Supplementary material for Dependent Normalized Random Measures

Changyou Chen[1,3] CHANGYOU.CHEN@NICTA.COM.AU

[1]RSISE, Australian National University, Australia

Vinayak Rao[2] VRAO@GATSBY.UCL.AC.UK

[2]Dept. Statistical Science, Duke University, USA

Wray Buntine[3,1] WRAY.BUNTINE@NICTA.COM.AU

[3]National ICT, Canberra, Australia

YeeWhye Teh[4] Y.W.TEH@STATS.OX.AC.UK

[4]Dept. Statistics, University of Oxford, UK

Abstract

This is the supplementary material for the ICML 2013 paper *Dependent Normalized Random Measures* by the same authors.

A. Notation & Preliminary

We list some of the notation used in this paper in Table 1 for reminder.

A.1. Definitions

For completeness, we restate the definition of MNRM and TNRM. We are given a Poisson process on a product space $\mathbb{R}^+ \times \Theta \times \mathcal{R}$ with intensity measure $\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a)$ (we will use the notation $\nu_r(\mathrm{d}w, \mathrm{d}\theta) = \int_{\tilde{R}_r} \nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a)$), denote the corresponding Poisson random measure as $\mathcal{N}(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a)$, the constructions are then defined as follow:

Mixed Normalized Random Measures (MNRM)

$$
\begin{aligned}
\tilde{\mu}_r(\mathrm{d}\theta) &= \int_{\mathbb{R}^+ \times \tilde{R}_r} w\mathcal{N}(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a), && r = 1, \cdots, \#\mathcal{R} \\
\tilde{\mu}_t(\mathrm{d}\theta) &= \sum_{r=1}^{\#\mathcal{R}} q_{rt}\tilde{\mu}_r(d\theta) && t = 1, \cdots, T \\
\mu_t(\mathrm{d}\theta) &= \frac{1}{Z_t}\tilde{\mu}_t(\mathrm{d}\theta) \text{ , where } Z_t = \tilde{\mu}_t(\Theta)
\end{aligned} \tag{1}
$$

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

Table 1. List of notation.

Notation	Description
I or $\#\mathcal{R}$	#regions for $\mathcal{R}$, indexed by r
T	#times for the observations
L_t	#observations in time t
s_{tl}	latent variable indexes which atom the l-observation in time t belongs to.
g_{tl}	latent variable indexes which region the l-observation in time t belongs to.
(w_{rk}, θ_{rk})	points/atoms in the Poisson process in region $\tilde{R}_r$. When used to construct an NRM, sometimes we also call w_{rk} as the jumps and θ_{rk} as the atoms
K_r	#atoms with observation in the NRM in region $\tilde{R}_r$
M_r	mass parameter for the NRM in region $\tilde{R}_r$
$\vec{X}_t$	observations in the NRM in time t
n_{trk}	#observations in time t attached to the k-th jump of the NRM in region $\tilde{R}_r$
N_t	total number of observations in time t
$n_{\cdot rk}$	$= \sum_t n_{trk}$
u_t	auxiliary variable for the NRM in time t
$F(\cdot\|\theta_{rk})$	likelihood function under atom θ_{rk}
$\mathcal{N}(w, \theta)$	a Poisson random measure on $\mathcal{W} \times \Theta$
$\nu(\mathrm{d}w, \mathrm{d}\theta)$	Lévy measure for the NRM on $\mathbb{R}^+ \times \Theta$, we assume it is decomposed as $\rho(\mathrm{d}w)H(\mathrm{d}\theta)$, where $H(\cdot)$ is a probability measure on Θ. We use $\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a)$ to denote the Lévy measure on the augmented space $\mathbb{R}^+ \times \Theta \times \mathcal{R}$ and is assume to be factorized as $\nu'(\mathrm{d}w, \mathrm{d}\theta)Q(\mathrm{d}a)$

Thinned Normalized Random Measures (TNRM)

$$\begin{aligned}
\tilde{\mu}_r(\mathrm{d}\theta) &= \int_{\mathbb{R}^+ \times \tilde{R}_r} w\mathcal{N}(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a), && r = 1, \cdots, \#\mathcal{R} \\
z_{rtk} &\sim \text{Bernoulli}(q_{rt}), && k = 1, 2, \cdots \\
\hat{\mu}_t(\mathrm{d}\theta) &= \sum_{k=1}^{\infty} z_{rtk} w_{rk} \delta_{\theta_{rk}}, && t = 1, \cdots, T \\
\mu_t(\mathrm{d}\theta) &= \frac{1}{Z_t}\hat{\mu}(\mathrm{d}\theta), \text{ where } Z_t = \tilde{\mu}_t(\Theta)
\end{aligned} \tag{2}$$

A.2. Preliminary Lemmas

We give three lemmas used in analyzing the properties and deriving the posteriors for the proposed MNRM, TNRM and their variants.

Lemma 1 below is a celebrated formula for Lévy processes know as the *Lévy-Khintchine formula*.

Lemma 1 (Lévy-Khintchine Formula) *Given a completely random measure $\tilde{\mu}$ (we consider the case where it only contains random atoms) constructed from a Poisson process on a produce space $\mathbb{R}^+ \times \Theta$ with intensity measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$. For any measurable function $f : \mathcal{W} \times \Theta \longrightarrow \mathbb{R}^+$, the following formula holds:*

$$\begin{aligned}
\mathbb{E}\left[e^{-\tilde{\mu}(f)}\right] &\triangleq \mathbb{E}\left[e^{-\int_\Theta f(w,\theta)\mathcal{N}(\mathrm{d}w,\mathrm{d}\theta)}\right] \\
&= \exp\left\{-\int_{\mathcal{W}\times\Theta}\left(1 - e^{-f(w,\theta)}\right)\nu(\mathrm{d}w, \mathrm{d}\theta)\right\},
\end{aligned} \tag{3}$$

where the expectation is taken over the space of bounded finite measures. Using (3), the characteristic functional of $\tilde{\mu}$ is given by

$$\varphi_{\tilde{\mu}}(u) \triangleq \mathbb{E}\left[e^{\int_\Theta iu\tilde{\mu}(\mathrm{d}\theta)}\right] = \exp\left\{-\int_{\mathcal{W}\times\Theta}\left(1 - e^{iuw}\right)\nu(\mathrm{d}w, \mathrm{d}\theta)\right\}, \tag{4}$$

where $u \in \mathbb{R}$ and i is the imaginary unit.

Lemma 2 is about the disintegration property of a Poisson random measure $\mathcal{N}$ and some fixed points $\theta_k \in \Theta$. This is a specific result derived using either the Poisson process partition calculus (James, 2005), or the well known Palm formula.

Lemma 2 *Let $\mathcal{N}$ be a Poisson random measure defined on $\mathbb{R}^+ \times \Theta$ with intensity measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$, $\tilde{\mu}$ be the CRM constructed from $\mathcal{N}$. Given samples $\{\varphi_n\}$ with ties $(\theta_k)_{k=1}^K$ and the corresponding counts $(n_1, \cdots, n_K)$, for any nonnegative function $f : \mathbb{R}^+ \times \Theta \mapsto \mathbb{R}^+$, the following formula holds:*

$$\mathbb{E}\left[e^{-\mathcal{N}(f)} \prod_{k=1}^{K} \tilde{\mu}(\theta_k)^{n_k}\right] = \mathbb{E}\left[e^{-\mathcal{N}(f)}\right] \prod_{k=1}^{K} \int_{\mathbb{R}^+} w_k^{n_k} e^{-f(w_k, \theta_k)} \nu(\mathrm{d}w_k, \theta_k) \ , \tag{5}$$

where $\mathcal{N}(f) = \int_{\mathbb{R}^+ \times \Theta} f(w, \theta) \mathcal{N}(\mathrm{d}w, \mathrm{d}\theta)$.

Lemma 3, originally from Proposition 2.1 of (James, 2005), gives the posterior intensity measure of the Poisson process under an exponential tilting operation. It is used in the proof of the posterior Lévy measure for MNRM and TNRM.

Lemma 3 *Let $\mathcal{N}$ denotes a Poisson random measure with intensity measure ν, taking values in space of boundedly finite measures $\mathcal{M}$ with sigma-field denoted as $\mathcal{B}(\mathcal{M})$. $BM_+(\mathcal{W})$ denotes the collection of Borel measurable functions of bounded support on $\mathcal{W}$. Then for each $f \in BM_+(\mathcal{W})$ and each g on $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$,*

$$\int_{\mathcal{M}} g(\mathcal{N}) e^{-\mathcal{N}(f)} P(\mathrm{d}\mathcal{N}|\nu) = \mathcal{L}_{\mathcal{N}}(f|\mathcal{N}) \int_{\mathcal{M}} g(\mathcal{N}) P(\mathrm{d}\mathcal{N}|e^{-f}\nu) \ ,$$

where $P(\mathrm{d}\mathcal{N}|e^{-f}\nu)$ is the law of a Poisson process with intensity $e^{-f(w)}\nu(\mathrm{d}w)$, $\mathcal{L}_{\mathcal{N}}(f|\mathcal{N}) = \exp\left\{-\int_{\mathcal{W}} \left(1 - e^{-f(w)}\right) \nu(\mathrm{d}w)\right\}$ denotes the Laplace functional of $\mathcal{N}$. In other words, exponential tilting of a Poisson random measure as $e^{-\mathcal{N}(f)} P(\mathrm{d}\mathcal{N}|\nu)$ is equivalent to dealing with a Poisson random measure with intensity $e^{-f}\nu$.

A.3. Normalized Generalized Gamma Processes

In this subsection we briefly introduce a special class of normalized random measures called the normalized generalized Gamma process (NGG), and list some of its well known properties. A NGG is defined by normalizing a generalized Gamma process (GGP), whose Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$ is defined on the produce space $\mathbb{R}^+ \times \Theta$ with the following form[1]:

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = \frac{\sigma M}{\Gamma(1-\sigma)} w^{-\sigma-1} e^{-w} \mathrm{d}w H(\theta) \mathrm{d}\theta \ , \tag{6}$$

where $0 < \sigma < 1$ is called the index parameter, $M \in \mathbb{R}^+$ is called the mass parameter, and $H(\cdot)$ is a probability measure on space Θ, called the base distribution. We will use $\mathrm{NGG}(\sigma, M, H(\cdot))$ to denote a NGG in the rest of the paper.

We give the Laplace functional and the marginal posterior of the NGG below. These results can be used in the following sections.

Lemma 4 (Laplace Functional of a GGP) *For a generalized Gamma process $\tilde{\mu}_g$ with Lévy measure defined in (6), let $f : \mathbb{R}^+ \longmapsto \mathbb{R}^+$ be a measurable function, the Laplace functional of $\tilde{\mu}_g$ is given by*

$$\mathcal{L}(f|\tilde{\mu}_g) \triangleq \mathbb{E}\left[e^{-\tilde{\mu}_g(f)}\right] = \exp\left\{-\int_{\mathcal{W}\times\Theta} \left(1 - e^{-f(w)}\right) \nu(\mathrm{d}w, \mathrm{d}\theta)\right\}$$
$$\xrightarrow{f(w) \triangleq uw} \exp\left\{-M\left((1+u)^\sigma - 1\right)\right\} \ ,$$

where $\tilde{\mu}_g(f) = \int_{\mathbb{R}^+ \times \Theta} f(w) \mathcal{N}(\mathrm{d}w, \mathrm{d}\theta)$, and $u > 0$ is a real constant.

