Bayesian density estimation and model selection using nonparametric hierarchical mixtures

Bayesian density estimation and model selection using nonparametric hierarchical mixtures

Raffaele Argiento[a], Alessandra Guglielmi[b] and Antonio Pievatolo[a]

IMATI-CNR[a] and Politecnico di Milano[b]

20th October 2008

Abstract

We consider mixtures of parametric densities on the positive reals with a normalized generalized gamma process (Brix, 1999) as mixing measure. This class of mixtures encompasses the Dirichlet process mixture (DPM) model, but it is supposedly more flexible in the detection of clusters in the data. With an almost sure approximation of the posterior distribution of the mixing process we can run a Markov chain Monte Carlo algorithm to estimate linear and nonlinear functionals of the predictive distributions. The best-fitting mixing measure is found by minimizing a Bayes factor for parametric against non-parametric alternatives. We illustrate the method with simulated and hystorical data, finding a tradeoff between the best-fitting model and the correct identification of the number of components in the mixture.

Keywords: Bayes factor, Bayesian nonparametrics, MCMC algorithms, mixture models, normalized completely random measures.

AMS 2000 Mathematics Subject Classification: 62F15, 62G07, 65C40.

1 Introduction

This paper considers the problem of density estimation on the positive reals from a Bayesian nonparametric viewpoint, using hierarchical mixture models. A typical choice, which dates back to Ferguson (1983) and Lo (1984), is to assume the unknown density as a mixture of a parametric family with a (discrete) random probability P as mixing distribution; for instance, if P is a Dirichlet process, the hierarchical mixture is the well-known Dirichlet process mixture (DPM) model. DPM models proved flexible enough to provide good density estimation, at least when assuming one more step in the hierarchy, i.e., when the total mass or (possibly) the parameters of the base measure are random. However, the prior distribution on the number of components resulting from DPM has been often criticized. Thus, with the aim of providing

valid alternatives to DPMs which could have a more elaborate clustering structure, some papers have appeared very recently, where the unknown density is modelled as a hierarchical mixture with different discrete random mixing probability P. For instance, Lijoi, Mena and Prünster (2005, 2007); Griffin (2006); Navarrete, Quintana and Müller (2008). All these papers assume, as a prior for P, a species sampling model (introduced in Pitman, 1996), or a normalized random measure with independent increments (see James, 2002, and Regazzini, Lijoi and Prünster, 2003).

In this paper, we assume a hierarchical mixture of densities on the positive real line, namely a gamma mixture over the scale and the shape parameters. Another commonly used family of distributions on $\mathbb{R}^+$ is the Weibull, which was considered in a nonparametric hierarchical mixture to model the error distribution in the accelerated failure time regression in Argiento, Guglielmi and Pievatolo (2008). As mixing measure P, here we assume a (normalized) generalized gamma process (as denoted in Brix, 1999). This family of processes, when the "mean" measure is non-atomic, is a species sampling model, but it is not stick-breaking. Moreover, as a (normalized) random measure with independent increments (RMI), P can be represented as a discrete probability with a denumerable infinite support, thus inducing an exchangeable random partition on the positive integers (see Pitman, 2006). Among normalized RMIs, those characterized by the Gibbs product form play a very important role in Bayesian nonparametrics, because of ease of modeling and computational tractability, still keeping flexibility; see Gnedin and Pitman (2006). Both Lijoi, Prünster and Walker (2008) and Cerquetti (2008) have characterized, through different derivations, the normalized generalized gamma processes as the unique family, among the normalized RMI, with exchangeable random partitions of Gibbs form with type parameter less than 1.

The family of normalized generalized gamma (NGG) processes we consider for P is indexed by the two-dimensional parameter (σ, κ), $\sigma \in [0, 1]$ and $\kappa > 0$; the "mean" probability measure P_0 is assumed fixed and absolutely continuous. The σ parameter, as emphasized in Lijoi, Mena and Prünster (2007), controls the clustering property of the nonparametric hierarchical mixture in this case, and stands for the "type" of the exchangeable Gibbs partition. On the other hand, κ plays the role of the total mass parameter as in the Dirichlet mixture case, which is recovered here by $\sigma = 0$. Moreover, both parameters control the overall variance of the process about its mean, and therefore they are quite difficult to specify; anyhow, their prior elicitation is a key issue. One way of dealing with (σ, κ) is to formulate a full Bayesian model, but we choose (σ, κ) via a model selection procedure, where Bayes factors (BF) are computed on a grid, in order to assess the effect of changes of (σ, κ) more precisely. The BF considered is the ratio between the marginal distribution of the data when $P = P_0$ a.s. and under the nonparametric model. In other words, we compare the parametric

mixture, which plays the role of a benchmark, or of the "origin of the axes", to the (vaguer) nonparametric one, and find the parameters (σ, κ) yielding the best fit to the data. Then we do a little sensitivity analysis in the areas where the BF is smaller. In case one wants to stick with the full Bayesian analysis, the surface of the BF over the (σ, κ)-space can be used, in an empirical Bayes fashion, to build a prior for (σ, κ).

In this paper, the emphasis is on modelization (especially the choice of a conjugate P_0 vs. a nonconjugate P_0 and the resulting prior marginal distributions) and computational issues.

With regards to modelization, we are going to study a mixture of gamma kernels, mixed on both the shape and the rate. A crucial point is the choice of the mean P_0. When P_0 is conjugate to the kernel, the computation of the posterior quantities of interest is simplified, since a closed-form expression for the univariate prior marginal is available. On the other hand, since P_0 gives the marginal prior of the data, P_0 should be flexible enough in order to actually represents our prior belief. Therefore we assume P_0 as the product of two gamma distributions yielding both conjugate and nonconjugate models, and achieving an acceptable degree of flexibility.

Turning to computation, in the nonparametric Bayesian literature, posterior analysis is very often achieved by the integration of the nonparametric component P. In this way computations are drastically simplified, and the posterior estimates of linear functionals are computed through predictive distributions induced by P (*e.g.* generalized Polya urn schemes). This approach is analytically convenient, but restricts the analysis to point estimation. In order to pursue a full nonparametric Bayesian inference, as in Gelfand and Kottas (2002), we build a Gibbs sampler algorithm, similarly to Nieto-Barajas and Prünster (2008). We do not integrate out the process P, but simulate a finite dimensional approximation of its posterior trajectories to obtain the entire posterior distributions of general functionals of P, such as population distribution functions, density functions or quantiles. We also show almost sure convergence of the approximated functionals to the true ones as the dimension of the approximation increases.

As far as computations of Bayes factors are concerned, we evaluate the joint marginal density under the nonparametric mixture using a generalized weighted Chinese restaurant (GWCR) algorithm (Ishwaran and James, 2003). The GWCR method is a sequential importance sampler where the marginal nonparametric density is computed as the mean of the importance weights. We point out that the computational cost of the GWCR algorithm grows substantially when nonconjugate models are considered.

The paper is organized as follows. Section 2 introduces the nonparametric hierarchical mixture model and recalls the properties of the NGG process, the predictive structure of a sample from it and its posterior distribution, which are useful for computations. In Sections

3 and 4 we illustrate the Gibbs sampler algorithm and the simulation of finite dimensional approximations of P to be used therein. In Section 5 we show the convergence of the finite dimensional approximation, with proofs deferred to the Appendix. Section 6 presents the computation of BFs via the GWCR algorithm. After examining the computational implications of having a nonconjugate P_0 in Section 7, we illustrate the method with simulated and real data (Section 8), finding a tradeoff between the best fitting model according to the BF and the correct identification of the number of components in the mixture. However our experiments show that many (σ, κ) pairs lead to both a well-fitting model and a correct identification of the number of components, so that there is scope for assuming a bivariate full support prior distribution for (σ, κ).

2 Nonparametric hierarchical mixtures using NGG processes

The model we consider can be hierarchically expressed as follows

$$\begin{aligned} V_i|\theta_i &\overset{ind}{\sim} k(\cdot;\theta_i), \\ \theta_i|P &\overset{iid}{\sim} P, \\ P \sim q, \quad P_0(A) &:= E_q(P(A)),\ A \in \mathcal{B}(\Theta), \end{aligned} \tag{1}$$

where $k(\cdot;\theta_i)$ is a family of densities on $\mathbb{R}^+$, depending on a vector of parameters θ_i belonging to a Borel subset Θ of $\mathbb{R}^s$, and q is the prior distribution on the random distribution function P; P_0 is a non-atomic distribution function (d.f.) on Θ, expressing the "mean" of P, and $\mathcal{B}(\Theta)$ denotes the Borel σ-field as usual. Model (1) will be addressed here as *nonparametric hierarchical mixture* model. The Bayesian model specification is completed assuming that P_0 depends on s hyperparameters $\gamma_1, \ldots, \gamma_s$ (possibly random and distributed according to $\pi(\gamma_1, \ldots, \gamma_s)$). In the paper, we will assume that P is a (normalized) generalized gamma measure on Θ, following the notation in Brix (1999).

2.1 Definition of the NGG process P

Let μ be a random measure on $(\Theta, \mathcal{B}(\Theta))$, let $\sigma \in [0, 1]$, $\omega \geq 0$ be nonnegative parameters, and $\kappa(\cdot)$ a (non-negative) non-atomic finite measure on Θ; Θ can be any Polish space. We say that μ is a *generalized gamma measure* if μ is *completely random*, i.e., $\mu(B_1), \ldots, \mu(B_k)$ are mutually independent if $B_1, \ldots, B_k \in \mathcal{B}(\Theta)$ are disjoint, and for any $B \in \mathcal{B}(\Theta)$, $\mu(B)$ has moment generating function

$$\mathbb{E}(\mathrm{e}^{-s\mu(B)}) = \exp\Big\{ -\frac{\kappa(B)}{\sigma}[(\omega + s)^\sigma - \omega^\sigma]\Big\},\ s \geq 0.$$

By Kingman's representation theorem of completely random measures, μ can be represented as follows (see, for instance, Brix, 1999)

$$\mu(B) = \int_{[0,+\infty)} yN(dy, B), \quad B \in \mathcal{B}(\Theta),$$

where N is a Poisson random measure on $[0, +\infty) \times \Theta$ with mean measure ν defined by

$$\nu(A \times B) = \kappa(B) \int_A \rho(ds), \quad A \in \mathcal{B}([0, +\infty)),\ B \in \mathcal{B}(\Theta),$$

and

$$\rho(ds) = \frac{1}{\Gamma(1-\sigma)} s^{-\sigma-1} \mathrm{e}^{-\omega s} ds, \quad s > 0. \tag{2}$$

Moreover μ has no fixed atoms (since κ is non-atomic), i.e. $\mathbb{P}(\mu(\{x\}) > 0) = 0$ for all x and it is almost surely purely atomic.

A random probability P can be built from a generalized gamma random measure μ according to a standard construction via normalization of completely random measures, which dates back to Kingman (1975); see also James (2002), Regazzini *et al.* (2003), or Pitman (2003). In fact, since $\int_{[0,+\infty)\times B} \min(s,1)\nu(ds,dy) = \kappa(B) \int_{[0,+\infty)} \min(s,1)\rho(ds) < +\infty$, then $\mathbb{P}(\mu(\Theta) =: T < +\infty) = 1$, so that

$$P(\cdot) := \frac{\mu(\cdot)}{T}$$

defines a random probability measure (r.p.m.) on Θ, which will be called *normalized generalized gamma process*, $P \sim NGG(\sigma, \kappa(\Theta), \omega,\ P_0)$, with parameters $(\sigma, \kappa(\Theta), \omega,\ P_0)$, where $0 \le \sigma \le 1$, $\omega \ge 0$, $P_0(\cdot) := \kappa(\cdot)/\kappa(\Theta)$. This parameterization is not unique, as the scaling property in Pitman (2003) shows, since $(\sigma, \kappa(\Theta), \omega,\ P_0)$ and $(\sigma, s^\sigma \kappa(\Theta), \omega/s, P_0)$ (for any $s > 0$) yield the same distribution for P.

