Heavy-Tailed Density Estimation

Surya T Tokdar*
Department of Statistical Science, Duke University
and
Sheng Jiang
Department of Statistics, University of California, Santa Cruz
and
Erika L Cunningham
Department of Statistical Science, Duke University

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.

Keywords: Semiparametric estimation, logistic Gaussian processes, posterior contraction, tail index estimation, regular variation.

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, e.g., precipitation (Katz et al., 2002), financial returns or insurance loss (Embrechts et al., 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

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

rates. Estimation methods with rich shape flexibility and guaranteed L^1 estimation consistency may provide no meaningful inference on the tails of the distribution; see Markovich (2007); Li et al. (2019) for detailed discussions and cautionary results on kernel mixture models.

When interest focuses on estimating only tail features, e.g., 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 *peaks-over-threshold* estimation methods, where a generalized Pareto distribution (GPD) is fitted to the subsample of observations exceeding a high threshold; see de Zea 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 $\bar{F}(y) = \int_y^\infty f(t)dt$ is regularly varying, i.e.,

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

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

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 MacDonnald, 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 (Tancredi et al., 2006; MacDonald et al., 2011; do Nascimento et al., 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

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

where $g_\theta(y) = \sigma^{-1}\{1 + y/(\alpha\sigma)\}^{-(\alpha+1)}$, $G_\theta(y) = \int_0^y g_\theta(z)dz$, $y > 0$, are the density and distribution functions of a GPD with location 0, scale σ and shape $1/\alpha$; here $\theta = (\alpha, \sigma) \in (0, \infty)^2$ is an unknown vector, and, ψ is an unknown density function on $(0, 1)$. Under this model, $Y \sim f(y)$ if and only if $U := G_\theta(Y) \sim \psi$, and, $\alpha_+(f) = \alpha_+(g_\theta) = \alpha$ under a regularity condition on $\psi(u)$ as $u \to 1$ (Lemma 1). Markovich (2007) offers a thorough analysis of an estimation approach where one first obtains an estimate $\hat{\theta}$ of θ by thresholding data $Y_1, \ldots, Y_n$ at a high quantile, and then a nonparametric estimate $\hat{\psi}$ of ψ is obtained based on the transformed data $\hat{U}_i = G_{\hat{\theta}}(Y_i)$, $i = 1, \ldots, n$. With $\hat{\psi}$ estimated by a variable kernel mixture, the back-transformed density $\hat{f} = p_{\hat{\theta},\hat{\psi}}$ 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 θ and ψ 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 non-parametric density ψ? 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 ψ (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 $f^* = p_{\theta^*,\psi^*}$ with a sufficiently smooth ψ^* . 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 et al., 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 data sets 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 expressive of an entire range of polynomial tail decay rates under a regularity assumption on ψ.

Let $\mathcal{P}$ denote the class of densities ψ on $(0,1)$ satisfying $\bar{\Psi}(1-u) = u\tilde{L}(1/u)$ for some slowly varying function $\tilde{L}$; here Ψ denotes the distribution function of ψ and $\bar{\Psi} = 1 - \Psi$. Note that if $L(y)$ is slowly varying then

$$\lim_{y\to\infty} \frac{L(a(y)y)}{L(y)} = 1 \tag{3}$$

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 $\psi \in \mathcal{P}$ then $f = p_{\theta,\psi}$ is regularly varying with tail index α. 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 $\psi \in \mathcal{P}$.*

Proof. If $f = p_{\theta,\psi}$ then $\bar{F}(y) = \bar{\Psi}(1-\bar{G}_\theta(y)) = \bar{G}_\theta(y)\tilde{L}(1/\bar{G}_\theta(y))$, with $\bar{G}_\theta(y) = 1-G_\theta(y) = y^{-\alpha}L_\theta(y)$, $L_\theta(y) = \{1/y+1/(\alpha\sigma)\}^{-\alpha} \to c_\theta := (\alpha\sigma)^\alpha$ as $y \to \infty$. Therefore, $\bar{F}(y) = y^{-\alpha}L(y)$ where $L(y) = L_\theta(y)\tilde{L}(y^\alpha/L_\theta(y))$ 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

$$\psi(u) = \frac{f(G_\theta^{-1}(u))}{g_\theta(G_\theta^{-1}(u))}, \quad u \in (0,1). \tag{4}$$

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

Lemma 1 says, with $\psi \in \mathcal{P}$ the semiparametric model (2) is fully expressive of all regularly varying densities on $(0,\infty)$ with tail index uniquely identified by the model parameter α. It also says that the pair (σ,ψ) is not uniquely identifiable. Although one could fix σ and have both α and ψ 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 ψ. The LGP prior introduced below concentrates around ψ 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 σ creates an important counterbalance. It offers an entire arc of equivalent (σ,ψ) 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 $f = g_{(2,1)}$ the pair $(\sigma = 1, \psi \equiv 1)$ 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 π_σ is the half-Cauchy distribution on $(0,\infty)$ and π_α is the distribution of $\alpha = \underline{\alpha} + (2-\underline{\alpha}) \cdot e^{\zeta/1.5}$ with ζ distributed according to the standard logistic distribution on the real line. The choice of π_α restricts $\alpha > \underline{\alpha}$ with probability one, where $\underline{\alpha} > 0$ is treated as a hyperparameter to be fixed by the modeler. The numerical analyses presented in Section 4 were carried out with $\underline{\alpha} = 0.5$, for which the extreme value index $\xi = \frac{1}{\alpha}$ has unimodal density on $(0,2)$ with a gentle peak at $\xi = 0.5$ (i.e., $\alpha = 2$). We also experimented with $\underline{\alpha} = 0.1$ and all posterior estimates were found to be essentially the same as with $\underline{\alpha} = 0.5$.

2.2 LGP prior for ψ

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

We adopt a hierarchical LGP prior for ψ in (2). Let $c_\lambda(u,v) = \exp\{-\lambda^2(u-v)^2\}$ denote the unit variance Gaussian covariance kernel with inverse length-scale parameter $\lambda > 0$. It

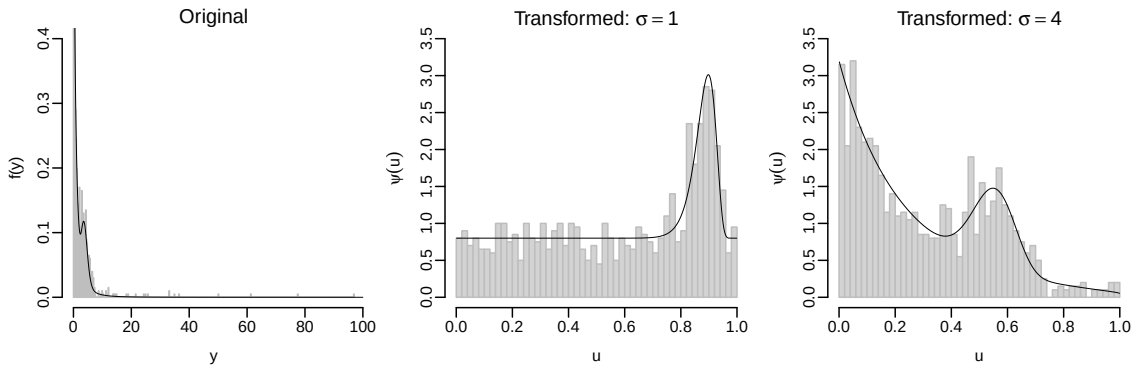

Figure 1: Lack of identifiability of (σ, ψ) and the importance of scale expressiveness. Left: graph of $f(y) = 0.8 \times g_{(2,1)}(y) + 0.2 \times \tilde{g}(y)$ where $\tilde{g}(y)$ is the density of a gamma distribution with mean 4 and variance 1; overlaid on the histogram of a sample of size $n = 1000$ 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 $\alpha = 2$ and two choices of the scale: $\sigma = 1$ (middle) and $\sigma = 4$ (right). The larger scale value produces a flatter ψ 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 σ.

is well known that if ω is a mean zero Gaussian process with covariance $\kappa^2 c_\lambda$ for some $\kappa > 0$, then $\omega \in C[0,1]$ with probability one, and hence, the probability distribution $\mathrm{LGP}(0, \kappa^2 c_\lambda)$ is well defined for every $\kappa > 0, \lambda > 0$. The prior on ψ is implicitly defined by the hierarchy

$$\psi \sim \mathrm{LGP}(0, \kappa^2 c_\lambda), \quad (\kappa^2, \lambda) \sim \pi_{\kappa^2} \times \pi_\lambda, \tag{5}$$

with the distributions π_{κ^2} and π_λ on $(0, \infty)$ described below.

It is clear that if $\omega \in C[0,1]$ and $\psi = \mathcal{L}(\omega)$, then $\psi(1) := \lim_{u \to 1} \psi(u)$ exists and $\psi(1) \in (0, \infty)$. By mean value theorem $\bar{\Psi}(1-u) = u\psi(t(u))$ for some $t(u) \in [1-u, 1]$. Consequently, $\tilde{L}(y) = \psi(t(1/y))$, is slowly varying because $\lim_{y \to \infty} \tilde{L}(y) = \psi(1) \in (0, \infty)$. Therefore, under the hierarchical LGP prior adopted here, $\Pr(\psi \in \mathcal{P}) = 1$. Of course, the prior support of ψ is actually smaller than $\mathcal{P}$, because $\lim_{u \downarrow 0} u^{-1} \bar{\Psi}(1-u) \in (0, \infty)$ almost surely under the prior, whereas $\mathcal{P}$ contains densities ψ 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 $0 \leq \underline{\alpha} < \overline{\alpha} \leq \infty$ are such that $\alpha \in (\underline{\alpha}, \overline{\alpha})$ with probability one under the prior π_θ.

Theorem 2. *Let f^* be any bounded, continuous, regularly varying density on $(0, \infty)$ with tail index $\alpha^* \in (\underline{\alpha}, \overline{\alpha})$. If $(\theta, \psi) \sim \pi_\theta \times \mathrm{LGP}(0, \kappa^2 c_\lambda)$ for some $\kappa > 0, \lambda > 0$, then for every $\epsilon > 0$, $\Pr(d_{\mathrm{KL}}(f^*, p_{\theta,\psi}) < \epsilon) > 0$, where $d_{\mathrm{KL}}(f, g) = \int f(y) \log\{f(y)/g(y)\} dy$ 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 $\underline{\alpha} < \alpha_0 < \alpha^*$ and $\sigma_0 > 0$. Let ψ_0 be defined as in (4) so that $f^* = p_{\theta_0, \psi_0}$. Since $\alpha_0 < \alpha^*$, g_{θ_0} has heavier tails than f^* and hence ψ_0 is bounded and continuous with

$\psi_0(1) = 0$. Consequently, the density $\psi_1(u) := (1-\delta)\psi_0(u) + \delta, \quad u \in (0,1)$, is bounded and continuous, and is bounded above δ, and therefore, $\omega_1 = \log \psi_1$ can be extended to an element of $C[0,1]$. Now, for any $\psi = \mathcal{L}(\omega)$ with $\omega \in C[0,1]$, $d_{\mathrm{KL}}(f^*, p_{\theta_0,\psi}) = d_{\mathrm{KL}}(\psi_0, \psi) \le d_{\mathrm{KL}}(\psi_0, \psi_1) + \int_0^1 \psi_0(u) \log \frac{\psi_1(u)}{\psi(u)} du \le -\log(1-\delta) + 2\|\omega - \omega_1\|_\infty$. Therefore, $\Pr(d_{\mathrm{KL}}(f^*, p_{\theta,\psi}) < \epsilon \mid \theta = \theta_0) \ge \Pr(\|\omega - \omega_1\|_\infty < \frac{\epsilon}{2})$, where the latter probability, calculated for a Gaussian process ω with mean zero and covariance $\kappa^2 c_\lambda$, must be positive because such a Gaussian process has the entire $C[0,1]$ in its uniform topology support (Tokdar 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 λ is of critical importance here because of its direct influence on the range of smoothing; though in our experience, a prior on κ also helps with model fit and posterior computation via Markov chain Monte Carlo. We take π_{κ^2} to be a convenient inverse-gamma distribution with shape a_κ and rate b_κ, i.e., $(1/\kappa^2) \sim Ga(a_\kappa, b_\kappa)$ which is partially conjugate to the likelihood function in κ^2 and allows this parameter to be integrated out during model fitting. No such conjugate choice exists for λ and formal subjective or objective principles are difficult to apply in selecting π_λ; however, see Paulo (2005); Gu et al. (2018) for relevant discussions.

