A Data-Retaining Framework for Tail Estimation

by

Erika L. Cunningham

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Surya Tokdar, Advisor

_____
Merlise Clyde

_____
Alexander Volfovsky

_____
Daniel Cooley

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2020

Abstract

A Data-Retaining Framework for Tail Estimation

by

Erika L. Cunningham

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Surya Tokdar, Advisor

_____
Merlise Clyde

_____
Alexander Volfovsky

_____
Daniel Cooley

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2020

Abstract

Modeling of extreme data often involves thresholding, or retaining only the most extreme observations, in order that the tail may "speak" and not be overwhelmed by the bulk of the data. We describe a transformation-based framework that allows univariate density estimation to smoothly transition from a flexible, semi-parametric estimation of the bulk into a parametric estimation of the tail without thresholding. In the limit, this framework has desirable theoretical tail-matching properties to the selected parametric distribution. We develop three Bayesian models under the framework: one using a logistic Gaussian process (LGP) approach; one using a Dirichlet process mixture model (DPMM); and one using a predictive recursion approximation of the DPMM. Models produce estimates and intervals for density, distribution, and quantile functions across the full data range and for the tail index (inverse-power-decay parameter), under an assumption of heavy tails. For each approach, we carry out a simulation study to explore the model's practical usage in non-asymptotic settings, comparing its performance to methods that involve thresholding.

Among the three models proposed, the LGP has lowest bias through the bulk and highest quantile interval coverage generally. Compared to thresholding methods, its tail predictions have lower root mean squared error (RMSE) in all scenarios but the most complicated, e.g. a sharp bulk-to-tail transition. The LGP's consistent underestimation of the tail index does not hinder tail estimation in pre-extrapolation to moderate-extrapolation regions but does affect extreme extrapolations.

An interplay between the parametric transform and the natural sparsity of the DPMM sometimes causes the DPMM to favor estimation of the bulk over estimation of the tail. This can be overcome by increasing prior precision on less sparse (flatter) nonparametric density shapes. A finite mixture model (FMM), substituted for the DPMM in simulation, proves effective at reducing tail RMSE over thresholding methods in some, but not all, scenarios and quantile-levels.

The predictive recursion marginal posterior (PRMP) model is fast and does the best job among proposed models of estimating the tail-index parameter. This allows it to reduce RMSE in extrapolation over thresholding methods in most scenarios considered. However, bias from the predictive recursion contaminates the tail, casting doubt on the PRMP's predictions in tail regions where data should still inform estimation. We recommend the PRMP model as a quick tool for visualizing the marginal posterior over transformation parameters, which can aid in diagnosing multimodality and informing the precision needed to overcome sparsity in the mixture model approach.

In summary, there is not enough information in the likelihood alone to prevent the bulk from overwhelming the tail. However, a model that harnesses the likelihood with a carefully specified prior can allow both the bulk and tail to speak without an explicit separation of the two. Moreover, retaining all of the data under this framework reduces quantile variability, improving prediction in the tails compared to methods that threshold.

To my Eric and my Nora

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols used across chapters

Note: Model-specific symbols are introduced within chapters.

Y	random variable, random variate
y_i	data or observations
n	sample size of data
f	sampling distribution or data likelihood
$\mathcal{F}$	family of sampling distributions
g	pdf of transformation distribution
G	transformation, itself a CDF
h	continuous nonparametric density function over $[0, 1]$
$\mathcal{H}$	family of continuous nonparametric density functions over $[0, 1]$
τ	variable typically used to index domain of h
p	quantile level of sampling distribution
θ	vector of parameters used to index g, either (σ, ν) or (σ, ξ)
Θ	space of possible θ parameters
ν	power decay parameter for g
ξ	tail index parameter for g, equivalent to $1/\nu$
σ	scale parameter of g
k	a family of kernels with domain $[0, 1]$

Abbreviations

CDF	cumulative distribution function
DP	Dirichlet Process
DPM	Dirichlet Process Mixture
DPMM	Dirichlet Process Mixture Model
EGPD	Extended Generalized Pareto Distribution
FM	Finite Mixture
FMM	Finite Mixture Model
GP	Gaussian Process
GPD	Generalized Pareto Distribution
HC	Half Cauchy
HPD	highest posterior density
HT	Half-t
LGP	Logistic Gaussian Process
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate
MSE	mean squared error
MVN	Multivariate Normal
pdf	probability density function
PR	Predictive Recursion
PRML	Predictive Recursion Marginal Likelihoods
PRMP	Predictive Recursion Marginal Posterior
PSRF	Gelman-Rubin potential scale reduction factor
RMSE	root mean squared error
SE	squared exponential

Acknowledgements

This material was based upon work indirectly supported by the National Science Foundation under Grant DMS-1638521 awarded to the Statistical and Applied Mathematical Sciences Institute (SAMSI). My thanks go out to members of the SAMSI extremes working group (2017-2018), who introduced me to many concepts from extreme value analysis and seeded the idea for this dissertation.

I would like to extend my sincere thanks to Dr. Surya Tokdar, my advisor, who made it possible for me to continue my graduate studies even after I moved cross country, who expressed confidence in me when I doubted my own abilities, and who provided numerous valuable insights that have been incorporated into this dissertation. Thanks for caring not just about my research but about my personal success and happiness.

I acknowledge my committee members—Dr. Merlise Clyde, Dr. Alexander Volfovsky, and Dr. Daniel Cooley—for their support and thank them.

I cannot begin to express my thanks to my family, immediate and extended, who made this possible. Nora, thank you for your prayers, for your understanding when I have been unavailable, and for your praise over my "pretty" plots. Eric, thank you for all your support: your encouragement, your sacrifices to keep our family running, our late-night math discussions, and your careful editing.

And I would be remiss not to thank God, who provided the energy, the inspiration, and the reason for all of this.

1

Introduction

How high do we need to build this levee to withstand flooding over its 50-year lifetime? How often do we expect to see an 8% drop of the S&P 500? What is the largest claim payout that we will have to make over the next 5 years? Each of these questions attempts to explain rare, extreme phenomena by characterizing how large measurements could be and how frequently they get that big. Being able to predict extreme rainfall, wind gusts, temperatures, wave heights, and traffic conditions is critical to those building infrastructure robust enough to withstand the extremes. Understanding large gains or drops in the stock market is important for those seeking to grow their portfolios and diversify their risk. With applications in environmental science, geology, finance, and insurance, to name a few, this study of rare-but-extreme events falls under the statistical subfield of extreme value analysis.

The set of extreme values that are either very large or very small in comparison to other realizations from the same process are sometimes called the "tails" or the "tail distribution." Values that are not large or small may be termed the "bulk" or "bulk distribution." These terms are adopted throughout this dissertation as general terms rather than as names of precise regimes. The term "light" is often used to describe

tail distributions with exponential decay rate (i.e. those that have a survival function that goes to zero at an exponential rate), while the term "heavy" is used to describe tails that decay at a rate slower than an exponential distribution. A few examples of heavy-tailed distributions are the Pareto, lognormal, and t distributions. This dissertation focuses on univariate density estimation of heavy-tailed distributions, specifically those with polynomial decay rates, with an aim to capture behavior in both the bulk and the tails simultaneously.

Chapter 1 provides some background from extreme value theory and puts forward a framework for heavy-tailed univariate density estimation. Subsequent chapters implement the framework using different modeling approaches.

1.1 Classic Extreme Value Theory

Models which do a good job of characterizing the bulk distribution may not do a good job explaining the tail values. When modeling an entire dataset, the relatively-abundant information in the bulk can overwhelm the relatively-sparse information in and about the tails. Because extreme values are often also rare values, it is possible that a particular sample contains very few large (or small) observations to provide information about behavior at the extreme.

When desired inference is only on the tails, it is common in practice to exclude all but the most extreme values from analysis. Leaving out the bulk of the distribution allows the tail to "speak" for itself. This exclusion is well-supported by the extremes literature. The Fisher-Tippett-Gnedenko theorem (Fisher and Tippett, 1928; Gnedenko, 1943) says that the limiting distribution of the maximum of a sequence of independent-and-identically-distributed variables, properly normalized, is the Generalized Extreme Value Distribution (GEVD), provided a limiting distribution exists.

The distribution function of the GEVD is

$$G(y) = P(Y \leq y) = \begin{cases} e^{-(1+\xi(\frac{y-\mu}{\sigma}))^{-1/\xi}} & \mu - \sigma/\xi \leq y < \infty, \qquad \xi > 0 \\ e^{-e^{-(\frac{y-\mu}{\sigma})}} & -\infty < y < \infty, \qquad \xi = 0 \\ e^{-(1+\xi(\frac{y-\mu}{\sigma}))^{-1/\xi}} & -\infty < y \leq \mu - \sigma/\xi, \quad \xi < 0 \end{cases}$$

In the GEVD, the location μ determines support and may need to be estimated, as the support of the maximum is not usually known ahead of time. The parameter $\sigma > 0$ controls scale, and ξ modifies shape. The $\xi > 0$ case is also known as the Fréchet Distribution; the $\xi = 0$ case is called a Gumbel Distribution; and the $\xi < 0$ case is a Reversed Weibull Distribution. Practitioners may use the GEVD to model a dataset of maxima, obtained by grouping observations into blocks (e.g. years or months) and retaining the maximum of each block.

Another important result from extreme value theory says that for most data, appropriately normalized, the excesses over a high threshold can be well approximated by the Generalized Pareto Distribution (GPD) (Pickands (1975) and Balkema and de Haan (1974) theorem). That is, for random variable Y and threshold u, the distribution function Y under truncation $F_u(y) = P(Y - u < y | Y > u) = \frac{F(u+y)-F(y)}{1-F(u)} \rightarrow G_{0,\sigma,\xi}(y),$ as $u \rightarrow \infty$, where $G_{\mu,\sigma,\xi}$ is the distribution function of the GPD:

$$G_{\mu,\sigma,\xi}(y) = P(Y \leq y) \begin{cases} 1 - (1 + \xi(\frac{y-\mu}{\sigma}))^{-1/\xi} & \mu < y, \qquad \xi > 0 \\ 1 - e^{-(\frac{y-\mu}{\sigma})} & \mu < y, \qquad \xi = 0 \\ 1 - (1 + \xi(\frac{y-\mu}{\sigma}))^{-1/\xi} & \mu < y < \mu - \sigma/\xi, \ \xi < 0 \end{cases}$$

In the GPD, threshold μ determines support, $\sigma > 0$ controls scale, and ξ modifies shape. When $\xi > 0$ the GPD has heavy tails with polynomial decay. When $\xi = 0$ the tails of the GPD are light, i.e. follow an exponential decay. Finally, with $\xi < 0$ a GPD exhibits bounded or short tails. This result allows practitioners to retain all data above some high threshold and model it with a GPD. This approach is sometimes called "peaks over thresholds" or POT.

1.2 Thresholding Methods

Despite the clean, theoretical, asymptotic results supporting truncation for extreme-value estimation, in practice it is often unclear where to threshold the tail. Truncation too far into the tail leaves little data for estimation and increases variance of estimates. Retaining too much of the data and applying a parametric model, on the other hand, can lead to biased parameter and quantile estimation. A review paper written by Scarrot and MacDonnald (2012) enumerates, classifies, and critiques many of the truncation methods proposed and used in extreme value theory, including literature around rule-of-thumb truncations and automated-truncation methods.

The Scarrot and MacDonnald (2012) paper also covers methods for density estimation which attempt to model the bulk and the tail simultaneously. Many of these existing approaches fall into the category of "mixture" models, meaning that they stitch together a truncated model for the bulk distribution with a separate (usually GPD) model for the tail distribution. Some models require continuity of the probability density function (pdf) at the stitch-point; others additionally require first-order smoothness at the join-point. Some models are fully parametric and others are nonparametric mixtures of these stitched distributions, giving more flexibility to an unknown bulk shape, as in the case of Carreau and Bengio (2009). Ultimately, these methods necessitate some sort of join-point estimation, whether implicitly or explicitly, at the point that the bulk distribution is stitched to the tail distribution. As there is no standard, accepted, best way to go about thresholding and this remains a difficult question, our goal is to skip the join-point estimation all together if possible. We would like to separate the estimation of the bulk from the estimation of the tail but without introducing an artificial boundary between the two.

Other full-distribution, data-retaining models have attempted to bypass the definition and estimation of a tail-threshold location. Frigessi et al. (2003) can be seen

as a continuous weighting between a Weibull bulk and a GPD tail. Naveau et al. (2016) employs a parametric transformation to a GPD, resulting in an "extended" generalized Pareto distribution with support over the same range as the data.

As the framework for tail estimation explored in this dissertation is itself a transformation method, a brief review of transformations in density estimation, including heavy-tailed density estimation, is warranted.

1.3 Transformations in Density Estimation

Transformations have long been used to improve univariate density estimation. Inducing some feature in the estimated distribution—such as reduced skewness, reduced kurtosis, or improved normality—was the aim of early work in transformations, such as that by Bartlett (1947) and Box and Cox (1964), and of much subsequent work. These features typify the bulk of the distribution, but transformations have also been appreciated for their ability to smooth estimates in the tail; see Silverman (1986), Wand et al. (1991), and Yang and Marron (1999) for early uses in kernel density estimation.

In a typical univariate transformation setup, data y_i come from an unknown density f, which takes the specific form $f(y) = G'_\theta(y)h(G_\theta(y))$. The transformation family $G_\theta(y)$ is indexed by θ, which may be a single scalar parameter or may be a vector of parameters. The family of densities h might be parametric or nonparametric. The support for h must be consistent with the range of G, but need not match the support of f. For example, the transformation of Clements et al. (2003) takes $G : \mathbb{R} \to [-1, 1]$.

Those transformations, $G_\theta(y)$, which are themselves cumulative distribution functions (CDFs) represent a smaller class of transformations with an interesting inter-

pretation. The model becomes

$$f(y) = g_\theta(y)h(G_\theta(y)), \tag{1.1}$$

and g_θ, the pdf corresponding to CDF G_θ, shares support with f. Density h has support $[0, 1]$. Thus f can be thought of as density g_θ that gets scaled by a perturbation factor h (see Verdinelli and Wasserman (1998) for nice exposition). The closer f is to the original g_θ the less perturbation is needed and the more closely h follows a uniform distribution. Such transformations are used in Brunk (1978), where θ is pre-estimated and fixed and h is estimated by orthogonal polynomials; in Verdinelli and Wasserman (1998), where h is estimated by Legendre polynomials simultaneously to the estimation of parameters θ from g_θ; and in Buch-Larsen et al. (2005), where h is estimated by kernel density estimators, to name a few.

1.4 Transformations in Heavy-Tailed Density Estimation

The transformation method has been applied to heavy-tailed data in kernel density estimation (Wand et al., 1991; Bolancé et al., 2003; Clements et al., 2003; Buch-Larsen et al., 2005; Markovich, 2007; Gustafsson et al., 2009; Bolancé et al., 2010). These transformations either raise all Y to a power, or they induce a power-law only for large values of Y. Typically the parameter that dictates the power transformation, an element of θ, is estimated first and then fixed in subsequent kernel density estimation of the transformed data. Markovich (2007) proposes estimation of this parameter by the Hill estimator, implying a tail-respecting transformation. Markovich also includes some discussion about how over- or under-estimating this parameter affects subsequent estimation of h, implying that overestimating the heaviness is preferred in kernel density estimation because the overestimation can be corrected by inducing $h(\tau)$ to zero in specific ways as $\tau \to 1$. What is sometimes overlooked is that a transformation can not only improve estimation in the tail but be used to

define estimation in the tail, provided appropriate constraints on h are used. The following sections put forward those constraints.

1.5 Transformations and Their Tail Properties

Consider a setting in which the data range is (a, b) with the potential for either a or b to be $\pm\infty$. Let $\{g_\theta : \theta \in \Theta\}$ be a parametric family of pdfs on (a, b) with parameters θ and distribution function G_θ.

Lemma 1. *For any density function f on (a, b) and any $\theta \in \Theta$, there exists a pdf h on $[0, 1]$ such that* $f(y) = g_\theta(y)h(G_\theta(y)), \qquad y \in (a, b).$

Proof. Let $Y \sim f$ and let h be the pdf for $U = G_\theta(Y)$. □

Let F be the CDF over (a, b) corresponding to f, and let H be the CDF over $[0, 1]$ corresponding to h. If we further restrict our attention to random variables Y that do not admit the possibility of a point mass and let F^{-1}, G_θ^{-1}, and H^{-1} be the quantile functions for f, g_θ, and h respectively, each defined over $[0, 1]$, then the following relationships are equivalent:

$$f(y) = g_\theta(y)h(G_\theta(y)) \iff F(y) = H(G_\theta(y)) \iff F^{-1}(p) = G_\theta^{-1}(H^{-1}(p)).$$

Additionally, the inverse probability integral transform says that for random variable $U \sim Unif(0, 1)$, $Y = G_\theta^{-1}(H^{-1}(U))$ is distributed with pdf f, providing a specific form for a data generating mechanism.

With the general form of the transformation defined, the tail properties under such transformations can be explored. Define $h_0 := \lim_{u\downarrow 0} h(u)$ and $h_1 := \lim_{u\uparrow 1} h(u)$. Then under the transformation $f(y) = g_\theta(y)h(G_\theta(y))$,

$$\lim_{y\downarrow a} \frac{f(y)}{g_\theta(y)} = \lim_{u\downarrow 0} h(u) = h_0 \text{ and } \lim_{y\uparrow b} \frac{f(y)}{g_\theta(y)} = \lim_{u\uparrow 1} h(u) = h_1. \tag{1.2}$$

Lemma 2. *In the case* $\lim_{y\to a} f(y) = \lim_{y\to a} g_\theta(y) \in \{0,\infty\}$*, Equation 1.2 implies that* $h_0 \in (0,\infty)$ *iff* $f(y) \asymp g_\theta(y)$ *as* $y \to a$*. Similarly, when* $\lim_{y\to b} f(y) = \lim_{y\to b} g_\theta(y) \in \{0,\infty\}$*, 1.2 implies* $h_1 \in (0,\infty)$ *iff* $f(y) \asymp g_\theta(y)$ *as* $y \to b$*.*

Proof. L'Hopital's rule applies here, giving $\lim_{y\to a} \frac{f(y)}{g_\theta(y)} = \lim_{y\to a} \frac{f'(y)}{g'_\theta(y)} = h_0$. Thus, $f(y) \asymp g_\theta(y)$ as $y \to a$. A similar argument holds as $y \to b$. □

In words, the consequence of positive finite h_0 and h_1 is that g_θ and f approach their respective boundaries at identical rates.

The tail-matching property of Lemma 2 also implies the following weaker result:

Lemma 3. *For tails in which* g_θ *is regularly varying, if* $h_0, h_1 \in (0,\infty)$*, then the corresponding tail indices of* f *exactly equal those of* g_θ*.*

Proof. For exposition, suppose $y \in [0,\infty)$. A regularly varying function g_θ can be written as $g_\theta(y) = y^\alpha L(y)$ for some slow-varying function L and tail index $-\alpha$. A function L is slow-varying if $\lim_{y\to\infty} \frac{L(ty)}{L(y)} = 1$ for all $t > 0$. A general reference is Resnick (2007). Applying the transformation, the regular variation of g_θ implies that

$$f(y) = g_\theta(y)h(G_\theta(y)) = y^\alpha L(y)h(G_\theta(y)).$$

Therefore a necessary condition for g_θ and f to have equal tail index $-\alpha$ while $h_1 \in (0,\infty)$ is for $L(y)h(G_\theta(y))$ to be slowly-varying also, which it is:

$$\begin{aligned}\lim_{y\to\infty} \frac{L(ty)h(G_\theta(ty))}{L(y)h(G_\theta(y))} &= \lim_{y\to\infty} \frac{L(ty)}{L(y)} \lim_{y\to\infty} \frac{h(G_\theta(ty))}{h(G_\theta(y))} \\ &= \lim_{y\to\infty} \frac{h(G_\theta(ty))}{h(G_\theta(y))} = \frac{\lim_{y\to\infty} h(G_\theta(ty))}{\lim_{y\to\infty} h(G_\theta(y))} = \frac{h_1}{h_1} = 1. \quad (1.3)\end{aligned}$$

□

The takeaway is that tails of f can be constrained to follow the asymptotic decay pattern of parametric g_θ by ensuring that h_0 and h_1 are positive finite. We aim to

improve estimation of the tail, where little data are naturally available, by leveraging this constraint. Namely, we incorporate prior information about the tail shape via selection of g_θ and tail heaviness via a prior over the tail index parameter (i.e. the inverse polynomial decay rate).

The general approach taken in this dissertation is to first select a parametric family of distributions $\{g_\theta : \theta \in \Theta\}$. Then select a nonparametric family of distributions $\mathcal{H} := \{h(\cdot)\}$, constrained such that h are density functions on [0,1] with $||\log h||_\infty \ll \infty$. Having both of these, set $\mathcal{F} := \{f(\cdot) = g_\theta(\cdot)h(G_\theta(\cdot))\}$ for $\theta \in \Theta$ and $h \in \mathcal{H}$. Finally, model $Y_1, Y_2, ... \overset{iid}{\sim} f$ from $f \in \mathcal{F}$ using a fully Bayesian approach. That is, set priors on θ and h and estimate them simultaneously.

While the theory indicates that the tail decay rate is identifiable under the constraint, we hope to understand if this approach is practically feasible. Is it possible for a model to separate the bulk and tail estimation sufficiently to prevent the bulk from overwhelming the tail? Can we learn the true tail index? Do the estimates provide improved tail prediction and extrapolation compared to peaks-over-threshold approaches? How much information is in the likelihood? And what is the influence of the prior? The models developed and the numerical studies presented in this work are geared towards answering these practical questions.

The following three chapters detail three different Bayesian nonparametric approaches to modeling the framework: Chapter 2 models the framework using a logistic Gaussian process; Chapter 3 implements it using a mixture model; and Chapter 4 introduces an approximation to the mixture model using predictive recursion. Each chapter provides background for the specific nonparametric approach, details for the model, summaries of validating simulations, and discussions around the strengths and weaknesses of the approach. Chapter 5 illustrates the three models on a set of daily rainfall data. Chapter 6 summarizes the results as a whole, comparing between

the approaches and synthesizing answers to the above questions. It also describes future work and directions.

2

Logistic Gaussian Process Model

This chapter explores one approach to modeling the transformation framework, described in Chapter 1, for univariate density estimation in the presence of heavy tails. The framework is composed of two pieces: a parametric family of distributions $\{g_\theta : \theta \in \Theta\}$ and also a nonparametric family of distributions $\mathcal{H} := \{h(\cdot)\}$, which have been constrained such that h are density functions on $[0, 1]$ with $||\log h||_\infty \ll \infty$. It brings together both pieces through $\mathcal{F} := \{f(\cdot) = g_\theta(\cdot)h(G_\theta(\cdot))\}$ for $\theta \in \Theta$ and $h \in \mathcal{H}$. In this chapter, h is formulated as a Logistic Gaussian Process (LGP). Modeling of data $Y_1, Y_2, ... \overset{iid}{\sim} f$ from $f \in \mathcal{F}$ is performed under a fully Bayesian approach. That is, priors are set on θ and h, which are estimated simultaneously.

First, we describe the specific approach we take to modeling the pdf, including selection of g and priors for θ and h. Second, we give an overview of computational setup, including approximations used to reduce computing time. Third, we explore the realities of using the method under non-asymptotic, real-life sample sizes by presenting the results of a simulation study, comparing the LGP model to existing methods, and highlighting its strengths and weaknesses. We conclude by summariz-

ing results and future directions.

2.1 Model Setup

2.1.1 Parametric Distribution and Priors

Under the transformation framework Equation 1.1, any family of parametric distributions $\{g_\theta : \theta \in \Theta\}$ may be selected that matches the support of Y. However, in order to model phenomena with heavy tails, the parametric family needs to admit possibility of heavy tails (see Yang and Tokdar, 2017).

For Y with support $(-\infty, \infty)$, a possible choice would be to model g_θ as a t-distribution with ν degrees of freedom. The parameter ν is equal to the polynomial decay rate in the tails. Choosing this g_θ would constrain both tails to follow a single decay rate; whereas, choosing an asymmetric t-distribution, such as that of Jones and Faddy (2003) or Zhu and Galbraith (2010), would allow for the possibility of different lower and upper tail indices.

For Y with half-real support $(0, \infty)$, the parametric family g_θ could contain distributions such as the half-t with ν degrees of freedom or the GPD with tail index $\xi = 1/\nu > 0$ and location parameter fixed at 0. Whatever distribution is selected should reflect an *a priori* belief about the shape and behavior of the tails. Other support of the form (a, ∞) or $(-\infty, b)$ where a and b are known can be modeled by an appropriate shift and/or reflection of g_θ defined on the positive reals.

For each parametric g_θ above, the parameter $\sigma \in \theta$ may be included to perform a scaling role. For distributions with support on the full real-line, including location parameter $\gamma_0 \in \theta$ allows for a shift in g_θ.

For Y with bounded support (a, b), $Beta(\nu_1, \nu_2)$ can be used for the base parametric distribution g_θ. Note that boundaries a and b must be fixed and known, as the methods herein have not yet been extended to the case where support is unknown and needs to be estimated, such as in the case of GPD with negative tail parameter.

Table 2.1: Possible choices for the underlying parametric distribution g_θ

Support	g_θ	$\boldsymbol{\theta}$	Tail index prior	Other prior
$(0,\infty)$	Half-t	(ν,σ^2)	$\log((\nu-0.5)/5.5) \sim \text{Logis}(0, 1/2)$	$\pi(2\log(\sigma^2)) \propto \frac{1}{2\log(\sigma^2)}$
	GPD	$(1/\nu,\sigma^2)$	$\log((\nu-0.5)/5.5) \sim \text{Logis}(0, 1/2)$	$\pi(2\log(\sigma^2)) \propto \frac{1}{2\log(\sigma^2)}$

For the remainder of this chapter, we focus on data distributed over the positive reals. Table 2.1 shows the parametric distributions, g_θ, used in this chapter, along with the independent priors that compose $\pi(\theta)$. The prior for parametric θ is assumed to be independent of nonparametric h.

2.1.2 Nonparametric Distribution and Priors

There are various ways to obtain nonparametric estimates for density h subject to the appropriate limit constraints. We pursue the logistic form

$$h(\tau) = \frac{e^{w(\tau)}}{\int_0^1 e^{w(t)}dt} \tag{2.1}$$

where $w \in C[0,1]$, the set of continuous functions on $[0,1]$. Modeling $w \sim GP(0,c)$ as a zero-mean, stationary Gaussian process induces a LGP prior on h. For density estimation using LGPs see Leonard (1978); Lenk (1988) and Lenk (2003); Tokdar (2007); Tokdar and Ghosh (2007); van der Vaart and van Zanten (2008) and van der Vaart and van Zanten (2009).

Because $w(\tau)$, the GP curve upon which the likelihood depends, is not handled well by MCMC, an approximating model similar to that used in Tokdar (2007) is employed. We replace $w(\tau)$ by a finite-rank approximating GP $\tilde{w}(\tau)$ and thereafter run the MCMC on the approximating model. This is arguably one of the more scalable approaches for handling of the LGP prior.

Specifically, we let $w \sim GP(m,c)$ denote a Gaussian process with mean functional m and covariance function c. Then for any finite set of points $\{\tau_1, \tau_2, ..., \tau_L\}$,

$w(\tau_1, \tau_2, ..., \tau_L)$ has a finite, L-variate normal distribution with mean $(m(\tau_1), m(\tau_2), ..., m(\tau_L))$ and $L \times L$ covariance matrix with elements $c(\tau_i, \tau_j)$. By fixing a dense grid of L values $(\tau_1, \tau_2, ..., \tau_L)$ and evaluating the function between grid points with a linear interpolator, the curve w is reduced to a parameter of length L.

The GP prior specification can be expressed in the following hierarchical form:

$$w \sim GP(0, \kappa^2 c^{SE}(\cdot, \cdot|\lambda))$$

$$(\kappa^2, \lambda) \sim \pi_\kappa(\kappa^2)\pi_\lambda(\lambda)$$

where $c^{SE}(\tau, \tau'|\lambda) = \exp\{-\lambda^2(\tau - \tau')^2\}$ is the squared-exponential covariance function. The parameter λ acts as a bandwidth parameter, enabling more wavy paths to be realized as λ gets larger.

We specify a prior for λ by reparameterizing into correlation $\rho_\ell(\lambda) = \exp(-\ell^2\lambda^2)$ and letting $\rho_\ell(\lambda) \sim \text{Beta}(a_\lambda, b_\lambda)$. Setting $\ell = 0.1$, $a_\lambda = 6$, $b_\lambda = 4$ places 95% of prior mass of $\rho_\ell(\lambda) \in (0.3, 0.86)$. The prior for κ^2 is set to be an $\text{IG}(a_\kappa, b_\kappa)$, which allows κ to be integrated out of the prior at run time. Hyper-parameters $a_\kappa = b_\kappa = 3/2$ ensure that the resultant t-process has three degrees of freedom, prior expectation equal to 1, and finite variance, i.e. a reasonably diffuse prior on w.

2.2 Computational Considerations

2.2.1 Posterior Approximations

Even with a discretization of the likelihood to L locations, the use of a Gaussian process still requires inversion of an $L \times L$ matrix, a theoretically-feasible but computationally-expensive undertaking, which also produces slow-mixing chains. Rather we replace the L-rank approximation with an M-rank interpolating, predictive process, where M is much smaller than L. That is, for fixed knots $\{t_1, t_2, ..., t_M\}$ we replace each $w(\tau)$ by $\tilde{w}(\tau) := E(w(\tau)|(w(t_1), w(t_2), ..., w(t_M))$. For more on the use of low-rank predictive processes see Tokdar (2007) and Banerjee et al. (2008).

A second approximation, made by substituting $\pi_\lambda(\lambda)$ with a dense, discretized approximation over $\rho_{0.1}(\lambda)$, allows the GP covariance matrices to be pre-computed, significantly speeding up algorithm run time. Support points $\{\lambda_1, \lambda_2, ..., \lambda_G\}$ are selected to be more densely packed for smaller values of λ using the algorithm in Section 3.2 of Yang and Tokdar (2017).

Regarding the M-rank predictive process, we find that a grid of equally spaced knots is effective and computationally expedient, provided a sufficient number of knots is used. Using too few knots can introduce weak multimodality into the posterior, an artifact of poor estimation of the nonparametric density, $h(\tau)$, between sparsely placed knots. For most applications, eleven or more knots are needed to avoid this artifact.

For additional details on likelihood, priors, and approximations see Appendix A.

2.2.2 MCMC Sampler

The above approximations reduce the number of parameters in the model to $M + 2$: M parameterizing h and two for the parameters of θ. Markov chain samplers are used to obtain draws from the posterior, and Monte Carlo approximation is used to estimate posterior quantities of interest. Specifically, a blocked, random-walk Metropolis sampler is used on a transformed parameter space such that multivariate normal proposals can be used. Candidate proposal covariances are slowly adapted to achieve a 15% acceptance rate using Algorithm 4 of Andrieu and Thoms (2008). Results in this chapter were achieved by using one block containing the M knot parameters, one block updating $\theta = (\nu, \sigma)$, and one block updating all $M+2$ parameters simultaneously.

In practice, even with a small number of knots, (e.g. $M = 11$ or $M = 21$), the MCMC sampler has difficulty moving around the posterior. We find that multi-chain MCMC samplers are more likely to "converge" when the parametric base

distribution is additionally scaled by a large-α-level quantile. It is uncertain exactly why this happens; however, it does appear that the quantile scaling performs some orthogonalizing role, at least in the range of values being sampled. We scale our transformation distribution by the 0.9-level quantile under unit scale and unknown shape parameter ξ for the GPD or ν for the half-t distribution.

All computing was performed in R (R Core Team, 2018) with the majority of the sampling work being performed with calls to C.

2.3 Simulation Study

We present the results of a simulation study to compare the LGP model to a few truncated maximum-likelihood methods. Specifically we compare LGP to the maximum likelihood estimates (MLEs) of 1) a generalized Pareto distribution (GPD) under various thresholds, fit using the `extRemes` package (Gilleland and Katz, 2016) in R; and 2) the extended generalized Pareto distribution (EGPD) proposed by Papastathopoulos and Tawn (2013) under various thresholds, fit using the `mev` R package (Belzile et al., 2018). Thresholding proportions for the GPD and EGPD methods are selected to cover a practical range of thresholds one might pick based on visually-diagnosed cutoffs. When referring to the method used, plots and tables also refer to the percentage of data retained in truncation.

2.3.1 Simulation Setup

Six positive-real, heavy-tailed-density scenarios are considered:

1. *Standard GPD.* Data are generated from a GPD with unit scale and tail index $\xi = 0.25$.

2. *Half-t.* Data are generated from a half-t distribution with unit scale and tail parameter, $\nu = 4$.

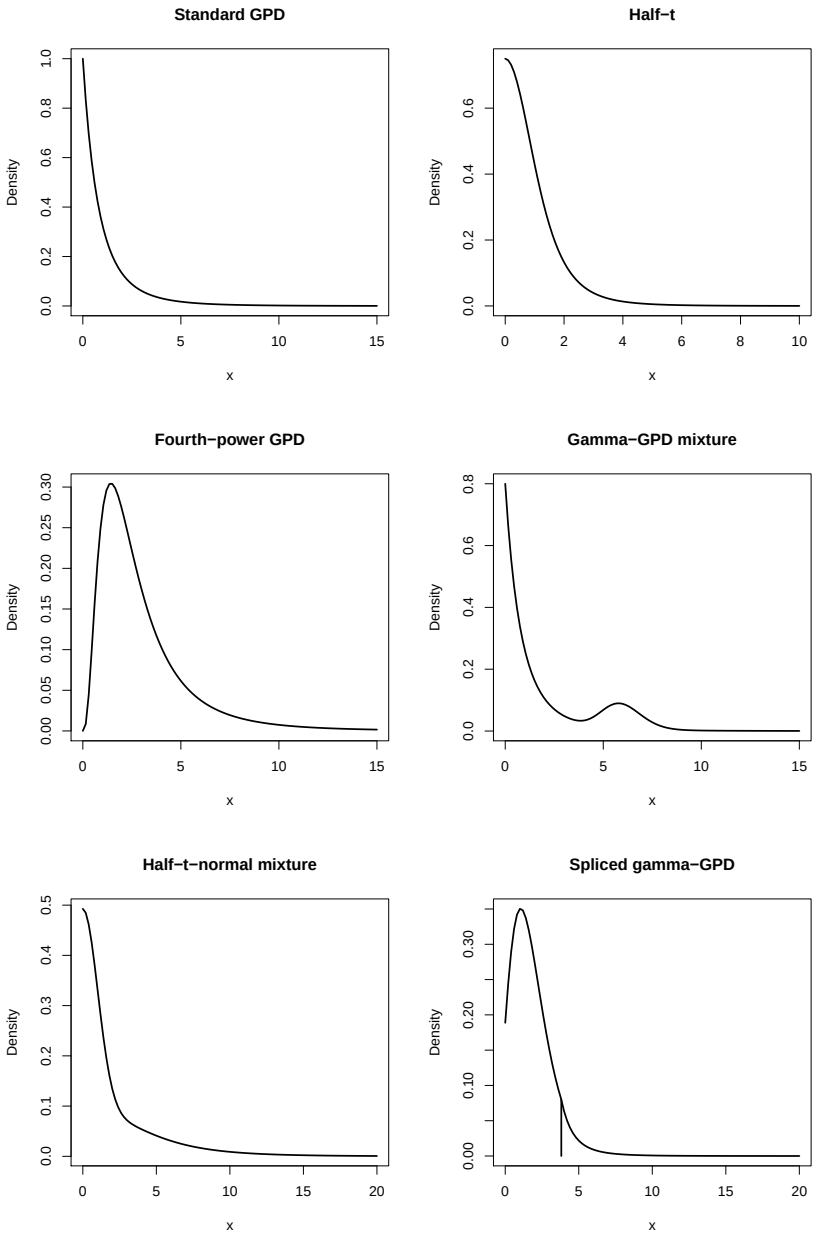


FIGURE 2.1: True densities from which data are simulated. Each has tail index $\xi = 1/4$. For the spliced gamma-GPD, the black vertical line denotes the location where the GPD tail meets the truncated gamma bulk.

Table 2.2: Simulation scenario sample size, number of LGP knots used, and summaries related to MCMC chains: number of iterations run in total, number discarded as burn in, number retained during thinning, percentage of simulations converging as determined by PSRFs, and time to run in minutes per 100,000 iterations using the GPD base and all data, i.e. LGP-GP 100%.

Simulation Scenario	n	knots	Iteration	Burn	Retain	Converge	Time
Standard GPD	1,000	11	600k	500k	1k	97%	2.2
Half-t	1,000	11	600k	500k	1k	92%	1.8
Fourth-power GPD	1,000	11	600k	500k	1k	89%	2.3
Gamma-GPD mixture	1,000	21	800k	500k	2k	82%	2.1
Half-t-normal mixture	2,000	21	600k	500k	1k	100%	3.6
Spliced gamma-GPD	5,000	21	800k	500k	3k	88%	8.2

3. *Fourth-power GPD.* Data are generated via the probability integral transform by raising uniform draws, U, to the fourth power (equivalent to transforming them by the quantile function of a Beta$(4, 1)$ distribution), then transforming those by the quantile function of a standard GPD with tail index $\xi = 0.25$, i.e. $Y = qGPD(qBeta(U; 4, 1); 0.25)$.

4. *Gamma-GPD mixture.* Data come from a mixture of gamma and GPD distributions: 80% of observations come from a standard GPD with tail index 0.25 and 20% come from a Gamma$(36, 6)$ distribution (mean 6, variance 1). This density is multimodal.

5. *Half-t-normal mixture.* Data come from a mixture of half-normal and half-t distributions: 50% of observations come from a half standard normal and 50% come from a half t distribution with $\nu = 4$ and standard deviation $\sigma = 4$.

6. *Spliced gamma-GPD.* Data are generated from a gamma distribution truncated on both the left and right, making up the bulk, "spliced on" to a GPD tail. The bulk distribution is truncated at its ninety-second quantile, $Q_{0.92} = 3.81331$, so that a fixed 8% of the distribution comes from the standard GPD$(\xi = 0.25)$ tail.

The scale parameter of the truncated-gamma is selected to ensure continuity in the density at the join-point for the two distributions.

All six densities have asymptotic tail index $\xi = 1/4$, and a representation of each density is plotted in Figure 2.1. The first four scenarios use sample size $n = 1{,}000$, the fifth uses $n = 2{,}000$, and the last uses $n = 5{,}000$ (see Table 2.2). One hundred independently generated datasets are used to calculate the comparison quantities of interest.

The methods are compared on the basis of their tail-index bias and interval coverage, as well as on their estimated, upper-quantile bias and root mean squared error (RMSE). For the sample sizes considered in simulation, we would expect to see, on average, a single observation above the quantile level $p = 0.999$ when $n = 1{,}000$; two observations for a dataset of $n = 2{,}000$; and five observations for a dataset of $n = 5{,}000$. With this in mind, quantile bias and RMSE are considered for $0.9 < p < 0.9999$, which includes both quantile levels where we expect to have seen observations (roughly $p < 0.999$) and levels where there are likely no observations ($p > 0.999$), a moderate to extreme extrapolation. For methods employing truncation, bias and RMSE are only displayed for levels p that were retained after thresholding.

In the LGP method, the first three simulation scenarios use eleven evenly spaced knots. This results in thirteen total parameters: eleven for knots capturing the nonparametric scaling density, h, and two for the parameters of the transformation distribution, ν and σ. The last three simulation scenarios use 21 evenly-spaced knots to capture slightly more difficult or subtle nonparametric densities, resulting in 23 total model parameters. For each scenario a GPD density is utilized for the base or transformation distribution, g_θ. These are labeled with the designation LGP-GP and the percentage of data used in estimation. The half-t-normal mixture scenario is additionally performed with a half-t density being used for g_θ in order to evaluate

the sensitivity of LGP to the choice of base distribution. These are designated with the label LGP-HT. For each simulated dataset, three chains with random "warm" starts are run. Their covariance matrices are also given "warm" starts: random MCMC draws from an adaptively-learned covariance matrix over a single, long-run chain, which have been scale-inflated. Convergence of the MCMC chains is assessed for each posterior separately by considering Gelman-Rubin potential scale reduction factors (PSRFs). When the upper limits for *all* parameters' univariate PSRFs are less than 1.1, we consider that posterior MCMC chain to have converged and include the results from that dataset in the aggregated results.

2.3.2 General Observations

Before delving into the results of each simulation scenario, we make some observations about the simulations as a whole.

Bulk Estimation

While the primary focus of comparison for these simulations is on the tails, we also mention that the LGP-GP 100% has low bias in the bulk of the estimated density. Simulations were essentially unbiased for the median—largest relative bias, i.e. (estimate - truth)/truth among the six simulations was 0.0074—and the largest biases for $p < 0.9$ consistently occurred at or near the lower "tail." The fourth-power GPD had the poorest relative bias, under-estimating the 0.00001-level quantile by nearly 100%. However, this behavior is to be expected given that our model inherits its left-hand tail properties, including strictly positive values at zero, from the GPD base g_θ; whereas, the fourth-power GPD scenario has true density $f(0) = 0$. With positive density at zero, LGP-GP 100% is going to estimate its 0.00001-level quantile much earlier (at smaller values) than the true fourth-power GPD. Among the remaining five simulations, the worst left-hand tail relative bias was 0.079, coming

from the gamma-GPD mixture.

Algorithm Speed

Unsurprisingly, the computational speed of the Bayesian LGP models is nowhere close to the computational speed of the maximum likelihood methods, each of which takes just a fraction of a second to run. The maximum likelihood methods are advantaged 1) by having fewer likelihood evaluations over truncation-reduced samples and 2) by employing an optimization algorithm rather than an algorithm aimed at full characterization of the posterior distribution.

Table 2.2 includes the time in minutes per 100,000 iterations it takes to run each of the simulation scenarios using the LGP-GP 100%. Mostly dependent on the sample size but also somewhat dependent on the number of knots, each chain took between 13 and 66 minutes to run. The run times for LGP-HT 100%, not included in the table and only used for the half-t-normal mixture scenario, were significantly longer, coming in at 64 minutes per 100,000 iterations, a seventeen-fold increase over the GPD-base run times for the same scenario. These differences arise because the half-t base density depends on computationally-expensive evaluations of t density and quantile functions. Therefore, we find it computationally expedient to use a GPD base distribution when possible.

