Bernoulli **25**(3), 2019, 1999–2028
https://doi.org/10.3150/18-BEJ1043

On posterior consistency of tail index for Bayesian kernel mixture models

CHENG LI[1], LIZHEN LIN[2] and DAVID B. DUNSON[3]

[1]*Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore. E-mail: stalic@nus.edu.sg*
[2]*Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, Notre Dame, IN 46556, USA. E-mail: lizhen.lin@nd.edu*
[3]*Department of Statistical Science, Duke University, Durham, NC 27708, USA. E-mail: dunson@duke.edu*

Asymptotic theory of tail index estimation has been studied extensively in the frequentist literature on extreme values, but rarely in the Bayesian context. We investigate whether popular Bayesian kernel mixture models are able to support heavy tailed distributions and consistently estimate the tail index. We show that posterior inconsistency in tail index is surprisingly common for both parametric and nonparametric mixture models. We then present a set of sufficient conditions under which posterior consistency in tail index can be achieved, and verify these conditions for Pareto mixture models under general mixing priors.

Keywords: heavy tailed distribution; kernel mixture model; normalized random measures; posterior consistency; tail index

1. Introduction

Datasets from a variety of fields, such as environmental science, finance, industrial engineering, and telecommunications, demonstrate heavy tailed behavior that can substantially influence statistical inference and decision making. It is of interest to develop estimation methods that can capture both the bulk of the data and the tails accurately. Bayesian kernel mixture models provide a flexible framework for density estimation with strong large sample guarantees. Some of the most popular models include finite mixtures (MFM, Richardson and Green [45], Green and Richardson [26]), Dirichlet process mixtures (DPM, Ferguson [17], Lo [38], MacEachern [39], Escobar and West [15], Neal [41]), and mixtures with mixing measures given by normalized random measures with independent increments (NRMI, Regazzini, Lijoi and Prünster [44], Lijoi, Mena and Prünster [36], James, Lijoi and Prünster [32], Lijoi and Prünster [37], Barrios et al. [1], Favaro and Teh [16]). However, most of the existing literature on Bayesian asymptotics for density estimation, including results on posterior consistency and convergence rates, assumes that the true density has either a compact support or exponentially decaying tails (Ghosal, Ghosh and Ramamoorthi [22], Ghosal, Ghosh and van der Vaart [23], Ghosal and van der Vaart [24], Kruijer, Rousseau and van der Vaart [33], Shen, Tokdar and Ghosal [47]). A few exceptions such as Tokdar [50] and Wu and Ghosal [54] have shown posterior consistency for some heavy tailed densities for specific kernel mixture models. There exist fundamental limitations and barriers in understanding the tail behavior of kernel mixture models and their large sample properties, especially for models with nonparametric mixing priors.

1350-7265

The current paper investigates theory on the tail behavior of popular Bayesian kernel mixture models, assessing whether they are suitable for modeling heavy tailed distributions. We focus on studying the tails of univariate continuous densities and assume that the true density has polynomially decaying tails. Such power law behavior has been observed in many real data applications (see Clauset, Shalizi and Newman [7] for a review). Denote f_0 and F_0 as the true density function and the true cumulative distribution function (cdf) on $\mathbb{R}$. Let $\overline{F}(x) = 1 - F(x)$ for $x \in \mathbb{R}$ be the survival function of F. For sufficiently large x, a distribution F with a polynomially decaying right tail can be described by the relation

$$\overline{F}(x) = x^{-\alpha_+(F)} L_F(x), \tag{1.1}$$

where $\alpha_+(F) > 0$ is the *right tail index*, and L_F is a *slowly varying function* that satisfies $\lim_{y\to+\infty} L_F(xy)/L_F(y) = 1$ for any $x > 0$. In this paper, we will only consider a true distribution F_0 that satisfies the relation (1.1). The decay rate in the right tail of F_0 can be characterized by the right tail index $\alpha_+(F_0)$, up to some slowly varying function L_{F_0}. The left tail index can be defined similarly. If $\alpha_+(F_0) \in (0, +\infty)$, then from extreme value theory, the distribution F_0 falls within the class of *Fréchet maximum domain of attraction* (FMDA, Beirlant et al. [2]). Examples of distributions satisfying (1.1) with $\alpha_+(F_0) \in (0, +\infty)$ include the Pareto distribution, the Student's t distribution, the F distribution, the inverse gamma distribution, the log-gamma distribution, the Burr distribution, etc.

We study theoretical properties of the posterior distribution of the right tail index $\alpha_+(F)$ in a Bayesian framework. In particular, we consider the kernel mixture model:

$$f(x) = \int k(x;\boldsymbol{\theta})\,\mathrm{d}G(\boldsymbol{\theta}), \qquad G \sim \pi(G;\boldsymbol{\xi}), \tag{1.2}$$

where $k(\cdot;\boldsymbol{\theta})$ is a univariate kernel function with parameter $\boldsymbol{\theta}$ such that $\int k(x;\boldsymbol{\theta})\,\mathrm{d}x = 1$ for all $\boldsymbol{\theta}$, G is a mixing measure of $\boldsymbol{\theta}$, and π is the prior on G with hyperparameters $\boldsymbol{\xi}$. This model is quite general, covering the aforementioned MFM, DPM and NRMI mixture models as special cases.

We will answer two critical questions for understanding how model (1.2) can handle heavy tailed densities: (i) what choices of kernels and priors for the mixing measure can generate density functions with *tail indices varying* in a reasonable range, and (ii) under what types of conditions can one guarantee that the tail indices from the posterior are close to the tail index of the true distribution. The first question is related to whether the Bayesian kernel mixture model is capable of flexibly fitting heavy tailed distributions with different decay rates. The second question is on the frequentist asymptotic properties of Bayesian models estimating the tail index, requiring substantial extension of the scope of existing theory for Bayesian density estimation.

There is a rich literature on frequentist estimation of the tail index. Most of the estimators are constructed from tail order statistics, such as the Hill's estimator (Hill [31], de Haan and Resnick [9]), the Pickands' estimator (Pickands [43]) and their variations. The Hill's estimator is consistent (Mason [40]) and asymptotically normal with appropriate choices of the tail order statistics for certain nonparametric classes of distributions (Hall [28], Haeusler and Teugels [27]). Minimax rates for the tail index have been obtained under different classes of distributions (Hall and Welsh [29], Drees [13], Drees [14], Novak [42], Carpentier and Kim [6]), and they are

attainable through adaptive estimators (Hall and Welsh [30], Carpentier and Kim [6], Boucheron and Thomas [4]).

However, there is a lack of understanding of the properties of likelihood-based approaches. The limited Bayesian literature has focused on a peak-over-threshold (POT) strategy, with the tail of the density over a high threshold t assumed to follow a generalized Pareto distribution. If F belongs to FDMA with right tail index $\alpha_+(F)$, then as the threshold t becomes large, the right excess distribution $\widetilde{F}(x) = F(t+x)/F(t)$ for $x > 0$ converges in law to a generalized Pareto distribution with tail index $\alpha_+(F)$. Posterior sampling schemes have been discussed in, for example, Frigessi, Haug and Rue [18], Bottolo et al. [3], Stephenson and Tawn [49], Diebolt et al. [10], do Nascimento, Gamerman and Lopes [11], Wang, Rodriguez and Kottas [52], Fúquene Patiño [21]. The POT strategy can be viewed as artificial in choosing different models below and above the threshold, with the restriction of a parametric tail. The kernel mixture model allows one to choose a single flexible model for all of the data including the tails. Tressou [51] argues in favor of such an approach in using DPMs of Pareto kernels.

The rest of the paper is organized as follows. In Section 2, we formally introduce the definition of tail index and the posterior consistency of tail index. In Section 3, we show that in general, location-scale kernel mixture models cannot generate densities with varying tail indices, even if the kernel is heavy tailed. In particular, our results reveal that in many cases, the posterior distribution under the mixture model can only generate distributions with a *singleton index*. In Section 4, we provide general sufficient conditions for Bayesian posterior consistency of tail index. These conditions are then verified for the example of Pareto kernel mixtures in Section 4.3. Section 5 concludes with discussions. Technical proofs are included in the appendix and the supplementary material.

2. Preliminaries for tail index in a Bayesian framework

2.1. Definition of tail index

We first describe a notion of tail index for any distribution F with density f defined on $\mathbb{R}$. For $x \in \mathbb{R}$, we define its right and left tail indices as

$$\begin{aligned} \alpha_+(F) &= \liminf_{x\to+\infty} \frac{-\log \overline{F}(x)}{\log x} = \liminf_{x\to+\infty} \frac{-\log P_F(X > x)}{\log x}, \\ \alpha_-(F) &= \liminf_{x\to-\infty} \frac{-\log F(x)}{\log(-x)} = \liminf_{x\to-\infty} \frac{-\log P_F(X \le x)}{\log(-x)}, \end{aligned} \tag{2.1}$$

where $P_F(\cdot)$ denotes the probability evaluated under the distribution F. In the following, we will mainly discuss properties related to $\alpha_+(F)$ and all the results can be generalized similarly to $\alpha_-(F)$. Both $\alpha_+(F)$ and $\alpha_-(F)$ take values in $[0, +\infty]$. For the right tail, $\alpha_+(F) = +\infty$ represents a thin tailed cdf such as the exponential distribution, and $\alpha_+(F) = 0$ typically represents a super heavy tailed cdf such as the log-Pareto distribution (Cormann and Reiss [8]).

The lim inf used in the definition (2.1) is to pick up the heaviest part in the tail of F. The slowest possible decaying rate in the right tail of F is roughly of order $O(x^{-\alpha_+(F)})$ as

$x \to +\infty$. Furthermore, if F belongs to FMDA (i.e. it satisfies (1.1)), then the limit of the ratio $-\log \overline{F}(x)/\log x$ exists as $x \to +\infty$, and one can replace the lim inf in (2.1) by lim.

The definition of (2.1) can be viewed as a generalization of the usual tail index for distributions in FMDA. Frequentist estimators such as the Hill's estimator (Hill [31], de Haan and Resnick [9]), the Pickands' estimator (Pickands [43]) and their variations, are known to be asymptotically consistent for the tail index defined in (2.1) for certain restricted classes of distributions, such as FMDA. In general, it is unknown whether $\alpha_+(F)$ and $\alpha_-(F)$ defined by (2.1) can be consistently estimated from the data. However, this generalized notion of tail index in (2.1) is useful in describing the tail behavior of potentially complicated distributions drawn from Bayesian nonparametric priors.

2.2. Bayesian estimation and posterior consistency

Let $\mathcal{F}$ be the set of all distributions that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. Let $\mathcal{F}$ be the set of all density functions with respect to the Lebesgue measure on $\mathbb{R}$. Suppose that we observe a sample of i.i.d. data $\mathbf{X}^n = \{X_1, \ldots, X_n\}$ from the true distribution F_0 with the density f_0 on $\mathbb{R}$. Then in the Bayesian paradigm, we can impose a prior distribution on the probability density function $f \in \mathcal{F}$. For generality, here we will denote such a prior as $\Pi_n(\mathrm{d}f)$, which explicitly allows the prior to depend on the sample size n. Equivalently, Π_n is also a prior distribution over the set $\mathcal{F}$. Then the posterior distribution of $\Pi_n(\cdot|\mathbf{X}^n)$ evaluated at some measurable set $A \subseteq \mathcal{F}$ is

$$\Pi_n\big(A|\mathbf{X}^n\big) = \frac{\int_A \prod_{i=1}^n f(X_i)\Pi_n(\mathrm{d}f)}{\int_{\mathcal{F}} \prod_{i=1}^n f(X_i)\Pi_n(\mathrm{d}f)}. \tag{2.2}$$

To study properties related to the tail index, we define the following notion of tail index neighborhood.

Definition 2.1. For any distribution F and $\varepsilon > 0$, the *ε-(right) tail index neighborhood* of F with $\alpha_+(F) \in [0, +\infty)$ is

$$B_{\alpha+}(F, \varepsilon) \equiv \big\{H \in \mathcal{F} : \big|\alpha_+(H) - \alpha_+(F)\big| < \varepsilon\big\},$$

where $\alpha_+(\cdot)$ is defined in (2.1). If $\alpha_+(F) = +\infty$, then the ε-(right) tail index neighborhood of F is defined as

$$B_{\alpha+}(F, \varepsilon) \equiv \big\{H \in \mathcal{F} : \alpha_+(H) = +\infty\big\}.$$

The difference between the tail indices of two distributions used in Definition 2.1 is only a pseudometric, since different distributions can have the same tail index. In general, the topology induced by this pseudometric can be different from the weak topology generated by the weak convergence of probability measures. However, in the next proposition, we show that $B_{\alpha+}(F, \varepsilon)$ is a Borel set on the space of all absolutely continuous distributions with respect to the Lebesgue measure on $\mathbb{R}$ associated to the weak topology. The proof is given in Appendix A.

Proposition 2.1. *$B_{\alpha+}(F,\varepsilon)$ is a Borel set on $\mathcal{F}$ under the weak topology, for any distribution F and any $\varepsilon > 0$.*

Because the true tail index $\alpha_{0+} = \alpha_+(F_0)$ is unknown *a priori*, we hope that a distribution F drawn from the posterior $\Pi_n(\cdot|\mathbf{X}^n)$ in (2.2) has a tail index $\alpha_+(F)$ sufficiently close to the truth α_{0+}, as the sample size n increases to infinity. This notion of asymptotics is usually stated as *consistency*.

