Dependent Normalized Random Measures

Changyou Chen

A thesis submitted for the degree of
Doctor of Philosophy
The Australian National University

February 2014

© Changyou Chen 2014

Except where otherwise indicated, this thesis is my own original work.

Changyou Chen
28 February 2014

Changyou Chen

To all my family, on earth and in heaven, for their selfless care and love.

Acknowledgments

It is time to say goodbye to my PhD research in the Australian National University. It has been an exciting and memorable experience and I would like to take this opportunity to thank everyone who has helped me during my PhD.

My utmost gratitude goes to my supervisor, Professor Wray Buntine, for both his extraordinary supervision in my research and selfless care in my daily life. Wray brought me into the Bayesian nonparametrics world, he has taught me how to perform quality research from the beginning of my PhD. He taught me how to do critical thinking, how to design professional experiments, how to write quality papers, and how to present my research work. I benefited greatly from his experience which has helped me develop as a machine learning researcher.

I would like to express my gratitude and respect to Professor Yee Whye Teh, who hosted me as a visiting student at UCL for a short time in 2012. My special thanks also goes to Vinayak Rao, who explained to me his spatial normalized Gamma processes, a precursor of our joint work together with Yee Whye. I was impressed by Yee Whye and Vinayak's knowledge in Bayesian nonparametrics, and their sophisticated ideas for Bayesian posterior inference. I was lucky to work with them and had our joint work published in ICML.

I also own my thanks to my friend and collaborator Nan Ding previously from Purdue University. I thank him for sharing his ideas in topic model and Bayesian nonparametrics when he visited NICTA in 2011. Since then we have had close collaborations, which lead to two publications in ICML and TPAMI respectively, as well as one manuscript to be submitted.

Thank you to Novi Quadrianto from University of Cambridge, who hosted my short visit to the machine learning group in University of Cambridge in 2012. My gratitude also goes to all members in Professor Zoubin Ghahramani's research group, who had shared with me their excellent research. The names cannot be put here because of the space limit.

My gratitude also goes to Jun Zhu, who hosted my visit to Tsinghua University in 2013. I thank him for sharing with me his ideas on Regularized Bayesian Inference, based on which our (together with Xinhua Zhang) joint work on Bayesian max-margin clustering is built. Thanks to Xinhua Zhang for teaching me optimization techniques and his enormous help in my research and our joint work.

I sincerely appreciate Professor Lancelot James from the Hong Kong University of Science and Technology to provide funding for my visit in 2013. His expertise in Bayesian nonparametrics has greatly inspired me in my research. It is my honor to have to the opportunity to work with him.

I thank Lexing Xie for giving me the opportunity as a teaching assistant for the undergraduate course *Advance Data Ming* and her assistance in our joint work of *differential topic models*. My thanks also goes to every one in ANU and NICTA who accompanied me during my PhD, including but not limited to Bob Williamson, Scott Sanner, Edwin Bonilla, Mark Reid, Stephen Gould, Tiberio Caetano, Chris Webb, Honglin Yu, Lan Du, Kar Wai Lim, Ehsan Abbasnejad, Avi Ruderman, Suvash Sedhain, Trung Nguyen, *etc*.

At last, my special thanks goes to all my family, including my parents Qihang Chen and Jianfeng Wu who brought me to this wonderful world, fed me and taught me; my grandparents Qingbo Chen and Weirong Huo who had given me all their loves during their lives but left me without my paying back during/before my PhD. I also owe my special thanks to my wife Xuefang He for her understanding, consideration and support of my PhD.

Abstract

Bayesian nonparametrics, since its introduction, has gained increasing attention in machine learning due to its flexibility in modeling. Essentially, Bayesian nonparametrics defines distributions over infinite dimensional objects such as discrete distributions and smooth functions. This overcomes the fundamental problem of model selection which is hard in traditional machine learning, thus is appealing in both application and theory. Among the Bayesian nonparametric family, random probability measures have played important roles in modern machine learning. They have been used as priors for discrete distributions such as topic distributions in topic models. However, a general treatment and analysis of the random probability measure has not been fully explored in the machine learning community.

This thesis introduces the normalized random measure (NRM), built on theories of Poisson processes and completely random measures from the statistical community. Then a family of dependent normalized random measures, including hierarchical normalized random measures, mixed normalized random measures and thinned normalized random measures, are proposed based on the NRM framework to tackle different kinds of dependency modeling problems, *e.g.*, hierarchical topic modeling and dynamic topic modeling. In these dependency models, various distributional properties and posterior inference techniques are analyzed based on the general theory of Poisson process partition calculus. The proposed dependent normalized random measure family generalizes some popular dependent nonparametric Bayesian models such as the hierarchical Dirichlet process, and can be easily adapted to different applications. Finally, more generalized dependent random probability measures and possible future work are discussed.

To sum up, the contributions of the thesis include:

- *Transfer the theory of the normalized random measure from the statistical to machine learning community.* Normalized random measures, which were proposed recently in the statistical community, generalize the Dirichlet process to a large extent, thus are much more flexible in modeling real data. This thesis forms the most extensive research to date in this area.

- *Explore different ideas about constructing dependent normalized random measures.* Existing Bayesian nonparametric models only explore limited dependency structures, with probably the most popular and successful being hierarchical construction, *e.g.*, the hierarchical Dirichlet process (HDP). The dependency models in the thesis not only extend the HDP to hierarchical normalized random measures for more flexible modeling, but also explore other ideas by controlling specific atoms of the underlying Poisson process. This results in many dependency models with abilities to handle dependencies beyond hierarchical

dependency such as the Markovian dependency. In addition, by constructing the dependency models in such ways, various distributional properties and posterior structures can be well analyzed, resulting in much more theoretically clean models. These are lacked of in the hierarchical dependent model.

- All the models proposed are extensively tested, through both synthetic data and real data, such as in topic modeling of documents. Experimental results have shown superior performance compared to realistic baselines, demonstrating the effectiveness and suitability of the proposed models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Probability modeling is a powerful tool and becoming increasing popular in modern machine learning. In probability modeling, we are given some observations denoted as $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}$, which are assumed to be generated from a model $\mathcal{M}$. A model is usually parameterized by a set of parameters, *e.g.*, a Gaussian model is represented by its mean and covariance; thus given a particular kind of model (sometimes called the structure of the model) we can think of the parameters as completing the model. For convenience, we often use the parameters to represent the model, leaving the kind of model as implicit. The conditional distribution of the observations given the model is called

$$\text{Likelihood: } p(\boldsymbol{X}|\mathcal{M}) \ .$$

The goal of the modeling is to find a suitable model $\mathcal{M}^*$ that explains the data $\boldsymbol{X}$. There are two popular ways of addressing this problem: one is to directly optimize the likelihood $P(\boldsymbol{X}|\mathcal{M})$, resulting in the maximal likelihood (ML) estimation of the model; the other is to employ priors for the model $p(\mathcal{M})$, and do maximum a posteriori (MAP) via point estimation or Bayesian inference on the conditional distribution $p(\mathcal{M}|\boldsymbol{X})$. The point estimation of MAP finds an optimal point of the posterior usually by optimization, thus lacks the flexibility of Bayesian inference on the whole posterior distribution. While in Bayesian inference, the posterior is related to the joint likelihood $p(\mathcal{M}, \boldsymbol{X}) = p(\mathcal{M})p(\boldsymbol{X}|\mathcal{M})$ via the well known *Bayes Theorem*:

$$p(\mathcal{M}|\boldsymbol{X}) = \frac{p(\mathcal{M}, \boldsymbol{X})}{p(\boldsymbol{X})} = \frac{p(\mathcal{M})p(\boldsymbol{X}|\mathcal{M})}{p(\boldsymbol{X})} \ .$$

This thesis focuses on the latter case, which is usually called *Bayesian methods for machine learning*. We can see that in this setting, the model prior $p(\mathcal{M})$ plays an important role in Bayesian inference, much research has therefore focused on defining suitable priors for different machine learning problems. Traditional Bayesian methods focus on cases where the model $\mathcal{M}$ is parameterized by a finite number of parameters, known as *parametric Bayesian methods*. However, these kinds of methods encounter the problem that model complexity does not scale with the data size, which would easily cause over-fitting. As a result, modern machine learning has embraced *nonparametric Bayesian methods* (or Bayesian nonparametrics), in which the model $\mathcal{M}$ is parameterized with an infinite number of parameters, and the model complexity would grow with data size[1] and is hopefully free of over-fitting. This is the setting considered in this thesis.

Given the appealing properties of Bayesian nonparametrics, this thesis will focus on a subclass of Bayesian nonparametric priors called discrete *random probability measures* (RPM). Specifically, the *normalized random measure* (NRM), a recently studied RPM in the statistical community, will be extended to construct *dependent normalized random measures*. Then their properties will be analyzed and applications to machine learning will be studied. As is well known, discrete random probability measures are fundamental in machine learning. For example, in topic models [Blei et al., 2003], this corresponds to topic distributions where a topic is defined to be a discrete distribution over vocabulary words; also, in social network modeling, the friendship of one person could be modeled by a random probability measure f, with entry f_i representing the probability of being friends with i. In addition, dependency modeling is ubiquitous in modern machine learning. For instance, in the topic modeling above, usually we are more interested in modeling how topics in each document correlate with each other instead of simply modeling topic distributions for a single document; and in friendship modeling in social network, we are more excited in modeling the correlation of friends between people. The reason is that by dependency modeling, we not only obtain information sharing within data, but also are able to do prediction for unseen data. This thesis focuses on constructing dependency models from the normalized random measure framework. It explores different ways of dependency modeling including the hierarchical dependency modeling and Markovian dependency modeling, which are commonly seen in real applications.

Before going into details of this topic, a brief overview of some popular Bayesian nonparametric priors is first given in the following in order to help the reader get a better understanding of Bayesian nonparametrics in machine learning.

1.1 Some Popular Nonparametric Bayesian Priors

Generally speaking, Bayesian nonparametric priors define distributions over infinite dimensional objects, such as discrete distributions, smooth functions, infinite dimensional binary matrices, *kd*-tree structures, continuous time Markov chains, *etc.* These distributions correspond to the stochastic processes called Dirichlet processes, Gaussian processes, Indian buffet processes, Mondrian processes and fragmentation-coagulation processes, respectively. A brief overview of these stochastic processes will be given in the following.

1.1.1 Dirichlet processes

The Dirichlet process is a distribution over discrete distributions, meaning that each draw/sample from it is a discrete distribution. Formally, let D be a random prob-

[1] It can be estimated from the data.

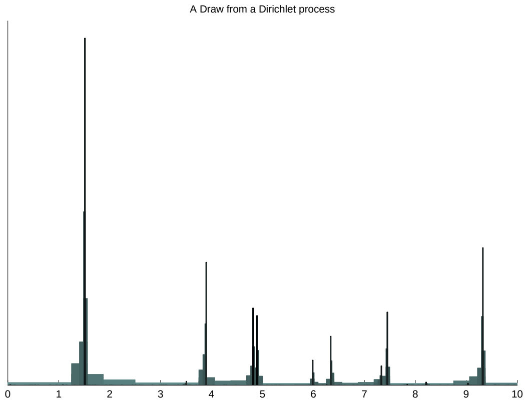


Figure 1.1: A draw from a Dirichlet process.

ability measure on $\mathbb{X}$, α be a positive measure on $\mathbb{X}$, $(X_1, X_2, \cdots, X_n)$ an arbitrary partition of $\mathbb{X}$, then D is called a Dirichlet process with base measure α if [Ferguson, 1973]

$$(D(X_1), D(X_2), \cdots, D(X_n)) \sim \mathrm{Dir}\left(\alpha(X_1), \alpha(X_2), \cdots, \alpha(X_n)\right) .$$

This consistent constructed definition of D meets the conditions of *Kolmogorov consistency theorem* [Çinlar, 2010], guaranteeing the existence of the Dirichlet process on space $\mathbb{X}$. Because each draw from D is a discrete distribution, it can be written as

$$D = \sum_{k=1}^{\infty} w_k \delta_{\theta_k} ,$$

where $0 < w_k < 1, \sum_k w_k = 1$, θ_k's are drawn *i.i.d.* from $\mathbb{X}$ and δ_θ is a point mass at θ. Also note that w_k's can be constructed from a stochastic process called the *stick-breaking process* [Sethuraman, 1994], which will be described in more detail in Chapter 4. Throughout the thesis the Dirichlet process will be denoted as $\mathrm{DP}(\alpha, H)$ where $\alpha(\mathbb{X})$ defined above is simplified as α, and H is the base distribution used to draw the samples θ_k's. A realization of the DP is illustrated in Figure 1.1.

1.1.2 Pitman-Yor processes

As a generalization of the Dirichlet process, the Pitman-Yor process (PYP) first arises in the statistics community from a comprehensive study of excursion lengths and related phenomena of a class of Bessel processes indexed by a parameter $0 \leq \sigma < 1$, see for example [Pitman and Yor, 1997]. Specifically, they were interested in studying the distribution of a decreasing sequence $\{w_k\}$ summing to one. An interesting distribution is the *two parameter Poisson-Dirichlet distribution* constructed via a stick-breaking process derived in [Perman et al., 1992]:

Definition 1.1 (Poisson-Dirichlet distribution). *For $0 \leq \sigma < 1$ and $\theta > -\sigma$, suppose that a probability $P_{\sigma,\theta}$ governs independent random variables V_k such that V_k has Beta$(1 -$*

$\sigma, \theta + k\,\sigma$) distribution. Let

$$w_1 \;=\; V_1, \quad w_k \;=\; (1 - V_1) \cdots (1 - V_{k-1}) V_k \quad k \geq 2\,, \tag{1.1}$$

yielding $\boldsymbol{w} = (w_1, w_2, ...)$. Define the Poisson-Dirichlet distribution *with parameters σ, θ, abbreviated PD(σ, θ) to be the $P_{\sigma,\theta}$ distribution of $\boldsymbol{w}$.*

Note this does assume a particular ordering of the entries in $\boldsymbol{w}$. Here our σ parameter is usually called the *discount parameter* in the literature, and θ is called the *concentration parameter*. The DP is the special case where $\sigma = 0$, and has some quite distinct properties such as slower convergence of the sum $\sum_{k=1}^{\infty} w_k$ to one. General results for the discrete case of the PYP are reviewed in Buntine and Hutter [2012]. Also note there are some other interesting processes of this class of distributions, for example, PD$(\frac{1}{2}, 0)$ corresponds to Brownian Motion and PD$(\frac{1}{2}, \frac{1}{2})$ corresponds to Brownian Bridge.

A suitable definition of a *Poisson-Dirichlet process* is that it extends the Poisson-Dirichlet distribution by attaching each weight w_k to a random point drawn from a space Θ with *base distribution* H. This is denoted as PYP$(\sigma, \theta, H(\cdot))$ and can be represented as $\sum_k w_k \delta_{\theta_k}$. Thus the PYP is a functional on distributions: it takes as input a base distribution and yields as output a discrete distribution with a finite or countable set of possible values on the same domain. The treatment of the PYP from a Bayesian perspective can be found in [Pitman, 1996, 2006], and Gibbs sampling methods for the PYP via stick-breaking prior were proposed by Ishwaran and James [2001]. From this the stick-breaking presentation of the PYP was popularized in machine learning community. Later Ishwaran and James [2003] showed how such processes could be used practically in complex Bayesian mixture models, also giving it its current name, the *Pitman-Yor process* (PYP).

1.1.3 Gaussian processes

The Gaussian process defines distributions over smooth functions–continuous functions. We use GP(μ, K) to denote this stochastic process, which is parameterized by a *mean function* $\mu : \mathbb{X} \mapsto \mathbb{R}$ and a *kernel function* $K : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}^+$, where $\mathbb{X}$ denotes its domain. Different from the DP, now each draw of GP is a random function f, such that for arbitrary $(X_1, X_2, \cdots, X_n) \in \mathbb{X}$, $(f(X_1), f(X_2), \cdots, f(X_n))$ is multivariate normal distributed, *i.e.*

$$\begin{gathered} f \sim GP(\mu, K) \Rightarrow \\ (f(X_1), f(X_2), \cdots, f(X_n)) \sim \mathrm{N}\left((\mu(X_1), \cdots, \mu(X_n))\,, \left(K(X_i, X_j)\right)_{i.j=1}^{N}\right)\,, \end{gathered}$$

where $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, $\left(K(X_i, X_j)\right)_{i,j=1}^{N}$ means a $N \times N$ kernel matrix with the (i, j)-element being $K(X_i, X_j)$.

From the definition we can see that different from the DP, the samples X_i's in GP are dependent via the kernel function K, thus is suitable for modeling correlated

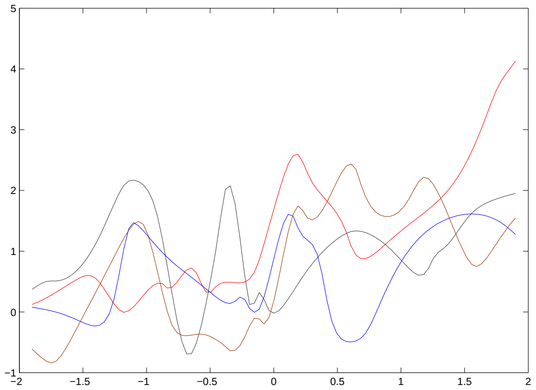


Figure 1.2: Sampled functions drawn from a Gaussian process with zero mean and squared exponential kernel.

observed data. Figure 1.2 shows some sampled functions drawn from a GP with zero mean and *squared exponential* kernel function [Rasmussen and Williams, 2006].

1.1.4 Indian buffet processes

The Indian buffet process (IBP) defines distributions over infinite dimensional binary matrices, or more precisely speaking, over infinite column binary matrices. It can be seen as compositions of Beta processes and Bernoulli processes [Thibaux and Jordan, 2007], but probably the most intuitive way of understanding the IBP is via the Indian buffet metaphor [Griffiths and Ghahramani, 2011], *i.e.*, the IBP defines distributions over the following Indian buffet seating process:

- In an Indian buffet restaurant, there are infinite number of dishes served in a line. N customers come into the restaurant one after another.

- The first customer starts at the left of the buffet, takes the first Poisson(α) dishes and stops as his plate becomes overburdened.

- The i-th customer moves along the buffet, chooses the k-dish with probability $\frac{m_k}{i}$, where m_k is the number of previous customers choosing dish k; having reaching the end of last customer, he tries a Poisson$(\frac{\alpha}{i})$ number of new dishes.

If we use a binary matrix X to record the choices of the customers where each row corresponds to one customer and each column corresponds to one dish, *i.e.*, $X_{ik} = 0$ means dish k was not chosen by customer i, $X_{ik} = 1$ means it was chosen. At the end of the process, we can get a distribution of the choices of the dishes by all the customers, which is essentially a distribution over infinite dimensional binary matrices. We will use IBP(α) to denote the Indian buffet process. Figure 1.3 illustrates a draw from the IBP with $\alpha = 5$.

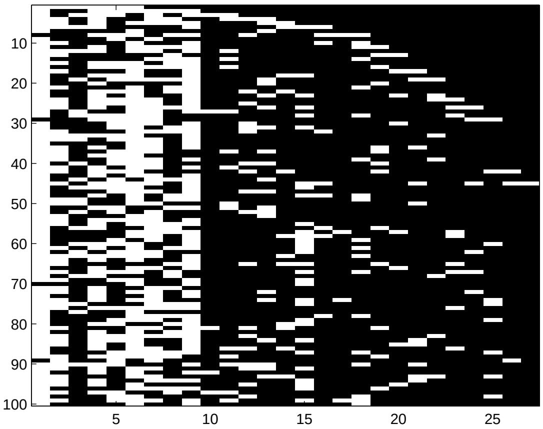


Figure 1.3: A draw from an Indian buffet process, where rows represent customers, columns represent dishes. Color "white" corresponds to value "1", "black" corresponds to "0".

1.1.5 Mondrian processes

The Chinese restaurant process (CRP) [Aldous, 1985], obtained by integrating out the random measure in the Dirichlet process, defines a distribution over a partition of the natural numbers $\mathbb{N} = \{1, 2, \cdots\}$. That is, if we use $z_x \in \mathbb{N}$ to denote the cluster assignment for data x, then the joint distribution of $(z_x : x \in \mathbb{X})$ is called the Chinese restaurant process. We can think of it as a distribution over a one-dimensional space, *e.g.*, the indexes are represented by positive integers which are one dimensional. The Mondrian process [Roy and Teh, 2009] generalizes this to arbitrary dimensional space. For example, in a two-dimensional case, suppose $x \in \mathbb{X}_1$ and $y \in \mathbb{X}_2$ are objects to be modeled ($\mathbb{X}_1$ and $\mathbb{X}_2$ do not necessary be the same space), we define a collocation cluster assignment variable z_{xy} for the pair (x, y) to denote which cluster the pair (x, y) is in. This can be model with a generalization of the CRP, called a two-dimensional Mondrian process, which defines the joint distribution over $(z_{xy} : x \in \mathbb{X}_1, y \in \mathbb{X}_2)$.

We can see from the above example that the Mondrian process is defined via partitioning the n-dimensional index space, this is equivalent to constructing a randomized n-dimensional *kd*-tree, where we sequentially and randomly choose one dimension each time, and then do a random cut on this dimension [Roy and Teh, 2009]. For instance, in the two dimension case where we want to construct a random *kd*-tree on the rectangle $(a, A) \times (b, B)$, we denote the resulting Mondrian process as $m \sim \text{MP}(\lambda, (a, A), (b, B))$, where λ is the parameter of the Mondrian process controlling the number of cuts in the process. Now we do the following cutting:

- Let $\lambda' = \lambda - E$ where $E \sim \text{Exp}(A - a + B - b)$:[2]

[2] $\text{Exp}(x)$ means the exponential distribution with parameter x.

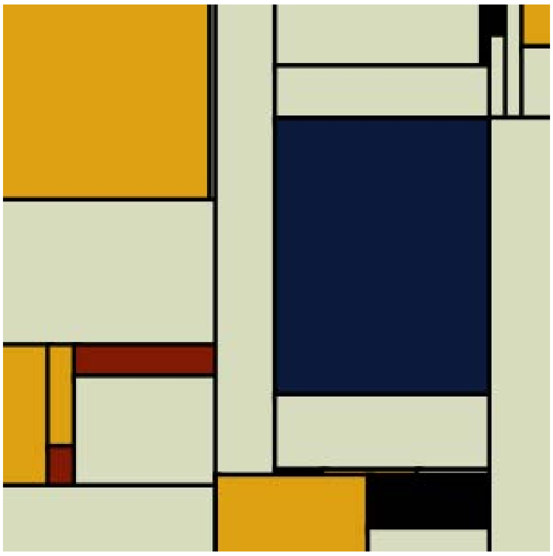


Figure 1.4: A draw from a Mondrian process on space $[0, 1] \times [0, 1]$.

- If $\lambda' < 0$: stop the cutting process, the resulting cutting configuration is a draw from the MP.
- Otherwise, uniformly choose a cutting point on $(a, A) \cup (b, B)$, and do a cut (vertically or horizontally) on the corresponding dimension. Assume the cutting point is on $x \in (a, A)$ (it is similar when $x \in (b, B)$), this results in two independent MPs after the cutting:

  $$m_< \sim \text{MP}(\lambda', (a, x), (b, B)), \quad m_> \sim \text{MP}(\lambda', (x, A), (b, B))$$

  and then we recurse on $m_<$ and $m_>$, respectively.

Roy and Teh [2009] show this construction results in nice theoretical properties including self-consistency, which is essential in defining valid stochastic processes [Çinlar, 2010]. Figure 1.4 shows a draw from a two dimensional Mondrian process, which defines a partition over the two dimensional space $[0, 1] \times [0, 1]$.

1.1.6 Fragmentation-Coagulation processes

The fragmentation-coagulation process (FCP) [Teh et al., 2011] is an instance of the more general Markov jump processes [Rao, 2012]. It defines distributions over continuous time Markov chains with states representing random split and merge operations. This is particularly interesting in modeling sequential data associated with random split and merge phenomena such as genetic variations.

The fragmentation-coagulation process is essentially the dynamic version of the Chinese restaurant process, where in each time, the tables can split or merge according to some probabilities. Formally, it is described via sequentially sitting N customers into the dynamic Chinese restaurant process [Teh et al., 2011] (i is used to

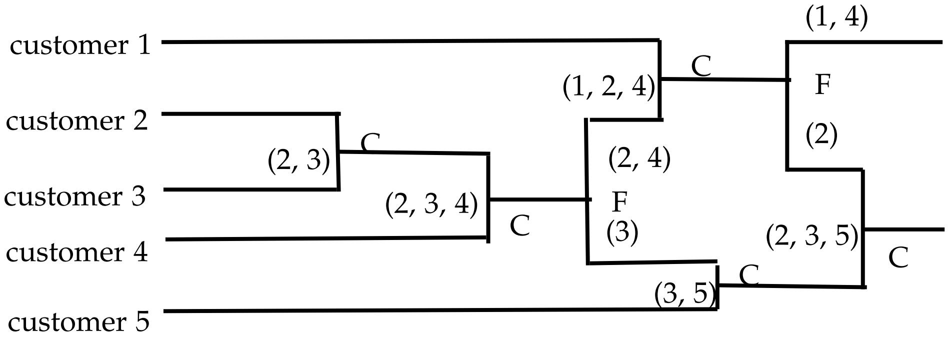


Figure 1.5: A draw from a Fragmentation-Coagulation process, where *F* means fragmentation and C means coagulation.

index customers and t to index time):

- $i = 1$: The first customer sits at a table for the duration of the process.
- $t = 0$: For the subsequent customers, at time $t = 0$, they sit at the tables according to the standard Chinese restaurant process. For the other times, *i.e.*, $t > 0$, the following circumstances might happen:
- If prior to time t, customer i is sitting with some other people:
  - Let i be sitting at table c prior to t, which will split into two tables at time t, then i will join one of the tables with probabilities proportional to the number of customers on those tables.
  - Otherwise if c is to merge with another table, then i will join the combined table.
  - Otherwise, with some probability i will fragment out to create a new table at time t.
- Otherwise the table where i is sitting at will have some probability to merge with other existing tables at time t.

The above description defines a distribution over the dynamic seating arrangements of the Chinese restaurant process, and is called the fragmentation-coagulation process. Figure 1.5 illustrates a realization of the FCP via fragmentation (F) and coagulation (C) events.

1.2 Nonparametric Bayesian Models for Machine Learning

Given well defined nonparametric Bayesian priors such as those described in the previous section, we can apply them to machine learning problems by identifying the likelihood terms and then perform Bayesian inference on the posterior. This section briefly introduces some typical machine learning problems with the nonparametric Bayesian priors defined in the last section.

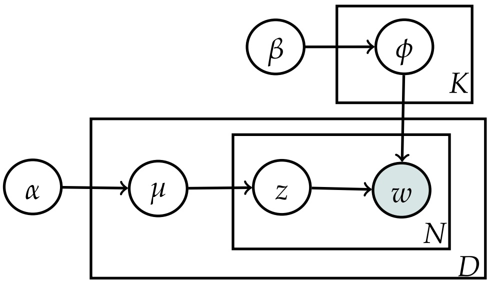


Figure 1.6: LDA topic model.

1.2.1 Hierarchical Dirichlet process latent Dirichlet allocation

This model, known as HDP-LDA, was first proposed by Teh et al. [2006] as a Bayesian nonparametric extension of the popular *latent Dirichlet allocation* (LDA) topic model [Blei et al., 2003]. By using the HDP as the prior for topic distributions in the LDA, it can automatically infer the number of topics in the corpus, allowing much more flexible modeling for documents.

Specifically, in the LDA model, a document is assumed to have the following generating process:

- Draw K topics distributions $\{\boldsymbol{\phi}_k\}$ *i.i.d.* from the Dirichlet distribution with parameter $\boldsymbol{\beta}$:
$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta}), k = 1, 2, \cdots, K$$
- For each document d:
  - Draw its topic distribution $\boldsymbol{\mu}_d$ from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$:
  $$\boldsymbol{\mu}_d \sim \text{Dirichlet}(\boldsymbol{\alpha}) .$$
  - For the ℓ-th word $w_{d\ell}$ in document d:
    - Draw its topic indicator $z_{d\ell}$ from the discrete distribution:
    $$z_{d\ell} \sim \text{Discrete}(\boldsymbol{\mu}_d) .$$
    - Generate the word $w_{d\ell}$ from the corresponding topic:
    $$w_{d\ell} \sim \text{Discrete}(\boldsymbol{\phi}_{z_{d\ell}}) .$$

The corresponding graphical model is shown in Figure 1.6. In HDP-LDA, the topic distribution for each document is replaced with a Dirichlet process, and these DPs are coupled with the same hierarchical structure as the LDA to achieve topic sharing, *e.g.*, they are drawn from a parent DP. This model in essence equals to letting the number of topics K in LDA go to infinite. In formulation, the generative process for the HDP-LDA is described as:

$$\begin{aligned} D_0 \mid \alpha_0, H &\sim \mathrm{DP}(\alpha_0, H) && \text{a common DP} \\ D_d \mid \alpha, D_0 &\sim \mathrm{DP}(\alpha, D_0) && \text{for each document } d \\ \varphi_{d\ell} \mid D_d \sim D_d, \quad w_{d\ell} &\mid \varphi_{d\ell} \sim \mathrm{Discrete}(\varphi_{d\ell}) && \text{for each word } w_{d\ell} \end{aligned}$$

where H is the Dirichlet base distribution, thus each draw $\varphi_{d\ell}$ from D_d is a discrete distribution, which is used to draw the words $w_{d\ell}$ for document d. More details of the model will be introduced in Chapter 4.

1.2.2 Language modeling with Pitman-Yor processes

For a long time since its proposal, the Pitman-Yor process did not find its justification in the modeling, *i.e.*, the distinction between DP and PYP in modeling has not been discovered, in fact, people even found DP is a much better process to use if one were trying to model logarithmic behavior [Arratia et al., 2003]. A first significant work demonstrating the advantage of Pitman-Yor processes over Dirichlet processes is done by Goldwater et al. [2006], where Pitman-Yor processes are introduced to model power-law phenomena in language models. At the same time, Teh [2006a,b] proposed to use Pitman-Yor processes for n-gram lauguage modeling in natural language processing.

Specifically, in n-gram language modeling, we are trying to model the probability of sequences of words $p(w_1, w_2, \cdots, w_n)$. In a k-gram model, it is assumed that word w_i only depends on its previous $k-1$ words, *e.g.*, $p(w_i | w_{i-k+1}, \cdots, w_{i-1}, \boldsymbol{\theta})$ where θ is the model parameter. Using the PYP to model these distributions, we can construct a hierarchy of word distributions, *i.e.*:

$$p(w_i | w_{i-k+1}, \cdots, w_{i-1}, \{\sigma_i, \theta_i\}) \sim \mathrm{PYP}\left(\sigma_i, \theta_i, p(w_{i-1} | w_{i-k}, \cdots, w_{i-2}, \{\sigma_{i-1}, \theta_{i-1}\})\right)$$
$$\text{for } i = 1, 2, \cdots, n\ .$$

In this way, the word distribution of w_i given its previous $k-1$ words is a smooth variant of the word distribution of w_{i-1} given its previous $k-1$ words. The Pitman-Yor process is shown to be superiors to the Dirichlet process in this setting because words in natural language follows the Zip-Law, which is in correspondence with the power-law property in the Pitman-Yor process. In term of posterior inference techniques, the most popular one is based on the Chinese restaurant process representation of the PYP, please refer to Teh [2006b] for details.

1.2.3 Gaussian process regression

The Gaussian process can be used as a prior for regression problems [Rasmussen and Williams, 2006], where the goal is to find a smooth function that fits the data well. The Gaussian process, in essence, meets the requirement of the problem. Furthermore, it provides a nonparametric Bayesian way to do the modeling where we do not need to explicitly specify the form of the regression function, thus is nonparametric. To utilize the conjugacy property, we usually choose a Gaussian likelihood, resulting in the following Gaussian process regression:

$$\begin{aligned} &f \sim \mathrm{GP}(\mu, K) && \text{draw a random function } f \\ &x_i \sim N(x_i; f(x_i), \sigma I), && \text{for } i = 1, 2, \cdots, N \end{aligned}$$

where $x_i \in \mathbb{R}, \sigma > 0$, and I is a $N \times N$ identity matrix. The advantage of the Gaussian likelihood is that it enables analytical prediction formulation under the framework. Other likelihood functions usually do not have this property, but we can use some approximation techniques such as the Laplacian approximation to make the problem analytically and efficiently computable [Rasmussen and Williams, 2006].

1.2.4 Infinite latent feature models

The Indian buffet process is usually used as a nonparametric Bayesian prior for the infinite latent feature models [Zhou et al., 2009; Griffiths and Ghahramani, 2011]. Specifically, with the infinite dimensional matrix Z drawn from the IBP, we take each row z_i as the feature indicator vector for an object, *e.g.*, '1' means the feature is on while '0' means off. In the image modeling [Griffiths and Ghahramani, 2011], image x_i is assumed be generated from a K-dimensional Gaussian distribution with mean $z_i A$ and covariance matrix $\Sigma_X = \sigma_X^2 I$, where A is a $K \times D$ weighting matrix and D is the dimension of the image. In other words, the whole generative process can be written as:

$$\begin{aligned} &Z \sim \mathrm{IBP}(\alpha) && \text{draw feature indicators for all the images} \\ &x_i \sim N(x_i; z_i A, \sigma_X^2 I) && \text{generate image } x_i \end{aligned}$$

To sum up, using the IBP in the latent feature modeling, the number of features for each object can be inferred from the data, providing an advance way of modeling than traditional latent feature models such as principal component analysis [Pearson, 1901].

1.2.5 Relational modeling with Mondrian processes

The Mondrian process defines partitions over n-dimensional positive integer-valued tensors, *i.e.*, it partitions a n-dimensional cubic into a set of blocks, with each block representing a unique cluster adopted with its own parameters, thus it is ideal to be used as a prior for relational data. For example, the author network data could be represented as a 2-dimensional relational data. We use X, Y to represent these two dimensions, if an author $i \in X$ co-authors with another author $j \in Y$ for some paper, we assume that they are in the same block/cluster in the partition, and use z_{ij} to denote the corresponding block index. Roy and Teh [2009] consider the case of 2-dimensional relational data with binary observations, *i.e.*, linked or unlinked. They use the following process for the generative model:

$$\begin{aligned}
M|\lambda &\sim \mathrm{MP}(\lambda, [0.1], [0,1]) && \text{a random partition on } [0,1] \times [0,1] \\
\phi_s &\sim \mathrm{Beta}(a_0, a_1) && \text{parameters for each block in } M \\
\xi_i &\sim \mathrm{Uniform}(0,1) && x\text{-axes for each data } i \\
\eta_i &\sim \mathrm{Uniform}(0,1) && y\text{-axes for each data } i \\
R_{ij}|\xi, \eta, \phi, M &\sim \mathrm{Bernoulli}(\phi_{z_{ij}}) && \text{connection between } i \text{ and } j
\end{aligned}$$

where z_{ij} is determined by matching its x and y values with the block locations in M. It can be seen that generalizing the model to arbitrary dimensional relational data is straightforward, and the Mondrian process represents as a powerful and flexible nonparametric Bayesian prior for the modeling.

1.2.6 Genetic variations modeling with fragmentation-coagulation processes

The construction of the fragmentation-coagulation process (FCP) facilitates it to model time evolving phenomena where latent cluster structures are assumed to be evolving over time. Teh et al. [2011] use the FCP to model SNP sequence (haplotype) evolution in generic applications. In the SNP sequence, there are M locations on a chunk of the chromosome, with simple binary observations. Furthermore, observations are assumed to be Markovian dependent, *e.g.*, the observation on location m on a SNP sequence depends on the observation on location $m-1$. Given a chunk of such sequence, the observations in each location are again assumed to exhibit latent cluster structures (*i.e.*, the observations at a specific location follow a Chinese restaurant process. Given the latent cluster structure at a location, the binary observations are then generated from Bernoulli distributions. Specifically, the whole model can be written as the following hierarchical construction:

$$\begin{aligned}
C &\sim \mathrm{FGP}(\mu, R) && \text{sample a fragmentation-coagulation configuration} \\
\beta_j &\sim \mathrm{Beta}(a, b) && \text{for each location } j \\
\theta_{cj} &\sim \mathrm{Bernoulli}(\beta_j) && \text{for each cluster } c \text{ in location } j \\
p(x_{ij} = \theta_{c_{ij}j}) &= 1 - \epsilon && \text{for each observation } i \text{ in location } j
\end{aligned}$$

where (μ, R) are hyperparameters in the FGP, ϵ is a noise probability, $c_{ij} \in C$ denotes the cluster that i is in at location j, the generative process can be illustrated in Figure 1.7. For more details on the construction, please refer to [Teh et al., 2011].

1.3 Dependent Random Probability Measures

The thesis focuses on a particular nonparametric Bayesian prior called the *random probability measure* (RPM), which is essentially a generalization of the Dirichlet process. In an RPM, each draw is a discrete distribution, and it has close relationship with the Poisson process, which will be explored in the thesis. The construction

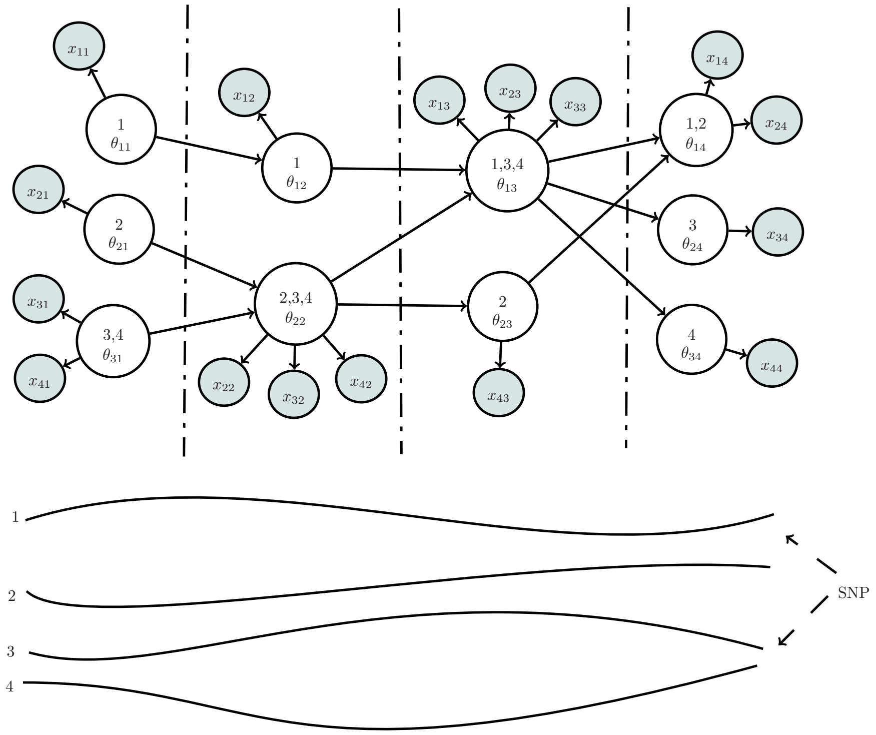


Figure 1.7: SNP sequence example. x_{ij} means the observation in cluster i at location j, θ_{ij} is the binary random variable for cluster i in location j, the numbers $(1,2,3,4)$ denote the sequences. The figure shows 4 SNP sequences of 4 locations.

of an RPM from Poisson processes results in a flexible and large class of random probability measures. Furthermore, due to the analytic distributional probability of the Poisson process, the RPMs discussed in the thesis allow feasible and efficient posterior inference algorithms to be designed.

1.3.1 Poisson processes

The Poisson process is a well studied stochastic process [Kingman, 1993], and is probably the simplest and most intuitive stochastic process in the literature. Because of the simplicity, it has many theoretically nice distributional properties such as complete randomness and analytical integration. This makes it the basis for many other more advanced stochastic processes such as the Dirichlet process [Ferguson, 1973] and the Markov process [Kolokoltsov, 2011; Rao, 2012].

A concise introduction of the Poisson process and its distributional properties will be given in Chapter 2. It will also be clear how the thesis is built on the Poisson process in the rest of the chapters.

1.3.2 Completely random measures

Completely random measures (CRM) are also fundamental in the development of models in this thesis. The CRM is quite similar to the Poisson process in definition but more general. Actually it can be shown that completely random measures can be constructed from Poisson processes as well [Kingman, 1993]. Note similar to the Poisson process, samples from the CRM are also discrete random measures. Thus the CRM also forms the basis for random probability measures studied in this thesis, *e.g.*, an NRM is obtained by doing a normalization step on the CRM. Subsequently, the resulting RPM could inherit some nice properties from the CRM, as will be shown in the thesis.

1.3.3 Random probability measures

Finally, the random probability measure (RPM) forms the main stream of the thesis. It will be shown how to obtain a RPM with the following transformations in the thesis:

$$\text{Poisson process} \overset{\text{linear functional}}{\Longrightarrow} \text{CRM} \overset{\text{transformation}}{\Longrightarrow} \text{RPM}$$

Basically, we first take a linear functional of the Poisson random measure to get a CRM, then we can apply suitable transformations on the CRM to get a RPM, *e.g.*, to get a *normalized random measure* (NRM), the transformation is simply a normalization operation. We can of course perform other more complicated transformations, for example, first apply some transformations on the individual atoms before doing the normalization. Chapter 8 will discuss this idea in more detail.

1.3.4 Dependent random probability measures

A further development on *random probability measures* is the family of *dependent random probability measures*, which has found various important applications in machine learning. From the statistical aspect, dependent random probability measures have also been well studied. After developing the notation of *dependent nonparametric processes* in [MacEachern, 1999], MacEachern [2000] specified a model called *dependent Dirichlet process*, where atoms of a set of Dirichlet processes are characrized by a stochastic process. The idea was further developed in his following papers such as [MacEachern, 2001; MacEachern et al., 2001]. Since then a lot of followup works had been proposed by others. For instance, James [2003] specified the random probability measure with a Gamma process and developed dependent semiparametric intensity models. A related approach but for hazard rate mixtures has been developed recently by Lijoi and Bernardo [2014]. Also, De Iorio et al. [2004] proposed a construction of dependent random measures with an ANOVA model. Finally, a recent work using dependent Bayesian nonparametric models for species data application can be found in [Arbel et al., 2014].

We note that most of the above mentioned works were done from the statistical aspect, this thesis aims to develop dependent random probability measures from the machine learning aspect, where tractability and computational feasibility are two of the major issues. To sum up, the thesis makes a contribution towards this by bridging the gap between statistical tools for RPMs and real applications.

1.4 Thesis Contributions

To be more precise, parts of the thesis are based on my following published joint work

- Parts of Chapter 3 appeared in our technical report [Chen et al., 2012a], where I did the related theory of normalized random measures by borrowing some ideas from existing work in the statistical community. Wray Buntine contributed by simplifying some theorems in the paper and derived the computational formula for $T_{\sigma,M}^{N,K}$ in Section 3.4.1. Both the coauthors helped with polishing the draft and helpful discussions.

- Chapter 4 is unpublished, and built partly based on some theory of hierarchical modeling with normalized random measures in [Chen et al., 2013b, Appendix. D]. All are my individual work.

- Chapter 5 is based on our published paper at International Conference on Machine Learning (ICML). I did the theory and implementation of the model, the other two authors participated in the work with helpful discussion and paper polish.

- Most content of Chapter 6 and Chapter 7 appeared in our published work in International conference on Machine Learning [Chen et al., 2013a], a joint work started from my visit to Yee Whye Teh at UCL. With very inspiration and helpful discussions with Yee Whye Teh and Vinayak Rao, I did the related theory and implementation of the models. All the coauthors especially Vinayak Rao helped with polishing the paper.

- Section 8.3 of Chapter 8 is based on discussion with Yee Whye Teh and Vinayak, as well as Vinayak Rao's unpublished technical report on Poisson-Kingman processes [Rao, 2013]. Other parts are my own work.

- Except wherever stated above, all the other parts of the thesis are done by my own.

Furthermore, some other work during my PhD not included in the thesis include four of our published papers [Chen et al., 2011; Du et al., 2012; Chen et al., 2014a; Lim et al., 2013; Chen et al., 2014b] and one submitted manuscript [Ding et al., 2014].

The contributions of the thesis include:

- *R*eview the theory of Poisson processes and a recent technique called Poisson process partition calculus, and relates it to the well known Palm formula.

- *Transfer the theory of the normalized random measure from the statistical to machine learning community.* The normalized random measure generalizes the Dirichlet process (DP) to a large extend, overcomes some limitations of the DP such as the incapability for power-law distribution modeling, thus is much more flexible in modeling real data. This thesis forms the most extensive research to date in this area.

- *Explore different ideas about constructing dependent normalized random measures.* Existing Bayesian nonparametric models only explore limited dependency structures such as the hierarchical dependency by, for example the hierarchical Dirichlet process (HDP). The dependency models in the thesis not only extend the HDP to hierarchical normalized random measures for more flexible modeling, but also explore other ideas by controlling specific atoms of the underlying Poisson process. This results in many dependency models with abilities to handle dependencies beyond hierarchical dependency such as the Markovian dependency. In addition, by constructing the dependency models in such ways, various distributional properties and posterior structures can be well analyzed, resulting in much more theoretically clean models. These properties cannot be achieved with the hierarchical dependent model.

- *T*est all the models proposed in the thesis extensively with both synthetic data and real data such as documents for topic modeling. Experimental results have shown superior performance compared to realistic baselines, demonstrating the effectiveness and suitability of the proposed models.

1.5 Thesis Outline

The rest of the chapters go as follows:

Chapter 2 This chapter introduces the Poisson process on a general probability space. It starts by relating some real life phenomena with the Poisson process, then introduces the definition of Poisson process, how to define probability distributions on Poisson random measures, and reviews some nice theoretical distributional properties of the Poisson process. Finally, it introduces a recently proposed calculus tool–the Poisson process partition calculus, and relates this with the well known *Palm formula* for Poisson processes, which will be used frequently in the rest of the thesis.

Chapter 3 This chapter introduces the normalized random measure (NRM). Specifically, it shows explicitly how an NRM is constructed from Poisson processes. Some distributional properties and posterior will be analyzed using the Poisson process partition calculus described in Chapter 2. Several forms of the posteriors for the NRM will also be compared and several versions of the posterior sampling algorithms are derived. Finally, experiments on NRM mixtures will be presented.

Chapter 4 This chapter builds hierarchical models to deal with hierarchical correlated data based on the NRM–hierarchical normalized random measures (HNRM). The construction is the same as the hierarchical Dirichlet process (HDP) [Teh et al., 2006] except that NRMs are used instead of DPs. This on the one hand makes the modeling more flexible, *i.e.*, allows modeling of power-law distributions; on the other hand, it complicates the model's posterior structure, requiring more sophisticated methods for posterior inference. Fortunately, it is shown that by introducing auxiliary variables, posterior inference can be performed almost as easily as for the HDP.

Chapter 5 This chapter constructs a time evolving hierarchical normalized random measure for dynamic topic modeling. It borrows ideas from some existing work by allowing topics to be born, to die and vary at each time. Some theoretical analysis of the model as well as the correlation structure induced by the correlated operators are derived, then posterior inference is performed based on an approximate Gibbs sampling algorithm. Finally, experimental results are shown.

Chapter 6 This chapter introduces a more theoretically clean dependency model called *mixed normalized random measures* (MNRM). It extends the work of Rao and Teh [2009] by generalizing the Gamma process to any completely random measure. The idea of the construction is to augment the domain space of the underlying Poisson process with an additional *spatial* space to index the *region* information. Based on such a construction, we can obtain some theoretically nice properties of the MNRM, for example, the dependent RPMs are marginally NRMs, and after marginalization, they form a generalized Chinese restaurant process. The MNRM can also be used to model time evolving documents for topic modeling.

