Stat Methods Appl (2018) 27:231–238
https://doi.org/10.1007/s10260-017-0397-8

ORIGINAL PAPER

Computational challenges and temporal dependence in Bayesian nonparametric models

Raffaele Argiento[1] · **Matteo Ruggiero**[1]

Accepted: 27 August 2017 / Published online: 9 September 2017

Abstract Müller et al. (Stat Methods Appl, 2017) provide an excellent review of several classes of Bayesian nonparametric models which have found widespread application in a variety of contexts, successfully highlighting their flexibility in comparison with parametric families. Particular attention in the paper is dedicated to modelling spatial dependence. Here we contribute by concisely discussing general computational challenges which arise with posterior inference with Bayesian nonparametric models and certain aspects of modelling temporal dependence.

Keywords Bayesian dependent model · Conjugacy · Computation · Dirichlet · Transition function

1 Computational challenges in Bayesian nonparametric models

One of the most successful strategies of the Bayesian nonparametric approach to statistical inference has arguably been semiparametric mixture modelling, which has proved to be extremely flexible and widely applicable. Semiparametric modelling assumes the observations are generated by parametric densities conditionally on the value of a set of parameters, which in turn are assigned a nonparametric distribution. More formally, we have the hierarchical representation

$$Y_i|\theta_i \overset{\text{ind}}{\sim} \varphi(y_i|\theta_i), \quad \theta_i|G \overset{iid}{\sim} G, \quad G \sim q(G^*, \zeta). \tag{1}$$

✉ Matteo Ruggiero
matteo.ruggiero@unito.it

[1] University of Torino and Collegio Carlo Alberto, Turin, Italy

Here $Y_1, \ldots, Y_n$ are the observations, $\theta_1, \ldots, \theta_n$ are a set of latent variables that parametrise the densities $\varphi(y_i|\theta_i)$, and G is a nonparametric distribution with prior q. The latter is in turn parametrised by a *baseline* distribution G^* on the parameter space $\Theta \subset \mathbb{R}^d$ and by a vector of reals ζ.

Upon observation of a dataset, posterior inference requires evaluating the conditional distribution of the parameters given the data. As is typically the case in absence of a fully conjugate model, i.e. such that the family of distributions assigned to the parameters is closed upon conditioning to the data, one needs to resort to Markov chain Monte Carlo methods to sample from the posterior. Early contributions dealing with this problem date back to Escobar and West (1995), MacEachern and Müller (1998) and Neal (2000), and the large use of computer-aided inference has since boosted the investigation of new and efficient algorithms to deal with posterior analysis under a variety of modelling assumptions, generating a very lively literature. A brief introduction to such methods can focus on models as in (1) under the assumption that the mixing distribution G is almost surely a *discrete probability measure* with representation $G := \sum_{h\geq 1} w_h \delta_{\theta_h^*}$, where $\{w_h\}$ are random weights that sum up to one and $\{\theta_h^*\}$ are *iid* random points, taken independent of the weights, from the baseline distribution G^*. When the weights are obtained by normalising the increments of a time-changed subordinator, or more generally of a completely random measure (Kingman 1967), this specification coincides with the relevant class of random probability measures given by (homogeneous) *normalised random measures with independent increments* (Regazzini et al. 2003), which have recently been object of intense research and in turn include the celebrated Dirichlet process (Ferguson 1973). See Lijoi and Prünster (2010) for a recent review.

A broad classification of algorithms which enable to perform posterior inference under the above specifications divides them into *marginal* and *conditional* Gibbs sampling methods. Marginal Gibbs samplers are so called because they integrate out of (1) the random probability measure G. This entails sampling from

$$\mathcal{L}(\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_n|\mathbf{y}) \propto \prod_{i=1}^{n} \varphi(y_i|\theta_i)\mathcal{L}(\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_n) \tag{2}$$

where $\mathcal{L}(\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_n)$ is the prior marginal distribution of a sample from G and $\mathbf{y} = (y_1, \ldots, y_n)$ are the data. This marginal distribution can typically be characterised in terms of the predictive laws $\mathcal{L}(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$, which gives rise to Pólya urn schemes, in the case of the Dirichlet process (Blackwell and MacQueen 1973), or generalisations thereof. Accordingly, Gibbs samplers with invariant distribution (2) are often called *generalised Pólya urns* Gibbs samplers.