Thanks to the above-mentioned Kingman's theorem, the process P can be represented as

$$P = \sum_{i=1}^{+\infty} P_i \delta_{\tau_i} = \sum_{i=1}^{+\infty} \frac{J_i}{T} \delta_{\tau_i}, \tag{3}$$

where $P_i := \frac{J_i}{T}$, $(J_i)_i$ are the ranked points of a Poisson process on $[0, +\infty)$ with mean intensity $\rho(ds)$, and $T = \sum_i J_i$; the sequences $(P_i)_{i\ge1}$ and $(\tau_i)_{i\ge1}$ in (3) are independent, and τ_i are i.i.d. from P_0. Observe that $P_1 \ge P_2 \ge P_3 \ge \ldots$ and $\sum_{i=1}^{+\infty} P_i = 1$ with probability 1. The distribution of $(P_i)_i$ is known as the *Poisson-Kingman distribution* with Lévy density ρ from Pitman (2003). The NGG process is also (when its parameter measure is non-atomic) a special case of *species sampling* models, introduced by Pitman (1996).

Since the NGG process selects discrete distributions with probability one, sampling from P induces a random partition $\mathbf{\Pi}$ on the positive integers; in fact, if we consider an infinite

sequence $(\theta_i)_i$ such that, for any n, $(\theta_1, \ldots, \theta_n)$ is a sample from P, and $\underline{\psi} := (\psi_1, \ldots, \psi_{n(\mathbf{\Pi})})$ denotes the distinct values in $(\theta_1, \ldots, \theta_n)$, then the marginal prior distribution of $(\theta_1, \ldots, \theta_n)$ is identified by the joint distribution of $\mathbf{\Pi}$ and $\underline{\psi}$. If $\mathbf{\Pi_n}$ is the restriction of $\mathbf{\Pi}$ to $\{1, \ldots, n\}$, then

$$\begin{aligned} \mathbb{P}(\mathbf{\Pi_n} = \pi_n, \psi_1 \in B_1, \ldots, \psi_{n(\pi)} \in B_{n(\pi)}) &= \mathbb{P}(\mathbf{\Pi_n} = \pi_n) \cdot \prod_{j=1}^{n(\pi)} P_0(B_j) \\ &= p(e_1, \ldots, e_{n(\pi)}) \cdot \prod_{j=1}^{n(\pi)} P_0(B_j), \end{aligned} \tag{4}$$

where

$$\pi = \pi_n = \{C_1, \ldots, C_{n(\pi)}\}, \quad C_j = \{i : \theta_i = \psi_j\}, \quad e_j := \#C_j \geq 1,\ j = 1, \ldots, n(\pi),$$

and of course $\sum_1^{n(\pi)} e_j = n$. The symmetric and non-negative function p in (4) is known as *exchangeable partition probability function* (EPPF) determined by $\mathbf{\Pi}$; see, for instance, Pitman (2006).

In the rest of the paper we will frequently use the notation $\kappa := \kappa(\Theta)$. Generally, the finite dimensional distributions of P are not available in closed analytic form, but the first two moment measures of P are given (see James, Lijoi and Prünster, 2006) by

$$\mathbb{E}\left(P(B)\right) = P_0(B), \quad \mathrm{Var}\left(P(B)\right) = P_0(B)(1 - P_0(B))\mathcal{I}(\sigma, \kappa), \tag{5}$$

while

$$\mathrm{Cov}(P(B_1), P(B_2)) = \Big(P_0(B_1 \cap B_2) - P_0(B_1)P_0(B_2)\Big)\mathcal{I}(\sigma, \kappa), \tag{6}$$

where

$$\mathcal{I}(\sigma, \kappa) := \left(\frac{1}{\sigma} - 1\right)\left(\frac{\kappa}{\sigma}\right)^{1/\sigma} \exp(\frac{\kappa}{\sigma})\Gamma\left(\frac{1}{\sigma}, \frac{\kappa}{\sigma}\right) = \left(\frac{1}{\sigma} - 1\right)\int_1^{+\infty} \mathrm{e}^{-\frac{\kappa}{\sigma}(y-1)} y^{-\frac{1}{\sigma}-1} dy \tag{7}$$

and $\Gamma(\alpha, x) := \int_x^{+\infty} \mathrm{e}^{-t} t^{\alpha-1} dt$ denotes the incomplete gamma function. The factor $\mathcal{I}(\sigma, \kappa)$ is decreasing as a function of each variable alone, and goes to 0 as $\sigma \to 1$ or $\kappa \to +\infty$, so that $P(B)$ converges in distribution to $P_0(B)$ for any B in $\mathcal{B}(\Theta)$. On the other hand, considering the integral expression in (7), it can be shown that $\lim_{\sigma \to 0, \kappa \to 0} \mathcal{I}(\sigma, \kappa) = 1$ and $P(B) \xrightarrow{d} \delta_\tau(B)$, where τ is an r.v. with distribution P_0. If $\sigma = 0$ and $\kappa > 0$ we recover the Dirichlet process with parameter κP_0, while P is a normalized inverse-gaussian (NIG) process (Lijoi *et al.*, 2005) for $\sigma = 1/2$. If $\omega = 0$ and $0 < \sigma < 1$, P is the Poisson-Dirichlet process with two parameters $(\sigma, 0)$. We will recall these limit behaviours when considering Bayes factors in Section 6.

2.2 Predictive distributions of a sample from P

If $P \sim NGG(\sigma, \kappa, \omega, P_0)$, the predictive distributions of θ_{n+1}, given $(\theta_1, \ldots, \theta_n)$, where $\theta_1, \theta_2, \ldots$, conditioning on P, are i.i.d. from P, can be represented as

$$P(\theta_{n+1} \in B | \theta_1, \ldots, \theta_n) = w_0(n, k; \sigma, \kappa) P_0(B) + w_1(n, k; \sigma, \kappa) \sum_{j=1}^{k} (e_j - \sigma) \delta_{\psi_j}(B) \tag{8}$$

where k is the number of distinct observations in $(\theta_1, \ldots, \theta_n)$ and

$$\begin{aligned} w_0(n, k; \sigma, \kappa) &= \frac{p(e_1, \ldots, e_k, 1)}{p(e_1, \ldots, e_k)} = \frac{\sigma}{n} \frac{\sum_{i=0}^{n} \binom{n}{i} (-1)^i (\frac{\kappa}{\sigma})^{i/\sigma} \Gamma(k + 1 - i/\sigma; \frac{\kappa}{\sigma})}{\sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i (\frac{\kappa}{\sigma})^{i/\sigma} \Gamma(k - i/\sigma; \frac{\kappa}{\sigma})} \\ w_1(n, k; \sigma, \kappa) &= \frac{p(e_1, \ldots, e_{j+1}, \ldots, e_k)}{p(e_1, \ldots, e_k)} = \frac{1}{n} \frac{\sum_{i=0}^{n} \binom{n}{i} (-1)^i (\frac{\kappa}{\sigma})^{i/\sigma} \Gamma(k - i/\sigma; \frac{\kappa}{\sigma})}{\sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i (\frac{\kappa}{\sigma})^{i/\sigma} \Gamma(k - i/\sigma; \frac{\kappa}{\sigma})}, \end{aligned} \tag{9}$$

for any $k = 1, \ldots, n$. The NGG process "prediction mechanism" is quite interesting and exploits the available information about the partition associated with the sample $\theta_1, \ldots, \theta_n$; indeed, the next observation θ_{n+1} is different from the previous ones with probability $w_0(n, k; \sigma, \kappa)$ and, for any $1 \leq j \leq k$, coincides with ψ_j with probability $w_1(n, k; \sigma, \kappa)(e_j - \sigma)$. Moreover, the (prior) distribution of P induces a (prior) distribution on $n(\mathbf{\Pi}_n)$, the number of distinct observations in a sample of size n from the NGG process,

$$\mathbb{P}(n(\mathbf{\Pi}_n) = k) = \mathbf{S}(n, k; \sigma) \frac{\mathrm{e}^{\frac{\kappa}{\sigma}}}{\sigma \Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i (\frac{\kappa}{\sigma})^{\frac{i}{\sigma}} \Gamma\left(k - \frac{i}{\sigma}; \frac{\kappa}{\sigma}\right), k = 1, \ldots, n, \tag{10}$$

where $\mathbf{S}(n, k; \sigma)$ denotes the generalized Stirling numbers of the first kind. See Lijoi, Mena and Prünster (2007) or Cerquetti (2007) for derivations of formulas (9) and (10).

As recalled in the Introduction, NGG processes are the unique family, among the (normalized) RMIs, with exchangeable random partitions of Gibbs form. In fact, the EPPF and the functions w_0, w_1 are represented in terms of nonnegative weights $V_{n,k}$, which are the solution of the backward equation

$$V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1}, \quad k = 1, \ldots, n, \quad V_{1,1} = 1. \tag{11}$$

For instance, as showed in Gnedin and Pitman (2006), w_0 coincides with the ratio $V_{n+1,k+1}/V_{n,k}$. We will make use of formula (11) in the GWCR algorithm when computing Bayes factors, since it avoids direct evaluation of $w_0(n^*, k; \sigma, \kappa)$ and $w_1(n^*, k; \sigma, \kappa)$ through (9), at least for $n^* < n$.

2.3 The posterior distribution of P

Unfortunately, the distribution of P is not conjugate; however, a posterior characterization of the NGG process is given in James, Lijoi and Prünster (2008), in terms of the latent

variable $U := \Gamma_n/T$, where $\Gamma_n \sim gamma(n,1)$. Here we briefly describe the distribution of $P = \mu/T \sim NGG(\sigma,\kappa,\omega,P_0)$, given a sample $\theta_1,\ldots,\theta_n$ from P and the latent variable $U=u$, in order to give details on the MCMC algorithm of Section 3. The posterior distribution of μ, given u, the vector $\underline{\psi} = (\psi_1,\ldots,\psi_{n(\pi)})$ of distinct values in $\underline{\theta}$ and $\pi = \{C_1,\ldots,C_{n(\pi)}\}$ is equal to the distribution of

$$\mu^* := \mu_{n(\pi)} + \sum_{j=1}^{n(\pi)} L_j \delta_{\psi_j}, \tag{12}$$

where $\mu_{n(\pi)}$ is a generalized gamma process with parameters $(\sigma,\kappa,\omega+u,P_0)$, and $(L_j)_1^{n(\pi)}$, conditionally on $\underline{\theta}$ and $U=u$, are independent of $\mu_{n(\pi)}$, each L_j being $gamma(e_j-\sigma,\omega+u)$-distributed. Therefore the posterior distribution of P, given u and the vector $\underline{\psi} = (\psi_1,\ldots,\psi_{n(\pi)})$, is the distribution of

$$P^* := \frac{\mu^*}{T^*} = \frac{1}{T_{n(\pi)} + \sum_{j=1}^{n(\pi)} L_j} \sum_{j=1}^{+\infty} J_j \delta_{\tau_j} + \frac{1}{T_{n(\pi)} + \sum_{j=1}^{n(\pi)} L_j} \sum_{j=1}^{n(\pi)} L_j \delta_{\psi_j}, \tag{13}$$

where

$$T^* := \mu^*(\Theta) = T_{n(\pi)} + \sum_{j=1}^{n(\pi)} L_j, \quad T_{n(\pi)} := \mu_{n(\pi)}(\Theta).$$

The $(J_j)_j$'s in (13) are the jumps from representation (3) of a $NGG(\sigma,\kappa,\omega+u,P_0)$ process.

3 Algorithm for posterior estimates

Under (1), the Bayesian estimate of the "true" density is

$$f_{V_{n+1}}(v|V_1,\ldots,V_n) = \mathbb{E}(g_P(v)|V_1,\ldots,V_n) = \int_{\mathcal{P}} \left\{ \int_\Theta k(v;\theta) P(d\theta) \right\} \mathcal{L}(dP|V_1,\ldots,V_n), \tag{14}$$

where $g_P(v) = \int_\Theta k(v;\theta)P(d\theta)$, and $\mathcal{P}$ denotes the space of all probability measures on Θ, endowed with the σ-algebra of the weak convergence. Of course, simple analytic expressions for (14) do not exist, so that it should be evaluated by a Markov Chain Monte Carlo integration. If we sample a Markovian sequence of trajectories $\left\{P^{(b)}\right\}_{b=1}^B$ from $\mathcal{L}(dP|V_1,\ldots,V_n)$ with B large enough, then an estimate of $f_{V_{n+1}}(v|V_1,\ldots,V_n)$ is the ergodic mean

$$\frac{1}{B}\sum_{b=1}^B \int_\Theta k(v;\theta) P^{(b)}(d\theta) = \frac{1}{B}\sum_{b=1}^B \left\{ \sum_{j=1}^{\infty} k(v;Y_j^{(b)}) P_j^{(b)} \right\}.$$

The last equality follows since each $P^{(b)} = \sum_1^{+\infty} P_j^{(b)} \delta_{Y_j^{(b)}}$; see (3).