An alternative track is to seek π_λ 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 λ 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 $\pi_\lambda \sim Ga(a_\lambda, b_\lambda)$. Our numerical experiments were carried out with $a_\kappa = b_\kappa = 3/2$, $a_\lambda = 16$ and $b_\lambda = 2.2$. The latter choices could be appreciated in several ways. Consider $\rho = c_\lambda(0, \Delta) = e^{-\lambda^2 \Delta^2}$ which gives the correlation of the Gaussian process at a distance Δ. With $\Delta = 10\%$, our choice of π_λ assigns 95% prior probability to $\rho \in (0.28, 0.84)$ with prior mean and median ≈ 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 of up-crossings of zero of a mean zero Gaussian process on the unit interval with covariance $\kappa^2 c_\lambda$ is $\lambda/(\pi\sqrt{2}) \approx 0.22\lambda$ (Rice, 1944; Adler and Taylor, 2009). With our choice of π_λ, 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 ψ only through the finite vector $\psi_U = (\psi(u_1), \ldots, \psi(u_n))^\top$. Unfortunately, the joint prior density of ψ_U, given model hyper-parameters (λ, κ), is not available in closed form. This necessitates involving the latent Gaussian process ω in the representation $\psi = \mathcal{L}(\omega)$ in posterior computation. However, ψ_U depends on both the corresponding vector ω_U and the scalar $\omega_{\mathrm{norm}} = \int_0^1 e^{\omega(u)} du$ giving the normalization in the logistic trans-

form 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 ω 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 ω and ψ are to be represented, respectively, as the vector $\omega_T = (\omega(t_1), \ldots, \omega(t_L))^\top$ and the corresponding vector ψ_T. Given ω_T, a very accurate approximation to ω_{norm} could be obtained by applying the trapezoidal rule of numerical integration to the pair (T, ω_T), readily producing the vector ψ_T. To evaluate ψ_U, which is needed for likelihood evaluation, it is useful to formally express the trapezoidal approximation to ω_{norm} as the exact integration of the function $h(u)$ that linearly interpolates the points $(t_l, e^{\omega(t_l)})$, $l = 1, \ldots, L$. We may now view ψ_T as the evaluation over the grid T of the (normalized) density function $\bar{h}(u) = h(u)/\int_0^1 h(t)dt$. Consequently, ψ_U could be readily equated with the corresponding vector $\bar{h}_U$. 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 (θ, ω_T). The prior density of ω_T, given (λ, κ), is a multivariate normal density with mean zero and covariance $\kappa^2 C_T(\lambda)$ where $C_T(\lambda) = ((c_\lambda(t_l, t_k)))_{l,k=1}^L$. 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 ω_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 ω 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 $\tilde{\omega}(u) = \mathbb{E}[\omega(u) \mid \omega(s_1), \ldots, \omega(s_m)]$, gives a smooth interpolation of the graph of (S, ω_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 (λ, κ), but a hyper-parameter marginalized extension proposed in Yang and Tokdar (2017) and described below offers considerable additional speed up.

Integrate out κ^2 from the model and express the prior distribution of ω_S given λ as the multivariate Student-t distribution with pdf $p(\omega_S|\lambda) \propto \{1 + \omega_S^\top C_S(\lambda)^{-1}\omega_S/(2b_\kappa)\}^{-(a_\kappa + m/2)}$ where $C_S(\lambda)$ is the analogue of $C_T(\lambda)$ over the knot set S. It is impossible to analytically integrate out λ, but a discrete integration could be carried out by replacing π_λ with a discrete approximation over a dense set of support points $\{\lambda_1, \ldots, \lambda_G\} \subset (0, \infty)$. With $\pi^*_\lambda(g) := \Pr(\lambda = \lambda_g)$, the prior distribution of ω_S is the mixture density $p(\omega_S) = \sum_g \pi^*_\lambda(g) p(\omega_S|\lambda_g)$.

The vector $\tilde{\omega}_T$, which is the predictive process replacement of ω_T, can be computed analytically from ω_S as $\tilde{\omega}_T = \sum_g \pi^*_\lambda(g|\omega_S) A_g \omega_S$ where $A_g = C_{TS}(\lambda_g) C_S(\lambda_g)^{-1}$ with $C_{TS}(\lambda)$ denoting the $L \times m$ matrix with elements $c_\lambda(t,s)$, $t \in T, s \in S$, and, $\pi^*_\lambda(g|\omega_S) \propto \pi^*_\lambda(g) p(\omega_S|\lambda_g)$. We select the support points $0 < \lambda_1 < \cdots < \lambda_G$ of π^*_λ based on the knots set S. First λ_1 is fixed such that $\rho_1 := c_{\lambda_1}(0, 0.1) = 0.95$ and then successive λ_g values are chosen so that $d_{\mathrm{KL}}(N(0, C_S(\lambda_{g-1})), N(0, C_S(\lambda_g))) = 0.5$ until we get $\rho_{G+1} := c_{\lambda_{G+1}}(0, 0.1) < 0.2$. This gradual stepping down ensures successive $N(0, \kappa^2 C_S(\lambda_g))$ distributions maintain considerable overlap, eliminating any major gaps in the prior distribution of ω_S due to the discretization of λ. In our experience, posterior calculation is not sensitive to exact choices of the bookending values of ρ_1 and ρ_{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 $(\alpha, \sigma, \omega_S)$. 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 paper were achieved by using one block containing ω_S, one block updating $\theta = (\alpha, \sigma)$, and one block updating all $(m+2)$ parameters simultaneously. An important consequence of the discretization of λ is that all relevant matrices, namely $\{(A_g, R_g) : g = 1, \ldots, G\}$, where R_g is a Cholesky factor of $C_S(\lambda_g)$, 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 end points equalling 0 and 1. This choice of S leads to a discretization of π_λ with $G = 30$ support points. In analyzing Fort Collins precipitation data with sample size $n = 6180$ (Section 5), it took 9.8 minutes 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) minutes respectively. For the original set with $n = 6180$, it took 12.2 minutes (1.2x) 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 minutes (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 Π denote the induced prior measure on $f = p_{\theta,\psi}$, with $(\theta, \psi) \sim \pi_\theta \times \text{LGP}(0, \kappa^2 c_\lambda)$ where we treat $\kappa = 1$ as fixed, and work with a gamma prior on λ. We allow π_θ to be arbitrary but assume it has a compact support $\Theta = [\underline{\alpha}, \overline{\alpha}] \times [\underline{\sigma}, \overline{\sigma}]$ for some $0 < \underline{\alpha} < \overline{\alpha} < \infty$, $0 < \underline{\sigma} < \overline{\sigma} < \infty$, with a strictly positive density in the interior of Θ. Compactness of Θ 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 Θ arbitrarily without virtually altering the posterior contraction rates.

We may view Π 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 $\Pi(df \mid y_1, \ldots, y_n) \propto \{\prod_{i=1}^n f(y_i)\}\Pi(df)$. Below $\mathcal{H}^\beta[0,1]$ denotes the Hölder-β space consisting of functions on $[0,1]$ that are $\lfloor\beta\rfloor$ times continuously differentiable with the $\lfloor\beta\rfloor$-th derivative being Hölder continuous of exponent $\beta - \lfloor\beta\rfloor$ where $\lfloor\beta\rfloor$ is the largest integer smaller than β. The minimax density estimation rate over Hölder-β classes is $n^{-\frac{\beta}{2\beta+1}}$ (Stone, 1982).

Theorem 3. *If $Y_1, \ldots, Y_n \overset{IID}{\sim} f^*$ where f^* is a bounded, continuous, regularly varying density on $(0, \infty)$ with tail index $\alpha^* \in (\underline{\alpha}, \overline{\alpha})$, then $\text{plim}_{n\to\infty} \Pi(\{f : \|f - f^*\|_1 > \epsilon\} \mid Y_1, \ldots, Y_n) = 0$ for every $\epsilon > 0$. Additionally, if $f = p_{\theta^*,\psi^*}$ with θ^* in the interior of Θ and $\phi^* = \log \psi^* \in \mathcal{H}^\beta[0,1]$ for some $\beta > 2$, then the fixed error margin ϵ may be replaced with the vanishing sequence $\epsilon_n = Bn^{-\frac{\beta}{2\beta+1}}(\log n)^{\frac{4\beta+1}{2\beta+1}}$ for some large constant B.*

Proof. To prove the first claim, we only need to establish (Ghosal et al., 1999, Theorem 2)

C1. $\Pi(\{f : d_{\text{KL}}(f^*, f) < \epsilon\}) > 0$ for every $\epsilon > 0$,

C2. for any $\epsilon > 0$, there exist constants $c, C > 0$ and sets $\mathcal{F}_1, \mathcal{F}_2, \ldots \subset \mathcal{F}$, such that $\Pi(\mathcal{F}_n^c) \le ce^{-Cn}$ and $\log N(\epsilon, \mathcal{F}_n, d_H) \le n\epsilon^2$ for all large n,

where $N(\delta, \mathcal{F}_n, d_H)$ denotes the covering number of $\mathcal{F}_n$ by balls of radius δ in the Hellinger metric $d_H(p,q) = [\int\{\sqrt{p}(y) - \sqrt{q}(y)\}^2 dy]^{1/2}$. 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 $\Pi(\mathcal{F}_n^c) \le e^{-(C+4)n\bar{\epsilon}_n^2}$ and $\log N(\bar{\epsilon}_n, \mathcal{F}_n, d_H) \le n\epsilon_n^2$ for all large n, where $\bar{\epsilon}_n = Bn^{-t}(\log n)^s$ for some $B > 0$ and $\epsilon_n = \bar{\epsilon}_n \log n$.

