Supplementary Material: Slice sampling normalized kernel-weighted completely random measure mixture models

Anonymous Author(s)
Affiliation
Address
email

1 Slice sampler

In this section we provide an unabridged derivation of the slice sampler for normalized kernel CRMs. We also present the Rao-Blackwellized estimator for the predictive density.

1.1 Derivation

A common approach to simplify sampling from infinite mixture models is to introduce another set of auxiliary variables, $u_{g,i}$, such that conditional on the $u_{g,i}$ each P_g will be a finite measure [1]. We can then write the mixture model for a single data point, $y_{g,i}$, as

$$f(y_{g,i}, u_{g,i}, s_{g,i}|\pi, \mu, \theta, \phi) = B_{+g}^{-1}\mathbf{1}\left(u_{g,i} < K(x_g^*, \mu_{s_{g,i}})\pi_{s_{g,i}}\right) q(y_{g,i}|\theta_{s_{g,i}}, \phi_{s_{g,i}}) \tag{1}$$

where where we have introduced the notation $B_{+g} = B_{x_g^*}(\Theta)$. From Equation 1 we write the density of all points with covariate x_g^* as

$$f(\mathbf{y}_g, \mathbf{u}_g, \mathbf{s}_g|\pi, \mu, \theta, \phi) = B_{+g}^{-n_g}\prod_{i=1}^{n_g}\mathbf{1}\left(u_{g,i} < K(x_g^*, \mu_{s_{g,i}})\pi_{s_{g,i}}\right) q(y_{g,i}|\theta_{s_{g,i}}, \phi_{s_{g,i}}). \tag{2}$$

Following [1, 2, 3] we can replace $B_{+g}^{-n_g}$, which involves an infinite sum, by relating it to the gamma distribution yielding

$$f(\mathbf{y}_g, \mathbf{u}_g, \mathbf{s}_g|\pi, \mu, \theta, \phi) = V_g^{n_g-1}\exp\left(-V_g B_{+g}\right)\prod_{i=1}^{n_g}\mathbf{1}\left(u_{g,i} < K(x_g^*, \mu_{s_{g,i}})\pi_{s_{g,i}}\right) q(y_{g,i}|\theta_{s_{g,i}}, \phi_{s_{g,i}}) \tag{3}$$

where $V_g \sim \text{Ga}(n_g, B_{+g})$. The joint distribution for all data, $f(\mathbf{y}, \mathbf{u}, \mathbf{s}|\pi, \mu, \theta, \phi)$, is then the product of Equation 3 for each unique covariate x_g^*

$$f(\mathbf{y}, \mathbf{u}, \mathbf{s}|\pi, \mu, \theta, \phi) = \prod_{g=1}^{G} V_g^{n_g-1}\exp\left(-V_g B_{+g}\right)\prod_{i=1}^{n_g}\mathbf{1}\left(u_{g,i} < K(x_g^*, \mu_{s_{g,i}})\pi_{s_{g,i}}\right) q(y_{g,i}|\theta_{s_{g,i}}, \phi_{s_{g,i}}). \tag{4}$$

A MCMC sampler is still hard to construct for Equation 3, since the infinite sum B_{+g} still makes an appearance. To alleviate this difficulty we define a truncation level according to the auxiliary $u_{g,i}$ variables introduced earlier [4]. Specifically, let $0 < L = \min\{u_{s_{g,i}}\}$ and assume that there are M_g atoms such that $K(x_g^*, \mu_m)\pi_m \geq L$ for some g, and $M = \sum_{g=1}^{G} M_g$. We can then rewrite

$B_{+g} = \sum_{m=1}^{M_g} K(x_g^*, \mu_m)\pi_m + \sum_{m=M_g+1}^{\infty} K(x_g^*, \mu_m)\pi_m$, where we denote the second term as B_g^*, i.e. the portion of the total mass of B_g from kernlized atoms with mass less than L. With this new notation we rewrite Equation 4 as

$$f(\mathbf{y}, \mathbf{u}, \mathbf{s}|\pi, \mu, \theta, \phi) = \prod_{g=1}^{G} V_g^{n_g - 1} \prod_{i=1}^{n_g} \mathbf{1}\left(u_{g,i} < K(x_g^*, \mu_{s_{g,i}})\pi_{s_{g,i}}\right) q(y_{g,i}|\theta_{s_{g,i}}, \phi_{s_{g,i}}) \\ \times \exp\left(-V^T B_+\right) \exp\left(-V^T B_*\right) \tag{5}$$