Chapter 7 In addition to the MNRM, this chapter proposes an alternative construction of dependent normalized random measures called *thinned normalized random measures* (TNRM). Instead of weighting the atoms of the Poisson process in each region, TNRM explicitly deletes some of the atoms, resulting in a thinned version of the original NRM and achieving the goal of sparse Bayesian nonparametric modeling. This construction also preserves the property that the DNRMs are marginally NRMs, however, it will result in a much more complex posterior structure than the MNRM. Thanks to the Poisson process partition calculus, a slice sampler can be designed to sample from the posterior. One advantage of the TNRM compared to the MNRM is its sparse representation, making it a preferable choice in modeling.

Chapter 8 This chapter discusses two extensions of the above dependent NRM models. One is to combine the MNRM and TNRM to strengthen the advantages of both; the other is to apply the dependency operators on a more general class of random probability measures called the Poisson-Kingman process. Ideas of the two constructions are described, posterior inferences are also discussed in this chapter.

Chapter 9 This chapter summarizes the thesis and discusses some potential directions for future research.

Chapter 2

Poisson Processes

2.1 Introduce

This chapter introduces the concept of Poisson processes, a basic tool arising in probability theory, then reviews some of the properties of Poisson processes that are proved useful in developing the statistical models in this thesis, and finally introduces some profound calculation methods based on Poisson processes called the *Poisson process partition calculus*.

Generally speaking, the Poisson process generalizes the Poisson distribution, a well studied probability distribution over integers, by extending it to be a distribution over integer valued measures. The Poisson process endows the property of complete randomness [Kingman, 1967], acting as a basic tool for constructing a variety of stochastic processes. A formal introduction of the Poisson process can be found, for example in [Kingman, 1993; Çinlar, 2010].

Turning to practical applications, we note that all the phenomena revealing complete randomness can be reasonably modeled by the Poisson process. For example: the arrival of customers in a bank during the work time can be taken as a Poisson process on the real line $\mathbb{R}^+$; the location of stars over the sky seen from the earth can be deemed as a Poisson process over the surface of a sphere; and more abstractly, the words in a document can also be modeled as a Poisson process on an abstract probability space so that the word count is equal to the value of the random measure induced by the Poisson process.

In the following, the theory of Poisson processes will be built up based on *measure theory*, starting from measurable spaces. The reason for resorting to measure theory is that most of the stochastic processes do not endow density functions, the evaluation of a stochastic process is represented by *measure*, thus we need to borrow tools from *measure theory* to study their distributional properties. Generally speaking, a measurable space is a pair $(E, \mathcal{E})$ where E is a set and $\mathcal{E}$ is a σ-algebra on E, which composes of all the subsets of E that is closed under complements and countable unions [Çinlar, 2010]. Actually, from [Çinlar, 2010], the σ-algebra of a set E can be taken as composed of all the measurable functions: $f : E \longrightarrow \mathbb{R}$, thus the goal of *measure theory* is to study the properties of the measurable functions in that space.

In the rest of this chapter, some basic definitions about probability are reviewed, following by a detailed introduction of Poisson processes and their distributional properties. Most of the proofs are given in the text for completeness and easier understanding.

2.2 Poisson Processes

This section formally defines the Poisson process in a modern probability setting, where everything is defined based on the concept of *probability spaces*.

Definition 2.1 (Probability Spaces [Çinlar, 2010])**.** A probability space is composed of three parts, denoted as $(\Omega, \mathcal{H}, \mathcal{P})$, where the set Ω is called the sample space, its elements are called outcomes. The σ-algebra $\mathcal{H}$ is composed of all subsets of Ω, with elements $H \in \mathcal{H}$ called the events, and finally $\mathcal{P}$ is a probability measure on $(\Omega, \mathcal{H})$.

Given a *probability space*, we can define random variables and their distributions on it. It differs from the traditional view in that now a random variable is defined as a mapping. The random variable may project to a different space to the probability space $(\Omega, \mathcal{H}, \mathcal{P})$ one starts with. This different space is denoted $(\mathcal{S}, \mathbb{S})$ below. Formally:

Definition 2.2 (Random Variables and their Distributions [Çinlar, 2010])**.** Let $(\mathcal{S}, \mathbb{S})$ be a measurable space, a random variable is defined as a mapping $X : \Omega \rightarrow \mathcal{S}$ taking values in $\mathcal{S}$ satisfying[1] that for $\forall A \in \mathbb{S}$,

$$X^{-1}A \triangleq \{X \in A\} \equiv \{\omega \in \Omega : X(w) \in A\} \in \mathcal{H} \ ,$$

thus $X^{-1}A$ is the subset of Ω mapping into A. Under this setting, define the distribution of X as a probability measure P on $(\mathcal{S}, \mathbb{S})$ such that for $\forall A \in \mathbb{S}$,

$$P(A) = \mathcal{P}(X^{-1}A) \equiv \mathcal{P}(\omega \in \Omega : X(\omega) \in A) \ .$$

The notation $P(X = A)$ will be sometimes simplified as $P(A)$ in the thesis and the corresponding density function is denoted as $p(X)$.

A familiar instance of the distributions is the *Poisson distribution*, which is the most elemental object in developing the theory for Poisson processes.

Definition 2.3 (Poisson Distributions)**.** A random variable X taking values in $\mathbb{N} = \{0, 1, \cdots, \infty\}$ is said to have the Poisson distribution with mean λ in $(0, \infty)$ if

$$p(X = k|c) = \frac{e^{-\lambda}\lambda^k}{k!}, k \in \mathbb{N}, \tag{2.1}$$

then $X < \infty$ almost surely and $\mathbb{E}[X] = \mathrm{Var}[X] = \lambda$. In such setting, the Poisson distribution is denoted as $\mathrm{Poi}(\lambda)$.

[1] Where $\triangleq$ means "defined as" and this notation will be used for the rest of the thesis.

Note that a distribution is uniquely identified by its characteristic function, *i.e.*, it is a one to one mapping. Furthermore, the characteristic functional–the generalization of the characteristic function for random variables, is a powerful tool in studying properties of stochastic processes. The following specifies the characteristic function of a Poisson random variable.

Proposition 2.1 (Characteristic Function of Poi(λ))**.** *For a Poisson random variable $X \sim$ Poi(λ), the corresponding characteristic function is given by*

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right] = e^{-\lambda\left(1-e^{it}\right)} ,$$

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

Proof. Let $\theta \in \mathbb{C}$ be any complex numbers, then we have

$$\begin{aligned}\mathbb{E}\left[e^{\theta X}\right] =& \sum_{k=0}^{\infty} e^{\theta k}\frac{\lambda^k}{k!}e^{-\lambda} \\ =& e^{-\lambda}\sum_{k=0}^{\infty}\frac{\left(\lambda e^{\theta}\right)^k}{k!} \\ =& e^{-\lambda}e^{\lambda e^{\theta}} = e^{-\lambda\left(1-e^{\theta}\right)} .\end{aligned}$$

Letting $\theta = it$ completes the proof. □

Now one can define *Poisson processes* and the induced *Poisson random measures*. Basically this extends the Poisson distribution to an arbitrary measurable space.

Definition 2.4 (Poisson Processes & Poisson Random Measures [Çinlar, 2010])**.** Let $(\mathcal{S}, \mathbb{S})$ be a measure space (called the *state space*). Let $\nu(\cdot)$ be a measure on it. A Poisson process on $\mathcal{S}$ is defined to be a random subset $\Pi \in \mathbb{S}$ such that if $\mathcal{N}(A)$ is the number of points of Π in the measurable subset $A \subseteq \mathcal{S}$, then

a) $\mathcal{N}(A)$ is a random variable having the Poisson distribution with mean $\nu(A)$,

b) whenever $A_1, \cdots, A_n$ are in $\mathcal{S}$ and disjoint, the random variables $\mathcal{N}(A_1), \cdots, \mathcal{N}(A_n)$ are independent.

The induced integer-value random measure $\mathcal{N}(\cdot)$ is called a *Poisson random measure* and the Poisson process is denoted as $\Pi \sim$ PoissonP(ν), where ν is called the mean measure of the Poisson process.

Figure 2.1 shows a realization of the Poisson process on the 2D space with mean measure ν as uniform distribution.

Note that the Poisson random measure $\mathcal{N}$ is an infinite dimensional object, *i.e.*, it is a random integer-valued function and can be evaluated for $\forall A \in \mathcal{S}$. To describe a Poisson random measure, we construct a product space based on the state space $\mathcal{S}$: $(\mathcal{F}, \mathbb{F}) \triangleq (\bigcup_{n=1}^{\infty}\mathcal{S}^n, \bigcup_{n=1}^{\infty}\mathbb{S}^n)$, where $\mathcal{S}^n = \otimes_{i=1}^{n}\mathcal{S}$ and $\mathbb{S}^n = \otimes_{i=1}^{n}\mathbb{S}$. A realization of

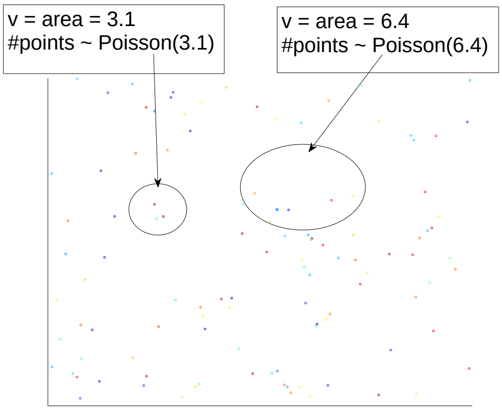


Figure 2.1: A draw from a Poisson process on space $\mathbb{R} \times \mathbb{R}$ with mean measure as the uniform distribution, can can be interpreted as *area*.

a Poisson process can thus be regarded as a random element on $(\mathcal{F}, \mathbb{F})$, represented as $\Pi = (\Pi_1, \Pi_2, \cdots)$, *i.e.*, it is an infinite sequence. Now for each element $\Pi \in \mathcal{F}$, define $\mathcal{N}_\Pi(A) = \#\{A \cap \Pi\}$ for $\forall A \in \mathbb{S}$ to be the #points Π_k in A, this is an integer value measure on $\mathcal{S}$. Considering all elements in the space $\mathcal{F}$, the construction then induces a probability measure $\mathcal{P}$ on a set of measures $\{\mathcal{N}_\Pi : \Pi \in \mathcal{F}\}$, *i.e.*, a distribution over the realization of the Poisson process (see [Theorem 2.15 Çinlar, 2010] for detailed assignment of measures for each realization, but note that the assignment should satisfy *Kolmogorov's consistency condition*). This probability measure is called the probability laws of the Poisson random measure $\mathcal{N}$. As is shown in [Daley and Vere-Jones, 1998], the space where $\mathcal{N}$ lies on is the space of boundedly finite measures, we denote it as $(\mathcal{M}, \mathbb{M})$ thus $\mathcal{P}$ is a probability measure on space $\mathcal{M}$.

Definition 2.5 (Probability Laws of Poisson Random Measures [Çinlar, 2010])**.** The distribution of elements (each corresponds to one realization of a Poisson process) of space $(\mathcal{F}, \mathbb{F})$ is called the probability law of the Poisson random measure $\mathcal{N}$, denoted by $\mathcal{P}(\mathcal{N}|\nu)$,[2] where ν is the mean measure of the corresponding Poisson process.

Given the definition of probability law of a (Poisson) random measure, it is now sensible to talk about the expectation over a stochastic process (Poisson process). Specifically, let $g : \mathcal{M} \to \mathbb{R}$ be a measurable function, then the expectation of $g(\mathcal{N})$ over $\mathcal{N}$ is defined as

$$\mathbb{E}\left[g(\mathcal{N})\right] \triangleq \int_{\mathcal{M}} g(\mathcal{N})\mathcal{P}(\mathrm{d}\mathcal{N}|\nu) \ . \qquad (2.2)$$

2.3 Properties

The Poisson process has many attractive properties. Below some representative ones are listed. Most of the proofs for the theorems can be found, for example in [Kingman, 1993; Çinlar, 2010].

[2] Sometimes it is also written as $\mathcal{P}_\nu(\mathcal{N})$.

The first theorem is about the disjointness of Poisson processes, which states that the points in two independent Poisson processes are disjoint almost surely. Intuitively this is true since the mean measure of the Poisson process is non-atomic, leading to that the probability of assigning a point $w \in \mathcal{S}$ with counts two (*i.e.*, the two Poisson processes both contain the point w) is equal to zero.

Theorem 2.2 (Disjointness Theorem [Kingman, 1993]). *Let ν_1, ν_2 be diffuse measures*[3] *on $\mathcal{S}$, $\Pi_1 \sim PoissonP(\nu_1)$ and $\Pi_2 \sim PoissonP(\nu_2)$ be independent Poisson processes on $\mathcal{S}$, let $A \in \mathbb{S}$ be a measurable set such that $\nu_1(A)$ and $\nu_2(A)$ are finite. Then Π_1 and Π_2 are disjoint with probability 1 on A:*

$$\mathcal{P}(\Pi_1 \cap \Pi_2 \cap A = \emptyset) = 1 \ .$$

The next theorem, the *Restriction Theorem*, discloses a "symmetric" property of the Poisson process, which states that restricting a Poisson process to a measurable set results in another Poisson process with mean measure restricted to that set.

Theorem 2.3 (Restriction Theorem [Kingman, 1993; Çinlar, 2010]). *Let $\Pi \sim PoissonP(\nu)$ be a Poisson process on $\mathcal{S}$ and let $\mathcal{S}_1$ be a measurable subset of $\mathcal{S}$. Then the random countable set $\Pi_1 = \Pi \cap \mathcal{S}_1$ can be regarded either as a Poisson process on $\mathcal{S}$ with mean measure $\nu_1(A) = \nu(A \cap \mathcal{S}_1), \forall A \in \mathbb{S}$, or as a Poisson process on $\mathcal{S}_1$ whose mean measure is the restriction of ν to $\mathcal{S}_1$.*

An informal proof of the above theorem goes as follows: Consider a Poisson process on $\mathcal{S}$, which is essentially a random set of points, according to the definition of Poisson processes, the average number of points in an infinitesimal region $\mathrm{d}w$ of $\mathcal{S}$ would be $\nu(\mathrm{d}w)$, thus the average number of points in $\mathcal{S}_1$ would be $\nu(\mathcal{S}_1)$ by integrating over $\mathcal{S}_1$, which is the mean measure of Π_1. Furthermore the randomness property of Π_1 follows from Π.

The *Restriction Theorem* can be thought of as an *intersection* operation. On the other hand, it is natural to define a *joint* operation, which is essentially the idea behind the following *Superposition Theorem*, stating that by *joining* a countable set of independent Poisson process results in another Poisson process with an updated mean measure.

Theorem 2.4 (Superposition Theorem [Kingman, 1993; Çinlar, 2010]). *Let $(\Pi_i \sim PoissonP(\nu_i))$ be a countable set of independent Poisson processes on $\mathcal{S}$. Then their superposition: $\Pi = \cup_{i=1}^{\infty} \Pi_i$ is a Poisson process with measure measure $\nu = \sum_{i=1}^{\infty} \nu_i$. Consequently, the corresponding Poisson random measure of Π is $\mathcal{N} = \sum_{i=1}^{\infty} \mathcal{N}_i$, where $\mathcal{N}_i$ is the Poisson random measure corresponding to Π_i.*

The intuition behind the proof of the *Superposition Theorem* is straightforward: by the *disjointness* of independent Poisson processes in Theorem 2.2, the #points of Π in $A \in \mathbb{B}$ would be $\sum_i \mathcal{N}_i(A)$, resulting in the same form for the mean measure. The Poisson distributed property of $\mathcal{N}(A)$ follows from the additive property of Poisson random variables and the randomness of Π follows from the randomness of Π_i's.

[3] A measure is called diffuse measure if and only if there are no minimal non-empty sets.

Given a Poisson process, it would be worthwhile to ask what is the resulting process if we take a transformation $f(w)$ for each point $(w \in \Pi)$ of the Poisson process. Interestingly, under certain conditions, the resulting process is still a Poisson process with different mean measure. Now let us denote the transformed space be $\mathcal{T}$ and let $B \subseteq \mathcal{T}$, by transformation f, this would induce an integer random measure $\mathcal{N}^*$ on $\mathcal{T}$ denoting #points in B, defined by

$$\mathcal{N}^*(B) = \#\{f(\Pi) \cap B\} \ . \tag{2.3}$$

Now if $f(w)(w \in \Pi)$ are distinct, we have

$$\mathcal{N}^*(B) = \#\{w \in \Pi : f(w) \cap B\} = \mathcal{N}(f^{-1}(B)) \ ,$$

where $\mathcal{N}$ is the Poisson random measure for Π. Thus the probability law of $\mathcal{N}^*(B)$ is $\mathcal{P}(\mathcal{N}^*(B)) = \mathcal{P}(\mathcal{N}(f^{-1}(B)))$. This is shown to be a Poisson random measure in the following *Mapping Theorem*.

Theorem 2.5 (Mapping Theorem [Kingman, 1993; Çinlar, 2010])**.** *Let Π be a Poisson process with σ-finite mean measure ν on space $(\mathcal{S}, \mathbb{S})$, let $f : \mathcal{S} \to \mathcal{T}$ be a measurable function such that the induced random measure (2.3) has no atoms. Then $f(\Pi)$ is a Poisson process on $\mathcal{T}$ with mean measure $\nu_f(B) \triangleq \nu(f^{-1}(B)), \forall B \in \mathbb{T}$ where $\mathbb{T}$ is the σ-algebra of $\mathcal{T}$.*

As is discussed in [Rao, 2012], a common mapping function f is the projection operator $\pi : \mathcal{S} \times \mathcal{T} \to \mathcal{T}$, which maps a Poisson process on a product space $\mathcal{S} \times \mathcal{T}$ down to a subspace $\mathcal{T}$, forming a new Poisson process. Another direction, *lifting* a Poisson process onto a higher dimensional space through a transition probability kernel, is done by the following *Marking Theorem*. Note the transition kernel in the *Marking Theorem* in [Kingman, 1993] is deterministic, while it is possible to generalize it to be a random probability kernel, see for example [Çinlar, 2010]. The following *Marking Theorem* adapts to this case.

Theorem 2.6 (Marking Theorem [Çinlar, 2010])**.** *Let $\Pi \sim PoissonP(\nu)$ be a Poisson process on $\mathcal{W}$, Q a transition probability kernel from $(\mathcal{W}, \mathbb{W})$ into $(\Theta, \mathcal{B}(\Theta))$. Assume that given $\Pi \triangleq \{w_i\}$, the variables $\theta_i \in \Theta$ are conditionally independent and have the respective distribution $Q(w_i, \cdot)$. Then*

1) $\{\theta_i\}$ forms a Poisson process on $(\Theta, \mathcal{B}(\Theta))$ with mean measure $(\nu Q)(\mathrm{d}\theta) \triangleq \int_{\mathcal{W}} \nu(w) Q(w, \mathrm{d}\theta) \mathrm{d}w$;

2) (w_i, θ_i) forms a Poisson process on $(\mathcal{W} \times \Theta, \mathbb{W} \otimes \mathcal{B}(\Theta))$ with mean measure $\nu \times Q$, defined as $(\nu \times Q)(\mathrm{d}w, \mathrm{d}\theta) \triangleq \nu(\mathrm{d}w) Q(w, \mathrm{d}\theta)$.

Remark 2.7. Note the usual notation for *Marking Theorem* is the statement 2) in Theorem 2.6, statement 1) can be seen as a generalization of the *Mapping Theorem* in Theorem 2.5 by letting the function f to be random. Also 1) can be obtained from 2) by applying the *Restriction Theorem* of Theorem 2.3.

The *Marking Theorem* 2.6 is quite general. By defining different transformation kernels Q, we can get many variants of it. There are two special cases of Q that are interesting and will be used in constructing dependent random measures in the following chapters, which are called *subsampling* and *point transition*. It will be shown here that these are defined via two specific transition kernels Q's, for the definition, see for example [Lin et al., 2010; Chen et al., 2012a] for details.

Definition 2.6 (Subsampling of Poisson processes). Subsampling of a Poisson process with sampling rate $q(w)$ is defined to be selecting the points of the Poisson process via independent Bernoulli trials with acceptance rate $q(w)$. It forms a new Poisson process with atoms (w_i, z_i) where $z_i \in \{0, 1\}$. This is equivalent to defining the transition kernel in Theorem 2.6 as

$$\begin{cases} \Theta = \{0, 1\} \\ Q(w, 1) = q(w) \\ Q(w, 0) = 1 - q(w) \end{cases}$$

Definition 2.7 (Point transition of Poisson processes). Point transition of a Poisson process Π on space $(\mathcal{W}, \mathbb{W})$, denoted as $T(\Pi)$, is defined as moving each point of the Poisson process independently to other locations following a probabilistic transition kernel $Q : \mathbb{W} \times \mathcal{W} \to [0, 1]$ such that Q in Theorem 2.6 satisfying

$$\begin{cases} \Theta = \mathcal{W} \\ Q(w, \cdot) \text{ is a probability measure} & \forall w \in \mathcal{W} \\ Q(\cdot, A) \text{ is integrable} & \forall A \in \mathbb{W} \end{cases}$$

It follows directly from Theorem 2.6 that subsampling and transition form new Poisson processes with modified mean measures:

Corollary 2.8 (Subsampling Theorem). *Let $\Pi \sim PoissonP(\nu)$ be a Poisson process on the space $\mathcal{W}$ and $q : \mathcal{W} \to [0, 1]$ be a measurable function. If we independently draw $z_w \in \{0, 1\}$ for each $w \in \Pi$ with $P(z_w = 1) = q(w)$, and let $\Pi_k = \{w \in \Pi : z_w = k\}$ for $k = 0, 1$, then Π_0 and Π_1 are independent Poisson processes on $\mathcal{W}$ with $S^{1-q}(\Pi) \triangleq \Pi_0 \sim PoissonP((1 - q)\nu)$ and $S^q(\Pi) \triangleq \Pi_1 \sim PoissonP(q\nu)$.*

Corollary 2.9 (Transition Theorem). *Let $\Pi \sim PoissonP(\nu)$ be a Poisson process on space $(\mathcal{W}, \mathbb{W})$, $\mathcal{T}$ a probability transition kernel*[4]*, and denote $T(\Pi)$ the resultant Poisson process after transition, then*

$$T(\Pi) \sim PoissonP(\nu\mathcal{T}). \tag{2.4}$$

where $\nu\mathcal{T}$ can be considered as a transformation of measures over $\mathcal{W}$ defined as $(\nu\mathcal{T})(A) := \int_{\mathcal{W}} \mathcal{T}(w, A)\nu(\mathrm{d}w)$ for $A \in \mathcal{W}$.

The proof of Theorem 2.6 is done through an application of the Laplace functional formula for (w_i, θ_i). An important result of the Laplace function related to the

[4] We replace Q in Theorem 2.6 with $\mathcal{T}$ to emphasis it as a transition kernel.

Poisson process is known as *Campbell's Theorem* below, which is particularly useful in calculations involving Poisson random measures. For completeness a proof of the theorem is included following [Çinlar, 2010, Theorem 3.2], but will be delayed after introducing Theorem 2.10, which will be used in the proof of Theorem 2.6.

Theorem 2.10 (Campbell's Theorem [Kingman, 1993; Çinlar, 2010])**.** *Let* Π *be a Poisson process on* $\mathcal{S}$ *with mean measure* ν*,* $f : \mathcal{S} \longrightarrow \mathbb{R}$ *be a measurable function. Denote the Poisson random measure of* Π *as* $\mathcal{N}$ *with probability law* $\mathcal{P}(\mathrm{d}\mathcal{N}|\nu)$*,* $\mathcal{M}$ *as the space of boundedly finite measures. Define the following sum*

$$\Sigma = \sum_{w\in\Pi} f(w) \equiv \int_{\mathcal{S}} f(w)\mathcal{N}(\mathrm{d}w) \tag{2.5}$$

1. Σ *is absolutely convergent with probability one if and only if*

$$\int_{\mathcal{S}} \min\left(|f(w)|, 1\right) \nu(\mathrm{d}w) < \infty\,. \tag{2.6}$$

2. *Under the condition (2.6):*

$$\mathbb{E}\left[e^{\theta\Sigma}\right] \triangleq \int_{\mathcal{M}} e^{\theta\Sigma}\mathcal{P}(\mathrm{d}\mathcal{N}|\nu) \tag{2.7}$$

$$= \exp\left\{\int_{\mathcal{S}} \left(e^{\theta f(w)} - 1\right) \nu(\mathrm{d}w)\right\}\,, \tag{2.8}$$

where $\theta \in \mathbb{C}$*. Moreover,*

$$\mathbb{E}\left[\Sigma\right] = \int_{\mathcal{S}} f(w)\nu(\mathrm{d}w)\,, \tag{2.9}$$

$$var(\Sigma) = \int_{\mathcal{S}} f(w)^2\nu(\mathrm{d}w)\,. \tag{2.10}$$

Proof. We first prove the formula (2.7). To do this, first assume f to be simple, *i.e.,*

$$f = \sum_{k=1}^{K} f_k 1_{S_k}\,,$$

where $\cup_{k=1}^{K} S_k = \mathcal{S}$ and $S_k \cap S_{k'} = \emptyset (\forall k \neq k')$ is a partition of $\mathcal{S}$. Then it is easily seen that

$$\Sigma = \sum_{w\in\Pi} f(w) = \sum_{k=1}^{K} f_k N(S_k) \triangleq \sum_{k=1}^{K} f_k N_k\,.$$

Now for $\theta \in \mathbb{C}$, we have

$$\begin{aligned}
\mathbb{E}\left[e^{\theta\Sigma}\right] &= \prod_{k=1}^{K} \mathbb{E}\left[e^{\theta f_k N_k}\right] \\
&\overset{(1^*)}{=} \prod_{k=1}^{K} \exp\left\{(e^{\theta f_k} - 1)\nu(S_k)\right\} \\
&\overset{(2^*)}{=} \exp\left\{\sum_{k=1}^{K} \int_{S_k} (e^{\theta f(x)} - 1)\nu(\mathrm{d}x)\right\} \\
&= \exp\left\{\int_{\mathcal{S}} (e^{\theta f(x)} - 1)\nu(\mathrm{d}x)\right\} \ , \qquad (2.11)
\end{aligned}$$

where (1^*) follows by the fact that N_k is a Poisson random variable with mean $\nu(S_k)$ and by applying Proposition 2.1, (2^*) follows by the definition of f.

Since we know that any positive measurable function f can be expressed as the limit of an increasing sequence (f^j) of simple functions, moreover, any measurable function can be expressed as the difference between two positive functions, thus by monotone convergence argument we conclude that for any measurable function f, (2.7) follows. This completes the proof of (2.7).

To prove condition (2.6), let $\theta = -u$ where $u > 0$ in (2.7), which results in

$$\mathbb{E}\left[e^{-u\Sigma}\right] = \exp\left\{-\int_{\mathcal{S}} \left(1 - e^{-uf(w)}\right)\nu(\mathrm{d}w)\right\} \ . \qquad (2.12)$$

If (2.6) holds, from (2.12) we see that for small u we have

$$\begin{aligned}
\lim_{u\to 0} \mathbb{E}\left[e^{-u\Sigma}\right] &= \exp\left\{-\int_{\mathcal{S}} uf(w)\nu(\mathrm{d}w)\right\} \\
&\xrightarrow{u\to 0} \exp(0) = 1 \ ,
\end{aligned}$$

meaning that Σ is a finite random variable. On the other hand, if (2.6) does not hold, for a fixed $u > 0$, we can find a f satisfying $uf(x)$ is small enough for $\forall x \in \mathcal{S}$ such that

$$\begin{aligned}
1 - e^{-uf(x)} &\longrightarrow uf(x) \ , \\
\Rightarrow \mathbb{E}\left[e^{-u\Sigma}\right] &\longrightarrow 0 \ .
\end{aligned}$$

This implies $\Sigma = \infty$ with probability 1. Thus (2.6) is proved.

To prove (2.9), take the derivative on two sides of (2.7) with respect to θ we have

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}\left[e^{\theta\Sigma}\right] = \frac{\mathrm{d}}{\mathrm{d}\theta}\exp\left\{\int_{\mathcal{S}} \left(e^{\theta f(w)} - 1\right)\nu(\mathrm{d}w)\right\}$$

$$\begin{aligned}\Rightarrow \mathbb{E}\left[\Sigma e^{\theta\Sigma}\right] &= \exp\left\{\int_{\mathcal{S}}\left(e^{\theta f(w)}-1\right)\nu(\mathrm{d}w)\right\}\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathcal{S}}\left(e^{\theta f(w)}-1\right)\nu(\mathrm{d}w)\\ &= \exp\left\{\int_{\mathcal{S}}\left(e^{\theta f(w)}-1\right)\nu(\mathrm{d}w)\right\}\int_{\mathcal{S}} f(w)e^{\theta f(w)}\nu(\mathrm{d}w)\end{aligned}$$

Letting $\theta = 0$ we obtain the formula (2.9).

To prove (2.10), similar to the proof of (2.9), taking the second derivative of (2.7) with respect to θ and letting $\theta = 0$, we get

$$\mathbb{E}\left[\Sigma^2\right] = \left(\int_{\mathcal{S}} f(w)\nu(\mathrm{d}w)\right)^2 + \int_{\mathcal{S}} f(w)^2\nu(\mathrm{d}w)\ .$$

Using the relation $\mathrm{var}(\Sigma) = \mathbb{E}\left[\Sigma^2\right] - \mathbb{E}\left[\Sigma\right]^2$ we obtain the formula (2.10). □

To show how Campbell's Theorem is used, I prove Theorem 2.6 with the above theorem in the following.

Proof of Theorem 2.6. The second statement will be proved here since the first one can be easily derived from it. Let M be the random measure formed by (w_i, θ_i), then for positive real-value $\tilde{f}$ in $\mathcal{W}\otimes\mathcal{B}(\Theta)$,

$$\mathcal{L}(\Pi_1, (\theta_i)) \triangleq e^{-M\tilde{f}} = e^{-\sum_i \tilde{f}(w_i,\theta_i)} = \prod_i e^{-\tilde{f}(w_i,\theta_i)}\ .$$

So the conditional expectation of $e^{-M\tilde{f}}$ given Π_1 is

$$\begin{aligned}\mathcal{L}(\Pi_1) &\triangleq \int_{\mathcal{M}} \mathcal{L}(w_i, \mathrm{d}\theta)\\ &= \prod_i \int_{\mathcal{M}} Q(w_i, \mathrm{d}\theta)e^{-\tilde{f}(w_i,\theta)}\end{aligned}$$

Let $f(w) = log(\int_{\mathcal{M}} Q(w, \mathrm{d}\theta)e^{-\tilde{f}(w,\theta)})$ in the Campbell's Theorem, applying the result of (2.7) we have

$$\begin{aligned}\mathbb{E}\left[e^{-Mf}\right] &= \mathbb{E}\left[\mathcal{L}(\Pi_1)\right]\\ &= \exp\left\{-\int_{\mathcal{W}}(1-\int_{\mathcal{M}} Q(w,\mathrm{d}\theta)e^{-\tilde{f}(w,\theta)})\nu(\mathrm{d}w)\right\}\\ &= \exp\left\{-\int_{\mathcal{W}}(1-\int_{\mathcal{M}} Q(w,\mathrm{d}\theta)e^{-\tilde{f}(w,\theta)})\nu(\mathrm{d}w)\right\}\\ &= \exp\left\{-\int_{\mathcal{W}}(\int_{\mathcal{M}} Q(w,\mathrm{d}\theta)-\int_{\mathcal{F}} Q(w,\mathrm{d}\theta)e^{-\tilde{f}(w,\theta)})\nu(\mathrm{d}w)\right\}\\ &= \exp\left\{-\int_{\mathcal{W}\times\mathcal{M}}(1-e^{-\tilde{f}(w,\theta)})Q(w,\mathrm{d}\theta)\nu(\mathrm{d}w)\right\}\end{aligned}$$

According to the unity of a Poisson random measure and its Laplace functional, we conclude M is a Poisson process with mean measure $\nu \times Q$. □

Real applications involving Poisson processes often require evaluating the expectation of a functional form under Poisson random measures, *e.g.*, sometimes we want to calculate the posterior of a model constructed from Poisson processes on $\mathcal{S}$ given observations. In such cases, we can think of the expectation to be taken with respect to the *joint distribution* of a Poisson random measure $\mathcal{N}$ and some fixed points in $\mathcal{S}$, we denote this joint distribution as $\mathcal{P}(\cdot|\nu, \boldsymbol{w})$. To define the joint distribution, we should rely on measure theory by first defining a *joint measure* for a Poisson process and some fixed points.

Formally, let us define this joint distribution from the following construction. Recall that essentially the Poisson process defines a mapping from the probability space $(\Omega, \mathcal{H}, \mathcal{P})$ to space $(\mathcal{F}, \mathbb{F})$ in Definition 2.5, with the state space being $(\mathcal{S}, \mathbb{S})$. Now for an event $A \in \mathbb{F}$ (which is essentially a functional defined on the space $\mathcal{F}$) and an event $W \in \mathbb{S}$ (similarly is a function on $\mathcal{S}$), define Campbell's measure $\mathcal{C}(A, W)$ as

$$\mathcal{C}(A, W) = \mathbb{E}\left[\mathbb{I}_A(\Pi)\Pi(W)\right] \ ,$$

where $\Pi \in \mathcal{F}$ means the measure induced by a realization of a Poisson process, $\mathbb{I}_A(\cdot)$ is an indicator function. Essentially, the above measure gives the expected #points in W given the Poisson processes in the even A. Since for any A, $\mathcal{C}(A, \cdot)$ is absolutely continuous *w.r.t.* the mean measure ν, this allows us to define the Radon-Nikodym derivative of $\mathcal{C}(A, \cdot)$:

$$\mathcal{P}_A(\cdot|\nu, w) \triangleq \frac{\mathrm{d}\mathcal{C}(A, \cdot)}{\mathrm{d}\nu}(w) \ .$$

$\mathcal{P}_A(\cdot|\nu, w)$ defines a distribution with respect to (ν, w) (up to a constant), and is called the *Palm distribution* [Bertoin, 2006; Daley and Vere-Jones, 1998]. The subscript A will be omitted when A is taken to be the whole space $\mathcal{F}$ for simplicity, written as $\mathcal{P}(\cdot|\nu, w)$. This describes how the joint measure changes with respect to the mean measure ν, and is exactly the *joint distribution* of $\mathcal{N}$ and one point $w \in \mathcal{S}$ (up to a constant) that we want to defined above. We can also think of $\mathcal{P}_A(\cdot|\nu, w)$ as the posterior distribution of the Poisson process given an observation w. Note that by the independence of the Poisson process and w, $\mathcal{P}(\cdot|\nu, w)$ can be written by disintegrating as $\mathcal{P}(\mathrm{d}\mathcal{N}|\nu, w) = \mathcal{N}(\mathrm{d}w)\mathcal{P}(\mathrm{d}\mathcal{N}|\nu)$. Based on this, the *Palm formula* [Bertoin, 2006] or called *Slivnyak's theorem* [Slivnyak, 1962] and *Mecke's Formula* [Serfoso, 1999] states that the distribution of a prior Poisson random measure $\mathcal{N}$ under $\mathcal{P}(\cdot|\nu, w)$ is identical to the distribution of $\mathcal{N} + \delta_w$. Specifically, we have

Theorem 2.11 (Palm Formula [Slivnyak, 1962; Bertoin, 2006]). *Let Π be a Poisson process on $\mathcal{S}$ with mean measure ν and Poisson random measure $\mathcal{N}$, $f : \mathcal{S} \to \mathbb{R}$ and $G : \mathcal{S} \times \mathcal{N} \to \mathbb{R}$ are functions of $w \in \mathcal{S}$ and of $(w, \mathcal{N})$ respectively, then the following formula holds:*

$$\mathbb{E}\left[\int_{\mathcal{S}} f(w)G(w, \mathcal{N})\mathcal{N}(\mathrm{d}w)\right] = \int_{\mathcal{S}} \mathbb{E}\left[G(w, \delta_w + \mathcal{N})\right] f(w)\nu(\mathrm{d}w) \ . \tag{2.13}$$

In another form, it can be represented as

$$\int_{\mathcal{M}}\int_{\mathcal{S}} f(w)G(w,\mathcal{N})\mathcal{N}(\mathrm{d}w)\mathrm{P}(\mathcal{N}|\nu) = \int_{\mathcal{S}}\int_{\mathcal{M}} G(w,\delta_w+\mathcal{N})\mathcal{P}(\mathrm{d}\mathcal{N}|\nu)f(w)\nu(\mathrm{d}w) \ . \tag{2.14}$$

The case of representing the function G in the above *Palm formula* as a positive process can be found in [Çinlar, 2010, Theorem 6.2]. Now instead of starting from the *Palm distribution* to prove the *Palm formula* in Theorem 2.11, an alternative proof is given in the next section from a general results of the *Poisson process partition calculus*, which can be found in Theorem 2.13.

2.4 Poisson Process Partition Calculus

Poisson process partition calculus [James, 2002, 2005] is a general framework targeted at analyzing the posterior structure of random measures constructed from Poisson random measures. It has a number of general results as shown in [James, 2002, 2005], but I will only list a few below for the purpose of this thesis.

Let $\mathcal{N}$ be a Poison random measure defined on a complete and separable space $\mathcal{S}$ with mean measure ν, the Laplace functional of $\mathcal{N}$ obtained by applying Theorem 2.10 is denoted as follows

$$\mathcal{L}_{\mathcal{N}}(f|\nu) = \int_{\mathcal{M}} e^{-\mathcal{N}(f)}\mathcal{P}(\mathrm{d}\mathcal{N}|\nu) \triangleq \int_{\mathcal{M}} e^{-\int_{\mathcal{S}} f(w)\mathcal{N}(\mathrm{d}w)}\mathcal{P}(\mathrm{d}\mathcal{N}|\nu) \tag{2.15}$$

$$= \exp\left\{-\int_{\mathcal{S}}(1-e^{-f(w)})\nu(\mathrm{d}w)\right\} \ , \tag{2.16}$$

where $f : \mathcal{S} \to \mathbb{R}^+$ is a measurable function.

Theorem 2.12 ([James, 2005])**.** *Let $f : \mathcal{S} \to \mathbb{R}^+$ be measurable and $g : \mathcal{M} \to \mathbb{R}$ be a function on $\mathcal{M}$, then the following formula holds*

$$\int_{\mathcal{M}} g(\mathcal{N})e^{-\mathcal{N}(f)}\mathcal{P}(\mathrm{d}\mathcal{N}|\nu) = \mathcal{L}_{\mathcal{N}}(f|\nu)\int_{\mathcal{M}} g(\mathcal{N})\mathcal{P}(\mathrm{d}\mathcal{N}|e^{-f}\nu) \ , \tag{2.17}$$

where $\mathcal{P}(\mathrm{d}\mathcal{N}|e^{-f}\nu)$ is the law of a Poisson process with intensity $e^{-f(w)}\nu(\mathrm{d}w)$. In other words, the following absolute continuity result holds:

$$e^{-\mathcal{N}(f)}\mathcal{P}(\mathrm{d}\mathcal{N}|\nu) = \mathcal{L}_{\mathcal{N}}(f|\nu)\mathcal{P}(\mathrm{d}\mathcal{N}|e^{-f}\nu) \ , \tag{2.18}$$

meaning that exponentially tilting (with $e^{-\mathcal{N}(f)}$) a Poisson random measure with mean measure ν ends up another Poisson random measure with an updated mean measure $e^{-f}\nu$.

Proof. By the unity of Laplace functionals for random measure on $\mathcal{S}$, it suffices to check this result for the case $g(\mathcal{N}) = e^{-\mathcal{N}(h)}$ for $h : \mathcal{S} \to \mathbb{R}^+$. Then we have

$$\int_{\mathcal{M}} g(\mathcal{N}) e^{-\mathcal{N}(f)} \mathcal{P}(\mathrm{d}\mathcal{N}|\nu) = \int_{\mathcal{M}} e^{-\mathcal{N}(f+h)} \mathcal{P}(\mathrm{d}\mathcal{N}|\nu) \tag{2.19}$$

$$= \exp\left\{ - \int_{\mathcal{S}} \left(1 - e^{-(f+h)(w)}\right) \nu(\mathrm{d}w) \right\} \tag{2.20}$$

$$= \exp\left\{ - \int_{\mathcal{S}} \left((1 - e^{-f(w)})\nu(\mathrm{d}w) + (e^{-f(w)} - e^{-(f+h)(w)})\nu(\mathrm{d}w) \right) \right\} \tag{2.21}$$

$$= \exp\left\{ - \int_{\mathcal{S}} \left(1 - e^{-f(w)}\right) \nu(\mathrm{d}w) \right\} \exp\left\{ - \int_{\mathcal{S}} \left(1 - e^{-h(w)}\right) e^{-f(w)} \nu(\mathrm{d}w) \right\} \tag{2.22}$$

$$= \mathcal{L}_{\mathcal{N}}(f|\nu) \int_{\mathcal{M}} g(\mathcal{N}) \mathcal{P}(\mathrm{d}\mathcal{N}|e^{-f}\nu) \ . \tag{2.23}$$

□

I develop the following theorem to reveal the relation between Theorem 2.12 and the celebrated Palm formula in Theorem 2.11. Specifically, the Palm formula is seen to be a special case of the result in Theorem 2.12.

Theorem 2.13. *The Palm formula is a special case of Theorem 2.12.*

Proof. In Theorem 2.12, let

$$g(\mathcal{N}) = \int_{\mathcal{S}} f(w) \mathcal{N}(\mathrm{d}w) \ ,$$

according to (2.17), we have

$$\int_{\mathcal{M}} g(\mathcal{N}) e^{-\mathcal{N}(f)} \mathcal{P}(\mathrm{d}\mathcal{N}|\nu) = \mathbb{E}\left[\int_{\mathcal{S}} f(w) e^{-\mathcal{N}(f)} \mathcal{N}(\mathrm{d}w) \right] \tag{2.24}$$

$$= \mathbb{E}\left[e^{-\mathcal{N}(f)} \right] \int_{\mathcal{M}} \int_{\mathcal{S}} f(w) \mathcal{N}(\mathrm{d}w) \mathcal{P}(\mathrm{d}\mathcal{N}|e^{-f}\nu) \tag{2.25}$$

$$= \mathbb{E}\left[e^{-\mathcal{N}(f)} \right] \int_{\mathcal{S}} f(w) e^{-f(w)} \nu(\mathrm{d}w) \tag{2.26}$$

$$= \int_{\mathcal{S}} \mathbb{E}\left[e^{-\mathcal{N}(f)} \right] f(w) e^{-f(w)} \nu(\mathrm{d}w) \tag{2.27}$$

$$= \int_{\mathcal{S}} \mathbb{E}\left[e^{-(\mathcal{N}+\delta_w)(f)} e^{f(w)} \right] f(w) e^{-f(w)} \nu(\mathrm{d}w) \tag{2.28}$$

$$= \int_{\mathcal{S}} \mathbb{E}\left[e^{-(\mathcal{N}+\delta_w)(f)} \right] f(w) \nu(\mathrm{d}w) \tag{2.29}$$

$$= \int_{\mathcal{S}} \int_{\mathcal{M}} e^{-(\mathcal{N})(f)} \mathcal{P}(\mathcal{N}|\nu, w) f(w) \nu(\mathrm{d}w) \ , \tag{2.30}$$

where (2.24) follows by definition, (2.25) by applying (2.17), (2.26) by applying (2.9) of Theorem 2.10, (2.27) by moving the expectation term into the integral, ((2.29)) by moving $e^{f(w)}$ out and cancellation, (2.30) by a change of measure: $\tilde{\mathcal{N}} = \mathcal{N} + \delta_w$.

Now if we let

$$G(w, \mathcal{N}) = e^{-\mathcal{N}(f)} \ ,$$

we get the Palm formula as in Theorem 2.11:

$$\int_{\mathcal{M}}\int_{\mathcal{S}} f(w)G(w,\mathcal{N})\mathcal{N}(\mathrm{d}w)\mathrm{P}(\mathcal{N}|\nu) = \int_{\mathcal{S}}\int_{\mathcal{M}} G(w,\delta_w+\mathcal{N})\mathcal{P}(\mathrm{d}\mathcal{N}|\nu,w)f(w)\nu(\mathrm{d}w) \ .$$

□

Under the framework of the Poisson process partition calculus, the Palm formula of Theorem 2.11 can be easily extended to the case of conditioning on a set of points, *e.g.*, $\boldsymbol{w} = (w_1,\cdots,w_n) \in \underbrace{\mathcal{S}\otimes\cdots\otimes\mathcal{S}}_{n} \triangleq \mathcal{S}^n$. I call it the *extended Palm formula*, which is stated in Theorem 2.14 and adapted from [James, 2002, Lemma 2.2].

Theorem 2.14 (Extended Palm Formula [James, 2002])**.** *Let* Π *be a Poisson process on* $\mathcal{S}$ *with mean measure* ν *and Poisson random measure* $\mathcal{N}$*,* $f : \mathcal{S}^n \to \mathbb{R}$ *and* $G : \mathcal{S}^n \times \mathcal{N} \to \mathbb{R}$ *are functions of* $\boldsymbol{w} \in \mathcal{S}^n$ *and of* $(\boldsymbol{w},\mathcal{N})$ *respectively, then the following formula holds:*

$$\mathbb{E}\left[\int_{\mathcal{S}^n} f(\boldsymbol{w})G(\boldsymbol{w},\mathcal{N})\prod_{i=1}^{n}\mathcal{N}(\mathrm{d}w_i)\right] = \sum_{\mathbf{p}}\int_{\mathcal{S}^{n(\mathbf{p})}}\mathbb{E}\left[G(\boldsymbol{w},\sum_{i=1}^{n(\mathbf{p})}\delta_{w_i}+\mathcal{N})\right]\prod_{i=1}^{n(\mathbf{p})} f(w_i)\nu(\mathrm{d}w_i) \ , \tag{2.31}$$

where $\mathbf{p}$ *means a partition over integers* $(1,2,\cdots,n)$*,* $n(\mathbf{p})$ *means the #ties in this partition configuration, and* $\sum_{\mathbf{p}}$ *means sum over all the partitions of* $(1,2,\cdots,n)$*.*

Theorem 2.14 is seen true by iteratively applying the *Palm formula* on the joint distribution of the Poisson random measure $\mathcal{N}$ and the points $\boldsymbol{w}$. A proof within the Poisson process partition calculus framework can be found in [James, 2002].

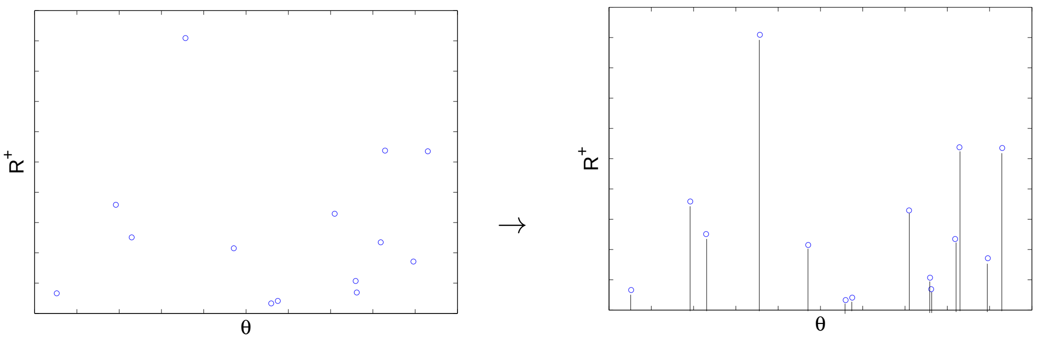
Chapter 3

Normalized Random Measures

3.1 Introduction

In Bayesian nonparametric modeling, a particularly important prior is the prior for discrete distributions since there are a lot of interesting applications involving the inference of such distributions, *e.g.*, topic distributions in topic models, word distributions for topics in text modeling, as well as mixing distributions in general mixture models. By employing priors on these discrete distributions, they form random probability measures (PRM). The Poisson process introduced in the last chapter constitutes the foundation for constructing random probability measures. This chapter introduces normalized random measures (NRM) based on this framework, reviews and extends related underlying theory for the NRM. Generally speaking, an NRM can be regarded as a random discrete distribution represented as $\sum_{k=1}^{\infty} w_k \delta_{\theta_k}$, where both w_k and θ_k are appropriate random variables with flexible probability distribution functions. Based on the NRM, an instance called *normalized generalized Gamma processes* (NGGs) is specified where the distribution of (w_k, θ_k) has a specific form. The NGG is a particular kind of the NRM that is theoretical nice (*e.g.*, with the power-law property) and computationally feasible (*e.g.*, efficient Gibbs samplers with an analytic posterior), and will be frequently used in the rest of the thesis.

