$ against which to integrate. The latter two involve setting some *hyperparameters* that will govern the behaviour of the Gaussian process; this will be deferred to Section 2.5.

Recall that the diagnostic is intended to quickly — and somewhat heuristically — determine whether a given function $f$ is "sufficiently Gaussian" to justify the LA for its integral. In particular, it should expend only as much computational effort as is necessary to reliably make this determination, with actual estimation of the integral $F$ being a *secondary* goal. In this respect, its objectives are different from those of "traditional" BQ, in which interrogation points may be chosen to minimize the posterior variance of the integral [215, 200, 138] or the entropy of the integrand [116]; and hyperparameters may be chosen by some goodness-of-fit criterion [26, 234] or approximately marginalized [217], with both approaches depending on the "observations" $r(\boldsymbol{s})$. The computational costs arising from such methods would be antithetical to the "moderately fast" nature of the diagnostic. Instead, it should be "one-size-fits-all" so that it can be quickly applied to any suitable function. Although "ad hoc" design choices are made in some BQ papers [e.g 150], the fact remains that the usual goal is to obtain an accurate integral estimate with low uncertainty. Beyond the issue of computation, there is a more fundamental difference between our goals and those of "traditional" BQ, or, indeed, the usual principles of inference in a more general sense. Typically, one may wish to maximize the *power* of their inference, ensuring that any true deviation from some null hypothesis will be found with sufficient data. In the present context, this would mean embracing the standard BQ goal of high accuracy and low uncertainty, so that even the smallest deviation from the LA could be rejected if there are enough well-placed interrogation points. However, such a diagnostic would not be very useful in practice. Hearkening back to the SSM example from Section 2.1, in all but the simplest models it

will be known in advance that the joint likelihood is not *exactly* Gaussian, and the LA not exactly met. The pertinent question is whether the joint likelihood is Gaussian *enough*, and a diagnostic that answered this question in the negative for every nonlinear model would be trivial and useless. Thus, the usual aim of high "power" is actually *not* desirable here: the diagnostic should be calibrated such that it *fails to reject* any function which is "close enough" to Gaussian, in a sense explained below. In this way, the design choices detailed in the following sections target an unconventional notion of "*good-enough-ness of fit*".

### 2.4.1 Placement of interrogation points

The selection of interrogation points (or "nodes", as they are commonly known in the literature) is the defining feature of any quadrature method. Much has been written about the asymptotic error rates (as the number of points $n \to \infty$) of various quadrature methods, and the ways in which they depend on the dimensionality of the domain $d$ and the smoothness of the integrand [e.g. 146, 26]. However, none of these considerations are relevant to the development of a quick, one-size-fits-all tool intended to determine if a function is "Gaussian enough" for the LA to be reasonable. Thus, the grid of interrogation points must provide as much pertinent information as possible about the *shape* of $f$, and (particularly in high dimensions, as explained in Chapter 3) how this shape influences the validity of the LA. Importantly, it must do this with as small a grid as possible in order to be "medium-effort"; in particular, the grid size must grow at a reasonable rate with respect to $d$. One hopes that the goals of the diagnostic can be accomplished with less computation than it takes to conduct a more accurate BQ.

To begin with, let $\boldsymbol{s}^* = (s_1^*, \ldots, s_n^*)^\top \in \mathbb{R}^{n \times d}$ be a grid of "preliminary" interrogation points. Ostensibly the preliminary grid should not depend on any properties of the function $f$, but considerations such as dimensionality can certainly inform its construction. We will assume that the grid is a union of *fully symmetric sets*, as considered by Karvonen and Särkkä [150]. Briefly, this means that if we take an arbitrary vector $s_i^*$ from the grid, any vector obtained via permutation or sign changes of its coordinates is also in the grid [ibid.]. We also assume that the grid contains multiples of the standard basis vectors of $\mathbb{R}^d$ (i.e. points are placed "along the axes") and that its centroid is the origin (the origin may be included in the grid, but this is not strictly necessary). No further restrictions will be placed on the preliminary grid, but some type of sparsity is desirable for the computational reasons mentioned above. The sparse grid methods described by Karvonen and Särkkä [150], or modifications thereof, are particularly useful to this end.

Now, recalling that $H$ is negative-definite, consider the eigendecomposition $-H^{-1} = VDV^\top$ (where $V$ is orthogonal and $D$ is diagonal) and let $T := V\sqrt{D}$. The vectors comprising the actual interrogation grid $\boldsymbol{s}$ used in the diagnostic will be affine transformations of the preliminary grid vectors: $s_i = Ts_i^* + \hat{x}$, $i = 1, \ldots, n$. This transformation serves three purposes. The first is a translation so that the centroid of the grid is $\hat{x}$, the mode of $f$.

Since $f$ will be a density or likelihood in most applications, it makes sense for the grid to be oriented around the region of highest density. In contrast, a grid centered at the origin may be "off-center" for some integrands, capturing only limited tail behaviour and certainly not enough "shape information". The second purpose for the transformation is a rotation, as $T$ maps standard basis vectors to eigenvectors of $H$ (which are the same as those of $-H^{-1}$). Thus, by placing some of the preliminary points along the "standard axes" of $\mathbb{R}^d$, we ensure that the corresponding interrogation points are aligned along the directions in which the "curvature" of $f$ at the mode is most extreme[3]. Because $H$ completely characterizes the shape of $f$ under the "null hypothesis" that it (approximately) satisfies the assumptions of the LA, heuristically it makes sense to say that, *a priori*, one would expect such interrogation points to contain the most pertinent "shape information". Finally, the transformation "stretches" its inputs in the direction of each eigenvector $V_i$ by a factor of $\sqrt{D_{ii}}$ ($D_{ii}$ being the eigenvalue associated with $V_i$). Thus, if $H$ is such that the Gaussian approximation to $f$ (and, presumably, $f$ itself) has different scales in different directions, the grid will capture this appropriately. In summary, this transformation turns a preliminary grid of the type stipulated above into an interrogation grid that is adapted to the contours of the Gaussian approximation to $f$. In this respect, it can be assumed — *a priori* or "under the null hypothesis" of Gaussian shape — that the grid so obtained is, in some informal sense, "optimal" for obtaining the necessary information about $f$.

There is another, perhaps more intuitive interpretation of interrogation grids generated in this way. Let $X$ be a multivariate normal random variable with density proportional to the Gaussian approximation to $f$, i.e. $X \sim \mathcal{N}\left(\hat{x}, -H^{-1}\right)$. Then the $i^{\text{th}}$ component of the vector $VX$ is the $i^{\text{th}}$ *principal component*, or PC, of $X$, and has marginal variance equal to $D_{ii}$ [142]. Thus, the affine transformation of the preliminary grid is centered at the mean of $X$, aligned with its "principal axes", and scaled according to the scales of its PC's. For example, recall that for $i = 1, \ldots, d$, the preliminary grid contains points of the form $\pm m e_i$, where $m > 0$ and $e_i$ is the $i^{\text{th}}$ standard basis vector of $\mathbb{R}^d$. The corresponding interrogation points, $\pm m\sqrt{D_{ii}}V_i + \hat{x}$, are "$m$ standard deviations (of the $i^{\text{th}}$ PC of $X$) away from the mode (in the direction of that PC)".

### 2.4.2 Form of covariance kernel

The covariance structure of the diagnostic will be based on the *squared exponential kernel*:

$$\kappa\left(x, z\right) = \alpha^{-d} \exp\left[-\frac{\|x - z\|^2}{2\lambda^2}\right], \tag{2.13}$$

[3]This point can be formalized and made clear with some linear algebra and multivariate calculus. First note that the second directional derivative of $\log f$ at the mode is always negative and is maximized (resp. minimized) in the direction of the first (resp. last) eigenvector of $H$. Finally observe that this statement must also be true for $f$ itself since is always positive and its gradient is zero at $\hat{x}$.

a common choice in BQ [e.g 215, 150, 26]. The hyperparameter $\alpha$ controls the *precision* of the GP, serving as a scaling factor for its variance and for that of its integral. It is more common in literature to parameterize the kernel in terms of scale as opposed to precision, replacing $\alpha^{-d}$ in (2.13) with $\alpha^2$ [e.g. 215, 116], but the practical difference between these choices is purely notational. The parameterization in (2.13) is the same as that used by Chkrebtii et al. [50], and the fact that $\alpha$ is raised to the power of $-d$ in (2.13) reflects their notion that the $d$-dimensional kernel can be viewed as a pointwise product of $d$ univariate kernels. The hyperparameter $\lambda$ is the *length-scale*, which controls the size of fluctuations in GP values between distinct points [234]. In informal terms[4], $\lambda$ therefore controls the "smoothness" of the GP.

The actual covariance function used in the diagnostic is a modification of (2.13) based on the function of interest $f$. It is

$$C_0^x(x,z) = f(\hat{x})^2 \det\left(-H^{-1}\right) \kappa\left(T^{-1}x, T^{-1}z\right), \tag{2.14}$$

where the transformation matrix $T$ was defined in Section 2.4.1. Because $\|T^{-1}x - T^{-1}z\|^2 = (x-z)^\top(-H)(x-z)$, the prior covariance of the GP at distinct points depends on the distance between these points in a linear transformation of Euclidean space, with the transformation depending on the "curvature" of $\log f$ at $\hat{x}$. Equivalently, the prior GP covariance function (2.14) is a (scaled) *Mahalanobis kernel* [1].

### 2.4.3 Choice of measure

In Section 2.3, we used an importance re-weighting trick to express $F$ as an integral w.r.t. a Gaussian measure $G$. O'Hagan [215] and Kennedy [151] considered BQ for $r = f/g$ with a constant GP prior mean and noted that results would be most accurate if the density $g$ closely approximated the shape of $f$, i.e. if $r$ was roughly constant. The latter noted an analogy with importance sampling (IS), in which $F$ is also modelled as the integral of $r$ w.r.t. G and the shape of $g$ should match that of the integrand [e.g. 297]. Although our GP prior mean (2.6) is not constant, we still found in preliminary experiments that $g$ had to be a fairly good "fit" to $f$ in order for the diagnostic to behave reasonably. Within the convenient class of Gaussian measures, remarks by O'Hagan and Kennedy suggest that $g$ proportional to the Gaussian approximation to $f$, i.e. $G = \mathcal{N}(\hat{x}, -H^{-1})$, would be a reasonable "starting point". The measure ultimately used for the diagnostic is a slight modification of this:

$$G = \mathcal{N}\left(\hat{x}, -\gamma^2 H^{-1}\right), \tag{2.15}$$

[4]In *formal* terms, a GP with squared exponential covariance kernel is infinitely differentiable, in the mean square sense, regardless of the value of $\lambda$ [234]. "Smoothness" as informally used above simply means an absence of "wiggles" at small scales in functions sampled from the GP.

where the new hyperparameter $\gamma > 0$ controls the "spread" of $G$ and will be discussed in Section 2.5.

### 2.4.4 Invariance of diagnostic behaviour

At first glance, it may seem that these function-specific design choices are antithetical to the intended "one-size-fits-all" nature of the diagnostic. On the contrary, our design ensures a few kinds of advantageous "invariance". Recall that the interrogation points are obtained from the function-agnostic preliminary grid as $s_i = Ts_i^* + \hat{x}$, $i = 1, \ldots, n$. Plugging any two interrogation points $s_i, s_j$ into (2.14) therefore gives $C_0^x(s_i, s_j) \propto \kappa\left(s_i^*, s_j^*\right)$. Note also that analogous results can be shown to hold for the integral terms[5] in (2.11 – 2.12) and for the prior mean interrogations $m_0^x(\boldsymbol{s})$. Therefore, in principle the interrogations should provide the same quality and quantity of "information" for *any* $f$. Now, recall that the diagnostic rejects the LA for $f$ iff it is not contained in the central 95% interval of the integral posterior, i.e. iff $L(f) \notin (m_1 - 1.96\sqrt{C_1}, m_1 + 1.96\sqrt{C_1})$. Note that $\sqrt{C_1}$ is equal to $L(f) \propto f(\hat{x})\sqrt{\det(-H^{-1})}$ times a factor depending only on $\boldsymbol{s}^*$ and the hyperparameters $(\lambda, \alpha, \gamma)$ (by (2.12) and (2.14)); similarly, $m_1$ is equal to $L(f)$ times a factor depending only on $\boldsymbol{s}^*$, the hyperparameters, and the "normalized" function values $f(\boldsymbol{s})/f(\hat{x})$ (by (2.6), (2.11), and the definition of $r$). Thus, the necessary and sufficient condition for rejection does not depend on the actual values of $\hat{x}$, $f(\hat{x})$, and $\det(-H^{-1})$: *it is invariant to any scaling of the function or affine transformation of its domain.* More formally, for a fixed set of hyperparameters, the diagnostic rejects the LA when applied to $f$ iff it rejects the LA when applied to any function of the form $f_{\text{Trans}} : x \mapsto af(Ax + b)$ with $a > 0$, $A \in \mathbb{R}^{d \times d}$ with $\det(A) \neq 0$, and $b \in \mathbb{R}^d$. The only way in which $f$ affects the result of the diagnostic is through the *relative* differences between its values at the interrogation points and those of its Gaussian approximation. Because the diagnostic seeks only to determine whether $f$ is "sufficiently Gaussian in shape", this is precisely the appropriate behaviour for it to have.

Note the "standardized" design developed in Sections 2.4.1–2.4.3 is not without precedent in the BQ literature. For instance, Särkkä et al. [256] adopted the idea of *stochastic decoupling* from sigma-point methodology: to integrate a function $r$ against some Gaussian measure $\mathcal{N}(\mu, P)$, they placed a GP prior with the standard squared exponential covariance kernel (2.13) on the function $r_{\text{Trans}} : x \mapsto r\left(\mu + \sqrt{P}x\right)$ and used a standardized set of "unit" interrogation points. Such an approach is essentially equivalent (possibly up to variance scaling factors) to our design; indeed, the authors made note of its invariance to affine transformations. However, their main interest was in deriving BQ-based methods for filtering and smoothing in nonlinear SSM's, in which $\mu$ and $P$ are computed for each necessary integral according to their algorithms [256].

[5] To see this, note that the density $g$ has a multiplicative factor of $\sqrt{\det(-H)} = \left|\det\left(T^{-1}\right)\right|$, and integrate (2.14) w.r.t. $G$ by substitution. This is another reason why the choice of measure (2.15) makes sense.

## 2.5 Hyperparameter calibration

It remains to select values for $(\lambda, \alpha, \gamma)$. As discussed above, the design of the interrogation grid and covariance kernel serve to "standardize" the input and output scales of the GP, so it is not necessary to consider these factors when setting the hyperparameters. Indeed, for a given dimension $d$ and preliminary grid $\boldsymbol{s}^*$, the same hyperparameter values should be used for *any* $f$ to ensure the aforementioned diagnostic invariance. Recall from the beginning of Section 2.4 that the intent is to test "good-enough-ness of fit": the diagnostic should reject the LA for functions with a substantially non-Gaussian shape, but should *not* be so "powerful" that it rejects functions which are close enough to Gaussian. With this in mind, we propose to set the hyperparameters in a somewhat heuristic way based on a predetermined *calibration* or *test function* $\tau$. Such a function should have a shape fairly close to Gaussian in order to serve as the "edge case" for the diagnostic. Specifically, given a preliminary grid $\boldsymbol{s}^*$ and test function $\tau$, the hyperparameters for the $d$-dimensional diagnostic should be set such that the following conditions are met when the diagnostic is applied to $\tau$.

(1) The LA $L(\tau)$ should be on the boundary of the rejection region (i.e. equal to one of the endpoints of the 95% central interval for the integral posterior); and

(2a) the discrepancy between $\tau$ and the "un-weighted" posterior GP mean, $m_1^x \cdot g$, should be as small as possible throughout the domain; or at the very least

(2b) the posterior integral mean $m_1$ should be as close as possible to the true integral of $\tau$.

Either version of the second condition should ensure that the diagnostic is reasonably accurate when applied to $\tau$. Of course, accurate estimation is still an ancillary goal in general, but at the very least it should be achieved for the test function to ensure that the diagnostic uses interrogations in a sensible way. Condition (2a) is the more desirable version since it directly targets the shape of the function and also implies (2b) by design, but in high dimensions with large interrogation grids it may only be possible to ensure that (2b) is met (see Section 3.1). The first condition establishes $\tau$ as the "borderline" function: any function that is "less Gaussian" will have its LA rejected, and any function "at least as Gaussian" will not. To see this, consider the normalized posterior "correction term"[6]

$$\Delta(f) := \frac{\sqrt{\det(-H)}}{f(\hat{x})} \left[ \int_{\mathbb{R}^d} C_0^x(z, \boldsymbol{s}) \mathrm{d}G(z) \right]^\top [C_0^x(\boldsymbol{s}, \boldsymbol{s})]^{-1} (r(\boldsymbol{s}) - m_0^x(\boldsymbol{s})), \tag{2.16}$$

[6]To avoid any possible confusion, it should be reiterated that all of the quantities in these definitions — namely, $G, r, \boldsymbol{s}, m_0^x, m_1^x, C_0$, and $C_1$ — technically depend on $f$ through the constructions detailed in Sections 2.3–2.4.3. More accurate notation would reflect this explicitly, but such notation would be cumbersome.

which, as per (2.11), is (up to the scaling factors in front) the difference between the prior and posterior integral means when the diagnostic is applied to a function $f$. It can be shown that the rejection criterion for the diagnostic is equivalent to $f(\hat{x})\sqrt{\det(-H^{-1})}\,|\Delta(f)| > 1.96\sqrt{C_1}$. Recall from Section 2.4.4 that $C_1$ only depends on $f$ through scaling factors $f(\hat{x})^2$ and $\det(-H^{-1})$, so the rejection criteria is equivalent to $|\Delta(f)| > \epsilon$, where the number $\epsilon > 0$ depends only on $\boldsymbol{s}^*, \lambda, \alpha$, and $\gamma$. Now, to meet condition (1) for the test function $\tau$ is to have $|\Delta(\tau)| = \epsilon$. Therefore, with this calibration scheme a function $f$ will have its LA rejected iff $|\Delta(f)| > |\Delta(\tau)|$. Again, all that matters are the *relative differences* between a function and its Gaussian approximation at the interrogation points — specifically, whether the weighted sum of these as given by (2.16) (with the weights depending on $\boldsymbol{s}^*, \lambda$, and $\gamma$) is larger in magnitude than it is for the predetermined "borderline Gaussian" $\tau$.

A natural choice for a test function is the density of a $d$-dimensional multivariate Student's $t$ distribution with $\nu$ degrees of freedom, mean at the origin, and scale matrix equal to the identity. Denote this density by $\tau_{\nu,d}$, so

$$\tau_{\nu,d}(x) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{\|x\|^2}{\nu}\right)^{-\frac{\nu+d}{2}}, \tag{2.17}$$

and note that it has heavier tails than a $d$-dimensional Gaussian density, so the LA, given by the formula

$$L(\tau_{\nu,d}) = \left(\frac{2}{\nu+d}\right)^{\frac{d}{2}}\frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \tag{2.18}$$

underestimates the true integral (which is always equal to 1). However, $\tau_{\nu,d}$ approaches a standard multivariate normal density in the limit $\nu \to \infty$, and therefore $L(\tau_{\nu,d}) \to 1$ as well. Therefore, for some large value of $\nu$, the shape of $\tau_{\nu,d}$ may be said to be "sufficiently Gaussian" to warrant non-rejection of the LA. Denote such a value by $\nu_d$ to reflect the fact (discussed further in Section 3.1) that the specific choice of test function should depend on the dimension $d$. One option that works reasonably well is to let $\nu_d$ be the smallest integer such that $L(\tau_{\nu_d,d}) \geq 0.95$. The densities of multivariate $t$ variables with more than $\nu_d$ degrees of freedom are close enough in shape to Gaussians that their Laplace approximations are within 5% of the true integral value; conversely, those with lower degrees of freedom have heavier tails and LA's that underestimate the true integral by over 5%.

With the family of test functions established, it is now possible to discuss how one may set the hyperparameters to satisfy the conditions listed above. First note that the precision parameter $\alpha$ does not actually affect the posterior mean; as a scaling factor, it serves only to ensure that condition (1) is met. Thus, it suffices to find good values for $\lambda$ and $\gamma$, after which $\alpha$ can simply be chosen to scale the posterior variance $C_1$ such that $|\Delta(\tau_{\nu_d,d})| = \epsilon$.

The fact that $\lambda$ affects the shape of the GP mean is obvious since, as noted in Section 2.4.2, it determines the "smoothness" of functions sampled from the GP and is therefore a "shape parameter" in some sense. What is perhaps more surprising is the effect of $\gamma$, the scaling factor for the underlying measure $G$. Recall from Section 2.4.3 that $G$ is analogous to the proposal distribution in IS. It is well-known that the performance of an importance sampler will be poor if the density $g$ has lighter tails than $f$, and it is therefore better to err on the side of caution by taking $g$ to have slightly heavier tails [e.g. 297]. In our context, this corresponds to setting $\gamma$ slightly larger than 1, and in our experiments we use a value of

$$\gamma = \sqrt{1.5 \frac{\nu_d + d}{\nu_d + d - 3}}. \tag{2.19}$$

The heuristic motivation for this choice is as follows. Consider $d$-dimensional random vectors $Y \sim \tau_{\nu_d,d}$ and $X \sim g$, where $g = g(\tau_{\nu_d,d})$ is the density corresponding to (2.15) for the choice of function $f = \tau_{\nu_d,d}$. The $\gamma$-value given by (2.19) ensures that $1.5 \times \mathrm{Var}[Y_1 \mid Y_2 = 0, \ldots, Y_d = 0] = \mathrm{Var}[X_1 \mid X_2 = 0, \ldots, X_d = 0]$ — in words, the univariate conditional densities (with all other coordinates fixed at the origin) of the $t$ distribution used for calibration have variance equal to two thirds of those of the "approximating Gaussian density" $g$ [151]. Here the analogy with IS becomes somewhat strained, as it can be shown that *any* Gaussian proposal distribution will result in an importance sampler with infinite variance when applied to a $t$ density. In fact, taking $G$ itself as a $t$ distribution is often a good choice in IS due to the heaviness of the tails [279, and references therein]. Prüher et al. [229] considered this choice of $G$ in BQ, but noted that the kernel integrals in (2.11–2.12) would not have closed forms. For computational convenience we will retain our choice of a Gaussian measure, but note that, unlike IS, the posterior variance of the integral is still guaranteed to be finite here.

### 2.5.1 Calibrating in two dimensions

Using these ideas, we will now demonstrate how calibration can work for the diagnostic in $d = 2$ dimensions. The test function will be a bivariate $t$ density with $\nu_2 = 38$ degrees of freedom, as $L(\tau_{38,2}) = 0.95$. The preliminary interrogation grid $\boldsymbol{s}^*$ will consist of evenly-spaced points in a "cross-shaped" formation "on the axes" of $\mathbb{R}^2$:

$$\boldsymbol{s}^* = \{(0,0)\} \cup \{\pm m e_i : m = 1, 2, 3,\ i = 1, 2\}, \tag{2.20}$$

where $e_i$ is the $i^{\text{th}}$ standard basis vector of $\mathbb{R}^2$. Such "cross-shaped grids" are appealing, at least in low dimensions, because the number of points $n$ scales linearly with $d$. Here, we have $n = 13$.

In order to heuristically understand how hyperparameter choices affect the behaviour of the diagnostic, it will be useful to plot the difference between the test function $\tau_{38,2}$ and the "un-weighted" GP posterior mean $m_1^x \cdot g$ for various $(\lambda, \gamma)$-values. Note that the "optimal" hyperparameters will depend on the dimensionality of the domain, the specific test function used, and the preliminary grid chosen. In particular, if one wishes to use the diagnostic in 2 dimensions with a different preliminary grid from the one considered here, it should not necessarily be assumed that the $\lambda$ value given below is suitable for the new grid.

Choosing $\gamma$ according to (2.19) with $d = 2$ and $\nu_d = 38$ results in a value of $\gamma = 1.2734$. In this low-dimensional setting with a small interrogation grid, it is possible to crudely approximate an analytic method to find an "optimal" $\lambda$: given the aforementioned $\gamma$-value, we approximate the "$L^2$ error" $\int_{\mathbb{R}^2} \left(m_1^x(x)g(x) - \tau_{38,2}(x)\right)^2 \mathrm{d}x$ and its derivative w.r.t. $\lambda$ by simple Riemann sums over the grid of points $\{-10, -9.99, -9.98, \ldots, 9.99, 10\}^2$. This approximate error is then minimized w.r.t. $\lambda$ using the BFGS algorithm as implemented in the `fminunc` function in the MATLAB Optimization Toolbox [196], resulting in a value of $\lambda = 4.2241$.

Figure 2.1 shows results for the diagnostic applied to $\tau_{38,2}$ with these design choices. The difference $(m_1^x \cdot g - \tau_{38,2})$ is very small among the lines defined by the interrogation grid, but there are deep valleys centered around the "main diagonals" of the plane and within the boundaries of the interrogation grid. Since the heavy-tailed $t$ density is larger than its Gaussian approximation in these regions, it is clear that there is not much difference between the prior and posterior GP means there. The interrogation points are too far from these regions to exert much influence on the posterior mean there - in this respect, one may say that the GP is failing to *interpolate* to these areas. A more mathematical explanation of this behaviour can be extracted from (2.8), the definition of $m_1^x$. By this definition, it holds that $m_1^x(\boldsymbol{s})\, g(\boldsymbol{s}) = f(\boldsymbol{s})$ for any $f$ and any combination of hyperparameter values. However, at any other point $x$, the extent to which $m_1^x(x)$ updates from the prior GP mean $m_0^x(x)$ is determined by the "weights" $C_0^x(x, \boldsymbol{s})^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}$. These weights tend to decrease in magnitude as $x$ moves away from the points in $\boldsymbol{s}$, to an extent determined by $\lambda$ and $\gamma$. When $\lambda$ is small, there is almost no prior dependence between GP values at distinct points, so these weights are close to zero for $x \notin \boldsymbol{s}$. This can be seen in Figure 2.2: the posterior GP mean is forced to equal $\tau_{38,2}$ at the interrogation points, but everywhere else it is virtually unchanged from the prior mean $m_0^x$). Thus, in this case $m_1$ is very close to the prior value $m_0 = L(\tau_{38,2}) = 0.95$. In contrast, the "optimal" $\lambda$-value results in a posterior integral estimate of $m_1 = 0.99095$, quite close to the true value of 1. Note that in each case, the integral of $(m_1^x \cdot g - \tau_{38,2})$ (the surface in the left plot) over $\mathbb{R}^2$ is equal to the difference between $m_1$ and the true integral (in the right plot, the horizontal distance between the peak of the bell curve and the red line). As mentioned above, $\alpha$ is chosen to ensure that the test function is on the boundary between rejection and non-rejection, resulting in a

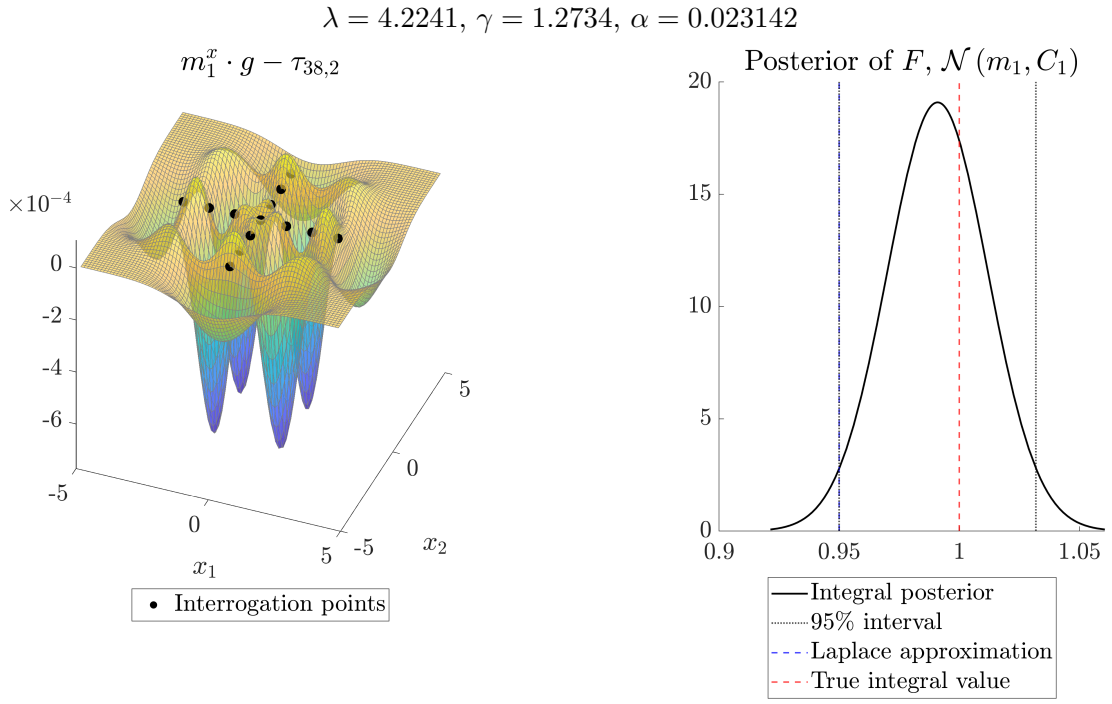


Figure 2.1: Results for the diagnostic applied to the 2-dimensional test function $\tau_{38,2}$, with an "optimal" $\lambda$, $\gamma$ obtained from (2.19), and $\alpha$ set to ensure that the LA is on the boundary of the "rejection region". Left: the difference between the un-weighted posterior GP mean and the true function. Right: the posterior distribution for the integral $F$.

posterior variance of $C_1 = 4.3653 \times 10^{-4}$ for the "optimal" $\lambda$ and $5.7369 \times 10^{-8}$ for the lower one.

The effect of $\gamma$ is less easily explained than that of $\lambda$. In fact, their effects counterbalance each other to some degree: we found that it was still possible to approximate an "optimal" $\lambda$ with the method described above even for different fixed values of $\gamma$, with lower $\gamma$-values resulting in higher required $\lambda$-values and vice-versa. In principle, this suggests that the diagnostic will not be too sensitive to the use of different $\gamma$-values, since any possible negative effect on its performance could be mitigated by adjusting $\lambda$ in the opposite direction. However, there is a limit to this in practice, and $\gamma$-values that are either too low or too high can still be problematic. With a lower value of $\gamma = 1$, it became difficult to find an optimal $\lambda$, as the BFGS algorithm was quite sensitive to the choice of initial value. Although the results of differently-initialized BFGS runs were not consistent with each other, they all resulted in final $\lambda$-values over 9. At length-scales this large, the *Gram matrix* $C_0^x(\boldsymbol{s}, \boldsymbol{s})$ is poorly conditioned (for instance, with $\boldsymbol{s}^*$ given by (2.20), its reciprocal condition number is $7.7885 \times 10^{-14}$ when $\lambda = 9$, as opposed to $7.1579 \times 10^{-10}$ when $\lambda = 4.2241$), so numerical stability becomes a concern. Furthermore, even with $\lambda$-values this high, the posterior integral mean $m_1$ was around 0.986: not as close to 1 as it was with the slightly larger $\gamma$-value and its "optimal" $\lambda$. The fact that these difficulties exist for $\gamma = 1$ is noteworthy since this

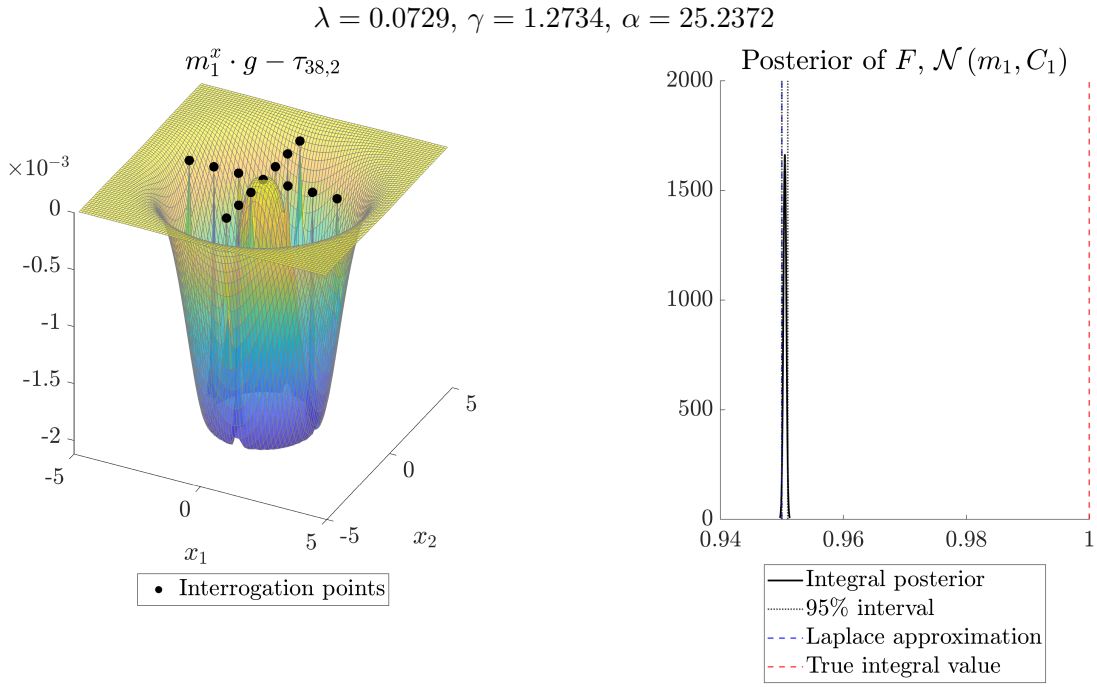


Figure 2.2: Results for the diagnostic applied to $\tau_{38,2}$ with a low $\lambda$-value, $\gamma$ obtained from (2.19), and $\alpha$ set to ensure that the LA is on the boundary of the "rejection region". Note the spikes created by "undersmoothing".

corresponds to using an integrating measure whose density is proportional to the Gaussian approximation to the true function.

Concerns about numerical accuracy do not exist with an even larger $\gamma$-value, as the corresponding optimal $\lambda$-value will be smaller and the Gram matrix will therefore be better conditioned. However, sensitivity becomes a problem in this situation: when $\gamma$ is high, even a relatively small deviation from the optimal $\lambda$ can change the diagnostic's behaviour quite dramatically. This will be of particular concern in higher dimensions, in which it is not viable to approximate and optimize the $L^2$ error numerically. In the current 2-dimensional setting, with $\gamma = 3$, the approximately-optimal $\lambda$-value is 1.1953, and the results with these hyperparameters (not shown) are fairly similar to those in Figure 2.1. A modest increase to $\lambda = 1.3$ creates a noticeably different outcome, as shown in Figure 2.3. The "interpolation valleys" seen in Figure 2.1 are slightly smaller in size, as the larger length-scale increases dependence between distinct points in the GP, thereby allowing the interrogations to exert more influence at faraway points. However, this slight improvement in interpolation comes at a cost: undesirable *extrapolation* effects due to oversmoothing. Indeed, in all four directions just beyond the extremal interrogation points, $m_1^x$ dips well below the true function $\tau_{38,2}$. As a result, $m_1 = 0.98108$ is farther from the true integral than it was with the hyperparameter values in Figure 2.1. Oversmoothing causes the weights $C_0^x(x, \boldsymbol{s})^\top \left[C_0^x(\boldsymbol{s}, \boldsymbol{s})\right]^{-1}$ to have unpredictable effects at $x$ beyond the boundaries of the interrogation grid, depending

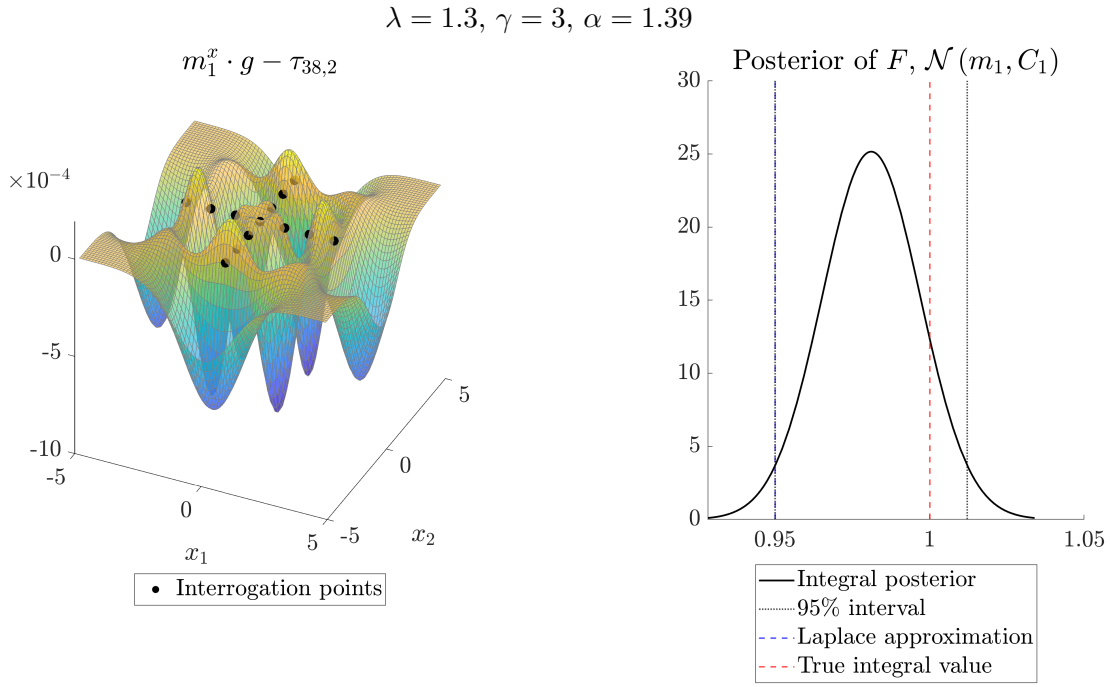


Figure 2.3: Results for the diagnostic applied to $\tau_{38,2}$ with a high $\gamma$-value, $\alpha$ set to ensure that the LA is on the boundary of the "rejection region", and a $\lambda$-value that is only slightly larger than the approximate "$L^2$ optimum" for this $\gamma$-value (which, in this case, was $\lambda = 1.1953$). Contrast with Figure 2.1 to see the excessive sensitivity to $\lambda$ caused by a high $\gamma$-value.

on the spread and density of $\boldsymbol{s}$ as well as the shape of the integrand. In some cases, the "extrapolation valleys" seen in Figure 2.3 may be replaced by large "hills", causing $m_1$ to significantly overestimate the value of $F$ (not shown). It is now clear that the original hyperparameter values in Figure 2.1 provide the best "tradeoff", balancing the interpolation errors of undersmoothing with the extrapolation errors of oversmoothing.

## 2.6 Example: a banana-shaped function

In a paper on MCMC algorithms, Haario et al. [118] considered a function with "banana-shaped" contours, defined by "twisting" one coordinate of a Gaussian density. Letting $\varphi(\cdot; \Sigma)$ denote a bivariate Gaussian density with mean at the origin and covariance matrix $\Sigma$, the version of the function used here is

$$\beta(x) := \varphi\left(x_1, x_2 - \frac{1}{2}\left(x_1^2 - 3\right); \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

It turns out that the Laplace approximation is true for this function: $L(\beta) = \int_{\mathbb{R}^2} \beta(x)\mathrm{d}x = 1$. As discussed in Section 2.2, this may be viewed as "coincidence", as it is clear from Figure 2.4 that $\beta$ is not well-approximated by a Gaussian shape. In this way, the function $\beta$

represents an interesting test case for the diagnostic: although its LA is technically valid, it is *not* "Gaussian enough" and should therefore be rejected. Indeed, with the preliminary interrogation grid (2.20) and corresponding (approximately) "optimal" hyperparameters (see Figure 2.1), this is precisely what the diagnostic does, as shown in Figure 2.5. The un-weighted GP posterior mean now accurately captures the light tails of $\beta$ along the line $x_2 = 0$, although it does not capture the large ridges defining the "banana" shape since there are no interrogation points along these ridges. As a result, the posterior integral estimate $m_1$ is 0.3658 — well below the true value and the LA. Note also that there are small oscillations between the interrogation points along the $x_1$-axis, perhaps signifying a small amount of oversmoothing. Finally, observe that the posterior variance is small enough to result in a rejection of the LA, which is well above the 97.5% quantile for the posterior distribution of $F$. These design choices would certainly be poor ones if accurate integral estimation was the main goal. In this framework, however, they are clearly suitable — the shape information captured by the diagnostic suggests that $\beta$ is not Gaussian enough to justify using the LA outright. In this type of scenario, a practitioner could subsequently employ a different method to estimate the integral. Presumably, they would then discover that the LA was correct all along — but *not* because of the quality of the Taylor approximation (2.3) underpinning its use.

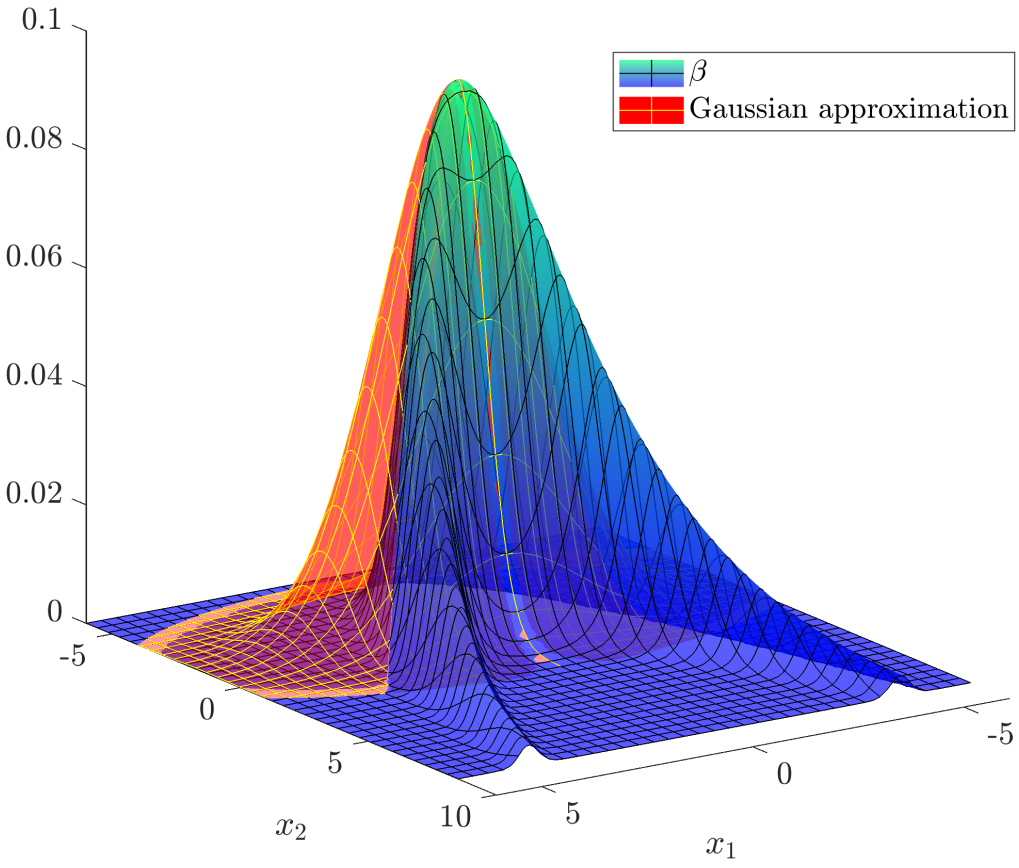


Figure 2.4: A two-dimensional "banana-shaped" function alongside its Gaussian approximation.

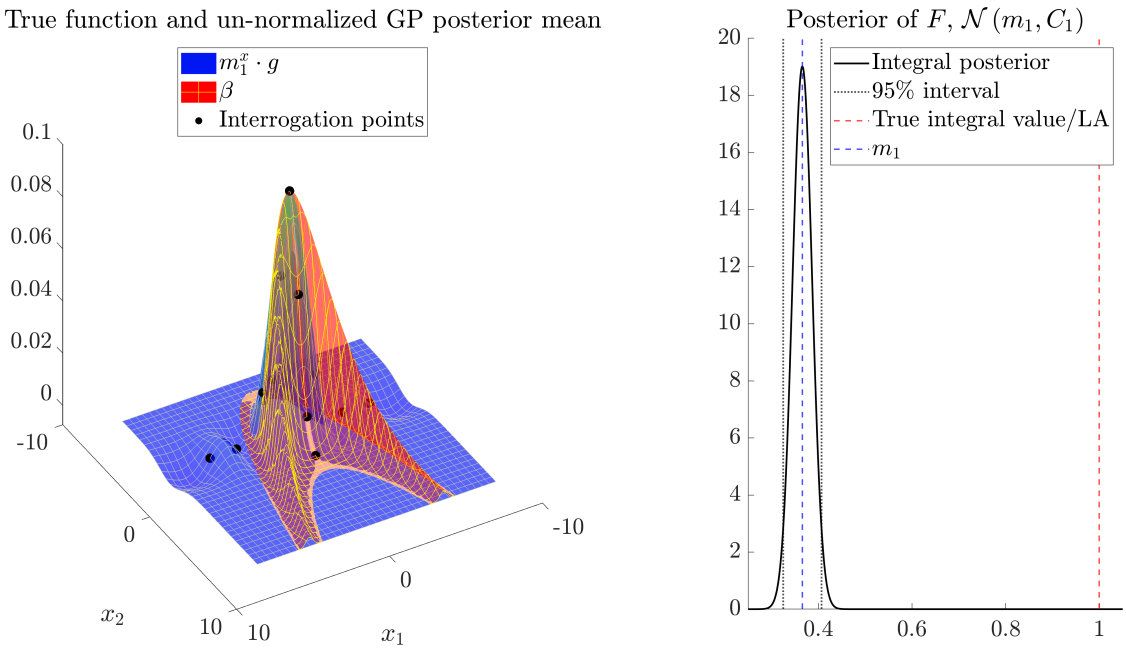


Figure 2.5: Results from applying the diagnostic to the two-dimensional banana-shaped function, using the same design choices as in Figure 2.1. Note that the colours in the left plot are reversed from those in Figure 2.4 for easier visualization.

# Chapter 3

# The Laplace approximation diagnostic in high dimensions: considerations and applications

## 3.1 Overview

The low-dimensional LA diagnostic experiments of the previous chapter (Sections 2.5.1 and 2.6) were useful for exposition, but ultimately our main interest is in applying the diagnostic to higher-dimensional functions. Unsurprisingly, for large dimensions $d$ it is more challenging to ensure that the diagnostic behaves well. Recall from Section 2.5.1 that we found an approximately "optimal" length-scale $\lambda$ by minimizing a type of $L^2$ error associated with the calibration function $\tau_{\nu_d,d}$. This required the numerical approximation of an integral over $\mathbb{R}^d$, which is not computationally feasible in high dimensions (if it was, there would be no need for the LA or for this very diagnostic). It is also not viable to seek a closed-form expression for the $L^2$ error: doing so would, in turn, require an analytic expression for the inverse of the Gram matrix $C_0^x(\boldsymbol{s}, \boldsymbol{s})$, which will be prohibitively complicated for all but the smallest of interrogation grids. With respect to the conditions for hyperparameter calibration listed in Section 2.5, condition (2a) can be assessed with a heuristic visual approach for moderate dimensions $d > 2$: viewing a 2-dimensional "slice" of the difference $m_1^x \cdot g - \tau_{\nu_d,d}$ with $x_3, \ldots, x_d$ all set to 0 (exploiting the symmetry of the $t$ density and the fact that its mode is at the origin), one can adjust $\lambda$ so as to make this difference appear as uniformly small as possible, attempting to balance issues with interpolation and extrapolation. Unfortunately, even this approach ceases to be viable when $d$ is large, so that Condition (2b) is all that can be ensured. The reasons for this depend on the structure of the preliminary grid $\boldsymbol{s}^*$; in turn, this structure should be chosen to mitigate the challenges that arise in high dimensions. More details on some possible choices are given below. We found in preliminary experiments that grids of the form (2.20) — that is, those with multiple evenly-spaced points along each axis — did not work very well when generalized to higher dimensions. Note that, although the points along any given axis are equally spaced in such

grids, the distances between points on *different* axes will be larger. We conjecture that this variation in interrogation point distances becomes problematic in high dimensions as more axes and points are added.

Fundamentally, the issue in high dimensions is that a function's "shape information" — of the type described in the preceding chapter — becomes more divorced from the value of its integral, making it more difficult to test the notion of "sufficiently Gaussian shape to justify the LA". There are a few different possible causes for this. The first is a well-known "curse of dimensionality" affecting certain high-dimensional probability density functions: most of their mass is in the tails, far away from the high-density region directly surrounding the mode [e.g. 39, 18]. Essentially, this happens because the neighbourhood around the mode is of a much smaller (Lebesgue) volume than the region encompassing the tails, so that most of the mass contributing to the integral is in a "shell" where the *product* of density and volume is high [ibid.]. For instance, if $X$ is a $d$-dimensional standard normal random variable, the *Gaussian annulus theorem* [22, Theorem 2.9] states that, with high probability, $X$ will be in a spherical shell of width $\mathcal{O}(1)$ and distance $\mathcal{O}\left(\sqrt{d}\right)$ from the origin.

This poses an unfortunate challenge for the diagnostic: when the integrand $f$ is a high-dimensional density, its shape is easiest to visually assess around the mode where its values are relatively large, but its integral (and its LA, which is the integral of the Gaussian approximation to $f$) may be determined farther away where $f$ is much smaller. For example, consider the case $d = 72$ (the dimensionality of the real-data examples in Section 3.2), for which (as explained in Section 2.5) we take the calibration function $\tau$ to be a multivariate $t$ density with $\nu_{72} = 25921$ degrees of freedom because $L\left(\tau_{25921,72}\right) = 0.95$. The top plot of Figure 3.1 shows the integral of this density — and that of its Gaussian approximation ($m_0^x \cdot g$, in the notation of Section 2.3) — over the 72-dimensional ball $\{x : \|x\| < r\}$ as the radius $r$ varies. Observe that both $\tau$ and its Gaussian approximation have most of their mass between distances 7–10 from the origin. Furthermore, the difference between the integrals does not start to become apparent until the radius of integration is at least 8 (note that, as $r \to \infty$, the integrals of $\tau$ and its Gaussian approximation converge to 1 and the LA, respectively). This affirms the idea that most of the important information about the integral (in particular, its closeness to the LA) is quite far from the mode, in a region that authors such as Betancourt [18] call the *typical set*. In contrast, the region of maximal *shape difference* between $\tau$ and its Gaussian approximation occurs much closer to the origin, where there is almost no mass. This can be seen in the bottom plot of Figure 3.1, which shows that $\tau$ differs most from its Gaussian approximation at a distance of approximately 2 from the origin. Even there, the largest difference between them is only about 0.002% of $\tau$'s value at the mode. Further out in the aforementioned "typical set", the two functions are visually indistinguishable.

There is another interesting point to be made here about the high-dimensional diagnostic. It was stated in Section 2.5 that $\nu_d$, the degrees of freedom for the calibration function

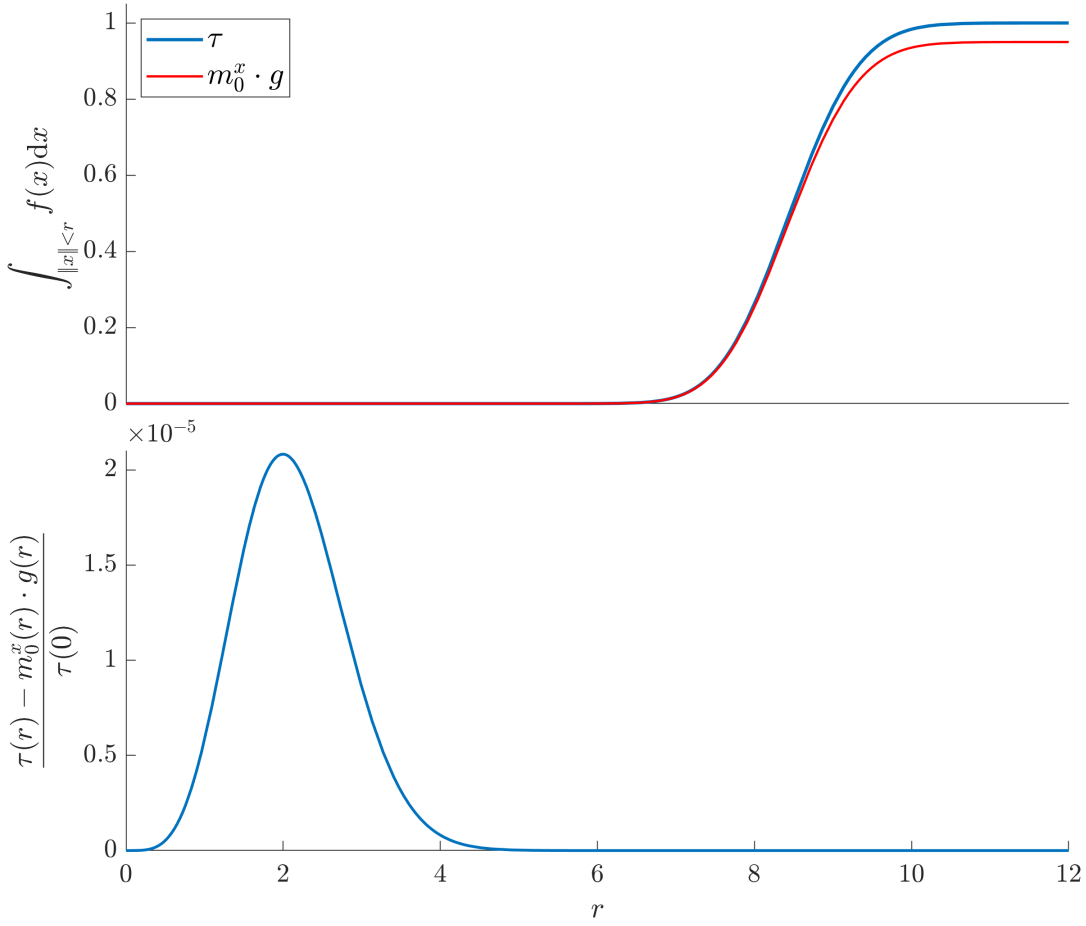


Figure 3.1: Top: the amount of mass enclosed by $\tau = \tau_{25921,72}$ and its Gaussian approximation over a ball of radius $r$ centered at the origin. Bottom: the difference between $\tau$ and its Gaussian approximation at a distance of radius $r$ from the origin, normalized by the value of $\tau$ at the origin.

in $d$ dimensions, would depend on $d$ itself. Indeed, the Laplace approximation (2.18) for a multivariate $t$ density is decreasing in $d$ for fixed $\nu$. Thus, if $\nu_d$ is defined, as previously suggested, to be the smallest integer such that $L(\tau_{\nu_d,d}) \geq 0.95$, then $\nu_d$ is necessarily an increasing function of $d$. Put another way, in higher dimensions a $t$ density must be closer in shape to a Gaussian for its LA to be within 5% of the true integral value. Indeed, using this definition of $\nu_d$ in 72 dimensions resulted in the extremely high value $\nu_{72} = 25921$. The difference between the resulting $t$ density and its Gaussian approximation is small enough to be virtually invisible, but because this difference is compounded over a (typical) set of extremely high volume, it results in a sizable difference between integrals.