[1] The Lévy measure of GGP can be formulated in different ways (Favaro & Teh, 2012), some via two parameters while some via three parameters, but they can be transformed to each other by using a change of variable formula. We only consider the form (6) in this paper for simplicity.

The following posterior result of a NGG is taken from (Corollary 2 Chen et al., 2012a), similar results can also be found in other references such as (James et al., 2009; Favaro & Teh, 2012).

Lemma 5 (Posterior of a NGG) *Let $\vec{X} = (x_1, \cdots, x_N)$ be samples from the NGG$(\sigma, M, H(\cdot))$ with distinct values (ties) $(x_1^*, \cdots, x_K^*)$ and the corresponding counts $(n_1, \cdots, n_K)$. Introduce a latent variable u (called* latent relative mass *(Chen et al., 2012a)), the marginal posterior is given by:*

$$p\left(\vec{X}, u, K \mid \sigma, M\right) = \frac{u^{N-1}}{\Gamma(N)(1+u)^{N-K\sigma}} (M\sigma)^K e^{M-M(1+u)^\sigma} \prod_{k=1}^{K} (1-\sigma)_{n_k-1} H(x_k^*) \ ,$$

where $(1-\sigma)_{n_k-1} = (1-\sigma)\cdots(n_k-1-\sigma)$ if $n_k > 1$, and 1 if $n_k \leq 1$.

B. Properties of MNRMs and TNRMs

We have the following property for the MNRM.

Proposition 6 (Proposition 1 in the main text) *Conditioned on the weights q_{rt}'s, each random probability measure μ_t defined in (1) is marginally distributed as a NRM with Lévy intensity $\sum_{r=1}^{\#\mathcal{R}} \nu_r(w/q_{rt}, \theta)/q_{rt}$.*

Proof First, from the definiton we have

$$\tilde{\mu}_t = \sum_{r=1}^{\#\mathcal{R}} q_{rt}\tilde{\mu}_r \ .$$

Because each $\tilde{\mu}_r$'s is a CRM, we have for any collection of disjoint subsets $(A_1, \cdots, A_n)$ of Θ, the random variables $\tilde{\mu}_r(A_n)$'s are independent. Moreover, since the $\tilde{\mu}_r$'s are independent, we have that $\{\tilde{\mu}_t(A_i)\}_{i=1}^n$ are independent. Thus $\tilde{\mu}_t$ is a completely random measure. To work out its Lévy measure, we calculate the characteristic functional of each random measure $q_{rt}\tilde{\mu}_r$ using Lemma 1:

$$\begin{aligned}\varphi_{q_{rt}\tilde{\mu}_r}(u) &= e^{-\int_{\mathbb{R}^+\times\Theta}(1-e^{iuq_{rt}w})\nu_r(w,\theta)\mathrm{d}w\mathrm{d}\theta}, \\ &= e^{-\int_{\mathbb{R}^+\times\Theta}(1-e^{iuw})\nu_r(w/q_{rt},\theta)\mathrm{d}w/q_{rt}\mathrm{d}\theta} \ ,\end{aligned}$$

where the last step follows by using a change of variable $w' = q_{rt}w$. Because $q_{rt}\tilde{\mu}_r$'s are independent, we have that the characteristic functional of $\tilde{\mu}_t$ is

$$\begin{aligned}\varphi_{\tilde{\mu}_t}(u) &= \prod_{r=1}^{\#\mathcal{R}} \varphi_{q_{rt}\tilde{\mu}_r}(u) \\ &= e^{-\int_{\mathbb{R}^+\times\Theta}(1-e^{iuw})\sum_{r=1}^{\#\mathcal{R}}\nu_r(w/q_{rt},\theta)\mathrm{d}w/q_{rt}\mathrm{d}\theta} \ ,\end{aligned} \tag{7}$$

The Lévy intensity of $\tilde{\mu}_t$ is thus $\sum_{r=1}^{\#\mathcal{R}} \nu_r(w/q_{rt}, \theta)/q_{rt}$.

■

The following two properties are proved for TNRMs.

Proposition 7 (Proposition 2 in the main text) *Conditioned on the set of q_{rt}'s, each random probability measure μ_t defined in (2) is marginally distributed as a normalized random measure with Lévy measure $\sum_r q_{rt}\nu_r(\mathrm{d}w, \mathrm{d}\theta)$.*

Proof One approach is to follow the proof of Lemma 11 in (Chen et al., 2012a), here we give a simplified proof using the characteristic function of a CRM (4).

Denote $\mathcal{B} = \{0,1\}^{\#R\times T}$, from the definition of $\tilde{\mu}_t$, the underlying point process can be considered as a Mark-Poisson process in the product space $\mathbb{R}^+ \times \Theta \times \mathcal{R} \times \mathcal{B}$, where each atom (w, θ) in region $\mathcal{R}_r$ is associated with a

Bernoulli variable z with parameter q_{rt}. From the *marking theorem* of a Poisson process we conclude that $\tilde{\mu}_t$'s are again CRMs. To derive the Lévy measures, denote $\mathrm{d}z$ as the infinitesimal of a Bernoulli random variable z, using the Lévy-Khintchine formula for a CRM as in Lemma 1, the corresponding characteristic functional can be calculated as

$$\begin{aligned}
\mathbb{E}\left[e^{\int_\Theta iu\tilde{\mu}_t(\mathrm{d}\theta)}\right] &= \exp\left\{-\int_{\mathbb{R}^+\times\Theta\times\mathcal{R}\times\mathcal{B}}\left(1-e^{iuw}\right)\nu(\mathrm{d}w,\mathrm{d}\theta,\mathrm{d}a)\mathrm{d}z\right\} \\
&= \exp\left\{-\int_{\mathbb{R}^+\times\Theta\times\mathcal{R}}\left(1-e^{iuw}\right)q_{r_a t}\nu(\mathrm{d}w,\mathrm{d}\theta,\mathrm{d}a)\right\} \qquad (8)\\
&= \exp\left\{-\int_{\mathbb{R}^+\times\Theta}\left(1-e^{iuw}\right)\left(\sum_{r=1}^{\#\mathcal{R}} q_{rt}\nu_r(\mathrm{d}w,\mathrm{d}\theta)\right)\right\}\ , \qquad (9)
\end{aligned}$$

where (8) follows by integrating out the Bernoulli random variable z with parameter $q_{r_a t}$, (9) follows by integrating out the *region space*. Again according to the uniqueness property of the characteristic functional, μ_t's are marginally normalized random measure with Lévy measures $\sum_{r=1}^{\#\mathcal{R}} q_{rt}\nu_r(\mathrm{d}w,\mathrm{d}\theta)$.

■

Proposition 8 (Proposition 3 in the main text) *Denote the Lévy measure in region R_r as $\nu_r(\mathrm{d}w,\mathrm{d}\theta)$, and fix the subsampling rates q_{rt}. Given observations associated with a set of weights W, and auxiliary variables u_t for each $t \in \mathcal{T}$, the remaining weights in region R_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 The independence of the atoms with and without observations directly follows from the property of the completely random measures (James et al., 2009). It remains to proof the Lévy measure of the random measure formed by the random atoms of the corresponding Poisson process.

The way to prove the posterior Lévy measure is to apply Lemma 3, where the idea is to formulate the joint distribution of the Poisson random measure and the observations into an exponential tilted Poisson random measure. Note it suffices to consider one region case because the CRM between regions are independent. For notational simplicity we omit the subscript r in all the statistics related to r, *e.g.*, n_{trw} is simplified as n_{tw}.

Now denote the base random measure as $\tilde{\mu}$, then construct a set of dependent NRMs μ_t's by thinning $\tilde{\mu}$ with different rates q_j. Given observations for μ_t's, by the Poisson partition calculus (James, 2005) it follows that the joint distribution for $\{\mu_t\}$ and observations with statistics $\{n_{tw}\}$ is

$$p(\{n_{tw}\},\{\mu_t\}) = \prod_t \frac{\prod_k w_k^{n_{tw_k}}}{\left(\sum_{k'} z_{tk'}w_{k'}\right)^{N_t}} P(\mathcal{N}|\nu)\ .$$

Now we introduce an auxiliary variable u_t for each t via Gamma identity, and the joint becomes

$$p(\{n_{tw}\},\{\mu_t\},\{u_t\}) = \prod_t \frac{\prod_{k:n_{tw_k}>0} w_k^{n_{tw_k}}}{\Gamma(N_t)} \prod_k e^{-\sum_t z_{tk}u_t w_k} P(\mathcal{N}|\nu)\ .$$

Now integrate out all the z_{tk}'s in the exponential terms we have:

$$\begin{aligned}
& \mathbb{E}_{\{z_{tk}\}}\left[\prod_k e^{-\sum_j z_{jik}u_j w_k}\right] \\
= & \prod_k \prod_j \left(1 - q_j + q_j e^{-u_j w_k}\right) \\
= & \exp\left(-\sum_k \sum_j -\log\left(1 + q_j\left(e^{-u_j w_k} - 1\right)\right)\right)
\end{aligned}$$

Let $f = -\sum_k \sum_j \log\left(1 + q_j\left(e^{-u_j w_k} - 1\right)\right)$, $g(\mathcal{N}) = 1$ in Lemma 3, then by applying Lemma 3, we conclude that the Poisson process has posterior intensity of

$$e^{-f(w)}\nu(\mathrm{d}w, \mathrm{d}\theta) = \prod_j \left(1 - q_j + q_j e^{-u_j w}\right)\nu(\mathrm{d}w, \mathrm{d}\theta) \ ,$$

which is the conditional Lévy measure of $\tilde{\mu}$ by the relationship between a Poisson process and the CRM constructed from it.

