

A Bayesian nonparametric model for density and cluster estimation: the ε -NGG process mixture

Un modello bayesiano nonparametrico per la stima di densità e l'analisi di cluster: il modello mistura attraverso il processo ε -NGG

Raffaele Argiento, Ilaria Bianchini and Alessandra Guglielmi

Abstract We define a new class of random probability measures, approximating the well-known normalized generalized gamma (NGG) process. Our new process is defined from the representation of NGG processes as discrete measures where the weights are obtained by normalization of the jumps of a Poisson process, and the support consists of iid points, however considering only jumps larger than a threshold ε . Therefore, the number of jumps of this new process, called ε -NGG process, is a.s. finite. A prior distribution for ε can be elicited. We will assume the ε -NGG process as the mixing measure in a mixture model for density and cluster estimation. Moreover, a efficient Gibbs sampler scheme to simulate from the posterior is provided. Finally, the performance of our algorithm on the Galaxy dataset will be illustrated.

Abstract Introduciamo una nuova classe di misure di probabilità aleatorie che approssimano il noto processo gamma generalizzata normalizzato (NGG). Il nuovo processo qui definito è costruito a partire dalla rappresentazione del processo NGG come misura discreta, in cui i pesi sono ottenuti dalla normalizzazione dei punti di un processo di Poisson, e il cui supporto è costituito da punti iid, ma considerando nella somma che definisce il nuovo processo solo salti più grandi di una certa soglia ε . Pertanto, questo processo è una misura di probabilità aleatoria discreta con un numero finito (q.c.) di punti di supporto. Alla soglia ε può essere assegnata una prior. Considereremo tale nuovo processo come la misura integrale in un modello mistura per la stima di densità e l'analisi di cluster. Uno dei principali risultati di questo lavoro è la costruzione di un algoritmo Gibbs sampler per simulare dalla posterior del modello considerato. Infine, illustreremo il nostro algoritmo utilizzando il dataset Galaxy.

Raffaele Argiento
CNR-IMATI, via Bassini 15, 20133 Milano (ITALY), e-mail: raffaele@mi.imati.cnr.it

Ilaria Bianchini, Alessandra Guglielmi
Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano (ITALY), e-mail: alessandra.guglielmi@polimi.it

Key words: Bayesian nonparametric mixture models, normalized generalized gamma process, Dirichlet process mixture model, Gibbs sampler, finite dimensional approximation

1 Introduction

The first aim of this work is the definition of a new class of nonparametric priors, which can be considered as an approximation of the distribution of a homogeneous normalized random measure, namely the normalized generalized gamma process. Any homogeneous normalized random measure can be represented as a discrete random probability measure: the weights are obtained by normalization of the jumps (a countable set) of a Poisson process, while the support consists of a countable number of random points from some distribution. In this case, posterior inference is made difficult by the presence of infinite unknown parameters. There are two main approaches to deal with this problem, namely marginal and truncated algorithms. The former integrate out the infinite dimensional parameter (i.e. the random probability), resorting to generalized Polya urn schemes (see[9] for a review on the subject). Dating back to [8], who built their *blocked Gibbs sampler*, the latter approach includes fixed ([2]), random ([12]) or adaptive ([6]) stopping rules.

The solution we propose here can be classified as a truncation algorithm, since we are going to consider, in the infinite sum representing the discrete random probability measure, only jumps larger than a threshold ε , which turns out to control the approximation to the infinite-dimensional parameter prior. In this way, we define a new “nonparametric” prior, called ε -NGG process prior; observe that the number of jumps of this process is a.s. finite. A prior distribution for ε can be given, as well as for all the other potential parameters defining the new process.

As often done in Bayesian Nonparametrics, we will consider this new discrete random probability measure as the mixing measure in a mixture model, which is a very flexible tool in density and cluster estimation problems. Among the main achievements of this work, there is the construction of a Gibbs sampler scheme to simulate from the posterior; in particular, we have built a conditional algorithm that uses a finite random number of jumps, but, on the other hand, it is easy to implement.