In light of these ideas, our first suggested design for a high-dimensional diagnostic uses a preliminary grid $\boldsymbol{s}^* = \{\mathbf{0}\} \cup \left\{\pm\sqrt{d}e_i : i = 1, \ldots, d\right\}$, where $\mathbf{0}$ denotes the origin and $e_i$ once again denotes the $i^{\text{th}}$ standard basis vector of $\mathbb{R}^d$. This will result in $2d + 1$ interrogation points: one at the mode, and two at distances of $\mathcal{O}\left(\sqrt{d}\right)$ away from it along each "principal axis" (see Section 2.4.1). Per the discussion above, if a function $f$ is assumed *a priori* to have similar shape to a Gaussian density, then it is reasonable to expect this type of design to provide the most pertinent information about its integral. As described by Särkkä et al. [256], this choice of $\boldsymbol{s}^*$ in BQ creates a connection with *sigma-point methods*, in which such grids are used to estimate integrals for filtering and smoothing in nonlinear SSM's [e.g. 145]. In particular, aside from the inclusion of the origin this choice of $\boldsymbol{s}^*$ is identical to the point set used in the *cubature Kalman filter* (CKF) of Arasaratnam and Haykin [5].

With this preliminary grid in $d = 72$ dimensions, we use $\tau_{25921,72}$ as our calibration function and once again take the hyperparameter $\gamma$ as in (2.19), resulting in a value of $\gamma = 1.2248$. As alluded to above, here $\lambda$ cannot be selected to visually ensure that Condition (2a) is met as in the low-dimensional experiments of Section 2.5.1. Because the differences between the calibration function and its Gaussian approximation are so small at the chosen interrogation points, adjusting $\lambda$ does not produce any visible change in the difference $m_1^x \cdot g - \tau_{25921,72}$ (not shown). Thus, we must rely on the weaker Condition 2(b): selecting $\lambda$ to produce a reasonable posterior integral estimate $m_1$. We found $\lambda = 3.7$ to be a good choice for this, giving a posterior integral mean of $m_1 = 0.998$. Finally, $\alpha = 0.1565$ is once again chosen to ensure that the calibration curve's LA (equal to 0.95) is on the boundary of the rejection region. Note that, although we were unable to use shape information as directly as we did in the low-dimensional experiments, the diagnostic's rejection criterion still depends solely on the "correction term" (2.16), itself a measure of deviation between a function and its Gaussian approximation. It could be said that the high-dimensional diagnostic, as it is configured here, determines whether a function is sufficiently Gaussian *in the tails* to justify the LA.

## 3.2 Example: North Sea cod modelling

This section returns to the SSM discussed at the beginning of the previous chapter (Sections 2.1 – 2.2). Recall that, given observed data $\boldsymbol{y}$, such a model can be fit by maximizing the Laplace-approximated marginal likelihood (integrating over hidden states $\boldsymbol{x}$) with respect to parameters $\theta$. These methods are increasingly common in fisheries science, where they are used for *stock assessment*: to infer population dynamics for various species of fish given observations from surveys and commercial catches [2]. SSM's applied to stock assessment are often called *state-space assessment models (SSAM's)* [ibid.] and serve as a natural context to test our diagnostic: although the LA is commonly used in practice for these models, if the joint likelihood (2.1) is not "sufficiently Gaussian" in shape, then the LA may not be a suitable proxy for the marginal likelihood (2.2) and the resulting inferences may be incorrect.

To investigate the performance of our diagnostic in this "real-world" setting, we use a dataset containing multiyear measurements of cod stocks in the North Sea and fit SSAM's to various subsets of this data following Aeberhard et al. [2]. The observations $y_t$ are taken on an annual basis over the span of several decades ($t = 1963, \ldots, 2015$). Briefly, for a given year $t$, $y_t$ is a vector comprising the amounts of cod of different ages observed during surveys and commercial catches conducted that year[1]. The hidden state $x_t$ contains, for each age group, the "true" abundance and fishing mortality rate for cod in that age group in year $t$. Finally, $\theta$ represents a variety of "global" parameters such as scaling factors and variances. The SSAM used here [see 208, and references therein] is highly nonlinear, with complex dependencies between the age-specific components of $x_t$ and $x_{t-1}$. For the sake of brevity further details are omitted here, but they are available in the appendix of Aeberhard et al. [2]. All models were fit using the stockassessment R package [208, 16], which is in turn built on the TMB package [165].

Two SSAM's are considered here, each corresponding to a different subset of the available data: one fit to the data collected from 1970 to 1975 (hereafter the "1970 model"), and another to the data from 2005–2011 (the "2005 model"). Since each hidden state $x_t$ is of dimension 12, using these six-year "windows" results in a latent dimensionality of $d = 12 \times 6 = 72$ for each model: fairly modest (and computationally convenient) compared to the 636 dimensions associated with the full dataset [2], but still large enough that any non-LA approach to marginalizing the likelihood would be far from trivial[2]. As stated above,

[1]Note that the dimensionality of $y_t$ is not constant with $t$, as the time ranges of the commercial catches and surveys only partially overlap. However, "missing observations" are not a problem for either model fitting or the diagnostic.

[2]We also found that, with smaller time windows, there was not sufficient data to guarantee model convergence. Even six-year windows besides the ones used here typically did not converge without careful selection of algorithm settings.

the Laplace-approximated marginal likelihood $L(p_{xy})$ is maximized numerically w.r.t. $\theta$, and ideally we would like to use our diagnostic at each step of this optimization to ensure it remains accurate throughout. For simplicity in these experiments, we only apply the diagnostic at the last optimization step, seeking to determine *only for the final parameter values* $\hat{\theta}$ whether $p_{xy}\left(\cdot, \boldsymbol{y} \mid \hat{\theta}\right)$ is "Gaussian enough" to justify the LA.

In order to assess the performance of the diagnostic, it is desirable to have some other estimate of the marginal likelihood $p_y\left(\boldsymbol{y} \mid \hat{\theta}\right)$ to serve as an approximate "ground truth". Since standard numerical integration is completely nonviable in 72 dimensions, we instead obtain such estimates via importance sampling [e.g. 95, and references therein]. For both models, samples were taken from a noncentral multivariate $t$ distribution with mean $\hat{x}$, scale matrix $-H^{-1}$, and 5 degrees of freedom [76]. The joint likelihoods of both models appear to have light tails in $\boldsymbol{x}$ (see below), so this choice of importance distribution should mitigate the risk of infinite variance in theory [297, 279]. However, because we can only assess the tail behaviour of the models in finitely many directions, we cannot rule out the possibility that, somewhere in the 72-dimensional space, they have a tail even heavier than that of a $t$ density. We conjecture that this is not the case, although the existence of such a tail could result in a sampler with infinite variance. A more pressing concern is that poor finite-sample performance can still occur even with theoretical guarantees. Nevertheless, importance sampling is not the main concern here — it is intended only as a convenient, if somewhat informal, check on the LA diagnostic.

This diagnostic is not the only way to check the LA for a SSM — the `checkConsistency` function in the TMB package [165] provides another method[3]. It is essentially a *score test* [231] for the Laplace-approximated marginal likelihood: by simulating many separate datasets $\boldsymbol{y}^* \sim p_y\left(\cdot \mid \hat{\theta}\right)$ (which can be done by simulating $\boldsymbol{x}^* \sim p_x$, then $\boldsymbol{y}^* \sim p_{y|x}$), it constructs a test statistic to test the hypothesis $\mathbb{E}_y\left[\nabla_\theta \log L(p_{xy})\,|_{\hat{\theta}}\right] = 0$, under which the statistic would be asymptotically $\chi^2$-distributed. Since the true marginal score function has mean zero, a rejection of this hypothesis means that the LA is *not* a suitable approximation for the marginal likelihood $p_y$. It will be useful to compare this method to our diagnostic, but it should be noted that there is a key conceptual difference between them. The `checkConsistency` methodology views $L(p_{xy})$ and $p_y$ as *functions of* $\boldsymbol{y}$; with this view, it seeks to determine whether the marginal likelihood is well approximated by the LA, and what effects this approximation could have on the bias of the estimated $\hat{\theta}$. In contrast, our diagnostic is focused on shape of the joint likelihood $p_{xy}$ when viewed *as a function of* $\boldsymbol{x}$: in particular, whether this shape warrants the use of the LA to fit the model *for the observed* (i.e. *fixed*) $\boldsymbol{y}$.

[3]Refer to the source code at `https://github.com/kaskr/adcomp/blob/master/TMB/R/checker.R` for further detail. Notes provided by Anders Nielsen in personal correspondence also helped to inform this discussion.

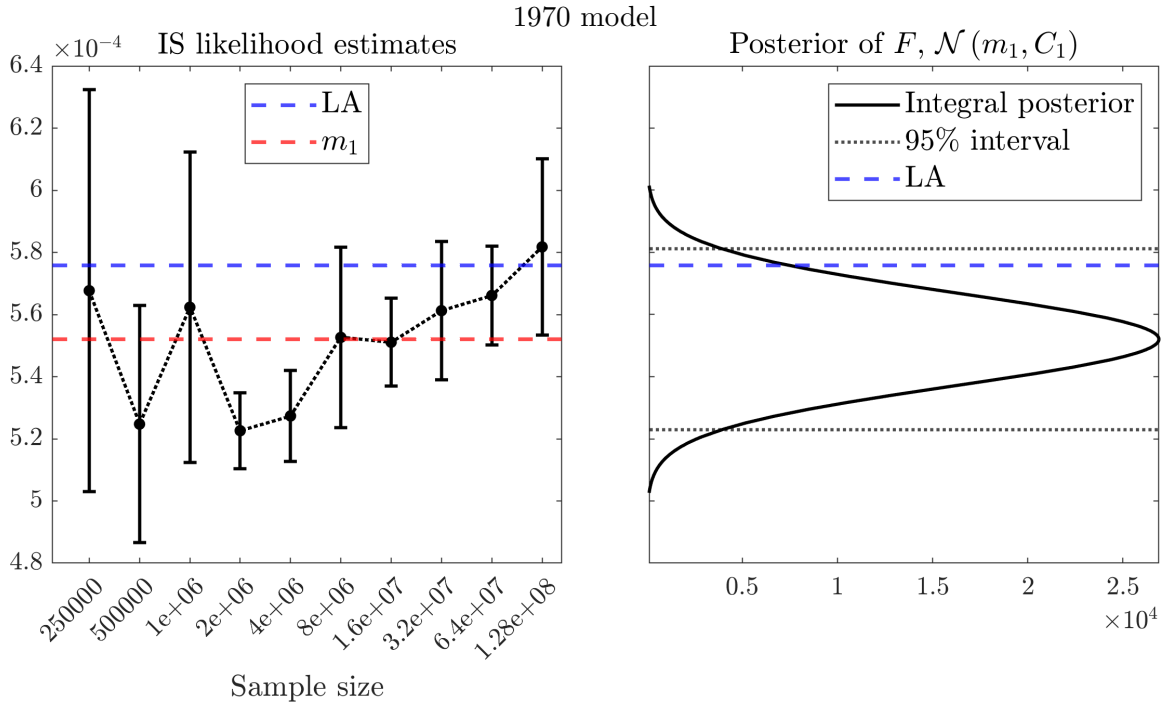


Figure 3.2: Results of the diagnostic applied to the 1970 SSAM. Left: IS estimates of $p_y\left(\boldsymbol{y} \mid \hat{\theta}\right)$ at various sample sizes (black dots) with estimated 95% confidence intervals (vertical line segments), with the Laplace approximation (blue dashed line) and the posterior integral mean (red dashed line) for reference. Right: the posterior distribution for the marginal likelihood, obtained from the diagnostic (rotated 90 degrees for ease of comparison with IS estimates).

Figure 3.2 shows results (from the diagnostic, as well as the aforementioned importance sampler with differing numbers of samples) for the 1970 model. For the importance samplers, 95% confidence intervals were obtained with a Gaussian approximation, using the sample variance of the IS weights. The central limit theorem dictates that for a well-behaved importance sampler, the width of these intervals should be roughly $\mathcal{O}\left(S^{-1/2}\right)$, where $S$ is the number of samples. The left plot of Figure 3.2 indicates that this may not be the case. Indeed, the score test of Koopman et al. [161] rejected the hypothesis that these samplers had finite variance. These rejections are typically the result of a few large weights, which seemingly indicate that in a few directions the tails of $p_{xy}\left(\cdot, \boldsymbol{y} \mid \hat{\theta}\right)$ are too heavy relative to those of the proposal density. However, further numerical evidence indicated that the tails of the squared joint likelihood were eventually dominated by its Gaussian approximation in those directions. In mathematical terms, at all sampled points $x \in \mathbb{R}^d$ for which the importance weights were large, it appeared that, for sufficiently large $r > 0$,

$$\left[p_{xy}\left(\hat{x} + rz, \boldsymbol{y} \mid \hat{\theta}\right)\right]^2 = o\left(\phi\left(\hat{x} + rz\right)\right) \tag{3.1}$$

as functions of $r$, where $\phi$ is the Gaussian approximation to $p_{xy}\left(\cdot, \boldsymbol{y} \mid \hat{\theta}\right)$ and $z$ is a unit vector in the direction of $x - \hat{x}$. Since the ratio of a Gaussian density and a Student's $t$ density is certainly integrable over $\mathbb{R}^d$, this provides some limited indication that the integral defining the variance of the importance sampler [e.g. 76] may indeed be finite after all. This is a very informal check on the validity of IS, and it does not guarantee finite-sample stability. However, their use as a heuristic reference against which to check the diagnostic does not seem unreasonable here.

Most of the importance samplers include the LA within their 95% confidence intervals, suggesting it is not excessively far from the true marginal likelihood value. The fact that most of the IS estimates are below the LA suggests that the latter is perhaps a slight overestimate of the true value (i.e. that the tails of the joint likelihood, as a function of $x$, tend to be lighter than those of its Gaussian approximation). Our diagnostic produces a similar conclusion: the posterior integral mean is slightly lower than the LA, but not to a degree that warrants rejection. With respect to our notion of "good-enough-ness-of-fit", it seems that the LA is a reasonable approximation to the marginal likelihood for this model, at least for the parameter values $\hat{\theta}$.

Since the diagnostic is based on a Gaussian "confidence interval" for the integral (see Section 2.3), its behaviour can be equivalently described in terms of "p-values": recalling from (2.10) that the integral posterior is $F \mid r(\boldsymbol{s}) \sim \mathcal{N}\left(m_1, C_1\right)$, it is straightforward to show that the diagnostic rejects the LA iff

$$2\left[1 - \Phi\left(\frac{|m_1 - L(f)|}{\sqrt{C_1}}\right)\right] < 0.05,$$

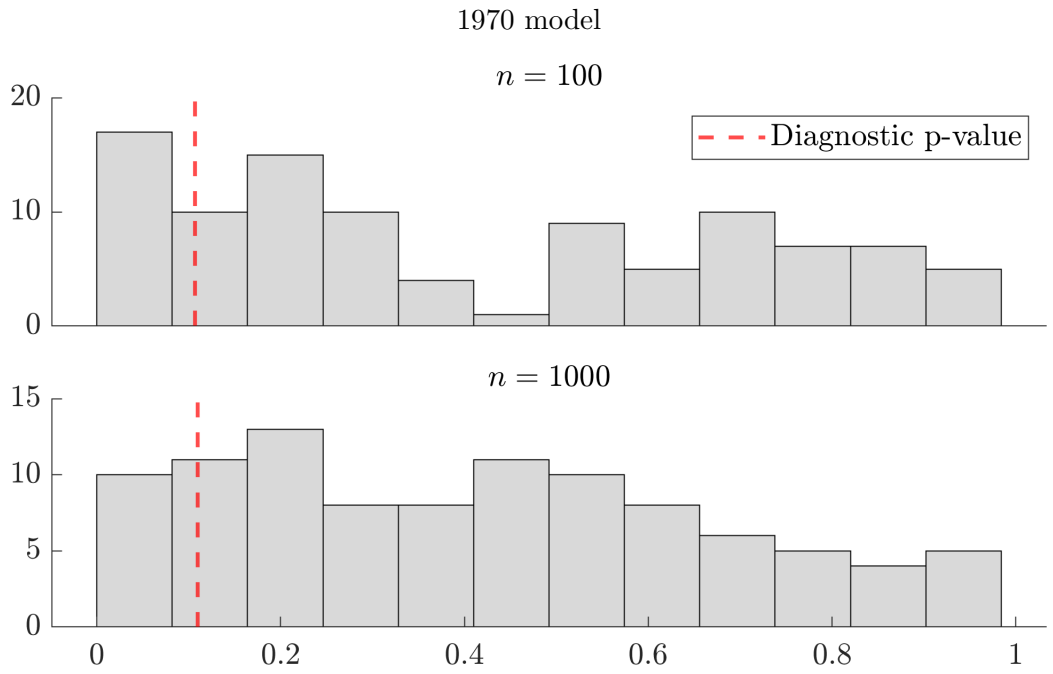


Figure 3.3: Histograms of p-values from repeated runs (100 runs each for simulation sizes $n = 100$ and $n = 1000$) of `checkConsistency` on the fitted 1970 SSAM. The "p-value" given by the diagnostic is shown as a dashed red line on each histogram.

where $\Phi$ is the c.d.f. of a standard normal random variable, and the quantity on the left-hand side has a natural interpretation as a sort of "p-value". This facilitates some comparison between the diagnostic and the `checkConsistency` method. Recall that the latter simulates $n$ separate datasets to construct a test statistic that is asymptotically $\chi^2$-distributed when $\mathbb{E}_y \left[ \nabla_\theta \log L\left(p_{xy}\right) \big|_{\hat{\theta}} \right] = 0$. This test statistic induces a p-value; if this is below some threshold (say, 0.05), we reject the hypothesis that the marginal likelihood and the LA are the same (as functions of $\boldsymbol{y}$). In Figure 3.3, we have performed the `checkConsistency` test 100 times each for two simulation sizes ($n = 100$ and $n = 1000$) in order to see how the p-value distribution changes with the number of simulated datasets and how it relates to the p-value of the diagnostic. If the null hypothesis of `checkConsistency` is true (i.e. the LA is the true marginal likelihood), then the p-value of the corresponding test should be uniformly distributed over $(0, 1)$. Although the histograms in Figure 3.3 show some deviation from uniformity, it is not severe. The p-value associated with the diagnostic is just above 0.1, consistent with non-rejection of the LA (see Figure 3.2). It is interesting to see from Figure 3.3 that the diagnostic and `checkConsistency` seem to lead to similar conclusions — that the LA may deviate slightly from the true marginal likelihood, but not to a problematic extent — despite the fundamental difference in the questions addressed by each method.

The results are markedly different for the 2005 model, as shown in Figure 3.4. IS stability considerations apply here as they did for the 1970 model: Koopman et al.'s score test [161] rejected the hypothesis of finite variance for the largest sample sizes, but (3.1) appeared to

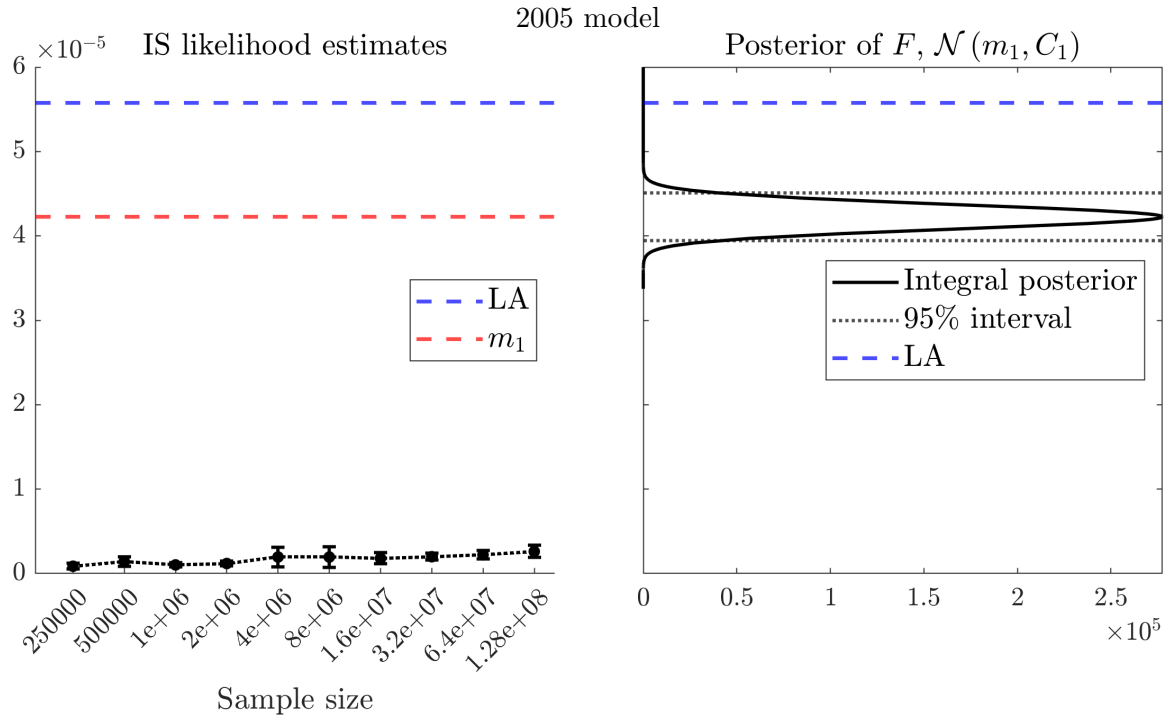


Figure 3.4: Results of the diagnostic applied to the 2005 SSAM. Left: IS estimates of $p_y\left(\boldsymbol{y} \mid \hat{\theta}\right)$ at various sample sizes (black dots) with estimated 95% confidence intervals (vertical line segments), with the Laplace approximation (blue dashed line) and the posterior integral mean (red dashed line) for reference. Right: the posterior distribution for the marginal likelihood, obtained from the diagnostic (rotated 90 degrees for ease of comparison with IS estimates).

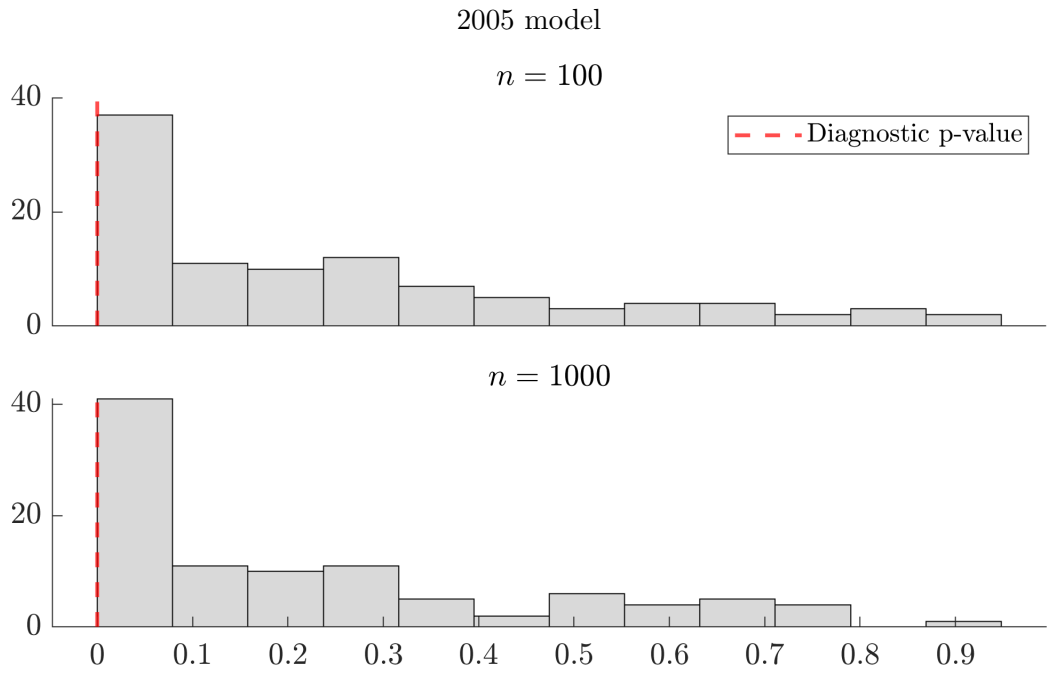


Figure 3.5: Histograms of p-values from repeated runs (100 runs each for simulation sizes $n = 100$ and $n = 1000$) of `checkConsistency` on the fitted 2005 SSAM. The "p-value" given by the diagnostic is shown as a dashed red line on each histogram.

hold in the directions of all the largest weights, potentially indicating a finite (but quite large) variance. All IS estimates are far lower than the LA, suggesting that the joint likelihood is, for the most part, substantially lighter-tailed than its Gaussian approximation. Accordingly, the diagnostic strongly rejects the LA, which is well above the upper bound of the posterior 95% confidence interval. Note that there is still substantial disagreement between the diagnostic and the importance samplers as it pertains to estimation of the true marginal likelihood. Thus, the posterior integral mean from the diagnostic should not be taken as a high-quality estimate, but what is important is that both methods agree on rejection of the LA.

As before, we also conduct repeated runs of `checkConsistency` and compare the resulting p-value distributions to the one associated with the diagnostic. The latter is numerically indistinguishable from zero, and for both simulation sizes the p-value distribution is decidedly non-uniform. As was the case with the 1970 model, both methods appear to agree that the LA is an unsuitable approximation to the marginal likelihood, despite asking this question in different ways.

Differing philosophies notwithstanding, one clear advantage the diagnostic has over `checkConsistency` is computation time. Using the `checkConsistency` replications shown in Figures 3.3 and 3.5, as well as 100 repeated computations of the diagnostic itself, Table 3.1 shows median computation times — along with median absolute deviations — for each method applied to each model. All computations were performed on a computer with 64

| Time (seconds) | 1970 model | 2005 model |
|---|---|---|
| `checkConsistency`, $n = 100$ | $2.511 \pm 0.035$ | $7.367 \pm 0.136$ |
| `checkConsistency`, $n = 1000$ | $25.115 \pm 0.152$ | $73.584 \pm 0.489$ |
| Diagnostic | $0.009 \pm 0.007$ | $0.012 \pm 0.0003$ |

Table 3.1: Table showing median computation times (along with median absolute deviations) of each method, applied to each model.

GB of RAM and eight Intel i7-6400K 4GHz CPU cores. Note that the time cost for the diagnostic includes the evaluation of function interrogations, the eigendecomposition of the Hessian, and the calculation of all the necessary kernel terms for BQ (the latter step was sped up substantially using the methods of Karvonen and Särkkä [150], as explained in Section 3.2.1). It is also interesting to note the differences in computational times between models: across all methods, the times for the 2005 model are longer than those for the 1970 model. Presumably, this is because of the "inner" numerical optimization [165] used to calculate the mode $\hat{x} = \hat{x}\left(\boldsymbol{y}, \hat{\theta}\right)$, which may require more iterations for the 2005 model than the 1970 model due to differences in their respective joint likelihoods. This would also explain why the difference is so much more pronounced for the `checkConsistency` runs, which require repeated (and possibly even more demanding) inner optimizations to find $\hat{x} = \hat{x}\left(\boldsymbol{y}^*, \hat{\theta}\right)$ for each simulated dataset $\boldsymbol{y}^*$. In any case, the diagnostic is by far the fastest method of assessing the LA[4].

### 3.2.1 Higher-order interrogation grids

The interrogation grids used thus far have been quite simple, consisting of $\mathcal{O}(d)$ preliminary points placed along the axes of $\mathbb{R}^d$ in a $d$-dimensional "cross" shape. As noted in Section 3.1, there is precedent in the literature for the use of such simple grids [256, 5]. They seem to be a reasonable choice here as well, allowing us to calibrate the diagnostic in such a way that appropriate results are obtained for a variety of "toy" and real-world examples. However, one potential drawback of such grids is that they only allow the diagnostic to use information about a function's shape along its "principal axes" (see Section 2.4.1). If this is not indicative of the function's behaviour in the rest of the domain, it is conceivable that the diagnostic could produce misleading results. For instance, consider the $d$-dimensional function

$$f_{\nu,d}(x) = \prod_{i=1}^{d} \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1 + \frac{x_i^2}{\nu}\right)^{-\frac{\nu+d}{2}}. \tag{3.2}$$

[4]IS computation times are not shown, as these were not replicated. However, they behaved largely as expected: computation times were roughly linear in the number of samples, and universally longer than those for the diagnostic.

Like the multivariate $t$ density (2.17), this function has a mode at the origin. The functions are equal there, as are the Hessians of their logarithms. Furthermore, they are equal along the axes of $\mathbb{R}^d$. Thus, their LA's are the same, and the diagnostic would give the same results for both functions using any of the "cross-shaped" interrogation grids considered above. However, the functions differ on the rest of their domain, and their integrals are different as a result. Whereas the integral of $\tau_{\nu,d}$ over $\mathbb{R}^d$ is equal to 1 for all $(\nu, d)$, the integral of $f_{\nu,d}$ is

$$\frac{\Gamma\left(\frac{\nu+d-1}{2}\right)^d}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\nu+d}{2}\right)^{d-1}}.$$

In particular, for $d = 72,\ \nu = \nu_{72} = 25921$ (the values used to calibrate the 72-dimensional diagnostic at the beginning of this chapter), $\int_{\mathbb{R}^{72}} f_{25921,72}(x)\mathrm{d}x = 0.952$. Thus the integral of $f_{25921,72}$ is quite a bit closer to the LA (0.95) than that of the calibration function $\tau_{25921,72}$, but the diagnostic calibrated with a "cross-shaped" grid will treat both of them identically, so that the LA is on the boundary of the rejection region for each function. One could argue that this is undesirable: the values of $f_{25921,72}$ "off the axes" are lower (and therefore, closer to the Gaussian approximation) than those of the calibration function, causing its integral to be closer to the LA, so perhaps the diagnostic should produce a more definitive non-rejection for this function. For this to be possible, we must be able to capture the differences between $f_{\nu,d}$ and $\tau_{\nu,d}$, for which a *higher-order* interrogation grid is required.

A grid of "order" $s$ is one whose size scales as $\mathcal{O}\left(d^s\right)$ for some fixed power $s > 1$ (the grids used throughout the chapter thus far and in Chapter 2 had $s = 1$). In order to use such grids without an excessive increase in computation time (which would defeat the purpose of the diagnostic), we use *fully symmetric kernel quadrature* (FSKQ), as detailed by Karvonen and Särkkä [150]. Briefly, because the squared exponential kernel is isotropic, using fully symmetric preliminary grids (as described in Section 2.4.1) reduces the number of *unique* quadrature weights that need to be calculated, allowing for significant algebraic and computational simplifications in BQ.

Here, we conduct a few experiments with higher-order grids, showing difficulties associated with their use. We recalibrate the 72-dimensional diagnostic using a *sparse Gauss-Hermite grid of order 2* — the two-dimensional version of which is shown in Figure 3.6 — as the preliminary grid. Following Karvonen and Särkkä [150], we remove the origin, as its quadrature weight tends to be a large negative value for most hyperparameter combinations. Furthermore, because a function is always equal to its Gaussian approximation at the mode, the origin does not actually contribute to the diagnostic beyond its effect on the inverted Gram matrix. We also multiply each point in the Gauss-Hermite grid by 3.6, thereby ensuring that they are far enough away from the origin to cover the "typical set" discussed in Section 3.1. The final preliminary grid in 72 dimensions is of size $n = 10512$,

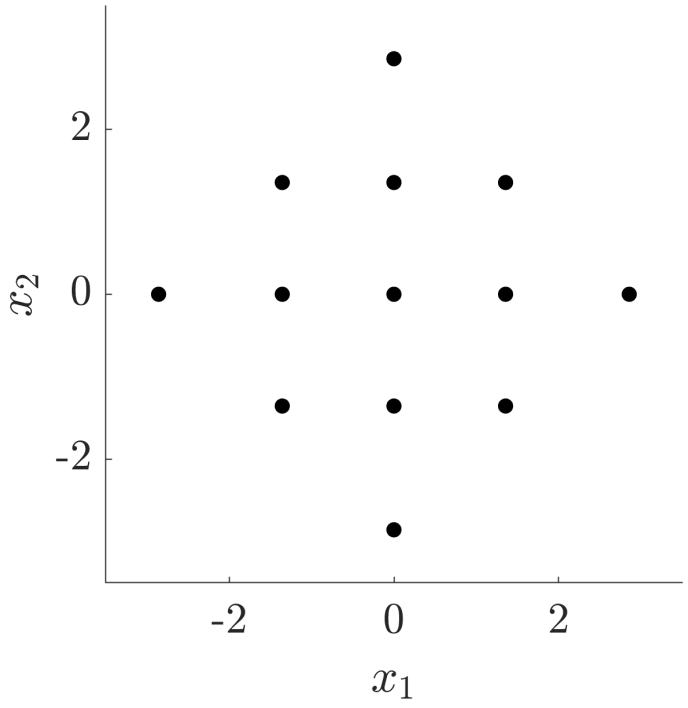


Figure 3.6: A sparse Gauss-Hermite quadrature grid of order 2 in $d = 2$ dimensions.

and as with the original "cross-shaped" preliminary grid (which, for reference, contained $n = 145$ points) we calibrate the diagnostic using the $t$ density $\tau_{25921,72}$ and taking the hyperparameter $\gamma = 1.2248$. As before, it is not possible to calibrate with respect to Condition (2a) from Section 2.5. Here, this is because of the size of the grid: the computational simplifications of FSKQ are only applicable to the integral of the GP, not to the GP posterior mean function (2.8) itself. As such, the visual calibration of Section 2.5.1 is not viable: even though we would only need to view a 2-dimensional slice of $m_1^x \cdot g - \tau_{25921,72}$, every change to the hyperparameter $\lambda$ would still necessitate the recalculation and inversion of the $10512 \times 10512$ Gram matrix, which is too slow for minute visual adjustments. Instead, we once again calibrate with respect to Conditions (1) and (2b), resulting in hyperparameters $(\lambda, \alpha) = (3.7, 0.1349)$ and a posterior integral mean of $m_1 = 0.9945$ for the calibration function.

Applying the new calibrated diagnostic with the larger preliminary grid to the SSAM's from Section 3.2 reveals that the use of higher-order grids does not necessarily cause an improvement in the diagnostic's behaviour in practice — indeed, the opposite may occur. The left plot of Figure 3.7 shows that, in contrast to the results in Section 3.2, this version of the diagnostic rejects the LA for the 1970 model. Initially, this may suggest that the tails of the joint likelihood are substantially lighter than those of its Gaussian approximation in directions besides its "principal axes", which would not have been observable using the

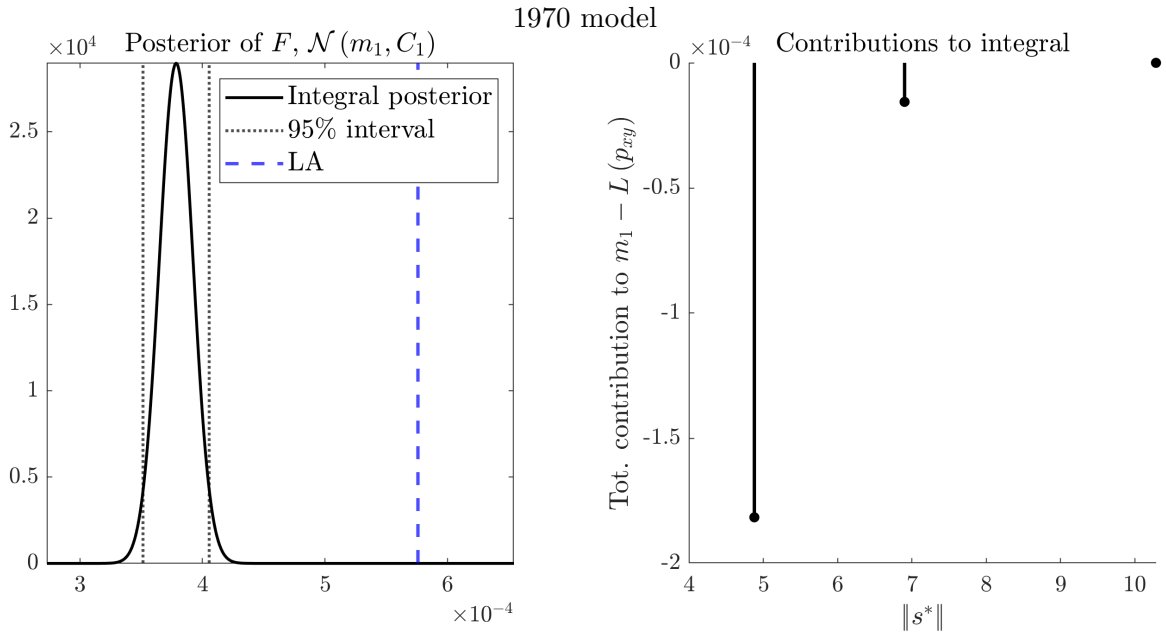


Figure 3.7: Results of the diagnostic with a higher-order interrogation grid applied to the 1970 SSAM. Left: the posterior distribution for the marginal likelihood, obtained from the diagnostic. Right: the total mass contributions to the quadrature estimate made by interrogations as a function of the distance between the corresponding preliminary points and the origin.

smaller grid. However, this is at odds with the results of the importance samplers and `checkConsistency`, both of which suggested that the LA was *not* very far from the true marginal likelihood and neither of which is constrained to the use of information on the principal axes of the joint likelihood. Furthermore, the right plot of Figure 3.7 reveals that the largest overall contribution to the lowered integral estimate comes from the interrogation points which are closest to the mode. This is despite the fact that there are only 144 such points in the Gauss-Hermite grid. In contrast, the points further from the origin — of which there are 10368 — collectively contribute a much smaller amount to the estimate. As discussed in Section 3.1, the integral of a high-dimensional function is mainly determined by the behaviour of its tails; ideally this would be reflected when using a preliminary grid with most of its points far away from the origin. In light of these considerations, it seems reasonable to conclude that this version of the diagnostic is not providing accurate inference on the integral, or on the function shape information most pertinent to it.

The new diagnostic exhibits a different problem when applied to the 2005 model, as seen in Figure 3.8. The left plot shows that the LA is once again definitively rejected, although the actual integral posterior differs quite noticeably from the one in Figure 3.4. However, as it turns out, there is one interrogation point $s$ where the weighted difference $r(s) - m_0^x(s)$ is far larger than it is for any of the other points. Removing this point from

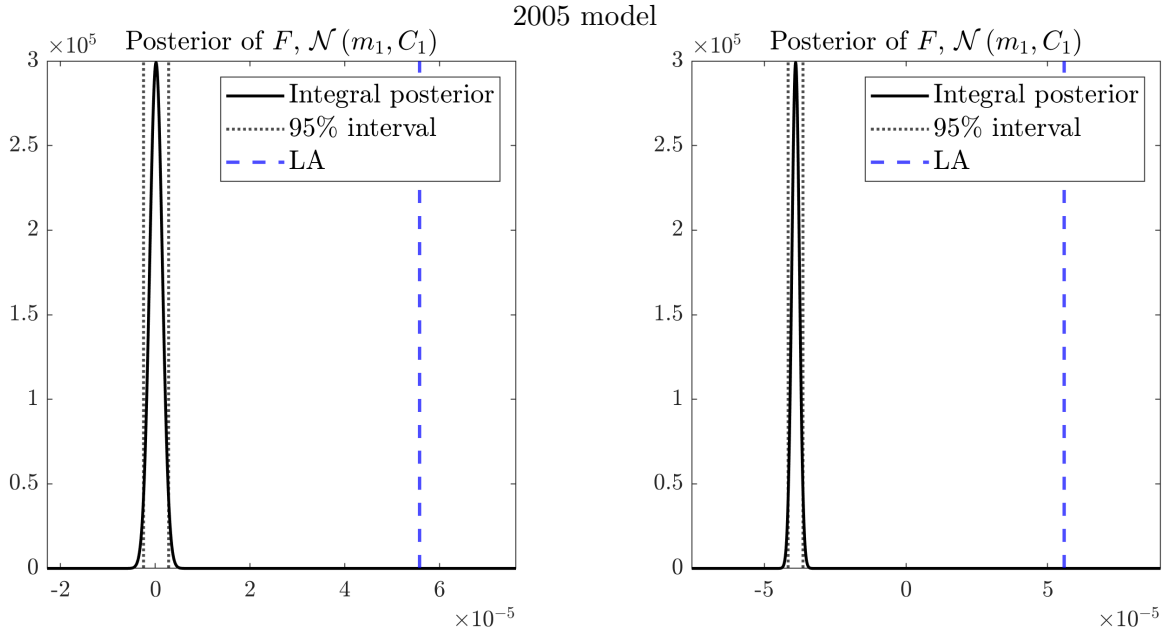


Figure 3.8: Results of the diagnostic with a higher-order interrogation grid applied to the 2005 SSAM. Left: the posterior distribution for the marginal likelihood, obtained from the diagnostic. Right: the same, but with a single interrogation point having been removed.

the grid, but keeping the hyperparameters fixed[5], results in a surprisingly large change in the posterior, shifting its mean from a small positive value to a larger negative value (which is nonsensical, given that the integral is a likelihood and must therefore be nonnegative). Although the diagnostic achieves its primary goal in both cases for this model — namely, determining that the joint likelihood's shape (as a function of $\boldsymbol{x}$) is too non-Gaussian to justify the LA — it is certainly undesirable for one interrogation point to have such a large impact. If this were allowed, a given function's LA could be rejected based solely on the inclusion or exclusion of a single point at which it deviates significantly from its Gaussian approximation, thereby rendering the diagnostic too sensitive to be useful for nontrivial high-dimensional applications (see the discussion at the beginning of Section 2.4).

At first, the failure of the diagnostic with high-order interrogation grids seems illogical. Intuitively, one would expect more accurate quadrature with larger grids. Indeed, several convergence theorems in the BQ liteature suggest that the addition of more points should be an asset [e.g 25, 150, 26]. However, these results tend to assume that the kernel and integrating measure are fixed. Here, we change both through our calibration of the hyperparameters, a necessary step in fulfilling the goals of the diagnostic. In this instance, asymptotics fail to guarantee the type of practical, finite-sample behaviour we require. De-

[5]Note that deleting the corresponding preliminary interrogation point did not produce a sizeable change in the diagnostic's behaviour when applied to the calibration function (not shown), despite not adjusting the hyperparameters for the altered grid. Thus, there is no concern about miscalibration here.

spite the potential shortcomings of lower-order grids, they seem to be a better choice in terms of ensuring a usable diagnostic, unless great care is taken with higher-order grids.

The computation times for the diagnostic with the higher-order grid are predictably higher than they were for the original diagnostic, although it is still much faster than `checkConsistency`. The median time was 0.4154 seconds for the 1970 model (MAD: 0.0105 seconds) and 0.5422 seconds for the 2005 model (MAD: 0.0104 seconds). Nevertheless, given the difficulties encountered above, the simpler CKF-style grid used in Section 3.2 seems to be a better choice.

## 3.3 Discussion

In this chapter and the last, we have built on the work of Zhou [301] to develop a non-asymptotic diagnostic tool for assessing the viability of Laplace approximations to integrals. More specifically and accurately, the diagnostic assesses whether a function's shape is close enough to the Gaussian approximation that is used to motivate the LA. It does so using the method of Bayesian quadrature, but in multiple ways it is structured differently than a more "conventional" BQ application. Namely, we avoid design choices that would ensure accurate, low-uncertainty estimates for the integral of a specific function, opting instead for a "one-size-fits-all" approach: relatively simple interrogation grids intended to capture the most pertinent information about a function's behaviour, hyperparameters chosen heuristically using calibration functions, and a covariance structure that ensures the diagnostic is invariant to all properties of the integrand besides its shape. More broadly, the diagnostic is based on a notion of "good-enough-ness-of-fit" that stands in stark contrast to a more conventional, power-focused approach to statistical inference. Indeed, such an approach would render the diagnostic useless, causing it to prioritize the detection of *any* deviation from Gaussian shape and likely producing rejections in almost all non-trivial applications.

As shown in this chapter, challenges arise when using the diagnostic in high dimensions, although they are not insurmountable. Compared to the low-dimensional settings of Chapter 2, it is more difficult to make conclusions about a function's integral given limited information about its shape — either because a high-dimensional function's mass tends to be far away from the regions with the most notable "shape information" (the curse of dimensionality), or because a single direction of non-Gaussian shape (which, intuitively, seems more likely to occur in high dimensions) can affect the diagnostic's behaviour to an unreasonable extent. Because of these challenges, more consideration must be given in high-dimensional spaces when choosing the preliminary interrogation grid and setting the hyperparameters, and the focus must be on the function's shape in its tail regions, assumed to correspond to its "typical set". If this is done carefully, the diagnostic can be calibrated to produce reasonable and useful results on real-world examples, as shown in Section 3.2.

Given SSAM's that had already been fit (producing parameter estimates $\hat{\theta}$), we applied the diagnostic to their joint likelihoods $p_{xy}\left(\cdot, \boldsymbol{y} \mid \hat{\theta}\right)$. While this served the purposes of this thesis (namely, a proof-of-concept for the diagnostic itself), it ignores the fact that the parameter estimate itself depends on the use of Laplace approximations: specifically, that it is obtained by maximizing the LA $L\left(p_{xy}\left(\cdot, \boldsymbol{y} \mid \theta\right)\right)$ with respect to $\theta$. Given the low computational cost of the diagnostic, it would be desirable to fold it directly into a model-fitting workflow, checking at each iteration of numerical optimization whether or not the LA is justified, thereby indicating if other methods need to be invoked to correct any incurred bias in the estimated model parameters.

Despite the promising initial performance of the diagnostic, there are opportunities for future potential improvements. The difficulties of using higher-order grids encountered in Section 3.2.1 should be further explored, as their resolution could result in improved diagnostic behaviour on a wider variety of functions. The methods of choosing interrogation points cited in the introduction of Section 2.4 may be a useful starting point to this end, but care must be taken to modify these methods in a way that preserves the quick, "one-size-fits-all" nature of the diagnostic. Another aspect of the diagnostic that remains unaddressed is the prior structure: specifically, that our use of a GP prior is *technically* inappropriate given that most applications involve likelihoods, which are nonnegative. It is worth investigating other prior specifications proposed in the BQ literature [e.g. 116, 45], which preserve nonnegativity of the integrand at the expense of inducing a non-analytic distribution on the integral which must be approximated.

As a final note, we conjecture that the methods developed here may be more broadly applicable beyond the assessment of Laplace approximations. Indeed, a great deal of statistical methods are based on an assumption that some function is well approximated by a Gaussian shape, which is precisely the assumption that the diagnostic is designed to check. The general idea of using non-asymptotic methods to diagnose the use of asymptotic methods is one that warrants further consideration and study.

# Chapter 4

# A review of uncertainty quantification for density estimation

This chapter is an adaptation of a standalone manuscript originally published in Statistics Surveys [197].

## 4.1 Introduction

Density estimation is one of the seminal examples of nonparametric statistical modelling. There are a litany of methods spread across decades of literature, from more "classical" approaches [223] to the most advanced modern techniques [38]. Estimation, however, is only one piece of the puzzle: as in any statistical problem, it is desirable to also conduct *inference*, providing some quantification of uncertainty in addition to single estimates. Broadly speaking, uncertainty is quantified using sets of "plausible" values — for example, confidence intervals for frequentist methods and credible intervals for Bayesian ones. Although not as abundant as other areas in nonparametric statistics, there is a sizeable body of literature on uncertainty quantification (UQ) for density estimation, ranging from rigorously theoretical to extremely practical.

The following sections provide more detail on various types of "uncertainty sets", then outline several density estimation methods and review available literature dealing with UQ for each one. Although some combinations of estimation and inference ideas are not represented in the literature (in particular, a substantial gap exists between theoretical and practical UQ developments in many cases), in principle, one could *always* obtain some kind of uncertainty bounds on a density estimate, either by bootstrapping a frequentist method or taking quantiles of MCMC output for a Bayesian one. Whether or not such bounds have suitable coverage properties or otherwise perform adequately is another question for which the answers are not always known. Despite some of these limitations, this chapter presents

a comprehensive review of the work done thus far in unknown density UQ, and suggests promising areas to extend the research or "fill in the gaps".

## 4.2 Overview and notation

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a set of samples from some unknown "true density" $f_0$. The majority of discussion here will assume i.i.d. samples, but other data structures will also be considered as warranted. One structure that is common enough to justify mentioning here is the case of "noisy" observations $Y_i = X_i + Z_i$, where the errors $Z_i$ have known distribution and the true $X$-values are unknown. Estimating the density of $\boldsymbol{X}$ in this case is called *deconvolution density estimation*. In the present context, $\hat{f}$ will denote a *specific* "point" estimate (in the sense that it is a *single element of a function space*) of $f_0$, such as a MLE or posterior mean; while $f$ will typically be used to discuss classes or function spaces of estimators in more generality.

As mentioned in the introduction, UQ arises by considering "uncertainty sets". Such sets are random through their dependence on $\boldsymbol{X}$, but for brevity the notation here does not reflect this. As $f_0$ is a function, there are several ways to define uncertainty sets, each with different implications and advantages. Perhaps the most obvious examples are *pointwise intervals* $C_x = [L(x), U(x)]$, defined separately for each point $x$ in the domain of $f_0$. A common special case is when the intervals are symmetric about an estimator $\hat{f}$: $C_x = \left[\hat{f}(x) - \epsilon_x, \hat{f}(x) + \epsilon_x\right]$. The goal with pointwise intervals is to achieve (possibly only approximately or asymptotically) $\mathbb{P}\left(f(x) \in C_x\right) \geq 1 - \alpha$ for all $x \in \mathrm{Dom}\left(f_0\right)$, where $1 - \alpha$ is the usual predetermined level. The meaning of the generic placeholders $\mathbb{P}$ and $f$ depends on whether the inference is frequentist or Bayesian.

Pointwise intervals tend to be easy to implement, and also have nice theoretical properties for some (but not all) density estimation techniques. However, they are fundamentally limited in their ability to make "global" uncertainty statements. For a given level $1 - \alpha$, even if $\mathbb{P}\left(f(x) \in C_x\right) \geq 1 - \alpha \ \ \forall x$, the stronger and perhaps more meaningful statement $\mathbb{P}\left(f(x) \in C_x \ \ \forall x\right) \geq 1 - \alpha$ cannot necessarily be deduced. However, in some cases the "simultaneous" statement *does* hold, in which case the set $C = \{(x, y) : x \in \mathrm{Dom}\left(f_0\right), y \in C_x\}$ is called a *confidence* or *credible band*. Like pointwise intervals, bands are often centered at a specific estimator, although this need not be the case. For instance, Hall and Titterington [125] proposed to construct frequentist bands for univariate densities based on simultaneous multinomial confidence intervals for the probability masses within consecutive subintervals of the domain. Classical approximation results allowed them to construct such intervals without a specific density estimator, and they constructed the density bands by interpolation, with modifications depending on $f_0$ being once or twice differentiable. Such bands are not smooth, but were shown to have suitable coverage properties and optimal asymptotic widths without making further assumptions or restrictions. Hengartner

and Stark [130] devised conservative confidence bands for shape-restricted densities (either monotonic or having $\leq k$ modes for known $k$, possibly relative to some weight function), also obtained without an estimator. To derive their bands, they started with a confidence region for the c.d.f $F_0$ comprised of distributions with densities having the same shape restriction as $f_0$, and subsequently showed how to reduce the determination of the band for $f_0$ to a finite-dimensional linear program while conservatively preserving coverage probability.