■

C. Inference

C.1. Mixed Normalized Random Measures

C.1.1. Posterior Inference for Mixed Normalized Generalized Gamma Processes with Marginal Sampler

We first derive the posterior of MNRMs. Given observations $\vec{X}$, denote μ_r as the NRM in region $\tilde{R}_r$, the likelihood can be expressed as

$$p(\vec{X}|\{\mu_r\}, \{q_{rt}\}) = \frac{\prod_{t=1}^T \prod_{r=1}^I \prod_{k=1}^{K_r} (q_{rt} w_{rk})^{n_{trk}}}{\prod_{t'=1}^T \left(\sum_{r'=1}^I \sum_{k'=1}^\infty q_{r't'} w_{r'k'}\right)^{N_{t'}}} \prod_{t=1}^T \prod_{l=1}^{L_t} F(x_{tl}|\theta_{g_{tl}s_{tl}}) \tag{10}$$

Now introduce auxiliary variables $\{u_t\}$ using Gamma identity, the joint becomes

$$\begin{aligned} & p(\vec{X}, \vec{u}|\{\mu_r\}, \{q_{rt}\}) \\ = \ & \left(\prod_{r=1}^I \prod_{t=1}^T q_{rt}^{n_{rt\cdot}}\right)\left(\prod_{r=1}^I \prod_{k=1}^{K_r} w_{rk}^{n_{\cdot rk}} \exp\left\{-\left(\sum_{t=1}^T q_{rt}u_t\right) w_{rk}\right\}\right) \\ & \left(\prod_{t=1}^T \frac{u_t^{N_t - 1}}{\Gamma(N_t)}\right) \exp\left\{-\sum_{r=1}^I \sum_{k=1}^\infty \left(\sum_{t=1}^T q_{rt}u_t\right) w_{rk}\right\}\left(\prod_{t=1}^T \prod_{l=1}^{L_t} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \end{aligned} \tag{11}$$

We assume the Lévy measure is factorized as

$$\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a) = \nu'(\mathrm{d}w, \mathrm{d}\theta) Q(\mathrm{d}a) \ ,$$

where $Q(\cdot)$ is a measure on $\mathcal{A}$. Now it is easily seen μ_r's are normalized generalized Gamma processes with Lévy measures

$$\nu_r(\mathrm{d}w, \mathrm{d}\theta) = \int_{R_r} \nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a) = \frac{\sigma M_r Q(R_r)}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-w} \mathrm{d}w H(\theta)\mathrm{d}\theta \ ,$$

re-writing $Q_r = Q(R_r)$ and integrating out μ_r's by applying Lemma 2 we get:

$$\begin{aligned} & p(\vec{X}, \vec{u}|\sigma, \{M_r\}, \{q_{rt}\}) = \mathbb{E}_{\{\mu_r\}}\left[p(\vec{X}, \vec{u}|\{\mu_r\}, \{q_{rt}\})\right] \\ \propto \ & \left(\prod_{t=1}^T \prod_{r=1}^{\#\mathcal{R}} q_{rt}^{n_{rt\cdot}}\right)\left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{K.}\left(\prod_{r=1}^{\#\mathcal{R}} (Q_r M_r)^{K_r}\right)\left(\prod_{r=1}^{\#\mathcal{R}} \prod_{k=1}^{K_r} \frac{\Gamma(n_{\cdot rk} - \sigma)}{(1 + \sum_t q_{rt}u_t)^{n_{\cdot rk} - \sigma}}\right) \\ & \left(\prod_{t=1}^T \frac{u_t^{N_t - 1}}{\Gamma(N_t)}\right)\left(\prod_{r=1}^{\#\mathcal{R}} e^{-Q_r M_r\left(\left(1 + \sum_t q_{rt}u_t\right)^\sigma - 1\right)}\right)\left(\prod_{t=1}^T \prod_{l=1}^{L_t} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \ , \end{aligned} \tag{12}$$

since Q_r and M_r always appear together, thus we omit Q_r and only use M_r to represent $Q_r M_r$, this applies to TNRM without further statement.

The variables needed to be sampled are $C = \{\{s_{tl}\}, \{g_{tl}\}\{M_r\}, \{u_t\}, \{q_{rt}\}\}$, based on (12), these can be iteratively sampled as follows:

Sampling (s_{tl}, g_{tl}): The posterior of (s_{tl}, g_{tl}) is

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

where $F_{rk}^{\setminus tl}(x_{tl}) = \frac{\int F(x_{tl}|\theta_{rk}) \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} F(x_{t'l'}|\theta_{rk}) H(\theta_{rk}) \mathrm{d}\theta_{rk}}{\int \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} F(x_{t'l'}|\theta_{rk}) H(\theta_{rk}) \mathrm{d}\theta_{rk}}$ is the conditional density.

Sampling M_r: The posterior of M_r follows a Gamma distribution:

$$p(M_r | C - M_r) \sim \text{Gamma} \left(K_r + a_m, \left(1 + \sum_t q_{rt} u_t \right)^\sigma + b_m - 1 \right) ,$$

where a_m, b_m are parameters of Gamma prior for M_r.

Sampling u_t: The posterior distribution of u_t is:

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

which is log-concave if we use a change of variables: $v_t = \log(u_t)$.

Sampling q_{rt}: Note we should introduce priors for $\{q_{rt}\}$'s, here we use a Gamma prior with parameter q_a and q_b, then the posterior of q_{rt} has the following posterior:

$$p(q_{rt} | C - q_{rt}) \propto \frac{q_{rt}^{n_{tr\cdot} + q_a - 1} \exp\left\{ - M_r \left(1 + \sum_{t'} q_{rt'} u_{t'}\right)^\sigma - q_b q_{rt} \right\}}{\left(1 + \sum_{t'} q_{rt'} u_{t'}\right)^{n_{\cdot r \cdot} - \sigma K_r}} ,$$

which is also log-concave in interval $[-\infty, 0]$ with a change of variables: $Q_{rt} = \log(q_{rt})$.

Sampling σ: From (12), we first instantiate a set of jumps $\{w_{rk}\}$ as

$$w_{rk} \sim \text{Gamma} \left(n_{\cdot rk} - \sigma, 1 + \sum_t q_{rt} u_t \right) ,$$

then the posterior of σ is proportional to:

$$p(\sigma | C - \sigma) \propto \left(\frac{\sigma}{\Gamma(1 - \sigma)} \right)^{K_\cdot} \left(\prod_{r=1}^{\#\mathcal{R}} \prod_{k=1}^{K_r} w_{rk} \right)^{-\sigma} \left(\prod_{r=1}^{\#\mathcal{R}} e^{-M_r \left(1 + \sum_t q_{rt} u_t\right)^\sigma} \right) \tag{13}$$

which is log-concave as well.

C.1.2. Posterior Inference for Mixed Normalized Generalized Gamma Processes with Slice Sampler

The idea of slice sampling MNGG is similar to that of TNGG but with different detailed techniques, readers unfamiliar with the slice sampler are encouraged to first refer to the slice sampler for TNRM in Appendix C.2.2 for more detailed introduction of the underlying ideas.

Starting from (11), we introduce a slice auxiliary variable v_{tl} for each observation such that

$$v_{tl} \sim \text{Uniform}(w_{g_{tl} s_{tl}}) .$$

Now (11) can be rewritten as

$$
\begin{aligned}
& p(\vec{X}, \vec{u}, \{\vec{v_{tl}}\}, \{s_{tl}\}, \{g_{tl}\} | \{\mu_r\}, \{q_{rt}\}) \\
= & \left(\prod_t \prod_l 1\left(w_{g_{tl}s_{tl}} > v_{tl}\right) q_{g_{tl}s_{tl}} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\exp\left\{-\sum_t \sum_r \sum_k q_{rt} u_t w_{rk}\right\}\right) \qquad (14)
\end{aligned}
$$

Now the joint distribution of observations, related auxiliary variables and the corresponding Poisson random measure $\{\mathcal{N}_r\}$ becomes

$$
\begin{aligned}
& p(\vec{X}, \vec{u}, \{\vec{v_{tl}}\}, \{\mu_r\}, \{s_{tl}\}, \{g_{tl}\} | \{q_{rt}\}) \\
= & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) q_{g_{tl}s_{tl}} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\exp\left\{-\sum_t \sum_r \sum_k q_{rt} u_t w_{rk}\right\}\right) \prod_r P(\mathcal{N}_r) \\
\overset{\text{slice at } \mathcal{L}_r}{=} & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) q_{g_{tl}s_{tl}} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \underbrace{\exp\left\{-\sum_t \sum_r \sum_k q_{rt} u_t w_{rk}\right\}}_{\text{jumps larger than } \mathcal{L}_r} \\
& \prod_r p(\{(w_{r1}, \theta_{r1})\}, \cdots, \{(w_{rK'_r}, \theta_{rK'_r})\}) \qquad (K'_r \text{ is \# jumps larger than } \mathcal{L}_r) \qquad (15) \\
& \underbrace{\prod_r \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_0^{\mathcal{L}_r} \left(1 - e^{-\sum_t q_{rt} u_t x}\right) \rho'(\mathrm{d}x)\right\}}_{\text{jumps less than } \mathcal{L}_r}, \qquad (16)
\end{aligned}
$$

where $\rho'(\mathrm{d}x) = x^{-1-\sigma} e^{-x}$ and (15) has the following form based on the fact that $\{(w_{rk}, \theta_{rk})\}$ are points from a compound Poisson process:

$$
p(\{(w_{r1}, \theta_{r1})\}, \cdots, \{(w_{rK'_r}, \theta_{rK'_r})\}) = \left(\frac{\sigma M_r}{\Gamma(1-\sigma)}\right)^{K'_r} \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x)\right\} \prod_k w_{rk}^{-1-\sigma} e^{-w_{rk}},
$$

see Appendix C.2.2 for the derivation.

Now the sampling goes as:

Sample (s_{tl}, g_{tl}): (s_{tl}, g_{tl}) are jointly sampled as a block, it is easily seen the posterior is:

$$
p(s_{tl} = k, g_{tl} = r | C - \{s_{tl}, g_{tl}\}) \propto 1(w_{rk} > v_{tl}) q_{rk} F(x_{tl}|\theta_{rk}) . \qquad (17)
$$

Sample v_{tl}: v_{tl} is uniformly distributed in interval $(0, w_{g_{tl}s_{tl}}]$, so

$$
v_{tl} | C - v_{tl} \sim \text{Uniform}(0, w_{g_{tl}s_{tl}}) . \qquad (18)
$$

Sample w_{rk}: There are two kinds of w_{rk}'s, one is with observations, the other is not, because they are independent, we sample these separately:

- **Sample w_{rk}'s with observations:** It can be easily seen that these w_{rk}'s follow Gamma distributions as

$$w_{rk}|C - w_{rk} \sim \text{Gamma}\left(\sum_t n_{trk} - \sigma, 1 + \sum_t q_{rt} u_t\right) ,$$

- **Sample w_{rk}'s without observations:** These w_{rk}'s are Poisson points in a Poisson process with intensity

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = \rho(\mathrm{d}w) H(\mathrm{d}\theta) = e^{-\sum_t q_{rt} u_t w} \nu_r(\mathrm{d}w, \mathrm{d}\theta) ,$$

where $\nu(\mathrm{d}w, \mathrm{d}\theta)$ is the Lévy measure of μ_r. This is a generalization of the result in (James et al., 2009). In regard of sampling, we use the adaptive thinning approach used in (Favaro & Teh, 2012) with a proposal adaptive Poisson process intensity as

$$\gamma_x(s) = \frac{\sigma M_r}{\Gamma(1-\sigma)} e^{-(1+\sum_t q_{rt} u_t) s} x^{-1-\sigma} \tag{19}$$

See Appendix C.2.2 for the detailed description of this approach and the case for TNGG.