A proof is given in Appendix C. A key step is Lemma 8 (Appendix A) which states $d_H(p_{\theta_1,\psi_1}, p_{\theta_2,\psi_2}) \le c_2\|\omega_1\|_{C^2}^{1/2}\|\theta_1 - \theta_2\| + \|\omega_1 - \omega_2\|_\infty e^{\|\omega_1 - \omega_2\|_\infty/2}$ if θ_1, θ_2 are interior points

in Θ and $\psi_1 = \mathcal{L}(\omega_1)$, $\psi_2 = \mathcal{L}(\omega_2)$ with $\omega_1, \omega_2 \in C^2[0,1]$, the space of twice continuously differentiable functions on $[0,1]$ with norm $\|\omega\|_{C^2} := \|\omega\|_\infty + \|\dot{\omega}\|_\infty + \|\ddot{\omega}\|_\infty$. Clearly, $\mathcal{H}^\beta[0,1] \subset C^2[0,1]$ 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 ω on $[0,1]$ with covariance function c_λ may be viewed as a Borel measurable random element with a Gaussian measure ν^λ 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 C2* with $t = \frac{\beta}{2\beta+1}$, $s = 2t$, in conjunction with the following sharper version of C1:

C1*. $\Pi(\{f : d_{\mathrm{KL}}(f^*, f) \le \bar{\epsilon}_n^2, V(f^*, f) \le \bar{\epsilon}_n^2\}) \ge e^{-Cn\bar{\epsilon}_n^2}$ 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 non-trivial 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 $\underline{\alpha} \le \frac{1}{2}$ and $\overline{\alpha} > 1$. As in Theorem 3, assume that the true density is some $f^* = p_{\theta^*,\psi^*}$ where $\theta^* = (\alpha^*, \sigma^*)$ is in the interior of Θ and $\phi^* = \log \psi^* \in \mathcal{H}^\beta[0,1]$ with $\beta > 2$. Denote $\gamma = \frac{\beta}{2\beta+1} \in (0, 1/2)$ so that the posterior contraction rate in L^1 topology equals a constant multiple of $n^{-\gamma}(\log n)^{2\gamma+1}$. The lower bound assumption on β implies that both $\bar{\rho}(\xi) := \frac{2\xi}{2\xi+1}\gamma - \frac{3(1-2\gamma)}{2\alpha^*(2\xi+1)}$ and $\hat{\rho}(\xi) := \xi\gamma - \frac{3}{2}(1-2\gamma)$ are strictly positive for every $\xi \in [\frac{\underline{\alpha}}{\alpha^*}, 1]$.

Theorem 4. *If* $Y_1, \ldots, Y_n \overset{IID}{\sim} f^*$ *and* $\underline{\alpha}/\alpha^* < \xi < \min(1, 1/\alpha^*)$ *is such that* $\beta\xi > 3/2$ *then* $\operatorname{plim}_{n\to\infty} \Pi(\{f : |\alpha_+(f) - \alpha^*| > B_1 n^{-\rho}(\log n)^s\} \mid Y_1, \ldots, Y_n) = 0$ *for all large* B_1 *where* $\rho = \min\{\bar{\rho}(\xi), \hat{\rho}(\xi)\}$ *and* $s = 2\rho + \frac{4}{\alpha^*(2\xi+1)}$ *if* $\rho = \bar{\rho}(\xi)$, $s = 2\rho + 4$ *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 $|\alpha^* - \alpha| > B_1 n^{-\rho}(\log n)^s$ 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 et al. (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 ξ may be found with $\beta\xi > 3/2$. This condition demands that a relatively rough density (small β) must have a sufficiently heavy tail (small

α^*) 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, i.e., only when the tail is heavy.

Additionally, multiple factors control the value of ρ which determines the posterior contraction rate. Notice that both $\bar{\rho}(\xi)$ and $\hat{\rho}(\xi)$ are strictly increasing in ξ and hence sharpest rates are obtained by taking ξ as close as possible to the maximum allowed value of $\min(1, \frac{1}{\alpha^*})$. Since $\bar{\rho}(\xi) < \hat{\rho}(\xi)$ if and only if $\beta\xi(2\xi - 1) > \frac{3}{2}(2\xi + 1 - 1/\alpha^*)$, the following observations can be made on the fastest possible rate. If $\alpha^* \in (\underline{\alpha}, \frac{2-\underline{\alpha}}{1+\underline{\alpha}}]$ then Theorem 4 holds with any ξ arbitrarily close to $\min(1, \frac{1}{\alpha^*})$, $\rho = \bar{\rho}(\xi)$ and $s = 2\rho + \frac{4}{\alpha^*(2\xi+1)}$. On the other hand, if $\alpha^* \geq 2$, the theorem holds with any ξ arbitrarily close to $\frac{1}{\alpha^*}$, $\rho = \hat{\rho}(\xi)$ and $s = 2\rho + 4$. In the intermediate case of $\alpha^* \in (\frac{2-\underline{\alpha}}{1+\underline{\alpha}}, 2)$, ξ can be arbitrarily close to $\frac{1}{\alpha^*}$ with $\rho = \bar{\rho}(\xi)$, $s = 2\rho + \frac{4}{\alpha^*(2\xi+1)}$ if $\beta > \frac{3\alpha^*}{2} \times \frac{1+\alpha^*}{2-\alpha^*}$ and $\rho = \hat{\rho}(\xi)$ and $s = 2\rho + 4$ otherwise.

When $\alpha^* \in (\underline{\alpha}, \frac{2-\underline{\alpha}}{1+\underline{\alpha}}]$, the choice of $\rho = \bar{\rho}(\xi) = \frac{2\xi}{2\xi+1}\gamma - \frac{3(1-2\gamma)}{2\alpha^*(2\xi+1)}$ compares favorably to the optimal rates obtained by Hall and Welsh (1984, 1985). In particular, whenever ψ^* is infinitely smooth, e.g., f^* is a GPD itself, the density estimation contraction rate has $\gamma \approx \frac{1}{2}$ and hence $\rho \approx \frac{\xi}{2\xi+1}$ with $\xi \approx \min(1, \frac{1}{\alpha^*})$; here $\approx$ indicates "arbitrarily close from below". Since $|y^{\alpha^*}\bar{F}^*(y)/\zeta(f^*) - 1| \asymp y^{-\min(1, \frac{1}{\alpha^*})}$, with $\zeta(f) = \lim_{y\to\infty} y^{\alpha_+(f)}\bar{F}(y)$, f^* belongs to a suitable Hall-Welsh class of heavy tailed densities $\mathcal{D}(\alpha^*, C_0, \epsilon, \xi, A) := \{f : \bar{F}(y) = Cy^{-\alpha}\{1 + R(y)\}, |R(y)| < Ay^{-\xi\alpha}, |\alpha - \alpha^*| < \epsilon, |C - C_0| < \epsilon\}$ for which the minimax rate of tail index estimation is precisely $n^{-\frac{\xi}{2\xi+1}}$. 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 quantities depends on how quickly the actual tail starts resembling the corresponding Pareto tail $y^{-\alpha_+(f)}$. When sample size n is only moderately large, the Pareto shape may only be partially established within the range of the observed data, posing 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, (i) GPD: $f_\alpha(y) = g_{(\alpha,1)}(y)$, (ii) GPD4: $f_\alpha(y) = 4g_{(\alpha,1)}(y)\{G_{(\alpha,1)}(y)\}^3$, and (iii) Half-t: $f_\alpha(y) = 2c(\alpha)(1+y^2/\alpha)^{-(\alpha+2)/2}$, $c(\alpha) = \Gamma(\frac{\alpha+1}{2})/\{\sqrt{\alpha\pi}\Gamma(\frac{\alpha}{2})\}$; each giving a regularly varying density with tail index α. Table 1 reports performance statistics of our semiparametric estimation of the corresponding extreme value index $\xi = \alpha^{-1}$, averaged across 100 data sets of size $n = 1000$ each, with the true value of ξ varying over $\{0.1, 0.2, 0.3, 0.5, 1.0\}$. For comparison, we include corresponding figures from a thresholding estimation of ξ, where the threshold is determined by the adaptive technique of Durrieu et al. (2015), followed by a Bayesian fit of a GPD model to

Model	EVI	Method	Estimating ξ			Estimating $\bar{Q}(p)$ ($\text{rMAE}_{\text{Cover}}$)			
			Bias	RMSE	Cover	$p = 0.01$	0.001	10^{-4}	10^{-5}
GPD	0.1	Semi	0.03	0.05	99	0.06_{93}	0.11_{94}	0.18_{96}	0.26_{98}
		Thresh	0.14	0.16	84	0.06_{95}	0.17_{93}	0.44_{91}	0.90_{89}
	0.2	Semi	0.01	0.06	100	0.07_{97}	0.14_{99}	0.23_{100}	0.34_{100}
		Thresh	0.12	0.15	90	0.07_{96}	0.20_{97}	0.50_{94}	1.03_{92}
	0.3	Semi	0.02	0.06	99	0.08_{94}	0.18_{96}	0.30_{96}	0.46_{97}
		Thresh	0.12	0.17	85	0.08_{96}	0.29_{93}	0.75_{88}	1.59_{86}
	0.5	Semi	0.00	0.09	97	0.13_{93}	0.29_{93}	0.49_{95}	0.76_{96}
		Thresh	0.09	0.15	89	0.14_{91}	0.41_{91}	0.92_{90}	1.81_{91}
	1	Semi	-0.01	0.12	97	0.23_{95}	0.49_{97}	0.85_{97}	1.45_{97}
		Thresh	0.04	0.15	89	0.25_{94}	0.62_{92}	1.23_{91}	2.33_{91}
GPD4	0.1	Semi	0.03	0.09	97	0.05_{95}	0.12_{91}	0.25_{90}	0.48_{91}
		Thresh	0.13	0.15	83	0.04_{100}	0.15_{95}	0.40_{93}	0.80_{88}
	0.2	Semi	0.01	0.06	98	0.06_{95}	0.13_{95}	0.22_{96}	0.34_{98}
		Thresh	0.08	0.11	92	0.06_{96}	0.17_{97}	0.38_{95}	0.71_{91}
	0.3	Semi	-0.00	0.07	96	0.08_{97}	0.17_{94}	0.29_{95}	0.43_{96}
		Thresh	0.07	0.12	93	0.08_{95}	0.23_{97}	0.51_{94}	0.97_{91}
	0.5	Semi	0.02	0.08	92	0.13_{94}	0.30_{93}	0.53_{91}	0.83_{92}
		Thresh	0.06	0.13	93	0.15_{93}	0.39_{95}	0.79_{93}	1.42_{92}
	1	Semi	0.05	0.10	96	0.17_{97}	0.40_{96}	0.76_{96}	1.36_{96}
		Thresh	0.04	0.12	97	0.19_{94}	0.44_{96}	0.82_{96}	1.38_{96}
Half-t	0.1	Semi	-0.06	0.06	87	0.06_{79}	0.12_{72}	0.16_{83}	0.18_{91}
		Thresh	0.09	0.11	96	0.04_{92}	0.13_{90}	0.31_{86}	0.59_{87}
	0.2	Semi	-0.10	0.11	56	0.06_{92}	0.11_{94}	0.17_{97}	0.27_{95}
		Thresh	0.06	0.11	94	0.06_{91}	0.16_{94}	0.37_{95}	0.71_{96}
	0.3	Semi	-0.12	0.13	75	0.07_{96}	0.13_{97}	0.23_{96}	0.35_{90}
		Thresh	0.04	0.11	99	0.06_{98}	0.20_{96}	0.44_{96}	0.78_{97}
	0.5	Semi	-0.10	0.13	87	0.10_{95}	0.23_{96}	0.39_{96}	0.54_{91}
		Thresh	0.02	0.10	95	0.10_{96}	0.26_{96}	0.51_{96}	0.85_{95}
	1	Semi	-0.06	0.14	96	0.19_{96}	0.36_{96}	0.57_{96}	0.82_{96}
		Thresh	0.02	0.11	100	0.21_{95}	0.45_{98}	0.76_{99}	1.19_{99}

Table 1: Estimating the extreme value index (EVI) $\xi = \alpha^{-1}$ and high tail quantiles $\bar{Q}(p) = \bar{F}^{-1}(p)$ from synthetic data of sample size $n = 1000$. Estimation accuracy and coverage of 95% credible intervals are averaged across 100 data sets for each experimental group. For $\bar{Q}(p)$, estimation accuracy is measured via relative mean absolute error as a fraction of the true quantile value. Additional keys: RMSE = root mean squared error, Cover = coverage.

the excess data with the GPD location parameter set at the threshold, and the scale σ 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 with higher coverage than those from the thresholding method (figure included in supplementary material). This improvement is unsurprising; the true density f_α matches the model specification in a very strong way. A similar match between the model and the truth is absent in the GPD4 sets for which f_α may be expressed as $p_{(\alpha,\sigma),\psi}$ but only with a $\psi = \mathcal{L}(\omega)$ for which $\lim_{u\to 0}\omega(u) = -\infty$. But this misspecification at the left does not appear to affect estimation of the right tail, where the semiparametric method performs as well or better than the thresholding method, especially when the tail is not too heavy.

