Heavy-Tailed Density Estimation

Surya T. Tokdar\textsuperscript{a}, Sheng Jiang\textsuperscript{b}, and Erika L. Cunningham\textsuperscript{a}

\textsuperscript{a}Department of Statistical Science, Duke University, Durham, NC; \textsuperscript{b}Department of Statistics, University of California, Santa Cruz, CA

\section*{ABSTRACT}
A novel statistical method is proposed and investigated for estimating a heavy tailed density under mild smoothness assumptions. Statistical analyses of heavy-tailed distributions are susceptible to the problem of sparse information in the tail of the distribution getting washed away by unrelated features of a hefty bulk. The proposed Bayesian method avoids this problem by incorporating smoothness and tail regularization through a carefully specified semiparametric prior distribution, and is able to consistently estimate both the density function and its tail index at near minimax optimal rates of contraction. A joint, likelihood driven estimation of the bulk and the tail is shown to help improve uncertainty assessment in estimating the tail index parameter and offer more accurate and reliable estimates of the high tail quantiles compared to thresholding methods. Supplementary materials for this article are available online.

\section*{1. Introduction}
For a heavy-tailed density with subexponential tail decay, the exceedance probabilities of a sample sum and a sample maximum are of the same order. A random sample drawn from such a density is likely to contain a small fraction of extreme observations whose magnitudes overshadow the sum total of the remaining magnitudes. This property is expressive of many naturally occurring phenomena, for example, precipitation (Katz, Parlange, and Naveau 2002), financial returns or insurance loss (Embrechts, Klüppelberg, and Mikosch 2013), and material or fatigue strength (Castillo 2012). However, statistically estimating a heavy tailed density from a random sample could be challenging if estimation was sought under only smoothness conditions. Two densities can be arbitrarily close in total variation distance while displaying entirely different tail decay rates. Estimation methods with rich shape flexibility and guaranteed $L^1$ estimation consistency may provide no meaningful influence on the tails of the distribution; see Markovich (2007) and Li, Lin, and Dunson (2019) for detailed discussions and cautionary results on kernel mixture models.

When interest focuses on estimating only tail features, for example, extrapolating to high quantiles from limited data, it is common to exclude all but the most extreme observations so that the tail speaks for itself. The Pickands-Balkema-de Haan Theorem (Balkema and de Haan 1974; Pickands 1975) justifies the so-called peak-over-threshold estimation methods, where a generalized Pareto distribution (GPD) is fitted to the subsample of observations exceeding a high threshold; see de Zee Bermudez and Kotz (2010) for a review. It also motivates nonparametric methods (Hill 1975; Pickands 1975; Dekkers et al. 1989; Alves 2001) based on only high sample quantiles for estimating the asymptotic tail decay rate of densities $f(y)$ whose survival function $\hat{F}(y) = \int_y^\infty f(t)\,dt$ is regularly varying, that is,

$$\hat{F}(y) = y^{-\alpha}L(y), \quad y > 0, \quad (1)$$

for some $\alpha > 0$ where $L(y)$ is a slowly varying function, that is, $\lim_{y \to \infty} L(ay)/L(y) = 1$ for every $a > 0$. We shall call such an $f(y)$ a regularly varying density with tail index $\alpha$, which may be recovered from $f$ as $\alpha = \alpha_+ := -\lim_{y \to \infty} \log \hat{F}(y)/\log y$.

A data driven threshold selection is critical to the analysis, but an optimal choice proves a steep challenge in practice. Diagnostic plots may point to multiple regimes of transition to the tail. Automatic threshold estimation methods gloss over such ambiguity with unverifiable tail assumptions and fail to account for the associated uncertainty in subsequent analyses (Scarrot and MacDonall 2012). Several methods have been proposed to estimate the entire density function by splicing together a mixture model for the bulk with a GPD tail attachment (Tancred, Anderson, and O’Hagan 2006; MacDonald et al. 2011; do Nascimento, Gamerman, and Lopes 2012). Although, in theory, these methods partially account for threshold uncertainty, they employ heuristic estimation methods supported by little mathematical analysis.

Toward a more formal statistical methodology we consider the semiparametric model

$$f(y) = p_{\psi,\varphi}(y) := g_\psi(y)\psi(G_\varphi(y)), \quad y > 0, \quad (2)$$

where $g_\psi(y) = \sigma^{-1}(1 + y/(\alpha\sigma))^{-(\alpha+1)}$, $G_\varphi(y) = \int_0^y g_\psi(z)\,dz$, $y > 0$, are the density and distribution functions of a GPD with location 0, scale $\sigma$ and shape $1/\alpha$; here $\theta = (\alpha, \sigma) \in (0, \infty)^2$ is an unknown vector, and, $\psi$ is an unknown density function
on \((0, 1)\). Under this model, \(Y \sim f(y)\) if and only if \(U \sim G_{0}(Y) \sim \psi\), and, \(\alpha_{+}(f) = \alpha_{+}(G_{0}) = \alpha\) under a regularity condition on \(\psi(u)\) as \(u \to 1\) (Lemma 1). Markovitch (2007) offers a thorough analysis of an estimation approach where one first obtains an estimate \(\hat{\theta}\) of \(\theta\) by thresholding data \(Y_{1}, \ldots, Y_{n}\) at a high quantile, and then a nonparametric estimate \(\hat{\psi}\) of \(\psi\) is obtained based on the transformed data \(\tilde{U}_{i} = G_{0}(Y_{i})\), \(i = 1, \ldots, n\). With \(\hat{\psi}\) estimated by a variable kernel mixture, the back-transformed density \(\tilde{f} = \tilde{p}_{\hat{\psi}}\) offers optimal estimation of \(f\) under the mean integrated square error loss. Such a two-stage approach does not account for threshold choice uncertainty in the estimation of \(f\) or any subsequent analyses. It also fails to take advantage of the estimate of the bulk to improve tail estimation.

We consider a likelihood-based alternative approach where \(\theta\) and \(\psi\) are jointly estimated under a Bayesian extension of (2). A Bayesian formulation immediately facilitates information sharing between the bulk and the tail and offers a joint assessment of uncertainty of the extreme and non-extreme features. But important new questions arise on both Bayesian and frequentist sides. What is a principled way to choose a prior distribution on the nonparametric density \(\psi\)? What are the statistical properties of the resulting estimates? These questions could be partially addressed by examining asymptotic concentration properties of the posterior distribution resulting from a specific prior allocation. We show that with a logistic Gaussian process (LGP) prior on \(\psi\) (Leonard 1978; Lenk 1988, 1991; Tokdar 2007), the posterior distribution on \(f\) given a random sample \(Y_{1}, \ldots, Y_{n}\) from an \(f^*\) concentrates around \(f^*\) whenever the latter is continuous and regularly varying. Moreover, the posterior distributions on \(f\) and \(\alpha_{+}(f)\) simultaneously concentrate around \(f^*\) and \(\alpha_{+}(f^*)\) at polynomially fast contraction rates that are nearly minimax optimal, whenever \(f^* = p_{\psi^*, \psi^*}\) with a sufficiently smooth \(\psi^*\). It is significant that the LGP prior enables the likelihood function to preserve relevant information on tail quantities; no other example has been worked out before (Li, Lin, and Dunson 2019). Moreover, guaranteeing posterior contraction across a large model subspace is tantamount to adopting the principle of intersubjective prior allocation to facilitate asymptotic merger of beliefs (Diaconis and Freedman 1986).

Computational details are provided for an efficient and streamlined implementation making it feasible to analyze datasets consisting of several thousand records. Finite sample properties are examined with an extensive simulation study which corroborates the asymptotic analysis result of accurate tail index estimation under strong GPD tail match, and complements it by revealing that even under deviations from a GPD tail, estimates of high tail quantiles are much superior compared to those obtained from thresholding methods. An analysis of daily precipitation records is presented to highlight potential benefits of the joint semiparametric estimation in mitigating ambiguity regarding threshold choice and providing tight but robust estimates of high tail quantiles.

2. Estimation Model

2.1. Tail Index Expression

We restrict to the case where the support of \(f\) is \([a, \infty)\) for a known finite number \(a\), which is set to be zero without any loss of generality. The primary goal of the analysis is taken to be estimating the entire density \(f\) accurately in \(L^1\) or comparable metrics, while also accurately estimating its heavy right tail. Toward this, we first show that the GPD-transformation model (2) is expressible of an entire range of polynomial tail decay rates under a regularity assumption on \(\psi\). Let \(P\) denote the class of densities \(\psi\) on \((0, 1)\) satisfying \(\tilde{\Psi}(1 - u) = uL((1/u)^{1/\alpha})\) for some slowly varying function \(L\); here \(\Psi\) denotes the distribution function of \(\psi\) and \(\tilde{\Psi} = 1 - \Psi\). Note that if \(L(y)\) is slowly varying then

\[
\lim_{y \to \infty} \frac{L(a(y)y)}{L(y)} = 1
\]

for any function \(a(y)\) with a limit \(a_{\infty} := \lim_{y \to \infty} a(y) \in (0, \infty)\).

Lemma 1. If \(\theta = (\alpha, \sigma) \in (0, \infty)^2\) and \(\psi \in P\) then \(f = p_{\theta, \psi}\) is regularly varying with tail index \(\alpha\). Conversely, if \(f\) is a regularly varying density on \((0, \infty)\) with tail index \(\alpha > 0\) then for every \(\sigma > 0\), \(f = p_{(\alpha, \sigma)}, \psi\) for some \(\psi \in P\).