For illustration purposes, we apply this model to a popular dataset, the galaxy data, since it is nowadays the favorite test dataset for any new nonparametric model in a density estimation context. We illustrate the performances of our algorithm for different sets of hyperparameters, in order to understand in particular the effect of the “approximation” parameter ε on posterior estimates.

2 ε -NGG processes

Any homogeneous normalized random measure with independent increments (see [11]) on Θ ($\subset \mathbb{R}^m$) can be represented as a discrete probability where the weights are obtained by normalization of the jumps of a Poisson process, and the support is a denumerable set of iid random points from a distribution P_0 on Θ :

$$P = \sum_{j=1}^{\infty} P_j \delta_{\tau_j} = \sum_{j=1}^{\infty} \frac{J_j}{T} \delta_{\tau_j}. \quad (1)$$

Here $(J_j)_j$ are the points of a Poisson process on $\mathbb{R}^+$ with mean intensity $\rho(ds)$, such that $0 < T := \sum_{j=1}^{\infty} J_j < +\infty$ a.s., and $\{\tau_j\}$ and $\{J_j\}$ are independent. See [3].

Of course, when using this process as an ingredient in Bayesian nonparametric (or semiparametric) models, posterior inference is difficult because of infinite unknown parameters. Our proposal is to define a random probability measure (r.p.m.) as in (1), but restricting the sum to jumps larger than a threshold $\varepsilon > 0$; it will turn out that only a finite number of jumps has to be considered.

Let us fix $\varepsilon > 0$ and let $J_0, J_1, \dots, J_{N_\varepsilon}$ iid random variables from the density ρ_ε , where $N_\varepsilon \sim \text{Poisson}(\Lambda_\varepsilon)$,

$$\Lambda_\varepsilon := \frac{\kappa \omega^\sigma}{\Gamma(1-\sigma)} \Gamma(-\sigma, \omega \varepsilon),$$

and

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

Here $\kappa > 0$, $\omega > 0$ and $0 < \sigma < 1$, and $\Gamma(a, x) = \int_x^{+\infty} t^{a-1} e^{-t} dt$ is the incomplete gamma function. Observe that N_ε is a.s. finite. Then we define

$$P_\varepsilon = \sum_{j=0}^{N_\varepsilon} P_j \delta_{\tau_j} = \frac{1}{T_\varepsilon} \sum_{j=0}^{N_\varepsilon} J_j \delta_{\tau_j}, \quad (2)$$

where $T_\varepsilon = \sum_{j=0}^{N_\varepsilon} J_j$, $\tau_j \stackrel{\text{i.i.d.}}{\sim} P_0$, $\{\tau_j\}$ and $\{J_j\}$ independent. We denote P_ε in (2) by ε -NGG($\sigma, \kappa, \omega, P_0$) process.

Observe that P_ε is a proper species sampling model ([10]) with a random number $N_\varepsilon + 1$ of different species. The main “theoretical” result on P_ε is the following:

Proposition 1. *The law of P_ε converges weakly to the law of P as ε goes to 0, where P is a NGG($\sigma, \kappa, \omega, P_0$) process.*

By a NGG($\sigma, \kappa, \omega, P_0$) process P we mean a normalized generalized gamma process (see [7]), that is

$$P = \sum_{j=1}^{+\infty} \frac{J_j}{\sum_j J_j} \delta_{\tau_j}, \quad (3)$$

where $\{\tau_j\}$ and $\{J_j\}$ independent, $\tau_j \stackrel{\text{i.i.d.}}{\sim} P_0$ and $\{J_j\}$ are the points of a Poisson process on $\mathbb{R}^+$ with mean intensity $\rho(ds) = \kappa/\Gamma(1-\sigma)s^{-1-\sigma}e^{-s\omega}ds, s > 0$. When $\sigma = 0$ the Dirichlet process is recovered.