Sample M_r: M_r follows a Gamma distribution as

$$M_r|C - M_r \sim \text{Gamma}\left(K_r' + 1, \frac{\sigma}{\Gamma(1-\sigma)} \int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x) + \int_0^{\mathcal{L}_r} \left(1 - e^{-\sum_t q_{rt} u_t x}\right) \rho'(\mathrm{d}x)\right) ,$$

where K_r' is the number of jumps larger than the threshold $\mathcal{L}_r$ and the integrals can be evaluated using numerical integration or via the incomplete Gamma function as described in (Chen et al., 2012a).

Sample u_t: From (16), we sample u_t using rejection sampling by first sample from the following proposal Gamma distribution

$$u_t|C - u_t \sim \text{Gamma}\left(N_t, \sum_r \sum_k q_{rt} w_{rk}\right) ,$$

then do the rejection step by evaluating it on the posterior (16).

Sample q_{rt}: q_{rt} can also be rejection sampled by using the following proposal Gamma distribution:

$$p(q_{rt}|C - q_{rt}) \propto\sim \text{Gamma}\left(n_{tr\cdot} + a_q, \sum_k u_t w_{rk} + b_q\right) ,$$

where a_q, b_q are the hyperparameters of the Gamma prior.

Sample σ: Based on (16), the posterior of σ is proportional to:

$$p(\sigma|C - \sigma) \propto \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r'} \left(\prod_r \prod_k w_{rk}\right)^{-\sigma}$$
$$\exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \left(\int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x) + \int_0^{\mathcal{L}_r} \left(1 - e^{-\sum_t q_{rt} u_t x}\right) \rho'(\mathrm{d}x)\right)\right\} ,$$

which can be sampled using the slice sampler (Neal, 2003).

C.2. Thinned Normalized Random Measures

C.2.1. Marginal Posterior for Thinned Normalized Generalized Gamma Processes

Though Proposition 3 in the main text shows us the posterior intensity of the Poisson process in region R_r, unfortunately marginalization over this Poisson random measure usually does not end up a simple form. The following proposition gives the marginal posterior of the TNRM under a specific class of the normalized random measure–the normalized generalized Gamma process.

Proposition 9 *Given observations $\vec{X}$ for all times, introduce a set of auxiliary variables $\{u_t\}$. Using the notation and statistics defined in Table 1, the marginal posterior for the TNGG is given by*

$$\begin{aligned}
& p(\vec{X}, \vec{u}, \{s_{tl}\}, \{g_{tl}\} | \sigma, \{M_r\}, \{z_{rtk}\}_{k:n_{\cdot rk}>0}, \{q_{rt}\}) && (20)\\
= & \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r} \left(\prod_r M_r^{K_r}\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\prod_r \prod_{k:n_{\cdot rk}>0} \frac{\Gamma(n_{\cdot rk}-\sigma)}{(1+\sum_t z_{rtk} u_t)^{n_{\cdot rk}-\sigma}}\right) \left(\prod_t \prod_l F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) && (21)\\
& \prod_r \exp\left\{-M_r \left[\sum_{\substack{z'_{rt}\in\{0,1\} \\ for\ t=1\cdots T}} \left(\left(\prod_{t'} q_{rt'}^{z'_{rt'}} (1-q_{rt'})^{1-z'_{rt'}}\right)\left((1+\sum_{t'} z'_{rt'} u_{t'})^{\sigma} - 1\right)\right)\right]\right\},
\end{aligned}$$

where in the last line $\sum_{\substack{z'_{rt}\in\{0,1\} \\ for\ t=1\cdots T}} = \sum_{z'_{r1}=0}^{1} \sum_{z'_{r2}=0}^{1} \cdots \sum_{z'_{rT}=0}^{1}$.

Proof Let $G_{rt} = \sum_k \frac{z_{trk} w_{rk}}{\sum_{k'} z_{trk'} w_{rk'}} \delta_{\theta_{rk}}$, from the property of Poisson process we see that it is a CRM in the augmented space $\mathbb{R}^+ \times \Theta \times \{0,1\}$. Given the observed data, the likelihood is given by

$$\begin{aligned}
& p(\vec{X}, \{s_{tl}\}, \{g_{tl}\} | \{G_{rt}\}) \\
= & \frac{\prod_{t=1}^T \prod_{r=1}^I \prod_{k=1}^{K_r} w_{rk}^{n_{trk}}}{\prod_{t'=1}^T \left(\sum_{r'} \sum_{k'} z_{r't'k'} w_{r'k'}\right)^{N_{t'}}} \prod_{t=1}^T \prod_{l=1}^{L_t} F(x_{tl|\theta_{g_{tl}s_{tl}}}), && (22)
\end{aligned}$$

where $z_{rtk} \sim \text{Bernoulli}(q_{rt}), 0 \le q_{rt} \le 1$.

Now introducing auxiliary variables $\vec{u}$ via the Gamma identity, we have

$$\begin{aligned}
& p(\vec{X}, \vec{u}, \{s_{tl}\}, \{g_{tl}\} | \{G_{rt}\}) \\
= & \left(\prod_{t=1}^T \prod_{r=1}^I \prod_{k=1}^{K_r} w_{rk}^{n_{trk}}\right) \left(\prod_{t=1}^T \prod_{l=1}^{L_t} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\exp\left\{-\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk}\right\}\right) && (23)
\end{aligned}$$

Denote $\Upsilon = \underbrace{\{0,1\} \otimes \cdots \otimes \{0,1\}}_{T}$, $\mathrm{d}\vec{R}_r = \mathrm{d}z_{r1} \cdots \mathrm{d}z_{rT}$, since $\{G_{rt}\}$'s are CRMs, now integrate out $\{G_r\}$'s with Lévy-Khintchine formula (3) and Lemma 2 we have

$$\begin{aligned}
& p(\vec{X}, \vec{u}, \{s_{tl}\}, \{g_{tl}\} | \sigma, \{M_r\}) = \mathbb{E}_{\{G_{rt}\}}\left[p(\vec{X}, \vec{u}, \{s_{tl}\}, \{g_{tl}\} | \{G_{rt}\})\right] \\
= & \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r} \left(\prod_r M_r^{K_r}\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\prod_r \prod_{k:n_{\cdot rk}>0} \frac{\Gamma(n_{\cdot rk}-\sigma)}{(1+\sum_t z'_{rtk} u_t)^{n_{\cdot rk}-\sigma}}\right) \left(\prod_{t=1}^T \prod_{l=1}^{L_t} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \\
& \prod_r \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_\Upsilon \int_\Theta \int_{\mathbb{R}^+} \left(1 - e^{-\sum_t z_{rtx} u_t x}\right) \frac{e^{-x}}{x^{1+\sigma}} \mathrm{d}x \mathrm{d}\theta \mathrm{d}\vec{R}_r\right\} && (24)
\end{aligned}$$

$$
\begin{aligned}
\overset{\text{Taylor}}{\underset{\text{expansion}}{=}} \quad & \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r} \left(\prod_r M_r^{K_r}\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\prod_r \prod_{k:n_{\cdot rk}>0} \frac{\Gamma(n_{\cdot rk}-\sigma)}{(1+\sum_t z'_{rtk} u_t)^{n_{\cdot rk}-\sigma}}\right) \left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \\
& \prod_r \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_\Upsilon \int_\Theta \int_{\mathbb{R}^+} \sum_{n=1}^{\infty} (-1)^{n-1} \frac{\left(\sum_t z_{rtx} u_t\right)^n x^n}{n!} \frac{e^{-x}}{x^{1+\sigma}} \mathrm{d}x \mathrm{d}\theta \mathrm{d}\vec{R}_r\right\} \\
\overset{\text{Integrate out}}{\underset{\text{all } z_{rtx}}{=}} \quad & \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r} \left(\prod_r M_r^{K_r}\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\prod_r \prod_{k:n_{\cdot rk}>0} \frac{\Gamma(n_{\cdot rk}-\sigma)}{(1+\sum_t z'_{rtk} u_t)^{n_{\cdot rk}-\sigma}}\right) \left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \\
& \prod_r \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \left[\sum_{\substack{z'_{rt}\in\{0,1\} \\ \text{for } t=1\cdots T}} \sum_{n=1}^{\infty} (-1)^{n-1} \frac{\left(\sum_{t'} z'_{rt'} u_{t'}\right)^n}{n!} \right.\right. \\
& \qquad \left.\left. \left(\prod_{t'} q_{rt'}^{z'_{rt'}} (1-q_{rt'})^{1-z'_{rt'}} \int_{\mathbb{R}^+} x^{n-\sigma-1} e^{-x} \mathrm{d}x\right)\right]\right\} \\
= \quad & \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r} \left(\prod_r M_r^{K_r}\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\prod_r \prod_{k:n_{\cdot tk}>0} \frac{\Gamma(n_{\cdot rk}-\sigma)}{(1+\sum_t z'_{rtk} u_t)^{n_{\cdot rk}-\sigma}}\right) \left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \qquad (25) \\
& \prod_r \exp\left\{-M_r \left[\sum_{\substack{z'_{rt}\in\{0,1\} \\ \text{for } t=1\cdots T}} \left(\left(\prod_{t'} q_{rt'}^{z'_{rt'}} (1-q_{rt'})^{1-z'_{rt'}}\right) \left((1+\sum_{t'} z'_{rt'} u_{t'})^\sigma - 1\right)\right)\right]\right\}
\end{aligned}
$$

where z_{rtx} in (24) means a Bernoulli random variable drawn at atom x with parameter q_{rt}. Furthermore, the last equation follows by applying the following result

$$
\begin{aligned}
& \sum_{n=1}^{\infty} (-1)^{n-1} \frac{\lambda^n}{n!} \Gamma(n-\sigma) \\
= \quad & \sum_{n=1}^{\infty} (-1)^{n-1} \lambda^n \frac{\Gamma(n-\sigma)}{n!} \\
= \quad & \frac{1}{\sigma} \left(\sum_{n=1}^{\infty} \frac{(-1)^{n-1}\sigma\Gamma(n-\sigma)}{n!} \lambda^n\right) \\
= \quad & \frac{\Gamma(1-\sigma)}{\sigma} \left(\sum_{n=1}^{\infty} \frac{\sigma(\sigma-1)\cdots(\sigma-n+1)}{n!} \lambda^n\right) \qquad (26) \\
= \quad & \frac{\Gamma(1-\sigma)}{\sigma} \left[(1+\lambda)^\sigma - 1\right] ,
\end{aligned}
$$

where the summation in (26) is the Taylor expansion of $(1+\lambda)^\sigma - 1$.

