Chapter 2
A New Finite Approximation for the NGG Mixture Model: An Application to Density Estimation

Ilaria Bianchini

Abstract A new class of random probability measures, approximating the well-known normalized generalized gamma (NGG) process, is defined. The new process is built from the representation of the NGG process as a discrete measure, where the weights are obtained by normalization of points of a Poisson process larger than a threshold ε. Consequently, the new process has an as surely finite number of location points. This process is then considered as the mixing measure in a mixture model for density estimation; we apply it to the popular Galaxy dataset. Moreover, we perform some robustness analysis to investigate the effect of the choice of the hyperparameters.

Key words: Bayesian nonparametric mixture models, A-priori truncation method, Normalized generalized gamma process

2.1 Introduction to Bayesian Nonparametric Mixture Models

In this first section we deal with the problem of density estimation from a Bayesian nonparametric point of view. The nonparametric approach is very useful because it allows a rich class of models for the data, considering infinite dimensional families of probability models. Priors on such families are known as nonparametric Bayesian priors and prevent misleading decisions and inference that may result for a parametric approach, which requires a strong assumption about the investigated

I. Bianchini (✉)
Department of Mathematics, Politecnico di Milano, Milan, Italy

Istituto di Matematica Applicata e Tecnologie Informatiche (IMATI), National Research Council, Milan, Italy
e-mail: ilaria.bianchini@polimi.it

S. Frühwirth-Schnatter et al. (eds.), *Bayesian Statistics from Methods to Models and Applications*, Springer Proceedings in Mathematics & Statistics 126,
DOI 10.1007/978-3-319-16238-6_2

phenomenon, cf. [11]. We will see how a particularly flexible class of nonparametric priors within the family of normalized random measures with independent increments (NRMI) can be applied for density estimation problems.

Mixture models provide a statistical framework for modeling a collection of continuous observations $(X_1,\dots,X_n)$ where each measurement is supposed to arise from one of k possible unknown groups and each group is modeled by a density from a suitable parametric family.

This model is usually represented hierarchically in terms of a collection of independent and identically distributed latent random variables $(\theta_1,\dots,\theta_n)$ as follows:

$$\begin{cases} X_i|\theta_i \overset{ind}{\sim} K(\cdot|\theta_i), & i=1,\dots,n, \\ \theta_i|P \overset{iid}{\sim} P, & i=1,\dots,n, \\ P \sim Q, & \end{cases} \tag{2.1}$$

where Q denotes the nonparametric prior distribution and $K(\cdot|\theta)$ is a probability density function parameterized by the latent random variable θ.

Model (2.1) is equivalent to assume $X_1,\dots,X_n$ i.i.d. according to a probability density that is a mixture of kernel functions:

$$X_1,\dots,X_n \overset{iid}{\sim} f(x) = \int_\Theta K(x|\theta)P(d\theta), \tag{2.2}$$

where P is called mixing measure. Note that if Q selects discrete probability measures, P is almost surely (a.s.) discrete and the mixture model can be written as a sum with a countably infinite number of components:

$$f(x) = \sum_{j=1}^{\infty} p_j K(x|\theta_j),$$

where the weights $(p_j)_{j\geqslant 1}$ represent the relative frequencies of the groups in the population indexed by θ_j. This approach provides a flexible model for clustering items in a hierarchical setting without the necessity to specify in advance the exact number of clusters; therefore, it can also be adopted in cluster analysis. In the next section, the normalized generalized gamma (NGG) prior is introduced, starting from its construction via normalization of a discrete random measure. As we will see, it is very flexible and still mathematically tractable at the same time, making it a suitable choice for Q in the mixture model.