The Half-t sets pose a far more serious challenge to the semiparametric method. Al-

though the averaged error of the estimates are comparable between the two methods, 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 $\dot\phi(1-t) = \frac{(\alpha+1)t^{\xi-1}}{\alpha} \times \frac{(1+\alpha\sigma^2)t^\xi - \alpha\sigma^2}{t^{2\xi}+\alpha\sigma^2(1-t^\xi)^2}$, and hence $\lim_{u\to 1}\dot\phi(u)$ 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 ξ 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 $\bar Q(p) = \bar F^{-1}(p) = F^{-1}(1-p)$ associated with excess tail probabilities $p \in \{0.01, 0.001, 0.0001, 0.00001\}$ and compare these against the true values for the 3×5 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 ξ, for which the semiparametric model is strongly misspecified at the right tail. However, a closer inspection of these cases reveals that while the semiparametric method overestimates the high quantiles, it still gives a credible interval that is comparable in magnitude to the true quantile value. In contrast, the thresholding method may minimally contain the true value at the lower end of its interval but usually produces a very wide interval with the upper end of the interval being several orders of magnitudes larger than the truth (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 et al. (2002) present an analysis of total daily precipitation measurements (in inches) between 1900-1999 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.

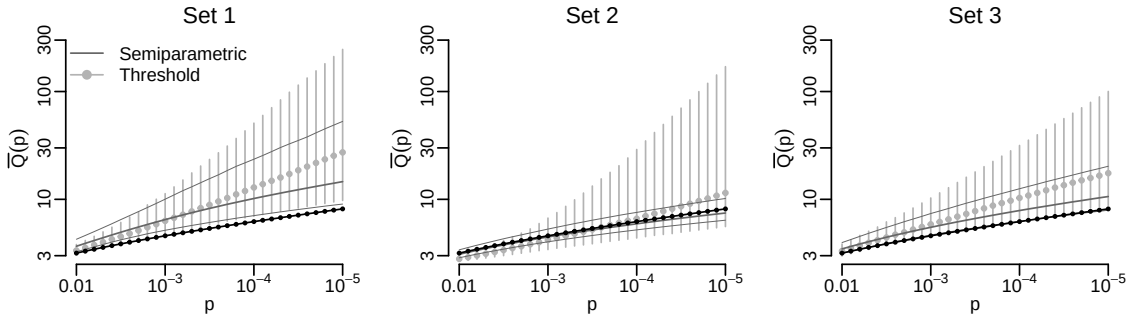

Figure 2: A comparison of the 95% posterior credible intervals for $\bar{Q}(p) = \bar{F}^{-1}(p)$ from the semiparametric and the thresholding methods, for three randomly chosen Half-t sets with $\xi = \alpha^{-1} = 0.1$, for which the semiparametric method serisouly underestimates ξ. True quantile values are shown as connected black beads.

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 data set[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% 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 ≥ 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 ξ 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 et al. (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 ξ 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 ξ 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

[1]Taken from the `extRemes` package in `R` (Gilleland and Katz, 2011).

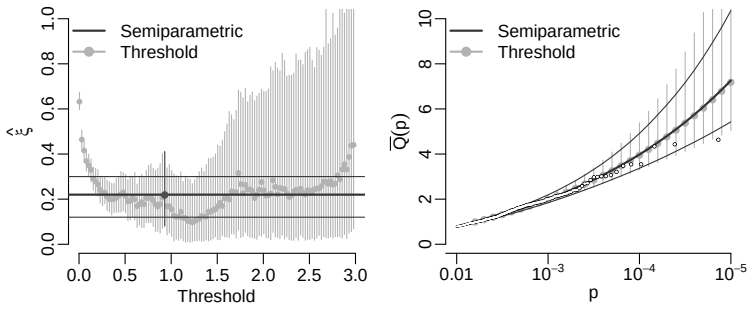

Figure 3: Estimation of tail heaviness ($\xi = \alpha^{-1}$) and high tail quantiles for Fort Collins daily precipitation. Left panel shows thresholding estimates and 95% credible intervals of ξ corresponding to a grid of threshold values between 0.005 and 3.0 with an increment of 0.025; the adaptive 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 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 $\{(\frac{i-0.5}{N}, Y_{(i)}), 1 \leq i \leq N\}$ is included to visualize empirical quantiles, where $Y_{(i)}$ denotes the i-th order statistic of the original data.

was reported by Katz et al. (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 et al. (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 28 years (95% PCI = [14.9, 59.9]). We note that in the 100 year observation period, there were 10 instances with 3 inches or more daily precipitation (1902, '04, '38, '49, '51, '51, '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 ≥ 0.1 inches (or ≥ 0.4 inches), the estimate of ξ 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

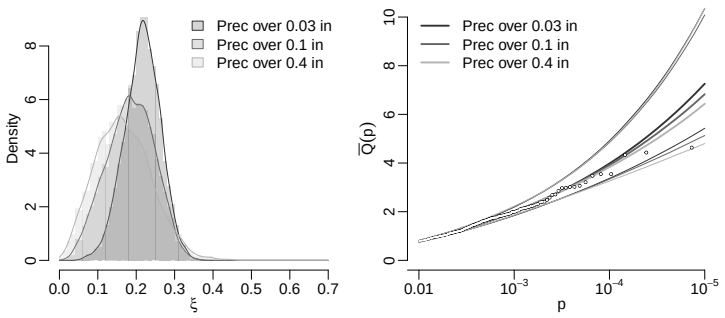

Figure 4: Semiparametric estimation of tail heaviness ($\xi = \alpha^{-1}$) and high tail quantiles under further left truncation of the Fort Collins precipitation data. Posterior distribution of ξ (left) widens and moves to the left slightly but maintaining overlap. High tail quantile estimates (right)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

was presented in Katz et al. (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 non-zero 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 $\mathcal{D}(\alpha, C_0, \xi, \epsilon, A)$ with $\xi = \min(1, \frac{1}{\alpha})$, albeit Hall and Welsh (1984, 1985) make stringent assumptions on other quantities such as $\zeta(f)$ (Carpentier 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 et al. (2009). It could also be feasible to sharpen our posterior contraction rate in Theorem 4 by utilizing 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 zeros 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, e.g., 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 $\zeta = \Psi^{-1}$ 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 $Q(p) = Q_\theta(\zeta(p)) = \int_0^{\zeta(p)} q_\theta(u)du$ where $Q_\theta = G_\theta^{-1}$ and $q_\theta = \dot{Q}_\theta$. To accommodate a predictor vector $x \in \mathbb{R}^d$, consider a quantile regression formulation $Q(p|x) = \int_0^{\zeta(p)} q_\theta(u)\{1 + x^\top h(\omega(u))\}du$ where $\omega : u \mapsto (\omega_1(u), \ldots, \omega_d(u))^\top \in \mathbb{R}^d$ is unknown and $h(b)$ is a suitably chosen, fixed transformation that ensures $1 + x^\top h(b) \geq 0$ 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 (θ, ζ, ω) by adopting a hierarchical LGP prior on the quantile density $\dot{\zeta}$ and smooth Gaussian process priors on $\omega_1, \ldots, \omega_d$. The theoretical analysis presented in the current paper is likely to yield a sharper understanding of asymptotic properties of the method by Yang and Tokdar (2017), especially with respect to tail estimation.

Appendix

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.

A Auxiliary results for density estimation

Mixed partial derivatives of functions $\ell(\theta)$, $\theta = (\alpha, \sigma)$, are denoted by $D^k\ell := \frac{\partial^{|k|}\ell(\alpha,\sigma)}{\partial\alpha^{k_1}\partial\sigma^{k_2}}$ for any bi-index $k = (k_1, k_2) \in \{0, 1, 2, \ldots\}^2$ of order $|k| := k_1 + k_2$. Below $\Theta = [\underline{\alpha}, \overline{\alpha}] \times [\underline{\sigma}, \overline{\sigma}]$

with $0 < \underline{\alpha} < \overline{\alpha} < \infty$, $0 < \underline{\sigma} < \overline{\sigma} < \infty$ and any constants appearing in statements and proofs may implicitly depend on the boundary values of Θ. Let $\mathring{\Theta}$ denote the interior of Θ.

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

Lemma 6. *Fix a density $\psi = \mathcal{L}(\omega)$ with $\omega \in C^2[0,1]$. If $\theta_1, \theta_2 \in \mathring{\Theta}$ then*

1. *$d_{\mathrm{KL}}(p_{\theta_1,\psi}, p_{\theta_2,\psi}) \leq c_2\|\omega_1\|_{C^2}\|\theta_1 - \theta_2\|^2$.*

Moreover, there exist positive numbers c_3, t_0 such that if $\|\theta_1 - \theta_2\| < t_0$ then

2. *$\int p_{\theta_2,\psi}(y)(\frac{p_{\theta_1,\psi}(y)}{p_{\theta_2,\psi}(y)} - 1)^2 dy \leq c_3 e^{3t_0 c_0\|\omega\|_{C^2}}\|\theta_1 - \theta_2\|^2$, and*

3. *$V(p_{\theta_1,\psi}, p_{\theta_2,\psi}) \leq c_3 e^{3t_0 c_0\|\omega\|_{C^2}}\|\theta_1 - \theta_2\|^2$.*

Lemma 7. *Fix $\theta_1 \in \mathring{\Theta}$, $\psi_1 = \mathcal{L}(\omega_1)$ with $\omega_1 \in C^2[0,1]$ and $\epsilon \in (0, t_0)$. There exists a constant K depending on $\|\omega_1\|_{C^2}$ such that $d_{\mathrm{KL}}(p_{\theta_1,\psi_1}, p_{\theta_2,\psi_2}) \leq K\epsilon^2, V(p_{\theta_1,\psi_1}, p_{\theta_2,\psi_2}) \leq K\epsilon^2$, for every $\theta_2 \in \mathring{\Theta}$ with $\|\theta_1 - \theta_2\| \leq \epsilon$ and every $\psi_2 = \mathcal{L}(\omega_2)$ with $\|\omega_1 - \omega_2\|_\infty < \epsilon$.*

Lemma 8. *If $\theta_1, \theta_2 \in \mathring{\Theta}$ and $\psi_1 = \mathcal{L}(\omega_1)$, $\psi_2 = \mathcal{L}(\omega_2)$ with $\omega_1, \omega_2 \in C^2[0,1]$ then $d_H(p_{\theta_1,\psi_1}, p_{\theta_2,\psi_2}) \leq c_2\|\omega_1\|_{C^2}^{1/2}\|\theta_1 - \theta_2\| + \|\omega_1 - \omega_2\|_\infty e^{\|\omega_1-\omega_2\|_\infty/2}$.*

B Proof of Condition C1*

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

C Proof of Condition C2*

Suppose there exist sets $\mathcal{B}_1, \mathcal{B}_2, \ldots \subset C^2[0,1]$ such that for all large n,

$$\bar{\nu}(\mathcal{B}_n^c) \leq e^{-(C+4)n\bar{\epsilon}_n^2} \tag{6}$$

$$\log N(\bar{\epsilon}_n, \mathcal{B}_n, \|\cdot\|_{C^2}) \leq n\epsilon_n^2/2 \tag{7}$$

$$\sup\{\|\omega\|_{C^2} : \omega \in \mathcal{B}_n\} \leq (n\bar{\epsilon}_n)^b, \tag{8}$$

for some $b \geq 1$. Then, $\mathcal{F}_1, \mathcal{F}_2, \ldots$ could be simply constructed as $\mathcal{F}_n = \{p_{\theta,\psi} : \theta \in \Theta, \psi = \mathcal{L}(\omega), \omega \in \mathcal{B}_n\}$. To see that these sets satisfy the requirements of C2*, notice that $\Pi(\mathcal{F}_n^c) = \bar{\nu}(\mathcal{B}_n^c) \leq e^{-(C+4)n\bar{\epsilon}_n^2}$, and, by Lemma 8, $\log N(\bar{\epsilon}_n, \mathcal{F}_n, d_H) \leq \log N(\bar{\epsilon}_n/(c_2(n\bar{\epsilon}_n)^{b/2}), \Theta, \|\cdot\|) + \log N(\bar{\epsilon}_n, \mathcal{B}_n, \|\cdot\|_{C^2}) \leq \log(\mathrm{diam}(\Theta)\cdot n^b) + n\epsilon_n^2/2 \leq n\epsilon_n^2$ 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 $\mathcal{B}_n$ for $(C[0,1],\|\cdot\|_\infty)$ by VZ09 to $(C^2[0,1],\|\cdot\|_{C^2})$. Our adaptation also produces a smaller exponent b in (8) than what is possible with the original construction of VZ09. Although a smaller 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_λ has infinitely differentiable sample paths with probability one. Therefore the Gaussian measures ν^λ introduced in the preceding section could also be viewed as probability measures with respect to the refined Borel σ-algebra of $(C^2[0,1],\|\cdot\|_{C^2})$. A more formal treatment is outlined below. Hereafter, $\Re(z)$ and $\Im(z)$ denote the real and imaginary parts of a complex number z.

