Dependent Normalized Random Measures

Changyou Chen[1,3] CHANGYOU.CHEN@NICTA.COM.AU
Vinayak Rao[2] VRAO@GATSBY.UCL.AC.UK
Wray Buntine[3,1] WRAY.BUNTINE@NICTA.COM.AU
YeeWhye Teh[4] Y.W.TEH@STATS.OX.AC.UK
[1]RSISE, Australian National University, Australia; [2]Dept. Statistical Science, Duke University, USA; [3]National ICT, Canberra, Australia; [4]Dept. Statistics, University of Oxford, UK

Abstract

In this paper we propose two constructions of dependent *normalized random measures*, a class of nonparametric priors over dependent probability measures. Our constructions, which we call mixed normalized random measures (MNRM) and thinned normalized random measures (TNRM), involve (respectively) weighting and thinning parts of a shared underlying Poisson process before combining them together. We show that both MNRM and TNRM are marginally normalized random measures, resulting in well understood theoretical properties. We develop marginal and slice samplers for both models, the latter necessary for inference in TNRM. In time-varying topic modeling experiments, both models exhibit superior performance over related dependent models such as the hierarchical Dirichlet process and the spatial normalized Gamma process.

1. Introduction

In recent years there has been growing interest in extending models for random probability measures (RPMs) like the Dirichlet process (DP) to *dependent* random probability measures (MacEachern, 1999). A popular class is the hierarchical Dirichlet process (HDP) (Teh et al., 2006), which introduces dependencies in an exchangeable set of DPs by having them share the same random base measure. A number of alternate approaches exist in the literature, many indexing DPs by more structured sets, and allowing more refined control of dependency. Examples include (Srebro & Roweis, 2005; Griffin & Steel, 2006; Caron et al., 2007; Ahmed & Xing, 2008; MacEachern et al., 2001; Gelfand et al., 2005). Of relevance to this paper is the *spatial normalized Gamma process* (Rao & Teh, 2009), which exploits the representation of the DP as a normalized Gamma process, and constructs dependent DPs from overlapping restrictions of a common Gamma process. (Lin et al., 2010) considered additional operations to introduce dependencies between DPs, viz. subsampling and perturbing atoms of a common Gamma process.

There has also been a growing body of work extending the DP to more expressive RPMs. A flexible framework is the class of normalized random measures (NRMs) (James et al., 2009), which includes the DP, the normalized inverse Gaussian process and the normalized generalized Gamma process.

In this paper we consider constructions for dependent random measures, all of which are marginally distributed as a specific NRM. We propose two approaches which we call *mixed normalized random measures* (MNRM) and *thinned normalized random measures* (TNRM), and study these models using tools from the Poisson process partition calculus of (James, 2005). Our framework encompasses and extends work such as (Rao & Teh, 2009), (Griffin et al., 2012), (Chen et al., 2012), (Lin et al., 2010), and (Lin & Fisher, 2012), and is an alternative to (Williamson et al., 2010). One contribution of this work is a systematic comparison of these related models, and we find that on a number of real world datasets, the MNRM (a novel and simple model) perform best. Additionally, many of the listed works propose approximate posterior MCMC samplers; here we develop and compare two exact samplers, a marginal Gibbs sampler and a slice sampler. We find the latter preferable in most cases. We also provide faster approximations to this sampler, as well as bounds on the approximation error.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

Proofs are provided in the appendices in the supplementary material.

2. Normalized Random Measures

In this section we provide a concise review of normalized random measures (NRMs). Consider a Poisson process $\mathcal{N}$ on a product space $\mathbb{S} = \mathbb{R}^+ \times \Theta$, with intensity $\nu(w, \theta)$. A completely random measure (CRM) on Θ is defined as a linear functional of $\mathcal{N}$:

$$\tilde{\mu}(\mathrm{d}\theta) = \int_{\mathbb{R}^+} w\mathcal{N}(\mathrm{d}w, \mathrm{d}\theta) = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}(\mathrm{d}\theta) \quad (1)$$

Here $\{(w_i, \theta_i)\}$ are the atoms of $\mathcal{N}$. The Poisson intensity $\nu(w, \theta)$ is the density of the *Lévy measure* of $\tilde{\mu}$ (called Lévy intensity), and is defined so that the total measure $Z = \int_\Theta \tilde{\mu}(\mathrm{d}\theta) = \int_{\mathbb{R}^+ \times \Theta} w\mathcal{N}(\mathrm{d}w, \mathrm{d}\theta)$ is finite and positive almost surely. A CRM is so called because of its property that the random masses assigned to disjoint sets are independent, this follows from the properties of the Poisson process (which itself is a CRM)[1]. When $\nu(w, \theta) = Mw^{-1}e^{-w}H(\theta)$, for some probability density H, we get a homogeneous Gamma process with concentration parameter M and base distribution with density H. Other examples include the stable process, the inverse Gaussian process and the generalized Gamma process, which has Lévy intensity $\nu(w, \theta) = Mw^{-\sigma-1}e^{-w}H(\theta)$ and which includes the other examples as subclasses.

The total mass is finite and nonzero. An NRM is obtained by normalizing $\tilde{\mu}$ to a random *probability* measure: $\mu(\mathrm{d}\theta) = \frac{1}{Z}\tilde{\mu}(\mathrm{d}\theta)$. From the Poisson construction, it follows that like the DP, an NRM is a discrete RPM with a countably infinite number of atoms:

$$\mu = \frac{1}{Z}\sum_{k=1}^{\infty} w_k \delta_{\theta_k}, \qquad Z = \sum_{k=1}^{\infty} w_k \quad (2)$$

3. Dependent NRMs

In many applications, one has observations at a finite collection of indices $\mathcal{T} = (t_1, \cdots, t_T)$, we refer to the indices as 'times' from now on. The observations at each time t are modeled as i.i.d. draws from a random probability measure μ_t. By allowing dependencies in the measures μ_t, one allows the sharing of statistical information between observations at different times. We use the term *dependent normalized random measures* (dNRMs) to refer to a dependent set of random measures $\{\mu_t\}$, each distributed marginally as a NRM.