If $C_x$ has the same width for all $x$ in a band, then it is uniform and is therefore a $L^\infty$-ball in a suitable function space $\mathcal{F}$. Thus, a uniform band is a special case of a more general idea: using a ball $C$ in some pseudo-metric space of functions $(\mathcal{F}, d)$ as an uncertainty set. Analogously to pointwise intervals and bands, here the goal is to have $\mathbb{P}(f \in C) \geq 1-\alpha$. For choices of $d$ such as the Hellinger or $L^2$ distances, such sets arise in nonparametric literature due to their satisfying theoretical properties. However, their practicality is somewhat limited: an $L^2$-ball of functions, for instance, does not provide error bounds that can be easily visualized or understood, short of simply plotting a large number of functions from the ball alongside $\hat{f}$. For example, Szabó et al. [276] visualized $L^2$-balls from an empirical Bayesian model for nonparametric regression by sampling functions from the posterior and plotting the $100(1-\alpha)\,\%$ of draws closest in the $L^2$ sense to the posterior mean. In a discussion of this paper, Low and Ma [187] suggested using this procedure to generate bands for the regression function whose boundaries are simply the pointwise maxima and minima of these closest posterior draws. Their simulations showed that the bands thus obtained performed quite well with respect to the framework of Cai et al. [35]. Beyond the aforementioned examples and those in Section 4.4.4, discussion of these "uncertainty balls" is limited, although Chapter 6 of Csörgo and Révész [58] contains theorems on the asymptotic distributions of the $L^2$-errors of several "classical" frequentist estimators (KDE's, histograms, and certain orthonormal basis expansions). These results *could* be relevant towards the construction of confidence balls, but this seems not to have been done in practice, likely due to their limited visual utility. On the other hand, if $d$ is the $L^\infty$ distance, one recovers the meaningful and easily-visualized UQ given by bands, at the expense of nice theory in some cases. As before, for *any* pseudo-metric a common special case arises by taking the associated sets to be centered at some estimator:

$$C(\epsilon) = \left\{ f \in \mathcal{F} : d\left(f, \hat{f}\right) < \epsilon \right\}. \tag{4.1}$$

In frequentist inference, uncertainty quantification relies on *confidence sets* of any of the forms described above, typically obtained in practice using asymptotic arguments and/or bootstrapping. Confidence sets are designed in view of the "ground truth" $\boldsymbol{X} \sim f_0$: letting $\mathbb{P}_0$ denote the probability law associated with $f_0$, the goal is to achieve *coverage probability* $\mathbb{P}_0(C_x \ni f_0(x)) \geq 1-\alpha \;\; \forall x$ in the pointwise case, or $\mathbb{P}_0(C \ni f_0) \geq 1-\alpha$ for bands or function balls. The Bayesian approach employs *credible* sets instead: using $\Pi(\cdot \mid \boldsymbol{X})$ as

generic notation for the posterior over a space of densities $f$, the sets of interest are either pointwise intervals such that $\Pi(f(x) \in C_x \mid \boldsymbol{X}) \geq 1-\alpha \ \ \forall x$, or bands/balls such that $\Pi(f \in C \mid \boldsymbol{X}) \geq 1-\alpha$. To facilitate validation and comparison, it is possible to view Bayesian methods through a frequentist lens by acknowledging the existence of the "ground truth" $f_0$, in which case the posterior $\Pi(\cdot \mid \boldsymbol{X})$ is considered as a random measure due to its dependence on $\boldsymbol{X} \sim \mathbb{P}_0$. This leads to a similar interpretation of credible sets as functions of the data. It is then natural to ask if they achieve coverage in the aforementioned frequentist sense. Put another way, can credible sets also serve as valid confidence sets? The difficulty of answering this question for nonparametric Bayesian methods is well-known and an active area of research; discussion of coverage therefore tends to be easier in the frequentist paradigm.

Naturally, the best possible inference produces small sets with high coverage probability. To this end, the concepts of *honesty* and *adaptivity* are relevant. Consider a confidence set $C_n$, where the subscript $n$ is added to emphasize limiting behaviour with respect to sample size. The remainder of this section ignores the distinction between pointwise intervals, bands, and balls.

In the context of density estimation, $C_n$ is *honest* at level $1-\alpha$ if

$$\liminf_n \inf_{f_0 \in \mathcal{F}} \mathbb{P}_0(C_n \ni f_0) \geq 1-\alpha, \tag{4.2}$$

where $\mathcal{F}$ is once again a suitable function space of interest [133]. In words, an honest confidence set asymptotically achieves the desired coverage level *uniformly* over all possible "ground truths". Honesty is crucial for practical finite-sample inference: without it, it is possible in some cases for the infimum of coverage probability over $\mathcal{F}$ to be zero for *any* $n$ [174].

The precise definitions and presentations underpinning the notion of *adaptivity* vary throughout the nonparametric literature [e.g. 34, 100, 133]. The present discussion will focus as narrowly as possible on material relevant to density UQ. Suppose $\mathcal{F} = \cup_{s \in \mathcal{S}} \mathcal{F}_s$ for some ordered index set $\mathcal{S}$, where for $s > t$ it holds that $\mathcal{F}_s \subseteq \mathcal{F}_t$ and the elements of $\mathcal{F}_s$ are smoother than those of $\mathcal{F}_t \setminus \mathcal{F}_s$. Typically each subset $\mathcal{F}_s$ is, say, a ball in a suitable Besov space of regularity $s$, with an associated minimax-optimal contraction rate $r_n(s)$ decreasing in both $n$ and $s$ [101]. Following Hoffmann and Nickl [133], call $C_n$ *adaptive* if there exists $L > 0$ such that, for all $s \in \mathcal{S}$ and for all $n$ large enough,

$$\sup_{f_0 \in \mathcal{F}_s} \mathbb{E}_0 |C_n| \leq L r_n(s), \tag{4.3}$$

where the expectation is with respect to $\mathbb{P}_0$, and $|C_n|$ is the diameter of $C_n$ with respect to the metric by which it is defined (typically $L^2$ or $L^\infty$ in this context). Naturally, less uncertainty is expected in the estimation of smoother functions. Adaptive confidence sets

take advantage of this fact: they are optimal in the sense that, for every level of smoothness under consideration, their "maximum" expected size contracts at the optimal rate. This is especially useful since the actual smoothness of the true density is likely to be unknown, and it does not have to be specified for adaptive $C_n$. Unfortunately, adaptivity is an elusive goal which cannot be achieved without caveats, especially if honesty is also desired. As it pertains to density estimation, one of the earliest results to this effect came from Low [186], who considered pointwise inference for $f_0$ with uniformly bounded $k^{\text{th}}$ derivatives. They showed that, over this space, an honest confidence interval could achieve the worst-case contraction rate for *any* $f_0$, regardless of its true smoothness. Confidence sets in $L^\infty$ are particularly tricky: full adaptivity over finitely many smoothness levels can only be achieved by swapping the $\liminf$ and $\inf$ in (4.2) (i.e. considering "dishonest" bands) [101, 133], but Bull [29] showed that even with this modification it is still impossible to adapt over a continuous range of smoothness levels in the white noise model. Dümbgen [68] defined density confidence bands using a test statistic depending on the c.d.f values at order statistics and showed some adaptivity results based on local smoothness, but they are only valid over sets of shape-restricted (e.g. unimodal or monotonic) densities. Such difficulties are pervasive for all types of confidence sets: to achieve honesty and adaptivity together, it is necessary to assume additional restrictions on the smoothness classes under consideration or the functions therein. The theory shows that $L^2$ confidence sets are less restrictive in this regard than confidence bands, but neither are without their difficulties. Section 8.3 of Giné and Nickl's textbook [101] is an excellent and comprehensive discussion of these ideas, and the references in their notes provide further details. The authors explored adaptation theory for the white noise model, but noted that it can be made to apply to density estimation.

Adaptivity and honesty are central to the theory of nonparametric inference, but to many practitioners they may ultimately be less important than the aforementioned visual aspect of UQ. Figure 4.1 shows how to graphically represent the uncertainty associated with a density estimate by plotting multiple estimators and corresponding UQ methods, all based on the same simulated dataset. The figure includes both frequentist and Bayesian inference methods, and demonstrates the differences between pointwise (P.W.) intervals and simultaneous bands (in particular, the latter are wider than the former, as one would expect to be necessary for this stronger type of inference). The methods shown in Figure 4.1 are among the many described in the following sections, each of which explores UQ in terms of the concepts described above. The figure itself is discussed in more detail in Section 4.9 and Appendix A, as well as the supplementary material from the original publication of this work [198].

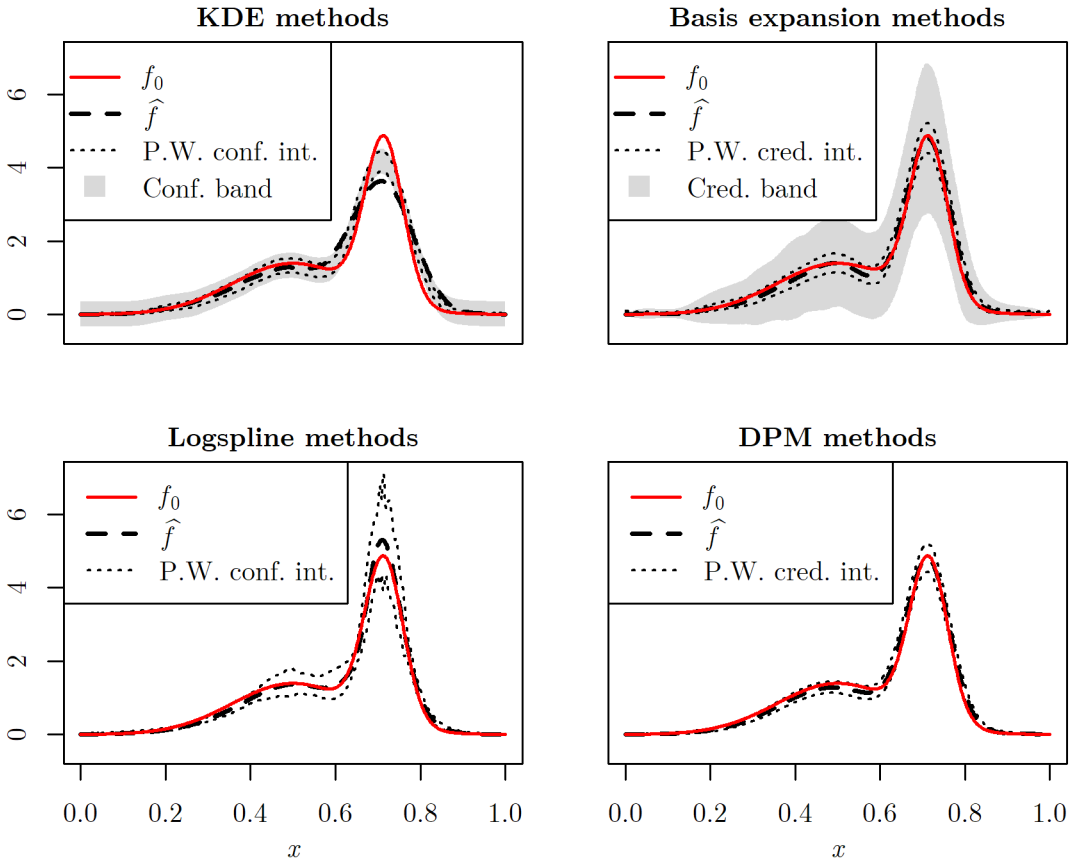


Figure 4.1: Different combinations of density estimation and UQ methods applied to the same sample. The true density is overlaid as a red line in each plot.

## 4.3 Kernel density estimators

KDE's are one of the most used and well-studied density estimation methods, at least in the frequentist literature. They are ubiquitous enough that their properties are arguably "common knowledge", receiving extensive documentation in textbooks, undergraduate course material, and review papers unto themselves [e.g. 47, whose review informs much of the discussion in this section]. Recall that a kernel density estimate for a density on $\mathbb{R}^d$ is of the form

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{4.4}$$

where $K$ is some (typically symmetric) kernel function and $h$ is a bandwidth which controls smoothing, or bias/variance tradeoff. Asymptotic theory for estimation is typically based on $h$ decaying to zero in some "big-O" relationship with $n$ that optimizes MSE, or integrated MSE. Practical methods for obtaining $h$ include cross-validation, plug-in methods, rules of thumb, and bootstrapping [144]. Note that, as the estimator is little more than a sample mean, it is equivalent to a conditional expectation with respect to the random measure $F_n$, the empirical distribution function of $\boldsymbol{X}$.

### 4.3.1 Pointwise inference

For pointwise inference, it is well-known that KDE's are asymptotically normal: with $\hat{f}$ as defined in (4.4), for all $x \in \mathbb{R}^d$ it holds that

$$\sqrt{\frac{nh^d}{f_0(x) \int K^2(t)\mathrm{d}t}} \left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right) \xrightarrow{d} \mathcal{N}(0, 1). \tag{4.5}$$

Furthermore, the distributions for a finite collection of points are asymptotically independent [33]. Using this fact, it follows that the endpoints for pointwise confidence intervals should be roughly of the form $\hat{f}(x) \pm z_{1-\alpha/2}\sigma_x$, where $\sigma_x^2$ is the variance which is asymptotically equal to $\frac{f_0(x) \int K^2(t)\mathrm{d}t}{nh^d}$. In practice, intervals can be computed by estimating $\sigma_x$: either by using one of its asymptotically-equivalent formulae ("plugging in" $\hat{f}$ in place of $f_0$ [47, 123] or replacing expectations with sample averages [134, 85]) or bootstrapping [47]. Many papers also replace the standard normal quantiles with those of bootstrap $t$-statistics [e.g. 134, 120]. Such studentized or "percentile-$t$" confidence intervals seem to be the most commonly-discussed in the literature, but any method of bootstrap confidence interval construction should be valid — for instance, Chen [47] discussed a bootstrap interval based entirely on the percentiles of absolute deviations. Hall and Kang [123] showed that re-calculating the bandwidth for each bootstrap sample does not provide worthwhile im-

provements to the accuracy of inference[1], so computational difficulty is avoided by using the same bandwidth across all replications. In the univariate case with a compactly-supported kernel, Chen [46] considered the construction of confidence intervals based on *empirical likelihood*, a nonparametric analogue to the standard methods of profile log-likelihood ratios. The theory is similar to the parametric case: viewing $\hat{f}(x)$ as a sample mean of random variables $K\left(\frac{x-X_i}{h}\right)/h$, Chen derived a limiting chi-squared distribution for $\ell\left(\mathbb{E}\left[\hat{f}(x)\right]\right)$, where $\ell$ is the profile empirical log-likelihood ratio. This allows for pointwise intervals of the form $\{y : \ell(y) \leq c_{1-\alpha}\}$, where $c_{1-\alpha}$ is the $1-\alpha$ quantile of the $\chi^2_1$ distribution. Chen showed that such intervals have asymptotic performance comparable to the percentile-$t$ bootstrap and can outperform it in simulations, especially with Bartlett correction.

Note that everything discussed thus far is based on (4.5), which is centered about $\mathbb{E}\left[\hat{f}\right]$ instead of $f_0$. This poses a problem for inference when choosing an "optimal" bandwidth minimizing (integrated) MSE or some proxy. The aforementioned intervals provide asymptotically-correct coverage for the expectation of $\hat{f}(x)$; in order for this to hold for $f_0(x)$ instead, the quantities in the numerator of (4.5) must be interchanged. This is only possible if the ratio of bias and asymptotic standard deviation

$$\sqrt{\frac{nh^d}{f_0(x)\int K^2(t)\mathrm{d}t}}\left(f_0(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)$$

goes to zero. However, the optimal asymptotic error rate is achieved when $h$ is set proportional to some power of $n$ such that the squared bias and variance decay at equal rates [47, 223]. Thus, with an "optimal" bandwidth, the ratio above tends to a nonzero constant so that confidence intervals do *not* have the correct coverage properties[2]. There are two main ways to handle this. The first is *undersmoothing*, where a lower-than-optimal bandwidth is selected. In the univariate case, Horowitz [134] suggested taking $h$ proportional to a higher power of $n$ than usual, thereby allowing the squared bias to decay faster than the variance; while Hall [120] multiplied a rule-of-thumb bandwidth by a constant $c \in (0, 1)$. Chen [46] used a version of the former when a confidence interval at only one point is desired: first obtaining kernel estimates $f_0$ and its second derivative at the point with approximations of the (local) MSE-optimal bandwidths, then using these to estimate the bias. Chen suggested simply using the optimal bandwidth in confidence interval construction when the estimated relative bias is small, or an estimate of a coverage-optimal undersmoothing bandwidth when it is large. Because a smaller $h$ means higher variance, confidence intervals based on un-

[1]In simulations, they found that recalculating the bandwidths can provide higher coverage, but at the expense of more conservative intervals. They showed that it doesn't asymptotically make a difference for compactly-supported kernels, but used the Gaussian kernel in simulations since its tails are light enough that it is "almost compact".

[2]This is one example of inference being at odds with the goal of optimal estimation. This will become a familiar refrain in theoretical ideas discussed throughout this chapter.

dersmoothing may be wider than one would prefer [47]. The second method is therefore to estimate the bias term with $\hat{b}$ and replace $\hat{f}$ with the bias-corrected estimator $\hat{f} - \hat{b}$. Assuming a kernel of order[3] $r$ is used, the bias depends on the $r^{\text{th}}$-order derivatives of $f_0$, assuming these are bounded and continuous [44]. These derivatives can also be estimated with kernel methods, but require higher bandwidths for optimality than the density estimator itself; for this reason, traditional bias correction uses an *oversmoothed* KDE to obtain $\hat{b}$ [120, 47]. Hall [120] showed through both asymptotics and simulations that undersmoothing with a higher-order kernel results in percentile-$t$ bootstrap confidence intervals with smaller coverage errors than those based on such "oversmoothing" bias corrections. However, Calonico et al. [36] developed a "robust" bias correction, in which the variance estimate used in confidence interval construction is modified to account for the correction, and showed that it can perform as well as undersmoothing-based intervals, with more robustness to bandwidth selection. Notably, their results hold when using the MSE-optimal bandwidth and second-order kernels for both $\hat{f}$ and the bias correction, which they noted to be convenient automatic choices. While lower error rates and narrower intervals are desirable, it should be noted that the bias-corrected centers $\hat{f} - \hat{b}$ are not necessarily nonnegative. Also note that the aforementioned results are based on a kernel with compact support.

Hall and Horowitz [122] devised another novel bootstrap approach. Starting with the original KDE $\hat{f}$, they repeatedly drew "bootstrap" samples *from* $\hat{f}$ (some papers call this the *smoothed bootstrap*, e.g. [207]) and used these to create Gaussian plug-in intervals at each point $x$ in the domain, with some nominal confidence level $1 - \beta(x)$. They set each $\beta(x)$ to ensure that the actual coverage (as estimated with bootstrap replicates) achieved the desired level $1 - \alpha$. Letting $\hat{\beta}_\delta$ be the $\xi$-quantile, for some low $\xi$, of the $\beta(x)$-values over a fine grid of $x$'s with edge width $\delta$, they took $\hat{\beta}$ as the limit of $\hat{\beta}_\delta$ as $\delta \to 0$ and finally used standard normal quantiles $z_{1-\hat{\beta}/2}$ to construct pointwise plug-in intervals centered at $\hat{f}$. Their theory and simulation studies focused on nonparametric regression, and in this case they showed asymptotic pointwise coverage of at least $1 - \alpha$ at roughly $(1 - \xi)100\%$ of points in the domain. However, they suggested that all results would translate to KDE's as well.

Similar ideas to those discussed above extend to situations besides a single i.i.d. sample $\boldsymbol{X}$. Louani [184] derived theoretical results for the case of randomly right-censored data: when there exists another sample $\boldsymbol{Y}$ of size $n$, and only $Z_i = \min\{X_i, Y_i\}$ and $\mathbb{1}(X_i \leq Y_i)$ are observed. They considered a modified KDE defined by integrating with respect to a Kaplan-Meier estimate of the c.d.f., rather than the usual e.d.f. $F_n$. Relaxing some of the conditions required for previous similar results [199, 181] (in particular, assuming only one bounded continuous derivative of $f_0$), Louani showed pointwise asymptotic normality for this estimator when using a kernel of compact support. The asymptotic standard deviation is

[3]The *order* $r$ of a kernel $K$ is the smallest positive integer such that the $r^{\text{th}}$ moment of $K$ is nonzero.

similar to the left-side factor in (4.5), but with an extra factor of $\sqrt{1 - G(x)}$, where $G$ is the c.d.f. of $\boldsymbol{Y}$. Another theoretical extension came from Giné and Mason [99], who considered kernel-based U-statistic estimators for the densities of functions $g(X_1, \ldots, X_m)$ with $m > 1$. Analogously to other results described here, they derived central limit theorems for such estimators, noting the bias can be eliminated if the bandwidth decays appropriately in the special case where $g$ is additive in its arguments (see also [258] for related results). Schick and Wefelmeyer [259] studied the case when the data is a linear process: $X_i = \sum_{s=0}^{\infty} a_s \epsilon_{i-s}$ for zero-mean $\epsilon_s$ and absolutely convergent $\{a_s\}$. The asymptotic mean in their limiting normal distribution is the convolution of the true density with the kernel. For a more practically-oriented extension, Wang and Wertelecki [293] considered data observed with rounding errors. They proposed a multi-step process to estimate the density of $\boldsymbol{X}$: first deriving a rough, convolution-based estimate for the c.d.f. of the non-rounded data; then using this to generate a sample from the estimated distribution of the rounding errors; and finally subtracting the simulated errors from the rounded data and constructing a KDE from the resulting quantities. Because the procedure involves simulated sampling, it naturally lends itself to bootstrap-style uncertainty quantification, which the authors showed in the form of pointwise confidence intervals for real data.

Further results exist for deconvolution density estimation. In this case, it is common to use a specialized *kernel deconvolution density estimator*, replacing the "standard" kernel in (4.4) with a *deconvolution kernel*: the Fourier transform of the ratio between the characteristic functions of some kernel function and the known error distribution. Fan [77] provided asymptotic normality results for such estimators in two cases: *ordinary smooth* deconvolution (where the tails of the error characteristic function decay at a polynomial rate) and *supersmooth* deconvolution (where they decay exponentially). In addition to the usual corollary of bias removal with undersmoothing, Fan also showed that the asymptotic variance (which depends on the true unknown density of the noisy data in the ordinary smooth case, and does not have a general expression in the supersmooth case) in the pivotal quantity could be replaced by a sample-dependent term: either the sum of squared deconvolution kernel values, or their sample variance (only the former was considered for the supersmooth case). Zhang [299] showed similar results for a similar estimator. Fan and Liu [78] later relaxed the conditions assumed in [77] for the ordinary smooth case, allowing the asymptotic normality results to apply to a wider variety of commonly-used error distributions. van Es and Uh [284] showed for a subset of supersmooth error densities that, under certain conditions on the kernel, the asymptotic variance of the estimator does not depend on the data or the true density. They noted that this allows in this case for the construction of pointwise confidence intervals without data-dependent standardization, although they do not address the issue of bias. Further asymptotic normality results with known variances are given in van Es and Uh [283] and Uh [280] for somewhat more general kernels and subsets of supersmooth error densities. Masry [194] generalized the classical

results of Fan and Zhang to inference on the joint density of stationary process data based on observations with i.i.d. additive noise. They showed asymptotic normality for various types of mixing with both ordinary smooth and supersmooth error distributions, but only considered undersmoothing-based bias removal and sample-based standardization for the former. For both i.i.d. and strongly-mixing data, Zu [302] proved asymptotic normality of the estimator when the noise is logarithmic chi-squared, a case not covered by the assumptions in the previous literature. The asymptotic variance in this case depends once again on the true density of the noisy data; Zu suggested that it could be consistently estimated by a classical KDE to facilitate construction of (biased) confidence intervals.

Returning once more to the case of an i.i.d. sample, a final extension is the *adaptive kernel density estimator* implemented in Stata by Van Kerm [285]. This method starts with a "pilot" density estimate of fixed bandwidth; its values at the sample points are used to assign individual bandwidths to each of the kernels in (4.4), which can also be given individual weights. These variable bandwidths reduce variance in regions where data is sparse, and bias in regions where it is dense. As with the normal KDE, it is an easy matter to get a plug-in estimate of standard error; this is how Van Kerm's software implements simple pointwise inference.

### 4.3.2 Simultaneous inference

Moving beyond the pointwise case, consider simultaneous UQ on the entire support or some subset thereof. The aforementioned undersmoothing and/or bias-correction principles still apply, and the rest of this section will largely take the application of such principles for granted. Bickel and Rosenblatt [20] provided perhaps the first results to this effect for univariate KDE's, showing that, under suitable technical conditions,

$$\mathbb{P}\left[A_n\left(\sqrt{\frac{nh}{\int K^2(t)\mathrm{d}t}}\sup_x\left|\frac{\hat{f}(x)-\mathbb{E}\hat{f}(x)}{\sqrt{f_0(x)}}\right|-d_n\right)<z\right]\to e^{-2e^{-z}} \tag{4.6}$$

for suitable sequences $A_n$ and $d_n$ (the latter being a function of the former), with the supremum taken over a compact interval (say, $[0,1]$ w.l.o.g.) on which $f_0$ is bounded away from 0. They further showed that with moderate undersmoothing, it is possible to replace $\mathbb{E}\hat{f}$ and $\sqrt{f_0}$ in (4.6) by $f_0$ and $\sqrt{\hat{f}}$, respectively, thereby justifying variable-width confidence bands $\hat{f}(x)\pm\sqrt{\frac{\hat{f}(x)\int K^2(t)\mathrm{d}t}{nh}}\left(\frac{z}{A_n}+d_n\right)$ for $x\in[0,1]$, where $z$ is such that $e^{-2e^{-z}}=1-\alpha$. Note that using a differently-scaled $A_n$, the factor of 2 in the exponent of the limiting distribution can be eliminated, thereby turning it into the c.d.f. of a *standard* Gumbel random variable [e.g. 100, who derived such a result for an undersmoothed data-driven bandwidth choice and plug-in estimator for $\sqrt{f_0}$; see Section 4.4.4 for further details]. In either case, the limiting probability law is of the extreme value or "double exponential" form. Rosenblatt [246] expanded upon Bickel and Rosenblatt's results, slightly relaxing the conditions

under which (4.6) holds in the univariate case and generalizing to the multivariate case. However, their multivariate results required rather strong restrictions on the bandwidth, differentiability of $f_0$, and moments of $K$. Rio [240] gave another rather technical result on the limiting distributions of suprema over closed subsets of $(0,1)^d$ for $d$-dimensional densities. Additional generalizations of the Bickel-Rosenblatt results in the univariate case were provided by Giné et al. [102], who gave conditions for results similar to (4.6) to hold with a different weight function $\Psi$ replacing the factor $\left(\sqrt{f_0}\right)^{-1}$ or the supremum taken over a data-dependent set. The same authors provided further theory to this end in a companion paper, in which they considered suprema over the whole real line [103]. Sakhanenko [253] further modified and extended these results to multivariate densities. Using moderate deviations principles, Mokkadem and Pelletier [201] showed that it is actually possible to construct Bickel/Rosenblatt-style confidence bands with asymptotic coverage level equal to 1 by using separate bandwidths for the KDE's in the mean and variance estimates (i.e. in the quantities used to define, respectively, the centre and margins of the bands). Further technical refinements allowed them to achieve this with narrower bands, at the expense of a slower convergence rate. While remarkable, these results have not been applied in practice in literature to date.

A drawback of using these asymptotics in practice is that convergence to the extreme value distribution is known to be very slow [e.g. 119]. Thus, it may be advisable to use bootstrapping for confidence bands. In what follows, let $f^*$ denote a KDE based on a bootstrap resample of $\boldsymbol{X}$. Hall [121] considered bands (over compact intervals) of the type

$$\left\{(x,y) : 0 \leq x \leq 1, \hat{L} \leq \frac{\hat{f}(x) - y}{\sqrt{y}} \leq \hat{U}\right\}$$

$$\text{where } \mathbb{P}\left(\hat{L} \leq \inf_x \frac{f^*(x) - \hat{f}(x)}{\sqrt{\hat{f}(x)}} \leq \sup_x \frac{f^*(x) - \hat{f}(x)}{\sqrt{\hat{f}(x)}} \leq \hat{U} \mid \boldsymbol{X}\right) = 1 - \alpha.$$

This band is based on the "studentized" quantity $\left(\hat{f} - \mathbb{E}\hat{f}\right)/\sqrt{\mathbb{E}\hat{f}}$, but differs from others by not using any kind of estimator for the denominator. Hall found in simulations that this interval had better coverage for $\mathbb{E}\hat{f}$ than a bias-corrected translation had for $f_0$, presumably due to inaccuracy in the bias correction. Hall and Owen [124] recommended the bootstrap to construct simultaneous confidence bands on $[0,1]$ with profile empirical likelihood methods. Recalling the notation for empirical likelihood in Section 4.3.1, they found an extreme value limiting result for $\sup_x \sqrt{\ell\left(\mathbb{E}\left[\hat{f}(x)\right]\right)}$ similar to (4.6), but recommended using percentile bootstrap methods to find a suitable bound $\hat{c}$ for a band of the form $\{f : \ell(f(x)) \leq \hat{c} \ \forall x \in [0,1]\}$. The technicalities and variations they considered are too cumbersome to discuss further here; see [124] for the full details. They found their intervals to be disappointingly wide when applied to real data, but suspected that this was due to the inherent variability of the density estimation itself. Neumann [207] gave quite

general theoretical results for uniform-width percentile bootstrap bands of the form $\hat{f} \pm t^*_\alpha$, where $t^*_\alpha$ is the bootstrap quantile of $\sup_x |f^* - \mathbb{E}f^*|$. Their results are valid for multivariate densities, suprema over all of $\mathbb{R}^d$, and weakly-dependent data. Neumann used compactly-supported kernels and the smoothed bootstrap: generating $\boldsymbol{X}^*$ from a possibly-different KDE based on the original sample, rather than from the empirical distribution. In a recent paper, Cheng and Chen [48] used the debiased estimator of Calonico et al. [36] to derive asymptotically correct bands, via the bootstrap, of either uniform width (using quantiles of $\sup_x \left| f^* - \hat{f} \right|$) or variable width (using quantiles of $\sup_x \left| \left( f^* - \hat{f} \right) / \sigma^* \right|$ multiplied by $\hat{\sigma}(\cdot)$), where the bootstrap density and associated variance estimates were all computed based on the bias-correction approach. Their results extend to the multivariate case and assume a compactly-supported $f_0$. Their simulation study showed that their bands achieved better coverage and narrower width than those based on the standard KDE, although some undercoverage still occurred for small samples without undersmoothing.

Yeh [296] used the bootstrap to create a rather novel type of confidence band. Generating a large number of KDE's from bootstrap samples of $\boldsymbol{X}$, they retained the $100\,(1-\alpha)\,\%$ of them with the largest *curve depth* (a way of ranking functions "from the centre out" based on some distance from a central curve, in this case the KDE $\hat{f}$). Simulation studies showed that such bands had reasonable performance compared to the asymptotic methods discussed in this section.

As is the case for pointwise inference, for simultaneous bands there are analogous results for deconvolution KDE's, described by Bissantz et al. [21]. A necessary assumption for these results is that the characteristic function of the error density decays as $t^{-\beta}$ for large $|t|$ and some known constant $\beta > 0$. Their asymptotic results for bands over a compact interval are nearly equivalent to those derived from (4.6), although in the asymptotic standard deviation they divided by an extra factor of $h^\beta$ and replaced $\hat{f}$ with $\hat{g}$, where the latter is an estimate of the density $g$ of the observed $Y$-values (a standard KDE suffices). However, they noted slightly better coverage probability (especially in terms of robustness to model misspecification) can be achieved with percentile bootstrap confidence bands of variable width based on the quantiles of $\left[ f^*(x) - \hat{f}(x) \right] / \hat{g}(x)$, where $f^*$ is a deconvolution estimator from a bootstrap sample of the observed noisy data.

### 4.3.3 Miscellaneous

Aside from some technical considerations in the previous section, not much consideration is given to the support of the true density. Indeed, the issue of KDE boundary bias for $f_0$ of restricted support is well-known and several mitigating strategies exist [e.g. 143, and discussion therein], but this is rarely discussed in the context of uncertainty quantification. One exception is given by Bouezmarni and Rombouts [23], who considered the *gamma kernel estimator* for time series data on $[0, \infty)$. The gamma kernel has a shape parameter varying with $x$ and leads to an estimator (asymptotically) free of boundary bias. The authors showed

pointwise asymptotic normality analogously to the results discussed above, based on the behaviour of the gamma scale parameter which acts as a bandwidth. In practice, it can be selected by cross-validation; the authors did so and constructed confidence intervals for real data based on their asymptotic results.

This section concludes by discussing a paper on large-sample Bayesian methods by Lo [180]. The key observation for this discussion is to recall that one can view the KDE (4.4) as a (conditional) expectation with respect to $F_n$. Lo's ideas are based on replacing $F_n$ in this expectation with a different random distribution $F$ conditional on $\boldsymbol{X}$. One such example is the empirical distribution of a bootstrap sample; this is equivalent to a probability measure with atoms at the sample values and weights randomly selected from $\{1/n, 2/n, \ldots, 1\}$. Lo also considered the *Bayesian bootstrap*, where the weights on the atoms are drawn from a uniform Dirichlet distribution [250]. This is equivalent to a draw from the posterior when $\boldsymbol{X} \sim F$, and $F$ is given an improper Dirichlet Process prior with zero base measure. Finally, Lo generalized this to allow for a non-zero base measure in the DP prior. They showed an asymptotic result analogous to (4.6) for all three aforementioned KDE variants, where the limit holds for $f_0$-almost all $\boldsymbol{X}$. This allowed them to use extreme value asymptotics to derive appropriate *Bayesian* bands for $f$ centred at the usual KDE $\hat{f}$. In practice one may prefer not to do this, given the substantial developments in Bayesian computation since the time of Lo's paper.

## 4.4 Adaptive basis expansion methods

This section considers estimates for $f_0$ of the form

$$f(x) = \sum_{j=1}^{K} b_j B_{j,K}(x), \tag{4.7}$$

where the $B_{j,K}$'s are a suitable set of fixed nonnegative "basis functions" for a given $K$. The simplest choice is taking them to be indicator functions on disjoint subsets of the support, in which case $f$ is simply a histogram. Other options include Bernstein polynomials [e.g. 289, 224], B-splines [e.g. 264], and wavelets [e.g. 100]. The coefficient vector $\boldsymbol{b} \in \mathbb{R}^K$ is constrained such that $f$ is a valid density. The remainder of this section will use $\hat{\boldsymbol{b}}$ to denote the coefficients associated with a specific estimator $\hat{f}$.

The dimensionality $K$ is of particular interest, serving as a smoothing parameter that controls the bias-variance tradeoff of the estimator. The basis functions corresponding to higher $K$-values are typically "narrower", allowing for more intricate shape detail to be captured in estimates. For instance, taking a high $K$-value for the histogram corresponds to using a larger number of narrower bins. Conversely, a value that is too high will result in a high-variance estimator that is unacceptably noisy. In general, higher $K$-values are required for larger samples to capture the true density.

One can choose $K$ in a data-driven way. Many theoretical results for this approach rely on $K$ increasing with $n$, usually appealing to some "big-O" conditions on its growth [e.g. 9, 100, 277]. In practice, a value could be chosen by cross-validation [171], changepoint methods [115], or appealing to known asymptotic theory [9, who derived nice properties for a method with $K = o(n/\log n)$ and then simply used $K = n/\log n$ in a simulation study]. In the theoretical Bayesian context, Rousseau and Szabó [249] considered *maximum marginal likelihood* (MML) estimates for $K$, marginalizing over a prior for $\boldsymbol{b}$. Further discussion of such ideas is beyond the scope of this chapter; see van de Wiel et al. [281] for details on practical implementation of MML.

In the Bayesian literature, methods involving a data-driven choice of a single $K$-value are often called *empirical* [249]. All frequentist methods discussed in this section are of this type. On the Bayesian side, such methods contrast with *hierarchical* ones, which use a suitable discrete prior on $K$ and allow it to "vary" [249]. In general terms, the $K$-values obtained with any approach will reflect what is necessary to capture the true shape of $f_0$. It is in this respect that these estimators are said to be "adaptive".

### 4.4.1 Histograms

Perhaps the simplest of density estimators, a histogram (sometimes referred to as an *empirical density* [237]) is piecewise constant over some division of the support into disjoint subsets, or "bins". In the most general form with countably many bins $\{A_j\}$, it can be written as

$$f(x) = \sum_j c_j \mathbb{1}_{A_j}(x), \tag{4.8}$$

with the constants $c_j$ chosen to ensure that the estimator is a valid density. The regularity of a histogram is controlled by adjusting the sizes of the bins. For instance, a very common form for the univariate case is

$$\hat{f}(x) = \frac{K}{n} \sum_j n_j \mathbb{1}_{\left[\frac{j-1}{K}, \frac{j}{K}\right)}(x), \tag{4.9}$$

where $n_j$ is the number of sample values in the interval $\left[\frac{j-1}{K}, \frac{j}{K}\right)$ and $K$ provides the needed regularity control. Assuming a bounded support, say $[0, 1]$, the sum in (4.9) is over $j = 1, \ldots, K$ and is equivalent to (4.7) using a basis of indicator functions: $B_{j,K} = K\mathbb{1}_{\left[\frac{j-1}{K}, \frac{j}{K}\right)}$.

Temporarily ignoring the notion of empirical or hierarchical approaches to $K$, suppose for now that it is fixed at some arbitrary value irrespective of everything else. Then the histogram simply becomes a problem of multinomial inference: the coefficient $b_j$ in (4.7) is an estimate of the probability that $X \sim f_0$ falls in the $j^{\text{th}}$ bin, say $p_j$. In this respect, the "traditional" histogram, where $\hat{b}_j = n_j/n$ as in (4.9), is a MLE. Here the object of

inferential interest is not necessarily $f_0$, but rather the so-called *theoretical histogram* $\overline{f}$ [269], a piecewise-constant density equal to $Kp_j$ in the $j^{\text{th}}$ bin. With this view, (piecewise-constant) pointwise intervals arise by considering the single binomial proportion $p_j$, and simultaneous bands by considering the vector of multinomial probabilities $\boldsymbol{p} = (p_1, \ldots, p_K)$. Frequentist and Bayesian methods for both are well-studied; see Vermeesch [288] for some practical applications to histograms.

**Simultaneous frequentist inference**

To discuss inference for $f_0$ itself, it is necessary to return to the adaptive paradigm. Much of the frequentist literature for histogram density UQ is theoretical and predates developments such as the bootstrap, relying on extreme value asymptotics similar to those for KDE's. One of the first such papers is by Smirnov [269]. They derived a limiting result much like (4.6) for the normalized quantity

$$\sqrt{nK^{-1}} \sup_x \frac{\left|\hat{f}(x) - f_0(x)\right|}{\sqrt{\overline{f}(x)}}, \tag{4.10}$$

where the supremum is over a compact interval on which $f_0$ is bounded away from 0 and has total mass less than 1. Smirnov claimed that it was not possible to replace $\overline{f}$ in the denominator with $f_0$ due to the systematic difference between them dominating the error. However, they stated that it *is* possible to do so by replacing the histogram with a *frequency polygon* (a linear interpolation between the histogram values at the sample points) and imposing some extra conditions on the relationship between $K$ and $n$. Although Smirnov did not provide proofs for these results, they will be shown later to be a special case of proven results for wavelets [100]. For $f_0$ supported on a compact interval, Révész [237] was able to prove a somewhat modified extreme value limit for the distribution of a quantity similar to (4.10), except $\hat{f}$ can be either the traditional histogram or a frequency polygon (with a slightly different interpolation scheme than that considered by Smirnov), $f_0$ replaces $\overline{f}$ in the denominator, and the supremum is taken over an interval converging to the whole support of $f_0$. Révész also derived a similar result with the absolute value removed from the supremum. Further results to this effect were given by Freedman and Diaconis [87]. For everywhere-positive densities with a unique maximum, they considered a quantity similar to (4.10) without the absolute value (i.e. considering only the maximum *positive* deviation, although they claimed their proofs can be adapted to the maximum absolute deviation), the supremum taken over the whole real line, and the factor of $\overline{f}(x)$ in the denominator replaced by the maximal value of $f_0$ (a fixed constant). Their limiting results are quite similar to those of Révész.

The three papers just discussed allow for (using a moderate amount of algebra) the construction of asymptotically correct simultaneous confidence bands for univariate densities satisfying suitable technical conditions, provided $K$ increases at a suitably fast rate with respect to $n$ (this roughly corresponds to the notion of undersmoothing discussed in Section 4.3). However, these papers did not concern themselves with the practicality of these ideas applied to actual data. It seems reasonable to suspect that slow convergence could be an issue which could be rectified with bootstrap methods, as was the case with KDE's.

**Pointwise frequentist inference**

Consider now the issue of frequentist pointwise intervals. For the "traditional" histogram (of the form (4.9)), Laloë and Servien [166] showed conditions on $K$ and $f_0(x)$ for the quantity

$$\sqrt{nK^{-1}}\frac{\hat{f}(x) - f_0(x)}{\sqrt{f_0(x)}}, \tag{4.11}$$

to have a limiting distribution, which they proved to be a standard Gaussian when it does exist. Their proof applied the Lindberg-Feller Central Limit Theorem to the histogram values (recall that these are scaled binomial random variables for each $K$). Their conditions were more general (but also more technical) than those in many other papers: in particular, they did not even require $f_0$ to be continuous. Some literature provides results for non-traditional histogram variants with non-uniform bin spacing. For univariate densities, Kim and Van Ryzin [154] showed pointwise asymptotic normality for a histogram with randomly-spaced bins. They required bin spacings to meet certain conditions for their results to hold; one valid option is to fix the number of observations in each bin and determine their widths by the spacings of the sample's order statistics. The same authors showed analogous results for an extension to the bivariate case [153]. Another variant for the univariate case is the *maximum entropy histogram estimator* (MEHE), which works by dividing the real line into $K \leq n$ subintervals (with the first and $K^{\text{th}}$ respectively extending to $-\infty$ and $+\infty$ and $\hat{f}$ having some suitable tail behaviour there) and choosing the spacing of their boundaries to maximize entropy subject to preservation of sample means and mass in the subintervals. Rodriguez and Van Ryzin [242] considered this estimator and a "symmetrized" variant and showed pointwise asymptotic normality of the quantity (4.11) for both. Their conditions on continuity and growth of the number of subintervals were slightly different for the symmetrized version, and the limiting law does not concentrate around zero as it does for the regular MEHE, presumably necessitating bias correction. Stadtmüller [271] considered asymptotics for yet another variant of the form (4.9), first considered by Gawronski and Stadtmüller [88], in which the indicator functions in the summands are replaced by the values of *lattice distributions* to yield a smoothed estimate. They gave a few suitable examples: replacing $\mathbb{1}_{\left[\frac{j}{K}, \frac{j+1}{K}\right)}(x)$ by $\mathbb{P}(Y = j)$ with, say, $Y \sim \text{Bin}(K, x)$ for densities supported on $[0, 1]$, or with $Y \sim \text{Poi}(Kx)$ for those supported on $[0, \infty)$. Note that the lattice

distributions do *not* necessarily constitute probability distributions with respect to $x$. Thus, density estimators of this type may not integrate to 1, although some examples presented in [88, 271] certainly will. For these estimators, Stadtmüller [271] showed pointwise asymptotic normality of the quantity

$$\left(\frac{4\pi\sigma^2(x)n^2}{Kf_0^2(x)}\right)^{1/4}\left(\hat{f}(x)-\mathbb{E}\hat{f}(x)\right),$$

where $\sigma^2$ depends on the lattice distributions. Additionally, they showed extreme value limiting results [somewhat similar in form to those in 20, as usual] for the supremum of this quantity (as well as the supremum of its absolute value) over compact intervals, under some regularity conditions on $f_0$ and the lattice distributions. They also noted that it is possible, as usual, to replace $\mathbb{E}\hat{f}$ by $f_0$ (thereby achieving correct asymptotic coverage for confidence intervals or bands) by undersmoothing — in this case, increasing $K$ at a higher-than-optimal rate with respect to $n$.

**A Bayesian approach**

Recently, Rousseau and Szabó [249] discussed theory for Bayesian UQ of histogram estimators, assuming univariate $f_0$ supported on a compact interval. For this, return to the form (4.7), with the basis functions equal to indicators for equally-spaced bins. Rousseau and Szabó considered credible sets of the form (4.1), where $d$ is the *Hellinger distance* and $\hat{f}$ is a suitable centering point such as the posterior mean. They showed that, under some regularity conditions on $f_0$ (it must be bounded away from zero and sufficiently smooth, and satisfy a "general polished tail assumption" defined by the authors and briefly described below) and the prior (a suitable $K$-dimensional Dirichlet for $\boldsymbol{b} \mid K$, and others omitted here for brevity), posterior credible sets of this type have arbitrarily high asymptotic frequentist coverage if their diameter is increased by an appropriate factor. In mathematical terms, for any $\epsilon \in (0,1)$, there exists $L_\epsilon > 0$ such that

$$\begin{aligned}&\liminf_{n\to\infty}\mathbb{P}_0\left(C\left(L_\epsilon\sqrt{\log n}r_\alpha\right)\ni f_0\right)\geq 1-\epsilon, \\ &\text{where } \Pi\left(f\in C\left(r_\alpha\right)\mid \boldsymbol{X}\right)=1-\alpha.\end{aligned} \tag{4.12}$$

In fact, they showed the stronger honesty result that this limit inferior holds *uniformly* over a certain class of functions, and that the uninflated credible sets are also almost adaptive over this class (save for a logarithmic factor in the diameter contraction rate). The densities comprising this class are those in an arbitrary union of Hölder balls of equal radius and regularities in $(1/2, 1]$. They must also satisfy the aforementioned general polished tail assumption, which essentially controls their high-resolution behaviour. Further ideas of this type will emerge in Section 4.4.4. Rousseau and Szabó's results hold for both the empirical

and hierarchical approaches to $K$. For the latter case, a geometric or Poisson prior for $K$ satisfies the relevant conditions. The authors noted that the "blow-up factor" of $\sqrt{\log n}$ is unfortunate, but they believe it is necessary to prevent coverage from decaying to zero in certain cases. Although it is quite pleasant to have such theoretical guarantees, it may be a challenge to put them towards a practical end due to the blow-up factor. Given an MCMC method to generate posterior simulations from this model, a credible set can be roughly visualized by plotting the $(1-\alpha)100\%$ of $f$ draws closest in Hellinger distance to $\hat{f}$, but plotting draws from the blown-up set is another matter since we are not aware of any way to estimate $L_\epsilon$.

Given the popularity of histograms, it is somewhat surprising that practical implementations and demonstrations of UQ for them appear so rare in the literature. For practitioners thorough enough to quantify errors in density estimation, it is perhaps reasonable to conclude that histograms have been superseded by KDE's and other methods that produce smooth estimates. Certainly, smoothness is advantageous for interpretation, especially when one wishes to account for uncertainty.

The following sections will contain a few more results which are applicable to histograms, arising as special cases of other methods.

### 4.4.2 Bernstein polynomials

One of the earliest non-histogram methods of the type (4.7) was proposed by Vitale [289] for densities supported on $[0,1]$. They took

$$B_{j,K} = \text{Beta}\left(j, K-j+1\right),$$
$$\hat{b}_j = \frac{\#\left\{X_i \in \left(\frac{j-1}{K}, \frac{j}{K}\right]\right\}}{n}.$$

The basis functions are beta densities with integer parameters (equivalently, scaled *Bernstein polynomials*), and the coefficients are equal to the proportion of sample values in each interval $\left[\frac{j-1}{K}, \frac{j}{K}\right)$. In this respect, Vitale's estimator is essentially a smoothed histogram. In fact, aside from a different scaling factor it is almost the same as the lattice-smoothed histogram of Gawronski and Stadtmüller [88]; it therefore seems reasonable to suspect that one could derive confidence bands from similar asymptotic arguments as Stadtmüller [271]. An equivalent way to interpret this estimator is as a mixture of beta densities.

Babu et al. [9] provided pointwise asymptotic normality results for this estimator under mild conditions, from which one can presumably derive expressions for approximate pointwise confidence intervals (subject to the usual handling of bias and variance terms). At interior points $x \in (0,1)$, Vitale's estimator is quite similar to the KDE: optimal MSE behaviour occurs with $K$ such that the asymptotic orders of variance and squared bias match [289]. With this choice, confidence intervals — based on either plug-in or bootstrap

methods — will not have the correct asymptotic coverage, concentrating around $\mathbb{E}\left[\hat{f}(x)\right]$ instead of $f_0(x)$ [9]. "Correct" intervals could be obtained by undersmoothing, choosing a higher-than-optimal $K$ [asymptotic conditions in 9]. Alternatively, noting that the bias term is a known function of the first two derivatives of $f_0$, it may be reasonable to estimate a bias correction with plug-in methods, again in analogy with KDE's. Tenbusch [277] proved analogous results for Vitale-style estimates of bivariate densities defined on triangular or rectangular regions, with some generalizations for the latter provided by Babu and Chaubey [8]. As they are quite similar to the univariate case, they are not repeated here. The aforementioned pointwise results are valid for interior points $x$, but these estimators are known to have different asymptotic behaviour at the boundaries [289, 277] and so UQ may also work differently there.

There are methods besides Vitale's for estimating a density with Bernstein polynomials. For another frequentist method, take the coefficients $\hat{\boldsymbol{b}}$ to be MLE's. Guan [115] claimed pointwise asymptotic normality results for this approach, but it is not clear how to turn these results into appropriate pointwise intervals.

Theory for Bayesian estimates of this type typically depends on the idea of viewing the coefficients $\boldsymbol{b}$ as increments of some unknown c.d.f. $F$: $b_j = F\left(\frac{j}{K}\right) - F\left(\frac{j-1}{K}\right)$ (Vitale's estimator fits this framework for $F$ equal to the e.d.f. of $\boldsymbol{X}$). To that end, Petrone [224, 225] considered a hierarchical Bayesian formulation with a discrete prior on $K$ and a Dirichlet Process prior on $F$. For practical implementation, they devised an equivalent formulation making use of the aforementioned "mixture-of-betas" interpretation. They introduced a vector of latent variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \sim F$, which provide "mixture labels" for the samples conditional on $K$: $X_i \mid Y_i, K, F \sim \text{Beta}\left(\lceil KY_i \rceil, K - \lceil KY_i \rceil + 1\right)$. See Petrone [225] for more details on the properties of this construction, as well as a Gibbs sampling algorithm for posterior inference. In principle, this formulation gives everything needed to obtain, at the very least, pointwise credible intervals — indeed, Petrone did so in these papers. Note that practical implementations of this model require the truncation of the prior for $K$ for computations to be possible. This has theoretical implications, but is not an issue in practice provided the maximum value for $K$ is reasonably high. Petrone and Veronese [226] generalized these ideas for data not necessarily in $[0, 1]$; see Section 4.7.3 for elaboration on this.

Following the analogous KDE ideas in Lo [180], Ghosal [97] considered an alternative "posterior" based on a (generalized) Bayesian bootstrap approach, where it is assumed that $\boldsymbol{X} \sim F$ and $F$ is a random distribution from a Dirichlet Process with base measure $\alpha(\cdot) + \sum \delta_{X_i}$. They conjectured pointwise asymptotic normality (concentrating around the Bernstein density with coefficients from $F = F_0$, rather than $f_0$ itself), but could not adapt the results in Lo [180] to prove this.

### 4.4.3 B-splines

B-splines are another option for the basis functions in (4.7). They are piecewise polynomials, characterized by a set of points in the domain called *knots* at which the values of the piecewise functions and a certain number of their derivatives must match. The number of basis functions $K$ depends on both the number of knots and the polynomial degree chosen. Cubic splines are the most common choice, but there are others: for instance, a Bernstein basis of size $K$ is a special case of B-splines of degree $K-1$ and no interior knots [72]. Using interior knots allows B-splines to be sharper-peaked in general than Bernstein polynomials, even at lower degrees. Literature about B-splines abounds; see Dias [67] for one of many introductions.

Although there will be plenty of discussion of splines in Sections 4.5.1 and 4.6, their use in estimators of the form (4.7) is limited in the literature. UQ for such estimators appears limited to practically-oriented Bayesian papers, although we suspect that it may be possible to translate some of the theory pertaining to histograms or Bernstein polynomials to this type of basis. Note that it is necessary to normalize each B-spline so it integrates to 1, thereby preserving the "mixture-of-basis-densities" view of (4.7). As another technical note, here attention is restricted to compactly-supported densities and estimators.

Shen and Ghosal [264] considered the hierarchical Bayesian setup (as defined at the beginning of Section 4.4), with $K$ having a suitable discrete prior and $\boldsymbol{b} \mid K$ having a conditional $K$-dimensional Dirichlet prior. Like most practically-oriented papers with a hierarchical framework, they noted that the prior on $K$ must be truncated for computation. They gave a closed-form expression for the posterior mean of $f$ and claimed similar expressions existed for higher posterior moments, allowing them to construct approximate credible intervals (presumably by a Gaussian-style "mean $\pm$ 2*standard deviation" approximation). Their expression for the posterior mean is a ratio of sums, each of which has a number of terms increasing exponentially in $n$ for splines of degree $\geq 1$. Thus, the authors suggested randomly sampling a reasonable number of summands, say 3000, to approximate it. This is not an issue for splines of degree 0, as many terms cancel out due to the basis functions having non-overlapping supports. In this case, the estimator is simply a histogram and simplicity arises at the expense of smoothness. Shen and Ghosal found in their simulation study that the credible intervals were more appealing with cubic splines than with constant ones, although both had some difficulty capturing some of the true density's shape.

Edwards et al. [72] compared Petrone-style Bayesian formulations [225, although Edwards et al. modified the MCMC] using both the Bernstein basis [see also 52] and B-splines for estimating the spectral density of a stationary time series. This use case differs from the probability density estimation considered here, but some of their ideas are nevertheless interesting for our purposes. In addition to pointwise credible intervals, they also considered

simultaneous bands generated from median absolute deviations:

$$\hat{f}(x) \pm \xi_\alpha \text{MAD}[f(x)], \tag{4.13}$$

where $\hat{f}$ is the posterior median, the pointwise MAD's are taken over MCMC draws, and $\xi_\alpha$ is the $1-\alpha$-quantile (obtained via MCMC draws) of $\sup_x \left( \left| f(x) - \hat{f}(x) \right| / \text{MAD}[f(x)] \right)$. In a simulation study, they found that such bands had vastly superior coverage using B-splines instead of the Bernstein basis. Pointwise intervals for B-splines tended to be wider, but both these and simultaneous bands captured intricate shape details more effectively than when the Bernstein basis was used. This is because the compact support of B-splines allows them to more effectively capture sharp peaks. The authors noted, however, that B-splines resulted in longer computation times than the Bernstein basis. Lopes and Dias [183] used a semiparametric Bayesian model for densities, combining a mixture of normalized B-splines with Dirichlet-distributed coefficients with a mixture of parametric densities. As usual, a straightforward Gibbs sampler allowed them to obtain pointwise credible intervals from MCMC output.