MCMC Mixing Issues

The adaptive MCMC sampler has difficulty mixing for many datasets and across scenarios, moving slowly through a complex posterior space. Figure 2.2 gives an idea of the potential difficulty, showing the posterior samples for datasets 4 and 53 of the spliced gamma-GPD scenario. While this scenario tends to have nonparametric densities, h, that are single-modal and follow a "low-high-low" pattern, the upper-right panel of the figure shows that even under a "low-high-low" regime the mode

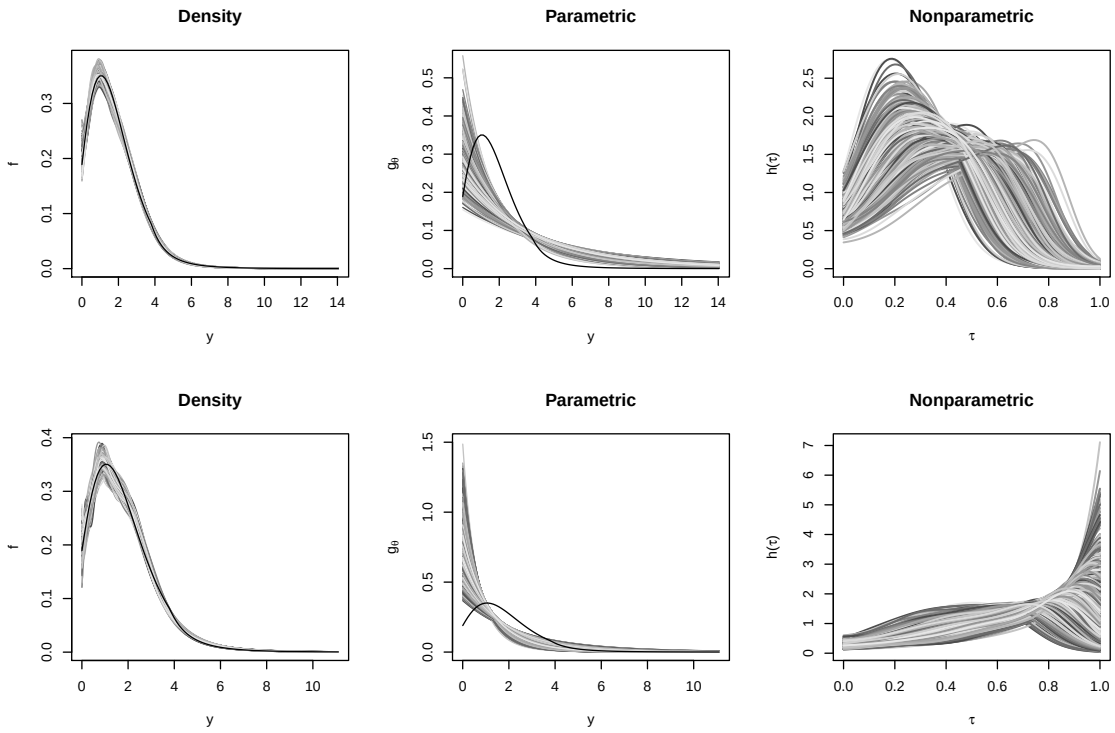


FIGURE 2.2: Spliced gamma-GPD posterior samples for dataset 4 (upper) and dataset 53 (lower). Left panels display sampled density estimates in variegated gray and truth in black. Middle panels show estimated parametric distributions in variegated gray and truth in black. Right panels show nonparametric scaling densities.

could be anywhere from $\tau = 0.2$ to $\tau = 0.8$. Modes near $\tau = 0.2$ correspond to large σ, and modes near $\tau = 0.8$ correspond to smaller σ estimates. Some chains for some datasets sample from part of the parameter space with even smaller σ, resulting in nonparametric densities closer to the "low-high" pattern shown in dataset 53. With σ this small, the needed nonparametric density h should have a rapid drop immediately before $\tau = 1$, but some samples fail to drop at all. This may be another example of multimodality being induced by too-few knots being placed at τ near 1. The covariance matrix between sampled parameters in the "low-high-low" pattern is different than the covariance matrix coming from a "low-high" pattern, making it difficult and slow for the adaptive sampler to move between these differing regimes.

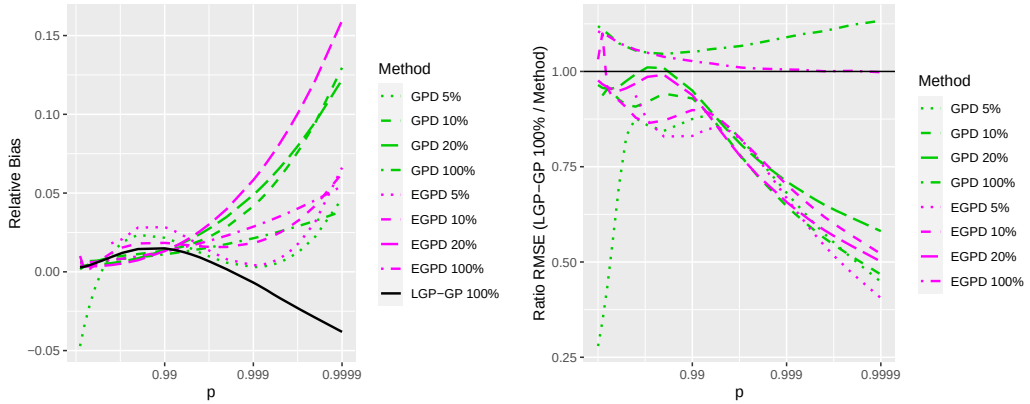


FIGURE 2.3: Standard GPD upper-tail, quantile-extrapolation with p on a log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of LGP-GP 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than LGP-GP 100%.

Whether a specific dataset *should* fall into one of these regimes or whether the sampler *should* be traversing between them is hard to know. Ultimately, we rely on long chains with long burn-in, combined with a strict triple-chain PSRF convergence criterion, to determine (hopefully) that we have faithfully sampled the posterior for each dataset. Convergence rates for each scenario are shown in Table 2.2.

2.3.3 Simulation Results

The results for tail-index estimation across scenarios and methods are included in Table 2.3. Across simulation scenarios, the LGP-GP 100% underestimates the true heaviness of the tail index parameter. The following subsections provide detailed results by scenario.

Standard GPD Scenario

Averaged across datasets, LGP-GP 100% estimated tail index $\hat{\xi} = 0.21$, meaning that it slightly underestimates the tail index (see Table 2.3); however, it maintains 99% interval coverage. Figure 2.3, which displays upper quantile levels on a log scale, shows that the LGP-GP 100% tail exhibits similar levels of quantile bias to

Table 2.3: Simulation tail-index results. Tables include mean tail-index estimates for each method and coverage of 95% confidence or credible intervals, i.e. proportion of intervals across simulated datasets that contain the true tail index.

Standard GPD

Method	Mean	Cover
Truth	**0.25**	
GPD 5%	0.19	0.94
GPD 10%	0.25	0.95
GPD 20%	0.26	0.97
GPD 100%	0.25	0.93
EGPD 5%	0.17	0.94
EGPD 10%	0.22	0.95
EGPD 20%	0.26	0.93
EGPD 100%	0.25	0.94
LGP-GP 100%	0.21	0.99

Half-t

Method	Mean	Cover
Truth	**0.25**	
GPD 5%	0.18	0.92
GPD 10%	0.18	0.91
GPD 20%	0.16	0.79
EGPD 5%	0.19	0.93
EGPD 10%	0.20	0.88
EGPD 20%	0.17	0.86
LGP-GP 100%	0.12	0.42

Fourth-power GPD

Method	Mean	Cover
Truth	**0.25**	
GPD 5%	0.18	0.88
GPD 10%	0.21	0.92
GPD 20%	0.21	0.90
EGPD 5%	0.19	0.88
EGPD 10%	0.20	0.89
EGPD 20%	0.22	0.89
LGP-GP 100%	0.20	0.93

Gamma-GPD mixture

Method	Mean	Cover
Truth	**0.25**	
GPD 5%	0.43	0.92
GPD 10%	0.25	0.87
GPD 20%	0.01	0.07
EGPD 5%	0.50	0.85
EGPD 10%	0.32	0.90
EGPD 20%	0.11	0.48
LGP-GP 100%	0.10	0.87

Half-t-normal mixture

Method	Mean	Cover
Truth	**0.25**	
GPD 5%	0.18	0.87
GPD 10%	0.17	0.80
GPD 20%	0.12	0.38
EGPD 5%	0.17	0.89
EGPD 10%	0.19	0.90
EGPD 20%	0.14	0.63
LGP-GP 100%	0.14	0.35
LGP-HT 100%	0.26	0.94

Spliced Gamma-GPD

Method	Mean	Cover
Truth	**0.25**	
GPD 5%	0.23	0.94
GPD 8%	0.24	0.95
GPD 13%	0.21	0.85
EGPD 5%	0.22	0.93
EGPD 8%	0.24	0.94
EGPD 13%	0.24	0.96
LGP-GP 100%	0.10	0.76

FIGURE 2.4: Half-t upper-tail, quantile-extrapolation with p on a log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of LGP-GP 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than LGP-GP 100%.

that of the GPD and EGPD methods. Note that GPD and EGPD are both unbiased for the standard GPD, so any biases seen in these graphs are due to small sample estimation over only 100 datasets. The ratio of RMSE (LGP-GP 100% divided by alternate method) shown in Figure 2.3 is also consistently below 1 when comparing the LGP-GP 100% to any of the truncation methods (GPD or EGPD, 5%, 10%, 20%), indicating that estimation with LGP on all data is better than the maximum likelihood methods under thresholding. Both the GPD 100% and the EGPD 100% have superior tail RMSE to the LGP-GP 100%, but that is not surprising, since they are specifically designed to estimate this scenario. Given that the LGP-GP 100% is not grossly inferior to the GPD 100%, i.e. RMSE ratios are not far above 1, the LGP is doing a reasonable job of capturing the standard GPD scenario.

Half-t Scenario

Table 2.3 shows that all methods underestimate the half-t tail-index parameter; however, at 0.12 the LGP-GP 100% has the most biased index as well as the poorest coverage (only 42%). Because the t-tail is not an exact GPD but converges to one

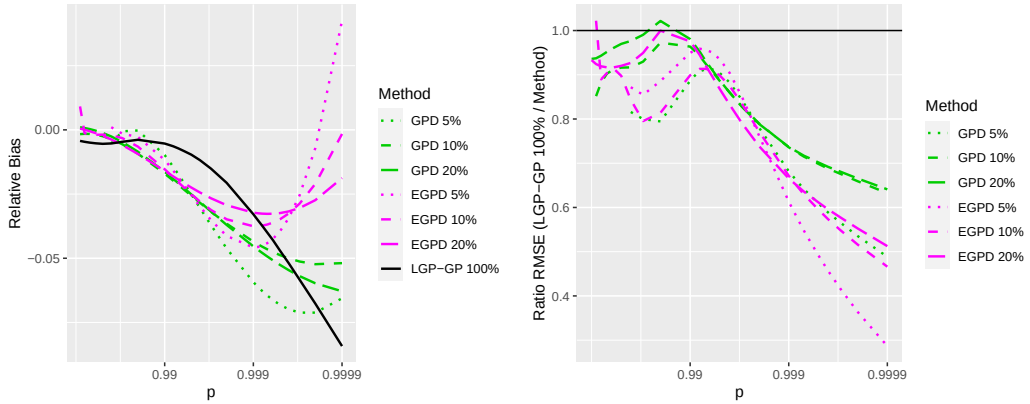


FIGURE 2.5: Fourth-power GPD upper-tail, quantile-extrapolation with p on a log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of LGP-GP 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than LGP-GP 100%.

as the threshold goes to infinity, the GPD methods tend to get better tail-index coverage when less data are retained. For all methods, the underestimation of the tail index results in underestimates of the tail quantiles; the GPD method bias is to be expected since the true tail comes from a t-distribution. The LGP-GP 100% has higher bias in the early tail range (quantile levels $0.9 < p < 0.995$) than do the other methods. Except for in a small region near $p = 0.99$ where the GPD 10% and GPD 20%-truncations have lower RMSE, the LGP-GP 100% has higher relative efficiency than all GPD methods considered. The gains are most apparent for $p > 0.999$, where data are scarce to non-existent. See Figure 2.4 for quantile bias and RMSE.

Fourth-power GPD Scenario

Table 2.3 shows that all methods are underestimating the fourth-power GPD tail-index parameter, but not grossly; LGP-GP 100% has similar bias and coverage to the other methods. For all methods, the underestimation of the tail index results in slight underestimates of the tail quantiles (see Figure 2.5); however, the LGP-GP 100% has lower bias in the early tail range (quantile levels $0.9 < p < 0.995$) than

FIGURE 2.6: Gamma-GPD mixture upper-tail, quantile-extrapolation with p on a log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of LGP-GP 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than LGP-GP 100%.

do the other methods. It also has lower tail quantile RMSE than any of the ML methods for most quantile levels $p > 0.9$. In the far-extrapolated tails, quantile RMSE is reduced by nearly half.

Gamma-GPD Mixture Scenario

All of the methods had a difficult time estimating the tail index in this tricky case (see Table 2.3). LGP-GP 100% maintained 87% coverage despite grossly underestimating the index. The LGP-GP 100% quantile estimates have similar magnitudes of bias to the ML methods, although not necessarily in the same direction. The RMSE results are mixed, with some methods doing better for some quantile levels and others doing well for other quantile levels (see Figure 2.6).

Half-T-Normal Scenario

When using the GPD base distribution, the results of this scenario are similar to those of the half-t scenario, including underestimation of the t tail-index parameter (see Table 2.3) and superiority of the LGP-GP 100% in quantile estimation RMSE despite tail quantile bias (see Figure 2.7). In this section we focus on LGP estimation

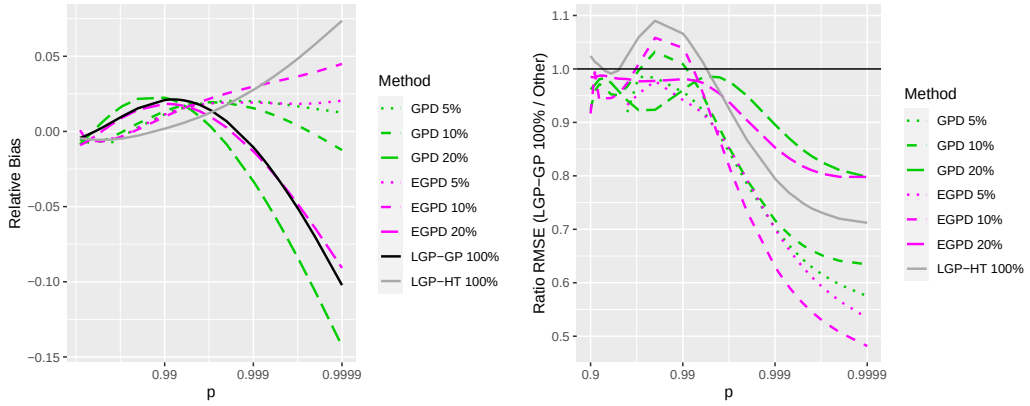


FIGURE 2.7: Half-t-normal mixture upper-tail, quantile-extrapolation with p on a log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of LGP-GP 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than LGP-GP 100%.

with a half-t distribution being used as the base or transformation density. Results for the LGP-HT 100% are included in the aforementioned table and figure.

The LGP-HT 100% has a near-unbiased estimate $\hat{\xi} = 0.26$ of tail index and also attains the best tail-index interval coverage of all methods considered with 94%. Additionally, it has the smallest quantile bias of all methods for large-but-not extreme quantile levels $p < 0.99$.

In comparing between LGP methods, the LGP-HT 100% has lowest RMSE for large quantiles that are still not particularly extreme (e.g. $0.9 < p < 0.995$) with ratio RMSE LGP-GP 100% / RMSE LGP-HT 100% maxing out near $p = 0.98$ at a value of 1.09. Moving into the extrapolated tail, however, LGP-GP 100% has lower RMSE than the LGP-HT 100%. With the target distribution having a t-like tail, this result was surprising. Further exploration shows that the estimated nonparametric densities h for the LGP-GP 100% have similar amounts of variability across replicates to the LGP-HT 100%. If anything, the LGP-HT 100% has slightly smaller variability in its nonparametric estimates into the upper tail. This implies that the LGP-GP 100% picks up its overall, quantile-RMSE advantage in the extreme

tails when it estimates the parametric g_θ. This was also surprising, since the LGP-HT 100% seemed to do so well at estimating the tail-index parameter. Although it is less biased than other methods, its variance is not sufficiently low to overcome the inherent quantile-RMSE advantage a method has when it underestimates a power-law tail index (e.g. a method which underestimates the true tail-index by 0.01 will have extreme quantiles with lower RMSE than does a method that overestimates the tail-index by the same amount). In fact, the LGP-HT 100% has a standard deviation of 0.047 among all simulated tail-index estimates while the LGP-GP 100% has a standard deviation of 0.033. Perhaps if the MCMC chains were allowed to run longer, the variance of the LGP-HT 100% tail-index could be reduced sufficiently to favor the LGP-HT 100% in extreme tail estimation, but there is no guarantee of this.

For the bulk distribution (not pictured), the LGP methods are similar in RMSE except for $p < 0.25$, where the LGP-HT 100% has smaller RMSE. There the RMSE ratio maxes out near $p = 0.01$ with a value of 1.26. This seems to point towards an advantage in estimating the lower quantiles when the shape of the parametric base distribution more closely mirrors the shape of the target distribution.

All of this leads us to believe that LGP is agnostic towards the choice of base distribution when estimating mid-range or bulk quantiles but that it can be sensitive to the choice of base distribution when estimating extreme quantiles. Unfortunately, these results do not provide clear direction as to how to pick the appropriate base distribution for any given dataset.

Lastly, we compare the LGP methods to the threshold methods as a whole. If we take a pessimistic view of the LGP, we say that it cannot guarantee superior tail-quantile estimation over a well-truncated GPD, even when using a base distribution with "correct" tails, as in the case of a t-tail being estimated by a t base distribution. But in practice for any given sample, we never know the optimal place to threshold. Since the LGP methods circumvents truncation all together, it is reasonable to take

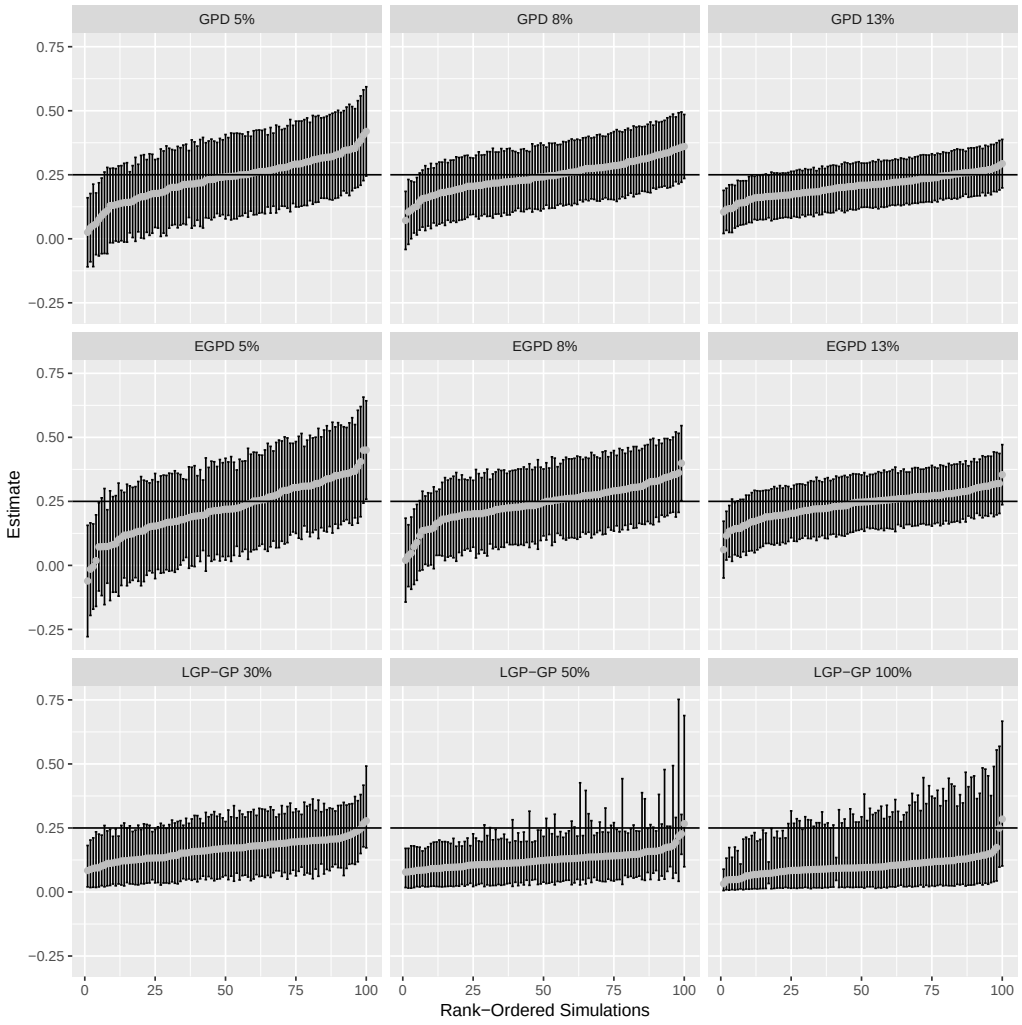


FIGURE 2.8: Spliced gamma-GPD tail-index intervals estimated under various methods and truncations. LGP-GP methods contain some non-converged simulations.

a more liberal view and recognize that both LGP methods have lower RMSE across all p in the tail than the worse-case-RMSE among the various threshold levels.

Spliced Gamma-GPD Scenario

Table 2.3 shows that, except for the GPD 13% method which has coverage of 85%, the maximum likelihood methods maintain unbiased and near-nominal 95% coverage for the tail-index parameter. At 76%, the LGP-GP 100% has the poorest tail-index

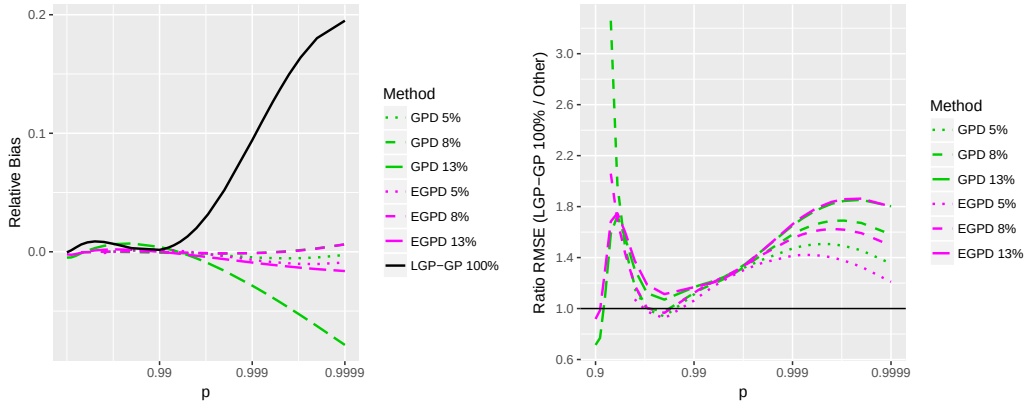


FIGURE 2.9: Spliced gamma-GPD upper-tail, quantile-extrapolation with p on a log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of LGP-GP 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than LGP-GP 100%.

interval coverage. Figure 2.8 plots interval estimates for this scenario under all of the estimation methods; the lower right panel of that figure shows that LGP-GP 100% consistently underestimates the heaviness of the true-GPD tail.

The tail quantile estimation plots in Figure 2.9 show that the GPD estimators that employ truncation at or after quantile-level $p = 0.92$ are unbiased into the tail, as expected, as are the EGPD 8% and EGPD 5% estimators. The GPD 13% reflects bias from a truncation that includes part of the gamma bulk. The EGPD 13%, on the other hand, is more robust to this truncation, retaining its low bias in the tails. The LGP-GP 100% estimator has the largest bias in the tail, overestimating 0.9999-level quantiles by nearly 20%.

RMSE is lower for all GPD methods in the tail than for the LGP-GP 100%, as shown in the right panel of Figure 2.9. While it is not surprising that the GPD ML methods do a better job of capturing an exactly-GPD-distributed tail than does the LGP, the magnitude of the differences is surprising. RMSE is 40-80% lower for the GPD and EGPD methods in the extrapolated tails than for the LGP-GP 100% method. At an oracle-truncation-level, the GPD 8% method has the lowest RMSE.

Except for right near $p = 0.92$, the EGPD estimator with generous data retention (EGPD 13%) has the lowest RMSE in the upper tail.

2.3.4 LGP as a Truncation Option

Despite having multiple chains converge to the same place, it is unclear whether the Markov chain samplers in many of the scenarios have explored the entirety of their posterior distributions. It is possible that some have learned the meaty bulk of the posterior but have failed to find a second mode or to round a difficult-to-traverse corner of the posterior, thereby failing to sample a small-but-relevant tail, rendering posterior inference biased.

The most difficult-to-learn posteriors seem to be those in which the target distribution departs most markedly from the base distribution. To explore this more fully, we revisited the spliced-gamma GPD scenario and truncated away part of the left-side bulk, leaving a greater proportion of the remaining data coming from the GPD tail. One truncation retained 50% of the data and a second set of truncations retained only 30%. We then ran the LGP method on the truncated data with a GPD base distribution. The LGP-GP with 50% truncation had even poorer convergence than the LGP-GP with 100% data retention; however, those that truncated at the 0.7-level quantile had markedly better convergence after 800,000 iterations than did the LGP-GP 100% with 96 of the 100 simulations showing adequate potential scale reduction factors.

Analyzing the converged chains, the LGP-GP 30% had 82% coverage of the tail-index parameter, better than the 76% coverage of the LGP-GP 100%. The bottom row of Figure 2.8 shows that this is due to the tail-index estimates and intervals being less biased. Additionally, the LGP-GP 30% had generally lower RMSE for $p > 0.9$ than did the LGP-GP 100%. While the GPD and EGPD methods still had lower RMSE estimating the GPD tail in pre-extrapolation regions, in extrapolation

(specifically $p \geq 0.997$) the LGP-GP 30% had the lowest RMSE of all methods and truncations considered. The GPD 13% and EGPD 13% had next lowest RMSE with ratios (LGP-GP 30% over method) near 0.95 for each. Though the reduction in RMSE is slight, any advantage in extrapolation is welcome. We also note that the LGP-GP 30% is beating the GPD and EGPD truncation methods in their optimal scenario, namely under an exact-GPD-tail case. This RMSE advantage cannot all be attributed to the difference in sample size. When 30% of the tail data is used for GPD or EGPD estimation, making it commensurate with the LGP-GP 30%, the GPD and EGPD quantile estimates become very biased, getting "contaminated" by the gamma bulk distribution, and thereby have even higher RMSE.

These results are promising. They suggest that if mixing issues crop up when used on all observations, the LGP may yet be used on truncated tail data to some advantage. The flexibility of the LGP can accommodate more generous data retention than the GPD or EGPD methods, resulting in quantile estimates that are closer on average to the truth as measured by RMSE. While the original goal was to move away from truncation altogether, we find that in difficult MCMC sampling scenarios truncation towards the tail may help after all.

2.4 Conclusions and Discussions

We have proposed a model that specifies a logistic Gaussian process form for the nonparametric density of our framework and accommodates various parametric transformation distributions. Bayesian priors have been proposed and a method for sampling from the posterior distribution has been implemented in R. While the theory promises a match of decay rates and tail-index parameters between the data likelihood f and the parametric distribution g_θ, simulations have not provided conclusive evidence that the logistic Gaussian process, as implemented, works in reasonable, finite sample sizes. Given the ambiguity that still remains, there are several directions that

future research could explore.

First, it is possible that longer MCMC chains are needed or that different MCMC samplers could be capable of traversing these complex posteriors better or more quickly than the adaptive MCMC. Once an adaptive MCMC sampler has learned a proposal covariance structure, it can have difficulty unlearning that structure, preventing it from rounding bends in "banana-shaped" posteriors. Other methods that might improve the MCMC mixing in this scenario include gradient-based samplers, such as the Metropolis-Adjusted-Langevian Algorithm (Roberts and Tweedie, 1996) or its successors, or some type of Hamiltonian Monte Carlo sampler (Duane et al., 1987).

Alternately, it may be that base distributions, g_θ, used in simulation were too different in shape from the target distribution. Preliminary results in Section 2.3.4 seem to support the idea that a base that more closely resembles the target distribution reduces the complexity of the posterior, improving the convergence of the adaptive MCMC sampler. If this avenue is pursued, the GPD could be replaced with some version of an EGPD, or it may be prudent to implement a base density that is capable of taking a value of zero at zero in instances where the target is suspected of having low probability for small values. Ultimately, we choose not to pursue this line of inquiry, since we want a method that is capable of estimating any sampling distribution, not just ones for which the bulk mimics some arbitrary transformation distribution.

Finally, it is possible that the nonparametric specification is simply not flexible enough to capture the necessary nonparametric density and that something different should be used altogether. One such approach might be to form the nonparametric density as a mixture of kernels with support $[0, 1]$. The next chapter will pursue this line of inquiry.

3

Mixture Model Approach

This chapter explores another approach to modeling the transformation framework, described in Chapter 1, for univariate density estimation in the presence of heavy tails. As a refresher, the framework requires that we select a parametric family of distributions $\{g_\theta : \theta \in \Theta\}$ and also a nonparametric family of distributions $\mathcal{H} := \{h(\cdot)\}$, which have been constrained such that h are density functions on $[0, 1]$ with $||\log h||_\infty \ll \infty$. Having both of these, we set $\mathcal{F} := \{f(\cdot) = g_\theta(\cdot)h(G_\theta(\cdot))\}$ for $\theta \in \Theta$ and $h \in \mathcal{H}$. Finally, we model $Y_1, Y_2, ... \overset{iid}{\sim} f, f \in \mathcal{F}$ using a fully Bayesian approach. That is, set priors on θ and h and estimate them simultaneously.

As mentioned before, there are various ways to obtain nonparametric estimates for $h(\tau)$ subject to the limit constraints $0 < h(0), h(1) < \infty$ outlined in Chapter 1. Chapter 2 had some success modeling h with a logistic Gaussian process, but the MCMC sampler was slow to explore the posterior space, and the densities produced by the LGP were limited in their flexibility. In response to both the slow mixing and inflexibility, this chapter explores modeling h with both a Dirichlet Process Mixture (DPM) and a Finite Mixture (FM) of kernels with support $[0, 1]$. By using a family

of kernels, k, that obeys the constraint $0 < k(0), k(1) < \infty$, h is also appropriately constrained at the boundaries.

3.1 Background

3.1.1 Mixture Models

It is proposed that $h(\cdot)$ takes the mixture form $h(\tau) = \int k(\tau|\Psi)dP(\Psi)$, where k is some family of kernels over $[0, 1]$ that obey the constraint $0 < k(0), k(1) < \infty$. Under this form, the CDF-transformed data $G_\theta(y_i) \sim h$; however, since θ is unknown and needs to be estimated simultaneous to h, a different data-likelihood specification is needed to model y_i directly. By noting that

$$\begin{aligned} f(y_i) =& g_\theta(y_i)h(G_\theta(y_i)) \\ =& g_\theta(y_i)\int k(G_\theta(y_i)|\Psi)dP(\Psi) \\ =& \int \underbrace{g_\theta(y_i)k(G_\theta(y_i)|\Psi)}_{\text{new kernel}} dP(\Psi), \end{aligned} \tag{3.1}$$

we see that unknown f can be represented as a mixture of kernels of the form $g_\theta(\cdot)k(G_\theta(\cdot)|\Psi)$ over an unknown mixing distribution P, establishing the direct connection $y_i \sim f$. Both families of pdfs g and k are known and prespecified.

3.1.2 Dirichlet Process Mixture Models

Assigning a Dirichlet process prior $P \sim \text{DP}(\alpha, H_0)$ to the mixing distribution P of Equation 3.1 results in a Dirichlet Process Mixture Model (DPMM). Dirichlet processes were first proposed by Ferguson (1973). Since then they have been used widely in density estimation and cluster detection. A standard reference for DPMMs is Escobar and West (1995), which models data as coming from normal kernels and assigns Dirichlet process priors to the means of those kernels, resulting in a nonparametric, flexible Bayesian estimate of the density.

Incorporating the specific kernel form of Equation 3.1, the full DPMM can be written as

$$\begin{aligned} y_i|\Psi_i, \theta \sim & g_\theta(y_i) \cdot k(G_\theta(y_i)|\Psi_i) \\ \theta \sim & \pi \\ \Psi_i \sim & P \\ P \sim & DP(\alpha, H_0). \end{aligned} \tag{3.2}$$

This form highlights that the data can be thought of as coming from a mixture of kernels indexed by some shared parameters, θ, and some observation-specific parameters, Ψ_i. As is true of other DPMMs, the discrete nature of the Dirichlet process lends itself to the possibility that atomic $\Psi_i = \Psi_j$ for some $i \neq j$, leading to a data-clustering interpretation. Though clusters can be obtained, this is not our primary goal, and in practice the kernels do not always lend themselves to strong separation over $[0, 1]$. Rather our goal is to create a sufficiently flexible prior on h for estimation of f.

3.1.3 Mixture Models for Extremes

Mixture models and even DPMMs have been used previously to model extremes. Some "mixture models" use a bifurcated approach, choosing one form for the bulk distribution below a threshold that is either explicitly or implicitly defined and another form for the tail distribution above the threshold. Scarrot and MacDonnald (2012) review these threshold approaches. Other mixture models come from weighting kernels that are defined over the full support of the data. For instance, Tressou (2008) proposes a DPM of Pareto distributions, and Carreau and Bengio (2009) propose a finite mixture of their five-parameter kernel, which they create from a truncated normal density stitched smoothly onto a GPD tail. Both of these approaches allow the kernels' tail indices to vary; the largest tail index determines the dominating rate of decay in the tail. Model 3.2, by contrast, specifies one polynomial

rate of decay or extreme value index, which will be a parameter of θ, and takes care of departures from the GPD tail through the nonparametric scaling of h.

Bean et al. (2016) is a mixture approach that uses a power-transformation to lighten the tails, a transformation they claim "can be directly related to the extreme tail index of the original density." It is similar to Model 3.2 in that it uses mixtures under transformation to model heavy-tailed data; however, the parameters of their transformation are pre-estimated and applied to the data prior to Bayesian modeling with DPMM of location-scale normals, i.e. the DPMM models $G_{\hat{\theta}}(y) \sim h$. The formulation of Model 3.2 allows estimation of θ to proceed simultaneously with the estimation of h.

3.2 Model Setup

3.2.1 Parametric Distribution and Priors

As in Chapter 2, the family of parametric distributions $\{g_\theta : \theta \in \Theta\}$ is selected to match the support of Y. Additionally, in order to model phenomena with heavy tails, $\{g_\theta\}$ needs to admit the possibility of heavy tails. For purposes of this research, we use a GPD with tail index $\xi = 1/\nu > 0$, scale parameter σ and location parameter fixed at 0. Additionally the GPD density is scaled by the 0.9-level-quantile of a unit-scale GPD with tail index $\xi = 1/\nu$, consistent with the approach of Chapter 2. Again the goal is to reduce dependence between ν and σ and induce some measure of orthogonalization.

The prior for ν, pictured in Figure 3.1, is chosen to be informative with mass over heavy-tailed values and support $\nu \in (0.5, \infty)$. Specifically, $\ln(\nu - 0.5) \sim \text{Logis}(2, 1)$. The independent prior on σ is set to be a scaled half-Cauchy distribution, with the scale elicited from the user's knowledge of the data mean, $\tilde{y}$. The prior $\sigma \sim \text{HC}(3/4 \times \tilde{y})$, also shown in Figure 3.1, is diffuse but proper. The prior for parametric θ is assumed to be independent of nonparametric h.

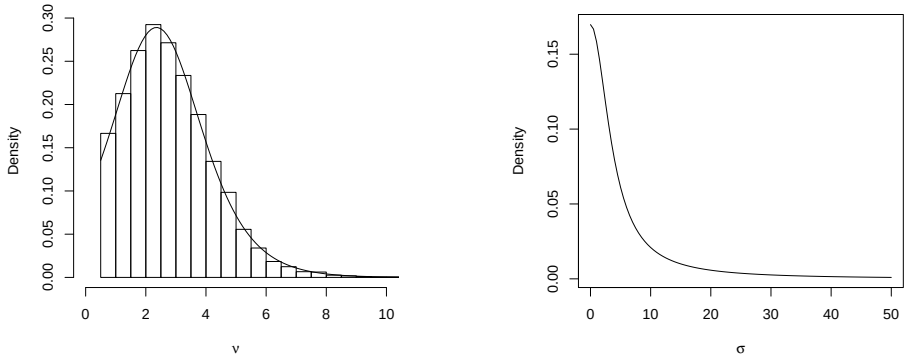


FIGURE 3.1: Priors for ν and σ

3.2.2 Nonparametric Distribution and Priors

Our mixture prior for h takes the general form $h(\tau) = \int k(\tau|\Psi)dP(\Psi)$. The kernel, $k(\tau|\Psi)$, is taken to be a normal distribution truncated to the $[0, 1]$ interval. Using a truncated normal kernel guarantees that $0 < k(0) < \infty$ and $0 < k(1) < \infty$, i.e. that both boundaries are finite and positive, as is needed for the framework constraints.

While a truncated normal kernel of itself does not look like a power law transformation, if the nonparametric kernel mixture distribution in aggregate mimics the shape of a power transformation, $b(1-\tau)^{b-1}$, then a confounding can exist between the GPD transformation and the nonparametric distribution form. Specifically,

$$g(y;\xi,\sigma) \times b(1 - G(y;\xi,\sigma))^{b-1} \equiv g\left(y;\frac{\xi}{b},\frac{\sigma}{b}\right), \tag{3.3}$$

where g and G represent the pdf and CDF of a GPD respectively. We aim to inject identifiability into this problem by placing priors on the nonparametric density that give prior probability to uniform-like shapes.

The priors for truncated normal kernel parameters Ψ, comprised of μ and β, are defined jointly by $\pi(\mu,\beta) = \pi(\mu|\beta) \cdot \pi(\beta)$. The prior over bandwidths, $\pi(\beta)$, is an InverseGamma(shape=0.01, rate=0.0005) truncated to an upper boundary of 1.

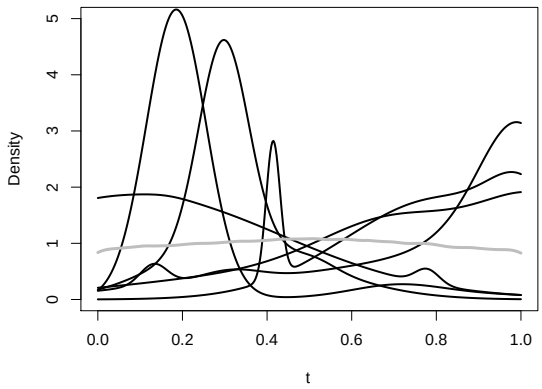


FIGURE 3.2: Nonparametric draws from prior for $h(\tau)$. Each is constrained such that $0 < h(\tau) < \infty$ for all τ. An estimate of the mean is show in gray.

Over the $[0, 1]$ interval, a truncated normal kernel with $\mu = 0.5$ looks nearly uniform when β is large. Truncating β at 1 allows there to be uniform-like truncated normals when β is near 1 without expanding the domain to an otherwise redundant infinity of uniform-like kernels. Given bandwidth β, the prior for center μ is taken to be $\mu|\beta \sim$ TrunNormal(mean=0.5, sd=0.5, $a = -\beta/2, b = 1 + \beta/2$). This truncation allows the kernel means to exist half a standard deviation outside the $[0, 1]$ interval, which in turn allows the nonparametric density estimates to "go up" at the boundaries. It also prevents computational loss of numerical significance that can occur if the mean, μ, falls many bandwidths, β, outside of the interval $[0, 1]$ where the data live.

For the precision parameter α of the Dirichlet process we use a Gamma$(1, 1)$ prior. Figure 3.2 shows several draws from the resultant prior for $h(\tau)$.

A finite mixture can be used in place of the more computationally-intensive infinite DPM, if desired. This is easily done through a latent-category likelihood construct, which can be represented as

$$y|P \sim \sum_{k=1}^{K} \rho_k \, TrunNorm(\mu_k, \beta_k) \text{ with } \sum_{k=1}^{K} \rho_k = 1 \text{ and } P = (\rho_1, \rho_2, ...\rho_K)$$

or as

$$y|w \sim \prod_{k=1}^{K} TrunNorm(\mu_k, \beta_k)^{\mathbb{1}(w=k)} \text{ with } w|P \sim Multinomial(\rho_1, \rho_2, ..., \rho_K). \quad (3.4)$$

When combined with a Dirichlet prior, these multinomial-Dirichlet likelihoods lend themselves well to Bayesian MCMC sampling because the probabilities can be updated with conjugate Dirichlet draws; latent categories can be simulated for each observation from a multinomial distribution; and updates to kernel parameters only depend on data within the associated latent category. Fixing the hyper-parameter $\alpha = 0.5$ in the Dirichlet prior $P = (\rho_1, \rho_2, ..., \rho_K) \sim \text{Dirichlet}(0.5, 0.5, ...0.5)$ encourages higher weights on fewer mixture kernels.

3.3 Computation

Bayesian posteriors over all model parameters are obtained through Gibbs sampling. Algorithm 8 of Neal (2000), appropriate in cases of non-conjugacy, is used to obtain group-membership updates for the infinite Dirichlet process mixture. For updates to the precision parameter of the Dirichlet process, a modification to the auxiliary variable scheme of Escobar and West (1995) is used. Namely, a sample drawn from $\alpha \sim \text{Gamma}(1 + K, 1 - \log(\eta))$ is followed by an update for η with a random draw from $\eta \sim \text{Beta}(\alpha, n)$.