where $V = [V_1, \ldots, V_G]^T$, $B_+ = [B_1, \ldots, B_G]^T$, and $B_* = [B_1^*, \ldots, B_G^*]^T$ and $B_g = \sum_{m=1}^{M_g} K(x_g^*, \mu_m)\pi_m$. We then marginalize out all kernelized atoms with mass less than L which allows us to write the joint distribution of the model as

$$\begin{aligned} p(\mathbf{y}, \mathbf{u}, \mathbf{s}, V, M, \pi, \mu, \theta, \phi, \alpha) = {} & p(\alpha)p(M|\alpha)p(\theta_{1:M})p(\pi_{1:M})p(\mu_{1:M}) \\ & \times \prod_{g=1}^{G} V_g^{n_g - 1} \prod_{i=1}^{n_g} \mathbf{1}\left(u_{g,i} < K(x_g^*, \mu_{s_{g,i}})\pi_{s_{g,i}}\right) q(y_{g,i}|\theta_{s_{g,i}}, \phi_{s_{g,i}}) \\ & \times \exp\left(-V^T B_+\right) \mathbb{E}[\exp\left(-V^T B_*\right)] \end{aligned} \tag{6}$$

We recognize the expectation in Equation 6 as the characteristic function of the Lévy process underlying the kernel-weighted CRM (see Section 2.1 of the main text). We can use the Lévy-Khintchine representation [5] of a Lévy process to simplify the expectation as

$$\mathbb{E}[\exp\left(-V^T B_*\right)] = \exp\left(-\alpha \int_A (1 - \exp\left(-V^T \mathcal{K}_\mu \pi\right))\nu(d\mu, d\pi)\right) \tag{7}$$

where $\mathcal{K}_\mu = [K(x_1^*, \mu), \ldots, K(x_G^*, \mu)]^T$ and $A = \{(\mu, \pi) : K(x_g^*, \mu)\pi < L\}$. Since we have a fixed kernel function ($K(\cdot, \cdot) \in [0, 1]$) and have assumed a finite dictionary of atom locations $\{\mu^*\}$, the integral in Equation 7 decomposes into two parts. The first part corresponds to atoms (π, μ) where $\pi < L$ which can be written as

$$\sum_{\mu^* \in \mathcal{Y}} \left(R_0(\mu^*) \int_0^L (1 - \exp\left(-V^T \mathcal{K}_{\mu^*} \pi\right))\nu(d\pi)\right) \tag{8}$$

and can be evaluated numerically for many CRMs including gamma and generalized gamma processes [1] by using the identity

$$\int_0^L (1 - \exp\left(-V^T \mathcal{K}_{\mu^*} \pi\right))\nu(d\pi)) = \psi(V^T \mathcal{K}_{\mu^*})/\alpha - \int_L^\infty (1 - \exp\left(-V^T \mathcal{K}_{\mu^*} \pi\right))\nu(d\pi). \tag{9}$$

The first term in Equation 9, $\psi(\cdot)$ is given by the exponent on the right side of Equation 7. Both terms of Equation 9 can be evaluated by numerical methods since they are one-dimensional integrals.

The second part of the integral in Equation 7 consists of realized atoms $\{(\pi_m, \mu_m)\}$ such that $\mathcal{K}_{\mu_m}\pi_m < L$ at covariate x_g^*. We evaluate this term with a Monte Carlo estimate

$$\frac{1}{Z} \sum_{g=1}^{G} \sum_{m=1}^{M} \mathbf{1}\left(K(x_g^*, \mu_m)\pi_m < L\right) \exp(-V_g K(x_g^*, \mu_m)\pi_m) \tag{10}$$

where $Z = \sum_{g=1}^{G} \sum_{m=1}^{M} \mathbf{1}\left(K(x_g^*, \mu_m)\pi_m < L\right)$. Recall that M is the number of instantiated atoms. In very simple cases the term in Equation 10 can be solved for analytically; in the case of a box kernel, it doesn't arise at all. In our experiments we consider both a box kernel and a square exponential kernel and we have found that the term contributes little to the accuracy of the sampler and very good results can be obtained by simply ignoring this term. However, for kernels that decay more slowly than the square exponential kernels we use this term will likely be more significant.