VZ09 show that the reproducing kernel Hilbert space $\mathbb{H}^\lambda$ associated with c_λ consists of functions $h(u)=\Re(\int e^{ut\sqrt{-1}}\eta(t)\mu_\lambda(t)dt)$ with $\|h\|_{\mathbb{H}^\lambda}=\|\eta\|_{L_2(\mu_\lambda)}$, where $\mu_\lambda(t)=e^{-t^2/4\lambda^2}/(2\lambda\sqrt{\pi})$ is the spectral density associated with c_λ. By applying Cauchy-Schwarz inequality, with differentiations under integration as needed, it follows that

$$\|h\|_\infty\le\|h\|_{\mathbb{H}^\lambda},\|\dot h\|_\infty\le\sqrt{2}\lambda\|h\|_{\mathbb{H}^\lambda},\text{ and }\|\ddot h\|_\infty\le\sqrt{12}\lambda^2\|h\|_{\mathbb{H}^\lambda}. \tag{9}$$

Clearly, $\|h\|_{C^2}\le(1+\sqrt{2}\lambda+\sqrt{12}\lambda^2)\|h\|_{\mathbb{H}}$ and $\mathbb{H}^\lambda$ can be continuously and densely embedded within the Banach space $(C^2[0,1],\|\cdot\|_{C^2})$, guaranteeing a Borel measure ν^λ on the embedding Banach space matching the law of a centered Gaussian process with covariance c_λ. As before, define $\bar\nu(\cdot)=\int\nu^\lambda(\cdot)\pi_\lambda(\lambda)d\lambda$.

Let $\mathbb{H}^\lambda_1$ and $\mathcal{B}_1$ denote the unit balls of $\mathbb{H}^\lambda$ and $C^2[0,1]$. Recall, $\bar\epsilon_n=Bn^{-t}(\log n)^s$, $\epsilon_n=\bar\epsilon_n\log n$ where we are free to choose $B>0$. To start off, take B large enough such that $r_n=n\bar\epsilon_n^2>1$ for all n. Let m_n be the smallest integer larger than $\log_2(r_n)$. Define

$$\mathcal{B}_n=[\{\cup_{j=1}^{m_n}(\sqrt{2}M_n\mathbb{H}_1^{2^j})\}\cup\{M_n\delta_n^{-1/2}\mathbb{H}_1^1\}\cup\{\cup_{\lambda<\delta_n}(M_n\mathbb{H}_1^\lambda)\}]+\bar\epsilon_n\mathbb{B}_1 \tag{10}$$

where $M_n=16Cr_n^{1/2}\log(r_n/\bar\epsilon_n)$ with C taken from Lemma 10 below, and $\delta_n=\bar\epsilon_n/(4M_n)$. Because of (14), for all large n, $\mathcal{B}_n\subset 5r_n^2M_n\mathbb{B}_1$ and hence $\mathcal{B}_n$ satisfies (8) with $b=2.5$. By Lemma 4.7 of VZ09, $\mathcal{B}_n\supset M_n\mathbb{H}^\lambda+\bar\epsilon_n\mathbb{B}_1$ for every $0<\lambda\le r_n$. Borell's inequality implies that $\nu^\lambda(\mathcal{B}_n^c)\le 1-\Phi(\Phi^{-1}(\nu^\lambda(\bar\epsilon_n\mathbb{B}_n))+M_n)\le 1-\Phi(\Phi^{-1}(\nu^{r_n}(\bar\epsilon_n\mathbb{B}_1))+M_n)$ where the second inequality follows since $\nu^\lambda(\epsilon\mathbb{B}_1)$ is decreasing in λ for every $\epsilon>0$ (Lemma 9 below). As $\nu^{r_n}(\bar\epsilon_n\mathbb{B}_1)\le\nu^1(\bar\epsilon_n\mathbb{B}_n)<1/4$ and $M_n\ge 4\sqrt{\log(1/\nu^{r_n}(\bar\epsilon_n\mathbb{B}_1))}$ for all large n (Lemma 10 below), it must be that $\nu^\lambda(\mathcal{B}_n^c)\le 1-\Phi(M_n/2)\le e^{-M_n^2/8}\le e^{-r_n}$ for every $\lambda\in(0,r_n)$, for all large n. This establishes (6), with B chosen suitably large, since $\pi_\lambda((r_n,\infty))\le e^{-C_3r_n}$ for all large n for some constant C_3.

To establish (7), first note that every $h\in\cup_{\lambda<\delta_n}(M_n\mathbb{H}_1^\lambda)$ satisfies $\|h-h(0)\|_{C^2}\le\bar\epsilon_n$ by (14), i.e., as an element of $C^2[0,1]$, the function $h(u)$ is within $\bar\epsilon_n$ distance of a constant function whose constant value ranges within $[-M_n,M_n]$. Clearly, $\log N(2\bar\epsilon_n,\cup_{\lambda<\delta_n}(M_n\mathbb{H}_1^\lambda)+\bar\epsilon_n\mathbb{B}_1,\|\cdot\|_{C^2})\le\log\frac{2M_n}{\bar\epsilon_n}$. Next, by Lemma 10 below, $\log N(2\bar\epsilon_n,M_n\delta_n^{-1/2}\mathbb{H}_1^1+\bar\epsilon_n\mathbb{B}_1,\|\cdot\|_{C^2})\le C\log^2(\frac{M_n}{\bar\epsilon_n}\delta_n^{-1/2})$ and $\log N(2\bar\epsilon_n,\sqrt{2}M_n\mathbb{H}_1^{2^j}+\bar\epsilon_n\mathbb{B}_1,\|\cdot\|_{C^2})\le C2^j\log^2(\frac{2^{j+1/2}M_n}{\bar\epsilon_n})\le 2Cr_n\log^2(\frac{r_nM_n}{\bar\epsilon_n})$ for each $1\le j\le m_n$ by the monotonicity of $\log x$. Consequently,

$$\log N(2\bar\epsilon_n,\cup_{j=1}^{m_n}(\sqrt{2}M_n\mathbb{H}_1^{2^j})+\bar\epsilon_n\mathbb{B}_1,\|\cdot\|_{C^2})\le\log(m_n)+2Cr_n(\log\tfrac{r_nM_n}{\bar\epsilon_n})^2,$$

concluding the proof of Condition C2*. Two auxiliary results used in the above prove are:

Lemma 9. *For any fixed $\epsilon > 0$, the small ball probability $\nu^\lambda(\epsilon\mathbb{B}_1)$ is decreasing in $\lambda > 0$.*

Lemma 10. *There exist C, ϵ_0 such that for all $\lambda \geq 1$ and all $\epsilon < \epsilon_0$, (a) $\log N(\epsilon, \mathbb{H}_1^\lambda, \|\cdot\|_{C^2}) \leq C\lambda \log^2(\lambda/\epsilon)$, and (b) $-\log \nu^\lambda(\epsilon\mathbb{B}_1) \leq C\lambda \log^2(\lambda/\epsilon)$.*

D Auxiliary results for tail estimation

If f is heavy tailed then $\lim_{y\to\infty} |y^{\alpha_+(f)}\bar{F}(y)/\zeta(f) - 1| = 0$. For our semiparametric analysis it is useful to consider classes of heavy tailed densities for which this convergence holds uniformly. Define $\mathcal{T}(t,\delta) = \{f : \sup_{y\geq t} |y^{\alpha_+(f)}\bar{F}(y)/\zeta(f) - 1| \leq \delta\}$, for any arbitrary $t > 0, \delta > 0$. For any $f \in \mathcal{F}$, let $\mathbb{P}_f^n$ denote the joint probability law of $(Y_1, \ldots, Y_n)$ with $Y_i \sim f$ independently of one another and $\mathbb{P}_f^n h$ denote expectation of $h(Y_1, \ldots, Y_n)$ under $\mathbb{P}_f^n$. For the following lemma, let f^* denote an arbitrary heavy tailed density with $\alpha^* = \alpha_+(f^*) \in (\underline{\alpha}, \overline{\alpha})$ and let $\epsilon_n \to 0$ be an arbitrary positive sequence satisfying $n\epsilon_n^2 \to \infty$. By a test function we mean any statistic that takes values in $[0,1]$.

Lemma 11. *Suppose there exist positive sequences $t_n \to \infty$, $\delta_n \to 0$ such that $f^* \in \mathcal{T}(t_n, \delta_n)$ and $\min\{\bar{F}^*(t_n), \bar{F}^*(t_n)^{1/2}\delta_n\} \geq 3\epsilon_n$ for all large n. Then there exist test functions $T_n = T_n(Y_1, \ldots, Y_n)$ satisfying $\mathbb{P}_{f^*}^n T_n \leq 4e^{-n\epsilon_n^2}$ and $\sup\{\mathbb{P}_f^n(1 - T_n) : f \in \mathcal{T}(t_n, \delta_n), \alpha_+(f) < \overline{\alpha}, |\alpha_+(f) - \alpha^*| > 2^{4+\overline{\alpha}}\delta_n\} \leq 4e^{-n\epsilon_n^2}$ for all large n.*

Lemma 12. *Suppose $\tau_n \to 0, D_n \to \infty$ are positive sequences and $t_n = (D_n/\tau_n)^{1/\min(1,A)}$ for some $A > 0$. Then, with $B_1 > 0$ chosen sufficiently large, $\{f = p_{\theta,\psi} : \theta = (\alpha, \sigma) \in \Theta, \alpha \geq A, \psi = \mathcal{L}(\omega), \|\dot{\omega}\|_\infty \leq D_n\} \subset \mathcal{T}(t_n, B_1\tau_n)$ for all large n.*