2.2 The NGG Process

Here, we briefly recall how to build a normalized random measure with independent increments (for an in-depth study, see Chapter 8 of [8]). Consider a (a.s.) discrete random measure $\mu(\cdot)$: it can be expressed as an infinite weighted sum of degenerate measures:

$$\mu(\cdot) = \sum_{i \geq 1} J_i \delta_{\tau_i}(\cdot). \tag{2.3}$$

The random elements $(J_i, \tau_i)_{i \geqslant 1}$ are the points of a Poisson process on $(\mathbb{R}^+, \mathbb{X})$ with mean measure ν that satisfies the following conditions:

$$\int_{(0,1)} s\nu(ds, \mathbb{X}) < \infty, \qquad \nu([1,\infty) \times \mathbb{X}) < \infty. \tag{2.4}$$

This construction produces the most general completely random measure (CRM) without fixed atoms and non-random measure parts: it selects discrete measures almost surely.

An important property which one could impose on a CRM is homogeneity, i.e. the underlying mean measure factorizes. Let P_0 be a non-atomic and σ-finite probability measure on $\mathbb{X}$: if $\nu(ds, dx) = \rho(ds)P_0(dx)$, for some measure ρ on $\mathbb{R}^+$, we call μ *homogeneous*: in this case, the jumps in the representation (2.3) are independent from the locations.

The sequence $(J_i)_{i \geq 1}$ represents the jumps controlled by the kernel ρ and $(\tau_i)_{i \geq 1}$ are the locations of the jumps determined by the measure P_0 on $\mathbb{X}$. Since μ is a discrete random measure almost surely, it is straightforward to build a discrete random probability measure by the normalization procedure, which yields NRMIs, first introduced by [14].

Obviously the procedure is well defined only if the total mass of the measure $T := \mu(\mathbb{X})$ is positive and finite almost surely:

$$\mathbb{P}(0 < T < \infty) = 1.$$

This requirement is satisfied if the measure ρ (in the homogeneous case) is such that

$$\int_{\mathbb{R}^+} \rho(ds) = \infty \quad \forall x \in \mathbb{X}. \tag{2.5}$$

This means that the jumps of the process form a dense set in $(0, \infty)$. However, since the second condition in (2.4) must hold, it turns out that infinite points of the Poisson process are very small. In fact, we find that the integral of intensity ρ over $\mathbb{R}^+$ is infinite while the subinterval over $[0, \infty)$ is finite. Now, we define an NRMI $P(\cdot)$ as $\mu(\cdot)/T$. It is important to highlight that NRMIs select, almost surely, discrete distributions, such that P admits a series representation as

$$P = \sum_{j \geq 1} p_j \delta_{\tau_j}, \tag{2.6}$$

where $p_j = J_j/T \ \forall j \geqslant 1$, where the weights J_j are those in (2.3).

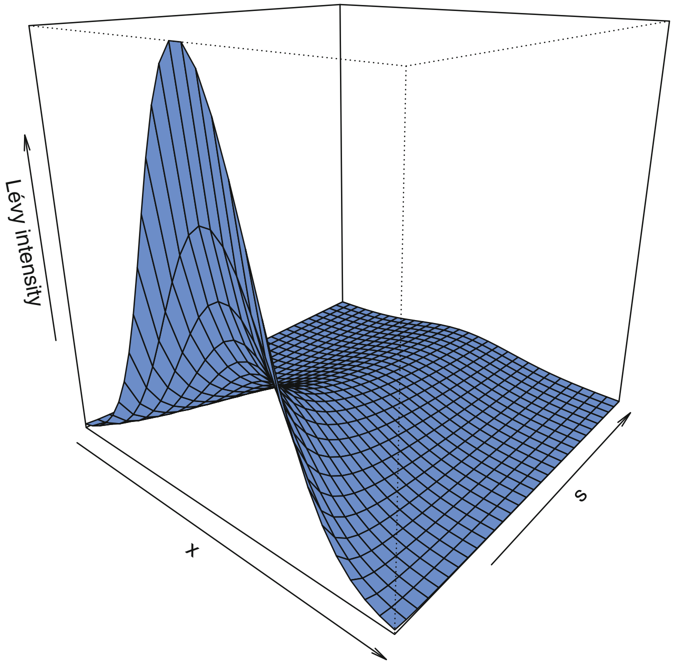

Fig. 2.1 Example of intensity measure $\nu(ds,dx) = 1/\Gamma(1-\sigma)e^{-s}s^{-1-\sigma}dsP_0(dx)$ where $\sigma = 0.1$ and P_0 is Gaussian with mean 0 and variance 1