This chapter is structured as follows: first some mathematical background of completely random measures (CRMs) and their construction from Poisson processes is introduced in Section 3.2. The CRM is fundamental because the concepts of NRMs and NGGs are built on it. Posterior inference for the NRM is developed in Section 3.3, then sampling formulas for the NGG is elaborated in Section 3.4 based on the marginal posterior results of [James et al., 2009]; in addition to the marginal sampler, slice sampler for NRMs is also introduced using techniques from [Griffin and Walker, 2011] in Section 3.5.2. Experiments for testing the NGG mixture with different samplers are described in Section 3.6. Proofs are given in the Appendix of Section 3.8.



Counting process:

$\mathcal{N}(\cdot) = \sum_k \delta_{(w_k,\theta_k)}(\cdot)$

Completely random measure:

$\tilde{\mu}(\cdot) = \sum_k w_k \delta_{\theta_k}(\cdot)$

Figure 3.1: Constructing a completely random measure from a counting random measure $\mathcal{N}(\cdot,\cdot)$ with points at (w_k, θ_k).

3.2 Completely Random Measures

This section briefly introduces background of completely random measures and the corresponding normalized random measures. Section 3.2.1 explains how to construct completely random measures from Poisson processes. Constructing normalized random measures (NRMs) from CRMs is discussed in Section 3.2.2 along with details of the NGG, a particular kind of NRM for which the details have been worked out.

First an illustration of the basic construction for an NRM for a target domain Θ is given. The Poisson process is used to create a countable (and usually) infinite set of points in a product space of $\mathbb{R}^+ \times \Theta$, as shown in the left of Figure 3.1. The resulting distribution is then a discrete one on these points, which can be pictured by dropping lines from each point (w, θ) down to $(0, \theta)$, and then normalizing all these lines so their sum is one. The resulting picture shows the set of weighted impulses that make up the constructed CRM on the target domain.

3.2.1 Constructing completely random measures from Poisson processes

The general class of completely random measure (CRM) [Kingman, 1967], usually admits a unique decomposition as the summation over three parts: 1) a deterministic measure, 2) a purely atomic measure with fixed atom locations and 3) a discrete measure with random jumps and atoms. However it suffices to consider only the part with random jumps and atoms in real applications. This is also called pure jump processes [Ferguson and Klass, 1972], which has the following form

$$\tilde{\mu} = \sum_{k=1}^{\infty} w_k \delta_{\theta_k} , \tag{3.1}$$

where w_k's and θ_k's are all random variables; $w_1, w_2, \cdots > 0$ are called the jump sizes of the process, and $\theta_1, \theta_2, \cdots$ are a sequence of independent random variables drawn from a base measurable space $(\Theta, \mathcal{B}(\Theta))$[1]. Note that (w_k, θ_k)'s drawn from the Poisson process will also be called *atoms* throughout the thesis.

It is shown that these kinds of CRMs can be constructed from Poisson processes with specific mean measures $\nu(\cdot)$. Formally, given a *Poisson random measure* $\mathcal{N}$ with mean measure ν, denote the corresponding Poisson process as $\Pi \sim \text{PoissonP}(\nu)$, a *completely random measure* can be constructed as

Definition 3.1 (Completely Random Measure)**.** A *completely random measure* $\tilde{\mu}$ defined on $(\Theta, \mathcal{B}(\Theta))$ is defined to be a linear functional of the Poisson random measure $\mathcal{N}$, with mean measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$ defined on a product space $\mathbb{S} = R^+ \otimes \Theta$:

$$\tilde{\mu}(B) = \int_{\mathbb{R}^+ \times B} w \mathcal{N}(\mathrm{d}w, \mathrm{d}\theta), \forall B \in \mathcal{B}(\Theta). \tag{3.2}$$

The mean measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$ is called the *Lévy measure* of $\tilde{\mu}$.

The general treatment of constructing random measures from Poisson random measures can be found in [James, 2005]. Note that the random measure $\tilde{\mu}$ in construction (3.2) has the same form as Equation (3.1) because $\mathcal{N}(\cdot)$ is composed of a countable number of points. It can be proven to satisfy the conditions of a completely random measure [Kingman, 1967] on Θ, meaning that for arbitrary disjoint subsets $\{A_i \in \Theta\}$ of the measurable space, the random variables $\{\tilde{\mu}(A_i)\}$ are independent.

For the completely random measure defined above to always be finite, it is necessary that $\int_{\mathbb{R}^+ \times \Theta} w\, \nu(\mathrm{d}w, \mathrm{d}\theta)$ be finite, and therefore for every $z > 0$, $\nu([z, \infty) \times \Theta) = \int_z^\infty \int_\Theta \nu(\mathrm{d}w, \mathrm{d}\theta)$ is finite [Kingman, 1993]. It follows that there will always be a finite number of points with jumps $w_k > z$ for that $z > 0$. Therefore in the bounded product space $[z, \infty) \otimes \Theta$ the measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$ is finite. So it is meaningful to sample those atoms (w_k, θ_k) with $w_k > z$ by first getting the count of points K sampled from a Poisson with (finite) mean $\nu([z, \infty) \times \Theta)$, and then to sample the K points according to the distribution of $\frac{\nu(\mathrm{d}w, \mathrm{d}\theta)}{\nu([z,\infty) \times \Theta)}$. This provides one way of sampling a completely random measure.

Without loss of generality, it is assumed that the Lévy measure of Equation (3.2) can be decomposed as

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = M\rho_\eta(\mathrm{d}w|\theta)H(\mathrm{d}\theta) ,$$

where η denotes the hyper-parameters if any[2], $H(\mathrm{d}\theta)$ is a probability measure on Θ so $H(\Theta) = 1$, and M is called the *mass* parameter of the Lévy measure. Note the total measure of $\rho_\eta(\mathrm{d}w|\theta)$ is not standardized in any way so in principle some mass could also appear in $\rho_\eta(\mathrm{d}w|\theta)$. The mass is used as a concentration parameter for the random measure.

A realization of $\tilde{\mu}$ on Θ can be constructed by sampling from the underlying Poisson process in a number of ways, either in rounds for decreasing bounds z using

[1] $\mathcal{B}(\Theta)$ means the σ-algebra of Θ, we sometimes omit this and use Θ to denote the measurable space.

[2] The subscript η in ρ_η might sometimes be omitted for simplicity in the rest of the thesis.

the logic just given above, or by explicitly sampling the jumps in order. The later goes as follows [Ferguson and Klass, 1972]:

Lemma 3.1 (Sampling a CRM). *Sample a CRM $\tilde{\mu}$ with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta) = M\rho_\eta(\mathrm{d}w|\theta)H(\mathrm{d}\theta)$ as follows.*

- *Draw i.i.d. samples θ_k from the base measure $H(\mathrm{d}\theta)$.*
- *Draw the corresponding weights w_k for these i.i.d. samples in decreasing order, which goes as:*
  - *Draw the largest jump w_1 from the cumulative distribution function $P(w_1 \leq j_1) = \exp\left\{-M\int_{j_1}^{\infty} \rho_\nu(\mathrm{d}w|\theta_1)\right\}$.*
  - *Draw the second largest jump w_2 from the cumulative distribution function $P(w_2 \leq j_2) = \exp\left\{-M\int_{j_2}^{j_1} \rho_\nu(\mathrm{d}w|\theta_2)\right\}$.*
  - *· · ·*
- *The random measure $\tilde{\mu}$ then can now be realized as $\tilde{\mu} = \sum_k w_k \delta_{\theta_k}$.*

As a random variable is uniquely determined by its Laplace transformation, the random measure $\tilde{\mu}$ is uniquely characterized by its *Laplace functional* or more precisely the *characteristic functional* through the Lévy-Khintchine representation of a Lévy process [Çinlar, 2010], which is stated in Lemma 3.2.

Lemma 3.2 (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 $\mathbb{R}^+ \otimes \Theta$ with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$. For any measurable function $f : \mathbb{R}^+ \times \Theta \longrightarrow \mathbb{R}^+$, the following formula holds:*

$$\begin{aligned} \mathbb{E}\left[e^{-\tilde{\mu}(f)}\right] &\triangleq \mathbb{E}\left[e^{-\int_\Theta f(w,\theta)\mathcal{N}(\mathrm{d}w,\mathrm{d}\theta)}\right] \\ &= \exp\left\{-\int_{\mathcal{W}\times\Theta}\left(1-e^{-f(w,\theta)}\right)\nu(\mathrm{d}w,\mathrm{d}\theta)\right\}, \end{aligned} \tag{3.3}$$

where the expectation is taken over the space of bounded finite measures. The formula can be proved using Campbell's Theorem in Chapter 2. Using (3.3), the characteristic functional of $\tilde{\mu}$ is given by

$$\varphi_{\tilde{\mu}}(u) \triangleq \mathbb{E}\left[e^{\int_\Theta iu\tilde{\mu}(\mathrm{d}\theta)}\right] = \exp\left\{-\int_{\mathcal{W}\times\Theta}\left(1-e^{iuw}\right)\nu(\mathrm{d}w,\mathrm{d}\theta)\right\}, \tag{3.4}$$

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

Now instead of dealing with $\tilde{\mu}$ itself, we deal with $\nu(\mathrm{d}w, \mathrm{d}\theta)$, which is called the Lévy measure of $\tilde{\mu}$, whose role in generating the measure via a Poisson process was explained above.

In the case where the jumps w_k's of the measure are independent on the data θ's, *i.e.*, $\rho_\eta(\mathrm{d}w|\theta) = \rho_\eta(\mathrm{d}w)$, $\tilde{\mu}$ is called homogeneous, which simplifies the calculations a lot and will be considered in this thesis. When f does not depend on θ, (3.3) simplifies to

$$\mathbb{E}\left[\exp\left\{-f\,\tilde{\mu}(B)\right\}\right] = \exp\left\{-M\,p(B)\int_{\mathbb{R}^+}\left[1-\exp\left\{-wf\right\}\right]\rho_\eta(\mathrm{d}w)\right\}\,. \tag{3.5}$$

Note the term inside the exponential plays an important role in subsequent theory, so it is given a name.

Definition 3.2 (Laplace exponent). The *Laplace exponent*, denoted as $\psi_\eta(f)$ for a CRM with parameters η is given by

$$\begin{aligned}\psi_\eta(f) &= \int_{\mathbb{R}^+\times\Theta}\left[1-\exp\left\{-wf\right\}\right]\nu(\mathrm{d}w,\mathrm{d}\theta)\\ &= M\int_{\mathbb{R}^+}\left[1-\exp\left\{-wf\right\}\right]\rho_\eta(\mathrm{d}w) \qquad \text{(homogeneous case)}\,.\end{aligned} \tag{3.6}$$

Note that to guarantee the positiveness of jumps in the random measure, $\rho_\eta(\mathrm{d}w)$ in the Lévy measure must satisfy $\int_0^\infty \rho_\eta(\mathrm{d}w) = +\infty$ [Regazzini et al., 2003], which leads to the following equations:

$$\psi_\eta(0) = 0, \qquad \psi_\eta(+\infty) = +\infty. \tag{3.7}$$

That $\psi_\eta(f)$ is finite for finite positive f implies (or is a consequence of) $\int_0^\infty w\rho_\eta(\mathrm{d}w)$ being finite.

Remark 3.3. There are thus four different ways to define or interpret a CRM:

1. via the linear functional of Equation (3.2),
2. through the Lévy-Khintchine representation of Equation (3.3) using the Laplace exponent,
3. sampling in order of decreasing jumps using Lemma 3.1, and
4. sampling in blocks of decreasing jump values as discussed before Lemma 3.1.

3.2.2 Normalized random measures

Given a completely random measure, a normalized random measure is obtained by simply transforming it to be a probability measure, as given by the definition below.

Definition 3.3 (Normalized Random Measure (NRM)). Based on (3.2), a normalized random measure on $(\Theta, \mathcal{B}(\Theta))$ is defined as[3]

$$\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\Theta)}. \tag{3.8}$$

[3] In this thesis, we always use μ to denote a normalized random measure, while use $\tilde{\mu}$ to denote its unnormalized counterpart if not explicitly stated.

The idea of constructing a random probability measure by normalizing a subordinator, in the spirit of [Ferguson, 1973] for the Dirichlet case, can be found in literature such as [Kingman, 1975; Perman et al., 1992; Pitman, 2003; James, 2005]. While James [2002] had considered constructions of general normalized random measures, Regazzini et al. [2003] studied the problem of normalizing a completely random measure defined on $\mathbb{R}$. This is termed normalized random measures with independent increment (NRMI) and the existence of such random measures is proved. This idea can be easily generalized from $\mathbb{R}$ to any parameter space Θ, *e.g.*, Θ being the Dirichlet distribution space in topic modeling. Also note that the idea of normalized random measures can be taken as doing a transformation, denoted as $Tr(\cdot)$, on completely random measures, that is $\mu = Tr(\tilde{\mu})$. In the normalized random measure case, $Tr(\cdot)$ is a transformation operator such that $Tr(\tilde{\mu}(\Theta)) = 1$. A concise survey of other kinds of transformations can be found in [Lijoi and Prunster, 2010].

Taking different Lévy measures $\nu(\mathrm{d}w, \mathrm{d}\theta)$ of (3.3), we can obtain different NRMs. Throughout the thesis, the following notation is used to denote a NRM:

$$\mu \sim \mathrm{NRM}(\eta, M, H(\cdot)) \ ,$$

where M is the total mass parameter, which usually needs to be sampled in the model, and $H(\cdot)$ is called the base probability measure where θ_k's are drawn from, η is the set of other hyper-parameters to the measure on the jumps, depending on the specific NRMs (in the DP case, η is empty). Note that a specific class of NRMs called *normalized generalized Gamma processes* (NGG) is interesting and has useful distributional properties. To introduce this, let us start with the unnormalized version–*generalized Gamma processes*:

Definition 3.4 (Generalized Gamma Process)**.** Generalized Gamma processes are completely random measures proposed by Brix [Brix, 1999] for constructing shot noise Cox processes. They have the Lévy measures as

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = \frac{e^{-bw}}{w^{1+\sigma}} \mathrm{d}w H(\mathrm{d}\theta), \text{ where } b > 0, 0 < \sigma < 1. \tag{3.9}$$

By normalizing the generalized Gamma process as in (3.8), we obtain the normalized generalized Gamma process (NGG).

Sometimes we also need the Gamma distribution. Because there are several parameterizations of this in use, for clarification, we define it here.

Definition 3.5 (Gamma distribution)**.** The Gamma distribution has two parameters, shape a and scale b, and is denoted $\mathrm{Ga}(a, b)$ with density function

$$p(x|a, b) \ = \ \frac{1}{\Gamma(a) b^a} x^{a-1} e^{-bx} \ .$$

The Lévy measure of the *generalized Gamma process* can be formulated in different ways [Favaro and Teh, 2013], some via two parameters while some via three parameters, but they can be transformed to each other by using a change of variable formula. For ease of representation and sampling, we convert the NGG into a unified form with two parameters using the following lemma.

Lemma 3.4. *Let a normalized random measure be defined using Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$. Then scaling w by $\lambda > 0$ yields an equivalent NRM up to a factor. That is, the normalized measure obtained using $\nu(\mathrm{d}w/\lambda, \mathrm{d}\theta)$ is equivalent to the normalized measure obtained using $\lambda\,\nu(\mathrm{d}t, \mathrm{d}\theta)$.*

By this lemma, without loss of generality, we can instead represent the NGG by absorbing the parameter b into the mass parameter M. Thus the case of $b = 1$ will be used in the rest of the thesis.

Definition 3.6 (Normalized Generalized Gamma)**.** The NGG with shape parameter σ, total mass (or concentration) parameter M and base distribution $H(\cdot)$, denoted $\mathrm{NGG}(\sigma, M, H(\cdot))$, has Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta) = M\rho_\sigma(\mathrm{d}w)H(\mathrm{d}\theta)$ where[4]

$$\rho_\sigma(\mathrm{d}w) \;=\; \frac{\sigma}{\Gamma(1-\sigma)}\frac{e^{-w}}{w^{1+\sigma}}\mathrm{d}w \; .$$

Note that similar to the two parameter Poisson-Dirichlet process (or Pitman-Yor process) [Pitman and Yor, 1997], the normalized generalized Gamma process with $\sigma \neq 0$ can also produce power-law phenomenon, making it different from the Dirichlet process and suitable for applications where long tail distributions are preferable, *e.g.*, *topic-word* distributions in topic modeling. The following power-law property had been well developed in statistical literature such as [Pitman, 2003; Lijoi et al., 2007; James, 2013], and the explict form of the random variable $S_{\sigma,M}$ below was first given in [Pitman and Yor, 1997].

Proposition 3.5 (Power-law of NGG)**.** *Let K_n be the number of components induced by the NGG with parameter σ and mass M or the Dirichlet process with total mass M. Then for the NGG, $K_n/n^\sigma \to S_{\sigma,M}$ almost surely, where $S_{\sigma,M}$ is a strictly positive random variable parameterized by σ and M. For the DP, $K_n/\log(n) \to M$.*

Figure 3.2 demonstrates the power law phenomena in the NGG compared to the Dirichlet process (DP). It is sampled using the generalized Blackwell-MacQueen sampling scheme [James et al., 2009]. Each data to be sampled can choose an existing cluster or create a new cluster, resulting in K clusters with N data points in total.

Many familiar stochastic processes are special/limiting cases of normalized generalized Gamma processes, *e.g.*, *Dirichlet processes* arise when $\sigma \to 0$. *Normalized inverse-Gaussian processes* (N-IG) arise when $\sigma = \frac{1}{2}$ and $b = \frac{1}{2}$. If $b \to 0$, we get the *σ-stable process*, and if $\sigma \to 0$ and b depends on x, we get the *extended Gamma process*. These special classes are listed below.

- **Dirichlet Processes:**

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = \frac{e^{-w}}{w}\mathrm{d}wH(\mathrm{d}\theta) \tag{3.10}$$

[4] Sometimes $\rho_\sigma(w)$ is simply written as $\rho(w)$ by dropping the subscript σ for notation simplicity without causing any confusion.

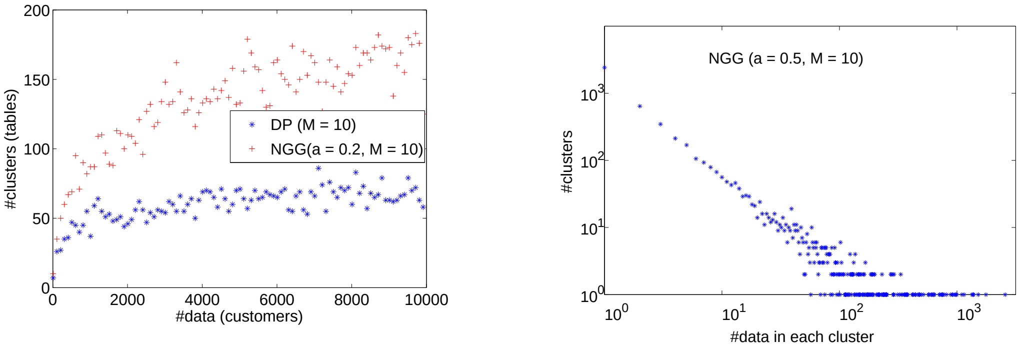


Figure 3.2: Power-law phenomena in NGG. The first plot shows the #data versus #clusters compared with DP, the second plot shows the size s of each cluster versus total number of clusters with size s.

- **Normalized inverse-Gaussian (N-IG) Processes:**

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}w}}{w^{\frac{3}{2}}} \mathrm{d}w H(\mathrm{d}\theta) \tag{3.11}$$

- **Normalized Generalized Gamma Processes:**

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = \frac{e^{-bw}}{w^{1+\sigma}} \mathrm{d}w H(\mathrm{d}\theta), b > 0, 0 \le \sigma < 1 \tag{3.12}$$

- **Extended Gamma Processes:**

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = \frac{e^{-\beta(\theta)w}}{w} \mathrm{d}w H(\mathrm{d}\theta), \beta(\cdot) \text{ is a function} \tag{3.13}$$

- **σ-Stable Processes:**

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = \frac{\sigma}{\Gamma(1-\sigma)w^{1+\sigma}} \mathrm{d}w H(\mathrm{d}\theta), 0 \le \sigma < 1 \tag{3.14}$$

Finally, the concept of *latent relative mass* is introduced here, which is particularly useful in posterior computation for the NRM. The reason for this is that the definition of a NRM naturally contains the divisor $\tilde{\mu}(\Theta) = \sum_{k=1}^{\infty} w_k$, and thus likelihoods of the NRM should involve powers of $\tilde{\mu}(\Theta)$. To facilitate the computation, a trick widely used is to do data augmentation via the Gamma identity [James, 2005].

Definition 3.7 (Latent relative mass)**.** Consider the case where N data are observed. By introducing the auxiliary variable, called *latent relative mass*, $U_N = \Gamma_N / \tilde{\mu}(\Theta)$

where $\Gamma_N \sim \text{Gamma}(N, 1)$, then it follows that

$$\frac{1}{\tilde{\mu}(\Theta)^N} p(\Gamma_N) \mathrm{d}\Gamma_N = \frac{U_N^{N-1}}{\Gamma(N)} e^{-U_N \tilde{\mu}(\Theta)} \mathrm{d}U_N$$

Thus the N-th power of the normalizer can be replaced by an exponential term in the jumps which factorizes, at the expense of introducing the new latent variable U_N. This treatment results in an exponential tilting of the NRM, making marginalization over the corresponding Poisson process feasible via the Poisson process partition calculus framework [James, 2005] described in Chapter 2 and will be seen more clear in the following sections. The idea of this latent variable originates from [James, 2005] and is further explicitly studied in [James et al., 2006, 2009; Griffin and Walker, 2011; Favaro and Teh, 2013], *etc*.

3.3 Posterior of the NRM

This section develops posterior formulas for general normalized random measures with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$. Specifically, the NRM hierarchical model is described as:

$$\begin{aligned} \mu &\sim \text{NRM}(\eta, M, H) && \text{sample a NRM} \\ x_i | \mu &\sim \mu && \text{for } i = 1, \cdots, N \end{aligned}$$

Now let $\theta_1, \cdots, \theta_K$ be the K unique values among $x_{1:N}$, with θ_k occurring n_k times. Let $\tilde{\mu}$ represents the corresponding CRM of μ. Intuitively we have the following conditional likelihoods:

$$\begin{aligned} \tilde{\mu} | x_{1:N} &= \tilde{\mu}^* + \sum_{k=1}^{K} w_k \delta_{\theta_k} \\ p(x_{1:N} | \tilde{\mu}) &= \frac{\prod_{k=1}^{K} \tilde{\mu}(\theta_k)^{n_k}}{\tilde{\mu}(\Theta)^N} , \end{aligned}$$

where $\tilde{\mu}^*$ represents the remaining CRM without observations.

It can be seen that integrating the NRM from $p(x_{1:N} | \tilde{\mu})$ is difficult with a normalizer, so a data augmentation is done here by introducing the *latent relative mass* auxiliary variable U_N defined in Definition 3.7, resulting in

$$p(x_{1:N}, U_N | \tilde{\mu}) = \frac{U_N^{N-1}}{\Gamma(N)} e^{-U_N \left(\tilde{\mu}^*(\Theta) + \sum_{k=1}^{K} w_k\right)} \prod_{k=1}^{K} w_k^{n_k} .$$

Now the posterior is obtained by applying the *Palm Formula* in Theorem 2.11 and the *Lévy-Khintchine Formula* in Lemma 3.2 to get

$$
\begin{aligned}
p(x_{1:N}, U_N|\eta, M) &= \mathbb{E}\left[p(x_{1:N}, U_N|\tilde{\mu})\right] \\
&= \mathbb{E}_{\tilde{\mu}^*}\left[\int_{\underbrace{\mathbb{R}^+ \times \cdots \mathbb{R}^+}_{K}} p(x_{1:N}, U_N|\tilde{\mu}) \underbrace{\mathcal{N}(\mathrm{d}w)\cdots\mathcal{N}(\mathrm{d}w)}_{K}\right] \qquad (3.15) \\
&= \frac{U_N^{N-1}}{\Gamma(N)}\int_{\mathbb{R}^+}\mathbb{E}_{\tilde{\mu}^*}\left[e^{-U_N\tilde{\mu}^*(\Theta\cup\theta_1)}\int_{\underbrace{\mathbb{R}^+ \times \cdots \mathbb{R}^+}_{K-1}}\prod_{k=2}^{K} e^{-U_N w_k} w_k^{n_k}\underbrace{\mathcal{N}(\mathrm{d}w)\cdots\mathcal{N}(\mathrm{d}w)}_{K-1}\right] \\
&\quad \cdots\cdots w_1^{n_1}e^{-U_N w_1}\mathrm{d}w_1 \qquad (3.16) \\
&= \frac{U_N^{N-1}}{\Gamma(N)}\mathbb{E}_{\tilde{\mu}^*}\left[e^{-U_N\tilde{\mu}^*(\Theta)}\int_{\underbrace{\mathbb{R}^+ \times \cdots \mathbb{R}^+}_{K-1}}\prod_{k=2}^{K} e^{-U_N w_k} w_k^{n_k}\underbrace{\mathcal{N}(\mathrm{d}w)\cdots\mathcal{N}(\mathrm{d}w)}_{K-1}\right] \\
&\quad \cdots\cdots \int_{\mathbb{R}^+} w_1^{n_1}e^{-U_N w_1}\mathrm{d}w_1 \qquad (3.17) \\
&= \frac{U_N^{N-1}}{\Gamma(N)}e^{-\psi_\eta(U_N)}\prod_{k=1}^{K}\kappa(U_N, n_k)h(\theta_k)\ , \qquad (3.18)
\end{aligned}
$$

where ψ_η is the *Laplace exponent* defined in Definition 3.2, $h()$ is the intensity of the probability measure H, $\kappa(u, m)$ is defined as

$$\kappa(u, m) = \int_0^\infty w^m e^{-uw}\rho_\eta(\mathrm{d}w)\ . \qquad (3.19)$$

In details, the above equations is obtained by the following argument: (3.15) follows by explicitly write out the expectation, which contains two parts: the first part is the expectation over the CRM $\tilde{\mu}^*$, the other part is the expectation over the observations; (3.16) follows by applying the *Palm formula* in Theorem 2.11 over the first atom (w_1, θ_1); (3.17) follows by simply taking out the integration term out of the expectation, which are independent each other; finally (3.18) follows by recursively apply the *Palm formula* for the rest of the atoms.

Thus by applying the *Poisson process partition calculus* in Theorem 2.12, the posterior of a NRM can be characterized by the following theorem:

Theorem 3.6 (Conditional posterior characterization of NRM [James et al., 2009]). *Given an NRM $\mu \sim NRM(\eta, M, H)$ with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$ and a set of samples $x_1, \cdots, x_N$ from it with distinct values $(\theta_1, \cdots, \theta_K)$ and counts $(n_1, \cdots, n_K)$, conditioned on the* latent relative mass U_N, *the posterior of $\tilde{\mu}$ is composed of a CRM with Lévy measure*

$$\nu'(\mathrm{d}w, \mathrm{d}\theta) = e^{-U_N w}\nu(\mathrm{d}w, \mathrm{d}\theta)\ ,$$

and K fixed point jumps distributed as

$$p(w_k|x_{1:N}, U_N) \propto w_k^{n_k} e^{-U_N w_k} \rho_\eta(w_k) \ .$$

3.4 Posteriors of the NGG

This section specializes posteriors for the NGG given observations $\{x_i\}$'s based on the general results in the previous section. Two versions of the posterior are developed, *e.g.*, a fully marginalized version[5] $p(\{x_i\}|\sigma, M)$ and an augmented version $p(\{x_i\}, U_N|\sigma, M)$ with the latent relative mass U_N. The second version is done because, as shown, the first version requires computing a complex recursive function thus is computationally more expensive. Throughout the thesis, the symbol $(x|y)_N$ is used to denote the Pochhammer symbol:

$$(x|y)_N \triangleq x(x+y)\cdots(x+(N-1)y) \ .$$

When $y = 1$, it is simply written it as $(x)_N$.

3.4.1 Marginal posterior

The marginal posterior of a NGG is obtained based on Theorem 3.6 above. James et al. [2009] have developed posterior analysis for the NGG, Theorem 3.7 below simplifies their results and specializes them to the NGG.

Theorem 3.7 (Posterior Analysis for the NGG)**.** *Consider the NGG*$(\sigma, M, H(\cdot))$. *For a data vector* $\{x_i\}$ *of length N there are K distinct values* $\theta_1, ..., \theta_K$ *with counts* $n_1, ..., n_K$ *respectively (where each* $n_k > 0$*). The posterior marginal is given by*

$$p(\{x_i\}|\sigma, M) \ = \ \frac{e^M T_{\sigma,M}^{N,K}}{\sigma^{N-K+1}} \prod_{k=1}^{K} (1-\sigma)_{n_k-1} h(\theta_k) \ . \tag{3.20}$$

where

$$T_{\sigma,M}^{N,K} \ = \ \frac{\sigma^{N-1}}{\Gamma(N)} \int_M^\infty \left(1 - \left(\frac{M}{t}\right)^{1/\sigma}\right)^{N-1} t^{K-1} e^{-t} \mathrm{d}t \tag{3.21}$$

is defined for $N, K \in \mathbb{Z}^+$ *such that* $K \leq N$ *and* $M \in \mathbb{R}^+$ *so* $M > 0$. *Moreover, the predictive posterior is given by:*

$$p(x_{N+1} \in \mathrm{d}x|\{\theta_i\}, \sigma, M) \ = \ \omega_0 H(\mathrm{d}x) + \sum_{k=1}^{K} \omega_k \delta_{\theta_k}(\mathrm{d}\theta)$$

[5] In an NGG, the hyperparameter η is represented as σ.

where the weights sum to 1 ($\sum_{k=0}^{K} \omega_k = 1$) are derived as

$$\begin{aligned}\omega_0 &\propto \sigma \frac{T_{\sigma,M}^{N+1,K+1}}{T_{\sigma,M}^{N+1,K}} \\ \omega_k &\propto (n_k - \sigma) \end{aligned} \tag{3.22}$$

Note, $T_{\sigma,M}^{N,K}$ is a strictly decreasing function of N and M, but an increasing function of K and σ. Moreover, an alternative definition of $T_{\sigma,M}^{N,K}$ derived using the transformation $t = M(1+u)^\sigma$ is

$$T_{\sigma,M}^{N,K} = \frac{\sigma^N M^K}{\Gamma(N) e^M} \int_{\mathbb{R}^+} \frac{u^{N-1}}{(1+u)^{N-K\sigma}} e^{M-M(1+u)^\sigma} \mathrm{d}u \,,$$

and various scaled versions of this integral are presented in the literature. Introducing a $\Gamma(b/\sigma, 1)$ prior on M and then marginalizing out M makes the term in $e^{M-M(1+u)^\sigma}$ disappear since the integral over M can be carried inside the integral over u. This parameterization is interesting because it highlights a connection between the NGG and the Pitman-Yor process. Note the following result is a transparent consequence of [Pitman and Yor, 1997, Proposition 21].

Corollary 3.8 (Relation between the NGG and Pitman-Yor process). *Let a NGG μ be $\mu \sim NGG(\sigma, M, H(\cdot))$ and suppose $M \sim \Gamma(b/\sigma, 1)$ then it follows that*

$$\mu \sim PYP(\sigma, b, H(\cdot)) \,,$$

where $PYP(\sigma, b, H(\cdot))$ denotes a Poisson-Dirichlet/Pitman-Yor process *[Pitman and Yor, 1997; Teh, 2006b,a; Buntine and Hutter, 2012] with discount parameter σ, concentration parameter b and base distribution H.*

For computation, the issue here will be computing the terms $T_{\sigma,M}^{N,K}$. Therefore we present some results for this. These use $\Gamma(x, y)$, the upper incomplete Gamma function, defined for $y > 0$ and all real x.

Lemma 3.9 (Evaluating $T_{\sigma,M}^{N,K}$). *Have $T_{\sigma,M}^{N,K}$ defined as in Theorem 3.7. Then the following formula hold:*

$$T_{\sigma,M}^{1,K} = \Gamma(K, M) \,, \tag{3.23}$$

$$T_{\sigma,M}^{N,K} \leq \frac{\sigma^{N-1}}{\Gamma(N)} T_{\sigma,M}^{1,K} \,, \tag{3.24}$$

$$T_{\sigma,M}^{N,K} = \frac{\sigma^{N-1}}{\Gamma(N)} \sum_{n=0}^{N-1} (-1)^n \binom{N-1}{n} \Gamma\left(K - \frac{n}{\sigma}, M\right) M^{n/\sigma} \,, \tag{3.25}$$

$$T_{\sigma,M}^{N-1,K-1} = T_{\sigma,M}^{N,K} + \left(\frac{N-1}{\sigma} - (K-1)\right) T_{\sigma,M}^{N,K-1} \qquad \forall N \geq 2 \,. \tag{3.26}$$

A variant of Equation (3.25) only applies for $K \in \mathbb{N}^+$,

$$T^{N,K}_{\sigma,M} = \frac{\sigma^{N-1}}{\Gamma(N)} \sum_{n=0}^{N-1} (-1)^n \binom{N-1}{n} \left(1 - \frac{n}{\sigma}\right)_{K-1} \Gamma\left(1 - \frac{n}{\sigma}, M\right) M^{n/\sigma} . \qquad (3.27)$$

Other recursions involve factors of $1/\sigma$ *and can be used when* $\sigma = 1/R$ *for some* $R \in \mathbb{N}^+, R > 1$. *Note the function* $T^{N,K}_{\sigma,M}$ *is well defined for non-integral K. Then*

$$\begin{aligned} T^{N,K}_{\sigma,M} &= \frac{\sigma}{N-1}\left(T^{N-1,K}_{\sigma,M} - M^{1/\sigma} T^{N-1,K-1/\sigma}_{\sigma,M}\right) && \forall N \geq 2\,, \quad (3.28) \\ T^{N,K}_{\sigma,M} &= (K-1) T^{N,K-1}_{\sigma,M} + M^{1/\sigma}\, T^{N-1,K-1-1/\sigma}_{\sigma,M} && \forall N \geq 2\,. \quad (3.29) \end{aligned}$$

Note the upper incomplete Gamma function becomes infinitesimal quickly for large y and negative x because $\Gamma(x,y) \to y^{x-1}e^{-y}$ as $y \to \infty$, and for positive y and $x \leq 1$, $\Gamma(x,y) \leq y^{x-1}e^{-y}$. As $y \to 0$ and $x < 0$, $\Gamma(x,y) \to -y^x/x$. Moreover, for $x < -1$ a good approximation is given by $\Gamma(x,y) \approx y^x e^{-y}/(y - x + 1)$. This implies the series summation of Equation (3.25) will be unstable for large N since to a first approximation it is a binomial expansion of $(1-1)^N$. Experiments show this can happen for $N > 20$, so the summation is not practically useful but good for checking small values.

The recursion of Equation (3.26) recurses down on K. The inverted version, recursing up with $T^{N,K}_{\sigma,M}$ on the left-hand side is unstable because it involves the subtraction of two terms, $T^{N-1,K-1}_{\sigma,M}$ and $\left(\frac{N-1}{\sigma} - (K-1)\right) T^{N,K-1}_{\sigma,M}$. Thus errors magnify and it is not practically useful for $N > 20$. However, the inverted version shows that $T^{N,K}_{\sigma,M}$ is related to a generalized Stirling number of the second kind.

Computing $T^{N,K}_{\sigma,M}$ would go as follows. Fix an upper bound on K to be used, denote in as K_{max}. Values of $T^{N,K}_{\sigma,M}$ need to be initialized for $K = K_{max} \& N \geq K_{max}$ and for $K < K_{max} \& N = K$. This can be done using either numerical integration or a saddle point approximation using Equation (3.21). The saddle point approximation requires an initial maximization step, which can be done using Newton-Raphson convergence, and typically has 6-decimal place accuracy for $N > 50$. Thereafter the recursion of Equation (3.26) can be applied to recurse down on K.

Remark 3.10. The Poisson-Dirichlet Process and Dirichlet Process are well known for their ease of use in a hierarchical context [Teh et al., 2006; Chen et al., 2011; Buntine and Hutter, 2012]. The NGG has the same general form, which comes from the fact that it belongs to the class of Gibbs-type priors whose conditional distribution has a convenient form [Favaro et al., 2013a].

The major issue with this posterior theory is that one needs to precompute the terms $T^{N,K}_{\sigma,M}$. While the Poisson-Dirichlet Process has a similar style, it has a generalized Stirling number dependent only on the discount σ [Buntine and Hutter, 2012]. The difference is that for the Poisson-Dirichlet Process we can tabulate these terms for a given discount parameter σ and still vary the concentration parameter (b above, but corresponding to M) easily. For the NGG, any tables of $T^{N,K}_{\sigma,M}$ would need to be recomputed with every change in mass parameter M. This might represent a significant computational burden.

3.4.2 Conditional posterior

To alleviate the computational issue of the above marginal posterior, a simplified conditional posterior is available for the NGG by introducing the *latent relative mass* auxiliary variable U_N. James et al. [2009] has also developed conditional posterior analysis for the NGG. Theorem 3.11 below simplifies their results and specializes them to the NGG.

Theorem 3.11 (Conditional Posterior Analysis for the NGG)**.** *Consider the NGG(σ, M, H) and the situation of Theorem 3.7. The conditional posterior marginal, conditioned on the auxiliary variable U_N, is given by*[6]

$$p\left(\{x_i\}|U_N, \sigma, M\right) \;=\; \frac{\left(M\sigma\left(1+U_N\right)^{\sigma}\right)^K}{\sum_{k=1}^{N} S_{k,\sigma}^{N}\left(M\sigma\left(1+U_N\right)^{\sigma}\right)^k}\prod_{k=1}^{K}(1-\sigma)_{n_k-1}h(\theta_k)\,. \tag{3.30}$$

Moreover, the predictive posterior is given by:

$$p\left(x_{N+1}\in \mathrm{d}x|\{\theta_i\}, U_N = u, \sigma, M\right) \;=\; \omega_0 H(\mathrm{d}x) + \sum_{k=1}^{K}\omega_k\delta_{\theta_k}(\mathrm{d}x)$$

where the weights sum to 1 ($\sum_{k=0}^{K}\omega_k = 1$) are derived as

$$\begin{aligned}\omega_0 &\propto M\sigma\left(1+u\right)^{\sigma}\\ \omega_k &\propto n_k - \sigma\,.\end{aligned} \tag{3.31}$$

The posterior for U_N is given by:

$$p\left(U_N = u|\{\theta\}, \sigma, M\right) \;=\; \frac{\sigma M^K}{T_{\sigma,M}^{N,K}}\frac{u^{N-1}}{\left(1+u\right)^{N-K\sigma}}e^{-M(1+u)^{\sigma}}\,. \tag{3.32}$$

A posterior distribution is also presented by James *et al.* as their major result of Theorem 1 [James et al., 2009]. It is adapted here to the NGG.

Theorem 3.12. *In the context of Theorem 3.11 the conditional posterior of the normalized random measure $\tilde{\mu}$ given data $\{x_i\}$ of length N and latent relative mass $U_N = u$ is given by*

$$\boldsymbol{\mu} \;=\; \frac{T}{T+W_+}\boldsymbol{\mu}' + \frac{W^+}{T+W_+}\sum_{k=1}^{K}p_k\delta_{\theta_k}$$

[6] Over the thesis, U_N is sometimes written as u for notational simplicity.

where

$$\begin{aligned}
\boldsymbol{\mu}' &\sim NGG\left(\sigma, \frac{M}{1+u}, H\right) , \\
T &\sim f_T(t) \;\; \textit{where Lévy measure of } f_T(t) = \frac{M\sigma}{\Gamma(1-\sigma)} t^{-\sigma-1} e^{-(1+u)t} , \\
W^+ &\sim \Gamma(N - K\sigma, 1+u) , \\
\boldsymbol{p} &\sim Dirichlet_K(\boldsymbol{n} - \sigma) .
\end{aligned}$$

Here, $\boldsymbol{\mu}'$, W_+ and $\boldsymbol{p}$ are jointly independent and T, W_+ and $\boldsymbol{p}$ are jointly independent.

Note in particular the densities given for $\boldsymbol{\mu}'$ and T are not independent from each other. While an explicit density is not given for T, its expected value is easily computed via the Laplace transform as $M\sigma(1+u)^{\sigma-1}$.

A joint form of the conditional posteriors presented in Theorem 3.11 can be developed, and can be derived from the general sampling form in Lemma 3.18. by marginalizing out jumps w_k and then taking the limit as $L \to 0$. This matches the conditionals of Theorem 3.11 so is seen to be correct.

Corollary 3.13 (Collapsed Sampling Posterior)**.** *In the context of Theorem 3.11, assume there are K jumps with attached data (w_k such that $n_k > 0$). The resultant posterior is as follows:*

$$\begin{aligned}
& p\left(\{x_i\}, U_N = u \,|\, \sigma, M\right) \\
&= \frac{u^{N-1}}{(1+u)^{N-K\sigma}} (M\sigma)^K e^{M - M(1+u)^\sigma} \prod_{k=1}^{K} (1-\sigma)_{n_k - 1} h(\theta_k) . \qquad (3.33)
\end{aligned}$$

Moreover, the posterior for jumps w_k with data count n_k given $\{\theta_i\}, U_N = u, K, N$ is

$$w_k \sim Ga(n_k - \sigma, 1+u) .$$

Remark 3.14. With the use of the latent relative mass U_N, the NGG lends itself to hierarchical reasoning without a need to compute the recursive series $T^{N,K}_{\sigma,M}$. This can be done with either the jumps integrated out, or the jumps retained.

3.5 Posterior Inference for NGG Mixtures

This section applies the posterior of the NGG for NGG mixture models. Specifically, given observations x_i's, the NGG mixture is defined as:

$$\begin{aligned}
\mu|\sigma, M &= \sum_{k=1}^{\infty} \bar{w}_k \delta_{\theta_k} \sim \text{NGG}(\sigma, M, H(\cdot)) \\
\theta_k | H &\sim H \\
s_i | \mu &\sim \text{Discrete}(\bar{w}_1, \bar{w}_2, \cdots) \\
x_i | \theta_{s_i} &\sim F(\cdot | \theta_{s_i}) \qquad (3.34)
\end{aligned}$$

where $\bar{w}_k = w_k / \sum_{l=1}^{\infty} w_l$, and $w_1, w_2, \cdots$ are the jumps of the corresponding CRM defined in (3.2), θ_k's are the components of the mixture model drawn *i.i.d.* from a base distribution $H(\cdot)$ with density denoted as $h(\cdot)$, s_i indexes which component x_i belongs to, and $F(\cdot|\theta_k)$ is the cumulative distribution function to generate data on component k, with the corresponding density function as $f(\cdot|\theta_k)$[7].

Given the posterior analysis for the NGG above, posterior inference for the NGG mixture can be done via Markov chain Monte Carlo (MCMC). Before proceeding to the detailed descriptions of the sampling algorithms for the NGG, some computational results related to the NGG such as the *Laplace exponent* in Theorem 3.15 below are first proved. These formula can be used in the slice sampler developed below.

Theorem 3.15 (Key formulas for the NGG). *Define for $0 < \sigma < 1$, $\Gamma(-\sigma) \triangleq \frac{\Gamma(1-\sigma)}{-\sigma}$. For the NGG, some key formula used in the posterior computation are as follows:*

$$\begin{aligned}
\psi_\sigma(v) &= M\left((1+v)^\sigma - 1\right) && (3.35)\\
\int_L^\infty \rho_\sigma(\mathrm{d}w) &= |Q(-\sigma, L)| && (3.36)\\
\int_L^\infty e^{-vw}\rho_\sigma(\mathrm{d}w) &= (1+v)^\sigma |Q(-\sigma, L(1+v))| && (3.37)\\
\int_0^L \left(1-e^{-vw}\right)\rho_\sigma(w)\mathrm{d}w &= ((1+v)^\sigma - 1) + (1+v)^\sigma |Q(-\sigma, L(1+v))| \\
&- |Q(-\sigma, L)| && (3.38)
\end{aligned}$$

where $Q(x,y) = \Gamma(x,y)/\Gamma(x)$ is the regularized upper incomplete Gamma function. Some mathematical libraries provide it for a negative first argument, or it can be evaluated using

$$Q(-\sigma, z) = Q(1-\sigma, z) - \frac{1}{\Gamma(1-\sigma)} z^{-\sigma} e^{-z}, \qquad (3.39)$$

using an upper incomplete Gamma function defined only for positive arguments.

Proof. For (3.35), according to (3.6), we have

$$\begin{aligned}
\psi_\sigma(v) &= M \int_{\mathbb{R}^+} \left[1 - \exp\{-wf\}\right] \rho_\eta(\mathrm{d}w)\\
&= \frac{M\sigma}{\Gamma(1-\sigma)} \int_{\mathbb{R}^+} \left(\sum_{n=1}^{\infty} (-1)^{n-1} \frac{(vx)^n}{n!}\right) x^{-\sigma-1} e^{-x} \mathrm{d}x\\
&= \frac{M\sigma}{\Gamma(1-\sigma)} \sum_{n=1}^{\infty} (-1)^{n-1} \frac{\lambda^n}{n!} \Gamma(n-\sigma)\\
&= \frac{M\sigma}{\Gamma(1-\sigma)} \sum_{n=1}^{\infty} (-1)^{n-1} \lambda^n \frac{\Gamma(n-\sigma)}{n!}\\
&= \frac{M}{\Gamma(1-\sigma)} \left(\sum_{n=1}^{\infty} \frac{(-1)^{n-1}\sigma\Gamma(n-\sigma)}{n!} \lambda^n\right)
\end{aligned}$$

[7] The same notation $h(\cdot)$ and $f(\cdot)$ will be used in the rest of the thesis if not otherwise stated.

$$\begin{aligned} &= M\left(\sum_{n=1}^{\infty}\frac{\sigma(\sigma-1)\cdots(\sigma-n+1)}{n!}\lambda^n\right) && (3.40)\\ &= M\left[(1+\lambda)^{\sigma}-1\right] \ , \end{aligned}$$

where the summation in (3.40) is the Taylor expansion of $(1+\lambda)^{\sigma}-1$.

For (3.36), we have

$$\begin{aligned} |Q(-\sigma,L)| &= \left|\frac{\Gamma(-\sigma,L)}{\Gamma(-\sigma)}\right| = \left|\frac{\Gamma(-\sigma,L)}{\frac{\Gamma(1-\sigma)}{-\sigma}}\right| \\ &= \frac{\sigma}{\Gamma(1-\sigma)}\int_L^{\infty} w^{-\sigma-1}e^{-w}\mathrm{d}w = \int_L^{\infty}\rho_{\sigma}(\mathrm{d}w) \ . \end{aligned}$$

(3.37) and (3.38) are easily obtained from (3.36) by using a change of variable as

$$w' \triangleq (1+v)w \ .$$

For (3.39), we have

$$\begin{aligned} \Gamma\left(-\sigma,z\right) &= \int_z^{\infty} w^{-\sigma-1}e^{-w}\mathrm{d}w \\ &= \frac{-1}{\sigma}w^{-\sigma}e^{-w}|_z^{\infty} - \int_z^{\infty}\frac{1}{\sigma}w^{-\sigma}e^{-w}\mathrm{d}w \\ &= \frac{1}{\sigma}\left(z^{-\sigma}e^{-z}-\Gamma\left(1-a,z\right)\right) \ . \end{aligned}$$

Thus

$$\begin{aligned} Q(-\sigma,z) &= \frac{-1}{\sigma\Gamma(\sigma)}\left(z^{-\sigma}e^{-z}-\Gamma\left(1-\sigma,z\right)\right) \\ &= Q\left(1-\sigma,z\right) - \frac{1}{\Gamma(1-\sigma)}z^{-\sigma}e^{-z} \ . \end{aligned}$$

□

3.5.1 Marginal sampler

Two versions of the posterior are developed for the NGG mixture in this section, one is based on the collapsed posterior extended from Theorem 3.7, called *collapsed Gibbs sampler*; the other is based on the conditional posterior extended from Corollary 3.13, called *conditional Gibbs sampler*.

