Bayesian Analysis (2008) **3**, Number 2, pp. 367–392

Bayesian Nonparametrics for Heavy Tailed Distribution. Application to Food Risk Assessment

Jessica Tressou*

Abstract. Based on the fact that any heavy tailed distribution can be approximated by a possibly infinite mixture of Pareto distributions, this paper proposes two Bayesian methodologies tailored to infer on distribution tails belonging to the Fréchet domain of attraction. Firstly, a Bayesian Pareto based clustering procedure is developed, where the mixing distribution is chosen to be the classical conjugate prior of the Pareto distribution. This allows the grouping of n objects into a certain number of clusters according to their extremal behavior and also exhibits a new estimator for the tail index. Secondly, a nonparametric extension of the model based clustering is proposed in which the parameter of interest is the mixing distribution. Estimation of the tail probability is conducted using a Dirichlet process prior for the unknown mixing distribution. To illustrate, both methodologies are applied to simulated data sets and a real data set concerning dietary exposure to a mycotoxin called Ochratoxin A.

Keywords: Dirichlet process, Model Based clustering, Ochratoxin A, Tail index estimation.

1 Introduction

In the food risk analysis field, it is accepted that dietary exposure to a contaminant is heavy tailed or at least assumed to be from a conservative perspective, see Tressou et al. (2004). Indeed, dietary exposure to a given contaminant is defined as the quantity of the contaminant one individual ingests when he consumes foods that are more or less naturally contaminated. Different consumption behaviors yield different levels of exposure which may present a health risk if those levels are too high. One particular contaminant is generally present in more that one food so that different consumption behaviors can yield a high exposure. In a given population, different risk levels exist and clustering may be a powerful tool to describe that population. Yet, we do not know *a priori* how many clusters there are and we would like to define the similarity between individuals based on their extremal behavior. Food safety is now a crucial public health concern in many countries (for example, it is one of the thematic top priorities of the 7th European Research Framework program, see http://ec.europa.eu/research/fp7/). This topic naturally interfaces with various disciplines, such as biology, nutritional medicine, toxicology, and of course applied mathematics with the aim of developing rigorous methods for quantitative risk assessment. Scientific literature devoted to prob-

*INRA-Mét@risk, UR1204 Food Risk Analysis Methodologies, F75005, France and HKUST-ISMT, Hong Kong University of Science and Technology, Hong Kong, mailto:tressou@agroparistech.fr

DOI:10.1214/08-BA314

abilistic and statistical methods for the study of dietary exposure to food contaminants is progressively carving out a place in applied probability and statistics journals (see Bertail et al. 2008; van der Voet et al. 2007; Tressou 2006; Bertail and Tressou 2006; Edler et al. 2002; Gibney and van der Voet 2003; Gauchi and Leblanc 2002).

The main idea of this paper is that heavy tailed distributions can be represented as mixtures of Pareto distributions so that most, if not all, heavy tailed distributions can be expressed as (possibly infinite) mixtures of Pareto distributions, where the mixing occurs on both parameters of the Pareto distributions. Two Bayesian methodologies are thus proposed to estimate the different components of this mixture: a Bayesian model-based clustering approach (Fraley and Raftery 2002) and a Bayesian nonparametric mixture approach (Petrone and Raftery 1997; Green and Richardson 2001), following ideas exposited in Lau and Lo (2007). For both approaches, the kernel is defined to be a Pareto distribution while most applications are realized with a Gaussian kernel (Lau and Green 2007; Lau and Lo 2007) since we are specifically interested in these mixtures to model heavy tailed distribution. In recent years, parametric and nonparametric Bayesian approaches have been developped for extreme value analysis (Coles and Powell 1996; Frigessi et al. 2002; Bottolo et al. 2003; Stephenson and Tawn 2004; Diebolt et al. 2005; Kottas and Sansó 2007). In this paper, estimators of the tail index and tail probability are derived from the posterior distribution. The tail index estimator is compared to a standard estimator (the Hill estimator).

The paper is organized as follows. Section 2 gives some background about Extreme Value Theory and emphasizes that heavy tailed distributions can be approximated by mixtures of Pareto distributions. Section 3 gives the general principle of Bayesian model-based clustering as well as one MCMC algorithm to find the best partition (Gibbs WCR) and presents the Pareto-based clustering. Section 4 introduces two key results for the nonparametric extension of the model-based clustering and details the quantities one may infer when extremes are at stake. The last section is dedicated to the implementation of both methodologies on simulated data first, with empirical validation and understanding perspectives, and on data concerning the French population's exposure to Ochratoxin A (OTA) in a purely applied perspective.

2 Characterization of the maximum domain of attraction of the Fréchet distribution as a general mixture of Pareto distributions

In Extreme Value Theory, one major breakthrough is the Fisher-Tippett theorem stating that there are only three possible limiting distributions for the properly normalized maximum: the Gumbel, the Weibull and the Fréchet distributions. These laws are called extreme value distributions and each one corresponds to a special tail behavior: the Gumbel distribution is related to light-tailed distributions such as normal, log-normal or exponential distributions; the Weibull distribution to finite support distributions such as the uniform distribution and the Fréchet distribution to heavy-tailed distributions such as Pareto, Cauchy or Student distributions. The latter one is of prime interest in the food risk analysis context since the distribution of exposure to a contaminant is often assumed to be heavy-tailed (Tressou et al. 2004).

The usual characterization of the Fréchet maximum domain of attraction (MDA) is the following (Embrechts et al. 1999). For sufficiently large x, the tail probability, $\mathbb{P}(X > x)$, is approximately equal to $Cx^{-\alpha^*}L(x)$, where C and α^* are non negative constants and $L(.)$ is a slowly varying function, that is, a function satisfying the condition

$$\forall t > 0, \lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1.$$

In this setting, the estimation of α^* is crucial and has been studied a lot as α^{*-1} may be interpreted as a risk indicator. Indeed the higher the α^{*-1}, the higher the probability of exceeding a fixed level x. A well known estimator for α^{-1} is the Hill estimator based on the k largest observations of a sample (Hill 1975). If $X_{1,n} \leq \ldots \leq X_{n,n}$ denotes the order statistic associated to a sample $(X_1, \ldots, X_n)$ then the Hill estimator is defined for $k = 1, \ldots, n-1$ as

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \ln X_{n-i+1,n} - \ln X_{n-k,n}.$$

The Hill estimator is obtained as the conditional maximum likelihood estimator in the exact Pareto model ($L(x) = 1$), given the number k of extreme values. This is very sensitive to the choice of k. Indeed, its bias increases with k while its variance decreases. Several authors proposed bias correction using more or less explicit forms of the slowly varying function L, see for example Beirlant et al. (1999); Feuerverger and Hall (1999).

These slowly varying functions naturally appear when considering mixtures of Pareto distributions.

Let $f_{\alpha,\tau}$ and $F_{\alpha,\tau}$ denote the density and cumulative distribution function of the Pareto distribution with tail index parameter α and precision parameter τ, abbreviated by $\mathcal{P}(\alpha, \tau)$, *i.e.*

$$\begin{aligned} 1 - F_{\alpha,\tau}(x) &= (\tau x)^{-\alpha}\mathbf{1}_{(\tau x > 1)} + \mathbf{1}_{(\tau x \leq 1)} \\ f_{\alpha,\tau}(x) &= \alpha\tau(\tau x)^{-(\alpha+1)}\mathbf{1}_{(\tau x>1)}, \end{aligned} \tag{1}$$

where $\mathbf{1}_{(A)}$ is the indicator function, equal to 1 if A is true, 0 otherwise.

If G is an unknown mixing distribution over the two dimensional parameter space $\Theta_1 \times \Theta_2 \subseteq \mathbb{R}_+^2$, then the tail probability is

$$\mathbb{P}(X > x) = \int_{\Theta_1}\int_{\Theta_2} \mathbb{P}(X > x|\alpha, \tau)\, G(d\alpha, d\tau) = \int_{\Theta_1}\int_{\Theta_2} [1 - F_{\alpha,\tau}(x)]\, G(d\alpha, d\tau). \tag{2}$$

In the case of a discrete mixing distribution, if $(\alpha, \tau) = (\alpha_j, \tau_j)$ with probability w_j,

$j = 1, \dots, J$, such that $\sum_{j=1}^{J} w_j = 1$, and $\alpha_1 \le \dots \le \alpha_J$, then

$$\mathbb{P}\left(X > x\right) = \sum_{j=1}^{J} w_j \times (\tau_j x)^{-\alpha_j} \mathbf{1}_{(\tau_j x > 1)} + \mathbf{1}_{(\tau_j x \le 1)} \sim_{x \to \infty} C x^{-\alpha^*} \left(1 + \sum_{j=2}^{J} D_{j-1} x^{-\beta_{j-1}}\right), \tag{3}$$

where $\alpha^* = \min_{j=1,\dots J} \alpha_j (= \alpha_1)$ and the (D_j, β_j) and C are non negative constants such that $\beta_1 \le \dots \le \beta_{J-1}$. More precisely, $C = w_1 \tau_1^{-\alpha_1}$, and for $j = 2, \dots, J$, $\beta_{j-1} = \alpha_j - \alpha_1$ and $D_{j-1} = w_j \tau_j^{-\alpha_j} / w_1 \tau_1^{-\alpha_1}$. The quantity $L(x) = (1 + \sum_{j=2}^{J} D_{j-1} x^{-\beta_{j-1}})$ is a slowly varying function, meaning that any discrete mixture of Pareto distributions is of the Fréchet type. Moreover, a natural estimator of the tail index α is the minimum tail index parameter of the Pareto components of the mixture.