One drawback to the DPMM is that cluster membership updates require sequential processing across all n observations and are slow. Replacing the Dirichlet-process prior with a Dirichlet prior and using a finite mixture approximation with $K = \lceil \log(n) \rceil$ groups, see Equation 3.4, is computationally expedient. Furthermore, we find that the finite version provides a sufficiently flexible form for most nonparametric densities but has the added advantage of being able to simultaneously update all n group memberships, insofar as the implementation software supports vector

operations. We perform all computing in R (R Core Team, 2018), and so the finite approximation and these vectorizations provide significant speedups.

Under the finite mixture model (FMM), the parameters dictating the observations' group memberships can be updated through conjugate draws. Given latent observation memberships w_i to the K truncated normal kernels, the probabilities in vector P, which dictate probabilistic membership among K groups, can be updated with a single conjugate Dirichlet draw. The latent memberships w_i use conjugate multinomial updates, which are vectorizable as previously mentioned.

Having used a truncated-inverse-gamma prior for β and a truncated-normal prior for μ, conjugate updates are also available for the clusters' truncated normal parameters through an augmented data approach, similar to the one employed in Kotecha and Djuric (1999). Essentially, data belonging to the cluster are first transformed by G_θ for the current θ iterate, then taken through the deterministic transformation

$$x_i = \mu_c + \beta_c \Phi^{-1}\left(\frac{\Phi(\frac{G_\theta(y_i)-\mu_c}{\beta_c}) - \Phi(\frac{0-\mu_c}{\beta_c})}{\Phi(\frac{1-\mu_c}{\beta_c}) - \Phi(\frac{0-\mu_c}{\beta_c})} \right), \tag{3.5}$$

where μ_c and β_c represent the current cluster mean and standard deviation, respectively. This generates a latent-variable representation of the data in "untruncated" normal form, i.e. over $(-\infty, \infty)$ instead of over (a, b). These normal x_i combined with the independent priors result in conjugate complete conditionals. Gibbs draws provide updates for β_c and μ_c in turn, using the inverse CDF method to obtain samples from either the inverse gamma or normal distribution truncated to their respective conditional domains. Numerical difficulties can arise in the inverse CDF samplers if the transformed x_i push (numerically) all of the posterior density for β or μ outside of the boundaries. In the rare instances where this happens, we take the draw to be some small ϵ inside the boundary.

The update for joint (ν, σ) is non-conjugate and uses an independence Metropo-

lis algorithm. After monotonically transforming ν and σ from $[0, \infty] \times [0, \infty]$ to $[-\infty, \infty] \times [-\infty, \infty]$, the complete conditional given all observation-level μ_i and β_i is optimized, using R's `optim` function. The proposal density is then taken to be a MVN_2 centered at the mode with covariance matrix equal to the negative inverse of the numerically-optimized Hessian upscaled by a factor of 1.1.

It is possible to use a "warm" start to speed up chain convergence. This can be obtained by numerically optimizing a simplified posterior that only uses a single truncated normal kernel, which amounts to a 4-dimensional optimization over a box-constrained space (2 GPD parameters, 2 truncated normal parameters with $0 < \mu < 1$ used for simplicity instead of the constraint depending on β). Observations can then be randomly assigned membership among the K groups for the finite case or into a single group in the infinite case, each of which group starts with the optimized truncated normal parameters.

After MCMC sampling, the posterior predictive is estimated by numerically integrating over Ψ and θ using MCMC draws from the posterior $\pi(\Psi, \theta | Y_{1:n})$:

$$\pi(Y_{n+1}|Y_{1:n}) = f(y) = \int \left\{ \int \tilde{k}_\theta(y|\Psi)\pi(\Psi, \theta|Y_{1:n}) d\mu(\Psi, \theta) \right\}. \tag{3.6}$$

The posterior predictive for the nonparametric, truncated-normal kernel mixture density, $h(\tau) = \int k(\tau|\Psi)\pi(\Psi)d\mu(\Psi)$, can also be approximated via numerical integration using the marginal Ψ posterior draws. These numerical integrations are straightforward for the finite mixture.

The DPM posterior predictive is more complicated and is estimated via sampling. For each iteration of the MCMC sampler, samples are drawn from the prior for Ψ and are assigned weights according to a truncated stick-breaking process. These samples are then weighted together with the posterior cluster parameters Ψ_c for the given MCMC iteration using a random draw from a Dirichlet$(\alpha, n_{c_1}, n_{c_2}, ..., n_{c_{K_s}})$, where n_{c_j} represent the number of observations in the j-th cluster of the given iteration.

These techniques are applied to get estimates for both the density and distribution functions.

Posterior estimates for the quantiles are obtained by linearly interpolating the *nonparametric* posterior predictive distribution function for the desired quantile level p and sending those inverse CDF estimates through the inverse θ-transformation function, i.e. the GPD quantile function with θ determined by the current MCMC iterate.

3.4 Simulation Study

A subset of the simulation scenarios described in Section 2.3.1 of Chapter 2 will be discussed here, namely the Standard GPD, Half-t, Fourth-power GPD, and Gamma-GPD mixture, each with sample size n = 1,000. Figure 2.1 has a representation of each density function. Simulations are again evaluated primarily on the basis of their tail-index estimation and upper-tail quantile bias and RMSE. With a sample size of n = 1,000 for each scenario, considering quantile level p out to 0.9999 reflects extreme extrapolation.

Similar to the simulations of Chapter 2, the mixture model is compared to maximum-likelihood estimates of 1) a generalized Pareto distribution (GPD) under various truncations fit using the `extRemes` package (Gilleland and Katz, 2016) in R; and 2) the extended generalized Pareto distribution (EGPD) proposed by Papastathopoulos and Tawn (2013) under various truncations fit using the `mev` R package (Belzile et al., 2018). Thresholding proportions for the GPD and EGPD methods are selected to span the range of visually-diagnosed cutoffs. For methods employing truncation, bias and RMSE are only displayed for levels p that were retained after thresholding. When referring to the method used, plots and tables also refer to the percentage of data retained.

For five datasets of each scenario, two DPMM chains are run for 200,000 itera-

Table 3.1: Simulation MCMC summaries for infinite DPMM (five datasets per scenario) and FMM (40 datasets per scenario), including total iterations run for each chain, number discarded as burn-in, number of draws retained after thinning, number of chains converging, and time (minutes) needed to run 1k iterations of one chain.

	DPMM Draws					FMM Draws				
Simulation Scenario	Total	Burn	Keep	Conv	Time	Total	Burn	Keep	Conv	Time
Standard GPD	200k	80k	3k	5/5	6.6	120k	20k	2k	38/40	1.9
Half-t	200k	80k	3k	3/5	6.9	120k	20k	2k	24/30	2.1
Fourth power GPD	200k	80k	3k	3/5	5.3	120k	20k	2k	30/40	1.7
Gamma-GPD mix	200k	80k	3k	4/5	5.3	120k	20k	2k	35/40	1.7

tions. The first comes from a "warm" start and the second has parameters drawn from the prior. For each chain, 80,000 iterations are discarded as burn-in and then further thinned (every 40) to retain only 3000 draws. These chains are meant to give an idea of the behavior of the model when run as an infinite mixture and are not included in simulation summaries. Instead, forty datasets from each scenario are run as FMMs with the number of groups fixed at $\lceil \log(n) \rceil = \lceil \log(1000) \rceil = 7$ and analyzed. Two chains, one with a "warm" start and one with a prior start, are run for 120,000 iterations. The first 20,000 draws are discarded, and thereafter 2,000 thinned samples are retained from the posterior. Table 3.1 summarizes the number of MCMC iterations obtained, burned, and retained.

Post burn-in, the convergence of MCMC chains is assessed for each posterior separately by considering Gelman-Rubin potential scale reduction factors (PSRFs) for ν and σ. PSRFs for mixture model parameters are not assessed since label-switching is accepted and not adjusted for, and PSRFs under this unidentifiable scenario would be meaningless.

3.4.1 Computational Speed

In order to draw 200,000 samples from each posterior, the DPMM takes an average of 17.7 hours (fourth-power GPD) to 23.0 hours (half-t) to run a single chain. This

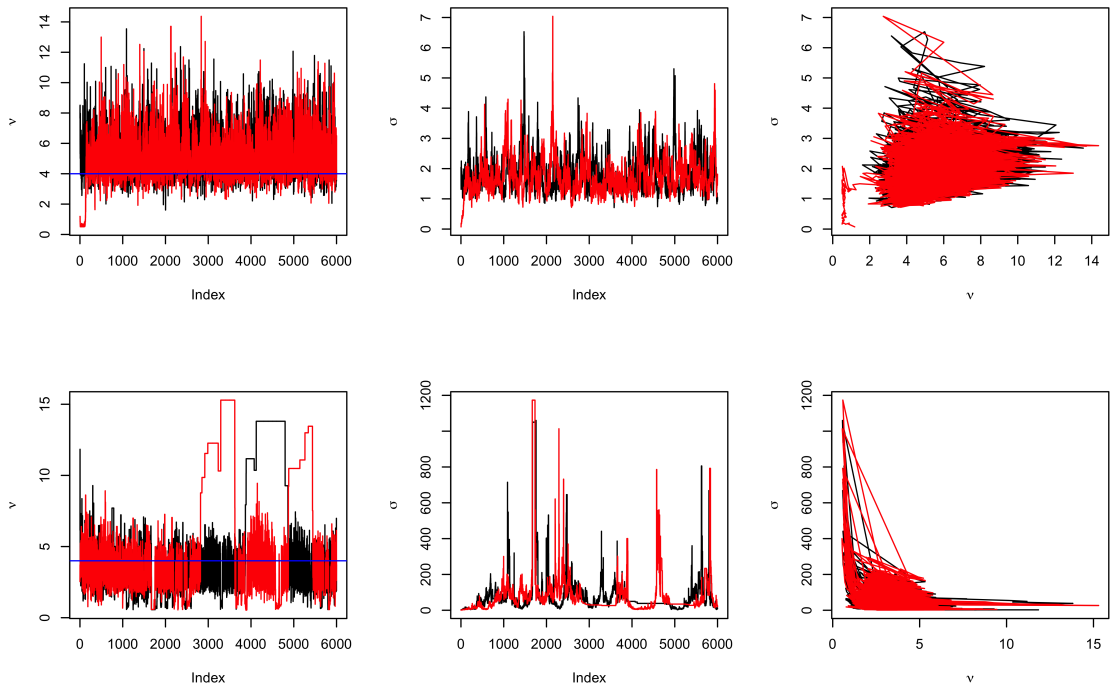


FIGURE 3.3: Traceplots over ν and σ separately and jointly for two example datasets from the half-t scenario fit using FMM. Black chains come from "warm" starts and red from random prior draws. Upper row is from dataset 26, and lower row is from dataset 19.

translates to 6.7 minutes per 1,000 draws when $n = 1{,}000$. The FMM takes between 3.3 hours (fourth-power GPD) and 4.2 hours (half-t) to obtain 120,000 draws, averaging 1.7 to 2.1 minutes per 1,000 iterations. That is, the DPM takes about five times as long to run as the FMM (see Table 3.1). The computational time for these chains is affected by the sample size, the time it takes for the numerical optimizer to converge for the θ independence sampler, and (in the case of the DPMM) the number of clusters, since each cluster requires parameter updates at each iteration.

3.4.2 Chain Convergence and Sampling Features

Similar to what is seen in Chapter 2, the MCMC samplers exhibit high autocorrelation, especially apparent in σ. This is because, as the MCMC sampler moves among different σ possibilities, potentially very different shapes of the nonparametric densities h are needed to accommodate the true density f, resulting in the need to adjust all of the mixing distribution parameters, Ψ.

Complicating the interplay between θ and h are complexities seen in some θ marginal posteriors: MCMC chains show indications of long tails, some of which are boomerang-shaped, and/or multimodality. The independence sampler used for θ, designed to adjust proposals to the locally needed covariance between ν and σ, is able to round bends in the marginal posteriors; however, it is unclear if the sampler's forays into these parts of the posterior provide representative marginal samples or if the sampler is getting caught in regions of lower density due to the slow adjustments in the mixing distribution parameters. For this model, the quantile scaling of the base distribution g_θ actually seems to be *inducing* curvature into the θ marginal. Post-hoc parameterization of the scale parameter to make it correspond to a non-scaled GPD base reduces some of the apparent curvature, indicating this model may have been better off without the quantile scaling.

Some chains propose and get stuck at unusual θ values, such as those with ν relatively much larger or smaller than other sampled values. The nonparametric density h quickly adjusts to these θ, but it can take dozens, hundreds, or even thousands of iterations before another proposal θ is accepted by the Metropolis-Hastings sampler. This sort of stickiness, which may be an indication of multimodality, can strongly affect PSRF metrics for assessing convergence. Knowing this, a generous PRSF limit is used; only replicates with upper-95% confidence limits below 1.5 for both ν and σ's univariate PSRFs are included in analysis. Table 3.1 summarizes the number of

Table 3.2: Summaries related to DPMM and FMM estimation: mean precision parameter (α); mean number of clusters with at least 5 observations (K); proportion of iterations where K stays at or below seven ($K \leq 7$); and tail index estimates.

	DPMM				FMM	
Simulation Scenario	α	K	$K \leq 7$	Index	K	Index
Standard GPD	1.23	5.2	0.79	0.29	5.9	0.28
Half-t	1.51	6.7	0.66	0.25	5.3	0.29
Fourth-power GPD	0.65	2.8	0.98	0.28	5.7	0.29
Gamma-GPD mixture	1.62	6.9	0.64	0.26	6.0	0.27

replicates that meet this criterion.

Figure 3.3 shows traceplots for two example datasets, each from the half-t scenario, run using FMM. The sampler for the first dataset has well-behaved, if slow-mixing, chains over the two θ parameters. The sampler for the second dataset, displayed in the lower panels, exhibits all three mixing difficulties mentioned above: a boomerang-shaped tail into small ν values; large σ spread, associated not only with small ν and but also mid-range ν values; and stickiness at certain unusual θ pairs. The fact that both chains appear to be experiencing all three of these difficult mixing situations may point towards these being legitimate features of the marginal posteriors under this model.

3.4.3 DPMM to FMM Comparison

Across scenarios, about 30% of DPMM clusters are small, containing fewer than five observations. The following discussion excludes small clusters and focuses on clusters with at least five allocated observations. The column labeled "K" in Table 3.2 shows that the DPMM uses on average only 2.8 clusters to explain the nonparametric density shape of the fourth-power GPD, while the other scenarios use more. The gamma-GPD mixture uses the most at 6.9. The α precision parameter reflects this spread, averaging 0.65 across fourth-power GPD datasets and ranging up to 1.62 in

the gamma-GPD mixture. The FMM precision parameter is not estimated but fixed at $\alpha = 0.5$.

FMM is given seven components among which to allocate observations. For comparison with DPMM, Table 3.2 shows the FMM average number of components (clusters) having at least five observations. DPMM uses many fewer clusters than did FMM (2.8 versus 5.7) for the fourth-power GPD scenario. Column "$K \leq 7$" of Table 3.2 shows that 98% of DPMM sampling iterations use seven or fewer clusters. For this scenario, FMM has ample components to describe the nonparametric density and may even use more than necessary. For the standard-GPD, FMM also uses slightly more clusters on average than does DPMM. The half-t and gamma-GPD scenarios, on the other hand, allocate to fewer clusters on average than their DPMM counterparts, with DPMM using more than seven components in more than a third of all iterations.

Table 3.2 includes the average tail indices over converged replicates of the DPMM and FMM. In all but the half-t scenario, indices are within 0.01 of each other. Estimates from nonparametric densities, displayed in Figure 3.4, show that DPMM and FMM estimates are similar to each other in form, though the FMM may show more downward cupping at either boundary than the DPMM. Not too much can be refined on either of these comparisons since the DPMM has few replicates. But overall the DPMM seems close enough to the FMM that we are comfortable proceeding with the FMM substitution for the more computationally intense DPMM.

3.4.4 Lower-tail and Bulk Results

The primary focus of analysis is on upper-tail estimation; however, Table 3.3 is included to summarize FMM relative bias (bias/truth) for each scenario at the median as well as at the smallest estimated quantile level, $p = 0.00001$. Simulations are essentially unbiased for the median: the largest relative bias among four scenarios

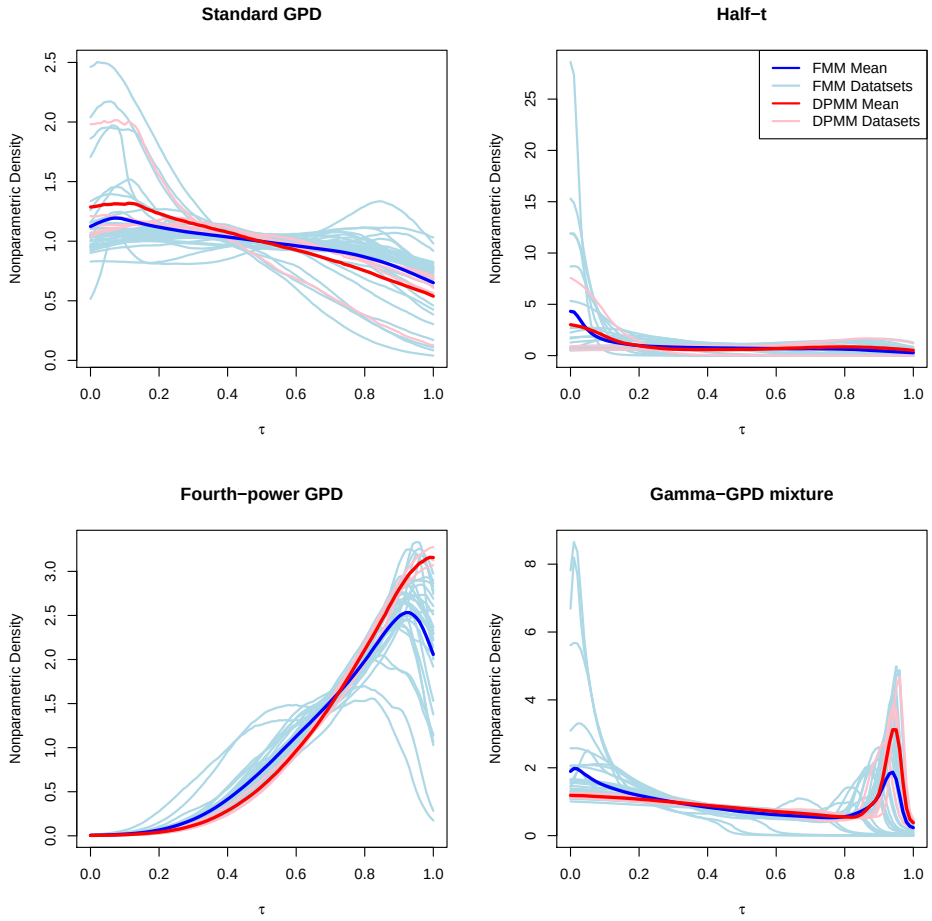


FIGURE 3.4: Estimated nonparametric densities for converged DPMM replicates (pink with mean in red) and FMM replicates (light blue with mean in blue).

is 0.013. The fourth-power GPD has the poorest lower-tail relative bias, underestimating the 0.00001 quantile by 95.5%. This happens, in part, because the FMM inherits its strictly positive density at zero from the GPD base g_θ; whereas, the true value $f(0) = 0$. This results in FMM hitting its 0.00001-level quantile at much smaller values than the true fourth-power GPD. Of note also is the half-t scenario, which had the smallest bias in the left-hand tail of the four scenarios. Unlike the other scenarios which all have largest relative bias in the bulk or lower tail at $p = 0.00001$, the half-t hit its largest relative bias at $p = 0.11$, underestimating the quantile by

Table 3.3: Relative bias (bias/truth), for each FMM scenario at the median, $p = 0.5$, representative of the bulk, and at the $p = 0.00001$ quantile level, representative of the lower tail.

Simulation Scenario	Median	Lower-tail
Standard GPD	0.010	0.115
Half-t	-0.002	0.002
Fourth-power GPD	-0.001	-0.955
Gamma-GPD mixture	0.013	0.113

Table 3.4: Simulation tail-index results for all scenarios across comparison methods. The table includes mean tail-index estimates (Mean); interval coverage (Cov), i.e. proportion of intervals across replicated datasets that contain the true tail-index; and width of 95% confidence or credible intervals (Wid).

	Standard GPD			**Half-t**			**Fourth-power**			**Gamma-GPD**		
Method	Mean	Cov	Wid	Mean	Cov	Wid	Mean	Cov	Wid	Mean	Cov	Wid
Truth	**0.25**			**0.25**			**0.25**			**0.25**		
GPD 5%	0.17	0.89	0.73	0.11	0.92	0.67	0.18	0.83	0.73	0.42	0.94	0.76
GPD 10%	0.25	0.97	0.53	0.14	0.88	0.48	0.20	0.90	0.50	0.23	0.86	0.42
GPD 20%	0.27	0.92	0.36	0.16	0.79	0.32	0.22	0.90	0.34	0.01	0.00	0.19
GPD 100%	0.25	0.92	0.15									
EGPD 5%	0.13	0.92	0.94	0.12	0.83	0.89	0.16	0.70	0.95	0.47	0.86	0.89
EGPD 10%	0.20	0.97	0.64	0.14	0.79	0.60	0.16	0.73	0.60	0.32	0.97	0.50
EGPD 20%	0.27	0.92	0.45	0.15	0.83	0.41	0.21	0.87	0.42	0.11	0.51	0.24
EGPD 100%	0.25	0.97	0.18									
FMM 100%	0.28	1.00	0.28	0.29	1.00	0.63	0.29	0.97	0.52	0.27	1.00	0.40

just 2%.

3.4.5 Tail-Index Results

Table 3.4 shows that when data come from a standard GPD, the best tail index coverage and narrowest widths come, not unsurprisingly, from ML estimation without thresholding; however, those intervals widen quickly with higher thresholds and less data. The FMM tail index overestimates ξ by 0.03 and has coverage too high for its nominal level (100% vs. 95%) but maintains this coverage with narrower intervals

than the methods employing truncation. Similar patterns hold across scenarios: the tail index is overestimated by FMM (but only at most by 0.04) while coverage remains high and intervals remain in the range of widths seen under GPD or EGPD with truncation. Among FMM estimates, the half-t scenario had the heaviest tail-index estimates and the widest interval widths.

All methods have difficulty in the gamma-GPD scenario dealing with the "contamination" coming from the gamma mixture component, as evidenced by their bias and/or poor interval coverage. The GPD 10% method gets closest to the true tail index of 0.25, but the method has coverage that is lower than desired, even with its wide intervals. The FMM on the other hand is able to use the full bulk of the GPD-gamma mixture to help in estimating the tail index.

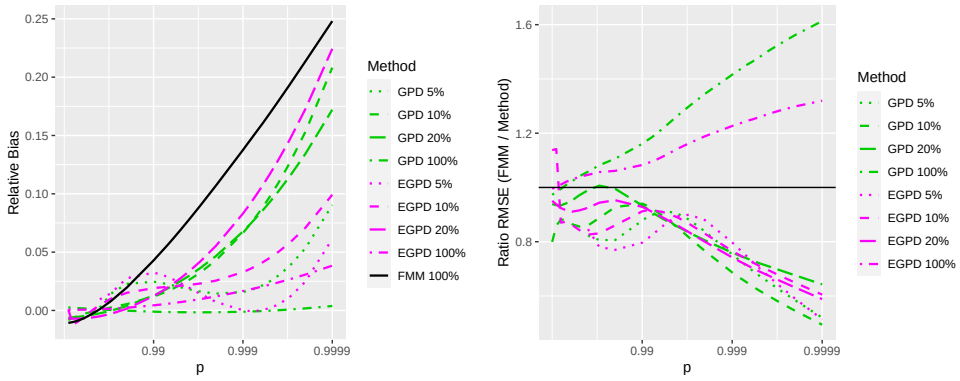
3.4.6 Upper-tail Quantile Estimation

In all four scenarios, FMM exhibits positive bias in the tails, as can be seen in the relative bias plots of Figure 3.5. The biases can be large, even for levels that should have some informing data, e.g. $p < 0.999$. The bias in GPD and EGPD over these data-informed quantile levels is small in comparison.

Despite these biases, the FMM performs better than GPD or EGPD in RMSE for several scenarios, as can be seen in the plots of Figure 3.5. In the standard GPD scenario, FMM has higher RMSE than either GPD or EGPD that retain all 100% of the data, but it has consistently lower RMSE across all tail levels $p > 0.9$ for any of the GPD or EGPD truncations considered. In the fourth-power GPD scenario, FMM has lower RMSE than all truncation methods at all tail quantile levels. For the gamma-GPD scenario, the FMM has higher RMSE than any of the MLEs for $0 < p < 0.999$, but in extrapolation ($p > 0.999$) it has slightly lower RMSE than the truncation methods, likely aided by its more accurate estimation of the tail index.

Figure 3.5b shows the results of the half-t simulation are particularly biased,

(a) **Standard GPD scenario**



(b) **Half-t scenario**

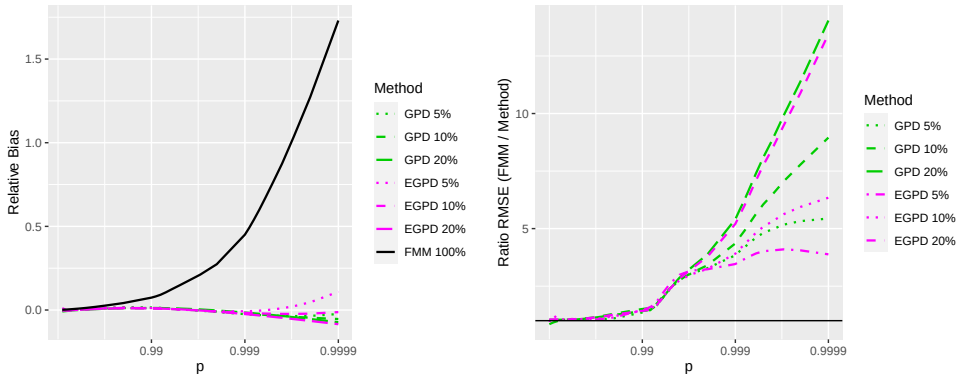


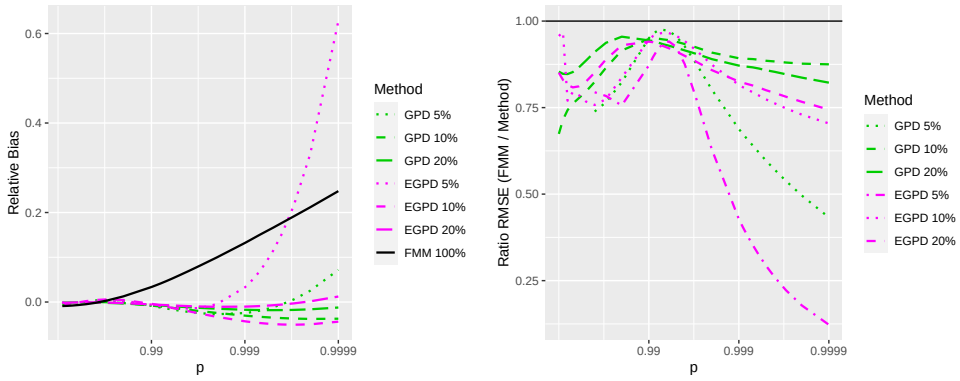
FIGURE 3.5: Upper-tail, quantile-extrapolation with p on a log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of FMM 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than FMM.

overestimating the 0.9999-level quantile by over 150%. This bias translates into RMSE being consistently and massively higher in the FMM than in the truncation methods.

3.4.7 Sensitivity of Model to Left-hand "Tail"

The half-t scenario is unique among the simulation scenarios in several ways: it has much lower bias at the zero boundary than do the other scenarios; it has higher proportions of datasets exhibiting difficulty in attaining MCMC convergence; and

(c) **Fouth-power GPD scenario**



(d) **Gamma-GPD mixture scenario**

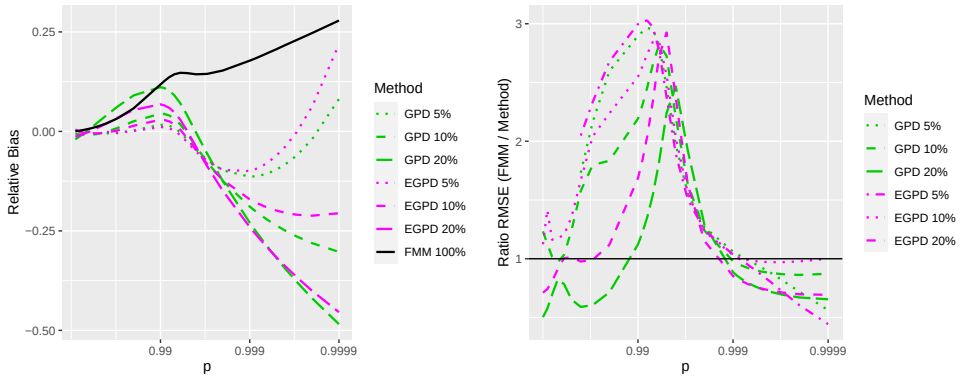


FIGURE 3.5: Upper-tail, quantile-extrapolation with p on a log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of FMM 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than FMM.

it has more datasets with large spread in the σ marginal posterior. Further exploration reveals that features of the data are informative about which MCMC chains might face difficulties converging. Datasets with lighter-than-usual left-hand tails—identified by comparing the empirical CDF to the true CDF for the smallest 20% of values using logged p-p plots—were more likely to have large sampled values for σ. Also, large sampled values for σ were associated with smaller values for ν, i.e. larger ξ and heavier tails, through the boomerang-shaped marginal ν-σ posterior.

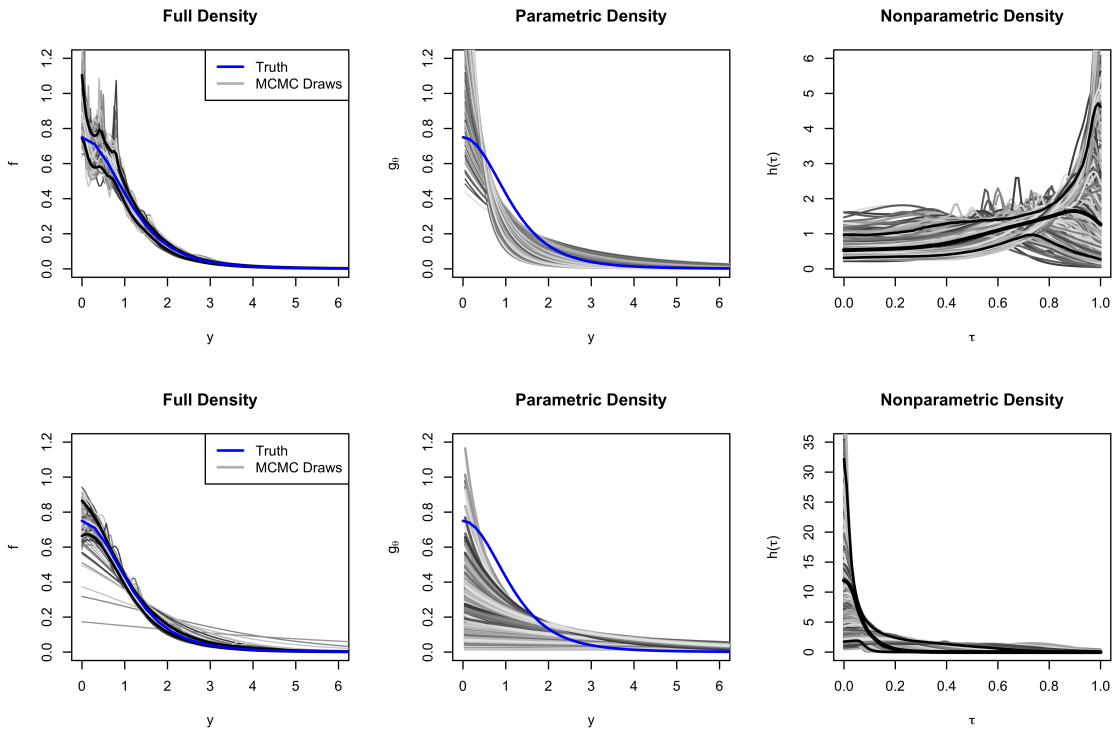


FIGURE 3.6: Estimated densities from same datasets as shown in Figure 3.3: dataset 26 in upper row and dataset 19 in lower row. Left panels show draws for the overall density in grays, truth in blue, and 95% intervals in black. Middle panels show draws for the transformation GPD density indexed by $\theta = (\nu, \sigma)$ in gray alongside truth in blue. Right panels show draws for the nonparametric density in grays with 95% credible intervals in black.

Revisiting the two half-t datasets highlighted in the discussion around sampler mixing (see Section 3.4.2), Figure 3.6 plots their full estimated sampling densities in the left panel plots, the parametric densities associated with sampled θ in the middle panels, and the sampled nonparametric densities in the right panels. Dataset 26, the example with well-behaved chains, is shown in the first row. Its full or sampling-density intervals have a negative slope at the left boundary. That is, they tend to follow the shape of the GPD base distribution (middle) near zero, requiring only small downscaling from the nonparametric density (right panel) at $h(0)$. Intervals for the nonparametric density of dataset 26 are greater than 0 at either boundary.

The estimates for dataset 19, which had chains slowly exploring more complex marginal θ spaces, are shown in the second row. Its full-density intervals flatten more at $f(0)$ than did dataset 26, capturing its slightly lighter-than-usual left-hand tail. The way that it achieves this effect is by accepting θ values that greatly flatten the parametric density (middle panel), i.e. with large σ and/or small ν, and pairing those with nonparametric densities h which pile all of their mass towards the zero end of its $[0, 1]$ support. The FMM is adept at building this nonparametric density shape because 1) the needed form is so like the truncated normal kernel itself and 2) the mixture model prefers to use as few kernels as possible. The overall effect across simulation replicates is that quantile estimation in the lower tail and bulk of the sampling distribution has low bias and RMSE for the half-t scenario, even more so than the other scenarios. But the upper-tail quantile estimation is very biased leading to high RMSE. This is clearly a case where the model has favored bulk estimation and the bulk estimation has overwhelmed the tail estimation. One takeaway is that the model form can make estimates sensitive to certain lower-tail and bulk shapes, which can adversely affect the estimation in the upper tail.

3.4.8 Fixing the Transform Scale

When considering simplifications to the model and in response to the mixing difficulty which is so strongly related to σ, it is worth asking whether σ needs to be estimated at all or if the value can be fixed. Nearly all literature that employs a transformation for heavy-tailed univariate density estimation does so by restricting the parameters θ of the transformation G_θ to a single parameter, the index of the power law. Even for those that have CDF transforms G_θ, most do not estimate a scaling parameter σ, unless they pre-estimate and fix it, thereafter estimating the nonparametric density (see for example Buch-Larsen et al., 2005; Bolancé et al., 2010). Those that do not include a scaling parameter directly in their model tend to scale the data by

some empirical measure, e.g. by the inter-quartile range, prior to estimation (see for example Bean et al., 2016) or back scale their power-transformed model to match the scale of the data as in Wand et al. (1991) or Bolancé et al. (2003).

Fixing σ is explored more fully in an alternative mixture model, which uses a kernel that is itself a mixture of uniform and beta distributions in place of the truncated normal distribution. Details about this model, its computation, and simulation results, including the fixed-σ analysis, are included in Appendix B. In short, the findings are that fixing σ improves chain mixing and speeds convergence across all scenarios, but that it may degrade the quality of tail-index and tail-quantile estimation. Namely, the tail index sees a drop in interval coverage. Additionally, the tail quantile estimates follow similar patterns in bias and relative RMSE to their σ-estimated counterparts but with slightly more bias and slightly less RMSE advantage. So while fixing σ is not generally advised, it may be expedient if difficulty is encountered with convergence of the MCMC sampler for a particular dataset.

3.5 Conclusions and Discussions

This dissertation chapter was a response to Chapter 2 and the LGP's inability to capture certain nonparametric density shapes and do so quickly. The FMM, while still plagued by slow mixing, is able to traverse the complex posterior shapes necessitated by varying σ. In most cases, FMM is able to reach steady states in fewer iterations and adjust its nonparametric form to changing σ more quickly than LGP. It has proven quite flexible and able to capture many bulk shapes while still giving form to the heavy tails, at least in three of the four scenarios considered.

An insight from this work that was not fully appreciated with the LGP model is how much the form of the nonparametric density can influence the estimation of parametric θ. The various shapes enabled and favored by the nonparametric model—shapes which may be sensitive to the lower-tail or bulk distribution—dictate where

the posterior mass of θ will reside (e.g. in long tail regions). Exploration of these marginal θ posteriors with MCMC samplers can be difficult, and the form of the model may end up inadvertently favoring θ that enables accurate bulk f estimation over θ that corresponds to accurate estimation of f's upper tail, as was seen in the half-t scenario.

The FMM has a few drawbacks. The curvature (and potential multimodality) in the marginal θ posteriors was anticipated, but perhaps not handled as well as it could have been. Future iterations of work might consider using a more traditional orthogonalization of ν and σ such as the one proposed in Chavez-Demoulin and Embrechts (2004), namely $\sigma = \tilde{\sigma}(1+\xi)$. Another drawback is that the model exhibits systemic bias in the tails. This bias may be related to the downward cupping shape that the nonparametric densities show at their boundaries, observed in the plots of Section 3.4.3.

The model's primary strength lies in its use of data. By retaining all data and incorporating prior information that the tails are heavy, the model can reduce RMSE of quantile extrapolations across many types of densities and tails.

Another strength of this work is the simplicity of the mixture density setup. With only two parameters indexing the truncated normal mixture kernel, it may be possible to represent the mixture through an approximation to the DPMM, namely via predictive recursion. The predictive recursion approximation allows the mixing distribution to be integrated out entirely. After integrating out the mixing distribution, all that is left is a marginal likelihood over the two θ parameters. This approach, termed Predictive Recursion Marginal Likelihood (PRML) and introduced by Martin and Tokdar (2011), should be much easier to maximize. This approximation may not only speed up the algorithm, which is admittedly very slow, but may also simplify the complex posterior σ-h relationships by integrating the mixture distribution h out of the model entirely. In theory this marginalization would also allow for exploration

of the θ space more fully. Predictive Recursion Marginal Likelihood will be utilized in the next chapter.

4

Predictive Recursion Marginal Posterior Model

This chapter explores a third modeling approach for the transformation framework, detailed in Chapter 1, for univariate density estimation in the presence of heavy tails. Within the framework, a family of distributions $\{g_\theta : \theta \in \Theta\}$ is selected and also a nonparametric family of distributions $\mathcal{H} := \{h(\cdot)\}$, which have been constrained such that h are density functions on $[0, 1]$ with $||\log h||_\infty \ll \infty$. Having both of these, set $\mathcal{F} := \{f(\cdot) = g_\theta(\cdot)h(G_\theta(\cdot))\}$ for $\theta \in \Theta$ and $h \in \mathcal{H}$. Then the data, $Y_1, Y_2, ... \overset{iid}{\sim} f, f \in \mathcal{F}$, are modeled using a Bayesian approach. That is, priors on θ and h and posteriors are estimated simultaneously.

The approach in this chapter uses Predictive Recursion (PR), an approximation to the Dirichlet process mixture presented in Chapter 3, to model nonparametric h and thereby f. For a fixed kernel, predictive recursion quickly produces both a mixing distribution over the kernel's parameters and an estimate of the data's generating distribution in the form of a kernel mixture distribution. An extension to PR called Predictive Recursion Marginal Likelihoods (PRML), first introduced by Martin and Tokdar (2011), allows additional parameters to be included in the kernel that are

not part of the mixing distribution. PRML provides a jumping off point from which we implement a Bayesian, approximate model capable of estimating the tail index of a heavy-tailed sampling density and obtaining predictions for the upper tail of that sampling density.

4.1 Background

As in Chapter 3, it is proposed that nonparametric $h(\cdot)$ takes the mixture form $h(\tau) = \int k(\tau|\Psi)dP(\Psi)$, where k is some family of kernels over $[0, 1]$ that obey the constraint $0 < k(0), k(1) < \infty$. Then using

$$\begin{aligned} f(y_i) =& g_\theta(y_i)h(G_\theta(y_i)) \\ =& g_\theta(y_i) \int k(G_\theta(y_i)|\Psi)dP(\Psi) \\ =& \int g_\theta(y_i)k(G_\theta(y_i)|\Psi)dP(\Psi), \end{aligned} \tag{4.1}$$

f represents a mixture of kernels of the form $\tilde{k}_\theta(\cdot|u) = g_\theta(\cdot)k(G_\theta(\cdot)|\Psi)$ over an unknown mixing distribution P. The data are assumed to be independent $y_i \sim f$. Both families of pdfs g and k are known and prespecified.

Predictive recursion is one approach to estimating the unknown mixing distribution P and its corresponding kernel mixture density f. The algorithm was introduced in Newton et al. (1998) and Newton and Zhang (1999) as an approximation to the posterior mean of the mixing distribution of a Dirichlet process mixture model. Its recursive evaluation provides a fast alternative to MCMC for Dirichlet process mixtures, and its form can be seen as having roots in empirical Bayes. Martin (2018) gives a review of the theory, properties, and extensions around predictive recursion. Presented here are the parts relevant in the context of the extreme density mixture model.

4.1.1 Predictive Recursion Algorithm

Given 1) the data sequence $y_1, ..., y_n$; 2) a family of kernel densities $\tilde{k}_\theta(\cdot|u)$ indexed by u and θ, parameters over which the mixing does and does not take place, respectively; 3) some initial guess at the mixing density over mixing parameters u, $p_{0,\theta}(u)$, acting as a prior; and 4) some sequence of weights $w_1, ..., w_n$, often taken to be of the form $w_i = (i + c)^{-\gamma}$; the mixing density $p_{i,\theta}$ can be found for $i = 1, ..., n$ by

$$p_{i,\theta}(u) = (1 - w_i)p_{i-1,\theta}(u) + w_i \frac{\tilde{k}_\theta(y_i|u)p_{i-1,\theta}(u)}{f_{i-1,\theta}(y_i)}, \tag{4.2}$$

where

$$f_{i-1,\theta}(y) = \int \tilde{k}_\theta(y|u')p_{i-1,\theta}(u')\mu(du'). \tag{4.3}$$

The quantity $f_{i,\theta}$ is an estimate of the mixture density from which the data are drawn, technically a conditional mixture density given θ. While useful as an online or sequential filter, we use PR here by recursing through all n observations and only retaining the final mixing density, $p_{n,\theta}$, and mixture density, $f_{n,\theta}$. Permuting the data many times, running the algorithm on each sequence, and averaging the mixing distributions over all orderings essentially eliminates PR's dependence on the ordering of the data (Tokdar et al., 2009).