Proof. If \(f = p_{\theta, \psi}\) then \(\tilde{F}(y) = \tilde{\Psi}(1 - \tilde{G}_{0}(y)) = \tilde{G}_{0}(y)\tilde{L}(1/\tilde{G}_{0}(y))\), with \(\tilde{G}_{0}(y) = 1 - \tilde{G}_{0}(y) = y^{-\alpha}\tilde{L}(y), L_{0}(y) = (1/y + 1/(\alpha \sigma))^{-\alpha}\to c_{\alpha} = (\alpha \sigma)^{\alpha}\) as \(y \to \infty\). Therefore, \(\tilde{F}(y) = y^{-\alpha}\tilde{L}(y) = L(y)\tilde{L}(y^{\alpha}/L_{0}(y))\) is slowly varying by (3). Conversely, if \(f\) is a regularly varying density on \((0, \infty)\) with tail index \(\alpha > 0\) and \(\theta = (\alpha, \sigma)\) for some \(\sigma > 0\), then \(f = p_{\theta, \psi}\) where

\[
\psi(u) = \frac{f(G_{0}^{-1}(u))}{G_{0}(y)}, \quad u \in (0, 1).
\]

It is trivial to check that \(\psi\) is a density on \((0, 1)\) with \(\tilde{\Psi}(1 - u) = \tilde{F}(G_{0}^{-1}(u) = \tilde{F}(\alpha \sigma(u^{-\frac{1}{\alpha}} - 1)) = \tilde{u}L(\frac{1}{\tilde{u}})\) where \(\tilde{L}(y) = \alpha \sigma(1 - 1/y^{\alpha - 1})^{-\alpha}\tilde{L}(\alpha \sigma(y^{1/\alpha} - 1))\), with \(L\) denoting the slowly varying component of \(\tilde{F}\). By (3), \(\tilde{L}\) itself is a slowly varying function.

Lemma 1 says, with \(\psi \in P\) the semiparametric model (2) is fully expressible of all regularly varying densities on \((0, \infty)\) with tail index uniquely identified by the model parameter \(\alpha\). It also says that the pair \((\psi, \sigma)\) is not uniquely identifiable. Although one could fix \(\sigma\) and have both \(\alpha\) and \(\psi\) uniquely identified under (2), no obvious choice presents itself. Instead, we find it more useful to retain the scale expressiveness of the model to adjust for implicit shape preferences of any nonparametric prior on \(\psi\). The LGP prior introduced below concentrates around \(\psi\) functions such that the derivatives of \(\log \psi\) are small in magnitude; a bias toward smooth functions being critical to statistical regularization. A flexible choice of the pairing \(\sigma\) creates an important counterbalance. It offers an entire arc of equivalent \((\sigma, \psi)\) pairs for a given \(f\), increasing the possibility that at least some of these pairs will be favorable to LGP shape bias and hence will enjoy high posterior concentration. For example, when \(f = g_{2.1}(\sigma = 1, \psi \equiv 1)\) presents a favorable representation. But if one now adds a little contamination a different pair with a \(\sigma \neq 1\) could be more suitable, even if the contamination does not alter the tail behavior. Figure 1 shows a concrete example with a gamma contamination.

Toward a Bayesian analysis, we choose a product prior distribution \(\pi_{\theta} = \pi_{\alpha} \times \pi_{\sigma}\) on \(\theta = (\alpha, \sigma)\), where \(\pi_{\alpha}\) is the
Figure 1. Lack of identifiability of \((\sigma, \psi)\) and the importance of scale expressiveness. Left: graph of \(f(y) = 0.8 \times g(2,y) + 0.2 \times \tilde{g}(y)\) where \(\tilde{g}(y)\) is the density of a gamma distribution with mean 4 and variance 1; overlaid on the histogram of a sample of size \(n = 1000\) drawn from the same. The tail of \(f(y)\) is completely dominated by that of \(g(2,1)\). Remaining panels show graphs of \(\psi(u)\) in (4) overlaid on the histogram of transformed data with \(\sigma = 2\) and two choices of the scale: \(\sigma = 1\) (middle) and \(\sigma = 4\) (right). The larger scale value produces a flatter \(\psi\) which is more favorable to the LGP prior. For the data displayed here, the posterior concentrates around \((\alpha, \sigma) = (2.1, 4)\) with [2.2, 7.2] giving a 95% interval for \(\sigma\).

half-Cauchy distribution on \((0, \infty)\) and \(\pi_{\alpha}\) is the distribution of \(\alpha = \tilde{c} + (2 - \tilde{c}) e^{\xi/1.5}\) with \(\xi\) distributed according to the standard logistic distribution on the real line. The choice of \(\pi_{\alpha}\) restricts \(\alpha > \tilde{c}\) with probability one, where \(\tilde{c} > 0\) is treated as a hyperparameter to be fixed by the modeler. The numerical analyses presented in Section 4 were carried out with \(\alpha = 0.5\), for which the extreme value index \(\xi = 0.5\) has unimodal density on \((0, 2)\) with a gentle peak at \(\alpha = 0.5\) (i.e., \(\alpha = 2\)). We also experimented with \(\alpha = 0.1\) and all posterior estimates were found to be essentially the same as with \(\alpha = 0.5\).

### 2.2. LGP Prior for \(\psi\)

Let \(C[0, 1]\) denote the space of real, continuous functions on \([0, 1]\). For any \(\omega \in C[0, 1]\), its logistic transform \(L(\omega)\), defined as \((L(\omega)u) = \frac{\psi_{\omega}(u)}{\int_0^1 \psi_{\omega}(t) dt}\), \(u \in (0, 1)\), is a well defined probability density function on \((0, 1)\). When \(\omega\) is a Gaussian process with \(E[\omega(u)] = \mu(u)\) and \(\text{Cov}(\omega(u), \omega(v)) = c(u, v)\) such that \(\omega \in C[0, 1]\) with probability one, the probability law of the random density function \(L(\omega)\) is called a Gaussian process distribution, denoted LGP \((\mu, c)\).

We adopt a hierarchical LGP prior for \(\psi\) in (2). Let \(c_2(u, v) = \exp[-\lambda^2 (u-v)^2]\) denote the unit variance Gaussian covariance kernel with inverse length-scale parameter \(\lambda > 0\). It is well-known that if \(\omega\) is a mean zero Gaussian process with covariance \(\kappa^2 c_1\) for some \(\kappa > 0\), then \(\omega \in C[0, 1]\) with probability one, and hence, the probability distribution function \(F_\omega(\kappa^2 c_2)\) is well defined for every \(\kappa > 0, \lambda > 0\). The prior on \(\psi\) is implicitly defined by the hierarchy

\[
\psi \sim \text{LGP}(0, \kappa^2 c_3), \quad (\kappa^2, \lambda) \sim \pi_\kappa \times \pi_\lambda,
\]

with the distributions \(\pi_\kappa\) and \(\pi_\lambda\) on \((0, \infty)\) described below.

It is clear that if \(\omega \in C[0, 1]\) and \(\psi = L(\omega)\), then \(\psi(1) := \lim_{u \to 1} \psi(u)\) exists and \(\psi(1) \in (0, \infty)\). By mean value theorem \(\tilde{\psi}(1) = \psi(t(u))\) for some \(t(u) \in [1 - u, 1]\). Consequently, \(L_\psi(1) = \psi(t(1/\psi))\), is slowly varying because \(\lim_{u \to 1} \tilde{\psi}(1-u) = \psi(1) \in (0, \infty)\). Therefore, under the hierarchical LGP prior adopted here, \(\text{Pr}(\psi \in \mathcal{P}) \equiv 1\). Of course, the prior support of \(\psi\) is actually smaller than \(\mathcal{P}\), because \(\lim_{u \to 0} u^{-1} \tilde{\psi}(1-u) \in (0, \infty)\) almost surely under the prior, whereas \(\mathcal{P}\) contains densities \(\psi\) where this limit may be zero, infinity or undefined. This may suggest that the induced prior distribution on \(p_{\theta, \psi}\) may not have full support within the class of regularly varying densities. The theorem below reassures that no loss is incurred in a probabilistic sense. Below we assume \(0 < \bar{\alpha} < \alpha < \tilde{\alpha} < \infty\) are such that \(\alpha \in (\bar{\alpha}, \tilde{\alpha})\) with probability one under the prior \(\pi_\theta\).

#### Theorem 2

Let \(f^*\) be any bounded, continuous, regularly varying density on \((0, \infty)\) with tail index \(\alpha^* \in (\bar{\alpha}, \tilde{\alpha})\). If \((\theta, \psi) \sim \pi_{\theta} \times \text{LGP}(0, \kappa^2 c_2)\) for some \(\kappa > 0, \lambda > 0\), then for every \(\epsilon > 0\), \(\text{Pr}(\text{dKL}(Y, \psi) < \epsilon) > 0\), where \(\text{dKL}(f, g) = \int f(y) \log f(y) / g(y) dy\) denotes the Kullback-Leibler divergence of \(f\) from \(g\).

**Proof.** Let \(\epsilon > 0\) be given. Fix a \(0 < \delta < 1 - e^{-\epsilon/2}\). Consider any \(\theta_0 = (\alpha_0, \sigma_0)\) where \(\alpha < \alpha_0 < \alpha^* \) and \(\sigma_0 > 0\). Let \(\psi_0\) be defined as in (4) so that \(f^* = p_{\theta_0, \psi_0}\). Since \(\alpha_0 < \alpha^*\), \(\psi_0\) has heavier tails than \(f^*\) and hence \(\psi_0\) is bounded and continuous with \(\psi_0(1) = 0\). Consequently, the density \(\psi_1(u) = (1 - \delta) \psi_0(u) + \delta\) is bounded and continuous, and is bounded above \(\delta\), and therefore, \(\omega_1 = \log \psi_1\) can be extended to an element of \(C[0, 1]\). Now, for any \(\psi = L(\omega)\) with \(\omega \in C[0, 1]\), \(\text{dKL}(f^*, p_{\theta_0, \psi}) = \text{dKL}(\theta_0, \psi) \leq \text{dKL}(\theta_0, \psi_1) + \int_0^1 \psi_0(u) < \epsilon/21(\omega - \omega_1)\). Therefore, \(\text{Pr}(\text{dKL}(f^*, p_{\theta_0, \psi}) < \epsilon) = \text{Pr}(\|\omega - \omega_1\|_\infty < \delta)\), where the latter probability, calculated for a Gaussian process \(\omega\) with mean zero and covariance \(\kappa^2 c_2\), must be positive because such a Gaussian process has the entire \(C[0, 1]\) in its uniform topology support (Todkar and Ghosh 2007; van der Vaart and van Zanten 2009). An application of the law of total probability completes the proof.

Although not apparent from the above result, the covariance parameters play an important role in determining how the prior mass is distributed within the broad support. The inverse length-scale parameter \(\lambda\) is of critical importance here because of its direct influence on the range of smoothing; though in our experience, a prior on \(\kappa\) also helps with model fit and posterior computation via Markov chain Monte Carlo. We take \(\pi_\kappa\) to be a convenient inverse-gamma distribution with shape \(a_\kappa\) and rate
b_k, that is, \( (1/k^2) \sim \text{Ga}(a_k, b_k) \) which is partially conjugate to the likelihood function in \( k^2 \) and allows this parameter to be integrated out during model fitting. No such conjugate choice exists for \( \lambda \) and formal subjective or objective principles are difficult to apply in selecting \( \pi_{\lambda} \); however, see Paulo (2005) and Gu, Wang, and Berger (2018) for relevant discussions.