Here we propose two approaches to model dependencies between the measures μ_t: *mixed normalized random measures* (MNRM) and *thinned normalized random measures* (TNRM). We start by defining $\mathcal{R} := \{1, \ldots, R\}$ for some positive integer R. We refer to the elements of $\mathcal{R}$ as regions, and define a collection of independent CRMs $\tilde{\mu}_r$ with Lévy intensity $\nu_r(w, \theta)$ for each $r \in \mathcal{R}$. At a high level, for each time t, our approach involves transforming and combining the CRMs $\tilde{\mu}_r$ (the nature of the transformation differing for MNRM and TNRM)[2]. This forms a new CRM $\tilde{\mu}_t$ at each t, which is then normalized to give probability measure μ_t. The shared regions make the CRMs $\{\tilde{\mu}_t\}$, and thus the NRMs $\{\mu_t\}$ dependent. In the following, we detail the operations used in the constructions.

3.1. Mixed Normalized Random Measures

In our first construction, the CRM at time t is a weighted combination of the independent CRMs $\tilde{\mu}_r$. Let q_{rt} be a nonnegative weight between region r and time t. We define μ_t simply as follows:

$$\mu_t(\mathrm{d}\theta) = \frac{1}{\tilde{\mu}_t(\Theta)}\tilde{\mu}_t(\mathrm{d}\theta), \quad \tilde{\mu}_t(\mathrm{d}\theta) = \sum_{r=1}^{R} q_{rt}\tilde{\mu}_r(d\theta) \quad (3)$$

where $\tilde{\mu}_t(\Theta) = Z_t$ is the normalizing constant at time t. Note in particular that μ_t is a mixture of the individual region-specific NRMs μ_r, with mixing weights given by $q_{rt}\tilde{\mu}_r(\Theta)/\tilde{\mu}_t(\Theta)$. We then have:

Proposition 1 *Conditioned on the q_{rt}'s, each random probability measure μ_t defined in (3) is marginally distributed as a NRM with Lévy intensity $\sum_{r=1}^{R} \frac{1}{q_{rt}}\nu_r(w/q_{rt}, \theta)$.*

This result follows from the facts that 1) a scaled CRM is still a CRM, and 2) a sum of independent CRMs is still a CRM. See Appendix B for a detailed proof using characteristic functionals of the CRMs. In our experiments, we placed independent Gamma priors on the q_{rt}'s, and inferred their values from the data.

3.1.1. Comparison with related work

The spatial normalized Gamma process (SNGP) of (Rao & Teh, 2009) is a special case of MNRM, with the weights fixed to be binary (i.e. $q_{rt} \in \{0, 1\}$, with the actual value determined *a priori*). Our MNRM is thus a generalization of the SNGP, from a normalized gamma process to a general NRM, and from fixed and binary q_{rt}'s to arbitrary positive values that will be inferred along with the rest of the model. On the other hand, the SNGP imposes a spatial structure to the q_{rt}'s which may allow better generalization.

[1] CRMs can also have random weights at fixed locations, like most work, we ignore this component.

[2] We emphasize that we use r to index the independent NRMs μ_r and t to index the dependent NRMs μ_t.

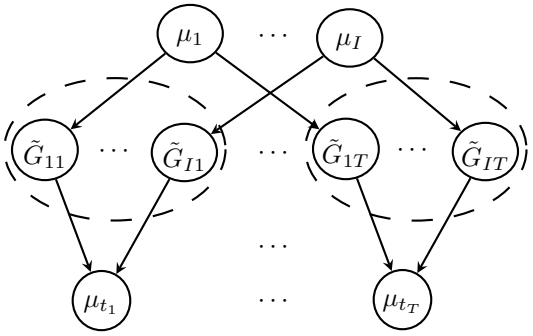

Figure 1. Construction of dependent normalized measure from R independent NRMs μ_r. In MNRM, $\tilde{G}_{rt}$ represents $q_{rt}\tilde{\mu}_r(\mathrm{d}\theta)$ defined in (3); while in TNRM it represents $\sum_{(\mathbf{b}\ s.t.\ b_t=1)} \tilde{G}_{r\mathbf{b}}(\mathrm{d}\theta)$ as defined in (5).

3.2. Thinned Normalized Random Measures

In our previous construction, a set of weights controlled the contribution of a set of CRMs to the NRM at any time, thus forming a 'softening' of (Rao & Teh, 2009) (where each of the CRMs is either present or absent). Our second construction is a different generalization of (Rao & Teh, 2009); rather than including or excluding entire CRMs, we control whether or not individual atoms in each of the CRMs are present in the NRM at a given time. More precisely, to each region-time pair (r, t) we associate a parameter q_{rt} taking values in $[0, 1]$.[3] q_{rt} is the subsampling rate of the atoms in region r for time t, with each atom of region r independently assigned to time t with probability q_{rt} (otherwise it is *thinned*). We call the resulting NRMs *thinned normalized random measures* (TNRM). Define a countably infinite sequence of Bernoulli variables $(z_{rt1}, z_{rt2}\cdots)$ for each region-time pair:

$$z_{rtk} \sim \text{Bernoulli}(q_{rt}) \quad k = 1, 2, \cdots$$

Then, the probability measure at time t is given by

$$\mu_t(\mathrm{d}\theta) = \frac{1}{\hat{\mu}_t(\Theta)}\hat{\mu}_t(\mathrm{d}\theta), \quad \hat{\mu}_t(\Theta) = \sum_{r=1}^{R}\sum_{k=1}^{\infty} z_{rtk} w_{rk} \quad (4)$$

Again, we can show the μ_t's are marginally NRMs:

Proposition 2 *Conditioned on the set of q_{rt}'s, each random probability measure μ_t defined in (4) is marginally distributed as a normalized random measure with Lévy measure $\sum_r q_{rt}\nu_r(\mathrm{d}w, \mathrm{d}\theta)$.*

The intuition behind this result is that independently thinning the atoms of a CRM maintains the property of complete randomness. Thus, $\hat{\mu}_t$ is a CRM, and μ_t, which is obtained by normalizing it is an NRM. For a formal proof, see Appendix B.