Definition 2.2. The posterior distribution $\Pi_n(\cdot|\mathbf{X}^n)$ is consistent for the (right) tail index if for any $\varepsilon > 0$, as $n \to \infty$,

$$\Pi_n\big(B^c_{\alpha+}(F_0,\varepsilon)|\mathbf{X}^n\big) \to 0, \qquad \text{in } P^{(n)}_{F_0} \text{ probability.}$$

Definition 2.2 is similar to the usual definition of posterior consistency for density estimation, but uses the tail index neighborhood in Definition 2.1. It requires that the posterior probability assigns almost zero mass to distributions outside ε-balls of F_0 as the sample size goes to infinity. On the other hand, although the weak consistency of density estimation is already well known for kernel mixture models (3.1) below (see, for example, Ghosal, Ghosh and Ramamoorthi [22], Tokdar [50], Wu and Ghosal [54]), posterior consistency of tail index does not follow directly from these results and requires further study, due to the non-equivalence between their topologies and neighborhoods.

3. Tail index of location-scale mixture models

In this section, we focus on a special case of Model (1.2), the location-scale mixture model

$$f(x) = \int \frac{1}{\sigma} k\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}G(\mu,\sigma), \qquad G \sim \pi(G;\boldsymbol{\xi}), \tag{3.1}$$

where $k(\cdot)$ is a kernel density function and the parameter $\boldsymbol{\theta} = (\mu,\sigma)$ consists of the location parameter μ and the scale parameter σ. We assume that the kernel $k(\cdot)$ has full support on $\mathbb{R}$. Frequentist asymptotic properties of this model have been extensively studied in the Bayesian nonparametrics literature. Both weak and strong posterior consistency of Model (3.1) have been discussed in Ghosal, Ghosh and Ramamoorthi [22], Tokdar [50], Wu and Ghosal [54], etc. Theorem 3.3 of Tokdar [50] established weak consistency of Model (3.1) when the true density f_0 has a very thick polynomially decaying tail, with the tail index in $(0,1)$. However, in the following, we will show that weak consistency, and even strong consistency based on L_1 or Hellinger distance, is insufficient for meaningful Bayesian inference of the tail index. Surprisingly, for many commonly used priors $\pi(G;\boldsymbol{\xi})$, the tail index of F generated from Model (3.1) can only take *one single value*, implying that there is no possibility of identifying the correct tail index unless we know the true α_{0+} *a priori*.

For the MFM model (Richardson and Green [45], Green and Richardson [26]), $f(x)$ in Model (3.1) is specified as a finite mixture of N components ($N \in \mathbb{Z}^+$), and a further prior distribution

is imposed on N. In more details, the model is given as,

$$\begin{aligned} f(x) &= \sum_{i=1}^{N} \frac{w_i}{\sigma_i} k\left(\frac{x-\mu_i}{\sigma_i}\right), \\ (\mu_i, \sigma_i)_{i=1}^{N} | N &\overset{\text{iid}}{\sim} G_0(\mu, \sigma), \\ (w_1, \ldots, w_N) | N &\sim \text{Dirichlet}(a, \ldots, a), \qquad \text{for some } a > 0, \\ N &\sim \pi(N) \text{ for } N = 1, 2, \ldots, \end{aligned} \tag{3.2}$$

The following theorem characterizes the tail index of a distribution F generated by Model (3.2).

Theorem 3.1. *Suppose that G_0 is a continuous distribution for (μ, σ). Then for any distribution F with density f drawn from Model (3.2), the range of $\alpha_+(F)$ is almost surely a singleton. In other words, almost surely all F's drawn from the MFM model have the same tail index.*

In the finite mixture model given in (3.2), the tail indices of different F's are all the same, since all of them are finite mixtures and their tail indices are solely determined by the tail heaviness of the kernel $k(\cdot)$. A heavy tailed kernel will only make the tails of F heavy, but not be able to generate varying tail heaviness. This limitation immediately indicates that we cannot obtain any meaningful posterior consistency in terms of tail index.

We now investigate the more complicated example where $G(\mu, \sigma)$ has a nonparametric NRMI prior. In the theorems to follow, we adopt similar NRMI notations as those in Lijoi, Mena and Prünster [36], James, Lijoi and Prünster [32], and Barrios et al. [1]. We consider a completely random measure $\widetilde{H}$, with $\widetilde{H}(x) = \sum_{i\geq 1} \widetilde{J}_i \delta_{X_i}(x)$ for $x \in \mathbb{R}$ such that $\{X_i\}_{i\geq 1}$ and nonnegative $\{\widetilde{J}_i\}_{i\geq 1}$ are independent sequences of random variables, ignoring jumps at nonrandom positions. The joint distribution of $\{\widetilde{J}_i\}_{i\geq 1}$ and $\{X_i\}_{i\geq 1}$ is characterized by the Lévy intensity $\nu(\mathrm{d}v, \mathrm{d}x)$ through the Laplace transformation of $\widetilde{H}$ (for $s > 0$):

$$E\left[e^{-s\widetilde{H}(A)}\right] = \exp\left\{-\int_{\mathbb{R}^+ \times A} \left(1 - e^{-sv}\right) \nu(\mathrm{d}v, \mathrm{d}x)\right\}, \qquad \text{for any } A \subseteq \mathbb{R}.$$

We consider the homogenous NRMI where the Lévy intensity can be factorized as $\nu(\mathrm{d}v, \mathrm{d}x) = \rho(\mathrm{d}v) H_0(\mathrm{d}x)$. $\rho(\mathrm{d}v)$ is the Lévy intensity for the nonnegative masses $\{\widetilde{J}_i\}_{i\geq 1}$, and $\{X_i\}_{i\geq 1}$ are independent draws from the nonatomic probability measure H_0, also called the "base measure". Then a NRMI H is defined as $H(x) = \sum_{i\geq 1} J_i \delta_{X_i}(x)$ with $J_i = \widetilde{J}_i / \sum_{i\geq 1} \widetilde{J}_i$ for any $x \in \mathbb{R}$. For all the theorems in this section, we assume that $\rho(\mathrm{d}v)$ satisfies $\int_0^\infty \rho(\mathrm{d}v) = +\infty$ and $\int_0^\infty (1 - e^{-v}) \rho(\mathrm{d}v) < +\infty$ which guarantees that $0 < \sum_{i\geq 1} \widetilde{J}_i < +\infty$ almost surely and the NRMI H is well defined; see equation (2.3) of Favaro and Teh [16].

The following theorem will be used as a fundamental tool in studying the tail behavior of a NRMI.

Theorem 3.2. *Suppose H is a homogeneous NRMI with the Lévy intensity measure $\rho(\mathrm{d}v)H_0(\mathrm{d}x)$ for $v \in \mathbb{R}^+, x \in \mathbb{R}$ where H_0 is a continuous probability measure on $\mathbb{R}$. Let $\Psi(s) = \int_0^{+\infty}(1 - e^{-sv})\rho(\mathrm{d}v)$ and let Ψ^{-1} be the inverse function of Ψ. Then*

(i) *If there exists a function h_γ defined as*

$$h_\gamma(x) = \frac{\log|\log x|}{\Psi^{-1}(\frac{\gamma \log|\log x|}{x})}, \tag{3.3}$$

with $\gamma > 1$ for $x \in (0, 1/e)$, such that $\liminf_{x\to+\infty} -\log h_\gamma(\overline{H}_0(x))/\log x = 0$, then $\alpha_+(H) = 0$ a.s.

(ii) *If there exists a function h such that*
   (a) *$h(x)$ is locally convex in $x \in [0, \varepsilon)$ for some small $\varepsilon > 0$;*
   (b) *$\int_0^\varepsilon \rho[h(x), +\infty)\,\mathrm{d}x < +\infty$;*
   (c) *$\liminf_{x\to+\infty} -\log h(\overline{H}_0(x))/\log x = +\infty$;*
   then $\alpha_+(H) = +\infty$ a.s.

The theorem follows from Fristedt [19] and Fristedt and Pruitt [20]; see Proposition A.1 and the subsequent proof of Theorem 3.2 in Appendix A. Proper choices of the functions h_γ in (i) and h in (ii) will lead to sharp lower and upper bounds for the tail index of a NRMI H.

The next theorem describes how these bounds for a NRMI can be used to characterize the tail behavior of a mixture density drawn from Model (3.1).

Theorem 3.3. *Suppose in Model (3.1), $G(\cdot,\cdot)$ is a homogeneous NRMI with Lévy intensity measure $\rho(\mathrm{d}v)G_0(\mathrm{d}\mu, \mathrm{d}\sigma)$ for $v \in \mathbb{R}^+$ and $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$, where $G_0(\mu, \sigma)$ is a continuous cdf on $\mathbb{R} \times \mathbb{R}^+$. Let $G_{0,\mu}$ and $G_{0,\sigma}$ be the marginal distributions of $G_0(\mu, \sigma)$ for μ and σ, respectively. Assume that $G_{0,\mu}$ is symmetric about zero. If both $G_{0,\mu}$ and $G_{0,\sigma}$ satisfy either* (i) *or* (ii) *in Theorem 3.2, that is, we can replace H_0 in either* (i) *or* (ii) *of Theorem 3.2 by $G_{0,\mu}$ and $G_{0,\sigma}$, then for any distribution F with density f drawn from Model (3.1), the range of $\alpha_+(F)$ as defined in (2.1) is almost surely a singleton.*

Theorem 3.3 indicates that the tail indices of all distributions F drawn from Model (3.1) are almost surely the same, if each of the two marginals $G_{0,\mu}$ and $G_{0,\sigma}$ satisfies either (i) or (ii) in Theorem 3.2. Again this indicates that there is no meaningful posterior consistency for tail index, by similar arguments after Theorem 3.1. Theorem 3.3 and its proof also lead to two other interesting implications. First, if the conditions for the two marginals of the base measure hold, then the tail index of F only depends on the tail indices of the two marginals, but not on the full joint distribution $G_0(\mu, \sigma)$. Second, whether $\alpha_+(F)$ is the same for all $F \sim \Pi_n(\cdot|\mathbf{X}^n)$ does not depend on the tail behavior of the kernel $k(\cdot)$, even if $k(\cdot)$ is a heavy tailed kernel.

Remark 3.1. The assumption of symmetric $G_{0,\mu}$ is only used as a sufficient condition for the case where $G_{0,\mu}$ satisfies (ii) of Theorem 3.2 and $G_{0,\sigma}$ satisfies (i) of Theorem 3.2, in other words, the case where G_μ has thin left and right tails and G_σ has a super heavy right tail. The assumption of symmetric $G_{0,\mu}$ is not necessary for the conclusions of Theorem 3.3 to hold when

both G_μ and G_σ are thin tailed, and when G_μ has a super heavy right tail. Details of the proof can be found in Appendix A.

Remark 3.2. The proof of Theorem 3.3 also relies on the moment techniques in Lemmas A.1–A.3 in Appendix A, which relate the tail index of F to the moments of F, and subsequently the moments of the kernel $k(\cdot)$ and the mixing distribution G. As a side product of this proof, we recovered the famous Breiman lemma in Breiman [5] about scale mixtures with heavy tailed mixing measures. Suppose in Model (3.1) we only have the scale mixture $f(x) = \int \sigma^{-1} k(x/\sigma)\,\mathrm{d}G(\sigma)$ and G has tail index $\alpha_+(G) \in (0, +\infty)$. The Breiman lemma says that if the kernel $k(\cdot)$ has a tail index larger than $\alpha_+(G)$, that is, it has a thinner tail than G, then the mixture $f(x)$ is also heavy tailed with tail index $\alpha_+(G)$. This is an immediate result of our Lemma A.3.

We make the tail conditions on $G_{0,\mu}$ and $G_{0,\sigma}$ more concrete for the special cases of Dirichlet process (DP) and normalized generalized Gamma process (NGGP, Lijoi, Mena and Prünster [36], James, Lijoi and Prünster [32], Lijoi and Prünster [37], Barrios et al. [1]) mixture models. It turns out that there is a large class of measures that satisfy the condition (i) or (ii) in Theorem 3.3, including both thin tailed distributions and heavy tailed distributions.

Theorem 3.4. *Suppose in Model* (3.1), $G(\cdot,\cdot) \sim \mathrm{DP}(a, G_0(\mu,\sigma))$ *with* $a > 0$ *and* $G_0(\mu,\sigma)$ *a continuous cdf on* $\mathbb{R} \times \mathbb{R}^+$. *Assume that* $G_{0,\mu}$ *is symmetric about zero. Consider the following two conditions for a generic distribution* H_0 *on* $\mathbb{R}$:

(i) $\limsup_{x\to+\infty} \overline{H}_0(x) \cdot (\log x / \log\log x) = +\infty$,
(ii) $\limsup_{x\to+\infty} \overline{H}_0(x) \cdot [(\log x) \cdot (\log\log x)^\delta] = 0$ *for some* $\delta > 1$.

If both $G_{0,\mu}$ *and* $G_{0,\sigma}$ *satisfy either one of the conditions* (i) *and* (ii), *that is, we can replace* H_0 *in either* (i) *or* (ii) *by* $G_{0,\mu}$ *and* $G_{0,\sigma}$, *then for any distribution* F *with density* f *drawn from Model* (3.1), *the range of* $\alpha_+(F)$ *as defined in* (2.1) *is almost surely a singleton.*

The proof of Theorem 3.4 involves the tail behavior of a DP, which has been studied in Doss and Sellke [12]. Conditions (i) and (ii) correspond to conditions (i) and (ii) in Theorem 3.3. As a result of the theorem, most distributions $G_{0,\mu}$ and $G_{0,\sigma}$ with either heavier or thinner tails than $1/\log x$ will lead to a single value of tail index for all F's in the DP mixture model, and therefore the posterior cannot estimate the truth α_{0+} consistently. For example, in the popular DP mixture of normals (Escobar and West [15]), the marginal distributions of the base measure for μ and σ^2 are the Student's t distribution and the inverse gamma distribution, both of which have much thinner tails than $1/\log x$. Therefore, the Bayesian posterior with the normal-inverse gamma prior for DP mixture of normals cannot consistently estimate the tail index. In contrast, Theorem 3.3 of Tokdar [50] has shown that such a normal-inverse gamma base measure is sufficient for posterior weak consistency, even if the true density is heavy tailed with a tail index in $(0, 1)$. This implies that the conditions required for consistent estimation of the tail index are more stringent than those for usual weak and strong posterior consistency. We emphasize again that the kernel here plays an inconsequential role due to Theorem 3.3, regardless of its tail thickness.