An alternative is to seek \( \pi_{\lambda} \) that guarantees optimal asymptotic frequentist convergence of the posterior distribution to the truth. In the setting of purely nonparametric density estimation with LGP, van der Vaart and van Zanten (2009) show that a gamma prior distribution on \( \lambda \) is critical to optimally spreading prior mass into various smoothness classes, which in turn is critical to guaranteeing adaptive and optimal concentration of the posterior distribution to the truth. We follow this recommendation to specify \( \pi_{\lambda} \sim \text{Ga}(a_k, b_k) \). Our numerical experiments were carried out with \( a_k = b_k = 3/2 \), and \( a_k = 16 \) and \( b_k = 2.2 \). The latter choices could be appreciated in several ways. Consider \( \rho = c_2(0, \Delta) = e^{-\lambda_2 \Delta^2} \) which gives the correlation of the Gaussian process at a distance \( \Delta \). With \( \Delta = 10\% \), our choice of \( \pi_{\lambda} \) assigns 95\% prior probability to \( \rho \in (0.28, 0.84) \) with prior mean and median \( \approx 0.6 \). Alternatively, one could look at the number of up-crossings at zero of the process sample paths; which could be taken as a proxy to the number of local modes. The well-known Rice formula states that the expected number \( \psi \) of up-crossings of zero of a mean zero Gaussian process on the unit interval with covariance \( \kappa \psi \) is \( \psi = \psi/\sqrt{2} \approx 0.22a_\psi \). (Rice 1944; Adler and Taylor 2009). With our choice of \( \pi_{\lambda} \), the prior probabilities of zero through five up-crossings, respectively, are 8\%, 37\%, 40\%, 13\%, and 2\%.

### 2.3. Posterior Computation

For data \( \{y_1, \ldots, y_n\} \), the likelihood function \( (\theta, \psi) \mapsto \prod_{i=1}^{n} g_\theta(y_i) \psi(u_i) \), where \( u_i = G_\theta(y_i) \), involves \( \psi \) only through the finite vector \( \psi = (\psi(u_1), \ldots, \psi(u_n))^\top \). Unfortunately, the joint prior density of \( \psi \), given model hyper-parameters \( (\lambda, \kappa) \), is not available in closed form. This necessitates involving the latent Gaussian process \( \omega \) in the representation \( \psi = L_\omega(\omega) \) in posterior computation. However, \( \psi \) depends on both the corresponding vector \( \omega_U \) and the scalar \( \omega_{\text{norm}} = \int_0^1 e^{w_{\text{norm}}(u)} \, du \) giving the normalization in the logistic transform and involving the whole function \( \omega(u) \). It is practically impossible to carry out any numerical analysis of the posterior when a function valued input variable is involved in the likelihood evaluation. We overcome this challenge by adopting a grid-based representation of \( \omega \) proposed and analyzed in Tokdar (2007).

#### 2.3.1. Likelihood Approximation

Specifically, a dense set of points \( T = \{0 = t_1 < t_2 < \cdots < t_L = 1\} \subset [0,1] \) is chosen as a grid over which both \( \omega \) and \( \psi \) are to be represented, respectively, as the vector \( \omega_T = (\omega(t_1), \ldots, \omega(t_L))^\top \) and the corresponding vector \( \psi_T \). Given \( \omega_T \), a very accurate approximation to \( \omega_{\text{norm}} \) could be obtained by applying the trapezoidal rule of numerical integration to the pair \( (T, \omega_T) \), readily producing the vector \( \psi_T \).

To evaluate \( \psi_T \), which is needed for likelihood evaluation, it is useful to formally express the trapezoidal approximation to \( \omega_{\text{norm}} \) as the exact integration of the function \( h(u) \) that linearly interpolates the points \( (t_i, \text{erf}(t_i)), i = 1, \ldots, L \). We may now view \( \psi_T \) as the evaluation over the grid \( T \) of the (normalized) density function \( \hat{h}(u) = h(u)/\int_0^1 h(t) \, dt \). Consequently, \( \psi_U \) could be readily evaluated with the corresponding vector \( \hat{h}_U \).

The overall computational complexity of this likelihood approximation is \( O(\max(n, L)) \) and can be carried out extremely fast in actual time with optimized codes. In the numerical experiments reported here we use an equally spaced grid with \( L = 101 \) and increment size 0.01.

#### 2.3.2. Low Rank Approximation and Marginalization of Hyper-parameters

With the availability of a grid based representation and the linear interpolation based approximation to the likelihood function, it is feasible to carry out a Markov chain Monte Carlo approximation the posterior distribution of \( (\theta, \omega, \psi) \). The prior density of \( \omega_T \), given \( (\lambda, \kappa) \), is a multivariate normal density with mean zero and covariance \( \kappa^2 C_T(\lambda) \) where \( C_T(\lambda) = ((c_i(t_i, t_j)))^{L}_{i,j=1} \).

An evaluation of this density involves factorizing \( C_T(\lambda) \) at \( O(L^3) \) computational complexity, which is practicable but slow at \( L = 101 \), and could be outright prohibitive for larger grid sizes. Additionally, running a Markov chain sampler on \( \omega_T \), which is a dense representation of a smooth function, produces slow-mixing chains.

Considerable efficiency gains can be made by replacing the smooth Gaussian process \( \omega \) with a low-rank Gaussian process (Snelson and Ghahramani 2006; Tokdar 2007; Banerjee et al. 2008). For a set of knots \( S = \{s_1, \ldots, s_m\} \subset [0,1] \), with \( m \) much smaller than \( L \), the so called predictive process \( \tilde{\omega}(u) = \frac{\sum_{j=1}^{m} \pi_{\lambda}(s_j) \omega(s_j)}{\sum_{j=1}^{m} \pi_{\lambda}(s_j)} \) gives a smooth interpolation of the graph of \( (S, \omega_S) \), and is fully determined by the random vector \( \omega_S = (\omega(s_1), \ldots, \omega(s_m))^\top \). Typically the predictive process conditioning is defined for given covariance parameters \( (\lambda, \kappa) \), but a hyper-parameter marginalized extension proposed in Yang and Tokdar (2017) and described below offers considerable additional speed up.

Integrate out \( \kappa^2 \) from the model and express the prior distribution of \( \omega_S \) given \( \lambda \) as the multivariate Student-t distribution with pdf \( p(\omega|\lambda) \propto \left[ 1 + \omega_S^2 C_S(\lambda)^{-1} \omega_S/(2\tilde{b}_k) \right]^{-(a_\lambda + m/2)} \) where \( C_S(\lambda) = \text{cov}(\omega_S) \) is the analogue of \( C_T(\lambda) \) over the knot set \( S \). It is impossible to analytically integrate out \( \lambda \), but a discrete integration could be carried out by replacing \( \pi_{\lambda} \) with a discrete approximation over a dense set of support points \( \{\lambda_1, \ldots, \lambda_G\} \subset (0, \infty) \).

With \( \pi^G_{\lambda}(g) := \Pr(\lambda = \lambda_g) \), the prior distribution of \( \omega_S \) is the mixture density \( p(\omega_S) = \sum_{g=1}^{G} \pi^G_{\lambda}(g)p(\omega_S|\lambda_g) \). The vector \( \tilde{\omega}_T \), which is the predictive process replacement of \( \omega_T \), can be computed analytically from \( \omega_S \) as \( \tilde{\omega}_T = \sum_{g=1}^{G} \pi^G_{\lambda}(g)A_g \omega_S \) where \( A_g = C_{TS}(\lambda_g)C_S(\lambda_g)^{-1} \) with \( C_{TS}(\lambda) \) denoting the \( L \times m \) matrix with elements \( c_{S}(t, s), t \in T, s \in S \), and, \( \pi^G_{\lambda}(g)p(\omega_S|\lambda_g) \). We select the support points \( 0 < \lambda_1 < \cdots < \lambda_G \) of \( \pi^G_{\lambda}(g) \) based on the knots set \( S \). First \( \lambda_1 \) is fixed such that \( \rho_1 := c_{S}(0,0.1) = 0.95 \) and then successive \( \lambda_g \) values are chosen so that \( d_{KL}(N(0, C_{S}(\lambda_g-1)), N(0, C_{S}(\lambda_g))) = 0.5 \) until we get \( \rho_{G+1} := c_{S}(0,0.1) < 0.2 \). This gradual stepping down ensures successful \( N(0, \kappa^2 C_S(\lambda_g)) \) distributions maintain considerable overlap, eliminating any major gaps in the prior distribution of \( \omega_S \) due to the discretization of \( \lambda \). In our expe-
rience, posterior calculation is not sensitive to exact choices of the bookkeeping values of $\rho_1$ and $\rho_{G+1}$, or the Kullback-Leibler stepping size.

### 2.3.3. Markov Chain Sampling and Runtimes

With the above approximations in place, the model parameters reduce to the $(m+2)$ dimensional vector $(\alpha, \sigma, \omega)$. An adaptive, blocked 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 presented in this article were achieved by using one block containing $\omega_S$, one block updating $\theta = (\alpha, \sigma)$, and one block updating all $(m+2)$ parameters simultaneously. An important consequence of the discretization of $\lambda$ is that all relevant matrices, namely $\{ (A_g, R_g) : g = 1, \ldots, G \}$, where $R_g$ is a Cholesky factor of $C_0(\lambda_g)$, could be precomputed and stored prior to Markov chain sampling. Subsequent evaluations of the log posterior density reduce to $O(m \cdot \max(n, L))$ computing complexity.