### 4.4.4 Orthonormal wavelets

Briefly, the idea behind estimation with orthonormal wavelets is to express a square-integrable function $f$ in the form

$$f(x) = \sum_{k \in \mathbb{Z}} \sum_{j \in \mathbb{Z}} c_{kj} \psi_{kj}(x), \tag{4.14}$$

where $\psi_{kj}(x) = 2^{k/2} \psi \left( 2^k x - j \right)$ for some suitable function $\psi$ called the *mother wavelet*. The mother wavelet is such that $\{\psi_{kj}\}$ is an orthonormal basis of $\text{L}^2(\mathbb{R})$, so that $c_{kj} = \int f \psi_{kj}$. For most of the literature discussed in this section, it can be assumed unless otherwise noted that the domain is indeed all of $\mathbb{R}$. However, in some cases it is desirable to modify the wavelets so that they form a basis of, say, $L^2([0,1])$, and there are multiple approaches to this [e.g. 55].

It is often more convenient to express $f$ as

$$f(x) = \sum_{j \in \mathbb{Z}} d_j \phi_{k_0 j}(x) + \sum_{k=k_0}^{\infty} \sum_{j \in \mathbb{Z}} c_{kj} \psi_{kj}(x), \tag{4.15}$$

where $\phi_{kj}(x) = 2^{k/2} \phi \left( 2^k x - j \right)$ for some *scaling function* or *father wavelet* $\phi$ such that $\{\phi_{k_0 j}, \psi_{kj} : j \in \mathbb{Z}, k \geq k_0\}$ is also an orthonormal basis for $\text{L}^2(\mathbb{R})$. The number $k_0$ corresponds to the "coarsest" level of detail under consideration. In most literature explored below, it is either left arbitrary or set to 0 when the domain is $\mathbb{R}$. When modifying the wavelets for use on $[0,1]$, the dimensionality of the basis will depend on $k_0$ and, for some

methods, the "regularity" of $\psi$ [55]. In this setting, $k_0$ may therefore be chosen to provide an appropriate set of basis functions [e.g. 27, 43].

The simplest wavelet example is the *Haar wavelet*, where $\phi = \mathbb{1}_{[0,1)}$ and $\psi = \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1)}$. In general, $\phi$ and $\psi$ must be selected to mutually satisfy certain functional equations. For further detail on wavelet theory and examples, refer to Kaiser's excellent book on the subject [147].

In practice, to estimate a density with wavelets, one must truncate the sum over $k$ in the second term of (4.15) to some upper limit $K$. In this respect, $K$ is a bandwidth or "resolution": higher values introduce thinner wavelets into the sum that capture finer details, thereby reducing bias and increasing variance. In this respect, wavelets differ from other basis expansion methods, in which the shapes of the basis functions themselves change with $K$. As previously mentioned, the coefficients in the wavelet expansion are simply inner products between the density and the basis functions. Thus, to obtain a point estimate $\hat{f}$, the natural choice is to estimate $d_j$ and $c_{kj}$ by their empirical versions: the sample means of $\phi_{k_0 j}(\boldsymbol{X})$ and $\psi_{kj}(\boldsymbol{X})$, respectively. It is now clear that density estimators based on the Haar wavelet are simply histograms with evenly-spaced bins.

**Frequentist $L^\infty$ inference**

Giné and Nickl [100] derived some theoretical results for confidence bands over a compact subinterval, taken w.l.o.g to be $[0, 1]$, by treating certain types of wavelet estimators in a unified framework with KDE's. In their approach, $\boldsymbol{X}$ is split into two subsamples: one of which is used for a data-driven bandwidth selection procedure (as always, their arguments involved undersmoothing to ensure correct coverage), with the other used to obtain the estimate $\hat{f}$ with this bandwidth. Letting $K$ denote the number obtained from the bandwidth selection procedure (the details of which can be read in [100]), their framework encompasses both

1. kernel density estimators with kernel $\kappa(x, y) = \kappa(x - y)$ and bandwidth $2^{-K}$; and
2. wavelet estimators in the form of (4.15) with $k_0 = 0$, and the sum over $k$ in the second term truncated to $K$ terms. To unify these estimators with KDE's, the authors invoked a *projection kernel* defined in terms of the father wavelet: $\kappa(x, y) = \sum_k \phi(x-k)\phi(y-k)$.

For a final piece of notation, let $c = \sup_x \int \kappa^2(x, y)\mathrm{d}y$. Giné and Nickl showed a result somewhat similar to the asymptotic KDE result in (4.6): for $f_0$ bounded away from zero on an open interval containing $[0, 1]$, under some technical conditions the estimators in their framework satisfy

$$\mathbb{P}\left[A_n\left(\sqrt{\frac{n2^{-K}}{c}}\sup_{x\in[0,1]}\left|\frac{\hat{f}(x) - f_0(x)}{\sqrt{\hat{f}(x)}}\right| - d_n\right) < z\right] \to e^{-e^{-z}} \tag{4.16}$$

for suitable (known) sequences $A_n$ and $d_n$. Just as in Section 4.3.2, it is straightforward to use this limit to get asymptotically-correct confidence bands. The authors further showed that these bands are honest and nearly[4] adaptive in a range of Hölder balls over all but a nowhere-dense (w.r.t. the Hölder norm) subset of the function space. However, they noted that their work is theoretically oriented and therefore cautioned against using these bands without assessing their finite-sample performance. Furthermore, the only wavelets they showed to fit into their framework were the *Battle-Lemarié* wavelets of order $1, 2, 3$, and 4. The scaling function for the Battle-Lemarié wavelet of order $r$ is a B-spline of order $r$ [61], so for $r = 1$ it reduces to the Haar wavelet.

Because (4.16) generalizes the histogram results first discussed by Smirnov [269] and the KDE results shown by Bickel and Rosenblatt [20], statements of this type are often called *Smirnov-Bickel-Rosenblatt theorems*. Bull [30] showed that a Smirnov-Bickel-Rosenblatt result holds in the white noise model using symlets and Daubechies wavelets. The Daubechies wavelet of order $r$ is a Haar wavelet for $r = 1$ and has increasing regularity for higher orders, but unlike the Battle-Lemarié wavelet it has the advantage of being compactly supported [30, 147]. Bull verified their results for orders $6 \leq r \leq 20$, using bases on both $\mathbb{R}$ and $[0, 1]$. Although the white noise model is not the focus of this review, they noted that these results could translate to the density estimation context via some of the Gaussian process theory in [100]. Indeed, the notion of equivalence between the white noise model and density estimation is established [e.g. 214], but the details are beyond the scope of this chapter.

It was noted above that a Smirnov-Bickel-Rosenblatt confidence band could achieve honesty and adaptivity under certain conditions and restrictions on the function space. More broadly, discussion of these concepts often uses wavelet theory as a starting point, due to the nice theoretical properties of an orthonormal basis. Hoffmann and Nickl [133] considered another approach to ensuring the existence of adaptive and honest confidence bands in *finitely many* nested Hölder balls: removing subsets of functions from the lower-regularity ones to ensure "separation" from the smoother classes. By connecting this idea to hypothesis tests for the smoothness of $f_0$, they showed that, in the case of finitely many smoothness levels, such separation conditions are necessary and sufficient for the existence of honest and adaptive bands, and that these conditions are weaker than those imposed by [100]. The constructive part of their argument involved a uniform band centered at an estimator satisfying certain properties; their paper and the references therein suggested that a wavelet estimator would be a good choice for both $L^2(\mathbb{R})$ and $L^2([0, 1])$. Unfortunately, the radii of these bands depend on properties of the Hölder balls that are unlikely to be known in practice, rendering application implausible. Nevertheless, these results are useful to inform theoretical discussion of the behaviour of confidence sets. Bull [29] considered

[4]Their diameters shrink at a rate which is nearly optimal, save for the presence of an extra logarithmic factor.

inference on a union of Hölder balls with diameters and regularities both varying over a continuum. The conditions they imposed on the function sets are similar to those in [100], but somewhat weaker. Specifically, they required the densities under consideration to be *self-similar*. Briefly, self-similarity is a property of a function's wavelet expansion ensuring that it exhibits similar regularity at both small and large scales. Note that the general polished tail condition of [249] (see Section 4.4.1) is a generalization of this. Bull showed that this restriction excludes only a "negligible" set of functions in both the topological and probabilistic[5] sense, and that it is necessary and sufficient to achieve honest and adaptive confidence bands over a continuous union of Hölder balls. Refer to Sections 8.3.3 – 8.3.4 of [101] for a more in-depth discussion of the role self-similarity plays in nonparametric inference.

Bull described a rather complex procedure to construct such a uniform band centered at a truncated empirical wavelet estimator, using Daubechies wavelets or symlets of order $6 \leq r \leq 20$, modified to form a basis of $L^2([0,1])$. The procedure exploits self-similarity to estimate the true smoothness of $f_0$. Unlike the construction of Giné and Nickl [100], it does not require sample-splitting.

### A practical approach

None of the literature discussed thus far in this section concerns itself with applications to real data. To the extent that there have been constructive results, they have tended in most cases to be rather complicated. For an example of somewhat more practically-oriented material, Chernozhukov et al. [49] developed $1-\alpha$ confidence bands of the form

$$\hat{f}_{\hat{l}}(x) \pm \hat{\sigma}_{\hat{l}}(x)\left(\hat{c}_n(\alpha) + c'_n\right) \tag{4.17}$$

over a compact subset of $\mathbb{R}^d$. The subscript $\hat{l}$ is a particular value of $l$, which is used to denote bandwidth ($l$ replaces the usual letter $K$ here for more streamlined notation as in the original paper). Much like Giné and Nickl [100], these authors cast both KDE's and wavelet estimators into the larger framework of estimators $\hat{f}_l$ based on some kernel $\kappa_l$ (note that, unlike [100], they folded the bandwidth into the definition of the kernel). In fact, their framework also encompasses estimators based on *nonwavelet* projection kernels using other orthonormal bases such as Legendre polynomials. They considered univariate kernels and wavelets, and extended to the multivariate case by using elementwise products. Returning to (4.17), $\hat{\sigma}_l$ is an estimate of the standard deviation of $\hat{f}_l$, obtained using sample mean analogues of the relevant expectations [e.g. 85]. Letting $\mathcal{L}_n$ denote the space of possible

[5] By considering a natural prior distribution on the space of functions.

bandwidths, $\hat{c}_n(\alpha)$ is an estimate of the $1-\alpha$ quantile of

$$\sup_{l \in \mathcal{L}_n, x} \left| \frac{\hat{f}_l(x) - \mathbb{E}\left[\hat{f}_l(x)\right]}{\sqrt{\mathrm{Var}\left[\hat{f}_l(x)\right]}} \right|.$$

The authors suggested obtaining $\hat{c}_n(\alpha)$ by using the *Gaussian multiplier bootstrap*: whereas the normal bootstrap takes repeated samples of size $n$ from the empirical distribution of $\boldsymbol{X}$, this version repeatedly samples $n$ i.i.d. standard normal variables $\xi_1, \ldots, \xi_n$. Subsequently,

$$\sup_{l \in \mathcal{L}_n, x} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \frac{\kappa_l\left(X_i, x\right) - \hat{f}_l(x)}{\hat{\sigma}_l(x)} \right| \tag{4.18}$$

is calculated, and $\hat{c}_n(\alpha)$ is taken to be the $1-\alpha$ quantile of this quantity over bootstrap replications. The numbers $c_n'$ and $\hat{l}$ are chosen based on a separate application of the Gaussian multiplier bootstrap: the former is a scaled quantile of a different Gaussian multiplier process, and the latter is based on a modified application of the popular *Lepskiĭ's method* [172]. Chernozhukov, Chetverikov and Kato showed that — under some conditions on the necessary intermediate quantities, $\mathcal{L}_n$, and $\kappa_l$ — the bands (4.17) constructed in this way are asymptotically honest and adaptive over a range of Hölder balls, subject to global upper and lower bounds on the densities and a modified version of the "self-similarity" notion mentioned previously. Furthermore, they showed that the worst-case coverage probability of their bands converges to the nominal level at a polynomial rate, asymptotically faster than the logarithmic rate associated with Smirnov-Bickel-Rosenblatt results. These theoretical results hold for KDE's with compactly-supported kernels and estimators using either compact or Battle-Lemarié wavelets, with regularity conditions based on the maximal degree of Hölder smoothness to which one wishes to adapt. In their supplementary material, the authors conducted a small simulation study. Although most of the intermediate quantities used to construct (4.17) must meet certain conditions (primarily in terms of their behaviour with respect to $n$), they simply experimented with predetermined numerical values for their simulations.

### Frequentist $L^2$ inference

Robins and van der Vaart [241] investigated the construction of $L^2$ confidence sets for conventional frequentist wavelet estimators. For a given wavelet basis[6], let $\theta(f)$ denote the expansion coefficients for an arbitrary $f$, and let $\hat{f}$ be the usual empirical wavelet density

[6]Actually, Robins and van der Vaart considered general orthonormal bases, not just wavelets. However, it seems appropriate to discuss their paper in this context, and to handwave some of the notation and technicalities for the sake of brevity.

estimator. Then their confidence sets are of the form

$$\left\{ f \in \mathcal{F} : \left\| \theta(f) - \theta\left(\hat{f}\right) \right\|_2 \leq \sqrt{\frac{\hat{\tau}_{K,n,\theta}}{\alpha} + \hat{R}_{K,n}\left(\theta\left(\hat{f}\right)\right)} + 2\hat{B}_K \right\}, \tag{4.19}$$

where $K = K(n)$ is a suitably-increasing bandwidth and the terms $\hat{\tau}$, $\hat{B}$, and $\hat{R}$ are estimates for variance, bias, and $\left\| \theta\left(f_0\right) - \theta\left(\hat{f}\right) \right\|_2$, respectively. They used sample splitting, assuming that the data used to calculate $\hat{f}$ is independent from that used for the other terms. These sets were shown to be honest at level $1-\alpha$, and adaptive to the fullest extent allowed by the theory without further restrictions[7]. Robins and van der Vaart mainly concerned themselves with the theoretical properties of such sets in various contexts. In practice, they may be difficult to construct due to the (likely unknown) quantities required for the calculation of the various terms in Equation 4.19. Bull and Nickl [31] further expanded upon the results of Robins and van der Vaart in the $L^2\left([0,1]\right)$ case, showing that honest and adaptive $L^2$ confidence sets over a wider range of regularity classes and Sobolev ball radii are possible by discretizing the smoothness range and using the "separation" approach from [100]. They constructed such a set in their proofs, somewhat similar in form to (4.19). Although they acknowledged the possibility of replacing some of the unknown terms in their construction by certain data-driven approximations, they did not consider applications to real data. Lerasle [173] provided a different approach to $L^2$ confidence balls, the full intricacies of which are omitted here. They used a model selection approach to determine the best approximation space (they dealt more generally with projection estimators on linear subspaces of an $L^2$ space, but for our purposes it suffices to consider the special case of wavelet estimators where the selection is for the truncation level) and a resampling method to estimate an $L^2$ norm needed in the radius of the set, thereby avoiding the sample splitting needed by some of the other literature discussed here. They showed that their confidence balls have the same adaptation properties as in [241], and that they are additionally *non-asymptotic*: they have correct coverage probability for *any* sample size $n$, not just in the limit.

### Some extensions and Bayesian ideas

Lounici and Nickl [185] defined a wavelet-based deconvolution density estimator analogous to the kernel one described in Section 4.3, based on a deconvolution kernel using Fourier transforms of the error density and wavelet basis functions. They used concentration inequalities and Rademacher processes to construct a confidence band, the radius of which is a complicated expression depending on the unknown density of the observed noisy data (although they noted that it can be replaced by the deconvolution estimator in practice).

[7] For instance, if $\mathcal{F}$ is a Sobolev ball of regularity $r$, the "fullest extent" in this $L^2$ context means adaptation over any nested Sobolev balls of regularity $s \in [r, 2r]$ [31, 101]. Recall that the $L^\infty$ context is even more restrictive than this [101].

Under some conditions on the bandwidth, error density, and smoothness of $f$, it is possible to control the probability with which the bands cover $f_0$ over all of $\mathbb{R}$.

For another somewhat unconventional theoretical example, Kerkyacharian et al. [152] developed an estimator for densities on homogeneous compact manifolds such as spheres. Their estimator is based on a *needlet* expansion of the density, where needlets form a basis with multiresolution properties similar to wavelets. They proposed a confidence band of random uniform width, discussed its (limited) adaptivity, and showed that its coverage probability can be controlled by undersmoothing.

Bayesian UQ literature for density estimators of this type is generally scarce, but some Bayesian results for histograms by Castillo and Nickl [43] can be more easily explained with the machinery of Haar wavelets. They used wavelet expansions of roughly the form (4.15) with $k_0 = 0$, the sums over $j$ restricted to $j = 0, \ldots, 2^k - 1$ to ensure the Haar system forms a basis of $L^2([0,1])$, and the sum over $k$ truncated to some upper limit $K$. This is equivalent to the basis function estimator (4.7) with $2^K$ piecewise constant basis functions instead of $K$. In the latter form, Castillo and Nickl placed a $2^K$-dimensional Dirichlet prior on the histogram coefficients $\boldsymbol{b}$, with $K = K_n$ chosen as a deterministic function of $n$ and the assumed Hölder regularity of $f_0$ (for regularities in the range $(1/2, 1]$). They proposed credible sets $C$ based on a "multiscale" approach:

$$C = \left\{ f : \max_{j,k} \frac{\left| \left\langle f - \hat{f}, \psi_{kj} \right\rangle \right|}{w_k} \leq \frac{R_n}{\sqrt{n}} \right\}, \tag{4.20}$$

where the inner product is the standard one on $L^2([0,1])$, $\hat{f}$ is the usual empirical wavelet estimator of $f_0$, $w_k$ is a sequence such that $w_k/\sqrt{k} \to \infty$ as $k \to \infty$, and $R_n$ is such that $\Pi(f \in C \mid \boldsymbol{X}) = 1 - \alpha$. Note that $R_n$ can be computed explicitly due to the conjugacy of the Dirichlet prior, since the likelihood depends only on the counts of observations in each "bin". Castillo and Nickl showed that the posterior over densities satisfies a sort of nonparametric Bernstein-von Mises property, and that these sets therefore have asymptotically correct frequentist coverage: $P_0(C \ni f_0) \to 1 - \alpha$ as $n \to \infty$. With a further refinement to their definition, their $L^\infty$-diameters also contract at a nearly-optimal rate in the "big-O in $\mathbb{P}_0$" sense. Unlike many of the other methods in this section, honesty and adaptivity are not implied here as the authors did not show the asymptotics to be uniform over all $f_0$ in some class. Although the choice of bandwidth $K$ depends on the regularity of the unknown $f_0$, they suggested that one could estimate a suitable bandwidth under a self-similarity assumption as in [100]. The geometry of these sets does not lend itself to visualizable error bounds. Instead, one can simulate from the posterior with MCMC, discard the 5% of function draws with the highest values for the multiscale quantity on the left-hand side of (4.20), and plot the remaining 95% to get a visual representation of the sets. This is the approach taken in, for example, the simulation study of [235], who considered similar theoretical ideas for

Bayesian UQ in the context of white noise and conjectured that they may be applicable to densities.

## 4.5 Adaptive basis expansion methods for log densities

An adaptive basis expansion does not have to be applied to the density itself as in the preceding section. Rather, it can serve as a model of the logarithm of the density, provided normalizing constant $c$ is incorporated:

$$\log f(x) = \sum_{j=1}^{K} b_j B_{j,K}(x) - c, \tag{4.21}$$
$$c = \log \int \exp \left[ \sum_{j=1}^{K} b_j B_{j,K}(t) \right] \mathrm{d}t.$$

Modelling the logarithm as a sum has a few nice consequences. In particular, it allows $f$ to be viewed as a member of an exponential family with sufficient statistics $\sum_i B_{j,K}(X_i)$, which makes it very easy to obtain an MLE $\hat{f}$ by maximizing $\sum_i \log f(X_i)$ with respect to $\boldsymbol{b}$ using (4.21) [e.g. 158]. Additionally, it is no longer necessary to constrain the coefficients.

### 4.5.1 Logsplines

One of the best-studied methods of this type is *logspline density estimation*. Assuming the density is supported on an interval and letting $L$ and $U$ denote its endpoints, let $\{B_{j,K} : j = 1, \ldots, K\}$ be a B-spline basis with knot sequence $L < t_1 < \ldots < t_m < U$ (recall Section 4.4.3). Although cubic splines are the most common choice, lower orders are possible; in particular, using splines of "order" 1 (equivalently, degree 0) corresponds to a histogram [275]. It is common to put some constraint on the tail behaviour of the estimate when using cubic splines, especially (but not exclusively) when $(L, U) = \mathbb{R}$, in which case the MLE $\log \hat{f}$ is typically required to be linear on $(L, t_1]$ and $[t_m, U)$ [127, 158]. If the support is a compact interval, another option is to require $\left(\log \hat{f}\right)''$ to be zero at $L$ and $U$ to reduce variance near the endpoints [160].

Stone [275] discussed some asymptotic theory for the maximum likelihood logspline density estimator, assuming the support is a compact interval ($[0, 1]$ w.l.o.g.) and the knots are equally spaced. They showed that, when $K$ increases to $\infty$ with $n$,

$$\frac{\hat{f}(x) - \overline{f}(x)}{\widehat{\mathrm{SE}}\left(\hat{f}(x)\right)} \xrightarrow{d} \mathcal{N}(0, 1)$$

for all $x \in [0, 1]$, where $\widehat{\mathrm{SE}}\left(\hat{f}(x)\right)$ is a standard error estimate involving values of the basis functions and derivatives of $c$ with respect to $\boldsymbol{b}$ (actual expression omitted for brevity), and $\overline{f}$ is the deterministic logspline density obtained by maximizing the *expected* (with respect

to $f_0$) log-likelihood. In a related technical report [274], Stone noted that this result can be used to obtain asymptotic confidence intervals for $f_0$, provided $K$ increases with respect to $n$ at a suitable rate depending on some underlying differentiability assumptions on $f_0$. As in many other cases, $K$ must increase faster than the error-optimizing rate for this, leading to undersmoothing.

A more comprehensive and practical treatment of pointwise inference for logsplines was given by Kooperberg and Stone [160]. They considered more involved knot placement schemes: one that involves stepwise selection, addition, and deletion, ultimately selecting the number of knots to optimize a generalized AIC [see 160, and references therein]; and a free knot placement scheme where knot locations and coefficients are *jointly* maximized, with the dimensionality again chosen by AIC. In either case, it is possible to estimate a standard error for $\log \hat{f}$ and use it to get approximate Gaussian pointwise intervals for the log density. By exponentiating the endpoints of these, Kooperberg and Stone obtained approximate confidence intervals for $f_0$. The only difference between the two knot selection procedures in this regard is the dimensionality of the gradients and Hessians required for the standard error estimate: the free knot procedure requires more components, since it is necessary to include derivatives with respect to knot locations. Additionally, for the stepwise procedure (in which the knots are considered fixed), the authors considered confidence intervals obtained via the bootstrap: either using percentile intervals, or plugging a bootstrap estimate of the standard error into the Gaussian interval approximation. The final UQ option they considered was a fully Bayesian approach, in which they put a hierarchical prior on $K$, knot placement (conditional on $K$), and coefficients $\boldsymbol{b}$ (conditional on knot placement and $K$). They simply took simulation quantiles from a reversible-jump MCMC procedure as pointwise credible intervals. In their simulation study, Kooperberg and Stone found that, for non-bootstrap methods, intervals based on the free knot procedure had higher coverage than those based on the stepwise procedure, but all non-bootstrap frequentist approaches consistently undercovered. Bootstrap methods based on the stepwise procedure were much better, although the percentile bootstrap tended to overcover (i.e. the intervals were perhaps too wide). Using a bootstrap standard error estimate with the stepwise procedure therefore appeared to be the best option, especially due to computational savings since fewer resamples were required than for the percentile bootstrap. They reserved analysis of the Bayesian approach for a real dataset, where they found that the credible intervals were much narrower than the "bootstrap standard error" confidence intervals, suggesting that the Bayesian approach may undercover. In a different publication [159], Kooperberg and Stone expanded somewhat on these results. There they found that the non-bootstrap frequentist intervals could achieve appropriate coverage on average when their widths were modified by some uniform scaling factor. Factors between 1.34 and 1.55 sufficed in their simulations depending on the specifics of the standard error calculations, but it was not clear how well these would generalize. They also found once again that the Bayesian intervals appeared too

small when applied to practical data, even with a larger prior covariance on the coefficients. Hansen and Kooperberg [127, in rejoinder to discussions] noted their challenges with UQ in Bayesian logspline estimation: they found it difficult to select priors that led to good point estimates *and* sensible credible intervals. More broadly, some authors have expressed skepticism about the usefulness of UQ for logspline density estimation, stating their view that pointwise confidence intervals do not generally provide useful shape information [158, 195].

### 4.5.2 General orthonormal bases

A few theoretical Bayesian papers discussed in Section 4.4 also provided analogous results for log density basis methods. Castillo and Nickl [43] modelled log densities with wavelets modified to form a basis of $L^2([0,1])$ instead of $L^2(\mathbb{R})$. The coefficients were given independent and identical priors — either Gaussian, Laplace, or something heavier-tailed — with a scale parameter depending on the Hölder regularity of $\log f_0$ (assumed to be $> 1$). Similarly to their histogram approach described in Section 4.4.4, the authors used a deterministic bandwidth choice and showed that multiscale credible sets of the form (4.20) have correct asymptotic frequentist coverage, with near-optimal diameter contraction possible with further refinements. The same comments about practicality made in Section 4.4.4 apply here. In a similar vein, Rousseau and Szabó [249] considered density estimators (supported on $[0,1]$) of the form (4.21) with an orthonormal basis of $L^2([0,1])$ such that $B_1 \equiv 1$, with the subscript $K$ removed since they did not consider basis functions changing with $K$. Among other technical conditions omitted here for brevity, they assumed $\log f_0$ has (up to a normalizing factor) an infinite series representation in terms of this basis; equivalently, that the true density is a member of an infinite-dimensional exponential family. With a suitable prior on $\boldsymbol{b} \mid K$ (a normal distribution with independent components is one example satisfying their conditions), the authors showed that (4.12) holds with Hellinger balls for both empirical and hierarchical approaches to $K$, just as it does for the histogram model. As in that case, honesty and near-adaptivity (up to a logarithmic factor) results hold over functions in a Sobolev ball of regularity $> 1/2$ satisfying their general polished tail condition. Unfortunately, their results remain difficult to put into practice due to the existence of the "blow-up factor" in the diameter of the sets.

## 4.6 Roughness penalty methods

Some of the frequentist estimators considered in Sections 4.4 – 4.5 were MLE's. In the i.i.d. case, one chooses $\hat{f}$ to maximize

$$\sum_{i=1}^{n} \log f(X_i)$$

over all $f$ in some predetermined class of possible estimators — generally those that can be expressed in the form of (4.7) or (4.21) — so that obtaining the estimate is simply a matter of optimizing the coefficients. In some cases it is advantageous to impose a further restriction on $\hat{f}$ to reduce variance or otherwise impose some desirable "baseline" shape properties. In this case, instead choose $\hat{f}$ to maximize

$$\sum_{i=1}^{n} \log f(X_i) - \lambda J(f) \tag{4.22}$$

over the estimator class, where the functional $J$ is some *roughness penalty*. This term forces $\hat{f}$ or $\log \hat{f}$ (depending on the context) to more closely resemble a function in the null space of $J$ to an extent controlled by the *smoothing parameter* $\lambda$. A common choice for $J$ is the integrated square of some linear differential operator: for instance, if $J : f \mapsto \int \left(\mathrm{D}^3 \log f\right)^2$, then as $\lambda \to \infty$, $\log \hat{f}$ is forced towards a quadratic shape, and therefore $\hat{f}$ towards a Gaussian [267]. For brevity, this case may be described as "penalizing [the size of] the third derivative" [232] on the log scale.

As indicated above, roughness penalties most commonly appear in the context of basis expansion methods, particularly spline fitting. When using splines with equally-spaced knots that do not repeat at the endpoints [74], an integrated squared $k^{\text{th}}$-order derivative penalty can be approximated by the sum of squared $k^{\text{th}}$-order differences between the coefficients. This simpler penalty gives rise to so-called *P-splines*, devised by Eilers and Marx [73]. In any case, such penalties are equivalent to quadratic forms in the basis function coefficients — for instance, the associated matrix for the aforementioned third derivative penalty consists of inner products between the third derivatives of the basis functions.

A Bayesian approach to roughness penalties is quite natural: it comes from viewing (4.22) as a log-posterior, with the first and second terms respectively corresponding to likelihood and prior. In this respect, the Bayesian methods of the previous two sections technically fit into this framework, but the focus in this section is on literature with a stronger emphasis on specific shape and smoothness restrictions imposed by the prior or penalty. The benefits of expressing penalties as quadratic forms as described above is now apparent: such a penalty is equivalent to an improper Gaussian prior on the spline coefficients (e.g. a P-spline penalty of order $k$ corresponds to a $k^{\text{th}}$-order Gaussian random walk; see Section 5.2.2), with $\lambda$ commonly given a Gamma hyperprior [e.g. 169]. Note that this type of prior is only suitable when modelling the log-density with basis functions — when using a basis expansion for the density itself, care must be taken to ensure that it is nonnegative and integrates to one. Some examples of this approach are given in Section 4.6.2.

### 4.6.1 Penalty methods for log-scale basis expansions

Although roughness penalty density estimators had already been developed by Good and Gaskins [106], Silverman [267] appears to have provided some of the earliest results for the

approximate distributions of such estimators. Letting $g = \log f$ and taking $J(g)$ (in a slight abuse of notation) to be the integrated square of some $m^{\text{th}}$-order linear differential operator on $g$, they considered the estimator $\hat{g} := \log \hat{f}$ which minimized (4.22) over all $g$ such that

1. $g$ has piecewise differentiable $(m-1)^{\text{th}}$ derivatives,
2. $J(g) < \infty$, and
3. $\int e^g < \infty$.

Silverman showed that, for bounded $f_0$ on a bounded univariate domain, $\hat{g}$ is asymptotically normal under suitable conditions on the higher-order derivatives of $\log f_0$ and the rate at which $\lambda \to 0$ as a function of $n$ and $m$. In principle this result could lead to some type of pointwise confidence intervals, but Silverman did not pursue this further. The mean and covariance functions for the limiting Gaussian process depend on eigenvalues of an inner product space of estimators, and it is not clear how to approximate these in practice. O'Sullivan [218] expanded further on Silverman's original ideas for univariate densities on compact intervals, and justified approximating $\hat{g}$ by cubic B-splines with knots at order statistics of $\boldsymbol{X}$. They proposed to calculate $\lambda$ by approximations to either a cross-validation score or an AIC-type quantity, and penalized the second derivatives of the log-densities. For uncertainty quantification, O'Sullivan adapted an idea from the non-parametric regression setting [291]: treating (4.22) as a log-posterior for the coefficients $\boldsymbol{b}$ in order to obtain "approximate Bayesian pointwise intervals". In the density case, O'Sullivan took a second-order Taylor series approximation of the unpenalized likelihood component $\sum \log f(X_i)$. This lead to an approximate Gaussian log-posterior, from which they derived pointwise intervals on the log scale of the form

$$\log \hat{f}(x) \pm 2\sqrt{\frac{2}{n}\boldsymbol{B}(t)^{\mathrm{T}}\left[\hat{\boldsymbol{H}} + 2\lambda\boldsymbol{\Omega}\right]^{-1}\boldsymbol{B}(t)}, \tag{4.23}$$

where $\boldsymbol{B}(t)$ is a vector of basis function evaluations, $\hat{\boldsymbol{H}}$ is the Hessian (with respect to $\boldsymbol{b}$) of the unpenalized likelihood at $\hat{\boldsymbol{b}}$, and $\Omega$ is the matrix of inner products associated to the roughness penalty. Presumably, confidence intervals for $f_0$ could be obtained by exponentiating the above expression. O'Sullivan did not comment on the performance of these intervals in their simulation study, but noted that they were found to have good coverage properties in the nonparametric regression setting by Wahba [291].

There are other formulations besides the Silverman approach for density estimation with roughness penalties. One such Bayesian approach came from Lambert and Eilers [168], who essentially used logistic regression to produce a smoothed estimate of a histogram. Suppose the density is supported on a bounded interval, which is partitioned into $J$ bins. Let $u_j$ and $m_j$ respectively denote the center of, and number of observations in, the $j^{\text{th}}$ bin $I_j$. Then

Lambert and Eilers proposed the model

$$(m_1, \ldots, m_J) \sim \text{Multinomial}(n, \boldsymbol{\pi}), \tag{4.24}$$

$$\pi_j = \frac{\exp\left[\sum_{k=1}^{K} b_k B_k(u_j)\right]}{\sum_{l=1}^{J} \exp\left[\sum_{k=1}^{K} b_k B_k(u_l)\right]}, \tag{4.25}$$

$$\boldsymbol{b}_{-K} \sim \mathcal{N}\left(0, (\tau\Lambda)^{-1}\right);$$

where the $B_k$'s are B-splines with equally-spaced knots, $b_K = -\sum_{k=1}^{K-1} b_k$ for identifiability, $\Lambda$ is a matrix of finite difference coefficients encoding a P-spline penalty, and $\tau$ is a precision parameter with a gamma hyperprior. For $x \in I_j$, one can take $f(x) = \pi_j / \ell(I_j)$ as a density estimate, where $\ell$ denotes the length of the interval. This penalized spline structure, combined with a high number of reasonably narrow bins, ensures the appearance of smooth estimates. Lambert and Eilers proposed this framework as a flexible way to handle grouped data by dividing the support into a smaller number of "wide bins" and replacing (4.24) with a multinomial model for wide bin counts, the probabilities for which are sums of the corresponding fine-grid $\pi$-values. Using a modified Langevin-Hastings algorithm to generate posterior samples, Lambert and Eilers applied this model to simulated and real data, using a moderately-sized cubic spline basis ($K = 20$). Unsurprisingly, their pointwise credible intervals (obtained from MCMC draws) exhibited higher variance when using larger "wide bins". In an earlier technical report, the same authors considered extensions of this model to multivariate densities by simply using products of B-spline bases, possibly allowing different dimensionalities and roughness penalties in each dimension [167].

### 4.6.2 Penalty methods for direct basis expansions

Roughness penalties can also be applied when modelling the density itself, rather than the log density, with basis functions. Komárek et al. [156] considered such a formulation to estimate the error density in accelerated failure time models. Rather than splines, they used Gaussian densities at fixed locations, which they noted to be the limiting case for B-splines as their degree tends to infinity. The number of basis functions in their model is determined by the desired distance between their means (which serve the same purpose as equally-spaced knots for splines), as is their standard deviation. To ensure their estimates were valid densities, the authors used a softmax transformation to obtain the coefficients $\boldsymbol{b}$:

$$b_k = \frac{e^{a_k}}{\sum_{l=1}^{K} e^{a_l}}. \tag{4.26}$$

For identifiability, it is necessary to fix, say, $a_K = 0$; a few other constraints on $\boldsymbol{a}$ are also necessary to ensure identifiability of other parameters in the failure time model. The

roughness penalty, based on second- or third-order finite differences, is imposed directly on $\boldsymbol{a}$. Estimation and inference follow from similar ideas as in O'Sullivan [218]: Komárek, Lesaffre and Hilton took a penalized maximum-likelihood estimate choosing the smoothing parameter by an approximate cross-validation score, and used a second-order Taylor expansion to obtain approximate pointwise "posterior" intervals for the density. They noted that in a simulation study (which they did not show), this method of constructing pointwise intervals yielded better coverage results than asymptotic methods. Komárek and Lesaffre [155] used a Bayesian version of this construction to model the errors and random effects in an accelerated failure time model with interval-censored data. As one might expect, the "logistic-scale" coefficients $\boldsymbol{a}$ in (4.26) were given (aside from a single identifiability constraint) a Gaussian prior with a (third-order) finite difference covariance structure, the scale of which is controlled by a smoothing parameter with a diffuse Gamma prior. Specifying the model in this way leads to related closed forms for estimated survival functions and densities of onset and event times. These functions can be simulated in an MCMC run, leading to pointwise credible intervals and means corresponding to posterior predictive functions. The simulation studyconducted by Komárek and Lesaffre [155] showed that such credible intervals did a good job of capturing the true densities of event and onset times, although their smoothness varied with different combinations of true random effect and error densities. Sharef et al. [263] provided an even more flexible Bayesian approach of this type to estimate the frailty density in a proportional hazards frailty model. They used a mixture of normalized B-splines and an optional parametric term, constrained to ensure the density has mean one. The authors considered the use of fixed splines, as well as a reversible-jump MCMC procedure allowing the number and location of knots (and therefore, of basis functions) to vary adaptively. For the latter, they put some truncated discrete prior on the number of knots, with their locations given a discrete uniform prior over a larger set of "candidate knots". Conditioned on dimensionality, they expressed the coefficients for the spline part of the model as in (4.26). They considered multiple choices for a smoothness-imposing prior on $\boldsymbol{a} \mid K$, listed below.

1. Simply taking the components of $\boldsymbol{a}$ to be i.i.d. Gaussians. The authors used this prior with adaptive knot selection, since the latter procedure controls smoothness automatically.

2. Taking $\boldsymbol{a}$ to be Gaussian with a covariance structure corresponding to second-order finite differences. The authors noted that this is only guaranteed to enforce smoothness for equally-spaced (fixed) knots.

3. Directly penalizing the second derivative of the spline mixture. This amounts to using a log-prior that is a quadratic form in $\exp(\boldsymbol{a})$ (with an associated matrix of inner products between B-spline second derivatives), divided by $(\sum_k e^{a_k})^2$.

In all cases, the prior for $\boldsymbol{a}$ has a scale parameter with an inverse-Gamma prior to control smoothing.The authors applied their approach to both simulated and real data, quantifying uncertainty with pointwise credible intervals from MCMC quantiles. Their simulation study showed that the adaptive knot selection approach without parametric component effectively captured the true frailty densities, although it required a sufficient quantity of data to do so (in particular, too few data clusters lead to wide pointwise intervals that did not adequately capture true shape information). On a real dataset with a modest number of clusters, they compared the fixed-knot version of their model (with second derivative penalty) to the adaptive knot procedure with different prior choices for the parametric component weight and number of knots. They found that the adaptive version with parametric components encouraged more smoothness in the posterior mean density and its credible intervals, to an extent determined by the choices of priors. However, the fixed-knot version with second derivative penalty performed best in terms of a modified Deviance Information Criterion.

This section concludes with a rather novel frequentist approach from Sardy and Tseng [254] which is better-suited to densities that may not be smooth in the sense of piecewise differentiability. They used estimators which are either piecewise linear between the order statistics of $\boldsymbol{X}$, or piecewise constant between their midpoints (equivalently, splines of degree 1 or 0, respectively), and *total variation* as their roughness penalty. The penalty is easily computed since their estimators ensure piecewise monotonicity, so that total variation is simply the sum of absolute differences between function values at consecutive order statistics. The authors devised two approaches for selecting the smoothing parameter: a universal one (depending only on sample size, not sample values) engineered to control the behaviour of $\hat{f}$ when the true density is uniform; and one based on a sparsity $\ell_1$ information criterion, in which $\lambda$ and $\hat{f}$ are jointly estimated. They used the latter approach on real datasets with some tied values due to rounding, and obtained 95% pointwise confidence intervals by bootstrapping. The pointwise intervals had reasonable width and shape, and the authors noted that they may allude to the existence of additional modes not captured in the "point estimates" of the densities.

## 4.7 Random measure mixture methods

This section explores uncertainty quantification for the canonical nonparametric Bayesian method of density estimation. In the general case, this method employs (conditional) mixtures of the form

$$f(\cdot \mid G) = \int \kappa(\cdot \mid \theta, \phi) \, \mathrm{d}G(\theta), \tag{4.27}$$

where $\kappa$ is some kernel with parameters $\theta$ and $\phi$, and the integral constitutes a mixture over the domain of $\theta$ with respect to a *random* probability distribution $G$. The bulk of the

nonparametric Bayesian literature uses infinite-dimensional discrete mixing distributions:

$$G(\cdot) = \sum_{i=1}^{\infty} w_i \delta_{Z_i}(\cdot), \tag{4.28}$$

where the locations and weights of the atoms — respectively, $Z$ and $w$ — are random sequences. The centrepiece of this Bayesian mixture model is the infinite-dimensional prior on $G$: a "distribution on distributions". As it pertains to density inference, the locations and weights are usually independent, with the former distributed according to some continuous "base measure" and the latter having a prior from one of two commonly-used broad classes.

1. *Normalized random measures with independent increments*, or NRMI's [236], in which unnormalized weights are generated from a Poisson point process [141] and subsequently normalized. The measure with unnormalized weights is a *completely random measure* (CRM).

2. *Gibbs-type random measures* of type[8] $\sigma \in (0, 1)$, which are equivalent to *$\sigma$-stable Poisson-Kingman processes* [104, 177]. Briefly, these arise from NRMI's with intensity measure corresponding to the $\sigma$-stable subordinator [98, p. 604] by conditioning the distribution of the weights on their sum $T$, then mixing over an arbitrary distribution for $T$ [227].

Assuming independence between weights and locations, each approach is a special case of the larger set of *Poisson-Kingman models* [227, 177], which are in turn a type of *species sampling model*. The *normalized generalized gamma* (NGG) processes comprise the intersection of these approaches [177], whereas the *Pitman-Yor* process [228] is an example of the second but not the first, as noted by Favaro and Teh [81]. It is well-known that both the NGG and Pitman-Yor processes admit the *Dirichlet process* as a limiting case when the parameter $\sigma \to 0$ [as mentioned in 182, for instance]; many Bayesian density inference papers are specifically devoted to so-called Dirichlet process mixtures. For the interested reader, Chapter 14 of Ghosal and van der Vaart [98] is an excellent exploration of the relationships between such discrete nonparametric priors.

For any of these priors on $G$, it is easily seen that its specification in the form (4.28) leads to another expression equivalent to (4.27):

$$f(\cdot \mid G) = \sum_{i=1}^{\infty} w_i \kappa(\cdot \mid Z_i, \phi). \tag{4.29}$$

[8]Other Gibbs-type random measures are possible for different values of $\sigma$. For $\sigma < 0$, they are mixtures (over the dimensionality) of finite-dimensional symmetric Dirichlet distributions; for $\sigma = 0$, they are mixtures (over the concentration parameter) of Dirichlet processes [104]. However, these are not typically seen in the density inference literature.

Discussion of the theoretical aspects of UQ, such as asymptotic coverage probability, appears scarce in the literature for such estimators. Instead, the focus is on practical generation of uncertainty sets (usually pointwise credible intervals) from posterior samples obtained via MCMC. As one might expect, difficulty arises here due to the nonparametric nature of the quantity of interest — in particular, since the posterior distribution of $G$ (this section hereafter adopts the bracket notation of Gelfand and Smith [90], denoting this posterior by $[G \mid \boldsymbol{X}]$) will be infinite-dimensional. The key to most ideas for MCMC sampling of this model is to reformulate it in a hierarchical way:

$$\begin{aligned} X_i &\sim \kappa\left(\cdot \mid \theta_i, \phi\right), \\ \theta_i &\sim G, \\ G &\sim P\left(\cdot \mid \psi\right). \end{aligned} \tag{4.30}$$

If there *are* additional hyperparameters $\phi$ and $\psi$, they are typically given their own independent priors, but these are not a main focus here. The latter encodes all parameters of the prior for $G$: for instance, for a Dirichlet Process prior with Gaussian base measure it may include the concentration parameter, as well as the location and scale of said base. Note that by the almost-sure discreteness of $G$, there is positive probability that $\theta_i = \theta_j$ for some $i \neq j$. In this respect, the model imposes a random partitioning or clustering of the data, where each cluster is comprised of all observations with the same $\theta$ value. With this formulation in mind, the known MCMC strategies divide into two main groups: *marginal* and *conditional*, depending on the way in which the infinite-dimensional parameter $G$ is handled. The sections below briefly explain, and discuss the UQ implications for, each of these groups.

### 4.7.1 Marginal sampling methods

Marginal methods rely on integrating $G$ out of the model and being able to obtain approximate samples from $[\theta \mid \boldsymbol{X}]$. Algorithms for this purpose are readily available when using the Dirichlet Process prior; see Neal [206] for a seminal review of them. In this case, it is easy to obtain a Monte Carlo estimate of the posterior mean density (denoted here as $f\left(\cdot \mid \boldsymbol{X}\right)$, in keeping with the rest of the Bayesian notation in this section), as discussed by Escobar and West [75]. Letting $\theta*$ denote the parameter for a hypothetical new observation and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$, note that $f\left(\cdot \mid \boldsymbol{\theta}\right) = \int f\left(\cdot \mid \theta*\right) \mathrm{d}\Pi\left(\theta* \mid \boldsymbol{\theta}\right)$. The integrand is simply the kernel $\kappa$, and the distribution $[\theta* \mid \boldsymbol{\theta}]$ is readily available. Assuming the base measure of the Dirichlet process is conjugate to the kernel (as in the Gaussian case, for instance), this integral has an analytic closed form. From there, the Monte Carlo estimate of $f\left(\cdot \mid \boldsymbol{X}\right) = \int f\left(\cdot \mid \boldsymbol{\theta}\right) \mathrm{d}\Pi\left(\boldsymbol{\theta} \mid \boldsymbol{X}\right)$ is an average of the above quantity over posterior MCMC draws of $\boldsymbol{\theta}$. By the same token, it is easy in the conjugate case to quantify uncertainty with respect to $[f\left(\cdot \mid \boldsymbol{\theta}\right) \mid \boldsymbol{X}]$. This is essentially the approach suggested by Wang and Dunson

[294] to find pointwise confidence intervals, although they further simplified inference by using a greedy algorithm to find an optimal partition of the data. They noted that the deterministic nature of their algorithm results in an underestimation of uncertainty.

Inference of this nature ether ignores or marginalizes out uncertainty in the weights of $G$. For marginal samplers, this seems to be fairly standard practice when obtaining posterior density estimates to construct credible sets. Shi et al. [266] used one of Neal's nonconjugate algorithms [206] and obtained posterior density draws by taking the mean of the $\kappa(\cdot \mid \theta_i)$'s for each MCMC draw of $\boldsymbol{\theta}$ [see also their R package 265, and its source code[9]]. This is equivalent to taking a mixture of cluster-specific kernels, each weighted by the number of observations in its corresponding cluster. Using kernel- and data-specific scaling and a low-information prior, Shi et al. obtained pointwise credible intervals for simulated and real data. Their framework can accommodate censored data, with pointwise uncertainty increasing in the presence of censoring as expected. In their simulation studies, the credible intervals did a good job of capturing the true densities, covering them throughout the domains for all one-dimensional examples and at roughly 98% of domain points for their two-dimensional example. Favaro and Teh [81] and Favaro et al. [79] devised marginal Gibbs samplers for NRMI's and a specific subclass of $\sigma$-stable Poisson-Kingman models, respectively. For both classes, their density draw computations appear[10] to be based on truncation, and marginalization of the distribution of $G$ given $\boldsymbol{\theta}$ and the auxiliary variables of the sampler. To elaborate, the density draws are a sum of cluster-specific kernels, each given weight proportional to its (conditional) expected mass; and ten "new" kernels with parameters taken from the prior base measure, each given weight proportional the expected *total* mass divided by ten. In the latter paper, the authors showed a pointwise credible interval for the density of a dataset of galaxy velocities, noting that the results were satisfactory and consistent with previous work.

It could be argued that the aforementioned approaches to density inference are inherently "incomplete". Indeed, marginalizing or otherwise deterministically approximating the random weights of $G$ fails to account for some of the uncertainty in (4.29). If the goal is full uncertainty quantification in this regard, the focus must be on $[f(\cdot \mid G) \mid \boldsymbol{X}]$ if possible. As noted by Gelfand and Kottas [89], it holds that

$$[\boldsymbol{\theta}, G \mid \boldsymbol{X}] \propto [\boldsymbol{\theta} \mid \boldsymbol{X}]\,[G \mid \boldsymbol{\theta}]\,. \tag{4.31}$$

This reveals the key to fully meaningful inference with a marginal sampler: for each MCMC draw $\boldsymbol{\theta}_b \sim [\boldsymbol{\theta} \mid \boldsymbol{X}], b = 1, \ldots, B$, if it is possible to draw $G_b \sim [G \mid \boldsymbol{\theta}_b]$, then the quantities $\{f(\cdot \mid G_b)\}$ constitute a posterior sample from $[f(\cdot \mid G) \mid \boldsymbol{X}]$. Gelfand and Kottas [89] noted

[9] Available at `https://github.com/cran/DPWeibull`.

[10] Based also on their source code at `https://github.com/BigBayes/BNPMix.java`.

that this is easy for the Dirichlet process prior by conjugacy, since $[G \mid \boldsymbol{\theta}]$ is a Dirichlet process with updated parameters. Of course, in practice the infinite sum in (4.29) must be somehow truncated to obtain actual density draws. Gelfand and Kottas did this by choosing the number of terms to satisfy a predetermined expected error threshold, then replacing the final weight to ensure that the truncated sum integrates to one. Kottas [163] later used this approach in the context of survival analysis, as did Griffin [108] when comparing different approaches to hyperpriors in the Dirichlet process model. Such methodology is not typically used for more general random measure priors, despite relevant distributional results existing in the literature [81, 79]. This is likely a computational matter: to directly sample the weights $\boldsymbol{w}$ of a random measure, it is typically necessary to employ a *stick-breaking process*, in which they are represented as

$$w_i = V_i \prod_{j=1}^{i-1} (1 - V_j) \tag{4.32}$$

for certain continuous random variables $\{V_i\}$ on $[0, 1]$. It is well-known that the Dirichlet process with concentration parameter $M$ has a stick-breaking representation of the form (4.32) with $V_i \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, M)$ [262]. However, such representations for the general classes of random measure considered here are more recent developments, and the densities of the $V_i$'s are quite complicated [83, 80].

### 4.7.2 Conditional samplers

In contrast to the approaches described above, conditional methods *do* produce posterior samples of the weights in (4.28), allowing for "full" inference on functionals such as (4.29). There are several ways to avoid the problem of having to sample infinitely many weights. Early conditional samplers simply replaced $G$ by a finite approximation, choosing the deterministic truncation level *a priori*. Discussion of such methods is deferred to Section 4.7.4; this section focuses on alternatives that better incorporate the infinite-dimensional nature of the model. Perhaps the most common approach to this end is to introduce some auxiliary variables such that the full conditionals of $G$ are finite-dimensional. This ensures that Gibbs samplers target the correct posterior without the need for approximation, aside from the inevitable truncation to calculate the density draws themselves.

The retrospective sampler of Papaspiliopoulos and Roberts [220] was one of the earliest methods of this type for Dirichlet process mixtures. It involves the introduction of *allocation variables* $\boldsymbol{K} = (K_1, \ldots, K_n)$ such that $K_i = j$ iff $\theta_i = Z_j$, with $Z_j$ as in (4.28). At each step of the chain, first draw only $\max(\boldsymbol{K}) := \max_i \{K_1, \ldots, K_n\}$ of the atoms and weights in $G$. A certain condition involving auxiliary standard uniform variables and the full conditionals of $\boldsymbol{K}$ is then checked. If the condition is met, perform a Metropolis-Hastings update of $\boldsymbol{K}$ and resume sampling as normal; otherwise, simulate additional components of $G$ one

at a time (from their priors, as they represent clusters with no allocated observations) until the condition is met. Note that the number of components is therefore variable across iterations. The authors noted that posterior draws for any linear functional of $G$ are equal in distribution to a deterministic function of prior draws and the first $\max(\boldsymbol{K})$ components from one retrospective sampling iteration. Thus, full posterior inference for $f(\cdot \mid G)$ is quite straightforward.

Another popular approach which avoids some of the computational burden of the retrospective algorithm is *slice sampling*, first used in this context by Walker [292]. Briefly, Walker's original idea involved introducing new latent variables $U_i, i = 1, \ldots, n$ such that, with $K_i$ again denoting the allocation variable as above, the joint likelihood for observation $i$ is

$$f(X_i, U_i, K_i = j \mid G) = \kappa(X_i \mid \theta_j) \mathbb{1}(U_i < w_j). \tag{4.33}$$

Integrating out $K_i$ and $U_i$ reduces this to (4.29). Furthermore, these variables ensure that all full conditionals in the Gibbs sampler - including those for the necessary components of $G$ - are finite-dimensional. Numerous adaptations of the algorithm exist: for instance, Kalli et al. [148] altered it for greater efficiency. Technical details aside, the main point is to simulate the finitely (but randomly) many components of $G$ needed for the other sampler variables; this can exceed $n$, in which case some components will correspond to clusters with no data allocated. Favaro and Walker [82] adapted the algorithm of Kalli et al. to the larger class of $\sigma$-stable Poisson-Kingman models, using their stick-breaking representation to devise a method for sampling the weights. They applied their method with mixtures of Gaussians with common variance and means from the random measure. Density draws were calculated by first adding together the components obtained from the sampler, then allocating the remaining mass (which the authors noted was usually quite small) to a Gaussian kernel with the sampled posterior variance centered at the prior mean of the base distribution. One can then extract posterior sets from these draws in the usual way. In the same paper discussed in Section 4.7.1, Favaro and Teh [81] considered a slice sampler for NRMI mixtures; here they sampled the *unnormalized* masses of the random measure. They showed pointwise credible intervals for the densities of some real datasets that were reasonable in shape and variability. Their source code suggests that they used the same formula for density draws here as they did for their aforementioned marginal samplers, with ten additional "new" kernels as described in the previous section.