3.5.1.1 Collapsed Gibbs sampler

Extending the posterior of an NGG to an NGG mixture is straightforward. Based on Theorem 3.7, it is easy to see the posterior of the NGG mixture in Eq.(3.34) is:

Corollary 3.16. *The collapsed posterior for the NGG mixture defined in Eq.(3.34) is given by:*

$$p\left(\{x_i\}|\sigma, M\right) \;=\; \frac{e^M T^{N,K}_{\sigma,M}}{\sigma^{N-K+1}} \prod_{k=1}^{K} \left((1-\sigma)_{n_k - 1} \int \prod_{i:s_i=k} f(x_i|\theta_k) h(\theta_k) \mathrm{d}\theta_k \right) . \tag{3.41}$$

where s_i indexes which component x_i belongs to, $T^{N,K}_{\sigma,M}$ is the same as in Theorem 3.7.

Sampling: Given the posterior (3.41), the only latent variables are s_i's. Denote the whole variables in the model as C, the conditional distribution for s_i can be read from the posterior as:

$$p(s_i|C - s_i) \propto \begin{cases} T^{N,K}_{\sigma,M}(n_k - \sigma) \frac{\int \prod_{j:s_j=k} f(x_j|\theta_k) h(\theta_k) \mathrm{d}\theta_k}{\int \prod_{j:j\neq i, s_j=k} f(x_j|\theta_k) h(\theta_k) \mathrm{d}\theta_k}, & \text{if } k \text{ already exists} \\ \sigma T^{N,K+1}_{\sigma,M} \int f(x_i|\theta) h(\theta) \mathrm{d}\theta, & \text{if } k \text{ is new} \end{cases}$$

Unfortunately, sampling for other hyperparameters such as σ and M does not seem easy in this representation since they are coupled in $T^{N,K}_{\sigma,M}$, thus this representation is usually not used in real applications.

3.5.1.2 Conditional Gibbs sampler

This section presents a more tractable form for the NGG by extending the conditional posterior for the NGG in Corollary 3.13 to that of NGG mixtures. The following corollary states this:

Corollary 3.17. *The conditional posterior for the NGG mixture defined in Eq.(3.34) is given by:*

$$p\left(\{x_i\}, u|\sigma, M\right) \;=\; \frac{u^{N-1}}{(1+u)^{N-K\sigma}} \left(M\sigma\right)^K e^{M - M(1+u)^\sigma} \prod_{k=1}^{K} (1-\sigma)_{n_k-1} \int \prod_{i:s_i=k} f(x_i|\theta_k) h(\theta_k) \mathrm{d}\theta_k . \tag{3.42}$$

The above posterior is more favorable than (3.41) in that variables are decoupled and have simple conditional distributions. From the posterior we can easily get the following conditional distributions for $(\{s_i\}, u, M, \sigma)$ (also denote the whole variables in the model as C):

Sampling s_i: the conditional distribution for s_i follows:

$$p(s_i|C - s_i) \propto \begin{cases} (n_k - \sigma) \frac{\int \prod_{j:s_j=k} f(x_j|\theta_k) h(\theta_k) \mathrm{d}\theta_k}{\int \prod_{j:j\neq i, s_j=k} f(x_j|\theta_k) h(\theta_k) \mathrm{d}\theta_k}, & \text{if } k \text{ already exists} \\ \sigma M (1+u)^{1-\sigma} \int f(x_i|\theta) h(\theta) \mathrm{d}\theta, & \text{if } k \text{ is new} \end{cases}$$

Sampling u**:** the auxiliary variable u has posterior distribution as

$$p(u|C-u) \propto \frac{u^{N-1}e^{-M(1+u)^{\sigma}}}{(1+u)^{N-K\sigma}} ,$$

which can be shown to be log-concave, thus can be sampled with the adaptive-rejection sampler [Gilks and Wild, 1992] or the slice sampler [Neal, 2003].

Sampling M**:** the model performance is also sensitive to the mass parameter M. If we introduce a Gamma(a, b) prior for it, then the posterior is also a Gamma:

$$p(M|C-M) \sim \text{Gamma}\left(K+a, (1+u)^{\sigma}+b-1\right) .$$

Sampling σ**:** by collecting related terms, the posterior for σ is

$$p(\sigma|C-\sigma) \propto \frac{\sigma^{K}e^{-M(1+u)^{\sigma}}}{(1+u)^{-K\sigma}} \prod_{k=1}^{K} (1-\sigma)_{n_k-1} ,$$

which can also be proven to be log-concave, thus can also be sampled with the adaptive-rejection sampler or the slice sampler.

3.5.2 Slice sampling

A convenient method for posterior inference with MCMC is slice sampling, which is particularly useful when the posterior does not have a close form. Slice sampling an NRM has been discussed in several papers; this section follows the slice sampling method in [Griffin and Walker, 2011], and briefly introduces the underlying ideas.

Given the observations $\boldsymbol{X} = \{x_1, \cdots, x_N\}$, let us introduce a slice latent variable v_i for each x_i so that we only consider those components whose jump sizes w_k's are larger than the corresponding v_i's. Moreover, only jumps with sizes greater than L are considered, and this is maintained by setting $L \leq \min_i v_i$. Sampling of the NRM can then be done by only considering jumps greater than L. In the NRM, the count of such jumps K, is shown to have a Poisson distribution, $K \sim \text{Poisson}(M \int_L^{\infty} \rho_{\eta}(\mathrm{d}t))$, while each jump has density $\frac{\rho_{\eta}(w_k)}{\int_L^{\infty} \rho_{\eta}(s)\mathrm{d}t}$.

Furthermore, the auxiliary variable U_N (latent relative mass) is introduced to decouple each individual jump w_k and their infinite sum of the jumps $\sum_{l=1}^{\infty} w_l$ appeared in the denominators of $\bar{w}_k$'s. Based on [Griffin and Walker, 2011], the following posterior Lemma can be derived, see appendix for the derivation.

Lemma 3.18. *The posterior of the infinite mixture model (3.34) with the above auxiliary* *variables is proportional to*

$$
\begin{aligned}
&p(\theta_{1:K}, w_1, \cdots, w_K, K, \boldsymbol{v}, \boldsymbol{s}, U_N, \boldsymbol{X} | L, \eta, M) \propto \\
&U_N^{N-1} \exp\left\{-U_N \sum_{k=1}^{K} w_k\right\} \exp\left\{-M \int_0^L \left(1 - e^{-U_N t}\right) \rho_\eta(t) \mathrm{d}t\right\} \\
&M^K \exp\left\{-M \int_L^\infty \rho_\eta(t) \mathrm{d}t\right\} \prod_{k=1}^{K} \rho_\eta(w_k) h(\theta_k) \prod_{i=1}^{N} 1(w_{s_i} > v_i) f(x_i | \theta_{s_i}),
\end{aligned}
\tag{3.43}
$$

where $1(a)$ *is a indicator function returning 1 if a is true and 0 otherwise,* $h(\cdot)$ *is the density of* $H(\cdot)$*,* $L \leq \min\{\boldsymbol{u}\}$*.*

The integrals for the NGG needed to work with this lemma were given in Theorem 3.15 above. Thus the integral term in Equation (3.43) can be turned into an expression involving incomplete Gamma functions.

3.5.2.1 Sampling:

Similarly denote the whole parameters as $C = \{\theta_{1:K}, w_1, \cdots, w_K, K, \boldsymbol{v}, L, \boldsymbol{s}, U_N, M\}$. Based on the posterior above, the sampling goes straightforwardly as

- **Sampling $\boldsymbol{s}$:** From (3.43) we get

$$
p(s_i = k | C \backslash \{s_i\}) \propto 1(w_k > v_i) f(x_i | \theta_k) \tag{3.44}
$$

- **Sampling U_N:** Similarly

$$
\begin{aligned}
p(U_N | C \backslash \{U_N\}) \quad \propto \quad & U_N^{N-1} \exp\left\{-U_N \sum_{k=1}^{K} w_k\right\} \\
& \exp\left\{-M \int_0^L \left[1 - \exp\left\{-U_N t\right\}\right] \rho_\eta(\mathrm{d}t)\right\}.
\end{aligned}
\tag{3.45}
$$

  Similar to [Griffin and Walker, 2011], the rejection sampler is used to sample U_N, with proposal distribution Gamma $\left(n, \sum_{k=1}^{K} w_k\right)$.

- **Sampling θ:** The posterior of θ_k with prior density $p(\theta_k)$ is

$$
p(\theta_k | C \backslash \{\theta_k\}) \propto p(\theta) \prod_{i | s_i = k} f(x_i | \theta_k). \tag{3.46}
$$

- **Sampling $K, \{w_1, \cdots, w_K\}$:** Sampling for w_k can be done separately for those associated with data points (fixed points) and for those that are not. Based on [James et al., 2009], when integrating out $\boldsymbol{u}$ in (3.43), the posterior of the jump w_k with data attached ($n_k > 0$) is proportional to

$$
w_k^{n_k} \exp\left\{-U_N w_k\right\} \rho_\eta(w_k), \tag{3.47}
$$

While for those without data attached ($n_k = 0$), based on [Griffin and Walker, 2011], conditional on U_N, the number of these jumps follows a Poisson distribution with mean

$$M \int_L^\infty \exp\{-U_N t\}\rho_\eta(\mathrm{d}t),$$

while their lengths t have densities proportional to

$$\exp\{-U_N t\}\rho_\eta(\mathrm{d}t)1(t > L).$$

Note that this strategy of sampling w_k's by first sampling the number of jumps and then simulating the jump sizes is restricted to the case when the posterior density of the jumps are relatively easy to evaluate. A much more flexible method to simulate the jumps by thinning a Poisson process with well behaved mean measure will be introduced in Chapter 7.

- **Sampling v:** v are uniformly distributed in the interval $(0, w_{s_i}]$ for each i. After sampling the v, L is set to $L = \min\{v\}$.
- **Sampling M:** The posterior of M with prior $p(M)$ is

$$p(M|C\backslash\{M\}) \propto p_M(M)M^K \exp\left\{-M\left[\int_L^\infty \rho_\eta(\mathrm{d}t) + \int_0^L \left[1-\exp\{-U_N t\}\right]\rho_\eta(\mathrm{d}t)\right]\right\}.$$

$p(M)$ is usually taken to be Gamma distributed, in which case the posterior is also Gamma, thus can be sampled conveniently.

3.6 Experiments

3.6.1 Marginal vs. slice samplers

This experiment compares the marginal and slice samplers for the NRM mixture. First they are tested on the *galaxy* dataset, a standard dataset for testing Gaussian mixture models [Griffin and Walker, 2011]. In this case the base distribution H in (3.34) is taken as *Gaussian Wishart* distribution, which is a prior distribution for the joint distribution of the *mean* and *covariance* in a Gaussian distribution [Teh, 2007; Murphy, 2007]

$$\text{Wishart:} \quad p(R) = 2^{-\nu d/2}\pi^{-d(d-1)/4}|S|^{\nu/2}\prod_{i=1}^{d}\Gamma\left(\frac{\nu+1-i}{2}\right)^{-1}|R|^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}[RS]\right)$$

$$\text{Gaussian:} \quad p(\mu|R) = (2\pi)^{-d/2}|rR|^{1/2}\exp\left(-\frac{1}{2}\mathrm{Tr}\left[rR\left((\mu-m)(\mu-m)^T\right)\right]\right)$$

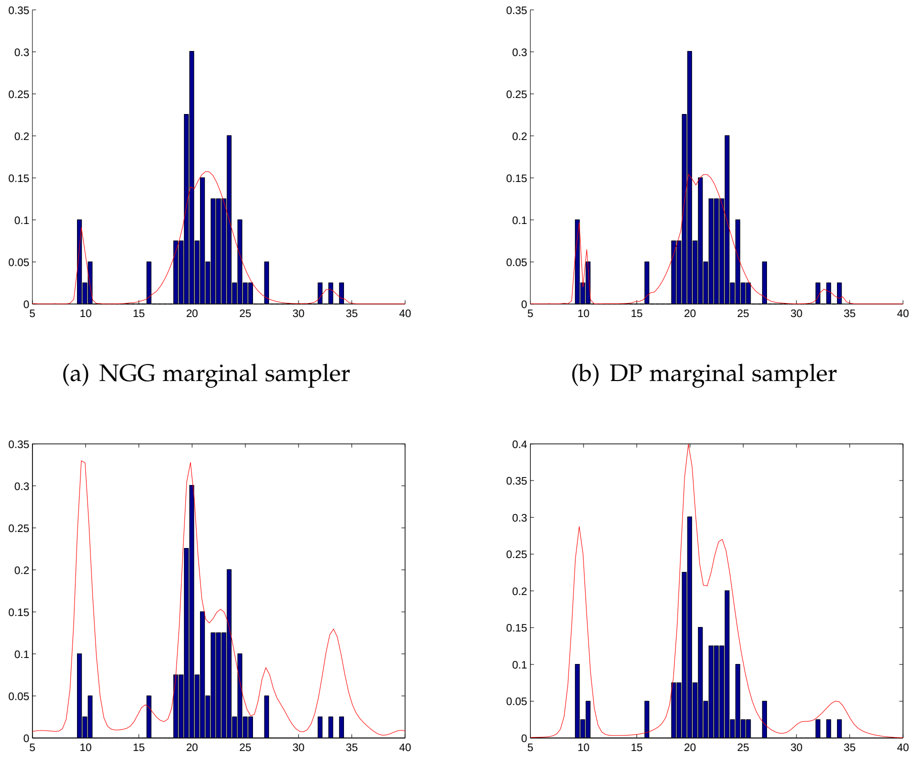
$$p(\mu, R) = \frac{1}{Z(d,r,\nu,S)}|R|^{(\nu-d)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}\left[R\left(r(\mu-m)(\mu-m)^T + S\right)\right]\right),$$

where $Z(d,r,\nu,S) = 2^{\frac{(\nu+1)d}{2}} \pi^{d(d+1)/4} r^{-d/2} |S|^{-nu/2} \prod_{i=1}^{d} \Gamma\left(\frac{\nu+1-i}{2}\right)$. When the dimension $d = 1$, this degenerates into the Gaussian Gamma prior [Teh, 2007]. Correspondingly, the data generation distribution $f(\cdot)$ is Gaussian distribution:

$$f(x_i|\mu,R) = (2\pi)^{-nd/2}|R|^{n/2}\exp\left(-\frac{1}{2}\mathrm{Tr}\left[R\left(\mu-x_i\right)\left(\mu-x_i\right)^T\right]\right) .$$

In the experiment, the hyperparameters are chosen as $\nu = 5, m = 0, r = 0.1, R = 1$, all the other parameters of the model are sampled during the inference. The marginal sampler are implemented in MATLAB, the code for slice sampler are borrowed from [Griffin and Walker, 2011]. The total number of MCMC iterations is set to 2,000 with 1,000 burn in iterations.

First, the NGG mixture is used for density estimation for the *galaxy* dataset, the estimated posteriors are plotted in Figure 3.3. The DP mixture with both marginal and slice samplers is also implemented for comparison. It can be seen from the figure that the slice sampler seems to be able to estimate the overall shape of the data better, but also overestimates the density in a relatively low density area (the first peak in the figure).



(a) NGG marginal sampler

(b) DP marginal sampler

(c) NGG slice sampler

(d) DP slice sampler

Figure 3.3: Density estimation on galaxy dataset

Table 3.1: Effective sample sizes over 1000 samples for the marginal and slice samplers

	min	max	mean	median
Marginal	$22.55_{\pm 15.06}$	$214.53_{\pm 105.27}$	$96.66_{\pm 39.44}$	$74.78_{\pm 29.94}$
Slice	$1.50_{\pm 0.00}$	$68.92_{\pm 33.23}$	$29.80_{\pm 9.53}$	$27.63_{\pm 12.31}$

To further compare the marginal and slice samplers, *effective sample sizes* (ESS) are calculated over 1000 samples. The statistics used for the evaluation are (σ, M, u) and sum of the data likelihood. The *minimum, maximum, mean* and *median* of the ESS are shown in Table 3.1. From the results it can be clearly seen that the marginal sampler beats the slice sampler in this simple model[8], obtaining much better mixing. This coincides with the discover in [Favaro and Teh, 2013].

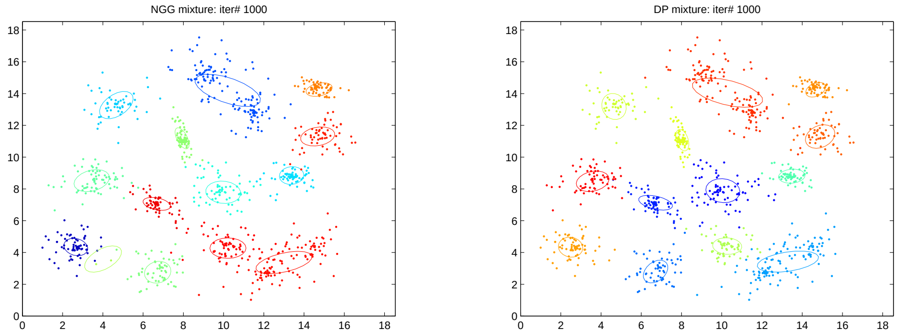
3.6.2 Clustering

Illustration This section demonstrates how to use the NGG mixture for clustering. As a mixture model, each component of the model is often taken as a cluster. To illustrate, the NGG mixture is first run on the S-set dataset [Franti and Virmajoki, 2006], which is a synthetic 2-d data with 5000 samples and 15 Gaussian clusters with different degree of cluster overlapping. 1000 data point are collected after 5000 burnin for ESS calculation. The hyperparameters are set as $\nu = 5, m = 0, r = 0.25, R = \mathbf{I}$ where $\mathbf{I}$ is a 2×2 identity matrix. σ in NGGM was set to 0.5, though it can be sampled during the inference. In the experiment NGGM model is compared with DPM. The models are randomly initialized with 10 components and 1000 Gibbs iterations are used. The learned clusters are plotted in Figure 3.4, from which we can see that the two models performs similarly, and are able to recover the mixture components in the data, though both seem to combine the two components (on top of the data) into one.

Clustering performance Next the NGG mixture (NGGM) model is tested on ten real datasets from the UCI repository [Bache and Lichman, 2013]. Table 3.2 lists some of the statistics of these datasets. For computational ease, subsets of the three large datasets, *e.g.*, Letter, MNIST and Segmentation datasets, are used.

To measure the clustering performance, the *normalized mutual information* (NMI) score, a standard measurement for evaluating clustering models [Vinh et al., 2010] is used. Let N be the number of data points in the dataset, $\Omega = \{\omega_1, \cdots, \omega_K\}$ denote the true cluster assignment of the data, where ω_k is the set of data points assigning to the k-th cluster. $\mathbb{C} = \{c_1, \cdots, c_J\}$ denote the cluster structure produced by a model and c_j is the set of data points assigning to the j-th cluster. $|\cdot|$ is the cardinality

[8] However, this is not always the case, see more complex models for example in Chapter 6 and Chapter 7.



(a) NGG mixture

(b) DP mixture

Figure 3.4: Clustering of a Gaussian mixture data with NGG mixture and DP mixture models.

Table 3.2: Statistics of the ten UCI datasets.

Data set	#instances	Dim	#clusters
Glass	214	10	7
half_circle	300	2	2
Iris	150	4	3
Letter	1000	16	10
MNIST	1000	784	10
Satimage	4435	36	6
Segmentation	1000	19	7
Vehical	846	18	4
Vowel	990	10	11
Wine	178	13	3

operator. The NMI is defined as

$$\mathrm{NMI}(\Omega, \mathbb{C}) = \frac{2\sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|}}{\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N} + \sum_j \frac{|c_j|}{N} \log \frac{|c_j|}{N}}$$

The model hyperparameters are set as $\nu = d, m = 0, r = 0.1, R = \mathrm{eye}(d)$ where d is the dimension of the data. To test the impacts of the σ parameter to the model performance, two version of the NGGM are instantiated, one is to sample σ during inference, denoted as NGGM^1, the other is to simply set σ as 0.1, denoted as NGGM^2. The models ares compared with the popular *kmeans* and *nCut* algorithms as well as the DPM model. For all the models, the experiments are repeated for 5 times with randomly initialization. The mean and standard deviations are reported in Table 3.3. It can be seen from the table that due to the flexibility, NGGM is slightly better than the more specific DPM in general; moreover, the one with σ sampled performs slight

Table 3.3: Comparison for different methods

Data set	kmeans	nCut	DPM	NGGM^1	NGGM^2
Glass	0.37±0.04	0.22±0.00	0.45±0.01	0.43±0.04	0.45±0.01
half_circle	0.43±0.00	1.00±0.00	0.39±0.07	0.48±0.03	0.48±0.03
Iris	0.72±0.08	0.61±0.00	0.73±0.00	0.73±0.00	0.73±0.00
Letter	0.33±0.01	0.04±0.00	0.26±0.07	0.22±0.05	0.20±0.07
MNIST	0.50±0.01	0.38±0.00	0.54±0.02	0.57±0.02	0.56±0.01
Satimage	0.57±0.06	0.55±0.00	0.29±0.06	0.32±0.00	0.32±0.00
Segmentation	0.52±0.03	0.34±0.00	0.33±0.03	0.37±0.09	0.33±0.02
Vehical	0.10±0.00	0.14±0.00	0.01±0.00	0.01±0.00	0.01±0.00
Vowel	0.42±0.01	0.44±0.00	0.27±0.02	0.26±0.01	0.28±0.03
Wine	0.84±0.01	0.46±0.00	0.56±0.01	0.56±0.02	0.56±0.01

better then the one with fixed σ, which is reasonable. Interestingly, in some cases both NGGM and DPM do not perform as well as the simplest *kmeans* method. The reason is that in real data, the underlying data distribution is usually not Gaussian, and NGGM usually generates much fragmented partition of the data than the true partition.

3.7 Conclusion

This chapter introduces the normalized random measure, a nonparametric Bayesian family of discrete random probability measures. It is built on the Poisson process, thus its distributional properties as well as the posterior can be analyzed by the theory of Poisson processes, or particularly via the Poisson process partition calculus. A concrete example of the normalized random measure called normalized generalized Gamma process is detailed studied in this chapter. Its posterior inference via marginal sampler and slice sampler are also developed. Note the slice sampler is adapted from the one proposed by Griffin and Walker [2011], and this is not the only slice sampler for normalized random measure mixtures. A more computationally efficient slice sampler can be developed using the Poisson process thinning technique to be discussed in Chapter 7. Finally, comparison of these two samplers and an application to clustering were presented, where the normalized random measure is shown to be more flexible than the Dirichlet process in real applications.

Finally, we note that in the experiments above, the NRM does not show obvious advantages compared with the DP. Indeed, the posterior structure of the NRM is much more complicated than the DP. However, this does not necessarily mean that NRM is not worth investigating because as will be seen in the rest of the thesis, different kinds of dependency models can be constructed based on the NRM framework, where posteriors of these dependency models are analytically tractable.

3.8 Appendix: Proofs

Proof of Lemma 3.4. We have $\nu(\mathrm{d}w, \mathrm{d}\theta/\lambda)$. Doing a change of variables $w' = w/\lambda$ and some rearranging of the Lévy-Khintchine formula yields the following:

$$\mathbb{E}\left[e^{-\int_{\Theta}(\lambda f(\theta))(\tilde{\mu}(\mathrm{d}\theta)/\lambda)}\right] = e^{-\int_{\mathbb{R}^+\times\Theta}\left(1-e^{-w',(\lambda f(\theta))}\right)\lambda\nu(\mathrm{d}w',\mathrm{d}\theta)}$$

Since $\tilde{\mu}(\mathrm{d}w)/\lambda$ normalizes to the same measure as $\tilde{\mu}(\mathrm{d}\theta)$, and saying something holds for any $f(\theta)$ is the same as saying something holds for any $\lambda f(\theta)$ (when $\lambda > 0$), the result follows. □

Proof of Theorem 3.7. The definition for $\tau_n(u)$ comes from [Proposition 1][James et al., 2009]. The posterior marginal of Equation (3.20) comes from [Proposition 3][James et al., 2009] and is simplified using the change of variables $t = M(1+u)^{\sigma}$. For the predictive posterior, the weights in Equation (3.22) are derived directly from the posterior. The posterior proportionality for $p(U_N = u|\boldsymbol{X}, \sigma, M)$ discards terms not containing u. □

Proof of Corollary 3.8. Marginalize out M from the posterior of Equation (3.20) using the alternative definition of $T_{\sigma,M}^{N,K}$. It can be seen this yields the posterior of a Poisson-Dirichlet distribution with discount parameter σ and concentration parameter b. Since the posteriors are equivalent for all data, the distributions are equivalent almost surely. □

Proof of Lemma 3.9. Equation (3.24) holds by noticing $T_{\sigma.M}^{N,K}$ is decreasing in N and then using the definition of the upper incomplete Gamma function. To prove Equation (3.25), expand the term $\left(1-\left(\frac{M}{t}\right)^{1/\sigma}\right)^{N-1}$ using the binomial expansion and absorbing the powers $t^{-n/\sigma}$ into t^{K-1} as an incomplete Gamma integral.

Now manipulate Equation (3.25). Expand $\Gamma\left(K-\frac{n}{\sigma}, M\right)$ using the recursion for the upper incomplete Gamma function, which can be applied for all first arguments when $M > 0$.

$$\begin{aligned}
&= \sum_{n=0}^{N-1}\binom{N-1}{n}\left(-M^{1/\sigma}\right)^n\left(\left(K-1-\frac{n}{\sigma}\right)\Gamma\left(K-1-\frac{n}{\sigma},M\right)+M^{K-1-\frac{n}{\sigma}}e^{-M}\right)\\
&= \sum_{n=0}^{N-1}\binom{N-1}{n}\left(-M^{1/\sigma}\right)^n\left(K-1-\frac{n}{\sigma}\right)\Gamma\left(K-1-\frac{n}{\sigma},M\right)\\
&\quad +M^{K-1}e^{-M}\sum_{n=0}^{N-1}\binom{N-1}{n}(-1)^n
\end{aligned}$$

The second sum is a binomial expansion of $(1-1)^{N-1}$ and therefore disappears. Apply this step repeatedly. For $K \in \mathcal{N}+$, this terminates after $K-1$ steps to get Equation (3.27).

Equation (3.28) holds by expanding

$$\left(1-\left(\frac{M}{t}\right)^{1/\sigma}\right)^{N-1} \;=\; \left(1-\left(\frac{M}{t}\right)^{1/\sigma}\right)^{N-2}-\left(1-\left(\frac{M}{t}\right)^{1/\sigma}\right)^{N-2}\left(\frac{M}{t}\right)^{1/\sigma}$$

inside the integral definition of $T_{\sigma,M}^{N+1,K}$ and then recognizing the terms.

Equation (3.26) and Equation (3.29) hold by applying the integration by parts formula on the terms $A(t) = \left(1 - \left(\frac{M}{t}\right)^{1/\sigma}\right)^{N-1}$ and $B(t) = t^{K-1}e^{-t}$. Rearranging the resultant integrals and recognizing the terms yields

$$0 \;=\; M^{1/\sigma}T_{\sigma,M}^{N-1,K-1-1/\sigma} + (K-1)T_{\sigma,M}^{N,K-1} - T_{\sigma,M}^{N,K} \,.$$

This proves Equation (3.29). Equation (3.26) follows by then applying Equation (3.28). □

Proof of Theorem 3.11. The posterior marginal of Equation (3.30) comes from [Proposition 4][James et al., 2009]. Although the denominator is difficult to evaluate, and it can be derived through a recursion, the easiest way is simply to normalize the enumerator. Sum over $\left(M\sigma\,(1+u)^{\sigma}\right)^{K}\prod_{k=1}^{K}(1-\sigma)_{n_k-1}$ for all length K partitions $(n_1, n_2, ..., n_K)$ yields $\left(M\sigma\,(1+u)^{\sigma}\right)^{K} S_{K,\sigma}^{N}$ and the result follows by again summing over K. The predictive posterior, as before, follows directly from the posterior marginal. The posterior proportionality for U_N, $p(U_N = u|\{\theta_i\}, \sigma, M)$, comes from [Proposition 4][James et al., 2009] after discarding terms not containing u. The normalizing constant is obtained using the methods of Theorem 3.7. □

Proof of Corollary 3.13. Equation (3.33) can be seen to hold true since conditioning it on $U_N = u$ and $\{\theta_i\}$ yields respectively Equation (3.30) and Equation (3.32). The posterior on w_k comes from [Griffin and Walker, 2011].

This can also be proven from [Griffin and Walker, 2011] at the end of Section 3, and includes the prior on $K_L, w_1, ..., w_K$ described in Section 4. The mixture model component $f(x_i|\theta_{s_i})$ has also been stripped and the slice sampling variables marginalized out. One then takes the limit as $L \to 0$. □

Proof of Lemma 3.18. First, for the infinite mixture model, we have infinite number of components, thus given the observed data $(x_1, \cdots, x_N)$ and their allocation indicators $\boldsymbol{s}$, the model likelihood is

$$p_{\mu}(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{\theta}, \boldsymbol{W}) = \prod_{i=1}^{N}\frac{w_{s_i}}{W_+}f(x_i|\theta_{s_i}),$$

where $W_+ = \sum_{k=1}^{\infty} w_k$. Now introduce the slice auxiliary variables $\boldsymbol{u}$ for each data, such that we only consider the components whose jumps are larger than a threshold

u_i for data x_i, in this auxiliary space we have

$$p_\mu(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{s}|\boldsymbol{\theta}, \boldsymbol{W}) = \frac{1}{W_+^N} \prod_{i=1}^N 1(u_i < w_{s_i}) g_0(x_i|\theta_{s_i}).$$

Now using the fact that

$$\frac{1}{W_+^N} = \frac{\int_0^\infty U_N^{N-1} \exp\{-U_N W_+\} \mathrm{d}U_N}{\Gamma(N)},$$

after introducing the auxiliary variable U_N, we have

$$p_\mu(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{s}, U_N|\boldsymbol{\theta}, \boldsymbol{W}) \propto U_N^{N-1} \exp\{-U_N W_+\} \prod_{i=1}^N 1(u_i < w_{s_i}) f(x_i|\theta_{s_i}).$$

Further decomposing W_+ as

$$W_+ = W^* + \sum_{k=1}^K w_k,$$

where K is the number of jumps which are large than a threshold L, $W^* = \sum_{k=K+1}^\infty w_k$, then we get

$$\begin{aligned} & p_\mu(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{s}, U_N|\boldsymbol{\theta}, w_1, \cdots, w_K, K) \\ \propto\ & U_N^{n-1} \exp\left\{-U_N \sum_{k=1}^K w_k\right\} \mathbb{E}\left[\exp\{-U_N W^*\}\right] \prod_{i=1}^N 1(u_i < w_{s_i}) f(x_i|\theta_{s_i}). \quad (3.48) \end{aligned}$$

Now use the Lévy-Khintchine representation of a Lévy process (3.3) to evaluate $\mathbb{E}\left[\exp\{-U_N W^*\}\right]$, we get

$$\begin{aligned} & p_\mu(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{s}, U_N|\boldsymbol{\theta}, w_1, \cdots, w_K, K) \propto U_N^{N-1} \exp\left\{-U_N \sum_{k=1}^K w_k\right\} \\ & \exp\left\{-M \int_0^L (1 - \exp\{-U_N t\}) \rho_\eta(t) \mathrm{d}t\right\} \prod_{i=1}^N 1(u_i < J_{s_i}) f(x_i|\theta_{s_i}). \quad (3.49) \end{aligned}$$

Now combining with the priors

$$p(w_1, \cdots, w_K) = \prod_{k=1}^K \frac{\rho_\eta(w_k)}{\int_L^\infty \rho_\eta(t)\mathrm{d}t},$$

$$K \sim \text{Poisson}(M \int_L^\infty \rho_\eta(\mathrm{d}t)), \qquad \theta_k \sim H(\theta_k),$$

the result follows. □

Chapter 4

Hierarchial Normalized Random Measures

4.1 Introduction

To a large extent the normalized random measure (NRM) described in the last chapter allows more modeling flexibility than the Dirichlet process. For example, by imposing the NRM as a nonparametric Bayesian prior for the mixing probability in a mixture model, it allows power-law distributions to be modeled properly with efficient MCMC samplers. However, it is clear the NRM fits well in modeling a single dataset, but is inadequate for multiple correlated datasets from different sources. In such cases, a set of correlated NRMs should be used instead of only one. This brings to the problem of building dependent normalized random measures.

In many machine learning tasks, modeling correlated datasets from different sources is a common setting, *e.g*,, blog articles from different websites (*e.g.*, Daily Kos and Right Wing News), papers from different journals (*e.g.*, JMLR, TPAMI and JAIR), genetic data from different groups (*e.g.*, Asian, African and European subpopulations). We want to jointly model these datasets such that:

- they should have information shared between each other.
- they should have their own variations.

One typical solution for the above requirements is by hierarchical modeling where nodes on top of the hierarchy represent the shared information, and those on the bottom represent their own specific variations. For example, in a book, a typical hierarchy is to treat the topic distribution[1] for the whole book as the top node in the hierarchical structure, the child nodes of the book as chapters, with their own topic distribution, a variation of their parent's topic distribution. This goes similarly for the paragraphs within each chapters [Du, 2012]. Information sharing and variations here means that different chapters or paragraphs tend to include the same topics but will have their own uses of words or specific opinions. Figure 4.1 gives an hierarchical structure for the structures of an introductory book on computer science.

[1] A topic in topic models [Blei et al., 2003] is simply a multinomial distribution over the vocabulary words.

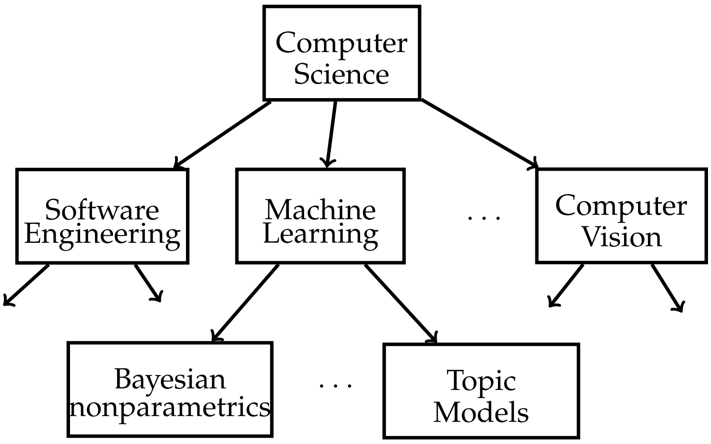


Figure 4.1: A hierarchical structure for an introductory book on *computer science*.

The popular tool in Bayesian nonparametrics for hierarchical modeling is the well known hierarchical Dirichlet process (HDP) [Teh et al., 2006]. This chapter extends the HDP by using the NRM inside the construction instead of DP to allow more flexible distributional modeling. For completeness, the HDP will be introduced first in the following sections.

4.2 Hierarchal Dirichlet Processes

The hierarchical Dirichlet process (HDP) proposed by Teh et al. [2006] is a popular tool for Bayesian nonparametric hierarchical modeling. This chapter starts from the introduction of Dirichlet processes (DP). Here, instead of describing it from the Poisson process point of view as for the NRM in Chapter 3, the original way of introduction by Ferguson [1973] is adopted for simplicity and easy understanding. Intuitively, a DP is an extension of the Dirichlet distribution to an infinite dimensional space. The definition of the Dirichlet distribution is given as:

Definition 4.1 (The Dirichlet Distribution)**.** The K-dimension Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$, denoted as $\mathrm{Dir}(\boldsymbol{\alpha})$, is the probabilistic distribution on the $(K-1)$-simplex, it has a probability density function with respect to Lebesgue measure on the Euclidean space R^{K-1}:

$$p(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K) = \frac{1}{Beta_K(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \tag{4.1}$$

for all $x_i > 0$ and $\sum_i x_i = 1$. $Beta_K(\alpha)$ is the K dimensional beta function that normalizes the Dirichlet[2], defined as:

$$Beta_K(\alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}, \tag{4.2}$$

where $\Gamma(\cdot)$ is the Gamma function.

Now the generalization of the Dirichlet distribution to DP can be clearly seen in the following definition:

Definition 4.2 (The Dirichlet Process [Ferguson, 1973]). Let $\alpha(\cdot)$ be a non-null finite measure (nonnegative and finitely additive) on $(\Theta, \mathcal{B}(\Theta))$. We say D is a Dirichlet process on $(\Theta, \mathcal{B}(\Theta))$ if for every finite or countably infinite measurable partition $(A_1, \cdots, A_k)$ of Θ (see Figure 4.2), the distribution of $(D(A_1), \cdots, D(A_k))$ is Dirichlet distributed as $\mathrm{Dir}(\alpha(A_1), \cdots, \alpha(A_k))$.

The Dirichlet process will be denoted as $\mathrm{DP}(\alpha, H)$ where with a little abuse of notation, α also represents the total mass $\alpha(\Theta)$, and is called the *concentration parameter*, $H = \frac{\alpha(\cdot)}{\alpha(\Theta)}$ is called the base distribution (or base probability measure).

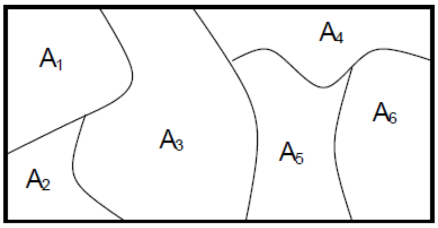


Figure 4.2: Partition of the measurable space

The above definition guarantees the existence of the DP, which follows directly from either Kolmogorov's extension theorem [Sato, 1990] or de Finetti's Theorem [Accardi, 2001]. In the following, two useful results of the DP are listed, proofs can be found in [Ferguson, 1973].

Proposition 4.1 (Discreteness of the DP). *Let D be a Dirichlet process on $(\Theta, \mathcal{B}(\Theta))$ with parameter α, and let $B \in \mathcal{B}(\Theta)$. If $\alpha(B) = 0$, then $D(B) = 0$ with probability 1. If $\alpha(B) > 0$, then $D(B) > 0$ with probability 1. Furthermore, $E(D(B)) = \frac{\alpha(B)}{\alpha(\Theta)}$.*

Remark 4.2. From Proposition 4.1, it can be seen that no matter whether the base measure α is discrete or not, the DP is always discrete, *i.e.*, $D(x) = 0$ for infinite many single points $x \in \Theta$. Thus a realization of the DP can be written as (exactly the same form as the normalized random measure):

$$D = \sum_{k=1}^{\infty} w_k \delta_{x_k} , \tag{4.3}$$

[2] However, $Beta(a, b)$ will be used to denote the Beta probability function with parameters a and b in the thesis.

where $w_k > 0$ and $\sum_k w_k = 1$, $\{x_k\}$ are drawn *i.i.d.* from Θ.

Proposition 4.3. *Let D be a Dirichlet process on $(\Theta, \mathcal{B}(\Theta))$ with parameter α, and let $x_1, \cdots, x_n \in \Theta$ be a sample of size n from D. Then the conditional distribution of D given $x_1, \cdots, x_n$ is again a Dirichlet process with an updated concentration parameter $\alpha + \sum_{i=1}^n \delta_{x_i}$.*

4.2.1 Chinese restaurant processes

Because each draw from the DP is a discrete distribution, in a sample $(x_1, \cdots, x_n)$ of length n, they could have duplicated values (also called *ties*), denoted as $\{\theta_1, \cdots, \theta_K\}$. Now introduce an indicator variable s_i for each data x_i such that $x_i = \theta_{s_i}$, clearly for a given sample $(x_1, \cdots, x_n)$, the indicator variables $(s_1, \cdots, s_n)$ define a partition over integers $\{1, \cdots, n\} \triangleq [n]$, *e.g.*, the i's with the same s_i value belong to the same cluster. The distribution over the partition is known as the Chinese restaurant process (CRP) [Pitman, 1995]. The CRP is so called because it adopts a Chinese restaurant metaphor, where the data are customers in a Chinese restaurant, clusters correspond to dishes. The distribution is equivalent to the seating arrangement induced by the following customer seating process:

- The first customer comes into the restaurant, opens a new table, and orders a dish θ_1 from a global menu.
- Subsequent customers come into the restaurant, and choose a table to sit as follows: for customer n,
  - With probability proportional to m_k to join table k, and share the dish with the other customers on that table, where m_k is the number of customers sharing the dish θ_k on that table.
  - With probability proportional to α to open a new table, and order a dish from the global menu θ.

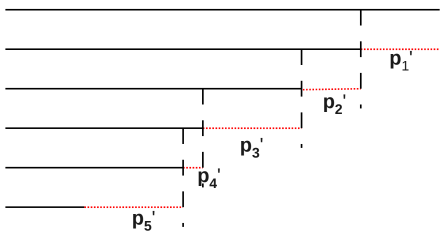
At the end, the customer indexes form a partition over integers $[n]$. It is easy to show that the probability for a specific partition $(m_1, \cdots, m_K)$ has the following form:

$$p(\{m_1, \cdots, m_K\} | \alpha) = \frac{\Gamma(\alpha)\alpha^K}{\Gamma(\alpha + N)} \prod_{k=1}^K \Gamma(m_k) \ . \tag{4.4}$$

The above is known as the Chinese restaurant distribution, and is actually the marginal distribution of the DP (with the random probability measure D marginalized out) given data $x_1, \cdots, x_n$, thus can be used to derive sampling algorithms for the DP.

4.2.2 Stick-breaking construction for the Dirichlet process

The DP also has a nice characterization as the *stick-breaking construction*. If we look at the realization formula of a DP in (4.3), we see that there are infinitely many



1. Draw θ_1 from H
2. Draw p'_1 from Beta$(1, \alpha)$
3. $w_1 = p'_1$
4. Draw θ_2 from H
5. Draw p'_2 from Beta$(1, \alpha)$
6. $w_2 = p'_2(1 - p'_1)$
7. $\cdots$

Figure 4.3: Stick-breaking construction for the Dirichlet process

sequences of positive weights $\{w_k\}$ with a summation of one. We can regard a realization of these weights as stick lengths by breaking a unit length stick into infinite intervals. The following theorem show how to break the stick to make it be a DP.

Theorem 4.4. *Assume the distributions for the random variables $(p'_k)_{k=1}^{\infty}$ and $(\theta_k)_{k=1}^{\infty}$:*

$$p'_k|\alpha_0 \sim Beta(1, \alpha), \qquad \theta_k|H \sim H, \tag{4.5}$$

where $Beta(a, b)$ is the Beta distribution with parameters a and b. Now define $(w_k)_{k=1}^{\infty}$ as

$$w_k = p'_k \prod_{l=1}^{k-1}(1 - p'_l). \tag{4.6}$$

Then the random measure $D = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$ is a Dirichlet process $DP(\alpha, H)$.

Proof. See Sethuraman [1994]. □

Figure 4.3 illustrates the stick-breaking construction for the Dirichlet process. The outcome of the weights $\{w_k\}$ is the same as breaking a unit length stick, thus named stick-breaking process. The detail of stick-breaking construction is described on the right of Figure 4.3.

4.2.3 Hierarchal Dirichlet processes

The hierarchical Dirichlet process puts a set of DPs into a hierarchical structure for correlation modeling as well as information sharing among different DPs [Teh et al., 2006]. Specifically, the output distribution of a DP is subsequently used as the base distribution for another DP, and so-forth. This creates a hierarchy of distributions/probability vectors. This situation is depicted in the graphical model of Figure 4.4.

Probability vector hierarchy: This depicts, for instance, that vectors $\boldsymbol{p}_1$ to $\boldsymbol{p}_K$ should be similar to $\boldsymbol{p}_0$. So for the j_2-th node branching off node j_1, $\boldsymbol{p}_{j_2} \sim \mathrm{DP}(\alpha_{j_1}, \boldsymbol{p}_{j_1})$.

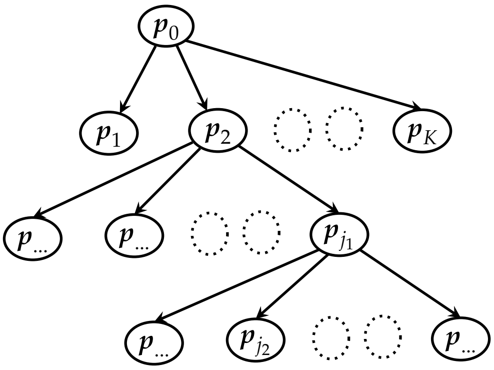


Figure 4.4: Probability vector hierarchy

In formula, these relationships are represented as[3]:

$$
\begin{aligned}
G_0|\alpha_0, H &\sim \mathrm{DP}(\alpha_0, H) && \text{draw a parent DP} \\
G_j|\alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0), && \text{hierarchical construction for } j = 1, \cdots, J
\end{aligned}
$$

The HDP is becoming more and more popular in modeling dependent probability measures since its proposal [Teh et al., 2006], and has found applications in different fields in machine learning such as topic modeling [Teh et al., 2006], n-gram language modeling [Teh, 2006a,b], image segmentation [Orbanz and Buhmann, 2007] and annotation [Du et al., 2009], scene learning [Sudderth et al., 2005], data compression [Wood et al., 2009], and relational modeling [Xu et al., 2006], *etc*. Though simple in construction, it does has some interesting interpretations. Below we briefly introduce the *Chinese restaurant franchise* interpretation of the HDP and the corresponding *stick-breaking construction*.

Chinese restaurant franchise The Chinese restaurant franchise (CRF) is an extension of the Chinese restaurant process to multiple restaurant scenario, where these restaurants are connected in a hierarchical structure and share a global menu. In all of these Chinese restaurants, they have an infinite number of tables, each of which has infinite seating capacity. Each table serves a dish, and multiple tables can serve the same dish. In the CRF, each restaurant connects to its parent restaurant and child restaurants in a tree-like structure. A newly arrived customer in a restaurant can choose to sit at an active table (*i.e.*, a table which at least has one customer), or choose a new table. If a new table is chosen (*i.e.* activated), this table will be sent as a new customer to the corresponding parent restaurant to order a dish, which means a table in any given restaurant reappears as a proxy customer [Mochihashi and Sumita, 2008] in its parent restaurant. This procedure is illustrated in Figure 4.5. The final seating arrangement of customers (including the proxy customers) in the Chinese restaurant franchise constitutes a hierarchical Dirichlet process distribution.

[3] A two level hierarchy case is considered here, generalizing to multiple levels is straightforward

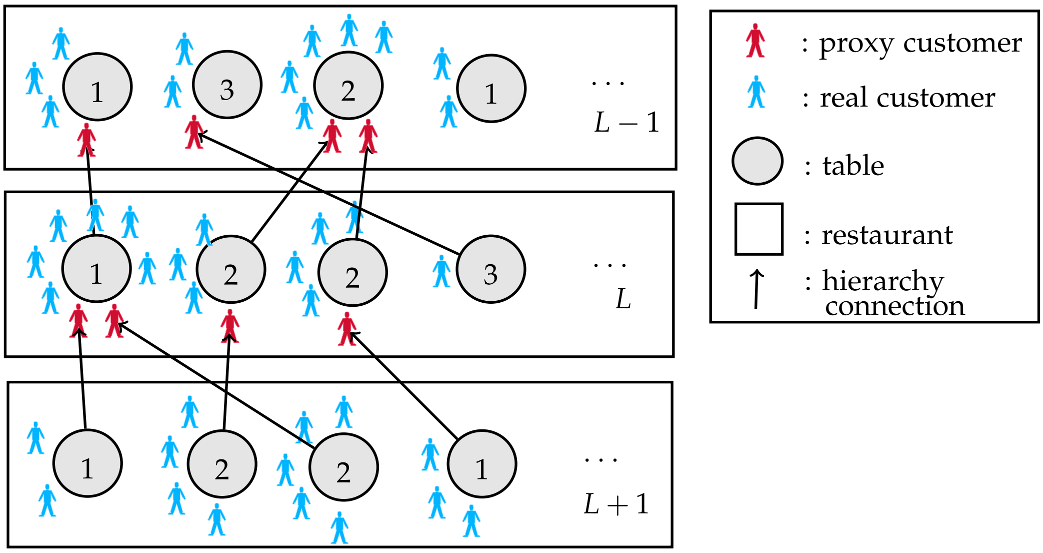


Figure 4.5: CRF representation of the HDP, where rectangles correspond to restaurants, circles correspond to tables, the numbers inside the tables represent dish indexes, and L means "level" in the hierarchy.