Experience suggests that the actual runtime of the sampler scales linearly in the sample size and sub-linearly in the number of knots $m$ or the grid size $L$. All numerical results reported in Sections 4 and 5 use $L = 101, m = 11$, with equally spaced points in $T$ and $S$ with endpoints equaling 0 and 1. This choice of $L$ leads to a discretization of $\pi_1$ with $G = 30$ support points. In analyzing Fort Collins precipitation data with sample size $n = 6180$ (Section 5), it took 9.8 min on a personal computer to carry out 500,000 iterations of the Markov chain. For two further subsamples with $n = 3645$ (0.6x with respect to the original set) and $n = 1061$ (0.2x), the same number of iterations took 6.3 (0.6x) and 2.2 (0.2x) min, respectively. For the original set with $n = 6180$, it took 12.2 min (1.3x) to run the same number of iterations when the knots set $S$ was doubled to $m = 21$ equally spaced knots ($G = 82$), keeping $L$ fixed at 101. Similarly, when the grid $T$ was doubled to $L = 201$ grid points (retaining $m = 11, G = 30$), it took 12.5 min (1.3x) to run the same number of iterations. We recommend $L = 101$ and $m = 11$ as default choices. But for any application, one should assess whether finer approximations are needed by repeating the analysis with larger values of $L$ and $m$ until posterior calculations stabilize.

### 3. Asymptotic Properties

In recent years, mathematical analyses of large sample concentration properties of the posterior distribution have proven useful to the question of prior allocation in Bayesian analysis of infinite dimensional models; see Ghosal and van der Vaart (2017) for a comprehensive overview. Here we focus on posterior consistency and posterior contraction rate properties of the semiparametric LGP prior. Our treatment involves distinct model space topologies suitable for assessing either density estimation accuracy or tail index estimation accuracy; keeping in mind that proximity of two densities in $L^1$ topology may not guarantee similar tail index values. Posterior consistency, a frequentist evaluation of a prior intended for Bayesian applications, guarantees intersubjective knowledge generation through asymptotic merger of beliefs (Diaconis and Freedman 1986).

#### 3.1. Density Estimation Consistency

Let $\Pi$ denote the induced prior measure on $f = \rho_0, \psi_0$, with $(\Theta, \psi_0) \sim \pi_\theta \times \mathbb{LG}(0, k^2 \zeta)$ where we treat $k = 1$ as fixed, and work with a gamma prior on $\lambda$. We allow $\pi_\theta$ to be arbitrary but assume it has a compact support $\Theta = [\alpha, \beta] \times [\sigma, \overline{\sigma}]$ for some $0 < \alpha < \beta < \infty, 0 < \sigma < \overline{\sigma} < \infty$, with a strictly positive density in the interior of $\Theta$. Compactness of $\Theta$ is assumed chiefly for technical reasons. An unbounded support adds layers of complication to critical function approximation results used below (e.g., Lemma 8 in Appendix A) with little gain in insight. One can enlarge $\Theta$ arbitrarily without virtually altering the posterior contraction rates.

We may view $\Pi$ as a probability measure on $\mathcal{F}$, the subspace of density functions in $L^1(0, \infty)$. Given data $(y_1, \ldots, y_n)$, the posterior measure equals $\Pi(d\theta \mid y_1, \ldots, y_n) = \{ \Pi_n(d\theta) \} \Pi(d\theta)$ below $\mathcal{H}_\phi[0, 1]$ denotes the Hölder-$\beta$ space consisting of functions on $[0, 1]$ that are \( \beta \) times continuously differentiable with the $[\beta]$th derivative being Hölder continuous of exponent $\beta - \| \beta \|$ where $\| \beta \|$ is the largest integer smaller than $\beta$. The minimax density estimation rate over Hölder-$\beta$ classes is $n^{-\frac{1}{2\beta}}$ (Stone 1982).

#### Theorem 3. If $Y_1, \ldots, Y_n \sim f^*$ where $f^*$ is a bounded, continuous, regularly varying density on $[0, \infty)$ with tail index $\alpha^* \in (0, 1)$, then $\lim_{n \to \infty} \Pi(f : \| f - f^* \| < \varepsilon) > 0$ for every $\varepsilon > 0$. Additionally, if $f = \rho_0, \psi_0$ with $\theta^0$ in the interior of $\Theta$ and $\phi^* = \log \psi^* \in \mathcal{H}_\phi[0, 1]$ for some $\beta > 2$, then the fixed error margin $\varepsilon$ may be replaced with the vanishing sequence $\varepsilon_n = Bn^{-\frac{1}{2\beta}}(\log n)^{\frac{1}{4(\beta+1)}}$ for some large constant $B$.

#### Proof.

To prove the first claim, we only need to establish (Ghosal, Ghosh, and Ramamoorthi 1999, Theorem 2)

1. $\Pi(f : d_{KL}(f^*, f) < \varepsilon) > 0$ for every $\varepsilon > 0$,

2. For every $\varepsilon > 0$, there exist constants $c, C > 0$ and sets $\mathcal{F}_1, \mathcal{F}_2, \ldots \subset \mathcal{F}$, such that $\Pi(\mathcal{F}_n) \leq ce^{-C\varepsilon}$ and $\log N(\varepsilon, \mathcal{F}_n, d_H) \leq n^c$ for all large $n$, where $N(\delta, \mathcal{F}, d_H)$ denotes the covering number of $\mathcal{F}$ by balls of radius $\delta$ in the Hellinger metric $d_H(p, q) = [\int (\sqrt{\rho(y)} - \sqrt{\psi(y)})^2 dy]^{1/2}$. C1 follows readily from Theorem 2. C2 follows from the following stronger condition necessary for the second part of the theorem.

#### C2*. For every $0 < t < 1/2, s > 0$, there exist a constant $C > 0$ and sets $\mathcal{F}_1, \mathcal{F}_2, \ldots \subset \mathcal{F}$, such that $\Pi(\mathcal{F}_n) \leq e^{-(C + 4t^2)n^2}$ and $\log N(\varepsilon_n, \mathcal{F}_n, d_R) \leq n^c_\varepsilon$ for all large $n$, where $\varepsilon_n = Bn^{-1}(\log n)^t$ for some $B > 0$ and $\varepsilon_n = \varepsilon_n^* \log n$.

A proof is given in Appendix A. A key step is Lemma 8 (Appendix A) which states $\log(p_0, \psi_0) - \log(p_0, \psi_0) \leq \epsilon \log \| \psi_0 \|_{C^2}^{1/2} \epsilon \| \psi_0 - \phi_0 \|_{C^2}^{1/2}$ if $\theta_1, \theta_2$ are interior points in $\Theta$ and $\psi_1 = \mathcal{L}(\phi_1), \psi_2 = \mathcal{L}(\psi_2)$ with $\phi_1, \phi_2 \in C^2[0, 1]$, the space of twice continuously differentiable functions on $[0, 1]$ with norm $\| \phi \|_{C^2} := \| \phi \|_{C^0} + \| \phi \|_{C^1} + \| \phi \|_{C^2}$. Clearly, $\mathcal{H}_\phi[0, 1] \subset C^2[0, 1]$ for all $\beta > 2$. Construction of the sets $\mathcal{F}_1, \mathcal{F}_2, \ldots$ relies on the observation that a separable, mean-zero Gaussian process $\omega$ on $[0, 1]$ with covariance function $\psi_0$ may be viewed as a Borel measurable random element with a
Gaussian measure \(\nu^h\) on the Banach space \((C^2[0,1], \|\cdot\|_{C^2})\). Our construction builds upon that of van der Vaart and van Zanten (2009) who embed the Gaussian measure in \((C[0,1], \|\cdot\|_{\infty})\). However, some key modifications are needed to address the change in the embedding space (Appendix C).

By Theorem 8.9 of Ghosal and van der Vaart (2017), under the additional assumption on \(f^*\), a proof of the second part of the theorem may be established by applying \(C^2\) with \(r = \frac{3}{2\alpha+1}\), \(s = 2t\), in conjunction with the following sharper version of C1:

\[
C1^*: \quad \Pi(\{f : dk(f^*, f) \leq \varepsilon_n^2, V(f^*, f) \leq \varepsilon_n^{-1}\}) > e^{-Cn^2\varepsilon_n^2} \quad \text{for all large } n,
\]

where \(V(f, g) = \int f(y) \log^2(f(y)/g(y))dy\). This sharper prior concentration bound can be proved via a nontrivial extension of Theorem 3.1 of van der Vaart and van Zanten (2009). A proof of possible independent interest is given in Appendix B.

### 3.2. Tail Index Estimation Consistency

In the following, assume without loss of generality that \(\alpha \leq \frac{1}{2}\) and \(\alpha > 1\). As in Theorem 3, assume that the true density is some \(f = p_{\theta^*, \sigma^*}\), where \(\theta^* = (\alpha^*, \sigma^*)\) is in the interior of \(\Theta\) and \(\sigma^* = \log \psi^* \in H^0[0,1]\) with \(\beta > 2\). Denote \(\gamma = \frac{\beta}{2\beta + 1} = (0,1/2)\) so that the posterior contraction rate in \(L^1\) topology equals a constant multiple of \(n^{-\gamma} (\log n)^{2r+1}\). The lower bound assumption on \(\beta\) implies that both \(\tilde{\rho}(\xi) := \frac{2\beta}{2\beta + 1} \gamma - 3(1-2\gamma)\) and \(\tilde{\rho}(\xi) := 2\gamma - \frac{3}{2}(1-2\gamma)\) are strictly positive for every \(\xi \in \left[\frac{1}{\alpha}, 1\right]\).

**Theorem 4.** If \(Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{\alpha^*}^*\) and \(\alpha/\alpha^* < \xi < \min(1, 1/\alpha^*)\) is such that \(\beta \xi > 3/2\) then \(\lim_{n \to \infty} \Pi(\{f : \alpha^* f^* \neq \sigma^* \}) > B_1 n^{-\beta} (\log n)^{\beta}\) if \(\alpha^* \neq 0\) for all large \(B_1\) where \(\rho = \min(\tilde{\rho}(\xi), \tilde{\rho}(\xi))\) and \(s = 2\rho + \frac{3}{2\beta + 1} \gamma\) if \(\rho = \tilde{\rho}(\xi), s = 2\rho + 4\) otherwise.

A proof is presented in Appendix E. The main argument relies on establishing existence of tests that can distinguish \(f^*\) from model elements \(f = p_{(\alpha, \sigma), \psi^*}\) with \(|\alpha^* - \alpha| > B_1 n^{-\beta} (\log n)^{\beta}\) with Type I and II error probabilities vanishing suitably rapidly. This line of argument directly follows the path laid out in the original work of Schwartz (1965); a modern presentation is Theorem 8.9 Ghosal and van der Vaart (2017). See also Kleijn (2021) for related recent developments. Li, Lin, and Dunson (2019) present a similar theoretical exploration with test functions derived from an exceedance probability based tail index estimator of Carpentier and Kim (2015). Our proof relies on a more complex test procedure which first tries to detect a difference between the exceedance probability of the empirical distribution at a high threshold and that of the true distribution, and if no significant difference is detected then repeats the process one more time at an even higher threshold but only to the conditional distributions to the right of the first threshold.