E Proof of Theorem 4

Our argument is based on the proof of Theorem 8.9 in Ghosal and van der Vaart (2017). Consider again the sets $\mathcal{F}_n = \{p_{\theta,\psi} : \theta \in \Theta, \psi = \mathcal{L}(\omega), \omega \in \mathcal{B}_n\}$ from the proof of Condition C2* where $\mathcal{B}_n$ is as in (10) with $\bar{\epsilon}_n = B\{(\log n)^2/n\}^{\frac{\beta}{2\beta+1}}$ for some large B. Recall that $\Pi(\mathcal{F}_n^c) \leq e^{-(C+4)n\bar{\epsilon}^2}$ for some constant C. Define

$$\mathcal{U}_n = \{p_{(\theta,\sigma),\psi} : \theta = (\alpha, \sigma) \in \Theta, |\alpha - \alpha^*| > B_1 n^{-\rho}(\log n)^s, \psi = \mathcal{L}(\omega), \omega \in C^2[0,1]\}.$$

It follows from Bayes' formula for $\Pi(\mathcal{U}_n \mid Y_1, \ldots, Y_n)$ that with $A_n := \{(y_1, \ldots, y_n) : \int_{\mathcal{F}} \prod_{i=1}^n \frac{f(y_i)}{f^*(y_i)}\Pi(df) \geq e^{-(2+C)n\bar{\epsilon}_n^2}\}$ and for any test function $T_n : \mathbb{R}^n \to [0,1]$,

$$\mathbb{P}_{f^*}^n \Pi(\mathcal{U}_n \mid Y_1, \ldots, Y_n) \leq \mathbb{P}_{f^*}^n T_n + \mathbb{P}_{f^*}^n(A_n^c) + e^{(2+C)n\bar{\epsilon}_n^2}\left[\sup_{f\in\mathcal{F}_n\cap\mathcal{U}_n} \mathbb{P}_f^n(1 - T_n) + \Pi(\mathcal{F}_n^c)\right]$$

Now, $\lim_{n\to\infty} e^{(2+C)n\bar{\epsilon}_n^2}\Pi(\mathcal{F}_n^c) = 0$ by construction and $\lim_{n\to\infty} \mathbb{P}_{f^*}^n(A_n^c) = 0$ by Lemma 8.10 of Ghosal and van der Vaart (2017). Therefore the proof of the theorem is complete once we have shown the existence of test functions $(T_n : n \geq 1)$ satisfying

$$\lim_{n\to\infty} \mathbb{P}_{f^*}^n T_n = 0, \qquad \sup_{f\in\mathcal{F}_n\cap\mathcal{U}_n} \mathbb{P}_f^n(1 - T_n) \leq e^{-(4+C)n\bar{\epsilon}_n^2} \text{ for all large } n. \tag{11}$$

We shall construct such a test function based on Lemmas 11 and 12.

Take $\epsilon_n = (4+C)^{1/2}\bar{\epsilon}_n$. For any $f = p_{\theta,\psi} \in \mathcal{F}_n$ it follows from (14) that if $\phi = \log\psi$ then $\|\dot{\phi}\|_\infty \le D_n := C_1 r_n^{3/2}\log n = C_1 n^{\frac{3}{2}(1-2\gamma)}(\log n)^{6\gamma+1}$ for some constant C_1. Set $\alpha_1 = \xi\alpha^*$ and note that $\underline{\alpha} < \alpha_1 < \min(1,\alpha^*)$ and partition $\mathcal{U}_n = \mathcal{U}_{1n} \cup \mathcal{U}_{2n}$ where $\mathcal{U}_{1n} = \mathcal{U}_n \cap \{f : \underline{\alpha} \le \alpha_+(f) < \alpha_1\}$ and $\mathcal{U}_{2n} = \mathcal{U}_n \cap \{f : \alpha_1 \le \alpha_+(f) \le \overline{\alpha}\}$. By Lemma 12 (with $A = \underline{\alpha}$), for any $\rho_1, s_1 > 0$,

$$\mathcal{F}_n \cap \mathcal{U}_{1n} \subset \mathcal{T}(t_{1n},\delta_{1n}) \cap \{f : \alpha_+(f) \le \overline{\alpha}, |\alpha_+(f) - \alpha^*| > 2^{4+\overline{\alpha}}\delta_{1n}\} \text{ for all large } n, \quad (12)$$

where $\delta_{1n} = B_{12}\underline{\tau}_n$, $\underline{\tau}_n = C_{12}n^{-\rho_1}(\log n)^{s_1}$, $t_{1n} = (D_n/\underline{\tau}_n)^{1/\underline{\alpha}}$ and B_{12}, C_{12} are large constants to be adjusted. We next show that $\rho_1, s_1 > 0$ could be chosen so that

$$\min\{\bar{F}^*(t_{1n}), \delta_{1n}\bar{F}^*(t_{1n})^{1/2}\} \ge 3\epsilon_n \text{ for all large } n. \quad (13)$$

Indeed, $\bar{F}^*(t_{1n}) \ge \frac{1}{2}\zeta(f^*)t_{1n}^{-\alpha^*} = \frac{1}{2}\zeta(f^*)(\frac{C_{12}}{C_1})^{1/\underline{\xi}} \times n^{-\{\rho_1+\frac{3}{2}(1-2\gamma)\}/\underline{\xi}}(\log n)^{(s_1-6\gamma-1)/\underline{\xi}}$ for all large n, where $\underline{\xi} = \underline{\alpha}/\alpha^* \in (0,1)$. Therefore, with a suitably large choice of C_{12} we can make $\bar{F}^*(t_{1n}) \ge 3\epsilon_n$ for all large n provided $\rho_1 \le \hat{\rho}(\underline{\xi})$, and in case of an equality, $s_1 = 2\rho_1 + 4$. On the other hand, in order to have $\delta_{1n}\bar{F}^*(t_n)^{1/2} \ge 3\epsilon_n$, we need to choose B_{12} suitably large and $\rho_1 \le \bar{\rho}(\underline{\xi})$, and in case of an equality, $s_1 = 2\rho_1 + \frac{4}{\alpha^*(2\underline{\xi}+1)}$. With (12)-(13) established with $\rho_1 > 0$ chosen as the minimum of the above two bounds and $s_1 > 0$ set accordingly, apply Lemma 11 to conclude that there exist test functions $T_{1n} = T_{1n}(Y_1,\ldots,Y_n)$ such that $\mathbb{P}^n_{f^*}T_{1n} \le e^{-n\epsilon_n^2}$ and $\sup\{\mathbb{P}^n_f(1-T_{1n}) : f \in \mathcal{F}_n \cap \mathcal{U}_{1n}\} \le e^{-n\epsilon_n^2}$ for all large n.

Next we repeat the same arguments for testing $f = f^*$ versus $f \in \mathcal{F}_n \cap \mathcal{U}_{2n}$. Rewrite the target rate as $B_1 n^{-\rho}(\log n)^s = 2^{4+\overline{\alpha}}\delta_n$ where $\delta_n = B_{22}\tau_n$, $\tau_n = C_{22}n^{-\rho}(\log n)^s$, and $t_n = (D_n/\tau_n)^{1/\alpha_1}$, with B_{22}, C_{22} to be adjusted as needed. As argued in the preceding paragraph, the choices of ρ, s imply that $\min(\bar{F}^*(t_n), \delta_n\bar{F}^*(t_n)^{1/2}) \ge 3\epsilon_n$ and Lemma 12 (with $A = \alpha_1$) implies that $\mathcal{F}_n \cap \mathcal{U}_{2n} \subset \mathcal{T}(t_n,\delta_n) \cap \{f : \alpha_+(f) \le \overline{\alpha}, |\alpha_+(f) - \alpha^*| > \delta_n\}$. Therefore, by Lemma 11, there are test functions $T_{2n} = T_{2n}(Y_1,\ldots,Y_n)$ such that $\mathbb{P}^n_{f^*}T_{2n} \le e^{-n\epsilon_n^2}$ and $\sup\{\mathbb{P}^n_f(1-T_{2n}) : f \in \mathcal{F}_n \cap \mathcal{U}_{2n}\} \le e^{-n\epsilon_n^2}$ for all large n. The proof is now complete by taking $T_n = \max(T_{1n}, T_{2n})$.

SUPPLEMENTARY MATERIAL

Proofs of auxiliary results

Proof of Lemma 5. Clearly $\phi = \log\psi \in C^2[0,1]$ with $\dot\phi = \dot\omega$, $\ddot\phi = \ddot\omega$. Let ∇_θ and ∇_θ^2 denote the first and second order vector differential operators with respect to $\theta = (\alpha, \sigma)$. Then,

$$\nabla_\theta \log q_\theta(y) = \nabla_\theta \log g_\theta(y) + \dot\phi(G_\theta(y))\nabla_\theta G_\theta(y)$$
$$\nabla_\theta^2 \log q_\theta(y) = \nabla_\theta^2 \log g_\theta(y) + \dot\phi(G_\theta(y))\nabla_\theta^2 G_\theta(y) + \ddot\phi(G_\theta(y))\nabla_\theta G_\theta(y)\nabla_\theta G_\theta(y)^\top$$

which immediately proves the result because $\frac{\partial}{\partial\alpha}\log g_\theta(y)$ is bounded by a shifted and scaled version of $\log(1+y)$, and $\frac{\partial}{\partial\sigma}\log g_\theta(y)$ as well as every term in $\nabla_\theta^2 \log g_\theta(y)$, $\nabla_\theta G_\theta(y)$ and $\nabla_\theta^2 G_\theta(y)$ is uniformly bounded over $y \geq 0$ and $\theta \in \Theta$. For completeness we list below the first and second order partial derivatives of $\log g_\theta(y)$ and $G_\theta(y)$; expressed in terms of $z = (1 + \frac{y}{\alpha\sigma})^{-1} \in (0,1]$,

$\frac{\partial}{\partial\alpha}\log g_\theta(y) = \log z + \frac{1-z}{\alpha}$, $\frac{\partial}{\partial\sigma}\log g_\theta(y) = \frac{\alpha-(\alpha+1)z}{\sigma}$,
$\frac{\partial^2}{\partial\alpha^2}\log g_\theta(y) = \frac{(1-z)\{\alpha-1-(\alpha+1)z\}}{\alpha^2}$, $\frac{\partial^2}{\partial\sigma^2}\log g_\theta(y) = \frac{(\alpha+1)z^2-\alpha}{\sigma^2}$, $\frac{\partial^2}{\partial\alpha\partial\sigma}\log g_\theta(y) = \frac{\{\alpha-(\alpha+1)z\}(1-z)}{\alpha\sigma}$
$\frac{\partial G_\theta(y)}{\partial\alpha} = (\log z + 1 - z)z^\alpha$, $\frac{\partial^2 G_\theta(y)}{\partial\alpha^2} = \{(\log z + 1 - z)^2 + \frac{(1-z)^2}{\alpha}\}z^\alpha$
$\frac{\partial G_\theta(y)}{\partial\sigma} = \frac{\alpha(1-z)}{\sigma}z^\alpha$, $\frac{\partial^2 G_\theta(y)}{\partial\sigma^2} = \frac{\alpha(1-z)\{\alpha-1-z(\alpha+1)\}}{\sigma^2}z^\alpha$, $\frac{\partial^2 G_\theta(y)}{\partial\alpha\partial\sigma} = \frac{(1-z)\{\alpha(\log z+1-z)+1-z\}}{\sigma}z^\alpha$.