Although finite-sum approximations are always necessary for density estimation, the approaches described above are noteworthy because the samplers *themselves* introduce no truncation error; all of their full conditionals are truly finite-dimensional. This is not the case for *all* papers which use conditional samplers for density inference. For instance, Barrios et al. [13] used a conditional algorithm for NRMI's that does not induce a finite-dimensional

full conditional for $G$. Instead, they used a representation which allowed them to sample the masses in decreasing order. This allowed them to select the number of components sampled based on a relative error criterion, and to calculate density draws from only these (normalized by the sum of the sampled masses). They obtained pointwise credible intervals for a real dataset, demonstrating that the choice of both kernel and NRMI prior can moderately affect the smoothness of said intervals. Argiento et al. [6] folded random truncation into a modification of the NRMI prior itself by discarding all unnormalized weights smaller than some threshold $\epsilon$. The resulting random measures have finite and random dimension, and converge in distribution to the corresponding NRMI's as $\epsilon \to 0$. The authors recommended fixing some small value for $\epsilon$ (it is possible to place a prior on it, but they warned that the computational cost may be unreasonable). They derived a conditional sampling algorithm, introduced a new class of NRMI's with a Bessel function in the intensity measure, and applied their method to real and simulated data. Their pointwise credible intervals showed a pleasing degree of smoothness and reasonable faithfulness to the true density of a simulated sample. Griffin [109] proposed an adaptive truncation method based on sequential Monte Carlo. The method involves iteratively resampling and increasing the dimension of the approximate model until a discrepancy measure falls below some threshold. Griffin applied this approach to a variety of nonparametric models, including Dirichlet process mixtures. Although they did not show credible intervals for densities, they did so in the context of time series modelling, indicating that density inference is indeed possible in this framework.

### 4.7.3 Extensions

#### Feller-Dirichlet priors

This Bayesian model from Petrone and Veronese [226] generalizes the Dirichlet process mixture model, although it also serves as an extension of Petrone's ideas [224, 225] from Section 4.4.2. Recall from that section that Petrone put a prior on $K$ and introduced latent variables $Y_1, \ldots, Y_n$ from a random distribution $F$ with DP prior such that $X_i \mid Y_i, K, F \sim \text{Beta}\left(\lceil KY_i \rceil, K - \lceil KY_i \rceil + 1\right)$. The *Feller-Dirichlet prior* generalizes this by replacing the latter beta densities by some kernels $g_K\left(\cdot; Y_i\right)$, leading to a density model of the form

$$f\left(\cdot \mid K, F\right) = \int g_K\left(\cdot; \theta\right) \mathrm{d}F\left(\theta\right).$$

Petrone and Veronese provided several examples beyond the original Bernstein model that are suitable for data on $[0, \infty)$ or $\mathbb{R}$. For instance, take $g_K\left(\cdot; \theta\right)$ to be an inverse Gamma density with parameters $(K, K\theta)$ with a Gamma base measure for the prior on $F$, or use a $\mathcal{N}\left(\theta, \sigma^2/K\right)$ density for the kernel with a Gaussian base measure. These examples illuminate the idea that the Feller-Dirichlet prior — a "mixture of DP mixtures" — bridges the gap between Dirichlet process mixture models and the Bernstein polynomial models explored previously. For inference, Petrone and Veronese truncated the DP to a large finite number

of components and used a Gibbs sampler similar to that of Ishwaran and Zarepour [140] to obtain density estimates and pointwise credible intervals.

**Extensions for non-i.i.d. data**

Several extensions to the random measure density model also exist for data structures besides an i.i.d. sample $\boldsymbol{X}$, most of which are based on the Dirichlet process instead of the more general measures. Müller and Rodriguez [203] and the references therein provide an excellent overview of such extensions; this section details some examples for which uncertainty quantification has been done in literature. In broad terms, these examples all involve inference for a family of densities $\{f(\cdot \mid G_t) : t \in \mathcal{T}\}$, where the random measures are indexed by some set $\mathcal{T}$ and share some form of dependence. In many cases, this will mean modelling the density for a "response variable" $X$ with associated covariate $t$, effectively building yet another bridge between density estimation and nonparametric regression.

The dependent Dirichlet process (DDP) first introduced by MacEachern [189] is the basis for many useful models. DDP mixtures are similar in construction to (4.29), except that the weights $\{w_{tj}\}$ and locations $\{Z_{tj}\}$ may both vary with $t \in \mathcal{T}$. For instance, De Iorio et al. [64] considered a model for survival analysis when there are covariates $t_i$ associated with each observation $X_i$. The weights do not vary with $t$, but the $Z_{tj}$'s correspond to location-scale pairs with the former component equal to a linear model in $t$: for example, if $t = (u, v)$ for categorical $u$ and continuous $v$, then $Z_{tj} = (m_j, A_{uj}, B_j v, \sigma^2)$ for $j \in \mathbb{N}$. Inference proceeds by reformulating the model into the conventional DP mixture framework, replacing the top line of the hierarchy in (4.30) by

$$X_i \mid \theta_i, t_i \sim \mathcal{N}(\theta_i d_i) \tag{4.34}$$

where $d_i$ is a design vector so that $\theta_i d_i = (m_j + A_{uj} + B_j v, \sigma^2)$ when $\theta_i = j$ and $t_i = (u, v)$. De Iorio et al. used this so-called *linear DDP* to analyze the densities of log survival times with various combinations of treatments and other factors. They showed some pointwise credible intervals for survivor and hazard functions, and although they did not do so for densities, it should be no more difficult. However, their inferential approach [as described in 63] is the same as that suggested by Escobar and West [75], where the weights are marginalized so that inference is based on $[f(\cdot \mid t, \theta) \mid \boldsymbol{X}]$ as opposed to $[f(\cdot \mid G_t) \mid \boldsymbol{X}]$.

The above formulation is somewhat similar to the *density regression* model considered by Dunson et al. [70] for modelling the density of a continuous response variable. The model assumes a set of continuous covariates associated to each observation and a structure similar to (4.34), except that the random measure governing the $\theta$-value for an observation now depends on the corresponding covariate vector $t$: it is a finite mixture of $n$ i.i.d. Dirichlet processes, with the $i^{\text{th}}$ weight based on the distance between $t$ and $t_i$. Dunson et al. used a marginal sampler, so that posterior inference for the predictive density of a "new" obser-

vation (given some covariate vector) was once again based only on the finite-dimensional parameters. Draws for these densities have closed forms due to conjugacy: they are mixtures of cluster-specific kernels and one using posterior draws of the hyperparameters of the base distribution. For both real and simulated data, the authors showed pointwise credible intervals for such densities conditioned on various values for the covariates. In the latter case, the intervals did a good job of capturing the true densities. Dunson and Park [69] subsequently developed the *kernel stick-breaking process* (KSBP) to model an uncountable collection of probability distributions (with particular focus on the density regression application), generalizing and expanding upon some of the ideas in [70]. In this model, the covariate-dependent distribution for an observation's $\theta$-value is an *infinite* mixture of "basis" random measures (typically either point masses or draws from a Dirichlet process) with stick-breaking mixture weights. To induce dependence on the covariates, the beta random variables defining the stick-breaking process are weighted by kernels evaluated at the covariate value and centered at random locations with some arbitrary prior distribution. Dunson and Park's MCMC algorithm for pointwise UQ was a hybrid between marginal and conditional: like [70], they marginalized over the basis random measures; but at the $t^{\text{th}}$ step of the chain they sampled $M_t$ mixture weights, where $M_t$ is the highest index of an occupied cluster across the first $t$ iterations. The authors repeated the same simulation study as in [70], showing that the pointwise credible intervals from the KSBP model enveloped the true densities. Norets and Pelenis [213] explored the same simulated data model, showing how changes in the KSBP hyperparameters affected the quality of inference.

The formulation in the preceding paragraph directly model the conditional density of $X$ given $t$ by specifying a covariate-dependent random measure. Alternatively, it is possible to first model the *joint* distribution of $X$ and $t$ as a mixture of a kernel $\kappa(X, t \mid \theta, \psi) = \kappa(X \mid t, \theta)\,\kappa(t \mid \psi)$ with respect to a random distribution on the product space for $(\theta, \psi)$, then obtain the desired (conditional) density estimates by standard calculations. This approach is used by Park and Dunson [221], who put a Dirichlet process prior on the product measure; and Wade et al. [290], who gave separate DP priors to $G_\theta$ and $G_{\psi|\theta}$ to allow for greater flexibility. Both used marginal samplers for inference (again, with uncertainty only in terms of the finite-dimensional parameters), with the latter finding pointwise credible intervals to be much narrower and more accurate than those resulting from a DP on the product measure.

Returning to the DDP, note that it is also a suitable starting point when there are multiple sets of observations from different discrete time points, in which case the density is a random process evolving through time. Nieto-Barajas et al. [210] used this approach in such a context, making the atom locations independent of time but introducing dependence into the weights through their stick-breaking construction. They achieved the latter by introducing latent variables $Y_{tj}$ dependent on the stick-breaking proportion $V_{tj}$, such that $V_{(t+1)j}$ is in turn dependent on $Y_{tj}$ and the usual Dirichlet process is recovered by marginalizing out

the $Y$'s. They applied this construction in a mixed-effects model for protein activation over time, using a partially marginalized algorithm which exploited conjugacy to sample only atoms corresponding to clusters containing data. Müller and Rodriguez [203] showed densities with pointwise credible intervals from the same application, presumably using the same algorithm. Gutiérrez et al. [117] used a different approach to introduce dependence in the stick-breaking process: with random probability $p$ having a beta prior, they sampled $V_{(t+1)j}$ from its usual distribution, and set it equal to $V_{tj}$ otherwise. They used slice sampling for inference, but did not specify if their density draws incorporated any components beyond those sampled (recall that this was the case for the Favaro-authored papers in Section 4.7.2). Their simulation study showed that their method was much more effective than one based on spline regression at capturing the true shape of their density, but their pointwise credible intervals did a much better job at enveloping the true density at later time points than at earlier ones.

Finally, there may be multiple samples $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m$ for which it makes sense "share information", assigning mutually dependent densities to each sample. The hierarchical methods discussed in Müller and Rodriguez [203] and its references are perhaps the most natural ways of doing this, but there does not appear to be existing literature which specifically conducts UQ with these methods. Griffin et al. [110] developed an interesting model: starting with $p$ underlying i.i.d. CRM's, the mixing distribution for each density is the normalized sum of some sample-specific subset of the underlying measures. Griffin et al. called this the *correlated NRMI* model, and implemented it with a combination of slice sampling and a split-merge step (in which clusters are moved between the underlying measures to address posterior multimodality). Although the main purpose of their model was assessing differences between distributions, they did show pointwise intervals for survival functions fitted from interval-censored data; as always, it seems reasonable to assume that density inference is possible by similar means.

### 4.7.4 Finite mixtures

As previously mentioned, one way around the difficulties of infinite-dimensional models is to simply truncate the sum in (4.29) at some level $N$. This case leads to a vector of weights $\boldsymbol{w} = (w_1, \ldots, w_N)$ on the $N-1$-dimensional probability simplex. This was the approach taken by the early conditional samplers of Ishwaran and Zarepour [140] and Ishwaran and James [139], who considered generalized Dirichlet priors on $\boldsymbol{w}$ to approximate random measures with stick-breaking representations (namely, those for which the stick-breaking variables $V_j$ in (4.32) have beta distributions). For instance, to approximate a Dirichlet process mixture with concentration parameter $\alpha$, they would either give $\boldsymbol{w}$ a symmetric Dirichlet prior with parameters $\alpha/N$; or truncate its stick-breaking representation, setting $V_N = 1$ to ensure the $N$ weights summed to one. They gave asymptotic justifications (as $N$ grows large) for both options. With the conditional samplers devised in these papers,

approximate posterior inference is obviously possible. Of course, extensions to the types of data structures considered in Section 4.7.3 can also be considered. For instance, Chung and Dunson [53] modelled covariate-dependent densities using truncated random measures with stick-breaking weights derived from a probit model. Their structure for the weights incorporated a variable selection component, resulting in a rather flexible density regression framework. Finucane et al. [84] conducted a meta-analysis of child nutrition data by modelling the study-specific densities of interest with finite mixtures of normals, using probit model stick-breaking weights which incorporated individual time and location effects. Norets and Pelenis [212] modelled the joint distribution of a response variable and covariates with a finite Gaussian mixture, obtaining the conditional response densities with standard calculations. Their model allows for any number of discrete variables by mapping them to continuous latent variables. The pointwise credible intervals obtained in these papers showed reasonably good uncertainty quantification, although the choice of a fixed finite number of components naturally reduces their flexibility somewhat.

The focus thus far in this section has been overwhelmingly Bayesian. Frequentist approaches to mixture models do exist in the literature, but it is rare to see them consider density UQ as it is defined here. Roeder [245] provided one rather novel exception for mixture-of-Gaussians estimators with finitely supported mixing distributions. Given some bandwidth $h$ for the Gaussian kernel $\kappa$, the mixing distribution $\hat{G}_h$ is uniquely chosen to optimize an asymptotically normal statistic based on sample spacings[11]. This statistic is nonincreasing in $h$, and so it is possible to find a range of $h$-values such that the statistic falls within the $(\alpha/2)$- and $(1-\alpha/2)$-quantiles of the standard normal distribution. The confidence set defined by Roeder is then the set of all estimators $f\left(\cdot \mid \hat{G}_h\right)$ as $h$ varies through this range. This set is comprised entirely of finite mixtures (although the number of components for each is random), is easy to visualize, and provides correct coverage if the true density is assumed to be a mixture of Gaussians.

In addition to the KDE connection, it is easy to see parallels between finite mixtures and some of the basis expansion methods discussed earlier. Indeed, even if one were to put a prior on $N$ (e.g. Norets and Pati [211], whose inference involved modelling conditional densities using covariate-dependent multinomial logit mixture weights), the model would be similar in principle to the fully Bayesian approaches in Section 4.4. Thus, beyond what has already been explored, there is little else to discuss here. The interested reader may refer to Chapter 22 of Gelman et al. [93] for some more details on working with models of this type.

[11] Roeder noted the analogy between such estimators and KDE's, the difference being the sample-dependent mixing distribution used. Similar connections and generalizations were briefly explored in Section 4.3.3.

## 4.8 Other methods

This section explores uncertainty quantification for an assortment of density estimation methods for which literature is too scarce to warrant separate sections.

### 4.8.1 Nearest neighbour methods

This classical density estimator is closely related to the KDE and is applicable to any density on $\mathbb{R}^d$. Let $k = k(n)$ be an integer increasing with sample size $n$, let $\|\cdot\|$ be some norm on $\mathbb{R}^d$ (typically Euclidean, but some other norms also satisfy the required conditions for some of the results discussed here), and for an arbitrary point $x \in \mathbb{R}^d$ let $R(k, x)$ be the $\|\cdot\|$-distance between $x$ and the $k^{\text{th}}$-closest value in $\boldsymbol{X}$. Then for a kernel $K$, the *nearest neighbour density estimator* as defined by Mack [190] is

$$\hat{f}(x) = \frac{1}{R(k,x)^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{R(k,x)}\right). \tag{4.35}$$

Unless otherwise stated, all results in this section require $K$ to equal 0 outside of the unit $\|\cdot\|$-ball. A particularly common case arises from the uniform kernel:

$$\hat{f}(x) = \frac{k}{nV(k,x)}, \tag{4.36}$$

where $V(k, x)$ is the volume of the $\|\cdot\|$-ball centered at $x$ with radius $R(k, x)$. Nearly all of the literature on NN density inference is theoretical, and closely mirrors the results discussed previously for KDE's[12]. For instance, Theorem 9.3.7 in Csörgő [59] is essentially a Smirnov-Bickel-Rosenblatt result for univariate NN estimators. Unlike the KDE and wavelet theorems, their formulation would lead to confidence bands over a certain *random* interval defined by order statistics of the sample, but they noted that this interval converges to the full support as $n \to \infty$.

Moore and Yackel [202] provided what appear to be the first asymptotic normality results for (4.35), showing that the limiting distribution could be made to center at $f_0$ under some conditions on its properties and the asymptotic behaviour of $k$. They also noted that the asymptotic variance of the NN estimator is smaller than that of the KDE at points $x$ where $f_0(x)$ is small, claiming that this makes it more efficient for estimating density tails. Mack and Rosenblatt [191] expanded on this, noting that the NN estimator can be much more biased than the KDE in the tails, with the opposite relations holding for large values of $f_0(x)$. These observations, combined with the non-monotonic dependence of asymptotic bias on $k$, make error analysis here somewhat less straightforward than it is for the KDE.

[12] Unfortunately, even the drawbacks are similar: most asymptotic results relevant to inference require a choice of $k$ which is *not* optimal w.r.t. the mean square error [e.g. 59]

Mack [190] derived slightly different asymptotic normality results than Moore and Yackel, centering at $\mathbb{E}\left[\hat{f}\right]$ instead of $f_0$. This allows for less restrictive conditions: for instance, theirs are the only results here which do not require $K$ to vanish outside of the unit ball. Pointwise Gaussian limits centered on $f_0$ with some variant of usual conditions (among others, as needed) are also available for univariate NN density estimates with non-i.i.d. data structures, such as randomly right-censored data [199], observations from an $\alpha$-mixing sequence [179, only for the uniform kernel], or randomly left-truncated samples [298, who actually implemented confidence intervals in practice using a plug-in estimator of the asymptotic variance].

A technical report by Rodríguez [244] [see also 243] made an interesting connection between NN estimators of the form (4.36) and KDE's: the former allocates the fixed mass $k/n$ to the random volume $V(k, x)$, while the latter can be rewritten to show that it essentially does the opposite, spreading a random mass over a fixed volume. This observation motivated Rodríguez to view the two estimators as endpoints on a "continuum" of estimators of the form

$$\hat{f}(x) = \frac{c \int K\left(\frac{x-t}{\mu}\right) \mathrm{d}F_n(t)}{\int_0^1 h^d(t) \mathrm{d}\omega(t)},$$

where $\omega$ is a distribution on $[0, 1]$ with mean $c$, and the (possibly random) number $\mu$ and function $h$ meet certain technical conditions. Rodríguez showed how KDE's and uniform NN estimators arise as special cases and described everything in-between as "double smoothing": in the numerator (resp. denominator), the mass (resp. volume) given by $F_n$ (resp. $h^d$) is smoothed with $K$ (resp. $\omega$). Rodríguez proved asymptotic normality for certain subclasses of these estimators in this report, as did Biau et al. [19] for another variant. Both cases are generalized NN estimators, and the results hold even using the "optimal" $k(n)$ with given asymptotic biases. It is possible to eliminate the bias and center at $f_0$ with a suboptimal $k_n$, although the conditions for this are less restrictive here than in [202] at the expense of stricter smoothness assumptions on $f_0$.

### 4.8.2 Logistic Gaussian process estimators

This approach is usually Bayesian and involves density estimates of the form

$$f(x) = \frac{e^{g(x)}}{\int e^{g(u)} \mathrm{d}u}, \tag{4.37}$$

where the latent function $g$ is given a zero-mean *Gaussian process (GP)* prior with hyperparameters $\gamma$ governing the covariance kernel. The "logistic" transformation of $g$ ensures that the estimates are valid densities: nonnegative and integrating to one. Riihimäki and Vehtari [239] explored some approaches for approximate Bayesian inference with this model with 1-

or 2-dimensional densities. Technically, they assumed that $g$ would be the sum of a Gaussian process and a parametric polynomial component, but they integrated out the coefficients for the latter so that the basis function values and hyperparameters could simply fold into the mean and variance of the GP. Similarly to Lambert and Eilers [168] (see Section 4.6.1), Riihimäki and Vehtari discretized the model, replacing the actual data with observation counts in a fine, equally-spaced partition of the domain. Assuming that the partition consists of $J$ subregions and letting $\boldsymbol{m}$ and $\boldsymbol{g}$ respectively denote the vectors of observation counts and latent function values within each subregion, the likelihood $\mathbb{P}(\boldsymbol{m} \mid \boldsymbol{g})$ is essentially the same as (4.24 – 4.25) [168], except the B-spline values in (4.25) are replaced by the latent function values $g_j$ for $j = 1, \ldots, J$. In turn, the prior $\Pi(\boldsymbol{g} \mid \gamma)$ for the latent values is simply the multivariate normal distribution induced by evaluating the GP prior at the center points of the subregions. The main object of inference is then the conditional posterior of $\boldsymbol{g}$ given the observation counts and hyperparameters (and, technically, conditioned on the chosen partition as well),

$$\Pi(\boldsymbol{g} \mid \boldsymbol{m}, \gamma) \propto \mathbb{P}(\boldsymbol{m} \mid \boldsymbol{g}) \Pi(\boldsymbol{g} \mid \gamma). \tag{4.38}$$

This posterior is not analytically tractable, so approximate methods must be used to employ this model in practice. As an alternative to MCMC, Riihimäki and Vehtari proposed the use of a *Laplace approximation* (see Section 2.2) to $\Pi(\boldsymbol{g} \mid \boldsymbol{m}, \gamma)$, thereby obtaining a Gaussian distribution for $\boldsymbol{g}$. In order to quantify uncertainty in $f$, samples must be drawn from this approximate Gaussian posterior and transformed via (4.37). To this end, the authors showed that importance sampling can improve performance, and rejection sampling can also be incorporated to ensure appropriate tail behaviour if necessary. The model is completed by putting a prior on $\gamma$, but Riihimäki and Vehtari also considered the possibility of ignoring the uncertainty in these hyperparameters: marginalizing the approximate Laplace posterior over $\boldsymbol{g}$, maximizing it with respect to $\gamma$, and simply plugging in the resulting approximate MAP point estimate for $\gamma$. They found that their method performed (in terms of mean log predictive density, evaluated with cross-validation for real data or w.r.t. the true distribution for simulated data) comparably with MCMC targeting the true joint posterior of $(\boldsymbol{g}, \gamma)$, as well as the Dirichlet process mixture models of Griffin [108]. The pointwise credible intervals for real and simulated data provided reasonable practical visualization for uncertainty quantification. However, one of their simulations showed that densities with varying amounts of smoothness throughout the domain can be challenging, as the MAP parameters needed to capture more narrow features can result in excessive roughness elsewhere. The authors also showed how their method can extend to density regression, modelling densities conditional on covariate values.

### 4.8.3 Pólya trees

The Pólya tree (PT) prior is a Bayesian nonparametric method for constructing a random probability measure, discussed in [170] and the first few references therein. The construction is based on a recursive partitioning of the domain and is most easily explained when the domain is an interval in $\mathbb{R}$. At the $m^{\text{th}}$ level of partitioning, the interval is split into $2^m$ subintervals. It is common to set the partition boundaries to the dyadic quantiles of some base measure $G_0$ (i.e. $G_0^{-1}\left(j/2^m\right), j = 0, \ldots, 2^m$), thus "centering" the random measures drawn from the PT prior around this base [170, 203]. Associate to each $m^{\text{th}}$-level subinterval a binary number $\epsilon = \epsilon_1 \ldots \epsilon_m \in \{0,1\}^m$, and define a set of beta random variables $\{Y_\epsilon : \epsilon \in \{0,1\}^m\}$ such that the $Y_{\epsilon_1 \ldots \epsilon_{m-1} 0}$'s are mutually independent and $Y_{\epsilon_1 \ldots \epsilon_{m-1} 1} = 1 - Y_{\epsilon_1 \ldots \epsilon_{m-1} 0}$. Finally, consider a random probability measure that assigns mass

$$\prod_{j=1}^{m} Y_{\epsilon_1 \ldots \epsilon_j}.$$

to $B_\epsilon$, where $B_\epsilon$ is the subinterval associated to binary number $\epsilon = \epsilon_1 \ldots \epsilon_m$. Iterating this process over all $m \in \mathbb{N}$ results in a draw from the Pólya tree prior (so named because the recursive partitioning defines a tree with nodes corresponding to subsets), defined by the sequence of partitions and beta parameters. A special case for the latter gives rise to the Dirichlet process, but they can also be tailored to almost surely produce absolutely continuous distributions [e.g. 170, 204], which is obviously more appealing for density inference.

It is possible to extend this construction to $d$-dimensional domains, for instance by using the construction of Hanson [129]. At the $m^{\text{th}}$ level, the domain is partitioned into $2^{md}$ subsets, indexed by base-$2^d$ numbers $\epsilon = \epsilon_1 \ldots \epsilon_m \in \{0, \ldots, 2^d - 1\}^m$ [203]. These subsets are formed by taking Cartesian products of the subintervals used in the univariate construction, then applying a suitable affine transformation. Probabilities are assigned to each subset in an analogous way to the univariate case, except that for a fixed $\epsilon_1 \ldots \epsilon_{m-1}$, the variables $\left\{Y_{\epsilon_1 \ldots \epsilon_{m-1} e}, e = 0, \ldots, 2^d - 1\right\}$ have a $2^d$-dimensional Dirichlet distribution. Literature on multivariate PT's rarely entails any density UQ, so the remainder of this section focuses primarily on the univariate case.

Castillo [42] provided theoretical results for posterior inference with such priors on the unit interval, with partition boundaries at the dyadic rationals. In particular, they showed that, when $f_0$ is Hölder with regularity $\beta \in (0, 1]$ and bounded away from zero, a type of *Bernstein-von Mises result* holds (i.e. the posterior weakly [43] converges in $\mathbb{P}_0$-probability to a Gaussian process) when the beta parameters of the prior grow suitably fast with $m$. The posterior must be centered at some estimator for $f_0$ for this to hold: either the posterior mean or, when the beta parameters grow suitably slowly depending on $\beta$ (note that this corresponds to "undersmoothing" of the posterior), a "canonical" estimate based on the Haar wavelet expansion of the empirical measure. Castillo noted that this result can lead

to similar results to some of those discussed earlier for wavelet estimators [43]: namely, multiscale credible bands similar to (4.20) with Pólya trees should have correct frequentist coverage.

Practical implementations of Pólya tree models involve truncating the partitioning at some finite "terminal" level, rather than continuing it infinitely. By a well-known conjugacy result [e.g. 204, 129], the posterior for the PT prior is simply an updated PT, with the same partition and updated beta (or Dirichlet, in the multivariate case) parameters for the $Y_\epsilon$'s. With the aforementioned truncation, density samples from this posterior can be obtained by allocating the mass proportion within each terminal subset either uniformly [98, chapter 3] or according to the density of the base measure [as in 129]. The resulting densities will be discontinuous at the partition boundaries [204, 98] and are therefore perhaps not as "well-behaved" as one may prefer. In a survival model with longitudinal data and a PT prior on event times, Zhang et al. [300] addressed this issue by applying kernel smoothing to the actual posterior PT draws to obtain event time densities. There are other ways around this which change the structure of the model itself: mixing the prior over the parameters of the base distribution [129], adding random "jitter" to the partition boundaries [219], or mixing a kernel with respect to a PT measure [24]. Surprisingly, literature employing such methods does not tend to address UQ for densities. On the other hand, Nieto-Barajas and Müller [209] did so for their *rubbery Pólya tree* (rPT) prior, introducing dependence amongst the $Y_{\epsilon_1 \ldots \epsilon_{m-1} 0}$'s at level $m$ (i.e. all "left nodes" in the tree at a given depth) through latent variables. The construction resembles that used to introduce dependence for time-series DDP's by Nieto-Barajas et al. [210] as discussed in Section 4.7.3, and recovers the usual PT prior by marginalizing over the latent variables. Conditional conjugacy allows for an easy Gibbs sampler, which Nieto-Barajas and Müller implemented by truncating the partitioning process at some depth (using a depth between 5–8 in all experiments) and allocating the mass uniformly within each of the terminal subsets. Pointwise credible intervals in their simulation study fully contained the true densities, but were not smooth. Indeed, the rPT only "smooths" the estimates in the sense of reducing jump sizes between masses in neighbouring partition sets. Its dependence structure addresses variability, not continuity. Nieto-Barajas and Müller suggested mixing (either over a kernel w.r.t. a rPT prior, or over the parameters of the rPT's base distribution) when more smoothness is desired, but did not attempt uncertainty quantification with such models.

A different extension of the model came from Hanson et al. [128], when each $X_i$ is observed at a spatial location $t_i$. Their object of interest was the predictive density (i.e. marginalizing over $G$) for a new $X$, and they proposed to modify the usual formula by weighting the contribution of each observation by some distance between their locations and that of the new $X$. Uncertainty was with respect to the (hyper)parameters of the PT prior and the distance function and was quantified with MCMC output. Their pointwise

credible intervals appeared quite smooth; it is unclear whether this is the result of an actual procedure or merely the plotting functions used.

### 4.8.4 Multiscale estimators

This rather novel Bayesian approach from Canale and Dunson [38] uses multiscale mixtures of Bernstein polynomials as estimates:

$$f(\cdot) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{sh} \text{Beta}(\cdot; h, 2^s - h + 1). \tag{4.39}$$

The weights $\pi_{sh}$ are constructed in terms of a stochastic process defined on an infinite binary tree. For the $h^{\text{th}}$ node at tree depth $s$ ($h = 1, \ldots, 2^s$), let $S_{sh}$ be the probability of stopping at that node and $R_{sh}$ be the probability (conditional on *not* stopping) of moving to the right daughter of node $(s, h)$. These probabilities define a sort of "random climb" on the branches of the tree, which at each step either stops with some probability or else moves on to the next depth, randomly choosing either the left or right path. The weight $\pi_{sh}$ is then the probability of the process taking the path to node $(s, h)$ (starting from the root of the tree) and then stopping there. For instance, $\pi_{12} = (1 - S_{00}) R_{00} S_{12}$, and $\pi_{23} = (1 - S_{00}) R_{00} (1 - S_{12}) (1 - R_{12}) S_{23}$. The specification of the model is completed with priors $S_{sh} \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, a)$ and $T_{sh} \overset{\text{i.i.d.}}{\sim} \text{Beta}(b, b)$, where $a$ and $b$ can be fixed or given their own hyperpriors.

Canale and Dunson noted that this model induces an interesting multiscale clustering on the data: two data points may be assigned to the same tree node at some depth $s$, meaning they are sufficiently similar to be clustered together at this scale; but may occupy different nodes at a depth $r > s$, so that they are separated at a higher "resolution". For practical inference, they truncated or "pruned" to a maximum tree depth $S$ by simply setting the stopping probabilities $S_{Sh} = 1$ for all $h$. Using a slice sampling approach, they devised an MCMC algorithm for inference, which alternates between two steps: assigning each observation to a tree node (equivalently, to a "multiscale cluster") given the $\pi_{sh}$'s, and updating the $S_{sh}$'s and $R_{sh}$'s given these allocations. Posterior density samples can then be obtained by plugging these probabilities into (4.39), truncated accordingly. Although Canale and Dunson did not show any credible intervals for densities in their paper, the corresponding R package for this type of model implements them readily [37].

### 4.8.5 Shape-restricted methods

If an *a priori* assumption can be made about the shape of the true density $f_0$ (for instance, that it is monotone or unimodal), one may wish to incorporate this into estimation and inference. A solid body of literature exists on the use of such shape constraints in non-

parametric estimation, but only a subset of this literature specifically considers UQ for densities.

Perhaps the best-studied shape constraint is monotonicity, in which $f_0$ is assumed to be non-decreasing. In the frequentist setting, the so-called *Grenander estimator* is the canonical choice for estimation of $f_0$, and is also the MLE subject to the monotonicity constraint [107]. Letting $F_n$ denote the empirical distribution function of $\boldsymbol{X}$, let $\hat{F}$ be the *least concave majorant* of $F_n$: the smallest concave c.d.f. such that $\hat{F} \geq F_n$ throughout the entire support (typically assumed w.l.o.g. to be $[0, 1]$, or $[0, \infty)$) [112]. The Grenander estimator $\hat{f}$ is then the left derivative of $\hat{F}$, which turns out to be a step function with jumps at sample values and $\hat{f}(x) = 0$ for $x \leq 0$ and $x > X_{(n)}$ [233]. Rao [233] derived a pointwise limiting distribution for this estimator, showing that with suitable standardization it is asymptotically equivalent to a particular functional of Brownian motion[13]. Groeneboom and Jongbloed [111] leveraged this fact to derive the asymptotic distribution of a likelihood ratio test statistic for $f_0(x)$ when $f_0$ has nonzero derivative in a neighbourhood of $x \in (0, \infty)$. The limiting distribution is that of a different functional of Brownian motion derived by Banerjee and Wellner [11]. The authors of that paper did not find an analytic form for this distribution, but provided estimates of its quantiles from simulation-based methods. Groeneboom and Jongbloed used these estimated quantiles to obtain pointwise confidence intervals with asymptotically correct coverage by inverting their likelihood ratio test. They also considered pointwise bootstrap intervals based on a boundary-corrected kernel (under-)smoothing of the Grenander estimator. The use of the bootstrap in this way is at least partially justified by an asymptotic normality result for this smoothed Grenander estimator [114] (indeed, there are a few modifications to this method that result in smooth, asymptotically normal estimators; see also [282]). Unfortunately, the bootstrap is unsuitable for inference with the unaltered estimator, due to the inconsistency results shown by Kosorok [162] and Sen et al. [260] and demonstrated in practice by the latter. However, both papers showed that consistency can be restored with a smoothed bootstrap (i.e. resampling from a modified kernel estimate of $f_0$, rather than from the empirical distribution). Kosorok further showed that smoothed bootstrap methods could be used to define an $L^1$-ball of functions centered at $\hat{f}$ with correct asymptotic coverage, based on the known asymptotic normality of the $L^1$-error [112]. Recall, however, that such sets are limited in visual interpretability. Uniform confidence bands were briefly considered by Durot et al. [71], who derived a Gumbel limiting distribution similar to the Smirnov-Bickel-Rosenblatt results in Section 4.4.4. However, they believed that the technicalities required for data-driven construction of such a band were not worth exploring further. Deng et al. [66] proposed another method to construct pointwise intervals, based on the adaptation of an analogous method for inference in iso-

[13]The full details of this functional are omitted here, but its distribution is commonly known as the *Chernoff distribution*, which commonly arises in shape-constrained nonparametric inference.

tonic regression (see their manuscript for details). They suggested that their method, which involves suitable estimates of nuisance parameters in the limiting distribution, could be tailored to adapt to the smoothness of $f_0$ more readily than the method of Groeneboom and Jongbloed [111], but both methods require simulation-based estimates for quantiles of the complicated limiting distribution.

As an alternative to frequentist inference methods based on the Grenander estimate (or some modification thereof), Bayesian methods are also available. For instance, Martin [193] proposed an empirical prior in which the density is modelled as a finite scale mixture of uniform densities. The mixture weights and uniform density scales are respectively given Dirichlet and Pareto priors, both of which are calibrated so the prior over densities is centered at some predetermined mode. This mode (and the dimensionality of the mixture) can either arise from a *sieve* MLE (i.e. the MLE over a space whose size depends on $n$) or the Grenander estimate. Using simulated data and MCMC, Martin compared the pointwise credible intervals from this model to those obtained from a Dirichlet process mixture, and found that the empirical model resulted in higher coverage probability and shorter intervals on average.

The second most common shape constraint explored in the literature is arguably log-concavity, in which $\log f_0$ is assumed to be concave. As in the monotone case, frequentist UQ for log-concave densities typically centers on the MLE $\hat{f}$. Rufibach [252] showed that the log of $\hat{f}$ is piecewise linear with breaks at sample values, and $\hat{f}$ supported on the range of $\boldsymbol{X}$. Balabdaoui et al. [10] obtained a limiting distribution for this estimator under some regularity conditions on $f_0$. Much like the monotone case, the MLE for log-concave densities converges in distribution to a certain functional of Brownian motion, scaled by nuisance parameters that depend on the value of $f_0$ and its derivatives. Azadbakhsh et al. [7] translated these results into practical means of constructing pointwise confidence intervals. They estimated the necessary quantiles of the limiting distribution by simulation, and considered several methods (kernel-based and plug-in) to estimate the $f_0$-dependent nuisance parameters. The intervals thus obtained performed reasonably well in a simulation study, although the best overall results came from standard bootstrap percentile intervals. Compared to the bootstrap intervals, the pointwise intervals based on asymptotics generally had a somewhat higher propensity for undercoverage in some parts of the domain, and for overcoverage (i.e. coverage probability exceeding the desired nominal level, leading to wider intervals than necessary) in other parts. Despite these promising empirical results, the authors cautioned that there were no theoretical results justifying bootstrap methods for this purpose.

Mariucci et al. [192] developed a Bayesian model for log-concave densities $f$:

$$f(x) = \frac{e^{w(x)} \mathbb{1}_{[a,b]}(x)}{\int_a^b e^{w(u)} \mathrm{d}u},$$
$$w(x) = \gamma_1 \sum_{j=1}^{m} p_j \frac{\min\{\theta_j, x-a\}}{\theta_j} - \gamma_2 (x-a). \tag{4.40}$$

The function $w$ is piecewise linear with $m$ break-points, where $m$ is a predetermined number dependent on sample size. The weights $p_1, \ldots, p_m$ can either be given a Dirichlet prior, or a prior based on truncating the stick-breaking representation of the Dirichlet process (4.32). The support $[a, b]$ can be deterministic (based on $n$), empirical ($a = X_{(1)}$, $b = X_{(n)}$), or hierarchical ($a$ and $b - a$ given their own priors). Priors on $\gamma_1 \geq 0$, $\gamma_2 \in \mathbb{R}$, and $\theta_1, \ldots, \theta_m \in [0, b-a]$ complete the model, and posterior density draws can be obtained from MCMC samples of these parameters using (4.40). See Mariucci, Ray and Szabó for technical details, as well as motivation for (4.40). Pointwise credible intervals obtained with MCMC did a good job capturing true densities in their simulation studies, although in some cases they underperformed somewhat around boundaries or modes. The authors also evaluated the coverage probability of the intervals in one example, showing a tendency for undercoverage in some parts of the domain but overall reasonable performance with increasing sample sizes.

Similarly to [233] and [10], complicated limiting distributions have been derived for density estimation under different shape constraints. Examples include monotonicity with right-censored data [137], convexity [113], and $s$-concavity [126]. In principle these limiting distributions could be used to derive practical UQ methods for densities as in the examples described above, but there does not appear to be any literature directly doing so.

### 4.8.6 Connections to nonparametric regression

Various parts of this chapter have suggested that some uncertainty quantification ideas from other nonparametric models could apply for density estimation. Indeed, there exist a great deal of theoretical results showing that many such models are "equivalent" in a sense involving asymptotic convergence of their risks [e.g. 214, 28, and references therein, especially those by Lucien Le Cam]. Brown et al. [27] offered a practical way of leveraging these ideas. They proposed the *root-unroot algorithm* to estimate a density on, say, $[0, 1]$ via nonparametric regression. The algorithm proceeds as follows.

1. Divide the domain, assumed w.l.o.g. to be $[0, 1]$, into $T$ equal subintervals.
2. For $j = 1, \ldots, T$, let $Y_j = \sqrt{Q_j + 1/4}$, where $Q_j$ is the count of $X_i$'s in the $j^{\text{th}}$ subinterval and the offset of $1/4$ gives optimal bias and variance properties.

3. Treat the $Y_j$'s as response variables and use any suitable method to fit the corresponding smooth regression function $\widehat{m}$.

4. Take $\hat{f}(\cdot) = [\widehat{m}(\cdot)]^2 / \int [\widehat{m}(t)]^2 \, \mathrm{d}t$ as the density estimate.

Wang [295] used the root-unroot algorithm for Bayesian density inference, using *integrated nested Laplace approximations* (INLA) for the posterior of the regression model. The details of INLA — first given by Rue et al. [251] — are omitted here, but it suffices to note that it uses Gaussian approximations and numerical integration to approximate the posterior, allowing for inference without MCMC being necessary. In the above algorithm, Wang took $\hat{m}$ to be the posterior mean from the INLA model. Letting $\gamma$ denote the normalizing integral in the denominator, they divided the INLA quantiles of $m$ by $\gamma$ to obtain approximate pointwise credible intervals for $f$. Such intervals did an excellent job capturing true density shapes in their simulation studies.

More broadly, one may exploit the connections described here to quantify density uncertainty with any number of methods originally devised for nonparametric regression. Examples include confidence bands based on coverage of surrogate functions [94], or on relaxed notions of coverage that still try to minimize the extent to which the band excludes the true function but allow for nice adaptivity properties [35].

## 4.9 Simulation study

Recall Figure 4.1 from Section 4.2, which shows select combinations of density estimation and UQ methods for a simulated dataset. Having described many such methods in the preceding sections, a more thorough discussion of the figure is presented here.

The dataset $\boldsymbol{X}$ is a sample of size $n = 1000$ from the mixture density $f_0 = 0.5\mathcal{N}\left(\frac{1}{2}, \frac{1}{49}\right) + 0.5\mathcal{N}\left(\frac{5}{7}, \frac{1}{490}\right)$. This is a bimodal, everywhere-positive density with almost all of its mass contained in the interval $[0, 1]$, and its magnitude and curvature are close to zero at the boundaries of this interval. Thus, it "approximately satisfies" the assumptions made by many of the UQ methods discussed here, while having a fairly interesting shape which provides a good test for UQ methods.

The methods applied to $\boldsymbol{X}$ and shown in Figure 4.1 are as follows.

1. KDE with pointwise bias-corrected confidence intervals as in Calonico et al. [36], and fixed-width bootstrap confidence bands based on the same bias correction [48]. The bandwidth was selected to minimize estimated integrated MSE (instead of pointwise MSE as in the former reference) in order to ensure a smooth estimator.

2. Adaptive basis expansion with Bernstein polynomials as in Petrone [224], with pointwise credible intervals and credible bands based on median absolute deviations [72].

3. Logspline estimation with stepwise knot selection [158] and exponentiated pointwise Gaussian confidence intervals using bootstrap standard error estimates [159].

4. A Dirichlet process mixture of Gaussians with a Normal-Inverse Gamma base measure. A marginal MCMC sampler was used (see Section 4.7.1) but pointwise credible intervals incorporated "full uncertainty" by using posterior draws of the Dirichlet process obtained by conditional conjugacy [89].

All Bayesian methods were based on output from an appropriate MCMC sampler, and the level for all UQ methods was taken to be $1-\alpha = 0.95$. The simulation study was conducted with the R programming language [230], and further details can be found in Appendix A and [198].

As noted in Section 4.2, the bands are expectedly wider than the pointwise intervals for both estimation methods shown on the top row of Figure 4.1. Note that the confidence sets for the KDE are not centered at the estimator due to the bias correction, and are in fact closer to the true density. However, they still fail to fully reach the height of the main mode. Certainly no conclusions can be made about the coverage probability of any UQ method when it is applied to only a single dataset, but further simulations (not shown; see [198]) suggested that this deficiency is typical for samples taken from the true density $f_0$, even when using pointwise instead of integrated MSE to select bandwidths. In fact, sample sizes in the millions were necessary to attain good coverage probability at the main mode, although the performance was much better at the smaller mode even for $n = 1000$. To some degree this is to be expected as the coverage error depends on higher-order derivatives of $f_0$ [36], but it is infeasible to fully predict this error in practice. This leads to an important point to be made about the difference between asymptotic and finite-sample behaviour: although Calonico et al. [36] showed that these confidence intervals have coverage error ultimately decaying at the optimal rate with respect to $n$, there are no concrete guarantees for any finite sample size when using data-driven methods.

Recall from Section 4.3.2 that Cheng and Chen [48] provided bootstrap methods for both fixed- and variable-width bias-corrected confidence bands for KDE's. Here the former was used, as the latter involves bootstrapping a quantity which can have a zero denominator when using a compact kernel, as was the case here [198]. In contrast, the credible band used for the Bernstein polynomial estimator has variable width (see (4.13)). However, the band shown in the top-right plot of Figure 4.1 extends over the subinterval $[0.01, 0.99]$, as the band taken over all of $[0, 1]$ was far too wide to be graphically meaningful. This is because a sizeable proportion of MCMC draws had absolute deviations near the boundaries that were much larger than the MAD there, so that the quantile $\xi_\alpha$ in (4.13) was very large. In turn, the MAD was comparatively small at the boundaries because, like $f_0$ itself, most MCMC draws had tail values near zero. These examples demonstrate that variable-width bands may not be an ideal choice unless $f_0$ is bounded suitably far away from zero.

Interpretation of the bottom row of Figure 4.1 is straightforward. The pointwise intervals for the logspline estimator are noticeably less smooth than those for the other estimation methods. Recall that the width of the interval (on the log scale) is determined by the pointwise sample variance of bootstrap density estimates [159]; evidently this induces some roughness. The pointwise credible intervals for the DP mixture are quite similar to those for the Bernstein polynomial estimator: both are quite narrow and smooth and encompass $f_0$ throughout nearly the entire domain.

## 4.10 Conclusion

There is a vast, sprawling body of work on density uncertainty quantification, dating back over half a century and spanning across many different methods for both estimation and inference. Reviewing the literature — from classical approaches like KDE's and histograms, to the spline methods of the late twentieth century, to modern nonparametric methods — one notices that the gap between theoretical and practical ideas seems to have widened over time. KDE's and related methods are extremely well-studied, with a litany of theoretical and practical results for all relevant types of UQ. Turning the focus to the past two decades of developments, one sees that UQ in the literature for random mixtures is entirely practical, with almost no regard for asymptotic properties; conversely, the advanced wavelet-based papers comprising the core of new theoretical developments often include no data studies whatsoever. It seems natural to wonder whether it is possible to "bridge the gap": perhaps introducing greater theoretical justification for some of the most commonly-used practical methods, or facilitating applications of some of the more obscure asymptotic arguments. However, such developments may be hampered by issues intrinsic to the problems at hand, such as the known complexities of asymptotics in nonparametric Bayesian inference [e.g. the review of 248]. The importance of these considerations is certainly a subjective matter, and as modern practitioners turn their focus to larger datasets and more overt "data science" approaches, there is perhaps a case to be made that applications could provide "all the proof we need".

Based on the simulation study described in Section 4.9, Figure 4.1 shows finite-sample results for a few of the methods discussed throughout this chapter. The code for these experiments is available in the original publication's supplementary material [198], and there is certainly merit to further comparative analysis beyond that considered here.

Of interest for future work are extensions to frameworks beyond a single i.i.d. sample, particularly hierarchical modelling of multiple related densities. Bayesian nonparametric methods are emerging as a promising approach to such frameworks, and we are eager to explore the improvements which further developments can provide.

# Chapter 5

# FRODO: a novel approach to micro-macro multilevel regression

## 5.1 Introduction

Hierarchically structured data is quite common in statistics, with a litany of resources and methodology available for almost every imaginable configuration. Books such as [105] provide comprehensive reviews on the subject of multilevel data. For the purposes of this chapter, it will suffice to consider data organized in a two-level hierarchy. Data will be observed from "groups", each of which is comprised of multiple "individuals", with variables measured at either the group level (i.e. one measurement per group) or individual level (i.e. one measurement for each individual within each group).

Multilevel data structures can be broadly categorized into two types: *macro-micro*, in which an individual-level outcome is predicted from group-level covariates; and *micro-macro*, which is the opposite [270]. Although substantial attention has been given to the former structure (random effects models being one example of the macro-micro framework), the micro-macro paradigm is the subject of much less discussion [86], despite the occurrence of such datasets in health sciences [60], sociology [14], and economics [4]. Among the relatively few papers on the subject is the one by Croon and van Veldhoven [57], one of the earliest papers to devise a method specifically for micro-macro regression. The data structure they considered (hereafter described as "classical") is as follows. Letting subscripts $i$ and $ij$ denote, respectively, the $i^{\text{th}}$ group and the $j^{\text{th}}$ individual within that group, the basic structure is

$$Y_i = \alpha + \beta\xi_i + \beta_Z Z_i + \epsilon_i, \tag{5.1}$$

$$X_{ij} = \xi_i + \nu_{ij}. \tag{5.2}$$

Assuming group $i$ contains $n_i$ individuals, the observed data corresponding to that group is $\{Y_i, Z_i, X_{i1}, \ldots, X_{in_i}\}$. In words, $Y_i$ is a group-level response variable (with regression error $\epsilon_i$), $Z_i$ is a group-level scalar covariate, and the $X_{ij}$'s are individual-level measurements

of some "latent" unobserved covariate $\xi_i$ with errors $\nu_{ij}$. One can think of the model as two "parts": a regression part specified by (5.1), and a covariate observation part specified by (5.2). The linearity of the regression and additivity of the covariate error justify the "classical" moniker for this structure.

Although micro-macro modelling literature is relatively scarce, the structure implied by (5.1–5.2) is essentially equivalent to (a version of) the much better-studied *classical measurement error model* [chapter 1 of 41, and references therein]. The main difference is conceptual: in a micro-macro model, replicate covariate measurements correspond to distinct individuals within a group; while in a measurement error model, they are merely repeated noise-corrupted observations of some true explanatory variable for the $i^{\text{th}}$ observational unit. There is another practical difference: most measurement error literature assumes smaller $n_i$'s (the number of covariate measurements per group) than one tends to encounter in a "true" micro-macro setting.

The simplest approach to modelling such data is the "naive" one: simply using the sample means $\bar{X}_i = n_i^{-1} \sum_j X_{ij}$ as proxies for the latent $\xi_i$'s. However, it is well-known [e.g. chapter 3 of 41, and references therein] that such a failure to account for the uncertainty in the $X_{ij}$'s biases estimates of the regression parameters. Most notably, it creates *attenuation* in the estimate of $\beta$: letting $\hat{\beta}$ denote such an estimate, we will have $|\widehat{\beta}| < |\beta|$, even as the number of groups grows asymptotically. In intuitive terms, this attenuation happens because the noise in the covariates stretches the regression line on the horizontal axis. Thus, a plethora of both frequentist and Bayesian methods have been proposed to account for covariate uncertainty in a way that produces less biased estimation and inference for the regression part of the model. A comprehensive review of measurement error methodology is beyond the scope of this chapter, but the interested reader may refer to books such as [32, 41] or the review paper of Schennach [257].

Many real-world datasets do not obey the "classical" framework of (5.1–5.2) [e.g. Section 6.4 of 32, and references therein], and there are two ways to transcend it: by replacing the linear terms $\beta\xi_i$ and $\beta_Z Z_i$ in (5.1) with arbitrary regression functions, or by generalizing the additive covariate structure in (5.2). There are few micro-macro modelling papers with generalizations of either type, aside from the discrete variable methods of Bennink et al. [14, 15]. Thus, we focus our attention here on the measurement error literature instead. Beyond the comprehensive review sources mentioned above, the most generalized framework which is relevant to this chapter is that of Hu and Schennach [136]. They assumed each observational unit $i$ only has a single covariate measurement $X_i \sim f_{X|\xi=\xi_i}$, but also has a single replicate measurement or *instrumental variable* $W_i$, assumed to provide further information about $\xi_i$. They also allowed a very general form for the regression function in which $Y$ only depends on the unobserved $\xi$, with only some technical assumptions on the distributions of $Y \mid \xi$, $X \mid \xi$, and $\xi \mid W$. Their assumptions on the covariate structure were very broad, requiring only that there exists a functional $M$ such that $M\left[f_{X|\xi}\left(\cdot \mid \xi\right)\right] \equiv \xi$

for all $\xi$. Examples of such functionals include the mode, as well as any quantile or moment. With this framework, the authors proposed a sieve likelihood estimator for the regression parameters and the densities of $X \mid \xi$ and $\xi \mid W$. To our knowledge, there are no established Bayesian methods that accommodate this level of generality. Sarkar et al. [255] proposed a Bayesian model which used Dirichlet Process mixtures to achieve a great deal of flexibility in modelling the regression function, latent covariates, and error terms; but it still assumed an additive error structure of the form (5.2).

Neither of the aforementioned papers (or, indeed, any measurement error literature we have seen) gives much consideration to the "unit-specific" covariate distributions $f_{X|\xi=\xi_i}$ — specifically, to any differences between them across units. This is understandable, as most errors-in-variables problems have no more than a single-digit number of covariate measurements available per unit, making any such differences irrelevant. However, in an explicitly multilevel setting, there are typically many more individuals per group [e.g. 57, 4], and it may be of interest to explicitly consider the group-specific covariate densities in inference. We believe that the Bayesian paradigm (or, at the very least, the empirical Bayesian paradigm) is the most natural setting in which to achieve this.

With all of the above considerations in mind, our goals in this chapter are threefold. First, we seek to develop a(n empirical) Bayesian model with generality comparable to that of Hu and Schennach [136]. Second, we wish to apply this model in the micro-macro multilevel setting, providing an ability to accommodate "non-classical" data structures which we believe is sorely missing in that literature. Our final goal is to leverage the data sizes characteristic of micro-macro situations in order to focus our inference not only on the regression part of the model, but also the distributions of "individual-level" covariates within each group.

To achieve these goals, we propose **FRODO** (Functional Regression On Densities of Observations), a method which unifies density estimation and functional regression in a joint empirical Bayesian model. Although the core idea of FRODO is a fairly straightforward combination of well-established methods in principle, it allows for a remarkable degree of generality in data structures, and its design proves to be far from trivial.

Before describing FRODO, we first give an overview of necessary functional data analysis concepts in Section 5.2. We then give a general overview of the FRODO model and its assumed data structure in Section 5.3, followed by a detailed description of its prior and likelihood components, as well as its practical implementation. In Sections 5.4 and 5.5, we show several simulation studies which demonstrate the potential generality of FRODO in both the regression and covariate observation parts of a micro-macro model.

## 5.2 A brief review of key functional data analysis concepts

Broadly speaking, *functional data analysis* (FDA) is a field of statistics in which the fundamental units of interest are (almost everywhere) smooth functions. A detailed overview of the field is beyond the scope of this chapter, but the interested reader may find one in the excellent book by Ramsay and Siverman [232]. Here we discuss only the concepts necessary to establish notation and motivation for FRODO.

### 5.2.1 Scalar-on-function functional regression

As the name implies, scalar-on-function regression concerns the modelling of a real-valued univariate (or "scalar") response variable with predictors that are functions [232, Section 12.3]. This is achieved by using integrals in place of the sums which define scalar regression models. For example, consider a simple case in which our data are pairs $\{Y_i, f_i^*\}$, $i = 1, \ldots, N$, where $Y_i$ is a real-valued (continuous) scalar response and $f_i^*$ is an almost everywhere continuous function on $[0, 1]$. For this data, a *functional linear model* would be of the form

$$Y_i = \alpha + \int_0^1 \beta^*(x) f_i^*(x) \mathrm{d}x + \epsilon_i, \tag{5.3}$$

with i.i.d. errors $\epsilon_i \sim \mathcal{N}(0, \sigma_Y)$. The *coefficient function* $\beta^*$ weighting the integral is analogous to regression coefficients in a fully scalar regression model.