The proof of the above proposition is in [1].

It is straightforward to check that definition (2) is equivalent to consider, in (3), only jumps larger than the threshold ε . This is true because definition (2) is given in terms of a Poisson process with intensity measure ρ_ε , which turns out to be finite (i.e. it is a density), and obtained from $\rho(ds)$ restricted to the interval $(\varepsilon, +\infty)$. Moreover, the definition is well-posed since, to avoid degenerate cases (i.e. when $N_\varepsilon = 0$), we add the extra point J_0 .

3 ε -NGG process mixtures

Often, in Bayesian nonparametric problems, it happens that discrete random probabilities as our ε -NGG process, appear as mixing measures in a mixture context. In fact, we are going to consider the following model:

$$\begin{aligned} X_i | \theta_i &\stackrel{\text{ind.}}{\sim} k(\cdot; \theta_i), \quad i = 1, \dots, n \\ \theta_1, \dots, \theta_n | P_\varepsilon &\stackrel{\text{i.i.d.}}{\sim} P_\varepsilon \\ P_\varepsilon &\sim \varepsilon - \text{NGG}(\sigma, \kappa, \omega, P_0) \text{ process prior,} \end{aligned} \quad (4)$$

where $k(\cdot; \theta_i)$ is a parametric family of densities on $\mathbb{X} \subset \mathbb{R}^p$, for all $\theta \in \Theta \subset \mathbb{R}^m$. Remember that P_0 is a non-atomic probability measure on Θ , expressing the “mean” of P . Model (4) will be addressed here as ε -NGG *hierarchical mixture* model. It is well known that this model is equivalent to data X_i which, conditionally on P_ε , are independently distributed according to the random density

$$f(x) = \int_{\Theta} k(x; \theta) P_\varepsilon(d\theta) = \sum_{j=0}^{N_\varepsilon} P_j k(x; \tau_j). \quad (5)$$

The Bayesian model specification may be completed assuming that $\varepsilon, \sigma, \kappa$, as well as potential hyperparameters of P_0 , are random and distributed according to some prior specification.

In general, computation of posterior inference for (4) when P_ε is substituted by a NGG process P is not straightforward, since this model assumes an infinite number of parameters. Indeed, we have proposed such ε -NGG *hierarchical mixture* to overcome this issue: we were motivated to introduce the ε -approximation since we were just looking for a straightforward and simple Gibbs scheme for sampling from the posterior of a NGG mixture model.

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ be a sample from P_ε and set $U := \Gamma_n/T_\varepsilon$, where $\Gamma_n \sim \text{gamma}(n, 1)$. The following proposition illustrates a “finite dimensional” version of a characterization of the posterior law of a NGG process, given in [4]. As for

the infinite dimensional case, the posterior distribution of an ε -NGG(σ, κ, ω) process, conditionally on U and θ , can be expressed as the law of a random probability measure, which is a mixture between a ε -NGG process and a discrete probability measure with support given by the (observed) distinct values $\theta^* = (\theta_1^*, \dots, \theta_k^*)$.

Proposition 2. *If P_ε is an ε -NGG($\sigma, \kappa, \omega, P_0$), then the posterior distribution of P_ε coincides with the law of the random probability measure*

$$P_\varepsilon^*(\cdot) = wP_{\varepsilon,u}(\cdot) + (1-w) \sum_{j=1}^k P_j^{(a)} \delta_{\theta_j^*}(\cdot) \quad (6)$$

where

- (a) $P_{\varepsilon,u}(\cdot)$ is distributed according to an ε -NGG($\sigma, \kappa, \omega + u, P_0$) process;
- (b) the jumps $\{P_1^{(a)}, \dots, P_k^{(a)}\}$ assigned to the fixed points of discontinuity $\theta^* = (\theta_1^*, \dots, \theta_k^*)$ of P_ε^* are obtained by normalization of $J_j^{(a)} \stackrel{\text{ind.}}{\sim} \text{gamma}(n_j - \sigma, u + \omega)$, for $j = 1, \dots, k$;
- (c) $P_{\varepsilon,u}(\cdot)$ and $\{J_1^{(a)}, \dots, J_k^{(a)}\}$ are independent.