An important implication of Theorem 3.4 is that the bounds in (i) and (ii) are not far from each other. As a result, not many distributions have been left out by (i) and (ii). Basically, only those base measures that decay at a similar rate to $1/\log x$ are not covered by the conditions (i) and (ii). As a result, the only combination that is not covered by Theorem 3.4 is the case where both $G_{0,\mu}$ and $G_{0,\sigma}$ decay at rates similar to $1/\log x$. When this happens, the tail index of F drawn from Model (3.1) can possibly vary in $[0, +\infty]$. In this case, whether the posterior consistency of tail index holds or not remains unknown.

The next theorem shows a similar posterior behavior for the general NGGP mixture model, denoted by NGGP$(a, \kappa, \tau, G_0(\mu, \sigma))$. Its Lévy intensity measure is given by $\rho(\mathrm{d}v)\, dG_0(\mu, \sigma) = \frac{a}{\Gamma(1-\kappa)} v^{-\kappa-1} e^{-\tau v}\, \mathrm{d}v\, dG_0(\mu, \sigma)$, where $a > 0$, $\kappa \in [0, 1)$ and $\tau > 0$. The NGGP class includes most of the discrete random probability measures in the Bayesian nonparametric literature. For example, the class includes DP as NGGP$(a, 0, 1, G_0)$, the normalized-inverse Gaussian process as NGGP$(1, 1/2, \tau, G_0)$, and the N-stable process as NGGP$(1, \kappa, 0, G_0)$ as special cases. See Lijoi, Mena and Prünster [36] and Barrios et al. [1] for discussions. The cases of $\kappa = 0$ (DP) and $\kappa > 0$ are different in nature, so the conclusion of Theorem 3.5 is also different from Theorem 3.4.

Theorem 3.5. *Suppose in Model* (3.1), $G(\cdot, \cdot) \sim \mathrm{NGGP}(a, \kappa, \tau, G_0(\mu, \sigma))$ *with* $a > 0$, $\kappa \in (0, 1)$, $\tau \geq 0$ *and* $G_0(\mu, \sigma)$ *is a continuous cdf on* $\mathbb{R} \times \mathbb{R}^+$. *Assume that* $G_{0,\mu}$ *is symmetric about zero. Consider the following two conditions for a generic distribution* H_0 *on* $\mathbb{R}$:

(i) $\limsup_{x \to +\infty} \overline{H}_0(x) \cdot x^{\delta} = +\infty$ *for all* $\delta > 0$,
(ii) $\limsup_{x \to +\infty} \overline{H}_0(x) \cdot x^{\delta} = 0$ *for all* $\delta > 0$.

If both $G_{0,\mu}$ *and* $G_{0,\sigma}$ *satisfy either one of the conditions* (i) *and* (ii), *that is, we can replace* H_0 *in either* (i) *or* (ii) *by* $G_{0,\mu}$ *and* $G_{0,\sigma}$, *then for any distribution* F *with density* f *sampled from Model* (3.1), *the range of* $\alpha_+(F)$ *as defined in* (2.1) *is almost surely a singleton.*

Similar to Theorem 3.4, here we also provide two conditions for the tail decaying rates of $G_{0,\mu}$ and $G_{0,\sigma}$, where (i) gives heavier than polynomial tails and (ii) gives thinner than polynomial tails. The gap between the base measures that satisfy (i) or (ii) in the current theorem is now larger than that in the DP case, but the theorem still has ruled out many possibilities for consistent estimation of tail index. For example, when both $G_{0,\mu}$ and $G_{0,\sigma}$ have exponentially decaying tails, the tail index generated from the posterior of a NGGP is always the same as the tail index of the kernel $k(\cdot)$ (see the proof of Theorem 3.3 in Appendix A). It remains unknown how the tail indices of F from a NGGP mixture model behave in the posterior when at least one of $G_{0,\mu}$ and $G_{0,\sigma}$ have a polynomially decaying tail.

4. Sufficient conditions for tail index consistency

4.1. Schwartz's theorem for posterior consistency

In this section, we provide a series of conditions that guarantee the posterior consistency of tail index for the most general model $f \sim \Pi_n$. These conditions are built on the classic Schwartz's argument in Schwartz [46] for posterior consistency, and therefore they are simple and intuitive.

We will then demonstrate the application of these sufficient conditions on Model (1.2) using the Pareto kernel in Section 4.3.

The definition of tail index in (2.1) applies to any distribution but may be too general so that no consistent frequentist estimator exists. Therefore, we will limit our scope to those priors that only generate candidate distributions from the class of FMDA, that is, distributions that satisfy (1.1). These distributions have a well defined tail index, that is, we can replace all the lim inf in (2.1) by lim. Throughout the entire Section 4, we assume that the true distribution has a tail index $\alpha_{0+} \in (0, +\infty)$, and the prior Π_n satisfies Condition (PT).

(PT) For almost surely all $F \sim \Pi_n$, F satisfies the relation (1.1) with $\alpha_+(F) \in (0, +\infty)$ and a slowly varying function L_F, and its right tail index is given by (2.1) with all lim inf replaced by lim.

The Schwartz consistency theorem relies on two key conditions: the Kullback–Leibler (KL) support of the prior, and the existence of a uniformly consistent test. For two distributions F_1 and F_2 (with densities f_1 and f_2), let the KL divergence between F_1 and F_2 be $KL(F_1, F_2) \equiv E_{F_1} \log(f_1/f_2)$. Define the ε-KL neighborhood of the true distribution F_0 as $\mathcal{K}(F_0, \varepsilon) \equiv \{F \in \mathcal{F} : KL(F_0, F) < \varepsilon\}$. The condition on the KL support of the prior is stated as follows:

(KL) The true distribution F_0 is in the KL support of Π_n, if for any $\varepsilon > 0$, $\liminf_{n\to\infty} \Pi_n(\mathcal{K}(F_0, \varepsilon)) > 0$.

We allow the prior Π_n to depend on the sample size n, since this can be conveniently incorporated into the standard posterior consistency argument (see Section 5 of Ghosal, Ghosh and Ramamoorthi [22]). It is well known that the condition (KL) implies weak consistency, and is therefore a very basic requirement for useful Bayesian models.

The other condition required in the Schwartz consistency theorem is the existence of uniformly consistent tests. For our purpose, we need a test for tail index that is able to separate F_0 from all the distributions *outside a tail index neighborhood of* F_0. A set $\mathcal{F}_n$ with large prior probability (called "sieve") helps when Π_n has a non-compact support and the uniform test can be found on a sufficiently large set.

(UT) Uniform testing condition: There exists a test $\Phi_n \equiv \Phi_n(X_1, \ldots, X_n)$ and a sieve $\mathcal{F}_n$ such that
(i) $\Pi_n(\mathcal{F}_n^c) \le e^{-bn}$ for some constant $b > 0$;
(ii) For any $\varepsilon > 0$, as $n \to \infty$,

$$E_{F_0}\Phi_n \to 0, \qquad \sup_{F \in B_{\alpha+}^c(F_0,\varepsilon)\cap\mathcal{F}_n} E_F(1 - \Phi_n) \to 0. \tag{4.1}$$

Based on Schwartz's consistency theorem, one can show posterior consistency of tail index under the conditions (KL) and (UT).

Theorem 4.1. *If both* (KL) *and* (UT) *hold true, then the posterior distribution* $\Pi_n(\cdot|\mathbf{X}^n)$ *is consistent for the* (*right*) *tail index.*

The proof follows the same thread as the usual weak consistency (see for example Ghosal, Ghosh and Ramamoorthi [22], Ghosh and Ramamoorthi [25]) and is therefore omitted. Note that

the uniform test in (UT) can be made exponentially fast by an argument using the Hoeffding's inequality (Theorem 2 of Ghosal, Ghosh and Ramamoorthi [22], Proposition 4.4.1 of Ghosh and Ramamoorthi [25]). However, a key unanswered question is whether such a uniformly consistent test Φ_n for tail index exists. One cannot directly apply the Le Cam theory because Φ_n will depend on the new tail index neighborhood of $B_{\alpha+}(F_0, \varepsilon)$ and the pseudometric about tail index difference. We instead proceed in a constructive way and pursue sufficient conditions for (UT) to hold.

4.2. Existence of tests

In the representation (1.1) for a generic distribution $F \sim \Pi_n$, let $h_F(x) = xL'_F(x)/L_F(x)$ and hence $L_F(x) = L_F(x_0)\exp(\int_{x_0}^{x} \frac{h_F(t)}{t}\,\mathrm{d}t)$ for some fixed x_0. Alternatively, $h_F(x)$ can be written as

$$h_F(x) = \alpha_+(F) - \frac{xf(x)}{\overline{F}(x)}.$$

For any given F from FMDA, the von-Mises theorem (see, for example, Proposition 2.1 of Beirlant et al. [2]) says that

$$\lim_{x\to+\infty} \frac{xf(x)}{\overline{F}(x)} = \alpha_+(F),$$

i.e. $\lim_{x\to+\infty} h_F(x) = 0$. Bounding the magnitude of $h_F(x)$ is crucial in showing the existence of uniform tests for $\alpha_+(F)$. In the Bayesian framework, $h_F(x)$ with $F \sim \Pi_n$ needs to be controlled in a uniform way on a sieve with large prior probability. In light of this, we have the following theorem on the existence of tests. Throughout the rest of the paper, for two positive sequences $\{x_n\}$ and $\{y_n\}$ that depend on the sample size n, $x_n \prec y_n$ means $x_n = o(y_n)$, $x_n \succ y_n$ means $y_n = o(x_n)$, $x_n \preceq y_n$ means $x_n = O(y_n)$, and $x_n \succeq y_n$ means $y_n = O(x_n)$.

Theorem 4.2. *Suppose that* $\alpha_{0+} \in (0, +\infty)$, *and* (PT) *holds. In addition, suppose the following conditions hold*:

(i) *There exist finite constants* $x_0 \geq e$ *and* $c_L \in (0, 1)$, *such that for all sufficiently large* n, $L_F(x_0) \geq n^{-c_L}$ *uniformly for all* $F \in \mathcal{F}_{1n}$, *where* $\mathcal{F}_{1n}$ *is a sieve satisfying* $\Pi_n(\mathcal{F}_{1n}^c) < e^{-c_1 n}$ *for some constant* $c_1 > 0$;

(ii) *There exists an envelope function* $\overline{h}_n(x) = B_n(\log x)^{-(1+\tau_n)}$ *for some positive* n*-dependent sequences* B_n *and* τ_n, *such that for all sufficiently large* n, $|h_F(x)| \leq \overline{h}_n(x)$ *for all* $F \in \mathcal{F}_{2n}$ *and all* $x \geq x_0$, *where* $\mathcal{F}_{2n}$ *is a sieve satisfying* $\Pi_n(\mathcal{F}_{2n}^c) < e^{-c_2 n}$ *for some constant* $c_2 > 0$;

(iii) *The prior* Π_n *satisfies* $\Pi_n(\mathcal{F}_{3n}^c) < e^{-c_3 n}$ *for some constant* $c_3 > 0$ *for* $\mathcal{F}_{3n} = \{F \in \mathcal{F} : \alpha_+(F) \leq \overline{\alpha}_n\}$ *and some sequence* $1 \prec \overline{\alpha}_n \prec \log n$, *for all sufficiently large* n;

(iv) B_n, τ_n *and* $\overline{\alpha}_n$ *satisfy* $1 \preceq B_n \prec \min(\overline{\alpha}_n^{-1}\log n, \tau_n \log n)$ *and* $\tau_n \preceq 1$;

then (UT) *holds*.

The proof of Theorem 4.2 uses a recently proposed tail index estimator in Carpentier and Kim [6] defined as

$$\hat{\alpha}_{s_n} = \log(\hat{p}_{s_n}) - \log(\hat{p}_{s_n+1}), \tag{4.2}$$

where $\hat{p}_{s_n} = n^{-1}\sum_{i=1}^{n} I(X_i > e^{s_n})$ and s_n is taken as a positive sequence that satisfies $B_n \prec s_n \prec \overline{\alpha}_n^{-1}\log n$ (see the proof of Theorem 4.2 in Appendix A). Such a sequence s_n exists given Condition (iv) in Theorem 4.2. Carpentier and Kim [6] has shown that when $\alpha_{0+} \in (0, +\infty)$, $\hat{\alpha}_{s_n}$ is a consistent estimator of $\alpha_+(F)$ for F from various classes of distributions, such as the first order and the second order approximately Pareto distributions. Carpentier and Kim [6] has also given the explicit choice of s_n (as well as a data-dependent version) such that $\hat{\alpha}_{s_n}$ converges at a minimax rate to $\alpha_+(F)$ for a certain class of distributions (adaptively). Therefore, a test for $H_0 : \alpha_+(F) = \alpha_{0+}$ can be $\Phi_n = I(|\hat{\alpha}_{s_n} - \alpha_{0+}| > \varepsilon)$ given some $\varepsilon > 0$. For our purposes, it is easier to work with $\hat{\alpha}_{s_n}$ than the Hill's estimator.