Stick-breaking construction for the HDP The stick-breaking construction for the set of DPs in a HDP is also interesting and useful for posterior inference. Teh et al. [2006] prove the formula for the construction. Assume there is an L-level HDP, denote the index of the parent of the current DP G_j as $p(j)$, also let G_j has concentration parameter α_j. Then the stick-breaking construction for the whole set of DPs goes as:

- For the top level DP: $G_0 = \sum_{k=1}^{\infty} w_{0k} \delta_{\theta_k}$:

$$w'_{0k} \sim \text{Beta}\left(1, \alpha_0\right) \Rightarrow w_{0k} = w'_{0k} \prod_{l=1}^{k-1} \left(1 - w'_{0l}\right) .$$

- For the subsequent DP: $G_j = \sum_{k=1}^{\infty} w_{jk} \delta_{\theta^*_k}$:

$$w'_{jk} \sim \text{Beta}\left(\alpha_j w_{p(j)k}, \alpha_j \left(1 - \sum_{l=1}^{k} w_{p(j)l}\right)\right) \Rightarrow w_{jk} = w'_{jk} \prod_{l=1}^{k-1} \left(1 - w'_{jl}\right) .$$

where all the DPs share the same atoms $\{\theta_k\}$, which are drawn *i.i.d.* from the base distribution H of the top level DP G_0.

Posterior inference for the HDP Approximated posterior inference for the HDP can be done via several ways such as MCMC or variational inference. In MCMC, there are mainly three sampling algorithms for the HDP based on the Chinese restaurant franchise and the stick-breaking representations, namely, *sampling in the Chinese restaurant franchise*, *sampling with an augmented representation* and *sampling by direct* *assignment*. These are all presented in [Teh et al., 2006], and will be omitted here for simplicity.

4.3 Hierarchal Normalized Random Measures

By applying the same idea as the hierarchical Dirichlet process, it is straightforward to generalize the HDP to the hierarchical normalized random measure (HNRM). Recall that a NRM is parameterized by some hyperparameters η (if any), a *mass parameter* M, and a *base distribution* H, denoted as $\mathrm{NRM}(\eta, M, H)$. A realization of a NRM is a discrete probability measure, denoted as

$$\mu \sim \mathrm{NRM}(\eta, M, H) \ .$$

Now suppose we have a set of NRMs, say $\{\mathrm{NRM}(\eta_0, M_0, H_0), \mathrm{NRM}(\eta_1, M_1, H_1), \cdots\}$. To define a hierarchical structure on these NRMs, we can replace the base distribution H_i's of $\mathrm{NRM}(\eta_i, M_i, H_i)$ with a realization of its parent NRMs, this construction applies recursively down to other NRMs in a tree structure, and finally we get a set of NRMs which correlate to each other via their base distributions. Specifically, a two layer HNRM mixture is defined as

$$\begin{aligned} \mu_0 &\sim \mathrm{NRM}(\eta_0, M_0, H_0) && \text{a parent NRM} \\ \mu_j &\sim \mathrm{NRM}(\eta_j, M_j, \mu_0) && \text{for each child NRM } j = 1, \cdots, J \\ \psi_{ji} \sim \mu_j, \qquad & x_{ji} \sim F(\cdot|\psi_{ji}) && \text{generate observations } x_{ji} \text{ for } i = 1, \cdots, N_j \ , \end{aligned} \tag{4.7}$$

where $F(\cdot|\psi_{ji})$ is the cumulative density function for generating observations x_{ji}, and we denote the corresponding probability density function as $f(\cdot|\psi_{ji})$, which will be used below.

4.3.1 A generalized Chinese restaurant franchise interpretation of the HNRM

A stick-breaking representation for the general class of HNRM does not seem to be available, thought there exists some work on constructing stick-breaking processes for some specific classes of single NRMs such as the *normalized inverse Gaussian process* [Favaro et al., 2012] and the general *Poisson-Kingman process* [James, 2013]. Fortunately, as a single NRM can be explained as a generalized Chinese restaurant process conditioned on *latent relative mass* James et al. [2009], the HNRM can also be interpreted as a generalized Chinese restaurant franchise. Specifically, in addition to the notion of customers, tables, dishes in traditional CRP, the *latent relative mass* variable u_j for each NRM (each restaurant) now represents the *popularity* variable for the corresponding restaurant, because the seating arrangement of the restaurant is related to this variable. Specifically, the seating process goes as:

- The customer for restaurant μ_j comes into the restaurant:

  - if she is the first customer, she sits at an empty table, orders a dish θ from the global menu.
  - otherwise, she updates the *popularity* u_j of restaurant μ_j using the following posterior:

$$p(u_j|\text{others}) \propto u_j^{n_{j\cdot}} e^{-\psi_{\eta_j}(u_j)} \prod_{k=1}^{K} \kappa(u_j, n_{jk}) ,$$

    where n_{jk} denotes the number of customers in restaurant μ_j sharing dish θ_k, $\psi_{\eta_j}(u)$ and $\kappa(u, m)$ are defined in (3.6) and (3.19), respectively. She then chooses the following options:

    * with probability proportional to $\frac{\kappa(u_j, n_{jk}+1)}{\kappa(u_j, n_{jk})}$ joins an exist table serving dish θ_k;
    * with probability $\kappa(u_j, 1)$ opens a new table and order a dish θ from the global menu.

- Whenever a new table is opened in restaurant μ_j, it serves as a new customer coming into μ_j's parent restaurant $\mu_{p(j)}$, and the seating arrangement in this restaurant follows exactly the same process as its child restaurant μ_j. This process recurses up to the hierarchy until reaching the top level.

The above procedure summarizes the generalized Chinese restaurant franchise interpretation of the HNRM. It can be understood by noting that given customers[4] in each restaurant μ_j, the NRMs represented by these restaurants are independent. As a result, we can use the results from the posterior of a single NRM described in Chapter 3 to get the prediction rules for each restaurant, which correspond to the seating rules described in the above generalized Chinese restaurant franchise.

4.3.2 A marginal sampler for the HNGG

As mentioned above, no closed form stick-breaking construction exists for the general HNRM family, thus posterior inference does not seem to be available from the stick-breaking point of view. However, the marginal sampler based on the above generalized Chinese restaurant franchise interpretation can be derived. This section describes a marginal sampler for a specific class of the HNRM–hierarchical normalized generalized Gamma processes (HNGG). Samplers for the general HNRM follow similar strategies as the HNGG and will be omitted. The auxiliary variable we need to introduce is the *latent relative mass* defined Chapter 3 (corresponds to the *popularity* variables defined above). This section considers the case of two layer HNGG as in (4.7) with NRM replaced by NGG, multiple layer case can be generalized straightforwardly.

Following the Chinese restaurant process metaphor, denote n_{jk} as the number of customers eating dish θ_k in restaurant μ_j where θ_k's are distinct values among all

[4] Note for the internal restaurants their customers are actually tables in their child restaurants.

ψ_{ji}'s in (4.7), t_{jk} as the number of tables serving dish θ_k in restaurant μ_j, K as the total number of distinct dishes currently served in all restaurants. The sampling procedure then mimics the direct assignment sampler for the HDP [Teh et al., 2006]. After integrating out the random measures G_j's (resulting in the generalized Chinese restaurant franchise representation), the variables needed to be sampled are dish index s_{ji} for each customer x_{ji}, the number of tables t_{jk} in restaurant μ_j, the popularity variable u_j for restaurant μ_j and mass hyperparameters M_0 and M for the NGGs in the first and second levels, respectively. Denote the whole parameter set as C, the sampling procedure then goes as:

- **Sampling dish index s_{ji} for customer x_{ji}:** this adopts a similar formula with the HDP, this can be summarized with the following proposition, see appendix in Section 4.7 for the proof.

  Proposition 4.5. *Define β as the prediction probability vector for μ_0 conditioned on other variables, e.g.,*

  $$\beta \propto \left(t_{\cdot 1} - \sigma, \cdots, t_{\cdot K} - \sigma, \sigma M_0 (1 + U_0)^\sigma\right) , \tag{4.8}$$

  such that β is a probability vector. The conditional posterior of s_{ji} satisfies

  $$p(s_{ji} = k | C - s_{ji}) \propto \begin{cases} \left(n_{jk}^{/ji} + \sigma\left(M(1+U_j)^\sigma \beta_k - 1\right)\right) f_k^{\backslash jl}(x_{ji}) & \text{if } k \text{ already exists} \\ \sigma M (1+U_j)^\sigma \beta_k \int_\Theta f(x_{ji}|\theta) h(\theta) \mathrm{d}\theta & \text{if } k = K+1 \text{ is new} \end{cases} , \tag{4.9}$$

  where $^{/ij}$ means the statistics after removing the (i, j) term, h is the density of H; $f_k^{\backslash jl}(x_{jl}) = \frac{\int f(x_{jl}|\theta_k) \prod_{j'l' \neq jl, s_{j'l'}=k, g_{j'l'}=r} f(x_{j'l'}|\theta_k) h(\theta_k) \mathrm{d}\theta_k}{\int \prod_{j'l' \neq jl, s_{j'l'}=k} f(x_{j'l'}|\theta_k) h(\theta_k) \mathrm{d}\theta_k}$ is the conditional density.

  Note since s_{ji}'s change the statistics, the random vector β should be updated in each iteration with μ_0's posterior prediction probabilities according to (4.8) with the current statistics.

- **Sampling restaurant popularity variable u_j:** conditioned on other variables, sampling for u_j is similar to a single NRM case. Based on results from Section 3.5.1.2, the posterior of u_j is:

  $$p(u_j | C - u_j) \propto \frac{u_j^{n_{j\cdot} - 1}}{(1 + u_j)^{n_{j\cdot} - t_{j\cdot}\sigma}} e^{-M(1+u_j)^\sigma} , \tag{4.10}$$

  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 and Wild, 1992] or the slice sampler [Neal, 2003]. Similarly the posterior of u_0

is given by:

$$p(u_0|C - u_0) \propto \frac{u_0^{t_{\cdot\cdot}-1}}{(1+u_0)^{t_{\cdot\cdot}-K\sigma_0}} e^{-M_0(1+u_0)^{\sigma_0}} ,$$

- **Sampling #tables t_{jk} in restaurant μ_t:** this follows by simulating a generalized Chinese restaurant process [Chen et al., 2012b] described in Section 4.3.1. Conditioned on all other statistics, in restaurant μ_j, the probability of creating a new table for dish θ_k is proportional to $(n_{jk} - \sigma)$, while the probability of creating a new table is proportional to $\sigma M(1 + u_j)^{\sigma}$. At the beginning, t_{jk} is initialized to 0, then the customers are added in one by one. If a new table is created, t_{jk} is increased by one. At the end of this generating process, t_{jk} is equal to the number of tables created.

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

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

  where K_j denotes the number distinct dishes served in restaurant μ_j, (a_M, b_M) and (a_0, b_0) are hyperparameters for the Gamma prior of M amd M_0, respectively.

Finally, conditioned on all other variables, sampling σ can be done similarly as the NGG mixture case in Chapter 3, it is omitted here for simplicity. In practice, we can simply choose a suitable value and fix it during inference.

4.4 Experiments

In this section, the use of HNGG in topic modeling [Blei et al., 2003], as well as the efficiency of the sampling algorithms and comparison with some related models are demonstrated

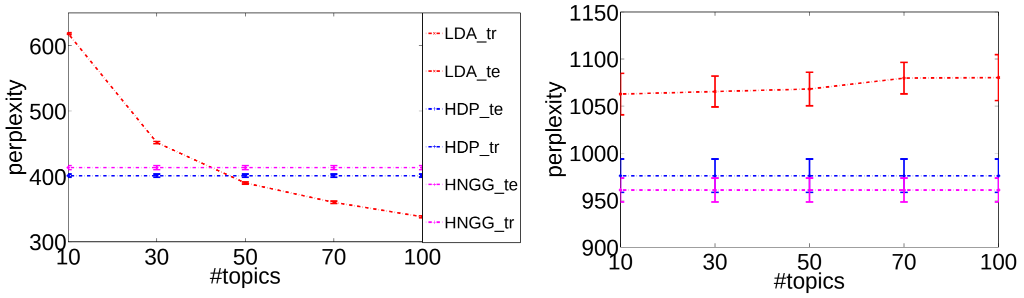
4.4.1 On a small dataset

First, a small dataset of 5 years' abstracts of papers published in ICML (2007–2011) is used. After simple removal of stop words, 764 documents are left with a total of 715,127 words and a vocabulary of size 1,918. In all the experiments, 80% of the documents are randomly chosen for training and 20% for testing. The number of iterations used for burn in is set to 2,000, then another 500 iterations are used to collect related statistics. Each experiment is repeated for 10 times with random initializations, means and variances are reported.

Comparisons The HNGG is compared with LDA [Blei et al., 2003] and HDP [Teh et al., 2006], the standard topic model and the corresponding Bayesian nonparametric extension. All the models are implemented in C/C++. In topic modeling, the base distribution H is a symmetry Dirichlet distribution with parameter β, *e.g.*, each draw ϕ_k from H is a topic-word distribution distributed as

$$\phi_k \sim \text{Dirichlet}_V(\beta) \ ,$$

where V is the vocabulary size. In the experiments, β is set to 0.1, all other parameters are sampled during inferences. Note the *latent relative mass* in HNGG can be sampled using the slice sampler or the adaptive rejection sampler. The slice sampler is used in this experiment, however, these two samplers will be compared in the next experiment. In addition, the σ is chosen as 0.1 in this experiment, other values will be studied later as well. For quantitative evaluation, the *perplexity* measure used in topic models [Wallach et al., 2009] is adopted. Figure 4.6 plots the training and test perplexities for different models, where for LDA, the number of topics varies from 10 to 100. From the figure we can see that with increasing number of topics, training perplexity in LDA keeps dropping, and finally achieves a little bit lower than the HDP and HNGG[5]; while for the test perplexity, HDP and HNGG are consistently better than LDA, with HNGG slight outperforms HDP due to the flexibility in the distributional modeling.



(a) training perplexities

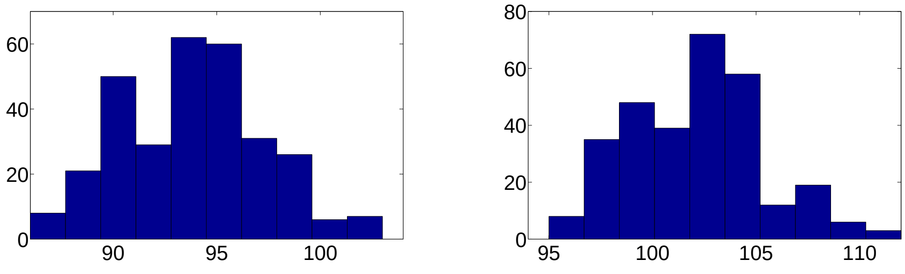
(b) test perplexities

Figure 4.6: Training and test perplexities for different models.

The posterior number of topics learned from HNGG and HDP are plotted in Figure 4.7. It can be seen that with $\sigma = 0.1$, HNGG generates less number of topics, indicating word distributions in HNGG tend to be more compact than those in HDP.

Adaptive rejection sampler VS. slice sampler Note due to the log-concave property of the restaurant popularity variables u_j's, they can be sampled with either slice or adaptive rejection samplers. In this experiment these two sampling schema are compared. The two samplers are implemented based on codes from [Johnson]

[5] In practice training perplexity for different models is not a fair metric for comparison, usually test perplexity is used instead to evaluate the generalization ability of the models.

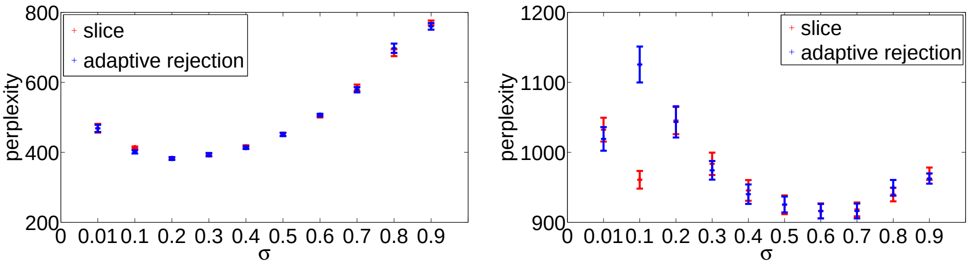


(a) HNGG

(b) HDP

Figure 4.7: Posterior number of topics learned in HNGG and HDP.

and [Gilks], respectively. Training and test perplexities are compared with σ ranging among $(0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$. Figure 4.8 plots the compared results for the two schema. From the figure we can find that these two sampling schema are comparable in term of perplexity, but statistically the adaptive rejection sampler seems slightly better.



(a) training

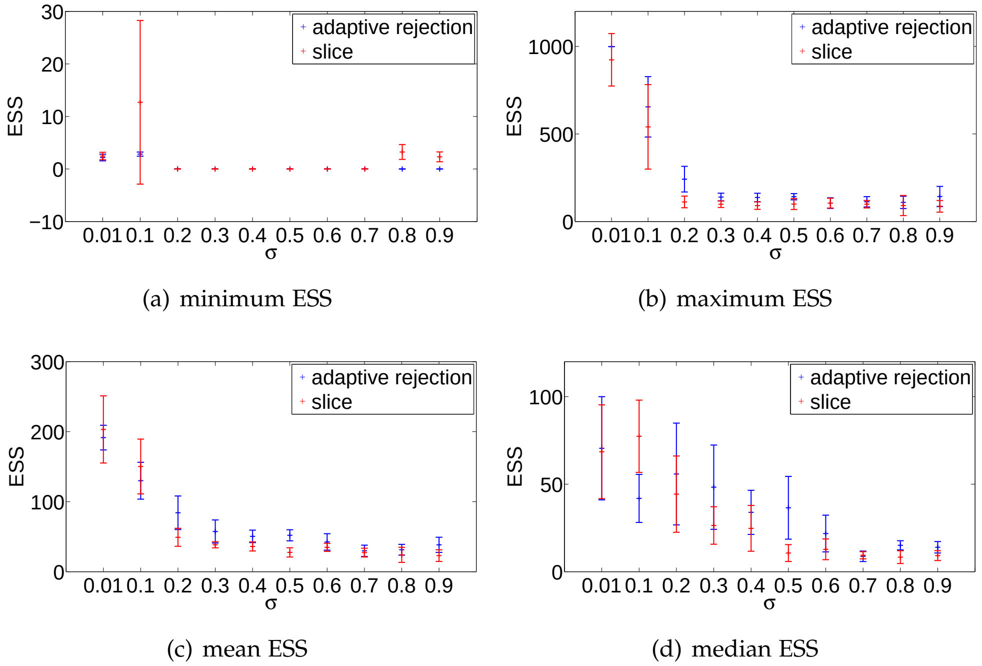
(b) testing

Figure 4.8: Comparison with the slice sampler and adaptive rejection sampler.

To further test how these two sampling schema impact the model, the *effective sample size* (ESS) among 1000 iterations are calculated. The mass parameters M_j, summation of u_j, the total number of topics K, training and test likelihoods are chosen as the statistics for ESS calculation. Figure 4.9 shows the minimum, maximum, mean and median ESS for the two samplers. We can see from the figures that although the two samplers perform quite similarly, again statistically the adaptive rejection sampler seems to has higher ESS than the slice sampler.

4.4.2 On a larger dataset

Now the HNGG is applied to a larger dataset–a political blog dataset containing six political blogs about the U.S. presidential election [Eisenstein et al., 2011]. The same procedure is performed as in [Eisenstein et al., 2011] to pre-process this data. Finally, the dataset is composed of 9461 documents with vocabulary size 13,644 and 2,051,708 words. The parameter setting is the same as in the small dataset in the last section.



(a) minimum ESS

(b) maximum ESS

(c) mean ESS

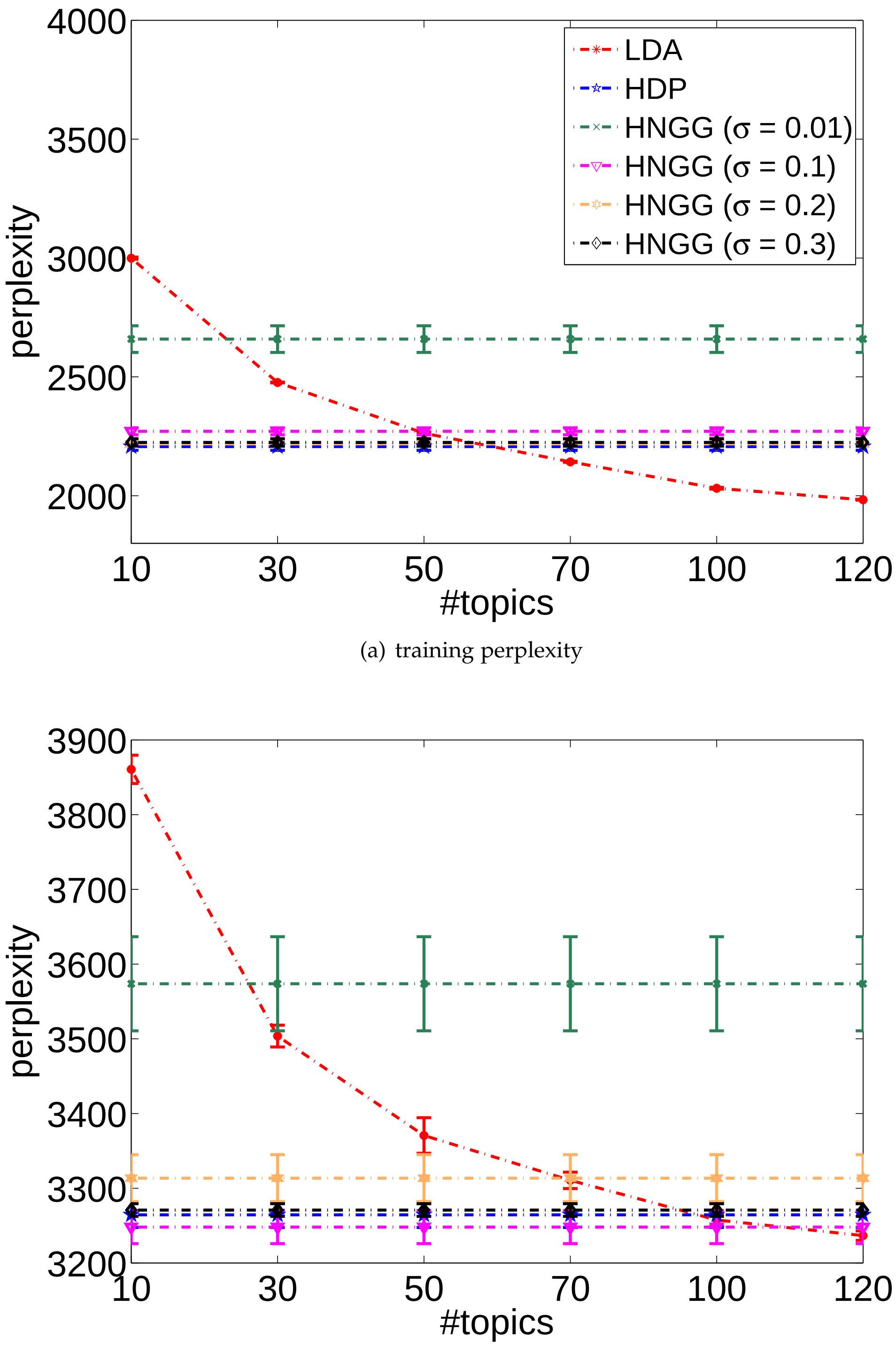
(d) median ESS

Figure 4.9: Effective sample sizes for the slice sampler and adaptive rejection sampler.

First, the training and test perplexities are compared using HNGG, HDP and LDA. For HNGG, four settings of the σ parameter are tested, *i.e.*, $\sigma = (0.01, 0.1, 0.2, 0.3)$, respectively, while all other parameters are sampled during inference. Also, the adaptive rejection sampler is used since it is shown to be better than the slice sampler in the last section. For the LDA model, the number of topics varies among (10, 30, 50, 70, 100), 120). Each experiment is repeated for 5 times, means and variances of the results are reported. Figure 4.10 plots the training and test perplexities for the three models. We can see that HDP and HNGG are able to generate comparable perplexities to the best LDA model (with number of topics being 120), whereas HNGG with parameter $\sigma = (0.1, 0.2, 0.3)$ is comparable to the HDP (the HNGG is better than the HDP in test perplexity with $\sigma = 0.1$), indicating more flexibility of the HNGG than the HDP in term of power-law distribution modeling. Next, as an illustration, Table 4.1 shows top 10 words from 10 randomly chosen topics learned by the HNGG. We can see clearly that HNGG successfully recovers some interesting topics hidden in the dataset. Finally, note that the inference algorithm for the HNGG is fairy fast. In the experiments, it is observed comparable running time with the HDP model, demonstrating the efficiency of the proposed sampling algorithm.

4.5 Topic-word distribution modeling

To show the flexibility of HNRM (specifically HNGG) over models like HDP, we compare them in topic models where HNGG/HDP are used to model topic-word



(a) training perplexity

(b) test perplexity

Figure 4.10: Comparison of LDA, HDP and HNGG on the CMU dataset

Table 4.1: 10 topics learned from HNGG on CMU dataset

"taxes"	"health care"	"voting"	"people & family"	"middle east"	"climate" change"	"financial crisis"
tax	health	Obama	women	Obama	global	financial
governm't	care	mccain	life	Israel	warming	crisis
economic	union	percent	family	Jewish	climate	bailout
economy	workers	voters	us	Barack	change	governm't
taxes	insurance	poll	people	policy	gore	market
spending	auto	among	young	American	science	wall
money	labor	democrats	day	Jews	scientific	street
billion	industry	polls	men	foreign	ice	fannie
pay	unions	points	American	middle	earth	treasury
jobs	companies	election	children	east	scientists	mortgage

distributions over all documents instead of topic distributions for each document. The generative process can be described as:

- Draw a global topic-word distribution:

$$\phi^0 \sim \mathrm{NGG}(\sigma_0, M_0, H) \ ,$$

  where H is a uniform distribution over vocabulary words.

- For each topic, draw its topic-word distribution:

$$\phi_k \sim \mathrm{NGG}(\sigma, M, \phi^0), \quad k = 1, 2, \cdots, K \ .$$

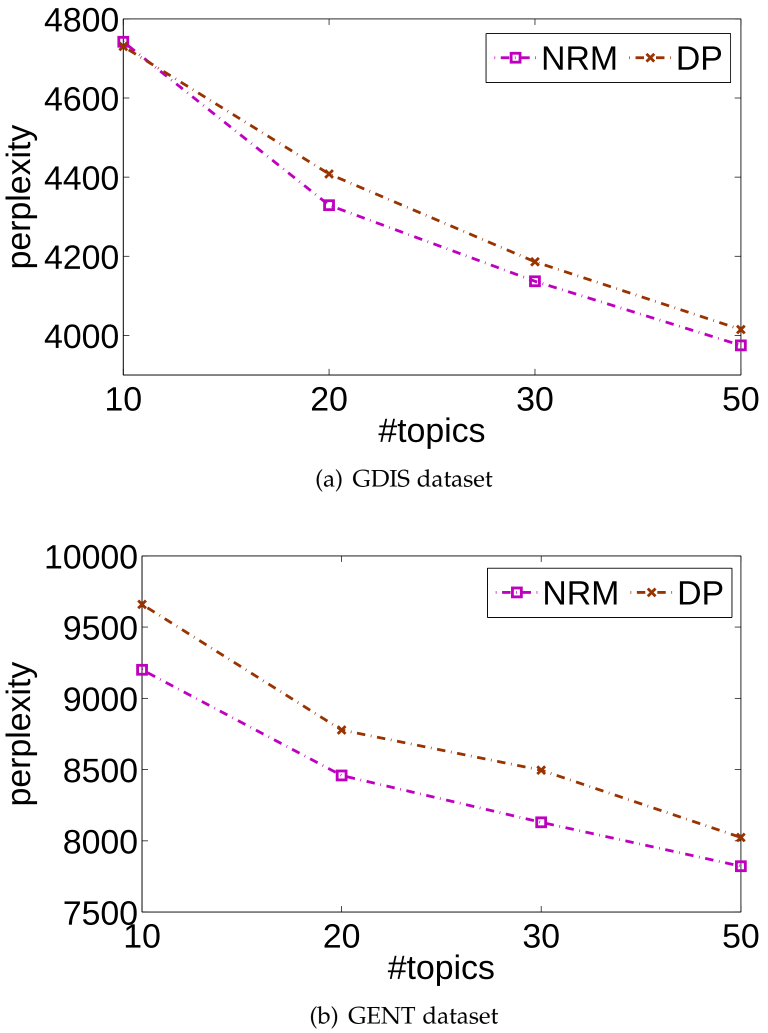
- For each document d:
  - draw its topic distribution: $\theta_d \sim \mathrm{Dir}(\alpha)$.
  - for each word index ℓ in document d:
    - ∗ draw its topic indicator: $z_{d\ell} \sim \mathrm{Dir}(\theta_d)$.
    - ∗ draw the observed word: $w_{d\ell} \sim \mathrm{Dir}(\phi_{z_{d\ell}})$.

The HDP version of the model is obtained by replacing HNGG with HDP in the above generative process. As it is known that natural language exhibits Zipf's law which is in correspondence with the power-law property in NGG, we expect models using the NGG to perform better than those using DP. For this experiment, we extract two datasets from the *Reuters RCV1* collection[6] about disasters and entertainment, the Reuters categories GDIS and GENT respectively. Sentences were parsed with the C&C Parser[7], then lemmatised and function words discarded. GDIS has a vocabulary of size 39534 and total documents of 9097; while the GENT dataset has 4126 documents

[6] Reuters Corpus, Volume 1, English language, 1996-08-20 to 1997-08-19 (Release date 2000-11-03).

[7] http://svn.ask.it.usyd.edu.au/trac/candc

and 43990 words in the vocabulary. The number of topics in the models are set between $\{10, 20, 30, 50\}$, the Dirichlet prior for the topic distribution is set to 0.1, while sampling all other parameters during inference. A burn in of 1000 iterations is used in the experiment followed by 200 iterations to collect samples from the posterior. Test perplexities are then calculated based on a 20% held-out data, and are plotted in Figure 4.11. It can be seen from the figure that models using NGG are consistently better than those using DP in term of test perplexity, demonstrating the advantages of NGG over DP.



(a) GDIS dataset

(b) GENT dataset

Figure 4.11: Comparison of HDP and HNGG for topic-word distribution modeling on GDIS and GENT datasets.

4.6 Conclusion

This chapter presents the first dependent normalized random measure to model hierarchical dependency, thus called hierarchical normalized random measures (HNRM).

The construction follows the same procedure with the hierarchical Dirichlet process (HDP) [Teh et al., 2006]. Due to the complicated posterior structure, the HNRM does not endow convenient properties as with the HDP, such as the stick-breaking construction. However, by applying the auxiliary trick as in the normalized random measure framework, posterior inference for the HNRM can still be done via efficient marginal Gibbs sampler[8]. A generalized Chinese restaurant franchise interpretation is also proposed to describe the posterior sampling for the HNRM. Typical topic model experiments are tested on the HNRM, which is compared with the LDA and HDP-LDA topic models. Experimental results show that HNRM together with the HDP overcome the model selection difficulty in traditional Bayesian models, *i.e.*, choosing the right number of topics, and obtain comparable results to the best LDA model. Furthermore, experimental results also show an improvement in perplexity of the HNRM over the HDP in topic distribution modeling, indicating the greater flexibility of the HNRM. Note that the HNRM is reminiscent of the Coag-Frag duality of a class of stable Poisson-Kingman mixtures James [2010, 2013] but with distinct posterior inference techniques. However, it would be interesting to borrow ideas from James [2010] to construct continuous time generalized Chinese restaurant processes, like the fragmentation-coagulation process does Teh et al. [2011].

4.7 Appendix: Proofs

Proof of Proposition 4.5. To see how the above formula is obtained, note that for x_{ji} in restaurant μ_j, we have

$$\mu_0 \sim \text{NGG}(\sigma_0, M_0, H), \qquad \mu_j \sim \text{NGG}(\sigma, M, \mu_0) \; .$$

Furthermore, according to Corollary 3.12, conditioned on other statistics, the posterior of μ_0 is a NRM μ'_0 expressed as:

$$\mu'_0 = \mu_0 + \sum_{k=1}^{K} \frac{w_{0k}}{\sum_{k'} w_{0k'}} \delta_{\theta_k} = \mu_0 + \sum_{k=1}^{K} \beta_k \delta_{\theta_k} \; . \tag{4.11}$$

The last equation satisfies because w_{0k}'s ($k \leq K$) represent the jump sizes of the atoms with observations, thus are the posterior prediction probabilities of μ_0. As a result, according to Corollary 3.17, the predicted distributions can be written as

$$p(s_{ji} = k | C - s_{ji}) \propto \begin{cases} \left(n_{jk}^{/ji} - \sigma\right) \mu'_0(\theta_k) & \text{if } k \text{ already exists} \\ \sigma M (1 + U_j)^{\sigma} \mu'_0(\Theta / \{\theta_{k'}\}_{k'=1}^{K}) & \text{if } k \text{ is new ,} \end{cases} \tag{4.12}$$

Substituting (4.11) into the above predicted probabilities and simplifying results in (4.9). □

[8] Almost the same running time on average compared to HDP.

Chapter 5

Dynamic Topic Modeling with Dependent Hierarchical Normalized Random Measures

5.1 Introduction

The hierarchical normalized random measure (HNRM) introduced in Chapter 4 is a flexible and powerful tool for hierarchical dependency modeling. However, real data might not only exhibit hierarchical dependency; other kinds of dependencies such as the Markovian dependency are also desired. This chapter presents a time dependent hierarchical model by extending the HNRM with Markovian dependent structure to describe time evolving phenomena in dynamic topic modeling.

The motivation for the proposed model is to describe topic evolution in topic models – dynamic topic models. In dynamic topic models, we want both hierarchical and Markovian dependencies, where the former models documents in the same time span, whereas the later models topic evolution over time: current topics depend on topics from previous time and will influence future topics as well. This chapter combines the ideas of HDP/HNRM with the Markovian dependency operations [Lin et al., 2010], and proposes a Markovian dependency hierarchical normalized random measures by manipulating the underlying Poisson processes and the corresponding completely random measures [Kingman, 1967]. These operators in Markovian dependency modeling are intuitive and allow flexibly control of topic correlations. Note a related construction in the statistical literature is made by A. Lijoi and B. Nipoti and I. Prunster [2013a], but it deals only with modeling two groups of data.

As hierarchical modeling via HNRM has been studied in Chapter 4, the Markovian dependency modeling will be the focus of this chapter. As is shown in previous chapters, an NRM is constructed from the Poisson process, thus to construct *Markovian dependent NRMs*, it suffices to construct *Markovian dependent Poisson processes*. This is achieved by defining some dependent operations on the Poisson process. Such a construction not only achieves more flexible modeling, but also allows a dependency structure to be theoretically analyzed. In the following sections, the dependency operations on Poisson processes, *e.g., superposition, subsampling and point* *transition* stated in Theorem 2.4, Corollary 2.8 and Corollary 2.9, are adapted to CRMs and NRMs in Section 5.2. Properties when applying these dependency operations to NRMs are then given in Section 5.3. The dynamic topic model based on these operations is presented in Section 5.4 with experiments given in Section 5.5. Finally proofs are given in the Appendix, Section 5.7.

5.2 Markovian Dependency Operations

This section introduces the dependency operations used in this chapter. These are developed for CRMs and NRMs adapted from those in Poisson processes introduced in Chapter 2.

5.2.1 Operations on CRMs

The dependency operations defined on Poisson processes in Chapter 2 can be naturally generalized to the completely random measures given the construction of (3.2). Formally, we have

Definition 5.1 (Superposition of CRMs)**.** Given n independent CRMs $\tilde{\mu}_1, \cdots, \tilde{\mu}_n$ on Θ, the superposition ($\tilde{\oplus}$) of the CRMs is defined as:

$$\tilde{\mu}_1 \tilde{\oplus} \tilde{\mu}_2 \tilde{\oplus} \cdots \tilde{\oplus} \tilde{\mu}_n := \mu_1 + \mu_2 + \cdots + \mu_n \, .$$

Definition 5.2 (Subsampling of CRMs)**.** Given a CRM $\tilde{\mu} = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$ on Θ, and a measurable function $q : \Theta \to [0,1]$. If we independently draw $z(\theta) \in \{0,1\}$ for each $\theta \in \Theta$ with $p(z(\theta) = 1) = q(\theta)$, the subsampling of $\tilde{\mu}$, is defined as

$$\tilde{S}^q(\tilde{\mu}) := \sum_k z(\theta_k) w_k \delta_{\theta_k}, \tag{5.1}$$

Definition 5.3 (Point transition of CRMs)**.** Given a CRM $\tilde{\mu} = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$ on Θ, the point transition of $\tilde{\mu}$, is to draw atoms θ'_k from a transformed base measure to yield a new random measure as

$$\tilde{T}(\tilde{\mu}) := \sum_{k=1}^{\infty} w_k \delta_{\theta'_k} \, ,$$

where $\theta'_k \sim \tilde{T}(\theta_k)$ and $\tilde{T} : \Theta \mapsto \Theta$ is a transition kernel.

5.2.2 Operations on NRMs

The operations on NRMs can be naturally generalized from those on CRMs by doing a normalization step:

Definition 5.4 (Superposition of NRMs)**.** Given n independent NRMs $\mu_1, \cdots, \mu_n$ on Θ, the superposition ($\oplus$) of NRMs is defined as:

$$\mu_1 \oplus \mu_2 \oplus \cdots \oplus \mu_n := c_1 \mu_1 + c_2 \mu_2 + \cdots + c_n \mu_n \, .$$

where the weights $c_m = \frac{\tilde{\mu}_m(\Theta)}{\sum_j \tilde{\mu}_j(\Theta)}$ and $\tilde{\mu}_m$ is the unnormalized random measures corresponding to μ_m.

Definition 5.5 (Subsampling of NRMs)**.** Given a NRM $\mu = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$ on Θ, and a measurable function $q : \Theta \to [0,1]$. If we independently draw $z(\theta) \in \{0,1\}$ for each $\theta \in \Theta$ with $p(z(\theta) = 1) = q(\theta)$, the subsampling of μ, is defined as

$$S^q(\mu) := \sum_{k:z(\theta_k)=1} \frac{w_k}{\sum_j z(\theta_j) w_j} \delta_{\theta_k} , \tag{5.2}$$

Definition 5.6 (Point transition of NRMs)**.** Given a NRM $\mu = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$ on Θ, the point transition of μ, is to draw atoms θ'_k from a transformed base measure to yield a new NRM as

$$T(\mu) := \sum_{k=1}^{\infty} w_k \delta_{\theta'_k} ,$$

where $\theta'_k \sim \tilde{T}(\theta_k)$ and $\tilde{T} : \Theta \mapsto \Theta$ is a transition kernel. The definitions are constructed so the following simple lemma holds.

Lemma 5.1. *Superposition, subsampling or point transition of NRMs is equivalent to superposition, subsampling or point transition of their underlying CRMs.*

Thus one does not need to distinguish between whether these operations are on CRMs or NRMs.

5.3 Dependencies and Properties of Operations

Based on the dependency operators defined above, this section presents a number of results to do with these operations applied to the NRMs. First, dependencies such as covariances are presented. Then some further properties are developed for when the operations are used in a network.

5.3.1 Dependencies between NRMs via operations

Properties of the NRMs here are given in terms of the Laplace exponent $\psi(v)$ defined in (3.2) and its derivatives. In the Dirichlet process case, we have $\psi(v) = M \log(1+v)$, while in the normalized generalized Gamma process case, we have $\psi_a(v) = M\left((1+v)^a - 1\right)$. Because the dependencies involve the total mass significantly, a modified version of the Laplace exponent is used in all these results, defined as $\tilde{\psi}_\eta(v) = \frac{1}{M}\psi_\eta(v)$ with the mass parameter M removed.

Different from the Dirichlet process, the total masses M are no longer independent from their normalized jumps in general normalized random measures. However, the correlations between different NRMs can still be derived. The following Theorems summarize these results.

Lemma 5.2 (Mean and Variance of an NRM). *Given a normalized random measure μ on Θ with the underlying Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta) = M\rho_\eta(\mathrm{d}w)H(\mathrm{d}\theta)$, for $\forall B \in \mathcal{B}(\Theta)$. The mean of this NRM is given by*

$$\mathbb{E}[\mu(B)] = H(B)\ . \tag{5.3}$$

The variance of this NRM is given by

$$\begin{aligned} Var(\mu(B)) &= H(B)(H(B)-1)M \\ &\int_0^\infty v\tilde{\psi}''_\eta(v)\exp\left\{-M\tilde{\psi}_\eta(v)\right\}\mathrm{d}v\ . \end{aligned} \tag{5.4}$$

Remark 5.3. For DP, the corresponding variances are:

$$\mathrm{Var}_{DP}(\mu(B)) = \frac{H(B)(1-H(B))}{M+1}.$$

For NGG, it is

$$\mathrm{Var}_{NGG}(\mu(B)) = H(B)(1-H(B))\frac{1-\sigma}{\sigma}e^M M^{\frac{1}{\sigma}}|\Gamma(-\frac{1}{\sigma}, M)|.$$

For large M the upper incomplete Gamma function used here has the property that $e^M M^{1+\frac{1}{\sigma}}|\Gamma(-\frac{1}{\sigma}, M)| \to 1$ and so we get for large M

$$\mathrm{Var}_{NGG}(\mu(B)) \to H(B)(1-H(B))\frac{1-\sigma}{M\sigma}\ .$$

Theorem 5.4 (Dependency via superposition). *Suppose $\mu_i, i = 1, \cdots, n$ are n independent normalized random measures on Θ with the underlying Lévy measures $\nu_i(\mathrm{d}w, \mathrm{d}\theta) = M_i\rho_\eta(\mathrm{d}w)H(\mathrm{d}\theta)$, let $\mu = \mu_1 \oplus \cdots \oplus \mu_n$, $B \in \mathcal{B}(\Theta)$, then the covariance between $\mu_k (k < n)$ and μ is*

$$\begin{aligned} &Cov\left(\mu_k(B), \mu(B)\right) = \\ &H(B)M_k\int_0^\infty \gamma(M_k, H(B), v)\exp\left\{-(\sum_{j\neq k} M_j)\tilde{\psi}_\eta(v)\right\}\mathrm{d}v \\ &+H(B)^2\left(\frac{2M_k}{\sum_j M_j} - 1\right)\ . \end{aligned} \tag{5.5}$$

where

$$\begin{aligned} &\gamma(M_k, H(B), v) = \\ &\int_0^v \left(H(B)M_k\tilde{\psi}'_\eta(v_1)^2 - \tilde{\psi}''_\eta(v_1)\right)\exp\left\{-M_k\tilde{\psi}_\eta(v_1)\right\}\mathrm{d}v_1 \end{aligned} \tag{5.6}$$

Theorem 5.5 (Dependency via subsampling). *Let $\tilde{\mu}$ be a completely random measure on Θ with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta) = M\rho_\eta(\mathrm{d}w)H(\mathrm{d}\theta)$, $\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\Theta)}$. The covariance between μ*

and its subsampling version $S^q(\mu)$, denoted as μ^q, with sampling rate $q(\cdot)$ on $B \in \mathcal{B}(\Theta)$ is

$$\begin{aligned} &Cov\left(\mu^q(B), \mu(B)\right) = \\ &H(B)M_q \int_0^\infty \gamma(M_q, H(B), v) \exp\left\{-(M - M_q)\tilde{\psi}_\eta(v)\right\} \mathrm{d}v \\ &+ \;\; H(B)^2 \left(\frac{2M_q - M}{M}\right), \end{aligned} \tag{5.7}$$

where $M_q := (q\tilde{\mu})(\Theta) = \int_\Theta q(\theta)\tilde{\mu}(\theta)\mathrm{d}\theta$.

Theorem 5.6 (Dependency via point transition). *Let $\tilde{\mu}$ be a random measure on Θ with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta) = M\rho_\eta(\mathrm{d}w)H(\mathrm{d}\theta)$, $\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\Theta)}$. Let $B \in \mathcal{B}(\Theta)$, $A = \mathcal{T}(B) := \{x : x \sim \mathcal{T}(y, \cdot), y \in B\}$ be the set of points obtained after the point transition on B, thus $H(A) = \int_B H(\mathcal{T}(x))\mathrm{d}x$. Suppose A and B are disjoint (which is usually the case when the transition operator T is appropriately defined), the covariance between μ and its point transition version $T(\mu)$ on $B \in \mathcal{B}(\Theta)$ is*

$$\begin{aligned} &Cov\left(\mu(B), (T\mu)(B)\right) = H(A)H(B) \\ &\left(M^2 \int_0^\infty \int_0^{v_1} \tilde{\psi}'_\eta(v_2)^2 \exp\left\{-M\tilde{\psi}_\eta(v_2)\right\} \mathrm{d}v_2 \mathrm{d}v_1 - 1\right) \end{aligned} \tag{5.8}$$

5.3.2 Properties of the three dependency operations

To start with, the following two Lemmas about superposition and subsampling of CRMs are proven. First, a straightforward extension of [Theorem 1 James et al., 2009] leads to the following Lemma about the posterior of CRMs under superposition.

Lemma 5.7 (Posterior of CRMs under superposition). *Let $\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_n$ be n independent CRMs defined on space Θ, with Lévy measures $\nu_i(\mathrm{d}w, \mathrm{d}\theta) = M_i\rho_i(\mathrm{d}w)H(\mathrm{d}\theta)$ for $i = 1, \cdots, n$. Let*

$$\tilde{\mu} = \oplus_{i=1}^n \tilde{\mu}_i. \tag{5.9}$$

Then given observed data $X = (x_i \in \Theta)_{i=1}^N$ with distinct values $\{\theta_k\}$ from $\tilde{\mu}/\tilde{\mu}(\Theta)$, and a latent relative mass u for $\tilde{\mu}'$, the posterior of $\tilde{\mu}$, denoted as $\tilde{\mu}'$, is given by

$$\tilde{\mu}' = \tilde{\mu}_0 + \sum_{k=1}^K w_k \delta_{\theta_k}, \tag{5.10}$$

where

1. *$\tilde{\mu}_0$ is a CRM with Lévy measure*

$$\nu(\mathrm{d}w, \mathrm{d}\theta) = e^{-uw}\left(\sum_{i=1}^n \nu_i(\mathrm{d}w, \mathrm{d}\theta)\right),$$

2. *θ_k ($k = 1, \cdots, K$) are the fixed points of discontinuity and w_k's are the corresponding*

jumps with densities proportional to

$$p(w_k|\cdot) \propto w_k^{n_k} e^{-uw_k} \left(\sum_{i=1}^{n} \nu_i(w_k, \Theta) \right) ,$$

where n_k is the number of data attached at jump w_k.