The NRMI addressed here is the NGG process. As stated in [1], a generalized gamma measure is an NRMI μ with intensity measure equal to

$$\nu(A \times B) = P_0(B)\int_A \rho(ds), \qquad A \in \mathscr{B}(\mathbb{R}^+), B \in \mathscr{B}(\mathbb{X})$$

where

$$\rho(ds) = \frac{\kappa}{\Gamma(1-\sigma)} s^{-1-\sigma} e^{-s\omega} ds, \qquad s > 0. \tag{2.7}$$

Figure 2.1 displays $\nu(s,x)$, where $\omega = \kappa = 1$, $\sigma = 0.1$ and P_0 is Gaussian with mean 0 and variance 1. It is straightforward to define the homogeneous random probability measure $P(\cdot) = \mu(\cdot)/T$ as in (2.6), by the name of NGG process

$$P \sim NGG(\sigma, \kappa, \omega, P_0),$$

with parameters $(\sigma, \kappa, \omega, P_0)$, where $0 \leqslant \sigma \leqslant 1$, $\omega \geq 0$, $\kappa \geq 0$. Within this wide class of priors one finds the following special cases:

1. The Dirichlet process $DP(\kappa, P_0)$ which is an $NGG(0, \kappa, P_0)$ process;
2. The normalized inverse Gaussian process that corresponds to a $NGG(1/2, \kappa, P_0)$.

One could wonder why to choose this process instead of using directly the popular Dirichlet process. The main reason lies in the greater flexibility of the clustering behavior, achieved by the additional parameter, σ, which tunes the variance of the number of distinct observations in a sample from P (if σ increases, the variance increases too; see, for instance, [9]).

2.3 The ε-NGG Approximation

The model we are going to approximate in this section is the so-called NGG mixture model,

$$\begin{cases} X_i|\theta_i \overset{ind}{\sim} K(\cdot|\theta_i), & i = 1, \dots, n, \\ \theta_i|P \overset{iid}{\sim} P, & i = 1, \dots, n, \\ P \sim NGG(\sigma, \kappa, \omega, P_0). \end{cases} \tag{2.8}$$

From now on, we will consider kernels $K(\cdot|\theta)$ defined on $\mathbb{X} \subseteq \mathbb{R}^p$, where p represents the dimension of the data, and the prior NGG is defined on $\Theta \subseteq \mathbb{R}^m$, the space of the parameters of the kernel. For instance, if K is the univariate Gaussian distribution, $N(\mu, \sigma^2)$, the latent variable θ could be the couple (μ, σ^2), hence $\Theta = (\mathbb{R} \times \mathbb{R}^+)$. The main problem when dealing with nonparametric mixture models is the presence of an infinite dimensional parameter P, which makes these models computationally difficult to handle.

In the literature, one can find two ways to tackle this problem, namely marginal and conditional methods: on the one hand, the first ones integrate out the infinite dimensional parameter, leading to generalized Polya urn schemes (see, for instance, [10] and [12]). This approach has one main limitation: We cannot obtain information about the latent variables, since the posterior inference involves only the predictive distribution $f(X_{n+1}|X_1, X_2, \dots, X_n)$. On the other hand, conditional methods build a Gibbs sampler which does not integrate out the nonparametric mixing measure but update it as a part of the algorithm itself. The reference papers on conditional algorithms for Dirichlet process mixtures are the retrospective sampler of [13] and the slice sampler of [15] (extended in the more general NRMI case in [5]). Conditional methods can also be based on truncation of the sum defining the mixing measure P in (2.6): it can be performed both a-posteriori, as in [6] and [1], or a-priori, as in [7] and [4]. The driving motivation for using conditional methods is that they provide a "full Bayesian analysis," i.e. it is possible to estimate either posterior mean functional or linear and nonlinear functionals, such as quantiles.