Moreover $w = T_{\varepsilon,u} / (T_{\varepsilon,u} + \sum_{j=1}^k J_j^{(a)})$ where $T_{\varepsilon,u}$ is the normalization variable in the representation of $P_{\varepsilon,u}(\cdot)$ as in (2).

The proof of the above proposition is given in [1].

To simulate from the posterior of (4), when $\varepsilon, \sigma, \kappa$ are a priori independent according to some distributions, we have proposed a Gibbs sampler algorithm, which includes the latent variables θ and the auxiliary variable U (of Proposition 2) in the state space. The algorithm sequentially draws from the following full conditionals: $\mathcal{L}(d\varepsilon, d\sigma, d\kappa, dP_\varepsilon | \theta, u, X_1, \dots, X_n)$, $\mathcal{L}(d\theta | P_\varepsilon, u, X_1, \dots, X_n)$, $\mathcal{L}(du | P_\varepsilon, \theta, X_1, \dots, X_n)$. Further details are illustrated in [1].

4 Galaxy data

This super-popular dataset contains $n = 82$ measured velocities of different galaxies from six well-separated conic sections of space. Values are expressed in Km/s, scaled by a factor of 10^{-3} . We report some of the posterior estimates (mainly density and number of clusters estimates) for different sets of hyperparameters of model (4) when $k(\cdot; \theta)$ is the Gaussian density on $\mathbb{R}$ and $\theta = (\mu, \sigma^2)$ stands for its mean and variance, and $P_0(d\mu, d\sigma^2) = \mathcal{N}(d\mu; m_0, \sigma^2/\kappa_0) \times \text{inv-gamma}(d\sigma^2; a, b)$; here $\mathcal{N}(m_0, \sigma^2/\kappa_0)$ is the Gaussian distribution with m_0 mean and σ^2/κ_0 variance, while $\text{inv-gamma}(a, b)$ is the inverse-gamma distribution with mean $b/(a-1)$ (if $a > 1$). We have fixed $m_0 = \bar{x}_n = 20.8315$, $\kappa_0 = 0.01$, $a = 2$, $b = 1$ as proposed first in [5]. We report posterior inference for three sets of hyperparameters, to understand sensitivity of the estimates when ε varies, but it is not random, then when

σ varies (but it is not random), and finally when ε is assumed random (and σ and κ are fixed).

With reference to the first set of hyperparameters, we have fixed $\sigma = 0.4$ and $\kappa = 0.45$, and $\varepsilon = 10^{-6}, 10^{-3}, 10^{-1}, 1$. Figure 1 shows the density estimates (posterior

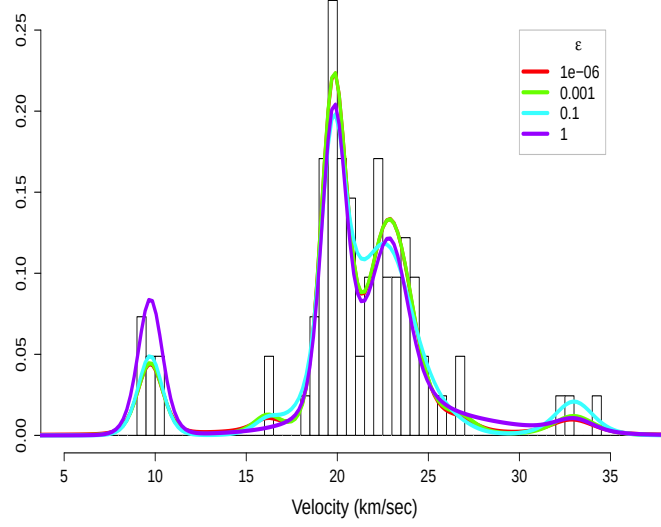


Fig. 1 Density estimates for different values of ε , while $\sigma = 0.4$ and $\kappa = 0.45$.

mean of the random density “parameter”) under the different values of ε : all the estimates are quite similar and detect the “right” number of clusters.