This argument does not prove any relation between the Fréchet MDA and the set of all possibly infinite mixtures of Pareto distributions but motivates the approximation of the Fréchet MDA with such mixtures.

3 Bayesian model based clustering

3.1 General principle

For statistical clustering of n objects, it is assumed that the numerical measurements, $\mathbf{x} = (x_1, \dots, x_n)$, of the n objects have a joint model density given a certain partition of the n objects. Given a partition $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$ of the indices $\{1, \dots, n\}$ of the n objects, the measurements of the objects are modeled by a *classification likelihood* that, given $\mathbf{p}$, has a product form

$$f(\mathbf{x}|\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} k(x_i, i \in C_j),$$

where $k(x_i, i \in C_j)$ is the joint density of the measurements for objects in cluster C_j, $k(x_i, i \in \{1, \dots, n\}) = k(\mathbf{x})$ being the joint density of the whole data $\mathbf{x}$. Typically, in a Bayesian framework, these joint densities result from a former parametric inference in which, given an unknown parameter θ with prior distribution $\pi_0(\theta)d\theta$, the x_i are assumed to be i.i.d. from a model density f_θ. Then $k(x_i, i \in C_j)$ is just the normalization constant of the posterior distribution of θ given the measurements of cluster C_j, given by

$$k(x_i, i \in C_j) = \int \prod_{i \in C_j} f_\theta(x_i) \pi_0(\theta) d\theta.$$

Alternatively, they can be directly assigned to some chosen function of the $x_i, i \in C_j$ that measures the homogeneity within the cluster, see Lau and Green (2007) for more details. When the first option is retained, direct calculation of the $k(x_i, i \in C_j)$ is easily achievable if the prior for θ is chosen to be the conjugate prior for the model density f_θ. Most applications of model-based clustering relate to the Normal model, where the conjugate prior is the Gamma-Normal distribution: this results in marginal t-densities for the $k(x_i, i \in C_j)$, see Lau and Lo (2007) for a gene clustering application. In the present paper, we focus on the Pareto kernel as detailed in Section 3.3.

As soon as the "classification likelihood" $f(\mathbf{x}|\mathbf{p})$ is chosen, the partition is the unknown parameter for which a prior-posterior analysis is required. A conjugate prior for $\mathbf{p}$ can be any distribution that has the product form, namely

$$\pi(\mathbf{p}) \propto \prod_{j=1}^{n(\mathbf{p})} g(C_j). \tag{4}$$

In this case, the posterior distribution of $\mathbf{p}$ given the data is also of the product form

$$\pi(\mathbf{p}|\mathbf{x}) \propto \prod_{j=1}^{n(\mathbf{p})} g^*(C_j),$$

where $g^*(C_j) = g(C_j) \times k(x_i, i \in C_j)$.

Finally, an estimator of the optimal clustering is the one that maximizes the posterior distribution, which can be approximated by MCMC techniques (the usual Gibbs sampler is used in this paper and described in Section 3.2). Lau and Green (2007) also propose other estimators based on the minimization of loss functions.

For the prior choice, the only requirement is the product form given in Eq. (4) so that many prior distributions can be used. A very convenient one is the Chinese Restaurant Process with parameter e_0, CRP(e_0), for which $g(C_j) = e_0 \times (e_j - 1)!$, where e_j is the size of cluster C_j. The parameter e_0 can be interpreted as the expected number of clusters.

3.2 Implementation: Gibbs Weighed Chinese Restaurant Process

In this section, the Gibbs sampler used in the application is described for a CRP(e_0) prior distribution on partitions. This algorithm is precisely the computational strategy described for normal kernels in MacEachern (1994) and more generally in MacEachern (1998). The first work on Gibbs sampling for such models is Escobar (1994) (see also his 1988 dissertation). This is only one of several possible algorithms (see Lau and Lo 2007; Lau and Green 2007; Heard et al. 2006; Quintana and Iglesias 2003, and the references therein).

Algorithm 1. Choose an initial partition $\mathbf{p}_0$ (the one with n clusters $\mathbf{p}_0 = \{\{1\}, ..., \{n\}\}$ is the default choice).
Then, repeat $L + M$ times (L times for burn in / warm up and M times for estimation of any function $h(\mathbf{p})$) the following Gibbs cycle:

For $i = 1, ...n$, do

- Remove $\{i\}$ from the current partition $\mathbf{p}$ of $\{1, ..., n\}$ to get a partition $\mathbf{p}^{(-i)}$ of $\{1, ..., i-1, i+1, ..., n\}$ ($n - 1$ elements)

- $\{i\}$ is then assigned to the cluster j, $j = 1, ..., n(\mathbf{p}^{(-i)})$ with probability proportional to

$$\frac{g^*(C_j \cup \{i\})}{g^*(C_j)} = e_j \times \frac{k(x_l, l \in C_j \cup \{i\})}{k(x_l, l \in C_j)} = e_j \times k(x_i|x_l, l \in C_j) \tag{5}$$

  and to a new one with probability proportional to $e_0 \times k(x_i)$.

The assignment of $\{n\}$ completes a Gibbs cycle and the last partition is stored and used as the initial one in the next cycle.

The $L + M + 1$ partitions, $\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_L, \mathbf{p}_{L+1}, ..., \mathbf{p}_{L+M}$, are then used to compute estimators for quantities such as $\xi = \sum_{\mathbf{p}} \pi(\mathbf{p}|\mathbf{x})h(\mathbf{p})$ or $\mathbf{p}^* = \arg\max_{\mathbf{p}} \pi(\mathbf{p}|\mathbf{x})$, namely

$$\widetilde{\xi_M} = \frac{1}{M} \sum_{m=L+1}^{L+M} h(\mathbf{p}_m), \qquad \widetilde{\mathbf{p}^*} = \arg \max_{m=0,...,L+M} \pi(\mathbf{p}_m|\mathbf{x}).$$

As suggested during the review of this paper, a "polishing" stage can be added to this MCMC algorithm. This consists of adding extra deterministic cycles until a global fixed point is hit. In these cycles, observation $\{i\}$ is deterministically assigned to the most likely cluster (including the new cluster) instead of being randomly assigned as it was in the Gibbs cycle.

3.3 Pareto-based clustering

In the Pareto-based clustering, the model density is $f_{\alpha,\tau}$ given in Eq. (1), and a conjugate prior for (α, τ) is retained. The classical conjugate family for the Pareto model is the Gamma-Pareto(a, b, c, d), such that $\alpha \sim \Gamma(a, b)$, and $\tau|\alpha \sim \mathcal{P}(c\alpha, d)$ with a, b, c, and $d > 0$, that is,

$$\pi_0(\alpha, \tau) \propto \alpha^{a-1} e^{-b\alpha} \alpha d(d\tau)^{-(c\alpha+1)} 1_{(d\tau>1)} \tag{6}$$

Straightforward computations yield the following marginal densities

$$k(x_i, i \in C_j) = \int\int \prod_{i \in C_j} f_{\alpha,\tau}(x_i)\pi_0(\alpha, \tau) d\alpha d\tau = \left(\prod_{i \in C_j} x_i\right)^{-1} \frac{\Gamma(a_j^*)}{\Gamma(a)} \frac{cb^a}{c_j^* \left(b_j^*\right)^{a_j^*}} \tag{7}$$

with

$$a_j^* = a + e_j, \quad c_j^* = c + e_j, \quad d_j^* = \min\left\{d, \min_{i \in C_j} x_i\right\}, \quad b_j^* = b + \sum_{i \in C_j} \ln x_i + c \ln d - c_j^* \ln d_j^*, \tag{8}$$

where e_j is the size of cluster C_j.

Then, the model driven part of the seating probabilities of the Gibbs sampler (cf. Eq. (5)) is such that

$$k(t|x_i, i \in C_j) = (t^{-1}) \times \frac{c_j^* a_j^* \left(b_j^*\right)^{a_j^*}}{(c_j^*+1)\left(b_j^*(t)\right)^{a_j^*+1}}, \tag{9}$$

where $b_j^*(t) = b + \sum_{i \in C_j} \ln x_i + \ln t + c \ln d - \left(c_j^*+1\right) \ln\left(\min\left\{d_j^*, t\right\}\right)$.

For this Pareto-based model, the optimal clustering, $\mathbf{p}^* = \arg\max_{\mathbf{p}} \pi(\mathbf{p}|\mathbf{x})$, allows us to characterize the studied objects in terms of extreme behavior. For example, in the food safety context, an analysis of the cluster composition would help food safety authorities to target their consumption recommendation campaigns at those most at risk. An interesting quantity to compute for the cluster description is the expected value of the tail index within each cluster $\mathbb{E}(\alpha \mid \{x_i, i \in C_j\})$. Since the posterior marginal of $\alpha \mid \{x_i, i \in C_j\}$ is a Gamma distribution with parameters (a_j^*, b_j^*), $\mathbb{E}(\alpha \mid \{x_i, i \in C_j\}) = a_j^*/b_j^*$.

Using the relationship between discrete mixtures of Pareto distributions and tail index estimation, given in Eq. (3) , an estimator of the "global" tail index α^* can be derived given a partition $\mathbf{p}$ based on the fact that

$$\alpha(\mathbf{p}) = \min_{j=1,\ldots,n(\mathbf{p})} \mathbb{E}(\alpha \mid \{x_i, i \in C_j\}) = \min_{j=1,\ldots,n(\mathbf{p})} \frac{a_j^*}{b_j^*}, \tag{10}$$

where a_j^* and b_j^* are the quantities defined in Eq. (8) for the partition $\mathbf{p}$.