The aim is then to simulate a draw P from $\mathcal{L}(dP|V_1,\ldots,V_n)$. If $\underline{\theta}=(\theta_1,\ldots,\theta_n)$ is a sample from P and $U:=\Gamma_n/T$ is the auxiliary variable introduced in Section 2.3, of course we have

$$\mathcal{L}(dP|V_1,\ldots,V_n)=\int_{\Theta^n\times(0,+\infty)}\mathcal{L}(dP,d\underline{\theta},du|V_1,\ldots,V_n)\ .$$

Hence we will build a Gibbs sampler, sequentially drawing from the full conditionals

a : $\mathcal{L}(dP|\underline{\theta},u,V_1,\ldots,V_n)$

b : $\mathcal{L}(d\underline{\theta}|P,u,V_1,\ldots,V_n)$

c : $\mathcal{L}(du|P,\underline{\theta},V_1,\ldots,V_n)$.

Sampling from $\mathcal{L}(dP|\underline{\theta},u,V_1,\ldots,V_n)$.

First of all, conditionally on $\underline{\theta}$, the posterior law of P does not depend on $V_1,\ldots,V_n$, i.e., $\mathcal{L}(dP|\underline{\theta},u,V_1,\ldots,V_n)=\mathcal{L}(dP|\underline{\theta},u)$. Let $\underline{\theta}$ be a sample from $P\sim NGG(\sigma,\kappa,\omega,P_0)$, with $P=(1/T)\sum_{j\geq1}J_j\delta_{\tau_j}$, and let $\underline{\psi}$ the vector of distinct observations in $\underline{\theta}$. Clearly the values in $\underline{\psi}=(\psi_1,\ldots,\psi_{n(\pi)})$ are a finite subset of $\{\tau_j\}_{j\geq1}$; we denote with $\mathbf{J}^{(a)}=\{J_1^{(a)},\ldots,J_{n(\pi)}^{(a)}\}$ the set of *assigned* weights, i.e., the set of the weights $\{J_j\}_{j\geq1}$ corresponding to some δ_{ψ_i} in the representation of P, for any $i=1,\ldots,n(\pi)$. Similarly let $\mathbf{J}^{(un)}=\{J_j\}_{j\geq1}\setminus\mathbf{J}^{(a)}$ be the set of *unassigned* weights. In this way we can express the posterior distribution of a process $NGG(\sigma,\kappa,\omega,P_0)$, conditionally on u and $\underline{\theta}$, as the law of a random probability measure, which is a mixture between a $NGG(\sigma,\kappa,\omega+u,P_0)$ process and a discrete probability measure with support given by the (observed) distinct values $\underline{\psi}$, i.e., as the law of

$$P^*=\frac{T_{n(\pi)}}{T_{n(\pi)}+\sum_{j=1}^{n(\pi)}J_j^{(a)}}\sum_{j=1}^{+\infty}P_j^{(un)}\delta_{\tau_j}+\frac{\sum_{j=1}^{n(\pi)}J_j^{(a)}}{T_{n(\pi)}+\sum_{j=1}^{n(\pi)}J_j^{(a)}}\sum_{j=1}^{n(\pi)}P_j^{(a)}\delta_{\psi_j}. \tag{15}$$

where $T_{n(\pi)}$ is defined immediately after (13).

Sampling from $\mathcal{L}(d\underline{\theta}|P,u,V_1,\ldots,V_n)$.

By Bayes' theorem, we have

$$\mathcal{L}(d\underline{\theta}|P,u,V_1,\ldots,V_n)\propto f(V_1,\ldots,V_n\mid\underline{\theta},P,u)\mathcal{L}(d\underline{\theta}\mid V_1,\ldots,V_n,P,u)=\prod_{i=1}^{n}k(V_i;\theta_i)\mu^*(d\theta_i)\ .$$

Then, a posteriori, $(\theta_1,\ldots,\theta_n)$ are independent, each with distribution proportional to

$$k(v_i;\theta_i)\mu^*(d\theta_i)=\sum_{j=1}^{\infty}J_j^{(un)}k(v_i;\theta_i)\delta_{\tau_j}(d\theta_i)+\sum_{j=1}^{n(\pi)}J_j^{(a)}k(v_i;\theta_i)\delta_{\psi_j}(d\theta_i)\ ;$$

see (12)-(13).

Sampling from $\mathcal{L}(du|P, \underline{\theta}, V_1, \ldots, V_n)$.

The conditional law $\mathcal{L}(du|P, \underline{\theta}, V_1, \ldots, V_n)$ of the latent random variable u is absolutely continuous with density

$$f(u|\theta) \propto (u+\omega)^{n(\pi)\sigma - n}\ u^{n-1} \exp\left\{-\frac{\kappa}{\sigma}(u+\omega)^{\sigma}\right\};$$

see James, Lijoi and Prünster (2008).

We observe that in the algorithm just described the ψ_j value corresponding to an assigned weight changes only when the weight disappears (i.e. no more θ_i's are equal to ψ_j). This phenomenon slows down the convergence of the algorithm dramatically; then, as in Bush and McEachern (1996) for the Polya-urn type algorithms, we introduced an acceleration step by updating the ψ_j components via their posterior distributions, which are proportional to

$$P_0(d\psi_j) \prod_{i \in C_j} k(v_i; \psi_j), \quad j = 1, \ldots, n(\pi).$$

4 Simulation of the trajectories of a NGG process

In the algorithm we have just described (in particular in step **a** of the Gibbs sampler) we should simulate the trajectories $P(d\theta) = \frac{1}{T}\sum_{j=1}^{\infty} J_j \delta_{\tau_j}(d\theta)$ of a NGG process exactly, for any choice of the parameters. This is an infinite sum, so that we can only simulate a finite number M of J_j's and τ_j's. Consequently, a criterion for choosing M and an assessment of convergence for functionals of P based on this approximation will be necessary.

As mentioned in Section 2.1, $(\tau_j)_j$ is an i.i.d. sequence from P_0, independent from $(J_j)_j$, which are the points of a Poisson process with intensity $\rho(ds)$ in (2), and $T = \sum_1^{+\infty} J_j$. As Ferguson and Klass (1972) proposed for the Dirichlet process case, let $R(x) = \int_x^{\infty} \rho(ds) = (\kappa\omega^{\sigma}/\Gamma(1-\sigma))\Gamma(-\sigma; \omega x)$. It is well known that, if $(\eta_j)_j$ is a sequence of points from a homogeneous Poisson process with unit intensity, then

$$J_j = R^{-1}(\eta_j) = \frac{1}{\omega}\Gamma^{-1}\left(-\sigma; \frac{\eta_j \Gamma(1-\sigma)}{\omega^{\sigma}\kappa}\right) \qquad j = 1, 2, \ldots,$$

(Γ^{-1} denotes the inverse of $x \mapsto \Gamma(-\sigma, x)$), where η_j are obtained as the cumulative sum of i.i.d. standard exponential random variables. Since R^{-1} is non-increasing, the sequence $(J_j)_j$ is sorted in decreasing order. The efficient inversion of the incomplete Gamma function is not a trivial task, so we built an algorithm for this purpose. In practice we simulate only a finite number of random variables $(J_1, \ldots, J_M)$: in the algorithm,

$$\mu_M(d\theta) = \sum_{j=1}^{M} J_j \delta_{\tau_j}(d\theta)$$

denotes the approximation of $\mu(d\theta)$.

As far as the choice of M is concerned, similarly as in Brix (1999), let c_j be the $\epsilon 2^{M-j}$– quantile of the distribution of η_j (clearly $\eta_j \sim gamma(j,1)$), $j = M+1, M+2, \ldots$ and $\epsilon > 0$. First of all, since $R(\cdot)$, as well as $\Gamma(-\sigma;\cdot)$, are non-increasing, $R^{-1}(\cdot)$ and $\Gamma^{-1}(-\sigma;\cdot)$ are non-increasing too. Then $\{\eta_j \geq c_j \, \forall j > M\} = \{R^{-1}(\eta_j) \leq R^{-1}(c_j) \, \forall j > M\} \subseteq \{\sum_{M+1}^{+\infty} R^{-1}(\eta_j) \leq \sum_{M+1}^{+\infty} R^{-1}(c_j)\}$, so that

$$\mathbb{P}\left(\sum_{M+1}^{+\infty} R^{-1}(\eta_j) \leq \sum_{M+1}^{+\infty} R^{-1}(c_j)\right) \geq \mathbb{P}\left(\eta_j \geq c_j \; \forall j > M\right) = 1 - \mathbb{P}\left(\cup_{M+1}^{+\infty}\{\eta_j < c_j\}\right)$$
$$\geq 1 - \sum_{M+1}^{+\infty} \mathbb{P}(\eta_j < c_j) \geq 1 - \sum_{j>M} \epsilon 2^{M-j} = 1 - \epsilon.$$

Moreover, observe that, since $\Gamma^{-1}(-\sigma;u) \leq (\sigma u)^{-1/\sigma}$, then

$$\sum_{j>M} R^{-1}(c_j) = \sum_{j>M} \frac{1}{\omega}\Gamma^{-1}\left(-\sigma; c_j \frac{\Gamma(1-\sigma)}{\omega^\sigma \kappa}\right) \leq \sum_{j>M} \frac{1}{\omega}\left(\sigma c_j \frac{\Gamma(1-\sigma)}{\omega\kappa}\right)^{-1/\sigma}$$
$$= \left(\frac{\sigma\Gamma(1-\sigma)}{\kappa}\right)^{-1/\sigma} \sum_{j>M} c_j^{-1/\sigma}.$$

We will choose M such that

$$\left(\frac{\sigma\Gamma(1-\sigma)}{\kappa}\right)^{-1/\sigma} \sum_{j>M} c_j^{-1/\sigma} < \tilde{\eta}\, \mathbb{E}(T),$$

for a suitably small $\tilde{\eta}$, and approximate P in (1) by

$$P_M := \frac{\mu_M}{T_M} = \sum_{j=1}^{M} \frac{J_j}{T_M}\delta_{\tau_j}\,, \qquad T_M := \sum_{j=1}^{M} J_j\,. \tag{16}$$

5 Inference for functionals

If $K(\cdot;\theta)$ denotes the d.f. of the density $k(\cdot;\theta)$, we rewrite the model we are studying as

$$V_1, \ldots, V_n | P \overset{iid}{\sim} G_P, \;\; \text{where } G_P(v) := \int_\Theta K(v;\theta)P(d\theta), \quad P \sim NGG(\sigma,\kappa,\omega,P_0).$$

Clearly, G_P is a random distribution on $\mathbb{R}^+$; let $H(\cdot)$ be a functional on the space of the distributions on $\mathbb{R}^+$. Our aim is the computation of Bayesian inferences for G_P through its posterior distribution via the Gibbs sampler outlined in Section 3. The only step there where the simulation of P (and its functionals) is involved is step **a**. We will be essentially interested in three particular functionals:

- the "distribution function-at-a-point" functional defined as

$$H_c(G_P) = G_P(c) \quad \text{for each } c \in \mathbb{R}^+,$$

- the "density-at-a-point":

$$H_c(G_P) = g_P(c) = \int_\Theta k(c;\theta)P(d\theta) \quad \text{for each } c \in \mathbb{R}^+,$$

- the quantile functional

$$Q_p(G_P) = \inf_{v\in\mathbb{R}^+} \{v : G_P(v) \geq p\} \quad \text{for each } p \in [0,1].$$

Full Bayesian statistical inference on the random functionals defined above needs in general the knowledge of the posterior distribution of the infinitely dimensional parameter P; we avoid this obstacle by approximating P with some finite dimensional parameter P_M. To simplify the exposition we will illustrate our convergence results *a priori*, and Remark 2 below shows how the results can be applied *a posteriori*.

For any fixed positive integer M, let

$$G_{P_M}(\cdot) = \int_\Theta K(\cdot;\theta)P_M(d\theta), \quad v > 0,$$

where P_M and T_M are defined as in (16). Our aim is to approximate the random distribution G_P by G_{P_M}; however, does $H(G_{P_M})$ converge, in some sense, to $H(G_P)$, as M goes to $+\infty$, at least when H is one of the functionals of interest here? The answer is given by the following Propositions, which hold for any kernel $k(\cdot;\theta)$ on $\mathbb{R}^+$; the proofs are in the Appendix.