Observe that, when ε increases, more jumps J_j are cut off from the sum defining the process P_ε (see (2)) and, consequently, less components in the mixture (5) are considered. Therefore the posterior estimate of the number of groups k will be concentrated on smaller integer values as ε increases (see Figure 2). It is worth underlining that, as another consequence of the smaller number of components in the mixture (5) when ε increases, we have observed a huge gain in run-time: for instance, with our machine, the run-time ranges from approximately 7 minutes ($\varepsilon = 10^{-6}$) to less than 1 minute ($\varepsilon = 1$).

The second set of hyperparameters is specified by $\varepsilon = 10^{-6}$ and $\kappa = 0.45$, while σ ranges in $\{0.001, 0.1, 0.2, \dots, 0.8\}$. The posterior density estimates are similar to those obtained before, and for this reason they are not reported here. On the other hand, we are interested to understand the effect of σ on the posterior number of clusters, as in Table 1. First of all, note that we are also including the Dirichlet process mixture model here, since, when $\sigma = 0.001$ (i.e. $\sigma \simeq 0$), and ε is small, this

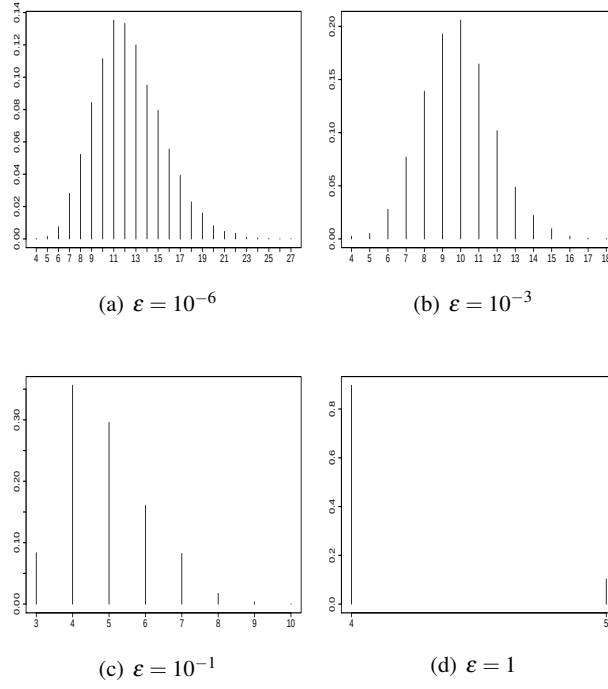


Fig. 2 Posterior distribution of the number of clusters k under the first set of hyperparameters.

Table 1 Posterior (and prior) summaries of the number of clusters k under the second set of hyperparameters

σ	Prior mean	Posterior mean	Posterior variance
0.001	3	6.13	1.73
0.1	4.06	7.18	2.39
0.2	5.6	8.74	4.25
0.3	7.8	10.49	6.39
0.4	10.9	12.36	9.30
0.5	15.3	14.06	11.49
0.6	21.5	15.90	14.61
0.7	30.2	17.67	17.66
0.8	42.3	19.05	20.16

well-known mixture is recovered. As expected, the posterior mean of k , as well as its variance, increases with σ .

We also point out that increasing σ has a positive effect on the MCMC chains, for example on the posterior of U : see Figure 3 where there is a clear improvement in the autocorrelation plot of the chain from the left ($\sigma = 0.001$) to the right ($\sigma = 0.8$).