From this, using the optimal partition $\mathbf{p}^*$, we get a first estimator of α^* given by

$$\alpha(\mathbf{p}^*) = \min_{j=1,\ldots,n(\mathbf{p}^*)} \frac{a_j^{**}}{b_j^{**}} \tag{11}$$

if a_j^{**} and b_j^{**} are the quantities defined in Eq. (8) for the optimal partition $\mathbf{p}^*$.

Another estimator for α^* is the one obtained by a Monte Carlo simulation in which the function given in Eq. (10) is computed for the M partitions $(\mathbf{p}_m)_{m=1,\ldots,M}$ sampled from $\pi(\mathbf{p}|\mathbf{x})$, and averaged, that is

$$\widetilde{\alpha}_M = \frac{1}{M} \sum_{m=L+1}^{L+M} \alpha(\mathbf{p}_m). \tag{12}$$

Remark 1. The chosen conjugate prior family is the one defined as the modified Lwin Priors in Arnold and Press (1989). A larger one is described in Arnold et al. (1998), which also includes one prior such that $\alpha|\tau \sim \Gamma(a(\tau), b(\tau))$, and the independent Gamma and Pareto priors. It is a 6-parameter family which could also be used in this model-based clustering. However the nonparametric methodology introduced in the next section is even more general.

Remark 2. From a practical point of view, the computation of the driven part of the seating probability in Eq. (9) needs to be carefully checked since overflow problems often occur in the presence of terms such as b^a with large values of a. The solution is therefore to use logarithm and exponential functions to avoid any undefined values (NaN).

Remark 3. One can easily compute the tail probability of the Gamma-Pareto predictive distribution as

$$\mathbb{P}\left(X > x\right) = \frac{cb^a}{(1+c)b_0(x)^a}$$

where $b_0(x) = b + \ln x + c \ln d - (c+1) \ln\left(\min\left\{d, x\right\}\right).$ For large x $(x > d)$, $b_0(x) = b + \ln x - \ln d$ and

$$\lim_{x\to\infty} \frac{\mathbb{P}\left(X > tx\right)}{\mathbb{P}\left(X > x\right)} = \left(1 + \ln t\right)^{-a},$$

which belongs to the Fréchet MDA.

4 Bayesian nonparametric mixture methods

In this section, a general mixture of Pareto distributions is considered. The unknown mixing distribution G is now an infinite dimensional parameter of the model and quantities of the form $\mathbb{E}\left[h(G)|\mathbf{x}\right]$, such as the tail probability given in Eq. (2), are of interest.

4.1 Two key results

Let us first recall two key results of Bayesian nonparametric statistics (see Theorems 1 and 2 in Lo 1984, and the references therein) in a general framework before considering the mixture of Pareto distributions.

The model assumption for a mixture model is $f(x \mid G) = \int k(x \mid u)G(du)$, where G is an unknown distribution (the parameter) and k is a known kernel density in x with parameter $u \in U \subset \mathbb{R}^k$, so that $\int k(x \mid u)dx = 1$.

The natural prior distribution for G is the Dirichlet process (Ferguson 1973) with a nondecreasing shape function γ such that $\gamma(U) < \infty$. It is denoted $G \sim \mathcal{D}(dG \mid \gamma)$.

Theorem 1. If $G \sim \mathcal{D}(dG \mid \gamma)$ and $\mathbf{x} = (x_1, ..., x_n) \mid G$ are i.i.d. $f(x \mid G)$, then for any nonnegative function h

$$\mathbb{E}\left[h(G)|\mathbf{x}\right] = \int \cdots \int \left[\int h(G)\mathcal{D}\left(dG \mid \gamma + \sum_{i=1}^{n} \delta_{u_i}\right)\right] \kappa_n\left(d\overrightarrow{\mathbf{u}}\right) \tag{13}$$

where $\overrightarrow{\mathbf{u}} = (u_1, ..., u_n)$, $\kappa_n\left(d\overrightarrow{\mathbf{u}}\right) = \frac{\prod_{i=1}^{n} k(x_i|u_i)\chi_n\left(d\overrightarrow{\mathbf{u}}\right)}{\int \cdots \int \prod_{i=1}^{n} k(x_i|u_i)\chi_n\left(d\overrightarrow{\mathbf{u}}\right)}$, with

$$\chi_n\left(d\overrightarrow{\mathbf{u}}\right) = \prod_{i=1}^{n} \left(\gamma + \sum_{j=1}^{i-1} \delta_{u_j}\right)(du_i), \text{ and } \int \cdots \int_n \chi\left(d\overrightarrow{\mathbf{u}}\right) = \frac{\Gamma(\gamma(U)+n)}{\Gamma(\gamma(U))}.$$

Remark 4. $\kappa_n\left(d\overrightarrow{\mathbf{u}}\right)$ can be seen as a weighted Blackwell-MacQueen urn distribution since $B_n\left(d\overrightarrow{\mathbf{u}}\right) = \frac{\chi_n\left(d\overrightarrow{\mathbf{u}}\right)}{\int \dots \int_n \chi\left(d\overrightarrow{\mathbf{u}}\right)}$ is called the Blackwell-MacQueen urn distribution (Blackwell and MacQueen 1973).

This first theorem reduces an infinite dimensional integral (on G) to a n-folded one (on $\mathbf{u}$). The second result reduces the n-folded integral to a sum over partitions which allows the use of the same MCMC technique as the one described in the previous section.

Theorem 2. Denoting $\int h(G)\mathcal{D}\left(dG \mid \alpha + \sum_{i=1}^{n} \delta_{u_i}\right) = \mathbb{E}(h(G) \mid \overrightarrow{\mathbf{u}}) = \overline{h}(\overrightarrow{\mathbf{u}})$, and

$$w(\mathbf{p}) = \prod_{j=1}^{n(\mathbf{p})} (e_j - 1)! \int \prod_{i \in C_j} k(x_i \mid u)\gamma(du), \tag{14}$$

then $\mathbb{E}\left(h(G)|\mathbf{x}\right) = \int \dots \int \mathbb{E}(h(G) \mid \overrightarrow{\mathbf{u}})\kappa_n\left(d\overrightarrow{\mathbf{u}}\right) = \sum_{\mathbf{p}} w\left(\mathbf{p}\right) \mathbb{E}\left[\overline{h}(\overrightarrow{\mathbf{u}}) \mid \mathbf{p}\right]$, where the distribution of $\overrightarrow{\mathbf{u}} | \mathbf{p}$ is the product of the distribution of $\left(\overrightarrow{\mathbf{u}} \mid \overrightarrow{\mathbf{u}^*}, \mathbf{p}\right)$ and the distribution of $\left(\overrightarrow{\mathbf{u}^*} \mid \mathbf{p}\right)$ if $\overrightarrow{\mathbf{u}^*}$ denotes the vector of distinct values in vector $\overrightarrow{\mathbf{u}}$, that is

- For $j = 1, \dots, n(\mathbf{p})$, u_j^* are i.i.d. $\pi(du \mid C_j)$, with

$$\pi(du \mid C_j) \propto \prod_{i \in C_j} k(x_i \mid u)\gamma(du) = \frac{\prod_{i \in C_j} k(x_i \mid u)\gamma(du)}{\int \prod_{i \in C_j} k(x_i \mid u)\gamma(du)}, \tag{15}$$

- For $j = 1, \dots, n(\mathbf{p})$, $u_i = u_j^*$ if $i \in C_j$.

This result is used in different manners to conduct Monte Carlo approximations of the quantity $\mathbb{E}\left(h(G)|\mathbf{x}\right)$ depending on the form of $h(G)$. If the density $h(G) = f(t|G)$ or the mixing distribution $h(G) = G(t)$ are to be estimated, further simplifications occur since $\overline{h}(\overrightarrow{\mathbf{u}})$ has an explicit form, as we shall see in the Pareto kernel case in the next section.

4.2 General mixture of Pareto distributions

Let us now turn back to the case of the mixture of Pareto distributions and the model assumption given by

$$f(x \mid G) = \int \int f_{\alpha,\tau}(x)G(d\alpha, d\tau),$$

where $f_{\alpha,\tau}$ is the pareto density given in Eq. (1).

By analogy, $u = (\alpha, \tau) \in \mathbb{R}_+^2$, $k(. \mid u) = f_{\alpha,\tau}(.)$, the prior distribution for G is chosen to be a Dirichlet process with shape $\gamma = \Pi_0$ such that $\gamma(d\alpha, d\tau) = \Pi_0(d\alpha, d\tau) = \pi_0(\alpha, \tau)d\alpha d\tau$, where $\pi_0(\alpha, \tau)$ is the Gamma-Pareto density defined in Eq. (6) so that expressions in Eq. (14) and Eq. (15) are easily computed from the prior-posterior analysis done in Section 3.3. Indeed, the expression in Eq. (14) exactly matches the posterior distribution of partitions in the Pareto-based clustering. The expression in Eq. (15) is the Gamma-Pareto distribution with parameters $(a_j^*, b_j^*, c_j^*, d_j^*)$ since it is the posterior distribution of (α, τ), when the $\{x_i, i \in C_j\}$ given (α, τ) are assumed to be $\mathcal{P}(\alpha, \tau)$, with prior $\pi_0(\alpha, \tau)$.