The theorem requires sufficient smoothness of the true density via the assumption \(\beta \cdot \min(1, 1/\alpha^*) > 3/2\) so that a suitable \(\xi\) may be found with \(\beta \xi > 3/2\). This condition demands that a relatively rough density (small \(\beta\)) must have a sufficiently heavy tail (small \(\alpha^*\)) to insure accurate estimation of the latter with our semiparametric estimation model. This requirement may be understood in the light that with a density function lacking in smoothness, the bulk of the density carries less information about the tail, and hence an accurate estimation of the tail is possible only when more observations are available directly from the tail itself, that is, only when the tail is heavy.

Additionally, multiple factors control the value of \(\rho\) which determines the posterior contraction rate. Notice that both \(\tilde{\rho}(\xi)\) and \(\tilde{\rho}(\xi)\) are strictly increasing in \(\xi\) and hence sharpest rates are obtained by taking \(\xi\) as close as possible to the maximum allowed value of \(\min(1, 1/\alpha^*)\). Since \(\tilde{\rho}(\xi) < \tilde{\rho}(\xi)\) if and only if \(\beta \xi(2\xi - 1) < 1\), the following observations can be made on the fastest possible rate. If \(\alpha^* \in \left(\frac{\alpha - 2}{\alpha + 2}, \frac{\alpha - 2}{\alpha + 2}\right) then\)

**Theorem 4** holds with any \(\xi\) arbitrarily close to \(\min(1, 1/\alpha^*)\), \(\rho = \tilde{\rho}(\xi)\) and \(s = 2\rho + \frac{3}{2\beta + 1} \gamma\). On the other hand, if \(\alpha^* \geq 2\), the theorem holds with any \(\xi\) arbitrarily close to \(\frac{1}{\alpha^*}\), \(\rho = \tilde{\rho}(\xi)\) and \(s = 2\rho + 4\). In the intermediate case of \(\alpha^* \in \left(\frac{\alpha - 2}{\alpha + 2}, 2\right)\), \(\xi\) can be arbitrarily close to \(\frac{1}{\alpha^*}\) with \(\rho = \tilde{\rho}(\xi), s = 2\rho + \frac{3}{2\beta + 1} \gamma\) if \(\beta > \frac{3}{2\beta + 1}\) \(\leq \frac{\alpha - 2}{\alpha + 2}\) and \(\rho = \tilde{\rho}(\xi)\) and \(s = 2\rho + 4\) otherwise.

When \(\alpha^* \in \left(\frac{\alpha - 2}{\alpha + 2}, 2\right)\), the choice of \(\rho = \tilde{\rho}(\xi) = \frac{\beta}{2\beta + 1} - \frac{3(1-2\gamma)}{2\beta(2\beta + 1)}\) compares favorably to the optimal rates obtained by Hall and Welsh (1984, 1985). In particular, whenever \(\psi^*\) is infinitely smooth, for example, \(f^*\) is a GPD itself, the density estimation contraction rate has \(\gamma \approx \frac{1}{2}\) and hence \(\rho \approx \frac{k}{2\beta + 1}\) with \(\xi \approx \min(1, 1/\alpha^*)\); hence \(\approx \min(1, 1/\alpha^*)\) indicates “arbitrarily close from below.” Since \(|\alpha^* f^*(\xi) - 1| \approx \gamma^{-\min(1, 1/\alpha)}\), \(f^* \in \{f : \tilde{F}(\xi) = 1 + R(\xi), |R(\xi)| < Ay^{-\min(1, 1/\alpha)}\}) for which the minimax rate of tail index estimation is precisely \(n^{-\gamma}\). See Section 6 for further discussion.

### 4. Finite Sample Behavior

#### 4.1. Tail Index Estimation

From Hall and Welsh (1984), statistical performance of any estimator of tail quantiles depends on how quickly the actual tail starts resembling the corresponding Pareto tail \(y^{-\alpha^*} / 2\). When sample size \(n\) is only moderately large, the Pareto shape may only be partially established within the range of the observed data, positing a serious challenge to any thresholding method in detecting if and where a bulk-to-tail transition takes place. A similar challenge is posed to our joint semiparametric estimation which must balance a likelihood function that receives little information from a partially established tail against a model specification that idealizes a generalized Pareto-like tail.

Consider three different choices of the shape of \(f\), namely,

1. GPD: \(f_{\alpha}(y) = \frac{\alpha}{\gamma}(\gamma(y)), (ii)\) GPD: \(f_{\alpha}(y) = \frac{\alpha}{\gamma}(\gamma(y)) [\Gamma(\alpha, y)]^3\), and
2. Half-t: \(f_{\alpha}(y) = \frac{2\alpha}{\gamma} y^2 - (\alpha + 2)/2\),

\(\alpha \in (0, \infty)\); each giving a regularly varying density with tail index \(\alpha\). Table 1 reports performance statistics of our semiparametric estimation of the corresponding extreme value index \(\xi = \alpha^{-1}\), averaged across 100 datasets.

---

S. T. TOKDAR, S. JIANG, AND E. L. CUNNINGHAM
of size $n = 1000$ each, with the true value of $\xi$ varying over $\{0.1, 0.2, 0.3, 0.5, 1.0\}$. For comparison, we include corresponding figures from a thresholding estimation of $\xi$, where the threshold is determined by the adaptive technique of Durrieu et al. (2015), followed by a Bayesian fit of a GPD model to the excess data with the GPD location parameter set at the threshold, and the scale $\sigma$ and shape $1/\alpha$ estimated under the same prior as used in our semiparametric estimation.

For the GPD sets, in addition to smaller bias and averaged error for the point estimates, the 95% posterior credible intervals from the semiparametric method are much narrower than the parametric method. Although the averaged error of the semiparametric method incurs a strong negative bias for $\xi = \alpha^{-1}$ (i.e., underestimates the tail heaviness) with fairly tight posterior credible intervals resulting in poor coverage when true $\xi < 1$. For a half-t density, we may use (4) to write $f_{\alpha} = p(\alpha, \sigma, \psi)$ and verify that $\phi = \log \psi \in C[0, 1]$ but $\phi(1-t) = (\alpha+1)\rho^{t-1} \times (1+\alpha \rho^2 - \sigma^2) \bar{F}^{t-1}(\rho)$, and hence $\lim_{t \to 1^-} \phi(t)$ equals $-\infty$, $-2$ or 0, according to whether $\xi < 1$, $\xi = 1$ or $\xi > 1$. There is a strong mismatch between the idealized shape and the truth on the right tail when $\xi < 1$, causing the posterior distribution on $\xi$ to be biased downward.

### 4.2. Estimation of Tail Quantiles

Although an accurate estimation of the tail index parameter is conceptually appealing, practical interest usually focuses on estimating tail quantiles of $f$. By extending the numerical analysis presented above, we find that the semiparametric joint estimation is substantially more effective at this task than the thresholding approach. Specifically, we look at the estimates and the 95% posterior credible intervals of the tail quantiles $\hat{Q}(p) = F^{-1}(p)$ associated with excess tail probabilities $p \in \{0.01, 0.001, 0.0001, 0.00001\}$ and compare these against the true values for the $3 \times 5$ experimental sets reported above. Both methods produce credible intervals with coverage at or above the nominal 95% level in most cases, but the semiparametric estimate is typically more accurate than the thresholding estimate, with up to 400% improvement in some cases for very high quantiles (Table 1). The semiparametric posterior credible interval is also much tighter than the threshold based interval (not shown).

The only concern about coverage of the semiparametric credible interval arises in the Half-t sets with a small $\xi$, for which the
A Bayesian expression of joint uncertainty of the bulk and the pure statistical uncertainty resulting from sampling variability. This kind of ambiguity about the tail index is distinct from method produces reasonably accurate and meaningful estimates bias in estimating asymptotic tail heaviness, the semiparametric at threshold \(0.85\) and 
\[
\hat{\xi} = 0.21 \pm 0.04 \quad (\text{standard error}) \quad \hat{\xi} = 0.21 \pm 0.04 \quad (\text{standard error})
\]
but usually produces a very wide interval with the upper end of the interval being several orders of magnitudes larger than
\[
\text{the truth} \quad (\text{Figure 2}).
\]
In other words, in spite of the persistent bias in estimating asymptotic tail heaviness, the semiparametric method produces reasonably accurate and meaningful estimates of the tail itself.

5. Fort Collins Precipitation

Katz, Parlange, and Naveau (2002) present an analysis of total daily precipitation measurements (in inches) between 1900 and 1999 from a single rain gauge in Fort Collins, CO, estimating a heavy-tailed distribution with \(\xi = \alpha^{-1} = 0.18\) at the threshold of 0.4 inches. Scarrot and MacDonnald (2012) estimate \(\hat{\xi} = 0.21 \pm 0.04\) (standard error) at a similar threshold, and identify two additional candidates for the threshold value at which usual GPD diagnostics plots appear to stabilize, each leading to a different estimate of the tail index parameter: \(\hat{\xi} = 0.13 \pm 0.07\) at threshold 0.85 and \(\hat{\xi} = 0.003 \pm 0.09\) at threshold 1.2. This kind of ambiguity about the tail index is distinct from pure statistical uncertainty resulting from sampling variability. A Bayesian expression of joint uncertainty of the bulk and the tail could be particularly useful in mitigating between multiple distinct GPD tails offering partial match.

The original dataset\(^1\) contains \(N = 36,524\) daily measurements with 78% of the records being zero; the rest are recorded to the nearest hundredth of an inch. We remove all records with a precipitation measurement below 0.03 inches and jitter the remaining data \((n = 6180, 17\% \text{ of all records})\) with a small uniform noise between \(-0.005\) and \(0.005\) to break ties while preserving original precision. With a smooth LGP prior at the core, the semiparametric method is sensitive to the presence of strong discontinuous features in the data histogram. A large number of ties in the records is one such feature, which necessitates the random jittering. The presence of excess zeros is another such feature, which cannot be overcome by jittering alone, since the distribution of the jittered data still presents a big jump discontinuity near zero. In fact, we find that such an effect persists up to measurements of 0.02 inches, whose inclusion in the data analysis significantly distorts the posterior inference from what is obtained when analyzing all or some subset of records \(\geq 0.03\) inches. We return to this point below after presenting our results.