Proposition 1. *The sequence of random densities* $\left(g_{P_M}\right)_M$ *converges to* g_P *a.s. in the* $L^1(\mathbb{R}^+)$*-metric, i.e.*

$$\int_{\mathbb{R}^+} |g_P(v) - g_{P_M}(v)|\, dv \to 0 \quad a.s. \quad \text{for } M \to +\infty.$$

Moreover, the sequence of random distributions $\left(G_{P_M}\right)_M$ *converges to* G_P *a.s. in the uniform metric, i.e.*

$$\sup_{v\in\mathbb{R}^+} |G_{P_M}(v) - G_P(v)| \to 0 \quad a.s. \quad \text{for } M \to +\infty.$$

Corollary 1. *For each fixed* $v > 0$ *the sequence of random variables* $\left(G_{P_M}(v)\right)_M$ *converges a.s. to* $G_P(v)$*.*

From Corollary 1 and Lemma 1 in the Appendix we have the following:

Proposition 2. *The sequence of random variables* $\left(Q_p(G_{P_M})\right)_M$ *converges in probability to* $Q_p(G_P)$ *for almost all* p *in* $(0,1)$ *(except for a set of null Lebesgue measure).*

Proposition 3. *For any fixed $v > 0$, $\left(g_{P_M}(v)\right)_M$ converges in mean to $g_P(v)$.*

Remark 1. Proposition 1 and the majorisation result in the proof of Proposition 3 give

$$\mathbb{E}\left(\|g_{P_M}(v) - g_P(v)\|_{L^1(\mathbb{R}^+)}\right) \leq 2\int_{\mathbb{R}^+} f_V(v)\left[1 - \mathbb{E}\left(\frac{T_M}{T}\right)\right]dv = 2\left[1 - \mathbb{E}\left(\frac{T_M}{T}\right)\right].$$

Therefore $\mathbb{E}\left(\frac{T_M}{T}\right)$ quantifies how close g_{P_M} approximates g_P. The following proposition gives a bound to the error $1 - \mathbb{E}\left(\frac{T_M}{T}\right)$.

Proposition 4. *If $T_M^+ = T - T_M = \sum_{j=M+1}^{\infty} J_j$, then*

$$\mathbb{E}\left(\frac{T_M}{T}\right) \geq \mathbb{E}\left(\frac{T_M}{T_M + \xi(J_M)}\right),$$

where $\xi(s) = \mathbb{E}(T_M^+ | J_M = s) = \dfrac{\kappa}{\omega^{1-\sigma}}\left[1 - \dfrac{\Gamma(1-\sigma, \omega s)}{\Gamma(1-\sigma)}\right]$, $s > 0$.

A recursive implicit description of the joint density of T and $J_1, \ldots, J_M$ is given in Perman (1993).

Remark 2. By (15), conditionally on the auxiliary variable u, the posterior of a $NGG(\sigma, \kappa, \omega, P_0)$ process P can be expressed as a mixture between a $NGG(\sigma, \kappa, \omega+u, P_0)$ process $P^{(un)}$ (the *unassigned* term) and a discrete random probability $P^{(a)}$ with a finite number of jumps (the *assigned* term). This conditional *quasi-conjugacy* property allow us to use, for the *conditional a posteriori* distributions of the linear functionals under consideration, the approximation results in Proposition 1, 3 and 4. In fact, if H is a linear functional and P^* is a draw of the posterior law of P, conditionally on u, then

$$H(G_{P^*}) = W_1 H(G_{P^{(un)}}) + W_2 H(G_{P^{(a)}}),$$

where W_1 and W_2 are random weights which can be computed from (15). Hence we can apply the convergence results to $P^{(un)}$, while $P^{(a)}$ is a finite sum and does not need any approximation. On the other hand, as far as the posterior distribution of the quantile functional is concerned, since $G_{P^*_M}(v)$ converges a.s. to $G_{P^*}(v)$ for each $v > 0$, then, by Lemma 1, $Q_p(G_{P^*_M})$ converges in probability to $Q_p(G_{P^*})$ for almost all p in $(0, 1)$, except for a set of null Lebesgue measure.

Remark 3. We studied convergences for the three functionals of r.p.m.s of interest in our statistical analysis, but the results can be slightly generalized. For instance, Corollary 1 can be easily extended to essentially bounded linear functionals, while Proposition 3 holds whenever H is linear and such that $\int_\Theta |H(k(.;\theta))| P_0(d\theta)$ is finite.

6 Best-fit parameters vi a Bayes factors

As mentioned in the Introduction, we are looking for the best-fit parameters (σ, κ) in $[0,1) \times (0, +\infty)$ as the values which achieve the minimum of the Bayes factor between the parametric and the nonparametric alternatives. If $P \sim NGG(\sigma, \kappa, P_0)$, for a given pair (σ, κ), we evaluate the fit of the nonparametric model (1) to the data through the ratio between the marginal density of the observed data under the parametric mixing measure P_0 and under the nonparametric P. Then we choose the model corresponding to the pair (σ, κ) minimizing the BF, if the minimum BF is smaller than 1. On the contrary, BFs greater than 1 for all values of (σ, κ) denote a better fit of the parametric mixture. Examples of this type of comparison, where the parametric model is embedded into the nonparametric one, can be found in Florens, Richard and Rolin (1996), Carota and Parmigiani (1996), Berger and Guglielmi (2000). The parametric model in the numerator is simply a "benchmark", since it is the "mean" model, obtained by letting P be a.s. equal to its mean P_0. Equivalently, it can be obtained taking the limit as $\kappa \to +\infty$ or as $\sigma \to 1$, since the factor $\mathcal{I}(\sigma, \kappa)$ in the variance of $P(B)$ decreases to 0, for any $B \in \mathcal{B}(\Theta)$; see (5)-(7). There are other interesting limits of the BF on the closure of the set $R = [0,1) \times (0, +\infty)$ in the (σ, κ)-space. When $\kappa = 0$ or $\sigma = 0$, P is a Poisson-Dirichlet process with parameters $(\sigma, 0)$ or the Dirichlet process with parameter measure κP_0, respectively. Finally, if (σ, κ) tends to $(0,0)$, the process converges to a random probability measure degenerate on a random point extracted from P_0, that is, the hierarchical mixture reduces to the usual exchangeable parametric model, where P_0 is the de Finetti measure of the observations.

To set up notation, we will compute the Bayes factor of model (1) when $q = \delta_{P_0}$ versus the nonparametric model (1),

$$BF(\underline{v}; \sigma, \kappa) = \frac{m_0(v_1, \ldots, v_n)}{m(v_1, \ldots, v_n; \sigma, \kappa)}; \tag{17}$$

here m_0 is the marginal of the data under the assumption $P = P_0$ a.s.:

$$m_0(v_1, \ldots, v_n) = \int_{\Theta^n} \prod_{i=1}^{n} k(v_i; \theta_i) \prod_{i=1}^{n} P_0(d\theta_i) = \prod_{i=1}^{n} \int_{\Theta} k(v_i; \theta) P_0(d\theta).$$

while $m(v_1, \ldots, v_n; \sigma, \kappa)$ is the marginal of the data when $P \sim NGG(\sigma, \kappa, \omega, P_0)$, $\sigma \in [0,1)$, $\kappa > 0$,

$$m(v_1, \ldots, v_n; \sigma, \kappa) = \int_{\mathcal{P}} \left\{ \int_{\Theta} \prod_{i=1}^{n} k(v_i; \theta_i) \prod_{i=1}^{n} P(d\theta_i) \right\} q(dP).$$

As far as computation of (17) is concerned, similarly to Basu and Chib (2003) and Ishwaran, James and Sun (2001), we resort to a sequential importance sampling (SIS) algorithm to evaluate $m(v_1, \ldots, v_n; \sigma, \kappa)$. In particular we consider a GWCR algorithm for

species sampling model introduced by Ishwaran and James (2003). The GWCR algorithm draws values $\pi_n = \{C_1, \ldots, C_{n(\pi)}\}$ from a distribution $Q(\pi_n)$ over the space of the partitions of the set $\{1, \ldots, n\}$, where $Q(\pi_n)$ acts as an importance function for approximating $p(\pi_n)f(V_1, \ldots, V_n|\pi_n)$, i.e.,

$$p(\pi_n)f(V_1, \ldots, V_n|\pi_n) = \Lambda(\pi_n)Q(\pi_n),$$

and $\Lambda(\pi_n)$ are the importance weights. In particular, the last equality implies that

$$m(v_1, \ldots, v_n; \sigma, \kappa) = \mathbb{E}_Q\left(\Lambda(\Pi_n)\right) = \sum_{\pi_n} \Lambda(\pi_n)Q(\pi_n) \tag{18}$$

Thus, by the strong law of large numbers, if $\pi_n^{(1)}, \ldots, \pi_n^{(B)}$ is a sequence of random partition draws from Q, then

$$m(\underline{v}; \sigma, \kappa) \simeq \frac{1}{B}\sum_{i=1}^{B} \Lambda(\pi_n^{(i)}).$$

The GWCR works by building the random partition sequentially. The first observation is associated with the first cluster (i.e the index of the first observation will be in the set C_1); then after the assignment of r observations, the $r+1$-th observation is assigned to a new cluster with probability

$$\frac{w_{0,r}}{\lambda_{r+1}} \times \int_\Theta k(v_{r+1}; \theta)P_0(d\theta), \tag{19}$$

and to an existing cluster C_j, containing n_j indexes, with probability

$$\frac{w_{1,r} \times (n_j - \sigma)}{\lambda_{r+1}} \frac{\int_\Theta k(v_{r+1}; \theta) \prod_{i \in C_j} k(v_i; \theta)P_0(d\theta)}{\int_\Theta \prod_{i \in C_j} k(v_i; \theta)P_0(d\theta)}, \qquad j = 1, \ldots, n(\pi_r), \tag{20}$$

where λ_{r+1} is the appropriate normalizing constant and $w_{1,r},\ w_{0,r}$ are the weights in the prediction rule (8) for a sample of size r. The resulting partition $\pi_n = \left\{C_1, \ldots, C_{n(\pi)}\right\}$ is a draw from the GWCR density Q and the corresponding importance weight is

$$\Lambda(\pi_n) = \lambda_1 \times \cdots \times \lambda_n\ .$$

Finally, since the draws Π_n from Q depend on the order of the data, randomizing the order of the data by a *permutation step* can be very useful to reduce the variance of the weight $\Lambda(\Pi_n)$, and therefore to improve the computation of $m(\underline{v}; \sigma, \kappa)$ (see equation (18)).

7 Prior marginal distributions

As mentioned in the Introduction, we are going to consider hierarchical mixtures of gamma densities $k(\cdot; \theta)$, $\theta = (\vartheta_1, \vartheta_2)$, with mean ϑ_1/ϑ_2, and to choose the centering distribution P_0

on $\mathbb{R}^+ \times \mathbb{R}^+$ as the product of two independent gamma distributions, i.e. ϑ_1 and ϑ_2 under P_0 are independently gamma distributed with parameter (ζ_1, γ_1) and (ζ_2, γ_2) respectively. When at least one among ζ_1 and ζ_2 is different from 1, then the model is nonconjugate, that is, the computation of the integrals in (20) and (19) is not achievable analytically and we must resort to a numerical integration method. With this aim, we observe that if C is a subset of n_C indexes in $\{1, \ldots, n\}$, then

$$
\begin{aligned}
(21) \quad & \int_\Theta \prod_{i \in C} k(v_i; \theta) P_0(d\theta) \\
& = \frac{\gamma_1^{\zeta_1} \gamma_2^{\zeta_2}}{\Gamma(\zeta_2)} \frac{1}{\left(\prod_{i \in C} v_i\right) \left(\sum_{i \in C} v_i + \gamma_2\right)^{\zeta_2} (r(\underline{v}))^{\zeta_1}} \cdot \int_0^{+\infty} \frac{\Gamma(n_C \vartheta_1 + \zeta_2)}{\Gamma^{n_C}(\vartheta_1)} \Gamma(\vartheta_1 | \zeta_1, r(\underline{v})) d\vartheta_1,
\end{aligned}
$$

where $\Gamma(\cdot|s, r)$ is the density of a gamma distributed random variable with shape parameter s and rate parameter r, and $r(\underline{v}) = \gamma_1 + n_C \log\left(\left(\sum_{i \in C} v_i + \gamma_2\right) / \left(\prod_{i \in C} v_i\right)^{1/n_C} \right)$.