The proposed method is based on an *a-priori truncation* of P: in particular, we consider only jumps greater than a threshold $\varepsilon > 0$, which turns out to control the approximation to the infinite dimensional prior: conditionally on ε, only a finite number of jumps has to be considered, hence we resorted to a finite dimensional problem. In particular, the number of jumps J_j greater than a threshold value ε is $N_\varepsilon + 1$, where N_ε is a random variable distributed as

$$N_\varepsilon \sim Poisson(\Lambda_\varepsilon), \qquad \Lambda_\varepsilon = \int_\varepsilon^\infty \rho(ds) = \frac{\kappa\omega^\sigma}{\Gamma(1-\sigma)}\Gamma(-\sigma, \omega\varepsilon),$$

so that its expectation increases as ε decreases. Furthermore, the jumps $(J_0, J_1, \ldots, J_{N_\varepsilon})$ turn out to be i.i.d. from

$$\rho_\varepsilon(s) = \frac{\rho(s)}{\Lambda_\varepsilon}\mathbb{1}_{(\varepsilon,\infty)}(s) = \frac{1}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} s^{-\sigma-1} e^{-\omega s}\mathbb{1}_{(\varepsilon,\infty)}(s).$$

We consider location points $(\tau_0, \tau_1, \ldots, \tau_{N_\varepsilon})$ i.i.d. from the base measure P_0 and define the following discrete (a.s.) random probability measure on Θ:

$$P_\varepsilon(\cdot) = \sum_{j=0}^{N_\varepsilon} \frac{J_j}{T_\varepsilon}\delta_{\tau_j}(\cdot) \tag{2.9}$$

where $T_\varepsilon = \sum_{j=0}^{N_\varepsilon} J_j$. P_ε in (2.9) is denoted as ε-NGG(σ,κ, ω, P_0) process. This process can be seen as an approximated version of the NGG process of Sect. 2.2, provided that ε is small, since the convergence to the NGG process holds true provided that ε tends to 0. The main advantage compared to the corresponding NGG is that in this case the sum defining P_ε is finite: We moved from an infinite dimensional process to a finite dimensional one, which eventually (when ε assumes a very small value) approximates the NGG.

The mixture model we are going to consider can be expressed as follows:

$$\begin{cases} X_1, \ldots, X_n | \theta_1, \ldots, \theta_n \sim \prod_{i=1}^n K(X_i|\theta_i), \\ \theta_1, \ldots, \theta_n | P_\varepsilon \sim P_\varepsilon \text{ i.i.d.}, \\ P_\varepsilon \sim \varepsilon\text{-}NGG(\sigma, \kappa, \omega, P_0), \\ \varepsilon, \sigma, \kappa \sim \pi(\varepsilon, \sigma, \kappa). \end{cases}$$

It can be either considered as an approximation of the NGG mixture model (2.8) or as a separate model when ε is random. In the latter case, we let data “drive” the degree of approximation and the model can be significantly different with respect to its nonparametric counterpart, because ε may assume relatively large values.

Before proceeding to the application of Sect. 2.4, it is useful to remember that the Bayesian estimate of the true density is

$$f_{X_{n+1}}(x|X_1, \ldots, X_n) = \int \sum_{j=0}^{N_\varepsilon} \frac{J_j}{T_\varepsilon} K(x|\tau_j) \mathscr{L}(d\varepsilon, d\sigma, d\kappa, dP | X_1, \ldots, X_n)$$

which will be estimated through Monte Carlo methods.

A more detailed description of the ε-NGG mixture model, providing also a proof of convergence and an MCMC algorithm to sample from the posterior distribution of the model, can be found in [2].