Figure 3 (left panel) shows thresholding estimates of \(\xi = \alpha^{-1}\) obtained from a Bayesian fit of a GPD tail to excess data over the threshold, as described in Section 4.1. These estimates of \(\xi\) are different from those reported in Scarrot and MacDonnald (2012), who employ maximum likelihood estimation without restricting \(\xi > 0\) and without any regularization via a prior. However, the detailed analysis of Katz, Parlange, and Naveau (2002) offers strong evidence of a heavy tail (i.e., \(\xi > 0\)), and thus a Bayesian estimation with a relatively flat prior on \(\xi \in [0, 2]\) appears a better alternative. In spite of a weak prior specification, the posterior estimate and interval of \(\xi\) are heavily influenced by the prior choice for large threshold values at which little excess data is left for parameter estimation. The adaptive threshold choice method of Durrieu et al. (2015) gives a threshold value of 0.93, for which \(\xi\) is estimated to be 0.22 with a 95% posterior credible interval [0.08, 0.41].

The semiparametric method offers a comparable estimate of \(\xi = 0.22\) with a tighter 95% credible interval [0.12, 0.30]. Both methods point to a slightly heavier tail than what was reported by Katz, Parlange, and Naveau (2002), but their estimate of \(\xi = 0.18\) lies well within the 95% credible intervals. The estimated high tail quantiles from the semiparametric method and the threshold method (threshold = 0.93) are very similar to one another and they line up well against empirical quantiles, but the 95% credible intervals from the semiparametric method are considerably tighter (Figure 3, right). However, the difference is much less stark than what we see in simulation studies.

The maximum daily precipitation during the observation period was 4.63 inches, recorded in the year 1997. The semiparametric method estimates the corresponding return period to be 47.6 years, with a 95% posterior credible interval (PCI) of [23, 122.3]; the thresholding method gives similar estimates. These estimates are close to the estimated return period of 50.8 years reported by Katz, Parlange, and Naveau (2002), who did not report an interval. The estimated return periods for 3 inches and 4 inches of precipitation are, respectively, 10 years (95% PCI = [6.5, 16.5]) and 25 years (95% PCI = [14.9, 59.9]).

\(^{1}\)Taken from the extRemes package in R (Gilleland and Katz 2011).
Figure 3. Estimation of tail heaviness ($\xi = \alpha^{-1}$) and high tail quantiles for Fort Collins daily precipitation. Left panel shows thresholding estimates and 95% credible intervals of $\xi$ corresponding to a grid of threshold values between 0.005 and 3.0 with an increment of 0.025; the adaptive choice of threshold = 0.93 is highlighted with a darker shade. The horizontal lines give the estimate and the 95% credible interval for the semiparametric analysis. Right panel shows estimates of high tail quantiles $\bar{Q}(p) = \bar{F}^{-1}(p)$ from the semiparametric method and the thresholding method (threshold = 0.93). The exceedance probability $p$ corresponds to the original data of size $N = 36,524$, without any truncation or thresholding. A graph of the points $\{(i-N/2, Y_i), 1 \leq i \leq N\}$ is included to visualize empirical quantiles, where $Y_i$ denotes the $i$th order statistic of the original data.

Figure 4. Semiparametric estimation of tail heaviness ($\xi = \alpha^{-1}$) and high tail quantiles under further left truncation of Fort Collins precipitation data. Posterior distribution of $\xi$ (left) widens and moves to the left slightly but maintaining overlap. High tail quantile estimates (right) remain robust.

'61, '77, '90, '97), of which three had more than 4 inches of rain ('02, '77, '97). More speculatively, we estimate the return period of 5 inches of rain to be 64.2 years (95% PCI = [28.7, 178.8]).

The estimates from the semiparametric method remain reasonably robust when analyzing further subsets of the data. When data analysis is restricted to records $\geq 0.1$ inches (or $\geq 0.4$ inches), the estimate of $\xi$ is 0.19 with 95% PCI = [0.06, 0.30] (or 0.16 with 95% PCI = [0.04, 0.32]). For these further truncations, the estimated tail heaviness is slightly lower with greater uncertainty, but the upper end of the credible interval remains essentially the same. The same is reflected in high tail quantile estimates (Figure 4). It appears that there is no strong evidence in the data pointing to a substantially lower tail heaviness than what was presented in Katz, Parlange, and Naveau (2002). The possible lower estimates at higher threshold values discussed by Scarrot and MacDonnald (2012) are likely spurious.

However, the semiparametric method is not completely robust to the issue of truncation. When data analysis is expanded to include all nonzero records, the posterior shifts substantially and results in a heavier tail estimate (0.3 with 95% PCI = [0.25, 0.36]) with high tail quantiles being significantly larger than the estimates reported above (not shown). The same shift is noticed also when expanding the analysis only slightly to include records of 0.02 inches, or records of 0.01 and 0.02 inches. As indicated earlier, this discrepancy is likely an artifact of excess of zero and other tiny measurements which cannot be fully mitigated by jittering alone. See Section 6 for further discussion.

6. Concluding Remarks

The semiparametric method analyzed here makes a case for likelihood based joint estimation of the bulk and the tail, with potential benefits that such joint estimation may improve estimation accuracy of high tail quantiles and provide a better uncertainty quantification of tail heaviness. Our asymptotic analysis reassures that sparse tail information does not get washed away by the bulk in such likelihood based estimation, however, suitable prior distributions are needed to strike a balance while also retaining full expressiveness of the bulk shape and tail decay rate. The transformation model (2) appears to deliver the right theoretical platform especially when combined with the hierarchical LGP prior on the nonparametric density of the transformed data. A crucial element of the model is the choice of the Gaussian covariance kernel for the LGP prior. With a gamma prior on the inverse length-scale parameter of the kernel, the adaptive estimation accuracy of the LGP prior (van der Vaart and van Zanten 2009) transfers seamlessly to our semiparametric setting.

The semiparametric model (2) adopts a GPD like tail and hence covers only the special Hall-Welsh class $D(\alpha, C_0, \xi, \epsilon, A)$ with $\xi = \min(1, \frac{1}{\alpha})$, albeit Hall and Welsh (1984, 1985) make
stringent assumptions on other quantities such as \(\xi(f)\) (Carpenter and Kim 2015). More ground might be recovered by using a more flexible parametric component, such as the three parameter extended-GPD family of Beirlant, Janssen, and Segers (2009). It could also be feasible to sharpen our posterior contraction rate in Theorem 4 by using test functions that specifically exploit the idealized shape. Theorem 4 intimately connects tail index estimation rate with the smoothness level of \(f\). It will be interesting to examine whether this connection is intrinsic to the statistical task or simply an artifact of the proof technique adopted here. We leave these extensions to a future study.

In applying the methodology developed here, an important consideration is whether one should fit the semiparametric model to the whole dataset, or only to data to the right of a low threshold. Our analysis of Fort Collins precipitation data indicates that while estimates are robust when data is truncated at or slightly over 0.02 inches, the estimates are sensitive to the presence of a massive number of excess zeroes as well as a relative over- abundance of measurements at 0.01 and 0.02 which cannot be fully addressed by a simple jittering operation. Theorem 4 sheds light on this issue. Critical to the success of the joint estimation is the assumption of smoothness of the entire density function. In applications, it may be useful to threshold the data at a point above which the density function is believed to maintain a common level of smoothness. An alternative approach will be to apply a suitable smooth jitter to the data in the lower tail.

A full estimation of the density function is also appealing with respect to model extension, for example, in accounting for serial correlation or incorporating covariate information. For the latter task, we note that the transformation based density estimation model investigated here is closely related to the joint quantile regression model of Yang and Tokdar (2017). Let \(\xi = \Psi^{-1}\) denote the quantile function of the transformed data \(U = G_\theta(Y)\). Then the quantile function \(Q(p)\) of the original data \(Y\) could be expressed as \(Q(p) = Q_\theta(\xi(p)) = \int_0^p q_\theta(u)du\) where \(Q_\theta = G_\theta^{-1} = \hat{Q}_\theta\). To accommodate a predictor vector \(X\in\mathbb{R}^d\), consider a quantile regression formulation \(Q(p|x) = \int_0^p q_\theta(u)(1 + x^T h(\omega(u)))du\) where \(\omega: u \mapsto (\omega_1(u), \ldots, \omega_d(u))^T \in \mathbb{R}^d\) is unknown and \(h(b)\) is a suitably chosen, fixed transformation that ensures \(1 + x^T h(b) \geq 0\) for all \(b\in\mathbb{R}^d\) and all \(x\) within a given bounded convex domain. This formulation is a special case of the joint linear quantile regression model proposed in Yang and Tokdar (2017) who jointly estimate \((\theta, \xi, \omega)\) by adopting a hierarchical LGP prior on the quantile density \(\xi\) and smooth Gaussian process priors on \(\omega_1, \ldots, \omega_d\). The theoretical analysis presented in the current article is likely to yield a sharper understanding of asymptotic properties of the method by Yang and Tokdar (2017), especially with respect to tail estimation.

Appendices

We present here proofs of the main results stated in Section 3. Several auxiliary technical results are stated whose proofs may be found in supplementary material. Several arguments build upon van der Vaart and van Zanten (2009) which we abbreviate below as VZ09.

Appendix A: Auxiliary Results for Density Estimation

Mixed partial derivatives of functions \(\ell(\theta), \theta = (\alpha, \sigma)\), are denoted by \(\partial^k \ell = \partial^k \ell(\alpha, \sigma)\) for any bi-index \(k = (k_1, k_2)\in\{0, 1, 2, \ldots\}^2\) of order \(|k| := k_1 + k_2\). Below \(\Theta = \{\alpha, \sigma\} \times (\alpha, \sigma) \in (0, \alpha, \sigma \leq \infty, 0 < \alpha < \sigma < \infty\) and any constants appearing in statements and proofs may implicitly depend on the boundary values of \(\Theta\). Let \(\Theta\) denote the interior of \(\Theta\).