[3] Note that this q_{rt} is different from that in MNRM.

3.2.1. Comparision with related work

The idea of thinning atoms is similar to (Lin et al., 2010) for DPs and to (Chen et al., 2012) for NGGs, but these were restricted to random probability measures with chain-structured dependence. In addition, posterior samplers developed in these prior works were approximate. The TNRM is also a generalization of a very recent work (Lin & Fisher, 2012). This model is restricted to dependent DPs, and again, the proposed sampler has an incorrect equilibrium distribution (more details in Section 4.2 and Appendix E). The TNRM is also related to an unpublished report by (Foti et al., 2012), where they focus on different thinning constructions of dependent CRMs. Our focus is on NRMs; the normalization provides additional challenges. Their posterior inference is also approximate, being based on truncated representations of the CRMs (which are restricted only to Beta and Gamma CRMs). Finally, the TNRM can be viewed as an alternative to the IBP compound Dirichlet Process (Williamson et al., 2010). These are finite dimensional probability measures constructed by selecting a finite subset of an infinite collection of atoms (via the Indian buffet process (IBP)). Our model allows this to be infinite, allowing it to be used as a convenient building block in deeper hierarchical models. By treating the atoms present at each time as features, we can contrast the TNRM with the Indian buffet process (Griffiths & Ghahramani, 2011): in addition to allowing an infinite number of possible features, TNRM allows the number of active features to display phenomena like power-law behaviour; this is not possible in the IBP (Teh & Gorur, 2009; Broderick et al., 2012).

3.2.2. Interpretation as Mixture of NRMs

It is possible to represent the TNRM construction as a mixture of NRMs. Associate the kth atom in a region $r \in \mathcal{R}$ with a binary vector $\mathbf{b}^r(k)$ of length T. $b^r_t(k) = 1$ means this atom is inherited by the NRM μ_t of time t (i.e. $z_{trk} = 1$). Accordingly, we can split each region r into 2^T further subregions, each associated with atoms with a particular configuration of $\mathbf{b}^r$. It is easy to see that with subregion $\mathbf{b}^r = b^r_1 \cdots b^r_t$ of region r is associated a CRM $\tilde{G}_{r\mathbf{b}}$ with Lévy measure $\prod_{t=1}^{T} q_{rt}^{b_t}(1 - q_{rt})^{1-b_t}\nu_r(\mathrm{d}w, \mathrm{d}\theta)$, so that

$$\mu_t(\mathrm{d}\theta) \propto \hat{\mu}_t(\mathrm{d}\theta) = \sum_{r\in\mathcal{R}} \sum_{(\mathbf{b}\ s.t.\ b_t=1)} \tilde{G}_{r\mathbf{b}}(\mathrm{d}\theta) \quad (5)$$

Thus, the NRM at any time t can be expressed as a mixture of a number of NRMs $G_{r\mathbf{b}}(\mathrm{d}\theta) =$

$\tilde{G}_{r\mathbf{b}}(\mathrm{d}\theta)/\tilde{G}_{r\mathbf{b}}(\Theta)$, this number being exponential in the number of times T. We can also see from this interpretation that TNRMs can be seen as fixed-weight (binary) MNRMs but with many more regions (2^T). The number of components also grows linearly with the number of regions R; we will see that this flexibility improves the performance of our model without too great an increase in complexity.

4. Posterior Inference

In the following, we consider a specific NRM viz. the *normalized generalized Gamma process* (NGG)[4] to demonstrate posterior inference, generalization to other NRMs is straightforward. The generalized Gamma process (GGP) is a CRM whose Lévy measure is given by $\nu(\mathrm{d}w, \mathrm{d}\theta) = \frac{\sigma M}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-w} \mathrm{d}w H(\theta) \mathrm{d}\theta$, where $0 < \sigma < 1$ is known as the *index parameter*, $M \in \mathbb{R}^+$ is the *mass parameter*, and $H(\cdot)$ is the base probability density. Normalizing this CRM gives a flexible class of NRMs called NGGs, which includes the DP as a special case, and is preferable in applications where one wishes to place less informative priors on the number of clusters, power-law distributions on the cluster sizes etc. Its flexibility comes without a loss of computational tractability: the NGG is a so-called Gibbs-type prior, whose partition probability function (the clustering probability with the RPM integrated out) has a convenient closed form that generalizes the Chinese restaurant process (CRP) (see Appendix A.3). A consequence of this is that marginal samplers are available for both MNGG and TNGG. However, we saw in the previous section that the number of mixture components for TNGG grows exponentially with the number of times, and this can make the marginal sampler impractical. Consequently, we also develop slice samplers that instantiate the underlying RPMs. Since the marginal sampler for TNGG is impractical in most cases, and since the slice sampler for MNGG is an easy modification of the more complex slice sampler for TNGG, we move their descriptions to the appendix.

In the following, $\mathcal{T}$ denotes the set of times with observations, n_{trk} denotes the number of observations from time t associated with the k-th atom in region r. The superscript in $n_{trk}^{\backslash tl}$ indicates the previous count excluding the lth observation at time t. s_{tl} indexes the atom to which observation l at time t is attached, and g_{tl} indexes the corresponding region. Dots in the subscript denote sums over the corresponding index, for example, $n_{\cdot rk} = \sum_t n_{trk}$. For simplicity, denote $N_t = n_{t\cdots}$ $F(x|\theta)$ is the likelihood function.