When the quantity of interest is the tail probability, namely when

$$h(G) = \mathbb{P}(X > x) = \int\int \mathbb{P}\left(X > x|\alpha, \tau\right) G(d\alpha, d\tau),$$

simple Dirichlet calculation and integration yield

$$\begin{aligned}
\overline{h}(\overrightarrow{\alpha}, \overrightarrow{\tau}) &= \mathbb{E}(h(G) \mid \overrightarrow{\alpha}, \overrightarrow{\tau}) \\
&= \int \left[\int\int \mathbb{P}\left(X > x|\alpha, \tau\right) G(d\alpha, d\tau)\right] \mathcal{D}\left(dG \mid \Pi_0 + \sum_{i=1}^{n} \delta_{\alpha_i, \tau_i}\right) \\
&= \frac{1}{(1+n)} \frac{cb^a}{(1+c)\left(b_0^*(x)\right)^a} + \frac{1_{x<d}}{(1+n)} \left[1 - \frac{b^a}{\left(b + c\ln\left(d/x\right)\right)^a}\right] \\
&+ \frac{1}{(1+n)} \left(\sum_{i=1}^{n} \left(\tau_i x\right)^{-\alpha_i} 1_{(\tau_i x > 1)} + \sum_{i=1}^{n} 1_{(\tau_i x \leq 1)}\right),
\end{aligned} \tag{16}$$

where $\overrightarrow{\alpha} = (\alpha_1, \ldots, \alpha_n)'$, $\overrightarrow{\tau} = (\tau_1, \ldots, \tau_n)'$, and $b_0^*(x) = b + \ln(x) + c\ln(d) - (1 + c)\ln(\min\{d, x\})$.

This can even be further simplified in case of ties among the $(\alpha_i, \tau_i)_i$, that is, given the fact that the distribution of $\overrightarrow{\alpha}, \overrightarrow{\tau} \mid \mathbf{p}$ is the product of the distribution of $\left(\overrightarrow{\alpha}, \overrightarrow{\tau} \mid \overrightarrow{\alpha^*}, \overrightarrow{\tau^*}, \mathbf{p}\right)$ and the distribution of $\left(\overrightarrow{\alpha^*}, \overrightarrow{\tau^*} \mid \mathbf{p}\right)$. Taking the expectancy of Eq. (16) with respect to this product distribution yields

$$\begin{aligned}
\mathbb{E}\left[\overline{h}(\overrightarrow{\alpha}, \overrightarrow{\tau}) \mid \mathbf{p}\right] = &\frac{1}{(1+n)} \frac{cb^a}{(1+c)\left(b_0^*(x)\right)^a} + \frac{1_{x<d}}{(1+n)} \left[1 - \frac{b^a}{\left(b + c\ln\left(d/x\right)\right)^a}\right] \\
&+ \frac{1}{(1+n)} \sum_{j=1}^{n(\mathbf{p})} \frac{e_j \quad c_j^* (b_j^*)^{a_j^*}}{(1 + c_j^*)(b_j^*(x))^{a_j^*}} \\
&+ \frac{1}{(1+n)} \sum_{j=1}^{n(\mathbf{p})} e_j 1_{x<d_j^*} \left[1 - \frac{(b_j^*)^{a_j^*}}{(b_j^* + c_j^* \ln(d_j^*/x))^{a_j^*}}\right],
\end{aligned}$$

where $b_j^*(x) = b_j^* + \ln(x) + c_j^* \ln(d_j^*) - (1 + c_j^*)\ln(\min\left\{d_j^*, x\right\})$ and $(a_j^*, b_j^*, c_j^*, d_j^*)$ are given in Eq. (8).

Algorithm 2. **Estimation of the probability tail** $\mathbb{P}(X > x)$

1. Sample M partitions from the distribution $w(\mathbf{p})$ (cf. using the Gibbs sampler provided in Section 3.2).

2. For each partition $\mathbf{p}_m$, given $x > 0$, compute the quantity

$$\begin{aligned} h_m(x) = {} & \frac{1}{(1+n)} \frac{cb^a}{(1+c)\left(b_0^*(x)\right)^a} + \sum_{j=1}^{n(\mathbf{p}_m)} \frac{e_j}{(1+n)} \frac{c_j^*(b_j^*)^{a_j^*}}{(1+c_j^*)(b_j^*(x))^{a_j^*}} \\ & + \frac{1_{x<d}}{(1+n)} \left[1 - \frac{b^a}{\left(b + c\ln\left(d/x\right)\right)^a} \right] \\ & + \frac{1}{(1+n)} \sum_{j=1}^{n(\mathbf{p})} e_j 1_{x<d_j^*} \left[1 - \frac{\left(b_j^*\right)^{a_j^*}}{(b_j^* + c_j^* \ln(d_j^*/x))^{a_j^*}} \right], \end{aligned} \tag{17}$$

   where e_j is the size of cluster C_j of $\mathbf{p}_m$, and all $^*_j$ quantities are computed with respect to cluster C_j of $\mathbf{p}_m$ as in Eq. (8).

3. Compute the tail probability estimator as the mean of the $(h_m(x))_{m=1,\ldots M}$.

5 Application

In this section, the Pareto based clustering is first applied to simulated data and then to a real data set related to dietary exposure to ochratoxin A (OTA).

In both applications, the Gibbs WCR was run from a Gauss routine (cf. the Gauss software webpage, http://www.aptech.com, for information) such that

- a burn-in of $L = 10000$ iterations is used,
- $M = 20000$ Monte Carlo iterations are computed
- a diffuse prior choice for the Gamma-Pareto hyperparameters: $a = b = c = 0$ and $d = \infty$, which is improper. In practice, the following setting is used: $a = b = c = 0.001$ and $d = \max_i x_i \times 1.1$.
- the parameter of the Chinese Restaurant Process is fixed to $e_0 = 1$.

5.1 Simulated data

5.1.1 Description

Four sets of data are generated based on discrete mixtures of four Pareto distributions: $\sum_{j=1}^{4} w_j \mathcal{P}(\alpha_j, \tau_j)$ with the settings given in Table 1.

The size of each simulated data set is fixed at $n = 200$. For example, 100 values are randomly selected from a $\mathcal{P}(3,1)$ and 100 from a $\mathcal{P}(6,1)$ to constitute data set 2. For all of these simulated data sets, the true tail index is 3: the main goal of this simulation study is to determine whether the proposed methodology provides a good estimation of this tail index or not. Figure 1 gives examples of histograms obtained with the different settings.

	w_1	w_2	w_3	w_4
Pareto Parameters	(3,1)	(6,1)	(3,3)	(6,3)
Data set 1	1	0	0	0
Data set 2	1/2	1/2	0	0
Data set 3	1/2	0	1/2	0
Data set 4	1/4	1/4	1/4	1/4

Table 1: Description of the simulated datasets.

5.1.2 Results

Table 2 gives a description of the resulting optimal partition as well as a few outputs of the two proposed approaches. A bias corrected Hill estimator is also computed for comparison's sake. The methodology here is similar to the one used in Tressou et al. (2004), adapted from Beirlant et al. (1999) and Feuerverger and Hall (1999). Comparison to other estimators of the tail index, namely the one proposed by Beirlant et al. (2005), is conducted in a forthcoming study.

Data set	1	2	3	4
$\max_i x_i$	4.8	3.9	3.4	2.8
$n(\mathbf{p}^*)$	1	1	2	2
$\pi(\mathbf{p}^*\|\mathbf{x})$	800.1	880.8	803.3	836.7
$\alpha(\mathbf{p}^*)$	3.120	4.280	3.517	4.091
$\widetilde{\alpha}_M$	3.130	4.280	3.507	4.082
$\Pr(X > \max_i x_i)$	0.81%	0.32%	0.77%	0.85%
Bias Corrected Hill	2.896	4.300	3.564	4.911

Table 2: Results on simulated data.

The main findings of these simulations are the following:

1. Mixtures over the location parameter τ are easily detected (cf. data set 3) whereas mixtures over the tail index parameter α are a lot more difficult to detect (cf. data sets 2 and 4) even if one considers data sets involving two tail indexes with a huge difference.

2. The tail index estimator referred to as $\alpha(\boldsymbol{p}^*)$ in Table 2 is defined in Eq. (11), and the one referred to as $\widetilde{\alpha}_M$ is defined in Eq. (12). When both parameters are mixed over, the two proposed Tail Index Estimators are less biased than, or equivalent to, the Bias Corrected Hill estimator. However, our estimators tend to overestimate α which is not desirable in risk analysis since one certainly does not want to underestimate the risk.

3. The methodology also allows to compute any tail probability as exemplified by the probability of exceeding the observed maximum, $\Pr(X > \max_i x_i)$, given in Table 2. It is computed as the mean of the $(h_m(\max x_i))_{m=1,\dots,M}$ as defined in Eq. (17).

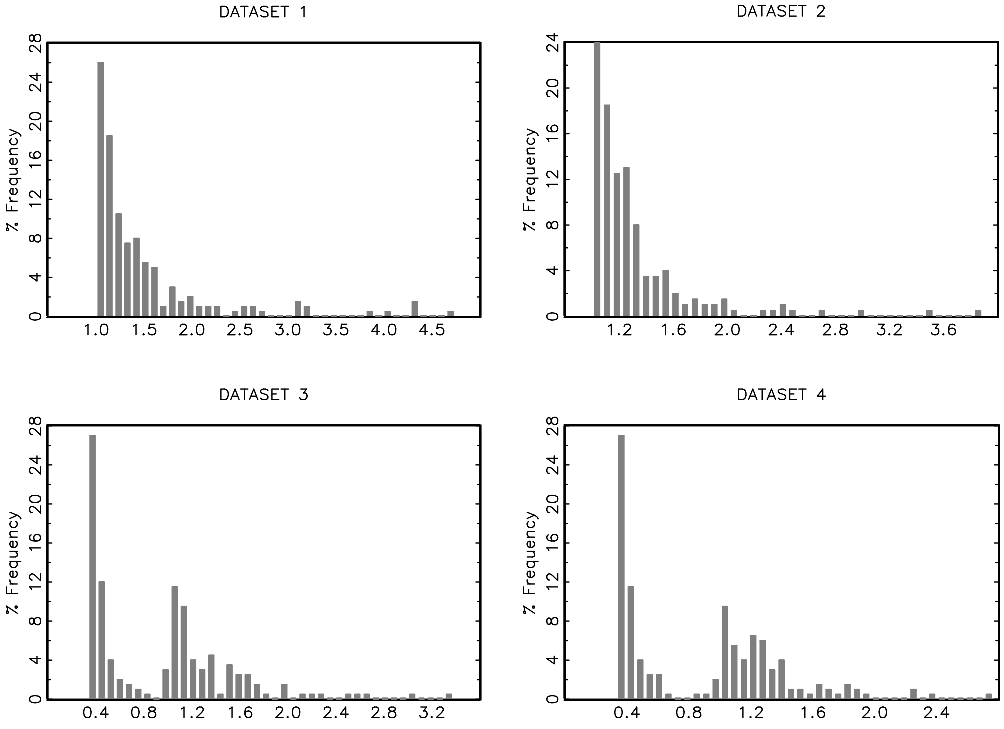

Figure 1: Example histograms of the 4 simulated datasets ($n = 200$).