4.1.2 Predictive Recursion Marginal Likelihood

As functions dependent on θ, the algorithm presented in equations 4.2 and 4.3 can be classified under the Predictive Recursion Marginal Likelihood umbrella introduced in Martin and Tokdar (2011). Their paper discusses approximating the marginal likelihood for θ by

$$L_n^M(\theta) = \prod_{i=1}^{n} f_{i-1,\theta}(y_i) \tag{4.4}$$

and then maximizing the marginal likelihood $\hat{\theta} = argmax_{\theta \in \Theta} L_n^M(\theta)$. Their estimate for the standard error of θ is obtained via Laplace approximations, where the curvature of $L_n^M(\theta)$ at its maximum is approximated as part of their PRML-gradient algorithm. We do not take an optimization approach, but rather set a prior on θ and incorporate the approximate likelihood into a marginal, Bayesian posterior over θ.

To our knowledge predictive recursion has not been used in a heavy-tailed context thus far. Nor are we aware of previous usage of the predictive recursion algorithm over bounded spaces, as our $[0, 1]$ bounded interval will need.

4.2 Model

4.2.1 Kernel

The kernel $\tilde{k}$ used in the mixture model is defined by the product of two kernels, $\tilde{k}_\theta(\cdot|u) = g_\theta(\cdot)k(G_\theta(\cdot)|\Psi)$. A generalized Pareto distribution (GPD) is used for g_θ; its corresponding distribution function is represented by G_θ. The parameters of $\theta = \{\nu = 1/\xi, \sigma\}$ correspond to the power of the decay (the reciprocal of the tail index) and the scale parameter of the GPD respectively. The family of normal distributions truncated to the fixed interval $[0, 1]$ is used for the second kernel, $k(\cdot|\Psi)$. Truncated normal densities have a range strictly between $(0, \infty)$ and therefore, when mixed together to create h, maintain the boundary constraint necessary for implementation of the extreme value framework explained in Section 1.5. Each truncated normal kernel has parameters $\Psi = (\mu, \beta)$ representing its pre-truncation mean and standard-deviation (also called bandwidth in this chapter), respectively.

4.2.2 Priors

The priors for θ, comprised of ν and σ, of the GPD are the same as those used in Chapter 3 and pictured in Figure 3.1 of that chapter. The prior for ν, $\ln(\nu - 0.5) \sim$ Logis$(2, 1)$, puts prior mass over heavy-tailed values. The prior for σ is independent

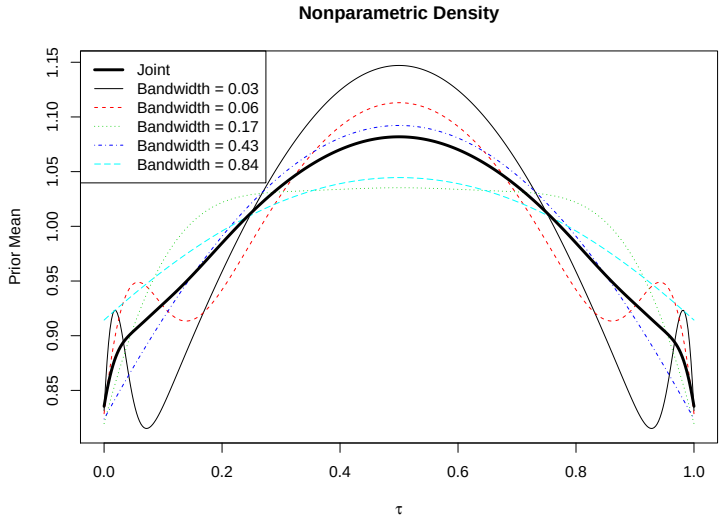


FIGURE 4.1: Prior predictive mean for nonparametric density, jointly over μ and β and also conditionally for a few bandwidth, β, values.

of ν and is a half-Cauchy distribution with scale elicited from the user's knowledge of the data mean, $\tilde{y}$, specifically $\sigma \sim \text{HC}(3/4 \times \tilde{y})$. As before, the prior for parametric θ is assumed to be independent of nonparametric $h(\tau)$.

The priors for truncated normal kernel parameters Ψ, comprised of μ and β, are defined jointly by $\pi(\mu, \beta) = \pi(\mu|\beta) \cdot \pi(\beta)$. As in chapter 3, the prior over bandwidths is an InverseGamma(shape=0.01, rate=0.0005) truncated to an upper boundary of 1. This results in a median bandwidth of 0.166. Over its $[0, 1]$ support, a truncated normal kernel with $\mu = 0.5$ begins to look very uniform when β is large. Truncating β at 1 helps to avoid having many integration evaluation nodes with redundant uniform-like truncated normals feeding into the nonparametric density estimate. Given bandwidth β, the prior for center μ is $\mu|\beta \sim$ TrunNormal(mean=0.5, sd=0.5, a$= -\beta/2$, b$=1+\beta/2$). Conditioning this truncated normal on β allows the kernel centers (their modes) to exist half a standard-deviation outside the $[0, 1]$ interval, which in turn allows the nonparametric density estimates to "go up" at the boundaries. It also prevents computational loss of numerical sig-

nificance that can occur if the mean, μ, falls many bandwidths, β, outside of the interval $[0, 1]$ where the data live. Figure 4.1 shows that the resultant mixture density prior is not quite uniform, though it is close, peaking in the middle and biasing slightly downward at the boundaries. Coupling this near-uniform mean with high prior density over small bandwidths should encourage global flatness while allowing for local adaptation.

4.2.3 Joint Posterior

From an empirical Bayes view, the PR mixing distribution $p_{n,\theta}(\Psi)$ can be thought of as the posterior for Ψ conditional on θ, $\pi(\Psi|\theta, Y_{1:n})$. As mentioned in Section 4.1.2, the marginal likelihood is approximated by $L_n^M(\theta) = \prod_{i=1}^n f_{i-1,\theta}(y_i)$. Incorporating prior $\pi(\theta)$ into a simple application of Bayes theorem gives the approximation to the marginal posterior for θ up to a normalizing constant:

$$\pi(\theta|Y_{1:n}) \propto L_n^M(\theta) \cdot \pi(\theta). \tag{4.5}$$

The joint posterior across all transformation and mixing parameters is then easily obtained by $\pi(\Psi, \theta|Y_{1:n}) = \pi(\Psi|\theta, Y_{1:n}) \times \pi(\theta|Y_{1:n})$. And a posterior predictive can be found via

$$\pi(Y_{n+1}|Y_{1:n}) = f_n(Y) = \int \left\{ \int \tilde{k}_\theta(y|\Psi)\pi(\Psi|\theta, Y_{1:n})d\mu(\Psi) \right\} \pi(\theta|Y_{1:n})d\nu(\theta) \tag{4.6}$$

Integrating the joint posterior over θ provides insight into the marginal posterior over mixing parameters: $\pi(\Psi|Y_{1:n}) = p_n(\Psi) = \int_\theta \pi(\Psi|\theta, Y_{1:n})\pi(\theta|Y_{1:n})d\nu(\theta)$. This becomes useful for looking at the posterior predictive of the nonparametric, truncated-normal kernel mixture density, $h_n(\tau) = \int k(\tau|\Psi)p_n(\Psi)d\mu(\Psi)$, if desired.

4.3 Computation

Computation is performed in R with calls to C to perform the PR recursions. This section details and justifies additional computational decisions related to the PR

algorithm, such as the method of numerical integration and the choice of recursion weights; approximations made to the marginal posterior over θ; and decisions related to estimating posterior quantities along with their uncertainty intervals.

4.3.1 Numerical Integration

Each recursive update of the mixture density f_i in PR requires an integral to be calculated over the kernel's mixing parameters, in this case over the truncated normal mean and standard deviation (μ, β). Numerical integration via Gaussian quadrature—or "cubature" as it is sometimes called for more than one dimension, as we have—is used.

The `mvQuad` package from R is used to obtain Gauss-Leguerre knots or grid-points and associated weights over $[0, 1]$ intervals for each parameter. Transformations coming from the prior quantiles (or inverse CDFs) take the integrals from $[0, 1] \times [0, 1]$ back to the joint domain: $0 < \beta < 1$ with $-\beta/2 < \mu < 1 + \beta/2$. The choice of the prior-quantile transformation also allows the prior described in Section 4.2.2 to be defined simply as a uniform distribution over the $[0, 1] \times [0, 1]$ domain.

While a two-dimensional product grid is easy to setup for quadrature, it is not particularly efficient for our setting. In order to capture spikes or multimodality in the nonparametric density, it is desirable to have many narrow kernels evaluated in the grid; however, wide kernels evaluated at the same centers as the narrow kernels would overlap a lot with other wide kernels and be redundant. Consequently, we implement an irregular grid to increase computing efficiency: a denser grid is used for smaller β bandwidths and a sparser grid is used for larger β bandwidths. The irregular spacing of the Gauss-Leguerre knots prevents precise discussion of the placement of kernel means, μ, but generally speaking, for given β, the number of grid points is selected so that kernels overlap by about two standard deviations or bandwidths.

4.3.2 Predictive Recursion Weights

Weight sequences of the class $w_i = (i + c)^{-\gamma}$ with $\gamma \in (1/2, 1]$ are suggested in the PR literature. Specifically, Martin and Tokdar (2009) prove PR convergence in total variation to the true mixture distribution, when appropriately specified, and convergence to the mixture which minimizes Kullback-Leibler distance when the model is misspecified and does this for $\gamma \in (2/3, 1]$. With little guidance given in the literature about how to choose c or γ, empirical evaluation guides our choice of appropriate values.

The value c performs a role similar to the precision parameter in a DPMM. If c is set too low (e.g. 0.5 or 1), the prior washes out quickly from the PR estimation and produces no shrinkage towards the prior. If set too high (e.g. 10^6), the posterior is biased towards the mean shown in Figure 4.1. Moderate values ($c = n/10$) encourage shrinkage towards the nonparametric prior mixing values. An empirical evaluation across a variety of target shapes shows that $\gamma = 1$ weights increase the variability in estimated h across the $[0, 1]$ interval. Variance decreases for decreasing γ up to a point, but by $\gamma = 0.5$ unwanted bias and some associated variability are introduced. Ultimately, $\gamma = 0.66$ is used as a compromise in the bias-variance trade-off.

4.3.3 Discretization of Marginal Prior

The notion of an approximate marginal posterior over $\theta = \{\nu, \sigma\}$, $\pi(\theta|Y_{1:n})$, was introduced in Section 4.2.3. For practical computing reasons (e.g. to avoid MCMC or otherwise evaluating the normalizing constant for the posterior), the marginal prior and posterior are evaluated over a finite, discretized two-dimensional product grid. The default evaluation points are determined by prior quantiles of ν and σ, and prior probabilities are determined at each evaluation point from the cumulative prior probabilities within a bounding box surrounding the point. Heat maps of the marginal posteriors provide insight into where posterior probability lies and may

prompt changes to the evaluation grid to "zoom in" on areas of higher probability. In addition to the default-prior quantile grid, the R code supports a square product grid over user-supplied independent sequences for ν and σ, which may be equispaced or non-equispaced.

As a consequence of using a square grid for ν and σ, it is possible that large regions of the grid have very low posterior probability. To speed up computation of posterior quantities (see next section), after the marginal posterior has been be computed, grid points with low posterior probability can be removed and the remaining posterior renormalized. This thresholding ability is also built into the R code.

4.3.4 Estimation Including Uncertainty Intervals

Estimates for the posterior predictive sampling density and distribution functions can be found as outlined in Section 4.2.3 by integrating over θ and Ψ. The discretization of θ makes this straightforward, as posterior mean estimates are probability-weighted sums of conditional $f_{n,\theta}(y)$ or $F_{n,\theta}(y)$ over all θ.

Conditional quantiles $F^{-1}_{n,\theta}(p)$ for quantile level, p, are found by linear interpolation at the nonparametric-distribution level. That is, for some dense τ-prediction grid and corresponding estimated $H_{n,\theta}(\tau)$, the linear interpolant $H^{-1}_{n,\theta}(p)$ is sent through the parametric transformation, $G^{-1}_{\theta}(\cdot)$, to get composite $F^{-1}_{n,\theta}(p) = G^{-1}_{\theta}(H^{-1}_{n,\theta}(p))$. Interpolation error is reduced by performing the interpolation over nonparametric H rather than the power-law-influenced F directly. Posterior means of predicted quantiles can then be found by weighting conditional $F^{-1}_{n,\theta}(p)$ by marginal posterior probabilities $\pi(\theta|Y_{1:n})$. Separate marginal estimates for ν and σ can be calculated as $\pi(\theta|Y_{1:n})$-weighted sums of ν and σ over θ, if desired.

Interval estimates for the quantities mentioned above—marginal estimates for ν and σ and posterior predictives for the sampling density, distribution, and quantile functions—also require weighting across the discrete marginal posterior probabili-

ties $\pi(\theta|Y_{1:n})$. We choose a sample-based, weighted-quantile form which linearly-interpolates (p_k, x), where x is a vector of ordered observed values; w is a vector of x-ordered observed weights; k is a vector of cumulative (x-ordered) weights; and $p_k = (k - w/2)/n$. This weighted quantile function, in some ways the weighted analogue to R's sample quantile function with "type = 5", provides a conservative (wide) interval while still providing low bias under a variety of weighting scenarios. However, these weighted quantiles only account for posterior uncertainty across θ and not uncertainty across the mixing parameters Ψ from the predictive recursion.

Estimating uncertainty intervals from PR is not as straightforward as it would be if Ψ were estimated via MCMC; PR, by its nature, only provides the mean of the approximated DPM distribution. While the PR theory proves asymptotic consistency for the mixing distribution Ψ, it technically does not yet cover the pointwise consistency for the mixture distribution. Dixit and Martin (2019) argue that permutation-based approximations to the sampling distribution, which are already calculated to reduce the algorithm's dependence on the data ordering, can also be used to provide interval estimates for the PR mixture distribution, i.e. through approximate $100(1 - \alpha)\%$ confidence intervals based on repeated sampling. Their numerical results show that nominal interval coverage is approximately valid in regions where the sampling density has low curvature but that it can depart significantly from nominal levels in regions where the sampling density has high curvature.

Posterior uncertainty from θ is incorporated using weighted quantiles and Ψ-PR uncertainty is captured using permutation variability as suggested in Dixit and Martin (2019). There are several ways in which marginal θ posterior probabilities and permutations can be incorporated into joint mixture-density interval estimates:

1. Weighted-Quantiles of Mixture-Means. The conditional mixture distributions are averaged across permutations, and the marginal θ posteriors are averaged

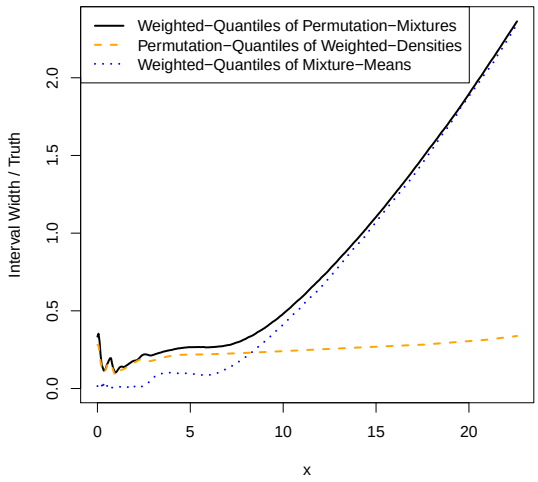


FIGURE 4.2: Comparison of three options for obtaining posterior credible interval estimates of the density function, each incorporating the marginal-θ variability and the conditional-PR-mixture-density variability.

across permutations. Joint mixture density intervals are calculated, pointwise across x, as weighted quantiles of the mean mixing distribution, weighted by the mean marginal posterior.

2. Permutation-Quantiles of Weighted-Densities. The per-permutation conditional mixture distributions are weighted by their respective per-permutation marginal θ posteriors to get per-permutation joint mixture distributions. Joint mixture density intervals are calculated, pointwise across x, as (unweighted) quantiles across the per-permutation joint mixture distributions.

3. Weighted-Quantiles of Permutation-Mixtures. Joint mixture density intervals are calculated, pointwise across x, as weighted quantiles of the per-permutation conditional mixture distributions, weighted by their respective per-permutation marginal θ posteriors.

Figure 4.2 shows interval widths divided by truth under the three proposed

Table 4.1: Simulation scenario sample size, ν-σ bounding box, number of evaluation points in ν-σ grid, number of integration nodes in μ-β grid (number of β bandwidths), and time to run each dataset in minutes while distributing ν-σ pairs across six cores.

Simulation Scenario	n	ν-σ box	ν-σ grid	μ-β (β)	time
Standard GPD	1,000	$(1, 9) \times (0.2, 2.4)$	$41 \times 45 = 2214$	250 (8)	0.98
Half-t	1,000	$(2, 13) \times (0.3, 1.7)$	$45 \times 29 = 1305$	250 (8)	0.67
Fourth-power GPD	1,000	$(1.5, 12) \times (0.3, 4.8)$	$36 \times 31 = 1116$	250 (8)	0.58
Gamma-GPD mixture	1,000	$(1.2, 14) \times (1, 6)$	$33 \times 34 = 1122$	498 (15)	1.09
Half-t-normal mixture	2,000	$(1.5, 12) \times (0.5, 4.5)$	$36 \times 27 = 972$	250 (8)	2.47
Spliced gamma-GPD	5,000	$(3.5, 18) \times (0.9, 2.2)$	$30 \times 27 = 810$	250 (8)	2.80

joint interval methods for a single jointly-estimated dataset coming from a standard GPD distribution, using 200 permutations. The permutation-quantiles of weighted-densities have wider relative intervals in the bulk of the distribution than do the weighted-quantiles of mixture-means; however, in the tails, their relative interval widths are narrower than the weighted-quantiles of mixture-means. The weighted-quantiles of permutation-mixtures seem to capture the width in both the bulk and the tail of the density, so we use this approach throughout.

4.4 Simulation Study

Chapter 2 introduces six simulation scenarios that are reused here to test the predictive recursion model for the extreme value transformation framework. As in previous chapters, the method under consideration (PRMP) is compared to maximum likelihood estimates for a GPD or an extended GPD (EGPD) under various truncations primarily on the basis of their tail-index bias and interval coverage and on their upper-tail quantile bias and RMSE.

4.4.1 Simulation Settings

For each dataset, 20 permutations were employed to reduce dependence on data ordering and to estimate variability. Several datasets were fit for each scenario using

the default, prior-quantile grid in order to give an idea of the bounding box needed to capture the posterior mass without spending computing power on regions of near-zero probability, and an appropriate ν-σ region was selected and fixed for all datasets within the scenario. Grid resolutions over the integrating Ψ were also modified per scenario. Most scenarios were evaluated with eight β bandwidths and 250 total integration evaluation nodes; however, fifteen bandwidths with corresponding 498 μ-σ nodes were used for the gamma-GPD mixture, where the nonparametric density was anticipated to be multimodal. Table 4.1 summarizes the settings used. In practice, a user can increase the grid size for either Ψ or θ until the results don't change between resolutions or until computational resources are exhausted.

4.4.2 Simulation Results

In this section we consider the aggregate results across all six simulation scenarios by topic: computational speed, ν-σ marginal posteriors, lower-tail and bulk results, tail-index results, upper-tail quantile-estimation results, and full-distribution quantile interval coverage. Maximum likelihood results for GPD and EGPD are included in plots and tables for reference when comparing tail results but discussed only lightly, as compared to PRMP. For elaboration on why maximum likelihood does well or poorly in various scenarios, see discussions in Subsection 2.3.3 of Chapter 2.

Computational Speed

Computational speed for the PRMP is affected by sample size, number of permutations, number of integration evaluation points over the Ψ grid, and number of evaluation points over the discretized θ grid. Table 4.1 includes the average time in minutes that the PRMP takes to run the scenario for a single dataset, distributing the conditional θ evaluations across six cores. These range from 0.58 minutes for the fourth-power GPD scenario ($n = 1{,}000$ and approximately 1,100 θ grid points) to 2.8

minutes for the spliced gamma-GPD scenario (n = 5,000 and approximately 800 θ grid points).

Marginal Posteriors

The marginal posteriors across scenarios are sensitive to the choice of PR weight sequences. Specifically, when c, the analogue to a DPMM precision parameter described in 4.3.2, is fixed at too-small or too-large of values, undesirable effects appear in both the estimation of θ and h. The sequence of plots in Figure 4.3 (split across two pages) elucidate these effects using a single representative dataset from the half-t scenario by considering $c \in \{0.5, 10, 50, 100, 1000, 100000\}$.

In the heat maps for marginal θ (left plots of Figure 4.3), large spread is seen across σ when $c = 0.5$. This spread decreases and the 99.9% HPD region for the θ marginal contracts as c is increased to 10 and again as c is increased to 50. The conditional h densities needed for large-σ tend to put high density at the zero end of the τ range and low density near $\tau = 1$. This shows up in the corresponding marginal h posteriors (right plots) as intervals that go well above 1 at $\tau = 0$ and extend down to 0 at $\tau = 1$. Conversely, smaller values of σ can result in conditional h pushing its mass towards the $\tau = 1$ boundary. In practice, for these small values of c, so much variability exists across the parts of the joint model that quantile intervals widen to the point of becoming essentially meaningless for extrapolation purposes.

When c is small (e.g. $c = 0.5$ or $c = 10$ in the illustrated example), the prior over Ψ washes out quickly from the PR conditional estimates, and h is heavily influenced by default DPM behavior, favoring the fewest clusters possible. Increasing c starts to put more weight on the Ψ prior. The effect of this can be seen in Figure 4.3 (see $c = 50$ and $c = 100$) as estimates for marginal h begin to tighten and variability decreases across σ in the marginal θ posterior. For even larger c (see c = 1,000 and c = 100,000), the conditional mixture density estimates start to bias towards the prior

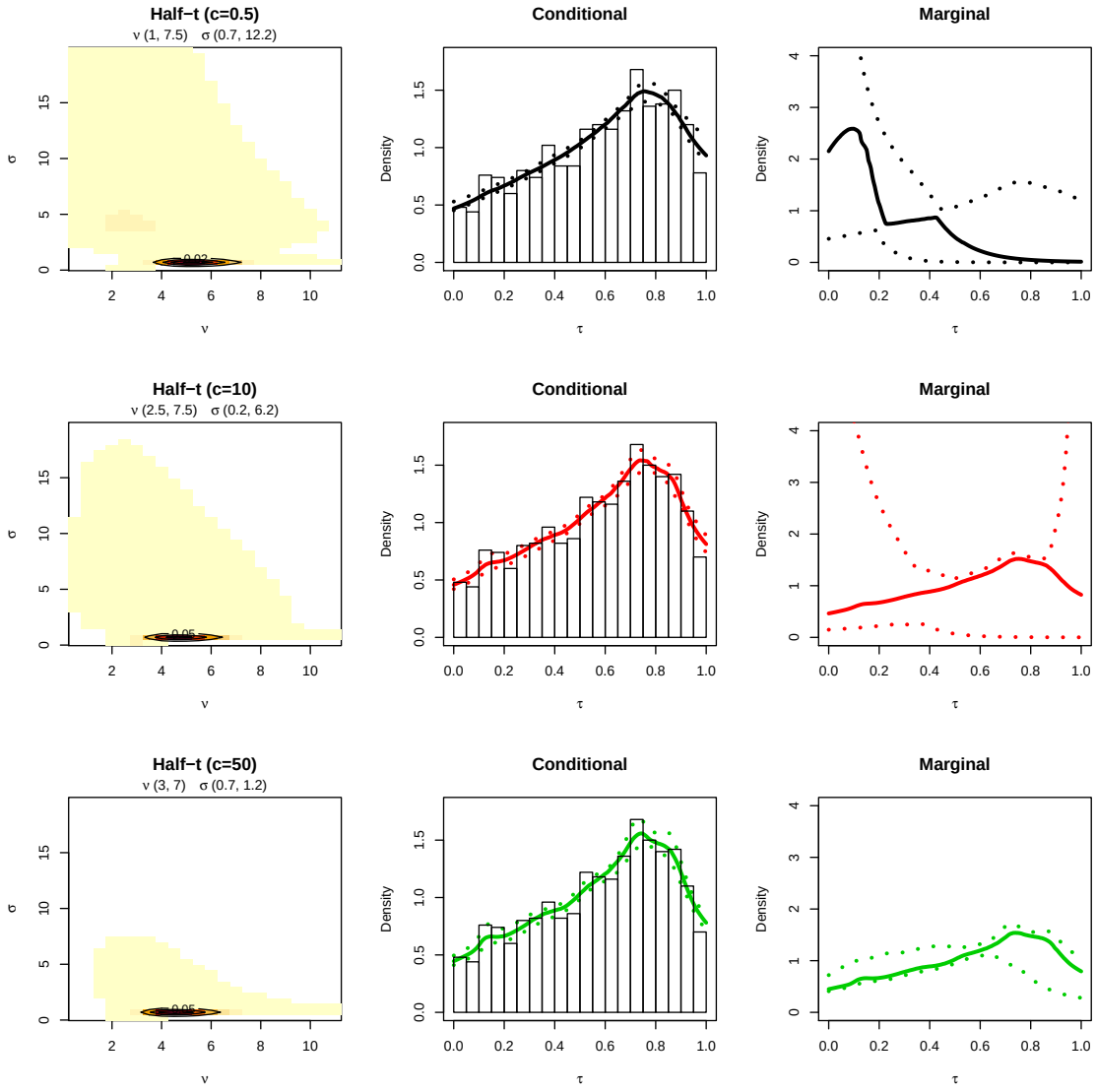


FIGURE 4.3: Each row of plots corresponds to estimates for dataset 1 of the half-t scenario under different values of c used in the PR weights w_i. (Figure continued on next page.) Left: Discrete marginal posteriors over decay power, ν, and scale, σ. Reds show areas of highest probability and yellow shows regions of lower probability. Subtitles show marginal 95% credible intervals; color displays 99.9% bivariate HPD region; and contours ring areas of highest probability. Middle: Medians and 95% intervals for h when conditioned on MAP of marginal θ. Right: Marginal medians and 95% intervals for h, integrating across θ.

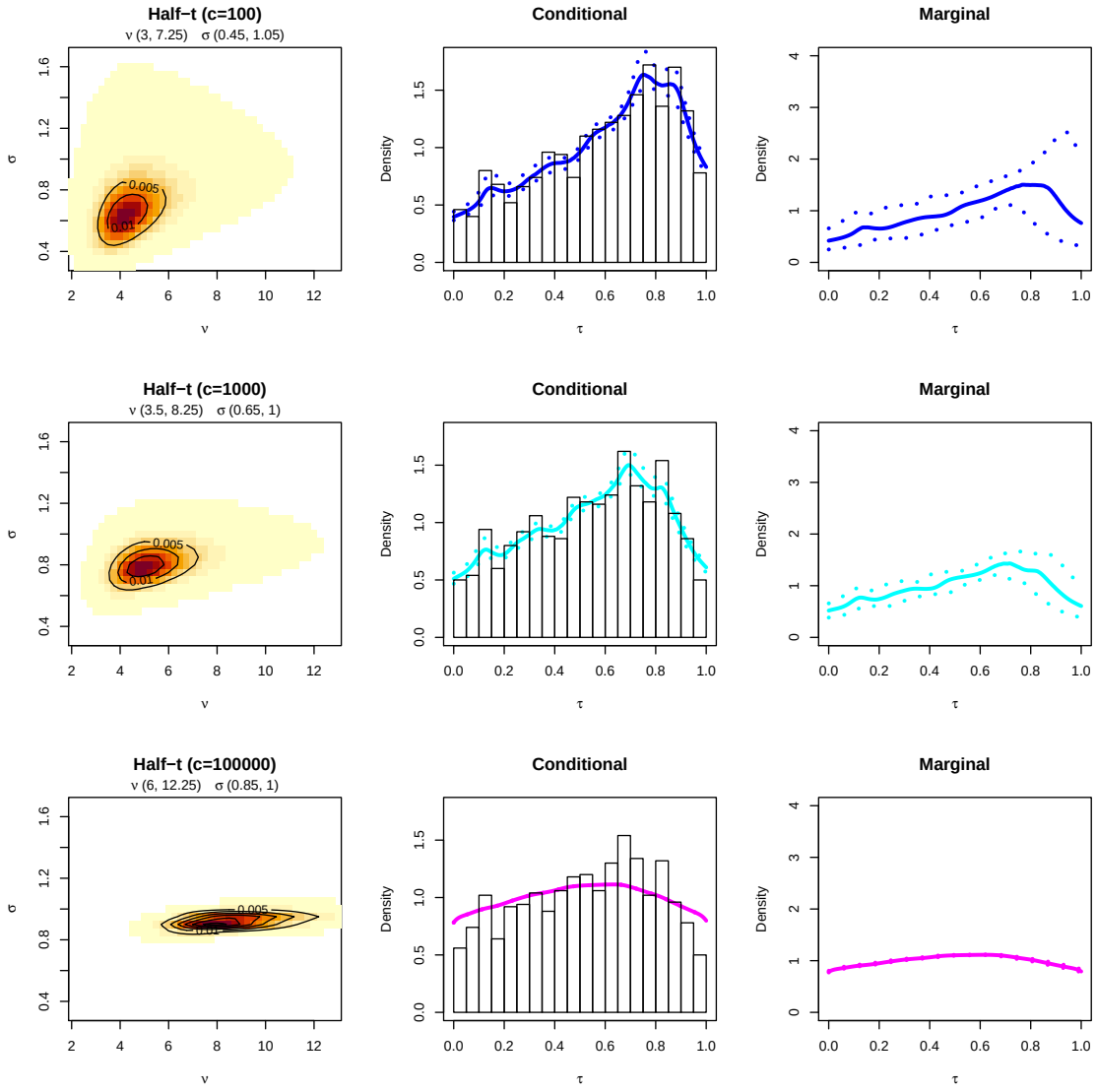


FIGURE 4.3: Each row of plots corresponds to estimates for dataset 1 of the half-t scenario under different values of c used in the PR weights w_i. (Figure continued from previous page.) Left: Discrete marginal posteriors over decay power, ν, and scale, σ. Reds show areas of highest probability and yellow shows regions of lower probability. Subtitles show marginal 95% credible intervals; color displays 99.9% bivariate HPD region; and contours ring areas of highest probability. Middle: Medians and 95% intervals for h when conditioned on MAP of marginal θ. Right: Marginal medians and 95% intervals for h, integrating across θ.

Table 4.2: Relative bias (bias / truth) for each PRMP scenario at the median, $p = 0.5$, representative of the bulk, and at $p = 0.00001$ quantile level, which for each scenario had the largest relative bias in the lower "tail."

Simulation Scenario	Median	Lower-tail
Standard GPD	0.0061	0.132
Half-t	0.0022	0.082
Fourth-power GPD	-0.0022	-0.994
Gamma-GPD mixture	-0.0019	0.119
Half-t-normal mixture	0.0000	0.074
Spliced gamma-GPD	0.0030	-0.032

mixture density mean, resulting in biased marginal h densities and correspondingly biased marginal θ posteriors.

Evaluation of each simulation scenario at $c = 0.5$ showed multimodality or long-σ tails in the marginal-θ posteriors of the half-t, gamma-GPD mixture, and spliced gamma-GPD scenarios. Aided by these empirical evaluations, c was set to $n/10$ across scenarios, acknowledging that c is only large or small relative to sample size n. This results in unimodal marginal-θ posteriors for all scenarios. Figure 4.4 contains a marginal posterior over discrete ν and σ for each scenario, using a single, representative dataset displayed over its bounding box. Permutations within a dataset tend to follow the pattern of the overall average, as opposed to having certain permutations with high probability in one area and other permutations with high probability in another area.

PRMP Lower-tail and Bulk Results

The primary focus of analysis is on tail estimation; however, Table 4.2 is included to summarize PRMP relative bias (bias / truth) for each scenario at the median as well as at the lower quantiles' point of highest relative bias, which in each scenario occurs at the smallest estimated quantile level, $p = 0.00001$. Simulations are essentially unbiased for the median; the largest relative bias among the six scenarios is 0.0061.

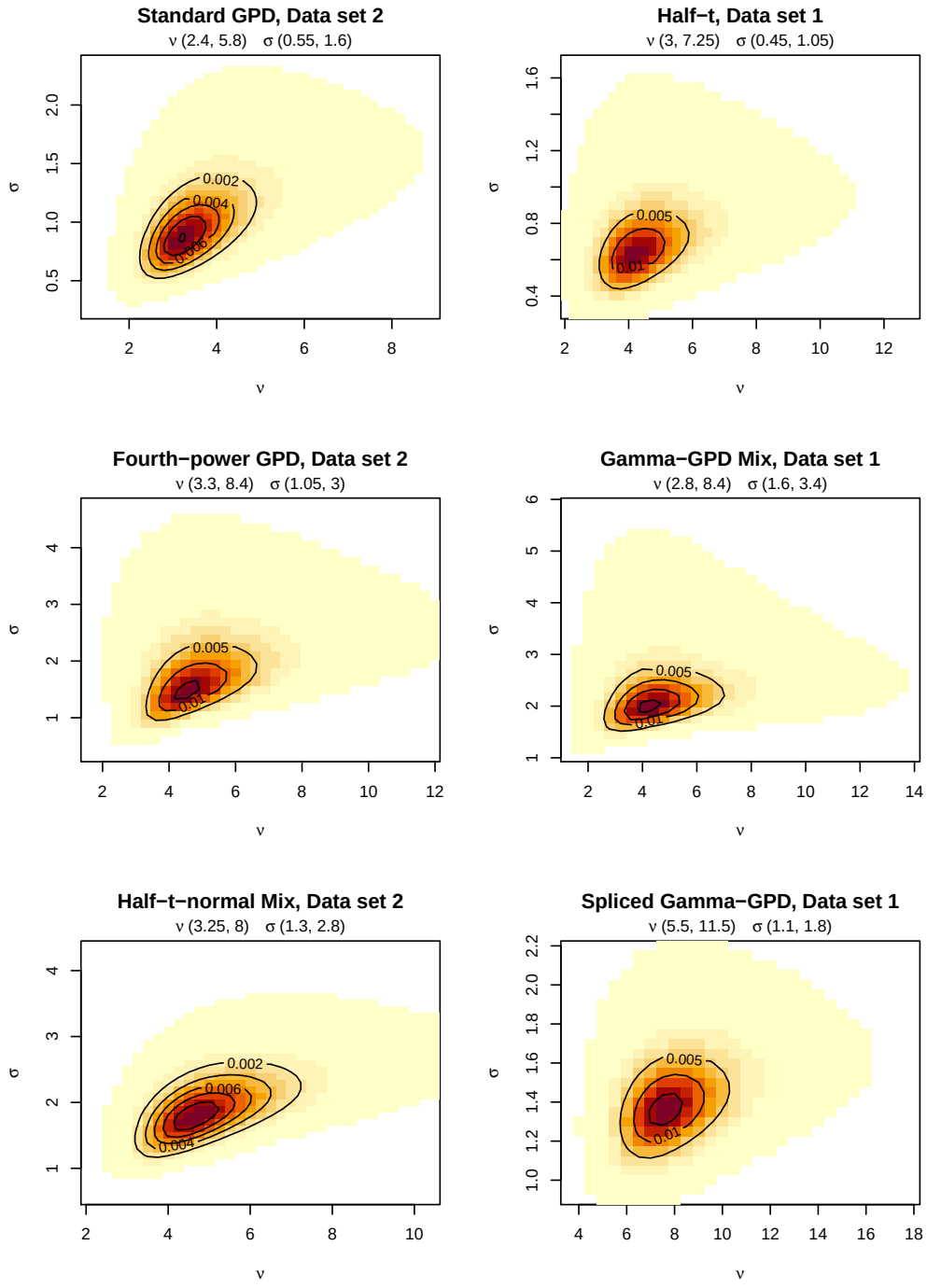


FIGURE 4.4: Discrete marginal posterior examples over decay power ν and scale σ.

The fourth-power GPD has the poorest lower-tail relative bias, under-estimating the 0.00001 quantile by nearly 100%. This happens, in part, because the PRMP inherits its strictly positive density at 0 from the GPD base g_θ, whereas the true value $f(0) = 0$. This results in PRMP hitting its 0.00001-level quantile at much smaller values than the true fourth-power GPD. A similar under-estimation of quantile bias can be seen in the spliced gamma-GPD, which truth also has a single mode pulled away from the boundary. While both true $f(0) > 0$ and also $g_\theta(0) > 0$ for this scenario, the nonparametric density does an insufficient job scaling down the GPD transformation density in the lower tail, resulting in too-large estimates of the density and correspondingly too-small estimates of the lower quantiles. The standard GPD and gamma-GPD mixture, both with bulk GPD distributions, seem to have the opposite problem; they underestimate the density near 0 leading to overestimates of the quantiles with 13.2% and 11.9% relative bias, respectively.

Tail-Index Estimation

Turning to tail-index estimation, Table 4.3 shows that in all but the standard GPD scenario, PRMP underestimates the tail index. In all but the spliced gamma-GPD scenario, where the PRMP is biased with too-narrow widths leading to 0% coverage, the PRMP tail index interval widths are narrower *and* have better coverage than their ML counterparts that require thresholding.

A special note is due to the gamma-GPD mixture scenario, where PRMP does a good job of capturing the tail index in comparison to the GPD or EGPD truncations. This is likely because the true "bulk," which is actually a GPD, informs the tail-index estimation despite its "contamination" by gamma-component observations. This could be indicative of how the model would perform if applied only to truncated tail values in a data-rich scenario where some of those tail observations are suspected to come from an alternative non-heavy-tailed processes.

Table 4.3: Simulation tail-index results. Tables includes mean tail-index estimates for each method and coverage of 95% confidence or credible intervals, i.e. proportion of intervals across simulated datasets that contain the true tail-index.

Standard GPD

Method	Mean	Width	Cover
Truth	**0.25**		
GPD 5%	0.19	0.73	0.94
GPD 10%	0.25	0.52	0.95
GPD 20%	0.26	0.35	0.97
GPD 100%	0.25	0.15	0.93
EGPD 5%	0.17	0.94	0.94
EGPD 10%	0.22	0.63	0.95
EGPD 20%	0.26	0.43	0.93
EGPD 100%	0.25	0.18	0.94
PRMP 100%	0.29	0.26	0.97

Half-t

Method	Mean	Width	Cover
Truth	**0.25**		
GPD 5%	0.18	0.72	0.92
GPD 10%	0.18	0.47	0.91
GPD 20%	0.16	0.32	0.79
EGPD 5%	0.19	0.92	0.93
EGPD 10%	0.20	0.59	0.88
EGPD 20%	0.17	0.39	0.86
PRMP 100%	0.19	0.18	0.91

Fourth-power GPD

Method	Mean	Width	Cover
Truth	**0.25**		
GPD 5%	0.18	0.72	0.88
GPD 10%	0.21	0.50	0.92
GPD 20%	0.21	0.34	0.90
EGPD 5%	0.19	0.93	0.88
EGPD 10%	0.20	0.61	0.89
EGPD 20%	0.22	0.41	0.89
PRMP 100%	0.22	0.21	1.00

Gamma-GPD mixture

Method	Mean	Width	Cover
Truth	**0.25**		
GPD 5%	0.43	0.78	0.92
GPD 10%	0.25	0.43	0.87
GPD 20%	0.01	0.78	0.07
EGPD 5%	0.50	0.95	0.85
EGPD 10%	0.32	0.51	0.90
EGPD 20%	0.11	0.25	0.48
PRMP 100%	0.21	0.25	1.00

Half-t-normal mixture

Method	Mean	Width	Cover
Truth	**0.25**		
GPD 5%	0.18	0.48	0.87
GPD 10%	0.17	0.32	0.80
GPD 20%	0.12	0.21	0.38
EGPD 5%	0.17	0.59	0.89
EGPD 10%	0.19	0.40	0.90
EGPD 20%	0.14	0.26	0.63
PRMP 100%	0.23	0.19	0.98

Spliced gamma-GPD

Method	Mean	Width	Cover
Truth	**0.25**		
GPD 5%	0.23	0.31	0.94
GPD 8%	0.24	0.25	0.95
GPD 13%	0.21	0.18	0.85
EGPD 5%	0.22	0.38	0.93
EGPD 8%	0.24	0.30	0.94
EGPD 13%	0.24	0.22	0.96
PRMP 100%	0.10	0.08	0.00

Upper-tail Quantile Estimation

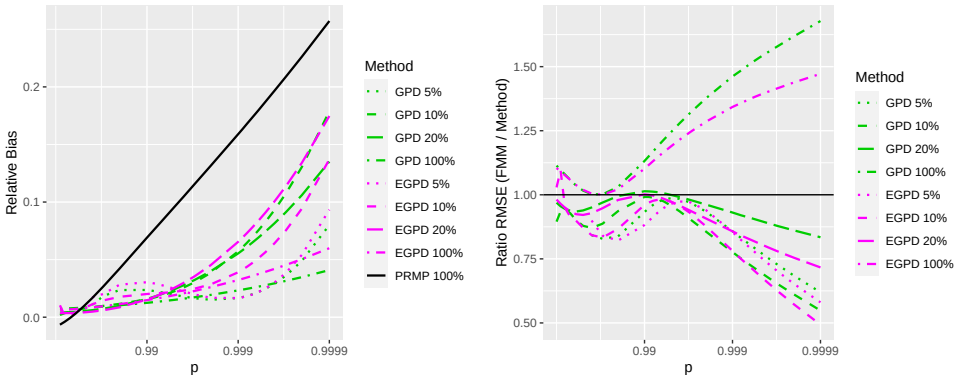
In all six of the simulations, PRMP exhibits positive bias in the tails, as can be seen in the relative bias plots of Figure 4.5. The biases can be large, even for levels where there should be some informing data, e.g. $p < 0.999$. The bias in GPD and EGPD over these data-informed quantile levels is small in comparison.

Despite these biases, the PRMP performs better than GPD or EGPD in RMSE for several scenarios, as can be seen in the plots of Figure 4.5. In the standard GPD scenario, PRMP does not do as well as ML for either the GPD or EGPD that retain all 100% of the data, but it has consistently lower RMSE across all tail levels $p > 0.9$ for any of the truncations considered. The bias in the half-t scenario translates to high RMSE in the $0 < p < 0.999$ range, but in extrapolation ($p > 0.999$) PRMP still has lower RMSE than the truncation methods. The fourth-power GPD scenario has lower RMSE than the truncation methods across essentially all tail levels.