[4] We use dNGG, MNGG and TNGG for dNRM, MNRM and TNRM.

4.1. The marginal sampler for MNGG

For this sampler, we follow (James, 2005) and introduce a set of auxiliary variables u_t $\forall t \in \mathcal{T}$, each conditionally distributed as a Gamma distribution with shape parameter N_t, and inverse scale parameter Z_t:

$$p(u_t|Z_t, \text{others}) \propto u_t^{N_t - 1} \exp(-u_t Z_t) \tag{6}$$

Integrating out u_t gives us $(1/Z_t)^{N_t}$, exactly the normalization constant corresponding to N_t independent draws from μ_t (see equation (3)). Thus the variables u_t allow us to move the normalization constants Z_t from the denominator to the exponent. Now, integrating out the w's and Z_t of each μ_t involves not much more that looking up the characteristic functional of a CRM, and it is not hard to see that

$$p(u_t|\text{others}) \propto \frac{u_t^{N_t-1} \exp\left\{-\sum_r M_r \left(1 + \sum_{t'} q_{rt'} u_{t'}\right)^{\sigma}\right\}}{\prod_r \left(1 + \sum_{t'} q_{rt'} u_{t'}\right)^{n_{\cdot r \cdot} - \sigma K_r}}$$

If we set $v_t = \log(u_t)$ we get a log-concave function, allowing easy sampling of the u_t's with a slice sampler (Neal, 2003). Additionally, conditioned on the u_t's, the cluster (and simultaneously, region) assignment of each observation can be sequentially resampled by a generalization of the CRP:

$$p(s_{tl} = k, g_{tl} = r|\text{others}\}) \propto \tag{7}$$
$$\begin{cases} \frac{q_{rt}(n_{\cdot rk}^{\backslash tl} - \sigma)}{1 + \sum_{t'} q_{rt'} u_{t'}} F_{rk}^{\backslash tl}(x_{tl}), & \text{if } k \text{ already exists,} \\ \sigma \left(\sum_{r'} \frac{M_{r'}}{\left(1 + \sum_{t'} q_{r't'} u_{t'}\right)^{1-\sigma}} \right) \int_\Theta F(x_{tl}|\theta) H(\theta) \mathrm{d}\theta \ , \end{cases}$$

where $F_{rk}^{\backslash tl}(x_{tl}) = \int F(x_{tl}|\theta_{rk}) \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} F(x_{t'l'}|\theta_{rk}) H(\theta_{rk}) \mathrm{d}\theta_{rk} / \int \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} F(x_{t'l'}|\theta_{rk}) H(\theta_{rk}) \mathrm{d}\theta_{rk}$ is the conditional density. Sampling the other variables is easy, see Appendix C.1.1 for details.

4.2. The slice sampler for TNGG

Our second sampler is a conditional sampler that instantiates (rather than integrates out) the underlying RPM. Recall that the Lévy measure in region r is $\nu_r(\mathrm{d}w, \mathrm{d}\theta)$, and that q_{rt} is the subsampling rate for atoms from this region for time t. In addition to the atoms to which observations are assigned (call this set W), we must also consider the remaining infinite atoms of each ν_r. The following result, involving the same auxiliary variables u_t as before, tells us that these atoms are distributed as independent Lévy processes:

Proposition 3 *Given observations associated with weights W, and auxiliary variables u_t for each $t \in \mathcal{T}$,*

the remaining weights in region r are independent of W, and are distributed as a CRM with Lévy measure

$$\nu'_r(\mathrm{d}w, \mathrm{d}\theta) = \prod_t \left(1 - q_{rt} + q_{rt}e^{-u_t w}\right) \nu_r(\mathrm{d}w, \mathrm{d}\theta) \ .$$

Proof See Appendix B, building on (James, 2005).

Remark Proposition 3 indicates that conditioned on observations, the remaining weights are distributed as a CRM from a different class than the original. The marginal samplers in (Lin et al., 2010; Lin & Fisher, 2012) implicitly assume these are the same, and are incorrect. We elaborate on this in the appendix.

Now, we deal with the infinite atoms associated with ν'_r. We follow (Griffin & Walker, 2011), and introduce auxiliary slice variables v_{tl} for each observation x_{tl}. If x_{tl} is assigned to atom s_{tl} in region g_{tl}, then v_{tl} is defined to be uniformly distributed in $[0, w_{g_{tl}s_{tl}}]$. Consequently, conditioned on v_{tl}, x_{tl} can only be assigned to atoms with weights greater than $w_{g_{tl}s_{tl}}$. From the properties of the NGG, this is a finite set, reducing posterior inference, conditioned on the v's, to that of a finite mixture model. Thus, at each iteration, for each region r, we only have to simulate atoms (w_{rk}, θ_{rk}) with weights larger than the smallest v_{tl} in that region (in fact, we simulate weights larger than a smaller number $\mathcal{L}_r$, as we explain later). As detailed in Appendix C.2.2, sampling the above variables (as well as indicators z_{rtk}) proceeds as follows:

- Jointly sample $\{(s_{tl}, g_{tl}) \ \ \forall\, t, l\}$ as:

$$p(s_{tl} = k, g_{tl} = r | \text{others}) \propto 1(w_{rk} > v_{tl}) 1(z_{rtk} = 1) F(x_{tl}|\theta_{rk}) \tag{8}$$

- Sample v_{tl} uniformly on $(0, w_{g_{tl}s_{tl}}]$:

$$v_{tl}|\text{others} \sim \text{Uniform}(0, w_{g_{tl}s_{tl}}) \tag{9}$$

- In each region r, we sample two kinds of w_{rk}'s, those associated with observations (the set W), and those larger than the $\mathcal{L}_r$ (call this set W^c).
  - $w_{rk} \in W$: these are Gamma distributed as

$$w_{rk}|\text{others} \sim \text{Gamma}\left(n_{\cdot rk} - \sigma, 1 + \sum_t z_{rtk} u_t\right) ,$$