2.4 An Application to Density Estimation for the Galaxy Data

In this section, we apply the model proposed in Sect. 2.3 to a very popular dataset in the literature, the Galaxy dataset, exploiting the Gibbs sampler scheme of [2]. These data are observed velocities of $n = 82$ different galaxies, belonging to six well-separated conic sections of space. Specifically, we use Gaussian kernel densities $K(x|\theta) = N(x|\mu, \sigma^2)$. Hence, P_0, the parameter of the nonparametric prior, is a normal inverse-gamma distribution,

$$N\left(\mu|\bar{X}, \frac{\sigma^2}{0.01}\right) IG\left(\sigma^2|2,1\right),$$

where $\bar{X}$ stands for the sample mean, 20.83. This set of hyperparameters, first proposed by [3], is standard in the literature.

We perform a robustness analysis through a lot of experiments which highlight the relationship between the posterior estimates and the prior choice of the parameters. In fact, the choice of a value (or a prior in the random case) for these parameters is the most complicated part of the model, since it strongly influences the posterior inference.

Here, we present some results corresponding to different sets of hyperparameters: we report in Table 2.1 nine combinations of (σ, κ) together with three values for the a-priori expected values for the number of groups K_n, namely $\{3, 5, 20\}$, that we used for our experiments.

Obviously, as mentioned in Sect. 2.2, as σ increases, the variance of K_n increases. In addition, we consider three different priors for ε, in order to study their influence on posterior inference. In what follows, we call (A) the case where the prior is degenerate on a value, i.e. $\varepsilon = 10^{-6}$, (B) where $\varepsilon \sim Unif(0, 0.1)$ and (C)

Table 2.1 Combinations of parameters (σ, κ) chosen for the numerical examples: we selected three different couples for each prior mean number of groups in the data

Index	$\mathbb{E}(K_n)$	σ	κ
1	3	0.001	0.45
2	3	0.1	0.25
3	3	0.2	0.05
4	5	0.001	1.0
5	5	0.2	0.35
6	5	0.3	0.09
7	20	0.2	5.0
8	20	0.4	2.2
9	20	0.6	0.3

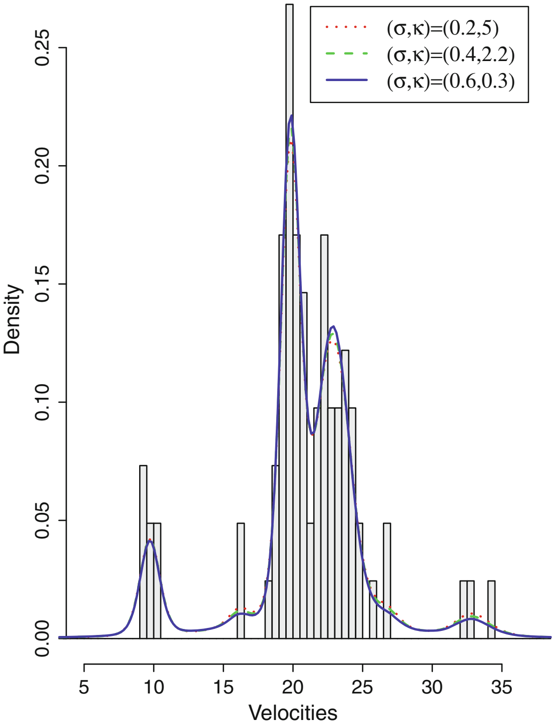

Fig. 2.2 Density estimates for test cases $A7$, $A8$, and $A9$

where ε is a $Beta(0.69, 2.06)$ scaled to the interval $(0, \delta = 0.1)$. In case (C), we chose an informative prior for ε (with mean 0.25δ and variance $0.05\delta^2$) which is concentrated over very small values, since our goal is to approximate the NGG mixture model. Overall, we will have 27 test cases named $A1, \ldots, A9$, $B1, \ldots, B9$, $C1, \ldots, C9$.