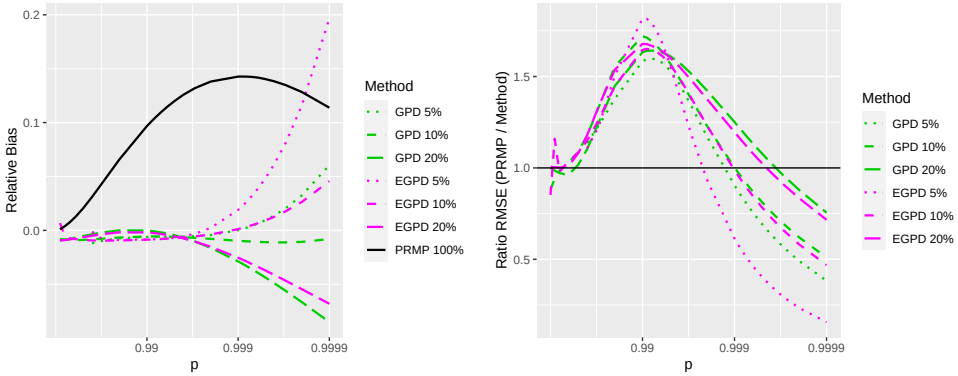
The remaining three scenarios (Figures 4.5d, 4.5e, 4.5f) do poorly in comparison to the GPD and EGPD methods across most tail levels, except in the highest levels of extrapolation ($\approx p > 0.999$), where they start to exhibit slight advantage. This is true, even for the gamma-GPD mixture scenario, which had success in estimating the tail-index parameter, indicating that the problem might lie with the scaling of the nonparametric density estimation. This requires further exploration.

Table 4.4 shows the proportion of mean squared error (MSE) that is attributable to squared bias at $p = 0.9999$ in each scenario. Contrary to what might be expected for estimation in a scarce-data, tail scenario, variance does not necessarily dominate the MSE, but rather bias contributes a large proportion. This is true for all but the fourth-power GPD, which is unique among the scenarios in that its preferred nonparametric shape has a positive slope at the right-hand boundary. Taking a closer look at the conditional nonparametric densities, we notice PR can have difficulty

(a) **Standard GPD scenario**



(b) **Half-t scenario**



(c) **Fouth-power GPD scenario**

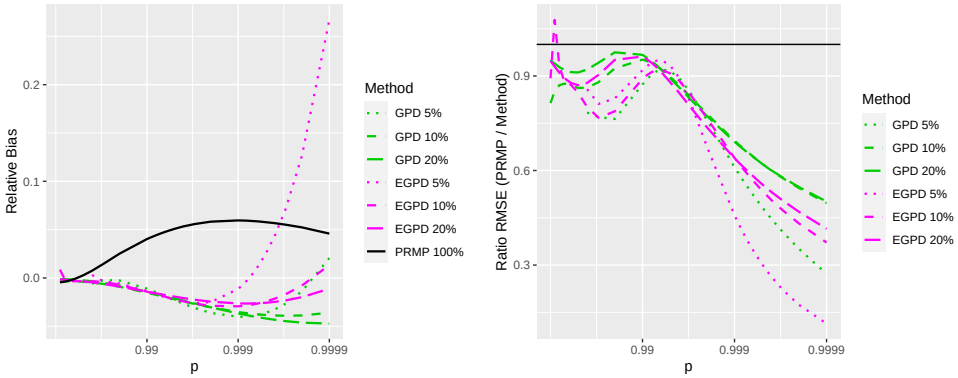
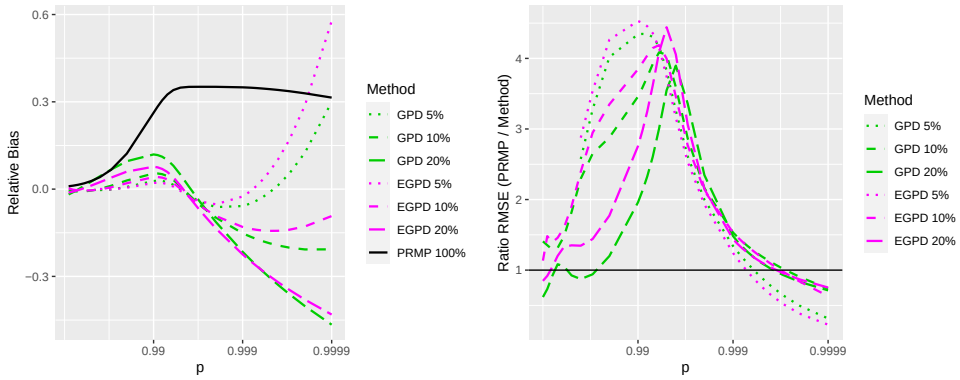
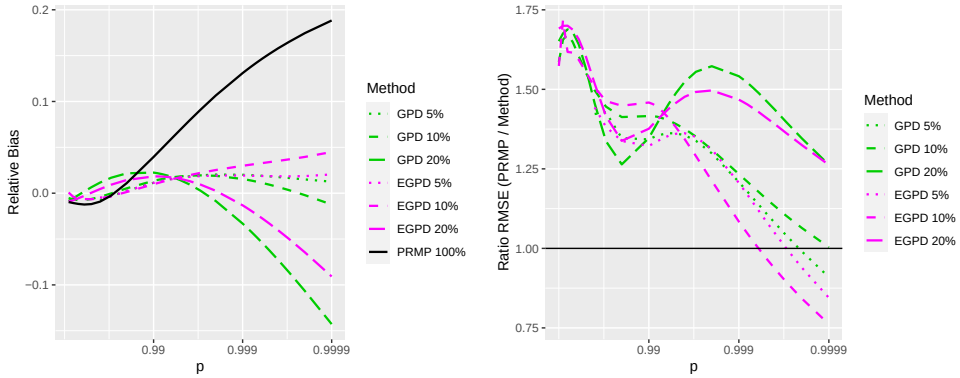


FIGURE 4.5: Upper-tail, quantile-extrapolation with p on log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of PRMP 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than PRMP.

(d) **Gamma-GPD mixture scenario**



(e) **Half-t-normal scenario**



(f) **Spliced gamma-GPD scenario**

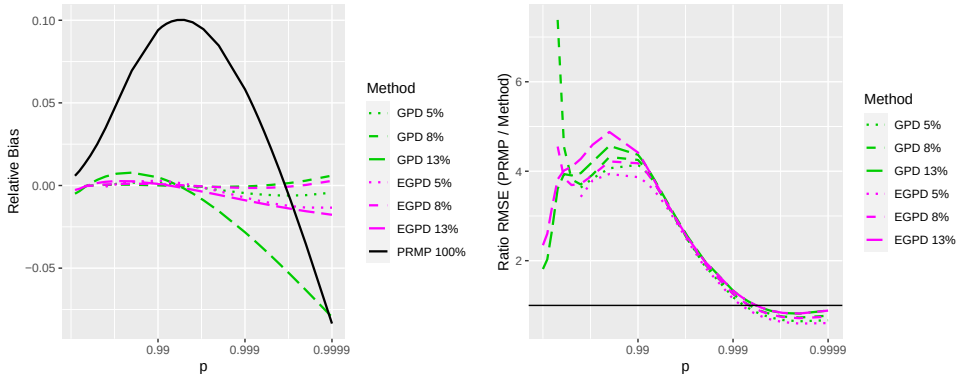


FIGURE 4.5: Upper-tail, quantile-extrapolation with p on log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of PRMP-GP 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than PRMP.

Table 4.4: Proportion of mean squared error attributable to (squared) bias at $p = 0.9999$.

Simulation Scenario	Proportion
Standard GPD	0.450
Half-t	0.415
Fourth-power GPD	0.061
Gamma-GPD mixture	0.822
Half-t-normal mixture	0.437
Spliced gamma-GPD	0.415

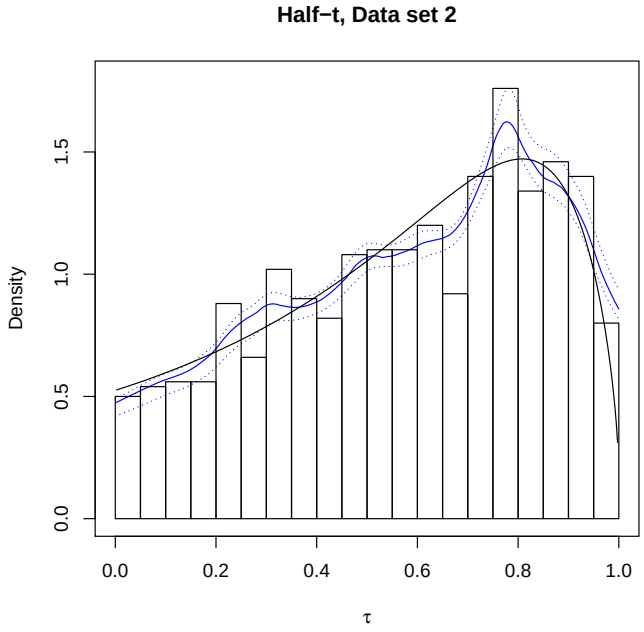


FIGURE 4.6: Example nonparametric estimates along with permutation-based 95% intervals (blue), given ν and σ at their marginal MAPs. True nonparametric distribution, given the same ν-σ pair, is given in black.

capturing rapid drop-offs in the nonparametric density in the right-hand tail. For instance, Figure 4.6 shows an example of an estimated nonparametric PR density (blue) conditioned on the ν and σ pairs which maximize the marginal-θ posterior. For comparison, the true nonparametric density that is needed, again given the MAP ν and σ estimates, is plotted in black. The boundary bias decreases with increasing sample size, but at sample sizes that we might want to consider (e.g. $500 < n < 50{,}000$) the biases are still readily apparent. If each conditional nonparametric density is poorly estimated, some positive bias will cancel with negative bias when weighted across the θ marginal posterior, but the joint estimates may still be biased, as seen across scenarios in Figure 4.5.

The PR boundary bias may be due in part to 1) the locally-symmetric nature of the truncated normal mixing kernel and 2) the inherent nature of there being less data near the boundary. These two problems are simultaneously visible in the half-t plot of Figure 4.6, where the estimate (blue) follows the truth (black) well in the 0.8 to 0.9 region but poorly in the 0.9 to 1.0 region. PR gets its shape near 1 from the wide kernels defining the global shape of the nonparametric density, failing to use local narrow-bandwidth kernels to capture the boundary drop. In other words, PR favors the shape needed to fit the data-rich interior values over the shape needed in data-poor boundary region. Additionally, because the kernel is locally symmetric about μ, there is no dispensation for the density to the right of the mode to fall off differentially to the density left of the mode: the tail is again overwhelmed by the bulk. We address potential remedies to this in the conclusions.

Quantile Interval Coverage

While quantile interval coverage has not been a highlighted objective of this work, we take a moment to make some general observations about the intervals. For mid-range quantile levels ($0.20 < p < 0.80$), the quantile interval coverage falls in the

90% to 100% range across scenarios. The lower quantiles $p < 0.20$ tend to have the worst coverage, dropping to essentially 0% at $p = 0.00001$ for the fourth-power GPD and into the 55% to 90% range for the other scenarios. The upper tails fare better in quantile coverage than the lower tails but still drop from their mid-quantile-level highs. After an initial drop, they rebound into the extreme-most tails. This rebound is likely due to a boost in variance from estimating the tail-index parameter. Two of the scenarios have notable drops and rebounds. The spliced gamma-GPD drops from 97% at $p = 0.9$ to 6% at $p = 0.98$ before rebounding to 77% at $p = 0.9999$, and the gamma-GPD mixture drops from 92% at $p = 0.9$ to 0% at $p = 0.98$ before rebounding to 92% at $p = 0.9999$. Some of the decreased coverage across quantile levels may be due to reduced permutation variability in areas with high curvature in the densities being estimated (see Section 4.3.4); but if this were fully the case, a scenario such as the gamma-GPD mixture should have a local coverage minimum near $p = 0.86$, where the true density hits its local mode from the gamma component of the mixture. Instead it sees its poorest coverage at $p = 0.98$. Ultimately, we believe the decreased coverage in both tails is, again, attributable to the finite-sample boundary biases of the nonparametric density estimates.

4.5 Discussion and Conclusions

In this chapter, we have developed a fully-Bayesian model for univariate density estimation in the presence of heavy tails using a parametric transformation framework coupled with a nonparametric predictive recursion approximation to the Dirichlet process mixture model. The PR model proposed for h, which depends upon a truncated normal mixing kernel over $[0, 1]$, exhibits undesirable boundary bias for the finite sample sizes considered in simulation, affecting all aspects of tail estimation.

One might think that excluding larger bandwidth β values from the prior for the truncated normal kernel (i.e. truncating the inverse-gamma prior at a value less than

1) would encourage the nonparametric density to use its small bandwidth kernels to capture boundary drops in the nonparametric density. Preliminary results show that while this restricted prior does reduce some boundary bias, it also increases the variability of the nonparametric estimator, $h(\tau)$, across the rest of τ. A second consequence is that the prior, and thereby the posterior, loses mass over its most uniform-like kernels. Nonparametric h, without this guidance, piles onto one side of τ or another, resulting in problems similar to those seen in Section 4.4.2 with small c. Giving weight to an appropriately-crafted prior turns out to be important to avoid this problem when estimating under the extreme value transformation framework.

An alternative approach to addressing the bias while staying within the PR construct may be to use an asymmetric kernel for the mixture distribution. For example, the uniform-beta-mixture distribution $p + (1-p)\text{Beta}(a, b)$ has the requisite boundary constraints needed for the extreme-value framework, provided $p > 0$, $a > 1$ and $b > 1$. Unfortunately, that kernel uses three parameters, complicating the Gaussian quadrature used for numerical integration. Alternately, a simple $\text{Beta}(a, b)$ kernel with $a \geq 1, b \geq 1$ could be used, provided that the mixing distribution induces non-zero posterior mass at $a = b = 1$ through a point-mass in the prior. Better yet, independent point masses at $a = 1$ and $b = 1$ would allow for kernels that exceed 1 at the boundaries, as well as for the uniform kernel. The PR theory supports continuous-atomic-mixture priors, though the implementation would be tricky over two dimensions.

A strength of the PRMP approach lies in its ability to characterize the marginal posterior over θ and to do so quickly. This helped elucidate the interplay between θ and h in the joint estimation and clarify the need for a strong prior over the nonparametric densities. While all scenarios in this chapter are unimodal in θ, it is also conceivable that a marginal density might be multimodal. For instance, if applying the transformation framework to data over the full real line, one might wish to add

a location parameter to the transformation density. Multimodality in the sampling distribution could easily lead to multimodality in the marginal distribution with separation occurring as the location parameter locates each sampling-density mode. PRMP marginal posterior plots would be helpful in diagnosing such multimodality.

The computational speed of the PRMP makes it a powerful tool for first-blush analysis under the data-retaining, extreme-value transformation framework. While this analysis makes us wary of relying exclusively on PRMP for tail quantile estimates because of bias in the lower tail, the overall PRMP approach shows promise, especially in reducing RMSE in extreme-value extrapolations. Additionally, the marginal posterior plots can be utilized as a tool to diagnose posterior multimodality. This information may prove useful in informing alternative models, especially those relying on computationally-intensive MCMC sampling methods. Chapter 5 applies this and the other models of this dissertation to rainfall data, and Chapter 6 compares the results of PRMP to the MCMC-sampling-based approaches of the LGP and DPMM/FMM models.

5

Rainfall Application

This chapter provides a short illustration of how the models of this dissertation can be applied to real data. The LGP model (Chapter 2), the FMM approach (Chapter 3), and the PRMP approach (Chapter 4) are used to model rainfall accumulation data collected daily at a site in south-west England. For comparison, the tail estimates from the three dissertation models are compared to three other methods: 1) maximum likelihood GPD estimation using a visually-selected threshold, 2) maximum likelihood EGPD estimation of Papastathopoulos and Tawn using a visually-selected threshold, and 3) EGPD2 of Naveau et al. (2016), which was specifically designed to model rainfall with all data included.

5.1 Data

We consider the `rain` data from R's `ismev` package. The full dataset represents daily accumulations of rainfall (mm) at a location in southwest England from 1914 to 1962. Forty seven percent of the observations are zeroes, representing days with no rainfall. Models are fit on the subset of 9,287 non-dry days, i.e. those with greater than 0 mm of accumulation. Return-levels are adjusted post-hoc to reflect full years

of dry and non-dry days. Though the data are temporally dependent, they will be treated as independent observations.

As is often the case with environmental data, these daily accumulations are measured with rounding error. Many days have just 0.3, 0.5, 0.8, or 1 mm of accumulation, and other small rounded values are very common. The framework itself does not admit the possibility of point masses, and difficulties arise in implementation with large number of ties, as the models try to pick up those point masses as well as the empty space between them, resulting in inadequately-smoothed estimations of the density function. To overcome this difficulty, we add a small amount of jitter to the data, uniformly distributing data between their recorded value and the next largest recorded value (or between recorded value and recorded value + 0.3 mm for values greater than 0.5 mm from the next largest record) and run all methods on the jittered data. If the rounding mechanism were known, it could be built into the likelihood formula for each data point.

5.2 Methods

The `mev` R package (Belzile et al., 2018) was used to estimate GPD, EGPD, and EGPD2 tail indices and standard errors, as well as return-levels. The GPD threshold was determined by fitting multiple GPD thresholds and settling on the largest threshold for which the estimates had stabilized yet still had narrow intervals. For this estimator, the threshold was set to 27 mm, representing 230 observations or 2.5% of non-dry days. Similarly the EGPD threshold was set by visual diagnostics over EGPD fits at multiple thresholds. For the EGPD a threshold of 25 mm was used, representing 290 observations or 3.1% of non-dry days. The EGPD2 requires no thresholding; the three-parameter Method 1, which uses a Beta-distribution carrier, is the model used from among Naveau's several methods.

Intervals for the tail index of both EGP methods are formed using the stan-

dard errors provided in the `mev` package, assuming asymptotic normality. The `return.level` function of the R `extRemes` package (Gilleland and Katz, 2016) is additionally used to find return-level interval estimates for GPD. Interval estimates for return levels are not available for the EGPD or EGPD2 in either package.

With plentiful data, $n \approx 9{,}300$, we elect to use 11 knots for the LGP model with the usual quantile-scaled GPD acting as the parametric transformation distribution. Otherwise, the same settings used in simulation are again used here. The LGP method takes 17 minutes to sample 300,000 posterior draws. The first half of these are removed and estimates are based on 1,500 thinned samples.

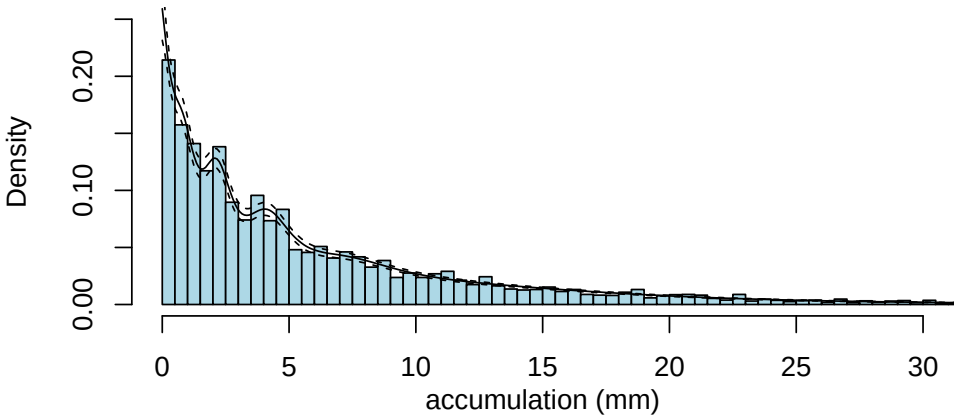
With a large sample size, the FMM is more computationally expedient than the DPMM and therefore used here. Ten mixture model components were employed, consistent with the $\log(n)$ simulations of Chapter 3. The FMM, run under its standard configuration and parameter settings, takes approximately 8.8 hours to sample 200,000 posterior draws. The first half of these were removed and estimates were based on 1,500 thinned posterior samples.

The PRMP takes just under 7 minutes to run distributed across 6 cores, using 8 bandwidths (250 integration grid nodes) and a grid of 756 marginal ν-σ combinations.

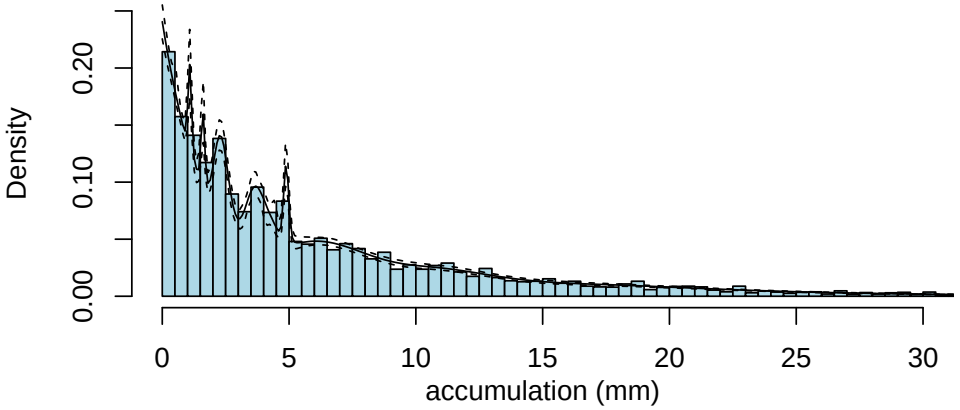
5.3 Estimates

Density estimates. Figure 5.1 displays a histogram of the data overlaid against density estimates, including 95% credible intervals, for each of the methods presented in this dissertation. The plots exclude the tail in order to give a closer look at how each estimator is behaving in the bulk distribution. The LGP and PRMP models employ more smoothing over the bulk distribution than does the FMM, which picks up many peaks in the zero-to-five mm accumulation range.

(a) **LGP density estimates of daily rainfall for non-dry days**



(b) **FMM density estimates of daily rainfall for non-dry days**



(c) **PRMP density estimates of daily rainfall for non-dry days**

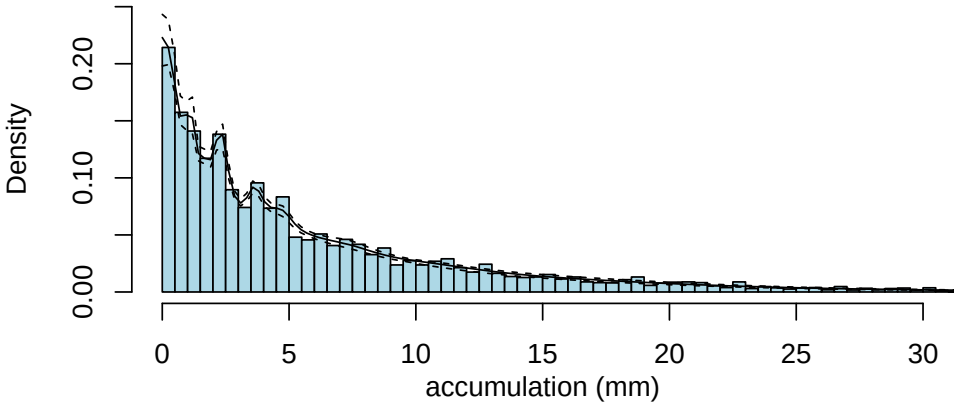


FIGURE 5.1: Application (`rain` data): histogram overlaid with density estimators

Table 5.1: Tail-index estimates and 95% intervals for the `rain` data. Intervals for the GPD, EGPD, and EGPD2 methods are based off asymptotic normality.

Method	Estimate	95% Interval
GPD	0.150	(-0.000, 0.301)
EGPD	0.172	(0.013, 0.332)
EGPD2	0.090	(0.053, 0.120)
LGP	0.132	(0.108, 0.156)
FMM	0.196	(0.160, 0.302)
PRMP	0.154	(0.105, 0.200)

Tail index. The tail-index parameters for each method are included in Table 5.1. The index for EGPD2 (Naveau's data-retaining method) is the smallest at 0.090, and FMM is the largest at 0.196. The GPD and EGPD intervals are wide, spanning from tail-index values near 0 to near 0.3. The LGP method admits narrowest interval estimates of the three methods presented in this dissertation, followed by PRMP and then FMM. FMM puts its mass over the heaviest tail indices, not even overlapping 95% intervals with either EGPD2 or LGP at all.

Return levels. Figure 5.2 shows the estimated 1-, 10-, 100-, and 200-year return levels (quantile levels p =0.99726, 0.99973, 0.99997, and 0.99999 respectively), along with 95% prediction intervals, where available. Prediction estimates and intervals are

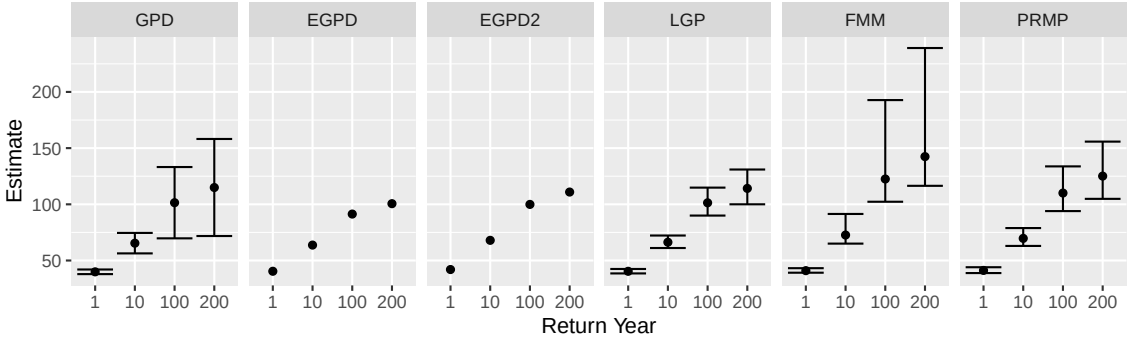


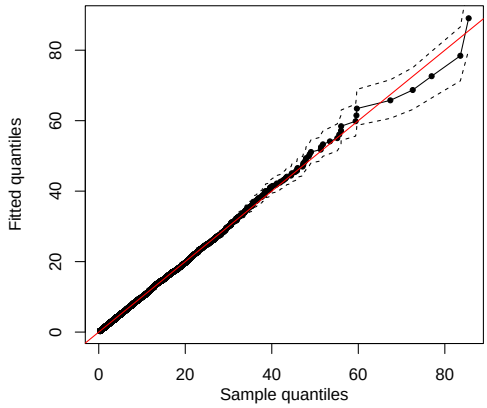
FIGURE 5.2: Application: return levels under all methods

similar at the 1-year return level across all methods, and point estimates of GPD, EGPD2, and LGP are all very similar to each other across all return-levels. The PRMP and FMM point estimates are slightly higher than the other methods for the longer return times, reflecting their heavier tail-index estimates. The LGP and PRMP have narrower prediction intervals than the GPD for the 100- and 200-year returns, both of which require extrapolation beyond the 48 years of observed data. At those same return years, the FMM intervals are wider than all other methods, reflecting the skewed heavy tail-index estimates seen previously. Without knowing the precise location or region that data were collected, we state these estimates without making effort to compare them to actual daily rainfall in the years subsequent to collection; however, we do make brief comparison to the in-sample data.

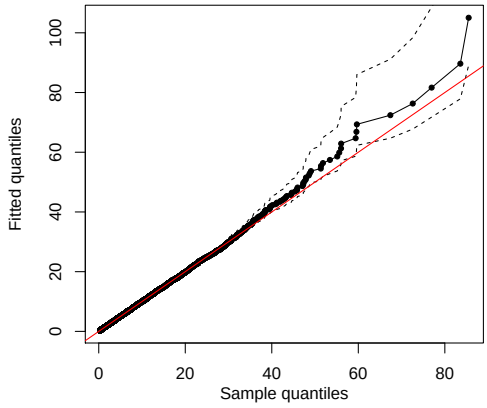
5.4 Model Fit

The empirical quantiles for the sampled non-dry days are compared to their predicted quantiles in Figure 5.3 for each of the three data-retaining models. For the lower quantiles, each model follows the 45-degree line closely indicating good fit through the bulk of the distribution. The LGP predictions maintain their fidelity to the data in the tail as well, while the FMM and PRMP seem to be slightly overestimating the upper quantiles; however, the overestimates are not egregious, as a 45-degree line is still for the most part contained within the 95% prediction intervals. We conclude that each of the three models does a reasonable job of fitting the in-sample data and accounting for the uncertainty in its estimates.

(a) **LGP model quantile-quantile plot**



(b) **FMM quantile-quantile plot**



(c) **PRMP model quantile-quantile plot**

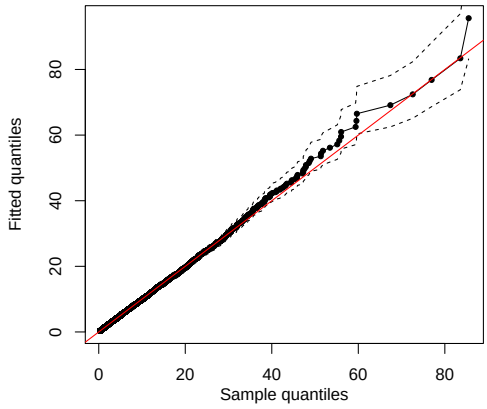


FIGURE 5.3: Application quantile-quantile plots for the three data-retaining models with sampled quantiles plotted against estimated quantiles and their associated 95% prediction intervals.

6

Conclusions

A main question of this work has been whether there is enough information in the likelihood, as specified under the heavy-tailed transformation framework described in Chapter 1, to estimate the entire sampling density without the tail getting overwhelmed by the bulk. The results of the previous chapters have shown that the information in the likelihood is not strong enough by itself for this estimation. However, when coupled with a carefully specified prior, models employing the framework have the potential to lower RMSE and improve estimation of predicted tail quantiles compared to threshold maximum likelihood methods, which allow the tail to "speak" by separating it from the bulk.

This chapter provides a comparison of the three models developed in this dissertation: the Logistic Gaussian Process (LGP) model of Chapter 2, the Dirichlet Process Mixture (DPM) model of Chapter 3, and the Predictive Recursion Marginal Posterior (PRMP) model of Chapter 4. After a high-level review of the similarities and differences between the models and their computation, we compare the results of their simulations. The conclusion highlights model features which are needed for successful implementation of the transformation framework in heavy-tailed univariate

density estimation and talks about future directions that this work could go.

6.1 Model Similarities and Differences

This section covers the high-level similarities and differences between the LGP, DPM, and PRMP models. All three models proceed under a transformation framework for univariate density estimation in the presence of heavy tails. Namely, a parametric family of distributions $\{g_\theta : \theta \in \Theta\}$ and a nonparametric family of distributions $\mathcal{H} := \{h(\cdot)\}$, constrained such that h are density functions on $[0, 1]$ with $||\log h||_\infty << \infty$ are brought together through $\mathcal{F} := \{f(\cdot) = g_\theta(\cdot)h(G_\theta(\cdot))\}$ for $\theta \in \Theta$ and $h \in \mathcal{H}$. Though the three differ in their handling of h, all three model data $Y_1, Y_2, ... \overset{iid}{\sim} f, f \in \mathcal{F}$ using a Bayesian approach. That is, priors are set on θ and h, which then get estimated simultaneously.

The density g_θ, corresponding to the CDF transformation G_θ, must match the support of the data and admit the possibility of heavy tails. The focus of this dissertation has been on data over the positive half-real line. Both the LGP and DPM models use the location-zero GPD($\sigma,\ \xi = 1/\nu$) scaled by the 0.9-level quantile of GPD($1,\ \xi$) for g_θ. The PRMP model uses GPD($\sigma,\ \xi$) without the quantile scaling. The DPM and PRMP use the same priors as each other but different priors from LGP, placing higher probability on slightly larger ν values than LGP and replacing the improper LGP prior on σ with a proper half-Cauchy prior. Details about these priors are available in Section 2.1.1 and Section 3.2.1.

For modeling of the nonparametric density, LGP assigns h a logistic Gaussian process prior over $[0, 1]$. The prior centers the LGP around a uniform density. Local adaptivity is achieved by estimating λ, the GP bandwidth parameter. The DPM and PRMP model h with a mixture density using a truncated-normal kernel over $[0, 1]$. Priors are selected for the truncated normal parameters so that the mixture density

prior mean is close to a uniform distribution. Local adaptivity is encouraged by placing prior probability over small truncated normal bandwidths, β. Both formulations guarantee that h is strictly positive and finite across the entire domain. Details about priors for these models are available in Section 2.1.2 for the LGP, Section 3.2.2 for the DPM, and Section 4.2.2 for the PRMP.

6.2 Computational Differences

Both the LGP and DPM models (also the finite mixture model, FMM, substituted in simulation) use MCMC methods to sample from the posterior distribution. The LGP approximates its infinite curves with a low-rank predictive processes over a discretized grid and uses an adaptive multivariate sampler to learn the posterior covariance between parameters, slowly adapting over many iterations to the learned information. The DPMM/FMM posterior curves are of themselves infinite; however, they too rely on approximations through truncation of infinite mixture components to a finite number of clusters. The PRMP employs a predictive recursion approximation, allowing it to avoid computationally-intensive MCMC sampling methods; however, it still requires numerical integration over the two parameters from the truncated normal mixing distribution.

Multimodality of the posteriors is seen across all three models; however, the reason for multimodality differs between models and is tied to the form and parameters of the nonparametric model. Multimodality of the LGP model is induced when insufficient knots are used in the low-rank predictive process. With too few knots, θ adjusts until the knots of h can settle near points of low h variability and can sandwich regions of high h variability. If multiple such sets of θ exist, multimodality is seen in the posterior, and moving between modes using the adaptive MCMC sampler is nearly impossible because it requires major adjustments to the mean and covariance of the Metropolis-Hastings proposal MVN distribution. The problem can

be eliminated, however, by increasing the number of knots used in the low-rank predictive process until multiple chains agree to one posterior mode. Usually ten to twenty knots is sufficient, but more may be necessary. For instance, some multimodality still exists across spliced gamma-GPD replicates of the LGP, indicating that 21 knots may yet be insufficient to capture the needed drop in h near the right boundary for datasets in this scenario.

The DPMM and FMM samplers do not have this same form of multimodality, as the modes in their nonparametric densities can easily shift by changing the kernel centers μ. Instead their multimodality (or sometimes just long tail regions) is caused by the mixture model's preference for sparsity and its adeptness at moving the mass of $h(\tau)$ from one end of τ to the other. This shaping induces multimodality or long-tails into the posterior because large σ tends to pair with h having mass pushed up against its zero boundary, while (relatively) small σ tends to pair with h having probability pushed near 1 and away from 0. Exploration of the PRMP shows that the mixture-model long-tail spread and multimodality can be overcome by increasing the prior precision parameter for the Dirichlet process, giving just enough weight to the prior to overcome the mixture's natural tendency towards sparsity and to encourage a flattening of the nonparametric density curve. The LGP approach does not require this additional injection of prior information since its natural tendency is toward non-sparsity and uniform nonparametric h shapes.

For LGP, the quantile-scaling of the base density improves convergence of the MCMC sampler, but the mechanism for this improvement remains an outstanding question. It could be that the quantile scaling works in harmony with the LGP to reduce the curvature of the posterior and condense the posterior mass into a single region, easily located and sampled by LGP's adaptive sampler. Or it could be that the quantile scaling induces *more* curvature into the posterior, similar to the curvature seen in the mixture models, making the posterior more difficult to traverse

by the adaptive sampler and ensuring that it only samples a small region of the posterior. This would be a problem that needs to be remedied. The same quantile scaling that (at its surface) appears to aid LGP may have a deleterious effect on the DPMM/FMM, enabling and exaggerating the natural tendency of the mixture model formulation towards splitting between one end of σ or the other.

6.3 Simulation Comparisons

In previous chapters, simulation studies were used to compare each proposed model to two maximum likelihood methods (GPD and EGPD) under various thresholds. Those simulations show the potential for the transformation framework to reduce RMSE in prediction of tail quantiles beyond what is possible using thresholding, though with better effect in some scenarios than in others.

In this section, we compare the results of the simulations to each other, seeking additional insights. The LGP and PRMP were each run on the full suite of six simulation scenarios (see a full description in Section 2.3.1 of Chapter 2). The FMM used a reduced set of scenarios, excluding the half-t-normal and spliced gamma-GPD scenarios; hence, tables and figures throughout this section do not have FMM results for these two cases.

6.3.1 Marginal Posteriors

Plots of the θ-marginal posteriors from three example datasets for each scenario are included in Figure 6.2. The posteriors from each of the three models are overlaid for comparison. Concentric contours, sometimes called isobands, delimit areas enclosing deciles of posterior probability. The LGP and FMM scale estimates are back transformed using $\sigma = \tilde{\sigma}/Q_{0.9}(\tilde{\nu})$, where $\tilde{\sigma}$ are sampled posterior estimates of scale obtained under the quantile-scaled base g_θ and $Q_{0.9}(\tilde{\nu})$ are the estimated quantile scalars. This back transformation makes the LGP and FMM marginals comparable

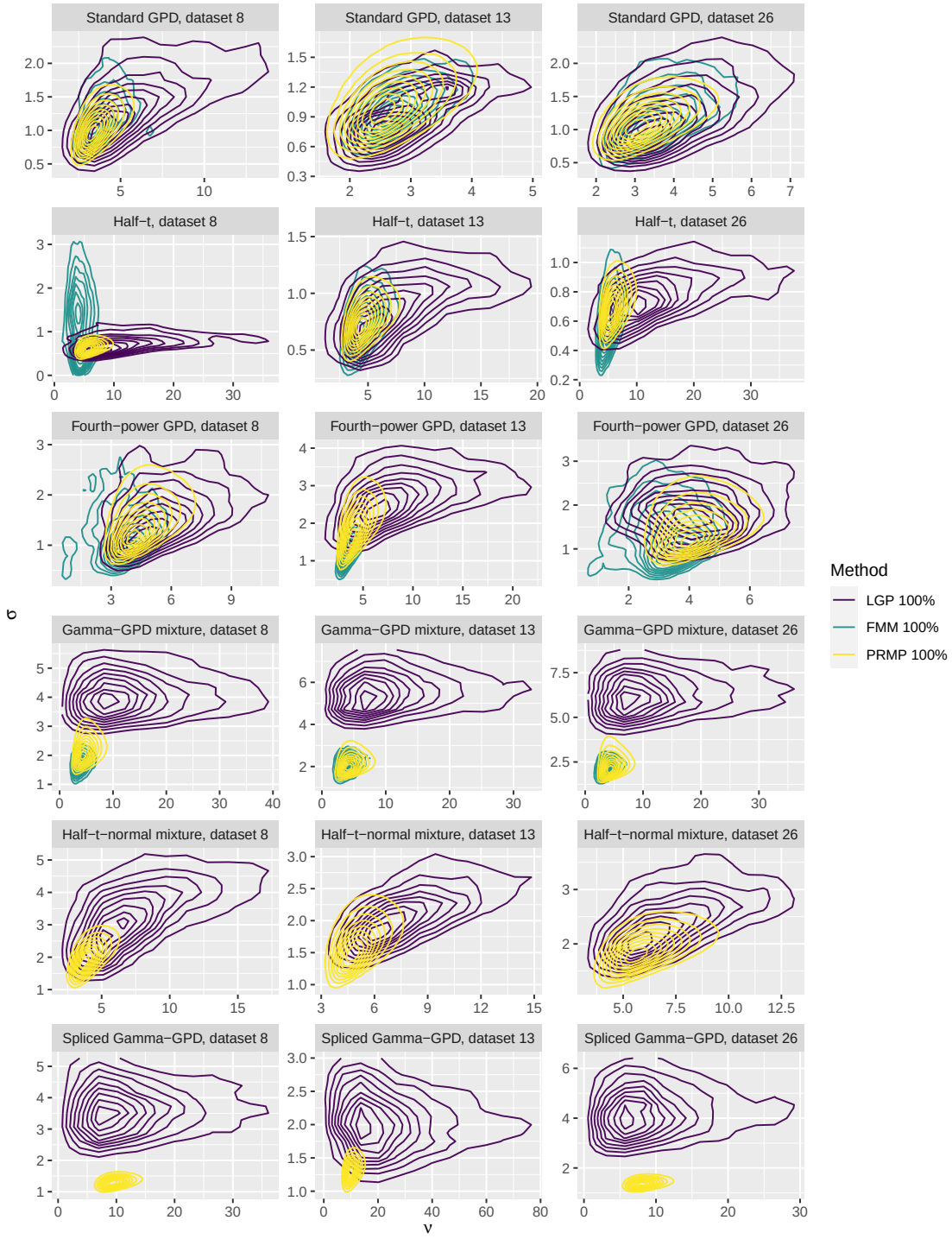


FIGURE 6.1: Posterior marginal distributions over θ for three datasets from each scenario, comparing the models run for that scenario. Concentric contour lines (isobands) enclose deciles of posterior probability.

Table 6.1: Estimated tail indices for each of the three models across scenarios along with bias, RMSE, interval width, estimated proportion of intervals containing the truth (Cov), and count of included or converged datasets. LGP and PRMP were run on 100 datasets, whereas FMM was run on only 40 datasets.

Scenario & Model	Estimate	Bias	RMSE	Width	Cov	Count
Standard GPD						
LGP 100%	0.21	-0.04	0.07	0.30	0.99	97
FMM 100%	0.28	0.03	0.05	0.28	1.00	38
PRMP 100%	0.29	0.04	0.06	0.26	0.97	100
Half-t						
LGP 100%	0.12	-0.13	0.14	0.21	0.42	92
FMM 100%	0.29	0.04	0.16	0.63	1.00	24
PRMP 100%	0.19	-0.06	0.08	0.18	0.91	100
Fourth-power GPD						
LGP 100%	0.20	-0.05	0.08	0.26	0.93	89
FMM 100%	0.29	0.04	0.05	0.52	0.97	30
PRMP 100%	0.22	-0.03	0.04	0.21	1.00	100
Gamma-GPD mixture						
LGP 100%	0.10	-0.15	0.15	0.28	0.87	82
FMM 100%	0.27	0.02	0.03	0.40	1.00	35
PRMP 100%	0.21	-0.04	0.05	0.25	1.00	100
Half-t-normal mixture						
LGP 100%	0.14	-0.11	0.11	0.18	0.35	99
PRMP 100%	0.23	-0.02	0.04	0.19	0.98	100
Spliced Gamma-GPD						
LGP 100%	0.10	-0.15	0.15	0.30	0.76	88
PRMP 100%	0.10	-0.15	0.15	0.08	0.00	100

to the PRMP model, which did not employ quantile scaling in its transformation distribution.