1.2 Prediction

For a new observation y^* with covariate x^*, using the slice sampler described above we can simulate from the predictive distribution $p(y^*|y)$ and propose a Rao-Blackwellized estimate of it without any truncation error. Analogously to [1], we introduce a new auxiliary variable u_* and allocation variable s_* for the new observation which we describe how to sample below. Then, the predictive density estimate is defined as

$$\hat{f}(y^*) = \frac{1}{T}\sum_{i=1}^{T} \frac{\sum_{m=1}^{M^{(i)}} \mathbf{1}\left(K(x^*, \mu_m^{(i)})\pi_m^{(i)} > u_*^{(i)}\right) q\left(y^*|\theta_m^{(i)}, \phi_m^{(i)}\right)}{\sum_{m=1}^{M^{(i)}} \mathbf{1}\left(K(x^*, \mu_m^{(i)})\pi_m^{(i)} > u_*^{(i)}\right)} \tag{11}$$

where $M^{(i)}$ is the number of used clusters in sample i and $u_*^{(i)}$ and $s_*^{(i)}$ are the i'th sample of u_* and s_* respectively. We sample s_* from a discrete distribution with

$$p(s_* = m) \propto \mathbf{1}\left(u_* < K(x^*, \mu_m)\pi_m\right) \tag{12}$$

The only other changes to the sampler is that when sampling V_g, $\{u_{g,i}\}$, M and $\{\pi_m\}$ with data allocated to them, a sample size of n_g is used, rather than $n_g - 1$. The same sampling methods can be used for each of these variables. We use this estimator in the experiments to estimate the predictive density on a fine grid of values.

2 Finite normalized KGaP

A gamma process (GaP) on a measurable space Θ, denoted GaP(H_0, β), is a CRM with Lévy measure $\nu(d\theta, d\pi) = \pi^{-1}e^{-\beta\pi}B_0(d\theta)$, where we have included the scale parameter for generality. Considered as a Poisson process on $\Theta \times \mathbb{R}^+$, a random measure drawn from a GaP, $X \sim$ GaP(H_0, β), is a discrete measure with an infinite number of atoms [6] where

$$X = \sum_{m=1}^{\infty} \pi_m \delta_{\theta_m^*} \tag{13}$$

We can approximate the countably infinite random measure X with a finite version X_M, where we restrict the measure to only have M atoms. We introduce the finite measure

$$\nu_\delta(d\theta, d\pi) = \pi^{\delta-1}e^{-\beta\delta}H_0(d\theta) \tag{14}$$

for $\delta > 0$. As δ gets smaller more mass is placed on smaller values π and so M will need to be large to obtain atoms with significant mass. Since ν_δ is proportional to the density of a Ga(δ, β) random variable it is easy to compute $\nu_\delta(\Theta \times \mathbb{R}^+)$ as

$$\int_{\Theta\times\mathbb{R}^+} \pi^{\delta-1}e^{-\beta\delta}H_0(d\theta) = H_0(\Theta)\frac{\Gamma(\delta)}{\beta^\delta} \tag{15}$$

which for $\delta > 0$ is finite. In fact, using 14 as the rate measure of a finite Poisson process on $\Theta \times \mathbb{R}^+$ and defining X_M as in Equation 13 one has that $M = \nu_\delta(\Theta \times \mathbb{R}^+)$ [6].

It is easy to see that as $\delta \to 0$ the finite rate measure converges to that of a GaP. Consider

$$\frac{\nu_\delta(d\theta, d\pi)}{\nu(d\theta, d\pi)} = \frac{\pi^{\delta-1}e^{-\beta\pi}H_0(d\theta)}{\pi^{-1}e^{-\beta\pi}H_0(d\theta)} = \pi^\delta \to 0 \tag{16}$$

in the limit $\delta \to 0$. In practice we choose the number of desired atoms, M and then set $\delta = 1/M$. In what follows we only consider the case $\beta = 1$ and so we drop it from our notation. One could also design reversible-jump moves [7] using this approximation.