■

A Marginal Sampler for TNGG We derive a marginal sampler for TNGG based on the posterior (20). To sample the topic allocation variables (s_{tl}, g_{tl}), we need to further integrated out the Bernoulli random variables

z_{rtk}'s for the fixed jumps in (21). Thus we augment the terms in the first parenthesis of (21) by instantiating a set of jump size variables w_{rk}'s distributed as

$$w_{rk} \sim \text{Gamma}\left(n_{\cdot rk} - \sigma, 1 + \sum_t z_{rtk} u_t\right) \ . \tag{27}$$

Further denote $\mathbf{u} = (u_1, \cdots, u_T)$, and $\mathbf{b}$ as a length T binary vector, and denote

$$\sum_{\mathbf{b}} = \sum_{b_1=0}^{1} \sum_{b_2=0}^{1} \cdots \sum_{b_T=0}^{1} \ ,$$

then the first parenthesis in (21) can be rewritten as

$$\begin{aligned}
&\prod_r \prod_{k: n_{\cdot rk}>0} w_{rk}^{n_{\cdot rk}-\sigma} e^{-w_{rk}} \prod_t e^{-z_{rtk} u_t w_{rk}} \\
\xrightarrow{\text{integrate out } z_{rtk}} &\prod_r \prod_{k: n_{\cdot rk}>0} w_{rk}^{n_{\cdot rk}-\sigma} e^{-w_{rk}} \prod_t \left(1 - q_{rt} + q_{rt} e^{-u_t w_{rk}}\right) \\
&= \prod_r \prod_{k: n_{\cdot rk}>0} w_{rk}^{n_{\cdot rk}-\sigma} \sum_{\mathbf{b}} \left(\prod_t q_{rt}^{b_t} (1-q_{rt})^{b_t}\right) e^{-(1+<\mathbf{u},\mathbf{b}>) w_{rk}} \\
\xrightarrow{\text{integrate out } w_{rk}} &\prod_r \prod_{k: n_{\cdot rk}>0} \sum_{\mathbf{b}} \left(\prod_t q_{rt}^{b_t} (1-q_{rt})^{b_t}\right) \frac{\Gamma(n_{\cdot rk} - \sigma)}{(1+ <\mathbf{u},\mathbf{b}>)^{n_{\cdot rk}-\sigma}} \ ,
\end{aligned}$$

where $< \cdot, \cdot >$ denotes the inner produce. Based on this, the sampling goes as

Sample (s_{tl}, g_{tl}): for the current time t, the corresponding b_t value is equal to 1, thus the conditional probability for (s_{tl}, g_{tl}) is proportional to

$$\begin{aligned}
&p(s_{tl} = k, g_{tl} = r | C - s_{tl} - g_{tl}) \\
&\propto \begin{cases} q_{rt}(n_{\cdot rk}^{\backslash tl} - \sigma) \left(\sum_{\mathbf{b}: b_t=1} \frac{\prod_{t' \neq t} q_{rt'}^{b_{t'}} (1-q_{rt})^{1-b_{t'}}}{1+<\mathbf{u},\mathbf{b}>}\right) F_{rk}^{\backslash tl}(x_{tl}), & \text{if } k \text{ already exists,} \\ \sigma \left(\sum_{r'} q_{r't} M_{r'} \sum_{\mathbf{b}: b_t=1} \frac{\prod_{t' \neq t} q_{r't'}^{b_{t'}} (1-q_{r't'})^{1-b_{t'}}}{(1+<\mathbf{u},\mathbf{b}>)^{1-\sigma}}\right) \int_\Theta F(x_{tl}|\theta) H(\theta) \mathrm{d}\theta, & \text{if } k \text{ is new.} \end{cases}
\end{aligned}$$

When $T = 2$ this becomes:

$$\propto \begin{cases} q_{rt}(n_{\cdot rk}^{\backslash tl} - \sigma) \left(\frac{1-q_{r\tilde{t}}}{1+u_{\tilde{t}}} + \frac{q_{r\tilde{t}}}{1+u_1+u_2}\right) F_{rk}^{\backslash tl}(x_{tl}), & \text{if } k \text{ already exists,} \\ \sigma \left(\sum_{r'} q_{r't} M_{r'} \left(\frac{1-q_{r'\tilde{t}}}{(1+u_{\tilde{t}})^{1-\sigma}} + \frac{q_{r'\tilde{t}}}{(1+u_1+u_2)^{1-\sigma}}\right)\right) \int_\Theta F(x_{tl}|\theta) H(\theta) \mathrm{d}\theta, & \text{if } k \text{ is new} \end{cases}$$

where $\tilde{t} = 1$ when $t = 2$, and $\tilde{t} = 2$ when $t = 1$. $F_{rk}^{\backslash tl}(x_{tl}) = \frac{\int F(x_{tl}|\theta_{rk}) \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} F(x_{t'l'}|\theta_{rk}) H(\theta_{rk}) \mathrm{d}\theta_{rk}}{\int \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} F(x_{t'l'}|\theta_{rk}) H(\theta_{rk}) \mathrm{d}\theta_{rk}}$ is the conditional density.

Sample M_r: M_r has a Gamma distributed posterior as

$$M_r | C - M_r \sim \text{Gamma}\left(K_r + a_m, \sum_{\mathbf{b}} \left(\prod_t q_{rt}^{b_t} (1-q_{rt})^{1-b_t}\right) ((1+ <\mathbf{u},\mathbf{b}>)^\sigma - 1) + b_m\right) \ ,$$

where (a_m, b_m) are parameters of the Gamma prior for M_r.

To sample $(\{u_t\}, \{q_{rt}\}, \sigma)$, we first instantiate the fixed jumps w_{rk} as in (27), and sample the latent Bernoulli variables z_{rtk} for $(k : n_{\cdot rk} > 0)$ using the following rule

$$p(z_{rtk} = 1 | C - z_{rtk}) = \begin{cases} 1, & \text{if } n_{trk} > 0 \ , \\ \frac{q_{rt} e^{-u_t w_{rk}}}{1-q_{rt}+q_{rt} e^{-u_t w_{rk}}}, & \text{if } n_{trk} = 0 \ . \end{cases}$$

With these latent variables, sampling for other parameters goes as

Sample u_t: the posterior of u_t has the following form:

$$p(u_t|C-u_t) \propto u_t^{N_t-1} e^{-\left(\sum_r \sum_{k:n_{\cdot rk}>0} z_{rtk} w_{rk}\right) u_t} e^{-\sum_r M_r \sum_{\mathbf{b}} \left(\prod_{t'} q_{rt'}^{b_{t'}} (1-q_{rt'})^{1-b_{t'}}\right)(1+<\mathbf{u},\mathbf{b}>)^{\sigma})}, \quad (28)$$

this is log-concave after using a change of variable $v_t = \log(u_t)$. Another possible way to sample is to first sample u_t from a Gamma distribution Gamma $\left(N_t, \sum_r \sum_{k:n_{\cdot rk}>0} z_{rtk} w_{rk}\right)$, then use a rejection step evaluated on the true posterior (28), though the acceptance rate would probably be low.

Sample q_{rt}: the posterior of q_{rt} follows:

$$p(q_{rt}|C-q_{rt}) \propto q_{rt}^{\sum_{k:n_{\cdot tk}>0} 1(z_{rtk}=1)+a_q-1} (1-q_{rt})^{\sum_{k:n_{\cdot tk}>0} 1(z_{rtk}=0)+b_q-1} \quad (29)$$

$$e^{-M_r \sum_{\mathbf{b}} \left(\prod_{t'} q_{rt'}^{b_{t'}} (1-q_{rt'})^{1-b_{t'}}\right)((1+<\mathbf{u},\mathbf{b}>)^{\sigma}-1)}, \quad (30)$$

where (a_q, b_q) are parameters of the Beta prior for q_{rt}'s. This is again log-concave, and can be sampled using the slice sampler. Also, similar to sampling u_t, we can also first sample q_{rt} from a Beta $\left(\sum_{k:n_{\cdot tk}>0} 1(z_{rtk}=1)+a_q, \sum_{k:n_{\cdot tk}>0} 1(z_{rtk}=0)+b_q\right)$ proposal distribution and do a rejection step based on the true posterior (29).

Sample σ: From (20), σ has the following posterior:

$$p(\sigma|C-\sigma) \propto \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{K.} \left(\prod_r \prod_{k:n_{\cdot rk}>0} w_{rk}\right)^{\sigma} \prod_r e^{-M_r \sum_{\mathbf{b}} \left(\prod_t q_{rt}^{b_t}(1-q_{rt})^{1-b_t}\right)(1+<\mathbf{u},\mathbf{b}>)^{\sigma})},$$

this is log-concave as well and can be sampled with the slice sampler.

We can see from the above marginal sampler for TNGG that it is computationally infeasible even for a moderately large time T. The reason being that the marginal posterior contains a 2^T summation term, thus computation complexity grows exponentially with the number of times. Alternatively, based on the recent development of sampling for normalized random measures (Griffin & Walker, 2011; Favaro & Teh, 2012), we are able to develop a slice sampler for TNGG that greatly reduces the computational cost. This is described in the next section.

C.2.2. Posterior Inference for the TNGG via Slice Sampling

This section describes a slice sampler for a specific class of the NRM, *i.e.*, thinned-spatial normalized generalized Gamma process (TNGG). The idea behind the slice sampler is to introduce auxiliary slice variables such that conditioned on these, the realization of normalized random measures only have a finite set of jumps larger than a threshold, thus turning the inference from infinite parameter spaces to finite parameter spaces.

To derive the slice sampling formula, we first introduce a slice auxiliary variable v_{tl} for each observation such that

$$v_{tl} \sim \text{Uniform}(w_{g_{tl}s_{tl}}) \ .$$

Based on (23), now the joint posterior of observations and related auxiliary variables becomes

$$\begin{aligned} & p(\vec{X}, \vec{u}, \{\vec{v_{tl}}\}, \{s_{tl}\}, \{g_{tl}\}|\{\mu_r\}, \{z_{rtk}\}, \{q_{rt}\}) \\ = \quad & \left(\prod_t \prod_l 1\left(w_{g_{tl}s_{tl}} > v_{tl}\right) F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\ & \left(\exp\left\{-\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk}\right\}\right) \end{aligned} \quad (31)$$