Conditions (i)–(iv) are sufficient for the existence of such tests. Among them, (i) and (ii) are mainly intended to control the slowly varying function L_F, where we allow exceptions on sets with exponentially small prior probabilities. The choice of $x_0 \geq e$ is mainly for convenience since $\log x > 1$ for all $x \geq x_0$. Alternatively, one can replace it with any finite $x_0 \in \mathbb{R}$ and modify the definition of logarithm function with a shift accordingly. In (ii) we specify the envelope function $\overline{h}_n(x)$ to be decaying in the logarithm of x. In the frequentist tail index literature, such control over the exponent in a slowly varying function has appeared in Drees [13] and Drees [14] for showing minimax rates in certain classes of distributions. The logarithmically decaying $\overline{h}_n(x)$ is not restrictive because we allow $B_n \to \infty$ and $\tau_n \to 0$ as $n \to \infty$. As an envelop function, it also includes all $h_F(x)$ that decays polynomially in x.

Condition (iii) restricts the largest possible tail index on a large sieve, but the sieve will eventually cover the true F_0 as the sample size n increases. Condition (iv) determines the choice of B_n, τ_n in (ii) and $\overline{\alpha}_n$ in (iii). For posterior consistency, we only require the existence of such sequences $B_n, \tau_n, \overline{\alpha}_n$. Conditions (i)–(iv) will be verified for Pareto mixtures in Section 4.3.

Remark 4.1. We would like to emphasize that in our Bayesian setup, the class of distributions for which $\hat{\alpha}_{s_n}$ in (4.2) gives a uniform test depends on the conditions on the prior, for example Conditions (i)–(iv) in Theorem 4.2. These conditions impose restrictions on the class of distributions and densities that can be consistently fitted by our posterior (2.2) in the sense of weak consistency. In fact, they can result in a relatively smaller KL support of the prior, which may only include a subclass of FDMA. This is partly due to the basic requirement that our Baysian posterior should achieve consistency for both fitting the density and fitting the tail index at the same time. Although in general it is difficult to describe exactly which distributions are included in the prior KL support given those conditions in Theorem 4.2, we will shed light on this for the example of Pareto mixtures in Theorem 4.4 in Section 4.3.

The following theorem is a consequence of Theorem 4.1 and Theorem 4.2.

Theorem 4.3 (Posterior Consistency of Tail Index). *Under all assumptions of Theorem 4.2 and* (KL), *the posterior distribution* $\Pi_n(\cdot|\mathbf{X}^n)$ *is consistent for the* (*right*) *tail index.*

4.3. Example of consistency: Mixtures of Paretos

The failure of tail index consistency in Section 3 is partly due to the structure of the location-scale mixture model (3.1), in which we have no control over how the mixing measure $G(\mu, \sigma)$ affects the tail index of the mixture distribution. A possible remedy is to introduce an explicit mixture on the tail index parameter. An example of this type is the DPM of Paretos used in Tressou [51]. In this section, we study the mixture of simple Pareto distributions with kernel density $k(x; \alpha) = \alpha x^{-(\alpha+1)}$ whose support is $[1, +\infty)$. We will take the mixing measure from a homogenous NRMI prior, such as DP and NGGP. Because a general discrete mixture distribution takes the form $\overline{F}(x) = \sum_{i=1}^{\infty} w_i x^{-\alpha_i}$, the right tail index is $\alpha_+(F) = \inf\{\alpha_1, \alpha_2, \ldots\}$. To make this tail index more explicit, in the following Bayesian model, we are going to first pick α_1 as the tail index of F together with its weight w_1, and then draw the other α_i and their weights w_i $(i = 2, 3, \ldots)$ from a mixture model conditional on α_1 and w_1. In this way, we can guarantee that $\alpha_i > \alpha_1$ for all $i \geq 2$ such that we can conveniently control the behavior of $\alpha_+(F)$ through α_1. The model is specified as follows.

$$
\begin{aligned}
f(x)|\alpha_1, w_1, H &= w_1 k(x; \alpha_1) + (1 - w_1) \int k(x; \alpha)\, \mathrm{d}H(\alpha), \\
\alpha_1 &\sim G_\alpha \cdot I_{[0, \overline{\alpha}_n]}, \qquad \operatorname{supp}(G_\alpha) = [0, +\infty),\ G_\alpha \text{ has no point mass at zero}, \\
w_1 &\sim G_w \cdot I_{[\underline{w}_n, 1]}, \qquad \operatorname{supp}(G_w) = [0, 1], \\
H_1 &\sim \Pi(H_1; \boldsymbol{\xi}, H_0), \qquad \operatorname{supp}(H_0) = [0, +\infty),\ H_0 \text{ has no point mass at zero}, \\
H(\alpha) &= H_1(\alpha - \alpha_1), \qquad \text{for any } \alpha > \alpha_1.
\end{aligned}
\tag{4.3}
$$

The notation "supp" stands for the support of a distribution. For a generic distribution G and a set A, $G \cdot I_A$ denotes the renormalized probability distribution of G truncated to the set A. The density f has two mixing components. The first component $w_1 k(x; \alpha_1)$ explicitly controls the tail index of F, and the second component is a general mixture of Paretos. α_1 in the first component determines $\alpha_+(F)$, and is drawn from G_α truncated to $[0, \overline{\alpha}_n]$. Here the deterministic positive sequences $\underline{w}_n$ and $\overline{\alpha}_n$ satisfy that $\underline{w}_n \to 0$ and $\overline{\alpha}_n \to +\infty$ as $n \to \infty$, so asymptotically the supports of w_1 and α_1 covers any number in $(0, 1]$ and $\mathbb{R}^+$. The second component in the mixture involves a mixing probability measure H, which is drawn from a prior Π. $\boldsymbol{\xi}$ contains all the hyperparameters of Π, such as the parameter a in a DP and the parameters a, κ, τ in a NGGP. Given the value of α_1, H is a right-shifted version of the distribution H_1 drawn from the prior Π. For the ease of presentation, we assume that G_α, G_w and Π do not depend on n.

The deterministic sequences $\underline{w}_n$ and $\overline{\alpha}_n$ introduced here are mainly designed to separate the leading component $w_1 k(x; \alpha_1)$ from the other mixing components, such that the sufficient conditions in Theorem 4.2 are satisfied. In particular, condition (PT) can be conveniently verified for Model (4.3) with the help from the leading component. $\overline{\alpha}_n$ is used such that α_1 has an increasingly large support and meanwhile Condition (iii) of Theorem 4.2 is satisfied. In fact, the way of isolating the leading Pareto component in Model (4.3) is similar to some well studied nonparametric classes of distributions in the frequentist tail index literature, such as the Hall and Welsh

class (Hall and Welsh [30], Carpentier and Kim [6], Boucheron and Thomas [4]) that satisfies $|\overline{F}(x) - Cx^{-\alpha}| \le C'x^{-\alpha(1+\beta)}$ for $\alpha, \beta, C, C' > 0$.

A function g on the interval I is called completely monotone if the mth derivative of g satisfies $(-1)^m g^{(m)}(x) \ge 0$ for all $m \in \mathbb{Z}^+$. Let

$$\mathcal{CM}_e = \left\{F : \operatorname{supp}(F) = [1, +\infty), \overline{F}\left(e^t\right) \text{ is completely monotone on } t \in [0, +\infty)\right\},$$

$$\mathcal{P}_2 = \left\{F : \operatorname{supp}(F) = [1, +\infty), \overline{F}(x) = Cx^{-\alpha} + O\left(x^{-(1+\beta)\alpha}\right), \right.$$
$$\left. \text{for some constant } \alpha > 0, \beta > 0, C > 0\right\},$$

where $\mathcal{P}_2$ is the class of second-order Pareto distributions. We can characterize the class of distributions described by Model (4.3).

Theorem 4.4. *Suppose in Model* (4.3), $\underline{w}_n \to 0$ *and* $\overline{\alpha}_n \to +\infty$ *as* $n \to \infty$. *If* $F \in \mathcal{CM}_e \cap \mathcal{P}_2$ *and the prior* $\Pi(H; \xi, H_0)$ *is a homogeneous NRMI, then* F *is in the KL support of Model* (4.3).

The KL support of Model (4.3) is related to the class of completely monotone functions. This is not surprising because the mixtures of Paretos are related to the mixtures of exponential distributions by the transformation $x = e^t$ in the Pareto kernel $k(x; \alpha)$. The KL support of the mixtures of exponentials includes the class of completely monotone functions (Theorem 16 in Wu and Ghosal [54]), by the Hausdorff–Bernstein–Widder theorem. In fact, it is proved in Lemma S.1 in the supplementary material that any distribution F from $\mathcal{CM}_e \cap \mathcal{P}_2$ has a density with a similar form to that in Model (4.3).

The following theorem imposes further conditions on $\underline{w}_n, \overline{\alpha}_n$ and the prior G_α, G_w, Π, such that Model (4.3) achieves posterior consistency of tail index.

Theorem 4.5. *Suppose the following conditions hold for Model* (4.3):

(i) $F_0 \in \mathcal{CM}_e \cap \mathcal{P}_2$;

(ii) *The prior* Π *on the mixing measure* H *satisfies one of the following conditions*:

(a) Π *is* $\mathrm{DP}(a, H_0)$ *where* $a > 0$ *and* H_0 *is a probability distribution on* $\mathbb{R}^+$, *and there exist positive constants* $0 < c_1 < 1, D_1 > 0, d_1 > 0$, *such that* $H_0(x) \le D_1[\log(1/x)]^{-(1+d_1)}$ *for all* $x \in (0, c_1)$;

(b) Π *is* $\mathrm{NGGP}(a, \kappa, \tau, H_0)$ *where* $a > 0$, $\kappa \in (0, 1)$, $\tau > 0$ *and* H_0 *is a probability distribution on* $\mathbb{R}^+$, *and there exist positive constants* $0 < c_2 < 1, D_2 > 0, d_2 > 0$, *such that* $H_0(x) \le D_2 x^{1+d_2}$ *for all* $x \in (0, c_2)$;

(iii) $1 \prec \overline{\alpha}_n \prec \log n, \overline{\alpha}_n / \log n \prec \underline{w}_n \prec 1$;

then the posterior distribution $\Pi_n(\cdot|\mathbf{X}^n)$ *of Model* (4.3) *is consistent for the tail index.*

Condition (ii) in Theorem 4.5 requires sufficient decay for the base measure H_0 near zero, though the decaying rate could be different for a DP prior and a NGGP prior. For a NGGP prior, the decaying rate of $H_0(x)$ near $x = 0$ needs to be in polynomials of x, while the rate for a DP prior can be slower, in polynomials of $\log(1/x)$ for x close to zero. This is due to the difference

in the tail behavior of DP and NGGP. Condition (iii) describes the orders of $\underline{w}_n$ and $\overline{\alpha}_n$. They can be taken as, for example, $\underline{w}_n = (\log n)^{-1/3}$ and $\overline{\alpha}_n = (\log n)^{1/2}$.

Remark 4.2. The densities in $\mathcal{CM}_e \cap \mathcal{P}_2$ always have nonnegative mixing coefficients, since $w_1 > 0$ and H is a probability measure. As a result, the KL support of Model (4.3) includes mixtures such as $\overline{F}(x) = \frac{1}{2x} + \frac{1}{2x^2}$, but also has excluded some other mixtures of Paretos, such as $\overline{F}(x) = \frac{2}{x} - \frac{1}{x^2}$ in which some components may have negative coefficients. To enlarge the KL support of Model (4.3) and allow negative mixing coefficients, the mixing measure can be characterized as a bounded signed measure $w_1\delta_{\alpha_1} + (1 - w_1)H = H_+ - H_-$, where δ_a denotes the Dirac measure at a. Similar priors to those in Model (4.3) can be imposed on both H_+ and H_- and they need further restrictions to guarantee that the density f is nonnegative. For example, if $\overline{F}(x) = \frac{2}{x} - \frac{1}{x^2}$, then $H_+ = 2\delta_1$ and $H_- = \delta_2$. According to Theorem 4.3 of Watanabe [53], the Pareto kernel mixture representation by using bounded signed mixing measure includes all distributions F that satisfy $\sum_{i=1}^{\infty} |\overline{F}^{(i)}(e^t)|t^i/i! < +\infty$.

5. Discussion

We have explored the theory behind the posterior consistency/inconsistency of tail index for Bayesian kernel mixture models, extending the scope of the vast literature on Bayesian consistency with respect to the weak and strong topology. We have shown that examples of inconsistency are extremely common, among the location-scale mixture models with MFM, DPM and NRMI mixture priors. There are special cases in which posterior consistency remains unknown in the DPM and NRMI mixture examples when the marginal base measures of the location and scale parameters meet certain restrictions.

We have also proposed a set of sufficient conditions that lead to posterior tail index consistency, and verified them in a Pareto mixture example. The simple Pareto mixture model is mainly used for illustration, as other heavy tailed kernels with an explicit tail index parameter can also be implemented in a similar manner, such as the inverse gamma kernel, the half Student's t kernel, and the F kernel, although their consistency theory involves extra technical complexity in verifying all those sufficient conditions. It is less obvious to see how models like (4.3) can be generalized to mixing models with two-sided kernels, since ideally one wants to estimate both the left tail index and the right tail index of a distribution, which can be possibly different. It will be an interesting topic to further study the posterior convergence rates for Model (4.3) when the true $F_0(x)$ comes from certain nonparametric classes such as the Hall and Welsh class, and compare them with the frequentist adaptive estimators such as Carpentier and Kim [6] and Boucheron and Thomas [4], which achieve the minimax rates.