4. When computing the tail index estimator and posterior log likelihood associated to the simulated partition (the original one generically denoted $\mathbf{p}_0$ in the sequel, that is, the one with 4 clusters in the case of data set 4 as an example), we obtain the following results:

   - For data set 2, $\alpha(\mathbf{p}_0) = 3.400$, $\pi(\mathbf{p}_0|\mathbf{x}) =$730.4.
   - For data set 3, $\alpha(\mathbf{p}_0) = 3.433$, $\pi(\mathbf{p}_0|\mathbf{x}) =$790.3.
   - For data set 4, $\alpha(\mathbf{p}_0) = 3.223$, $\pi(\mathbf{p}_0|\mathbf{x}) =$659.0.

   This illustrates the well known identifiability problem of mixture models (see for example Marin et al. 2005) and the fact that maximizing the posterior likelihood is not always the right approach. Indeed, the optimal partition described in Table 2 enjoys a higher posterior likelihood than the one generating the data for the three data sets 2, 3 and 4. Furthermore, the tail index estimator associated with this "generating" partition is still biased but (not shown) simulations empirically show that it goes to zero for large values of n. For example for $n = 3000$ (OTA data set size) in the setting of data set 2, we get $\alpha(\mathbf{p}_0) = 3.147$ on one particular simulation and 3.006 if averaging on 100 independent simulation results.

5. The "polishing" stage described at the end of Section 3.2 was applied to these simulated data but does not change any of the results. Indeed, the optimal partition $\mathbf{p}^*$ already has a larger likelihood than the one that generated the data.

5.2 OTA data set

5.2.1 Food risk assessment context, description of the data

Ochratoxin A (OTA) is a mycotoxin produced by fungi *Aspergillus Ochraceus* and *Penicillium Viridicatum*. This mycotoxin can be detected in several food items: cereals, coffee, grapes, pork meat, wine, beer, and so on. Ochratoxin A is nephrotoxic, genotoxic, teratogenic, carcinogenic and immunosuppressive. The compound has been linked to Balkan Endemic Nephropathy, a kidney disease frequently observed in the Balkan countries (Boižić et al. 1995, for a review). Such a disease can appear after long and excessive exposure to the contaminant. This exposure is not directly observed but is assessed from food consumption surveys that record the quantity of different foods consumed and contamination data mostly derived from national surveillance plans in which foods are analyzed and contaminant levels are measured. This exposure assessment step can be conducted in different ways which are not the concern here but are described in Kroes et al. (2002) and the reference therein.

The motivating real data set is composed of the possible extreme OTA exposure of $n = 3003$ French individuals. More precisely, for each of the 3003 individuals, food consumption is observed in the INCA data (CREDOC-AFSSA-DGAL 1999) and individual distribution of exposure is built by a Monte Carlo simulation using the individual consumption and the empirical distribution of several independently available OTA contamination data (cf. Bertail and Tressou 2006; Tressou 2006; Counil et al. 2005, 2006, for a full description of the data and examples of OTA exposure assessments.). Then the 95^{th} percentile of this simulated distribution is retained as evidence of possible extreme exposure to OTA. This is expressed on a body weight basis (quantity of contaminant divided by body weight). A histogram of the observations is given in Figure 2.

5.2.2 Results

Table 3 introduces and describes the resulting optimal partitions comprising 11 clusters respectively before and after the polishing stage (that hit a fixed point after only 3 iterations, and detailed at the end of Section 3.2). The two resulting optimal partitions do not differ much, but the polished log-likelihood ($LL = 4629$) is 2.5 times higher than the non polished one ($LL = 1850$). We observe that the cluster sizes are heterogeneous (Cluster 11 only comprises 2 or 3 individuals, respectively in the polished and non polished cases). Analysis of the clusters is not obvious: a few socioeconomic variates were considered here and a comparison of the lower (AP) and upper part (BP) of Table 3 shows the consistency of our findings before and after polishing. The proportion of female adults and under-reporting individuals (who declare insufficient consumption in relation to their nutritional needs) decreases with an average of the 95^{th} percentile of

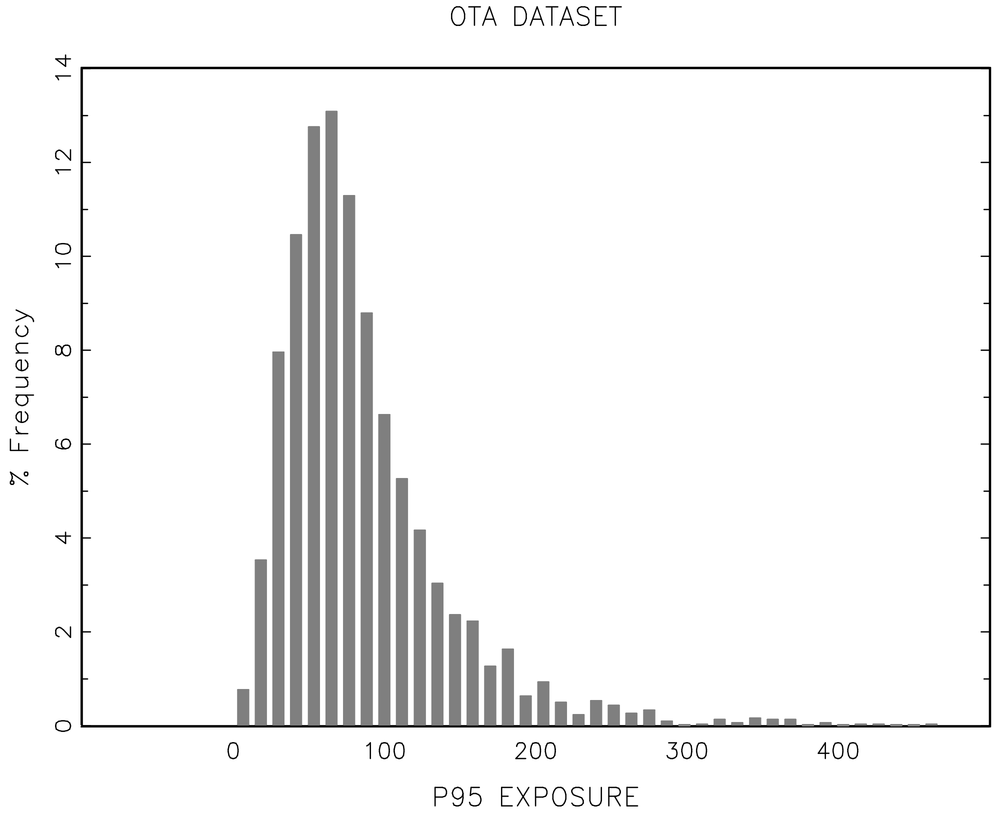

Figure 2: Histogram of the 95^{th} percentile of individual exposure (expressed in ng/kg bw/w).

exposure whereas the proportion of children increases with this average. The body mass index (BMI: body weight divided by squared height) also decreases with the average 95^{th} percentile of exposure, which is in accordance with the usual fish consumer typology. Cluster 9 is mostly comprised of children and this cluster enjoys the highest average 95^{th} percentile of exposure and the lowest BMI. These two features are consistent with the child population in most food risk assessments.

The Cluster Tail Index (CTI), computed as the ratio a_j^*/b_j^* for each cluster j (see Eq. (11)) allows classification of the clusters according to risk levels, the larger the CTI, the less serious the risk. The entire population tail index is 1.440 in the polished case, and 0.622 in the non polished case, when the estimator based on the optimal partition $\alpha(\boldsymbol{p}^*)$ is used, see Eq. (11) . Indeed, this is the minimum tail index among all cluster tail indices reached for Cluster 11 in both cases. This is not satisfactory because of the very small size of this cluster. Indeed, we can question here whether the estimation of the α_{11} is consistent with only 2 or 3 observations in this cluster. If $\widetilde{\alpha}_M$, defined in Eq. (12) , is used instead, the tail index estimator is equal to 0.863 (see Figure 3) and does not depend on the polishing stage. This last estimation is certainly much closer to the actual general tail index for the extreme exposures to OTA. Note that the bias corrected Hill Estimator would be 11.52, which totally misses the heaviest part of the tail.

As in the simulation, the tail probability was computed using Eq. (17) and is plotted in Figure 4. The proposed methodology provides a nonparametric estimator of the tail probability on the half line so that any tail probability (even an extremely small one) can be estimated.