  - $w_{rk} \in W^c$: From Proposition 3, these are distributed as a finite Poisson process on $[\mathcal{L}_r, \infty) \times \Theta$ with intensity $\prod_t (1 - q_{rt} + q_{rt}e^{-u_t w})\nu_r(\mathrm{d}w, \mathrm{d}\theta)$. We do this by thinning samples of a Poisson process whose intensity is pointwise larger that this intensity, for efficiency, we adaptively construct this upper-bound following (Favaro & Teh, 2012).
- Sampling z_{rtk}: z_{rtk}'s are Bernoulli variables, they equal to 1 with probability 1 if $n_{trk} > 0$, otherwise this probability is equal to $\frac{q_{rt}e^{-u_t w_{rk}}}{1 - q_{rt} + q_{rt}e^{-u_t w_{rk}}}$.

The remaining variables (M_r, u_t and q_{rt}), can be sampled exactly using the pseudo-marginal Metropolis-Hastings algorithm (Andrieu & Roberts, 2009), details in Appendix C.2.2. However by setting $\mathcal{L}_r$ to a small value, we can sample from an accurate approximation and gain significant computational savings. In Appendix C.2, we describe the exact sampler, perform a bound analysis on the approximation, and derive the approximate update rules listed below:

- M_r: the posterior of M_r is a Gamma distribution with shape parameter $K'_r + 1$ and rate parameter $\xi_\sigma \kappa_r + \zeta_\sigma \mathcal{L}_r^{1-\sigma} \sum_t q_{rt} u_t$, where K'_r is number of jumps larger than the threshold $\mathcal{L}_r$, $\kappa_r = \int_{\mathcal{L}_r}^{\infty} x^{-1-\sigma} e^{-x} \mathrm{d}x$ and $\zeta_\sigma = \frac{\xi_\sigma}{(1-\sigma)}$.
- u_t: the posterior of u_t is also Gamma with shape parameter N_t and rate parameter $\sum_r \sum_k z_{rtk} w_{rk} + \zeta_a \sum_r q_{rt} M_r \mathcal{L}_r^{1-\sigma}$.
- q_{rt}: the posterior of q_{rt} is approximately a Beta distribution with the two shape parameters as $\sum_k 1(z_{rtk} = 1) + a_q$ and $\sum_k 1(z_{rtk} = 0) + b_q$ respectively, where a_q and b_q are hyperparameters for the Beta prior of q_{rt}.

We did not find any significant difference in accuracy between this and the true sampler, although the computational benefits were significant.

5. Experiments

In the following, we applied our ideas to modelling text documents organized in time. We focused on six models: MNGG, TNGG, HMNGG, HMNGP, HTNGG and HSNGG. The first two are based on the mixed and thinned constructions respectively, with each document is assigned to its own 'time', thus TNGG resembles focused topic models (Williamson et al., 2010). On one hand, this disregards statistical information that might be shared across documents from the same true time period, on the other hand, this affords more flexibility, since each document can have its own set of q_{rt} parameters. Letting G be the Dirichlet distribution, F the multinomial distribution, and t span all documents in the corpus, the generative process is as follows:

$$(\mu_t) \sim dNGG(\sigma_0, M_0, G, \{q_{rt}\}) \tag{10}$$

$$\theta_i^t \sim \mu_t, \qquad x_i^t|\theta_i^t \sim F(\cdot|\theta_i^t) \ , \tag{11}$$

where $dNGG(\sigma_0, M_0, G, \{q_{rt}\})$ denotes the dependent NGG constructed via MNGG or TNGG with index

parameter σ, mass parameter M_0, base distribution G and the set of weights/subsampling rates $\{q_{rt}\}$.

The remaining models specify the organization of documents into time-periods by adding another layer to the hierarchy. In particular, we used our dNGG constructions to produce an RPM μ_t for each time-period t; each document in time period t then had a distribution over topics drawn from an NGG with base-measure μ_t, replacing (11) by

$$\begin{aligned} \{\mu_{ti}\}|\mu_t &\sim NGG(\sigma, M, \mu_t) \\ \theta^t_{ij} \sim \mu_{ti}, \quad x^t_{ij}|\theta^t_{ij} &\sim F(\cdot|\theta^t_{ij}) \ , \end{aligned} \tag{12}$$

Both HMNGG and HTNGG follow this construction, with the dependent NGGs produced by mixing and thinning respectively. HMNGP is the same as HMNGG but with the NGG replaced with a Gamma process (GP). HSNGG denotes the spatial normalized generalized Gamma process (Rao & Teh, 2009), a special case of HMNGG with $q_{rt} \in \{0, 1\}$. We also compare our models with the popular hierarchical Dirichlet process (HDP), furthermore, we generalize the HDP to the HNGG (Appendix D), where the construction is the same as HDP but using NGGs instead of DPs.