Since G is discrete, the θ_i's will induce a partition $\rho = \{C_1, \ldots, C_k\}$ of $\{1, \ldots, n\}$ with $C_j = \{i : \theta_i = \theta_j^*\}$, where $j = 1, \ldots, k$ and $\theta_1^*, \ldots, \theta_k^*$'s are the distinct values in $\theta_1, \ldots, \theta_n$. Given that θ_h^* are *iid* and independent of $\{w_h\}$ in G, the law $\mathcal{L}(\theta_1, \ldots, \theta_n)$ is equivalent to

$$\mathcal{L}(C_1, \ldots, C_k, d\theta_1^*, \ldots, d\theta_k^*) = p(n_1, \ldots, n_k) \prod_{j=1}^{k} G^*(d\theta_j^*),$$

where $n_i = \text{card}(C_i)$. The function p is called *exchangeable partition probability function* (EPPF), and represents the law of the random partition ρ. Marginalising over $\theta_1^*, \ldots, \theta_k^*$ allows to express (2) as

$$\mathcal{L}(C_1, \ldots, C_n | \mathbf{y}) \propto \prod_{j=1}^{k} m(y_{C_j}) p(n_1, \ldots, n_k) \tag{3}$$

where $m(y_{C_j}) := \int_\Theta \prod_{i \in C_j} \varphi(y_i|\theta) \mathrm{d}\theta$ is the marginal distribution of the data in group C_j. The computation now relies on the availability of efficient strategies for sampling from the EPPF, which are model-specific. Efficient solutions have been found based on the so called *Chinese restaurant process* (Pitman 2006) and its generalisations. Furthermore, expression (3) is the starting point for setting up inference under *product partition models* with regression on covariates, a class of models introduced in Müller et al. (2011) and extended to the spatial setting by Page and Quintana (2016) (cf. also Section 5.3 of Müller et al. 2017).

The above, briefly described, marginal sampling methods are extremely useful since they allow to reduce an infinite-dimensional computation to a finite number of operations, entailed by integrating out the random probability measure. However, a downside is that inference is limited to point estimation of linear functionals of the population such as, e.g., predictive distributions, without allowing to quantify the associated uncertainty.

Alternative strategies retain the random probability measure G as part of the model, to be updated within the Gibbs sampling routine, and are therefore called *conditional* methods. Given the series representation of G, this strategies then shift the problem to that of simulating G, conditional on the data, with small or no approximation error. Truncation methods are the most intuitive option, and entail finding an appropriate N in $G_N := \sum_{h=1}^{N} w_h \delta_{\theta_h^*}$ which guarantees certain desired minimal requirements. Several ways to achieve these have been proposed, among others, in Muliere and Tardella (1998), Ishwaran and James (2001), Barrios et al. (2013), Argiento et al. (2016a, b), Arbel and Prünster (2017). These truncation methods are generally fairly easy to implement, but need to fix a priori, implicitly or explicitly, some notion of distance between the approximating and the target measure.

Other, very successful, stochastic truncation methods allow to perform exact sampling from the random probability measure and have proven to be reliable and relatively easy to implement. These include the *slice sampler* (Walker 2007; Kalli et al. 2011) and the *retrospective sampler* (Papaspiliopoulos and Roberts, 2008) together with their adaptations and generalisations. The slice sampler requires introducing appropriate latent [0,1]-valued variables u_i so that

$$\mathcal{L}(\mathrm{d}u_i, \mathrm{d}\theta_i | G) = \sum_{h=1}^{\infty} \mathbf{1}_{(0, w_h)}(\mathrm{d}u_i) \delta_{\theta_h^*}(\mathrm{d}\theta_i),$$