First observe that PRMP follows closely the posterior patterns of FMM, except in cases where the FMM sampler catered to lower-bulk data features (i.e. mostly occurring with the half-t scenario). This concordance was anticipated, as the PRMP and FMM are both approximations to the DPMM.

For all but the spliced gamma-GPD scenario, LGP estimates of ν tend to center at larger values than their FMM or PRMP counterparts. This can also be seen in the tail-index summary of Table 6.1, where LGP underestimates the tail, leading

to more tail-index bias and RMSE than the other two models. The LGP also has greater spread in its ν estimates; however, when inverted to estimate tail index, ξ, these large-ν values do not translate to much wider intervals, as the tail index is bounded below by 0. In contrast, the FMM, which often has small-ν-large-σ tails, has heavier estimated tail indices on average and wider intervals than either LGP or PRMP. Its bias is still reasonable though, at most overestimating the true tail index by 0.04 across scenarios. The PRMP has slightly lighter tail indices than its FMM relative, resulting in the lowest bias among all models. Generally, its tail-index intervals are also narrowest among the three models, while it also retains high interval coverage. The one exception to this is the spliced gamma-GPD scenario where the PRMP completely misses the tail index, underestimating ξ by 0.15 and having 0% tail interval coverage.

Again visiting Figure 6.2, the standard GPD posteriors have a great deal of overlap in their marginal posterior isobands across all three models and in both ν and σ. The overlap in the σ margin is also great in the half-t, fourth-power GPD, and half-t-normal mixture scenarios, though departures can and do exist on a per-data-set basis. The models have notable separation across σ in the gamma-GPD mixture and spliced gamma-GPD scenarios, with the LGP generally preferring larger σ than either FMM or PRMP. In both of these scenarios, small-σ corresponds to true nonparametric densities that require precipitous drops (large-in-magnitude negative slopes) immediately before or at the right boundary. Large-σ, on the other hand, pushes the true nonparametric density mass away from the right boundary, thereby requiring smaller negative slopes of h near 1.

Figure 6.2, which displays nonparametric density estimates for each dataset alongside model averages, highlights the large-σ and small-σ differences for the gamma-GPD and spliced gamma-GPD scenarios. (Note that the values of the nonparametric densities of FMM and LGP are not directly comparable to PRMP because their g_θ

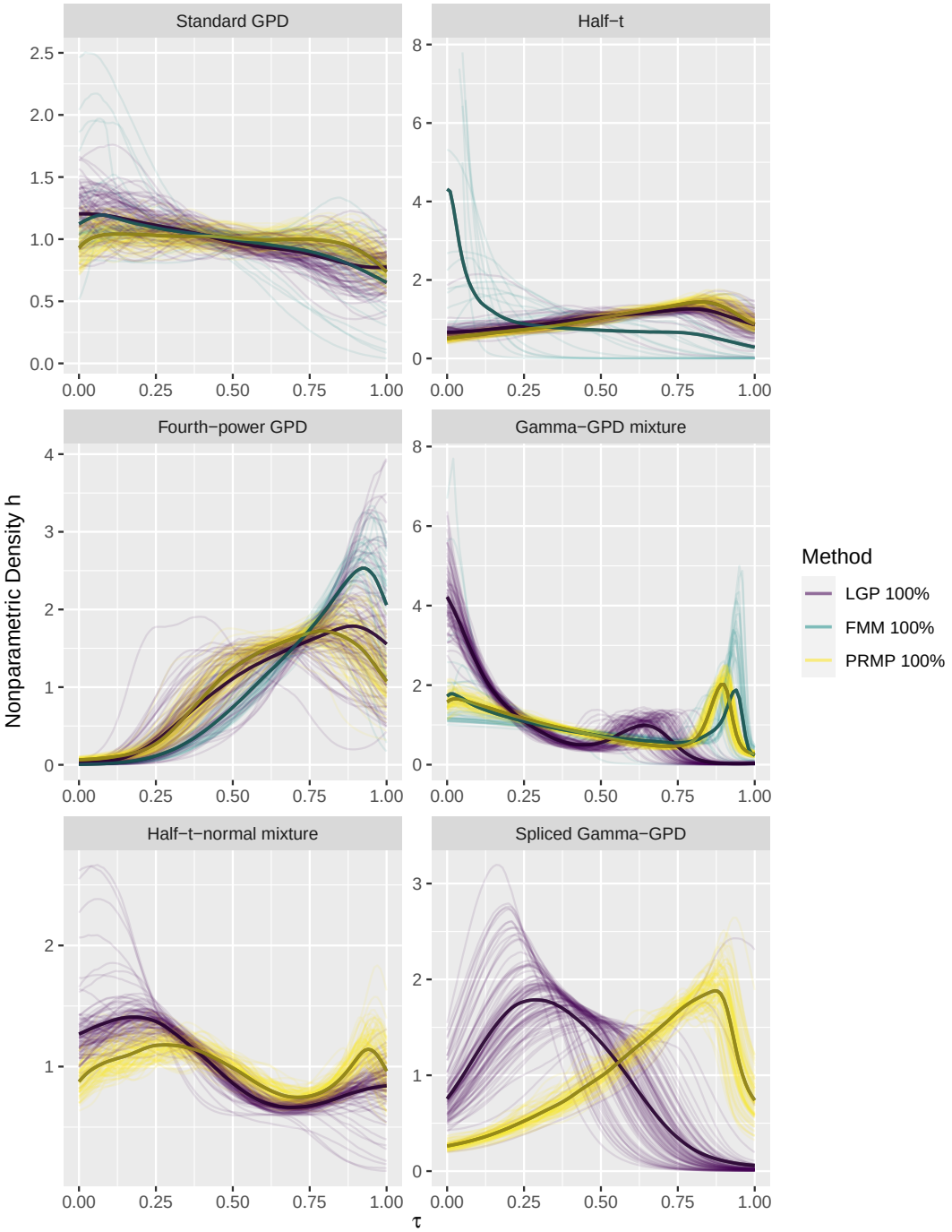


FIGURE 6.2: Estimated nonparametric densities for each dataset and (in darker hue) averaged across datasets within a model. LGP and FMM are directly comparable, having used the same transformation density. PRMP should only be compared in form, having used an alternative transformation density.

was quantile-scaled and the PRMP was not; however, the form can generally be compared.) The LGP estimates push mass away from $\tau = 1$, seeing their modes in values $\tau < 0.9$ and values of $h(\tau)$ near zero for $\tau > 0.9$. The PRMP and FMM, with their small-σ estimates, have local nonparametric density modes near $\tau = 0.9$. Thereafter, they try to capture the true nonparametric density's precipitous drop with varying degrees of success, as will become apparent in the subsequent section where quantile bias and RMSE are explored. There tends to be more variability across dataset replicates in the LGP nonparametric density estimates of these two scenarios than in either the FMM or PRMP. This corresponds to more varied shapes in the θ marginals across datasets in the LGP than in FMM and PRMP (see per-dataset 80% isobands for each model in Appendix C.) The FMM nonparametric density estimates in the half-t scenario also have a lot of variability, reflective of the spread in estimating σ, so much so that the mean does not follow the form of any of its constituent datasets.

The nonparametric density estimates in the remaining scenarios are similar in form within a model and across models, mirroring the similarity of θ estimates. The PRMP and FMM densities in the standard GPD scenario both tend to bow downwards at the boundaries. The LGP shrinks toward its mean, $h(\tau) = 1$, at the boundaries where data are scarce, a behavior inherent to the LGP. Estimates of the fourth-power GPD are similar across models, though the FMM does extend higher at the right boundary than the others, more closely following the true, generating Beta$(4, 1)$ nonparametric density than the other two. The LGP estimates across replicates have quite a bit of variability in this scenario, some favoring an interior mode and others placing their mode at $\tau = 1$. The LGP estimates of the half-t-normal scenario also have more variability than the PRMP estimates, though the form of both models is similar. For that scenario, PRMP again exhibits downward cupping at the right-hand boundary.

6.3.2 Quantile Estimation

Bulk bias. Figure 6.3 displays the bias relative to the true quantile, averaged across datasets, in the bulk of the distribution. The relative bias is similar in magnitude among all three models but the LGP (purple) has slightly lower bias across levels p and across scenarios. As noted in Chapter 3, the FMM for the half-t scenario is sensitive to lower-tail data differences, fitting the extreme lower tail with a great degree of accuracy. In the fourth-power GPD bulk, all three models do their worst job of estimating the quantile at the lower boundary, though the bias for PRMP extends into larger quantile levels p than for either FMM or LGP. In the gamma-GPD mixture, all three models have a peak in bias near $p = 0.73$, the locale where true density h starts to increase to accommodate gamma mixture-component observations. This bias indicates that all three models have difficulty making a sufficiently sharp transition in the nonparametric density. The LGP, which favored larger σ, flattening the peak needed for h, has the lowest bias among the three models through this region. The bulk quantile bias considered on a per-dataset basis is similar in magnitude across datasets and models (see Figure C.4 in Appendix C).

Tail bias. The relative quantile bias in the tails of estimated sampling distribution f (Figure 6.4) are displayed per dataset as well as averaged across datasets. Generally speaking, LGP has the smallest tail quantile bias. It maintains low levels of bias in the lower tails, where data should still be plentiful, unlike both FMM and PRMP which exhibit bias in these lower tails. Within a scenario, the models exhibit similar amounts of intra-dataset variability. The one exception is LGP estimates of the gamma-GPD mixture scenario, which show more variability among estimated datasets than the other two models do with their replicates.

The relative quantile bias reflects the over- and under-estimation trends of the tail-index parameter: LGP tends to underestimate the tail heaviness, leading to

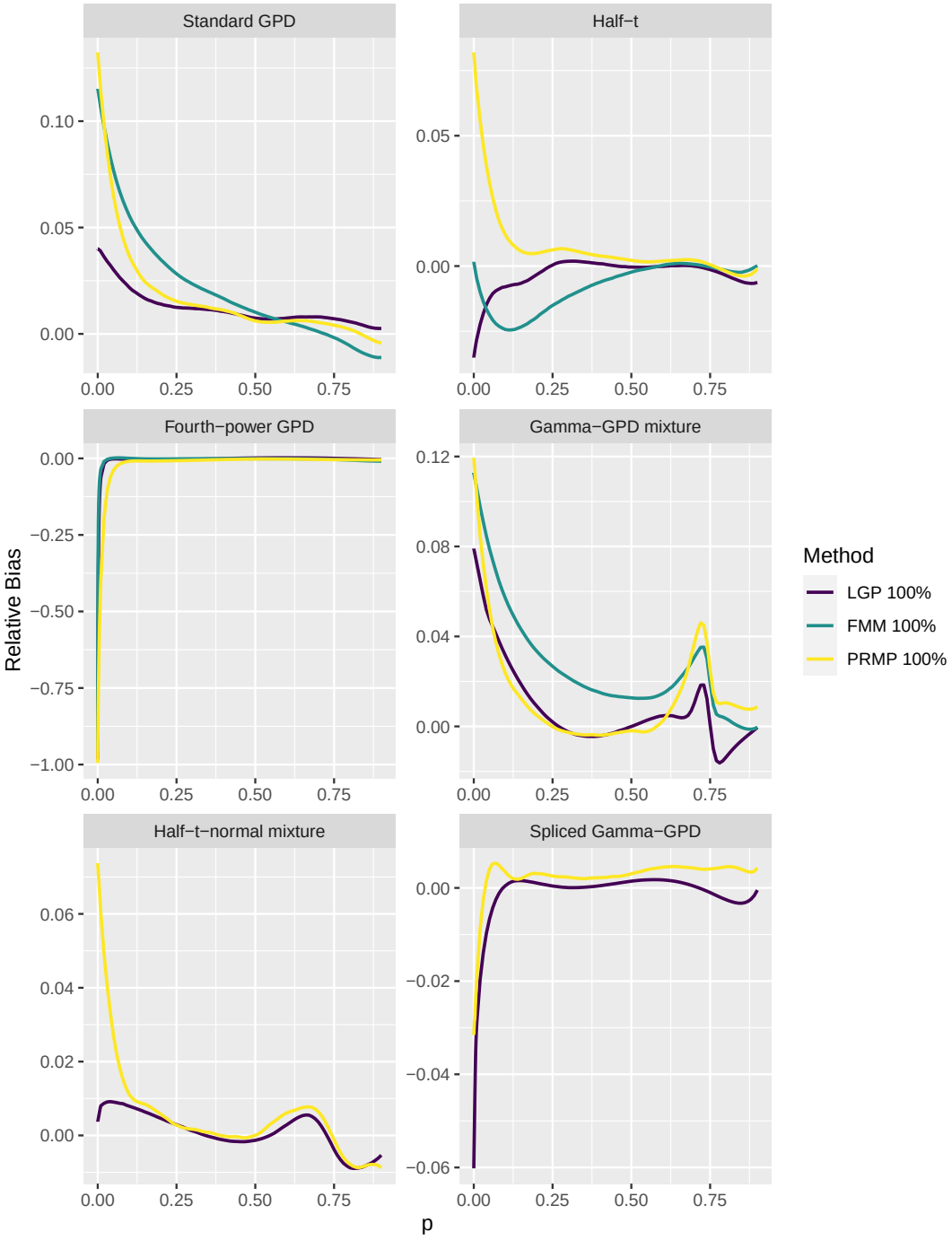


FIGURE 6.3: Estimated relative quantile bias (bias / true quantile value) for the bulk portion of the sampling density ($p \leq 0.9$).

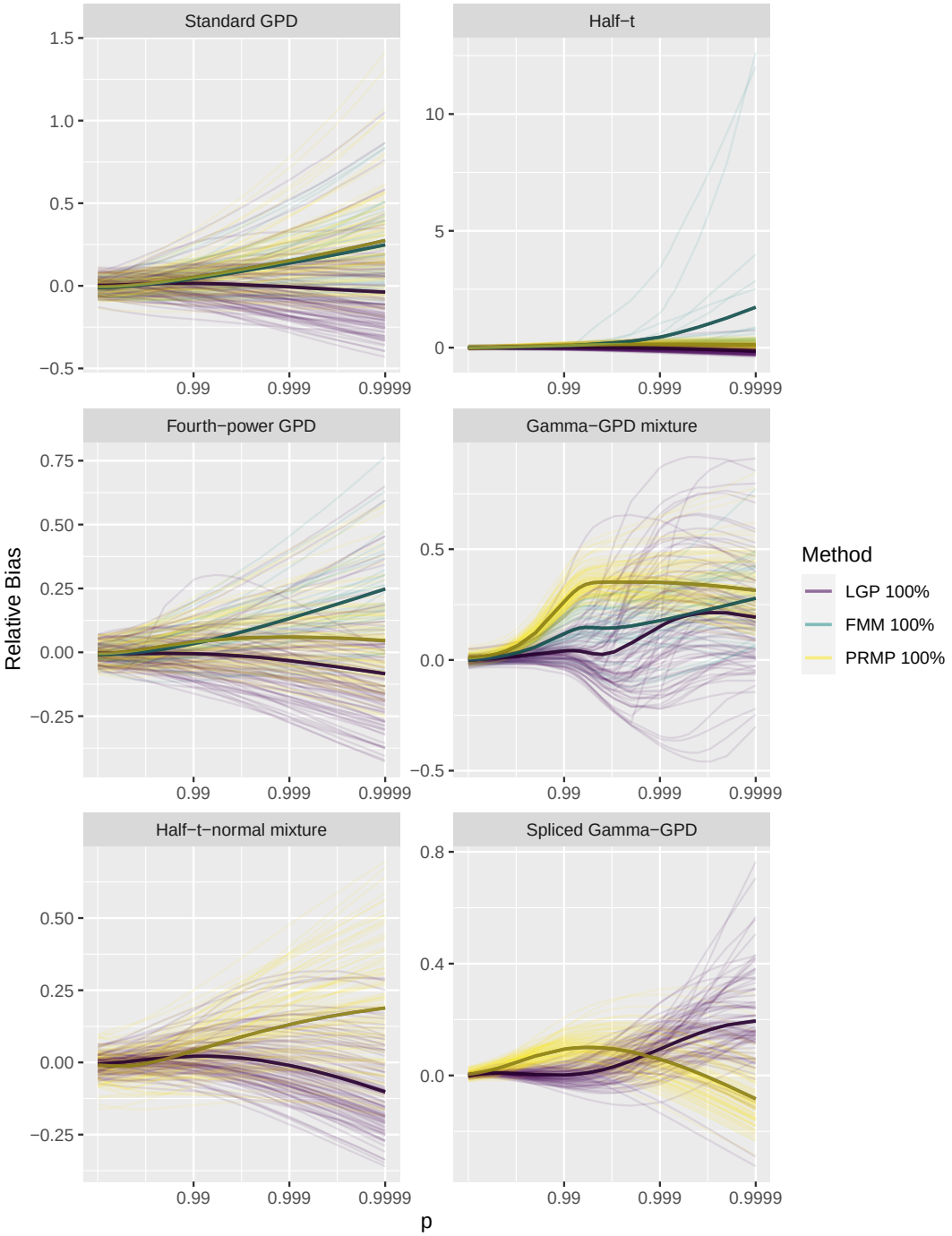


FIGURE 6.4: Estimated relative quantile bias (bias / true quantile value) for the tail portion of the sampling density ($p \geq 0.9$). Lighter thin lines represent estimates from a single dataset and darker broad lines represent estimates averaged across datasets within a model.

underestimates of the quantiles of f; and FMM and PRMP tend to overestimate the tail index, leading to overestimates of the quantiles. But this does not tell the full story, since the LGP ends up overestimating the extrapolated tails of both the gamma-GPD mixture and the spliced gamma-GPD. For both of these scenarios, the underestimation of the tail index is still apparent in the downward curvature of the mean bias lines at extreme $p = 0.9999$ levels, but the nonparametric density must also play a part to send quantiles into over-estimation territory. These two scenarios both exhibit LGP estimates of $h(1)$ near zero, which may provide a hint. By spreading the tail of the nonparametric density over larger τ-regions of h and more fixed-knot LGP estimation locations, the nonparametric density, aided by transformation G_θ, can induce much more curvature in the extrapolated tails of f than it does otherwise in scenarios where the upper tails of h are less spread out and therefore exhibit more local linearity.

Bulk RMSE. Figure 6.5 shows the bulk RMSE for the three models. RMSE is similar in magnitude across quantile levels $p < 0.9$ with perhaps some advantage going to the LGP and FMM models over the PRMP. The LGP-to-PRMP comparison is more stark in the half-t-normal and spliced gamma-GPD scenarios, where RMSE is clearly higher for the PRMP across all bulk quantile levels.

Tail RMSE. The tail RMSE does not clearly delineate which model is best (see Figure 6.6). LGP has lowest RMSE for the standard GPD, for the half-t-normal, and for most of the half-t. PRMP has lowest RMSE in the fourth-power GPD scenario. The gamma-GPD mixture and spliced gamma-GPD plots show that LGP has the lowest RMSE in the lower tails but higher RMSE in extrapolation than either the FMM (mixture scenario) or PRMP (spliced scenario). This may mean that the LGP strategy of sending nonparametric h to near-zero values at $\tau = 1$ does not aid but hinder quantile estimation after all.

Quantile interval coverage. Finally, Figure 6.7 displays the quantile interval cov-

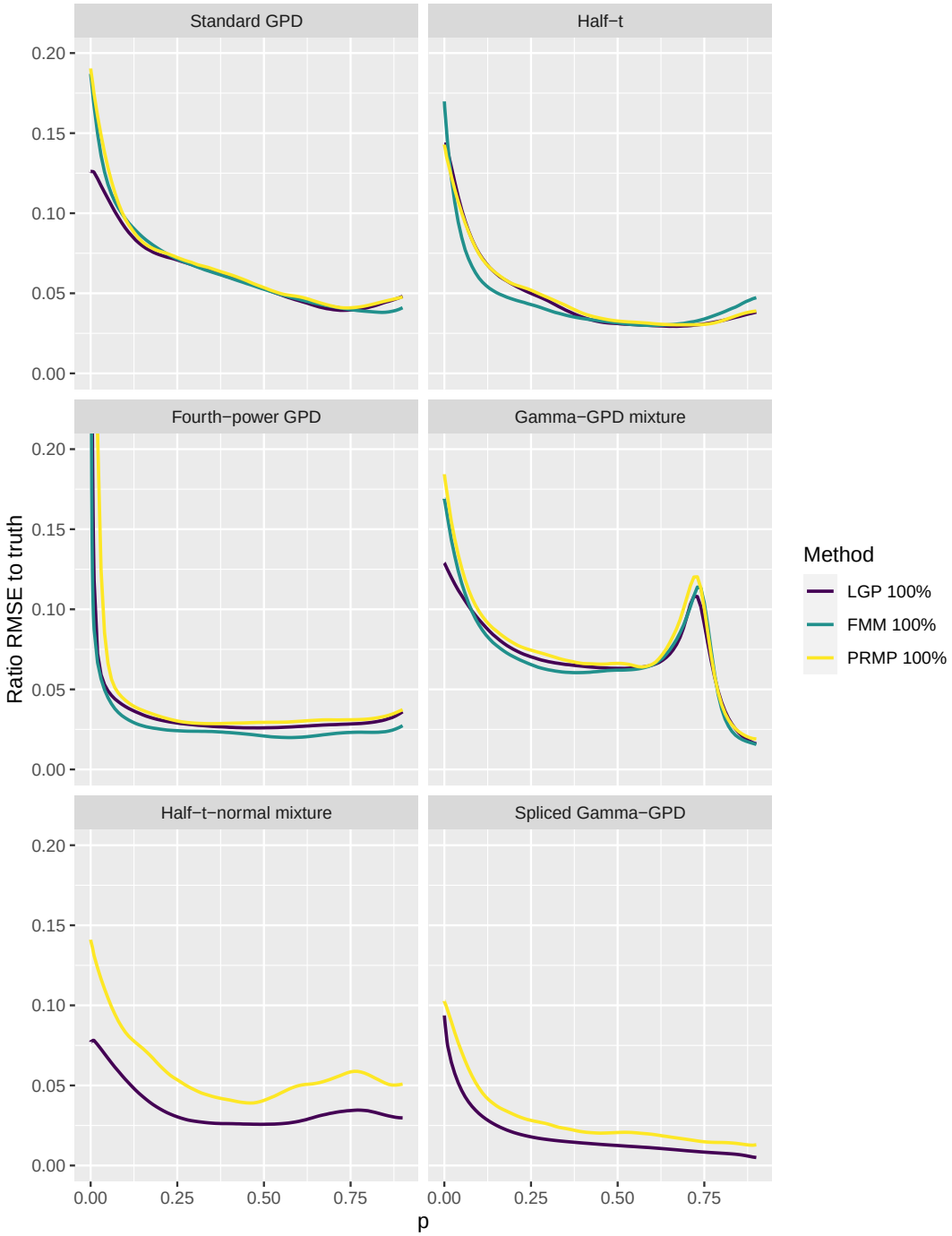


FIGURE 6.5: Estimated relative quantile RMSE (RMSE / truth) for the bulk portion of the sampling density ($p \leq 0.9$). Lighter thin lines represent estimates from a single dataset and darker broad lines represent estimates averaged across datasets within a model.

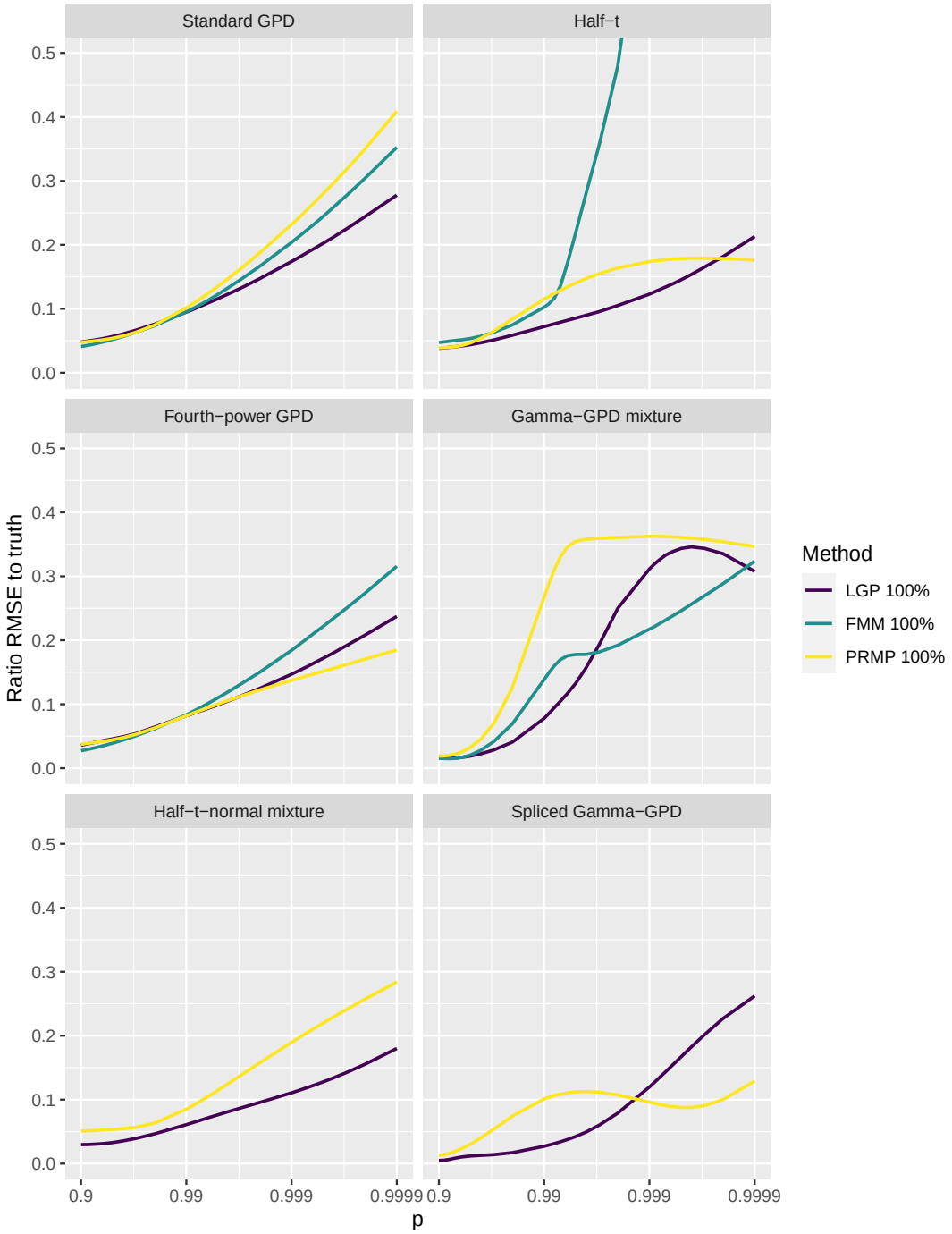


FIGURE 6.6: Estimated relative quantile RMSE (RMSE / truth) for the tail portion of the sampling density ($p \geq 0.9$).

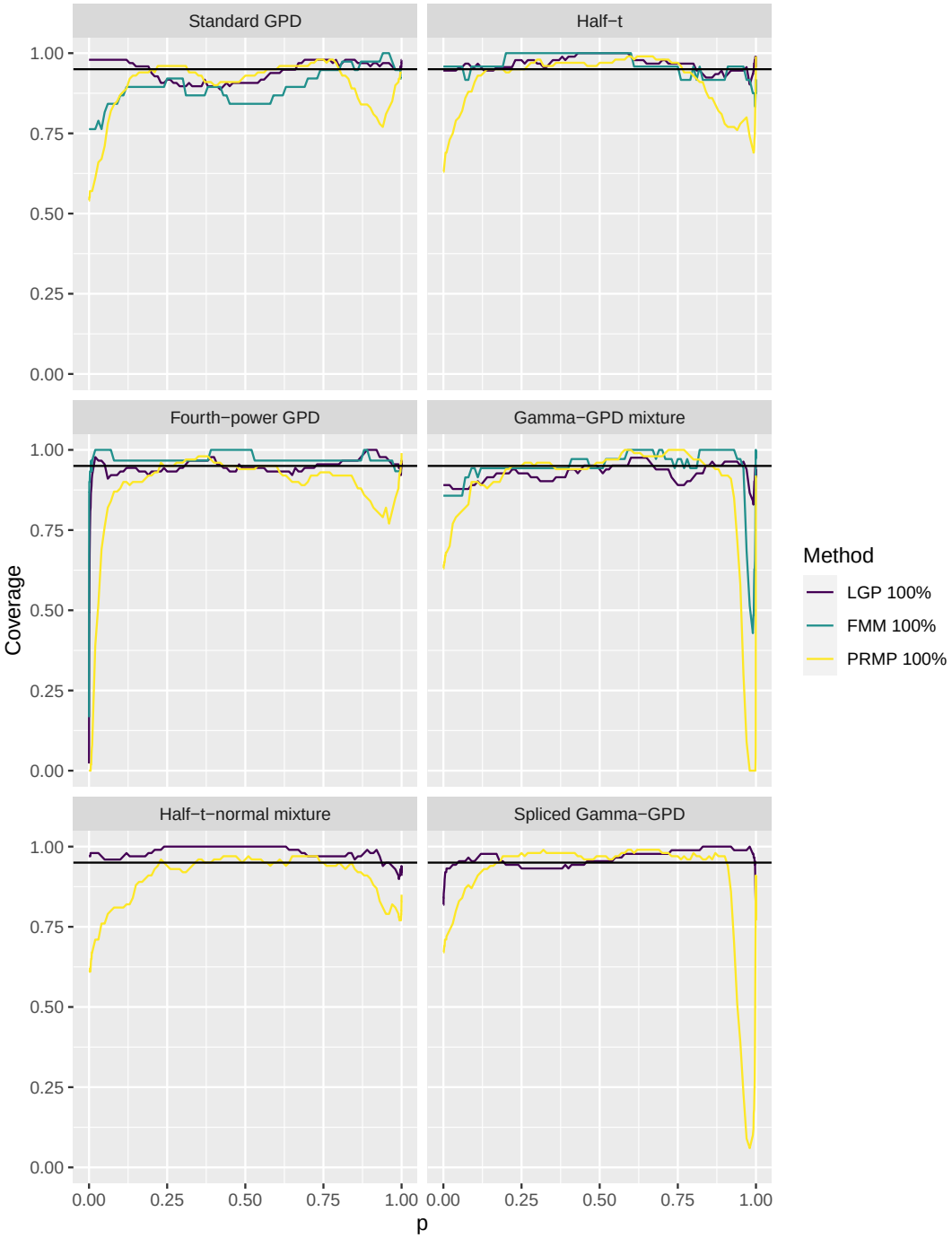


FIGURE 6.7: Estimated coverage of quantile intervals for the full sampling density.

erage for all p, for each model, and across all scenarios. The LGP keeps consistently high interval coverage across all quantile levels in both the bulk and the tail, only experiencing a small drop in upper-tail coverage in the hard-to-estimate gamma-GPD mixture and spliced gamma-GPD. The drop in coverage for PRMP at either tail is due to the bias of the PR estimates and is discussed in Chapter 4.

6.4 Conclusions and Future Work

Throughout this dissertation and across approaches, several patterns have repeated themselves. Small changes in transformation parameters, θ—primarily in the scale parameter, σ, but also to a lesser part the shape parameter, ξ—can vastly change the shape needed for the nonparametric distribution, h. Conversely, the way in which a model captures the nonparametric density has a pronounced effect on how it captures its transformation parameters. This interplay between the estimation of h and θ depends upon the form of the model, and elucidating that interplay can be both subtle and challenging. While model form affects the way that estimation plays out, commonalities exist across the approaches, pointing to features a model needs in order to successfully implement the transformation framework for heavy-tailed density estimation. This section briefly reviews the strengths and weaknesses of each model, summarizes the key take-aways and desirable features, and discusses some areas of future work for semi-parametric heavy-tailed density estimation under the transformation framework approach.

6.4.1 Model Strengths and Weaknesses

Of the approaches considered in this dissertation, LGP has many advantages over FMM or PRMP: it has lower bias through the bulk and lower tail for many simulation scenarios and exhibits consistently high quantile interval coverage. However, it has one drawback that leaves a question as to whether the model is functioning as

intended: LGP underestimates the tail-index parameter of the GPD transformation. This is not entirely surprising. For given σ, an underestimate of ξ results in less change to the form of the needed nonparametric density than does an overestimate, so an adaptive sampler learning the covariance directly on the nonparametric density might have an easier time settling into shape parameters that are too light. Another cause of the underestimation of ξ may be the exponentiation of the log-scale nonparametric density estimates, which allows $\hat{h}$ to scale up more quickly near a boundary than it allows $\hat{h}$ to scale down quickly near a boundary.

Despite this shape-parameter-estimation bias, the LGP model does an admirable job in tail estimation—often producing tail quantile estimates with lower RMSE than threshold methods—using excellent estimation of its nonparametric density. Our only indication that LGP's approach to estimating large h tail drop-offs—namely adjusting σ until the nonparametric density is spread out enough to capture the tail with a shape LGP is adept at capturing—is perhaps not optimal for estimating heavy tails is that our other models, which capture ξ better, exhibit lower RMSE in extrapolation for these two scenarios. Understanding the mechanism for why quantile scaling seemed to improve the convergence of the sampler *may* help uncover the source of this tail-index bias or at least eliminate MCMC sampling bias as a probable cause.

The DPMM and FMM models have MCMC sampling difficulty for some datasets and scenarios, but the good tail-index estimation and the reduction in tail quantile RMSE compared to the threshold methods make us confident that this model is worth pursuing and refining. We anticipate that strengthening the prior for the precision parameter will encourage the DPMM to adapt flatter nonparametric density shapes and keep the nonparametric density from catering to the bulk at the expense of its tail estimation. These may also aid in condensing the posterior distribution, reducing the banana-shaped θ curvature and making it easier to sample.

The PRMP does the best job of estimating the tail index or shape parameter among the models considered despite bias of quantile predictions at either tail. This tail-index estimation is itself is a success, as it helps to validate the semi-parametric density framework approach in finite samples. When the posterior is integrated appropriately over the nonparametric density space, information *is* teased out about tail heaviness. The PRMP also provides a quick and easy tool for looking at the marginal density and understanding how much weight needs to be given to a mixture density prior in order to balance the interplay between h and θ. An outstanding question is whether the PR nonparametric density boundary bias is due to the particular kernel chosen, is inherent to PR, or is somehow related to the interplay between h and θ.

While none of the three models has emerged as uniformly "best" across scenarios and quantile levels, *all* have shown promise as data-retaining estimators for heavy-tailed density estimation, reducing prediction error in quantile estimation.

6.4.2 Take-aways

From this dissertation, we have acquired knowledge about estimation under the semi-parametric Bayesian transformation framework for heavy-tailed density estimation.

First, there is not enough information in the likelihood alone to prevent the tail information from getting overwhelmed by the bulk. However, when the likelihood is harnessed with a carefully specified prior, there is potential to model both the bulk and the tail simultaneously, allowing both to speak, without an explicit separation of the two parts.

Second, the nonparametric prior needs to give strong weight to uniform-type shapes. If it does not, or if this prior is not appropriately weighted with respect to the sample size, the nonparametric density may end up serving the bulk distribution well at the expense of tail estimation. In practice, a nonparametric density estimate

that drops to near-zero values approaching but well before $\tau = 1$ may be an indication that the prior has not been given enough weight and that the bulk is being preferred over the tail. On the flip side, a nonparametric density that never approaches zero is not necessarily an indication that the model is doing the right thing, which leads to our next point.

Third, as the nonparametric prior is given more weight, the full-model estimates may become subject to the biases and the tendencies of the selected nonparametric model form. Bias in the nonparametric model may have an amplified effect when sent through the transformation framework. A seemingly small boundary bias will affect large swaths of the lower and upper quantile estimation.

Finally, the less curvature that is needed from the nonparametric density (i.e. the closer it can stay to a uniform shape), the easier it will be for a transformation model to capture the full form. This was seen in the uniform and half-t-normal scenarios as well as in the truncation exercise of Chapter 2. This is because less curvature in h is generally associated with a condensed θ posterior, making them easier to sample when using MCMC sampling approaches.

6.4.3 Future Directions

There are a variety of future directions that this work could take. With the transformation framework defined and some of its pitfalls and strengths brought to light, the area is rich for experimentation with alternative model forms as well as for modifying the approaches of this dissertation.

Nonparametric density modifications. One area begging immediate attention is the modification of LGP to increase knot placement near the nonparametric density boundaries, especially at the upper boundary. This should help ensure that models requiring drops at the boundary do not end up in bimodal situations induced by sparsity of knot placement at the boundary. This might also allow the model

to find and utilize a broader range of true-density shapes. Modifications allowing second-order nonstationarity in the LGP across the domain of the nonparametric density may also prove useful. Alternatively, to avoid the multimodality induced by knot sparsity altogether, we could try sampling the nonparametric density using the Gaussian process density sampler of Adams et al. (2009). This approach infers the unknown density of the data by framing the generative process (a CDF-transformed GP) in terms of a rejection sampler, bounded above by 1; augmenting the data with latent acceptance indicators at sampled locations; sampling posterior Bernoulli probabilities given the augmented data with its sampled locations as regressors; and backing out the GP through the probabilities. Because it is based on a rejection sampler, this may prove to be less efficient than the current LGP implementation.

The DPMM and FMM can be updated based on increased understanding, coming from the PRMP exploration, that stronger prior weight on the prior precision parameter can temper the mixture models' tendency towards sparsity. Additionally, a more comprehensive study of prior sensitivity could help calibrate the weight needed to give stability to the model. From a more practical standpoint, both the DPMM and FMM could benefit from being written in a faster programming language, e.g. DPM cluster allocation could use a call to C, as their current run times make them less user friendly.

Some alternative mixture model forms are already under construction or consideration. An alternative kernel may prove to be the solution to the boundary bias seen using the truncated normal distribution. Appendix B details a kernel that it is a mixture of one (or three) beta kernels and a uniform density, which keeps h always strictly greater than zero but may also allow more rapid drops in h near the right-hand boundary.

Transformation modifications. In each model of this dissertation, the primary parametric transformation considered is the CDF of a generalized Pareto distribu-

tion. One aspect of the GPD that was not fully appreciated when selecting it as the transformation distribution was how fully ξ is integrated with the shape of the bulk. It is easy to think of ξ as a parameter which dictates the shape of the tail, since the distribution often gets used to model tails. But as a transformation density, ξ operates in both the bulk and tail and is therefore influenced by the nonparametric density in both areas. This is not wholly undesirable, as we never really know where the tail "starts," and information about the power decay parameter might be retrievable from any part of the distribution given the right nonparametric density. For example, in the simulation scenarios considered, the shape parameter affects the full standard GPD; informs all parts of the fourth-power GPD, provided a nonparametric shape similar to the generating Beta$(4, 1)$ can be found; and affects the bulk shaping of the gamma-GPD mixture. This integration also allows ξ to get lost in or overwhelmed by the bulk information, as may be happening in the LGP model.

Another lesson learned about using a GPD as the transformation is that with density values $g_\theta(0)$ strictly greater than zero, the GPD proves difficult to scale sufficiently down to zero for sampling densities f that need it; the required nonparametric h has to be incredibly flexible to accommodate these forms. In those scenarios that did not need scaling down to zero, undesirable bias in the lower "tail" can still be found. Perhaps a more flexible transformation density, one capable of taking a variety of shapes, would keep the needed nonparametric form h closer to a uniform density more naturally, without the need to infuse the prior with additional information.

One alternative might be an extended generalized Pareto distribution (EGPD), i.e. one from Papastathopoulos and Tawn (2013) or Naveau et al. (2016). The EGPD formed by a Beta$(\kappa, 1)$ transformation of a GPD, common to both papers, results in a distribution with CDF $G_{\kappa,\sigma,\xi} = (1+1/\sigma(1-\xi x/\sigma)^{-1/\xi})^\kappa$. This density is zero at zero, i.e. $g_\theta(0) = 0$, has a non-boundary mode, and maintains right-hand tail-decay power

of $1/\xi$. The addition of a third parameter κ, which controls the left-hand power decay of the distribution, may be desirable from a bulk-tail parameter separation standpoint, but it also increases the dimension and complexity of any model beyond the two parameters of the standard GPD. Also, the constraint $g_\theta(0) = 0$ may not always be desirable in a candidate transformation density.

An alternative transformation distribution that has gained some traction in the kernel density estimation literature for heavy-tailed data is the modified Champernowne distribution, introduced by Buch-Larsen et al. (2005). The three-parameter distribution takes the form

$$G_{c,M,\nu}(x) = \frac{(x+c)^\nu - c^\nu}{(x+c)^\nu + (M+c)^\nu - 2c^\nu}. \tag{6.1}$$

This distribution has some appealing properties as a heavy-tailed transformation distribution. First, the tail converges to a Pareto tail with power-law decay ν. Second, in tandem with ν, the parameter c is capable of ensuring that $0 < g_\theta(0) < \infty$ and of inducing a mode at $x > 0$. Consequently, the Champernowne distribution can take a variety of shapes. Lastly, the parameter M can easily be estimated by the empirical median since $G_\theta^{-1}(0.5) = M$. Pre-estimating and fixing M while continuing to estimate c and ν, simultaneous to h, could take advantage of the flexibility of the modified Champernowne while keeping the number of transformation parameters at two.

Insofar as these transformation densities associate the shape of the bulk with one set of parameters and the tail power decay with a separate set of parameters, we believe this will aid in the semi-parametric transformation estimation.

Using a transformation density that exists over the full real line (e.g. a standard t-distribution or an asymmetric t-distribution) would also be interesting, though potentially complicated, especially if the sampling density is itself multimodal. De-

pending on the separation of the sampling density modes, multimodality could show up in the θ margins (PRMP might help diagnose this), and estimating h under changing transformation location parameter would not be inconsequential. Estimating these densities may require more than just strong priors on the nonparametric densities. They may require intervention in the transformation densities themselves, such as using a transformation which is itself a mixture or one which somehow enforces quantile matching, e.g. $H(0.5) = 0.5$.