Figure 2.2 shows the posterior estimates in test cases $A7$, $A8$, and $A9$, proving reasonable density estimates. We notice that there are only slight differences between the various density estimates, indicating robustness of the model. Figure 2.3 demonstrates that, when σ assumes larger values, the posterior distributions of K_n spread to a larger range of possible values. Since the model is more flexible the posterior mean is free to shift towards the “true” average, being more “sensitive” to the data. This fact is more evident in cases B and C, where ε is random: the posterior mode of the number of clusters is around 10, while in case A is around 16. Here, the data determine the degree of approximation such that unreasonable a-priori information impacts the resulting number of groups less.

Furthermore, we fix $\sigma = 0.1$ and $\kappa = 0.45$ but we consider $\varepsilon \sim Gamma(\alpha, \beta)$, with support over all positive real numbers; in particular, we choose $(\alpha, \beta) \in \{(0.5, 2), (0.01, 0.1), (1, 10)\}$. The first combination corresponds to a relatively large mean (0.25) and variance (0.125) for ε. However, a large mass of the distribution lies around 0 due to the presence of an asymptote in the prior distribution. The second and third combinations have the same mean (0.1) but the variance is 1 and

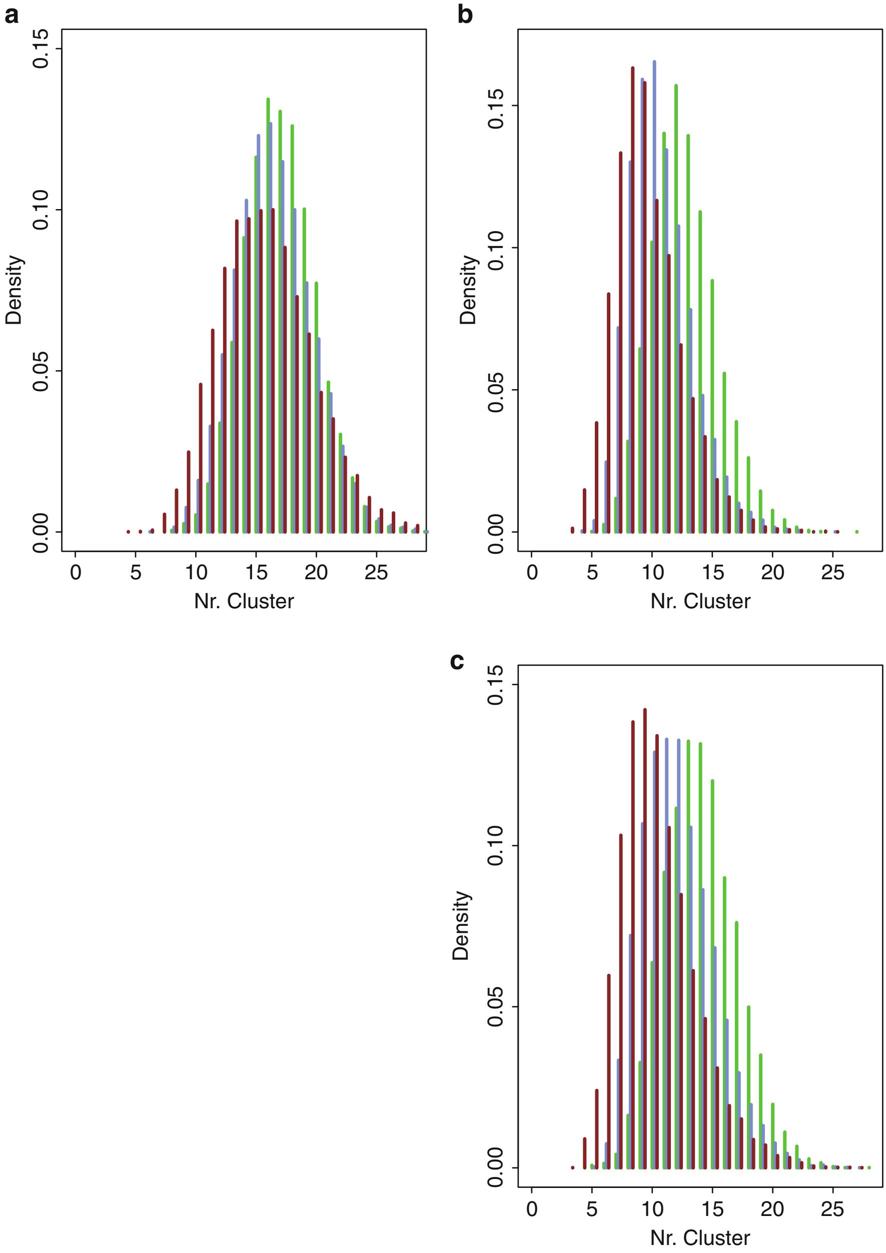

Fig. 2.3 Histograms of the posterior number of clusters in tests A, B, C. In *magenta* the tests with a bigger a-priori variance for K_n, in *green* the tests corresponding to a relatively small variance a-priori, in *blue* the intermediate ones. (**a**) $A7$, $A8$, $A9$. (**b**) $B7$, $B8$, $B9$. (**c**) $C7$, $C8$, $C9$

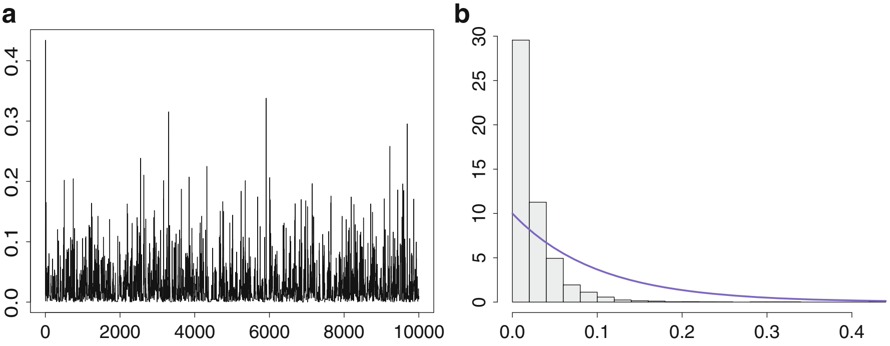

Fig. 2.4 (**a**) Traceplots and histogram for variable ε in the second test case. In panel (**b**) the *violet line* shows the prior distribution, i.e. $\varepsilon \sim gamma(1,10)$

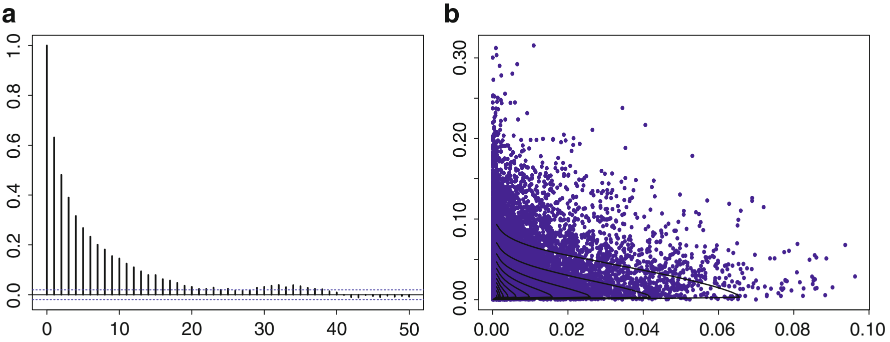

Fig. 2.5 Autocorrelation of variable ε, (**a**), and scatterplot of ε versus σ, (**b**): the *gray lines* represent the contour levels of the prior

0.01, respectively. We report for brevity only results for $(\alpha,\beta) = (1,10)$; however, we point out that some mixing problems in the chain for ε arise, when increasing the a-priori variance. Figure 2.4b shows that ε moves a posteriori towards smaller values with respect to the prior information. Besides, the traceplot of ε, Fig. 2.4a, exhibits a good mixing for the chain in this case.