### 5.2.2 Basis function expansions

Because function spaces are infinite-dimensional, a core component of FDA is the representation of functions of interest in finite-dimensional spaces [232]. Typically, this is achieved by modelling functions as linear combinations of finitely many *basis functions* [232, Section 3.3]. Throughout this chapter, we will use $f^*$ to denote a function of interest, and remove the asterisk to denote a relevant basis function approximation $f$.

Several types of functional bases exist, many of which were described in Sections 4.4–4.6. Attention here is restricted to splines, and in particular the P-splines of Eilers and Marx [73] mentioned briefly in Section 4.6. Recall from Section 4.4 that a basis function approximation of a function $f^*$ on a compact interval $[a, b]$ has the form

$$f(x) = \sum_{k=1}^{K} c_k B_k(x). \tag{5.4}$$

Here, the basis functions $B_k$ are splines: piecewise polynomials with supports defined by a set of equally-spaced "knots" in $[a, b]$. More detailed explanations of splines can be found in [232], Eilers and Marx [73], and Section 4.4.3 and the references therein. As discussed in Section 4.6, Eilers and Marx [73] used penalized likelihood optimization to fit the coefficients

$c = (c_1, \ldots, c_K)$, introducing a penalty based on *finite differences* between coefficients. Their penalty defines the notion of *P-splines* and is of the form

$$\lambda \sum_{k=r+1}^{K} \left[ (\Delta^r c)_{k-r} \right]^2, \tag{5.5}$$

for a positive integer $r$, where $\Delta^r$ denotes the $r^{\text{th}}$-order finite difference operator and $(\Delta^r c)_{k-r}$ denotes the $(k-r)^{\text{th}}$ element of the $(K-r)$-dimensional vector $(\Delta^r c)$. For instance,

$$\begin{aligned}
\left(\Delta^1 c\right)_1 &= c_2 - c_1, \\
\left(\Delta^2 c\right)_1 &= c_3 - 2c_2 + c_1, \text{ and} \\
\left(\Delta^3 c\right)_1 &= c_4 - 3c_3 + 3c_2 - c_1.
\end{aligned}$$

When the *smoothing parameter* $\lambda > 0$ is large, (5.5) dominates the penalized likelihood. Eilers and Marx noted that the sum in this penalty is a good approximation to the $r^{\text{th}}$ derivative of $f$ when the knots defining the spline basis are equally spaced, especially for large dimensionality $K$. Thus, for large $\lambda$ the estimated $f$ is forced to take the approximate shape of a polynomial of degree $r - 1$.

Lang and Brezger [169] devised a Bayesian version of P-splines, based on the notion that a penalized likelihood function is analogous to a posterior distribution on the log scale, with the penalty term assuming the role of the prior (see Section 4.6). The penalty (5.5) is the log density of an $r^{\text{th}}$-order Gaussian *random walk*:

$$(\Delta^r c)_{k-r} \sim \mathcal{N}\left(0, \frac{1}{\sqrt{2\lambda}}\right) \tag{5.6}$$

for $k = r, r+1, \ldots, K$. Lang and Brezger [169] gave the first $r$ components of $c$ (which we call "*free parameters*" in contrast with the last $K - r$ components, whose behaviour is restricted by (5.6)) flat priors. However, we adopt the philosophy that such priors are unreasonable because they give equal weight to all values, no matter how extreme [e.g. the case study of 17], and we have also found such priors to result in extremely poor MCMC sampling behaviour in our models. Our priors on the free parameters in the various P-spline components of FRODO are described in Sections 5.3.2–5.3.3.

As noted by Eilers and Marx [73] (see also Section 4.6.1), one can use P-splines to model a density $f^*$ by replacing $f$ with $\log f$ in (5.4). The imposition of a polynomial shape on $\log f$ then leads to a density estimate which is close to the exponentiation of the corresponding polynomial. For instance, recall from Section 4.6 that using a penalty of order $r = 3$ (in either the frequentist or Bayesian setting) forces $\log f$ towards a quadratic shape, and therefore the resulting density estimate will be similar in shape to a Gaussian.

## 5.3 The FRODO model

### 5.3.1 General overview

Having reviewed the necessary functional data analysis concepts, we are now ready to describe the FRODO approach to micro-macro modelling. Assume the data is organized into $N$ groups, with the $i^{\text{th}}$ group containing $n_i$ individuals. In the simplest case (assumed in the remainder of this section for ease of exposition), data $i^{\text{th}}$ group consists of a group-level response variable $Y_i$, and individual-level observations of a covariate $X$, $(X_{i1}, \ldots, X_{in_i})$. Although we assume real-valued Gaussian $Y_i$'s throughout this chapter for the sake of simplicity, in principle the following methodology could be extended to any response type for which generalized linear modelling is possible. As in Section 5.1, the model is comprised of both a regression part and a covariate observation part, but we assume a much greater level of generality than in (5.1–5.2). Our only assumption for the covariate density part is that, for the $i^{\text{th}}$ group, $X_i := (X_{i1}, \ldots, X_{in_i})$ (where an omitted subscript means the collection of all elements across that subscript) is an i.i.d. sample from an unobserved or "latent" group-specific covariate density $f_i^*$. The regression part of the model defines the "novel" idea at the core of FRODO: the use of these densities (technically, basis expansion estimators thereof) as predictors in a functional linear regression. In mathematical terms, the regression part of the model is

$$
\begin{aligned}
Y_i &= \alpha + \int \beta^*(x) f_i^*(x) \mathrm{d}x + \epsilon_i && (5.7)\\
&= \alpha + \mathbb{E}_i^* \left[\beta^*\left(X\right)\right] + \epsilon_i, && (5.8)\\
\epsilon_i &\overset{i.i.d}{\sim} \mathcal{N}\left(0, \sigma_Y\right),
\end{aligned}
$$

where $\mathbb{E}_i^* \left[\beta^*\left(X\right)\right]$ denotes the expectation of $\beta^*\left(X\right)$ with respect to the density $f_i^*$. The equivalence between (5.7) and (5.8) is the key to FRODO's utility: by simply using densities as predictors in a functional linear regression, the resulting model is essentially a GAM. Thus, FRODO allows for a fully nonparametric approach to both regression functions and covariate structures.

It must be noted that the regressor in (5.8), $\mathbb{E}_i^* \left[\beta^*\left(X\right)\right]$, is the "expectation of the regression function". In general, this is *not* equal to $\beta^*\left(\mathbb{E}_i^* \left[X\right]\right)$ — the "regression on the expectations" — unless $\beta^*$ is linear. Use of the latter is perhaps more "standard" in the measurement error literature, where it is typically assumed that the $X_{ij}$'s within each unit $i$ are noise-corrupted versions of some "true" covariate $\xi_i$ [see 41, or any standard reference on measurement error]. Although it is not always assumed that $\mathbb{E}_i^* \left[X\right] = \xi_i$ (e.g. the general linear error structures described in Section 6.4 of [32], and references therein), typically the target is estimation of $\beta^*\left(\xi_i\right)$, possibly marginalized over an estimate of the "posterior" $f_{\xi|X_i}$ [e.g. 135, 175]. We are not aware of any literature which explicitly uses "expectations of the regression" in the way that FRODO does.

In the next two subsections, we detail the priors and likelihoods comprising FRODO. Recall that we approximate $\beta^*$ and the $f_i^*$'s with basis function expansions, use of which will be denoted without asterisks. In a slight abuse of notation, we consider the model

$$Y_i = \alpha + \int \beta(x) f_i(x) \mathrm{d}x + \epsilon_i \tag{5.9}$$

$$= \alpha + \mathbb{E}_i \left[\beta\left(X\right)\right] + \epsilon_i, \tag{5.10}$$

as a proxy to (5.7–5.8), where $\beta$ and $f_i$ are the basis function approximations to their "true" counterparts ($\beta^*$ and $f_i^*$, respectively), and $\mathbb{E}_i$ denotes expectation w.r.t. $f_i$.

Before exploring the details of FRODO, some final technical and notational points are in order. We recommend standardizing the data so that default prior choices are weakly informative [93, Sections 2.9 and 16.3]. Keeping with our convention of using omitted subscripts to mean the collection of all elements across that subscript, let $Y = (Y_1, \ldots, Y_N)$ and $X = \{X_1, \ldots, X_N\}$, where $X_i$ was defined above. In what follows, we will assume that $Y$ and $X$ have both been standardized to have zero mean and unit variance. Note that for $X$, this standardization is "marginal", meaning that it is done across groups *and* individuals within groups. We will overload notation and use $f_i^*$ and $f_i$ to refer to, respectively, the true density and its basis function approximation for the standardized version of $X_i$. For technical reasons, it is necessary to assume that $\beta$ and the $f_i$'s are all defined on a common compact interval. This will be denoted by $[a, b]$ on the standardized scale, and when it is necessary to speak about the domain of the covariates on the original (unstandardized) scale, it will be denoted by $[a', b']$. Assuming $X$ has been standardized as recommended above, we have $a = \left(a' - \bar{X}\right)/\sigma(X)$ and $b = \left(b' - \bar{X}\right)/\sigma(X)$, and $[a', b']$ can be chosen so that its endpoints are (nearly) equal to the unscaled extrema of the covariates.

### 5.3.2 The density model

For computational convenience — and because it suffices for the ordinal covariates which are common in real micro-macro datasets [e.g. 57, 4, 60] — the $f_i$'s are modelled as histograms. In practical terms, this means that they are linear combinations of constant basis functions (see Section 4.4.1):

$$f_i(x) = \sum_{k=1}^{K} \phi_{ik} \mathbb{1}_{I_k}(x), \tag{5.11}$$

where $I_k$ is the $k^{\text{th}}$ equal-width subinterval $[a + (k-1)h, a + kh)$ of $[a, b]$, $\mathbb{1}_{I_k}$ is the indicator function of $I_k$, and $h = (b - a)/K$ is the bin width. The density coefficients $\phi_{ik}$ are scaled

"softmax" transformations of Gaussian random variables $\theta_{ik}$, $i = 1, \ldots, N$, $k = 1, \ldots, K$:

$$\phi_{ik} = \frac{e^{\theta_{ik}}}{h \sum_{j=1}^{K} e^{\theta_{ij}}}, \tag{5.12}$$

where, for all $i$, $\theta_{i1} \equiv 0$ to ensure identifiability. Equivalently, we may say that the $\phi_i$'s are (up to the scaling factor $h$) logistic normal random vectors [3].

The priors for the $\theta$'s are chosen in order to impose useful constraints on the behaviour of the densities. In particular, for some positive integer $r$ we will impose an $r^{\text{th}}$-order Gaussian random walk prior on $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$ for all $i$. Since the logarithms of the $f_i$'s are also piecewise constant, this structure means that $\log f_i$ is a Bayesian P-spline of degree zero, with $r^{\text{th}}$-order penalty, for all $i$. Recall from Section 5.2.2 that an $r^{\text{th}}$-order random walk prior on $\theta_i$,

$$\left(\Delta^r \theta_i\right)_{k-r} \sim \mathcal{N}\left(0, \tau_i\right), \, k \geq r+1 \tag{5.13}$$

forces $\log f_i$ towards the approximate shape of a $(r-1)^{\text{th}}$-degree polynomial when the smoothing parameter $\tau_i$ is small[1].

Note that (5.13) completely determines the conditional distributions of $\theta_{ik}$ for $k > r$ given $\theta_{ik}$ for $k \leq r$. In the case $r > 1$, it remains to set the priors on the "free parameters" $\theta_{ik}$ for $2 \leq k \leq r$: the "initial values" of the random walk. A seemingly sensible and simple choice would be diffuse, mean-zero, independent Gaussian priors. Unfortunately, this turns out not to be entirely suitable for FRODO. For $r > 1$, imposing fully independent priors on the densities[2] causes bias in the posterior mean coefficient function, $\widehat{\beta}$. For instance, if the true $\beta$ is a linear function, the magnitude of the slope of $\widehat{\beta}$ will be biased downward, just as in the "naive" approach to modelling described in Section 5.1. In the Bayesian hierarchical setting, this "attenuation" problem can be solved by putting priors on the covariates which introduce dependence between them and "pool" each group's measurements towards a latent group-level variable. The solution here is similar.

[1]Henceforth, the phrase "smoothing parameter" will refer to the standard deviation of the random walk prior ($\tau$), instead of its precision as in Section 5.2.2 (where it was denoted by $\lambda = \tau^{-2}/2$).

[2]When discussing the model itself, we will typically write "the densities" to refer to the histograms $f_i$ which are actually part of the model. When it is necessary to invoke the $f_i^*$'s, we will specify them as the "*true* densities".

To expand on this, first note that with $\theta_{i1} \equiv 0$ for all $i$, we have

$$\theta_{ik} = \log\left(h^{-1}\int_{a+(k-1)h}^{a+kh} f_i(x)\mathrm{d}x\right) - \log\left(h^{-1}\int_{a}^{a+h} f_i(x)\mathrm{d}x\right) \tag{5.14}$$

$$\approx \log f_i^*\left(a + h\left(k - \frac{1}{2}\right)\right) - \log f_i^*\left(a + \frac{h}{2}\right), \tag{5.15}$$

recalling that $f_i^*$ is the "true" density for group $i$.

Suppose $f_i^*$ is that of a $\mathcal{N}(\xi_i, \sigma_i)$ random variable[3]. This corresponds to the limiting case for $r = 3$ as $\tau_i \to 0$, and it can be shown that (5.15) in this case reduces to

$$\theta_{ik} \approx \frac{h(k-1)}{\sigma_i^2}\left(\xi_i - \left(a + \frac{kh}{2}\right)\right) \tag{5.16}$$

For $r = 3$, this approximation motivates our choice of priors for the "free parameters". For each $i$ and $k = 2, 3$, we take them to be Gaussian with mean given by the right side of (5.16) and standard deviation $\tau_i$. Thus, $\tau_i$ controls $f_i$'s adherence to the limiting Gaussian shape in two respects: by controlling the free parameters' deviations from their means, and by scaling the random walk behaviour in (5.13).

We now set priors on $\xi_i$ and $\sigma_i$. When the true covariate densities are Gaussian, the structure of the data is analogous to that of the "classical" micro-macro model, with $\xi_i$ being a "latent group-level covariate" and $\sigma_i$ controlling the level of Gaussian noise for each group's individual-level covariate measurements. In keeping with natural choices for that setting, we first assign the $\xi_i$'s a $\mathcal{N}(\mu_\xi, \sigma_\xi)$ prior. Recalling that $[a', b']$ denotes the assumed domain of the covariate densities on the original (unstandardized) scale, the mean $\mu_\xi$ is given a $\mathcal{N}\left((a'b - b'a)/(a' - b'), 15/K^2\right)$ hyperprior. This corresponds to a mean-zero hyperprior on the original covariate scale, with the empirically-determined standard deviation $15/K^2$ accounting for the discretization error from approximation (5.15). The scale $\sigma_\xi$ is given a standard half-normal prior, which will be fairly uninformative if the $X_{ij}$'s have been scaled to have unit marginal variance. It will often be reasonable that the covariate densities are homoscedastic: $\sigma_i \equiv \sigma_X$ for all $i$. A standard half-normal prior is a sensible choice in this case. If one wishes to explicitly model heterogeneity, then each $\sigma_i$ can be given its own half-normal prior, perhaps sharing a common scale parameter with its own hyperprior.

Now, suppose $f_i^*$ is instead a (shifted) Exponential$(\lambda_i)$ density. This corresponds to the limiting case for the random walk with $r = 2$, and here (5.15) reduces to

$$\theta_{ik} = -\lambda_i (k-1) h. \tag{5.17}$$

[3]Assuming the covariates have been standardized as recommended in Section 5.3.1, most of $f_i^*$'s mass presumably lies in $[a, b]$, and $a < \xi_i < b$.

Note that for an exponential density, there is no discretization error, so (5.14) and (5.15) are equal. Thus, analogously to the $r = 3$ case described above, when $r = 2$ we assume the "free parameters" $\theta_{i2}$ are Gaussian with mean given by the right side of (5.17) and standard deviation $\tau_i$. A natural choice of prior for the "latent rates" $\lambda_i$ is Gamma $(\alpha_\lambda, \alpha_\lambda/\mu_\lambda)$. The mean $\mu_\lambda$ is given a standard half-normal prior (which should be only weakly informative if the covariates have been standardized), while the shape parameter $\alpha_\lambda$ is given a more diffuse half-normal prior with scale 10. Note that this parameterization of the Gamma in terms of shape and mean, rather than the more conventional shape and rate, proved computationally advantageous.

By defining the "free parameters" in terms of latent group-level variables with their own hyperpriors, we introduce the necessary dependence and "pooling" to prevent bias in the regression part of the model, just as one might do in the scalar case. For any order $r$, the density model is completed with priors on the smoothing parameters $\tau_i$, which we take to be exponentials with rates $\delta_i^{-1}$. The scales are assumed to be fixed data, chosen empirically based on heuristics and the properties of the $X_{ij}$'s in the absence of more meaningful prior information. Such choices place FRODO in the category of "empirical Bayesian" methods, but we have found that sampling behaviour and posterior results can become poor when the $\delta_i$'s are not chosen carefully. If group sizes are moderate ($n_i$'s roughly between 20 and 60) and one doesn't expect any of the covariate densities to deviate too seriously from the shape implied by the $r^{\text{th}}$-order random walk prior, $\delta_i = 0.1$ for all $i$ seems to be a good default choice based on preliminary empirical results. Smaller groups tend to require smaller $\delta_i$'s, and it may also be advantageous to shrink them when the basis dimension $K$ is very large, especially relative to the $n_i$'s.

Finally note that, because the densities are piecewise constant, the likelihood $X_i \sim f_i$ is equivalent to $m_i := (m_{i1}, \ldots, m_{iK}) \sim$ Multinomial $(n_i, \phi_i)$, where $m_{ik}$ is the bin count $|\{j : X_{ij} \in I_k\}|$. In summary, the model for the densities, assuming an $r^{\text{th}}$-order random

walk prior structure (for $r \leq 3$), is

$$
\begin{aligned}
m_i &\sim \text{Multinomial}\left(n_i, \phi_i\right) \\
\phi_{ik} &= \frac{e^{\theta_{ik}}}{h \sum_{j=1}^{K} e^{\theta_{ij}}} \\
\theta_{i1} &\equiv 0
\end{aligned}
$$

$$
\left.
\begin{aligned}
\theta_{i2} &\sim \mathcal{N}\left(-\lambda_i h, \tau_i\right) \\
\lambda_i &\sim \text{Gamma}\left(\alpha_\lambda, \frac{\alpha_\lambda}{\mu_\lambda}\right) \\
\alpha_\lambda &\sim \text{Half-Normal}(0, 10) \\
\mu_\lambda &\sim \text{Half-Normal}(0, 1)
\end{aligned}
\right\} \quad r = 2
$$

$$
\left.
\begin{aligned}
\theta_{ik} &\sim \mathcal{N}\left(\frac{h(k-1)}{\sigma_i^2}\left(\xi_i - \left(a + \frac{kh}{2}\right)\right), \tau_i\right) \;\; (k = 2, 3) \\
\xi_i &\sim \mathcal{N}\left(\mu_\xi, \sigma_\xi\right) \\
\mu_\xi &\sim \mathcal{N}\left(\frac{a'b - b'a}{a' - b'}, \frac{15}{K^2}\right) \\
\sigma_\xi &\sim \text{Half-Normal}(0, 1) \\
\sigma_X &\sim \text{Half-Normal}(0, 1)
\end{aligned}
\right\} \quad r = 3
$$

$$
\begin{aligned}
\left(\Delta^r \theta_i\right)_{k-r} &\sim \mathcal{N}\left(0, \tau_i\right), \; k > r \\
\tau_i &\sim \text{Exp}(\delta_i^{-1})
\end{aligned}
$$

### 5.3.3 The regression model

Here we detail priors for the regression part of FRODO, the likelihood for which is defined by (5.9–5.10). Recall that we have restricted our attention in this chapter to continuous real-valued responses $Y_i$ with i.i.d. errors $\epsilon_i \sim \mathcal{N}\left(0, \sigma_Y\right)$, The following priors on $\alpha$ and $\beta$ would require only minor changes to accommodate more general response types (e.g. different scaling may be in order to ensure plausible effect sizes in a logistic regression; see Section 16.3 of Gelman et al. [93]), and the prior on the dispersion parameter could easily be changed as necessary.

The error scale $\sigma_Y$ is given a half-T prior with 4 degrees of freedom and scale $1/\sqrt{2}$, so that $\sigma_Y$ has a prior mean of $1/\sqrt{2}$. Recalling the assumption from Section 5.3.1 that $Y$ has been standardized to have unit variance, this scale (in informal terms) loosely corresponds to a prior expectation that roughly half of the variation in the response values is due to regression error (assuming that the errors and regressors are independent, which we do

here). This seems to be a sensible approach for a "default" prior, unless one has prior domain knowledge which would allow for context-specific prior beliefs about the regression error.

Both $\alpha$ and $\beta$ are given hierarchical priors with scales proportional to $\sigma_Y$. This can be shown to ensure unimodality in some penalized Bayesian regression models [222], and we also found that it improved sampling behaviour. The intercept $\alpha$ is given a diffuse $\mathcal{N}(0, 20\sigma_Y)$ prior.

We take the coefficient function $\beta$ to be piecewise constant, with the same dimensionality $K$ as the densities. This is quite computationally convenient, as the integral in (5.9) then reduces to the inner product between the coefficients of $\beta$ and $f_i$, scaled by the bin width $h$. Because the functional predictors all have unit integral, adding a constant shift to $\beta$ does not change the model: for any $c \in \mathbb{R}$, the model is identical if $\beta$ and $\alpha$ are replaced by $\beta + c$ and $\alpha - c$, respectively. Thus, we impose the identifiability constraint $\mathbb{E}[\beta(X)] := \int_a^b \widehat{f}_{\text{Cent}}(x)\beta(x)\mathrm{d}x = 0$, where $\widehat{f}_{\text{Cent}}$ is the *empirical central density*:

$$\widehat{f}_{\text{Cent}}(x) := \sum_{k=1}^{K} \frac{\sum_{i=1}^{N} m_{ik}}{\sum_{l=1}^{K}\sum_{i=1}^{N} m_{il}} \mathbb{1}_{I_k}(x). \tag{5.18}$$

Essentially, $\widehat{f}_{\text{Cent}}$ is the "marginal histogram" of all covariate data across groups. Presumably, the total number of covariate observations $\sum_i n_i$ will be large enough in most data sets to ensure that $\widehat{f}_{\text{Cent}}$ is reasonably "smooth", so that it is a good approximation to the "marginal" covariate density (i.e. marginalized across groups) for large $K$. Note that we use the *empirical* central density mainly for computational convenience: an "inferred central density" like $N^{-1}\sum_i f_i$ would certainly be "smoother", but this would add needless complexity to the gradients used in NUTS when the empirical version is sufficient to ensure identifiability.

This constraint amounts to centering the inferred regressors $\mathbb{E}_i[\beta(X)]$. In practice, the constraint is achieved by defining a piecewise constant function

$$\beta^0(x) := \sum_{k=1}^{K} \beta_k^0 \mathbb{1}_{I_k}(x) \tag{5.19}$$

and taking $\beta = \beta^0 - \int \widehat{f}_{\text{Cent}}\beta^0$. In keeping with Bayesian functional regression approaches such as [56], we put a second-order random walk prior on the coefficients of $\beta^0$, with the first coefficient set to 0 for identifiability:

$$\begin{aligned} \beta_1^0 &\equiv 0, \\ \beta_2^0 &\sim \mathcal{N}(0, 20h\sigma_Y), \\ \left(\Delta^2\beta^0\right)_{k-2} &\sim \mathcal{N}(0, \tau_\beta\sigma_Y). \end{aligned}$$

The smoothing parameter $\tau_\beta$ controls the extent to which $\beta$ deviates from the random-walk behaviour. As $\tau_\beta \to 0$, $\beta$ is forced towards a stepwise approximation to a straight line, and the regression model (5.9) is therefore forced towards a linear regression. In this limiting case, the "slope" of $\beta$, $h^{-1}\beta_2^0$, is equivalent to the regression coefficient in a scalar linear model. Thus, using a scale factor of $20\sigma_Y h$ in $\beta_2^0$'s prior can be considered roughly analogous to placing a $\mathcal{N}(0, 20\sigma_Y)$ prior on the coefficient in the scalar case, which should be reasonably diffuse if the covariates have been scaled as recommended above [e.g 272, Section 25.12 of User's Guide]. Finally, $\tau_\beta$ is given an exponential prior with rate 2 (equivalently, scale 0.5). In contrast to the smoothing parameters for the densities, we found that $\tau_\beta$ did not require a careful selection of prior scale in order to ensure good model performance.

### 5.3.4 Implementation

The FRODO model is implemented in the Stan programming language [40], which provides exceptional power, flexibility, and efficiency through its use of the No-U-Turns Sampling (NUTS) variant of Hamiltonian Monte Carlo [132]. For each of the below simulation studies, four parallel chains were run with fairly diffuse starting values, with sufficiently many sampling iterations to ensure effective sample sizes of at least 450 for all parameters [see 93, Section 11.5]. All model runs were devoid of divergent transitions [272], and the overwhelming majority of parameters in all simulations had $\hat{R}$ values (where $\hat{R}$ is a diagnostic which helps to assess model convergence, see Vehtari et al. [287]) below 1.01, with only a single parameter in each of the models of Sections 5.4.2 and 5.4.5 having a value very slightly above this threshold. All of the simulation studies below were conducted using R [230], interfacing with Stan via the RStan package [273]. More details are given in Appendix B.

## 5.4 Simulation studies

As discussed in previous sections, FRODO is uniquely powerful in theory because it is "doubly nonparametric": it can capture arbitrary unknown structures in both the covariate densities and the regression model. In the following subsections, we put this to the test with a wide variety of simulated datasets. We will assess FRODO's ability to harness location, scale, and shape information from covariate densities and use it to recover true regression relationships. In each study, FRODO will be compared to two simpler models:

1. a "naive" scalar regression model using only the sample means of the covariate measurements (or of some suitable transformation thereof, where applicable); and

2. a "hierarchical" scalar regression model, where the form of the regression function and covariate distributions are assumed known, with only the actual parameter values unknown.

More detail will be provided in the following subsections.

Because FRODO does not assume any parametric form for either the regression or covariate parts of the model, all that is required are choices of an appropriate random walk order $r$, dimensionality $K$, (unstandardized) density domain $[a', b']$, and set of density scaling factors $\delta = (\delta_1, \ldots, \delta_N)$. These choices must be made assuming that the true data-generating mechanisms are not known *a priori*. One could use subject-specific domain knowledge if it is available. Otherwise, an "empirical Bayesian" approach based on informal inspections of the data is acceptable, and this is the approach we will use for all simulation studies in this chapter. Visual inspection of default histograms or KDE's suffices to this end. From a strictly Bayesian perspective on inference, one could argue that this data dependence in the prior is not philosophically sound. However, an empirical Bayesian approach to nonparametric modelling is certainly not without precedent, as discussed at the beginning of Section 4.4 [see also 248, 281]. Serra and Krivobokova [261] devised an empirical Bayesian method for determining both the smoothing parameter and penalty order in spline fitting; our strategy could be viewed as a crude, heuristic approximation of such a method.

### 5.4.1 Gaussian covariate densities, linear regression model

We begin with the "classical" structure from Section 5.1, where the individual-level measurements within groups are Gaussian deviations from a latent group-level covariate, itself Gaussian:

$$\xi_i \sim \mathcal{N}(0, \sigma_\xi), \tag{5.20}$$

$$X_{ij} \sim \mathcal{N}(\xi_i, \sigma_X). \tag{5.21}$$

The regression model is also linear:

$$\begin{aligned} Y_i &= \alpha + \tilde{\beta}\xi_i + \epsilon_i, \\ &= \alpha + \mathbb{E}_i\left[\tilde{\beta}X\right] + \epsilon_i, \end{aligned} \tag{5.22}$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma_Y). \tag{5.23}$$

Note that the second line explicitly restates the regression model in the form of (5.8), with the true regression function $\beta^*(x) = \tilde{\beta}x$ being a line with slope $\tilde{\beta}$. Some clarification on notation is in order here. Throughout Sections 5.4–5.5, $\tilde{\beta} \in \mathbb{R}$ will denote a scalar which determines the magnitude and sign of the true regression function $\beta^*$. In turn, recall that the (piecewise constant) basis function approximation to $\beta^*$ is denoted as $\beta$.

The true parameter values *before* standardizing[4] the data as described in Section 5.3.2 are $\sigma_\xi = 2$, $\sigma_X = 3$, $\alpha = 0.3$, $\tilde{\beta} = 0.4$, and $\sigma_Y = 0.5$. The result is a dataset with moderate amounts of noise in both the regression and the covariate measurements. The number of groups is $N = 275$ and each group contains covariate samples for $n = 20$ individuals.

Upon inspecting the data as recommended in the introduction to this section (not shown), we find that an assumption of roughly Gaussian density shape (corresponding to $r = 3$) is reasonable for these data. Because the densities are moderately wide but relatively close together (as the between-density variability $\sigma_\xi$, is somewhat smaller than the within-density variability, $\sigma_X$), a modest basis of size $K = 10$ should suffice without substantial loss of information. For this simulated data we have $\min_{i,j} X_{ij} = -13.54922$ and $\max_{i,j} X_{ij} = 10.87845$, so we extend this range slightly by the same amount in each direction to arrive at an assumed density domain[5] of $[a', b'] = [-13.67077, 11]$. Finally, the default choice of $\delta_i = 0.1$ for all $i$ recommended in Section 5.3.2 is used here.

As stated at the beginning of this section, we compare FRODO to two simpler models. The first is simply a standard Bayesian linear regression, with (5.20) omitted and the group-level sample covariate means $\bar{X}_i$ treated as the "true" covariates. The second is a scalar micro-macro Bayesian regression, implemented in the "obvious" way: namely, (5.20)–(5.23) are assumed to be the known form of the model, with all parameters (including the latent $\xi_i$'s) unknown and inferred. Recall that the estimate of $\tilde{\beta}$ from the "naive" model will be smaller in magnitude than the "true" value, which the hierarchical scalar model will presumably recover more effectively.

Figure 5.1 shows results for the regression part of the model. In the left plot, the piecewise-constant estimator of the regression function from FRODO is shown with its pointwise (P.W.) 95% credible interval (C.I.). Superimposed on the plot are the true regression function, as well as the posterior means from the hierarchical and naive scalar models (both of which assume a known linear form for the regression unlike FRODO, which only controls adherence to a linear regression through $\tau_\beta$). Because the within-group variability is not too much larger than the across-group variability and the sample sizes are reasonable, only a small amount of attenuation is caused by using the naive model, so the estimated regression functions for both scalar models are entirely within the pointwise C.I. from FRODO. However, the "slope" of the mean regression function from FRODO seems to be closer to those of the true function and the hierarchical scalar estimate, rather than that of the naive estimate. We can formalize this observation by considering the secant line to the FRODO regression function which intersects it at the midpoints of the first and

[4]Throughout this section, all parameter values and results will be presented on the original (unstandardized) scale of the given data. The standardization only occurs "internally", during the fitting of the FRODO model.

[5]Henceforth, the "assumed domain" will be stated on the unstandardized scale of the original data (i.e. $[a', b']$), with the standardization to $[a, b]$ left unstated.

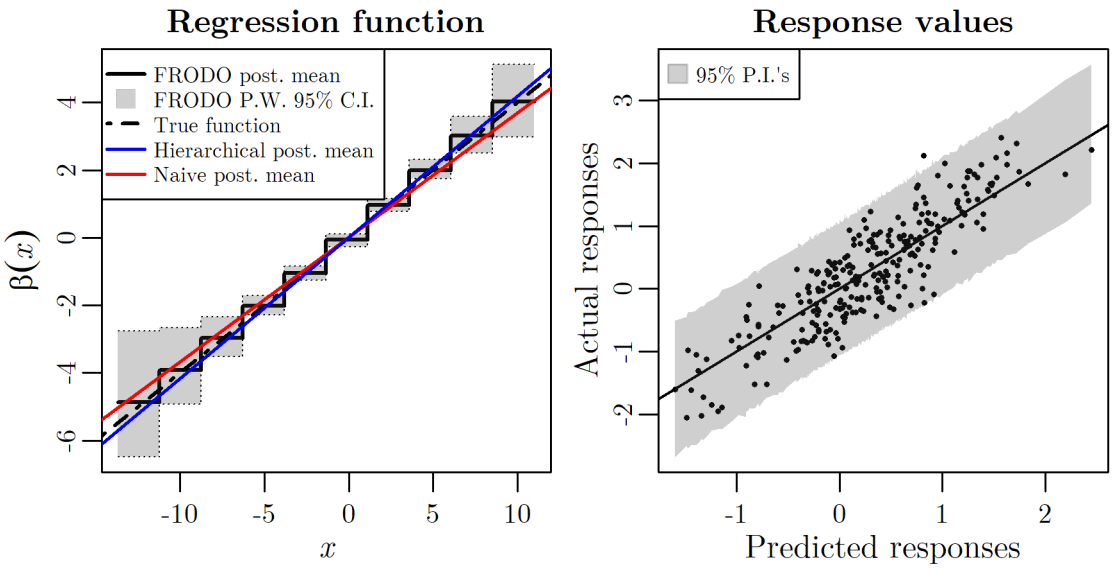


Figure 5.1: Results of FRODO applied to data with Gaussian covariates and a linear regression structure. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, and posterior mean estimates from hierarchical and naive scalar models. Right: responses $\hat{Y}_i$ predicted by FRODO (along with 95% prediction intervals) vs. true responses.

last bins. The slope of this line (which is roughly analogous to a notion of "slope" for the FRODO regression function) is 0.4002, whereas the slopes of the true, hierarchical scalar, and naive scalar regressions are 0.4, 0.4172, and 0.3678, respectively.

Another way to assess FRODO's ability to infer the "true" regression (rather than the incorrect one implied by the naive model) is by checking the posterior for the regression error scale, $\sigma_Y$. Because of the additional noise in the individual-level covariate measurements, the naive model's estimate for $\sigma_Y$ will be biased upward [e.g. 41, Section 3.2.1]. Indeed, the posterior mean for this parameter from the naive scalar model is 0.5559 (95% C.I. $(0.5104, 0.6015)$), while the posterior means from FRODO and the hierarchical scalar model are 0.4944 (95% C.I. $(0.4417, 0.5505)$) and 0.4901 (95% C.I. $(0.4363, 0.5494)$), respectively. Because the FRODO estimate is much closer to the true value of 0.5 than it is to the "naive estimate", we are satisfied that we have avoided the attenuation problem inherent in the naive model. Table B.2 contains summaries of the $\sigma_Y$ posteriors for every simulation study in this chapter.

On the right of Figure 5.1, we have plotted the posterior mean predicted responses $\hat{Y}_i$ against the observed responses. The shaded region is a visual representation of 95% posterior prediction intervals (P.I.'s) for each group.

Figure 5.2 shows the estimated $f_i$'s, along with their pointwise 95% C.I.'s, for the group with the smallest (left) and largest (right) $\xi_i$'s, as well as the group whose $\xi_i$ is closest to the sample mean (middle). The middle and right fits are satisfactory, with the inference

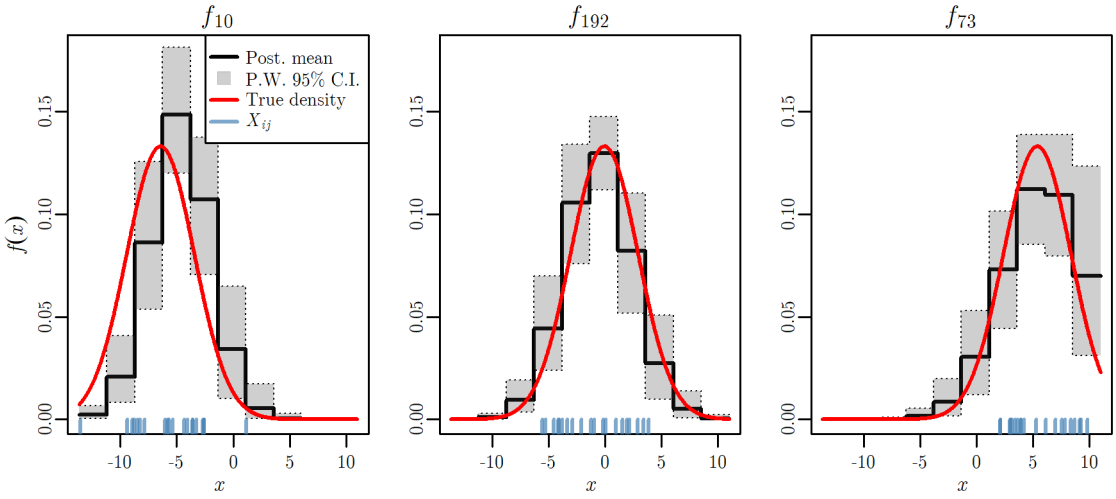


Figure 5.2: For a selection of groups (from the data with Gaussian covariates and a linear regression structure), the FRODO estimate of the group-specific covariate density, alongside its pointwise 95% credible regions. The true densities are superimposed as red lines, and the actual covariate samples are shown as rug plots.

effectively capturing the true covariate densities (shown in red). The left plot shows that there is something of a mismatch between the inferred and true densities for the group with the lowest $\xi_i$, with the former shifted slightly too far to the right. Given that the model appears to perform well in all other respects, this is not a significant concern, especially since the rug plot suggests consistency with the data. We did not observe this problem in other datasets generated with the same parameter values (not shown), and therefore assume it is simply an unfortunate quirk of this particular data.

### 5.4.2 Gaussian covariate densities, nonlinear regression model

Here, we test FRODO's ability to handle nonlinear regression functions. The covariates adhere to the same Gaussian structure as in Section 5.4.1, but the regression model is now quadratic:

$$\begin{aligned} Y_i &= \alpha + \tilde{\beta}\left(\xi_i^2 + \sigma_X^2\right) + \epsilon_i \\ &= \alpha + \mathbb{E}_i\left[\tilde{\beta}X^2\right] + \epsilon_i. \end{aligned}$$

Because the true covariate densities all have common variance $\sigma_X^2$, the difference between $\mathbb{E}_i\left[X^2\right]$ and $\left(\mathbb{E}_i\left[X\right]\right)^2$ is constant and can therefore be absorbed into the intercept. Here, the regression function is $\beta^*(x) = \tilde{\beta}x^2$.

The same parameter values $\left(\sigma_\xi, \sigma_X, \alpha, \tilde{\beta}, \sigma_Y\right) = (2, 3, 0.3, 0.4, 0.5)$ and number of groups $N = 275$ are used as in Section 5.4.1 although the data is not strictly the same as we used a different seed for pseudorandom number generation in this study. Because the values of

$\xi^2$ span a wider interval than those of $\xi$, the "relative" level of regression error is lower than in Section 5.4.1, since the "signal" is larger in scale than the "noise".

As before, we compare FRODO to two scalar models, one hierarchical and one naive. Here, however, it is assumed known in the hierarchical model that the regression is quadratic in the latent covariates $\xi$, with no linear term. The naive scalar model here is a GAM rather than a linear model, with the covariates taken to be the group-level sample means and the unknown regression function modelled as a cubic P-spline with second-order penalty.

Because the regression function is not one-to-one, an interesting difficulty arises in this framework when the group sizes $n_i$ are too small. On the regression side, the distributions are unchanged if $\xi_i$ is replaced with $-\xi_i$ in a given group. When $n_i$ is small and the true $\xi_i$ is close to zero, the available measurements $X_i$ may not be informative enough to distinguish between these possibilities[6]. This creates multimodality in the posterior (for the hierarchical scalar model, and for FRODO to a somewhat lesser extent) with all of its associated difficulties, including poor HMC sampling behaviour and posterior mean estimates that are not particularly meaningful. Thus, larger group sizes are required if one wants meaningful inference on the covariate parameters as well as the regression parameters. Here, we increase the group size in the simulated data from the $n = 20$ used in Section 5.4.1 to $n = 50$ for all $i$. The $X_{ij}$'s range from -13.76074 to 14.0043, and we expand this range by a small amount in each direction for an assumed density domain of $[-13.80644, 14.05]$. As before, we find $K = 10$ and $\delta_i = 0.1\ \forall i$ to be suitable choices here.

Results for the regression part of the model are shown in Figure 5.3. At first glance, it may appear as though the FRODO estimate of the regression function is too attenuated, as it is closer to the estimate from the naive scalar model at the endpoints than it is to the true function and the hierarchical scalar estimate. Note, however, that over 95% of the $X_{ij}$'s lie within the middle six bins, and over 95% of the true latent $\xi_i$'s within the middle four. In those regions, the FRODO estimate is quite close to the true quadratic regression function. Towards the endpoints where the $X_{ij}$'s are very sparse, there is much less information with which to estimate value of the regression function. This edge effect is readily seen in several examples in this manuscript by observing that the pointwise C.I.'s for $\beta$ are wider in regions with few covariate estimates. In the linear example of Section 5.4.1, this did not create noticeable bias in the actual posterior mean for $\beta$ near the endpoints. Presumably this is because — in somewhat informal terms — the covariates in the middle of the domain were sufficiently informative to constrain $\beta$ to a linear shape there with high posterior probability, which results in the smoothing parameter $\tau_\beta$ being small with high probability, which, in turn, enforces a linear shape in $\beta$ with fairly high probability throughout the rest of the domain. In this example, we do not penalize $\beta$ towards a quadratic shape — only away from

[6]This appears to also depend on the amount of covariate variability within the group relative to the size of its regression error, although it is not currently clear exactly how this dependence works.

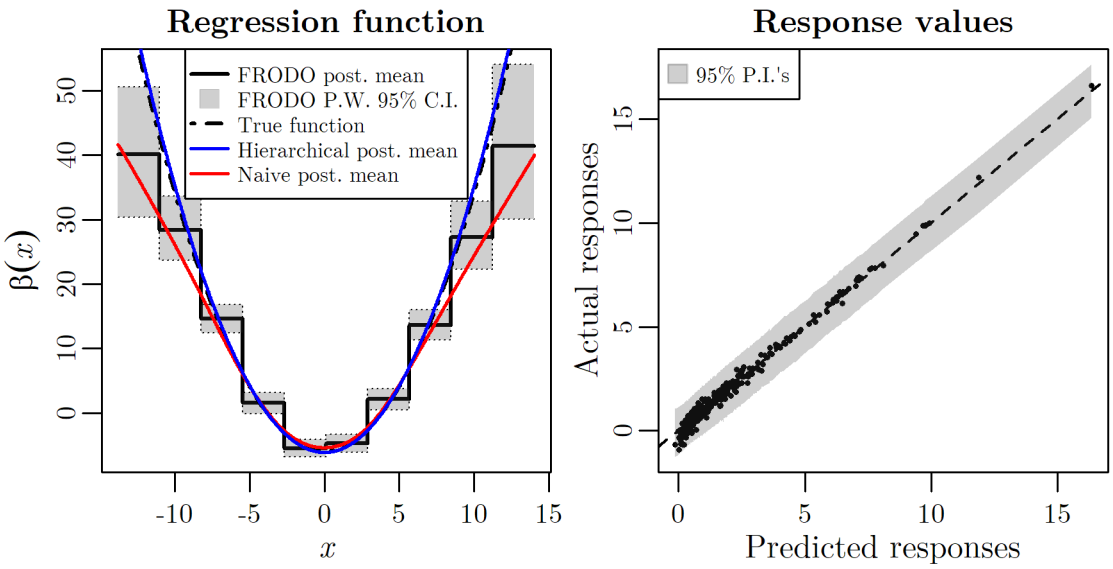


Figure 5.3: Results of FRODO applied to data with Gaussian covariates and a quadratic regression structure. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, and posterior mean estimates from hierarchical and naive scalar models. Right: responses $\hat{Y}_i$ predicted by FRODO (along with 95% prediction intervals) vs. the true response values.

a linear shape. As such, it is not surprising that the posterior for $\beta$ is biased away from the truth near the endpoints, as neither the prior nor the likelihood are very informative there. In principle, one could specify a third-order random walk prior for $\beta$ in order to ensure a more genuinely quadratic shape, provided one had sufficient reason *a priori* to assume this was an appropriate choice. However, we argue that the second-order random walk prior used here is more intuitive, as it is formulated in terms of deviations from a linear model. At any rate, the heightened bias and uncertainty in the FRODO regression function near the endpoints does not create any seriously adverse consequences for the rest of the inference. In particular, the FRODO posterior mean for $\sigma_Y$ is 0.4715 (95% C.I. $(0.3848, 0.6662)$), much closer to the true value of 0.5 than the estimate from the naive model (0.8848, 95% C.I. $(0.8150, 0.9620)$), suggesting that FRODO is successfully recovering the true regression model and not the biased naive version. The plot of estimate vs. true responses on the right of Figure 5.3 shows an overall good fit, although there is a small amount of bias in the estimates of the lowest responses.

Figure 5.4 shows a sample of covariate densities, once again for the group with the smallest and largest $\xi_i$'s, and the $\xi_i$ closest to the sample mean. With larger group sizes, FRODO successfully approximates the true densities for each group shown here.

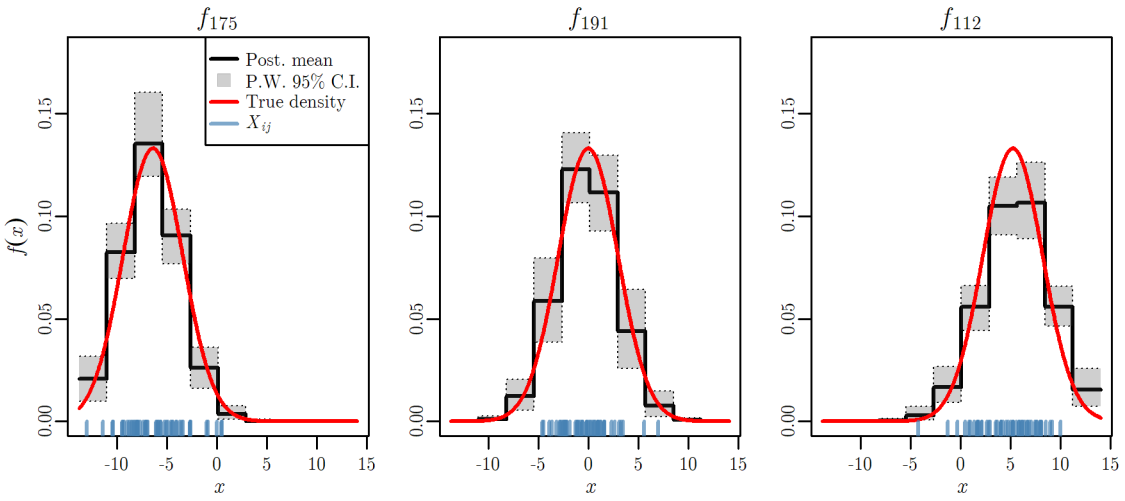


Figure 5.4: For a selection of groups (from the data with Gaussian covariates and a quadratic regression structure), the FRODO estimate of the group-specific covariate density, alongside its pointwise 95% credible region. The true densities are superimposed as red lines, and the actual covariate samples are shown as rug plots.

### 5.4.3 Exponential covariate densities, linear regression model

Although it is useful to model arbitrary regression functions, doing so with Gaussian covariate distributions is a capability shared by many methods. In fact, authors such as Sarkar et al. [255] have developed Bayesian methods which allow for even more general structures of the form $X_{ij} = \xi_i + \nu_{ij}$. The true advantage of FRODO lies in its ability to handle covariates that are not based on any kind of *additive* error structure. To demonstrate this, here we use an exponential covariate structure:

$$\begin{aligned} \lambda_i &\sim \text{Gamma}\left(10, 10\right) \\ X_{ij} &\sim \text{Exponential}\left(\lambda_i\right); \end{aligned}$$

and a linear regression model

$$\begin{aligned} Y_i &= \alpha + \tilde{\beta}\lambda_i^{-1} + \epsilon_i \\ &= \alpha + \mathbb{E}_i\left[\tilde{\beta}X\right] + \epsilon_i, \end{aligned}$$

where we have not restated the distribution for the error variance since it is identical to (5.23) for all subsequent studies.

It is worth contrasting this framework with that of Section 5.4.1. There, the true covariate densities were Gaussians with equal variances, so the group-level responses depended on their *locations*. With exponential covariate distributions, the linear regression model implies responses that instead vary with the *scales* of the densities. This turns out to be a somewhat

challenging type of model for FRODO, due to its treatment of $\beta$ and the $f_i$'s as piecewise constant functions on bins of equal width. When the true densities are exponential, for any group $i$ it is highly probable that most of the $X_{ij}$'s will be near 0, with a few very large measurements in the groups with small rates $\lambda_i$. If the dimensionality (equivalently, the number of bins) $K$ is taken too small, then the groups with large rates will all have estimated $f_i$'s with probability mass near one in the first bin, and mass near zero in the rest. Thus, it is necessary to use a fairly large $K$ in order to capture the differences between these densities. However, this introduces an opposing challenge due to the sparsity of large $X_{ij}$'s: near the right end of the domain, many of the bins will not contain any covariate measurements, so there is little information with which to estimate the densities — and therefore, the regression function — in that region. In summary, when the density scales differ to this extent, the "resolution" of the data varies throughout the domain.

The use of unequal-width bins would perhaps mitigate this problem, but recall from Section 5.2.2 that the P-spline constructions used here are predicated on an assumption of equally-spaced "knots" (which, with splines of degree zero, are simply the bin endpoints). Without these, the unaltered finite-difference penalties on the coefficients no longer serve as approximations to derivatives of a suitable order. It then becomes nontrivial to penalize the $f_i$'s towards some predetermined "smooth" shape, although Li and Cao [176] proposed a method of modifying the P-spline penalty in the presence of uneven knots. We do not pursue this here, acknowledging that FRODO in its current state has slightly more difficulty using scale information in the covariate densities than it does using location or shape information.

For this dataset ($N = 200$ groups, each of size $n = 50$), we use parameter values $\left(\alpha, \tilde{\beta}, \sigma_Y\right) = (0.1, -0.9, 0.1)$. A preliminary visual inspection of KDE's or histograms (not shown) of the covariate data — and the observation that they are all strictly positive and highly concentrated near zero — justifies a random walk prior of order $r = 2$ on the densities. In order to capture the "high-resolution" differences between covariate measurements near zero as described above, we use a moderately large basis of size $K = 20$. With no reason to suspect severe deviations from this shape we once again set $\delta_i = 0.1$ for all groups. The observed covariates range from $1.3232 \times 10^{-4}$ to 16.3810. Zero is a natural choice for the left endpoint of the assumed domain, and because there are so few large values, we simply take the right endpoint to be the overall sample maximum 16.3810.

The regression results in the left plot of Figure 5.5 represent the most significant example of the phenomenon discussed in Section 5.4.2; namely, the heightened uncertainty in the regression function in regions where covariate measurements are sparse. Here, 99.73% of the observed $X_{ij}$'s lie in the left half of the domain, while all of the latent $\lambda_i^{-1}$'s lie within the first 3 bins. Thus, the pointwise 95% credible interval for $\beta$ is quite narrow near zero — where most of the covariates are concentrated — and becomes significantly wider moving from left to right. Once again, we compare FRODO to two scalar models: a naive linear regression using the of the $\bar{X}_i$'s as fixed covariates, and a hierarchical linear model in which

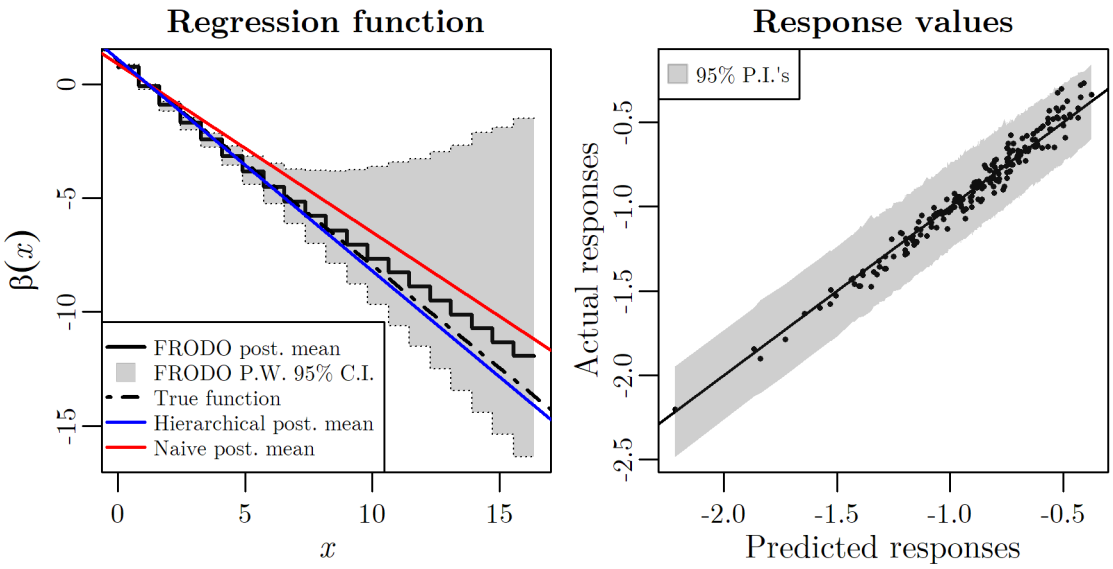


Figure 5.5: Results of FRODO applied to data with exponential covariates and a linear regression structure. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, and posterior mean estimates from hierarchical and naive scalar models. Right: responses $\hat{Y}_i$ predicted by FRODO (along with 95% prediction intervals) vs. the true response values.

the latent $\lambda_i$'s are jointly inferred with the regression parameters. As in previous studies, the estimated regression function from the hierarchical model is very close to the true function, and the FRODO estimate approximates it quite well. Some attenuation bias occurs in the right half of the domain, but because all of the covariate densities have such small mass in this region, this does not seem to adversely affect the regression inference in any other significant way. Indeed, the right plot of Figure 5.5 shows that the predicted responses closely align with the true $Y_i$'s.