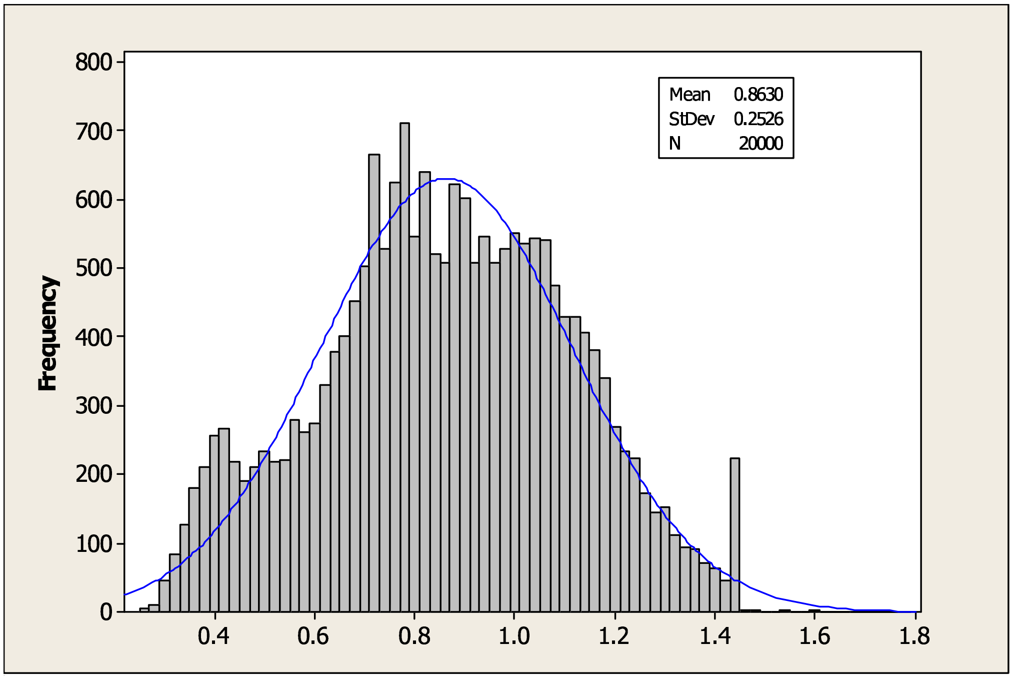

Figure 3: Empirical distribution of the Monte Carlo simulation for the tail index α, resulting in the $\widetilde{\alpha}_M$ estimator.

			Cluster				Observations (P95 of exposure)				Covariates		
	j	Size	MLL	CTI	Avg	StD	Min	Max	Avg.Age	Avg.BMI	P.Ch	P.AdF	P.UR
BP	1	549	-2652.61	2.582	122.4	53.9	77.7	467.5	24.0	19.5	51.7%	6.6%	2.2%
	2	259	-1121.28	1.542	50.5	50.4	20.7	369.4	39.2	24.1	17.0%	22.0%	44.0%
	3	104	-427.31	1.243	32.6	45.0	10.3	366.5	42.6	25.3	14.4%	24.0%	64.4%
	4	237	-863.487	4.538	59.9	16.1	46.8	161.7	36.5	23.0	18.6%	19.8%	17.7%
	5	205	-622.761	9.201	64.1	7.1	57.2	101.4	34.9	22.4	26.3%	11.7%	11.2%
	6	515	-2184.36	2.458	66.3	39.3	40.1	346.6	36.2	22.9	20.8%	16.7%	21.0%
	7	569	-2505.73	3.127	95.0	36.2	65.7	429.0	28.9	21.0	37.3%	7.6%	4.4%
	8	272	-1017.33	2.826	43.6	23.2	28.4	249.3	37.5	24.2	16.5%	19.5%	39.0%
	9	278	-1390.29	2.987	157.8	51.3	108.2	364.4	15.1	17.6	76.6%	2.2%	0.4%
	10	12	-42.5631	2.338	7.9	3.3	4.9	17.1	54.7	25.2	0.0%	25.0%	91.7%
	11	3	-22.5831	0.622	8.1	11.2	0.7	21.0	41.7	24.8	0.0%	0.0%	66.7%
AP	1	633	-2303.4501	6.689	90.8	8.6	77.8	108.1	31.0	21.0	33.5%	8.8%	2.7%
	2	118	-307.33299	5.856	24.7	2.2	20.7	28.3	46.2	25.9	5.1%	21.2%	63.6%
	3	67	-208.77301	2.474	15.7	3.0	10.3	20.6	43.4	25.8	10.4%	25.4%	76.1%
	4	337	-906.07401	10.173	51.7	3.0	46.8	57.2	38.0	23.7	14.2%	18.4%	22.6%
	5	293	-714.13941	15.589	61.0	2.4	57.2	65.6	37.1	22.8	21.2%	17.4%	16.0%
	6	206	-483.343	12.503	43.5	2.0	40.1	46.8	39.5	23.9	12.6%	19.4%	35.4%
	7	375	-1052.3487	12.391	71.3	3.5	65.7	77.7	33.1	22.3	24.5%	10.9%	8.8%
	8	251	-713.35581	5.804	33.8	3.4	28.3	39.9	40.6	24.9	11.6%	23.9%	49.8%
	9	710	-3559.2548	2.863	160.7	55.7	108.2	467.5	15.2	17.7	75.5%	3.5%	0.3%
	10	11	-37.305571	2.838	7.1	1.6	4.9	9.7	54.6	25.1	0.0%	27.3%	90.9%
	11	2	-15.761918	1.440	1.7	1.4	0.7	2.7	28.0	23.8	0.0%	0.0%	100.0%

Table 3: Description of the resulting partition for the OTA dataset (BP: before polishing; AP: after polishing).
Note: MLL=Marginal log-likelihood of the cluster; CTI=Cluster Tail Index; Avg.= Average; StD.=Standard Deviation; Min=Minimum; Max=Maximum; BMI=Body Mass Index (body weight divided by squared height); P.Ch= proportion of Children; P.AF=proportion of Female Adults; P.UR=proportion of under-reporting individuals.

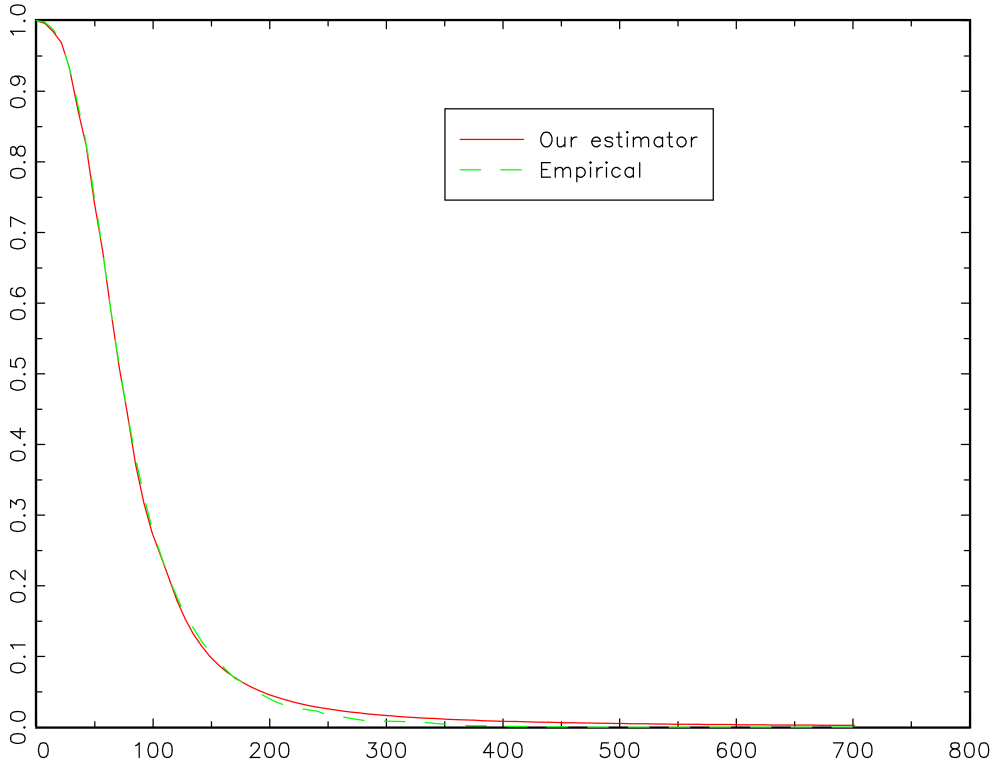

Figure 4: Tail estimation in the OTA dataset.

6 Discussion

The implementation of the two proposed methodologies together with classical extreme value approaches illustrates the difficulty of estimating the tail index if the data is generated from a mixture. Yet, in many applications, this assumption holds. The two proposed tail index estimators are actually at least as good as the Hill estimator even though the Monte Carlo approximation is preferable above all when cluster sizes are small. The proposed estimator for the tail probability is a good alternative to the basic empirical estimator: still nonparametric, it does not require any specific parametric assumption except the heavy tailed one, and has the advantage of being defined on the whole half line. The resulting clusters are not easy to describe and, surprisingly for univariate data, they do not correspond to a partition of the real line into disjoint intervals.

Several extensions or changes in the framework may be considered. First, in the parametric approach, other distributions may be considered for the Pareto parameters

α and τ as mentionned in Remark 1. In the nonparametric extension, a basic Dirichlet process was considered as the prior for the mixing distribution using Ferguson's original definition as in Lo (1984): $G \sim \mathcal{D}(dG \mid \gamma)$. One could also use the (θ, H) parametrization such that $G \sim \mathcal{D}(dG \mid \theta, H)$, where θ is the total mass of the base-line measure H, that is θ corresponds to $\gamma(U)$ in our setting. Going further in this direction the recent paper of Lijoi et al. (2007) provides interesting extensions. Furthermore, other processes, such as the Poisson-Dirichlet process, may be considered, see Lau and Green (2007) and the references therein.