Other directions. Perhaps a more fruitful and interesting approach than trying to make the transformation density multimodal would be to include additional variables as predictors in the model, for example by using LGP density regression (Tokdar et al., 2010). If additional variables can explain some of the underlying variability in the sampling density form, e.g. via linear prediction of the transformation location parameter, the scaling needed by the nonparametric density to explain the remaining variability may be brought closer to uniformity and be easier to estimate. Taking care of the bulk density in this way could, in turn, aid estimation of the tails.

Alternatively, instead of taking a conditional or regression approach to the heavy-tailed density estimation, two or more variables can be considered jointly. The LGP can be formulated in multiple dimensions (Tokdar, 2007) but its theory does not yet support the notion of multivariate tail dependence. A future direction of research could be to establish whether asymptotic tail dependence of some arbitrary family of multivariate transformations G_θ is maintained under a transformation framework of the multivariate LGP model.

A final question that could be of interest in future work is in regards to the tail index. Can posterior consistency of the tail-index estimate be proven for either the LGP model or the DPMM? We know by construction that ξ lies in the Kullback-Leibler support of our prior. If we could leverage an existing hypothesis test for the tail index parameter, showing that it can be formulated as a sequence of exponentially

continuous tests, then posterior consistency of the tail-index would follow directly from the Schwartz Theorem (Schwartz, 1965). Li et al. (2019) use such an approach when establishing necessary criteria for a mixture model to exhibit posterior tail-index consistency. Their work may be applied directly to determine if posterior consistency of the tail index applies to the DPMM proposed in this dissertation, and their work might also inform the steps needed to prove posterior consistency of the tail index for the LGP model.

This dissertation has laid out the framework necessary for heavy-tailed, univariate density estimation under a semi-parametric transformation framework. We have developed three models, complete with error uncertainty quantification, under the framework and evaluated the relative merits and weaknesses of each. Even so, we are just beginning to understand the subtleties of estimation under the framework, and the area is ripe for further development and future work.

Appendices

Appendix A

LGP Model Details and Approximations

This appendix provides details for the logistic Gaussian process model of Chapter 2, including the likelihood, priors, and posterior. It also includes the predictive process substitution and its subsequent finite approximation.

A.1 Likelihood

The likelihood follows the form $f(y|\theta, h) = g_\theta(y) \cdot h(G_\theta(y))$, and h takes the form $h(\tau) = \frac{e^{w(\tau)}}{\int_0^1 e^{w(t)}dt}$. Assuming independence of each y_i, the joint likelihood is

$$f(\boldsymbol{y}|\theta, w) = \prod_{i=1}^{n} \left[g_\theta(y_i) \cdot \frac{e^{w(G_\theta(y_i))}}{\int_0^1 e^{w(t)}dt} \right].$$

A.2 Prior

The prior is decomposed into two independent priors, one of which governs the nonparametric w and one of which governs the parametric θ: $\pi(w, \theta) = \pi(w) \cdot \pi(\theta)$. Further details of each prior follow.

The nonparametric prior $\pi(w)$ is specified via the hierarchical form

$$\begin{aligned} w|\kappa^2, \lambda &\sim GP(0, \kappa^2 c^{SE}(\cdot,\cdot|\lambda)) \\ \kappa^2 &\sim IG(a_\kappa, b_\kappa) \text{ with } a_\kappa = b_\kappa \\ \lambda &\sim Discrete(\alpha, \pi^*_\lambda), \text{ where } \alpha = (\lambda_1, ..., \lambda_G) \text{ and } \kappa^2 \perp \lambda, \end{aligned}$$

resulting in the marginal prior for w

$$\begin{aligned} \pi(w) &= \int_\lambda \int_{\kappa^2} \pi(w|\kappa^2, \lambda)\pi_\kappa(\kappa^2)\pi_\lambda(\lambda)d\kappa^2 d\lambda \\ &= \sum_{g=1}^{G} \pi(w|\lambda_g)\pi^*_\lambda(\lambda_g) \\ &= \sum_{g=1}^{G} \mathrm{TP}_{2a_\kappa}(w|0, c^{SE}(\cdot,\cdot|\lambda_g))\pi^*_\lambda(\lambda_g). \end{aligned}$$

This implies that w is distributed *a priori* as a π^*_λ-weighted mixture of t-processes, each of which has a unique covariance kernel depending on $\lambda_g \in \alpha$.

The parametric prior is defined for a location-zero generalized Pareto distribution $GPD(\sigma, 1/\nu)$ with scale σ and tail parameter $1/\nu$. The priors on θ are assumed to be independent: $\pi(\theta) = \pi(\nu, \sigma^2) = \pi(\nu)\pi(\sigma^2)$ with $\log((\nu - 0.5)/5.5) \sim \mathrm{Logis}(0, 1/2)$ and $\pi(2\log(\sigma^2)) \propto \frac{1}{2\log(\sigma^2)}$. Taken together we have

$$\pi(\nu, \sigma^2) \propto \frac{1}{2\log(\sigma^2)} \cdot \mathrm{Logis}\left(\log\left(\frac{\nu - 0.5}{5.5}\right)|0, 0.5\right). \tag{A.1}$$

A.3 Log Posterior

The log posterior is given as

$$\begin{aligned} \log(\pi(\nu, \sigma^2, w|Y)) \propto &\sum_{i=1}^{n} \log(g_\theta(y_i)) + \sum_{i=1}^{n} w(G_\theta(y_i)) - n \cdot \log\left(\int_0^1 e^{w(t)}dt\right) + \\ &\log\left(\sum_{g=1}^{G} \mathrm{TP}_{2a_\kappa}(w|0, c^{SE}(\cdot,\cdot|\lambda_g))\pi^*_\lambda(\lambda_g)\right) + \log(\pi(\theta)). \end{aligned}$$

A.4 Predictive Process Approximation

A low-rank predictive process replaces w in the likelihood. Specifically, $\tilde{w}(\tau) = E(w(\tau)|w(t_1^*), ..., w(t_M^*))$ replaces $w(\tau)$ using a finite-predetermined set of knots (t_1^*, ..., t_M^*). We further define $W^* = (w(t_1^*), ..., w(t_M^*))$. Recall that w is a π_λ^*-prior-weighted mixture of multivariate t-processes that depend on λ_g. The expectation of a t-process conditioned on a finite subset of its points follows the form $E(X_2|X_1) = \Sigma_{21}\Sigma_{11}^{-1}X_1$, which is reminiscent of the expectation of a partitioned and conditioned multivariate normal. Because this conditional expectation depends on the covariance kernels, which in turn depend on λ, further conditioning is needed (i.e. law of total expectation):

$$\tilde{w}(\tau) = E(w(\tau)|W^*) = E_{\lambda|W^*}\{E(w(\tau)|W^*, \lambda)\}$$

$$= \sum_{g=1}^{G} \pi(\lambda = \lambda_g|W^*) \cdot E(w(\tau)|W^*, \lambda).$$

By Bayes rule, for any given λ_g and realization ψ of W^*

$$\pi(\lambda = \lambda_g|W^* = \psi) = \frac{\pi(W^* = \psi|\lambda = \lambda_g) \cdot \pi(\lambda = \lambda_g)}{\sum_{g=1}^{G} \pi(W^* = \psi|\lambda = \lambda_g) \cdot \pi(\lambda = \lambda_g)}.$$

From the priors, $\pi(W^* = \psi|\lambda = \lambda_g) = \text{MVT}_{2a_k}^M(\psi|0, C_{**}(\lambda_g))$, where $C_{**}(\lambda_g)$ denotes $C(\cdot, \cdot|\lambda_g)$ evaluated at the knots, and $\pi(\lambda = \lambda_g) = \pi_\lambda^*(\lambda_g)$. Therefore the conditional probability is

$$\pi(\lambda = \lambda_g|W^* = \psi) = \frac{\text{MVT}_{2a_k}^M(\psi|0, C_{**}(\lambda_g))\pi_\lambda^*(\lambda_g)}{\sum_{g=1}^{G} \text{MVT}_{2a_k}^M(\psi|0, C_{**}(\lambda_g))\pi_\lambda^*(\lambda_g)}.$$

These can be recovered *a posteriori* even though λ is integrated out of the posterior.

Letting $C_{\tau*}(\lambda_g)$ denote the evaluation of the covariance kernel at input τ and the knots, $E(w(\tau)|W^* = \psi, \lambda = \lambda_g) = C_{\tau*}(\lambda_g) \cdot C_{**}(\lambda_g)^{-1}\psi$ by the aforementioned

expectation of a conditioned t-process. Bringing it all together,

$$\tilde{w}(\tau) = \sum_{g=1}^{G} \frac{\mathrm{MVT}_{2a_k}^{M}(\psi|0, C_{**}(\lambda_g))\pi_{\lambda}^{*}(\lambda_g)}{\sum_{g=1}^{G} \mathrm{MVT}_{2a_k}^{M}(\psi|0, C_{**}(\lambda_g))\pi_{\lambda}^{*}(\lambda_g)} C_{\tau *}(\lambda_g) C_{**}(\lambda_g)^{-1}\psi \tag{A.2}$$

provides a functional approximation to $w(\tau)$ given $W^* = \psi$ at the knots.

A.5 Finite Approximations to Posterior

To allow for pre-computation of $C_{\tau *}(\lambda)$, $\tilde{w}(\tau)$ is only tracked over a dense, finite grid at $\tau \in \{t_1, ..., t_L\}$, and functional likelihood evaluations $\tilde{w}(\cdot)$ are performed by interpolating between grid points. $C_{**}(\lambda)$ is also pre-computed.

To facilitate MCMC, we complete the infinite-to-finite transformation by replacing the prior on functional w by a prior on finite W^*. That is, we use the same M-dimensional W^* that facilitates the reduced-rank $\tilde{w}(\tau)$ in the likelihood evaluation to approximate w in the prior. The prior changes from being a mixture of t processes to being a mixture of M-dimensional multivariate t distributions. The final approximated log posterior is as follows

$$\log(\pi(\nu, \sigma, W^*|Y)) \propto \sum_{i=1}^{n} \log(g_\theta(y_i)) + \sum_{i=1}^{n} \tilde{w}(G_\theta(y_i)) - n \cdot \log\left(\int_0^1 e^{\tilde{w}(t)} dt\right) +$$
$$\log\left(\sum_{g=1}^{G} \mathrm{MVT}_{2a_\kappa}^{M}(W^*|0, C_{**}(\lambda_g))\pi_{\lambda}^{*}(\lambda_g)\right) + \log(\pi(\nu, \sigma^2)))$$

with $\pi(\nu, \sigma^2)$ defined in A.1 and $\tilde{w}(\tau)$ defined in A.2.

Appendix B

Alternative-kernel Mixture Model

This appendix provides details for an alternative mixture model, similar to the mixture approach of Chapter 3 but constructed with a customized kernel made of a mixture of four beta densities in lieu of the truncated normal kernel. Here, only a finite mixture model (FMM) and not a Dirichlet process mixture model (DPMM) is implemented.

B.1 Model Setup

B.1.1 Parametric Distribution and Priors

This beta mixture model uses a GPD, parameterized by shape parameter $\xi = 1/\nu$ and scale parameter σ, as its parametric transformation G_θ within the transformation framework. Priors and transformations for ν and σ are the same as those used in the truncated-normal-kernel model in Chapter 3.

B.1.2 Nonparametric Distribution and Priors

The mixture prior for h takes the general form $h(\tau) = \int k(\tau|\Omega, \Psi)dP(\Omega, \Psi)$. The kernel, $k(\tau|\Omega, \Psi)$, is constructed from a mixture of four components, each a beta

density as

$$k(\tau|\Omega,\Psi) = \omega_1 \text{Beta}(\tau|1,1) + \omega_2 \text{Beta}(\tau|1,\beta_2) + \omega_3 \text{Beta}(\tau|\alpha_3,1) + \omega_4 \text{Beta}(\tau|\alpha_4,\beta_4).$$

By constraining the shape parameters $\Psi = \{\beta_2, \alpha_3, \alpha_4, \beta_4\}$ so that each is greater than 1, we guarantee that none of the four beta density components goes to infinity on $[0,1]$. Additionally, constraining $\omega_1 > 0$ (or alternatively simultaneously constraining $\omega_2 > 0$ and $\omega_3 > 0$) guarantees that the beta-mixture kernel is also greater than 0 at the boundary. For simplicity, each weight in $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ is constrained to be strictly greater than 0 with $\omega_1 + \omega_2 + \omega_3 + \omega_4 = 1$.

We take a small diversion to note that finite mixtures can be represented by latent-category likelihoods. Consequently a single beta-mixture kernel can be represented marginally as

$$x|\Omega,\Psi \sim \omega_1 \text{Beta}(1,1) + \omega_2 \text{Beta}(1,\beta_2) + \omega_3 \text{Beta}(\alpha_3,1) + \omega_4 \text{Beta}(\alpha_4,\beta_4)$$

with $\omega_1+\omega_2+\omega_3+\omega_4 = 1$ or jointly given the underlying group/cluster memberships as

$$x|z,\Psi \sim \text{Beta}(1,1)^{\mathbb{1}(z=1)} \text{Beta}(1,\beta_2)^{\mathbb{1}(z=2)} \text{Beta}(\alpha_3,1)^{\mathbb{1}(z=3)} \text{Beta}(\alpha_4,\beta_4)^{\mathbb{1}(z=4)}$$

with $z|\Omega \sim \text{Multinomial}(\omega_1, \omega_2, \omega_3, \omega_4)$.

Similarly, a finite mixture of beta-mix densities, which we use in place of the more computationally-intense infinite Dirichlet process mixture, can be represented as

$$y|P \sim \sum_{k=1}^{K} \rho_k \text{BetaMix}(\Omega_k, \Psi_k) \text{ with } \sum_{k=1}^{K} \rho_k = 1 \text{ and } P = (\rho_1, \rho_2, ...\rho_K)$$

or as

$$y|w \sim \prod_{k=1}^{K} \text{BetaMix}(\Omega_k, \Psi_k)^{\mathbb{1}(w=k)} \text{ with } w|P \sim \text{Multinomial}(\rho_1, \rho_2, ..., \rho_K). \quad \text{(B.1)}$$

When combined with a Dirichlet prior, these multinomial-Dirichlet likelihoods lend themselves well to Bayesian MCMC sampling because the probabilities can be updated with conjugate Dirichlet draws; latent categories can be simulated for each observation from a multinomial distribution; and updates to kernel parameters only depend on data within the associated latent category. We use $\Omega_k \sim \text{Dirichlet}(4, 1.001, 1.001, 2)$ for the probability parameters that weight the four components of each beta-mixture kernel $k = 1...K$. This Dirichlet prior guarantees that all component weights are greater than 0 while placing somewhat higher weight on the uniform component. A Dirichlet prior $P = (\rho_1, \rho_2, ..., \rho_K) \sim \text{Dirichlet}(0.5, 0.5, ...0.5)$ for the weights that mix the K kernels favors higher weights on fewer kernels.

Finally, the shape parameters of the beta-mixture kernels, Ψ_k, are constrained to be greater than 1 using the following priors:

$$\beta_2 - 1 \sim \text{Gamma(shape=1, rate=1/2)}$$

$$\alpha_3 - 1 \sim \text{Gamma(shape=1, rate=1/2)}$$

$$\alpha_4 - 1 \sim \text{Gamma(shape=2, rate=1/3)}$$

$$\beta_4 - 1 \sim \text{Gamma(shape=2, rate=1/3)}$$

Figure B.1 shows several draws from the resultant prior for $h(\tau)$.

Before moving on to computational considerations, we stop to note that a complete confounding exists between GPD transformations on uniformly distributed Y with GPD transformations on $\text{Beta}(1, b_2)$ power-transformed Y. Specifically,

$$g(y; \xi, \sigma) \times \text{Beta}(G(y; \xi, \sigma); 1, b) \equiv g\left(y; \frac{\xi}{b}, \frac{\sigma}{b}\right) \times \text{Beta}(1, 1), \tag{B.2}$$

where g and G represent the pdf and CDF of a GPD respectively, and Beta is the pdf of a beta distribution. In separately estimated situations, this confounding would be a problem, but within the beta-mixture kernel there should be enough information to tease out b_2, ξ, and σ if the sample size is large enough. Potentially confounded

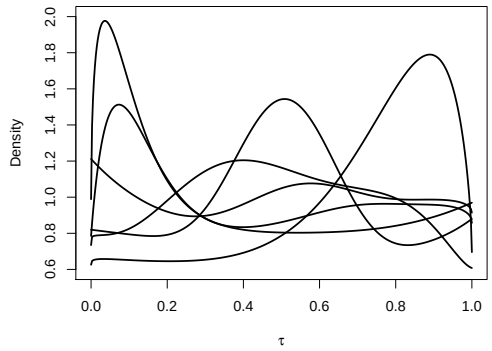


FIGURE B.1: Nonparametric draws from prior for $h(\tau)$. Each is constrained such that $0 < h(\tau) < \infty$ for all τ.

observations should either allocate to the latent uniform component with tail index ξ or to the latent Beta$(1, b_2)$ component with tail index ξ/b_2. However, because of the constraint that $b_2 > 1$, the tail heaviness for the overall kernel will be at most ξ. This should become estimable with large enough n, even if only a small proportion of the data belongs to the latent uniform component versus the latent Beta$(1, b_2)$ component.

B.2 Computation

Bayesian posteriors over all model parameters are obtained through Gibbs sampling. Algorithm 8 of Neal (2000), appropriate in cases of non-conjugacy, can be used to obtain group-membership updates from an infinite Dirichlet process mixture. However, these updates require sequential processing across all n observations and are slow. Instead, we recommend replacing the Dirichlet-process prior with a Dirichlet prior and using a finite mixture approximation with $K = \lceil \log(n) \rceil$ groups (see Equation B.1). This provides a sufficiently flexible form for most nonparametric scaling distributions but has the added advantage of being able to simultaneously update all n group memberships, insofar as the implementation software supports vector

operations. We perform all computing in R (R Core Team, 2018), and so this finite approximation and these vectorizations provide significant speedups.

Conjugate updates are available for many parameters. Given latent observation memberships w_i to the K beta-mixture kernels, the probabilities in vector P, which dictate probabilistic membership among K groups, are updated with a single conjugate Dirichlet draw. The latent memberships w_i use conjugate multinomial updates, which are vectorizable as previously mentioned. Given these memberships, w_i, and the memberships z_i among the four BetaMix components, the inter-beta-mixture probabilities Ω_k for $k = 1, ..., K$ can be updated with conjugate Dirichlet updates. The corresponding z_i, now given all Ω_k, are updated with conjugate, vectorizable, multinomial draws.

Non-conjugate draws within the Gibbs sampler are made via Metropolis Hastings updates. The parameters b_{2_k} and a_{3_k} each correspond to a beta-mixture component which is a power transformation distribution. Under full real support, these parameters would be conjugate with a gamma prior; however, the support for these parameters is shifted to be strictly greater than 1, so conjugacy does not apply. However, by taking the full-real-conjugate gamma and truncating greater than 1, we create a proposal density for the Metropolis Hastings algorithm that achieves "near-conjugacy," or a shape very close to the target density, that does not depend on the previous draw from the MCMC chain, i.e. an independence proposal. This allows for targeting of high acceptance rates without introducing as much problematic autocorrelation.

Parameters a_{4_k} and b_{4_k} are obtained jointly from a random-walk-centered sampler. The proposal density is a two-dimensional multivariate normal (MVN) distribution truncated to $[1, \infty] \times [1, \infty]$. The covariance is data-estimated from a beta information matrix if $\mathbb{N}(w_i = k \text{ and } z_i = 4) > 5$ and $10I_2$ otherwise. This creates a Laplace-like proposal for the joint parameter set.

Table B.1: Simulation MCMC summaries

Simulation Scenario	n	Chain length	Burn	Retain	Converged
Standard GPD	1,000	16,000	1,000	5,000	95%
Half-t	1,000	23,000	3,000	5,000	92%
Fourth-power GPD	1,000	23,000	5,000	5,000	36%
Gamma-GPD mixture	1,000	16,000	1,000	5,000	16%
Half-t-normal mixture	2,000	18,000	2,000	5,000	93%
Spliced Gamma-GPD	5,000	18,000	10,000	5,000	20%

The final non-conjugate update for joint (ν, σ) also uses a random-walk-centered sampler. After transforming ν and σ from $[0, \infty] \times [0, \infty]$ to $[-\infty, \infty] \times [-\infty, \infty]$, a MVN_2 proposal density with covariance obtained from a numerically-optimized Hessian is used. A true multivariate Laplace proposal, e.g. centered at the MAP, seems to miss the "corner" regions of the cornucopia-shaped posterior with its strictly-elliptical proposals. Centering at the previous MCMC draw allows the proposals to better explore these corners.

Finally, it is proposed that a "warm" start be used to speed up chain convergence. This can be obtained by numerically optimizing a simplified posterior that only contains a single beta-mixture kernel, which amounts to a 10-dimensional optimization over a box-constrained space (2 GPD parameters, 4 beta shape parameters, 4 component probabilities). Observations can then be randomly assigned membership among the K groups, each of which starts with these optimized beta-mixture parameters.

B.3 Simulation Study

The same suite of six simulation scenarios used in Chapter 2 are considered here against a similar set of comparison methods and using similar metrics.

For each simulated dataset under each scenario, two chains are run, one with a

warm start and one with parameters being drawn from the prior. Table B.1 summarizes the number of MCMC iterations obtained and the number of MCMC samples discarded after visual inspection of trace plots. Post burn-in, the convergence of MCMC chains is assessed for each posterior separately by considering Gelman-Rubin potential scale reduction factors (PSRFs) for ν and σ. PSRFs for mixture model parameters are not assessed since label-switching is accepted and not adjusted for, and PSRFs under this unidentifiable scenario would be meaningless. Only those replicates which have upper-95% confidence limits below 1.3 for both ν and σ's univariate PSRFs are included in subsequent analysis. The standard GPD, half-t, and half-t-normal mixture simulations have at least 90% of replicates achieving the PSRF standard (percentages in Table B.1), while the other three have considerably worse convergence rates. The results of the three convergent simulations are described in the next subsection. A discussion of the simulations experiencing poor convergence follows.

B.3.1 Convergent Simulation Results

Tail-Index Results. Table B.2 shows that when data come from a standard GPD, the best tail-index coverage and narrowest widths come, not unsurprisingly, from ML estimation without thresholding; however, those intervals widen quickly with higher thresholds and less data. The FMM tail index estimate is unbiased, even if its coverage is too high for the nominal level (e.g. 100% vs. 95%). While the interval widths may be unnecessarily wide, they are not unusually wide, falling between the widths of ML estimators that retain 100% and 20% of data.

Both the half-t scenario and the half-t-normal mixture scenario have the FMM performing well in tail-index comparisons to the maximum likelihood estimators. Because the tails of the true densities are *not* GPD, the ML bias increases when greater proportions of tail data are included, and the ML interval coverage decreases. Cov-

Table B.2: Simulation tail-index results for three converged simulation scenarios across comparison methods. Includes mean tail-index estimates (Mean); interval coverage (Cov), i.e. proportion of intervals across replicated datasets that contain the true tail-index; and width of 95% confidence or credible intervals (Wid).

	Standard GPD			**Half-t**			**Half-t-normal**		
Method	Mean	Cov	Wid	Mean	Cov	Wid	Mean	Cov	Wid
Truth	**0.25**			**0.25**			**0.25**		
GPD 5%	0.19	0.94	0.73	0.18	0.92	0.72	0.18	0.87	0.48
GPD 10%	0.25	0.95	0.52	0.18	0.91	0.47	0.17	0.80	0.32
GPD 20%	0.26	0.97	0.35	0.16	0.79	0.32	0.12	0.38	0.21
GPD 100%	0.25	0.93	0.15						
EGPD 5%	0.17	0.94	0.94	0.19	0.93	0.92	0.17	0.89	0.59
EGPD 10%	0.22	0.95	0.63	0.20	0.88	0.59	0.19	0.90	0.40
EGPD 20%	0.26	0.93	0.43	0.17	0.86	0.39	0.14	0.63	0.26
EGPD 100%	0.25	0.94	0.18						
FMM 100%	0.25	1.00	0.30	0.15	0.97	0.22	0.21	1.00	0.23

erage for the FMM tail index, on the other hand, stays high despite some bias. Additionally, FMM widths in both scenarios are among the narrowest of the comparison methods considered.

Quantile Estimation. Figure B.2, Figure B.3, and Figure B.4 show the results of tail-quantile estimation for the three simulations that converged well. With either $n = 1{,}000$ or $n = 2{,}000$, considering quantile level p out to 0.9999 reflects extreme extrapolation. For methods employing truncation, bias and RMSE are only displayed for p in the quantile-levels retained after truncation.

In Figure B.2, the quantiles from the FMM estimator on the standard GPD scenario show similar amounts of bias and RMSE to the GPD 100% and EGPD 100% estimators. It is promising that, despite calculating many additional nonparametric density parameters, the quantile estimates of the FMM behave close to the "gold-star," unbiased MLEs in the standard GPD case. Note that apparent bias in the extreme quantiles of the MLEs is likely due to small sample size, i.e. bias being

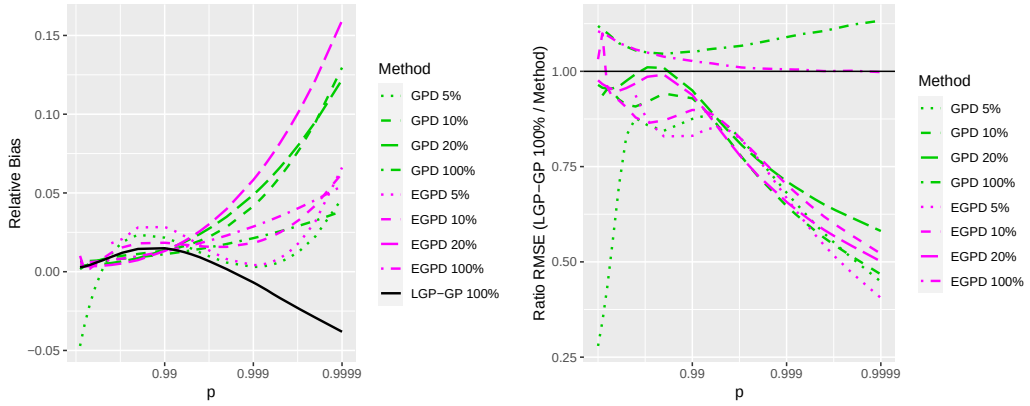


FIGURE B.2: Standard GPD simulation upper-tail, quantile-extrapolation with p on log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of FMM 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than FMM.

FIGURE B.3: Half-t simulation upper-tail, quantile-extrapolation with p on log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of FMM 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than FMM.

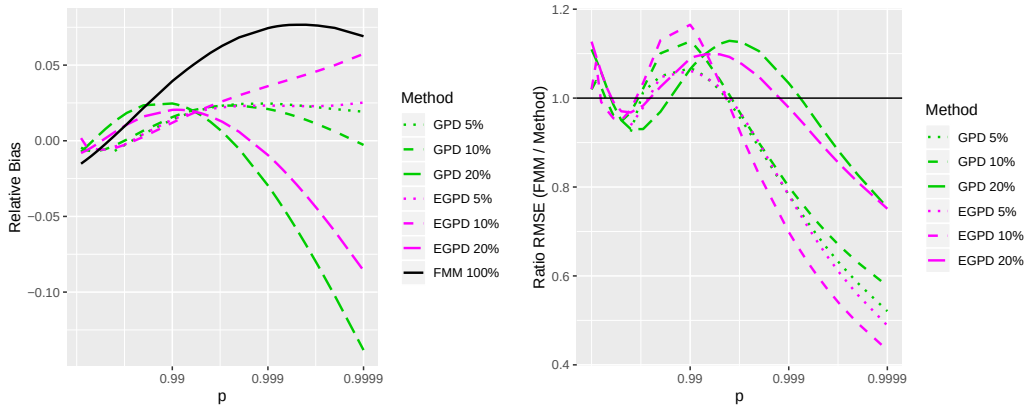


FIGURE B.4: Half-t-normal mixture simulation upper-tail, quantile-extrapolation with p on log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of FMM 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than FMM.

calculated over only 100 replicated datasets.

Figure B.3 shows the results of the half-t scenario. The FMM displays more bias and RMSE in the region where there is data; however, in extrapolation or beyond where data is available, the FMM has lower RMSE than any of the thresholding approaches. The same is true in the half-t-normal mixture simulation, displayed in Figure B.4, though with not as pronounced an effect.

B.3.2 Non-convergent Simulation Results

Similar to what is seen in Chapter 2, this FMM MCMC sampler can have difficulty moving through a complex posterior space, resulting in high autocorrelation or stickiness, especially apparent in σ. As the MCMC sampler moves among different σ possibilities, potentially very different shapes of the nonparametric densities h are needed to accommodate the true density f.

The plots in Figure B.5 illustrate a case of poor chain convergence occurring in the Gamma-GPD mixture simulation's first data-set replicate. The warm-start chain slowly wanders its way around a complex posterior space while the prior-start chain,

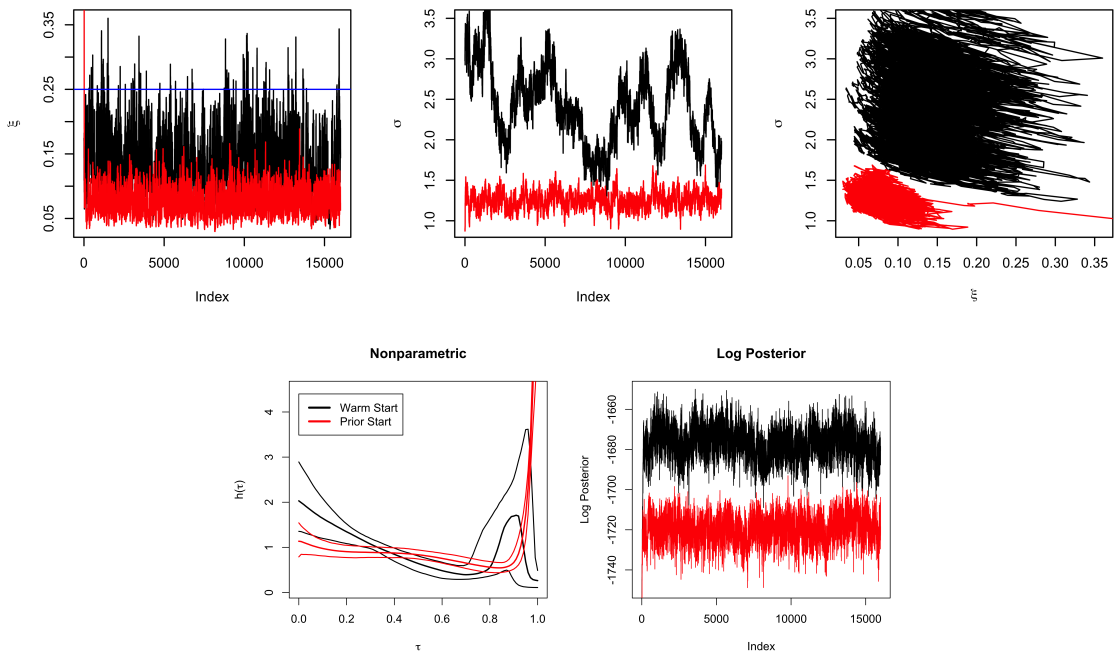


FIGURE B.5: Example of nonconvergent chains run on dataset 1 of the spliced gamma-GPD simulation: black chain comes from a warm start and red from a random draw from the prior. The upper row has traceplots for ξ and σ separately and jointly. The bottom left plot shows the 95% credible intervals for nonparametric density estimates, and the bottom right plot shows the unnormalized log posterior values at each iteration.

run on the same data, settles into a small σ value. The bivariate plot of ξ by σ shows multidimensional separation between the two chains' θ values, but it is not until plotting the two nonparametric density estimates against each other (see 95% posterior credible intervals of density estimates in Figure B.5) that it is apparent that these chains must lie in distant parts of the posterior space from each other. Figure B.5 also shows traceplots comparing unnormalized log posterior values of the two chains. While these posterior values do not give relative importance between the two chains, it is clear that the values coming from the warm-start chain have higher posterior likelihood than those from the prior-start chain.

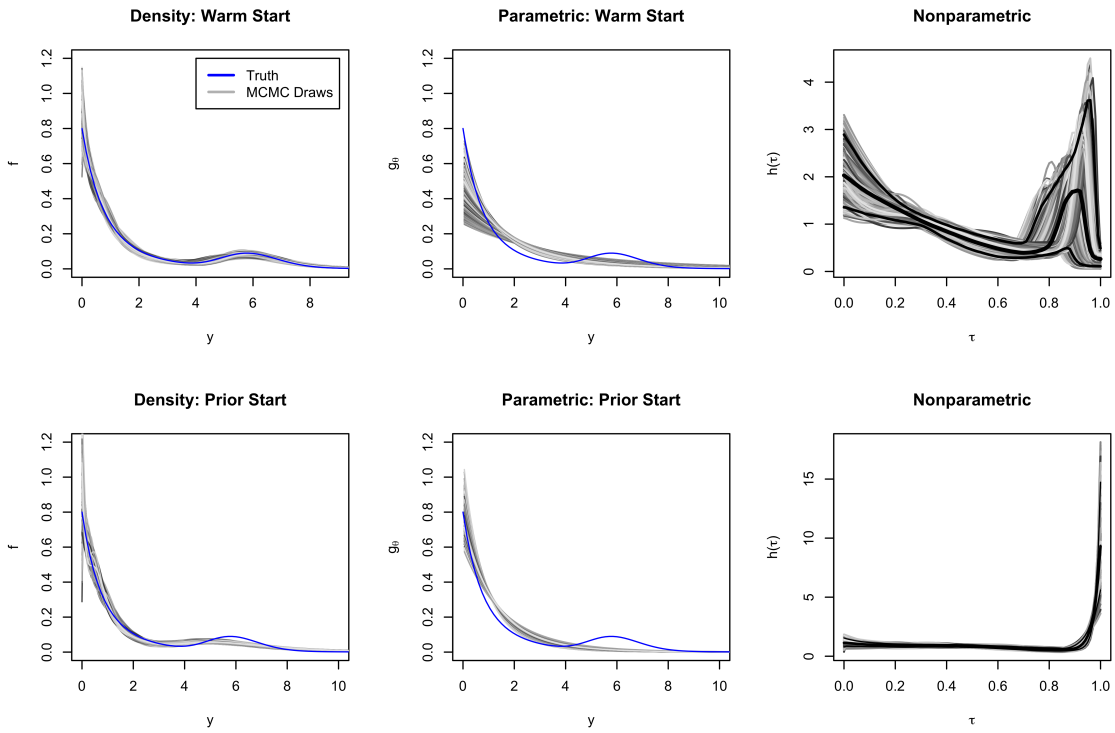


FIGURE B.6: Example of variance between two chains run on dataset 1 for the spliced gamma-GPD simulation. Left panels show draws for the overall density in grays and the truth in blue. Middle panels show draws for the transformation GPD density indexed by $\theta = (\xi, \sigma)$. Right panels show draws for the nonparametric density in grays with 95% credible intervals in bold black.

A second set of plots for the same simulation and the same dataset in Figure B.6 illustrate 1) the two chains' overall density estimates, 2) parametric or transformation density estimates under the sampled ξ and σ, and 3) density estimates for the nonparametric distributions. Notably, the "Density: Prior Start" completely misses the gamma mode from the gamma-GPD mixture. This is likely the reason that the prior-start chain has small log posteriors relative to the warm-start chain. Having this chain get stuck in such a distant, low-density area of the space is a stark example of the potential posterior sampling difficulty.

For each of the three "nonconvergent" scenarios (i.e. the fourth-power GPD, the

Gamma-GPD mixture, and the spliced gamma-GPD), there is great concordance between overall results when run on only the subset of chains that meet the PSRF criterion and when run on the entire set of chains with warm starts. This gives some reason to believe that the warm-start chains are in fact starting in and staying in the areas of highest posterior density and that poor random starts are what makes chains wander into low density areas. Given this concordance, we proceed with presenting the results for these three simulation scenarios, using only the warm-start chains, addressing each simulation separately.

Fourth-Power GPD Simulation Results

The true transformation, $G_\theta(\cdot)$, for the fourth power GPD simulation is the density of a Beta$(4, 1)$, which actually lies just outside of the model space because of the constraint that all kernels be strictly greater than 0. Therefore, it would be reasonable to expect posterior density on high probabilities associated with the kernels' third, boundary-mode, power-law component and at values near the true, generating $\sigma = 1$. However, under the generating transformation, the random variable's mass moves upward away from zero, and the prior for σ pulls its posterior upward in the direction of the data mean. Therefore a larger σ may be needed, and the corresponding "low-high-low" scaling pattern of the nonparametric density, associated with the kernels' fourth, central mode component, may point towards the existence of a true, second posterior mode.

Figure B.7 displays draws of the densities associated with dataset 95. The warm-start chain (upper three plots) successfully traverses between small values of σ with their "low-high" nonparametric pattern and larger values of σ with their "low-high-low" pattern. The random start chain, however, seems to get stuck in the small-σ regime.

Looking across dataset replicates, most chains appear to be wandering slowly

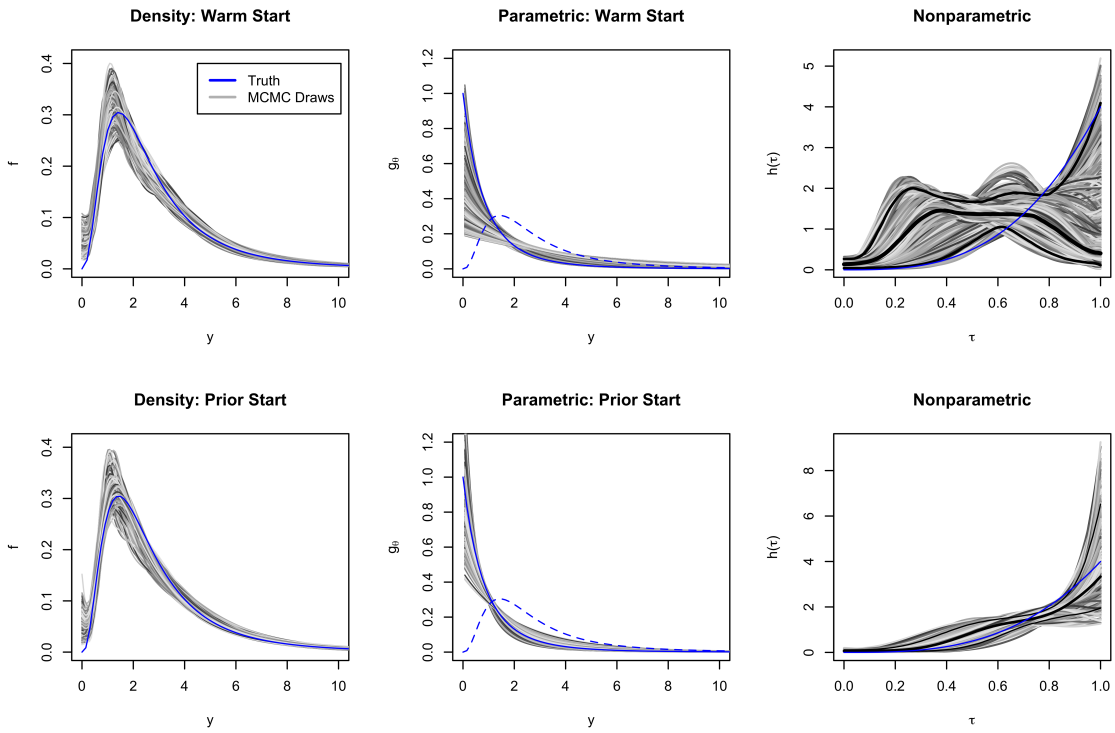


FIGURE B.7: Density plots for dataset 95 of the fourth power GPD simulation. Left panels show draws for the overall density. Middle panels show draws for the transformation GPD density indexed by $\theta = (\xi, \sigma)$. Right panels show draws for the nonparametric density in grays with 95% credible intervals in bold black. The truth for each is in blue.

through the mode with larger σ; however, some chains spend time in (or get stuck in) the small σ mode. There is no apparent difference between the warm-start chains and the prior-start chains as far as which chains end up where. A third set of chains was also run with starting values very close to the true nonparametric distribution and generating θ parameters, i.e. with $\sigma = 1$ and tail index $\xi = 0.25$. A visual comparison shows the new chains moving in similar patterns to the previous two chains. If for a given dataset the first two chains seem to jump between two modes, the new chain also moves between both modes. If the chains seem to stick to the large-σ mode, the new chain quickly and successfully migrates from the $\sigma = 1$ into

Table B.3: Fourth-power gamma scenario tail-index results across comparison methods. Includes mean tail-index estimates (Mean); interval coverage (Cov), i.e. proportion of intervals across replicated datasets that contain the true tail-index; and width of 95% confidence or credible intervals (Wid).

Method	Mean	Coverage	Width
GPD 5%	0.18	0.88	0.72
GPD 10%	0.21	0.92	0.50
GPD 20%	0.21	0.90	0.34
EGPD 5%	0.19	0.88	0.93
EGPD 10%	0.20	0.89	0.61
EGPD 20%	0.22	0.89	0.41
FMM 100%	0.20	0.96	0.55

the large σ mode. All of this gives confidence that, while the posterior for any of these datasets *may* in fact be multimodal and while the PSRFs may seem large due to finite sampling of *very* slow mixing chains, in aggregate across simulation replicates the chains have converged to the right place! The following results use (arbitrarily) the warm-start chains and include all 100 dataset replicates.

While the tail index for the true density f is preserved under the generating Beta$(4, 1)$ transformation, the "GPD-ness" of the tail is not preserved. This is apparent in the decreasing coverage of the ML estimators compared to what they would be if the tails were exactly GPD instead of just approximately GPD (see Table B.3). The FMM produces similar tail index estimates to the ML methods. It manages to get near nominal 95% coverage of the tail index parameter, though at the expense of slightly wider intervals than the other methods.

A similar pattern plays out in the quantile estimates as was seen previously in the convergent simulation scenarios, as can be seen in Figure B.8. Namely, the FMM quantiles are biased, and while they may have poorer RMSE in the parts of the tail where data is still plentiful compared to the other methods, in extrapolation they actually have reduced RMSE.