As in previous studies, we compare inferred and true covariate densities for multiple groups in Figure 5.6. FRODO appears to do a good job of capturing the true densities for small, moderate, and large $\lambda_i$'s, although with no real deviations from the shape imposed by the random walk prior, this is perhaps not surprising.

### 5.4.4 Beta covariate densities, linear regression model

In the following two sections, we demonstrate FRODO's ability to capture regression relationships that are encapsulated in the shapes of the covariate densities, rather than their locations or scales. Whereas the covariate densities in preceding examples were governed by group-level latent parameters which were random themselves, here those parameters are deterministic, allowing us to better control the range of shapes we see. In particular, for this section we take $\xi = (\xi_1, \ldots, \xi_N)$ to be a mesh of equally-spaced points from 1/10 to

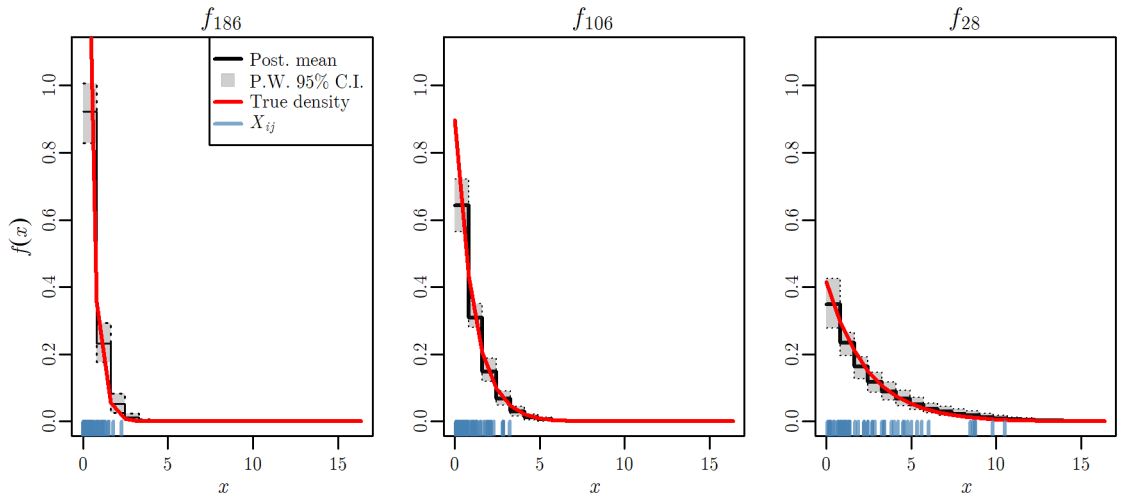


Figure 5.6: For a selection of groups (from the data with exponential covariates and a linear regression structure), the FRODO estimate of the group-specific covariate density, alongside its pointwise 95% credible region. The true densities are superimposed as red lines, and the actual covariate samples are shown as rug plots.

9/10, and

$$X_{ij} \sim \text{Beta}\left(\xi_i, 1-\xi_i\right).$$

The regression model is

$$\begin{aligned} Y_i &= \alpha + \tilde{\beta}\xi_i + \epsilon_i \\ &= \alpha + \mathbb{E}_i\left[\tilde{\beta}X\right] + \epsilon_i. \end{aligned}$$

The true densities $f_i^*$ are bimodal for all $i$, with peaks at 0 and 1 and minima at 1/2. For small $i$ with $\xi_i < 1-\xi_i$, the peak on the left is wider than the one on the right, so $f_i$ is skewed towards 0 and $\mathbb{E}_i[X] < 1/2$. The opposite is true for large $i$, and for $i$ near $N/2$ the densities are roughly symmetric.

For this simulation, we use $N = 250$ groups. Because the beta densities have relatively low variance (for the parameter values used here, all of them have variance below 1/8), we use relatively small groups of size $n_i = 15$ for all $i$, so that the difference between the "true" and "naive" regression functions is more pronounced[7]. The true regression parameters are $\left(\alpha, \tilde{\beta}, \sigma_Y\right) = (0.2, 1, 0.05)$.

Upon inspection of the available covariate data, one would see that all covariate measurements are constrained to the unit interval, with the minimum and maximum measurements

[7]With large groups, the "naive" regression with group-level covariate sample means would be quite close to the true model, making it difficult to tell which one FRODO was capturing.

being extremely close to 0 and 1, respectively. Thus, $[a', b'] = [0, 1]$ is a sensible choice for the assumed domain. Quick visual assessment of KDE or histogram estimates for the group-specific covariate densities reveals that they are neither Gaussian nor exponential. This observation, combined with the strong evidence that the densities are supported only on the unit interval, may lead one to believe that the covariates within each group are, indeed, roughly beta-distributed. This justifies a random walk prior of order $r = 1$ on the densities, for which the limiting shape is a uniform distribution. Note, however, that unlike the examples above for which we used second- and third-order random walk priors, here the limiting behaviour is *unique*, in the sense that there is only one uniform density on the chosen domain. Thus, if all groups had small smoothing parameter scales $\delta_i$ (corresponding to a prior assumption that no severe deviations from the limiting shape occurred), the FRODO estimates of the covariate densities all would be nearly identical, thereby suppressing the differences between groups and compromising the model's ability to extract meaningful regression information. With an assumed first-order random walk prior, one should therefore expect that the covariate densities will exhibit larger deviations from the limiting shape than they would in a situation where $r > 1$ was appropriate (especially since a bimodal shape will be apparent for at least some of the groups upon preliminary visual inspection). Thus, rather than the default $\delta_i = 0.1$ used in previous examples, here we take $\delta_i = 1$ for all groups. Finally, since several groups have most of their covariate measurements near the endpoints (necessitating bins which are narrow enough to capture differences in densities within these regions), we use $K = 12$ bins: more than the 10 used in the Gaussian examples, but less than the 20 used in Section 5.4.3 since we do not have enough covariate measurements per group to support such a large number of bins (especially since "roughness", or deviation from the random walk shape, is penalized less severely here).

Once again, the regression component of the model is visualized in Figure 5.7, alongside posterior mean estimates from naive and hierarchical scalar models. In contrast to previous datasets, here there are more covariate measurements at each endpoint of the domain than there are in the middle, leading to a slight "bulge" in the pointwise 95% credible interval around 0.5. However, each bin is relatively well-populated with observations, compared to the large differences in concentration seen in previous examples. It is visually obvious that FRODO captures the true regression function and not the naive one. The plot of predicted vs. true responses on the right of Figure 5.7 provides further confirmation that FRODO's regression inference is satisfactory here.

Figure 5.8 shows that FRODO has more difficulty inferring the true densities here than for previous examples. Although the asymmetrical shapes for $\xi_i$'s near 0.1 or 0.9 are captured, the steep curvature of the true densities near the endpoints in these cases results in them being near the edges of the model's pointwise 95% credible intervals — if not excluded altogether — in these regions. From the middle plot, we see that the model imposes a somewhat excessive degree of uniformity on the nearly-symmetric densities for

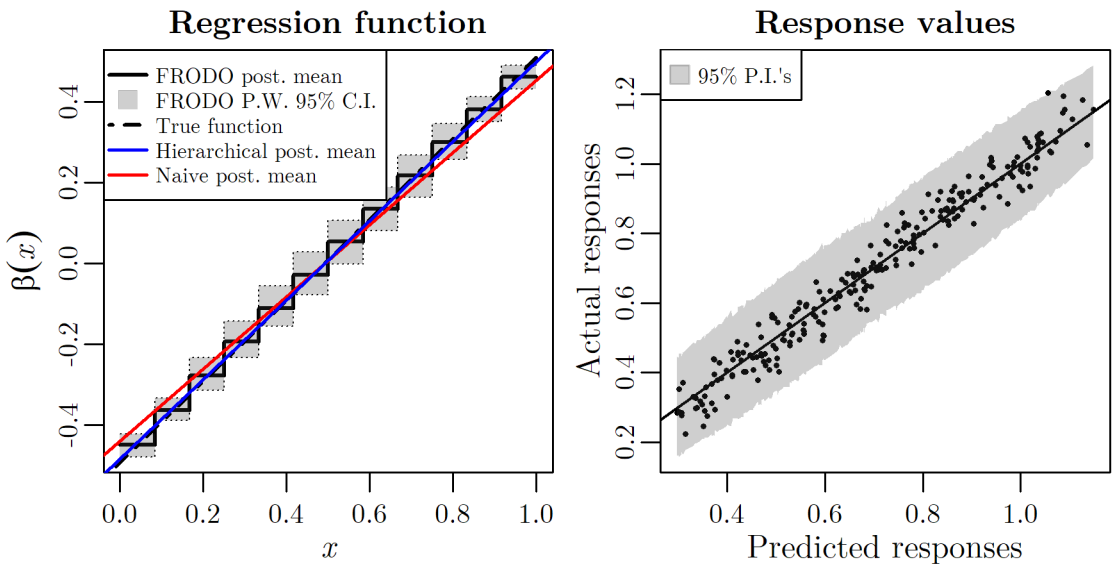


Figure 5.7: Results of FRODO applied to data with beta-distributed covariates and a linear regression structure. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, and posterior mean estimates from hierarchical and naive scalar models. Right: responses $\hat{Y}_i$ predicted by FRODO (along with 95% prediction intervals) vs. the true response values.

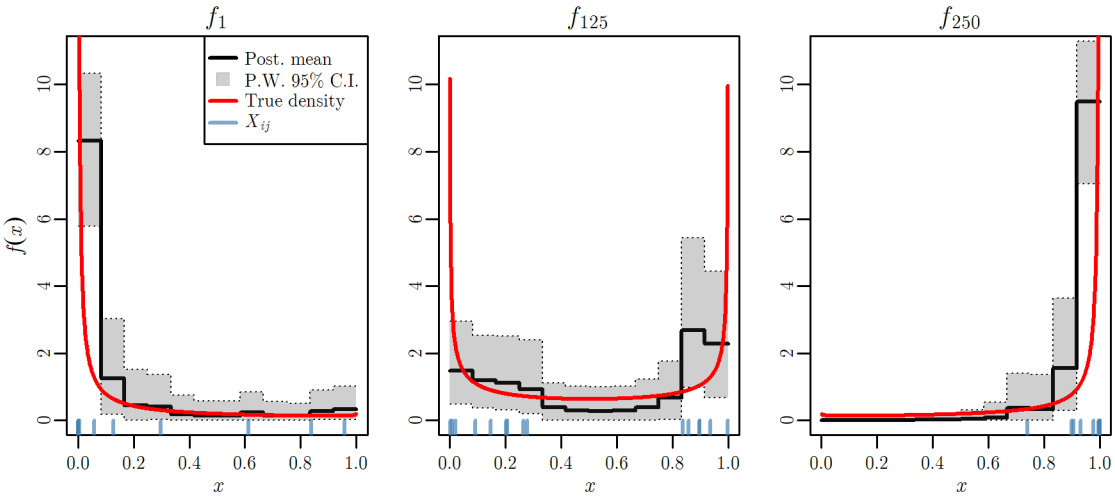


Figure 5.8: For a selection of groups (from the data with beta-distributed covariates and linear regression structure), the FRODO estimate of the group-specific covariate density, alongside its pointwise 95% credible region. The true densities are superimposed as red lines, and the actual covariate samples are shown as rug plots.

which $\xi_i$ is near 0.5. These difficulties are not surprising: given the small group sizes and the fairly large values used for $K$ and the $\delta_i$'s, neither the prior nor the likelihood make very strong implications about the density shapes. Aside from collecting more covariate measurements for each group (i.e. strengthening the likelihood), the only other possible mitigation for this would be to strengthen the prior: either by using smaller $\delta_i$'s to more strictly enforce the uniform shape, or by using a smaller $K$ to reduce the dimensionality of the problem. However, as discussed above, both of these options would result in an obfuscation of any information that does exist in the available covariate data. Thus, the most prudent choice seems to be accepting that FRODO's density inference in this example is necessarily limited to some degree. Fortunately, this limitation does not adversely affect any of the inference on the regression side of the model. Furthermore, despite the relative "roughness" of the FRODO density estimates[8], they are certainly improvements over, say, "raw" histograms (corresponding to $\delta_i \to \infty$), for which the low amount of covariate data would result in even less interpretable shapes.

### 5.4.5 Beta covariate densities, nonlinear regression model

Although the previous example shows that FRODO can extract relationships based on the shapes of covariate densities, the regression model itself still ultimately depended only on the means of the covariate measurements. The non-additive structure of the $X_{ij}$'s would pose a challenge for many established multilevel methods, but it is conceivable that one could devise a nonparametric, hierarchical Bayesian method which jointly inferred the $\mathbb{E}_i[X]$'s while using them to recover the correct regression parameters, subverting the need for full functional regression on the densities. When the regression is not linear, this may not be the case. Thus, in this section we combine a nonadditive covariate structure with a nonlinear regression model to demonstrate the full generality of FRODO. Once again $\xi$ is a mesh of equally-spaced points, this time from 1/10 to 2, and

$$\begin{aligned} X_{ij} &\sim \text{Beta}(\xi_i, \xi_i), \\ Y_i &= \alpha + \tilde{\beta}\left(1 + \frac{1}{2\xi_i + 1}\right) + \epsilon_i \\ &= \alpha + \mathbb{E}_i\left[4\tilde{\beta}\left(X - \frac{1}{2}\right)^2\right] + \epsilon_i. \end{aligned} \tag{5.24}$$

Here, the regression function is $\beta^*(x) = 4\tilde{\beta}(x - 1/2)^2$. The $f_i^*$'s are all symmetric: bimodal and U-shaped for $i$ near 1, roughly uniform for $i$ near $N/2$, and peaked at 1/2 for $i$ near $N$. For positive $\tilde{\beta}$, the expected response $\mathbb{E}_i[Y]$ is higher for "more bimodal" covariate densities

[8]Note that this is an inherent difficulty in any dataset for which the first-order random walk prior is justified, because imposing smoothness in this case is inseparable from forcing all of the densities towards being identical.

and lower for "more unimodal" ones. The regression is therefore entirely dependent on the shapes of the densities, not their locations or scales. Furthermore, because the densities are all symmetric it holds that $\mathbb{E}_i^* [X] = 1/2$ for all $i$. Thus, any modelling approach targeting $\beta(\mathbb{E}_i [X])$ ("regression on the expectaton") will be unsuitable here[9], as opposed to FRODO with its use of "the expectation of the regression", $\mathbb{E}_i [\beta(X)]$. In every aspect, this particular data structure is decidedly "non-classical", and FRODO seems uniquely well-suited to handle such a structure.

Because the true covariate densities all have expectation equal to 1/2, the regression function is actually not unique: indeed, when the $f_i^*$'s are all symmetric Beta densities, (5.24) is equivalent to $\alpha + \mathbb{E}\left[4\tilde{\beta}X^2\right] + \epsilon_i$, up to a term which is constant with respect to $i$. This does not seem to be a problem in practice, however: even when HMC chains are explicitly initialized such that $\beta$ is close to the latter form, they converge to a posterior which is consistent with (5.24). We conjecture that the FRODO posterior concentrates around the form of the regression function with "lowest error": empirically, we observed that the within-group sample means of $(X_{ij} - 1/2)^2$ values provide much more accurate estimates of their population analogues than the within-group sample means of the $X_{ij}^2$'s.

For this example, we simulated a dataset with $N = 250$ groups, each containing $n = 60$ covariate measurements. The true regression parameters were $\left(\alpha, \tilde{\beta}, \sigma_Y\right) = (0.7, 1, 0.1)$. As in Section 5.4.4, the observed range of the covariate measurements provides strong evidence that $[0, 1]$ is a good choice for the assumed density domain. Here, the range of shapes in preliminary histograms or KDE's (from bimodal, to roughly uniform, to unimodal) gives further justification for a random walk prior of order $r = 1$. As in the previous section, we take $\delta_i = 1$ for all $i$ to allow a greater degree of deviation from the limiting (uniform) shape of the prior. Because the data is highly concentrated near the endpoints for the groups whose $\xi$-values are low (even moreso than in Section 5.4.4's dataset), we use a basis of size $K = 15$.

Due to the aforementioned uselessness of methods involving "regression on expectations" here, constructing scalar models to compare with FRODO is nontrivial. We cannot use a "naive GAM" as we did for the Gaussian quadratic model in Section 5.4.2. There, $\mathbb{E}_i [X^2]$ and $(\mathbb{E}_i [X])^2$ differed by a constant, but this is not the case here. Thus, the naive scalar model we use for comparisons is somewhat contrived: a linear regression model, using the within-group sample means of the $(X_{ij} - 1/2)^2$ values as covariates. As always, the hierarchical scalar model assumes the true forms of the regression function and covariate densities are all known, jointly inferring the $\xi_i$'s and all regression parameters.

[9] In theory, one could invoke a measurement error method with more general assumptions on the covariate structure. Recall that the frequentist approach of Hu and Schennach [136] described in Section 5.1 assumed a general functional mapping the $f_i^*$'s to the $\xi_i$'s. Although higher-order moments should be permissible under their assumptions, the authors required a *known* functional. Thus, even with their level of generality it would still be necessary to assume quadratic regression *a priori*.

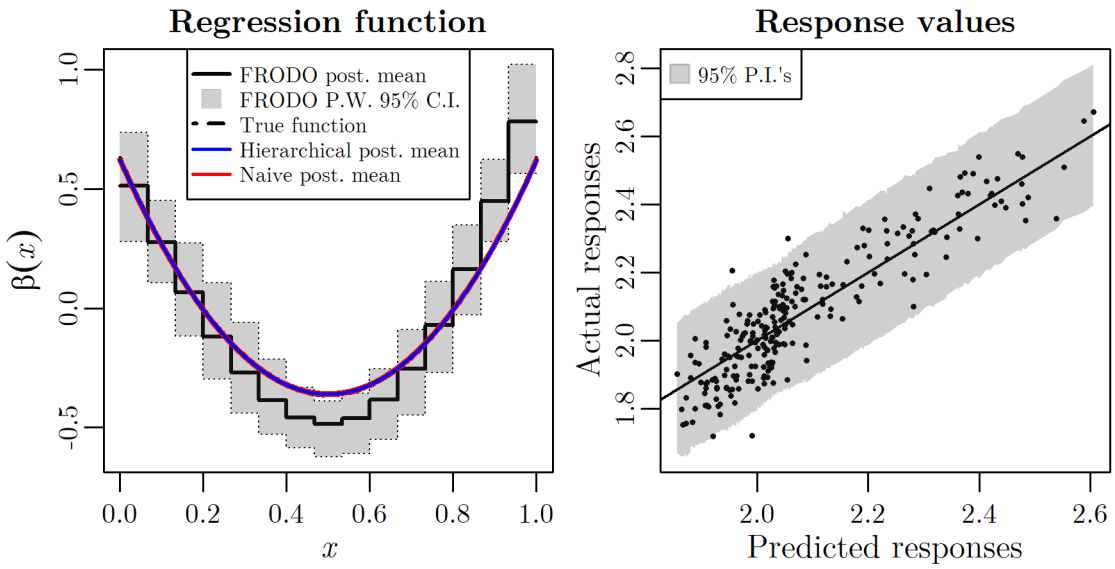


Figure 5.9: Results of FRODO applied to data with beta-distributed covariate data and a quadratic regression structure. Left: the regression function estimated by FRODO, alongside its pointwise 95% credible region, the true function, and posterior mean estimates from hierarchical and naive scalar models. Right: responses $\hat{Y}_i$ predicted by FRODO (along with 95% prediction intervals) vs. the true response values.

Because of the relatively large group sizes, and the fact that the quadratic form of the regression function was assumed known in both scalar models, the naive model does not suffer from any appreciable attenuation bias. As shown on the left of Figure 5.9, both it and the hierarchical scalar model approximate the true regression function almost perfectly. Some bias is apparent in the FRODO estimate, particularly near the vertex at 1/2, but its pointwise 95% credible interval almost completely captures the true function. On the right side of Figure 5.9, we see a moderate "clumping" of predicted responses just over 2.0, where the variability in the actual $Y_i$'s exceeds that of the mean predictions from FRODO. These values correspond to groups with $\xi$-values near 1 (i.e. those whose true covariate densities $f_i^*$ are close to uniform). For this dataset, it appears that FRODO has a small amount of difficulty capturing small shape differences between nearly-uniform densities. Note also that a few groups have posterior 95% prediction intervals which exclude their observed responses, although it seems reasonable to attribute this to mere random chance given the large number of groups. In any case, the overall fit appears largely satisfactory, especially considering that the true forms of the regression function and covariate densities are not known *a priori*.

Figure 5.10 shows that FRODO roughly captures all three types of density shapes present in this data, although some excess noise and bias is evident in the posterior estimates. This is particularly evident for the unimodal density in the right plot. Although the true density is fully contained in the pointwise 95% credible interval, the posterior mean

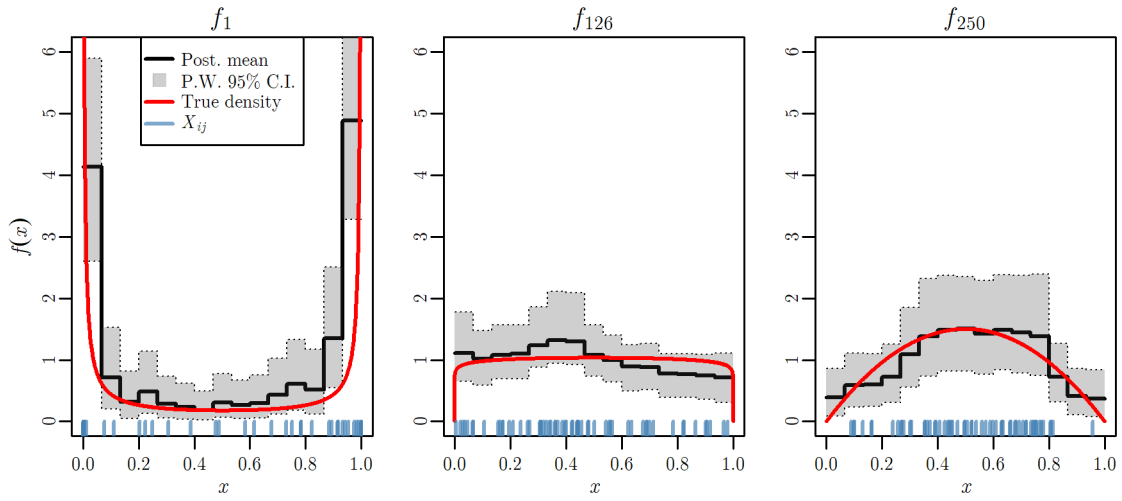


Figure 5.10: For a selection of groups (from the data with beta-distributed covariates and quadratic regression structure), the FRODO estimate of the group-specific covariate density, alongside its pointwise 95% credible region. The true densities are superimposed as red lines, and the actual covariate samples are shown as rug plots.

is perhaps somewhat too flat. The true unimodal densities in this dataset certainly differ more subtly from the uniform shape than the bimodal ones (contrast the true density in the left plot of Figure 5.10 with that on the right) — since the prior on densities here is structured only in terms of "deviations from uniformity", this slight deficiency is not entirely unexpected. As in Section 5.4.4, some of the excess noise in the density inference is an unavoidable consequence of the larger values of $K$ and $\delta$ necessary to capture the shapes and fine structure of the true densities with the first-order prior.

## 5.5 Extended simulation study: FRODO with varying group sizes and a group-level covariate

As a final "application" of FRODO, we recreate the simulated data considered by Croon and van Veldhoven [57]. This is very much a "classical" model, with Gaussian covariate data and a linear regression function much like the one considered in Section 5.4.1. However, there are three unique features here which were absent from the "toy" examples explored above. First (recalling the notation of (5.20–5.23)), the parameter values are $\left(\sigma_\xi, \sigma_X, \alpha, \tilde{\beta}, \sigma_Y\right) = \left(1, 3, 0.3, 0.3, \sqrt{0.35}\right)$: not only is the within-group variability of the $X_{ij}$'s much greater than the between-group variability of the true $\xi_i$'s, but the regression error is also quite high, accounting for just under 65% of the variability in the $Y_i$'s. Overall, the amount of "signal" in the data — at both the covariate and regression levels — is low relative to the amount of noise. Second, there are varying group sizes, some of which are quite small: out of $N = 100$ groups, roughly 50% (randomly selected with probability 1/2) contain $n_i = 10$ covariate

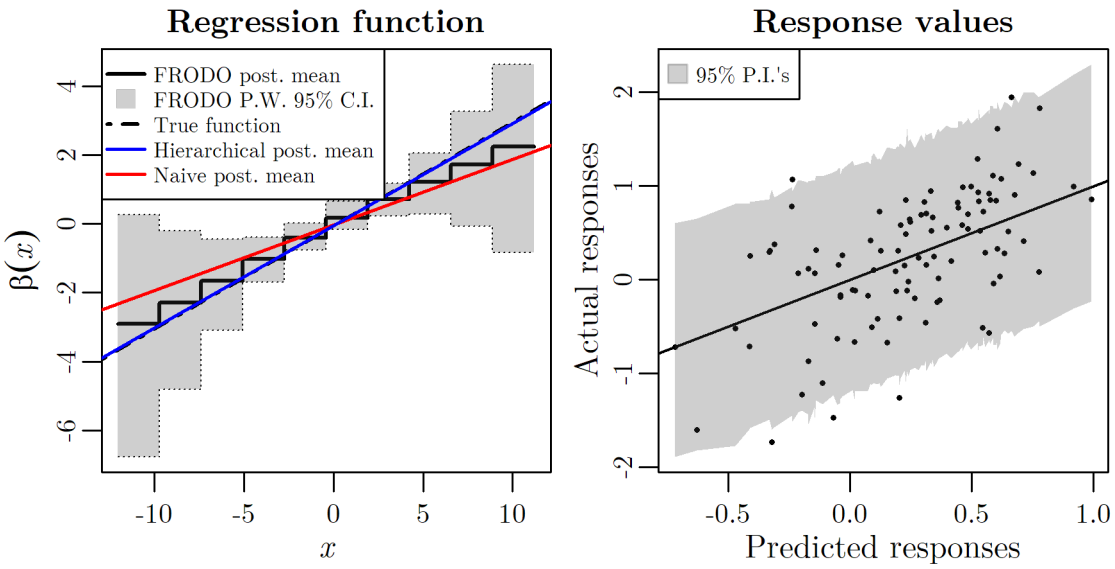


Figure 5.11: Results of FRODO applied to data with Gaussian covariates, a linear regression structure, and an additional group-level scalar covariate. Left: the regression function for the multilevel covariate estimated by FRODO, alongside its pointwise 95% credible region, the true function, and posterior mean estimates from hierarchical and naive scalar models. Right: responses $\hat{Y}_i$ predicted by FRODO (along with 95% prediction intervals) vs. the true response values.

measurements, and the rest contain $n_i = 40$. Finally, the actual regression model is altered from the basic FRODO form considered thus far, with the inclusion of a "scalar" group-level covariate $Z$ as in (5.1):

$$Y_i = \alpha + \tilde{\beta}\xi_i + \beta_Z Z_i + \epsilon_i. \tag{5.25}$$

The covariate values $Z_i$ are generated from a standard normal distribution, independently of $\xi$, and are treated as fixed observations.

It is straightforward to extend FRODO to accommodate for $Z$ by putting a $\mathcal{N}(0, 20\sigma_Y)$ prior on $\beta_Z$, conditionally independent from the prior for $\beta$ (which still denotes the regression function corresponding to the group-specific densities of the $X_{ij}$'s). We use a third-order random walk prior on the $f_i$'s with $K = 10$ bins as in Section 5.4.1, since the available data gives no reason to suspect that finer structures need to be captured. Due to the relatively small amount of covariate measurements, we simply take the assumed domain $[a', b']$ to be the range of observed $X_{ij}$-values, which in this case is $[-12.0365, 11.2258]$. For the groups of size $n_i = 40$, the default smoothing prior scale choice $\delta_i$ is appropriate, but with only $n_i = 10$ observations in the smaller groups, a tighter prior is necessary to ensure posterior density estimates with useful shape information. Thus, we set $\delta_i = 0.05$ for the small groups.

The actual method proposed by Croon and van Veldhoven [57] for micro-macro modelling is frequentist and involves a stepwise estimation procedure. An R implementation exists [188], but here we are only interested in comparing FRODO to analogous scalar Bayesian methods. Thus, as in the studies of Section 5.4 we compare it to both a naive and hierarchical scalar model, trivially extended to accommodate $Z$ with a suitable prior placed on $\beta_Z$. These results are shown in the left plot of 5.11. Note how much wider the pointwise 95% credible interval is — particularly near the endpoints — than the one in the similar model of Figure 5.1, owing to the higher noise and smaller amount of available covariate data here. It appears that the posterior for FRODO has concentrated somewhere in between the true and naive regressions. Indeed, FRODO's posterior mean for $\sigma_Y$ is 0.5975 (95% C.I. $(0.5152, 0.6945)$), in contrast with 0.5856 from the hierarchical scalar model (95% C.I. $(0.5014, 0.6835)$) and 0.6128 from the naive scalar model (95% C.I. $(0.5332, 0.7029)$). Given that the dataset is fairly small and high in noise, it is perhaps unsurprising that FRODO struggles more than it did in previous studies. However, this seems to be a problem of variability, not of bias: other simulated datasets with the exact same parameters, group sizes, and number of groups resulted in FRODO estimates with differing amounts of attenuation (not shown). Even the scalar hierarchical model proved quite variable with other datasets, as its estimate of the regression function did not always align as closely with the true function as it does here. Although the high degree of noise in the right plot of Figure 5.11 may appear troubling, this is reflective of the actual amount of noise in the data: a plot of predicted vs. actual responses from a frequentist multiple linear regression using the true $\xi_i$'s appears similar.

The usual density plots are shown in Figure 5.12. Note that the group in the left plot contains 40 individuals, and the other two contain only 10. It is intuitive that the smaller groups would have wider pointwise credible intervals for their densities (on further inspection, this pattern also seemed to hold for other groups not shown here), although it is somewhat noteworthy that the smaller $\delta_i$-values for these groups do not seem to neutralize this effect. Some bias in the model is evident, particularly in the middle plot, but overall the inference provided by FRODO seems reasonable.

## 5.6 Discussion and future work

In this chapter, we have presented a new approach for micro-macro modelling which combines density estimation and functional data analysis into a unified hierarchical Bayesian framework. Although FRODO is relatively simple in principle due to its use of step functions and only *linear* functional regression terms, it is deceptively powerful in its ability to use these elements for approximation of generalized additive models. Beyond the generality of the regression component of the model, FRODO is also quite flexible in terms of the individual-level covariate structures it can accommodate. Whereas many Bayesian methods

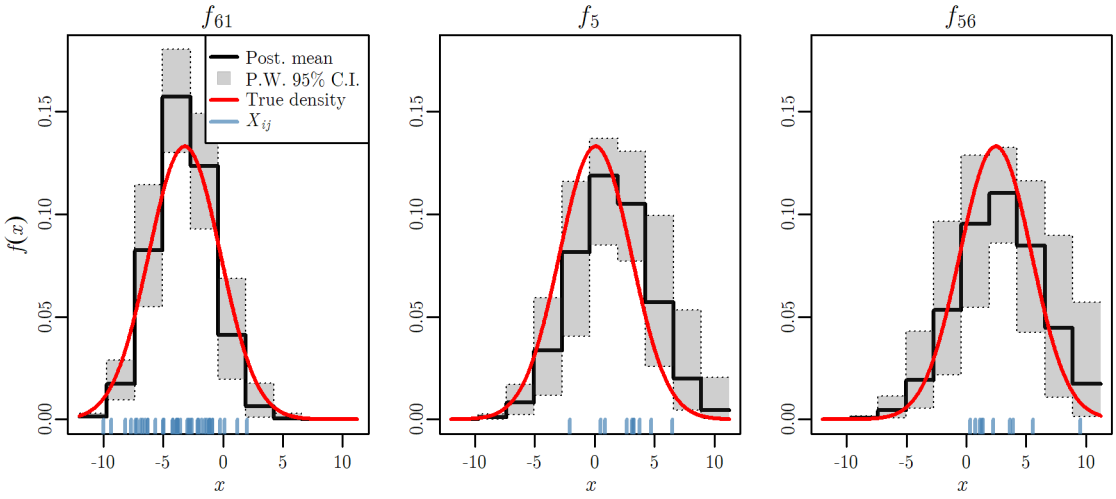


Figure 5.12: For a selection of groups (from the data with Gaussian covariate data, a linear regression structure, and an additional group-level covariate), the FRODO estimate of the group-specific covariate density, alongside its pointwise 95% credible region. The true densities are superimposed as red lines, and the actual covariate samples are shown as rug plots.

for GAM's with measurement error or micro-macro structure assume a Gaussian — or at the very least, additive — error structure in the $X_{ij}$'s, FRODO has no such limitation, allowing for covariate densities which influence the group-level regression responses through their locations, scales, or shapes. All that is required is the selection of a suitable prior structure for the densities, based on either prior domain knowledge, or — if this is not possible and an empirical Bayesian approach is required — a preliminary heuristic examination of the data. Although FRODO's inference on the covariate densities is generally more accurate when the true densities adhere to the specified "smooth shape" encoded in the prior, this is not a strict *requirement* provided hyperparameters are chosen carefully.

The simulation studies conducted above show that the power and generality of FRODO translate from theory to practice, providing reasonable inference for a variety of data structures. However, the potential for improvements and extensions to the model is vast. The most immediate potential for this is in the density part of the model, as described in Section 5.3.2. Here we have not considered $r^{\text{th}}$-order random walk priors for any integer $r > 3$. These would result in densities being penalized towards exponentiated polynomials of higher degree: with an $r^{\text{th}}$-order random walk prior, $\log f_i(x)$ is close to a polynomial of degree $r-1$ when the smoothing parameter $\tau_i$ is small. Such limiting smooth shapes correspond to *generalized error distributions* [286] (or folded versions thereof) with shape parameter $r-1$, of which the normal, Laplace, and uniform distributions are special cases. For $r > 2$, the generalized error distribution has lighter tails than a Gaussian. It is not certain how useful such higher-order random walk priors would be in practice (i.e. how often one might expect covariate densities to be similar to, say, an exponentiated quartic), but one challenge in

implementing these would be determining suitable distributions for the "free parameters" $\theta_{ik}$, $2 \leq k \leq r$. Equivalent derivations of the type carried out for $r = 2$ and 3 in Section 5.3.2 would be much more complex.

There is even room for generalization within the confines of the third-order (resp. second-order) random walk priors considered here. Although the construction in Section 5.3.2 was explicitly tailored in terms of Gaussian (resp. exponential) distributions, in principle it could be adapted for *any* densities whose logarithms are roughly quadratic (resp. linear) in shape. Folded or truncated normal distributions may be a useful shape to accommodate with a third-order random walk prior; one could even modify it to allow for densities $f$ such that $\log f$ is approximately quadratic with *positive* leading coefficient, not negative as for a Gaussian. This may be useful for modelling "U-shaped" densities, such as the Beta distributions considered in Section 5.4.4. Similarly, the second-order structure could be generalized to allow for positively-sloped densities (i.e. "reversed" exponentials), or Laplace densities whose logarithms are *piecewise* linear. Furthermore, it may be useful to combine differing random walk orders within the same model. For instance, the example in Section 5.4.5 might have benefited if we used a third-order random walk prior for the unimodal densities (since symmetric Beta densities are close to Gaussians in shape for large parameter values), a first-order R.W. prior for the flatter densities, and perhaps an "inverted" third-order R.W. prior for the U-shaped densities as suggested above.

Further investigation of the relationships between $n$, $r$, $K$, and $\delta$ would also be useful, particularly how best to set the latter two in terms of the former two. Although the empirical heuristic methods employed here worked well in practice, a more formal approach might result in better performance and generalization. Appeals to asymptotics could guide derivation of mathematical relationships between the hyperparameters: for instance, an expression for an "optimal" $\delta_i$ in terms of $r$, $K$, and $n_i$, based on the "big-O" relationships shown by Silverman [267] to guarantee convergence of penalized density estimators in the frequentist setting. The choice of the assumed domain for the densities may also have an effect on any such expressions.

There is also significant potential for generalizations on the regression side of the model. The most immediate of these is the realization of our proposed extension to non-Gaussian responses such as count or categorical data. Just as the regression part of FRODO for the Gaussian responses considered here is nothing more than a functional linear model, allowing for other response types is simply a matter of using functional GLM machinery.

Perhaps the most useful immediate extension to FRODO would be the incorporation of multiple multilevel covariates. Indeed, many real-world micro-macro datasets include several covariates measured at the individual level within groups [e.g 57, 4, 60]. Of course, this would increase the computational complexity of FRODO, as the number of parameters to infer grows roughly linearly in the number of multilevel covariates. Note, however, that real-world micro-macro datasets commonly include ordinal covariates with a small number

of levels [e.g. 4, 60]. Modelling the distributions for these covariates requires only as many basis functions as there are levels, which would mitigate computational difficulty to some extent in practice.

A powerful yet challenging improvement would be modelling more complex relationships amongst covariates. For instance, Croon and van Veldhoven [57] considered a version of the simulation study replicated in Section 5.5 where the latent and observed group-level covariates ($\xi$ and $Z$, respectively) were correlated [see also measurement error literature such as 238]. Accounting for dependence between multilevel and "scalar" covariates in FRODO will be highly nontrivial, especially if one wishes to maintain flexibility in the shapes of the inferred densities. For instance, if the multilevel data is Gaussian as in Section 5.5, the most obvious way to account for correlation between $\xi$ and $Z$ is to explicitly include it in the prior for the $\xi_i$'s (see Section 5.3.2). However, we have found in practice that the $\xi_i$'s inferred by FRODO are often poor approximations for the actual latent group means of the $X_{ij}$'s, unless a Gaussian shape is heavily enforced on the $f_i$'s by deliberately taking very small $\delta_i$'s. This was not a problem for the examples in Sections 5.4.1 and 5.4.2, as the posterior density estimates ended up being close enough to the true Gaussians that there were no major difficulties in the inference. If such latent density parameters are required more explicitly to model correlations with scalar covariates, this inaccuracy may become problematic. The potential for dependence between distinct *multilevel* covariates is arguably even more interesting. Presumably this would require regression on multiple integrals over their joint densities. However, even with the degree-zero splines considered here, this would result in a substantial increase in computational complexity. Indeed, the number of coefficients required to model the joint density of $d$ multilevel covariates for a single group in this way is exponential in $d$. Therefore, some type of simplification would likely be required to make interactions between multilevel covariates viable. See Lambert and Eilers [167] for a discussion of multivariate density estimation with splines in the case of a single density.

We conclude by acknowledging potential shortcomings in FRODO for which there are likely no solutions, either due to the inherent properties of the model or the excessive computational difficulty that would be required to solve them. First, one may question the use of piecewise constant basis functions, since higher-order splines would certainly result in smoother and better-behaved density estimates. However, recall from Section 5.3.3 that this choice was made partially for computational convenience: it ensures that the integral of $\beta \cdot f_i$ is simply the inner product of the two functions' coefficients. This is no longer the case with higher-order splines, for which the integrals are more complicated expressions involving products between neighbouring coefficients. Beyond the heightened complexity, we also found in preliminary experiments that the resulting posterior geometry was extremely difficult to navigate with NUTS. Note that these experiments modelled the densities themselves with higher-degree splines, requiring (among other things) a potentially costly softmax

transformation of each $\theta_i$ vector. The other possibility is modelling the *logarithms* of the densities with splines [e.g. 218]. These approaches are equivalent for degree-zero splines, but with higher degrees the logarithmic approach requires approximate numerical integration to normalize the $f_i$'s, which are exponentiated piecewise polynomials. These numerical integrals, in turn, depend on the spline coefficients in complex ways which would likely complicate the posterior geometry even further. Thus, unless a radically different approach is used to fit the model, higher-order splines do not seem to be worth the effort, given the satisfactory results obtained with piecewise constant functions and the prevalence of ordinal covariates in real-world micro-macro data.

In earlier experiments (not shown), we found problems with bias and sampling efficiency when the within-group covariate noise was large relative to either the regression noise or between-group covariate scale. In the notation of the Gaussian model, problems occurred when the $n_i$'s were small and $\sigma_X$ was large relative to either $\sigma_\xi$ or $\sigma_Y$, especially when the magnitude of the effect size $\tilde{\beta}$ was large. This problem also affected hierarchical scalar models — suggesting that there is innate difficulty in the posteriors induced by such datasets — but FRODO did seem slightly more sensitive to it, in the sense that some parameter combinations were problematic for FRODO but not for a scalar model. These problems could be mitigated with different prior choices such as a zero-avoiding prior for $\sigma_Y$, but these can create bias [91]. Fortunately, we suspect that the relative noise levels which tend to create problems are unlikely to occur in practice, as they imply either extremely low-error regression models or high-error covariate groups.

Finally, it bears repeating that FRODO only models responses in terms of *expectations of functions of covariates*: any regression relationship that cannot be expressed in the form (5.8), or some multivariate extension thereof, is incompatible with this methodology. In particular, responses which depend on the medians or modes of densities cannot be modelled with FRODO, requiring other methods specifically suited for those purposes [e.g. 136]. Its current inability to model *functions of expectations* may also be a shortcoming. For instance, if the data in Section 5.4.2 was modified so that the covariate densities had unequal variances and the group-level responses were proportional to these variances, FRODO would not be usable due to the nonconstant $(\mathbb{E}_i[X])^2$ term in the regression. One could potentially augment (5.8) with an "outer function", using terms of the form $g(\mathbb{E}_i[\beta(X)])$ with some unknown function $g$ to be modelled with a basis function expansion. However, this would likely create a litany of problems with unidentifiability.

Despite these challenges, we believe that FRODO's power and flexibility make it a strong addition to the field of micro-macro regression modelling, especially as improvements and extensions are developed to handle an even broader variety of data structures.

# Chapter 6

# Conclusion

In this thesis, we have presented an assortment of studies on uncertainty quantification for several types of nonparametric statistical methods. Chapters 2 and 3 detail a probabilistic numerical tool for use in state-space models. It uses Bayesian quadrature to determine whether the assumptions underpinning the Laplace approximation are justified — whether the likelihood of the model is "close enough" to a Gaussian shape to justify its use. The central philosophy of "good-enough-ness-of-fit" embodied by the diagnostic is relatively novel and unusual, but ensures that meaningful, useful results are obtained in high dimensions and with real data, and that practical performance is not sacrificed in the name of asymptotic guarantees.

Chapter 4 is a detailed overview of density inference methods, encompassing a full spectrum of practical and theoretical methods for many different types of density estimator. Several meaningful concepts in both Bayesian and frequentist nonparametric inference are discussed and put into context, and a simulation study is shown which compares a small assortment of the described methods.

Chapter 5 introduces FRODO, a combination of functional linear regression and density inference for generalized additive modelling of micro-macro data. The method allows a level of generality not often seen in comparable Bayesian literature, and was shown to provide meaningful and accurate inference with a wide variety of simulated data sets encompassing many types of regression and covariate structures. The potential for further improvements to FRODO is vast, even in spite of its promising results at present.

Collectively, these chapters demonstrate that better incorporation and quantification of uncertainty in complex situations can give rise to improved insights and powerful new methods. Implementation of the tools detailed in this thesis — and, more broadly, the principles motivating them — have the potential to elevate the quality of statistical inference performed in a variety of modern contexts.

# Bibliography


[1] Shigeo Abe. Training of support vector machines with mahalanobis kernels. In Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrożny, editors, *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*, pages 571–576. Springer Berlin Heidelberg, 2005.

[2] William H. Aeberhard, Joanna Mills Flemming, and Anders Nielsen. Review of State-Space Models for Fisheries Science. *Annual Review of Statistics and Its Application*, 5:215–235, 2018.

[3] J. Aitchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

[4] Adalgiso Amendola, Cristian Barra, and Roberto Zotti. Does graduate human capital production increase local economic development? An instrumental variable approach. *Journal of Regional Science*, 60(5):959–994, 2020.

[5] Ienkaran Arasaratnam and Simon Haykin. Cubature kalman filters. *IEEE Transactions on Automatic Control*, 54(6):1254–1269, 2009.

[6] Raffaele Argiento, Ilaria Bianchini, and Alessandra Guglielmi. Posterior sampling from $\epsilon$-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics*, 10(2):3516–3547, 2016.

[7] Mahdis Azadbakhsh, Hanna Jankowski, and Xin Gao. Computing confidence intervals for log-concave densities. *Computational Statistics and Data Analysis*, 75:248–264, 2014.

[8] G. Jogesh Babu and Yogendra P. Chaubey. Smooth estimation of a distribution and density function on a hypercube using Bernstein polynomials for dependent random vectors. *Statistics & Probability Letters*, 76(9):959–969, 2006.

[9] G. Jogesh Babu, Angelo J. Canty, and Yogendra P. Chaubey. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, 105(2):377–392, 2002.

[10] Fadoua Balabdaoui, Kaspar Rufibach, and Jon A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *The Annals of Statistics*, 37(3):1299–1331, 2009.

[11] Moulinath Banerjee and Jon A. Wellner. Likelihood ratio tests for monotone functions. *The Annals of Statistics*, 29(6):1699–1731, 2001.

[12] O E Barndorff-Nielsen, D R Cox, and H.F.D.R. Cox. *Asymptotic Techniques for Use in Statistics*. Asymptotic Techniques for Use in Statistics. Springer US, 1989.

[13] Ernesto Barrios, Antonio Lijoi, Luis E. Nieto-Barajas, and Igor Prünster. Modeling with Normalized Random Measure Mixture Models. *Statistical Science*, 28(3):313–334, 2013.

[14] Margot Bennink, Marcel A. Croon, and Jeroen K. Vermunt. Micro-Macro Multilevel Analysis for Discrete Data: A Latent Variable Approach and an Application on Personal Network Data. *Sociological Methods & Research*, 42(4):431–457, 2013.

[15] Margot Bennink, Marcel A. Croon, Brigitte Kroon, and Jeroen K. Vermunt. Micro-macro multilevel latent class models with multiple discrete individual-level variables. *Advances in Data Analysis and Classification*, 10:139–154, 2016.

[16] Casper W. Berg and Anders Nielsen. Accounting for correlated observations in an age-based state-space stock assessment model. *ICES Journal of Marine Science*, 73 (7):1788–1797, 2016.

[17] Michael Betancourt. How the shape of a weakly informative prior affects inferences, 2017. URL `https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes`.

[18] Michael Betancourt. A Conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434, 2017.

[19] Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, and Carlos Rodríguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237, 2011.

[20] P. J. Bickel and M. Rosenblatt. On Some Global Measures of the Deviations of Density Function Estimates. *The Annals of Statistics*, 1(6):1071–1095, 1973.

[21] Nicolai Bissantz, Lutz Dümbgen, Hajo Holzmann, and Axel Munk. Non-Parametric Confidence Bands in Deconvolution Density Estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(3):483–506, 2007.

[22] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

[23] Taoufik Bouezmarni and Jeroen V.K. Rombouts. Nonparametric density estimation for positive time series. *Computational Statistics & Data Analysis*, 54(2):245–261, 2010.

[24] Adam J. Branscum and Timothy E. Hanson. Bayesian Nonparametric Meta-Analysis Using Polya Tree Mixture Models. *Biometrics*, 64(3):825–833, 2008.

[25] François-Xavier Briol, Chris Oates, Mark Girolami, and Michael A. Osborne. Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees. In *Advances in Neural Information Processing Sections*, pages 1162–1170. MIT Press, 2015.

[26] François-Xavier Briol, Chris J. Oates, Mark Girolami, Michael A. Osborne, and Dino Sejdinovic. Probabilistic Integration: A Role in Statistical Computation? *Statistical Science*, 34(1):1–22, 2019.

[27] Lawrence Brown, Tony Cai, Ren Zhang, Linda Zhao, and Harrison Zhou. The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields*, 146:401–433, 2010.

[28] Lawrence D. Brown, Andrew V. Carter, Mark G. Low, and Cun-Hui Zhang. Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *The Annals of Statistics*, 32(5):2074–2097, 2004.

[29] Adam D. Bull. Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics*, 6:1490–1516, 2012.

[30] Adam D. Bull. A Smirnov-Bickel-Rosenblatt Theorem for Compactly-Supported Wavelets. *Constructive Approximation*, 37:295–309, 2013.

[31] Adam D. Bull and Richard Nickl. Adaptive confidence sets in $L^2$. *Probability Theory and Related Fields*, 156:889–919, 2013.

[32] John P Buonaccorsi. *Measurement error: models, methods, and applications*. Chapman and Hall/CRC, 2010.

[33] Theophilos Cacoullos. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1):179–189, 1966.

[34] T. Tony Cai and Mark G. Low. An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840, 2004.

[35] T. Tony Cai, Mark Low, and Zongming Ma. Adaptive Confidence Bands for Nonparametric Regression Functions. *Journal of the American Statistical Association*, 109 (507):1054–1070, 2014.

[36] Sebastian Calonico, Matias D. Cattaneo, and Max H. Farrell. On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. *Journal of the American Statistical Association*, 113(522):767–779, 2018.

[37] Antonio Canale. msBP: An R Package to Perform Bayesian Nonparametric Inference Using Multiscale Bernstein Polynomials Mixtures. *Journal of Statistical Software*, 78 (1):1–19, 2017.

[38] Antonio Canale and David B. Dunson. Multiscale Bernstein polynomials for densities. *Statistica Sinica*, 26(3):1175–1195, 2016.

[39] Bob Carpenter. Typical sets and the curse of dimensionality, 2017. URL `https://mc-stan.org/users/documentation/case-studies/curse-dims.html`.

[40] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[41] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, 2006.

[42] Ismaël Castillo. Pólya tree posterior distributions on densities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(4):2074–2102, 2017.

[43] Ismaël Castillo and Richard Nickl. On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics*, 42(5):1941–1969, 2014.

[44] José E. Chacón and Tarn Duong. *Multivariate Kernel Smoothing and Its Applications*. CRC Press, first edition, 2018.

[45] Henry R. Chai and Roman Garnett. Improving quadrature for constrained integrands. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2751–2759. PMLR, 2019.

[46] Song Xi Chen. Empirical Likelihood Confidence Intervals for Nonparametric Density Estimation. *Biometrika*, 83(2):329–341, 1996.

[47] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017.

[48] Gang Cheng and Yen-Chi Chen. Nonparametric inference via bootstrapping the debiased estimator. *Electronic Journal of Statistics*, 13(1):2194–2256, 2019.

[49] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818, 2014.

[50] Oksana A. Chkrebtii, David A. Campbell, Ben Calderhead, and Mark A. Girolami. Bayesian Solution Uncertainty Quantification for Differential Equations. *Bayesian Analysis*, 11(4):1239–1267, 2016.

[51] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer Cham, 2020.

[52] Nidhan Choudhuri, Subhashis Ghosal, and Anindya Roy. Bayesian Estimation of the Spectral Density of a Time Series. *Journal of the American Statistical Association*, 99(468):1050–1059, 2004.

[53] Yeonseung Chung and David B. Dunson. Nonparametric Bayes Conditional Distribution Modeling With Variable Selection. *Journal of the American Statistical Association*, 104(488):1646–1660, 2009.

[54] Jon Cockayne, Chris Oates, Tim Sullivan, and Mark Girolami. Bayesian Probabilistic Numerical Methods. *SIAM Review*, 61(4):756–789, 2019.

[55] Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1(1):54–81, 1993.

[56] Ciprian M. Crainiceanu and A. Jeffrey Goldsmith. Bayesian Functional Data Analysis Using WinBUGS. *Journal Of Statistical Software*, 32(11):195, 2010.

[57] Marcel A. Croon and Marc J.P.M. van Veldhoven. Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, 12(1):45–57, 2007.

[58] Miklos Csörgo and Pál Révész. *Strong approximations in probability and statistics*. Academic Press, first edition, 1981.

[59] Miklós Csörgő. An Invariance Principle for Nearest-Neighbor Empirical Density Functions. In *Quantile Processes with Statistical Applications*, chapter 9, pages 137–143. Society for Industrial and Applied Mathematics, 1983.

[60] Oumou Salama Daouda, Mounia N. Hocine, and Laura Temime. Determinants of healthcare worker turnover in intensive care units: A micro-macro multilevel analysis. *Plos One*, 16(5):1–13, 2021.

[61] Ingrid Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[62] Nicolaas Govert de Bruijn. *Asymptotic methods in analysis*. Courier Corporation, 1981.

[63] Maria De Iorio, Peter Müller, Gary L. Rosner, and Steven N. Maceachern. An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association*, 99(465):205–215, 2004.

[64] Maria De Iorio, Wesley O. Johnson, Peter Müller, and Gary L. Rosner. Bayesian Nonparametric Nonproportional Hazards Survival Modeling. *Biometrics*, 65:762–771, 2009.

[65] Perry de Valpine. Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear, non-Gaussian state-space models. *Bulletin of Marine Science*, 70(2):455–471, 2002.

[66] Hang Deng, Qiyang Han, and Cun-Hui Zhang. Confidence intervals for multiple isotonic regression and other monotone models. *The Annals of Statistics*, 49(4), 2021.

[67] Ronaldo Dias. Nonparametric Estimation: Smoothing and Visualization. Technical report, Universidade Estadual de Campinas, 2011.

[68] Lutz Dümbgen. New goodness-of-fit tests and their application to nonparametric confidence sets. *The Annals of Statistics*, 26(1):288–314, 1998.

[69] David B. Dunson and Ju-Hyun Park. Kernel Stick-Breaking Processes. *Biometrika*, 95(2):307–323, 2008.

[70] David B. Dunson, Natesh Pillai, and Ju-Hyun Park. Bayesian Density Regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(2): 163–183, 2007.