whereby integrating u out of the previous recovers $\mathcal{L}(\mathrm{d}\theta_i|G) = G(\mathrm{d}\theta_i)$. Define now

$$\mathcal{L}(\theta_i|u_i, G) = G_{u_i}(\mathrm{d}\theta_i) := \sum_{h\in A(u_i)} \delta_{\theta_h^*}(\mathrm{d}\theta_i)$$

where $A(u_i) := \{h : w_h > u\}$ is a finite set. From the latter it is clear that sampling G, conditional on $u_1, \ldots, u_n, \theta_1, \ldots, \theta_n$ and the data, entails updating only finitely-many of its components, namely the pairs (w_h, θ_h^*) for $h \in \cup_{i=1}^n A(u_i)$.

The retrospective sampler instead is based on the idea of exchanging the intuitive order of simulation for sampling from G. This would lead to sampling the infinite sequences $\{w_h\}, \{\theta_h^*\}$, then draw v_i uniformly distributed in $(0, 1)$ and set $\theta_i = \theta_l^*$ if $\sum_{j=1}^{l-1} w_l < v_i < \sum_{j=1}^{l-1} w_l$, The retrospective sampler instead first samples v_i and then draws as many w_h, θ_h^* as are needed to meet the above inequalities.

Gibbs sampling procedures described so far are very appealing strategies but still computationally intensive methods. This makes the use of mixtures such as (1) infeasible when dealing with large datasets, or when the computational resources are limited. Recently, variational Bayes methods have been proposed as an alternative (Blei and Jordan 2006). Acting essentially as optimisation algorithms, under these methods the posterior distribution of G is approximated by a distribution $\tilde{q}$, called *variational distribution*, of a finite dimensional process. The goal is then to adjust the parameters of $\tilde{q}$ in order to minimise the Kullback–Leibler divergence between $\tilde{q}$ and the posterior. Robustness of variational Bayes methods is currently one of the open problems in the Bayesian nonparametric literature, as it is known they can underestimate the model uncertainty.

2 Temporal dependence in Bayesian nonparametric models

An important line of research in Bayesian nonparametrics on the so called dependent processes has developed from the ideas introduced in MacEachern (1999), where collections of dependent random probability measures $\{G_z, z \in \mathcal{Z}\}$ are considered, and G_z encodes the current state of the problem in correspondence of the covariate value z. Cf. Müller et al. (2017, Section 3.2.1). Computational methods for dependent models are very often problem-specific extensions of those summarised in Sect. 1. Providing a general overview of these computational strategies would be a difficult task far beyond the scope and possibilities of this discussion. Since Section 5 of Müller et al. (2017) presents some applications of dependent models for spatial data, we choose here to briefly discuss some issues related to models with temporal dependence, with particular emphasis on the role of conjugacy.

A common setting for Bayesian inference with temporal dependence is that of partial exchangeability, whereby the available data are of the form $y_{t_i,1}, \ldots, y_{t_i,n_i}$, where the indices t_i are discrete data-collection times, $n_i \geq 1$ for all i, and the data $y_{t_i,j}$ are such that, as j varies,

$$y_{t_i,j} \mid G_{t_i} \sim^{iid} G_{t_i}.$$

Hence the data are exchangeable across the t_i-sections, but not overall. From a temporal modelling perspective, one ideally wants the correlation between pairs of random measures G_t and G_s to increase as the indices t and s get closer, and decay to zero as t and s grow farther apart.

A non exhaustive list of contributions, along this line of investigation, based on Dirichlet mixture models includes, among others, Dunson (2006), Caron et al. (2008), Griffin and Steel (2010), Caron et al. (2007), Rodriguez and Horst (2008), Taddy and Kottas (2009), Mena and Ruggiero (2016). Other contributions have explored models which go beyond the structure of the Dirichlet process or closely related constructions, aiming at modelling, for example: marginal measures of the dependent process of geometric stick-breaking type (Mena et al. 2011), of Pitman–Yor type (Caron et al. 2017), of GEM type (Gutierrez et al. 2016), or of gamma type (Caron and Teh 2012); evolving binary matrices for relational network structures (Durante and Dunson 2014), or for dynamic feature allocation (Perrone et al. 2017); monotonic functional time series (Canale and Ruggiero 2016); emission distributions for hidden Markov models (Yau et al. 2011; Linderman et al. 2016).

