A New Finite Approximation for the NGG Mixture Model: an Application to Density Estimation

Raffaele Argiento[1], Ilaria Bianchini[1], Alessandra Guglielmi[2]

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

[1] CNR-IMATI, via Bassini 15, 20133 Milano, Italy
raffaele@mi.imati.cnr.it, ilaria@mi.imati.cnr.it

[2] Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano, Italy
alessandra.guglielmi@polimi.it

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 NGG processes as discrete measures, where the weights are obtained by normalization of points of a Poisson process which are larger than a threshold ε. Consequently, the new process has a a.s finite number of location points. This process is then considered as the mixing measure in a mixture model for density estimation; a blocked Gibbs sampler scheme to simulate from the posterior is also given. We perform a thorough robustness analysis for the univariate Galaxy dataset and the multivariate Yeast Cell Cycle dataset with respect to the choice of hyperparameters.

Keywords: normalized generalized gamma process; hierarchical mixture models; model-based clustering; nonparametric density estimation.

1 Introduction

Sometimes in density estimation problems a parametric approach could be too restrictive: an approach allowing for a richer and larger class of models is needed. This is achieved using infinite dimensional families of probability models. We consider a mixture model having an homogeneous normalized random measure with independent increments as mixing measure, in particular the normalized generalized gamma (NGG) process. This random probability measure is more flexible than the popular Dirichlet process. In general, the main difficulty of the nonparametric approach is that posterior inference involves computation of infinite unknown parameters; there are two main approaches to deal with this problem, namely marginal and truncation algorithms. The former integrate out the infinite dimensional parameter, while the latter ones approximate the infinite dimensional process with a finite dimensional one: this turns out to be useful when interested in retrieving information about the variability of the estimates. The solution we propose can be classified as an a-priori truncation method, similar to the approach in [2] for the DPM model. In fact, we define a new discrete random probability measure, called ε-NGG process, which is a truncated version of the NGG process. In

particular, we build the new measure considering only jumps greater than a threshold $\varepsilon > 0$: only a finite number of jumps are kept by the approximation and it can be written in the form $\sum_{i=0}^{N_\varepsilon} (J_i/T_\varepsilon)\, \delta_{\tau_i}$, where N_ε is Poisson distributed. The jumps J_i are iid from a known density, due to the relationship between Poisson and Bernoulli processes. We prove that our process converges in distribution to the NGG process as ε goes to 0. The final model we consider, called ε-NGG process mixture, is:

$$\begin{aligned} & X_i|\theta_i \overset{ind}{\sim} k(\cdot;\theta_i) && i = 1,..,n \\ & \theta_i|P \overset{iid}{\sim} P && i = 1,..,n \\ & P \sim \varepsilon - NGG(\sigma,\kappa,P_0) \\ & \varepsilon,\sigma,\kappa \sim \pi(\varepsilon)\times\pi(\sigma)\times\pi(\kappa) \end{aligned} \tag{1}$$

where k is a parametric kernel and P_0 a non-atomic distribution. A Gibbs sampler scheme to simulate from the posterior of (1) is provided; when ε is small, our algorithm is meant as an extension of the well-known blocked Gibbs sampler in [2]. Note that a large range of models can be obtained as ε varies: if used as an approximation of NGG mixture models, on which many theoretical results are available in the literature (see [1]), the model is nonparametric, while when ε assumes large values it can be viewed as a model with a new finite dimensional prior.

2 Posterior Inference

We consider density estimation for two datasets: the Galaxy data and the Yeast cell cycle data, which is a multivariate dataset consisting of gene expression profiles measured at 9 different times. We carried on a deep robustness analysis of estimates and convergence of posterior chains for different prior choices; note that a prior on the parameters of the NGG process σ and κ and on ε is elicited in order to robustify the inference, as in (1): in this case, data "drive" the degree of approximation. We point out that the algorithm is quite fast and all density estimates are pretty good for both applications; see Figures 1 and 2. From the analysis we see that the model is robust with respect to the scalar parameters ε, σ and κ, while it strongly depends on the choice of the mean distribution P_0, as usual in nonparametric mixture models. Moreover, we recall that the computational cost under NGG process mixtures, i.e. ε-NGG mixtures when ε is very small, is pretty high when σ is close to 1, due to the huge number of components in the mixture in this case. This problem can be overcome considering ε-NGG mixtures with random ε. In fact, a "large" ε drives a lower number of components, trading-off the effect of large σ.

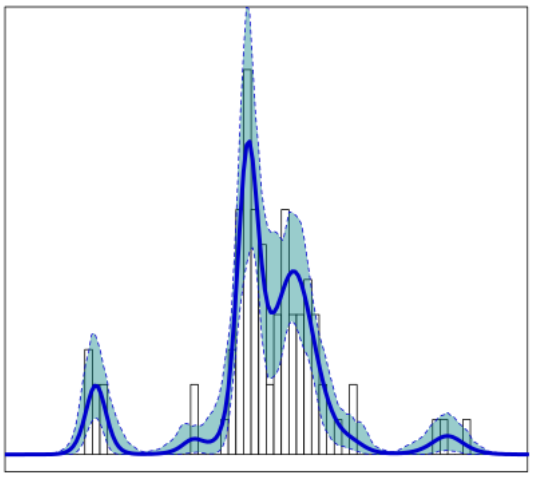

Figure 1: Example of density estimation and pointwise 90% credibility interval of Galaxy data.

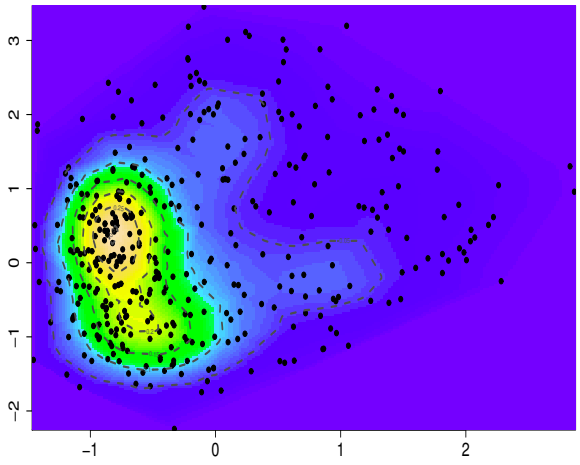

Figure 2: Bivariate density estimates (along the first 2 dimensions) of Yeast Cell Cycle data.

References

[1] Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Controlling the reinforcement in Bayesian non-parametric mixture models." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **69**(4), 715–740.

[2] Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *J. Amer. Statist. Assoc.*, **96**(453), 161–173.