[71] Cécile Durot, Vladimir N. Kulikov, and Hendrik P. Lopuhaä. The limit distribution of the $L_\infty$-error of Grenander-type estimators. *The Annals of Statistics*, 40(3):1578–1608, 2012.

[72] Matthew C. Edwards, Renate Meyer, and Nelson Christensen. Bayesian nonparametric spectral density estimation using B-spline priors. *Statistics and Computing*, 29(1): 67–78, 2019.

[73] Paul H. C. Eilers and Brian D. Marx. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11(2):89–121, 1996.

[74] Paul H. C. Eilers, Brian D. Marx, and Maria Durbán. Twenty years of P-splines. *Statistics & Operations Research Transactions SORT*, 39(2):149–186, 2015.

[75] Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

[76] Michael Evans and Tim Swartz. Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems. *Statistical Science*, 10(3): 254–272, 1995.

[77] Jianqing Fan. Asymptotic normality for deconvolution kernel density estimators. *Sankhya : The Indian Journal of Statistics*, 53(1):97–110, 1991.

[78] Yanqin Fan and Yanjun Liu. A Note on Asymptotic Normality for Deconvolution Kernel Density Estimators. *Sankhyā: The Indian Journal of Statistics, Series A*, 59 (1):138–141, 1997.

[79] S. Favaro, M. Lomeli, and Y. W. Teh. On a class of $\sigma$-stable Poisson-Kingman models and an effective marginalized sampler. *Statistics and Computing*, 25(1):67–78, 2015.

[80] S. Favaro, A. Lijoi, C. Nava, B. Nipoti, I. Prünster, and Y. W. Teh. On the Stick-Breaking Representation for Homogeneous NRMIs. *Bayesian Analysis*, 11(3):697–724, 2016.

[81] Stefano Favaro and Yee Whye Teh. MCMC for Normalized Random Measure Mixture Models. *Statistical Science*, 28(3):335–359, 2013.

[82] Stefano Favaro and Stephen G. Walker. Slice Sampling $\sigma$-Stable Poisson-Kingman Mixture Models. *Journal of Computational and Graphical Statistics*, 22(4):830–847, 2013.

[83] Stefano Favaro, Maria Lomeli, Bernardo Nipoti, and Yee Whye Teh. On the stick-breaking representation of $\sigma$-stable Poisson-Kingman models. *Electronic Journal of Statistics*, 8(1):1063–1085, 2014.

[84] Mariel M. Finucane, Christopher J. Paciorek, Gretchen A. Stevens, and Majid Ezzati. Semiparametric Bayesian Density Estimation With Disparate Data Sources: A Meta-Analysis of Global Childhood Undernutrition. *Journal of the American Statistical Association*, 110(511):889–901, 2015.

[85] Carlo V Fiorio. Confidence intervals for kernel density estimation. *The Stata Journal*, 4(2):168–179, 2004.

[86] Lynn Foster-Johnson and Jeffrey D. Kromrey. Predicting group-level outcome variables: An empirical comparison of analysis strategies. *Behavior Research Methods*, 50 (6):2461–2479, 2018.

[87] David Freedman and Persi Diaconis. On the Maximum Deviation Between the Histogram and the Underlying Density. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 58:139–167, 1981.

[88] W. Gawronski and U. Stadtmüller. Smoothing Histograms by Means of Lattice-and Continuous Distributions. *Metrika*, 28:155–164, 1981.

[89] Alan E. Gelfand and Athanasios Kottas. A Computational Approach for Full Nonparametric Bayesian Inference Under Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 11(2):289–305, 2002.

[90] Alan E. Gelfand and Adrian F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Source: Journal of the American Statistical Association*, 85(410): 398–409, 1990.

[91] Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[92] Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457 – 472, 1992.

[93] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2013.

[94] Christopher Genovese and Larry Wasserman. Adaptive confidence bands. *The Annals of Statistics*, 36(2):875–905, 2008.

[95] John Geweke. Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica*, 57(6):1317–1339, 1989.

[96] Zoubin Ghahramani and Carl Rasmussen. Bayesian monte carlo. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

[97] Subhashis Ghosal. Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics*, 29(5):1264–1280, 2001.

[98] Subhashis Ghosal and Aad van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017.

[99] Evarist Giné and David M. Mason. On local U-statistic processes and the estimation of densities of functions of several sample variables. *The Annals of Statistics*, 35(3): 1105–1145, 2007.

[100] Evarist Giné and Richard Nickl. Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170, 2010.

[101] Evarist Giné and Richard Nickl. Adaptive Inference. In *Mathematical Foundations of Infinite-Dimensional Statistical Models*, pages 607–666. Cambridge University Press, Cambridge, 2015.

[102] Evarist Giné, Vladimir Koltchinskii, and Lyudmila Sakhanenko. Convergence in distribution of Self-Normalized Sup-Norms of Kernel Density Estimators. In Jørgen Hoffmann-Jørgensen, Jon A. Wellner, and Michael B. Marcus, editors, *High-Dimensional Probability III, Progress in Probability*, volume 55, pages 241–253. Birkhäuser, Basel, 2003.

[103] Evarist Giné, Vladimir Koltchinskii, and Lyudmila Sakhanenko. Kernel density estimators: convergence in distribution for weighted sup-norms. *Probability Theory and Related Fields*, 130:167–198, 2004.

[104] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006.

[105] Harvey Goldstein. *Multilevel statistical models*. John Wiley & Sons, fourth edition, 2010.

[106] I. J. Good and R. A. Gaskins. Nonparametric Roughness Penalties for Probability Densities. *Biometrika*, 58(2):255–277, 1971.

[107] Ulf Grenander. On the theory of mortality measurement. *Scandinavian Actuarial Journal*, 1956(2):125–153, 1956.

[108] J E Griffin. Default priors for density estimation with mixture models. *Bayesian Analysis*, 5(1):45–64, 2010.

[109] J. E. Griffin. An adaptive truncation method for inference in Bayesian nonparametric models. *Statistics and Computing*, 26:423–441, 2016.

[110] J. E. Griffin, M. Kolossiatis, and M. F. J. Steel. Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):499–529, 2013.

[111] Piet Groeneboom and Geurt Jongbloed. Nonparametric confidence intervals for monotone functions. *The Annals of Statistics*, 43(5):2019–2054, 2015.

[112] Piet Groeneboom, Gerard Hooghiemstra, and Hendrik P. Lopuhaä. Asymptotic Normality of the $L_1$ Error of the Grenander Estimator. *The Annals of Statistics*, 27(4): 1316–1347, 1999.

[113] Piet Groeneboom, Geurt Jongbloed, and Jon A. Wellner. Estimation of a convex function: characterizations and asymptotic theory. *The Annals of Statistics*, 29(6): 1653–1698, 2001.

[114] Piet Groeneboom, Geurt Jongbloed, and Birgit I. Witte. Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *The Annals of Statistics*, 38(1):352–387, 2010.

[115] Zhong Guan. Efficient and robust density estimation using Bernstein type polynomials. *Journal of Nonparametric Statistics*, 28(2):250–271, 2016.

[116] Tom Gunter, Michael A. Osborne, Roman Garnett, Philipp Hennig, and Stephen J. Roberts. Sampling for inference in probabilistic models with fast bayesian quadrature. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[117] Luis Gutiérrez, Ramsés H. Mena, and Matteo Ruggiero. A time dependent Bayesian nonparametric model for air quality analysis. *Computational Statistics and Data Analysis*, 95:161–175, 2016.

[118] Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14:375–395, 1999.

[119] Peter Hall. On convergence rates of suprema. *Probability Theory and Related Fields*, 89:447–455, 1991.

[120] Peter Hall. Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density. *The Annals of Statistics*, 20(2):675–694, 1992.

[121] Peter Hall. On Edgeworth Expansion and Bootstrap Confidence Bands in Nonparametric Curve Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):291–304, 1993.

[122] Peter Hall and Joel Horowitz. A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics*, 41(4):1892–1921, 2013.

[123] Peter Hall and Kee-Hoon Kang. Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *The Annals of Statistics*, 29(5):1443–1468, 2001.

[124] Peter Hall and Art B. Owen. Empirical Likelihood Confidence Bands in Density Estimation. *Journal of Computational and Graphical Statistics*, 2(3):273–289, 1993.

[125] Peter Hall and D. M. Titterington. On Confidence Bands in Nonparametric Density Estimation and Regression. *Journal of Multivariate Analysis*, 27(1):228–254, 1988.

[126] Qiyang Han and Jon A. Wellner. Approximation and estimation of $s$-concave densities via Rényi divergences. *The Annals of Statistics*, 44(3):1332–1359, 2016.

[127] Mark H. Hansen and Charles Kooperberg. Spline Adaptation in Extended Linear Models. *Statistical Science*, 17(1):2–51, 2002.

[128] Timothy Hanson, Haiming Zhou, and Vanda Inácio De Carvalho. Bayesian Nonparametric Spatially Smoothed Density Estimation. In Yichuan Zhao and Ding-Geng Chen, editors, *New Frontiers of Biostatistics and Bioinformatics*, chapter 4, pages 87–105. Springer International Publishing, 2018.

[129] Timothy E. Hanson. Inference for Mixtures of Finite Polya Tree Models. *Journal of the American Statistical Association*, 101(476):1548–1565, 2006.

[130] Nicolas W. Hengartner and Philip B. Stark. Finite-Sample Confidence Envelopes for Shape-Restricted Densities. *The Annals of Statistics*, 23(2):525–550, 1995.

[131] Philipp Hennig, Michael A. Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179), 2015.

[132] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[133] Marc Hoffmann and Richard Nickl. On adaptive inerence and confidence bands. *The Annals of Statistics*, 39(5):2383–2409, 2011.

[134] Joel L. Horowitz. The bootstrap. In James J. Heckman and Edward Leamer, editors, *Handbook of Econometrics*, volume 5, pages 3159–3228. Elsevier, 2001.

[135] Cheng Hsiao. Consistent estimation for some nonlinear errors-in-variables models. *Journal of Econometrics*, 41(1):159–185, 1989.

[136] Yingyao Hu and Susanne M. Schennach. Instrumental Variable Treatment of Nonclassical Measurement Error Models. *Econometrica*, 76(1):195–216, 2008.

[137] Youping Huang and Cun-Hui Zhang. Estimating a Monotone Density from Censored Observations. *The Annals of Statistics*, 22(3):1256–1274, 1994.

[138] Ferenc Huszár and David Duvenaud. Optimally-weighted herding is bayesian quadrature. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 377–386. AUAI Press, 2012.

[139] Hemant Ishwaran and Lancelot F James. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[140] Hemant Ishwaran and Mahmoud Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

[141] Lancelot F. James, Antonio Lijoi, and Igor Prünster. Posterior Analysis for Normalized Random Measures with Independent Increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

[142] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag New York, second edition, 2002.

[143] M. C. Jones. Simple boundary correction for density estimation kernel. *Statistics and Computing*, 3:135–146, 1993.

[144] M. C. Jones, J. S. Marron, and S. J. Sheather. A Brief Survey of Bandwidth Selection for Density Estimation. *Journal of the American Statistical Association*, 91(433): 401–407, 1996.

[145] Simon Julier and Jeffrey K. Uhlmann. A General Method for Approximating Nonlinear Transformations of Probability Distributions. Technical report, University of Oxford, 1996.

[146] Vesa Kaarnioja. Smolyak Quadrature. Master's thesis, University of Helsinki, 2013.

[147] Gerald Kaiser. *A Friendly Guide to Wavelets*. Modern Birkhäuser Classics. Birkhäuser Boston, 2011.

[148] Maria Kalli, Jim E. Griffin, and Stephen G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21:93–105, 2011.

[149] Nikolas Kantas, Arnaud Doucet, Sumeetpal S. Singh, Jan Maciejowski, and Nicolas Chopin. On Particle Methods for Parameter Estimation in State-Space Models. *Statistical Science*, 30(3):328–351, 2015. doi: 10.1214/14-STS511.

[150] Toni Karvonen and Simo Särkkä. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):697–720, 2018.

[151] Marc Kennedy. Bayesian quadrature with non-normal approximating functions. *Statistics and Computing*, 8:365–375, 1998.

[152] Gerard Kerkyacharian, Richard Nickl, and Dominique Picard. Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probability Theory and Related Fields*, 153:363–404, 2012.

[153] B. K. Kim and J. Van Ryzin. A bivariate histogram density estimator: Consistency and asymptotic normality. *Statistics & Probability Letters*, 3(3):167–173, 1985.

[154] Bock Ki Kim and John Van Ryzin. On the asymptotic distribution of a histogram density estimator. In *Colloquia Mathematica Societatis Janos Bolyai, 32. Nonparametric Statistical Inference*, pages 483–499, 1980.

[155] Arnošt Komárek and Emmanuel Lesaffre. Bayesian Accelerated Failure Time Model With Multivariate Doubly Interval-Censored Data and Flexible Distributional Assumptions. *Journal of the American Statistical Association*, 103(482):523–533, 2008.

[156] Arnošt Komárek, Emmanuel Lesaffre, and Joan F. Hilton. Accelerated Failure Time Model for Arbitrarily Censored Data With Smoothed Error Distribution. *Journal of Computational and Graphical Statistics*, 14(3):726–745, 2005.

[157] Charles Kooperberg. *logspline: Routines for Logspline Density Estimation*, 2020. URL `https://CRAN.R-project.org/package=logspline`. R package version 2.1.16.

[158] Charles Kooperberg and Charles J. Stone. A study of logspline density estimation. *Computational Statistics & Data Analysis*, 12(3):327–347, 1991.

[159] Charles Kooperberg and Charles J. Stone. Confidence intervals for logspline density estimation. In David D. Denison, Mark H. Hansen, Christopher C. Holmes, Bani Mallick, and Bin Yu, editors, *Nonlinear Estimation and Classification*, pages 285–295. Springer New York, 2003.

[160] Charles Kooperberg and Charles J. Stone. Comparison of Parametric and Bootstrap Approaches to Obtaining Confidence Intervals for Logspline Density Estimation. *Journal of Computational and Graphical Statistics*, 13(1):106–122, 2004.

[161] Siem Jan Koopman, Neil Shephard, and Drew Creal. Testing the assumptions behind importance sampling. *Journal of Econometrics*, 149(1):2–11, 2009.

[162] Michael R. Kosorok. Bootstrapping the grenander estimator. In N. Balakrishnan, Edsel A. Peña, and Mervyn J. Silvapulle, editors, *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, pages 282–292. Institute of Mathematical Statistics, 2008.

[163] Athanasios Kottas. Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, 136(3):578–596, 2006.

[164] Shinsuke Koyama, Lucia Castellanos Pérez-bolde, Cosma Rohilla Shalizi, and Robert E. Kass. Approximate Methods for State-Space Models. *Journal of the American Statistical Association*, 105(489):170–180, 2010.

[165] Kasper Kristensen, Anders Nielsen, Casper W. Berg, Hans Skaug, and Bradley M. Bell. TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70(1):1–21, 2016.

[166] Thomas Laloë and Rémi Servien. A note on the asymptotic law of the histogram without continuity assumptions. *Brazilian Journal of Probability and Statistics*, 30 (4):562–569, 2016.

[167] Philippe Lambert and Paul H. C. Eilers. Bayesian multi-dimensional density estimation with P-splines. In John Hinde, Jochen Einbeck, and John Newell, editors, *Proceedings of the 21st International Workshop on Statistical Modelling*, pages 313–320, 2006.

[168] Philippe Lambert and Paul H. C. Eilers. Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, 53(4):1388–1399, 2009.

[169] Stefan Lang and Andreas Brezger. Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.

[170] Michael Lavine. Some Aspects of Polya Tree Distributions for Statistical Modelling. *The Annals of Statistics*, 20(3):1222–1235, 1992.

[171] Alexandre Leblanc. A bias-reduced approach to density estimation using Bernstein polynomials. *Journal of Nonparametric Statistics*, 22(4):459–475, 2010.

[172] O. V. Lepskiĭ. Asymptotically Minimax Adaptive Estimation. I: Upper Bounds. Optimally Adaptive Estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

[173] Matthieu Lerasle. Adaptive non-asymptotic confidence balls in density estimation. *ESAIM - Probability and Statistics*, 16:61–85, 2012.

[174] Ker-Chau Li. Honest Confidence Regions for Nonparametric Regression. *The Annals of Statistics*, 17(3):1001–1008, 1989.

[175] Tong Li. Robust and consistent estimation of nonlinear errors-in-variables models. *Journal of Econometrics*, 110(1):1–26, 2002.

[176] Zheyuan Li and Jiguo Cao. General P-Splines for Non-Uniform B-Splines. arXiv:2201.06808, 2022.

[177] Antonio Lijoi, Igor Prünster, and Stephen G. Walker. Investigating nonparametric priors with Gibbs structure. *Statistica Sinica*, 18(4):1653–1668, 2008.

[178] D. V. Lindley. The Use of Prior Probability Distributions in Statistical Inference and Decisions. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 453–468. University of California Press, 1961.

[179] Yanyan Liu and Yanli Zhang. The consistency and asymptotic normality of nearest neighbor density estimator under $\alpha$-mixing condition. *Acta Mathematica Scientia*, 30 (3):733–738, 2010.

[180] Albert Y. Lo. A Large Sample Study of the Bayesian Bootstrap. *The Annals of Statistics*, 15(1):360–375, 1987.

[181] S. H. Lo, Y. P. Mack, and J. L. Wang. Density and Hazard Rate Estimation for Censored Data via Strong Representation of the Kaplan-Meier Estimator. *Probability Theory and Related Fields*, 80:461–473, 1989.

[182] María Lomelí, Stefano Favaro, and Yee Whye Teh. A Marginal Sampler for $\sigma$-Stable Poisson-Kingman Mixture Models. *Journal of Computational and Graphical Statistics*, 26(1):44–53, 2017.

[183] Hedibert F. Lopes and Ronaldo Dias. Bayesian Mixture of Parametric and Nonparametric Density Estimation: A Misspecification Problem. *Brazillian Review of Econometrics*, 31(1):19–44, 2011.

[184] Djamal Louani. On the asymptotic normality of the kernel estimators of the density function and its derivatives under censoring. *Communications in Statistics - Theory and Methods*, 27(12):2909–2924, 1998.

[185] Karim Lounici and Richard Nickl. Global uniform risk bounds for wavelet deconvolution estimators. *The Annals of Statistics*, 39(1):201–231, 2011.

[186] Mark G. Low. On nonparametric confidence intervals. *The Annals of Statistics*, 25 (6):2547–2554, 1997.

[187] Mark G. Low and Zongming Ma. Discussion of "frequentist coverage of adaptive nonparametric bayesian credible sets". *Ann. Statist.*, 43(4):1448–1454, 2015.

[188] Jackson G Lu, Elizabeth Page-Gould, and Nancy R Xu. *MicroMacroMultilevel: Micro-Macro Multilevel Modeling*, 2017. R package version 0.4.0.

[189] Steven N MacEachern. Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, volume 1, pages 50–55, 1999.

[190] Y. P. Mack. Asymptotic Normality of Multivariate k-NN Density Estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, 42(1):53–63, 1980.

[191] Y. P. Mack and M. Rosenblatt. Multivariate k-Nearest Neighbor Density Estimates. *Journal of Multivariate Analysis*, 9:1–15, 1979.

[192] Ester Mariucci, Kolyan Ray, and Botond Szabó. A Bayesian nonparametric approach to log-concave density estimation. *Bernoulli*, 26(2):1070–1097, 2020.

[193] Ryan Martin. Empirical Priors and Posterior Concentration Rates for a Monotone Density. *Sankhyā : The Indian Journal of Statistics*, 81(2):493–509, 2019.

[194] Elias Masry. Asymptotic normality for deconvolution estimators of multivariate densities of stationary processes. *Journal of Multivariate Analysis*, 44(1):47–68, 1993.

[195] Benoît R. Mâsse and Young K. Truong. Conditional Logspline Density Estimation. *The Canadian Journal of Statistics*, 27(4):819–832, 1999.

[196] The MathWorks, Inc. *MATLAB Optimization Toolbox*. Natick, MA, USA, 2019.

[197] Shaun McDonald and David Campbell. A review of uncertainty quantification for density estimation. *Statistics Surveys*, 15:1 – 71, 2021.

[198] Shaun McDonald and David Campbell. Supplement to "A Review of Uncertainty Quantification for Density Estimation", 2021.

[199] Jan Mielniczuk. Some Asymptotic Properties of Kernel Estimators of a Density Function in Case of Censored Data. *The Annals of Statistics*, 14(2):766–773, 1986.

[200] Thomas P. Minka. Deriving quadrature rules from Gaussian processes. Technical report, Carnegie Mellon University, 2000.

[201] Abdelkader Mokkadem and Mariane Pelletier. Confidence bands for densities, logarithmic point of view. *Alea*, 2:231–266, 2006.

[202] David S. Moore and James W. Yackel. Large sample properties of nearest neighbor density function estimators. In Shanti S. Gupta and David S. Moore, editors, *Statistical Decision Theory and Related Topics*, pages 269–279. Academic Press, 1977.

[203] Peter Müller and Abel Rodriguez. Dependent Dirichlet Processes and Other Extensions. In *Nonparametric Bayesian Inference*, pages 53–75. Institute of Mathematical Statistics, 2013.

[204] Peter Müller and Abel Rodriguez. Pólya Trees. In *Nonparametric Bayesian Inference*, pages 43–51. Institute of Mathematical Statistics, 2013.

[205] Lawrence M. Murray. Bayesian state-space modelling on high-performance hardware using LibBi. *Journal of Statistical Software*, 67(10), 2015.

[206] Radford M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[207] Michael H. Neumann. Strong Approximation of Density Estimators from Weakly Dependent Observations by Density Estimators from Independent Observations. *The Annals of Statistics*, 26(5):2014–2048, 1998.

[208] Anders Nielsen and Casper W. Berg. Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research*, 158:96–101, 2014.

[209] Luis E. Nieto-Barajas and Peter Müller. Rubbery Polya Tree. *Scandinavian Journal of Statistics*, 39(1):166–184, 2012.

[210] Luis E. Nieto-Barajas, Peter Müller, Yuan Ji, Yiling Lu, and Gordon B. Mills. A Time-Series DDP for Functional Proteomics Profiles. *Biometrics*, 68(3):859–868, 2012.

[211] Andriy Norets and Debdeep Pati. Adaptive Bayesian estimation of conditional densities. *Econometric Theory*, 33(4):980–1012, 2017.

[212] Andriy Norets and Justinas Pelenis. Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, 168(2):332–346, 2012.

[213] Andriy Norets and Justinas Pelenis. Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory*, 30(3):606–646, 2017.

[214] Michael Nussbaum. Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics*, 24(6):2399–2430, 1996.

[215] A. O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 1991.

[216] Michael Osborne. *Bayesian Gaussian Processes for Sequential Prediction, Optimisation and Quadrature*. PhD thesis, University of Oxford, 2010.

[217] Michael A. Osborne, David Duvenaud, Roman Garnett, Carl E. Rasmussen, Stephen J. Roberts, and Zoubin Ghahramani. Active Learning of Model Evidence Using Bayesian Quadrature. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in neural information processing systems*, pages 46–54. Curran Associates, Inc., 2012.

[218] Finbarr O'Sullivan. Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators. *SIAM Journal on Scientific and Statistical Computing*, 9(2): 363–379, 1988.

[219] Susan M. Paddock, Fabrizio Ruggeri, Michael Lavine, and Mike West. Randomized Polya tree models for nonparametric Bayesian inference. *Statistica Sinica*, 13(2):443–460, 2003.

[220] Omiros Papaspiliopoulos and Gareth O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

[221] Ju-Hyun Park and David B. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20(3):1203–1226, 2010.

[222] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[223] Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[224] Sonia Petrone. Random Bernstein Polynomials. *Scandinavian Journal of Statistics*, 26(3):373–393, 1999.

[225] Sonia Petrone. Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, 27(1):105–126, 1999.

[226] Sonia Petrone and Piero Veronese. Non parametric mixture priors based on an exponential random scheme. *Statistical Methods & Applications*, 11:1–20, 2002.

[227] Jim Pitman. Poisson-Kingman Partitions. *Lecture Notes-Monograph Series*, 40:1–34, 2003.

[228] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[229] Jakub Prüher, Filip Tronarp, Toni Karvonen, Simo Särkkä, and Ondřej Straka. Student-t process quadratures for filtering of non-linear systems with heavy-tailed noise. In *20th International Conference on Information Fusion*, Xi'an, China, 2017. Institute of Electrical and Electronics Engineers Inc.

[230] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[231] C. Radhakrishna Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44(1):50–57, 1948.

[232] James O. Ramsay and Bernard W. Siverman. *Functional Data Analysis*. Springer Series in Statistics. Springer New York, 2005.

[233] B. L. S. Prekasa Rao. Estimation of a Unimodal Density. *Sankhyā: The Indian Journal of Statistics, Series A*, 31(1):23–36, 1969.

[234] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[235] Kolyan Ray. Adaptive Bernstein-von Mises theorems in Gaussian white nosie. *The Annals of Statistics*, 45(6):2511–2536, 2017.

[236] Eugenio Regazzini, Antonio Lijoi, and Igor Prünster. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585, 2003.

[237] Pal Révész. On empirical density function. *Periodica Mathematica Hungarica*, 2: 85–110, 1972.

[238] Sylvia Richardson and Walter R. Gilks. Conditional Independence Models for Epidemiological Studies with Covariate Measurement Error. *Statistics in Medicine*, 12 (18):1703–1722, 1993.

[239] Jaakko Riihimäki and Aki Vehtari. Laplace Approximation for Logistic Gaussian Process Density Estimation and Regression. *Bayesian Analysis*, 9(2):425–448, 2014.

[240] Emmanuel Rio. Local invariance principles and their application to density estimation. *Probability Theory and Related Fields*, 98:21–45, 1994.

[241] James Robins and Aad van der Vaart. Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253, 2006.

[242] C. Rodriguez and J. Van Ryzin. Large sample properties of maximum entropy histograms. *IEEE Transactions on Information Theory*, 32(6):751–759, 1992.

[243] C. C. Rodríguez. Optimal recovery of local truth. In *AIP Conference Proceedings*, volume 567, pages 89–115. AIP Publishing, 2003.

[244] Carlos C. Rodríguez. On a New Class of Density Estimators. Technical report, State University of New York at Albany, 1986.

[245] Kathryn Roeder. Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.

[246] M. Rosenblatt. On the Maximal Deviation of $k$-Dimensional Density Estimates. *The Annals of Probability*, 4(6):1009–1015, 1976.

[247] Gordon J. Ross and Dean Markwick. *dirichletprocess: Build Dirichlet Process Objects for Bayesian Modelling*, 2020. URL `https://CRAN.R-project.org/package=dirichletprocess`. R package version 0.4.0.

[248] Judith Rousseau. On the Frequentist Properties of Bayesian Nonparametric Methods. *The Annual Review of Statistics and Its Applications*, 3:211–231, 2016.

[249] Judith Rousseau and Botond Szabó. Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *Annals of Statistics (to appear)*, 2019.

[250] Donald B. Rubin. The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

[251] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

[252] Kaspar Rufibach. *Log-concave density estimation and bump hunting for IID observations*. PhD thesis, University of Bern, 2006.

[253] L. Sakhanenko. Asymptotics of suprema of weighted Gaussian fields with applications to kernel density estimators. *Theory of Probability & its Applications*, 59(3):415–451, 2015.

[254] Sylvain Sardy and Paul Tseng. Density Estimation by Total Variation Penalized Likelihood Driven by the Sparsity $\ell_1$ Information Criterion. *Scandinavian Journal of Statistics*, 37(2):321–337, 2010.

[255] Abhra Sarkar, Bani K. Mallick, and Raymond J. Carroll. Bayesian Semiparametric Regression in the Presence of Conditionally Heteroscedastic Measurement and Regression Errors. *Biometrics*, 70(4):823–834, 2014.

[256] Simo Särkkä, Jouni Hartikainen, Lennart Svensson, and Fredrik Sandblom. On the relation between Gaussian process quadratures and sigma-point methods. arXiv:1504.05994, 2015.

[257] Susanne M. Schennach. Recent Advances in the Measurement Error Literature. *Annual Review of Economics*, 8:341–377, 2016.

[258] Anton Schick and Wolfgang Wefelmeyer. Root $n$ consistent density estimators for sums of independent random variables. *Journal of Nonparametric Statistics*, 16(6): 925–935, 2004.

[259] Anton Schick and Wolfgang Wefelmeyer. Pointwise convergence rates and central limit theorems for kernel density estimators in linear processes. *Statistics & Probability Letters*, 76(16):1756–1760, 2006.

[260] Bodhisattva Sen, Moulinath Banerjee, and Michael Woodroofe. Inconsistency of bootstrap: the Grenander estimator. *The Annals of Statistics*, 38(4):1953–1977, 2010.

[261] Paulo Serra and Tatyana Krivobokova. Adaptive Empirical Bayesian Smoothing Splines. *Bayesian Analysis*, 12(1):219 – 238, 2017.

[262] Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[263] Emmanuel Sharef, Robert L. Strawderman, David Ruppert, Mark Cowen, and Lakshmi Halasyamani. Bayesian adaptive B-spline estimation in proportional hazards frailty models. *Electronic Journal of Statistics*, 4:606–642, 2010.

[264] Weining Shen and Subhashis Ghosal. Adaptive Bayesian Procedures Using Random Series Priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213, 2015.

[265] Yushu Shi. *DPWeibull: Dirichlet Process Weibull Mixture Model for Survival Data.*, 2020. R package version 1.5.

[266] Yushu Shi, Michael Martens, Anjishnu Banerjee, and Purushottam Laud. Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models. *Bayesian Analysis*, 14(3):677–702, 2019.

[267] B.W. Silverman. On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method. *The Annals of Statistics*, 10(3):795–810, 1982.

[268] Hans J. Skaug and David A. Fournier. Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, 51(2):699–709, 2006.

[269] Nikolai Vasilvich Smirnov. On the construction of confidence regions for the density of distribution of random variables. *Doklady Akad. Nauk SSSR*, 74:189–191, 1950.

[270] Tom A. B. Snijders and Roel J. Bosker. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE Publications Ltd., 2011.

[271] Ulrich Stadtmüller. Asymptotic Distributions of Smoothed Histograms. *Metrika*, 30 (1):145–158, 1983.

[272] Stan Development Team. Stan modeling language users guide and reference manual, version 2.30.0, 2018.

[273] Stan Development Team. *RStan: the R interface to Stan*, 2021. R package version 2.21.3.

[274] Charles J. Stone. Asymptotic properties of logspline density estimation. Technical report, Department of Statistics, University of California, 1986.

[275] Charles J. Stone. Large-Sample Inference for Log-Spline Models. *The Annals of Statistics*, 18(2):717–741, 1990.

[276] Botond Szabó, A. W. van der Vaart, and J. H. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4): 1391–1428, 2015.

[277] Axel Tenbusch. Two-Dimensional Bernstein Polynomial Density Estimators. *Metrika*, 41:233–253, 1994.

[278] Luke Tierney and Joseph B. Kadane. Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[279] Surya T. Tokdar and Robert E. Kass. Importance sampling: a review. *Advanced Review*, 2(1):54–60, 2009.

[280] Hae-won Uh. *Kernel Deconvolution*. PhD thesis, University of Amsterdam, 2003.

[281] Mark A. van de Wiel, Dennis E. Te Beest, and Magnus M. Münch. Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics*, 46(1):2–25, 2019.

[282] Aad W. van der Vaart and Mark J. van der Laan. Smooth estimation of a monotone density. *Statistics: A Journal of Theoretical and Applied Statistics*, 37(3):189–203, 2003.

[283] A. J. van Es and H.-W. Uh. Asymptotic normality of nonparametric kernel type deconvolution density estimators: crossing the Cauchy boundary. *Nonparametric Statistics*, 16(2):261–277, 2004.

[284] Bert van Es and Hae-Won Uh. Asymptotic Normality of Kernel-Type Deconvolution Estimators. *Scandinavian Journal of Statistics*, 32(3):467–483, 2005.

[285] Philippe Van Kerm. Adaptive kernel density estimation. *The Stata Journal*, 3(2): 148–156, 2003.

[286] Mahesh K. Varanasi and Behnaam Aazhang. Parametric generalized gaussian density estimation. *The Journal of the Acoustical Society of America*, 86(4):1404–1415, 1989.

[287] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: an improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.

[288] Pieter Vermeesch. Statistical uncertainty associated with histograms in the Earth sciences. *Journal of Geophysical Research*, 110, 2005.

[289] Richard A. Vitale. A Bernstein Polynomial Approach to Density Function Estimation. In *Statistical Inference and Related Topics*, pages 87–99. Academic Press, 1975.

[290] Sara Wade, David B. Dunson, Sonia Petrone, and Lorenzo Trippa. Improving Prediction from Dirichlet Process Mixtures via Enrichment. *Journal of Machine Learning Research*, 15(30):1041–1071, 2014.

[291] Grace Wahba. Bayesian "Confidence Intervals" for the Cross-Validated Smoothing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):133–150, 1983.

[292] Stephen G. Walker. Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics-Simulation and Computation*, 36(1):45–54, 2007.

[293] B. Wang and W. Wertelecki. Density estimation for data with rounding errors. *Computational Statistics and Data Analysis*, 65:4–12, 2013.

[294] Lianming Wang and David B. Dunson. Fast Bayesian Inference in Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011.

[295] Xiao-Feng Wang. Bayesian nonparametric regression and density estimation using integrated nested Laplace approximations. *Journal of Biometrics and Biostatistics*, 25(4), 2013.

[296] Arthur B. Yeh. Bootstrap percentile confidence bands based on the concept of curve depth. *Communications in Statistics Part B: Simulation and Computation*, 25(4): 905–922, 1996.

[297] Changhe Yuan and Marek J. Druzdzel. Theoretical analysis and practical insights on importance sampling in Bayesian networks. *International Journal of Approximate Reasoning*, 46(2):320–333, 2007.

[298] R. Zamini, V. Fakoor, and M. Sarmad. Asymptotic Behaviors of Nearest Neighbor Kernel Density Estimator in Left-truncated Data. *Journal of Sciences, Islamic Republic of Iran*, 25(1):57–67, 2014.

[299] Cun-Hui Zhang. Fourier Methods for Estimating Mixing Densities and Distributions. *The Annals of Statistics*, 18(2):806–831, 1990.

[300] Song Zhang, Peter Müller, and Kim-Anh Do. A Bayesian Semi-parametric Survival Model with Longitudinal Markers. *Biometrics*, 66(2):435–443, 2010.

[301] Haoxuan Zhou. Bayesian Integration for Assessing the Quality of the Laplace Approximation. Master's thesis, Simon Fraser University, 2017.

[302] Yang Zu. A Note on the Asymptotic Normality of the Kernel Deconvolution Density Estimator with Logarithmic Chi-Square Noise. *Econometrics*, 3(3):561–576, 2015.

# Appendix A

# Details of the density UQ simulation study

## A.1 Introduction

This appendix is adapted from the supplementary material of our publication "A Review of Uncertainty Quantification for Density Estimation [198]. It contains a detailed explanation of the methodology for the simulation study described in Section 4.9 and displayed in Figure 4.1. The source code for this simulation study is available in its entirety at `https://github.com/ShaunMcDonald1021/density_UQ_review_paper`.

As described in Section 4.9, the data $\boldsymbol{X}$ is a sample of size 1000 from a Gaussian mixture $f_0 = 0.5\mathcal{N}\left(\frac{1}{2}, \frac{1}{49}\right) + 0.5\mathcal{N}\left(\frac{5}{7}, \frac{1}{490}\right)$.

The following sections describe and implement the different UQ methods used for the simulation study, all with a nominal level of $1 - \alpha = 0.95$.

## A.2 KDE methods

This section explores frequentist UQ methods based on a standard kernel density estimator. First, we implement the pointwise bias-corrected confidence intervals of Calonico et al. [36]. In keeping with the theory of Calonico et al., we use the compactly-supported Epanechnikov kernel. Their theory also assumes that the KDE bandwidth is selected to minimize pointwise MSE, *separately at each point*. Here, we instead use a global bandwidth to minimize *integrated* MSE, which ensures smooth estimates. We conjecture that the same "big-O" asymptotics underpinning the theory for pointwise-optimal bandwidths should translate when using the globally optimal one.

Next, we implement the simultaneous bootstrap confidence bands of Cheng and Chen [48], using 1000 bootstrap replicates. Note that their theory assumes the true density has compact support, and it and its gradient are equal to zero at the boundaries. Our choice of $f_0$ does not satisfy these assumptions, but we argue that it is sufficiently flat and low at the endpoints

of the interval to be considered "close enough". For consistency with the pointwise inference described above, we continue to use the Epanechnikov kernel. Unfortunately, this means that Cheng and Chen's variable-width bands cannot be implemented, as they involve quantities of the form

$$\left| \frac{f^*(x) - f_0(x)}{\sigma^*(x)} \right|,$$

where $f^*$ is a KDE based on a bootstrap sample and $\sigma^*$ is the usual sample estimate of its standard deviation. The problem arises from the use of a compactly-supported kernel: $\sigma^*(x)$ can easily be zero for $x$ near the boundaries if there are no points around there in a given bootstrap sample. Thus, we must use Cheng and Chen's fixed-width bands instead.

## A.3 Bernstein polynomial methods

Next, we try Bayesian UQ methods based on the Bernstein polynomial model of Petrone [224, 225]. All UQ here is based on a run of the MCMC algorithm proposed by Petrone [225], ran for 5000 iterations and discarding the first 1000 as burn-in. Our implementation of the algorithm can be obtained from the Github repository listed in Section A.1.

Recall from Section 4.4.2 that a truncated discrete prior must be placed on the dimensionality $K$; here we use a uniform prior over the integers $\{1, \ldots, 150\}$. We chose $M = 10$ for the concentration parameter of the Dirichlet process prior, which is used to induce a prior on the coefficients of the basis expansion.

Pointwise 95% credible intervals are obtained in a straightforward way: by using the pointwise (0.025, 0.975) quantiles of the density draws from the MCMC run. We also implement variable-width simultaneous credible bands based on median absolute deviations (MADs) [72]. Edwards, Meyer and Christensen applied such bands to the spectral density of a time series, but the machinery easily translates to the (probability) density UQ context. Unfortunately, the simultaneous bands are quite wide relative to the other UQ methods (see Figure 4.1). In fact, in the figure we have only taken the bands over the interval [0.01, 0.99], as taking them over the entire interval [0, 1] renders them too wide to be visually useful. Recall from Section 4.4.3 that these bands (obtained entirely from MCMC output) are of the form

$$\hat{f}(x) \pm \xi_\alpha \text{MAD}[f(x)],$$

where $\hat{f}$ is the posterior median and $\xi_\alpha$ is the $1 - \alpha$-quantile of

$$\sup_x \left( \left| f(x) - \hat{f}(x) \right| / \text{MAD}[f(x)] \right).$$

Most density draws $f$ have absolute deviations $\left| f(x) - \hat{f}(x) \right|$ which are quite small near the boundaries, so that the MAD is also quite small there. However, a moderate proportion of the draws have higher tail values, and therefore $\xi_\alpha$ turns out to be fairly large.

## A.4 Logspline methods

Here we look at a logspline density estimate with the pointwise bootstrapped confidence intervals described by Kooperberg and Stone [159, 160]. For this, we used the logspline R package [157], which implements a stepwise knot addition/deletion algorithm to fit a logspline density estimate. We used most of its default settings, although we did increase the maximum number of knots to 20. This allows for consistency with the bootstrap estimates below, for which the default setting for maximum number of knots did not always result in convergence.

As described in Kooperberg and Stone [160], confidence intervals for $f_0$ can be constructed using a Gaussian approximation. First, we obtain an estimate of the standard error of $\log \hat{f}$ using a small number of bootstrap samples (25, in this case). With this standard error estimate $\hat{\sigma}$, the pointwise confidence intervals for $f_0$ are of the form $\exp\left[\log \hat{f}(x) \pm z_{\alpha/2}\hat{\sigma}(x)\right]$, where $z_{\alpha/2}$ is the $\alpha/2$-quantile of the standard normal distribution.

## A.5 Dirichlet process mixture methods

The final type of UQ method considered here is based on the Dirichlet process mixture (DPM) model, as implemented in the dirichletprocess package [247]. This package uses a marginal sampling algorithm from Neal [206], but full UQ is possible due to the conjugacy of the Dirichlet process, as described in Section 4.7.1.

The mixture kernel $\kappa(\cdot \mid \theta)$ is taken to be Gaussian with location-scale parameters $\theta = (\mu, \sigma^2) \sim G$, where $G$ is a Dirichlet process prior with Normal-Inverse-Gamma base measure. We use the package's default choices for the base measure hyperparameters, as well as those for the Gamma prior on the Dirichlet process concentration parameter.

Following the advice given in the dirichletprocess package documentation, we linearly transform the sample to have zero mean and unit variance, as this helps with MCMC convergence. We run the sampler for 5000 iterations, discarding the first 1000 as burn-in. Pointwise credible intervals are once again obtained from the pointwise (0.025, 0.975) quantiles of the MCMC density draws.

# Appendix B

# Details of the implementation of FRODO in Stan

This appendix expands on the brief discussion in Section 5.3.4 regarding the Stan implementation of FRODO. We explain our method of initializing HMC chains, detail the parameter values used in the NUTS sampler, and assess the sampling behaviour of the simulation studies in Sections 5.4–5.5. The reader may also refer to our source code at `https://github.com/ShaunMcDonald1021/FRODO`.

This appendix will assume the reader is familiar with Stan, and the terminology associated with implementation and assessment of models therein. However, references to relevant Stan documentation are included where appropriate.

## B.1 Reparameterizations

It is known that Stan's sampling behaviour can suffer in the presence of difficult posterior geometries: for instance, when the posterior has heavy tails or nonlinear correlations between parameters [272, Section 25.7 of the User's Guide and references therein]. Following standard advice [ibid.], we use *non-centered parameterizations* for various parameters. Briefly, this means restating the target distribution (i.e. the posterior) in terms of parameters which do not have the same hierarchical dependence structures as in the original parameterization, thereby inducing a posterior geometry more amenable to HMC. The parameters of interest (see Sections 5.3.2–5.3.3) are then recovered as deterministic functions of the ones actually sampled. Additionally, the error variance $\sigma_Y$ is expressed as the ratio of a half-normal random variable and a Gamma random variable with shape parameter 2, neither of which have the type of heavy tails which are often problematic in NUTS [272]. The full details of the reparameterizations used are described in the comments of the source code referenced above.

## B.2 Initialization of chains

By default, Stan initializes all parameters uniformly in the range $[-2, 2]$ (for positive parameters, this is done on the logarithmic scale) [40]. This proved to be a problem for the densities: the default scheme, in conjunction with the reparameterizations discussed in Section B.1, almost always resulted in initial density estimates for which the logarithm of the posterior was infinite. It is not known how often these were "genuine" infinities as opposed to mere numerical overflow, but in either case the result is an inability to obtain posterior samples.

The problem appears to be related to the random walk structure of the $\theta_i$'s, which are encoded into the Stan model through a linear transformation of "non-centered" parameters. This transformation tends to "magnify" the variability in the default initial values to the extent that the initial $\phi_i$'s are severely mismatched with the likelihood of their corresponding covariate data (see Section 5.3.2). Thus, we use a modified initialization strategy based on preliminary frequentist estimates for the $f_i$'s. These are obtained using P-splines and Poisson regression models for the bin counts in each group, as proposed by Eilers and Marx [73, Section 8]. These are then "inverse-transformed" to obtain initial values for the parameterization used in Stan. A modest amount of randomness — Gaussian noise for the $\theta_i$'s, and Gamma-distributed initial values for the $\tau_i$'s and scale components for the "free parameter" means defined in Section 5.3.2 — is injected into the initialization to ensure that the starting points of the HMC chains are reasonably diffuse [92].

## B.3 Parameters of NUTS samplers

Sampling in Stan depends on several "parameters[1]" which govern the behaviour of the NUTS algorithm. Section 15.2 of the Stan Reference Manual [272] explains these parameters, and further details on their implications for sampling performance are discussed in the vignette at `https://mc-stan.org/misc/warnings.html`.

Due to the complexity of FRODO's posterior geometry, we found it necessary to use maximum tree depths and target Metropolis acceptance rates which were higher than the defaults (10 and 0.8, respectively). In all of the simulation studies shown in Sections 5.4–5.5, we used a maximum tree depth of 12. The target acceptance rate was set to 0.99, except in the studies with Gaussian covariate data, where it was set to 0.985. For each study, we ran four NUTS chains in parallel. Each chain was run for 750 warmup iterations, then 1000 sampling iterations.

[1] Not to be confused with the "parameters" whose posterior is the target of inference. In Section B.3, the word "parameters" refers only to the "*sampling* parameters" discussed therein.

| Study | Max. warmup time | Max. sampling time | Min. $n_{\text{Eff}}$ | Max. $\hat{R}$ |
|---|---|---|---|---|
| 5.4.1 | 665.823 | 576.135 | 668.08 | 1.004 |
| 5.4.2 | 2021.06 | 2230.08 | 809.87 | 1.011 |
| 5.4.3 | 842.034 | 1163.82 | 463.62 | 1.005 |
| 5.4.4 | 572.697 | 483.391 | 655.660 | 1.007 |
| 5.4.5 | 657.588 | 762.475 | 643.639 | 1.010 |
| 5.5 | 230.505 | 100.251 | 450.481 | 1.009 |

Table B.1: Various quantities quantifying the performance and sampling behaviour of FRODO, for each of the simulated datasets in Chapter 5.

## B.4 Behaviour of simulation runs

In Table B.1, we summarize the performance of the samplers for each of the six simulation studies in Chapter 5. Each study is denoted by the section in which it appears, and the following information is included for each one.

1. The maximum warmup time (in seconds) for any of the four chains,
2. the maximum sampling time (in seconds) for any of the four chains,
3. the smallest estimated [287] effective sample size ($n_{\text{Eff}}$) for any parameter in the model, and
4. the maximum split $\hat{R}$ value for any parameter in the model [287].

Note that the reported $n_{\text{Eff}}$ (resp. $\hat{R}$) is the minimum (resp. maximum) over the actual sampled parameters *and* the "true" model parameters obtained with transformations (see Section B.1). All simulations were run on an Acer laptop with 16 GB of RAM and four Intel i5-9300H 2.40GHz CPU cores.

In every simulation study, all parameters had effective sample sizes exceeding 450. Vehtari et al. [287] recommend a threshold of at least 400 effective samples per parameter, so we are confident that ours are large enough for inference to be reasonably accurate. Each of the studies with a quadratic regression structure (Sections 5.4.2 and 5.4.5) had a single split $\hat{R}$ value above the threshold of 1.01 recommended by Vehtari et al. [287]. For the Gaussian covariate data, this maximal $\hat{R}$ occurred for the value of the log posterior; and for the beta covariate data, it occurred for one of the density smoothing parameters $\tau_i$. Although split $\hat{R}$ values above 1.01 are often considered indicative of convergence problems, we are not concerned by a single value slightly exceeding this threshold in a model with thousands of parameters, especially since the estimated Monte Carlo standard errors [e.g. 287] for these parameters are less than 5% of their posterior standard deviations. Trace plots for these parameters, shown in Figure B.1, also suggest that there are not any egregious convergence problems.

As one would expect given FRODO's complexity, warmup and sampling are several times slower than they are for the corresponding scalar models used in the simulation studies (not shown). The only study whose computation time we would consider problematic is

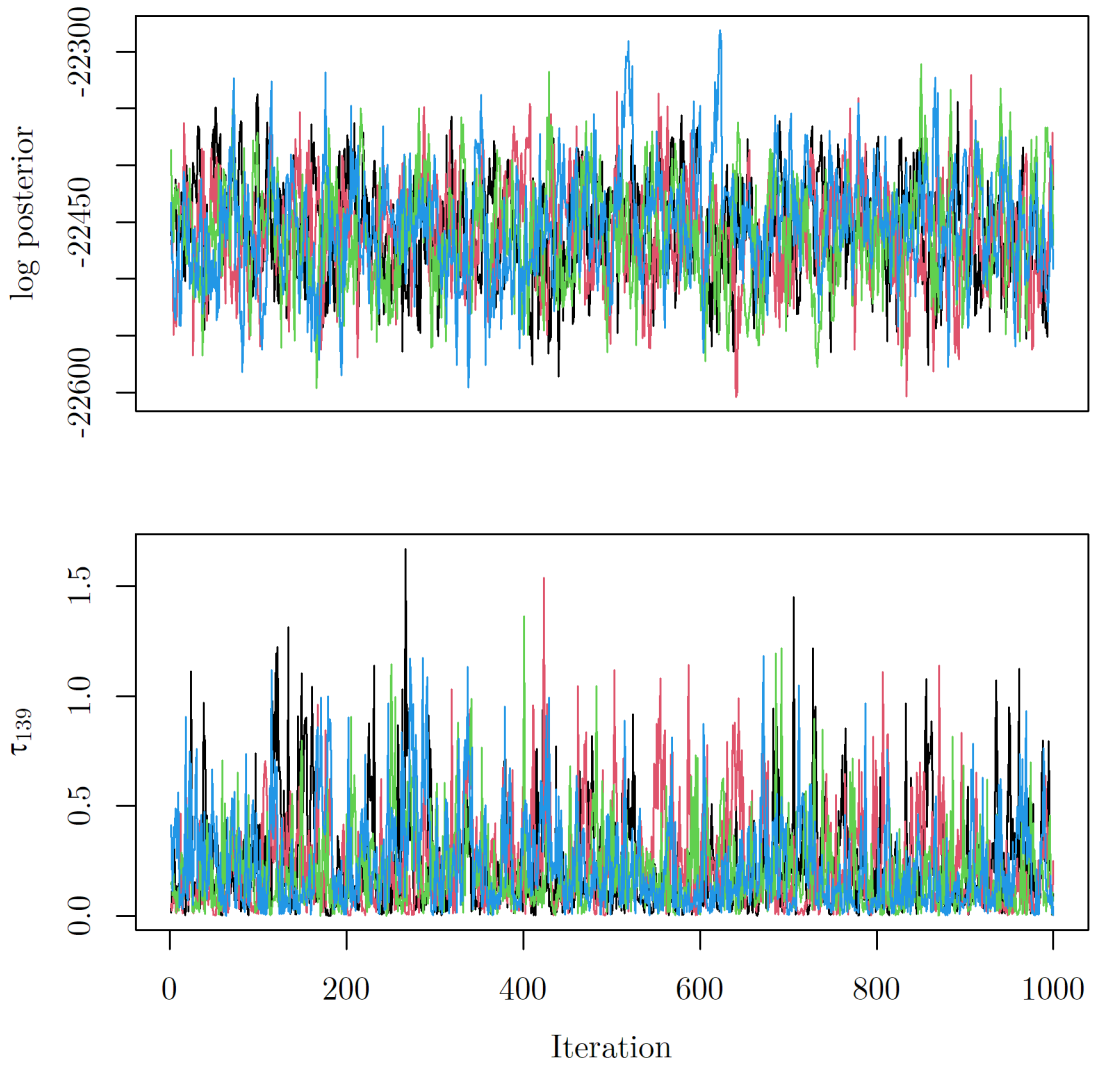


Figure B.1: Trace plots for simulation study parameters whose split $\hat{R}$ values were above 1.01. Top: the value of the log posterior from the model in Section 5.4.2. Bottom: the density smoothing parameter $\tau_{139}$ from the model in Section 5.4.5.

| Study | True | FRODO | Hierarchical scalar model | Naive scalar model |
|---|---|---|---|---|
| 5.4.1 | 0.5 | 0.494 (0.442, 0.551) | 0.490 (0.436, 0.549) | 0.556 (0.510, 0.602) |
| 5.4.2 | 0.5 | 0.472 (0.385, 0.566) | 0.479 (0.414, 0.550) | 0.885 (0.815, 0.962) |
| 5.4.3 | 0.1 | 0.099 (0.069, 0.126) | 0.096 (0.070, 0.122) | 0.165 (0.150, 0.183) |
| 5.4.4 | 0.05 | 0.065 (0.054, 0.077) | 0.055 (0.046, 0.065) | 0.089 (0.082, 0.098) |
| 5.4.5 | 0.1 | 0.094 (0.084, 0.105) | 0.099 (0.090, 0.109) | 0.150 (0.137, 0.164) |
| 5.5 | 0.592 | 0.595 (0.512, 0.695) | 0.585 (0.502, 0.681) | 0.614 (0.536, 0.709) |

Table B.2: Posterior inference for $\sigma_Y$ (the regression error) from FRODO and the scalar models within each simulation study. For each model, the posterior mean is reported, as is a 95% credible interval in parentheses. The second column from the left shows the true $\sigma_Y$.

the one from Section 5.4.2, with Gaussian covariate data and a quadratic regression structure. Including warmup and sampling, the Stan model for this study took over an hour to run. Most of the sampling iterations for this study had larger tree depths than in the other studies, meaning that the number of gradient evaluations involved in sampling was roughly higher by a factor of 2 or more [272, Section 15.2 of Reference Manual]. This is likely a consequence of posterior geometry, and the way in which the samplers adapt to it during warmup. However, it should be noted that we deliberately used a liberal number of warmup iterations, and chains appeared to have converged to the "typical set" [18] well before sampling began (not shown). Note also that the smallest effective sample size is over twice as large as the threshold of 400 recommended by Vehtari et al. [287] Therefore, reasonable posterior inference with acceptable computation time could likely be achieved by reducing the number of warmup and sampling iterations, provided the latter did not induce problematic $\hat{R}$ values.

Finally, recall from Section 5.4 that estimates of the regression variance, $\sigma_Y$, are biased upward in "naive" regression models, and this fact can be used to check whether or not FRODO is recovering "true" regression relationships. For each simulation study, Table B.2 shows the true value of $\sigma_Y$, as well as the posterior mean and 95% credible interval for this parameter from FRODO, the hierarchical scalar model, and the naive scalar model (see the beginning of Section 4.9). The endpoints of posterior intervals are simply 0.025- and 0.975-quantiles from the HMC samples. For each simulation study, the FRODO estimate for $\sigma_Y$ is much closer to the true value than the estimate from the naive model.