Here we are interested in highlighting two roles conjugacy can play in these approaches to inference. One is with the aim of constructing stationary temporal models with a built-in simulation scheme available, as done in Pitt and Walker (2005), Walker et al. (2007), Mena and Walker (2009). The kernel of the idea is to consider joint distributions, for some fixed $n \geq 1$,

$$\mathcal{L}(\mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_n, \mathrm{d}G) = \mathcal{L}(\mathrm{d}G) \prod_{i=1}^{n} \mathcal{L}(\mathrm{d}\theta_i \mid G)$$

where q is the nonparametric prior on G, and to construct transition functions through latent variables by writing

$$P(G, \mathrm{d}H) = \int \mathcal{L}(\mathrm{d}H \mid \theta_1, \ldots, \theta_n) \prod_{i=1}^{n} \mathcal{L}(\mathrm{d}\theta_i \mid G) \tag{4}$$

where $\mathcal{L}(\mathrm{d}H \mid \theta_1, \ldots, \theta_n)$ is the posterior of H given $\theta_1, \ldots, \theta_n$. For example, if $p := G(A) \in [0, 1]$ for some fixed set A, the law of $G(A)$ is a beta distribution and $\mathcal{L}(\mathrm{d}\theta_i \mid G(A))$ is Bernoulli with parameter p, then the above reduces to a beta-binomial transition

$$P(p, \mathrm{d}p') = \mathrm{beta}(\mathrm{d}p' \mid a + \theta, b + n - \theta)\mathrm{Binom}(\theta \mid n, p)$$

where (a, b) are prior beta hyperparameters. Note that this is in fact the transition function of the marginal state of a two-dimensional Gibbs sampler on the augmented space of (p, θ), which is stationary with respect to a beta. In a nonparametric framework, if $\mathcal{L}(\mathrm{d}G) = \Pi(\mathrm{d}G \mid \alpha)$ for some finite parameter measure α, and Π is conjugate in the sense that $G \sim \Pi$ and $X \mid G \sim G$ jointly imply $\Pi \mid X \sim \Pi(\cdot \mid f(\alpha, X))$ for some function f of the data and the prior parameter, then (4) yields

$$P(G, \mathrm{d}H) = \int \Pi(\mathrm{d}H \mid f(\alpha, \theta_1, \ldots, \theta_n)) \prod_{i=1}^{n} G(\mathrm{d}\theta_i). \tag{5}$$

Here Π can be shown to be the reversible measure of the process, so this strategy allows to construct stationary nonparametric processes. Lijoi et al. (2016) discuss along these lines the Bayesian interpretation of the dynamics of two families of continuous-time Dirichlet and gamma dependent models for Bayesian nonparametrics, the latter used for example in Caron and Teh (2012). See also Ruggiero and Walker (2009). The transition functions of such models can be obtained by randomising n in (4) and by introducing appropriate coefficients which make $P(G, \mathrm{d}H)$ satisfy the Chapman–Kolmogorov conditions in continuous time. For example, for these two families one has

$$P_t(G, \mathrm{d}H) = \sum_{n \geq 0} P(N_t = n) \int \Pi(\mathrm{d}H \mid f(\alpha, \theta_1, \ldots, \theta_n)) \prod_{i=1}^{n} G(\mathrm{d}\theta_i), \tag{6}$$

where N_t is an appropriate $\mathbb{Z}_+$-valued continuous-time process which determines the size of the latent sample $(\theta_1, \ldots, \theta_n)$ and complies with the requirements on the correlation mentioned at the beginning of the section. This approach has been followed explicitly in Walker et al. (2007). The resulting transitions are therefore infinite mixtures. Simulation of these transition functions can in principle be done by resorting to one of the methods outlined in the previous section, e.g. by using a slice sampler twice on the mixture (6) and on the infinite-dimensional random measure which is the state of process, as done for example in Mena et al. (2011). Model-specific hurdles however may make call for additional steps, e.g. Jenkins and Spanò (2017) develop an exact simulation scheme for (6) in some finite and infinite-dimensional Dirichlet cases, which deals efficiently with the non trivial expression for $P(N_t = n)$, which has itself a series representation.