The unidimensional integral in (21) can be easily computed by deterministic numerical integration. If $n_C = 1$, for $v > 0$, (21) reduces to the univariate marginal prior for the variable V

$$
\begin{aligned}
(22) \quad & f_V(v) = \int_\Theta k(v; \theta) P_0(d\theta) = g_{P_0}(v) \\
& = \frac{\gamma_1^{\zeta_1} \gamma_2^{\zeta_2}}{\Gamma(\zeta_2)} \frac{1}{v (v + \gamma_2)^{\zeta_2} (\gamma_1 + \log(\frac{v + \gamma_2}{v}))^{\zeta_1}} \cdot \int_0^{+\infty} \frac{\Gamma(\vartheta_1 + \zeta_2)}{\Gamma(\vartheta_1)} \Gamma(\vartheta_1 | \zeta_1, r(v)) d\vartheta_1.
\end{aligned}
$$

This distribution has an asymptote in zero, but admits a mode for $\zeta_2 > 1$. In Figure 1 the graphics of $f_V(\cdot)$ for some choices of the hyperparameters are depicted. Moreover, f_V also admits $j-$th moment for $\zeta_2 > j$. In this case, $\mathbb{E}(V^j) = \mathbb{E}(\mathbb{E}(V^j | \vartheta_1, \vartheta_2)) = \mathbb{E}(\vartheta_1^j)\mathbb{E}(1/\vartheta_2^j)$, and $\mathbb{E}(1/\vartheta_2^j)$ exists if and only if $\zeta_2 > j$. In particular

$$
\mathbb{E}(V) = \frac{\zeta_1 \gamma_2}{(\zeta_2 - 1)\gamma_1}, \quad \zeta_2 > 1.
$$

Of course, if $\zeta_2 = 1$, $\mathbb{E}(V)$ is infinite.

8 Data illustrations

In this section, by the scaling property mentioned in Section 2.1, we may assume $\omega = 1$ and use the notation $P \sim NGG(\sigma, \kappa, P_0)$ in place of the four-parameter notation.

Before discussing the results, we briefly describe how Bayesian estimates in the examples were obtained. We drew samples from the posterior law of the functionals of the random distribution function G_P considered in Section 5. If $P_M^{(b)}$, $b = 1, \ldots, B$, is a sequence of posterior trajectories given by the Gibbs sampler (see Section 3), then $H(G_{P_M^{(b)}})$, $b = 1, \ldots, B$,

is a Markov sequence with stationary distribution equal to the law of $H(G_{P_M})$. If H is linear, computation of $H(G_{P_M^{(b)}})$ is straightforward. On the other hand, if H is the quantile functional Q_p, $p \in (0,1)$, we consider a finite grid $x_0, \ldots, x_m$, with $x_0 = 0$ and x_m large enough, so that $G_{P_M^{(b)}}(x)$ is almost 1 for any $x \geq x_m$, and any $b = 1, \ldots, B$. At each iteration we evaluate $G_{P_M^{(b)}}(x_i)$, for $i = 0, \ldots, m$, and then consider their linear interpolation, $\hat{G}_{P_M^{(b)}}$, as an approximate realization of the distribution G_{P_M}. The quantile functional $Q_p(G_{P^{(b)}})$ is assumed equal to $\hat{G}^{-1}_{P^{(b)}}(q)$ for each $b = 1, \ldots, B$.

As far as computational details are concerned, since evaluation of the weights (9) or inversion of the incomplete gamma function in the simulations of posteriors trajectories of P is computationally heavy, we coded our programs in C, using GSL and PARI libraries when necessary. Moreover, in order to reduce the variance of the importance weights in the SIS algorithm for computing Bayes factors, we resorted to a GWCR algorithm, which is equivalent to the collapsed SIS algorithm by Basu and Chib (2003). Integrals in (20) could only be computed numerically, thus slowing down the speed of the algorithm.

8.1 Simulated data

We consider simulated data from a mixture of 3 gamma densities. We generated a random sample of size 100 from the density

$$0.2 \cdot \text{gamma}(40, 20) + 0.6 \cdot \text{gamma}(6, 1) + 0.2 \cdot \text{gamma}(200, 20),$$

(the mean is 6 and the variance is 10.12). This mixture was considered in Argiento *et al.* (2007) to compare the performances of two nonparametric hierarchical mixture models, both belonging to the class of NGG processes: the Dirichlet and the normalized inverse-gaussian.

First of all, in order to find the best-fit parameters, we computed the $\min BF(\sigma, \kappa)$ over a fine enough grid in $[0, 1) \times (0, +\infty)$. We set hyperparameters $\zeta_1 = \zeta_2 = 3$, $\gamma_1 = 0.075$ and $\gamma_2 = 0.3$ so that $\mathbb{E}(V) = 6$ (the true mean) and $\text{Var}(V) = 13.41$ (slightly larger than the truth). With this choice of ζ_1 and ζ_2, f_V assigns small mass to a neighborhood of the origin; otherwise there would be too much disagreement with the sampling density. In Figure 2 the BF of the parametric model against the nonparametric mixture is plotted on a grid of 11×11 points, with $\log(\kappa) \in \{-5, -4, \ldots, 5\}$ and $\sigma \in \{0.01, 0.1, 0.2, \ldots, 0.9, 0.99\}$. The minimum of the BF on the grid is reached for $\sigma = 0.01$ and $\log(\kappa) = 1$. Nevertheless, the BF function is substantially flat and small in a region of the $(\sigma, \log(\kappa))$-plane larger than just a neighborhood of the minimum. Then we did some sensitivity analysis for $(\sigma, \kappa) \in \{0.01, 0.1, 0.2, 0.3\} \times \{-1, 0, 1\}$, and $(\sigma, \kappa) \in \{0.3, 0.4\} \times \{-5, -4\}$. We ran a 10000-iteration chain for each set of hyperparameters with a thinning of 5 iterations, after a burn-in of 1000

iterations. Figure 3 shows the true density, the histogram from the simulated data and the density estimates, together with pointwise 95% highest probability density intervals, for some experiments. The posterior distributions of the number of clusters for the (σ, κ)-values used in Figure 3 are displayed in Figure 4.

There are no substantial differences in the shape of the density estimates; nevertheless a look at the posterior credibility intervals for the 99% percentile of the predictive distributions (see Table 1) highlights how the uncertainty on the right tails increases with σ and κ. We argue that this is due to a different dispersion, as a function of σ and κ, in the number of components the algorithm uses to build the posterior estimates. This happens because the prior distribution on the number of clusters induced by a NGG process with "high" values of σ is less informative with respect to the prior induced by σ values near zero; from Figure 5, which displays the probability mass functions of the prior number of clusters for some choices of the hyperparameters, it is clear that the tails of these distributions are heavier (less *informative*) for large values of σ. This effect is confirmed by the posterior estimates of the number of clusters for "high" values of σ – $\sigma = 0.3$ or 0.4 –, which are more effective than the posterior estimates for smaller values of σ ($\sigma \in \{0.01, 0.1, 0.2\}$). In particular, $\sigma = 0.1$ gives a prior for P which is close to the Dirichlet, and a more informative prior on the number of clusters, and this affects the posterior estimates. This result agrees with analogous experiments in Argiento *et al.* (2007), where the posterior modal number of clusters is consistently larger than the truth for the DPM model.

It is worth noting that the posterior mode of the distribution of the number of clusters with hyperparameters $\sigma = 0.01$, $\kappa = \exp(1)$, corresponding to the minimum of the Bayes factor, is 10, and this value is far from 3, the true number of clusters. We explain this apparent contradiction looking at the structure of the nonparametric mixture model. The variance of the process P is strictly related to the expected number of clusters in the mixture. Thus, (σ, κ) controls both the dispersion of P around the mean and the prior expected number of components of the mixture. For example, when $(\sigma, \kappa) \to (0, 0)$, the dispersion of P is maximal and the process P reduces to a degenerate distribution on a random point from P_0 (i.e, only one cluster), while if $(\sigma, \kappa) \to (1, \infty)$ the dispersion is zero and P collapses on its mean P_0 (i.e., no ties). We argue that a great dispersion of P does not correspond to a wealth of models; instead, such wealth is attainable by increasing the expected number of clusters, that is, by increasing κ or σ, and consequently reducing the variability of P.

We conclude that the optimal hyperparameters (σ, κ) lead to posterior density estimates that fit the true density well, but also may lead to an estimation of the number of components in the mixture far from the true one. However, some hyperparameters (σ, κ) corresponding to a *quasi-optimal* Bayes factor still lead to a well-fitting density estimate and improve con-

siderably the estimate of the number of components in the mixture. The surface of the Bayes factor plotted in Figure 2 represents a good direction to start building a prior distribution on (σ, κ).

8.2 Enzyme data

We analyzed a data set of 245 observations reporting the enzymatic activity in the blood of unrelated individuals. It is hypothized that there are groups of slow and fast metabolizers. The data set has been studied by Richardson and Green (1997) and Griffin (2006), testing Bayesian parametric and nonparametric models, respectively, with gaussian kernel mixture. Both papers underline similar features of the posterior of the number of components $n(\mathbf{\Pi}_n)$ in the mixture. On one hand Richardson and Green observe that the prior specification on the variance of the kernel densities influences the posterior distribution of the number of components; on the other hand Griffin obtains a posterior number of clusters shifted on values greater than 2 or 3 (suggested by the histogram), ascribing the overestimate to the skewness of the right tail of the data. We fixed $\zeta_1 = 0.39$, $\zeta_2 = 2.5$, $\gamma_1 = 0.2$ and $\gamma_2 = 0.5$, so that $\mathbb{E}(V) = 0.62$ (the sample mean). This choice affects also the (prior) mean and variance of the kernel, since $\mathbb{E}(\vartheta_1/\vartheta_2) = \mathbb{E}(V)$ and $\mathbb{E}(\vartheta_1/\vartheta_2^2) = \gamma_2/(\zeta_2 - 2)\mathbb{E}(V)$. We computed the $\min BF(\sigma, \kappa)$ over a smaller grid than in the previous example. Figure 6 displays the BF of the parametric model against the nonparametric mixture for $\log(\kappa) \in \{-5, -4, \ldots, 0, 1\}$ and $\sigma \in \{0.01, 0.1, 0.2, \ldots, 0.5\}$. The $\min(BF)$ is obtained when $\sigma = 0.01$, $\log(\kappa) = -2$ and the BF function seems slightly flatter than before. The magnitude of the BFs is much smaller than in Example 1, because the parametric model we assume is not perfectly fitted to the data. However, as already pointed out, here the parametric model plays only the role of a benchmark, while the emphasis is on the shape of the BF as a function of σ and κ.

As in Example 1, a sensitivity analysis of the density estimates and the posterior of $n(\mathbf{\Pi}_n)$ was made, using quasi-optimal values of hyperparameters; each MCMC chain was obtained after 10000 iterations with a thinning of 5 and a burn-in of 1000. A selection of the estimates we obtained is displayed in Figures 7 and 8. Although the posterior distribution of the number of clusters is more robust with respect to the choice of (σ, κ), "high" values of σ keep leading to a better estimate of $n(\mathbf{\Pi}_n)$. Moreover, since the gamma kernel can fit the skewness of the data well, the majority of our posteriors give higher probability to the 2-component mixtures than Griffin (2006). As before, there are no substantial differences among the density estimate plots we get.

However, if the hyperparameters express a prior opinion very different from the data, the posterior number of clusters might be inconsistent with the data, for any choice of P in the family of $NGG(\sigma, \kappa)$ processes; for istance, for some hyperparameters we obtained a posterior

of $n(\mathbf{\Pi}_n)$ giving zero mass to $\{2,3\}$. After several experiments it was confirmed that, when hyperparameters in P_0 are not data-consistent, the BF does not help in choosing the best nonparametric model, since optimal or sub-optimal values of (σ,κ) do not yield a reasonable posterior of the number of clusters. We found that, for some values of hyperparameters not consistent with the data, the posterior of $n(\mathbf{\Pi}_n)$ is very disperse and centered on $\{6,7,8\}$. However, the trace plot of the latent variable (θ_1,θ_2) in Figure 9 shows that there are only 3 components in the predictive density.