Denote $\rho'(\mathrm{d}x) = x^{-1-\sigma} e^{-x}$. In the slice sampler we want to instantiate the jumps larger than a threshold, say $\mathcal{L}_r$, for region $\tilde{R}_r$. As a result, the joint distribution of the observations, related auxiliary variables and the Poisson random measure $\{\mathcal{N}_r\}$ becomes

$$p(\vec{X}, \vec{u}, \{\vec{v_{tl}}\}, \{\mu_r\}, \{s_{tl}\}, \{g_{tl}\}|\{z_{rtk}\}, \{q_{rt}\})$$

$$
\begin{aligned}
= \quad & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\exp\left\{-\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk}\right\}\right) \prod_r P(\mathcal{N}_r) \\
\overset{\text{slice at } \mathcal{L}_r}{=} \quad & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \underbrace{\exp\left\{-\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk}\right\}}_{\text{jumps larger than } \mathcal{L}_r} \\
& \prod_r p(\{(w_{r1}, \theta_{r1})\}, \cdots, \{(w_{rK'_r}, \theta_{rK'_r})\}) \qquad (K'_r \text{ is \# jumps larger than } \mathcal{L}_r) \\
& \underbrace{\prod_r \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_0^{\mathcal{L}_r} \left(1 - \prod_t \left(1 - q_{rt} + q_{rt} e^{-u_t x}\right)\right) \rho'(\mathrm{d}x)\right\}}_{\text{jumps less than } \mathcal{L}_r \text{, according to Proposition 8}} \qquad (32) \\
\overset{\text{small } \mathcal{L}_r}{=} \quad & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \underbrace{\exp\left\{-\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk}\right\}}_{\text{jumps larger than } \mathcal{L}_r} \\
& \underbrace{\prod_r \left(\frac{\sigma M_r}{\Gamma(1-\sigma)}\right)^{K'_r} \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x)\right\} \prod_k w_{rk}^{-1-\sigma} e^{-w_{rk}}}_{p(\{(w_{1k},\theta_{1k})\},\{(w_{2k},\theta_{2k})\},\cdots,\{(w_{Ik},\theta_{Ik})\})} \qquad (33) \\
& \underbrace{\prod_r \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_0^{\mathcal{L}_r} \left((\sum_j q_{rt} u_t) x + O((u_t x)^2)\right) \rho'(\mathrm{d}x)\right\}}_{\text{jumps less than } \mathcal{L}_r} \qquad (34) \\
\approx \quad & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) F(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \underbrace{\exp\left\{-\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk}\right\}}_{\text{jumps larger than } \mathcal{L}_r} \\
& \underbrace{\prod_r \left(\frac{\sigma M_r}{\Gamma(1-\sigma)}\right)^{K'_r} \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x)\right\} \prod_k w_{rk}^{-1-\sigma} e^{-w_{rk}}}_{p(\{(w_{1k},\theta_{1k})\},\{(w_{2k},\theta_{2k})\},\cdots,\{(w_{Ik},\theta_{Ik})\})} \\
& \underbrace{\exp\left\{-\sum_r (\sum_t q_{rt} u_t) M_r \frac{\sigma \mathcal{L}_r^{1-\sigma}}{(1-\sigma)\Gamma(1-\sigma)}\right\}}_{\text{jumps less than } \mathcal{L}_r}, \qquad (35)
\end{aligned}
$$

where (33) is the joint density of a finite jumps from the Poisson process, since it is a compound Poisson process, so the density is:

$$\begin{aligned} & p((w_{r1}, \theta_{r1}), (w_{r2}, \theta_{r2}), \cdots, (w_{rK_r}, \theta_{rK_r})) \\ & = \text{Poisson}\left(K_r; \frac{\sigma M_r}{\Gamma(1-\sigma)} \int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x)\right) K_r! \prod_{k=1}^{K_r} \frac{\rho'(w_{rk})}{\int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x)} , \end{aligned}$$

where we assume the Lévy measure is decomposed as $\nu(\mathrm{d}w, \mathrm{d}\theta) = \rho(\mathrm{d}w)H(\mathrm{d}\theta)$, Poisson$(k; A)$ means the density of the Poisson distribution with mean A under value k.

Further integrate out all the $\{z_{rtk}\}$'s, we have

$$\begin{aligned} & p(\vec{X}, \vec{u}, \{\vec{v_{tl}}\}, \{w_{rk}\}, \{s_{tl}\}, \{g_{tl}\} | \sigma, \{M_r\}) \\ \approx \quad & \left(\prod_{t=1}^{T} \prod_{l=1}^{L_t} 1(w_{g_{tl}s_{tl}} > v_{tl}) F(x_{tl} | \theta_{g_{tl}s_{tl}})\right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)}\right) \\ & \underbrace{\left(\prod_t \prod_r \prod_{k: n_{trk=0}} \left(1 - q_{rt} + q_{rt} e^{-u_t w_{rk}}\right) \prod_{k: n_{trk>0}} e^{-u_t w_{rk}}\right)}_{\text{jumps larger than } \mathcal{L}_r} \\ & \underbrace{\prod_r \left(\frac{\sigma M_r}{\Gamma(1-\sigma)}\right)^{K_r'} \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x)\right\} \prod_k w_{rk}^{-1-\sigma} e^{-w_{rk}}}_{p(\{(w_{1k}, \theta_{1k})\}, \{(w_{2k}, \theta_{2k})\}, \cdots, \{(w_{Ik}, \theta_{Ik})\})} \\ & \underbrace{\exp\left\{-\sum_r (\sum_t q_{rt} u_t) M_r \frac{\sigma \mathcal{L}_r^{1-\sigma}}{(1-\sigma)\Gamma(1-\sigma)}\right\}}_{\text{jumps less than } \mathcal{L}_r} \end{aligned} \tag{36}$$

C.2.3. Bound analysis

Note that in the above derivation, we have used a linear approximation for an exponential function in (32) to make it become (35). Actually, this approximation is quite accurate given $u_t \ll 1/\mathcal{L}_r$, and this is easily satisfied by choosing an appropriate threshold $\mathcal{L}_r$ in the sampling (we chose $\mathcal{L}_r = \min\left\{0.001/\max_t\{u_t\}, \min_{(t,l):g_{tl}=r}\{v_{tl}\}\right\}$ in the experiments).

In this section we will give an analysis on the tightness of the bound in the approximation (35) with respect to $\mathcal{L}_r$, we analysis the lower bound and upper bound of the true posterior (32). First, we define the following notation:

$$t_{min}^r = \arg\min_{t: q_{rt} \neq 0} \{q_{rt}(1 - e^{-u_t \mathcal{L}_r})\},$$

$$t_{max}^r = \arg\max_t \{q_{rt} u_t\}.$$

Also denote the last term in (32) as $\tilde{Q}_r(\mathcal{L}_r)$, *i.e.*,

$$\tilde{Q}_r(\mathcal{L}_r) = \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_0^{\mathcal{L}_r} \left(1 - \prod_t \left(1 - q_{rt} + q_{rt} e^{-u_t x}\right)\right) \rho'(\mathrm{d}x)\right\}.$$

We use the following inequality:

$$1 - u_t x \le e^{-u_t x} \le 1 - \frac{1 - e^{-u_t L}}{L} x, \qquad \forall L \ge x. \tag{37}$$

Then we have the upper bound for $\tilde{Q}_r(\mathcal{L}_r)$:

$$\tilde{Q}_r(\mathcal{L}_r) \quad \le \quad \exp\left\{-\int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)} \left(1 - \prod_t \left(1 - \frac{q_{rt}(1 - e^{-u_t \mathcal{L}_r})}{\mathcal{L}_r} x\right)\right) \left(x^{-\sigma-1} - x^{-\sigma}\right) \mathrm{d}x\right\}$$

$$\begin{aligned}
&\leq \exp\left\{-\int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)}\left(1-\left(1-\frac{q_{rt^r_{min}}(1-e^{-u_{t^r_{min}}\mathcal{L}_r})}{\mathcal{L}_r}x\right)^T\right)\left(x^{-\sigma-1}-x^{-\sigma}\right)\mathrm{d}x\right\} \\
&\leq \exp\left\{-\int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)}\left(2-q_{rt^r_{min}}(1-e^{-u_{t^r_{min}}\mathcal{L}_r})\right)^{T/2}\left(\frac{q_{rt^r_{min}}(1-e^{-u_{t^r_{min}}\mathcal{L}_r})}{\mathcal{L}_r}\right)^{T/2}\right. \\
&\qquad \left. x^{T/2}\left(x^{-\sigma-1}-x^{-\sigma}\right)\mathrm{d}x\right\} \\
&= \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)}\left(\frac{q_{rt^r_{min}}(1-e^{-u_{t^r_{min}}\mathcal{L}_r})}{\mathcal{L}_r}\right)^{T/2}\left(2-q_{rt^r_{min}}(1-e^{-u_{t^r_{min}}\mathcal{L}_r})\right)^{T/2}\right. \\
&\qquad \left.\left(\frac{2}{T-2\sigma}-\frac{2\mathcal{L}_r}{T-2\sigma+2}\right)\mathcal{L}_r^{\frac{T}{2}-\sigma}\right\}. \qquad (38)
\end{aligned}$$

Similarly, we have the lower bound:

$$\begin{aligned}
\tilde{Q}_r(\mathcal{L}_r) &\geq \exp\left\{-\int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)}\left(1-\prod_t(1-q_{rt}u_t x)\right)\left(x^{-\sigma-1}-\frac{1-e^{-\mathcal{L}_r}}{\mathcal{L}_r}x^{-\sigma}\right)\mathrm{d}x\right\} \\
&\geq \exp\left\{-\int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)}\left(1-\left(1-q_{rt^r_{max}}u_{t^r_{max}}x\right)^T\right)\left(x^{-\sigma-1}-\frac{1-e^{-\mathcal{L}_r}}{\mathcal{L}_r}x^{-\sigma}\right)\mathrm{d}x\right\} \\
&\geq \exp\left\{-\int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)}2^{T/2}\left(q_{rt^r_{max}}u_{t^r_{max}}\right)^{T/2}x^{T/2}\left(x^{-\sigma-1}-\frac{1-e^{-\mathcal{L}_r}}{\mathcal{L}_r}x^{-\sigma}\right)\mathrm{d}x\right\} \\
&= \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)}\left(q_{rt^r_{max}}u_{t^t_{max}}\right)^{T/2}2^{T/2}\left(\frac{2}{T-2\sigma}-\frac{2(1-e^{-\mathcal{L}_r})}{T-2\sigma+2}\right)\mathcal{L}_r^{\frac{T}{2}-\sigma}\right\}. \qquad (39)
\end{aligned}$$

C.2.4. Sampling

The variables needed to be sampled include the jumps $\{w_{rk}\}$'s (with or without observations), the Bernoulli variables $\{z_{rtk}\}$'s, mass parameters $\{M_r\}$'s, atom assignment $\{s_{tl}\}$'s, source assignment $\{g_{tl}\}$'s and auxiliary variables u_t's as well as the index parameter σ. We denote the whole set as C, then the sampling goes as follows:

Sample (s_{tl}, g_{tl}): (s_{tl}, g_{tl}) are jointly sampled as a block, it is easily seen the posterior is:

$$p(s_{tl}=k, g_{tl}=r|C-\{s_{tl},g_{tl}\}) \propto 1(w_{rk}>v_{tl})1(z_{rtk}=1)F(x_{tl}|\theta_{g_{tl}s_{tl}}) \; . \qquad (40)$$

Sample v_{tl}: v_{tl} is uniformly distributed in interval $(0, w_{g_{tl}s_{tl}}]$, so

$$v_{tl}|C-v_{tl} \sim \mathrm{Uniform}(0, w_{g_{tl}s_{tl}}) \; . \qquad (41)$$

Sample w_{rk}: There are two kinds of w_{rk}'s, one is with observations, the other is not, because they are independent, we sample these separately:

- **Sample w_{rk}'s with observations:** It can easily be seen that these w_{rk}'s follow Gamma distributions as

$$w_{rk}|C-w_{rk} \sim \mathrm{Gamma}\left(\sum_t n_{trk}-\sigma, 1+\sum_t z_{rtk}u_t\right) \; ,$$