5.1. Synthetic data

In our first experiment, we generated 3000 observations from a hierarchical Pitman-Yor topic model (Du et al., 2010). We set the vocabulary size to 100, and used the following generative process:

$$G_0 \sim \mathcal{PY}(\alpha_0, d_0, G), G_t \sim \mathcal{PY}(\alpha_t, d_t, G_0) \quad t = 1, 2, 3$$
$$\theta_{tj} \sim G_t, \quad x_{tj} \sim F(\cdot|\theta_{tj}) \quad j = 1, \cdots, 3000$$

The base measure G over topic distributions was a 100-dimensional symmetric Dirichlet with parameter 0.1, while $F(\cdot|\theta)$ is the 100-dimensional discrete distribution. The concentration parameters $\alpha_i, i = 0, \cdots, 3$ were set to $1, 3, 4$ and 5 respectively, while all discount parameters d_i were set to 0.5. Following the generative process described above, we then split the data at each time into 30 documents of 100 words each, and modelled the resulting corpus using the HMNGG and HTNGG described in (12). The Pitman-Yor process (which is not an NRM) exhibits a power-law behavior, and the purpose of this experiment is to demonstrate the flexibility of the NGG over the DP. Accordingly, we compare the performance of HMNGG and HTNGG on this dataset against their dDP equivalences, the HMNGP and the HTNGP (obtained by replacing the generalized Gamma process with the Gamma process in the constructions). We set the number of regions equal to the number of times, and sampled all the model parameters (placing Gamma$(0.1, 0.1)$ priors on all scalars in $\mathbb{R}^+$, and Beta$(0.5, 0.5)$ priors on all scalars in $[0, 1]$).

We plot the predictive likelihood on a 20% held-out dataset in Figure 2. We see that both HMNGG and HTNGG outperform their non-power-law variants HMNGP and HTNGP in terms of predictive likelihoods. The inferred parameter σ is around 0.2 (a value of 0 recovers the Gamma process). Furthermore, HTNGG gets higher likelihoods than HMNGG, in this case, this follows from the added flexibility afforded by allowing the thinning of individual atoms.

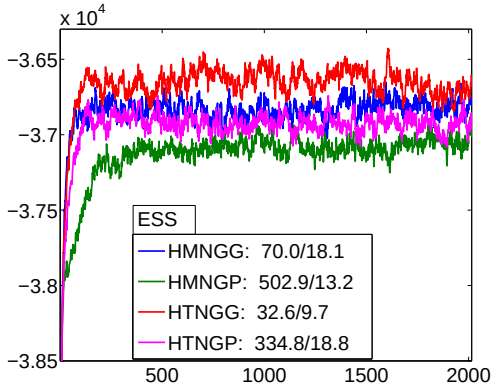

Figure 2. Power-law distribution modeling with different models. HMNGG and HTNGG have higher likelihoods than their non-power-law versions.

5.2. Topic Modelling

Datasets Next, we considered four real-world document datasets, viz. ICML, TPAMI, Person and NIPS. The first 2 corpora consisted of abstracts obtained from the ICML and PAMI websites; ICML contained 765 documents from 2007-2011 with a total of about 44K words, and a vocabulary size of about 2K; TPAMI had 1108 documents from 2006-2011, with total of 91K words and vocabulary size of 3K. The Person dataset was extracted from Reuters RCV1 using the query *person* under Lucene,and contained 8616 documents, 1.55M words and a vocabulary size of 60K. It spanned the period 08/96 to 08/97. The NIPS corpus consisted of proceedings over the years 1987 to 2003 (Globerson et al., 2007). It was not postprocessed, and has 2483 documents, 3.28M words and vocabulary size 14K.

Parameter Setting and Evaluation In modelling these datasets, for MNGG and TNGG (where we disregard the years associated with each document and assign it to its own time), we set the number of regions to be 20; in the other models these were set equal to the number of years. The Dirichlet base distribution was symmetric with parameter 0.3, and as in the previous section, weak Gamma and Beta priors were placed appropriately on all nonnegative scalars.

To evaluate the models, we computed perplexity scores

Table 1. Training perplexities and test perplexities for different models on ICML, TPAMI, Person and NIPS datasets.

Datasets	ICML		TPAMI		Person		NIPS	
Models	train	test	train	test	train	test	train	test
HDP	580 ± 6	1017 ± 8	671 ± 6	1221 ± 6	4541 ± 33	5962 ± 43	1813 ± 27	1956 ± 18
HNGG	575 ± 5	1057 ± 8	671 ± 6	1262 ± 11	4565 ± 60	5999 ± 54	1713 ± 13	1878 ± 11
TNGG	681 ± 23	1071 ± 6	701 ± 38	1327 ± 3	5815 ± 122	7981 ± 36	2990 ± 57	3231 ± 2
MNGG	569 ± 6	1056 ± 9	644 ± 6	1272 ± 12	4560 ± 63	6013 ± 66	1612 ± 3	1920 ± 5
HSNGG	550 ± 5	1007 ± 8	643 ± 3	1237 ± 22	4324 ± 77	5733 ± 66	1406 ± 5	1679 ± 8
HTNGG	572 ± 7	$\mathbf{945 \pm 7}$	642 ± 4	$\mathbf{1174 \pm 9}$	4196 ± 29	5527 ± 47	1377 ± 5	1635 ± 3
HMNGG	$\mathbf{535 \pm 6}$	1001 ± 10	$\mathbf{608 \pm 4}$	1199 ± 10	$\mathbf{4083 \pm 36}$	$\mathbf{5488 \pm 44}$	$\mathbf{1366 \pm 8}$	$\mathbf{1618 \pm 5}$
HMNGP	561 ± 10	995 ± 14	634 ± 10	1208 ± 8	4118 ± 45	5519 ± 41	1370 ± 3	1634 ± 4

on a held-out test dataset. In all cases, 20% of the original data sets was held-out, following the standard dictionary hold-out method (50% of the held-out documents was used to estimate topic probabilities) (Rosen-Zvi et al., 2004). Test perplexity was calculated over 10 repeated runs with random initialization, we report mean values and standard deviations. In each run 2000 cycles were used as burn-in, followed by 1000 cycles to collect samples for perplexity calculation. To avoid complications resulting from the different representations used by the marginal and slice sampler, we calculated perplexities after first transforming the representation of the slice sampler to those in the marginal sampler. In other words, given the state of the slice sampler, we determined the induced partition structure, and used this to calculate prediction probabilities (calling the same piece of code).