As a future work, from our analyses on both data sets, it seems reasonable to assume a bivariate full support prior distribution for (σ,κ). In Argiento et al. (2008) we assumed a prior distribution degenerate on a line of the $(\sigma,\kappa)-$plane, obtaining good results on the posterior distribution of the number of random effects in an AFT regression model. These conclusions are in accordance with the well-known effect of randomization of the total mass parameter in DPM models.

Appendix

Proof of Proposition 1. It is easy to show that

$$\begin{aligned}
\|g_P - g_{P_M}\|_{L^1(\mathbb{R}^+)} &= \int_{\mathbb{R}^+} |g_P(v) - g_{P_M}(v)|\, dv \\
&= \int_{\mathbb{R}^+} \left| \frac{1}{T}\sum_{j=1}^{\infty} J_j k(v;\tau_j) - \frac{1}{T_M}\sum_{j=1}^{M} J_j k(v;\tau_j) \right| dv \\
&\leq \int_{\mathbb{R}^+} \sum_{j=1}^{M} \left| \frac{1}{T} - \frac{1}{T_M} \right| J_j k(v;\tau_j) dv + \frac{1}{T}\int_{\mathbb{R}^+} \sum_{j=M+1}^{\infty} J_j k(v;\tau_j) dv \\
\text{(by Beppo Levi's theorem)} &= \left| \frac{1}{T} - \frac{1}{T_M} \right| \sum_{1}^{M} J_j \int_{\mathbb{R}^+} k(v;\tau_j) dv + \sum_{M+1}^{\infty} J_j \int_{\mathbb{R}^+} k(v;\tau_j) dv \\
&= 2\left(1 - \frac{T_M}{T}\right) \downarrow 0 \text{ a.s. for } M \to +\infty,
\end{aligned}$$

since $T_M \uparrow T = \sum_1^{+\infty} J_j$ a.s. and $\mathbb{P}(T < +\infty) = 1$.

On the other hand, since

$$\sup_{v\in\mathbb{R}^+} |G_{P_M}(v) - G_P(v)| \leq \int_{\mathbb{R}^+} |g_P(v) - g_{P_M}(v)|\, dv,$$

the result follows. □

Lemma 1. *Let $\left(G_M(\cdot)\right)_{M>1}$ be a sequence of random distributions on the positive reals such that, for each fixed $v>0$, $G_M(v) \xrightarrow{a.s.} G(v)$, where G is a random d.f. Then $Q_p(G_M(\cdot)) \xrightarrow{Prob} Q_p(G(\cdot))$ for almost all p in $(0,1)$, except for a set of null Lebesgue measure.*

Proof. We see that, if $Y_1 := Q_p(G(\cdot)) + \epsilon$ and $Y_2 := Q_p(G(\cdot)) - \epsilon$,

$$\begin{aligned}\mathbb{P}\left(|Q_p(G_M(\cdot)) - Q_p(G(\cdot))| \leq \epsilon\right) &= \mathbb{P}\left(Q_p(G_M(\cdot)) \leq Y_1\right) - \mathbb{P}\left(Q_p(G_M(\cdot)) < Y_2\right) \\ &= \mathbb{P}\left(G_M(Y_1) \geq p\right) - \mathbb{P}\left(G_M(Y_2) > p\right) \\ &\overset{M\to+\infty}{\longrightarrow} \mathbb{P}\left(G(Y_1) \geq p\right) - \mathbb{P}\left(G(Y_2) > p\right) = 1,\end{aligned}$$

whenever p is a point of continuity of the distribution function of the random variable $G(Y_1)$. □

Proof of Proposition 3.

$$\begin{aligned}&\mathbb{E}\left(|g_P(v) - g_{P_M}(v)|\right) = \mathbb{E}\left(\left|\sum_{j=1}^{\infty} \frac{J_j}{T} k(v;\tau_j) - \sum_{j=1}^{M} \frac{J_j}{T_M} k(v;\tau_j)\right|\right) \\ &\leq \sum_{j=1}^{M} \mathbb{E}\left(J_j \left|\frac{1}{T} - \frac{1}{T_M}\right| k(v;\tau_j)\right) + \sum_{j=M+1}^{\infty} \mathbb{E}\left(\frac{J_j}{T} k(v;\tau_j)\right) \\ &= f_V(v)\left(\mathbb{E}\left(\sum_{j=1}^{M} J_j\left(\frac{1}{T_M} - \frac{1}{T}\right)\right) + \sum_{j=M+1}^{\infty} \mathbb{E}\left(\frac{J_j}{T}\right)\right) = 2 f_V(v)\mathbb{E}\left(1 - \frac{T_M}{T}\right),\end{aligned}$$

recalling that $(\tau_j)_{j\geq 1}$ and $(J_j)_{j\geq 1}$ are independent, and $f_V(v) = \mathbb{E}(k(v;\tau_j))$ is the marginal distribution of V in (1), for any j. The result holds since $\mathbb{E}(T_M) \uparrow \mathbb{E}(T)$ as M goes to $+\infty$ by the Monotone Convergence Theorem. □

Proof of Proposition 4. By Jensen's inequality

$$\mathbb{E}\left(\frac{T_M}{T}\right) = \mathbb{E}\left\{\mathbb{E}\left(\frac{\sum_{j=1}^{M} J_j}{\sum_{j=1}^{M} J_j + T_M^+}\middle| J_1, \ldots, J_M\right)\right\} \geq \mathbb{E}\left(\frac{\sum_{j=1}^{M} J_j}{\sum_{j=1}^{M} J_j + \mathbb{E}\left(T_M^+ \middle| J_1, \ldots, J_M\right)}\right).$$

To evaluate the conditional mean of T_M^+, we observe that, since $J_1 \geq J_2 \geq \cdots \geq J_M$, by elementary properties of Poisson processes, $\mathbb{E}(T_M^+|J_1, \ldots, J_M) = \mathbb{E}(T_M^+|J_M)$. Moreover, conditionally on $J_1, \ldots, J_M$, the sequence $(J)_{j>M}$ is the ranked values of points of a Poisson process with mean intensity $\rho_{J_M}(x) = \rho(x)\mathbb{1}_{(0,J_M)}(x)$.

If $f_s(x)$ denotes the density of T_M^+, conditionally on $J_M = s$, for any $s > 0$, its Laplace transform is

$$\mathbb{E}(\mathrm{e}^{-\lambda T_M^+}) = \int_0^{\infty} \mathrm{e}^{-\lambda x} f_s(x)dx = \exp(-\psi_s(\lambda)), \quad \lambda \geq 0,$$

where, according to the Lévy-Khintchine formula,

$$\psi_s(\lambda) = \int_0^{\infty} (1 - \mathrm{e}^{-\lambda x})\rho_s(x)dx = \int_0^{s} (1 - \mathrm{e}^{-\lambda x})\rho(x)dx.$$

It is straightforward to verify that the hypothesis of the differentiation under the integral theorem (see, *e.g.*, Carter and van Brunt (2000), Theorem 6.3.2) are met, so that simple integral calculations and $\psi(0) = 0$ yield

$$\xi(s) = \mathbb{E}(T_M^+ | J_M = s) = -\left. \frac{d}{d\lambda} \exp(-\psi_s(\lambda)) \right|_{\lambda=0} = \left. \frac{d}{d\lambda} \psi_s(\lambda) \right|_{\lambda=0} = \frac{\kappa}{\omega^{1-\sigma}} \left[1 - \frac{\Gamma(1-\sigma, \omega s)}{\Gamma(1-\sigma)} \right],$$

for any $s > 0$. □

Acknoledgments

Raffaele Argiento was partially funded by "Sovvenzione Globale INGENIO", an initiative of Regione Lombardia financed by the European Social Fund 2000-2006.

References

Argiento, R., Guglielmi, A. and Pievatolo, A. (2007). A comparison of nonparametric priors in hierarchical mixture modelling of lifetime data. Technical report no. 07.2-MI, CNR IMATI, Milano.

Argiento, R., Guglielmi, A. and Pievatolo, A. (2008). Nonparametric Bayesian mixture modelling for failure time data. Technical report no. 08.3-MI, CNR IMATI, Milano.

Basu, S. and Chib, S. (2003). Marginal Likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, **98**, 224-235.

Berger, J.O. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, **96**, 174–184.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Prob.*, **31**, 929-953.

Bush, C. A. and MacEachern, S. N.(1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika*, **83**, 1013-1021,

Carota, C., Parmigiani, G. (1996). On Bayes factors for nonparametric alternatives. In: *Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics*, Vol. 5, Oxford University Press, Oxford, 507-511.

Carter, M. and van Brunt, B. (2000). *The Lebesgue-Stieltjes integral*. Springer, New York.

Cerquetti A. (2007). A note on Bayesian nonparametric priors derived from exponentially tilted Poisson. *Statist. Probab. Lett.*, **77**, 1705-1711.

Cerquetti, A. (2008). On a Gibbs characterization of normalized generalized Gamma processes. To appear in *Statist. Probab. Lett.*.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics, Eds. H. Rizvi and J. Rustagi, Academic Press, New York, 287–302.*

Ferguson, T. S. and Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics*, **43**, 1634–1643.

Florens, J.P., Richard, J.F. and Rolin, J.M. (1996). Bayesian encompassing specification tests of a parametric model against a nonparametric alternative. Technical Report 96.08, Institut de Statistique, Université Catholique de Louvain, Belgium.

Gelfand, A. E. and Kottas, A. (2002). A Computational Approach for Full Nonparametric Bayesian Inference Under Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **11**, 289–305.

Gnedin, A and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, **138**, 5674-5685.

Griffin, J.E. (2006). On the Bayesian analysis of species sampling mixture models for density estimation. Technical Report, University of Warwick.

Ishwaran, H. and James, L.F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13, 1211–1235.

Ishwaran, H., James, L.F. and and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Stat. Assoc.*, **96**, 1316–1332.

James, L.F. (2002). Poisson Process Partition Calculus with applications to Exchangeable models and Bayesian Nonparametrics. The arXiv, arXiv:math.PR/0205093, 2002. Available at http://arxiv.org/abs/math.PR/0205093.

James, L.F., Lijoi, A. and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, **33**, 105-120.

James L.F., Lijoi, A. and Prünster, I. (2008). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* (in press).

Kingman, J.F.C. (1975). Random discrete distributions. *J. Roy. Statist. Soc. Ser. B*, **37**, 1–22.

Lijoi, A., Mena, R. H. and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, **100**, 1278–1291.

Lijoi, A., Mena, R. H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. R. Stat. Soc. Ser.* B, **69**, 715–740.

Lijoi, A., Prünster, I., Walker, S.G. (2008). Investigating nonparametric priors with Gibbs structure. *Statist. Sin.* (in press).

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.*, **12**, 351–357.

Navarrete, C., Quintana, F.A. and Müller, P. (2008). Some issues in nonparametric Bayesian modeling using species sampling models. *Statistical Modelling*,**8**, 3–21.

Nieto-Barajas, L.E. and Prünster, I. (2008). A sensitivity analysis for Bayesian nonparametric density estimators. *Statistica Sinica* (in press).

Perman, M. (1993). Order statistics for jumps of normalised subordinators. *Stochastic Processes and their Applications*, **46**, 267–281.

Pitman J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In: *Statistics, Probability and Game Theory* (IMS Lecture Notes MonographSeries), Vol. 30, 245–267.

Pitman J. (2003). Poisson-Kingman partitions. In D.R. Goldstein, editor, *Science and Statistics: a Festschrift for Terry Speed*, volume 40 of *Lectures Notes - Monograph Series*, 1-34. Institute of Mathematical Statistics, Hayward, California.

Pitman J. (2006). *Combinatorial stochastic processes.* LNM n. 1875, Springer, New York.

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.*, **31**, 560–585.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. (Corr: 1998V60 p661) *J. R. Stat. Soc. Ser.* B, **59**, 731–758.