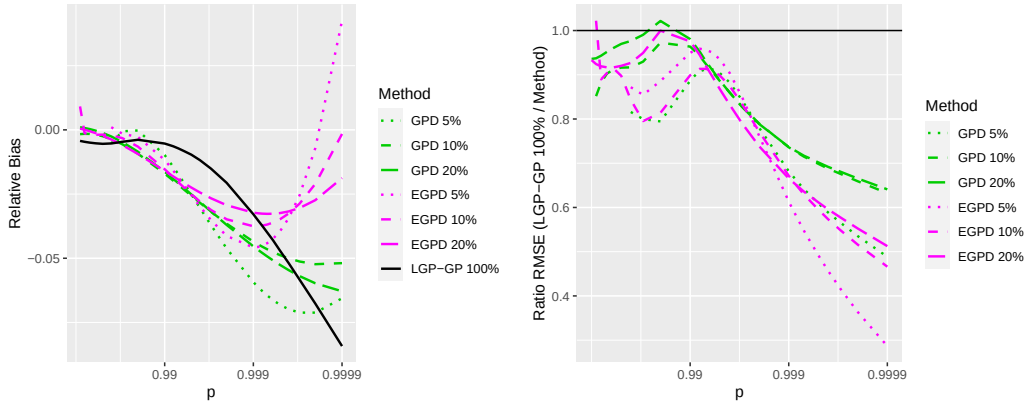


FIGURE B.8: Fourth-power GPD simulation upper-tail, quantile-extrapolation with p on log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of FMM 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than FMM.

Gamma-GPD Mixture Simulation Results

Reasoning might lead us to believe that because the true density for the gamma-GPD mixture scenario is a mixture of a $GPD(x; \xi = 1/4, \sigma = 1)$ with a gamma distribution and because the transformation family is a GPD that we should expect to see a nonparametric density that looks essentially uniform but with a spike coming from the gamma observations. However, that nonparametric density would need to pair with $\sigma = 1$, which σ would essentially ignore any contribution from the gamma observations. Instead, as can be seen when revisiting the "Parametric: Warm Start" plot of Figure B.6, the model prefers values of σ that allow the transformation density to cut through the middle of the multimodal truth. The corresponding nonparametric density (upper right plot of Figure B.6) is forced to first scale up and then scale down the parametric transformation (it "tilts" the uniform density), and the expected peak of the nonparametric density, which would be very close to 1 under $\sigma = 1$, is forced to move towards more central values of τ and away from 1. Across replicates the tilt in the left portion of the nonparametric density is apparent. It comes from high

Table B.4: Gamma-GPD mixture scenario tail-index results across comparison methods. Includes mean tail-index estimates (Mean); interval coverage (Cov), i.e. proportion of intervals across replicated datasets that contain the true tail-index; and width of 95% confidence or credible intervals (Wid).

Method	Mean	Coverage	Width
GPD 5%	0.43	0.92	0.78
GPD 10%	0.25	0.87	0.43
GPD 20%	0.01	0.07	0.19
EGPD 5%	0.50	0.85	0.95
EGPD 10%	0.32	0.90	0.51
EGPD 20%	0.11	0.48	0.25
FMM 100%	0.14	0.48	0.19

weight being given to Beta$(1, b_2)$ kernel components and is a manifestation of the confounding mentioned in Section B.1.2 for small sample sizes. That confounding will also affect tail index estimation. Table B.4 contains those results.

As expected, the tail index is biased down (too heavy) for the FMM, but most of the methods seem to be having difficulty dealing with the "contamination" coming from the gamma mixture component, as evidenced by their biased and/or poor coverage. The GPD 10% estimate gets closest to the true tail index of 0.25, but the method has coverage that is lower than desired, even with its wide intervals.

The quantiles in Figure B.9 tell a now-predictable story. FMM has the highest bias, and its RMSE is not as good as the other methods until exploring large quantile-levels, beyond the regions where we expect to have seen data with only $n = 1{,}000$.

Spliced Gamma-GPD Simulation Results

The replicates of the spliced gamma-GPD scenario were among the most likely to get stuck in local, small-σ modes when given random prior starts. For a single dataset, unnormalized log posteriors from the small-σ chains are clearly smaller than log posteriors coming from regions with larger σ values; however, there is not as distinct

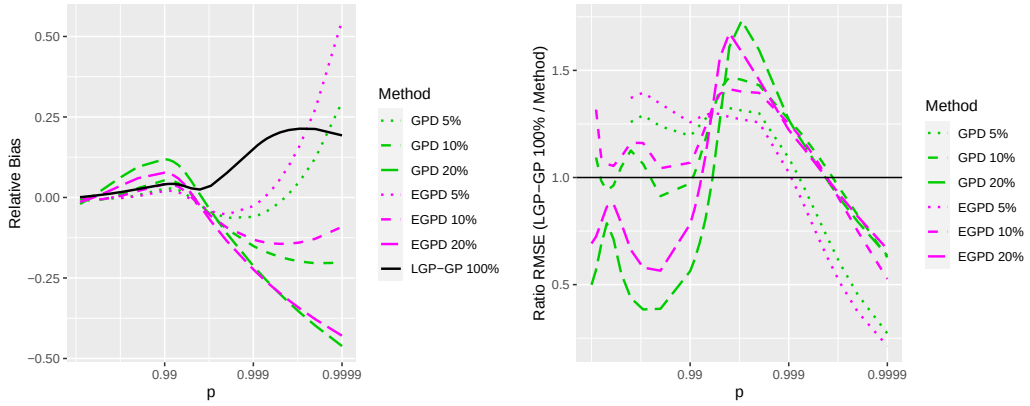


FIGURE B.9: Gamma-GPD mixture simulation upper-tail, quantile-extrapolation with p on log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of FMM 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than FMM.

Table B.5: Spliced gamma-GPD scenario tail-index results across comparison methods. Includes mean tail-index estimates (Mean); interval coverage (Cov); and width of 95% confidence or credible intervals (Wid).

Method	Mean	Coverage	Width
GPD 5%	0.23	0.94	0.31
GPD 8%	0.24	0.95	0.25
GPD 13%	0.21	0.85	0.18
EGPD 5%	0.22	0.93	0.38
EGPD 8%	0.24	0.94	0.30
EGPD 13%	0.24	0.96	0.22
FMM 100%	0.12	0.05	0.15

of separation in the log posteriors as for the gamma-GPD mixture case of Figure B.5. A small consolation is that across replicates the nonparametric distributions h produced by the warm-start chains all have similar shapes, which gives some credence to this warm-start-chain analysis.

With a mean tail index estimate across simulations of 0.12, the FMM is clearly biased (see Table B.5). Pairing that bias with narrow intervals, narrowest among all methods considered, there is correspondingly poor interval coverage. As seen in

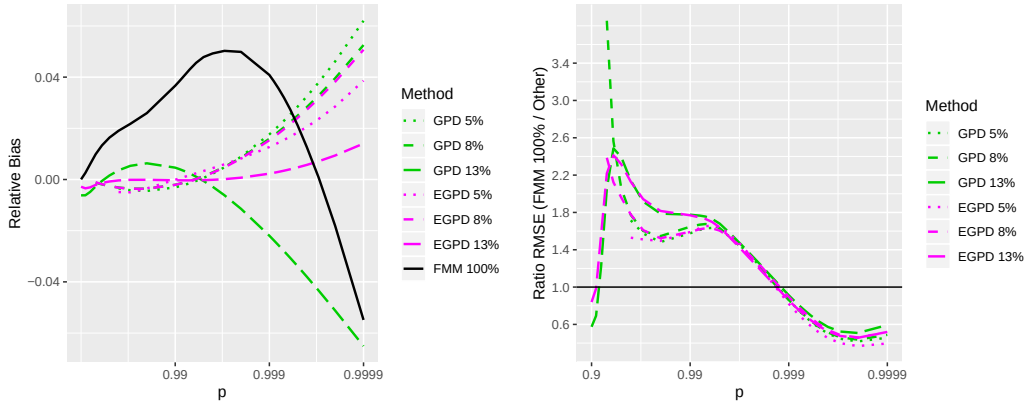


FIGURE B.10: Spliced gamma-GPD simulation upper-tail, quantile-extrapolation with p on log scale. Left panel shows relative bias (bias / true quantile value); right panel shows ratio of FMM 100% RMSE to other methods' RMSE, i.e. ratios greater than 1 indicate that other methods have lower RMSE than FMM.

Chapter 2, the maximum likelihood estimates with 8% and 5% of tail data retained do the best in terms of unbiasedness and tail-index coverage. The GPD with 13% of tail data retained begins to be biased by the gamma bulk and coverage falters, whereas the EGPD with 13% of data retained continues to do well. These are the scenarios for which the EGPD was designed.

Again, there is quantile bias through the lower part of the tail and reduced RMSE in extrapolation (Figure B.10). This is notable in that FMM does a better job of quantile estimation in extrapolation than the estimators *specifically designed to capture these exact tail cases*, and it does this *despite* grossly overestimating the heaviness of the tail. The nonparametric density must be performing a compensating roll. This is another testament to the power of retaining the data.

B.3.3 Simplifying the Kernel

A common mechanism may be hampering the convergence of all three "nonconvergent" simulations and could possibly even be contributing to the bias seen in the tails. Namely, if the sampler has difficulty sampling extremely large α_4 and simultaneously

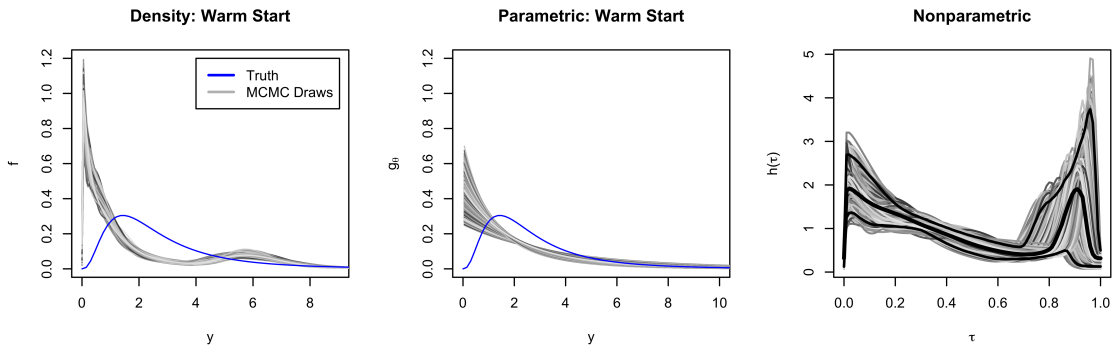


FIGURE B.11: Gamma-GPD mixture scenario densities using the modified kernel for Dataset 1. Left panel shows draws for the overall density in grays and the truth in blue. Middle panel shows draws for the transformation GPD density indexed by $\theta = (\xi, \sigma)$. Right panel shows draws for the nonparametric density in grays with 95% credible intervals in bold black.

small β_4 values for the Beta(α_4, β_4) component, that could either create a false barrier between observations that might otherwise move freely between the Beta(α_4, β_4) and Beta$(\alpha_3, 1)$ components or it could leave observations no better choice than to associate themselves with some accommodating boundary-mode Beta$(\alpha_3, 1)$ component, as was seen in the gamma-GPD mixture scenario. While this phenomenon was seen in the right tail values, it could also exist in left tail values if extremely large β_4 are unable to be sampled with small α_4 and observations permanently decamp to a Beta$(1, \beta_2)$ component.

This line of reasoning prompts the question: "If the sampler has difficulty converging in part because it gives inappropriate weight to the boundary-mode components of the beta-mixture kernel, i.e. those governed by α_3 and β_2 that allow the nonparametric density to go above 1 at the boundary, what happens if these two boundary-mode components are eliminated from the beta-mixture kernel?"

This was explored by removing the Beta$(1, \beta_2)$ and Beta$(\alpha_3, 1)$ components from the beta-mixture kernel and running the first four simulation scenarios again on all 100 datasets.

Table B.6: Tail-index results for the FMM across simulation scenarios using the modified beta-mixture kernel. Includes mean tail-index estimates (Mean); interval coverage (Cov); and width of 95% confidence or credible intervals (Wid).

Simulation Scenario	Mean	Coverage	Width
Standard GPD	0.28	1.00	0.34
Half-t	0.17	1.00	0.26
Fourth-power GPD	0.25	0.99	0.66
Gamma-GPD mixture	0.14	0.37	0.18

The most surprising result was that the form of each nonparametric density did not change much under the simplified kernel. For instance, Figure B.11 shows the densities for the same gamma-GPD mixture dataset that was discussed in Section B.3.2. The nonparametric density is able to recreate its characteristic tilt plus tall right mode without relying on any Beta$(1, b_2)$ components. The biggest change to the nonparametric density appears in τ near 0, where without the boundary-mode components to let the density be greater than 1 the nonparametric density is forced to drop below 1.

New tail index summaries for the four scenarios under the simplified kernel are in Table B.6. In the standard GPD scenario, the tail index went from 0.25 to 0.28 while also increasing in width from 0.3 to 0.34. The tail-index coverage remained unchanged at 100%. In the half-t scenario, the tail-index increased from 0.15 to 0.17, the width went from 0.22 to 0.26, and the coverage went from 97% to 100%. In the fourth-power GPD scenario, the tail index mean increased from 0.2 to 0.25, its 95% interval width increased from 0.55 to 0.66, and its coverage dropped from 99% to 96%. Finally, the gamma-GPD scenario kept its tail index of 0.14, decreased its tail-index coverage from 0.48 to 0.37, and decreased its width from 0.19 to 0.18. Overall, there were not drastic changes to the tail-index summaries.

No plots are included for the quantile bias and RMSE from these modified kernel

simulations. That is because each plot was nearly identical to the plots created for the full beta-mixture kernel. In short, when it comes to quantile estimation there is no additional benefit of using the full beta-mixture kernel over using the simplified beta-mixture kernel form; however, there are several benefits to giving up the full beta-mixture form in favor of the simplified kernel.

One benefit of eliminating these components is to reduce the total number of parameters in the model. With each beta-mixture kernel having 8 parameters (7 free parameters if accounting for the weights that are constrained to sum to 1), the total number of parameters estimated is $2 + K \times 8$ (correspondingly $2 + K \times 7$). Granted, the effective number of parameters is shrunk by the Dirichlet prior over the K kernels, but this is still a lot of parameters to be estimating. Dropping two of the four beta-mixture components reduces the total number of parameters per beta-mixture kernel to four plus a constraint on the weights and the total number of parameters in the model to $2 + K \times 4$ (correspondingly $2 + K \times 3$).

A second benefit to simplifying the kernel is that it may be possible to represent this simplified kernel mixture through an approximation to the Dirichlet process mixture model, namely via predictive recursion. The predictive recursion approximation allows the mixing distribution to be integrated out entirely. After integrating out the mixing distribution, all that is left is a marginal likelihood over the two θ parameters (Predictive Recursion Marginal Likelihood or PRML), which should be much easier to maximize. In theory this marginalization would be much faster and would allow for exploration of the θ space more fully.

B.3.4 Fixing the Transform Scale

While considering simplifications to the model and in response to the mixing difficulty which is so strongly related to σ, we also questioned whether σ need be estimated at all or if the value could be fixed. Nearly all literature that employs a transformation

Table B.7: Tail-index results for the FMM across simulation scenarios after fixing σ. Includes mean tail-index estimates (Mean); interval coverage (Cov); and width of 95% confidence or credible intervals (Wid).

Simulation Scenario	Mean	Coverage	Width
Standard GPD	0.26	0.97	0.21
Half-t	0.16	0.63	0.18
Fourth-power GPD	0.25	0.93	0.49
Gamma-GPD mixture	0.12	0.03	0.16

for heavy-tailed univariate density estimation does so by restricting the parameters θ of the transformation G_θ to a single parameter, the index of the power law. Even for those that have CDF transforms G_θ, most do not estimate a scaling parameter σ, unless they pre-estimate and fix it, thereafter estimating the nonparametric density. Those that do not include a scaling parameter directly in their model tend to scale their data by some empirical measure, e.g. by the inter-quartile range, prior to estimation. See references in Section 3.4.8.

Simulations were run over the first 30 datasets of the first four simulation scenarios, fixing σ. The value for σ was determined by estimating a ML tail index using 10% of the tail data, then using that index to back out an estimate for σ based on the empirical IQR under a GPD:

$$\hat{\sigma} = \frac{\hat{\xi}(\hat{Q}_{0.75} - \hat{Q}_{0.25})}{0.25^{-\hat{\xi}} - 0.75^{-\hat{\xi}}} \tag{B.3}$$

In each of the four scenarios, convergence of chains improved. After just 16,000 iterations, all 30 of the standard GPD and half-t datasets had upper limit PSRFs under 1.3. For the gamma-GPD mixture all but one replicate converged, and for the fourth-power GPD 21 (70%) had converged. That objective was met.

Fixing σ also resulted in the narrowing of all four tail-index interval widths (see Table B.7). Interval coverage was closer to the nominal level for both the standard-

GPD and the fourth-power GPD scenarios, but the interval coverage decreased for the half-t and gamma-GPD mixture simulations compared to when σ was estimated.

Although the quantile estimates followed very similar forms for bias and RMSE and still maintained their RMSE edge in the extrapolated-tails, most had some decline in quality when compared to the simulations where σ was estimated. The standard GPD and half-t scenarios both saw increased bias and RMSE in the tails and the gamma-GPD mixture had slightly higher RMSE as before. The fourth-power GPD quantile estimates were essentially unchanged.

While fixing σ may be a practically useful thing to do to improve convergence in tricky cases, it may impede the model's ability to correctly capture the tail index.

B.4 Conclusion

This FMM, while still plagued by slow mixing, is able to traverse the complex posterior shapes necessitated by varying σ. It has proven quite flexible and able to capture many bulk shapes while still giving form to the tails.

The model has some drawbacks. It still exhibits a bias in the lower part of the upper-level quantiles. It seems systemic as all simulations see the bias increase over the 0.9 to 0.99 region and perhaps a little beyond that. If the source of this bias is found, perhaps it could be remedied. The model also can get stuck in local modes in low-density parts of the parameter space; although this can largely be remedied using the proposed warm starts.

The model's strength lies in its use of data. By retaining all data and incorporating prior information that the tails are heavy, the model is able to significantly reduce the RMSE of quantile extrapolation across many types of densities and tails.

The most promising future direction lies in the simplified kernel. Index and quantile estimation does not seem significantly affected by removing four of the seven parameters of the beta-mixture kernel. Bringing the mixture kernel down to

three parameters puts it within the realm where predictive recursion approximations might be reasonable. Implementing this model using Predictive Recursion Marginal Likelihood seems like the next natural step. The approximation may not only speed up the algorithm, but may also simplify the complex posterior σ-h relationships by integrating the mixture distribution h out of the model entirely.

Appendix C

Model-Comparison Plots by Dataset

This appendix provides plots of marginal posteriors over the parametric transformation parameters for each method, for each scenario, and for dataset. One of the figures also provides dataset-level estimates for the relative bias through the bulk for each model and each scenario.

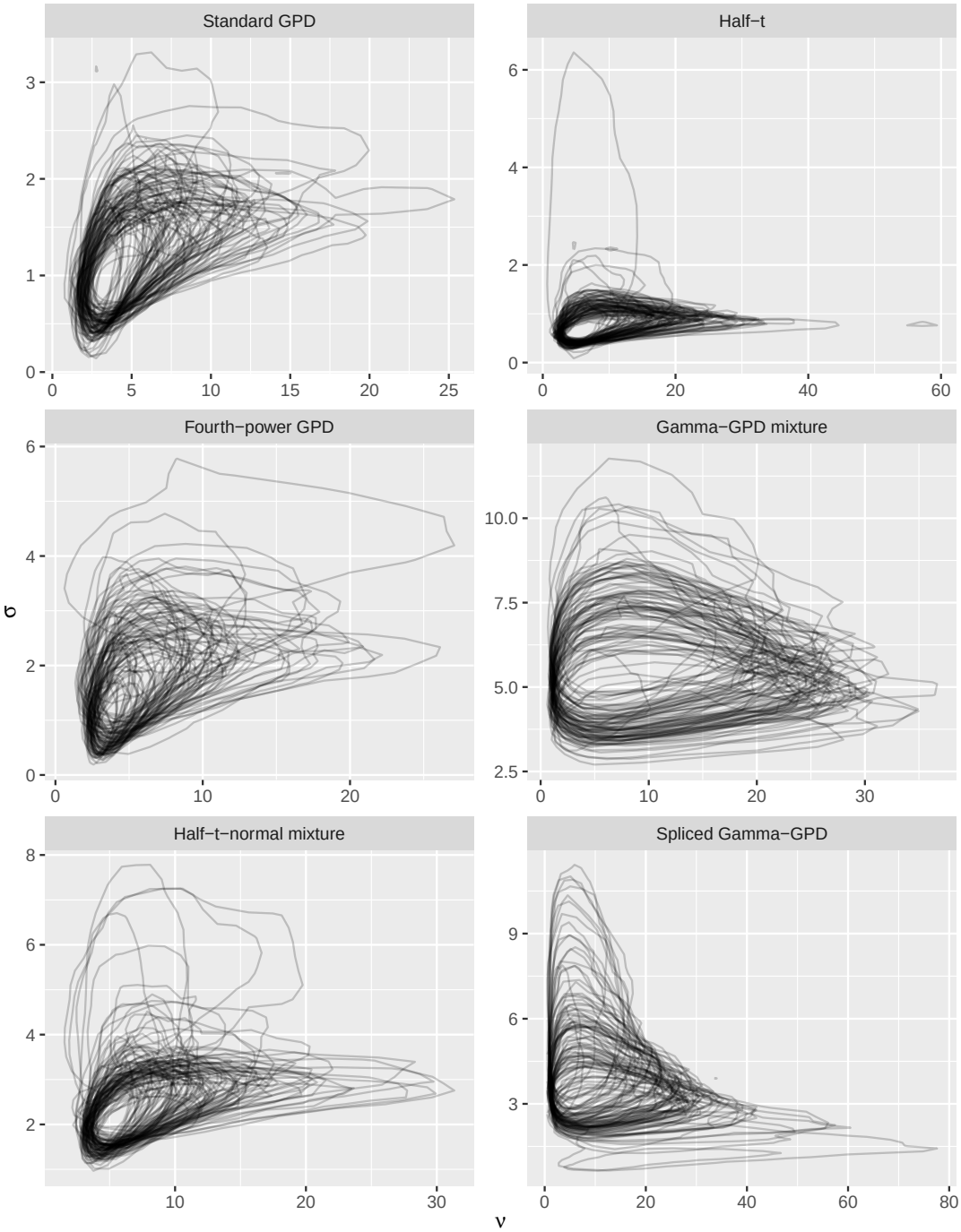


FIGURE C.1: LGP posterior marginal distributions over θ for each scenario and each dataset. Each black contour line is an isoband enclosing 80% of posterior mass. The shapes of the densities are similar to the 80% bands when extending out to enclose more of the mass.

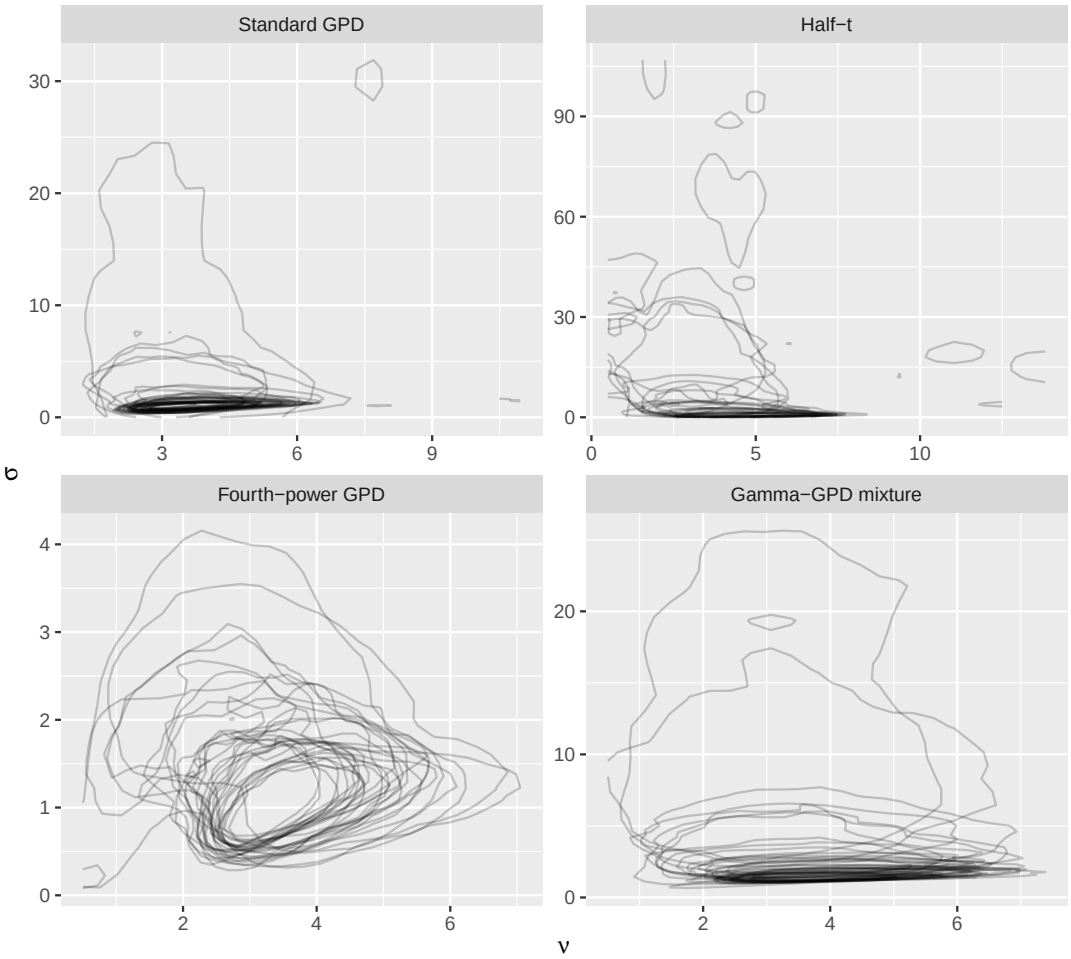


FIGURE C.2: FMM posterior marginal distributions over θ for each scenario and each dataset. Each black contour line is an isoband enclosing 80% of posterior mass. The shapes of the densities are similar to the 80% bands when extending out to enclose more of the mass.

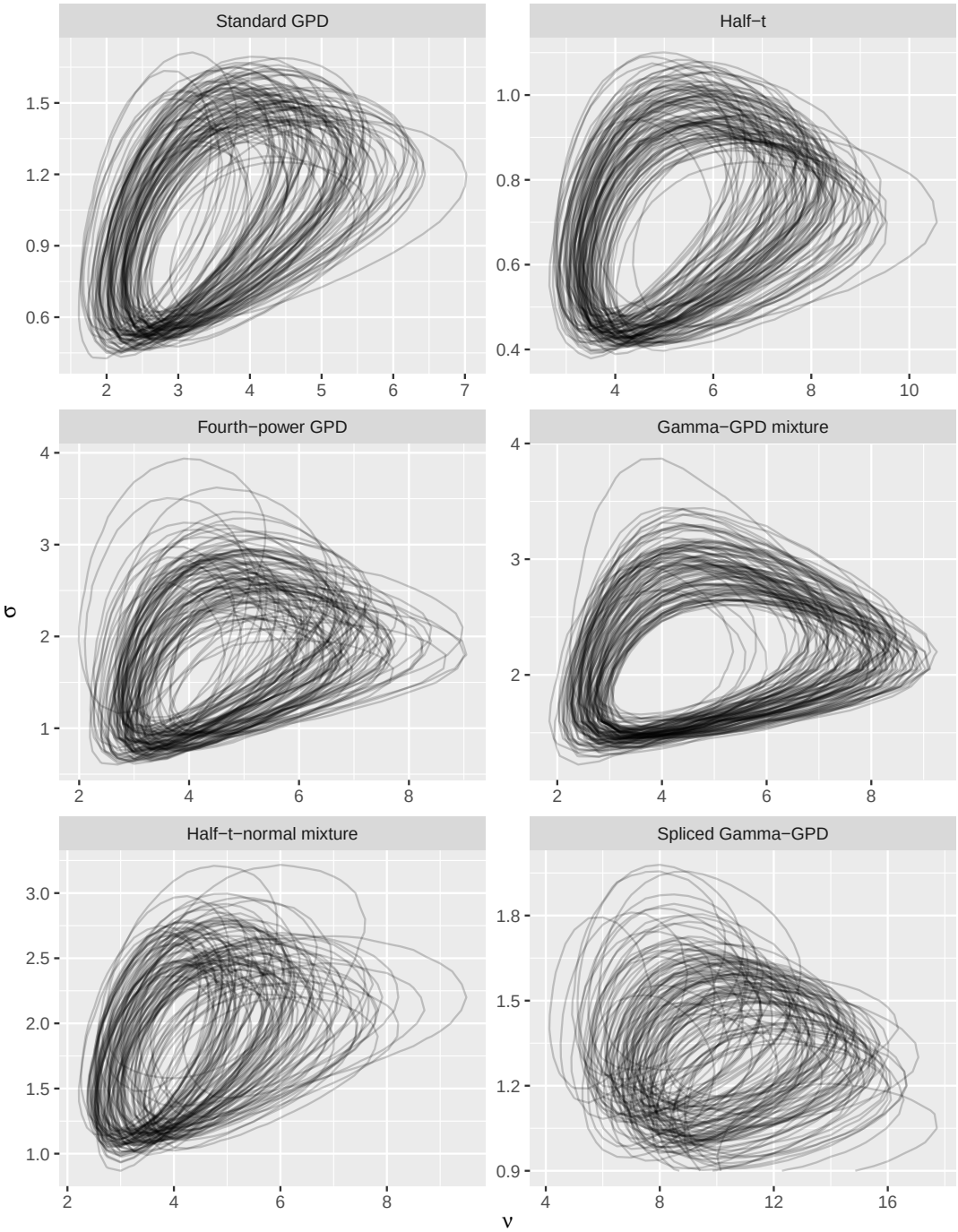


FIGURE C.3: PRMP posterior marginal distributions over θ for each scenario and each dataset. Each black contour line is an isoband enclosing 80% of posterior mass. The shapes of the densities are similar to the 80% bands when extending out to enclose more of the mass.

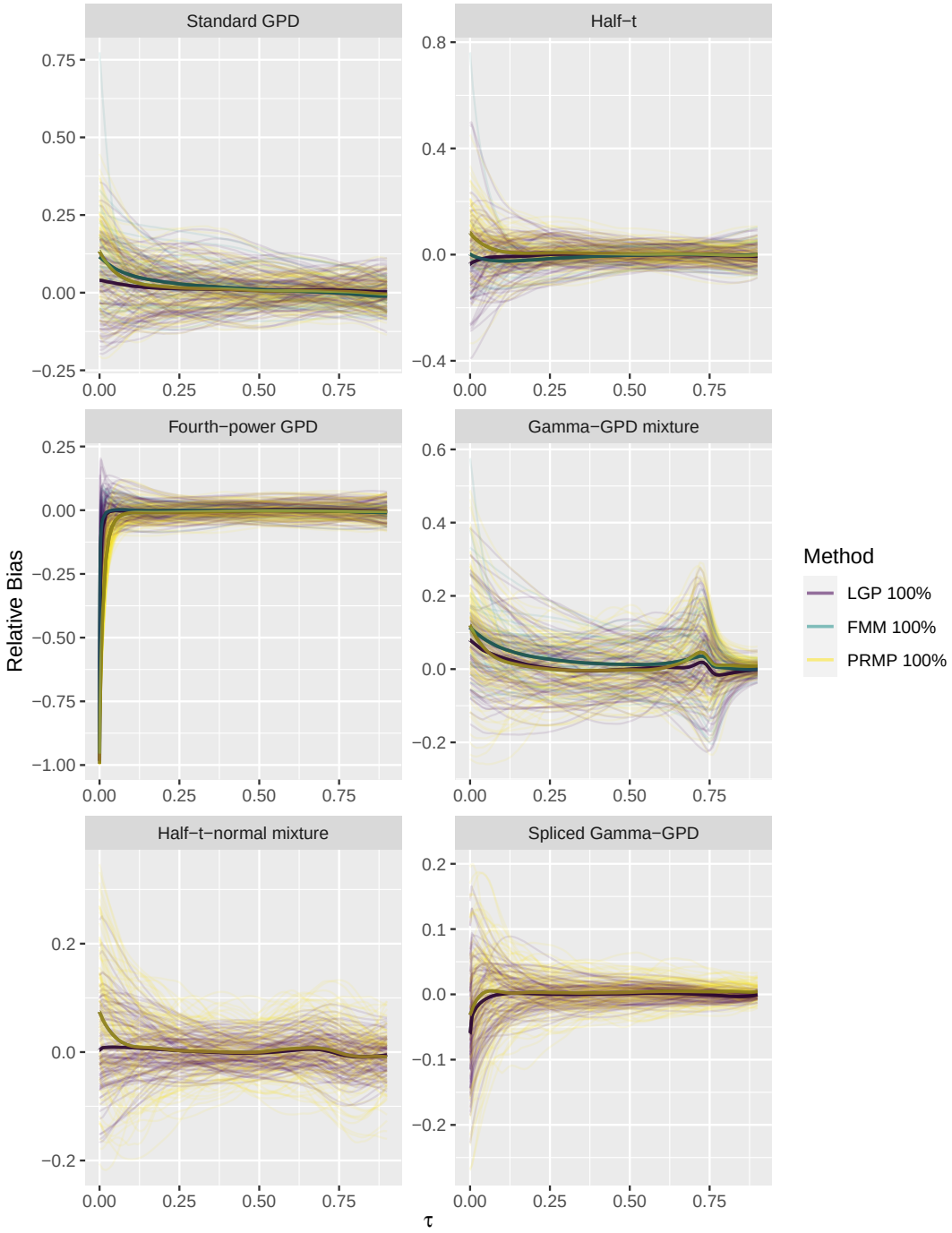


FIGURE C.4: Estimates of the relative quantile bias, (bias / true quantile value), for the bulk portion of the sampling density ($p \leq 0.9$) are displayed for each dataset. The averages, taken pointwise across p, are displayed in a darker hue.

Bibliography

Adams, R. P., Murray, I., and MacKay, D. J. (2009), "The Gaussian Process Density Sampler," in *Advances in Neural Information Processing Systems 21*, eds. D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, pp. 9–16, Curran Associates, Inc.

Andrieu, C. and Thoms, J. (2008), "A tutorial on adaptive MCMC," *Statistics and Computing*, 18, 343–373.

Balkema, A. and de Haan, L. (1974), "Residual life time at great age," *Annals of Probability*, 2, 792–804.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 825–848.

Bartlett, M. (1947), "The use of transformations," *Biometrics*, 3, 39–52.

Bean, A., Xu, X., and MacEachern, S. (2016), "Transformations and Bayesian density estimation." *Electronic Journal of Statististics*, 10, 3355–3373.

Belzile, L., Wadsworth, J. L., Northrop, P. J., Grimshaw, S. D., and Huser, R. (2018), *mev: Multivariate Extreme Value Distributions*, R package version 1.11.

Bolancé, C., Guillén, M., and Nielsen, J. P. (2003), "Kernel density estimation of actuarial loss functions," *Insurance: Mathematics and Economics*, 32, 19–36.

Bolancé, C., Guillén, M., and Nielsen, J. P. (2010), "Inverse beta transformation in kernel density estimation," *Statistics and Probability Letters*, 78, 1757.

Box, G. E. P. and Cox, D. R. (1964), "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 26, 211–252.

Brunk, H. (1978), "Univariate density estimation by orthogonal series," *Biometrika*, 65, 521–528.

Buch-Larsen, T., Guillen, M., Nielsen, J., and Bolancé, C. (2005), "Kernel density estimation for heavy-tailed distributions using the Champernowne transformation," *Statistics*, 39, 503–518.

Carreau, J. and Bengio, Y. (2009), "A hybrid Pareto model for asymmetric fat-tailed data: the univariate case," *Extremes*, 12, 53–76.

Chavez-Demoulin, V. and Embrechts, P. (2004), "Smooth extremal models in finance and insurance," *The Journal of Risk and Insurance*, 7, 183–199.

Clements, A., Hurn, A., and Lindsay, K. (2003), "Mobius-like mappings and their use in kernel density estimation," *Journal of the American Statistical Association*, 98, 993–1000.

Dixit, V. and Martin, R. (2019), "Permutation-based uncertainty quantification about a mixing distribution," arXiv preprint arXiv:1906.05349.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987), "Hybrid Monte Carlo," *Physics Letters B*, 195, 216–222.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230.

Fisher, R. A. and Tippett, L. H. (1928), "Limiting forms of the frequency distribution of the largest or smallest member of a sample." *Mathematical Proceedings of the Cambridge Philosophical Society*, 24, 180–190.

Frigessi, A., Haug, O., and Harvard, R. (2003), "A dynamic mixture model for unsupervised tail estimation without threshold estimation," *Extremes*, 5, 219–235.

Gilleland, E. and Katz, R. W. (2016), "extRemes 2.0: An Extreme Value Analysis Package in R," *Journal of Statistical Software*, 72, 1–39.

Gnedenko, B. (1943), "Sur la distribution limite du terme maximum d'une série aléatoeir," *Annals of Mathematics*, 44, 423–453.

Gustafsson, J., Hagmann, M., Nielsen, J. P., and Scaillet, O. (2009), "Local Transformation Kernel Density Estimation of Loss Distributions," *Journal of Business & Economic Statistics*, 27, 161–175.

Jones, M. C. and Faddy, M. J. (2003), "A Skew Extension of the t-Distribution, with Applications," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65, 159–174.

Kotecha, J. and Djuric, P. M. (1999), "Gibbs sampling approach for generation of truncated multivariate Gaussian random variables," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 3, 1757–1760.

Lenk, P. J. (1988), "The logistic normal distribution for Bayesian, nonparametric, predictive densities," *Journal of the American Statistical Association*, 83, 509–516.

Lenk, P. J. (2003), "Bayesian Semiparametric Density Estimation and Model Verification Using a Logistic-Gaussian Process," *Journal of Computational and Graphical Statistics*, 12, 548–565.

Leonard, T. (1978), "Density estimation, stochastic processes and prior information," *Journal of the Royal Statistical Society. Series B (Methodological)*, 40, 113–146.

Li, C., Lin, L., and Dunson, D. B. (2019), "On posterior consistency of tail index for Bayesian kernel mixture models," *Bernoulli*, 25, 1999–2028.

Markovich, N. (2007), *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, John Wiley & Sons, Ltd.

Martin, R. (2018), "On nonparametric estimation of a mixing density via the predictive recursion algorithm," unpublished, available at arXiv:1812.02149.

Martin, R. and Tokdar, S. T. (2009), "Asymptotic properties of predictive recursion: Robustness and rate of convergence," *Electronic Journal of Statistics*, 3, 1455–1472.

Martin, R. and Tokdar, S. T. (2011), "Semiparametric inference in mixture models with predictive recursion marginal likelihood," *Biometrika*, 98, 562–582.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016), "Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection," *Water Resources Research*, 52.

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of computational and graphical statistics*, 9, 249–265.

Newton, M. A. and Zhang, Y. (1999), "A recursive algorithm for nonparametric analysis with missing data," *Biometrika*, 86, 15–26.

Newton, M. A., Quintana, F. A., and Zhang, Y. (1998), "Nonparametric Bayes methods using predictive updating," in *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statistics*, eds. D. Dey, M. P., and S. D., vol. 133, pp. 45–61, Springer.

Papastathopoulos, I. and Tawn, A. J. (2013), "Extended generalised Pareto models for tail estimation," *Journal of Statistical Planning and Inference*, 143, 131–143.

Pickands, J. (1975), "Statistical inference using extreme order statistics," *Annals of Statistics*, 3, 119–131.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Resnick, S. I. (2007), *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*, Springer-Verlag New York.

Roberts, G. O. and Tweedie, R. L. (1996), "Exponential Convergence of Langevin Distributions and Their Discrete Approximations," *Bernoulli*, 2, 341–363.

Scarrot, C. and MacDonnald, A. (2012), "A review of extreme value threshold estimation and uncertainty quantification," *Statistical Journal*, 103, 33–60.

Schwartz, L. (1965), "On Bayes procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4, 10–26.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall.

Tokdar, S. T. (2007), "Towards a faster implementation of density estimation with logistic Gaussian process priors," *Journal of Computational and Graphical Statistics*, 16, 633–655.

Tokdar, S. T. and Ghosh, J. K. (2007), "Posterior consistency of logistic Gaussian process priors in density estimation," *Journal of Statistical Planning and Inference*, 137, 34–42.

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009), "Consistency of a recursive estimate of mixing distributions," *Annals of Statistics*, 37, 2502–2522.

Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010), "Bayesian density regression with logistic Gaussian process and subspace projection," *Bayesian Analysis*, 5, 319–344.

Tressou, J. (2008), "Bayesian nonparametrics for heavy tailed distribution. Application to food risk assessment," *Bayesian Analysis*, 3, 367–392.

van der Vaart, A. and van Zanten, J. (2008), "Rates of contraction of posterior distributions based on Gaussian process priors," *Annals of Statistics*, 36, 1435–1463.

van der Vaart, A. W. and van Zanten, J. H. (2009), "Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth," *Annals of Statistics*, 37, 2655–2675.

Verdinelli, I. and Wasserman, L. (1998), "Bayesian Goodness of Fit Testing using Infinite Dimensional Exponential Families," *Annals of Statistics*, 20, 1203–1221.

Wand, M. P., Marron, J. S., and Ruppert, D. (1991), "Transformations in density estimation," *Journal of the American Statistical Association*, 86, 343–353.

Yang, L. and Marron, J. S. (1999), "Iterated Transformation–Kernel Density Estimation," *Journal of the American Statistical Association*, 94, 580–589.

Yang, Y. and Tokdar, S. T. (2017), "Joint estimation of quantile planes over arbitrary predictor spaces," *Journal of the American Statistical Association*, 112, 1107–1120.

Zhu, D. and Galbraith, J. W. (2010), "A generalized asymmetric Student-t distribution with application to financial econometrics," *Journal of Econometrics*, 157, 297–305.