Quantitative comparison for different models We calculated both training and test perplexities for the models specified above, these are shown in Table 1. We see that HMNGG and HTNGG perform best, achieving significant lower perplexities than the others. While HTNGG is more flexible than HMNGG, it performs sightly worse when the datasets becomes large; this is more obvious when comparing MNGG and TNGG. Part of the reason for this is the complex posterior structure for the thinned models, so that the samplers are often stuck in local optima, resulting in much worse perplexities. Interestingly, HMNGP (without the power-law property) does not perform much worse than HMNGG, indicating topic distributions in topic models might not follow an obvious power-law behavior. This coincides with the sampled value of the index parameter σ (around 0.01). Thus it is not surprising that HDP is comparable to HNGG: slightly better in small datasets, but a bit worse in large datasets. Moreover, the simple MNGG and TNGG do much worse than HMNGG and HTNGG, emphasizing the importance of statistical information shared across documents in the same year.

Topic evolutions Figure 3 is a posterior sample, showing the evolution of 12 randomly selected topics on the NIPS dataset for HMNGG and HTNGG. In all cases, we calculated the proportion of words assigned to the topic k in region r at each time t (i.e. $\frac{n_{trk}}{n_{tr\cdot}}$), and the predictive probabilities for each topic at each time. The latter is defined for MNGG to be proportional to $\frac{q_{rt}(n_{\cdot rk}^{\setminus tl} - \sigma)}{1+\sum_{t'} q_{rt'} u_{t'}}$ (see equation 7), and for TNGG to be proportional to $q_{rt} w_{rk}$ (see equation 8) by integrating out v_{tl} and z_{rtk}. We see (as we expect) HMNGG generating smoother topic proportions over time (topics in HTNGG can die and then be reborn later because of the thinning mechanism).

Marginal vs slice sampler We next compare the performance of the marginal and slice samplers for MNGG, HMNGG and HTNGG. The marginal sampler for TNGG could not handle datasets with more than even 2 times. Instead, we had to divide each dataset into two times (the first and the second halves, call the resulting datasets as 2-*time* datasets), and treat these as the only covariates available. We emphasize that we do this only for a comparison with our slice sampler, which can handle more complex datasets (the slice sampler was used in the previous sections). Table 2 shows the average effective sample sizes and running times over 5 repeated runs for the two samplers on the original datasets and the 2-time datasets. On the original datasets, in MNGG, the marginal sampler generally obtains larger ESS values than the slice sampler; while it is opposite for HMNGG. Regarding to the running time, the marginal sampler is more efficient in small datasets (*i.e.*, ICML and TPAMI), while they are comparable in the other datasets. The reason for this is that in small datasets, a large amount of the running time in the slice sampler was used in sampling the extra atoms (which is unnecessary in the marginal sampler), while in large datasets, the time for sampling word allocations starts to become significant. In the 2-time datasets, we observe that the slice

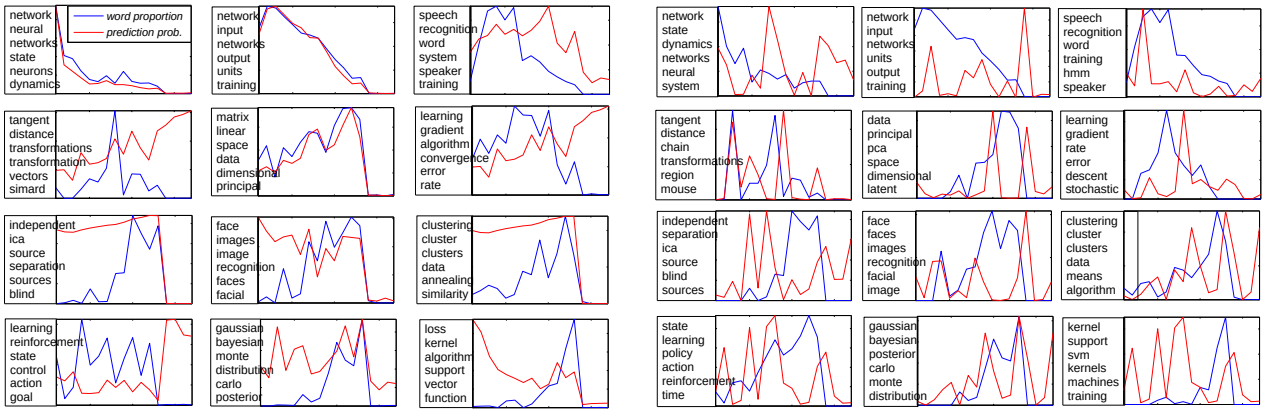

Figure 3. Topic evolutions on NIPS dataset for 12 randomly chosen topics learned by HMNGG (left) and HTNGG (right), respectively. The two curves correspond to word proportions within each topic (blue) and prediction probabilities (red) for each time. HTNGG tends to produce less smooth topic proportions over time.

Table 2. Comparison of effective sample sizes and run times for marginal and slice sampler (subscript s). Subscript *2* in the datasets means the 2-time datasets. over 5 repeated runs. $a/b/c \mid t$ in the table means the average ESS among all the chosen statistics is a, the median is b, the minimum is c, and the running time for the 1000 cycles is t.