- **Sample w_{rk}'s without observations:** We already know that these w_{rk}'s are Poisson points in a Poisson process, and from Proposition 8 we know the intensity of the Poisson process is

$$\nu(\mathrm{d}w,\mathrm{d}\theta)=\rho(\mathrm{d}w)H(\mathrm{d}\theta)=\prod_t(1-q_{rt}+q_{rt}e^{-u_t w})\nu_r(\mathrm{d}w,\mathrm{d}\theta) \; ,$$

where $\nu_r(\mathrm{d}w, \mathrm{d}\theta)$ is the Lévy measure of μ_r. So now sampling w_{rk}'s means instantiating a Poisson process with the above intensity, since such Poisson process has infinite points but we only need those points with w_{rk} larger than the threshold $\mathcal{L}_r$, this is finite and the instantiation can be done. An efficient way to do this is to use the adaptive thinning approach in (Favaro & Teh, 2012), as it does not require any numerical integrations but only the evaluation of the intensity $\rho(\mathrm{d}w)$. The idea behind this approach is to sample the points from a *nice* Poisson process with intensity pointwise larger than the intensity needed to be sampled. In another word, we need define a Poisson process with intensity $\gamma_x(s)$ that adaptively bounds ρ, *i.e.*:

$$\begin{cases} \gamma_x(x) = \rho(x) & \\ \gamma_x(s) \geq \rho(s) & \forall s > x \\ \gamma_x(s) \geq \gamma_{x'}(s) & \forall x' \geq x \end{cases}$$

Furthermore, it is expected both $\gamma_x(s)$ and the inversion are analytically tractable with $\int_x^\infty \gamma_x(s')\mathrm{d}s' < \infty$. Then the samples from the Poisson process with intensity $\rho(\mathrm{d}w)$ can be obtained by adaptively thinning some of the instantiated points in the Poisson process with intensity $\gamma_x(s)$. For TNGG, the following adaptive intensity is found to be a good one:

$$\gamma_x(s) = \frac{\sigma M_r}{\Gamma(1-\sigma)} \prod_t \left(1 - q_{rt} + q_{rt} e^{-u_t x}\right) e^{-s} x^{-1-\sigma} \tag{42}$$

Then the procedure goes similarly as in (Favaro & Teh, 2012).

Sample z_{rtk}: For those w_{rk}'s with observations from time t, clearly the posterior is

$$p(z_{rtk} = 1 | C - z_{rtk}) = 1 \ .$$

For those without observation, according to (22), given all the w_{rk}'s, the posterior of the Bernoulli random variable z_{rtk} is

$$p(z_{rtk} = 1 | C - z_{rtk}) = \frac{q_{rt} e^{-u_t w_{rk}}}{1 - q_{rt} + q_{rt} e^{-u_t w_{rk}}} \ .$$

Sample M_r, u_t, q_{rt} and σ: The simplest procedure to sample M_r, u_t and q_{rt} is to use an approximated Gibbs sampler based on the accurate approximated posterior (35) and (36):

- **Sample M_r:** M_r has a Gamma distribution as

$$M_r | C - M_r \sim \text{Gamma}\left(K'_r + 1, \frac{\sigma}{\Gamma(1-\sigma)} \int_{\mathcal{L}_r}^\infty \rho'(\mathrm{d}x) + \frac{\sigma \mathcal{L}_r^{1-\sigma}}{(1-\sigma)\Gamma(1-\sigma)} \sum_t q_{rt} u_t\right) \ ,$$

  where K'_r is the number of jumps larger than the threshold $\mathcal{L}_r$.
- **Sample u_t:** u_t also has a Gamma distribution as

$$u_t | C - u_t \sim \text{Gamma}\left(N_t, \sum_r \sum_k z_{rtk} w_{rk} + \frac{\sigma}{(1-\sigma)\Gamma(1-\sigma)} \sum_r q_{rt} M_r \mathcal{L}_r^{1-\sigma}\right) \ .$$

- **Sample q_{rt}:** the posterior of q_{rt} is proportional to:

$$p(q_{rt} | C - q_{rt}) \propto \prod_{k: n_{trk} = 0} \left(1 - q_{rt} + q_{rt} e^{-u_t w_{rk}}\right) e^{-\frac{\sigma M_r u_t \mathcal{L}_r^{1-\sigma}}{(1-\sigma)\Gamma(1-\sigma)} q_{rt}} \ , \tag{43}$$

  which is log-concave. Now if we use the construction (**??**), and we further employ a Beta prior with parameter a_q and b_q for each q_{rt}, then it can be easily seen that given z_{rtk}, the approximated conditional posterior of q_{rt} is

$$q_{rt} | C - q_{rt} \sim \text{Beta}\left(\sum_k 1(z_{rtk} = 1) + a_q, \sum_k 1(z_{rtk} = 0) + b_q\right) \ .$$

- **Sample σ:** based on (35), the posterior of σ is proportional to:

$$p(\sigma|C-\sigma) \propto \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K'_r} \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)}\int_{\mathcal{L}_r}^{\infty}\rho'(\mathrm{d}x)\right\}\left(\prod_r\prod_k w_{rk}\right)^{-\sigma}$$
$$\exp\left\{-\sum_r(\sum_t q_{rt}u_t)M_r\frac{\sigma\mathcal{L}_r^{1-\sigma}}{(1-\sigma)\Gamma(1-\sigma)}\right\} ,$$

which can be sampled using the slice sampler (Neal, 2003).

Sample M_r, u_t, q_{rt} using pseudo-marginal Metropolis-Hastings: Note the above sampler for M_r, u_t and q_{rt} is not exact because it is based on an approximated posterior. A possible way for exact sampling is by a Metropolis-Hastings schema. However, note that the integral in (34) is hard to evaluate, making the general MH sampler infeasible. A strategy to overcome this is to use the *pseudo-marginal Metropolis-Hastings* (PMMH) method (Andrieu & Roberts, 2009). The idea behind PMMH is to use an unbiased estimation of the likelihood which is easy to evaluate instead of the original likelihood.

Formally, assume we have a system with two sets of random variables M and J, in which J is closely related to M [2], *i.e.*,

$$p(M, J) = p(M)p(J|M) .$$

To sample M, we use the proposal distribution

$$Q(M^*, J^*|M, J) = Q(M^*|M)p(J^*|M^*) ,$$

the acceptance rate is:

$$\begin{aligned}
A &= \min\left(1, \frac{p(M^*, J^*, X)Q(M, J|M^*, J^*)}{p(M, J, X)Q(M^*, J^*|M, J)}\right) \\
&= \min\left(1, \frac{p(M^*, J^*, X)Q(M|M^*)p(J|M)}{p(M, J, X)Q(M^*|M)p(J^*|M^*)}\right) \\
&= \min\left(1, \frac{p(M^*)Q(M|M^*)p(X|M^*, J^*)}{p(M)Q(M^*|M)p(X|M, J)}\right) \qquad (44)
\end{aligned}$$

Here $p(X|M, J)$ is an approximation to the original likelihood. To make the PMMH correct, $p(X|M, J)$ is required to be unbiased estimation of the true likelihood $p^*(X|M, J)$, that is

$$\mathbb{E}[p(X|M, J)] = cp^*(X|M, J),$$

where c is a constant.

To sample M_r, u_t and q_{rt}, we can use the approximation (35), which is unbiased with respective to the random points w_{rk}'s, and also according to the bound analysis in Section C.2.3, the approximated likelihood is accurate if $\mathcal{L}_r$ is small enough. Note that to sample with the PMMH, we need to evaluate the approximated likelihood $p(X|\{u_t\}, \{M_r\}, \{q_{rt}\}, \{w_{rk}\})$ on the proposed M_r^*, u_t^* and q_{rt}^*, which usually has heavy computationally cost given a large number of simulated atoms. This procedure goes as in Algorithm 1.

We usually use Gamma priors for M_r, u_t and Beta prior for q_{rt}, *e.g.*:

$$p(M_r) \sim \text{Gamma}(a_M, b_M) = \frac{b_M^{a_M}}{\Gamma(a_M)} M_r^{a_M-1} e^{-b_M M_r} ,$$

$$p(u_t) \sim \text{Gamma}(a_u, b_u) = \frac{b_u^{a_u}}{\Gamma(a_u)} u_t^{a_u-1} e^{-b_u u_t} ,$$

$$p(q_{rt}) \sim \text{Beta}(a_q, b_q) = \frac{\Gamma(a_q+b_q)}{\Gamma(a_q)\Gamma(b_q)} q_{rt}^{a_q-1}(1-q_{rt})^{b_q-1} .$$

[2] In our case J corresponds to the random points $\{w_{rk}\}$ in the Poisson process, and M corresponds to M_r, u_t or q_{rt}.

Algorithm 1 PMMH sampling for M_r and u_t

1: **repeat**
2: Assume the current state as M_r, u_t, q_{rt}, use this state to simulate the jumps larger than $\mathcal{L}_r$ from a Poisson process, following ideas as in (Favaro & Teh, 2012).
3: Sample the Bernoulli variables z_{rtk}'s
4: Use these jumps and z_{rtk}'s to evaluate the approximated likelihood (35).
5: Propose a move

$$M_r^* \sim Q_M(M_r^*|M_r),$$
$$u_t \sim Q_u(u_t^*|u_t) \text{ , and}$$
$$q_{rt} \sim Q_q(q_{rt}^*|q_{rt}) \ .$$

6: Use this state to simulate the jumps larger than $\mathcal{L}_r$ from a Poisson process, following similar procedure as in (Favaro & Teh, 2012).
7: Use these jumps to evaluate the approximated likelihood (35).
8: Do the accept-reject step using (44).
9: **until** converged

Also we would choose a random walk proposal in the log spaces of M_r, u_t and q_{rt}, *i.e.*,

$$Q(\log(M_r^*)|\log(M_r)) = \frac{1}{\sqrt{2\pi}\sigma_M} \exp\left\{\frac{(\log(M_r^*) - \log(M_r))^2}{2\sigma_M^2}\right\}$$

$$Q(\log(u_t^*)|\log(u_t)) = \frac{1}{\sqrt{2\pi}\sigma_u} \exp\left\{\frac{(\log(u_t^*) - \log(u_t))^2}{2\sigma_u^2}\right\} \ .$$

$$Q(\log(q_{rt}^*)|\log(q_{rt})) = \frac{1}{\sqrt{2\pi}\sigma_q} \exp\left\{\frac{(\log(q_{rt}^*) - \log(q_{rt}))^2}{2\sigma_q^2}\right\} \ .$$

Now the acceptance rates are easily seen to be

$$A_m = \left(\frac{M_r^*}{M_r}\right)^{a_M} e^{-b_M(M_r^* - M_r)} \frac{p(X|M_r^*, \{M_j\}_{j\neq r}, \{u_t\}, \{q_{rt}\}, \{J^*\})}{p(X|\{M_r\}, \{u_t\}, \{q_{rt}\}, \{J\})} \ ,$$
$$A_u = \left(\frac{u_t^*}{u_t}\right)^{a_u} e^{-b_u(u_t^* - u_t)} \frac{p(X|u_t^*, \{u_i\}_{i\neq t}, \{M_r\}, \{q_{rt}\}, \{J^*\})}{p(X|\{u_t\}, \{M_r\}, \{q_{rt}\}, \{J\})} \ ,$$
$$A_q = \left(\frac{q_{rt}^*}{q_{rt}}\right)^{a_q} \left(\frac{1-q_{rt}^*}{1-q_{rt}}\right)^{b_q - 1} \frac{p(X|\{q_{rt}^*\}, \{u_t\}, \{M_r\}, \{J^*\})}{p(X|\{q_{rt}\}, \{u_t\}, \{M_r\}, \{J\})} \ ,$$

where $p(X|\{M_r\}, \{u_t\}, \{J\})$ is the evaluation of (35) with the current set of parameters $\{\{M_r\}, \{u_t\}, \{q_{rt}\}, \{w_{rk}\}\}$.