Lemma 5. Fix a density \(\psi = \ell(\omega)\) with \(\omega \in C^2[0, 1]\). Let \(\theta = \theta(\psi)\), \(\theta \in \Theta\). Then \(\max_{|k| = 1} sup_{\Theta \in \Theta} |D^\ell \log q_\theta(\psi)| \leq c_0 ||\omega||_{C^2} + c_1 \log(1 + y), \) \(\max_{|k| = 2} sup_{\Theta \in \Theta} |D^\ell \log q(\psi)| \leq c_2 ||\omega||_{C^2}\), for some constants \(c_0, c_1, c_2\).

Lemma 6. Fix a density \(\psi = \ell(\omega)\) with \(\omega \in C^2[0, 1]\). If \(\theta_1, \theta_2 \in \Theta\) then

1. \(d_{KL}(p_{\theta_1}, p_{\theta_2}, p_{\theta_1}) \leq c_3 ||\omega||_{C^2} ||\theta_1 - \theta_2||^2\).

Moreover, there exist positive numbers \(c_4, c_5\) such that \(\|\theta_1 - \theta_2\| < c_1\) then

1. \(\int p_{\theta_2}(\psi)(p_{\theta_2}(\psi) - 1)^2 dy \leq c_3 ||\omega||_{C^2} ||\theta_1 - \theta_2||^2\), and
2. \(V(p_{\theta_2}, p_{\theta_2}, \psi) \leq c_3 ||\omega||_{C^2} ||\theta_1 - \theta_2||^2\).

Lemma 7. Fix \(\theta_1 \in \Theta, \theta_1 = \ell(\omega_1)\) with \(\omega_1 \in C^2[0, 1]\) and \(\epsilon \in (0, t_0)\). There exists a constant \(K\) depending on \(\omega_1\) such that \(d_{KL}(p_{\theta_1}, p_{\theta_2}, p_{\theta_1}) \leq Ke^2\). \((p_{\theta_1}, p_{\theta_2}, p_{\theta_1}) \leq Ke^2\), for every \(\theta_1 \in \Theta\) with \(\|\theta_1 - \theta_2\| \leq \epsilon\) and \(\psi_2 = \ell(\omega_2)\) with \(\omega_1 - \omega_2 \|\|_\infty \epsilon^{1-n_2/2}|\omega_1 - \omega_2|^{1,n_2/2}.

Lemma 8. If \(\theta_1, \theta_2 \in \Theta\) and \(\psi_1 = \ell(\omega_1), \psi_2 = \ell(\omega_2)\) with \(\omega_1, \omega_2 \in C^2[0, 1]\) then \(d_{KL}(p_{\theta_1}, p_{\theta_2}, p_{\theta_1}) \leq c_4 ||\omega_1||^{1/2}_C ||\theta_1 - \theta_2|| + ||\omega_1 - \omega_2||^{1/2}_\infty e^{-n_1/2}.

Appendix B: Proof of Condition C1∗

For a mean-zero Gaussian process on \(C^2[0, 1]\) with covariance \(c_2\), let \(v^{\lambda}\) denote the Gaussian measure with respect to the Borel \(\sigma\)-algebra on \((C^2[0, 1], \|\|_\infty);\) see sec 2 of van der Vaart and van Zanten (2008) for necessary technical details. Define \(v^{\lambda}(\cdot) = f^{\lambda}(-\pi\lambda, \lambda)\) as the probability measure on \(C^2[0, 1]\) under the hierarchical Gaussian process prior specification with a prior density \(p_{\lambda}\) on the inverse length-scale parameter. In light of Lemma 7, to prove Condition C1∗ it is enough to show that for all large \(n\), both \(t_n = \pi_\theta(\theta; ||\theta - \theta^{*}\|_\infty \leq \epsilon)\) and \(w_n = \pi(\omega; ||\omega - \omega^{*}\|_\infty \leq \epsilon)\) are larger than \(e^{-CK}\), where \(C\) is a constant that may depend on \(\|\phi^{*}\|_{C^2}\). The bound on \(t_n\) follows trivially from the assumption on \(\pi\) and that on \(w_n\) follows directly from Theorem 3.1 of VZ09.

Appendix C: Proof of Condition C2∗

Suppose there exist sets \(B_1, B_2, \ldots \subset C^2[0, 1]\) such that for all large \(n\),

\[
\pi(B_n) \leq e^{-(C+4)n^2/2}.
\]

\[
\log N(e_n, B_n) \leq ne_n^2/2,
\]

\[
|sup\{||\omega||_{C^2}; \omega \in B_n\}| \leq (ne_n)^b,
\]

for some \(b \geq 1\). Then \(F_1, F_2, \ldots\) could be simply constructed as \(F_n = \{p_{\theta_1, \psi} ; \theta \in \Theta, \psi = \ell(\omega), \omega \in B_n\}\). To see that these sets satisfy the requirements of C2∗, notice that \(P(F_n) \rightarrow \pi(B_n) \leq e^{-(C+4)n^2/2}\), and, by Lemma 8, \(\pi(e_n, F_n, d\ell) \leq \log(N(e_n, c/2e_n)^b/2), \theta, \|\|_\infty) +\)
log $N(\varepsilon_n, B_n, \| \cdot \|_{C^2}) \leq \log(\text{diam}(\Theta) \cdot n^h) + n^2\varepsilon_n^2/2 \leq n^2\varepsilon_n^2$ for all large $n$.

Conditions (6)–(7) mirror conditions (3.6)–(3.7) of VZ09, but with the crucial technical difference that we need entropy calculation in $\| \cdot \|_{C^2}$ as opposed to $\| \cdot \|_{\infty}$. Accordingly, we adapt the construction of $B_{\lambda}$ for $(C(0,1), \| \cdot \|_{\infty})$ by VZ09 to $(C^2[0,1], \| \cdot \|_{C^2})$. Our adaptation also produces a smaller exponent $h$ in (8) than what is possible with the original construction of VZ09. Although a larger exponent is not critical to the current proof, it proves useful for tail index estimation.

Our adaptation builds on the well-known fact that a centered Gaussian process with covariance $c_j$ has infinitely differentiable sample paths with probability one. Therefore, the Gaussian measures $\nu^h$ introduced in the preceding section could also be viewed as probability measures with respect to the refined Borel $\sigma$-algebra of $(C^2[0,1], \| \cdot \|_{C^2})$. A more formal treatment is outlined below. Hereafter, $\Theta(z)$ and $\Xi(z)$ denote the real and imaginary parts of a complex number $z$.

VZ09 show that the reproducing kernel Hilbert space $\mathbb{H}^2$ associated with $c_j$ consists of functions $h(u) = \Theta(t) e^{\mu(t)} \mathbb{R}(t) d\tau$ with $\| h \|_{\mathbb{H}^2} = \| h \|_{L_2(\mu_j)}$, where $\mu_j(t) = -2^{-1}/2^{\lambda / \sqrt{2\pi}}$ is the spectral density associated with $c_j$. By applying Cauchy-Schwarz inequality, with differentiations under integration as needed, it follows that

$$\| h \|_{\mathbb{H}^2} \leq \| h \|_{\mathbb{H}^2} \leq \sqrt{2\lambda} \| h \|_{\mathbb{H}^2}, \quad \| h \|_{\mathbb{H}^2} \leq \sqrt{2\lambda} \| h \|_{\mathbb{H}^2}. \tag{9}$$

Clearly, $(h \|_{C^2} \leq (1 + \sqrt{2\lambda} + \sqrt{2\lambda}^2)^{\lambda / \sqrt{2\pi}} \| h \|_{\mathbb{H}^2}$ and $\mathbb{H}^2$ can be continuously and densely embedded within the Banach space $(C^2[0,1], \| \cdot \|_{C^2})$, guaranteeing a Borel measure $\nu^h$ on the embedding Banach space matching the law of a centered Gaussian process with covariance $c_j$. As before, define $\nu^h = \nu^h(\cdot, \lambda, \alpha, \tau, d\lambda)$.

Let $\mathbb{H}^2_1$ and $B_1$ denote the unit balls of $\mathbb{H}^2$ and $C^2[0,1]$. Recall, $\varepsilon_n = B_{\alpha}^{-1}(\log n), \varepsilon_n = \log n$ where we are free to choose $B > 0$. To start off, take $B$ large enough such that $(n \leq n^2\varepsilon_n^2 < 2$ for all $n$. Let $m_n$ be the smallest integer larger than $\log_2(n^2)$. Define

$$B_n = \left\{(m_{j=1}^{\min\{\log_2 n, 2\}}) \cup \left\{2^{\lambda / \sqrt{2\pi}} \Theta(t) d\tau \right\} : \Theta(t) \left(\Theta(t) + \Theta(t) \right) \right\} + \varepsilon_n B_1 \tag{10}$$

where $M_n = 16\varepsilon_n^{1/2} / \log(\varepsilon_n)$ is C taken from Lemma 10, and $\varepsilon_n = \log(n)$ is $C$. Because of (9), for all large $n, B_n \subset (0, M_n]^{\log_2 n} B_1$, and hence $B_n$ satisfies (8) with $b = 2$. By Lemma 4.7 of VZ09, $B_n \supseteq M_n B_1 + \varepsilon_n B_1$ for every $0 < \lambda \leq \log n$. Borell’s inequality implies that $\nu^h(B_n) \leq 1 - \Phi(\sqrt{W^h} \varepsilon_n B_1) + \varepsilon_n B_1 \leq 1 - \Phi(\varepsilon_n B_1) + \varepsilon_n B_1$ where the second inequality follows since $\nu^h(\varepsilon_n B_1)$ is decreasing in $\lambda$ for every $\varepsilon > 0$. (Lemma 9). As $\varepsilon_n B_1 / \varepsilon_n B_1 < 4 / \sqrt{\log(1 / \varepsilon_n B_1)}$ for all large $n$ (Lemma 10), it must be that $\nu^h(B_n) \leq 1 - \Phi(\varepsilon_n B_1) \leq \varepsilon_n^{2\lambda / \sqrt{2\pi}} / \log(\varepsilon_n B_1)$ for all large $n$. This establishes (6), with $B$ chosen suitably large, so that $(\varepsilon_n B_1) \leq \varepsilon_n^{2\lambda / \sqrt{2\pi}}$ for all large $n$.