Appendix A: Technical proofs

Proof of Proposition 2.1. Lange [34] Theorem 3.5 has proved that $\mathcal{F}$, the set of all distributions that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, is a Borel set. Below,

we show that the following sets

$$\begin{aligned}\mathcal{A}_1(a) &= \{F \in \mathcal{F} : \alpha_+(F) \le a\},\\ \mathcal{A}_2(a) &= \{F \in \mathcal{F} : \alpha_+(F) < a\},\end{aligned} \tag{A.1}$$

are Borel sets for any $a \in [0, +\infty]$. First let $a \in [0, +\infty)$. Then for a generic continuous function g on $\mathbb{R}$, we have the following relation

$$\begin{aligned}\Big\{g : \liminf_{x\to+\infty} g(x) \le a\Big\} &= \Big\{g : \sup_{k\in\mathbb{Z}^+,k\ge2} \inf_{r_j\in\mathbb{Q}^+} g(k+r_j) \le a\Big\}\\ &= \bigcap_{k=2}^{+\infty}\Big\{g : \inf_{r_j\in\mathbb{Q}^+} g(k+r_j) \le a\Big\}\\ &= \left(\bigcup_{k=2}^{+\infty}\Big\{g : \inf_{r_j\in\mathbb{Q}^+} g(k+r_j) > a\Big\}\right)^c\\ &= \left(\bigcup_{k=2}^{+\infty}\bigcup_{q_l\in\mathbb{Q}^+}\Big\{g : \inf_{r_j\in\mathbb{Q}^+} g(k+r_j) \ge a+q_l\Big\}\right)^c\\ &= \left(\bigcup_{k=2}^{+\infty}\bigcup_{q_l\in\mathbb{Q}^+}\bigcap_{r_j\in\mathbb{Q}^+}\{g : g(k+r_j) \ge a+q_l\}\right)^c, \quad \text{and}\\ \Big\{g : \liminf_{x\to+\infty} g(x) < a\Big\} &= \Big\{g : \sup_{k\in\mathbb{Z}^+,k\ge2} \inf_{r_j\in\mathbb{Q}^+} g(k+r_j) < a\Big\}\\ &= \bigcap_{k=2}^{+\infty}\Big\{g : \inf_{r_j\in\mathbb{Q}^+} g(k+r_j) < a\Big\}\\ &= \left(\bigcup_{k=2}^{+\infty}\Big\{g : \inf_{r_j\in\mathbb{Q}^+} g(k+r_j) \ge a\Big\}\right)^c\\ &= \left(\bigcup_{k=2}^{+\infty}\bigcap_{r_j\in\mathbb{Q}^+}\{g : g(k+r_j) \ge a\}\right)^c,\end{aligned} \tag{A.2}$$

where $\mathbb{Q}^+$ is the set of all positive rational numbers.

For any $F \in \mathcal{F}$, we have that $-\log\overline{F}$ is a continuous function on $\mathbb{R}$ (in case $\overline{F}(x) = 0$ for all $x \in [x_1, +\infty)$ with some finite number x_1, we can extend the concept of continuity by defining $-\log\overline{F}(x) = +\infty$ for all $x \ge x_1$). Since $1/\log x$ is continuous for $x \in (1, +\infty)$, we have that the product $-\log\overline{F}(x)/\log x$ is also continuous on $(1, +\infty)$. For given $x > 1, b \ge 0$, we

define

$$\begin{aligned}\mathcal{D}(x,b) &= \left\{F\in\mathcal{F}:\frac{-\log\overline{F}(x)}{\log x}\geq b\right\} = \left\{F\in\mathcal{F}:\overline{F}(x)\leq x^{-b}\right\}\\ &= \left\{F\in\mathcal{F}:F(x)\geq 1-x^{-b}\right\} = \left\{F\in\mathcal{F}:F(x)<1-x^{-b}\right\}^c.\end{aligned} \tag{A.3}$$

Then (A.1), (A.2) and (A.3) together imply that for any $a\in[0,+\infty)$,

$$\begin{aligned}\mathcal{A}_1(a) &= \left(\bigcup_{k=2}^{+\infty}\bigcup_{q_l\in\mathbb{Q}^+}\bigcap_{r_j\in\mathbb{Q}^+}\mathcal{D}(k+r_j,a+q_l)\right)^c,\\ \mathcal{A}_2(a) &= \left(\bigcup_{k=2}^{+\infty}\bigcap_{r_j\in\mathbb{Q}^+}\mathcal{D}(k+r_j,a)\right)^c.\end{aligned} \tag{A.4}$$

For any fixed $x>1$ and fixed $p\in(0,1]$, the set $\{F:F(x)<p\}$ is the pre-image of the Borel set $[0,p)$ under the mapping $T_A:F\mapsto F(A)$ for the given Borel set $A=(-\infty,x]$. On the other hand, we know that the Borel sigma-algebra on the space of all distributions F is defined as the smallest sigma-algebra that makes the mapping $F\mapsto F(A)$ measurable for any Borel set $A\subseteq\mathbb{R}$. Using this definition, we know that $\{F:F(x)<p\}$ for fixed $x>1$ and $p\in(0,1]$ is a Borel set. Therefore, $\mathcal{D}(x,b)$ in (A.3) is a Borel set, which further implies that in (A.4), both $\mathcal{A}_1(a)$ and $\mathcal{A}_2(a)$ are Borel sets for any $a\in[0,+\infty)$.

If $a=+\infty$, then $\mathcal{A}_1(+\infty)=\mathcal{F}$ is trivially Borel, and $\mathcal{A}_2(+\infty)=\bigcup_{l=2}^{+\infty}\mathcal{A}_2(l)$ is also Borel since every $\mathcal{A}_2(l)$ is Borel for $l=2,3,\ldots$. Therefore, both $\mathcal{A}_1(a)$ and $\mathcal{A}_2(a)$ are Borel sets for any $a\in[0,+\infty]$.

Finally, we can write $B_{\alpha+}(F,\varepsilon)=\mathcal{A}_2(\alpha_+(F)+\varepsilon)\cap(\mathcal{A}_1(\alpha_+(F)-\varepsilon))^c$ (in case $\alpha_+(F)-\varepsilon<0$, then $\mathcal{A}_1(\alpha_+(F)-\varepsilon)$ is understood as the empty set). Thus, $B_{\alpha+}(F,\varepsilon)$ is a Borel set. □

In the following, P_F and E_F represent the probability and the expectation under the probability distribution F. A random variable X has the decomposition $X=X_+-X_-$, where $X_+=\max(X,0)$ and $X_-=\max(-X,0)$.

Lemma A.1 (Shorack and Wellner [48], Theorem 1, Section 7 in Chapter 4). *Let F be an univariate distribution on $\mathbb{R}$ with right tail index $\alpha_+(F)$ as defined in (2.1). If a random variable X has the cdf $F(x)$, then*

$$E_F X_+^m=\begin{cases}<+\infty & \text{if } 0<m<\alpha_+(F),\\ =+\infty & \text{if } m>\alpha_+(F).\end{cases}$$

Lemma A.2. *Let $m>0$.*

(i) *For any $x,y\in\mathbb{R}$, there exists a constant C_m that only depends on m, such that*

$$\big[(x+y)_+\big]^m\leq C_m\big(x_+^m+y_+^m\big).$$

(ii) *For any $x \geq 0$, $y \geq 0$, there exists a constant c_m that only depends on m, such that*

$$(x+y)^m \geq c_m\big(x^m + y^m\big).$$

Proof of Lemma A.2.

(i) For $m \geq 1$, $C_m = 2^{m-1}$. For $m \in (0,1)$, $C_m = 1$.
(ii) Let $f(t) = t^m + (1-t)^m$ and $t \in [0,1]$. If $m \geq 1$, then $\max_{t\in[0,1]} f(t) = 1$ and set $c_m = 1$. If $m \in (0,1)$, then $\max_{t\in[0,1]} f(t) = 2^{1-m}$ and set $c_m = 2^{m-1}$. Now let $t = x/(x+y)$ and the conclusion follows. □

Lemma A.3. *Suppose f is a density drawn from Model (3.1) with cdf F. Let $K(\cdot)$ be the cdf of $k(\cdot)$. Then*

$$E_F X_+^m \geq c_m\big(E_{G_\mu}\mu_+^m \cdot \overline{K}(0) + E_K X_+^m \cdot E_{G_{\mu,\sigma}}\big[\sigma^m I(\mu \geq 0)\big]\big), \tag{A.5}$$

$$E_F X_+^m \geq C_m^{-1} E_K X_+^m \cdot E_{G_\sigma}\sigma^m - E_{G_\mu}\mu_-^m, \tag{A.6}$$

$$E_F X_+^m \leq C_m\big(E_{G_\mu}\mu_+^m + E_K X_+^m \cdot E_{G_\sigma}\sigma^m\big), \tag{A.7}$$

where $\overline{K}(0) = P_K(X \geq 0)$ (the probability of $X \geq 0$ if X has the density $k(x)$), G_μ and G_σ are the marginal distributions of $G_{\mu,\sigma}$, and c_m, C_m are defined in Lemma A.2.

Proof of Lemma A.3. Let $I(\cdot)$ be the indicator function. We have

$$\begin{aligned} E_F X_+^m &= \iint x^m I(x \geq 0)\frac{1}{\sigma}k\left(\frac{x-\mu}{\sigma}\right)\mathrm{d}G(\mu,\sigma)\,\mathrm{d}x \\ &= \iint (\mu+\sigma y)^m I(\mu+\sigma y \geq 0)k(y)\,\mathrm{d}G(\mu,\sigma)\,\mathrm{d}y. \end{aligned} \tag{A.8}$$

Then we give lower and upper bounds for (A.8). Notice that $\sigma \geq 0$ always holds and $I(\mu+\sigma y \geq 0) \geq I(\mu \geq 0)I(y \geq 0)$. Based on (A.8) and part (ii) of Lemma A.2, we have:

$$\begin{aligned} E_F X_+^m &\geq \iint (\mu+\sigma y)^m I(\mu \geq 0)I(y \geq 0)k(y)\,\mathrm{d}G(\mu,\sigma)\,\mathrm{d}y \\ &\geq \iint c_m\big(\mu_+^m + \sigma^m y_+^m\big)I(\mu \geq 0)I(y \geq 0)k(y)\,\mathrm{d}G(\mu,\sigma)\,\mathrm{d}y \\ &= c_m\big(E_{G_\mu}\mu_+^m \cdot \overline{K}(0) + E_K X_+^m \cdot E_{G_{\mu,\sigma}}\big[\sigma^m I(\mu \geq 0)\big]\big), \end{aligned}$$

which is (A.5).

On the other hand, since $(-\mu)_+ = \mu_-$, part (i) of Lemma A.2 implies

$$\begin{aligned} &(\sigma y)_+^m \leq C_m\big[(\mu+\sigma y)_+^m + (-\mu)_+^m\big] \\ \Longrightarrow\quad &(\mu+\sigma y)_+^m \geq C_m^{-1}\sigma^m y_+^m - \mu_-^m. \end{aligned}$$

This together with (A.8) gives

$$\begin{aligned} E_F X_+^m &= \iint (\mu + \sigma y)_+^m k(y)\,\mathrm{d}G(\mu,\sigma)\,\mathrm{d}y \\ &\geq \iint \left[C_m^{-1}\sigma^m y_+^m - \mu_-^m\right] k(y)\,\mathrm{d}G(\mu,\sigma)\,\mathrm{d}y \\ &\geq C_m^{-1} E_K X_+^m \cdot E_{G_\sigma}\sigma^m - E_{G_\mu}\mu_-^m, \end{aligned}$$

which is (A.6).

By part (i) of Lemma A.2

$$\begin{aligned} E_F X_+^m &= \iint \left[(\mu + \sigma y)_+\right]^m \mathrm{d}G(\mu,\sigma)\,\mathrm{d}y \\ &\leq \iint C_m\left(\mu_+^m + \sigma^m y_+^m\right) k(y)\,\mathrm{d}G(\mu,\sigma)\,\mathrm{d}y = C_m\left(E_{G_\mu}\mu_+^m + E_K X_+^m \cdot E_{G_\sigma}\sigma^m\right), \end{aligned}$$

which is (A.7). □

Proof of Theorem 3.1. For Model (3.2), the marginal distributions G_μ and G_σ are both finite mixtures at the points $\mu_{i=1}^N$ and $\sigma_{i=1}^N$ respectively. Because $G_0(\mu,\sigma)$ is a continuous distribution, we have $0 \leq E_{G_\mu}\mu_+^m < +\infty$, $0 \leq E_{G_\mu}\mu_-^m < +\infty$ and $0 < E_{G_\sigma}\sigma^m < +\infty$ for all $m > 0$. We can use Lemma A.1 to determine the relation between $\alpha_+(F)$ and $\alpha_+(K)$. According to Lemma A.3, whether $E_F X_+^m$ is finite or not for a given m is solely determined by whether $E_K X_+^m$ is finite or not. The analysis goes as follows:

(i) If $\alpha_+(K) = +\infty$, then by Lemma A.1 $E_K X_+^m < +\infty$ for all $m > 0$. The upper bound (A.7) implies that $E_F X_+^m < +\infty$ for all $m > 0$. Hence, $\alpha_+(F) = +\infty$ by Lemma A.1.

(ii) If $\alpha_+(K) = 0$, then by Lemma A.1, $E_K X_+^m = +\infty$ for all $m > 0$. The lower bound (A.6) implies that $E_F X_+^m = +\infty$ for all $m > 0$. Then by setting $m = 0$ in (ii) of Lemma A.1 we can see that $\alpha_+(F) = 0$.