C.2.5. Prediction in the Slice Sampler

Note that prediction from the slice sampler for TNRM in Section C.2.2 is not straightforward. To make it be able to do prediction, we need to introduce an extra slice variable $v_{t(L_t+1)}$ for the unseen data, an extra jump indicator variable $s_{t(L_t+1)}$, and an extra region indicator variable $g_{t(L_t+1)}$. These auxiliary variables are also sampled during the inference, sampling for $v_{t(L_t+1)}$ is the same as the other slice variables as in (41), while sampling for $(s_{t(L_t+1)}, g_{t(L_t+1)})$ is now modified as:

$$p(s_{tl} = k, g_{tl} = r|C - \{s_{tl}, g_{tl}\}) \propto 1(w_{rk} > v_{tl})1(z_{rtk} = 1) \tag{45}$$

because its observation $x_{t(L_t+1)}$ is unknown. Sampling for the other variables is the same as the previous version except that we need to using $L_t + 1$ observations instead of L_t.

D. Hierarchical Normalized Generalized Gamma Processes

We propose hierarchical normalized generalized Gamma processes (HNGG), a direct generalization of HDP (Teh et al., 2006), and develop a marginal sampler for it.

First, as in Section A.3, we denote an NGG with Lévy measure $\frac{\sigma M}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-w} \mathrm{d}w H(\theta)\mathrm{d}\theta$ as

$$\mu \sim \mathrm{NGG}(\sigma, M, H) \ .$$

An HNGG mixture is then defined as

$$\begin{aligned} &\mu_0 \sim \mathrm{NGG}(\sigma, M_0, H) & & \\ &\mu_j \sim \mathrm{NGG}(\sigma, M, \mu_0) & & j = 1, \cdots, J \\ \psi_{ji} \sim \mu_j, \quad & x_{ji} \sim F(\cdot|\psi_{ji}) & & i = 1, \cdots, N_j \ . \end{aligned}$$

D.1. Marginal Sampler for the HNGG

When marginalized out an NRM, it can be interpreted as a generalized Chinese process conditioned on an auxiliary variable (called *latent relative mass* in (Chen et al., 2012a)). Following the Chinese restaurant process metaphor, we denote n_{jk} as the #customer eating dish θ_k in restaurant μ_j (θ_k's are distinct values among all ψ_{ji}'s), t_{jk} as the #tables serving dish θ_k in restaurant μ_j, K as the #dishes currently activated. We develop an analogue of the direct assignment sampler for the HDP (Teh et al., 2006), where we introduce auxiliary variable β served as the predicted distribution of μ_0 so that μ_0 and μ_j's can be decoupled. We further introduce auxiliary variables U_j for $\mu_j (j = 0, 1, \cdots, J)$, denote the whole set of variables to be sampled as C, based on the conditional posterior of an NGG in Lemma 5, the sampling for the HNGG now goes as follows:

- **Sampling dish index s_{ji} for customer x_{ji}:** this follows a similar way as the HDP

$$p(s_{ji} = k | C - s_{ji}) \propto \begin{cases} \left(n_{j\cdot k}^{/ji} + \sigma \left(M(1+U_j)^\sigma \beta_k - 1\right)\right) F_{rk}^{\backslash tl}(x_{ji}) & \text{if } k \text{ already exists} \\ \sigma M (1+U_j)^\sigma \beta_k \int_\Theta F(x_{ji}|\theta) H(\theta) \mathrm{d}\theta & \text{if } k \text{ is new} \ , \end{cases} \tag{46}$$

  where $F_{rk}^{\backslash tl}(x_{tl}) = \frac{\int F(x_{tl}|\theta_{rk}) \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} F(x_{t'l'}|\theta_{rk}) H(\theta_{rk}) \mathrm{d}\theta_{rk}}{\int \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} F(x_{t'l'}|\theta_{rk}) H(\theta_{rk}) \mathrm{d}\theta_{rk}}$ is the conditional density.

- **Sampling the auxiliary variable U_j:** also based on (Corollary 2 Chen et al., 2012a), the posterior of U_j is

$$p(U_j | C - U_j) \propto \frac{U_j^{N_j - 1}}{(1+U_j)^{N_j - K_j \sigma}} e^{-M(1+U_j)^\sigma} \ ,$$

  where K_j is the #dishes in restaurant μ_j. This posterior is proved to be log-concave after a change of variable as $V_j = \log(U_j)$, thus can be efficiently sampled using the adaptive rejection sampler (Gilks & Wild, 1992) or the slice sampler (Neal, 2003).

- **Sampling #tables t_{jk} in restaurant μ_t:** this follows by simulating a generalized Chinese restaurant process (Chen et al., 2012b). Conditioned on all other statistics, in restaurant μ_j, the probability of creating a new table for dish θ_k is proportional to $(n_{jik} - \sigma)$, while the probability of creating a new table is proportional to $\sigma M (1+U_j)^\sigma$. At the end of this generating process, we get t_{jk} which is equal to the #tables created.

- **Sampling mass parameters M and M_0:** Using Gamma priors for M and M_0, the posterior are simply a Gammas as

$$\begin{aligned} M | C - M &\sim \mathrm{Gamma}\left(\sum_j K_j + a_M, \sum_j (1+U_j)^\sigma + b_M - J\right) \ , \\ M_0 | C - M_0 &\sim \mathrm{Gamma}\left(K + a_0, (1+U_0)^\sigma + b_0 - 1\right) \ , \end{aligned}$$

  where (a_M, b_M) and (a_0, b_0) are hyperparameters for the Gamma prior of M amd M_0, respectively.

- **Update β:** β can be updated using the prediction probabilities for an NGG as

$$\beta \propto (t_{\cdot 1} - \sigma, \cdots, t_{\cdot K} - \sigma, \sigma M_0 (1 + U_0)^\sigma) \ ,$$

such that β is a probability vector.

E. Comments on the correctness of the samplers from (Lin et al., 2010) and (Lin & Fisher, 2012)

As stated in Proposition 8 (Proposition 3 in the main text), given observations in different times, the atoms without observations form a CRM that is usually not in the same class of the original CRM. This has the consequence that marginalization using the Lévy measure of the original CRM is incorrect. There have been two models, *e.g.*, (Lin et al., 2010) and (Lin & Fisher, 2012) ignoring this fact and end up with incorrect marginal samplers. We will detail their problems in the following (we will use their notation and equation counter as in the corresponding papers).

E.1. (Lin et al., 2010)'s sampler

In (Lin et al., 2010), the authors construct a DP-valued Markov chain with a transition operator as follows: given D_t, a DP-distributed RPM at time t, the RPM at time $t+1$ is constructed by thinning D_t, perturbing its atoms, and mixing it with a new 'innovation' DP D_ν. For simplicity, consider only the last transformation, so that

$$D_{t+1} = c_1 D_t + c_2 D_\nu \tag{47}$$

For D_{t+1} to be DP distributed, it must be a convex combination of the other two, with weights drawn from a Dirichlet distribution whose parameters are determined by the concentration parameters of the 2 DPs, as is stated in their theorem:

Theorem 1 (Theorem 3 in (Lin et al., 2010)) *Let $D_1, \cdots, D_m$ be independent Dirichlet processes on Ω with $D_k \sim DP(\mu_k)$, and $(c_1, \cdots, c_m) \sim Dir(\mu_1(\Omega), \cdots, \mu_m(\Omega))$ be independent of $D-1, \cdots, D_m$, then*

$$c_1 D_1 + \cdots + c_m D_m \sim DP(\mu_1, \cdots, +\mu_m) \ .$$

Now, given n observations $\{x_i^t\}$ from D_t, the posterior is still a DP. In equation 19 of their paper, (Lin et al., 2010) apply the previous theorem, and claim that the posterior distribution of D_{t+1} given $\{x_i^t\}$ is still DP distributed. This is not true: the concentration parameter of the posterior DP D_t is $\alpha + n$, and no longer matches the distribution of the mixing parameters. By assuming the posterior of D_{t+1} is DP distributed, the authors are implicitly using a mixture parameter that has a $\mathrm{Dir}(\mu_1(\Omega) + n, \mu_2(\Omega))$ distribution, different from the model specification (which is $\mathrm{Dir}(\mu_1(\Omega), \mu_2(\Omega))$).

The result of this is that as n increases, the mixing coefficient tends to 0 (and thus D_{t+1} tends to D_t). We have a Markov chain whose innovation depends on the number of observations at earlier times, different from the model where the transition probability doesn't depend on the number of observations.

We can see this directly by looking at the cluster assignment rule for observations at time $t+1$ (eq (20) in their paper). This also says that the probability that, say, the first observation at time $t+1$ is assigned to a new cluster decreases to 0 as the number of observations at previous times increases (since the denominator tends to infinity). This cannot be the marginal cluster assignment rule for the proposed model, since this probability should remain $O(1)$ independent of the past.

E.2. (Lin & Fisher, 2012)

The marginal sampler in (Lin & Fisher, 2012) has the same problem. The consequence of these is that inference for these models appears to be much more straightforward than it actually is.

Specifically, from Proposition 8 (Proposition 3 in the main text), the conditional Lévy measure of the base CRMs

(H_s's in their notation for the DP case) is

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

This Lévy measure is not in the form of a DP (after marginalizing out u_t's) even if $v_r(\mathrm{d}w, \mathrm{d}\theta)$ is. As a consequence, conditioned on observations from other times, the sampling probabilities for the observations in the current time is not CRP (Chinese restaurant process) distributed, and thus the prediction rules of the CRP can not be used to resample the current data However, in Lin and Fisher's paper, they actually used the CRP prediction rules to do the resampling, *e.g.* Eq.(10) in their paper. This mis-usage makes their sampling method not consistent with their model and thus is not correct.

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. Rao is funded by DARPA MSEE. Rao and Teh were funded by the Gatsby Charitable Foundation.

References

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

Chen, C., Buntine, W., and Ding, N. Theory of dependent hierarchical normalized random measures. Technical Report arXiv:1205.4159, ANU and NICTA, Australia, May 2012a. URL http://arxiv.org/abs/1205.4159.

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

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

Gilks, W. R. and Wild, P. Adaptive rejection sampling for Gibbs sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 41(2):337–348, 1992.

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

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

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.