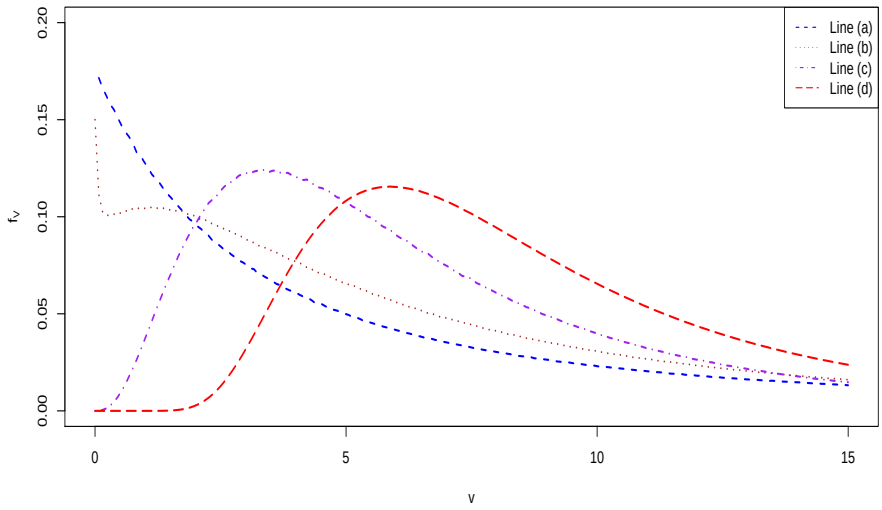

Figure 1: Graphics of the error marginal prior (22) for some choices of the hyperparameters. (*a*): $\zeta_1 = 1,\ \zeta_2 = 1,\ \gamma_1 = 0.002,\ \gamma_2 = 0.01$; (*b*): $\zeta_1 = 3,\ \zeta_2 = 2,\ \gamma_1 = 1,\ \gamma_2 = 4$; (*c*): $\zeta_1 = 4,\ \zeta_2 = 4,\ \gamma_1 = 0.007,\ \gamma_2 = 0.04$; (*d*): $\zeta_1 = 149,\ \zeta_2 = 4,\ \gamma_1 = 1,\ \gamma_2 = 0.2$.

Table 1: 95% credibility intervals and medians of the 99%-percentile of the predictive distribution for different values of (σ, κ).

$(\sigma, \log(\kappa))$	-1	0	1
0.01	(11.35,12.90) 11.91	(11.27,13.34) 11.95	(11.31,50.39) 12.09
0.1	(11.30,13.07) 11.85	(11.28,14.42) 11.88	(11.29,58.87) 12.06
0.2	(11.40,13.16) 12.00	(11.24,22.96) 11.89	(11.47,73.33) 12.25
0.3	(11.30,14.88) 11.95	(11.38,52.79) 12.09	(11.56,85.70) 12.51

$(\sigma, \log(\kappa))$	-5	-4	-3	-2	-1
0.3	(11.20,12.94) 11.78	(11.19,12.86) 11.81	(11.26,12.91) 11.88	(11.26,13.33) 11.89	(11.30,14.88) 11.95
0.4	(11.27,13.23) 11.86	(11.37,13.33) 11.99	(11.38,13.87) 12.00	(11.27,24.47) 11.89	(11.41,33.45) 12.11

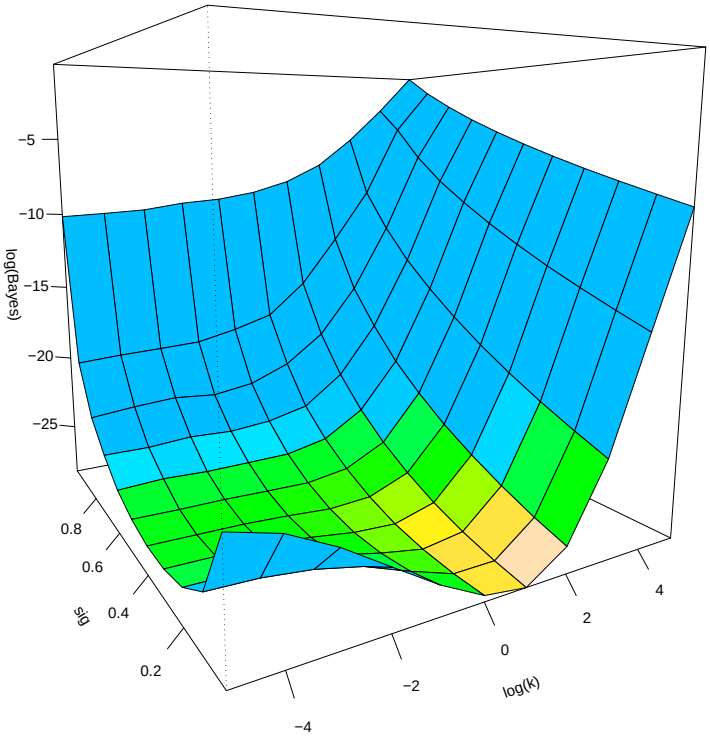

Figure 2: Plot of the log-Bayes factor (17) as a function of $(\sigma, \log(\kappa))$ for the simulated dataset.

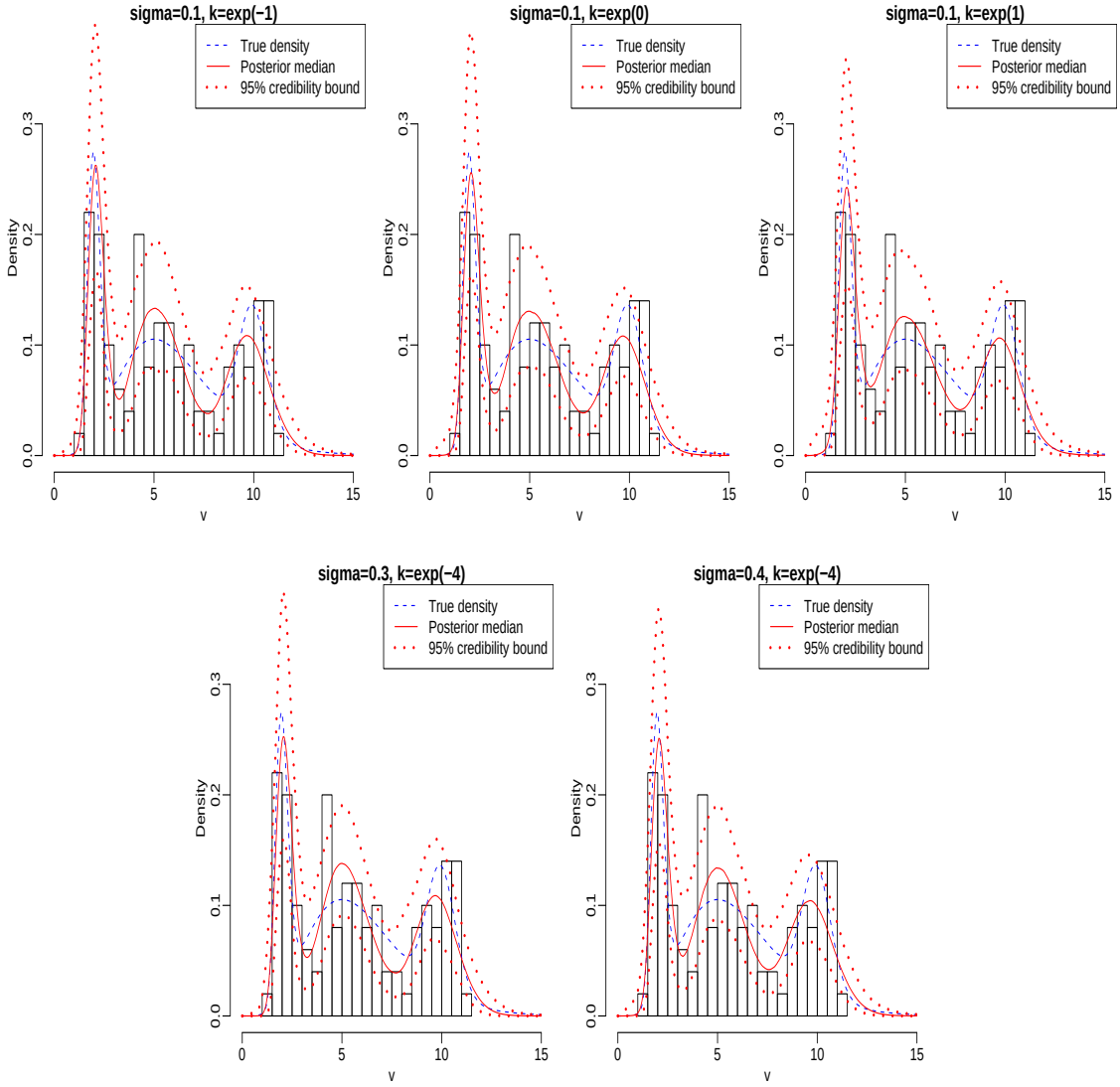

Figure 3: Histograms from the simulated data and density estimates with 95% credibility intervals under the NGG-mixture for some values of (σ, κ).

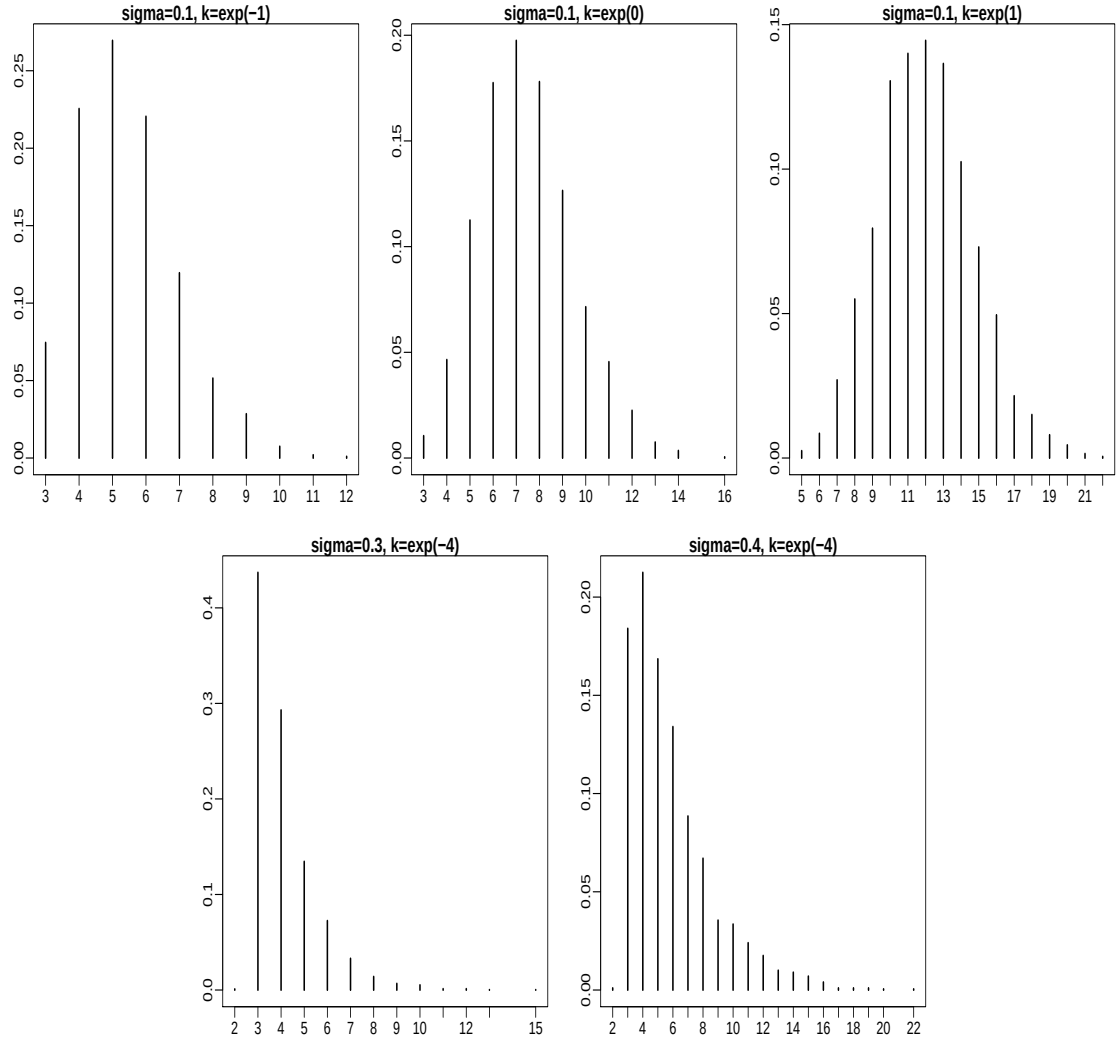

Figure 4: Posterior distributions of the number of clusters under the NGG-mixture. The (σ, κ) parameters are those in Figure 3.