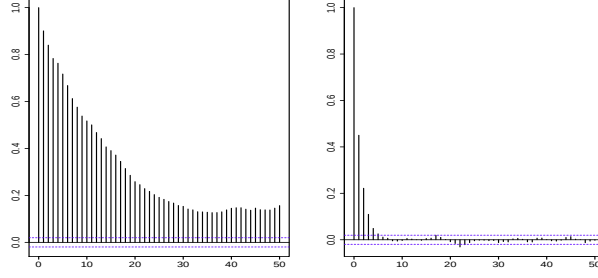


Fig. 3 Autocorrelation of the auxiliary variable U under the second set of hyperparameters, when $\sigma = 0.001$ (left) and $\sigma = 0.8$ (right).

For the last experiment, we have considered ε random, uniformly distributed on the interval $(0, \delta)$, with $\delta = \min(0.1, E(T_\varepsilon))$ and $\sigma \in \{0.001, 0.1, 0.2, \dots, 0.9\}$, while $\kappa = 0.45$. When ε is random, the model is expected to be more flexible, since it would “adjust” for the number of jumps of the process P_ε that must be considered. Furthermore, on one hand, if ε increases, the process will be significantly different from the NGG process, since, in this case, many small jumps will not be included in (2). On the other hand, when ε is large, the variable N_ε , which counts the number of weights in the mixture, will be generally smaller than when ε is fixed. As in the previous cases, density estimates are pretty good (see Figure 4); we also mention (but do not report results) that, as expected, the model is more parsimonious as far as the estimated number of groups k is concerned. The algorithm is faster in this case too, since it allows to sample less jumps and less points of support as mentioned before: indeed, this sampling is the most complex part of the algorithm.

As far as robustness with respect to σ is concerned, we should acknowledge that, as σ increases, more computational problems come up, because of the incomplete gamma function, appearing in ρ_ε , that is harder to be numerically evaluated.

Looking at the posterior distribution of ε in Figure 5, data suggest that small values of ε are the “best” fit, when the prior of ε is uniform. In particular, increasing σ , and consequently increasing the prior expected number of clusters, we get that the posterior of ε is concentrated on smaller values, which implies a larger posterior number of clusters. Finally, we also point out that there is a strong correlation between the variables U and ε , especially when σ is small: see the scatterplots in Figure 6.

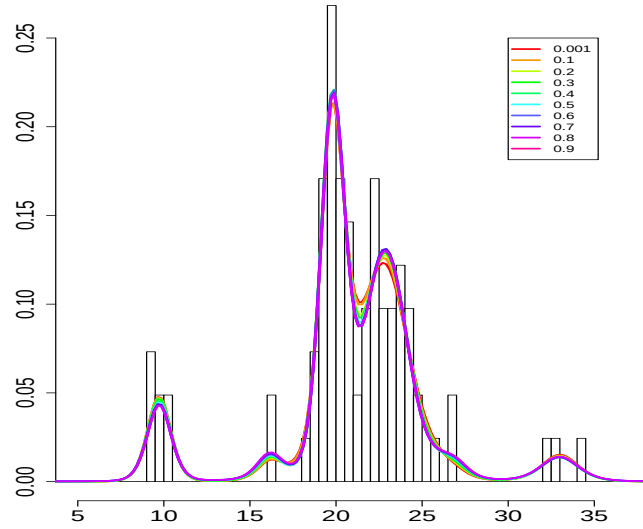


Fig. 4 Density estimation under the third set of hyperparameters with different values of σ .