Finally, we mention a further test, where all three parameters are random: in particular, we assume $\varepsilon \sim Beta(0.69, 2.06)$ with support on $(0,0.1)$, $\sigma \sim Beta(1.1,30)$ and $\kappa \sim Gamma(1.1,8)$. The density estimate is satisfying, the only issue to mention is the high autocorrelation of ε and the correlation between the two parameters σ and ε (Fig. 2.5). This result is even more pronounced under a less informative prior distribution for (σ,ε).

2.5 Conclusions

A method to deal with a particularly flexible nonparametric mixture model, namely the NGG mixture model, is presented. It is based on a-priori truncation of the infinite sum defining the random probability measure P and it allows to computationally handle the presence of an infinite dimensional parameter, P, in the mixture model. In fact, conditionally on a threshold value ε, we can define a new process P_ε, which consists of a finite sum. We showed an application to density estimation for the popular Galaxy dataset. Through the exposition of several choices of the hyperparameters we established the robustness of the model and studied the relationship between posterior estimates and prior elicitation. In particular, we illustrated some suitable priors for the threshold parameter ε, letting in this case, the data drive the degree of approximation. If there is no need to consider a fully nonparametric model, ε may be relatively far from 0, implying smaller computational effort. Overall, density estimates were satisfying in all the experiments.

References

[1] Argiento, R., Guglielmi, A., Pievatolo, A.: Bayesian density estimation and model selection using nonparametric hierarchical mixtures. Comput. Stat. Data Anal. **54**(4), 816–832 (2010)
[2] Argiento, R., Bianchini, I., Guglielmi, A.: A blocked Gibbs sampler for NGG-mixture models via a-priori truncation. Stat. Comput. Advance online publication. doi: 10.1007/s11222-015-9549-6 (2015)
[3] Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**(430), 577–588 (1995)
[4] Griffin, J.E.: An adaptive truncation method for inference in bayesian nonparametric models. arXiv preprint arXiv:1308.2045 (2013). doi:10.1007/s11222-014-9519-4
[5] Griffin, J.E., Walker, S.G.: Posterior simulation of normalized random measure mixtures. J. Comput. Graph. Stat. **20**(1), 241–259 (2011)
[6] Gelfand, A.E., Kottas, A.: A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. J. Comput. Graph. Stat. **11**(2), 289–305 (2002)
[7] Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**(453), 161–173 (2001)
[8] Kingman, J.F.C.: Poisson Processes. Oxford Studies in Probability, vol. 3. The Clarendon Press/Oxford University Press, New York (1993). Oxford Science Publications
[9] Lijoi, A., Mena, R.H., Prünster, I.: Controlling the reinforcement in Bayesian non-parametric mixture models. J. R. Stat. Soc. Ser. B Stat. Methodol. **69**(4), 715–740 (2007)
[10] MacEachern, S.N.: Computational methods for mixture of Dirichlet process models. In: Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statistics, vol. 133, pp. 23–43. Springer, New York (1998)
[11] Müller, P., Mitra, R.: Bayesian nonparametric inference—why and how. Bayesian Anal. **8**(2), 269–302 (2013)
[12] Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**(2), 249–265 (2000)

[13] Papaspiliopoulos, O., Roberts, G.O.: Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika **95**(1), 169–186 (2008)

[14] Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. Ann. Stat. **31**(2), 560–585 (2003). Dedicated to the memory of Herbert E. Robbins

[15] Walker, S.G.: Sampling the Dirichlet mixture model with slices. Commun. Stat. Simul. Comput. **36**(1–3), 45–54 (2007)