(iii) If $\alpha_+(K) \in (0, +\infty)$, then $E_K X_+^m < +\infty$ for $m < \alpha_+(K)$ and $E_K X_+^m = +\infty$ for $m > \alpha_+(K)$. Then by (A.7), $E_F X_+^m < +\infty$ for $m < \alpha_+(K)$, and by (A.6), $E_K X_+^m = +\infty$ for $m > \alpha_+(K)$. Apply Lemma A.1 and we can see that $\alpha_+(F) = \alpha_+(K)$.

In sum, $\alpha_+(F) = \alpha_+(K)$ in all three cases and thus $\alpha_+(F)$ is almost surely a singleton. □

The homogenous NRMI with Lévy intensity $\rho(\mathrm{d}v)$ and base measure H_0 defined in Sect. 3 can be expressed as $H(x) = \sum_{i\geq 1} J_i \delta_{X_i}(x)$ where $J_i = \widetilde{J}_i / \sum_{i\geq 1} \widetilde{J}_i$. Equivalently, the cdf $H(x)$ also has the representation $H(x) = S(H_0(x))/S(1)$, where $\{S(t), t \geq 0\}$ is a subordinator with Lévy intensity measure $\rho(\mathrm{d}v)$ (see for example Regazzini, Lijoi and Prünster [44]). As a result, the function Ψ defined in Theorem (3.2) is the Laplace exponent of the subordinator $S(t)$. The conditions $\int_0^\infty \rho(\mathrm{d}v) = +\infty$ and $\int_0^\infty (1 - e^{-v})\rho(\mathrm{d}v) < +\infty$ guarantees that $0 < S(1) < +\infty$ almost surely.

The following proposition is a combination of Theorem 1 in Fristedt [19] and Lemmas 4 and 5 in Fristedt and Pruitt [20].

Proposition A.1 (Fristedt [19], Fristedt and Pruitt [20]). *Suppose* $\{S(t), t \geq 0\}$ *is a subordinator with Lévy intensity measure* $\rho(\mathrm{d}v)$ *for* $v \in \mathbb{R}^+$. *Define the functionals* $R_L(h) = \liminf_{t\to 0+} S(t)/h(t)$ *and* $R_U(h) = \limsup_{t\to 0+} S(t)/h(t)$.

(i) *For* $\gamma > 0$, *let* $h_\gamma(x)$ *be the same as defined in* (3.3). *Then*

$$R_L(h_\gamma) \leq \gamma \qquad \text{a.s. if } \gamma < 1,$$
$$R_L(h_\gamma) \geq \gamma - 1 \qquad \text{a.s. if } \gamma > 1.$$

(ii) *If* $h(x)$ *is locally convex in* $x \in [0, \varepsilon)$ *for some small* $\varepsilon > 0$, *then*

$$R_U(h) = 0 \qquad \text{a.s. if } \int_0^\varepsilon \rho[h(x), +\infty)\,\mathrm{d}x < +\infty,$$
$$R_U(h) = +\infty \qquad \text{a.s. if } \int_0^\varepsilon \rho[h(x), +\infty)\,\mathrm{d}x = +\infty.$$

Proof of Theorem 3.2. The proof is a direct application of Proposition A.1.

(i) By the stationary increment property of subordinators, $S(1-t)$ has the same distribution as $S(1) - S(t)$ for $t \in (0, 1)$. Therefore for $\gamma > 1$ and h_γ defined in (3.3), part (i) of Proposition A.1 implies

$$\liminf_{t\to 1-} \frac{S(1) - S(t)}{h_\gamma(1-t)} \geq \gamma - 1 \qquad \text{a.s.}$$

Let $t = H_0(x)$ and we have

$$\liminf_{x\to+\infty} \frac{\overline{H}(x)S(1)}{h_\gamma(\overline{H}_0(x))} \geq \gamma - 1 \qquad \text{a.s.} \tag{A.9}$$

since $H(x) = S(H_0(x))/S(1)$. Our assumptions $\int_0^\infty \rho(\mathrm{d}v) = +\infty$ and $\int_0^\infty (1 - e^{-v})\rho(\mathrm{d}v) < +\infty$ guarantee that $0 < S(1) < +\infty$ almost surely. Therefore, we conclude from (A.9) that almost surely for all such NRMI H,

$$\liminf_{x\to+\infty} \frac{\overline{H}(x)}{h_\gamma(\overline{H}_0(x))} > 0 \qquad \text{a.s.}$$

As $x \to +\infty$, the function $\overline{H}(x)/h_\gamma(\overline{H}_0(x))$ is almost surely lower bounded by a positive constant, which implies that the function $\log[\overline{H}(x)/h_\gamma(\overline{H}_0(x))]$ is almost surely lower bounded by a finite constant. Hence, it follows that

$$\liminf_{x\to+\infty} \frac{\log[\overline{H}(x)/h_\gamma(\overline{H}_0(x))]}{\log x} \geq 0 \qquad \text{a.s.} \tag{A.10}$$

For the right tail index of H, we can use (A.10) and the condition on $h_\gamma(\cdot)$ to obtain that

$$\begin{aligned}
\alpha_+(H) &= \liminf_{x\to+\infty} \frac{-\log \overline{H}(x)}{\log x} \\
&= \liminf_{x\to+\infty} \left\{ \frac{-\log h_\gamma(\overline{H}_0(x))}{\log x} + \frac{\log[h_\gamma(\overline{H}_0(x))/\overline{H}(x)]}{\log x} \right\} \\
&\le \liminf_{x\to+\infty} \frac{-\log h_\gamma(\overline{H}_0(x))}{\log x} - \liminf_{x\to+\infty} \frac{\log[\overline{H}(x)/h_\gamma(\overline{H}_0(x))]}{\log x} \\
&\le \liminf_{x\to+\infty} \frac{-\log h_\gamma(\overline{H}_0(x))}{\log x} = 0.
\end{aligned}$$

Therefore, $\alpha_+(H) = 0$.

(ii) For such $h(x)$ that satisfies (a)(b)(c), by similar argument as above, we apply part (ii) of Proposition A.1 and obtain that

$$\limsup_{x\to+\infty} \frac{\overline{H}(x)S(1)}{h(\overline{H}_0(x))} = 0 \qquad \text{a.s.}$$

which implies that almost surely for all such NRMI H,

$$\limsup_{x\to+\infty} \frac{\overline{H}(x)}{h(\overline{H}_0(x))} = 0 \qquad \text{a.s.}$$

Therefore, we have

$$\liminf_{x\to+\infty} \log \frac{h(\overline{H}_0(x))}{\overline{H}(x)} = +\infty \qquad \text{a.s.}$$

and hence

$$\liminf_{x\to+\infty} \frac{\log[h(\overline{H}_0(x))/\overline{H}(x)]}{\log x} \ge 0 \qquad \text{a.s.}$$

We finally combine this with the condition (c) and conclude that

$$\begin{aligned}
\alpha_+(H) &= \liminf_{x\to+\infty} \frac{-\log \overline{H}(x)}{\log x} \\
&= \liminf_{x\to+\infty} \left\{ \frac{-\log h(\overline{H}_0(x))}{\log x} + \frac{\log[h(\overline{H}_0(x))/\overline{H}(x)]}{\log x} \right\} \\
&\ge \liminf_{x\to+\infty} \frac{-\log h(\overline{H}_0(x))}{\log x} + \liminf_{x\to+\infty} \frac{\log[h(\overline{H}_0(x))/\overline{H}(x)]}{\log x} \\
&\ge \liminf_{x\to+\infty} \frac{-\log h(\overline{H}_0(x))}{\log x} = +\infty.
\end{aligned}$$

which means $\alpha_+(H) = +\infty$. □

Proof of Theorem 3.3. First we note that because $G_0(\mu, \sigma)$ is a continuous probability measure, if $G(\cdot, \cdot)$ is a homogenous NRMI with Lévy intensity $\rho(\mathrm{d}v)G_0(\mathrm{d}\mu, \mathrm{d}\sigma)$, then using the stick-breaking representation, we have that the two marginal distributions G_μ and G_σ are also homogenous NRMIs with Lévy intensities $\rho(\mathrm{d}v)G_{0,\mu}(\mathrm{d}\mu)$ and $\rho(\mathrm{d}v)G_{0,\sigma}(\sigma)$ respectively. Given the conclusion of Theorem 3.2, we have that if $G_{0,\mu}$ or $G_{0,\sigma}$ satisfies (i) of Theorem 3.2, then $\alpha_+(G_\mu) = 0$ or $\alpha_+(G_\sigma) = 0$; if $G_{0,\mu}$ or $G_{0,\sigma}$ satisfies (ii) of Theorem 3.2, then $\alpha_+(G_\mu) = +\infty$ or $\alpha_+(G_\sigma) = +\infty$.

Since $k(\cdot)$ has part of the support in $\mathbb{R}^+$, $E_K X_+^m > 0$ for any $m > 0$. We will examine the existence of moments $E_F X_+^m$ with F from Model 3.1 for any $m > 0$, and use Lemma A.1 to determine $\alpha_+(F)$. Similar to the proof of Theorem 3.1, we can analysis $E_F X_+^m$ using the lower bounds and the upper bound from Lemma A.3.

(i) If $\alpha_+(G_\mu) = 0$, then $E_{G_\mu}\mu_+^m = +\infty$ for all $m > 0$. Also note that $\overline{K}(0) > 0$ and $E_K X_+^m > 0$ since $k(\cdot)$ has full support in $\mathbb{R}$. Therefore, by the lower bound (A.5), $E_F X_+^m = +\infty$ for all $m > 0$ since $\overline{K}(0) > 0$, $E_K X_+^m > 0$, and $E_{G_{\mu,\sigma}}[\sigma^m I(\mu \geq 0)] \geq 0$. This implies $\alpha_+(F) = 0$ by Lemma A.1.

(ii) If $\alpha_+(G_\mu) = +\infty$ and $\alpha_+(G_\sigma) = 0$, then for all $m > 0$, $E_{G_\mu}\mu_+^m < +\infty$ and $E_{G_\sigma}\sigma^m = +\infty$. Because we have assumed that $G_{0,\mu}$ is symmetric about zero, this implies that $E_{G_\mu}\mu_-^m < +\infty$ for all $m > 0$. Also $E_K X_+^m > 0$ for all $m > 0$. Therefore by the lower bound (A.6), $E_F X_+^m = +\infty$. This again implies $\alpha_+(F) = 0$ by Lemma A.1.

(iii) If $\alpha_+(G_\mu) = +\infty$ and $\alpha_+(G_\sigma) = +\infty$, then for all $m > 0$, $E_{G_\mu}\mu_+^m < +\infty$ and $E_{G_\sigma}\sigma^m < +\infty$. This can be further separated into three scenarios: (a) $\alpha_+(K) \in (0, +\infty)$, then if $m \in (0, +\infty)$ and $m < \alpha_+(K)$, $E_K X_+^m < +\infty$ and $E_F X_+^m < +\infty$ by the upper bound (A.7); if $m \in (0, +\infty)$ and $m > \alpha_+(K)$, $E_K X_+^m = +\infty$ and $E_F X_+^m = +\infty$ by the lower bound (A.6). Hence, $\alpha_+(F) = \alpha_+(K)$ by Lemma A.1. (b) $\alpha_+(K) = 0$, then $E_K X_+^m = +\infty$ for all $m \in (0, +\infty)$ and $E_F X_+^m = +\infty$ for all $m \in (0, +\infty)$ by the lower bound in (A.6). (c) $\alpha_+(K) = +\infty$, then $E_K X_+^m < +\infty$ for all $m \in (0, +\infty)$ and $E_F X_+^m < +\infty$ for all $m \in (0, +\infty)$ by the upper bound in (A.7). We conclude that in all three scenarios $\alpha_+(F) = \alpha_+(K)$.

The results from different scenarios can be summarized as

$$\alpha_+(F) = \min\{\alpha_+(G_\mu), \alpha_+(G_\sigma), \alpha_+(K)\},$$

which is always a fixed number. Therefore, $\alpha_+(F)$ is almost surely a singleton, if both $G_{0,\mu}(\mu)$ and $G_{0,\sigma}(\sigma)$ satisfy either (i) or (ii) in Theorem 3.2. □

Proof of Theorem 3.4. We will show that for a measure H_0 on $\mathbb{R}$

(a) If $\limsup_{x\to+\infty} \overline{H}_0(x) \cdot (\log x / \log\log x) = +\infty$, then part (i) of Theorem 3.2 holds;
(b) If $\limsup_{x\to+\infty} \overline{H}_0(x) \cdot [(\log x) \cdot (\log\log x)^\delta] = 0$ for some $\delta > 1$, then part (ii) of Theorem 3.2 holds.

If both $\overline{G}_{0,\mu}$ and $\overline{G}_{0,\sigma}$ satisfy either (a) or (b), that is, we can replace H_0 with $\overline{G}_{0,\mu}$ and $\overline{G}_{0,\sigma}$, then the right tail indices of G_μ and G_σ are either 0 or $+\infty$, and the conclusion of Theorem 3.4 follows directly from Theorem 3.3.

To show (a) and (b), we use the similar arguments as in Doss and Sellke [12]. We note that a cdf $H(x)$ on $\mathbb{R}$ drawn from $\mathrm{DP}(a, H_0)$ can be written as a normalized Gamma process with Lévy

intensity $\rho(\mathrm{d}v)H_0(\mathrm{d}x) = av^{-1}e^{-v}\,\mathrm{d}v H_0(\mathrm{d}x)$. The Laplace exponent for ρ is $\Psi(s) = a\log(1+s)$ and its inversion is $\Psi^{-1}(u) = e^{u/a} - 1$. Thus for any given $\gamma > 1$, the function (3.3) in Proposition A.1 is given by

$$h_\gamma(x) = \frac{\log|\log x|}{\exp(\frac{\gamma\log|\log x|}{ax}) - 1}.$$

We have $\lim_{x\to 0+} h_\gamma(x) = 0$, and $h_\gamma(x) \in (0, 1/2)$ for $x \in [0, \varepsilon)$ for small enough $\varepsilon > 0$.