3. *$\tilde{\mu}_0$ and w_k's are independent.*

Furthermore, the posterior of u is given by

$$p(u|others) \propto u^N e^{-\sum_i M_i \int_{\mathbb{R}^+}(1-\exp(-uw))\rho_i(\mathrm{d}w)} \prod_k \int_{\mathbb{R}^+} w^{n_k} e^{-uw} \rho_{r(k)}(\mathrm{d}w) ,$$

where $r(k)$ indexes which $\tilde{\mu}_i$ the atom θ_k comes from.

Second, the following formula of the Lévy measure under different dependency operations can also be proved.

Lemma 5.8 (Lévy measure under dependency operations)**.** *Let $\tilde{\mu} = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$ be a CRM with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta)$.*

- *Let $S^q(\tilde{\mu})$ be its subsampling version with acceptance rate $q(\cdot)$, then $S^q(\tilde{\mu})$ has the Lévy measure of $q(\theta)\nu(\mathrm{d}w, \mathrm{d}\theta)$.*
- *Let $T(\tilde{\mu})$ be its point transition version, where $\nu(\mathrm{d}w, \mathrm{d}\theta) = M\rho_\eta(\mathrm{d}w)H(\mathrm{d}\theta)$. Then its Lévy measure is $M\rho_\eta(\mathrm{d}w)T(H)(\mathrm{d}\theta)$ where $T(H)$ is the transformed base measure.*
- *Let $\tilde{\mu}_1 \oplus \tilde{\mu}_2$ be the superposition, then its Lévy measure is $\nu_1(\mathrm{d}w, \mathrm{d}\theta) + \nu_2(\mathrm{d}w, \mathrm{d}\theta)$.*

Now based on the above lemmas, some properties about compositions of the dependency operations are given, which follow simply.

Lemma 5.9 (Composition of dependency operators)**.** *Given CRMs $\tilde{\mu}$, $\tilde{\mu}'$ and $\tilde{\mu}''$, the following hold:*

- *Two subsampling operations are commutative. So with acceptance rates $q(\cdot)$ and $q'(\cdot)$, then $S^{q'}(S^q(\tilde{\mu})) = S^q(S^{q'}(\tilde{\mu}))$. Both are equal to $S^{q'q}(\tilde{\mu}))$.*
- *A constant subsampling operation commutes with a point transition operation. Thus $S^q(T(\tilde{\mu})) = T(S^q(\tilde{\mu}))$ where the acceptance rate q is independent of the data space.*
- *Subsampling and point transition operations distribute over superposition. Thus for acceptance rate $q(\cdot)$ and point transition $T(\cdot)$,*

$$S^q(\tilde{\mu} \oplus \tilde{\mu}') = S^q(\tilde{\mu}) \oplus S^q(\tilde{\mu}') , \qquad T(\tilde{\mu} \oplus \tilde{\mu}') = T(\tilde{\mu}) \oplus T(\tilde{\mu}') .$$

- *Superposition is commutative and associative. Thus $\tilde{\mu} \oplus \tilde{\mu}' = \tilde{\mu}' \oplus \tilde{\mu}$ and $(\tilde{\mu} \oplus \tilde{\mu}') \oplus \tilde{\mu}'' = \tilde{\mu} \oplus (\tilde{\mu}' \oplus \tilde{\mu}'')$.*

Thus when subsampling operations are all constant, a composition of subsampling, point transition and superposition operations admits a normal form where all the subsampling operations are applied first, then the transition operations and lastly the superposition operations.

Lemma 5.10 (Normal form for compositions). *Assume subsampling operations all have a constant acceptance rate. A normal form for a composition of subsampling, point transition and superposition operations is obtained by applying the following rules until no further can apply.*

$$\begin{aligned} S^q(S^{q'}(\tilde{\mu})) &\rightarrow S^{qq'}(\tilde{\mu})) \, , \\ S^q(T(\tilde{\mu})) &\rightarrow T(S^q(\tilde{\mu})) \, , \\ S^q(\tilde{\mu} \oplus \tilde{\mu}') &\rightarrow S^q(\tilde{\mu}) \oplus S^q(\tilde{\mu}') \, , \\ T(\tilde{\mu} \oplus \tilde{\mu}') &\rightarrow T(\tilde{\mu}) \oplus T(\tilde{\mu}') \, . \end{aligned}$$

The remaining top level set of superpositions are then flattened out by removing any precedence ordering.

Finally, note that Lemmas 5.7, 5.8, 5.9 and 5.10 all apply to NRMs as well due to Lemma 5.1, thus the specific results for the NRM will be omitted here.

5.4 A Dynamic Topic Model based on dependent NRMs

5.4.1 Motivation

As is shown before, dependency modeling with HDP or HNRM is appealing because of its ability of flexible dependency modeling as well as the ease of implementations. However, when modeling dynamic data, they do not fit well because the underlying assumption of HDP and HNRM is the full exchangeability of the DPs/NRMs, this violates the intuition that we might want the content of ICML literature depends on previous years' so order is important.

To overcome the full exchangeability limitation, several dependent Dirichlet process models have been proposed, for example, the dynamic HDP [Ren et al., 2008], the evolutionary HDP [Zhang et al., 2010], and the recurrent Chinese Restaurant process [Ahmed and Xing, 2010]. Dirichlet processes are used partly because of their simplicity and conjugacy which make the posterior inference easy [James et al., 2006]. These models are constructed by incorporating the previous DP's into the base distribution of the current DP. Markovian dependent DPs have also been constructed using the underlying Poisson processes [Lin et al., 2010]. However, recent research has shown that many real datasets have the power-law property, *e.g.*, in images [Sudderth and Jordan, 2008], in topic-word distributions [Teh, 2006a], in language models [Goldwater et al., 2006; Johnson et al., 2007] and in document topic (label) distributions [Rubin et al., 2011]. This makes the Dirichlet process an improper tool for modeling these datasets.

Although there also exists some dependent nonparametric models with power-law phenomena, their dependencies are limited. For example, Bartlett et al. [2010] proposed a dependent hierarchical Pitman-Yor process that only allows deletion of atoms, while Sudderth and Jordan [2008] construct the dependent Pitman-Yor process by only allowing dependencies between atoms.

In the following, using the dependency operations defined above, a time dependent hierarchical model for dynamic topic modeling based on NRMs is constructed. By this, the dependencies are flexibly controlled between atoms of the NRMs, resulting in more flexible dependency modeling.

5.4.2 Construction

The main interest of the model is to construct a dynamic topic model that inherits *partial* exchangeability, meaning that the documents within each time frame are exchangeable, while between time frames they are not. To achieve this, it is crucial to model the dependency of the topics between different time frames. In particular, a topic can either inherit from the topics of earlier time frames with certain transformation, or be a completely new one which is "born" in the current time frame. The above idea can be modeled by a series of hierarchical NRMs, one per time frame. Between the time frames, these hierarchical NRMs depend on each other through three dependency operators – *superposition*, *subsampling* and *point transition*, which are defined previously. The corresponding graphical model is shown in Figure 5.1(left) and the generating process for the model is as follows:

- Generating independent NRMs μ_m for time frame $m = 1, \cdots, n$:

  $$\mu_m | H, \eta_0 \sim \text{NRM}(\eta_0, M_0, H) \tag{5.11}$$

  where M_0 is the mass parameter for μ_m and H is the base distribution, η_0 is the set of hyperparameters of the corresponding NRM, *e.g.*, in NGG, $\eta_0 = \sigma$.

- Generating dependent NRMs μ'_m (from μ_m and μ'_{m-1}), for time frame $m > 1$:

  $$\mu'_m = T(S^q(\mu'_{m-1})) \oplus \mu_m \, . \tag{5.12}$$

  where the three dependency operators *superposition* ($\oplus$), *subsampling* ($S^q(\cdot)$) with acceptance rate q, and *point transition* ($T(\cdot)$) for NRMs have been defined above.

- Generating hierarchical NRM mixtures (μ_{mj}, θ_{mji}, x_{mji}) for time frame $m = 1, \cdots, n$, document $j = 1, \cdots, N_m$, word $i = 1, \cdots, W_{mj}$:

  $$\begin{aligned} &\mu_{mj} \sim \text{NRM}(\eta_m, M_m, \mu'_m), \\ &\theta_{mji} | \mu_{mj} \sim \mu_{mj}, \quad x_{mji} | \theta_{mji} \sim F(\cdot | \theta_{mji}) \end{aligned} \tag{5.13}$$

  where M_m is the total mass for μ_{mj}, $F(\cdot|\theta_{mji})$ denotes the cumulative density to generate data x_{mji} from atom θ_{mji} with the corresponding density function as $f(\cdot)$, which is essentially the multinomial distribution in topic modeling.

5.4.3 Reformulation of the model

Note it is found that directly dealing with the above model is challenging, thus the model is reformulated with the following theorem.

Theorem 5.11 (Equivalence Theorem). *Assume the subsampling rates $q(\cdot)$ are independent (constant)*[1] *for each point of the corresponding Poisson process, the following dependent random measures (5.14) and (5.15) are equivalent:*

- *Manipulate the normalized random measures:*

$$\mu'_m \sim T(S^q(\mu'_{m-1})) \oplus \mu_m, \qquad \textit{for } m > 1. \tag{5.14}$$

- *Manipulate the completely random measures:*

$$\begin{aligned} \tilde{\mu}'_m &\sim \tilde{T}(\tilde{S}^q(\tilde{\mu}'_{m-1})) \oplus \tilde{\mu}_m, \qquad \textit{for } m > 1. \\ \mu'_m &= \frac{\tilde{\mu}'_m}{\tilde{\mu}'_m(\Theta)}, \end{aligned} \tag{5.15}$$

Furthermore, both resulting NRMs μ'_m's correspond to:

$$\mu'_m = \sum_{j=1}^{m} \frac{\left(q^{m-j}\tilde{\mu}_j\right)(\Theta)}{\sum_{j'=1}^{m}\left(q^{m-j'}\tilde{\mu}_{j'}\right)(\Theta)} T^{m-j}(\mu_j), \qquad \textit{for } m > 1$$

where $q^{m-j}\tilde{\mu}$ is the random measure with Lévy measure $q^{m-j}(\theta)\nu(\mathrm{d}w, \mathrm{d}\theta)$, and $\nu(\mathrm{d}w, \mathrm{d}\theta)$ is the Lévy measure of $\tilde{\mu}$. $T^{m-j}(\mu)$ denotes point transition on μ for $(m-j)$ times .

Note that from the above theorem we can think of μ'_m as a linear combination of μ_j's, however, it does not necessary to say that the Lévy measure of μ'_m is a linear combination of μ_j's. Actually, their relationship is much more complicated due to the subsampling operation, as can be seen in the next two chapters. On the other hand, it provides us a reasonable approximation for posterior inference of such Markovian dependent models, where we first instantiate the atoms in μ_j's, and do the operations on these atoms to construct μ'_m. Such kind of approximation will be used in the dynamic topic model defined above for efficient posterior inference.

Specifically, Theorem 5.11 allows us to first take *superposition*, *subsampling*, and *point transition* on the completely random measures $\tilde{\mu}_g$'s and then do the normalization. Therefore, by using the theorem, the dynamic topic model in Figure 5.1(left) can be shown to be equivalent to the model in the right by expanding the recursive formula in (5.15).

As a result, the generating process of the reformulated model is:

[1] This assumption is to deal with the case when considering point transition, meaning we can drop this assumption if no point transition operation is considered.

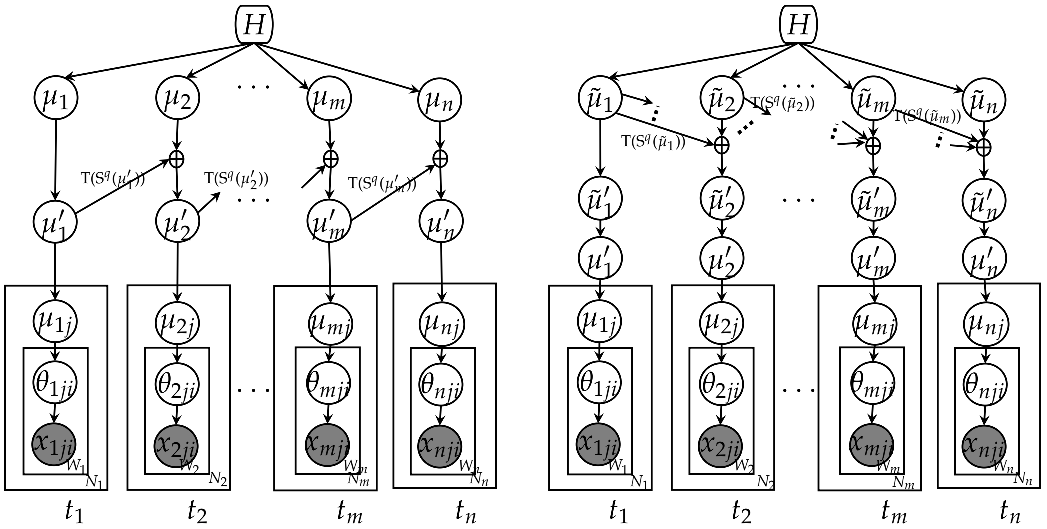


Figure 5.1: The time dependent topic model. The left plot corresponds to directly manipulating on normalized random measures (5.14), the right one corresponds to manipulating on unnormalized random measures (5.15). T: Point transition; S^q: Subsampling with acceptance rate q; $\oplus$: Superposition. Here $m = n - 1$ in the figures.

- Generating independent CRM's $\tilde{\mu}_m$ for time frame $m = 1, \cdots, n$, following (3.2).
- Generating μ'_m for time frame $m > 1$, following (5.15).
- Generating hierarchical NRM mixtures (μ_{mj}, θ_{mji}, x_{mji}) following (5.13).

The reason for this reformulation is because the inference on the model in Figure 5.1(left) appears to be complicated. In general, the posterior of an NRM introduce complex dependencies between jumps, thus sampling is difficult after taking the three dependency operators.

On the other hand, the model in Figure 5.1(right) is more amenable to computation because the NRMs and the three operators are decoupled. It allows us to first instantiate the dependent CRM's, then apply dependency operators on the corresponding atoms. As a result, the sampling procedure will be performed based on the model in Figure 5.1(right).

5.4.4 Sampling

To introduce the sampling method, the familiar Chinese restaurant metaphor (*e.g.* [Teh et al., 2006]) is used to explain key statistics. In this model customers for the variable μ_{mj} correspond to words in a document, restaurants to documents, and dishes to topics. In time frame m,

- x_{mji}: the customer i in the jth restaurant.
- s_{mji}: the index of dish that x_{mji} is eating.

- n_{mjk}: $n_{mjk} = \sum_i \delta_{s_{mji=k}}$, the number of customers in μ_{mj} eating dish k.
- t_{mjr}: the table r in the jth restaurant.
- ψ_{mjr}: the dish that the table t_{mjr} is serving.
- n'_{mk}: $n'_{mk} = \sum_j \sum_r \delta_{\psi_{mjr}=k}$, the number of customers[2] in μ'_m eating dish k.
- $\tilde{n}'_{mk}$: $\tilde{n}'_{mk} = n'_{mk}$, the number of customers in $\tilde{\mu}'_m$ eating dish k.
- $\tilde{n}_{mk}$: $\tilde{n}_{mk} = \sum_{m' \geq m} \tilde{n}'_{m'k}$, the number of customers in $\tilde{\mu}_m$ eating dish k.

The sampling is done by marginalizing out μ_{mj}'s. As it turns out, the remaining random variables that require sampling are s_{mji}, n'_{mk}, as well as

$$\tilde{\mu}_m = \sum_k w_{mk} \delta_{\theta_k}, \qquad \tilde{\mu}'_m = \sum_k w'_{mk} \delta_{\theta_k}$$

Note the t_{mjr} and ψ_{mjr} are not sampled as the n'_{mk} are directly sampled. Thus the sampler deals with the following latent statistics and variables: s_{mji}, n'_{mk}, w_{mk}, w'_{mk} and some auxiliary variables are sampled to support these.

Sampling w_{mk}. Given $\tilde{n}_{mk}$, the $\tilde{\mu}_m$'s are treated independently, thus the slice sampler introduced in [Griffin and Walker, 2011] is used to sample these jumps [3], with the posterior given in (3.43). Note that the mass parameters M_m's are also sampled conditioned on other variables, see Section 3.5.1.2 for the formula for a single NRM. The resulting $\{w_{mk}\}$ are those jumps that exceed a threshold defined in the slice sampler, thus the number of jumps is finite.

Sampling w'_{mk}. w'_{mk} is obtained by subsampling of $\{w_{m'k}\}_{m' \leq m}$[4]. By using a Bernoulli variable z_{mk},

$$w'_{mk} = \begin{cases} w_{m'k} & \text{if } z_{mk} = 1 \\ 0 & \text{if } z_{mk} = 0. \end{cases}$$

The posterior $p(z_{mk} = 1 | \tilde{\mu}_m, \{\tilde{n}'_{mk}\})$ is computed to decide whether to inherit this jump to $\tilde{\mu}'_m$ or not. The posterior $p(z_{mk} = 1 | \tilde{\mu}_m, \{\tilde{n}'_{mk}\})$ is given by the following corollary, please refer to the appendix for the proof.

Corollary 5.12 (Posterior acceptance rates in sampling w'_{mk}). *The posterior $p(z_{mk} = 1 | \tilde{\mu}_m, \{\tilde{n}'_{mk}\})$ is computed as:*

[2] The customers in μ'_m corresponds to the tables in μ_{mj}. For convenient, we also regard a CRM as a restaurant.

[3] Strictly speaking, an approximation is adopted here where we slice sample on the CRMs without considering the impact of the subsampling using techniques from [Griffin and Walker, 2011]. However, this results in slightly different posterior Lévy measure from $\tilde{\mu}_m$ to the true one. Detailed analysis of the true posterior Lévy measure with subsampling can be found in Chapter 7.

[4] Since all the atoms across $\{\tilde{\mu}_{m'}\}$ are unique, w'_{mk} is inherited from only one of $\{w_{m'k}\}$.

- *If* $\tilde{n}'_{mk} > 0$*, then* $p(z_{mk} = 1|\tilde{\mu}_m, \{\tilde{n}'_{mk}\}) = 1$.
- *Otherwise,*

$$p(z_{mk} = 1|\tilde{\mu}_m, \{\tilde{n}'_{mk}\}) = \frac{q^{m-m'}/J_m}{q^{m-m'}/J_m + (1 - q^{m-m'})/J_m^{-k}}, \tag{5.16}$$

where

$$J_m = \left(\sum_{m' \leq m} \sum_{k'} z_{m'k'} w_{m'k'} \right)^{\tilde{n}'_{m\cdot}}$$
$$J_m^{-k} = \left(\sum_{m' \leq m} \sum_{k' \neq k} z_{m'k'} w_{m'k'} \right)^{\tilde{n}'_{m\cdot}}$$
$$\tilde{n}'_{m\cdot} = \sum_{k'} \tilde{n}'_{mk'} \ .$$

In practice, the infinite sum J_m is hard to compute. There are two approximation strategies here: one is to rely on a truncated version; the other is to simply use the approximation of rescaling the jump sizes of the original CRM with a factor of q as:

$$w'_{mk} = q^{m-m'} w_{m'k}, \tag{5.17}$$

which are then used in the sampling. The intuition can be explained as follows: when $w_k \to 0$ (which is usually the case for the infinite many small jumps), $J_m^{-k} \to J_m$ as well, thus the posterior (5.16) approaches the prior: $p(z_{mk} = 1) = q^{m-m'}$. Now using the prior, and integrating out z_{mk}'s we get the rescaled formula (5.17). The implementation of the model adopts this approximation due to its simplicity.

After the sampling of $\{w'_{mk}\}$, they are further normalized, and the NRM μ'_m is obtained: $\mu'_m = \sum_k r_{mk} \delta_{\theta_k}$ where $r_{mk} = w'_{mk} / \sum_{k'} w'_{mk'}$

Sampling s_{mji}**,** n'_{mk}**.** This follows similar strategies as for the HNRM introduce in Chapter 4. The sampling method goes as follows:

- **Sampling** s_{mji}**:** A similar strategy called the *sampling by direct assignment algorithm* for the HDP [Teh et al., 2006] is used to sample s_{mji}, the conditional posterior of s_{mji} is:

  $$p(s_{mji} = k|\cdot) \propto (\omega_k + \omega_0 \tilde{M}_m r_{mk}) f(x_{mji}|\theta_k)$$

  where ω_0, ω_k and $\tilde{M}_m$ depend on the corresponding Lévy measure of μ_{mj} (see [Theorem 2 James et al., 2009]). When μ_{mj} is a DP, then $\omega_k \propto n_{mjk}$, $\omega_0 \propto 1$ and $\tilde{M}_m = \alpha$ known as the concentration parameter. When μ_{mj} is a NGG, $\omega_k \propto n_{mjk} - \sigma$, $\omega_0 \propto \sigma(1 + v_{mj})^{\sigma}$ and $\tilde{M}_m = M_m$ as in (4.9), where v_{mj} is the introduced auxiliary variables which can be sampled by an adaptive-rejection sampler using the posterior given in [Proposition 1 James et al., 2009].

dataset	vocab	docs	words	epochs
ICML	2k	765	44k	2007–2011
JMLR	2.4k	818	60k	12 vols
TPAMI	3k	1108	91k	2006–2011
NIPS	14k	2483	3.28M	1987-2003
Person	60k	8616	1.55M	08/96–08/97
Twitter$_1$	6k	3200	16k	14 months
Twitter$_2$	6k	3200	31k	16 months
Twitter$_3$	6k	3200	25k	29 months
BDT	8k	2649	234k	11/07–04/08

Table 5.1: Data statistics

- **Sampling n'_{mk}:** Using the similar strategy as in [Teh et al., 2006], n'_{mk}'s are sampled by simulating the (generalized) Chinese Restaurant Process, following the prediction rule (the probabilities of generating a new table or sitting on existing tables) of μ_{mk} in [Proposition 2 James et al., 2009].

Dealing with point transition The point transition operator applies on θ_k's, and depends on the specific operator used. Here a simple transition of random perturbation is used. Specifically, let θ_k be a Dirichlet distribution parameterized by the word counts, say $(\tilde{m}_{k1}, \tilde{m}_{k2}, \cdots, \tilde{m}_{kV})$ where V is the vocabulary size. Before sampling for time m, the counts $\{\tilde{m}_{kv}\}$ obtained from previous times are first randomly perturbed by a Gaussian noise, *e.g.*, $\tilde{m}'_{kv} \sim \mathrm{N}(\tilde{m}'_{kv}; \tilde{m}_{kv}, 1)$. The perturbed version $\{\tilde{m}'_{kv}\}$ is then used as initialized counts to sample θ_k for the current time.

5.5 Experiments

5.5.1 Datasets

The time dependent dynamic topic model is tested on 9 datasets, where stop-words and words appearing less than 5 times are removed. ICML, JMLR, TPAMI are crawled from their websites and the abstracts are parsed. The preprocessed NIPS dataset is from [Globerson et al., 2007]. The Person dataset is extracted from Reuters RCV1 using the query "person" under Lucene. The Twitter datasets are updates from three sports twitter accounts: `ESPN_FirstTake` (Twitter$_1$), `sportsguy33` (Twitter$_2$) and `SportsNation` (Twitter$_3$) obtained with the TweetStream API[5] to collect the last 3200 updates from each. The *Daily Kos* blogs (BDT) were pre-processed by Yano et al. [2009]. Statistics for the data sets are given in Table 5.1.

Illustration: Figure 5.2 gives an example of topic evolution in the Twitter$_2$ dataset. We can clearly see that the three popular sports in the USA, *i.e.*, basketball, football and baseball, evolve reasonably with time. For example, MLB starts in April each

[5] http://pypi.python.org/pypi/tweetstream.

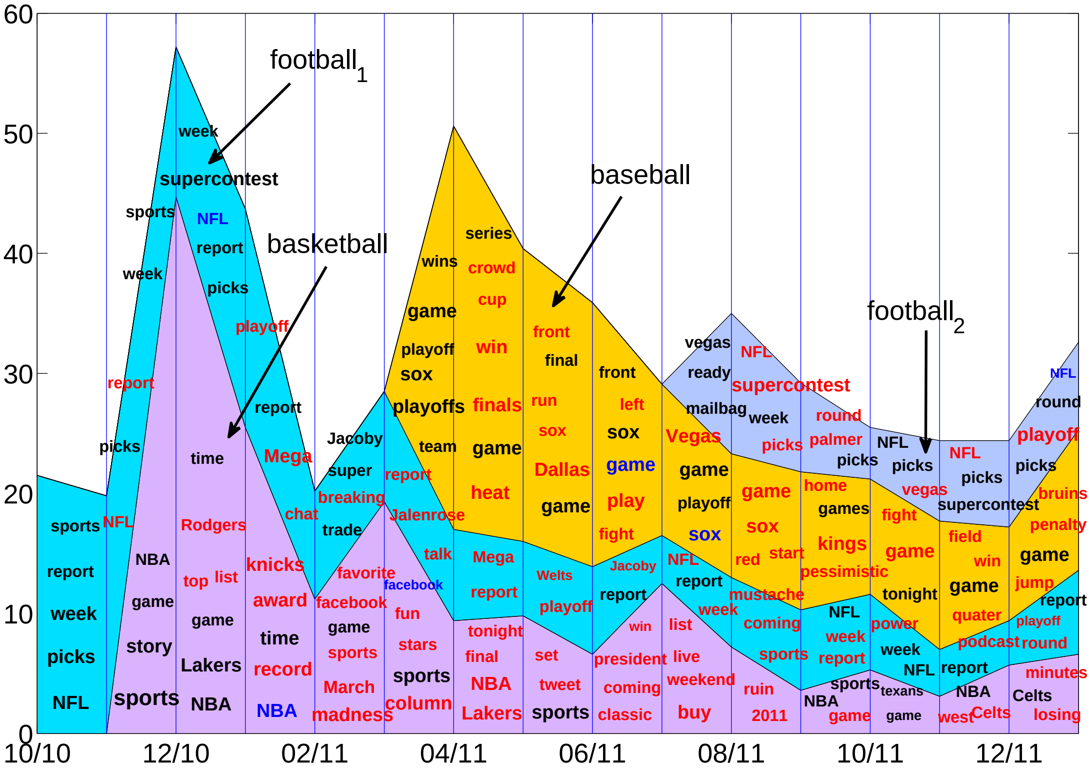


Figure 5.2: Topic evolution on Twitter. Words in red have increased, and blue decreased.

year, showing a peak in baseball topic, and then slowly evolves with decreasing topic proportions. Also, in August one football topic is born, indicating a new season begins. Figure 5.3 gives an example of the word probability change in a single topic for the JMLR.

5.5.2 Quantitative evaluations

Comparisons The model is first compared with two popular dynamic topic models where the author's own code is available for use: (1) the dynamic topic model by Blei and Lafferty [2006] and (2) the hierarchical Dirichlet process, where a three level HDP is used, with the middle level DP's representing the base topic distribution for the documents in a particular time. For fair comparison, similar to [Blei and Lafferty, 2006], the data in previous time are held out but their statistics are used to help the training of the current time data, this is implemented in the HDP code by Teh [2004]. Furthermore, the proposed model is tested without power-law, which is to use a DP instead of an NGG. The model is tested on the 9 datasets, for each dataset 80% are used for training and 20% are held out for testing. The hyperparameters for DHNGG is set to $\sigma = 0.2$ in this set of experiments with subsampling rate being 0.9, which is found to work well in practice. The topic-word distributions are symmetric Dirichlet with prior set to 0.3. Table 5.2 shows the test log-likelihoods for all these methods, which are calculated by first removing the test words from the topics and

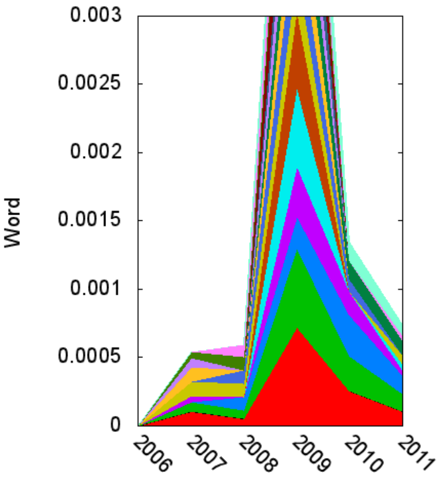


Figure 5.3: Topic evolution on JMLR. Shows a late developing topic on software, before during and after the start of MLOSS.org in 2008.

adding them back one by one and collecting the add-in probabilities as the test likelihood [Teh et al., 2006]. For all the methods 2000 burn in iterations are run, followed by 200 iterations for collecting posterior samples. The results are averages over these samples.

From Table 5.2 we see the proposed model *DHNGG* works best, with an improvement of 1%-3% in test log-likelihoods over the *HDP* model. In contrast the time dependent model *iDTM* of [Ahmed and Xing, 2010] only showed a 0.1% improvement over *HDP* on NIPS, implying the superiority of *DHNRM* over *iDTM*.

Hyperparameter sensitivity In NGG, there is the hyperparameters σ controlling the behavior of the power-law. This section studies the influences of this hyperparameter to the model. Specifically, σ is varied among $(0.1, 0.2, 0.3, 0.5, 0.7, 0.9)$, and the subsampling rate is fixed to 0.9 in this experiment. The models with these settings are run on all these datasets, the training likelihoods are shown in Figure 5.4. From these results $\sigma = 0.2$ is considered to be a good choice in practice.

Influence of the subsampling rate One of the distinct features of the model compared to other time dependent topic models is that the dependency comes partially from subsampling the previous time random measures, thus it is interesting to study the impact of subsampling rates to this model. In this experiment, σ is fixed to 0.2, and the subsampling rate q is varied among $(0.1, 0.2, 0.3, 0.5, 0.7, 0.9, 1.0)$. The results are shown in Figure 5.5. From Figure 5.5, it is interesting to see that on the academic datasets, *e.g.*, ICML,JMLR, the best results are achieved when q is approximately equal to 1: these datasets have higher correlations. While for the Twitter datasets, the best results are achieved when q is equal to $0.5 \sim 0.7$, indicating that people tend to discuss more variable topics in these datasets.

Table 5.2: Test log-likelihood on 9 datasets. *DHNGG*: dependent hierarchical normalized generalized Gamma processes, *DHDP*: dependent hierarchical Dirichlet processes, *HDP*: hierarchical Dirichlet processes, *DTM:* dynamic topic model (we set $K = \{10, 30, 50, 70\}$ and choose the best results).

Datasets	ICML	JMLR	TPAMI	NIPS	Person
DHNGG	**-5.3123e+04**	**-7.3318e+04**	**-1.1841e+05**	**-4.1866e+06**	**-2.4718e+06**
DHDP	-5.3366e+04	-7.3661e+04	-1.2006e+05	-4.4055e+06	-2.4763e+06
HDP	-5.4793e+04	-7.7442e+04	-1.2363e+05	-4.4122e+06	-2.6125e+06
DTM	-6.2982e+04	-8.7226e+04	-1.4021e+05	-5.1590e+06	-2.9023e+06
Datasets	Twitter$_1$	Twitter$_2$	Twitter$_3$	BDT	
DHNGG	**-1.0391e+05**	**-2.1777e+05**	**-1.5694e+05**	**-3.3909e+05**	
DHDP	-1.0711e+05	-2.2090e+05	-1.5847e+05	-3.4048e+05	
HDP	-1.0752e+05	-2.1903e+05	-1.6016e+05	-3.4833e+05	
DTM	-1.2130e+05	-2.6264e+05	-1.9929e+05	-3.9316e+05	

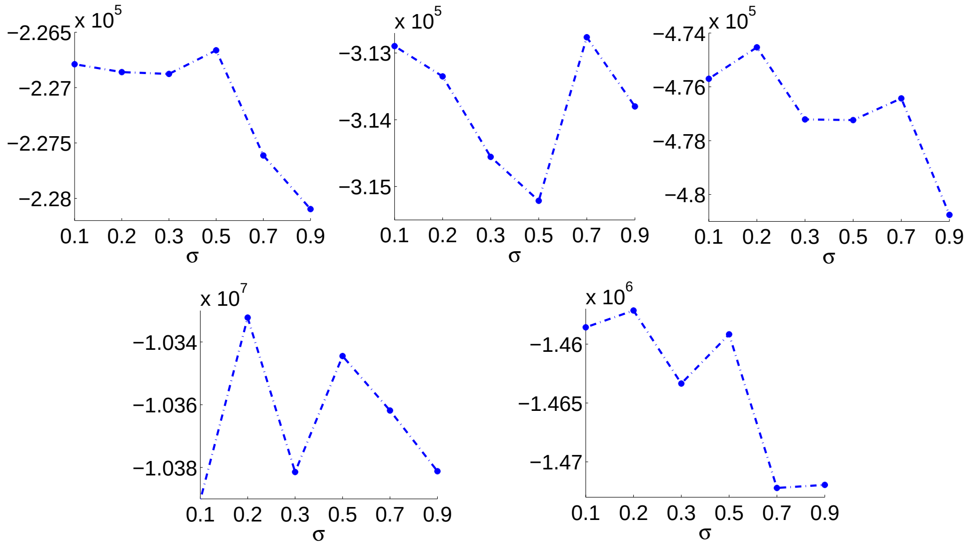


Figure 5.4: Training log-likelihoods influenced by hyperparameter σ. From left to right (top-down) are the results on ICML, JMLR, TPAMI, Person and BDT.

5.6 Conclusion

This chapter proposes dependent hierarchical normalized random measures. Specifically, the three dependency operations for the Poisson process are extended to hierarchical normalized random measures, and dependencies with these operators are also analyzed in detail. The dependency model is then applied to dynamic topic modeling. Experimental results on different kinds of datasets demonstrate the superior performance of the model over existing models such as DTM, HDP and iDTM. One drawback, as mentioned above, lies on the accuracy of the posterior inference, where several approximations have been made to design efficient sampling algo-

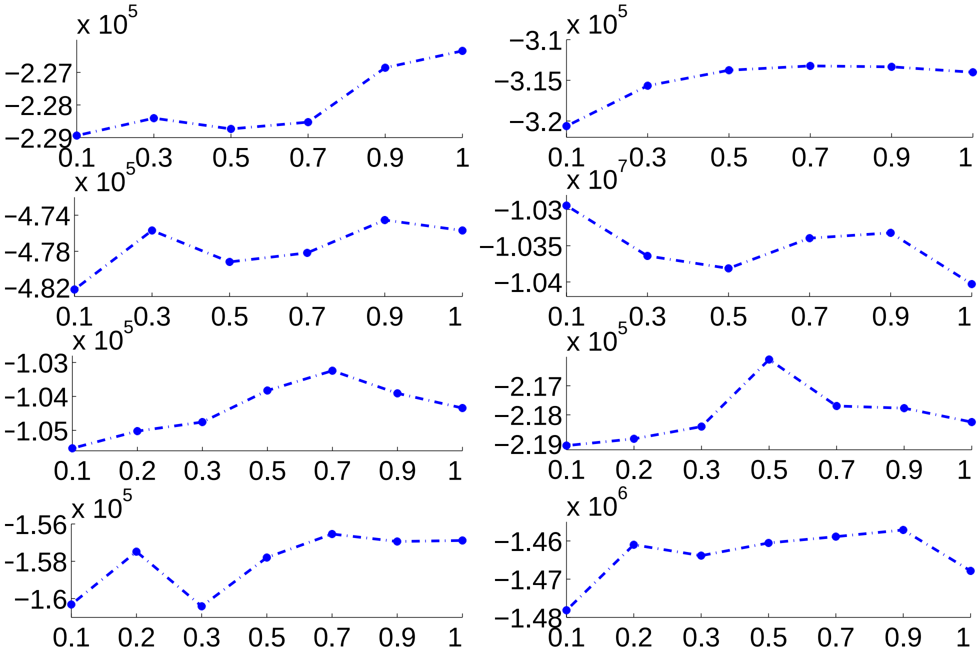


Figure 5.5: Training log-likelihoods influenced by the subsampling rate $q(\cdot)$. The x-axes represent q, the y-axes represent training log-likelihoods. From top-down, left to right are the results on ICML, JMLR, TPAMI, Person, Twitter$_1$, Twitter$_2$, Twitter$_3$ and BDT datasets, respectively.

rithm. How to design exact posterior inference for such kinds of Markovian and hierarchical structure is interesting future work, which could probably be solved by using the techniques to be introduced in Chapter 7.

5.7 Proofs

Proof of Lemma 5.2. This uses a similar technique to that of Theorem 1 in [Griffin et al., 2013]. Using the identity $1/b = \int_0^\infty e^{-v\,b}\mathrm{d}v$ we get

$$\begin{aligned}\mathbb{E}\left[\mu(B)\right] &= \mathbb{E}\left[\frac{\tilde{\mu}(B)}{\tilde{\mu}(\Theta)}\right] \\ &= \int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)\exp\left\{-v\tilde{\mu}(B)\right\}\right]\mathbb{E}\left[\exp\left\{-v\tilde{\mu}(\Theta\setminus B)\right\}\right]\mathrm{d}v\,. \qquad (5.18)\end{aligned}$$

According to the Lévy-Khintchine representation of $\tilde{\mu}$ and definition (3.6), we have

$$\begin{aligned}&\mathbb{E}\left[\exp\left\{-v\tilde{\mu}(B)\right\}\right] = \exp\left\{-P(B)M\tilde{\psi}_\eta(v)\right\} \qquad (5.19)\\ &\mathbb{E}\left[\tilde{\mu}(B)\exp\left\{-v\tilde{\mu}(B)\right\}\right] = -\mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}v}\exp\left\{-v\tilde{\mu}(B)\right\}\right]\\ =\ & H(B)M\tilde{\psi}'_\eta(v)\exp\left\{-H(B)M\tilde{\psi}_\eta(v)\right\} \qquad (5.20)\\ &\mathbb{E}\left[\tilde{\mu}(B)^2\exp\left\{-v\tilde{\mu}(B)\right\}\right] = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}v^2}\exp\left\{-v\tilde{\mu}(B)\right\}\right]\\ =\ & \left(H(B)^2M^2\left(\tilde{\psi}'_\eta(v)\right)^2 - H(B)M\tilde{\psi}''_\eta(v)\right)\exp\left\{-H(B)M\tilde{\psi}_\eta(v)\right\} \qquad (5.21)\end{aligned}$$

Substituting (5.19) and (5.20) into (5.18) and using the fact in (3.7), after simplifying we have

$$\mathbb{E}\left[\mu(B)\right] = H(B).$$

Since $\mathrm{Var}\left(\mu(B)\right) = \mathbb{E}\left[\mu(B)^2\right] - \left(\mathbb{E}\left[\mu(B)\right]\right)^2$, and the last term is equal to $\left(H(B)\right)^2$, we now deal with the first term.

$$\begin{aligned}\mathbb{E}\left[\mu(B)^2\right] &= \mathbb{E}\left[\frac{\tilde{\mu}(B)^2}{\tilde{\mu}(\Theta)^2}\right]\\ &= \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)^2\times\exp\left\{-v_1\tilde{\mu}(\Theta)-v_2\tilde{\mu}(\Theta)\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \qquad (5.22)\\ &= \int_0^\infty\int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)^2\exp\left\{-(v_1+v_2)\tilde{\mu}(B)\right\}\right]\mathbb{E}\left[\exp\left\{-(v_1+v_2)\tilde{\mu}(\Theta\setminus B)\right\}\right]\mathrm{d}v_1\mathrm{d}v_2\end{aligned}$$

Substituting (5.19)(5.21) into (5.22) we have

$$\begin{aligned}(5.22) \ =\ & \int_0^\infty\int_0^\infty\left[H(B)^2M^2\left(\tilde{\psi}'_\eta(v_1+v_2)\right)^2 - H(B)M\tilde{\psi}''_\eta(v_1+v_2)\right]\\ & \exp\left\{-M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2\,. \qquad (5.23)\end{aligned}$$

Furthermore, let $v = v_1+v_2, B = \Theta$ in (5.21), after integrating out v_1, v_2 in $[0,\infty]$, we have

$$\begin{aligned}&\int_0^\infty\int_0^\infty M^2\left(\tilde{\psi}'_\eta(v_1+v_2)\right)^2\exp\left\{-M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2 \\ =&1+\int_0^\infty\int_0^\infty M\tilde{\psi}''_\eta(v_1+v_2)\exp\left\{-M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2\end{aligned} \tag{5.24}$$

Substitute (5.24) into (5.23) and simplify we get

$$\begin{aligned}&\mathrm{Var}(\mu(B))= \\ &H(B)(1-H(B))M\int_0^\infty\int_0^\infty -\tilde{\psi}''_\eta(v_1+v_2)\exp\left\{-M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2\end{aligned} \tag{5.25}$$

Now use a change of variables, let $v'_1=v_1, v'_2=v_1+v_2$ and simplify we get the result of (5.4). □

Proof of Theorem 5.4. Let $\tilde{M}_k\triangleq\tilde{\mu}_k(\Theta)$, from the definition we have

$$\begin{aligned}\mathrm{Cov}\left(\mu_k(B),\mu(B)\right)=&\sum_{i=1}^n\mathrm{Cov}\left(\frac{\tilde{M}_i}{\sum_j\tilde{M}_j}\mu_i(B),\mu_k(B)\right) \\ =&\ \mathrm{Cov}\left(\frac{\tilde{M}_k}{\sum_j\tilde{M}_j}\mu_k(B),\mu_k(B)\right)+\sum_{i\neq k}\mathrm{Cov}\left(\frac{\tilde{M}_i}{\sum_j\tilde{M}_j}\mu_i(B),\mu_k(B)\right) \\ =&\ \mathbb{E}\left[\frac{\tilde{\mu}_k(B)^2}{\left(\sum_j\tilde{\mu}_j(\Theta)\right)\tilde{\mu}_k(\Theta)}\right]-\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\sum_j\tilde{\mu}_j(\Theta)}\right]\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\tilde{\mu}_k(\Theta)}\right] \\ +&\ \sum_{i\neq k}\left\{\mathbb{E}\left[\frac{\tilde{\mu}_i(B)\tilde{\mu}_k(B)}{\left(\sum_j\tilde{\mu}_j(\Theta)\right)\tilde{\mu}_k(\Theta)}\right]-\mathbb{E}\left[\frac{\tilde{\mu}_i(B)}{\sum_j\tilde{\mu}_j(\Theta)}\right]\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\tilde{\mu}_k(\Theta)}\right]\right\}\end{aligned} \tag{5.26}$$

Note that for the Dirichlet process, the last $n-1$ terms of (5.26) vanish because μ_i's are independent from their total mass $\tilde{M}_i$'s, but this is not the case for general NRMs. Now we calculate these term by term.

For the first term, we have

$$\begin{aligned}&\mathbb{E}\left[\frac{\tilde{\mu}_k(B)^2}{\left(\sum_j\tilde{\mu}_j(\Theta)\right)\tilde{\mu}_k(\Theta)}\right] \\ =&\ \int_0^\infty\int_0^\infty\mathbb{E}\left[\tilde{\mu}_k(B)^2\exp\left\{-v_1(\sum_j\tilde{\mu}_j)(\Theta)-v_2\tilde{\mu}_k(\Theta)\right\}\right]\mathrm{d}v_1\mathrm{d}v_2 \\ =&\ \int_0^\infty\int_0^\infty\mathbb{E}\left[\tilde{\mu}_k(B)^2\exp\left\{-(v_1+v_2)\tilde{\mu}_k(B)\right\}\right]\mathbb{E}\left[\exp\left\{-(v_1+v_2)\tilde{\mu}_k(\Theta\setminus B)\right\}\right] \\ &\mathbb{E}\left[\exp\left\{-v_1(\sum_{j\neq k}\tilde{\mu}_j(\Theta))\right\}\right]\mathrm{d}v_1\mathrm{d}v_2\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty \int_0^{v_2} \left(H(B)^2 M_k^2 \tilde{\psi}'_\eta(v_1)^2 - H(B) M_k \tilde{\psi}''_\eta(v_1) \right) \exp\left\{ -M_k \tilde{\psi}_\eta(v_1) \right\} \\
&\quad \exp\left\{ -(\sum_{j \neq k} M_j) \tilde{\psi}_\eta(v_2) \right\} \mathrm{d}v_1 \mathrm{d}v_2 \\
&= H(B) M_k \int_0^\infty \gamma(M_k, H(B), v) \exp\left\{ -(\sum_{j \neq k} M_j) \tilde{\psi}_\eta(v) \right\} \mathrm{d}v \qquad (5.27)
\end{aligned}
$$

For the second term, we have

$$
\begin{aligned}
\mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\sum_j \tilde{\mu}_j(\Theta)} \right] \mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\tilde{\mu}_k(\Theta)} \right] &= H(B) \int_0^\infty \mathbb{E}\left[\tilde{\mu}_k(B) \exp\left\{ -v \sum_j \tilde{\mu}_j(\Theta) \right\} \right] \mathrm{d}v \\
&= H(B)^2 M_k \int_0^\infty \tilde{\psi}'_\eta(v) \exp\left\{ -(\sum_j M_j) \tilde{\psi}_\eta(v) \right\} \mathrm{d}v \\
&= \frac{H(B)^2 M_k \exp\left\{ -\left(\sum_j M_j \right) \tilde{\psi}_\eta(0) \right\}}{\sum_j M_j} \\
&= \frac{H(B)^2 M_k}{\sum_j M_j} \qquad (5.28)
\end{aligned}
$$

For the third term, similarly

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\tilde{\mu}_i(B) \tilde{\mu}_k(B)}{\left(\sum_j \tilde{\mu}_j(\Theta) \right) \tilde{\mu}_k(\Theta)} \right] \\
&= \int_0^\infty \int_0^\infty \mathbb{E}\left[\tilde{\mu}_i(B) \tilde{\mu}_k(B) \exp\left\{ -v_1 (\sum_j \tilde{\mu}_j)(\Theta) - v_2 \tilde{\mu}_k(\Theta) \right\} \right] \mathrm{d}v_1 \mathrm{d}v_2 \\
&= \int_0^\infty \int_0^\infty \mathbb{E}\left[\tilde{\mu}_k(B) \exp\left\{ -(v_1 + v_2) \tilde{\mu}_k(B) \right\} \right] \mathbb{E}\left[\exp\left\{ -(v_1 + v_2) \tilde{\mu}_k(\Theta \setminus B) \right\} \right] \\
&\quad \mathbb{E}\left[\tilde{\mu}_i(B) \exp\left\{ -v_1 \tilde{\mu}_i(B) \right\} \right] \mathbb{E}\left[\exp\left\{ -v_1 \tilde{\mu}_i(\Theta \setminus B) \right\} \right] \\
&\quad \mathbb{E}\left[\exp\left\{ -v_1 (\sum_{j \neq \{i,k\}} \tilde{\mu}_j(\Theta)) \right\} \right] \mathrm{d}v_1 \mathrm{d}v_2 \\
&= \int_0^\infty \int_0^\infty H(B) M_k \tilde{\psi}'_\eta(v_1 + v_2) \exp\left\{ -M_k \tilde{\psi}_\eta(v_1 + v_2) \right\} \\
&\quad H(B) M_i \tilde{\psi}'_\eta(v_1) \exp\left\{ -M_i \tilde{\psi}_\eta(v_1) \right\} \\
&\quad \exp\left\{ -(\sum_{j \neq \{i,k\}} M_j) \tilde{\psi}_\eta(v_1) \right\} \mathrm{d}v_1 \mathrm{d}v_2 \\
&= H(B)^2 M_i M_k \int_0^\infty \tilde{\psi}'_\eta(v_1) \exp\left\{ -(\sum_{j \neq k} M_j) \tilde{\psi}_\eta(v_1) \right\}
\end{aligned}
$$

$$\begin{aligned}
&\int_0^{v_1} \tilde{\psi}'_\eta(v_2) \exp\left\{-M_k \tilde{\psi}_\eta(v_2)\right\} \mathrm{d}v_2 \mathrm{d}v_1 \\
= \quad & H(B)^2 M_i \left(\frac{1}{\sum_{j\neq k} M_j} - \frac{1}{\sum_j M_j} \right) \exp\left\{ -(\sum_j M_j) \tilde{\psi}_\eta(0) \right\} \\
= \quad & H(B)^2 M_i \left(\frac{1}{\sum_{j\neq k} M_j} - \frac{1}{\sum_j M_j} \right) \qquad (5.29)
\end{aligned}$$

The fourth term is similar to the second term, and is equal to

$$\begin{aligned}
\mathbb{E}\left[\frac{\tilde{\mu}_i(B)}{\sum_j \tilde{\mu}_j(\Theta)}\right] \mathbb{E}\left[\frac{\tilde{\mu}_k(B)}{\tilde{\mu}_k(\Theta)}\right] &= \frac{H(B)^2 M_i \exp\left\{-\left(\sum_j M_j\right) \tilde{\psi}_\eta(0)\right\}}{\sum_j M_j} \\
&= \frac{H(B)^2 M_i}{\sum_j M_j} \qquad (5.30)
\end{aligned}$$

The result follows. □

Proof of Theorem 5.5. By subsampling, we obtain two independent NRMs μ^q and μ_0^q, corresponding to those points selected and those rejected by the independent Bernoulli trials, respectively.