□

Proof of Lemma 6. Denote $q_\theta = p_{\theta,\psi}$, $\theta \in \Theta$. By Taylor's theorem, for $\theta, \theta+u$ in the interior of Θ,

$$\log\frac{q_{\theta+u}(y)}{q_\theta(y)} = R_1(\theta,u,y) = u^\top\nabla_\theta \log q_\theta(y) + R_2(\theta,u,y)$$

with $|R_j(\theta,u,y)| \leq \|u\|^j \max_{|k|=j}\sup_{\theta\in\Theta}|D^k \log q_\theta(y)|$, $j = 1,2$. The first claim now follows because $d_{\mathrm{KL}}(q_\theta, q_{\theta+u}) = \int q_\theta(y)\log\frac{q_\theta(y)}{q_{\theta+u}(y)}dy = 0 + \int R_2(\theta,u,y)q_\theta(y)dy \leq c_2\|\omega\|_{C^2}\|u\|^2$ by Lemma 5. Next, use the inequality $|e^x - 1| \leq |x|e^{|x|}$ to conclude

$$\left|\frac{q_{\theta+u}(y)}{q_\theta(y)} - 1\right| \leq |R_1(\theta,u,y)|e^{|R_1(\theta,u,y)|} \leq \|u\|\{c_0\|\omega\|_{C^2} + c_1\log(1+y)\}e^{\|u\|\{c_0\|\omega\|_{C^2}+c_1\log(1+y)\}}$$

by Lemma 5. Therefore, $\int q_\theta(y)(\frac{q_{\theta+u}(y)}{q_\theta(y)} - 1)^2 dy \leq c_4\|u\|^2$ where

$$c_4 = \sup_{\theta\in\Theta}\int\{c_0\|\omega\|_{C^2} + c_1\log(1+y)\}^2 e^{2t_0\{c_0\|\omega\|_{C^2}+c_1\log(1+y)\}}q_\theta(y)dy \leq c_3 e^{3t_0 c_0\|\omega\|_{C^2}}$$

with $c_3 := t_0^{-2}\sup_{\theta\in\Theta}\int(1+y)^{3t_0c_1}q_\theta(y)dy$ a finite number if $t_0 < \underline{\alpha}/(3c_0)$. This proves the second claim as well as the third claim since $V(q_\theta, q_{\theta+u}) = \int R_1(\theta,u,y)^2 q_\theta(y)dy \leq c_4\|u\|^2$. □

Proof of Lemma 7. Denote $p_{ij} = p_{\theta_i,\psi_j}$, $P_{ij}[g] := \int g(y)p_{ij}(y)dy$, for $i,j \in \{1,2\}$. Note that $d_{\mathrm{KL}}(p_{11},p_{22}) = d_{\mathrm{KL}}(p_{11},p_{21}) + P_{11}[\log \frac{p_{21}}{p_{22}}] \le c_2\|\omega_1\|_{C^2}\|\theta_1-\theta_2\|^2 + P_{11}[\log \frac{p_{21}}{p_{22}}]$ by Lemma 6. Use the fact that every p_{ij} has full support on $[0,\infty)$ to write

$$P_{11}[\log \tfrac{p_{21}}{p_{22}}] = P_{21}[\tfrac{p_{11}}{p_{21}} \log \tfrac{p_{21}}{p_{22}}] = P_{21}[(\tfrac{p_{11}}{p_{21}} - 1)\log \tfrac{p_{21}}{p_{22}}] + d_{\mathrm{KL}}(p_{21},p_{22}).$$

Notice, $d_{\mathrm{KL}}(p_{21},p_{22}) = d_{\mathrm{KL}}(\psi_1,\psi_2) \le K_0\epsilon^2$ for some constant K_0 that depends only on t_0; see Lemma 3.1 of van der Vaart and van Zanten (2008). An application of Cauchy-Schwarz inequality gives

$$P_{21}[(\tfrac{p_{11}}{p_{21}} - 1)\log \tfrac{p_{21}}{p_{22}}] \le \{P_{21}[(\tfrac{p_{11}}{p_{21}} - 1)^2]\}^{1/2}\{P_{21}[(\log \tfrac{p_{21}}{p_{22}})^2]\}^{1/2}.$$

Clearly $P_{21}[(\log \frac{p_{21}}{p_{22}})^2] = V(\psi_1,\psi_2) \le \|\log \frac{\psi_1}{\psi_2}\|_\infty^2 \le 4\|\omega_1-\omega_2\|^2$, and, by Lemma 6, $P_{21}[(\frac{p_{11}}{p_{21}} - 1)^2] \le c_3 e^{3t_0c_0\|\omega_1\|_{C^2}}\|\theta_1-\theta_2\|^2$. Additionally, $V(p_{11},p_{22}) \le 2V(p_{11},p_{21}) + 2\|\log \frac{\psi_1}{\psi_2}\|_\infty^2 \le c_3 e^{3t_0c_0\|\omega_1\|_{C^2}}\|\theta_1-\theta_2\|^2 + 4\|\omega_1-\omega_2\|_\infty^2$ by Lemma 6. This concludes the proof of the lemma with $K = \max(4, K_0, c_2\|\omega_1\|_{C^2}, c_3 e^{3t_0c_0\|\omega_1\|_{C^2}})$. □

Proof of Lemma 8. Denote $p_{ij} = p_{\theta_i,\psi_j}$, $i,j \in \{1,2\}$. By triangle inequality, $d_H(p_{11},p_{22}) \le d_H(p_{11},p_{21}) + d_H(p_{21},p_{22})$. The second term on the right equals $d_H(\psi_1,\psi_2)$ which is bounded by $\|\omega_1-\omega_2\|_\infty \exp\{\|\omega_1-\omega_2\|_\infty/2\}$ by Lemma 3.1 of van der Vaart and van Zanten (2008). The desired bound on the first term follows by the inequality $d_H(p_{11},p_{21}) \le d_{\mathrm{KL}}(p_{11},p_{21})^{1/2}$ and Lemma 6. □

Proof of Lemma 9. Let $W(t)$ be a centered Gaussian process on $\mathbb{R}$ with $\mathbb{C}\mathrm{ov}(W(s),W(t)) = e^{-(t-s)^2}$, $t,s \in \mathbb{R}$. Then ν^λ is the probability law of the rescaled process $W^\lambda = (W^\lambda(t) := W(\lambda t) : 0 \le t \le 1)$. The proof is complete by noting that

$$\|W^\lambda\|_{C^2} = \sup_{0\le t\le\lambda}|W(t)| + \lambda \sup_{0\le t\le\lambda}|\dot W(t)| + \lambda^2 \sup_{0\le t\le\lambda}|\ddot W(t)|$$

where, with probability one, the right hand side is non-decreasing in λ. □

Proof of Lemma 10. Fix $\lambda \ge 1$ and $\delta < 1/12$. Recall that $\mathbb{H}_1^\lambda$ consists of functions $\Re(h_\eta)$ where $h_\eta(u) = \int e^{ut\sqrt{-1}}\eta(t)\mu_\lambda(t)$ with $\|\eta\|_{L_2(\mu_\lambda)} \le 1$. By applying Cauchy-Schwarz inequality, with differentiations under integration as needed, it follows that

$$\|h\|_\infty \le 1, \|\dot h\|_\infty \le \sqrt{2}\lambda, \text{ and } \|\ddot h\|_\infty \le \sqrt{12}\lambda^2. \tag{14}$$

Any such h_η could be extended to an analytic function h_η on the complex plane $\mathbb{C}$ such that $|\frac{d^j}{dz^j}h_\eta(z)| \le 8\lambda^j e^{2|\Im(z)|^2\lambda^2}$, $z \in \mathbb{C}$ and $j \in \{0,1,2\}$. By Proposition C.9 of Ghosal and van der Vaart (2017), there is a collection $\mathcal{P} = \{P_1,\ldots,P_N\}$ of piecewise polynomials on $[0,1]$ with $\log N \le C_0\lambda(\log\frac{\lambda}{\delta})^2$ such that every $h \in \mathbb{H}_1^\lambda$ satisfies $\|\ddot h - P_n\|_\infty < \delta$ for some $1 \le n \le N$; here C_0 is a universal constant. Consider an expanded collection $\tilde{\mathcal{P}}$ of functions $\tilde P(u) = a + bu + \int_0^1 (u-t)_+P(t)dt$ where a belongs to a δ-net of $[-1,1]$, b belongs to a δ-net of $[-\sqrt{2}\lambda,\sqrt{2}\lambda]$ and $P \in \mathcal{P}$. Use (14) and Taylor's Theorem (second order, with residual in the integral form) to conclude every $h \in \mathbb{H}_1^\lambda$ satisfies $\|h-\tilde P\|_{C^2} < 6\delta$ for some $\tilde P \in \tilde{\mathcal{P}}$. This establishes the first claim because the cardinality $\tilde N$ of $\tilde{\mathcal{P}}$ satisfies

$\log \tilde{N} \leq \log N + \log(2/\delta) + \log(2\sqrt{2}\lambda/\delta) \leq C\lambda(\log \frac{\lambda}{6\delta})^2$ for all $\epsilon < 1/2$ and a new universal constant C. As shown in the proof of Lemma 4.7 of van der Vaart and van Zanten (2009), the second claim follows as a corollary to the first claim and Theorem 2 of Li and Linde (1999). □

Proof of Lemma 11. Let $S_n(t) = \sum_{i=1}^n I(Y_i > t)$ denote the sample exceedance count over a threshold t. Define the test functions

$$T_{1n} = I(|\tfrac{S_n(t_n)}{n} - \bar{F}^*(t_n)| > \epsilon_n), \quad T_{2n} = I(|\tfrac{S_n(2t_n)}{\max\{S_n(t_n),1\}} - \tfrac{\bar{F}^*(2t_n)}{\bar{F}^*(t_n)}| > \delta_n),$$

and take $T_n = \max(T_{1n}, T_{2n})$. Since $T_n \leq T_{1n} + T_{2n}$, we have $\mathbb{P}^n_{f^*} T_n \leq \mathbb{P}^n_{f^*} T_{1n} + \mathbb{P}^n_{f^*} T_{2n} \leq 2e^{-2n\epsilon_n^2} + \mathbb{P}^n_{f^*}[2e^{-2S_n(t_n)\delta_n^2}]$ by applications of Hoeffding's inequality where the second term is handled by the law of iterated expectation with an intermediate conditioning on $S_n(t_n)$. Now, for all large n, $\mathbb{P}^n_{f^*}[e^{-2S_n(t_n)\delta_n^2}] = [1 - \bar{F}^*(t_n)(1 - e^{-2\delta_n^2})]^n \leq [1 - \bar{F}^*(t_n)\delta_n^2]^n \leq e^{-n\bar{F}^*(t_n)\delta_n^2} \leq e^{-9n\epsilon_n^2}$; the last two inequalities hold because $1 - e^{-2x} \geq x$ for all small $x > 0$ and $1 + x \leq e^x$ for all x.