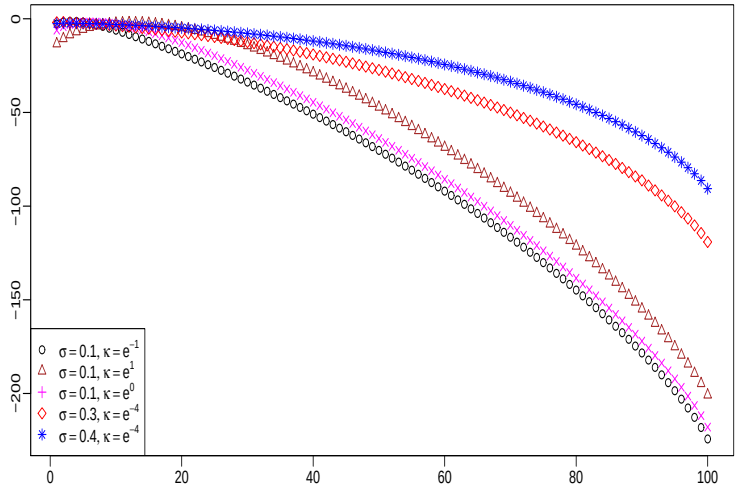

Figure 5: Probability mass functions (in the log scale) of the prior number of clusters when the sample size is 100, for some choices of the hyperparameters (see equation (10)).

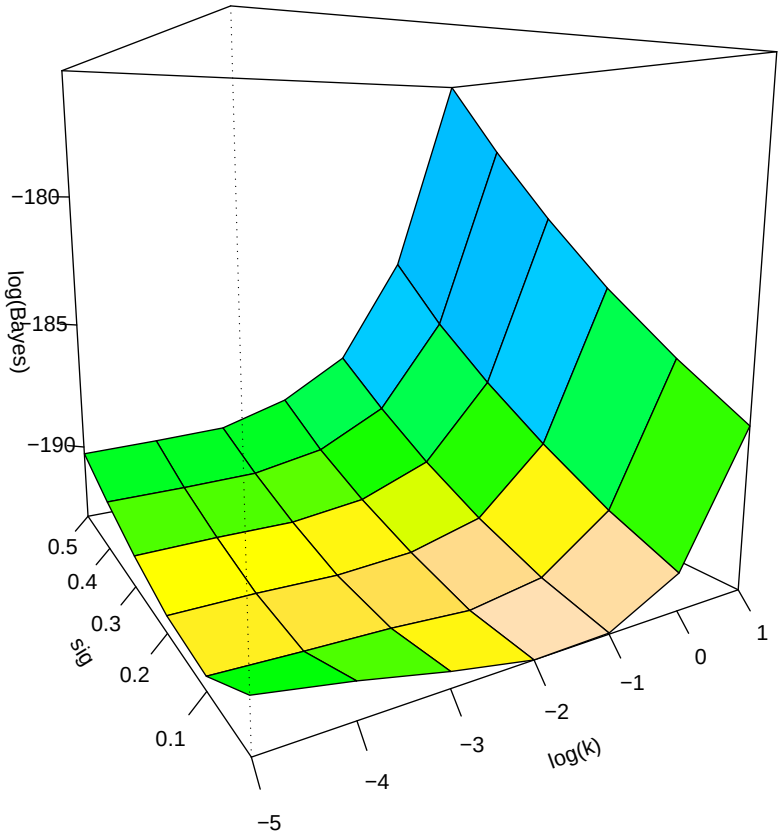

Figure 6: Plot of the log-Bayes factor (17) as a function of $(\sigma, \log(\kappa))$ for the Enzyme dataset.

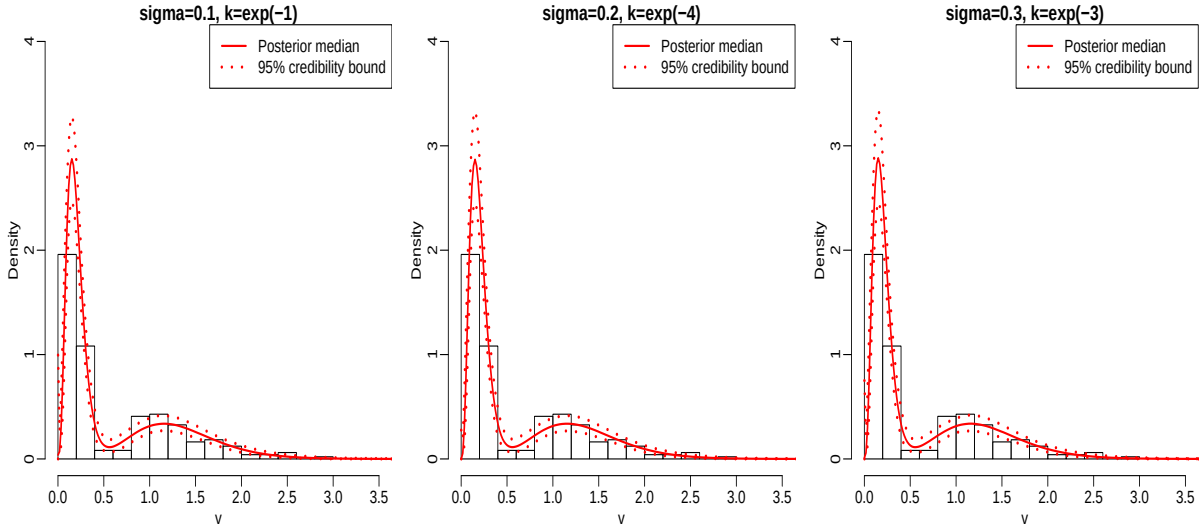

Figure 7: Histograms from the Enzyme data and density estimates with 95% credibility intervals under the NGG-mixture for some values of (σ, κ).

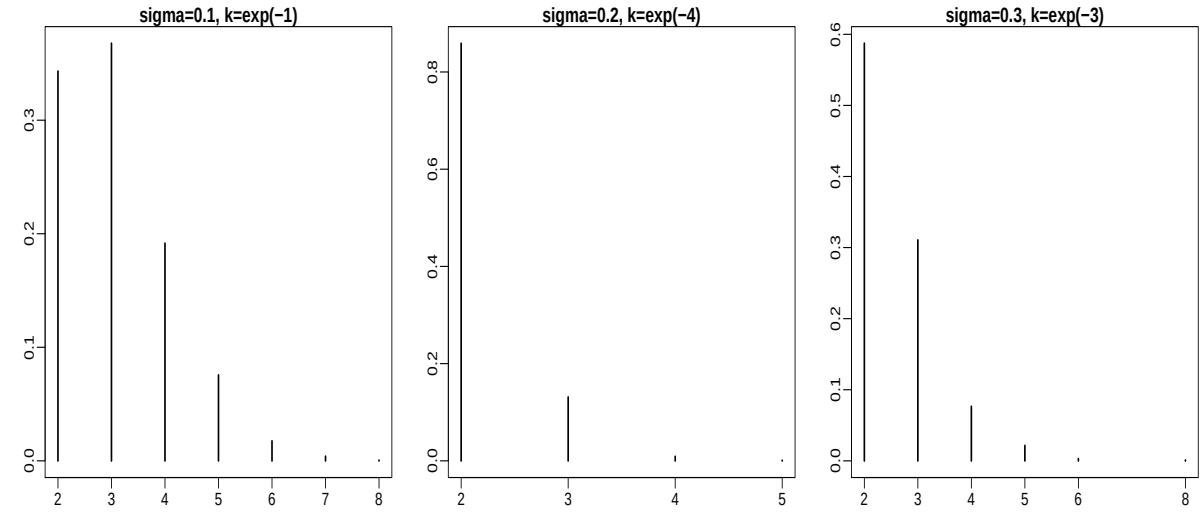

Figure 8: Posterior distributions of the number of clusters under the NGG-mixture. The (σ, κ) parameters are those in Figure 7.

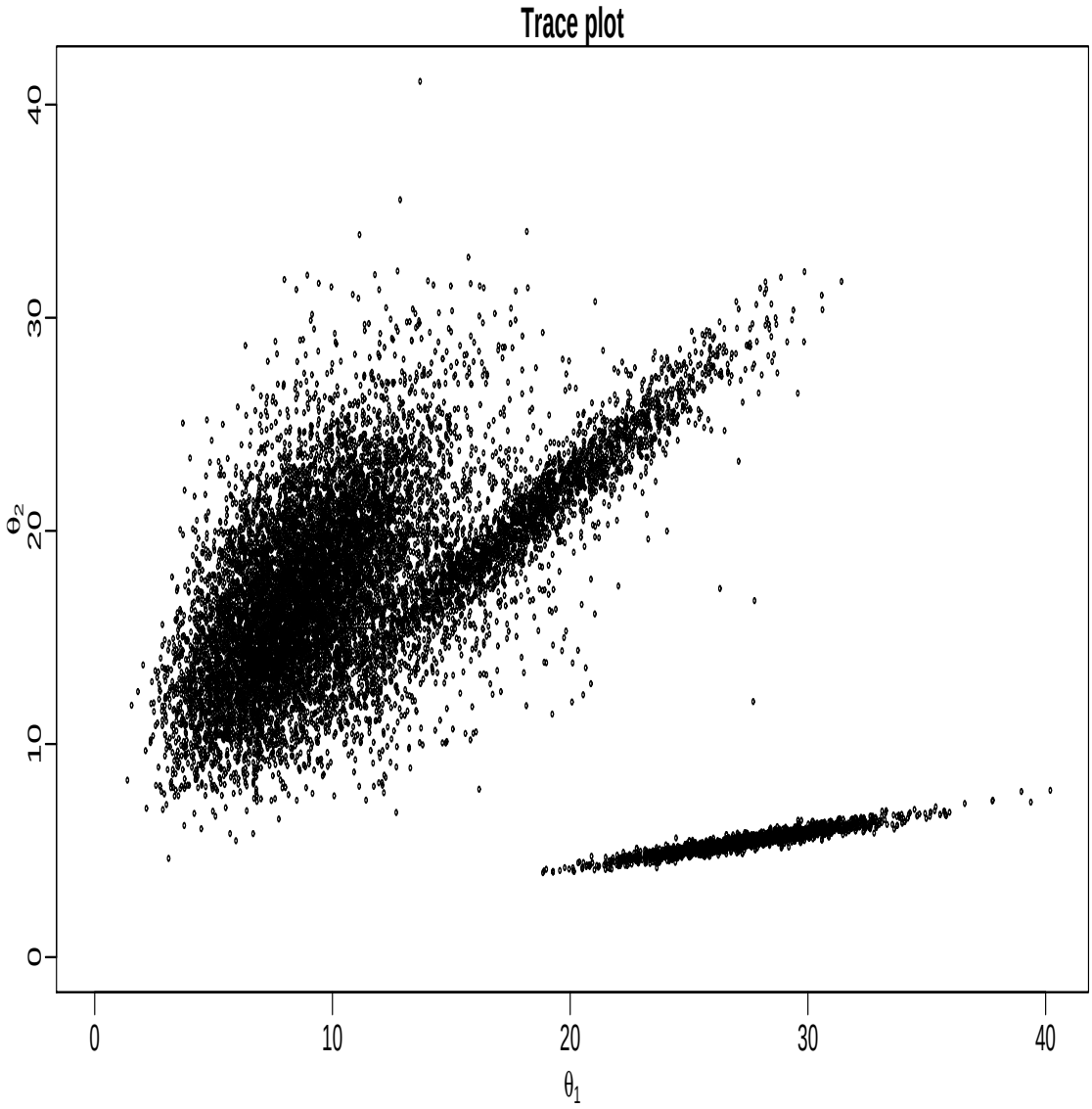

Figure 9: Trace plot of (θ_1, θ_2) for the Enzyme data, when $\zeta_1 = 13.94$, $\zeta_2 = 2.5$, $\gamma_1 = 0.75$, $\gamma_2 = 0.5$, $\sigma = 0.1$ and $\kappa = \exp(1)$.