We denote the total mass of the corresponding unnormalized μ^q as M_q, and M_q^0 for μ_0^q. From the definition of subsampling, we have

$$M_q := (q\tilde{\mu})(\Theta) = \int_\Theta q(x)\tilde{\mu}(x)\mathrm{d}x,$$

$$M_q^0 = M - M_q.$$

Furthermore, notice that the original NRM μ is the superposition of μ^q and μ_0^q. Thus according to Theorem 5.4, the covariance between μ and μ^q is

$$H(B)M_q \int_0^\infty \gamma(M_q, H(B), v) \exp\left\{-(M - M_q)\tilde{\psi}_\eta(v)\right\} \mathrm{d}v + H(B)^2 \left(\frac{2M_q - M}{M}\right),$$

□

Proof of Theorem 5.6. Note that $\tilde{\mu}$ and $\tilde{\mu}'$ are not independent, thus they cannot be separated when taking the expectation. Now let A and B are defined as in the theorem, then:

$$\begin{aligned}
&\mathbb{E}\left[\mu(B)\left((T\mu)(B)\right)\right] = \mathbb{E}\left[\frac{\tilde{\mu}(B)}{\tilde{\mu}(\Theta)} \frac{\tilde{\mu}'(B)}{\tilde{\mu}'(\Theta)}\right] = \mathbb{E}\left[\frac{\tilde{\mu}(B)}{\tilde{\mu}(\Theta)} \frac{\tilde{\mu}(A)}{\tilde{\mu}(\Theta)}\right] \\
= \quad & \int_0^\infty \int_0^\infty \mathbb{E}\left[\tilde{\mu}(B)\tilde{\mu}(A) \times \exp\left\{-(v_1 + v_2)\tilde{\mu}(\Theta)\right\}\right] \mathrm{d}v_1 \mathrm{d}v_2 \\
= \quad & \int_0^\infty \int_0^\infty \mathbb{E}\left[\tilde{\mu}(B) \exp\left\{-(v_1 + v_2)\tilde{\mu}(B)\right\}\right] \\
& \qquad \mathbb{E}\left[\tilde{\mu}(A) \exp\left\{-(v_1 + v_2)\tilde{\mu}(A)\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
& \mathbb{E}\left[\exp\left\{-(v_1+v_2)\tilde{\mu}(\Theta/\{A\cup B\})\right\}\right] \mathrm{d}v_1\mathrm{d}v_2 \\
= & \int_0^\infty\int_0^\infty H(B)M\tilde{\psi}_\eta'(v_1+v_2)\exp\left\{-H(B)M\tilde{\psi}_\eta(v_1+v_2)\right\} \\
& H(A)M\tilde{\psi}_\eta'(v_1+v_2)\exp\left\{-H(A)M\tilde{\psi}_\eta(v_1+v_2)\right\} \\
& H(\Theta/\{A\cup B\})M\tilde{\psi}_\eta'(v_1+v_2)\exp\left\{-H(\Theta/\{A\cup B\})M\tilde{\psi}_\eta(v_1+v_2)\right\}\mathrm{d}v_1\mathrm{d}v_2 \\
= & H(A)H(B)M^2\int_0^\infty\int_0^{v_1}\tilde{\psi}_\eta'(v_2)^2\exp\left\{-M\tilde{\psi}_\eta(v_2)\right\}\mathrm{d}v_2\mathrm{d}v_1
\end{aligned}
$$

Then the covariance is:

$$
\begin{aligned}
& \mathrm{Cov}\left(\mu(B),(T\mu)(B)\right) \\
= & \mathbb{E}\left[\mu(B)\left((T\mu)(B)\right)\right]-\mathbb{E}\left[\mu(B)\right]\mathbb{E}\left[(T\mu)(B)\right] \\
= & H(A)H(B) \\
& \left(M^2\int_0^\infty\int_0^{v_1}\tilde{\psi}_\eta'(v_2)^2\exp\left\{-M\tilde{\psi}_\eta(v_2)\right\}\mathrm{d}v_2\mathrm{d}v_1-1\right) \qquad (5.31)
\end{aligned}
$$

□

Proof of Lemma 5.7. From the existing of Poisson processes, each Lévy measure $\nu_i(\mathrm{d}w,\mathrm{d}\theta)$ corresponds to a Poisson random measure $N_i(\mathrm{d}w,\mathrm{d}\theta)$ with

$$\mathbb{E}\left[N_i(\mathrm{d}t,\mathrm{d}x)\right]=\nu_i(\mathrm{d}t,\mathrm{d}x).$$

Also we have $\forall i$,

$$\tilde{\mu}_i(\mathrm{d}\theta)=\int_0^\infty wN_i(\mathrm{d}w,\mathrm{d}\theta).$$

Thus from (5.9) we have

$$\tilde{\mu}(\mathrm{d}\theta)=\int_0^\infty w\left(\sum_{i=1}^n N_i(\mathrm{d}w,\mathrm{d}\theta)\right)=\int_0^\infty wN(\mathrm{d}w,\mathrm{d}\theta),$$

where $N(\cdot)=\sum_{i=1}^n N_i(\cdot)$ is again a Poisson random measure. Thus the Lévy intensity for $\tilde{\mu}(\cdot)$ is

$$\nu(\mathrm{d}w,\mathrm{d}\theta) \quad = \quad \sum_{i=1}^n \nu_i(\mathrm{d}w,\mathrm{d}\theta). \qquad (5.32)$$

Because Theorem 1 in [James et al., 2009] applies for any CRMs with Lévy measure $\nu(\mathrm{d}w,\mathrm{d}\theta)$, thus conclusion 2 and 3 in Lemma are proved.

Finally, by substituting the Lévy measure (5.32) into formula (3.18) in Chapter 3 and simplifying, we can get the posterior of u as shown in Lemma 5.7. □

Proof of Lemma 5.8. The case for point transition and superposition are developed similarly to the case for subsampling, so we only consider the later here.

The case for subsampling follows by merging the impact of the subsampling operation with the sampling step in Lemma 3.1. Suppose the Lévy measure is in the form $M\rho(\mathrm{d}w|\theta)H(\mathrm{d}\theta)$. The infinitesimal rate at data point θ_i when sampling the jump is now $q(\theta_i)M\rho(\mathrm{d}w|\theta)$. Thus the Lévy measure for the subsampled measure must be $M\rho(\mathrm{d}w|\theta)q(\theta)H(\mathrm{d}\theta)$.

This argument can also be seen from the detailed derivation below. First note that $S^q(\tilde{\mu})$ is equivalent to

$$S^q(\tilde{\mu}) = \int_{R^+ \times \Theta} z(\theta) s N(\mathrm{d}w, \mathrm{d}\theta) \ , \tag{5.33}$$

where $z(\theta)$ is a Bernoulli random variable with parameter $q(\theta)$. Let $B \in \Theta$, we divide B into n non-overlap patches and use A_{nm} to denote the m-th patch of them. So we have

$$\begin{aligned}
\mathbb{E}_{N(\cdot),z}\left[e^{-uS^q(\tilde{\mu})(B)}\right] &\overset{n\to\infty}{=} \mathbb{E}_{N(\cdot),z}\left[e^{-\sum_{A_{nm}\in B} uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}\right] \\
&= \mathbb{E}_{N(\cdot),z}\left[\prod_{A_{nm}\in B} e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}\right] \\
&= \prod_{A_{nm}\in B} \mathbb{E}_{N(\cdot),z}\left[e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}\right] \\
&= e^{\sum_{A_{nm}\in B} \log\left\{\mathbb{E}_{N(\cdot),z}\left[e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}-1\right]+1\right\}} \\
&\overset{(a)}{=} e^{\sum_{A_{nm}\in B} \mathbb{E}_{N(\cdot),z}\left[e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})}-1\right]} \\
&\overset{(b)}{=} e^{q\sum_{A_{nm}\in B} \mathbb{E}_{N(\cdot)}\left[e^{-us_{nm}N(A_{nm},s_{nm})}-1\right]} \\
&\overset{n\to\infty}{=} e^{-\int_{R^+\times B}(1-e^{-us})(q\nu(\mathrm{d}w,\mathrm{d}\theta))}
\end{aligned} \tag{5.34}$$

Here (a) above follows because $\mathbb{E}_{N(\cdot)}\left[\left(e^{-uz(A_{nm})s_{nm}N(A_{nm},s_{nm})} - 1\right)\right]$ is infinitesimal thus $\log(1+x) \overset{x\to 0}{\sim} x$ applies. (b) is obtained by integrating out $z(A_{nm})$ with Bernoulli distribution. Thus it can be seen from (5.34) that $S^q(\tilde{\mu})$ has the Lévy measure of $q(\theta)\nu(\mathrm{d}w, \mathrm{d}\theta)$. □

Proof of Theorem 5.11. We show that starting from (5.15) and (5.14), we can both end up the random measures defined in (5.16).

First, for the operations in (5.15), adapting from Theorem 2.17 of [Çinlar, 2010], a Poisson random measure with mean measure ν on the space $\mathbb{R}^+ \times \Theta$ has the form

$$N = \sum_{n=1}^{\infty} \sum_{i<K_n} \delta_{(w,\theta)}, \tag{5.35}$$

where K_n is a Poisson distributed random variable with mean ν, and $(w \in \mathbb{R}^+, \theta \in \Theta)$ are points in the corresponding Poisson processes. Then a realization of N composes of points in a Poisson process Π_1, and the corresponding Poisson random measure can be written as $N_1 = \sum_{(w,\theta)\in\Pi_1} \delta_{(w,\theta)}$.

Now consider doing a subsampling S^q and a point transition T on Π_1, by the definitions and (5.35) we get a new random measure

$$\begin{aligned}\tilde{N} &= T(S_q(N_1)) = T(S_q(\textstyle\sum \delta_{(w,\theta)}))\\ &\overset{(*)}{=} \textstyle\sum z(q(T(\theta)))\delta_{(w,T(\theta))} \overset{(**)}{=} \textstyle\sum z(q(\theta))\delta_{(w,T(\theta))},\end{aligned} \tag{5.36}$$

where $z(q(\cdot))$ means a Bernoulli random variable with acceptance rate $q(\cdot)$, $(*)$ follows from definitions, $(**)$ follows from the assumption of constant subsampling rate.

It is easy to show by induction that by subsampling and point transitioning i times of the Poisson process Π_1, we get a random measure as

$$\tilde{N}' = \textstyle\sum z(q^i(\theta))\delta_{(w,T^i(\theta))}. \tag{5.37}$$

By the definition, when superpositioning the this Poisson process $T^i(S_i^q(\Pi_1))$ with another Poisson process Π_2 with mean measure ν_2, we get another random measure as

$$N'' = \sum_{(w,\theta)\in\Pi_1} z(q^i(\theta))\delta_{(w,T^i(\theta))} + \sum_{(w,\theta)\in\Pi_2} \delta_{(w,T(\theta))}. \tag{5.38}$$

This Poisson random measure is then used to construct a completely random measure $\tilde{\mu}$ using (3.2) as:

$$\begin{aligned}\tilde{\mu}(A) &= \int_{\mathbb{R}^+\times\Theta} wN''(\mathrm{d}w,\mathrm{d}\theta)\\ &= \sum_{(w,\theta)\in\Pi_1} z(q^i(\theta))s\delta_{(w,T^i(\theta))} + \sum_{(w,\theta)\in\Pi_2} w\delta_{(w,\theta)}.\end{aligned} \tag{5.39}$$

By marginalize over r's and normalizing this random measure, we get

$$\begin{aligned}\mu(A) &= \frac{\tilde{\mu}(A)}{\tilde{\mu}(\Theta)}\\ &= \frac{M_1'}{M_1 + M_2'}\frac{\sum_{(w,\theta)\in\Pi_1\cap A} s\delta_{(w,T^i(\theta))}}{\sum_{(w,\theta)\in\Pi_1\cap\Theta} w\delta_{(w,T^i(\theta))}}\\ &+ \frac{M_2'}{M_1' + M_2'}\frac{\sum_{(w,\theta)\in\Pi_2\cap A} w\delta_{(w,T^i(\theta))}}{\sum_{(w,\theta)\in\Pi_2\cap\Theta} w\delta_{(w,T^i(\theta))}}\\ &= \frac{M_1'}{M_1' + M_2'}(T^i\mu_1)(A) + \frac{M_2'}{M_1' + M_2'}(T^i\mu_2)(A),\end{aligned} \tag{5.40}$$

where by apply Lemma 5.8 we conclude that $M_1' = \left(q^i\tilde{\mu}_1\right)(\Theta)$ is the total mass of the random measure with Lévy measure $q^j(\mathrm{d}\theta)\nu(\mathrm{d}w,\mathrm{d}\theta)$ and $M_2' = \tilde{\mu}_2(\Theta)$. We use the fact that $(T^k\tilde{\mu}_i)(\Theta) = \tilde{\mu}_i(\Theta)$ in the derivation of (5.40), because the point transition operation only moves the points (w,θ) of the Poisson process to other locations $(w,\theta+\mathrm{d}\theta)$, thus does not affect the total mass of the corresponding random measure.

This means by superposition after subsampling, the mass of the normalized random measure decays exponentially fast with respect to the distance i. Based on Eq. (5.40), when taking i from 1 to n, and taking superposition for all these random measure induced, the resulting normalized random measure is:

$$\mu_n' = \sum_{i=1}^{n} \frac{\left(q^{n-i}\tilde{\mu}_i\right)(\Theta)}{\sum_{j=1}^{n}\left(q^{n-j}\tilde{\mu}_j\right)(\Theta)} T^{n-i}(\mu_i). \tag{5.41}$$

Next, for the operations in (5.14), from the definition we have

$$\begin{aligned} \mu_2' &= T\left(S^q\left(\mu_1'\right)\right) \oplus \mu_2 \\ &= \frac{(q\tilde{\mu}_1)(\Theta)}{(q\tilde{\mu}_1 + \tilde{\mu}_2)(\Theta)} T(\mu_1) + \frac{(\tilde{\mu}_1)(\Theta)}{(q\tilde{\mu}_1 + \tilde{\mu}_2)(\Theta)} \mu_2 \end{aligned} \tag{5.42}$$

Now μ_2' has a total mass of $(q\tilde{\mu}_1 + \tilde{\mu}_2)(\Theta)$, by induction on i, we get the formula in (5.16) for $i = n$.

This completes the proof. □

Finally, to prove Corollary 5.12, we first prove the following lemma:

Lemma 5.13 (Posterior acceptance rates for subsampling)**.** *Let $\tilde{\mu}' = \sum_k w_k \theta_k$ be a completely random measure on Θ, $\tilde{\mu} = S^q(\tilde{\mu}') := \sum_k z_k w_k \delta_k$ be its subsampling version, where z_k's are independent Bernoulli random variables with acceptance rate q. Further define $\mu = \frac{\tilde{\mu}}{\tilde{\mu}(\Theta)}$. Given $n = \sum_k n_k$ observed data in μ, the posterior of z_k is:*

$$p(z_k = 1|\tilde{\mu}, n) = \begin{cases} 1 & \textit{if } n_k > 0, \\ \frac{q/J}{q/J + (1-q)/J^{-k}} & \textit{if } n_k = 0. \end{cases} \tag{5.43}$$

where $J = \left(\sum_{k'} z_{k'} w_{k'}\right)^n$, $J^{-k} = \left(\sum_{k' \neq k} z_{k'} w_{k'}\right)^n$.

Proof. Given the current data configuration $\{n_k, k = 1, 2, \cdots\}$, for a particular k,

- If $n_k > 0$, this means this jump w_k must exist in μ, otherwise it is impossible to have $n_k > 0$, thus $p(z_k = 1|\tilde{\mu}, n) = 1$.
- Otherwise,since $\mu = \sum_{k: z_k = 1} \frac{w_k \delta_k}{\sum_{k'} z_{k'} w_{k'}}$, we have the likelihood as:

  $$\prod_{k'': n_{k''} > 0} \frac{w_{k''}^{n_{k''}}}{\left(\sum_{k' \neq k} z_{k'} w_{k'} + z_k J_k\right)^{n_k}} = \frac{\prod_{k'': n_{k''} > 0} w_{k''}^{n_{k''}}}{\left(\sum_{k' \neq k} z_{k'} w_{k'} + z_k w_k\right)^n}.$$

  Furthermore, we know that the prior for z_k is $p(z_k = 1) = q$, thus the posterior is:

  $$p(z_k = 1|\tilde{\mu}, n) \propto \frac{q}{\left(\sum_{k' \neq k} z_{k'} w_{k'} + w_k\right)^n}.$$

  $$p(z_k = 0|\tilde{\mu}, n) \propto \frac{1-q}{\left(\sum_{k' \neq k} z_{k'} w_{k'}\right)^n}.$$

After normalizing, we get the posterior for the case $n_k = 0$ in (5.43).

□

Now the proof for Corollary 5.12 is an direct extension of the above lemma:

Proof of Corollary 5.12. Note that w'_{mk} is obtained by subsampling of $\{w_{m'k}, m' \leq m\}$, the number of data points in $\tilde{\mu}'_m$ is denoted as $\tilde{n}'_{m\cdot} = \sum_{k'} \tilde{n}'_{mk'}$.

Following the same arguments as in the proof of Theorem 5.13, when $\tilde{n}'_{mk} > 0$, $p(z_{mk} = 1|\tilde{\mu}_m, \tilde{n}'_{m\cdot}) = 1$. Otherwise, by subsampling, μ'_m can be written as:

$$\mu'_m = \sum_{m' \leq m} \sum_{k':z_{m'k'}=1} \frac{z_{m'k'} w_{m'k'} \delta_{\theta_{m'k'}}}{\sum_{m'' \leq m} \sum_{k''} z_{m''k''} w_{m''k''}}.$$

Now following the same proof of Theorem 5.13, if we define

$$J_m = \left(\sum_{m' \leq m} \sum_{k'} z_{m'k'} w_{m'k'} \right)^{\tilde{n}'_{m\cdot}}, J_m^{-k} = \left(\sum_{m' \leq m} \sum_{k' \neq k} z_{m'k'} w_{m'k'} \right)^{\tilde{n}'_{m\cdot}},$$

then we get the likelihood as

$$\frac{\prod_{k'':\tilde{n}'_{mk''}>0} {w'_{mk''}}^{n_{k''}}}{J_m}.$$

Furthermore, from subsampling, we know that the Bernoulli prior for z_{mk} is $q^{m-m'}$, and the posterior can then be derived using the Bayes rule as in the proof of Theorem 5.13. □

Chapter 6

Mixed Normalized Random Measures

6.1 Introduction

So far we have introduced dependent Bayesian nonparametric modeling by the hierarchical normalized random measure in Chapter 4, and the Markovian dependent hierarchical normalized random measures used for dynamic topic model in Chapter 5. One problem with the above models is that their posterior structures have not been fully explored due to their intrinsic complexities. This chapter proposes a simple family of dependent normalized random measures called *mixed normalized random measures* (MNRM). The MNRM yields nice analytical posterior structures, allowing efficient posterior inference algorithms to be developed. MNRM is the generalization of a recently proposed spatial dependent Bayesian nonparametric model called the *spatial normalized Gamma process* (SNGP) [Rao and Teh, 2009]. Note a subclass of the MNRM called *spatial normalized random measure* (SNRM) can also be straightforwardly generalized from the SNGP by replacing the Dirichlet process inside with the normalized random measure. It will be shown that in general MNRM, the posterior is simply a generalized Chinese restaurant process, thus posterior inference is as easy as the DP mixture while more complex dependencies can be well captured in MNRM.

The idea of the MNRM is to introduce dependencies on an augmented *spatial space*, say $\mathcal{R}$, while constructing completely random measures on the original product space, say $\mathbb{R}^+ \times \Theta$. That is, for each element $r \in \mathcal{R}$ (it is called a *region*), it is associated with a CRM in $\mathbb{R}^+ \times \Theta$. Consider a set of NRMs $\{\mu_t\}$ indexed by $t \in \mathcal{T}$, where $\mathcal{T}$ can be an arbitrary Borel space, but without loss of generality it can be simply considered as a *time space*, *e.g.*, $\mathbb{R}$. Now there are two ways to introduce dependencies between $\{\mu_t\}$'s:

- by making some of the μ_t's share some elements in space $\mathcal{R}$.
- by putting correlated weights between any region-time pair $(r \in \mathcal{R}, t \in \mathcal{T})$.

In the first way, if only adjacent μ_t's have shared regions, *e.g.*, μ_t and μ_{t+1} share region r, it corresponds to the spatial normalized random measure; whereas in the second way, which includes the first way by defining the "weights" to be either 0 or 1, we arrive at the mixed normalized random measure. Figure 6.1 illustrates the construction. Details of the construction and properties of the SNRM and MNRM will be presented below. To begin with, the special class SNRM will be introduced first.

6.2 Spatial Normalized Random Measures

The spatial normalized random measure (SNRM), as will be shown below, is a special case of the MNRM. Specifically, it follows the construction of spatial normalized Gamma process (SNGP) [Rao and Teh, 2009] by generalizing the normalized Gamma process to normalized random measures. However, note that the sampling method in [Rao and Teh, 2009] for the SNGP cannot be generalized to the general SNRM, thus a new sampler for the whole SNRM family is developed. It will be shown in Section 6.2.4 that the proposed sampler and the original SNGP sampler are equivalent. For completeness, a brief review of the SNGP is given in the next section.

6.2.1 Spatial normalized Gamma processes

The spatial normalized Gamma process (SNGP) [Rao and Teh, 2009] constructs a Bayesian nonparametric prior to model spatial dependency structures, *e.g.*, topic dependency over time. The construction is based on a global Gamma process $\tilde{\mu}$ defined on product space $\mathbb{R}^+ \times \Theta \times \mathcal{R}$ with Lévy measure

$$\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}r) = \alpha(\mathrm{d}\theta, \mathrm{d}r) w^{-1} e^{-w} \mathrm{d}w \ ,$$

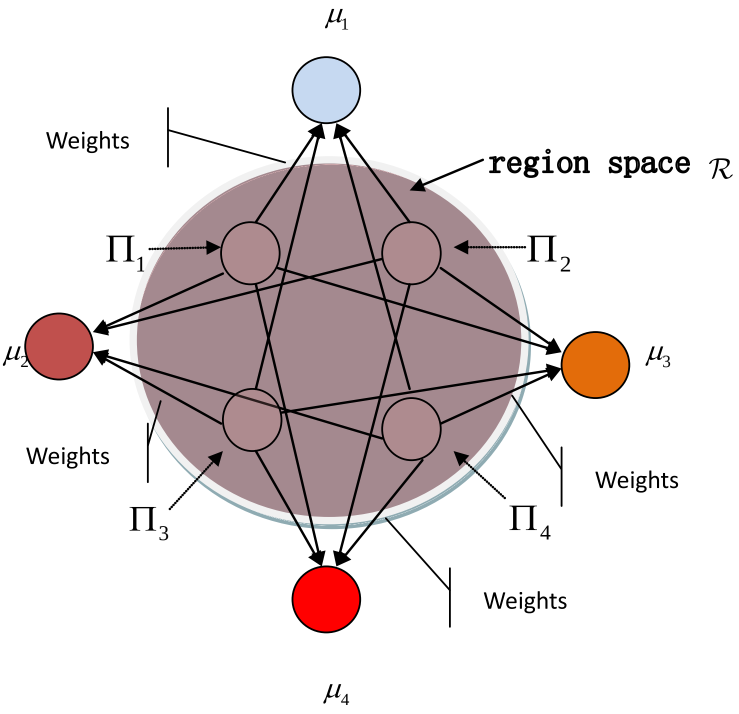
where α is the *mass measure* of the Gamma process. For each time t, a Gamma process is constructed by integrating over the associated regions R_t:

$$\tilde{\mu}_t(\mathrm{d}\theta) = \int_{R_t} \tilde{\mu}(\mathrm{d}\theta, \mathrm{d}r) = \tilde{\mu}(\mathrm{d}\theta, R_t) \ ,$$

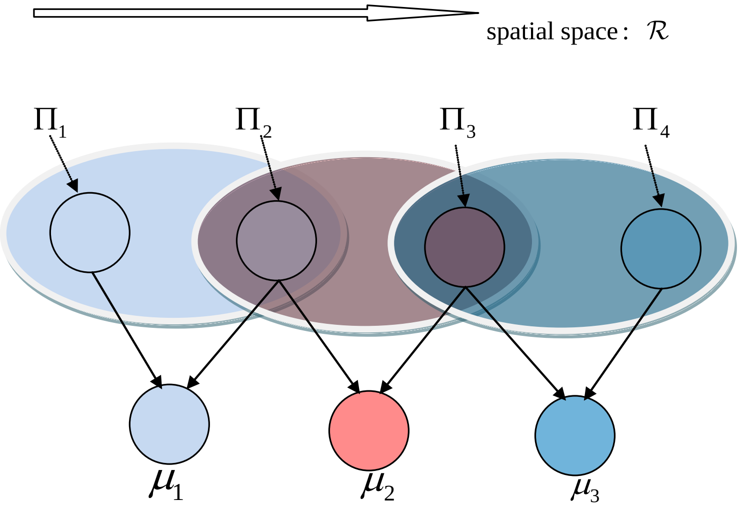
where R_t is the set of *regions* associated to *time t* (see the regions on top of Figure 6.1(b)). Finally, each $\tilde{\mu}_t$ is further normalized to get a set of dependent normalized Gamma processes which endow the spatial dependency by sharing some of their regions between adjacent NRMs:

$$\mu_t = \frac{\tilde{\mu}_t}{\tilde{\mu}_t(\Theta)} \ .$$

Note that when t lives on a dense Borel space, *e.g.*, $t \in \mathbb{R}$, μ_t's constitute a continuous time stochastic process, or measure-value stochastic process, thus the SNRM is quite appearing from both theoretical and application sides. Furthermore, it is easy to show that these μ_t's are marginally Dirichlet process distributed [Rao and Teh, 2009]:



(a) Mixed normalized random measure



(b) Spatial normalized random measure

Figure 6.1: Construction of mixed normalized random measures (MNRM) and a special class called spatial normalized random measures (SNRM). The idea is to correlate μ_t's by sharing and/or weighting of the Poisson processes in separated regions. MNRM takes both sharing and weighting while SNRM only consider sharing with weights being either 0 (unlinked) or 1 (linked).

Theorem 6.1. *Based on the notation and construction as above, each μ_t is marginally a Dirichlet process with corresponding Lévy measure $\nu_t(\mathrm{d}w, \mathrm{d}\theta) = \alpha(\mathrm{d}\theta, R_t) w^{-1} e^{-w} \mathrm{d}w$.*

6.2.2 Spatial normalized random measures

The spatial normalized random measure is defined exactly the same as the SNGP, except that it is generalized to the more general class of normalized random measures instead of the pure Dirichlet process. This not only allows more flexible modeling, *e.g.*, model power-law distributions, but also facilitates a more general posterior inference algorithm for the whole SNRM family to be developed.

In formulation, assume that a global completely random measure $\tilde{\mu}$ is defined on space $\mathbb{R}^+ \times \Theta \times \mathcal{R}$ with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}r)$,[1] which is assumed to be decomposed as

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

with Q being a Lebesgue measure on space $\mathcal{R}$. Now T dependent normalized random measures $\{\mu_t\}$ are constructed by assigning overlapping regions $R_t \subseteq \mathcal{R}$ for consecutive μ_t's, *e.g.*, $R_t \bigcap R_{t+1} \neq \emptyset$, and define the normalized random measure μ_t as:

$$\mu_t = \frac{\tilde{\mu}(\mathrm{d}w, \mathrm{d}\theta, R_t)}{\tilde{\mu}(\mathbb{R}^+, \Theta, R_t)} \ . \tag{6.1}$$

Now it is easy to show that

Corollary 6.2. *Based on the above notation and construction for SNRM, μ_t is a normalized random measure with Lévy measure $\nu_t(\mathrm{d}w, \mathrm{d}\theta) = \nu(\mathrm{d}w, \mathrm{d}\theta, R_t)$.*

Similar to the NGG, this again can be easily specified to the *spacial normalized generalized Gamma process* (SNGG), which has Lévy measure of

$$\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}r) = \frac{M\sigma}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-w} \mathrm{d}w Q(\mathrm{d}r) \mathrm{d}\theta \ . \tag{6.2}$$

Corollary 6.3. *Suppose the global Poisson process has a mean measure of the form as (6.2), following the construction of SNRM, μ_t is a normalized generalized Gamma process with Lévy measure $\nu_t(\mathrm{d}w, \mathrm{d}\theta) = MQ(R_t) w^{-1-\sigma} e^{-w} \mathrm{d}w \mathrm{d}\theta$, where $Q(\cdot)$ is the same as above.*

6.2.3 Posterior inference for spatial normalized generalized Gamma processes

Posterior inference formula for the whole SNRM family can be done with a marginal sampler by applying the Poisson process partition calculus framework introduced in Chapter 2 to integrate out the underlying Poisson process. For simplicity, this section only demonstrates the procedure on a specific class of the spatial normalized random measure – the spatial normalized generalized Gamma process (SNGG). Note

[1] which is equivalent to a Poisson process with mean measure $\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}r)$

similarly to the NRM, the slice sampler is also available, but will not be described until the general case of mixed normalized random measures is covered since the underlying ideas are basically the same.

First note that for a normalized generalized Gamma process with Lévy measure given in (6.2), the corresponding Laplace exponent is given by

$$\psi(u) = MQ(\mathcal{R})\left((u+1)^{\sigma} - 1\right) \tag{6.3}$$

In the following $\mathcal{R}_r$ is used to denote a subspace of the region space $\mathcal{R}$,[2] $\mathcal{T}_r$ to denote the set of time indexes connecting to region $\mathcal{R}_r$. K_r is the number of atoms associated with observations in region $\mathcal{R}_r$, n_{trk} is the number of observations in time t on atoms inherited from region $\mathcal{R}_r$. The superscript in $n_{trk}^{\backslash tl}$ indicates the previous count excluding the lth observation at time t. The observations are denoted as $\boldsymbol{X} = (x_{tl})_{t=1,l=1}^{T,L_t}$, each observation x_{tl} is associated with the s_{tl}-th atom $\theta_{g_{tl}s_{tl}}$ drawn from region g_{tl}. A "dot" is used for marginal sum, *e.g.*, $n_{tr\cdot} = \sum_{k=1}^{K} n_{trk}$. $F(x|\theta)$ denotes the cumulative density function where x is drawn from the corresponding density function denoted as $f(x|\theta)$. N_t denotes the number of observations in time t. Based on the notation, clearly the likelihood is given by[3]:

$$p(\boldsymbol{X}|\{\mu_r\}) = \frac{\prod_{t=1}^{T}\prod_{r\in\mathcal{R}_t}\prod_{k=1}^{K_r} w_{rk}^{n_{trk}}}{\prod_{t'=1}^{T}\left(\sum_{r'\in\mathcal{R}_{t'}}\sum_{k'=1}^{\infty} w_{r'k'}\right)^{N_{t'}}}\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}}) \tag{6.4}$$

Following similar idea as the posterior inference for the NRM, auxiliary variables $\{u_t\}$ are introduced using Gamma identity $\int u^{n-1}\exp\left(-uZ\right)\mathrm{d}u = \Gamma(n)/Z^n$, thus (6.4) is augmented as

$$\begin{aligned} p(\boldsymbol{X},\boldsymbol{u}|\{\mu_r\}) = &\left(\prod_{r=1}^{\#\mathcal{R}}\prod_{k=1}^{K_r} w_{rk}^{n_{\cdot rk}}\exp\left\{-\left(\sum_{t\in\mathcal{T}_r} u_t\right)w_{rk}\right\}\right) \\ &\left(\prod_{t=1}^{T}\frac{u_t^{N_t-1}}{\Gamma(N_t)}\right)\exp\left\{-\sum_{r=1}^{\#\mathcal{R}}\sum_{k=K_r}^{\infty}\left(\sum_{t\in\mathcal{T}_r} u_t\right)w_{rk}\right\}\left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right). \end{aligned} \tag{6.5}$$

Now, rewriting $Q_r = Q(R_r)$ for simplicity and integrating out the random weights with the *extended Palm formula* in Theorem 2.14 (following similar steps as in (3.18)), we get

$$\begin{aligned} &p(\boldsymbol{X},\boldsymbol{u}|\sigma,\{M_r\}) = \mathbb{E}\left[p(\boldsymbol{X},\boldsymbol{u}|\{\mu_r\})\right] \\ &\propto \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r}\left(\prod_{r=1}^{\#\mathcal{R}} M_r^{K_r}Q_r^{K_r}\right)\left(\prod_{r=1}^{R}\prod_{k=1}^{K_r}\frac{\Gamma(n_{\cdot rk}-\sigma)}{\left(1+\sum_{t\in\mathcal{T}_r} u_t\right)^{n_{\cdot rk}-\sigma}}\right) \end{aligned} \tag{6.6}$$

[2] Sometimes r is also used to indicate a subspace of $\mathcal{R}$ for simplicity.

[3] For notation cleanness, μ_r is used to denote the NRM formed in region $\mathcal{R}_r$, *e.g.*, $\tilde{\mu}(\mathrm{d}w, \mathrm{d}\theta, R_t)$ in (6.1), while use a different subscript t to denote the dependent NRMs constructed, *i.e.*, μ_t. This applies in the rest of the thesis.

$$\left(\prod_{t=1}^{T}\frac{u_t^{N_t-1}}{\Gamma(N_t)}\right)\left(\prod_{r=1}^{\#\mathcal{R}}e^{-M_rQ_r\left(\left(1+\sum_{t\in\mathcal{T}_r}u_t\right)^{\sigma}-1\right)}\right)\left(\prod_{t=1}^{T}\prod_{l=1}^{L_t}f(x_{tl}|\theta_{g_{tl}s_{tl}})\right)$$

Given this joint distribution, a Gibbs sampler for topic indicators $\{s_{tl}\}$, source indicators $\{g_{tl}\}$, mass parameters $\{M_r\}$ and the latent relative mass parameters $\{u_t\}$ can be easily derived. Denote the whole set of variables as C, the sampling then goes as:

Sampling (s_{tl}, g_{tl}):

$$p(s_{tl}=k, g_{tl}=r|C-\{s_{tl},g_{tl}\}) \propto$$
$$\begin{cases}\frac{(n_{\cdot rk}^{/tl}-\sigma)}{1+\sum_{t:r\in\mathcal{R}_t}u_t}f_{rk}^{\backslash tl}(x_{tl}), & \text{if } k \text{ already exists,}\\ \sigma\left(\sum_{r'\in\mathcal{R}_t}\frac{M_{r'}Q_{r'}}{\left(1+\sum_{t'\in\mathcal{T}_{r'}}u_{t'}\right)^{1-\sigma}}\right)\int_{\Theta}f(x_{tl}|\theta)h(\theta)\mathrm{d}\theta, & \text{if } k \text{ is new ,}\end{cases}$$

where h is the density of H, $f_{rk}^{\backslash tl}(x_{tl}) = \frac{\int f(x_{tl}|\theta_{rk})\prod_{t'l'\neq tl, s_{t'l'}=k, g_{t'l'}=r}f(x_{t'l'}|\theta_{rk})h(\theta_{rk})\mathrm{d}\theta_{rk}}{\int\prod_{t'l'\neq tl, s_{t'l'}=k, g_{t'l'}=r}f(x_{t'l'}|\theta_{rk})h(\theta_{rk})\mathrm{d}\theta_{rk}}$ is the conditional density [4].

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

$$M_r|C-M_r \sim \text{Gamma}\left(K_r+a_m, Q_r\left(1+\sum_{t:r\in\mathcal{R}_t}u_t\right)^{\sigma}+b_m-Q_r\right),$$

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

Sampling u_t: The posterior distribution of u_t is:

$$p(u_t|C-u_t) \propto \frac{u_t^{N_t-1}\exp\left\{-\sum_r M_rQ_r\left(1+\sum_{t':r\in\mathcal{R}_{t'}}u_{t'}\right)^{\sigma}\right\}}{\prod_r\left(1+\sum_{t':r\in\mathcal{R}_{t'}}u_{t'}\right)^{\sum_k n_{\cdot rk}-\sigma K_r}},$$

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

Sampling σ: To resample σ, first instantiate the weights w_{rk} associated with extant clusters:

$$w_{rk} \sim \text{Gamma}\left(n_{\cdot rk}-\sigma, 1+\sum_{t:r\in\mathcal{R}_t}u_t\right) \tag{6.7}$$

[4] This thesis assumes conjugacy between f and h so that the integration has a close form, though the non-conjugate case can be dealt with techniques from for example [Neal, 2000] or [Favaro and Teh, 2013]

Then, the conditional distribution of σ is given by:

$$p(\sigma|C-\sigma) \propto \left(\frac{a}{\Gamma(1-\sigma)}\right)^{\sum_r K_r} \left(\prod_{r=1}^{I}\prod_{k=1}^{K_r} w_{rk}\right)^{-\sigma} \left(\prod_r e^{-M_r Q_r\left(\left(1+\sum_{t:r\in\mathcal{R}_t} u_t\right)^{\sigma}-1\right)}\right) \tag{6.8}$$

It can be checked that now (6.8) is log-concave so that it can be sampled using the adaptive rejection sampler [Gilks and Wild, 1992] or slice sampler [Neal, 2003].

6.2.4 Relation to the original sampler for spatial normalized Gamma processes

The above marginal sampler is applicable to any homogeneous spatial normalized random measures, and is different from the one proposed for the spatial normalized Gamma process (SNGP) in [Rao and Teh, 2009]. This section develops a connection between these two samplers by showing how the sampler used for the SNGG can be transformed to the one used for the SNGP in [Rao and Teh, 2009].

For the SNGP, the Gamma process in region r has Lévy measure

$$\nu_r(\mathrm{d}w, \mathrm{d}\theta) = \nu(\mathrm{d}w, \mathrm{d}\theta, \mathcal{R}_t) = M_r Q_r w^{-1} e^{-w} \mathrm{d}w H(\theta) \mathrm{d}\theta \ .$$

After some algebra, the corresponding posterior (6.6) is equivalent to

$$\begin{aligned} &p(\boldsymbol{X}, \boldsymbol{u}|\{M_r\}) \\ &= \left(\prod_{r=1}^{\#\mathcal{R}} M_r^{K_r} Q_r^{K_r}\right) \left(\prod_{t=1}^{T} \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \\ &\left(\prod_{r=1}^{R} \frac{\prod_{k=1}^{K_r}\Gamma(n_{\cdot rk})}{\left(1+\sum_{t\in\mathcal{T}_r} u_t\right)^{n_{\cdot r\cdot}+M_r Q_r}}\right) \end{aligned} \tag{6.9}$$

Now introduce a set of auxiliary variables $\{g_r\}$ via the Gamma identity

$$\left(1+\sum_{t\in\mathcal{T}_r} u_t\right)^{-(n_{\cdot r\cdot}+M_r Q_r)} = \frac{1}{\Gamma\left(n_{\cdot r\cdot}+M_r Q_r\right)} \int_{\mathbb{R}^+} g_r^{n_{\cdot r\cdot}+M_r Q_r} e^{-(1+\sum_{t\in\mathcal{T}_t} u_t) g_r} \mathrm{d}g_r \ ,$$

we have the augmented posterior

$$\begin{aligned} &p(\boldsymbol{X}, \{u_t\}, \{g_r\}|\{M_r\}) \\ &= \left(\prod_{r=1}^{\#\mathcal{R}} M_r^{K_r} Q_r^{K_r}\right) \left(\prod_{t=1}^{T} \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \end{aligned}$$

$$\left(\prod_{r=1}^{\#\mathcal{R}} \frac{\left(\prod_{k=1}^{K_r} \Gamma(n_{\cdot rk})\right) g_r^{n_{\cdot r \cdot}+M_r Q_r} e^{-(1+\sum_{t\in\mathcal{T}_t} u_t) g_r}}{\Gamma\left(n_{\cdot r \cdot}+M_r Q_r\right)}\right)$$

Now $\{u_t\}$'s can be integrated out as

$$p(\mathbf{X},\{g_r\}|\{M_r\}) = \left(\prod_{r=1}^{\#\mathcal{R}} M_r^{K_r} Q_r^{K_r}\right) \left(\prod_{t=1}^{T} \left(\sum_{r\in\mathcal{R}_t} g_r\right)^{-N_t}\right) \left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right)$$
$$\left(\prod_{r=1}^{\#\mathcal{R}} \frac{\left(\prod_{k=1}^{K_r} \Gamma(n_{\cdot rk})\right) g_r^{n_{\cdot r \cdot}+M_r Q_r} e^{-g_r}}{\Gamma\left(n_{\cdot r \cdot}+M_r Q_r\right)}\right)$$

This is exactly the same posterior used to derive the sampler in [Rao and Teh, 2009] for the SNGP. The conditional probabilities as those given by [Rao and Teh, 2009] can be derived from this posterior, which are omitted here for simplicity.

Regarding the differences between the sampler in [Rao and Teh, 2009] (denoted as P_1) and the sampler for the SNRM (denoted as P_2), it can be seen that the number of auxiliary variables in P_1 equals to the number of regions, while it equals to the number of times in P_2. According to the construction of SNGP, the number of regions is usually larger than the number of times ($\#\mathcal{R} = O((\#\mathcal{T})^2)$), thus sampler P_2 is preferable because it contains less auxiliary variables so the sampling cost could be cheaper. Furthermore, note that P_1 is only applicable for the special case of SNGP, while P_2 is much more flexible and applicable for all classes of SNRMs.

6.3 Mixed Normalized Random Measures

This section extends the spatial normalized random measure to the *mixed normalized random measure* (MNRM). MNRM generalizes SNRM in the way that instead of explicitly defining a spatial sharing structure between the dependent NRMs μ_t's, *e.g.*, $R_t \bigcap R_{t+1} \neq \emptyset$, it assumes that μ_t connects to all the regions but with different *weights* q_{rt}'s for different regions $\mathcal{R}_r$.[5] In this way, the correlation between the pair $(\mu_t, \mu_{t'})$ is controlled by these weights. Specifically, let q_{rt} be a nonnegative weight between region $\mathcal{R}_r$ and time t.[6] The MNRM μ_t is simply defined as follows:

Definition 6.1 (Mixed Normalized Random Measure (MNRM))**.** Let $\mathcal{N}$ be the Poisson random measure on $\mathbb{R}^+ \times \Theta \times \mathcal{R}$. The MNRM is defined by the following construction:

- For each region $\mathcal{R}_r$, define a completely random measure:

$$\tilde{\mu}_r(\mathrm{d}\theta) = \int_{\mathbb{R}^+\times\tilde{R}_r} w\mathcal{N}(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}r) \ .$$

[5] A weight of 0 means absent of the connection, thus it is a generation of the SNRM mechanism.

[6] In the experiments, independent Gamma priors are placed on the q_{rt}'s, thus their values can be inferred from the data.

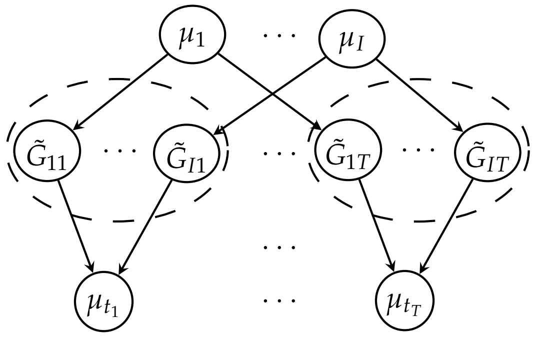


Figure 6.2: Construction of mixed normalized measure from R independent NRMs μ_r, $\tilde{G}_{rt}$ represents $q_{rt}\tilde{\mu}_r(\mathrm{d}\theta)$ defined in (6.10).

- For each time t, construct a dependent completely random measure:

$$\tilde{\mu}_t(\mathrm{d}\theta) = \sum_{r=1}^{\#\mathcal{R}} q_{rt}\tilde{\mu}_r(d\theta) \ . \tag{6.10}$$

- Normalize the completely random measure:

$$\mu_t(\mathrm{d}\theta) = \frac{1}{Z_t}\tilde{\mu}_t(\mathrm{d}\theta) \ , \text{ where } Z_t = \tilde{\mu}_t(\Theta) \ .$$

The construction is illustrated in Figure 6.2, where each μ_t is constructed by superpositioning a set of intermediate random measures to achieve dependencies.

Note in particular μ_t can be rewritten as:

$$\begin{aligned}\mu_t(\mathrm{d}\theta) &= \frac{\sum_{r=1}^{\#\mathcal{R}} q_{rt}\tilde{\mu}_r(\mathrm{d}\theta)}{\tilde{\mu}_t(\Theta)} \\ &= \sum_{r=1}^{\#\mathcal{R}} \frac{q_{rt}\tilde{\mu}_r(\Theta)}{\tilde{\mu}_t(\Theta)}\frac{\tilde{\mu}_r(\mathrm{d}\theta)}{\tilde{\mu}_r(\Theta)} = \sum_{r=1}^{\#\mathcal{R}} \frac{q_{rt}\tilde{\mu}_r(\Theta)}{\tilde{\mu}_t(\Theta)}\mu_r(\mathrm{d}\theta)\end{aligned}$$

Thus it can be easily seen 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:

Theorem 6.4. *Conditioned on the q_{rt}'s, each random probability measure μ_t defined in Definition 6.1 is marginally distributed as a NRM with Lévy intensity* [7] $\sum_{r=1}^{R}\frac{1}{q_{rt}}\nu_r(w/q_{rt},\theta)$.

Proof. 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. Specifically:

First, from the definition we have

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

[7] Lévy intensity is defined as the density of the Lévy measure. With a little abuse of notation, the Lévy measure ν is still used to denote its Lévy intensity but without the derivative operator "d".

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, by applying the Lévy-Khintchine Formula in Lemma 3.2, we calculate the characteristic functional of each random measure $q_{rt}\tilde{\mu}_r$ as:

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

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

$$\begin{aligned}
\varphi_{\tilde{\mu}_t}(u) &= \prod_{r=1}^{\#\mathcal{R}} \varphi_{q_{rt}\tilde{\mu}_r}(u) \\
&= e^{-\int_{\mathbb{R}^+ \times \Theta}(1-e^{iuw})\sum_{r=1}^{\#\mathcal{R}}\nu_r(w/q_{rt},\theta)\mathrm{d}w/q_{rt}\mathrm{d}\theta}, \qquad (6.11)
\end{aligned}$$

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

6.3.1 Comparison with the SNGP

The spatial normalized Gamma process (SNGP) of [Rao and 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.

6.3.2 Posterior Inference

In the following, again a specific NRM *viz.* the *normalized generalized Gamma process* (NGG) is studied to demonstrate the posterior inference. Generalization to other NRMs is straightforward. As is known, the NGG, which includes the DP as a special case, is attractive 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 [Favaro et al., 2013a], 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 Chapter 3 for review of the CRP). A consequence of this is that marginal samplers as well as slice samplers are available for MNGG, which will be derived in the following sections. In the following, the same notation as in SNGG will be used.