From the applied perspective, it would be interesting to work on the individual exposure curves instead of only considering the 95^{th} percentile of exposure for each individual. This could be conducted using a Hierarchical Dirichlet process, also called "Chinese Restaurant Franchise", see Teh et al. (2006). This way, there would be a double clustering of exposure values and individual exposure distributions. This would require some computational adaptation since the data set would be huge (from the OTA data set, we can actually get $n = 3003$ exposure distribution curves, described by $n \times M$ points if M exposure levels are simulated for each individual). The use of the Gibbs sampling methods for stick-breaking priors proposed in Ishwaran and James (2001) will be investigated in future work.

7 Appendix: Technical details

The notations from Eq. (8) are again used in this appendix and recalled here: $a_j^* = a + e_j$, $b_j^* = b + \sum_{i \in C_j} \ln x_i + c \ln d - c_j^* \ln d_j^*$, $c_j^* = c + e_j$, and $d_j^* = \min\left\{d, \min_{i \in C_j} x_i\right\}$.

7.1 Derivation of Eq. (7)

Eq. (7) is obtained by:

$$
\begin{aligned}
k(x_i, i \in C_j) &= \int\int \prod_{i \in C_j} f_{\alpha,\tau}(x_i) \pi_0(\alpha, \tau) d\alpha d\tau \\
&= \int\int \prod_{i \in C_j} \alpha\tau \left(\tau x_i\right)^{-(\alpha+1)} 1_{(\tau x_i > 1)} \times \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha} c\alpha d(d\tau)^{-(c\alpha+1)} 1_{(d\tau>1)} d\alpha d\tau \\
&= \frac{cb^a}{\Gamma(a)} \left(\prod_{i \in C_j} x_i\right)^{-1} \int \alpha^{a+e_j} \exp^{-\alpha\left(b + \sum_{i \in C_j} \ln x_i + c \ln d\right)} d\alpha \\
&\quad \times \int_{\tau = 1/\min\left\{d, \min_{i \in C_j} x_i\right\}}^{\infty} \tau^{-\alpha(e_j + c) - 1} d\tau \\
&= \frac{cb^a}{\Gamma(a) c_j^*} \left(\prod_{i \in C_j} x_i\right)^{-1} \int \alpha^{a_j^* - 1} \exp\left[-\alpha b_j^*\right] d\alpha = \left(\prod_{i \in C_j} x_i\right)^{-1} \frac{\Gamma(a_j^*)}{\Gamma(a)} \frac{cb^a}{c_j^* \left(b_j^*\right)^{a_j^*}}.
\end{aligned}
$$

7.2 Derivation of Eq. (9)

Eq. (9) is the model driven part of the seating probability, used to reassign a measurement t in one of the clusters C_j and denoted $k(t \mid \{x_i, i \in C_j\})$. It can be obtained in two ways:

1. First, the ratio of the marginal densities of the clusters $\{x_i, i \in C_j\} \cup \{t\}$ and $\{x_i, i \in C_j\}$, namely
$$k(t \mid \{x_i, i \in C_j\}) = \frac{k(\{x_i, i \in C_j\} \cup \{t\})}{k(\{x_i, i \in C_j\})}$$
$$= \frac{t^{-1}\left(\prod_{i\in C_j} x_i\right)^{-1} \frac{\Gamma(a_j^*+1)}{\Gamma(a)} \frac{cb^a}{(c_j^*+1)\left(b_j^*(t)\right)^{a_j^*+1}}}{\left(\prod_{i\in C_j} x_i\right)^{-1} \frac{\Gamma(a_j^*)}{\Gamma(a)} \frac{cb^a}{c_j^*\left(b_j^*\right)^{a_j^*}}} = (t^{-1}) \times \frac{c_j^* a_j^* \left(b_j^*\right)^{a_j^*}}{(c_j^*+1)\left(b_j^*(t)\right)^{a_j^*+1}},$$
where $b_j^*(t) = b + \sum_{i\in C_j} \ln x_i + \ln t + c \ln d - c_j^* \ln\left(\min\{d_j^*, t\}\right)$.

2. The predictive density of a new data t given observations $\{x_i, i \in C_j\}$ can also be directly computed by first computing the predictive density for no observation, namely $k(t) = \int\int f_{\alpha,\tau}(t)\pi_0(\alpha,\tau)d\alpha d\tau$, and then replacing all hyperparameters by their updated version $(^*_j)$ given in Eq. (8) since
$$k(t \mid \{x_i, i \in C_j\}) = \int\int f_{\alpha,\tau}(t)\pi(\alpha,\tau \mid \{x_i, i \in C_j\})d\alpha d\tau,$$
where $\pi(\alpha,\tau \mid \{x_i, i \in C_j\})$ is the posterior density in a Pareto model with Gamma Pareto prior, i.e. a Gamma Pareto $(a_j^*, b_j^*, c_j^*, d_j^*)$.
$$k(t) = \int\int f_{\alpha,\tau}(t)\pi_0(\alpha,\tau)d\alpha d\tau$$
$$= \frac{cab^a t^{-1}}{(1+c)(b + \ln t + c\ln d - (1+c)\ln(\min\{d,t\}))^{a+1}}$$
$$\Longrightarrow k(t|x_i, i \in C_j) = (t^{-1}) \times \frac{c_j^* a_j^* \left(b_j^*\right)^{a_j^*}}{(c_j^*+1)\left(b_j^*(t)\right)^{a_j^*+1}},$$
with $b_j^*(t) = b_j^* + \ln t + c_j^* \ln d_j^* - \left(c_j^*+1\right)\ln\left(\min\{d_j^*, t\}\right)$ which is the same as the one obtained using the ratio method in [Way 1].

In the Pareto case, both calculations are straightforward and may be used to check on the exactitude of the result, while for other kernel densities, the second approach may be simpler since calculations are exactly the same as the ones for the marginal densities.

From a computational point of view, remark that e_j, $\min_{i\in C_j} x_i$ and $\sum_{i\in C_j} \ln x_i$ are the only quantities needed to compute the marginal of cluster C_j and the seating probability to cluster C_j so that there is no need to store and manipulate all the $\{x_i, i \in C_j\}$ for $j = 1, ..., n(p)$ in the Gibbs cycle.

7.3 Derivation of Eq. (16)

Eq. (16) is obtained by first applying the Fubini result for Dirichlet processes (see Lemma 1 of Lo (1984)). Then, given $\overrightarrow{\alpha} = (\alpha_1, \ldots, \alpha_n)'$, $\overrightarrow{\tau} = (\tau_1, \ldots, \tau_n)'$ and considering $h(G) = P(X > x) = \int\int P(X > x|\alpha, \tau)\, G(d\alpha, d\tau)$, we have

$$\begin{aligned}
\mathbb{E}(h(G) \mid \overrightarrow{\alpha}, \overrightarrow{\tau}) &= \int \left[\int\int \mathbb{P}(X > x|\alpha, \tau)\, G(d\alpha, d\tau)\right] \mathcal{D}\left(dG \mid \Pi_0 + \sum_{i=1}^{n} \delta_{\alpha_i,\tau_i}\right) \\
&= \frac{1}{\left(\Pi_0 + \sum_{i=1}^{n} \delta_{\alpha_i,\tau_i}\right)(\mathbb{R}^{2+})} \left[\begin{array}{c} \int\int \mathbb{P}(X > x|\alpha, \tau)\, \Pi_0(d\alpha, d\tau) \\ +\sum_{i=1}^{n} \int\int \mathbb{P}(X > x|\alpha, \tau)\, \delta_{\alpha_i,\tau_i}(d\alpha, d\tau) \end{array}\right] \\
&= \frac{1}{(1+n)} \left[\begin{array}{c} \int\int \left[(\tau x)^{-\alpha}\, 1_{(\tau x > 1)} + 1_{(\tau x \le 1)}\right] \pi_0(\alpha, \tau) d\alpha d\tau \\ +\sum_{i=1}^{n} (\tau_i x)^{-\alpha_i}\, 1_{(\tau_i x > 1)} + 1_{(\tau_i x \le 1)} \end{array}\right] \\
&= \frac{1}{(1+n)} \frac{cb^a}{(1+c)\left(b_0^*(x)\right)^a} + \frac{1}{(1+n)} \int\int 1_{(\tau x \le 1)} \pi_0(\alpha, \tau) d\alpha d\tau \\
&+ \frac{1}{(1+n)} \left(\sum_{i=1}^{n} (\tau_i x)^{-\alpha_i}\, 1_{(\tau_i x > 1)} + \sum_{i=1}^{n} 1_{(\tau_i x \le 1)}\right),
\end{aligned}$$