Alternatively, conjugacy can be deliberately sought in order to reduce the overall Monte Carlo error and predictive uncertainty within a broader computation. Papaspiliopoulos et al. (2016) for example extend classical posterior characterisations for Dirichlet and gamma random measures to the two above-mentioned families of dependent processes, conditional on discretely collected data. In particular, sufficient conditions are identified for these models (cf. Papaspiliopoulos and Ruggiero 2014) that allow to write (6), conditional on $y_1, \ldots, y_m$ collected possibly at different times, as

$$P_t(G, \mathrm{d}H \mid y_1, \ldots, y_m) = \sum_{i=0}^{m} w_i(t) \Pi(\mathrm{d}H \mid f(\alpha, y_1, \ldots, y_i)).$$

This reduces (6), upon conditioning to the observed data, to a finite mixture of distributions in the same conjugate family. Note that the mixture components only consider $y_1, \ldots, y_i$ and not the entire sample. The $w_i(t)$'s are appropriate time-dependent weights which regulate how the posterior mass is reassigned at different times to the mixture components.

References

Arbel J, Prünster I (2017) A moment-matching Ferguson & Klass algorithm. Stat Comput 27:3–17

Argiento R, Bianchini I, Guglielmi A (2016a) A blocked Gibbs sampler for NGG-mixture models via a priori truncation. Stat Comput 26:641–661

Argiento R, Bianchini I, Guglielmi A (2016b) Posterior sampling from ε-approximation of normalized completely random measure mixtures. Electron J Stat 10:3516–3547

Barrios E, Lijoi A, Nieto-Barajas LE, Prünster I (2013) Modeling with normalized random measure mixture models. Stat Sci 28:313–334

Blackwell D, MacQueen JB (1973) Ferguson distributions via Pólya urn schemes. Ann Stat 1:353–355

Blei DM, Jordan MI (2006) Variational inference for Dirichlet process mixtures. Bayesian Anal 1:121–143

Canale A, Ruggiero M (2016) Bayesian nonparametric forecasting of monotonic functional time series. Electron J Stat 10:3265–3286

Caron F, Teh YW (2012) Bayesian nonparametric models for ranked data. In: Neural information processing systems (NIPS 2012), Lake Tahoe, USA

Caron F, Davy M, Doucet A (2007) Generalized Polya urn for time-varying Dirichlet process mixtures. In: Proceedings of 23rd conference on uncertainty in artificial intelligence, Vancouver

Caron F, Davy M, Doucet A, Duflos E, Vanheeghe P (2008) Bayesian inference for linear dynamic models with Dirichlet process mixtures. IEEE Trans Signal Process 56:71–84

Caron F, Neiswanger W, Wood F, Doucet A, Davy M (2017) Generalized Pólya urn for time-varying Pitman-Yor processes. J Mach Learn Res 18:1–32

Dunson DB (2006) Bayesian dynamic modelling of latent trait distributions. Biostatistics 7:551–568

Durante D, Dunson D (2014) Nonparametric Bayes dynamic modelling of relational data. Biometrika 101:883–898

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90:577–588

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230

Griffin JE, Steel MFJ (2010) Stick-breaking autoregressive processes. J Econom 162:383–396

Gutierrez L, Mena RH, Ruggiero M (2016) On GEM diffusive mixtures. In: Proceedings JSM (ed) Section on nonparametric statistics. American Statistical Association, Alexandria

Ishwaran H, James L (2001) Gibbs sampling methods for stick-breaking priors. J Am Stat Assoc 96:161–173

Jenkins PA, Spanò D (2017) Exact simulation of the Wright-Fisher diffusion. Ann Appl Probab 27:1478–1509