References

1. Argiento, R., Bianchini, I., Guglielmi, A.: A Bayesian nonparametric model for density estimation: the ε -NGG process, Unpublished Technical Report (2014)
2. Argiento, R., Guglielmi, A., Pievatoto, A.: Bayesian density estimation and model selection using nonparametric hierarchical mixtures. *Comput. Stat. Data An.*, **54**, 816–832 (2010).
3. James, L.F.: A simple proof of the almost sure discreteness of a class of random measures. *Statist. Probab. Lett.*, **65**, 363–368 (2003)
4. James, L., Lijoi, A., Prünster, I.: Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.*, **36**, 76–97 (2009)
5. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, **90**, 577–588 (1995)
6. Griffin, J. E.: An adaptive truncation method for inference in Bayesian nonparametric models. *arXiv:1308.2045 [stat.CO]*
7. Lijoi, A., Mena, R. H., Prünster, I.: Controlling the reinforcement in Bayesian nonparametric mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69**, 715–740 (2007)
8. Ishwaran, H., and James, L. F.: Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161–173 (2001).
9. Neal, R. M.: Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Stat.*, **9**, 249–265 (2000):
10. Pitman, J.: Some developments of the Blackwell-MacQueen urn scheme. In Ferguson, T. S., Shapley, L. S., MacQueen, J. B. (Eds.) *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, volume 30 of *IMS Lecture Notes-Monograph Series*, pp. 245–267. Institute of Mathematical Statistics, Hayward, USA (1996)
11. Regazzini, E., Lijoi, A., Prünster, I.: Distributional Results for Means of Normalized Random Measures with Independent Increments. *Ann. Statist.*, **31**, 560–585 (2003)

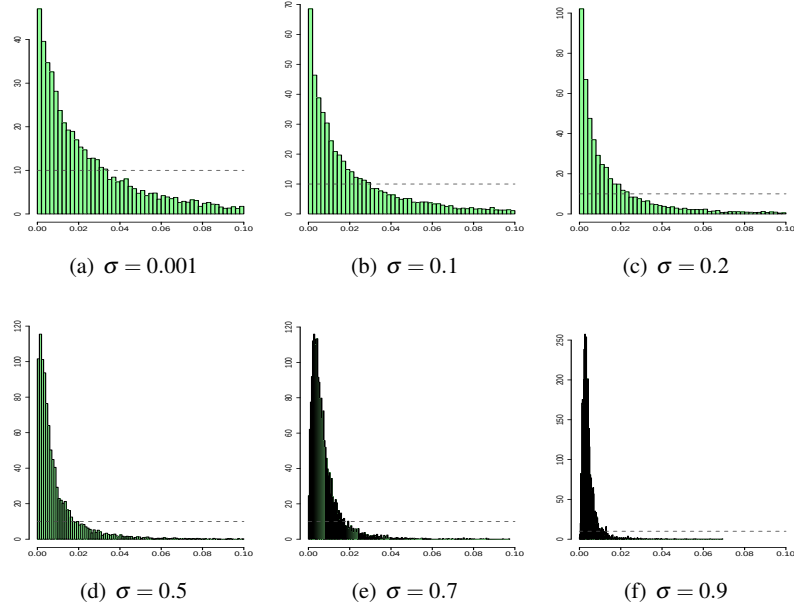


Fig. 5 Posterior distribution of ε under the third set of hyperparameters, together with the prior (dashed) $Unif(0, \delta)$.

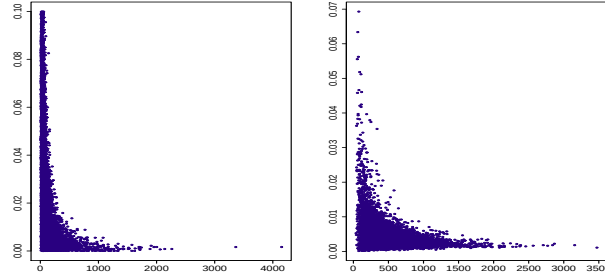


Fig. 6 Variable U versus ε for the third set of hyperparameters, when $\sigma = 0.001$ (left) and $\sigma = 0.9$ (right).

12. Walker, S. G.: Sampling the Dirichlet mixture model with slices. *Commun. Stat. Simulat.*, **36**, 45–54 (2007)