Now by the condition $\limsup_{x\to+\infty} \overline{H}_0(x)/(\log\log x/\log x) = +\infty$, there exists a positive sequence x_j that increases to $+\infty$ as $j \to +\infty$, such that for any $C > 2$, $1/16 > \overline{H}_0(x_j) > C\log\log x_j/\log x_j$ and $x_j > \exp(C^2)$ as long as $j > J(C)$ for some large integer $J(C)$. Therefore, for all $j > J(C)$,

$$\begin{aligned}
&\frac{-\log h_\gamma(\overline{H}_0(x_j))}{\log x_j} \\
&= \frac{\log[\exp(\frac{\gamma\log|\log\overline{H}_0(x_j)|}{a\overline{H}_0(x_j)}) - 1] - \log\log|\log\overline{H}_0(x_j)|}{\log x_j} \le \frac{\gamma\log|\log\overline{H}_0(x_j)|}{a\overline{H}_0(x_j)\log x_j} \\
&\le \frac{\gamma\log|\log\log x_j - \log C - \log\log\log x_j|}{aC\log\log x_j} \le \frac{\gamma\log\log\log x_j}{aC\log\log x_j}.
\end{aligned}$$

As $j \to +\infty$, this upper bound converges to 0. Together with the fact that $h_\gamma(x) \in (0, 1/2)$ for $x \in [0, \varepsilon)$, we obtain that $\liminf_{x\to+\infty} -\log h_\gamma(\overline{H}_0(x))/\log x = 0$. This is exactly the condition in part (i) of Theorem 3.2. Thus (a) is proved.

For (b), we set $h(x) = \exp[-(x|\log x|^{\delta'})^{-1}]$ for some $1 < \delta' < \delta$. This function is convex in $[0, \varepsilon)$ for small enough $\varepsilon > 0$. It also satisfies $\lim_{x\to 0+} h(x) = 0$, and $h(x) \in (0, 1/2)$ for $x \in [0, \varepsilon)$ for small enough $\varepsilon > 0$. Due to the lower and upper bounds $ae^{-1}\log 1/u \le \rho[u, +\infty) \le a\log(1/u) + ae^{-1}$ (see Doss and Sellke [12]) and $\delta' > 1$, we have $\int_0^\varepsilon \rho[h(x), +\infty)\,\mathrm{d}x < +\infty$. Furthermore, if $\limsup_{x\to+\infty} \overline{H}_0(x)\cdot[(\log x)\cdot(\log\log x)^\delta] = 0$, then for any $C > 2$ and all sufficiently large x, $\overline{H}_0(x) < \min(1/[C(\log x)\cdot(\log\log x)^\delta], 1/2)$. Therefore for sufficiently large x,

$$\begin{aligned}
\frac{-\log h(\overline{H}_0(x))}{\log x} &= \frac{1}{\overline{H}_0(x)|\log\overline{H}_0(x)|^{\delta'}\log x} \ge \frac{C(\log x)\cdot(\log\log x)^\delta}{\{\log[C\log x\cdot(\log\log x)^\delta]\}^{\delta'}\log x} \\
&\ge \frac{C(\log x)(\log\log x)^\delta}{(2\log\log x)^{\delta'}\log x} = \frac{C}{2^{\delta'}}(\log\log x)^{\delta-\delta'} \to +\infty,
\end{aligned}$$

which implies that $\liminf_{x\to+\infty} -\log h(\overline{H}_0(x))/\log x = +\infty$. This is exactly the condition in part (ii) of Theorem 3.2. Thus, (b) is proved. □

Proof of Theorem 3.5. We will show that for a measure H_0 on $\mathbb{R}$

(a) If $\limsup_{x\to+\infty} \overline{H}_0(x)\cdot x^\delta = +\infty$ for all $\delta > 0$, then part (i) of Theorem 3.2 holds;
(b) If $\limsup_{x\to+\infty} \overline{H}_0(x)\cdot x^\delta = 0$ for all $\delta > 0$, then part (ii) of Theorem 3.2 holds.

If both $\overline{G}_{0,\mu}$ and $\overline{G}_{0,\sigma}$ satisfy either (a) or (b), that is, we can replace H_0 with $\overline{G}_{0,\mu}$ and $\overline{G}_{0,\sigma}$, then the right tail indices of G_μ and G_σ are either 0 or $+\infty$, and the conclusion of Theorem 3.5 follows directly from Theorem 3.3.

To show (a), we note that for the Lévy process with intensity $\rho(\mathrm{d}v) = \frac{a}{\Gamma(1-\kappa)} v^{-\kappa-1} e^{-\tau v}\,\mathrm{d}v$, its Laplace exponent is $\Psi(s) = \frac{a}{\kappa}[(s+\tau)^\kappa - \tau^\kappa]$, and its inverse is $\Psi^{-1}(u) = [\kappa u/a + \tau^\kappa]^{1/\kappa} - \tau$. Thus for any given $\gamma > 1$, the function (3.3) is given by

$$h_\gamma(x) = \frac{\log|\log x|}{[\frac{\kappa \log|\log x|}{ax} + \tau^\kappa]^{1/\kappa} - \tau}.$$

We have $\lim_{x\to 0+} h_\gamma(x) = 0$, and $h_\gamma(x) \in (0, 1/2)$ for $x \in [0, \varepsilon)$ for small enough $\varepsilon > 0$.

Now by the condition $\limsup_{x\to+\infty} \overline{H}_0(x) \cdot x^\delta = +\infty$ for all $\delta > 0$, we have the following conclusion: for any given $\delta > 0$, there exists a positive sequence x_j that increases to $+\infty$ as $j \to +\infty$, such that $x_j > 16$ and $\min(\kappa \log\log 16/(a\tau^\kappa), 1/16) > \overline{H}_0(x_j) > x_j^{-\delta}$ as long as $j > J$ for some large integer J. Such choice of x_j guarantees that $\log\log(\delta \log x_j) > \log\log|\log \overline{H}_0(x_j)| > \log\log\log 16 > 0$, and

$$\frac{\kappa \log|\log \overline{H}_0(x_j)|}{a\overline{H}_0(x_j)} > \frac{\kappa \log\log 16}{a\overline{H}_0(x_j)} > \tau^\kappa.$$

Therefore for all $j > J$,

$$\begin{aligned}
\frac{-\log h_\gamma(\overline{H}_0(x_j))}{\log x_j} &\le \frac{\kappa^{-1}\log[\frac{2\kappa\log|\log\overline{H}_0(x_j)|}{a\overline{H}_0(x_j)}] - \log\log|\log\overline{H}_0(x_j)|}{\log x_j} \\
&= \frac{-\kappa^{-1}\log\overline{H}_0(x_j) + (\kappa^{-1}-1)\log\log|\log\overline{H}_0(x_j)| + \kappa^{-1}\log(2\kappa/a)}{\log x_j} \\
&\le \kappa^{-1}\delta + \big(\kappa^{-1}-1\big)\frac{\log\log(\delta\log x_j)}{\log x_j} + \frac{\log(2\kappa/a)}{\kappa\log x_j}.
\end{aligned}$$

In the last display, the second and the third terms converge to zero as $j \to +\infty$. The first term can be made arbitrarily small if δ is made small. Therefore, we have shown that $\liminf_{x\to+\infty} -\log h_\gamma(\overline{H}_0(x))/\log x = 0$. Thus (a) is proved.

For (b), we have the following bound for $\rho[u, +\infty)$:

$$\begin{aligned}
\rho[u,+\infty) &= \frac{a}{\Gamma(1-\kappa)}\int_u^1 t^{-\kappa-1}e^{-\tau t}\,\mathrm{d}t + \frac{a}{\Gamma(1-\kappa)}\int_1^\infty t^{-\kappa-1}e^{-\tau t}\,\mathrm{d}t \\
&\le \frac{a}{\Gamma(1-\kappa)}\int_u^1 t^{-\kappa-1}\,\mathrm{d}t + \frac{a}{\Gamma(1-\kappa)}\int_1^\infty e^{-\tau t}\,\mathrm{d}t \\
&= \frac{a}{\kappa\Gamma(1-\kappa)}\left(\frac{1}{u^\kappa} - 1\right) + \frac{ae^{-\tau}}{\tau\Gamma(1-\kappa)}.
\end{aligned}$$

Therefore if $\int_0^{\varepsilon} \frac{1}{h(x)^{\kappa}}\,\mathrm{d}x < +\infty$, then $\int_0^{\varepsilon} \rho[h(x), +\infty)\,\mathrm{d}x < +\infty$. Let $h(x) = (x|\log x|^{\eta})^{1/\kappa}$ for some $\eta > 1$. For $\kappa \in (0,1)$, this $h(x)$ is convex and increasing in $[0,\varepsilon)$ and satisfies $\lim_{x\to 0+} h(x) = 0$, and $h(x) \in (0, 1/2)$ for $x \in [0,\varepsilon)$ if ε is sufficiently small.

On the other hand, if $\limsup_{x\to+\infty} \overline{H}_0(x) \cdot x^{\delta} = 0$ for all $\delta > 0$, then $\overline{H}_0(x) \le x^{-\delta}$ for any given δ for all sufficiently large x, which implies that

$$\frac{-\log h(\overline{H}_0(x))}{\log x} \ge \frac{-\log h(x^{-\delta})}{\log x} = \frac{\delta}{\kappa} - \frac{\eta \log(\delta \log x)}{\kappa \log x}.$$

On the right-hand side of the last display, the second term converges to zero as $x \to +\infty$. The first term can be made arbitrarily large if δ is made large. Therefore, we have shown that $\liminf_{x\to+\infty} -\log h(\overline{H}_0(x))/\log x = +\infty$. Thus, (b) is proved. □

Proof of Theorem 4.2. Let s_n be a positive sequence such that $B_n \prec s_n \prec \overline{\alpha}_n^{-1} \log n$, whose existence is guaranteed by Condition (iv). For $\varepsilon > 0$, we define the test $\Phi_n = I(|\hat{\alpha}_{s_n} - \alpha_{0+}| \ge \varepsilon/2)$ with $\hat{\alpha}_{s_n}$ given by (4.2). Let $p_{s_n} = P_F(X > e^{s_n})$ be the population mean of $\hat{p}_{s_n}$ and let $\alpha_{s_n} = \log(p_{s_n}) - \log(p_{s_n+1})$. Note that p_{s_n} and α_{s_n} implicitly depend on F. We complete the proof in two steps.

Step 1: Show $E_{F_0}\Phi_n \to 0$ as $n \to \infty$.

We have

$$\begin{aligned} E_{F_0}\Phi_n &= P_{F_0}\left(|\hat{\alpha}_{s_n} - \alpha_{0+}| \ge \frac{\varepsilon}{2}\right) \\ &\le P_{F_0}\left(|\hat{\alpha}_{s_n} - \alpha_{s_n}| \ge \frac{\varepsilon}{4}\right) + P_{F_0}\left(|\alpha_{s_n} - \alpha_{0+}| \ge \frac{\varepsilon}{4}\right). \end{aligned} \tag{A.11}$$

The first term in (A.11) can be bounded by Lemma 2 and equation (4.2) of Carpentier and Kim [6]:

$$P_{F_0}\left(|\hat{\alpha}_{s_n} - \alpha_{s_n}| \ge \frac{\varepsilon}{4}\right) \le 2\exp\left(-\frac{n p_{s_n+1}\varepsilon^2}{576}\right), \tag{A.12}$$

where $p_{s_n+1} = P_{F_0}(X > e^{s_n+1}) = e^{-\alpha_{0+}(s_n+1)} L_0(e^{s_n+1})$. Since L_0 is slowly varying, as $n \to \infty$, eventually $L_0(e^{s_n+1}) \ge e^{-\delta(s_n+1)}$ for arbitrarily small $\delta > 0$. By Condition (iv) of Theorem 4.2, $s_n \prec \log n/\alpha_{0+}$ and hence $n p_{s_n+1} \ge \exp(\log n - (\alpha_{0+} + \delta)(s_n + 1)) \to +\infty$, which implies that the righthand side of (A.12) goes to zero as $n \to \infty$.

The second term in (A.11) is not stochastic. We have

$$\begin{aligned} |\alpha_{s_n} - \alpha_{0+}| &= \left|\log(p_{s_n}) - \log(p_{s_n+1}) - \alpha_{0+}\right| \\ &= \left|\log \frac{L_0(e^{s_n})}{L_0(e^{s_n+1})}\right| \to 0, \end{aligned}$$

because L_0 is slowly varying and $s_n \to \infty$. Therefore both terms on the righthand side of (A.11) converge to zero as $n \to \infty$.

Step 2: Show $\sup_{F\in B^c_{\alpha+}(F_0,\varepsilon)\cap\mathcal{F}_n} E_F(1-\Phi_n)\to 0$ as $n\to\infty$, where we let $\mathcal{F}_n=\mathcal{F}_{1n}\cap\mathcal{F}_{2n}\cap\mathcal{F}_{3n}$. By Conditions (i)–(iii), it is clear that $\Pi_n(\mathcal{F}_n^c)\le\Pi_n(\mathcal{F}_{1n}^c)+\Pi_n(\mathcal{F}_{2n}^c)+\Pi_n(\mathcal{F}_{3n}^c)\le e^{-c_1 n}+e^{-c_2 n}+e^{-c_3 n}\le e^{-c' n}$ where $c'=\min(c_1,c_2,c_3)/2$.