With a finite approximation to a gamma process, we construct a finite version of a kernel GaP and then normalize it. Let A be measurable on Θ and define similarly as the infinite version

$$B_x^M(A) = \sum_{m=1}^{M} K(x, \mu_m)\pi_m \delta_{\theta_m^*}(A) \tag{17}$$

where $\pi_m \sim$ Ga$(1/M, 1)$ and $\theta_m^* \sim H_0(d\theta)$. We can then define the finite KNRM

$$P_x^M(A) = \sum_{m=1}^{M} \frac{K(x, \mu_m)\pi_m}{\sum_{l=1}^{M} K(x, \mu_l)\pi_l}\delta_{\theta_m^*}(A) \tag{18}$$

We can use then use P_x^M as a prior for the mixture model described in the main text.

2.1 Gibbs sampler

Since there are only a finite number of atoms in the approximation described in Section 2, there is no need to perform the marginalization of small atoms required for the slice sampler. This allows a simple Gibbs sampler to be derived when using the finite approximation. We describe the sampling equations for the normalized KGaP below, the model parameters are sampled the same as with the slice sampler.

- **Cluster allocations** $s_{g,i}$: The conditional distribution for $s_{g,i}$ is given by (up to a constant)

  $$s_{g,i} = m \mid y_{g,i}, \pi_m, \mu_m, \theta_m, \phi_m \propto K(x_g^*, \mu_m)\pi_m q(y_{g,i}|\theta_m, \phi_m) \tag{19}$$

  for $1 \leq m \leq M$.
- **Raw atom sizes** π_m: The conditional distributions for the atoms sizes up to a constant is given by

  $$\pi_m \mid \{s\}, \{\mu\}, \{\pi_{-m}\} \propto \text{Ga}(1/M, 1) \prod_{g=1}^{G} \prod_{i=1}^{n_g} \frac{K(x_g^*, \mu_{s_{g,i}})\pi_{s_{g,i}}}{\sum_{l=1}^{M} K(x_g^*, \mu_l)\pi_l} \tag{20}$$

  This distribution could be sampled with Metropolis-Hastings, however we have found slice sampling [8] to be efficient.
- **Raw atom covariate locations** μ_m: Since we assume a finite set of covariate locations, the conditional distribution is give by

  $$\mu_m = \mu_p^* \mid \{s\}, \{\pi\}, \{\mu_{-m}\} \propto R_0(\mu_p^*) \prod_{g=1}^{G} \prod_{i=1}^{n_g} \frac{K(x_g^*, \mu_{s_{g,i}})^{\mathbf{1}(s_{g,i} \neq m)} K(x_g^*, \mu_p^*)^{\mathbf{1}(s_{g,i} = m)} \pi_{s_{g,i}}}{\sum_{l \neq m} K(x_g^*, \mu_l)\pi_l + K(x_g^*, \mu_p^*)\pi_m}. \tag{21}$$

3 Extra experimental results

In Table 1 we show the held-out predictive log-likelihoods obtained with the Rao-Blackwellized estimator for the slice sampler using both the box and square exponential (SE) kernels.

Table 1: Rao-Blackwellized estimates of held-out predictive log-likelihood.

	Synthetic	**CMB**	**Motorcycle**
Box	-2.23 (0.06)	-0.11 (0.004)	-0.28 (0.008)
SE	NA	-0.18 (0.004)	-0.27 (0.007)

References

[1] Jim E. Griffin and Stephen G. Walker. Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20(1):241–259, 2011.

[2] Lancelot F. James, Antonio Lijoi, and Igor Prnster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

[3] Stefano Favaro and Yee Whye Teh. Mcmc for normalized random measure mixture models. *Submitted to Statistical Science*, 2012.

[4] Maria Kalli, Jim E. Griffin, and Stephen G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, January 2011.

[5] B. Fristedt and L.F. Gray. *A Modern Approach to Probability Theory*. Probability and Its Applications. Birkhäuser, 1997.

[6] J.F.C. Kingman. *Poisson processes*. OUP, 1993.

[7] Peter J. Green and David I. Hastie. Reversible jump mcmc, 2009.

[8] P. Damien, J. Wakefield, and S. Walker. Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society Series B*, 61(2):331–344, 1999.