Kingman JFC (1967) Completely random measures. Pac J Math 21:59–78

Lijoi A, Prünster I (2010) Models beyond the Dirichlet process. In: Hjort NL, Holmes CC, Müller P, Walker SG (eds) Bayesian nonparametrics. Cambridge University Press, Cambridge, pp 80–136

Lijoi A, Ruggiero M, Spanò D (2016) On the transition function of some time-dependent Dirichlet and gamma processes. In: Proceedings JSM (ed) Section on nonparametric statistics. American Statistical Association, Alexandria

Linderman SW, Johnson MJ, Wilson MW, Chen Z (2016) A nonparametric Bayesian approach to uncovering rat hippocampal population codes during spatial navigation. J Neurosci Methods 263:36–47

Kalli M, Griffin JE, Walker SG (2011) Slice sampling mixture models. Stat Comput 21:93–105

MacEachern SN, Müller P (1998) Estimating mixture of Dirichlet process models. J Comput Graph Stat 7:223–238

MacEachern SN (1999). Dependent nonparametric processes. In: ASA proceedings of the section on Bayesian statistical science. American Statistical Association, Alexandria, VA

Mena RH, Ruggiero M (2016) Dynamic density estimation with diffusive Dirichlet mixtures. Bernoulli 22:901–926

Mena RH, Walker SG (2009) On a construction of Markov models in continuous time. Metron 67:303–323

Mena RH, Ruggiero M, Walker SG (2011) Geometric stick-breaking processes for continuous-time Bayesian nonparametric modelling. J Stat Plan Inf 141:3217–3230

Muliere P, Tardella L (1998) Approximating distributions of random functionals of Ferguson–Dirichlet priors. Can J Stat 26:283–297

Müller P, Quintana FA, Rosner GL (2011) A product partition model with regression on covariates. J Comput Graph Stat 20:260–278

Müller P, Quintana FA, Page G (2017) Nonparametric Bayesian inference in applications. Stat Methods Appl **(to be completed)**

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat 9:249–265
Page G, Quintana FA (2016) Spatial product partition models. Bayesian Anal 11:265–298
Papaspiliopoulos O, Ruggiero M (2014) Optimal filtering and the dual process. Bernoulli 20:1999–2019
Papaspiliopoulos O, Roberts GO (2008) Retrospective MCMC for Dirichlet process hierarchical models. Biometrika 95:169–186
Papaspiliopoulos O, Ruggiero M, Spanò D (2016) Conjugacy properties of time-evolving Dirichlet and gamma random measures. Electron J Stat 10:3452–3489
Perrone V, Jenkins PA, Spanò D, Teh YW (2017) Poisson random fields for dynamic feature models. Preprint available at arXiv:1611.07460
Pitman J (2006) Combinatorial Stochastic Processes. In: École d'Été de Probabilités de Saint-Flour XXXII-2002. Springer, Berlin
Pitt MK, Walker SG (2005) Constructing stationary time series models using auxiliary variables with applications. J Am Stat Assoc 100:554–564
Regazzini E, Lijoi A, Prünster I (2003) Distributional results for means of random measures with independent increments. Ann Stat 31:560–585
Rodriguez A, Ter Horst E (2008) Bayesian dynamic density estimation. Bayesian Anal 3:339–366
Ruggiero M, Walker SG (2009) Bayesian nonparametric construction of the Fleming–Viot process with fertility selection. Stat Sin 19:707–720
Taddy MA, Kottas A (2009) Markov switching Dirichlet process mixture regression. Bayesian Anal 4:793–816
Walker SG (2007) Sampling the Dirichlet mixture model with slices. Commun Stat Simul Comput 36:45–54
Walker SG, Hatjispyros SJ, Nicoleris T (2007) A Fleming–Viot process and Bayesian nonparametrics. Ann Appl Probab 17:67–80
Yau C, Papaspiliopoulos O, Roberts GO, Holmes C (2011) Bayesian non-parametric hidden Markov models with applications in genomics. J R Stat Soc Ser B 73:37–57