For every $F\in B^c_{\alpha+}(F_0,\varepsilon)\cap\mathcal{F}_n$, we have $|\alpha_+(F)-\alpha_{0+}|>\varepsilon$. Therefore,

$$\begin{aligned}
E_F(1-\Phi_n) &= P_F\left(|\hat\alpha_{s_n}-\alpha_{0+}|\le\frac{\varepsilon}{2}\right)\\
&= P_F\left(|\hat\alpha_{s_n}-\alpha_{0+}|\le\frac{\varepsilon}{2},\,\left|\hat\alpha_{s_n}-\alpha_+(F)\right|<\frac{\varepsilon}{2}\right)\\
&\quad + P_F\left(|\hat\alpha_{s_n}-\alpha_{0+}|\le\frac{\varepsilon}{2},\,\left|\hat\alpha_{s_n}-\alpha_+(F)\right|\ge\frac{\varepsilon}{2}\right)\\
&\le P_F\left(\left|\alpha_+(F)-\alpha_{0+}\right|<\varepsilon\right)+P_F\left(\left|\hat\alpha_{s_n}-\alpha_+(F)\right|\ge\frac{\varepsilon}{2}\right)\\
&= P_F\left(\left|\hat\alpha_{s_n}-\alpha_+(F)\right|\ge\frac{\varepsilon}{2}\right)\\
&\le P_F\left(|\hat\alpha_{s_n}-\alpha_{s_n}|\ge\frac{\varepsilon}{4}\right)+P_F\left(\left|\alpha_{s_n}-\alpha_+(F)\right|\ge\frac{\varepsilon}{4}\right).
\end{aligned}\tag{A.13}$$

We only need to show that both terms on the right-hand side of (A.13) converge to zero uniformly over all $F\in B^c_{\alpha+}(F_0,\varepsilon)\cap\mathcal{F}_n$ as $n\to\infty$. For a fixed F, the first term can be bounded by Lemma 2 and equation (4.2) of Carpentier and Kim [6] again as

$$P_F\left(|\hat\alpha_{s_n}-\alpha_{s_n}|\ge\frac{\varepsilon}{4}\right)\le 2\exp\left(-\frac{np_{s_n+1}\varepsilon^2}{576}\right).$$

To obtain uniform convergence for the right-hand side, we only need the quantity np_{s_n+1} to be uniformly bounded below for all $F\in B^c_{\alpha+}(F_0,\varepsilon)\cap\mathcal{F}_n$. Using Conditions (i)–(iii), we can obtain the following uniform lower bound:

$$\begin{aligned}
np_{s_n+1} &= ne^{-\alpha_+(F)(s_n+1)}L_F\left(e^{s_n+1}\right)=ne^{-\alpha_+(F)(s_n+1)}L_F(x_0)\exp\left(\int_{x_0}^{e^{s_n+1}}\frac{h_F(t)}{t}\,\mathrm{d}t\right)\\
&\ge\exp\left(\log n-\overline{\alpha}_n(s_n+1)-c_L\log n-\int_{x_0}^{e^{s_n+1}}\frac{B_n}{t(\log t)^{1+\tau_n}}\,\mathrm{d}t\right)\\
&=\exp\left((1-c_L)\log n-\overline{\alpha}_n(s_n+1)-\frac{B_n}{\tau_n(\log x_0)^{\tau_n}}+\frac{B_n}{\tau_n(s_n+1)^{\tau_n}}\right)\\
&\ge\exp\left((1-c_L)\log n-\overline{\alpha}_n(s_n+1)-\frac{B_n}{\tau_n}\right),
\end{aligned}$$

where we use $x_0\ge e$ and hence $\log x_0\ge 1$ in the last inequality. Condition (i) says $1-c_L>0$. By our choice of s_n, we have $\log n\succ\overline{\alpha}_n(s_n+1)$, and Condition (iv) implies $\log n\succ B_n/\tau_n$.

Therefore, we have obtained that uniformly over all $F \in B^c_{\alpha+}(F_0, \varepsilon) \cap \mathcal{F}_n$, $P_F(|\hat{\alpha}_{s_n} - \alpha_{s_n}| \geq \frac{\varepsilon}{4})$ converges to zero as $n \to \infty$.

For the second term in (A.13), we have

$$\begin{aligned}
\left|\alpha_{s_n} - \alpha_+(F)\right| &= \left|\log p_{s_n} - \log p_{s_n+1} - \alpha_+(F)\right| \\
&= \left|\log e^{-\alpha_+(F)s_n} L_F\left(e^{s_n}\right) - \log e^{-\alpha_+(F)(s_n+1)} L_F\left(e^{s_n+1}\right) - \alpha_+(F)\right| \\
&= \left|\log L_F\left(e^{s_n}\right) - \log L_F\left(e^{s_n+1}\right)\right| = \left|\int_{e^{s_n}}^{e^{s_n+1}} \frac{h_F(x)}{x}\,\mathrm{d}x\right| \\
&\leq \int_{e^{s_n}}^{e^{s_n+1}} \frac{\overline{h}_n(x)}{x}\,\mathrm{d}x = \int_{e^{s_n}}^{e^{s_n+1}} \frac{B_n}{x(\log x)^{1+\tau_n}}\,\mathrm{d}x \\
&= \frac{B_n}{\tau_n s_n^{\tau_n}}\left[1 - \left(\frac{s_n}{1+s_n}\right)^{\tau_n}\right] = \frac{B_n}{\tau_n s_n^{\tau_n}}\left[1 - \exp\left(-\tau_n \log\frac{1+s_n}{s_n}\right)\right] \\
&\leq \frac{B_n}{\tau_n s_n^{\tau_n}} \cdot \tau_n \log\left(1 + \frac{1}{s_n}\right) \leq \frac{B_n}{s_n^{1+\tau_n}},
\end{aligned}$$

where we have used $1 - e^{-t} \leq t$ for $t > 0$ and $\log(1+t) \leq t$ for $t > 0$. Since $1 \preceq B_n \prec s_n$, we have $B_n / s_n^{1+\tau_n} \to 0$ as $n \to \infty$. Therefore, the probability $P_F(|\alpha_{s_n} - \alpha_+(F)| \geq \frac{\varepsilon}{4})$ is zero for all large n uniformly over all $F \in B^c_{\alpha+}(F_0, \varepsilon) \cap \mathcal{F}_n$. □

Acknowledgements

We thank Professor Jayanta Ghosh for a comment on the challenges of Bayesian asymptotics for heavy-tailed densities, which served as inspiration for this work. We are grateful to the referees, Associate Editor, and Editor for their comments and suggestions. LL would like to thank Dong Quan Nguyen for useful discussions on the measurability of the tail index neighborhood. This work was partially supported by National University of Singapore start-up grant R155000172133, NSF grants IIS 1663870 and DMS CAREER 1654579.

Supplementary Material

Supplement to "On posterior consistency of tail index for Bayesian kernel mixture models" (DOI: 10.3150/18-BEJ1043SUPP; .pdf). We provide the technical proofs of Theorems 4.4 and 4.5 in Barrios et al. [35].

References

[1] Barrios, E., Lijoi, A., Nieto-Barajas, L.E. and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statist. Sci.* **28** 313–334. MR3135535

[2] Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004). *Statistics of Extremes*: *Theory and Applications*. *Wiley Series in Probability and Statistics*. Chichester: Wiley. With contributions from Daniel De Waal and Chris Ferro. MR2108013
[3] Bottolo, L., Consonni, G., Dellaportas, P. and Lijoi, A. (2003). Bayesian analysis of extreme values by mixture modeling. *Extremes* **6** 25–47. MR2021591
[4] Boucheron, S. and Thomas, M. (2015). Tail index estimation, concentration and adaptivity. *Electron*. *J*. *Stat*. **9** 2751–2792. MR3435810
[5] Breiman, L. (1965). On some limit theorems similar to the arc-sin law. *Teor*. *Veroyatn*. *Primen*. **10** 351–360. MR0184274
[6] Carpentier, A. and Kim, A.K.H. (2015). Adaptive and minimax optimal estimation of the tail coefficient. *Statist*. *Sinica* **25** 1133–1144. MR3410301
[7] Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Rev*. **51** 661–703. MR2563829
[8] Cormann, U. and Reiss, R.-D. (2009). Generalizing the Pareto to the log-Pareto model and statistical inference. *Extremes* **12** 93–105. MR2480725
[9] de Haan, L. and Resnick, S.I. (1980). A simple asymptotic estimate for the index of a stable distribution. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **42** 83–87. MR0567205
[10] Diebolt, J., El-Aroui, M.-A., Garrido, M. and Girard, S. (2005). Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling. *Extremes* **8** 57–78. MR2201901
[11] do Nascimento, F.F., Gamerman, D. and Lopes, H.F. (2012). A semiparametric Bayesian approach to extreme value estimation. *Stat*. *Comput*. **22** 661–675. MR2865043
[12] Doss, H. and Sellke, T. (1982). The tails of probabilities chosen from a Dirichlet prior. *Ann*. *Statist*. **10** 1302–1305. MR0673666
[13] Drees, H. (1998). Optimal rates of convergence for estimates of the extreme value index. *Ann*. *Statist*. **26** 434–448. MR1608148
[14] Drees, H. (2001). Minimax risk bounds in extreme value theory. *Ann*. *Statist*. **29** 266–294. MR1833966
[15] Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J*. *Amer*. *Statist*. *Assoc*. **90** 577–588. MR1340510
[16] Favaro, S. and Teh, Y.W. (2013). MCMC for normalized random measure mixture models. *Statist*. *Sci*. **28** 335–359. MR3135536
[17] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann*. *Statist*. **1** 209–230. MR0350949
[18] Frigessi, A., Haug, O. and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes* **5** 219–235. MR1995776
[19] Fristedt, B.E. (1967). Sample function behavior of increasing processes with stationary, independent increments. *Pacific J*. *Math*. **21** 21–33. MR0210190
[20] Fristedt, B.E. and Pruitt, W.E. (1971). Lower functions for increasing random walks and subordinators. *Z*. *Wahrsch*. *Verw*. *Gebiete* **18** 167–182. MR0292163
[21] Fúquene Patiño, J.A. (2015). A semi-parametric Bayesian extreme value model using a Dirichlet process mixture of gamma densities. *J*. *Appl*. *Stat*. **42** 267–280. MR3276944
[22] Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann*. *Statist*. **27** 143–158. MR1701105
[23] Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann*. *Statist*. **28** 500–531. MR1790007
[24] Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann*. *Statist*. **35** 697–723. MR2336864

[25] Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. *Springer Series in Statistics*. New York: Springer. MR1992245
[26] Green, P.J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Stat.* **28** 355–375. MR1842255
[27] Haeusler, E. and Teugels, J.L. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *Ann. Statist.* **13** 743–756. MR0790569
[28] Hall, P. (1982). On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc. Ser. B* **44** 37–42. MR0655370
[29] Hall, P. and Welsh, A.H. (1984). Best attainable rates of convergence for estimates of parameters of regular variation. *Ann. Statist.* **12** 1079–1084. MR0751294
[30] Hall, P. and Welsh, A.H. (1985). Adaptive estimates of parameters of regular variation. *Ann. Statist.* **13** 331–341. MR0773171
[31] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174. MR0378204
[32] James, L.F., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36** 76–97. MR2508332
[33] Kruijer, W., Rousseau, J. and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885
[34] Lange, K. (1973). Borel sets of probability measures. *Pacific J. Math.* **48** 141–161. MR0357723
[35] Li, C., Lin, L. and Dunson, D.B. (2019). Supplement to "On posterior consistency of tail index for Bayesian kernel mixture models." DOI:10.3150/18-BEJ1043SUPP.
[36] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 715–740. MR2370077
[37] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*. *Camb. Ser. Stat. Probab. Math.* **28** 80–136. Cambridge: Cambridge Univ. Press. MR2730661
[38] Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519
[39] MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23** 727–741. MR1293996
[40] Mason, D.M. (1982). Laws of large numbers for sums of extreme values. *Ann. Probab.* **10** 754–764. MR0659544
[41] Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804
[42] Novak, S.Y. (2014). Lower bounds to the accuracy of inference on heavy tails. *Bernoulli* **20** 979–989. MR3178524
[43] Pickands, J. III (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3** 119–131. MR0423667
[44] Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31** 560–585. Dedicated to the memory of Herbert E. Robbins. MR1983542
[45] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213
[46] Schwartz, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378
[47] Shen, W., Tokdar, S.T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441
[48] Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. New York: Wiley. MR0838963

[49] Stephenson, A. and Tawn, J. (2004). Bayesian inference for extremes: Accounting for the three extremal types. *Extremes* **7** 291–307. MR2212389

[50] Tokdar, S.T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā* **68** 90–110. MR2301566

[51] Tressou, J. (2008). Bayesian nonparametrics for heavy tailed distribution. Application to food risk assessment. *Bayesian Anal*. **3** 367–391. MR2407431

[52] Wang, Z., Rodriguez, A. and Kottas, A. (2012). A nonparametric mixture modeling framework for extreme value analysis. Technical report. https://www.soe.ucsc.edu/research/technical-reports/UCSC-SOE-11-26/download.

[53] Watanabe, T. (1960). A probabilistic method in Hausdorff moment problem and Laplace–Stieltjes transform. *J. Math. Soc. Japan* **12** 192–206. MR0120683

[54] Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat*. **2** 298–331. MR2399197

Received December 2016 and revised March 2018