To establish (7), first note that every $h \in \cup_{(1, \log n)} (M_n B_1)$ satisfies $\| h \|_{\mathbb{H}^2} \leq \varepsilon_n B_1$ by (9), that is, as an element of $C^2[0,1]$, the function $h(u)$ is within $\varepsilon_n B_1$ of a constant function whose constant value ranges within $[-M_n, M_n]$. Clearly, log $N(\log(1 / \varepsilon_n B_1), \| \cdot \|_{\mathbb{H}^2})$ satisfies $\| h \|_{\mathbb{H}^2} \leq \varepsilon_n B_1$ for all large $n$ and $M_n$. Hence, for all large $n$ and $M_n$, we have

$$\log N(\log(1 / \varepsilon_n B_1), \| \cdot \|_{\mathbb{H}^2}) \leq \log(2\pi\log(n)) \leq \log(2\pi\log(n)) \leq \log\left(\frac{2\pi\log(n)}{\log(1 / \varepsilon_n B_1)}\right).$$

By Lemma 10, $\log N(\log(1 / \varepsilon_n B_1), \| \cdot \|_{\mathbb{H}^2}) \leq \log(2\pi\log(n)) \leq \log(2\pi\log(n)) \leq \log\left(\frac{2\pi\log(n)}{\log(1 / \varepsilon_n B_1)}\right).$
Take εn = (4 + C)n1/2εn. For any f = ρ,ψ ∈ Fn it follows from (9) that if φ = log ψ then |φ|∞ ≤ Dn := C1n3/2(log n)n + 1 for some constant C1. Set α1 = εn and note that α2 ≤ min(1, α∗) and partition U1 = U1n ∪ U2n where U1n = U(τ1−εn) and U2n = U(τ1 + εn). By Lemma 12 (with A = εn), for any ρ1 > 0,0
\[
F_n ∩ U(τ_1, δ_n) \subset \bigcap \{ f : α_*(f) ≤ ε_n \} \bigcap \{ f : α_*(f) > 2ε_n \}
\]
Indeed, for large n, where εn = α/α∗ > 0. Therefore, with a suitably large choice of C1 we can make \( F_n(t_1) \geq ε_n \) for all large n provided ρ1 ≤ ρ(ε), and in case of an equality, s = 2ρ1 + 4. On the other hand, in order to have δ1n ∈ Fn(t1)/2 ≥ εn, we need to choose B12 suitably large and ρ1 ≤ ρ(ε), and in case of an equality, s = 2ρ1 + 4. With (12)–(13) established with ρ1 > 0 chosen as the minimum of the above two bounds and s > 0 set accordingly, apply Lemma 11 to conclude that there exist test functions T1n = T1n(Y1, ..., Yn) such that [p1n, T1n ≤ e−nε2] and sup[p1n(1 − T1n) : f ∈ F_n ∩ U(τ1, δ_n) ≤ e−nε2] for all large n.

Next, repeat the same arguments for testing f = f∗ versus f ∈ F_n ∩ U(τ1, δ_n). Rewrite the target rate as B1 n−ρ (log n) a = 24 + 5δ_n where δ_n = B22 τ_n, τ_n = C22 n−ρ (log n) a, and τ_n = D1n/τ_n a, with B22, C22 to be adjusted as needed. As argued in the preceding paragraph, the choices of ρ, a imply that \( \min(F_n(t_1), (F_n(t_1)/2) \geq ε_n \) and Lemma 14 (where A = α1) implies that \( F_n(t_1) \subset \bigcap \{ f : α_*(f) ≤ ε_n \} \bigcap \{ f : α_*(f) > 2ε_n \} \geq δ_n \). Therefore, by Lemma 11, there are test functions T2n = T2n(Y1, ..., Y_n) such that \[ p1n (T2n ≤ e−nε2] and sup[p1n(1 − T2n) : f ∈ F_n ∩ U(τ2, δ_n) ≤ e−nε2] for all large n. The proof is now complete by taking Tn = max(T1n, T2n).

Supplementary Materials

Additional proofs and graphics: are included as a single PDF document.

Codes: R codes (FortCollins.R) along with dataset and auxiliary codes required to reproduce Fort Collins precipitation analysis presented in this article (Zipped file). Included is a CRAN submission-ready R package "abide" containing code to fit the proposed semiparametric Bayesian density estimation method (GNU zipped tar file).

Funding

This research was partially supported by grants DMS1613173 and DMS2014861 from the National Science Foundation.

ORCID

Surya T. Tokdar http://orcid.org/0000-0001-5162-1155
Sheng Jiang http://orcid.org/0000-0002-1662-5106

References

Adler, R. J., and Taylor, J. E. (2009), Random Fields and Geometry, New York: Springer. [166]
Alves, M. F. (2001), "A Location Invariant Hill-Type Estimator," Extremes, 4, 199–217. [163]
Andrews, C., and Thoms, J. (2008), "A Tutorial on Adaptive MCMC," Statistics and Computing, 18, 343–373. [167]
 Balkema, A., and de Haan, L. (1974), "Residual Life Time at Great Age," Annals of Probability, 2, 792–804. [163]
 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, 70, 825–848. [166]
 Beiriant, J., Joossens, E., and Segers, J. (2009), "Second-Order Refined Peaks-Over-Threshold Modelling for Heavy-Tailed Distributions," Journal of Statistical Planning and Inference, 139, 2800–2815. [172]
 Carpenter, A., and Kim, A. K. (2015), "Adaptive and Minimax Optimal Estimation of the Tail Coefficient," Statistica Sinica, 25, 1133–1144. [168,172]
 Castillo, E. (2012), Extreme Value Theory in Engineering. Boston: Elsevier. [163]
 de la Peña, V., and Kotz, S. (2010), "Parameter Estimation of the Generalized Pareto Distribution? Part II," Journal of Statistical Planning and Inference, 140, 1374–1388. [163]
 Dekkers, A., Einmahl, J. H., and de Haan, L. et al. (1989), "A Moment Estimator for the Index of an Extreme-Value Distribution," The Annals of Statistics, 17, 1833–1855. [163]
 Diaconis, P., and Freedman, D. (1986), "On the Consistency of Bayes Estimates," The Annals of Statistics, 14, 1–26. [164,167]
 do Nascimento, F. F., Gardner, C., and Lopes, H. F. (2012), "A Semiparametric Bayesian Approach to Extreme Value Estimation," Statistics and Computing, 22, 661–675. [163]
 Durieu, G., Graña, I., Pham, Q.-K., and Tricot, J.-M. (2015), "Nonparametric Adaptive Estimation of Conditional Probabilities of Rare Events and Extreme Quantiles," Extremes, 18, 473–478. [169,170]
 Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013), Modelling Extremal Events: For Insurance and Finance (Vol. 33), Berlin: Springer. [163]
 Ghosal, S., Ghosh, J. K., and Ramamoorthy, R. V. (1999), "Posterior Consistency of Dirichlet Mixtures in Density Estimation," The Annals of Statistics, 27, 143–158. [167]
 Ghosal, S., and van der Vaart, A. (2017), "Fundamentals of Nonparametric Bayesian Inference (Vol. 44), Cambridge: Cambridge University Press. [167,168,173]
 Gilleland, E., and Katz, R. W. (2011), "New Software to Analyze How Extremes Change Over Time," Eos, Transactions American Geophysical Union, 92, 13–14. [170]
 Gu, M., Wang, X., and Berger, J. O. (2018), "Robust Gaussian Stochastic Process Emulation," The Annals of Statistics, 46, 3038–3066. [166]
 Hall, P., and Welsh, A. (1984), "Best Attainable Rates of Convergence for Estimates of Parameters of Regular Variation," Annals of Statistics, 12, 1079–1084. [168,171]
 Hall, P., and Welsh, A. H. (1985), "Adaptive Estimates of Parameters of Regular Variation," The Annals of Statistics, 13, 331–341. [168,171]
 Hill, B. M. (1975), "A Simple General Approach to Inference about the Tail of a Distribution," Annals of Statistics, 3, 1163–1174. [163]
 Katz, R. W., Parlange, M. B., and Naveau, P. (2002), "Statistics of Extremes in Hydrology," Advances in Water Resources, 25, 1287–1304. [163,170,171]
 Kleijn, B. (2021), "Frequentist Validity of Bayesian Limits," The Annals of Statistics, 49, 182–202. [168]
 Lenk, P. J. (1988), "The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities," Journal of the American Statistical Association, 83, 509–516. [164]
——— (1991), "Towards a Practicable Bayesian Nonparametric Density Estimator," Biometrika, 78, 531–543. [164]
 Leonard, T. (1978), "Density Estimation, Stochastic Processes and Prior Information," Journal of the Royal Statistical Society, Series B, 40, 113–146. [164]
 Li, C., Lin, L., and Dunson, D. B. (2019), "On Posterior Consistency of Tail Index for Bayesian Kernel Mixture Models," Bernoulli, 25, 1999–2028. [163,164,168]
MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., and Russell, G. (2011), "A Flexible Extreme Value Mixture Model," *Computational Statistics & Data Analysis*, 55, 2137–2157. [163]

Markovich, N. (2007), *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, Chichester: Wiley. [163,164]

Paulo, R. (2005), "Default Priors for Gaussian Processes," *The Annals of Statistics*, 33, 556–582. [166]

Pickands, J. (1975), "Statistical Inference Using Extreme Order Statistics," *Annals of Statistics*, 3, 119–131. [163]

Rice, S. O. (1944), "Mathematical Analysis of Random Noise," *The Bell System Technical Journal*, 23, 282–332. [166]

Scarrot, C., and MacDonnald, A. (2012), "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification," *Statistical Journal*, 103, 33–60. [163,170,171]

Schwartz, L. (1965), "On Bayes Procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4, 10–26. [168]

Snelson, E., and Ghahramani, Z. (2006), "Sparse Gaussian Processes using Pseudo-Inputs," in *Advances in Neural Information Processing Systems*, pp. 1257–1264. [166]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [167]

Tancredi, A., Anderson, C., and O’Hagan, A. (2006), "Accounting for Threshold Uncertainty in Extreme Value Estimation," *Extremes*, 9, 87–106. [163]

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. [164,166]

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. [165]

van der Vaart, A., and van Zanten, J. (2008), "Rates of Contraction of Posterior Distributions based on Gaussian Process Priors," *The Annals of Statistics*, 36, 1435–1463. [172]

van der Vaart, A. W., and van Zanten, J. H. (2009), "Adaptive Bayesian Estimation Using a Gaussian Random Field with Inverse Gamma Bandwidth," *The Annals of Statistics*, 37, 2655–2675. [165,166,168,171,172]

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. [166,172]