	ICML	TPAMI	Person	NIPS
Models	ESS | Time	ESS | Time	ESS | Time	ESS | Time
MNGG	243.3/202.5/4.4|234s	252.4/231.9/3.7| 285s	402.5/401.4/1.5|1.5h	314.8/376.1/1.5|3.3h
MNGG$_s$	201.2/122.0/26.9|760s	205.1/131.9/23.5|813s	321.5/291.8/11.3|2.9h	228.4/110.6/2.2|2.2h
TNGG$_s$	115.2/90.0/4.5|555s	135.7/113.0/11.1|592s	300.6/231.3/3.2|3.3h	223.8/107.7/1.1|1.4h
HMNGG	99.1/70.3/2.6|91s	171.5/80.4/5.1|176s	213.0/246.5/1.9|3.3h	282.1/198.2/4.3|9.4h
HMNGG$_s$	150.7/117.7/4.6|97s	194.3/180.9/6.5|227s	293.3/358.6/2.0|3.5h	346.1/467.2/1.7|10.4h
HTNGG$_s$	82.8/80.1/4.7|126s	92.5/105.1/5.4|312s	184.9/226.3/6.1|4.1h	225.4/210.2/3.4|11.9h
	ICML$_2$	TPAMI$_2$	Person$_2$	NIPS$_2$
HMNGG	57.4/52.5/7.3|66s	59.4/56.3/6.7|89s	119.4/102.0/3.1|1.0h	111.1/73.8/3.3|1.5h
HMNGG$_s$	125.4/112.5/15.0|69s	142.0/125.6/10.6|91s	212.9/212.0/5.9|1.1h	205.2/203.0/5.5|1.9h
HTNGG	50.3/46.9/3.0|71s	55.3/58.4/4.3|95s	144.8/170.6/4.2|1.3h	119.1/130.0/2.8|2.3h
HTNGG$_s$	94.9/90.9/4.0|76s	116.0/107.8/3.4|106s	153.2/113.5/2.7|1.1h	176.1/151.0/3.3|1.9h

sampler obtains larger ESS values than its marginal sampler in both HMNGG and HTNGG, with comparable running times. We repeat that for HTNGG, the slice sampler is applicable for any number of times, while the marginal sampler is computationally infeasible even for a moderately large number of times.

6. Conclusion

We proposed two classes of dependent normalized random measures for the nonparametric modeling of dependent probability measures, the *mixed normalized random measure* and the *thinned normalized random measure*. Our construction involves weighting and thinning independent CRMs, before combining and normalizing them. We developed two different MCMC algorithms for posterior inference, a marginal and a slice sampler. In our experiments, our models showed significantly superior performance compared to related dependent nonparametric models such as HDP and SNGP, with the simpler MNRM performing better on complex data. We also find the slice sampler generally mixes better than the marginal sampler in both models. Interesting future work includes extending our framework to allow each atom to have its own thinning probability, as well as allowing marginal RPMs in the broader class of *Poisson-Kingman processes*, which includes the Pitman-Yor process as a special case. Moreover, our models can be applied not just for time-series in topic modelling but also to allow sparsity of probabilities, for instance (Williamson et al., 2010).

Acknowledgments: NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Center of Excellence program. VR was funded by DARPA MSEE. YWT was funded by the Gatsby Charitable Foundation.

References

Ahmed, A. and Xing, E. P. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process. In *SDM*, 2008.

Andrieu, C. and Roberts, G. O. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725, 2009.

Broderick, T., Jordan, M. I., and Pitman, J. Beta processes, stick-breaking, and power laws. *Bayesian Anal.*, 7:439–476, 2012.

Caron, F., Davy, M., and Doucet, A. Generalized Polya urn for time-varying Dirichlet process mixtures. In *UAI*, 2007.

Chen, C., Ding, N., and Buntine, W. Dependent hierarchical normalized random measures for dynamic topic modeling. In *ICML*. 2012.

Du, L., Buntine, W., and Jin, H. A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Mach. Learn.*, 81:5–19, 2010.

Favaro, S. and Teh, Y. W. MCMC for normalized random measure mixture models. *Stat. Sci.*, 2012.

Foti, N. J., Futoma, J., Rockmore, D., and Williamson, S. A. A unifying representation for a class of dependent random measures. Technical Report arXiv:1211.4753, Dartmouth College and CMU, USA, 2012. URL http://arxiv.org/abs/1211.4753.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.*, 100(471):1021–1035, 2005.

Globerson, A., Chechik, G., Pereira, F., and Tishby, N. Euclidean embedding of co-occurrence data. *JMLR*, 8:2265–2295, 2007.

Griffin, J. E. and Steel, M. F. J. Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:179–194, 2006.

Griffin, J. E., Kolossiatis, M., and Steel, M. F. J. Comparing distributions using dependent normalized random measure mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 2012.

Griffin, J.E. and Walker, S.G. Posterior simulation of normalized random measure mixtures. *J. Comput. Graph. Stat.*, 20(1):241–259, 2011.

Griffiths, T. L. and Ghahramani, Z. The Indian buffet process: An introduction and review. *JMLR*, 12: 1185–1224, 2011.

James, L. F. Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Ann. Statist.*, 33(4):1771–1799, 2005.

James, L.F., Lijoi, A., and Prünster, I. Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.*, 36:76–97, 2009.

Lin, D., Grimson, E., and Fisher, J. Construction of dependent Dirichlet processes based on Poisson processes. In *NIPS*. 2010.

Lin, D. H. and Fisher, J. Coupling nonparametric mixtures via latent Dirichlet processes. In *NIPS*. 2012.

MacEachern, S. Dependent nonparametric processes. In *Proc. of the SBSS*, 1999.

MacEachern, S.N., Kottas, A., and Gelfand, A.E. Spatial nonparametric Bayesian models. In *Proc. of the 2001 Joint Statistical Meetings*, 2001.

Neal, R. M. Slice sampling. *Ann. Statist.*, 31(3):705–767, 2003.

Rao, V. and Teh, Y. W. Spatial normalized Gamma processes. In *NIPS*. 2009.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. The author-topic model for authors and documents. In *UAI*, 2004.

Srebro, N. and Roweis, S. Time-varying topic models using dependent Dirichlet processes. Technical report, University of Toronto, 2005.

Teh, Y.W. and Gorur, D. Indian buffet processes with power-law behavior. In *NIPS*. 2009.

Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101(476):1566–1581, 2006.

Williamson, S. A., Wang, C., Heller, K. A., and Blei, D. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*. 2010.