To bound the maximum type II error probability, first note that if $f \in \mathcal{F}_{1n} := \{f : |\bar{F}(t_n) - \bar{F}^*(t_n)| > 2\epsilon_n\}$ then $\mathbb{P}^n_f(1 - T_n) \leq \mathbb{P}^n_f(1 - T_{1n}) \leq 2e^{-2n\epsilon_n^2}$ by another application of Hoeffding's inequality. Next consider an $f \in \mathcal{T}(t_n, \delta_n) \setminus \mathcal{F}_{1n}$ with $\alpha_+(f) < \overline{\alpha}$ and $|\alpha_+(f) - \alpha^*| > 2^{4+\overline{\alpha}}\delta_n$. Let n be large enough so that $\delta_n < 1/2$. It follows from the definition of $\mathcal{T}(t, \delta)$ that $|\frac{\bar{F}(2t_n)}{\bar{F}(t_n)} - 2^{-\alpha_+(f)}| < 2^{1-\alpha_+(f)}\delta_n < 2\delta_n$ and hence

$$|\tfrac{\bar{F}(2t_n)}{\bar{F}(t_n)} - \tfrac{\bar{F}^*(2t_n)}{\bar{F}^*(t_n)}| \geq 2^{-\max(\alpha_+(f),\alpha^*)}\log(2)|\alpha_+(f) - \alpha^*| - 4\delta_n > 2\delta_n.$$

Consequently, $\mathbb{P}^n_f(1 - T_{2n}) \leq 2\mathbb{P}^n_f[2e^{-2\bar{S}_n(t_n)\delta_n^2}] \leq 2e^{-n\bar{F}_n(t_n)\delta_n^2}$. Since $f \notin \mathcal{F}_{1n}$, it follows that $\bar{F}(t_n) \geq \bar{F}^*(t_n) - 2\epsilon_n \geq \frac{1}{3}\bar{F}^*(t_n)$ and hence $\mathbb{P}^n_f(1 - T_{2n}) \leq 2e^{-n\epsilon_n^2}$. □

Proof of Lemma 12. Suppose $f = p_{\theta,\psi}$ with $\theta = (\alpha, \sigma) \in \Theta$, $\alpha \geq \alpha_1$, and $\psi = \mathcal{L}(\omega)$, $\|\dot{\omega}\|_\infty \leq D_n$. Denote $\phi = \log\psi$ and use Taylor's theorem to write $\bar{F}(y) = \psi(1)\bar{G}_\theta(y)\{1 - R_{\theta,\psi}(y)\}$ where $R_{\theta,\psi}(y) = \frac{\dot{\psi}(1-u)}{2\psi(1)}\bar{G}_\theta(y) = \frac{1}{2}e^{-u\dot{\phi}(1-v)}\dot{\phi}(1-u)\bar{G}_\theta(y)$ for some $0 < v < u < \bar{G}_\theta(y)$. Notice that $\bar{G}_\theta(y) = (\alpha\sigma/y)^\alpha\{1 + r_\theta(y)\}$ with $|r_\theta(y)| < \overline{\alpha}^2\overline{\sigma}/y \leq \overline{\alpha}^2\overline{\sigma}/t_n$ for all $y \geq t_n$ and consequently, $\bar{G}_\theta(y) \leq c_1 t_n^{-\alpha} \leq c_1\tau_n/D_n$ for all $y \geq t_n$, for some fixed constant c_1. Since $\|\dot{\phi}\|_\infty = \|\dot{\omega}\|_\infty \leq D_n$, it follows that for all large n, $|R_{\theta,\psi}(y)| \leq \frac{1}{2}e^{c_1\tau_n}c_1\tau_n \leq 2c_1\tau_n$ for all $y \geq t_n$ and consequently,

$$\frac{y^\alpha\bar{F}(y)}{\zeta(f)} = \{1 + r_\theta(y)\}\{1 - R_{\theta,\psi}(y)\} = 1 + \tilde{R}_{\theta,\psi}(y)$$

with $|\tilde{R}_{\theta,\psi}(y)| \leq 3\max(|R_{\theta,\psi}(y)|, |r_\theta(y)|) \leq B_1\tau_n$ for all $y \geq t_n$, for some constant B_1. This concludes the proof since the choice of B_1 does not depend on f. □

Additional summary of numerical experiments

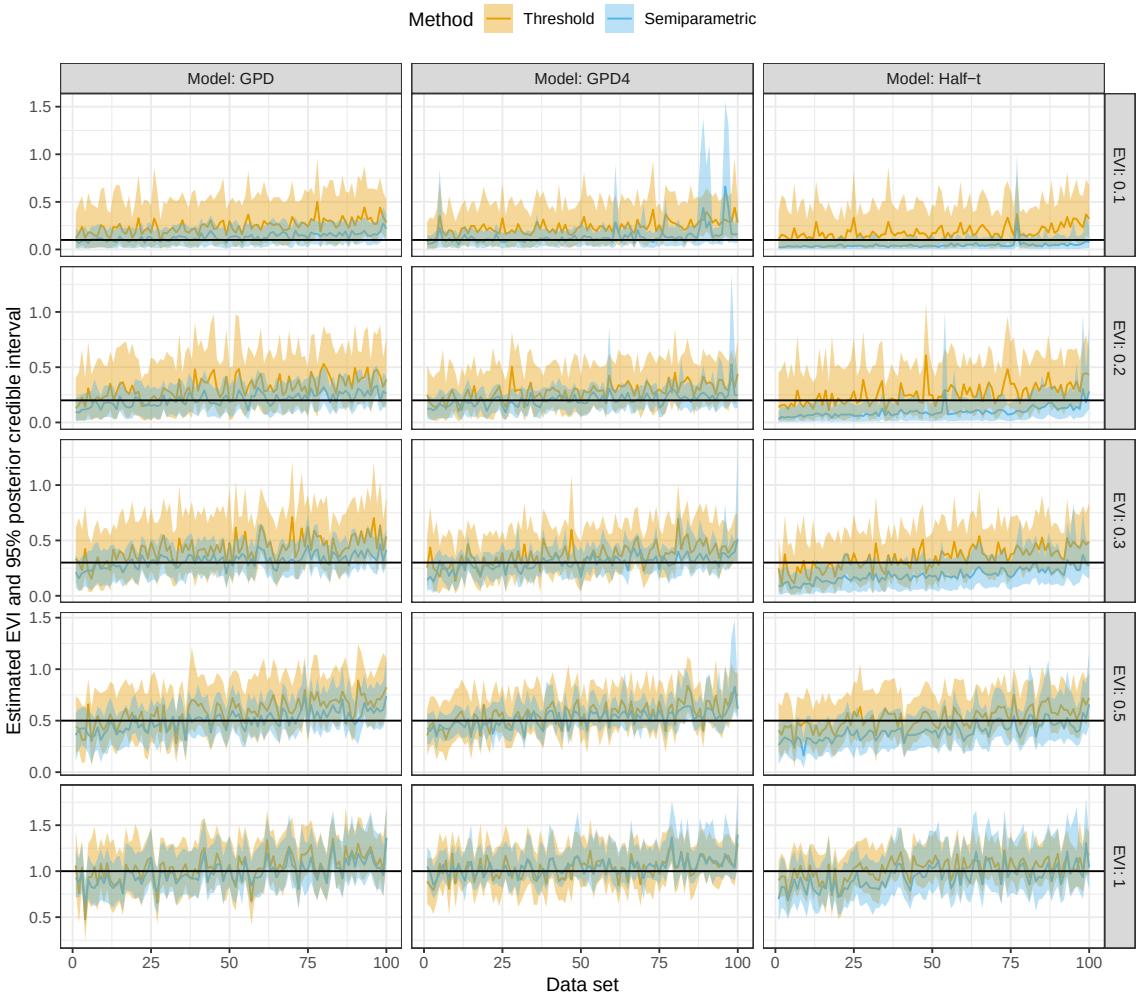

Figure 5: A comparison of the 95% posterior credible intervals for $\xi = \alpha^{-1}$ from the semiparametric and the thresholding methods. For each group, the 100 data sets are arranged in the ascending order of the maximum observation.

Codes

R package '`sbde`' can be downloaded from `https://CRAN.R-project.org/package=sbde`. Follow the link `https://www2.stat.duke.edu/~st118/Codes-FortCollins/` to access R codes along with dataset and auxiliary codes required to reproduce Fort Collins precipitation analysis presented in this article.

References

Adler, R. J. and J. E. Taylor (2009). *Random fields and geometry*. Springer Science & Business Media.

Alves, M. F. (2001). A location invariant hill-type estimator. *Extremes 4*(3), 199–217.

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive MCMC. *Statistics and Computing 18*(4), 343–373.

Balkema, A. and L. de Haan (1974). Residual life time at great age. *Annals of Probability 2*(5), 792–804.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Beirlant, J., E. Joossens, and J. Segers (2009). Second-order refined peaks-over-threshold modelling for heavy-tailed distributions. *Journal of Statistical Planning and Inference 139*(8), 2800–2815.

Carpentier, A. and A. K. Kim (2015). Adaptive and minimax optimal estimation of the tail coefficient. *Statistica Sinica 25*, 1133–1144.

Castillo, E. (2012). *Extreme value theory in engineering*. Elsevier.

de Zea Bermudez, P. and S. Kotz (2010). Parameter estimation of the generalized pareto distribution?part ii. *Journal of Statistical Planning and Inference 140*(6), 1374–1388.

Dekkers, A., J. Einmahl, L. De Haan, et al. (1989). A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics 17*(4), 1833–1855.

Diaconis, P. and D. Freedman (1986). On the consistency of bayes estimates. *The Annals of Statistics 14*(1), 1–26.

do Nascimento, F. F., D. Gamerman, and H. F. Lopes (2012). A semiparametric bayesian approach to extreme value estimation. *Statistics and Computing 22*(2), 661–675.

Durrieu, G., I. Grama, Q.-K. Pham, and J.-M. Tricot (2015). Nonparametric adaptive estimation of conditional probabilities of rare events and extreme quantiles. *Extremes 18*(3), 437–478.

Embrechts, P., C. Klüppelberg, and T. Mikosch (2013). *Modelling extremal events: for insurance and finance*, Volume 33. Springer Science & Business Media.

Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics 27*(1), 143–158.

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*, Volume 44. Cambridge University Press.

Gilleland, E. and R. W. Katz (2011). New software to analyze how extremes change over time. *Eos, Transactions American Geophysical Union 92*(2), 13–14.

Gu, M., X. Wang, and J. O. Berger (2018). Robust gaussian stochastic process emulation. *The Annals of Statistics 46*(6A), 3038–3066.

Hall, P. and A. Welsh (1984). Best attainable rates of convergence for estimates of parameters of regular variation. *Annals of Statistics 12*(3), 1079–1084.

Hall, P. and A. H. Welsh (1985). Adaptive estimates of parameters of regular variation. *The Annals of Statistics 13*(1), 331–341.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics 3*(5), 1163–1174.

Katz, R. W., M. B. Parlange, and P. Naveau (2002). Statistics of extremes in hydrology. *Advances in water resources 25*(8-12), 1287–1304.

Kleijn, B. (2021). Frequentist validity of bayesian limits. *The Annals of Statistics 49*(1), 182–202.

Lenk, P. J. (1988). The logistic normal distribution for bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association 83*(402), 509–516.

Lenk, P. J. (1991). Towards a practicable bayesian nonparametric density estimator. *Biometrika 78*(3), 531–543.

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

Li, C., L. Lin, and D. B. Dunson (2019). On posterior consistency of tail index for bayesian kernel mixture models. *Bernoulli 25*(3), 1999–2028.

Li, W. V. and W. Linde (1999). Approximation, metric entropy and small ball estimates for gaussian measures. *The Annals of Probability 27*(3), 1556–1578.

MacDonald, A., C. J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell (2011). A flexible extreme value mixture model. *Computational Statistics & Data Analysis 55*(6), 2137–2157.

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

Paulo, R. (2005). Default priors for gaussian processes. *The Annals of Statistics 33*(2), 556–582.

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

Rice, S. O. (1944). Mathematical analysis of random noise. *The Bell System Technical Journal 23*(3), 282–332.

Scarrot, C. and A. MacDonnald (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*(1), 10–26.

Snelson, E. and Z. Ghahramani (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pp. 1257–1264.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*, 1040–1053.

Tancredi, A., C. Anderson, and A. O'Hagan (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes 9*(2), 87–106.

Tokdar, S. T. (2007). Towards a faster implementation of density estimation with logistic Gaussian process priors. *Journal of Computational and Graphical Statistics 16*(3), 633–655.

Tokdar, S. T. and J. K. Ghosh (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference 137*(1), 34–42.

van der Vaart, A. and J. van Zanten (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics 36*(3), 1435–1463.

van der Vaart, A. W. and J. H. van Zanten (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics 37*(5B), 2655–2675.

Yang, Y. and S. T. Tokdar (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association 112*(519), 1107–1120.