where $b_0^*(x) = b + \ln(x) + c\ln(d) - (1+c)\ln(\min\{d, x\})$ and

$$\begin{aligned}
\int\int 1_{(\tau x \le 1)} \pi_0(\alpha, \tau) d\alpha d\tau &= \int\int 1_{(\tau x \le 1)} \left[\frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha}\right] \left[c\alpha d (d\tau)^{-(c\alpha+1)} 1_{(d\tau > 1)}\right] d\alpha d\tau \\
&= 1_{x<d} \frac{cb^a}{\Gamma(a)} \int \alpha^a e^{-(b + c\ln d)\alpha} \left[\int_{\tau = 1/d}^{1/x} \tau^{-(c\alpha+1)} d\tau\right] d\alpha \\
&= 1_{x<d} \frac{cb^a}{\Gamma(a)} \int \alpha^a e^{-(b + c\ln d)\alpha} \left[\frac{\tau^{-c\alpha}}{-c\alpha}\right]_{\tau = 1/d}^{1/x} d\alpha \\
&= 1_{x<d} \frac{b^a}{\Gamma(a)} \int \alpha^{a-1} e^{-(b + c\ln d)\alpha} \left[d^{c\alpha} - x^{c\alpha}\right] d\alpha \\
&= 1_{x<d} \left[\int \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha} d\alpha - \frac{b^a}{\Gamma(a)} \int \alpha^{a-1} e^{-(b + c\ln d - \ln x)\alpha} d\alpha\right] \\
&= 1_{x<d} \left[1 - \frac{b^a}{(b + c\ln d - \ln x)^a}\right],
\end{aligned}$$

so that finally,

$$\begin{aligned}
\mathbb{E}(h(G) \mid \overrightarrow{\alpha}, \overrightarrow{\tau}) &= \frac{1}{(1+n)} \frac{cb^a}{(1+c)\left(b_0^*(x)\right)^a} + \frac{1_{x<d}}{(1+n)} \left[1 - \frac{b^a}{(b + c\ln d - \ln x)^a}\right] \\
&+ \frac{1}{(1+n)} \left(\sum_{i=1}^{n} (\tau_i x)^{-\alpha_i}\, 1_{(\tau_i x > 1)} + \sum_{i=1}^{n} 1_{(\tau_i x \le 1)}\right).
\end{aligned}$$

References

Arnold, B., Castillo, E., and Sarabia, J. (1998). "Bayesian analysis for classical distributions using conditionally specified priors." *Sankhya: The Indian Journal of Statistics*, 60: 228–245. 373

Arnold, B. and Press, S. (1989). "Bayesian estimation and prediction for Pareto data." *Journal of the American Statistical Association*, 84(408): 1079–1084. 373

Beirlant, J., Dierckx, G., Goegebeur, Y., and Matthys, G. (1999). "Tail index estimation and an exponential regression model." *Extremes*, 2(2): 177–200. 369, 378

Beirlant, J., Dierckx, G., and Guillou, A. (2005). "Estimation of the extreme-value index and generalized quantile plots." *Bernoulli*, 11(6): 949–970. 378

Bertail, P., Clémençon, S., and Tressou, J. (2008). "A storage model with random release rate for modeling exposure to food contaminants." *Math. Biosc. Eng.*, 35(1): 35–60. 368

Bertail, P. and Tressou, J. (2006). "Incomplete generalized U-Statistics for food risk assessment." *Biometrics*, 62(1): 66–74. 368, 380

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *Annals of Statistics*, 1: 353–355. 375

Boižić, Z., Duančić, V., Belicza, M., Krausand, O., and Skljarov, I. (1995). "Balkan endemic nephropathy: still a mysterious disease." *European Journal of Epidemiology*, 11: 235–238. 380

Bottolo, L., Consonni, G., Dellaportas, P., and Lijoi, A. (2003). "Bayesian Analysis of Extreme Values by Mixture Modeling." *Extremes*, 6: 25–47. 368

Coles, S. and Powell, E. (1996). "Bayesian Methods in Extreme Value Modelling: A Review and New Developments." *International Statistical Review*, 64: 119–136. 368

Counil, E., Verger, P., and Volatier, J.-L. (2005). "Handling of contamination variability in exposure assessment: A case study with Ochratoxin A." *Food and Chemical Toxicology*, 43(10): 1541–1555. 380

— (2006). "Fitness-for-purpose of dietary survey duration: A case-study with the assessment of exposure to Ochratoxin A." *Food and Chemical Toxicology*, 44(4): 499–509. 380

CREDOC-AFSSA-DGAL (1999). *Enquête INCA (individuelle et nationale sur les consommations alimentaires)*. Lavoisier, Paris, TEC&DOC edition. (Coordinateur : J.L. Volatier). 380

Diebolt, J., El-Aroui, M.-A., Garrido, M., and Girard, S. (2005). "Quasi-Conjugate Bayes estimates for GPD parameters and Applications to Heavy tails modelling." *Extremes*, 8: 57–78. 368

Edler, L., Poirier, K., Dourson, M., Kleiner, J., Mileson, B., Nordmann, H., Renwick, A., Slob, W., Walton, K., and Würtzen, G. (2002). "Mathematical modelling and quantitative methods." *Food and Chemical Toxicology*, 40: 283–326. 368

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1999). *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics. Berlin: Springer-Verlag. 369

Escobar, M. D. (1994). "Estimating normal means with a Dirichlet process prior." *Journal of the American Statistical Association*, 89: 268–277. 371

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230. 374

Feuerverger, A. and Hall, P. (1999). "Estimating a tail exponent by modelling departure from a Pareto Distribution." *Annals of Statistics*, 27: 760–781. 369, 378

Fraley, C. and Raftery, A. (2002). "Model-based Clustering, Discriminant analysis, and density estimation." *Journal of the American Statistical Association*, 97(458): 611–631. 368

Frigessi, A., Haug, O., and Rue, H. (2002). "A dynamic mixture model for unsupervised tail estimation without threshold selection." *Extremes*, 5: 219–235. 368

Gauchi, J. P. and Leblanc, J. C. (2002). "Quantitative Assessment of Exposure to the Mycotoxin Ochratoxin A in food." *Risk Analysis*, 22: 219–234. 368

Gibney, M. J. and van der Voet, H. (2003). "Introduction to the Monte Carlo project and the approach to the validation of probabilistic models of dietary exposure to selected food chemicals." *Food Additives and Contaminants*, 20(Suppl. 1): S1–S7. 368

Green, P. and Richardson, S. (2001). "Modelling Heterogeneity With and Without the Dirichlet Process." *Scandinavian Journal of Statistics*, 28(2): 355–375. 368

Heard, N., Holmes, C., and Stephens, D. (2006). "Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves." *Journal of the American Statistical Association*, 101: 18–29. 371

Hill, B. (1975). "A simple general approach to inference about the tail of a distribution." *Annals of Statistics*, 3: 1163–1174. 369

Ishwaran, H. and James, L. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96: 161–173. 385

Kottas, A. and Sansó, B. (2007). "Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis." *Journal of Statistical Planning and Inference*, 37: 3151–3163. 368

Kroes, R., Müller, D., Lambe, J., Lowik, M. R. H., van Klaveren, J., Kleiner, J., Massey, R., Mayer, S., Urieta, I., Verger, P., and Visconti, A. (2002). "Assessment of intake from the diet." *Food Chemical and Toxicology*, 40: 327–385. 380

Lau, J. W. and Green, P. (2007). "Bayesian Model Based Clustering Procedures." *Journal of Computational and Graphical Statistics*, 16(3): 526–558. 368, 370, 371, 385

Lau, J. W. and Lo, A. (2007). "Model based clustering and weighted Chinese restaurant processes." *Advances in Statistical Modeling and Inference: Essays in Honor of Kjell A. Doksum*, 405–424. 368, 371

Lijoi, A., Mena, R., and Prünster, I. (2007). "Controlling the reinforcement in Bayesian nonparametric mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4): 715–740. 385

Lo, A. Y. (1984). "On a class of bayesian nonparametric estimates: I. Density Estimates." *Annals of Statistics*, 12(1): 351–357. 374, 385, 387

MacEachern, S. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics: Simulation and Computation*, 23: 727–741. 371

— (1998). "Computational methods for Mixture of Dirichlet process models." In Dey, D., Muller, P., and Sinha, D. (eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer-Verlag. 371

Marin, J., Mengersen, K., and Robert, C. (2005). "Bayesian modelling and inference on mixtures of distributions." In Dey, D. and Rao, C. (eds.), *Handbook of Statistics*, volume 25, 459–507. Elsevier. 379

Petrone, S. and Raftery, A. (1997). "A Note on the Dirichlet Process Prior in Bayesian Nonparametric Inference with Partial Exchangeability." *Statistics and Probability Letters*, 36: 39–83. 368

Quintana, F. and Iglesias, P. (2003). "Bayesian clustering and product partition models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 557–574. 371

Stephenson, A. and Tawn, J. (2004). "Bayesian Inference for Extremes: Accounting for the Three Extremal Types." *Extremes*, 7: 297–307. 368

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, 101(416): 1566–1581. 385

Tressou, J. (2006). "Non Parametric Modelling of the Left Censorship of Analytical Data in Food Risk Exposure Assessment." *Journal of the American Statistical Association*, 101(476): 1377–1386. 368, 380

Tressou, J., Crépet, A., Bertail, P., Feinberg, M. H., and Leblanc, J. C. (2004). "Probabilistic exposure assessment to food chemicals based on Extreme Value Theory. Application to heavy metals from fish and sea products." *Food and Chemical Toxicology*, 42(8): 1349–1358. 367, 369, 378

van der Voet, H., de Mul, A., and van Klaveren, J. D. (2007). "A probabilistic model for simultaneous exposure to multiple compounds from food and its use for risk-benefit assessment." *Food and Chemical Toxicology*, 45(8): 1496–1506. 368

Acknowledgments

The author would like to thank Professor Albert Y. Lo, Professor Lancelot F. James and Dr John W. Lau for their continuous support, and for the useful comments they made on earlier versions of this paper. She is also grateful to the Editor, Associate Editor and referee for their valuable suggestions. Any error however remains hers. The author's research was supported in part by Hong Kong RGC Grant #601906 when she was visiting Hong Kong University of Science and Technology.