6.3.2.1 Posterior inference for mixed normalized generalized Gamma processes with marginal sampler

For completeness, this section first gives the posterior of MNGGs, which is almost the same as the SNGG[8]. Given observations $\boldsymbol{X}$ and weights q_{rt}'s, denote μ_r as the NRM in region $\mathcal{R}_r$, the likelihood can be expressed as

$$p(\boldsymbol{X}|\{\mu_r\},\{q_{rt}\}) = \frac{\prod_{t=1}^{T}\prod_{r=1}^{I}\prod_{k=1}^{K_r}\left(q_{rt}w_{rk}\right)^{n_{trk}}}{\prod_{t'=1}^{T}\left(\sum_{r'=1}^{I}\sum_{k'=1}^{\infty} q_{r't'}w_{r'k'}\right)^{N_{t'}}}\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}}) \tag{6.12}$$

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

$$\begin{aligned} & p(\boldsymbol{X},\boldsymbol{u}|\{\mu_r\},\{q_{rt}\}) \qquad (6.13)\\ = \; & \left(\prod_{r=1}^{I}\prod_{t=1}^{T} q_{rt}^{n_{rt\cdot}}\right)\left(\prod_{r=1}^{I}\prod_{k=1}^{K_r} w_{rk}^{n_{\cdot rk}} \exp\left\{-\left(\sum_{t=1}^{T} q_{rt}u_t\right) w_{rk}\right\}\right)\\ & \left(\prod_{t=1}^{T}\frac{u_t^{N_t-1}}{\Gamma(N_t)}\right)\exp\left\{-\sum_{r=1}^{I}\sum_{k=1}^{\infty}\left(\sum_{t=1}^{T} q_{rt}u_t\right) w_{rk}\right\}\left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \end{aligned}$$

Using the factorized Lévy-measure of $\nu(\mathrm{d}w,\mathrm{d}\theta,\mathrm{d}r) = \nu'(\mathrm{d}w,\mathrm{d}\theta)Q(\mathrm{d}r)$, it is easily seen μ_r's are normalized generalized Gamma processes with Lévy measures

$$\nu_r(\mathrm{d}w,\mathrm{d}\theta) = \int_{R_r}\nu(\mathrm{d}w,\mathrm{d}\theta,\mathrm{d}r) = \frac{\sigma M_r Q_r}{\Gamma(1-\sigma)} w^{-1-\sigma}e^{-w}\mathrm{d}wH(\theta)\mathrm{d}\theta\ .$$

Integrate out μ_r's by applying the Poisson process partition calculus formula in Theorem 2.12 we get:

$$\begin{aligned} & p(\boldsymbol{X},\boldsymbol{u}|\sigma,\{M_r\},\{q_{rt}\}) = \mathbb{E}_{\{\mu_r\}}\left[p(\boldsymbol{X},\boldsymbol{u}|\{\mu_r\},\{q_{rt}\})\right] \qquad (6.14)\\ \propto \; & \left(\prod_{t=1}^{T}\prod_{r=1}^{\#\mathcal{R}} q_{rt}^{n_{rt\cdot}}\right)\left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{K.}\left(\prod_{r=1}^{\#\mathcal{R}}(Q_rM_r)^{K_r}\right)\left(\prod_{r=1}^{\#\mathcal{R}}\prod_{k=1}^{K_r}\frac{\Gamma(n_{\cdot rk}-\sigma)}{(1+\sum_t q_{rt}u_t)^{n_{\cdot rk}-\sigma}}\right)\\ & \left(\prod_{t=1}^{T}\frac{u_t^{N_t-1}}{\Gamma(N_t)}\right)\left(\prod_{r=1}^{\#\mathcal{R}} e^{-Q_rM_r\left((1+\sum_t q_{rt}u_t)^{\sigma}-1\right)}\right)\left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right)\ . \end{aligned}$$

Since Q_r and M_r always appear together, we simply omit Q_r and only use M_r to represent Q_rM_r in the above formula.

Now it is straightforward to derive the posterior sampler for MNRM. The variables needed to be sampled are $C = \{\{s_{tl}\},\{g_{tl}\}\{M_r\},\{u_t\},\{q_{rt}\}\}$, where the first four sets of parameters are the same as those in the SNRM, whereas the last set is called weighting parameters of the MNRM. Based on (6.14), these can be iteratively sampled as follows:

[8] Thus reader familiar with the SNGG could skip the derivation.

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

$$p(s_{tl} = k, g_{tl} = r | C - s_{tl} - g_{tl}) \tag{6.15}$$
$$\propto \begin{cases} \frac{q_{rt}(n_{\cdot rk}^{\setminus tl} - \sigma)}{1 + \sum_{t'} q_{rt'} u_{t'}} f_{rk}^{\setminus tl}(x_{tl}), & \text{if } k \text{ already exists,} \\ \sigma \left(\sum_{r'} \frac{q_{r't} M_{r'}}{(1 + \sum_{t'} q_{r't'} u_{t'})^{1-\sigma}} \right) \int_{\Theta} f(x_{tl}|\theta) h(\theta) \mathrm{d}\theta, & \text{if } k \text{ is new ,} \end{cases}$$

where $f_{rk}^{\setminus tl}(x_{tl})$ is the same as in SNGG.

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}**:** q_{rt} can be sampled by introducing appropriate priors. Here a Gamma prior with parameter q_a and q_b is used, so the posterior of q_{rt} has the following form:

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

which is also log-concave with a change of variables: $Q_{rt} = \log(q_{rt})$.

Sampling σ**:** From (6.14), to sample σ, a set of jumps $\{w_{rk}\}$ are first instantiated:

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

Based on these jumps, the posterior of σ is proportional to:

$$p(\sigma | C - \sigma) \propto \left(\frac{\sigma}{\Gamma(1 - \sigma)} \right)^{K.} \left(\prod_{r=1}^{\#\mathcal{R}} \prod_{k=1}^{K_r} w_{rk} \right)^{-\sigma} \left(\prod_{r=1}^{\#\mathcal{R}} e^{-M_r (1 + \sum_t q_{rt} u_t)^{\sigma}} \right) \tag{6.16}$$

which is log-concave as well.

6.3.2.2 Posterior inference for mixed normalized generalized Gamma processes with slice sampler

As an alternative sampling algorithm, this section describes the slice sampler for the MNRM. Similar to the case of single NRM, 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. This turns the inference from infinite parameter spaces to finite parameter spaces.

Starting from (6.13), a slice auxiliary variable v_{tl} is introduced for each observation such that

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

Now (6.13) can be augmented as

$$\begin{aligned} & p(\boldsymbol{X}, \boldsymbol{u}, \{v_{tl}\}, \{s_{tl}\}, \{g_{tl}\} | \{\mu_r\}, \{q_{rt}\}) \\ = \quad & \left(\prod_t \prod_l 1 \left(w_{g_{tl}s_{tl}} > v_{tl} \right) q_{g_{tl}s_{tl}} f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\ & \left(\exp\left\{ - \sum_t \sum_r \sum_k q_{rt} u_t w_{rk} \right\} \right) \end{aligned} \tag{6.17}$$

The joint distribution of observations, related auxiliary variables and the corresponding Poisson random measure $\{\mathcal{N}_r\}$[9] can be written as

$$\begin{aligned} & p(\boldsymbol{X}, \boldsymbol{u}, \{\boldsymbol{v_{tl}}\}, \{\mu_r\}, \{s_{tl}\}, \{g_{tl}\}, \{\mathcal{N}_r\} | \{q_{rt}\}) \\ = \quad & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) q_{g_{tl}s_{tl}} f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\ & \left(\exp\left\{ - \sum_t \sum_r \sum_k q_{rt} u_t w_{rk} \right\} \right) \prod_r P(\mathcal{N}_r) \\ \overset{\text{slice at } \mathcal{L}_r}{=} \quad & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) q_{g_{tl}s_{tl}} f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\ & \underbrace{\exp\left\{ - \sum_t \sum_r \sum_k q_{rt} u_t w_{rk} \right\}}_{\text{jumps larger than } \mathcal{L}_r} \\ & \prod_r p(\{(w_{r1}, \theta_{r1})\}, \cdots, \{(w_{rK'_r}, \theta_{rK'_r})\}) \quad (K'_r \text{ is \# jumps larger than } \mathcal{L}_r) \end{aligned} \tag{6.18}$$

$$\underbrace{\prod_r \exp\left\{ - \frac{\sigma M_r}{\Gamma(1-\sigma)} \int_0^{\mathcal{L}_r} \left(1 - e^{-\sum_t q_{rt} u_t x} \right) \rho'(\mathrm{d}x) \right\}}_{\text{jumps less than } \mathcal{L}_r} \,, \tag{6.19}$$

[9] As mentioned in Chapter 2, usually stochastic processes including Poisson process do not endow probability density functions, however, we use the notation $P(\mathcal{N})$ here to emphasis the joint distribution with the Poisson random measure $\mathcal{N}$. This applies in the rest of the thesis without further declaration.

where $\rho'(\mathrm{d}x) = x^{-1-\sigma}e^{-x}$ and (6.18) has the following form based on the fact that $\{(w_{rk}, \theta_{rk})\}$ are points from a compound Poisson process, so the density is:

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

where Poisson$(k; A)$ means the density of the Poisson distribution with mean A under value k.

Given the joint posterior, apart from the variables in the marginal sampler for the MNRM, additional random variables we are interested in include the threshold variables $\{v_{tl}\}$ as well as the jump sizes $\{w_{rk}\}$. As before, the whole set is denoted as C, then 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:

$$p(s_{tl} = k, g_{tl} = r | C - \{s_{tl}, g_{tl}\}) \propto 1(w_{rk} > v_{tl}) q_{rk} f(x_{tl}|\theta_{rk}) . \tag{6.20}$$

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

$$v_{tl}|C - v_{tl} \sim \text{Uniform}(0, w_{g_{tl}s_{tl}}) . \tag{6.21}$$

Sample w_{rk}: There are two kinds of w_{rk}'s, one is with observations, the other is not, because they are independent, they are sampled 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, the adaptive thinning approach introduced in [Favaro and Teh, 2013] is used with a proposal adaptive Poisson process mean measure as

$$\gamma_x(s) = \frac{\sigma M_r}{\Gamma(1-\sigma)} e^{-(1+\sum_t q_{rt}u_t)s} x^{-1-\sigma} \tag{6.22}$$

The idea of this method is to generate samples from the proposal Poisson process with mean measure $\gamma_x(\cdot)$, and reject some of the samples to make them be samples from the desired Poisson process. The details of the simulation of the Poisson points will be delayed to Section 7.4.2 in Chapter 7 because the slice sampler is not the most effective approach in MNRM, but it seems to be the only choice for posterior sampling of the new model–thinned normalized random measures in Chapter 7.

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 in Theorem 3.15.

Sample u_t: From (6.19), u_t is sampled 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 a rejection step is done by evaluating it on the posterior (6.19).

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 (6.19), the posterior of σ is proportional to:

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

Thought log-concaveness is not guaranteed, it still can be sampled using the slice sampler of [Neal, 2003].

6.4 Experiments

6.4.1 Illustration

This section illustrates how MNRM works via a Gaussian mixture example with MNRM as the prior. A 2D Gaussian mixture dataset consisting of 4 Gaussian components is first generated. The first three Gaussians have means around $(2,0)^T$ and covariance $0.3 \times \mathbf{I}$, where $\mathbf{I}$ is the 2-dimensional identity matrix; while the last Gaussian has mean around $(4,0)^T$ and covariance $0.6 \times \mathbf{I}$. So this dataset can be thought of as generated from 2 regions, where the first consists of the first three Gaussians, and the last consisting of the last Gaussian component. Then 3 groups of data points are generated (each groups corresponds to *one time t*), each consisting of 70 data points. To generate data points for time t, a region r is first chosen with probability proportional to $\frac{1}{|r-t|+1}$, then a Gaussian component in the region is randomly chosen for drawing the data.

After generating the data, the MNGG mixture model with $\sigma = 0.1$ and base distribution as *Gaussian-Wishart* distribution (see Chapter 3) is tested on this data. The hyperparameters for the *Gaussian-Wishart* are chosen as $(r = 0.25, \nu = 5, \boldsymbol{m} = (0,0)^T, \mathbf{S} = \text{eye}(2))$. The q is set to the one used in the data generation above during inference and other variables are sampled. The result of the MNGG is shown in Figure 6.3(a), where it is clear that MNRM successfully recovers the Gaussians in the 2 regions. The Dirichlet process Gaussian mixture model (DPGMM) [Rasmussen, 2000] is also tested and compared on this dataset, the result is shown in Figure 6.3(b). We can see that compared with MNGG, DPGMM seems having difficulty in correctly finding the large variance Gaussian component.

6.4.2 Two constructions for topic modeling

In the following, the ideas of modeling text documents organized in time are applied to the MNRM framework. Four specified models will be studied in this section (defined below): 1) *mixed normalized generalized Gamma process* (MNGG), 2) *hierarchical mixed normalized generalized Gamma process* (HMNGG), 3) *hierarchical mixed normalized Gamma process* (HMNGP) and 4) *hierarchical spatial normalized generalized Gamma process* (HSNGG).

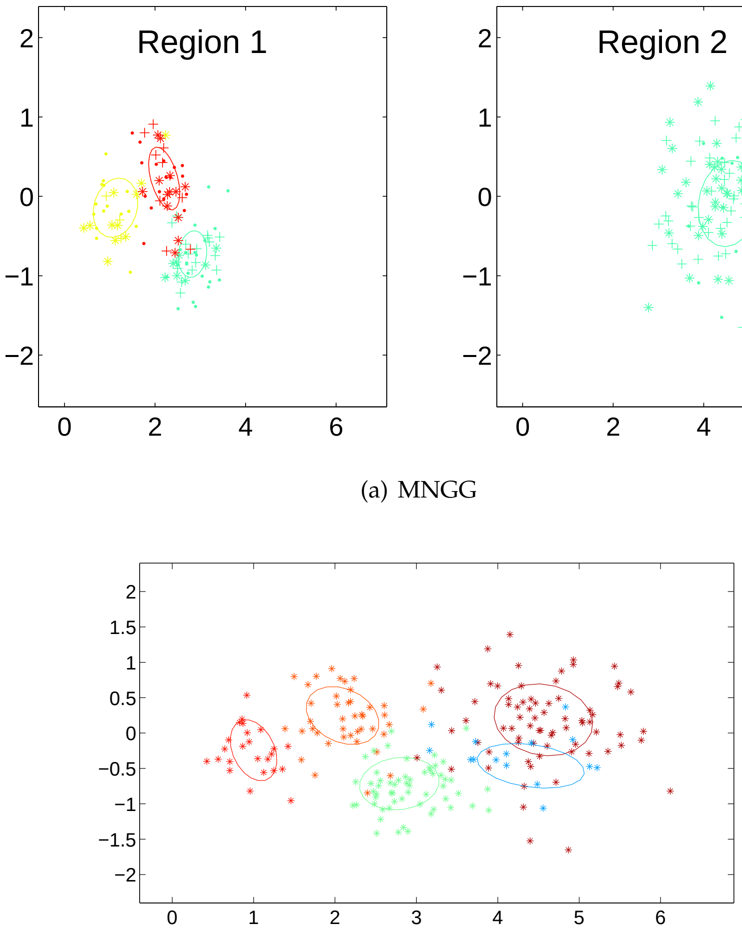
The first model is based on the mixed construction, with each document assigned to its own 'time', so there is no hierarchical sharing for the documents at the same time. This on the one hand, disregards statistical information that might be shared across documents from the same true time period, on the other hand, it affords more flexibility, since each document can have its own set of q_{rt} parameters. Specifically, let G be the Dirichlet distribution, F the multinomial distribution, and t span all documents in the corpus, the generative process is as follows:

$$(\mu_t) \sim \text{MNGG}(\sigma_0, M_0, G, \{q_{rt}\}), \qquad \text{for each doc } t \tag{6.23}$$

$$\theta_i^t \sim \mu_t, \qquad x_i^t | \theta_i^t \sim F(\cdot|\theta_i^t), \qquad \text{for each word } i\ , \tag{6.24}$$

where $\text{MNGG}(\sigma_0, M_0, G, \{q_{rt}\})$ denotes the dependent NGG constructed via MNGG 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, the MNRM constructions are



(a) MNGG

(b) DPGMM

Figure 6.3: A demo on MNGG mixture, compared with DPGMM. MNGG correctly recovers the two regions in the data generating process.

used to produce an RPM μ_t for each time-period t; each document in time period t then has a distribution over topics drawn from an NGG with base-measure μ_t:

$$\begin{aligned}(\mu_t)|\sigma_0, M_0, G, \{q_{rt}\} &\sim \text{MNGG}(\sigma_0, M_0, G, \{q_{rt}\})\\ \{\mu_{ti}\}|\mu_t &\sim NGG(\sigma, M, \mu_t)\\ \theta_{ij}^t \sim \mu_{ti}, \quad x_{ij}^t|\theta_{ij}^t &\sim F(\cdot|\theta_{ij}^t)\ ,\end{aligned} \tag{6.25}$$

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 and Teh, 2009], a special case of HMNGG with $q_{rt} \in \{0, 1\}$. The models are also compared with the popular hierarchical Dirichlet process (HDP), as well as the hierarchical normalized generalized Gamma process (HNGG) proposed in Chapter 4. All the algorithms are implemented in C/C++; the slice sampler needs careful memory management, requiring tricks such as memory pre-allocation and pooling to manage the computational requirements.

6.4.3 Synthetic data

In the first experiment, a dataset with 3000 observations are generated from a hierarchical Pitman-Yor topic model [Du et al., 2010] [10]. The vocabulary size is set to 100. The generative process is described as follows

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

The base measure G over topic distributions is a 100-dimensional symmetric Dirichlet with parameter 0.1, while $F(\cdot|\theta)$ is the 100-dimensional discrete distribution. The concentration parameters $\alpha_i, i = 0, \cdots, 3$ are set to $1, 3, 4$ and 5 respectively, while all discount parameters d_i are 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 model the resulting corpus using the HMNGG described in (6.25). 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, the performance of HMNGG on this dataset against its dependent DP equivalence–the HMNGP (obtained by replacing the generalized Gamma process with the Gamma process in the constructions) are compared. In the experiments, the number of regions is set equal to the number of times, and all the model parameters are sampled during inference (placing $\text{Gamma}(0.1, 0.1)$ priors on all scalars in $\mathbb{R}^+$).

Figure 6.4 plots the predictive likelihood on a 20% held-out dataset as well as the effective sample sizes of the models. It can be seen that HMNGG outperforms its non-power-law variants HMNGP in terms of predictive likelihoods. The inferred parameter σ is around 0.2 (a value of 0 recovers the Gamma process).

6.4.4 Topic modelling

Datasets Next, four real-world document datasets are used for topic modeling, viz. ICML, TPAMI, Person and NIPS. The first 2 corpora consist of abstracts obtained from the ICML and PAMI websites; ICML contains 765 documents from 2007-2011 with a total of about 44K words, and a vocabulary size of about 2K; TPAMI has 1108 documents from 2006-2011, with total of 91K words and vocabulary size of 3K. The Person dataset is extracted from Reuters RCV1 using the query *person* under Lucene [Otis et al., 2009], and contained 8616 documents, 1.55M words and a vocabulary size of 60K. It spans the period 08/96 to 08/97. The NIPS corpus consists of proceedings over the years 1987 to 2003 [Globerson et al., 2007]. It is not postprocessed, and has 2483 documents, 3.28M words and vocabulary size 14K. The statistics

[10] An HDP-LDA topic model [Teh et al., 2006] by replacing the DPs with PYPs.

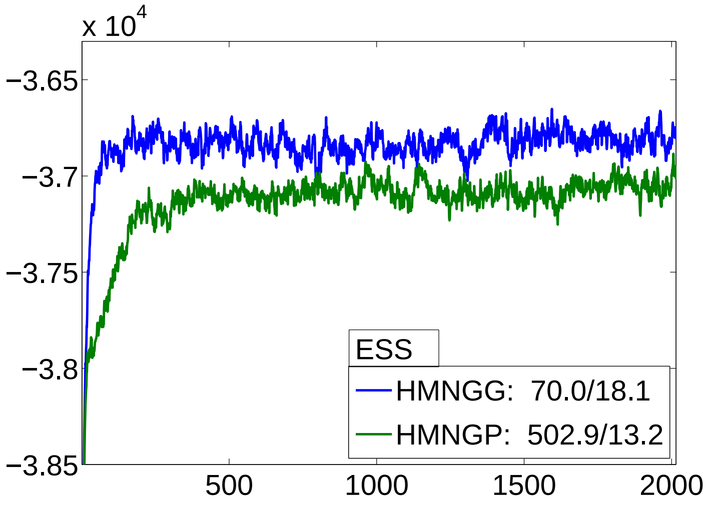


Figure 6.4: HMNGG VS. HMNGP.

dataset	vocab	docs	words	epochs
ICML	2k	765	44k	2007–2011
TPAMI	3k	1108	91k	2006–2011
Person	60k	8616	1.55M	08/96–08/97
NIPS	14kk	2483	3.28M	1987–2003

Table 6.1: Data statistics

are listed in Table 6.1.

Parameter setting and evaluation In modeling these datasets, for MNGG (where the years associated with each document are disregarded), the number of regions was set 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, perplexity scores are computed on a held-out test dataset. In all cases, 20% of the original data sets is held-out, following the standard dictionary hold-out method (50% of the held-out documents is used to estimate topic probabilities) [Rosen-Zvi et al., 2004]. Test perplexity is calculated over 10 repeated runs with random initialization, mean values and standard deviations are reported. In each run 2000 cycles are 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, perplexities are calculated after first transforming the representation of the slice sampler to those of the marginal sampler. In other words, given the state of the slice sampler, the induced partition structures are determined and used to calculate prediction probabilities (calling the

Table 6.2: Train perplexities and test perplexities for different models on ICML, TPAMI, Person and NIPS datasets.

Datasets	ICML		TPAMI	
Models	train	test	train	test
HDP	580 ± 6	1017 ± 8	671 ± 6	1221 ± 6
HNGG	575 ± 5	1057 ± 8	671 ± 6	1262 ± 11
MNGG	569 ± 6	1056 ± 9	644 ± 6	1272 ± 12
HSNGG	550 ± 5	1007 ± 8	643 ± 3	1237 ± 22
HMNGG	$\mathbf{535 \pm 6}$	1001 ± 10	$\mathbf{608 \pm 4}$	1199 ± 10
HMNGP	561 ± 10	995 ± 14	634 ± 10	1208 ± 8

Datasets	Person		NIPS	
Models	train	test	train	test
HDP	4541 ± 33	5962 ± 43	1813 ± 27	1956 ± 18
HNGG	4565 ± 60	5999 ± 54	1713 ± 13	1878 ± 11
MNGG	4560 ± 63	6013 ± 66	1612 ± 3	1920 ± 5
HSNGG	4324 ± 77	5733 ± 66	1406 ± 5	1679 ± 8
HMNGG	$\mathbf{4083 \pm 36}$	$\mathbf{5488 \pm 44}$	$\mathbf{1366 \pm 8}$	$\mathbf{1618 \pm 5}$
HMNGP	4118 ± 45	5519 ± 41	1370 ± 3	1634 ± 4

same piece of code).

Quantitative comparison for different models Both training and test perplexities for the models specified above are calculated, which are shown in Table 6.2.

It is seen that HMNGG performs best, achieving significant lower perplexities than the others. 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 does much worse than HMNGG, emphasizing the importance of statistical information shared across documents in the same year.

Topic evolution Figure 6.5 is a posterior sample, showing the evolution of 12 randomly selected topics on the NIPS dataset for HMNGG. The proportion of words assigned to the topic k in region r at each time t (i.e. $\frac{n_{trk}}{n_{tr\cdot}}$) are calculated, as well as the predictive probabilities for each topic at each time. The latter is defined for MNGG to be proportional to $\frac{q_{rt}(n^{\backslash tl}_{\cdot rk}-\sigma)}{1+\sum_{t'} q_{rt'} u_{t'}}$ (see equation 6.15).

Marginal vs slice sampler Next the performance of the marginal and slice samplers for MNGG and HMNGG are compared. Table 6.3 shows the average effective sample

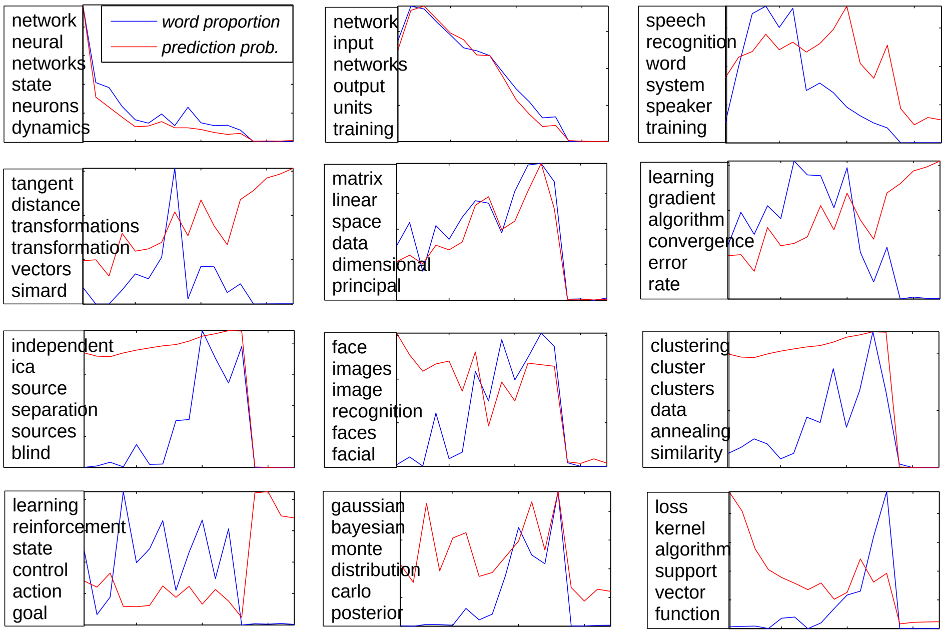


Figure 6.5: Topic evolution on NIPS dataset for 12 randomly chosen topics learned by HMNGG. The two curves correspond to word proportions within each topic (blue) and prediction probabilities (red) for each time.

sizes and running times over 5 repeated runs for the two samplers. It is seen that in MNGG, the marginal sampler generally obtains larger ESS values than the slice sampler; while it is opposite for HMNGG. Regarding 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.

6.5 Conclusion

This chapter proposes a more theoretically tractable dependent normalized random measure by operating directly on the underlying Poisson process. In the construction the Poisson process is defined on an augmented *spatial* space where each element in the *spatial space* is associated with a Poisson process. Key to the construction is to weight and superposition these Poisson process to get mixed normalized random measures (MNRM). Simple in the construction, MNRM maintains nice distributional properties as well as tractable posterior structure as a generalized Chinese restaurant process, thus posterior inference is easy. Meanwhile, the MNRM seems to be more flexible in dependency modeling by comparing with the HNRM because MNRM is not only designed for hierarchical modeling, but also is able to model other depen-

Table 6.3: 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
Models	ESS | Time	ESS | Time
MNGG	243.3/202.5/4.4|234s	252.4/231.9/3.7| 285s
MNGG_s	201.2/122.0/26.9|760s	205.1/131.9/23.5|813s
HMNGG	99.1/70.3/2.6|91s	171.5/80.4/5.1|176s
HMNGG_s	150.7/117.7/4.6|97s	194.3/180.9/6.5|227s
	Person	NIPS
Models	ESS | Time	ESS | Time
MNGG	402.5/401.4/1.5|1.5h	314.8/376.1/1.5|3.3h
MNGG_s	321.5/291.8/11.3|2.9h	228.4/110.6/2.2|2.2h
HMNGG	213.0/246.5/1.9|3.3h	282.1/198.2/4.3|9.4h
HMNGG_s	293.3/358.6/2.0|3.5h	346.1/467.2/1.7|10.4h

dencies such as the Markovian dependency, *e.g.*, the dependent operator used in the dynamic topic model in Chapter 5 can be replaced by the MNRM. One potential drawback is the dense representation of the dependent NRMs, which might not be favorable in real applications. This will be addressed in the next chapter.

Chapter 7

Thinned Normalized Random Measures

7.1 Introduction

The mixed normalized random measure proposed in the last chapter induces dependencies via sharing and weighting the entire points of the corresponding Poisson process. Though flexible in modeling and tractable in posterior computation, the MNRM induces a restriction in the construction: all the points in the same region will or will not be inherited to an NRM at time t. This property makes the resulting dependent NRMs in a dense representation, *e.g.*, the NRMs have a lot of atoms with small weights that might not be interesting to the problem. For example, in topic models, each document is represented by a topic distribution. The total number of topics might be large in the corpus, however, each document often contains just several topics, say 3 topics. In this case, we want the topic distribution vector for the document to be sparse, obviously using the MNRM to do the modeling is not a good choice. Another example is in the modeling of friendship in a social network, where the distribution of connections between a person and their friends is assumed to be modeled with an NRM. Again, in this case we do not want the person to connect to all the people in the network, thus dependent normalized random measures with a selection mechanism for the elements are desirable.

This chapter introduces another class of dependent normalized random measures called *thinned normalized random measures* (TNRM). It is constructed by independently thinning individual atoms of the Poisson process. The merits of TNRM over the MNRM include:

- by controlling individual atoms in the NRMs, it achieves a more flexible modeling of the dependency structure; and
- each resulting dependent NRM is sparsely represented, which is preferable for many applications.

Similarly to the MNRM, the construction of TNRM also results in some nice distributional properties. However, the flexibility of this construction is paid for with a more complicated posterior structure than the MNRM, which will be shown in this

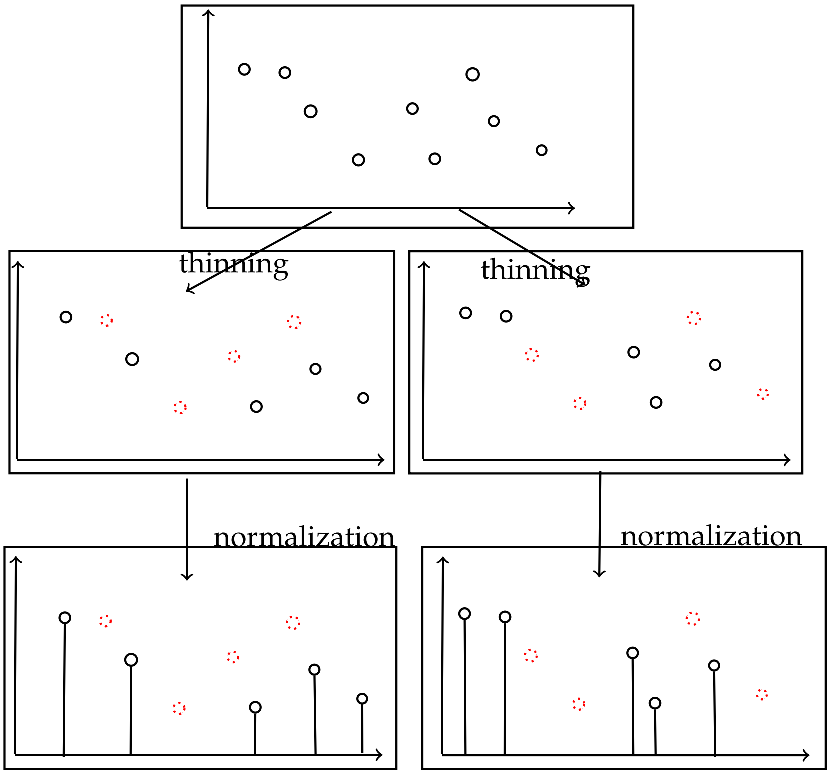


Figure 7.1: Construction of thinned normalized measures. The two NRMs in the bottom are dependent because they share some of the original atoms of the Poisson process from the top.

chapter. Fortunately, with advances of the MCMC, posterior inference can still be performed efficiently via a slice sampler.

7.2 Thinned Normalized Random Measures

As can be seen in a MNRM, a set of weights control the contribution of the independent CRMs to the dependent NRMs at any time, thus forming a 'softening' of the spatial normalized Gamma process [Rao and Teh, 2009] (where each of the CRMs is either present or absent). The thinned normalized random measure (TNRM) proposed in this chapter, however, is a different generalization in that rather than including or excluding entire CRMs, it chooses whether or not individual atoms in each of the CRMs are present in the NRM at a given time. This can be done by an operator called *thinning* or *subsampling*. More precisely, to each region-time pair (r, t) we associate a parameter q_{rt} taking values in $[0, 1]$.[1] 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). Figure 7.1 illustrates the construction of

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

two dependent NRMs by thinning a common Poisson process.

Formally, TNRMs can be constructed as follows:

Definition 7.1 (Thinned Normalized Random Measure)**.** Let a Poisson process on space $\mathcal{W} \otimes \Theta \otimes \mathcal{R}$ with Lévy measure $\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a)$. The thinned normalized random measure (TNRM) is defined by the following construction:

- For each region $\mathcal{R}_r$, define a completely random measure:

$$\tilde{\mu}_r(\mathrm{d}\theta) = \int_{\mathbb{R}^+ \times \mathcal{R}_r} w\mathcal{N}(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}a) \ .$$

- For each region-time pair (r, t), generate a countably infinite sequence of Bernoulli random variables:

$$z_{rtk} \sim \text{Bernoulli}(q_{rt}) \ .$$

- For each time t, construct a dependent completely random measure by thinning $\tilde{\mu}_r$'s:

$$\tilde{\mu}_t(\mathrm{d}\theta) = \sum_{r=1}^{\#\mathcal{R}} \sum_{k=1}^{\infty} z_{rtk} w_{rk} \delta_{\theta_{rk}} \ .$$

- Normalize the completely random measure:

$$\mu_t(\mathrm{d}\theta) = \frac{1}{Z_t} \tilde{\mu}_t(\mathrm{d}\theta) \ , \text{ where } Z_t = \tilde{\mu}_t(\Theta) \ . \tag{7.1}$$

Clearly the μ_t's can be shown to be marginally NRMs:

Theorem 7.1. *Conditioned on the set of q_{rt}'s, each random probability measure μ_t defined in (7.1) is marginally distributed as a normalized random measure with Lévy measure $\sum_r q_{rt} \nu_r(\mathrm{d}w, \mathrm{d}\theta)$.*

Proof. 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.

One approach for the proof is to follow the proof of Lemma 5.8 in Chapter 5. Here a simplified proof is given using the characteristic function of a CRM (3.4).

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 (Theorem 2.6), 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 3.2, the corresponding characteristic functional can be calculated as

$$\mathbb{E}\left[e^{\int_{\Theta} iu\tilde{\mu}_t(\mathrm{d}\theta)}\right] = \exp\left\{-\int_{\mathbb{R}^+\times\Theta\times\mathcal{R}\times\mathcal{B}}\left(1-e^{iuw}\right)\nu(\mathrm{d}w,\mathrm{d}\theta,\mathrm{d}a)\mathrm{d}z\right\}$$
$$= \exp\left\{-\int_{\mathbb{R}^+\times\Theta\times\mathcal{R}}\left(1-e^{iuw}\right)q_{r_a t}\nu(\mathrm{d}w,\mathrm{d}\theta,\mathrm{d}a)\right\} \qquad (7.2)$$
$$= \exp\left\{-\int_{\mathbb{R}^+\times\Theta}\left(1-e^{iuw}\right)\left(\sum_{r=1}^{\#\mathcal{R}} q_{rt}\nu_r(\mathrm{d}w,\mathrm{d}\theta)\right)\right\}, \qquad (7.3)$$

where (7.2) follows by integrating out the Bernoulli random variable z with parameter $q_{r_a t}$, (7.3) follows by integrating out the *region space*. According to the uniqueness property of the characteristic functional, μ_t's are marginally normalized random measure with Lévy measures $\sum_{r=1}^{\#\mathcal{R}} q_{rt}\nu_r(\mathrm{d}w,\mathrm{d}\theta)$. □

7.2.1 Comparision with related work

The idea of thinning atoms is similar to [Lin et al., 2010] for DPs and to [Chen et al., 2012b] for NGGs, but these were restricted to random probability measures with chain-structured dependency. In addition, posterior samplers developed in these prior works were approximate. The TNRM is also a generalization of a very recent work [Lin and Fisher, 2012]. This model is restricted to dependent DPs, and again, the proposed sampler has an incorrect equilibrium distribution (more details in Section 7.3). The TNRM is also related to a recently proposed model in [Foti et al., 2013], the main differences are: 1) they focus on different thinning constructions of dependent CRMs; the focus of TNRMs is on normalized random measures, where the normalization provides additional challenges. 2) Their posterior inference is approximated based on truncated representations of the CRMs (which are restricted only to Beta and Gamma CRMs), while TNRMs' is exact. 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)). TNRM makes 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, the TNRM can be contrasted with the Indian buffet process [Griffiths and 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 behavior; this is not possible in the IBP [Teh and Gorur, 2009; Broderick et al., 2012].

7.2.2 Interpretation as mixture of NRMs

The complication of the TNRM comes from the introduction of the latent selecting Bernoulli random variables z_{trk}'s. Incorrect posterior samplers are easily derived without carefully inspecting its posterior structure. This happens in [Lin et al., 2010] and [Lin and Fisher, 2012], detailed analysis can be found in [Chen et al., 2013b].

To reveal the posterior structure of a TNRM, a mixture of NRMs interpretation is first presented in this section, *e.g.*, it shows that μ_t is a mixture of NRMs that are formed by transforming the original $\tilde{\mu}_r$'s. To show this, associate the kth atom in a region $r \in \mathcal{R}$ with a binary vector $\mathbf{b}^r(k)$ of length T. Index the atoms in the common Poisson process by k, so $b^r_t(k) = 1$ means atom k is inherited by the NRM μ_t of time t (i.e. $z_{trk} = 1$). Accordingly, we can split each region $\mathcal{R}_r$ into 2^T smaller subregions, each associated with atoms with a particular configuration[2] of $\mathbf{b}^r$. That is, each configuration of $\mathbf{b}^r$ corresponds to a new sub-region, and there is totally 2^T configurations for each region $\mathcal{R}_r$. It is easy to see that with subregion $\mathbf{b}^r = b^r_1 \cdots b^r_t$ of region r, it associates a new 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 it is easy to see that

$$\hat{\mu}_t(\mathrm{d}\theta) = \sum_{r\in\mathcal{R}} \sum_{(\mathbf{b}\ s.t.\ b_t=1)} \tilde{G}_{r\mathbf{b}}(\mathrm{d}\theta) \tag{7.4}$$

$$\mu_t(\mathrm{d}\theta) = \frac{\hat{\mu}_t(\mathrm{d}\theta)}{\hat{\mu}_t(\Theta)} \tag{7.5}$$

Thus, the NRM at any time t can be expressed as a mixture of a number of NRMs defined as

$$G_{r\mathbf{b}}(\mathrm{d}\theta) = \tilde{G}_{r\mathbf{b}}(\mathrm{d}\theta)/\tilde{G}_{r\mathbf{b}}(\Theta) .$$

This number is exponential in the number of times T. The interpretation of a single thinned NRM is illustrated in Figure 7.2. We can also see from this interpretation that TNRMs can be seen as fixed-weight (binary) MNRMs but with many more regions (which is 2^T). The number of components also grows linearly with the number of regions $\#\mathcal{R}$; we will see that this flexibility improves the performance of the model without too great an increase in complexity.

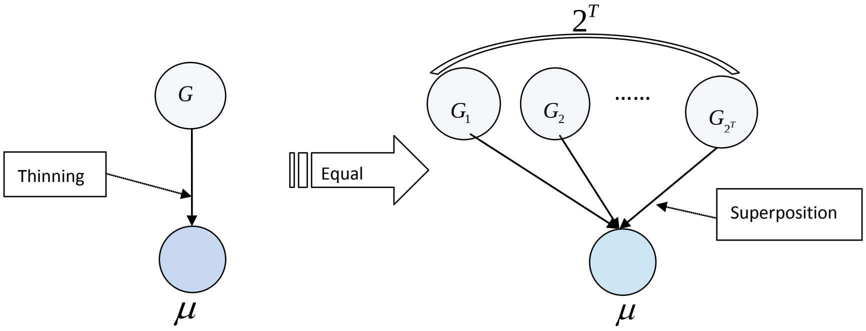


Figure 7.2: Interpretation of TNRM as a mixture of 2^T NRMs. The thinned version μ of the NRM G is equivalent to a mixture of 2^T NRMs G_i's, where each G_i represents the NRM corresponding to one configuration of the $\mathbf{b}$ defined in the text.

[2] Each configuration is an assignment of the vector $\mathbf{b}^r$, thus there are totally 2^T configurations.

7.3 Conditional Posterior of TNRM

One difficult problem with the TNRM is its conditional posterior distribution, *i.e.*, given observations from all μ_t's, what is the posterior Lévy measure of $\tilde{\mu}_r$ for the regions? One might argue that because μ_t is constructed by superpositioning thinned versions of the independent CRMs $\tilde{\mu}_r$, the posterior Lévy measure would also be the thinned Lévy measures of $\tilde{\mu}_r$'s, *i.e.*, $\nu_r^* = \left(\sum_t q_{rt}\right)\nu_r$. Unfortunately, this is not true. It actually ends up with a fairly complex Lévy measure that is beyond computational tractability, making the marginalization difficult and impractical. In the following theorem, as in the normalized random measure case, we need to condition on the *latent relative mass* auxiliary variable μ_t for each t, then we have

Theorem 7.2. *Given observations associated with atoms $W = \{(w_1, \theta_1), \cdots, (w_K, \theta_K)\}$ in region $\mathcal{R}_r$, and auxiliary variables u_t (the latent relative mass) for each $t \in \mathcal{T}$, the remaining atoms in the Poisson process in region $\mathcal{R}_r$ are independent of W, and are distributed as a CRM with Lévy measure*[3]

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

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 again to apply Theorem 2.12 of the Poisson process partition calculus, 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 CRMs 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, 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

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

[3] The auxiliary variables u_t's also have the corresponding posteriors, but are fairly complex. They can be obtained by inspecting the posterior (7.20) in the following sections with an accurate approximation given in Section 7.4.4.

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 Theorem 2.12, then by applying the theorem, we conclude that the Poisson process has posterior mean measure 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. □

Remark 7.3. Theorem 7.2 indicates that conditioned on observations, the remaining weights are distributed as a CRM from a different family than the original one. The marginal samplers in [Lin et al., 2010; Lin and Fisher, 2012] implicitly assume these are the same, and are incorrect. Please refer to [Chen et al., 2013b] for more details.

By looking at the posterior intensity of the Poisson process in region $\mathcal{R}_r$ in Theorem 7.2, we see that marginalization over this Poisson random measure is impractical for posterior inference. For example, the following theorem shows the complicated marginal posterior of the TNRM under a specific class of the normalized random measure – the normalized generalized Gamma process, denoted as TNGG.

Theorem 7.4. *Given observations* $\mathbf{X}$ *for all times, introduce a set of auxiliary variables* $\{u_t\}$*, the marginal posterior for the TNGG is given by*

$$\begin{aligned} & p(\boldsymbol{X}, \boldsymbol{u}, \{s_{tl}\}, \{g_{tl}\} | \sigma, \{M_r\}, \{z_{rtk}\}_{k:n_{\cdot rk}>0}, \{q_{rt}\}) && (7.6) \\ = \; & \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r} \left(\prod_r M_r^{K_r}\right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)}\right) \\ & \left(\prod_r \prod_{k:n_{\cdot rk}>0} \frac{\Gamma(n_{\cdot rk} - \sigma)}{(1 + \sum_t z_{rtk} u_t)^{n_{\cdot rk} - \sigma}}\right) \left(\prod_t \prod_l f(x_{tl} | \theta_{g_{tl} s_{tl}})\right) && (7.7) \\ & \prod_r \exp\left\{-M_r \left[\sum_{\substack{z'_{rt} \in \{0,1\} \\ for \; t = 1 \cdots T}} \left(\left(\prod_{t'} q_{rt'}^{z'_{rt'}} (1 - q_{rt'})^{1 - z'_{rt'}}\right) \left((1 + \sum_{t'} z'_{rt'} u_{t'})^{\sigma} - 1\right)\right)\right]\right\} \; , \end{aligned}$$

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 G_{rt} is a CRM in the augmented space $\mathbb{R}^+ \times \Theta \times \{0, 1\}$. Given the observed data, the likelihood is given by

$$\begin{aligned} & p(\boldsymbol{X}, \{s_{tl}\}, \{g_{tl}\} | \{G_{rt}\}) \\ = \quad & \frac{\prod_{t=1}^{T} \prod_{r=1}^{I} \prod_{k=1}^{K_r} w_{rk}^{n_{trk}}}{\prod_{t'=1}^{T} \left(\sum_{r'} \sum_{k'} z_{r't'k'} w_{r'k'}\right)^{N_{t'}}} \prod_{t=1}^{T} \prod_{l=1}^{L_t} f(x_{tl|\theta_{g_{tl}s_{tl}}}) \ , \end{aligned} \tag{7.8}$$

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

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

$$\begin{aligned} & p(\boldsymbol{X}, \boldsymbol{u}, \{s_{tl}\}, \{g_{tl}\} | \{G_{rt}\}) \\ = \quad & \left(\prod_{t=1}^{T} \prod_{r=1}^{I} \prod_{k=1}^{K_r} w_{rk}^{n_{trk}} \right) \left(\prod_{t=1}^{T} \prod_{l=1}^{L_t} f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\ & \left(\exp\left\{ - \sum_t \sum_r \sum_k z_{rtk} u_t w_{rk} \right\} \right) \end{aligned} \tag{7.9}$$

Denote $\Upsilon = \underbrace{\{0,1\} \otimes \cdots \otimes \{0,1\}}_{T}$, $\mathrm{d}\boldsymbol{R}_r = \mathrm{d}z_{r1} \cdots \mathrm{d}z_{rT}$, since $\{G_{rt}\}$'s are CRMs, now integrate out $\{G_r\}$'s with Lévy-Khintchine formula in Lemma 3.2 as well as the Poisson process partition calculus formula in Theorem 2.12 we have

$$\begin{aligned} & p(\boldsymbol{X}, \boldsymbol{u}, \{s_{tl}\}, \{g_{tl}\} | \sigma, \{M_r\}) = \mathbb{E}_{\{G_{rt}\}} \left[p(\boldsymbol{X}, \boldsymbol{u}, \{s_{tl}\}, \{g_{tl}\} | \{G_{rt}\}) \right] \\ = \quad & \left(\frac{\sigma}{\Gamma(1-\sigma)} \right)^{\sum_r K_r} \left(\prod_r M_r^{K_r} \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\ & \left(\prod_r \prod_{k: n_{\cdot rk} > 0} \frac{\Gamma(n_{\cdot rk} - \sigma)}{\left(1 + \sum_t z'_{rtk} u_t\right)^{n_{\cdot rk} - \sigma}} \right) \left(\prod_{t=1}^{T} \prod_{l=1}^{L_t} f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \\ & \prod_r \exp\left\{ - \frac{\sigma M_r}{\Gamma(1-\sigma)} \int_{\Upsilon} \int_{\Theta} \int_{\mathbb{R}^+} \left(1 - e^{-\sum_t z_{rtx} u_t x}\right) \frac{e^{-x}}{x^{1+\sigma}} \mathrm{d}x \mathrm{d}\theta \mathrm{d}\boldsymbol{R}_r \right\} (7.10) \\ \overset{\text{Taylor}}{\underset{\text{expansion}}{=}} \quad & \left(\frac{\sigma}{\Gamma(1-\sigma)} \right)^{\sum_r K_r} \left(\prod_r M_r^{K_r} \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\ & \left(\prod_r \prod_{k: n_{\cdot rk} > 0} \frac{\Gamma(n_{\cdot rk} - \sigma)}{\left(1 + \sum_t z'_{rtk} u_t\right)^{n_{\cdot rk} - \sigma}} \right) \left(\prod_{t=1}^{T} \prod_{l=1}^{L_t} f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \\ & \prod_r \exp\left\{ - \frac{\sigma M_r}{\Gamma(1-\sigma)} \int_{\Upsilon} \int_{\Theta} \int_{\mathbb{R}^+} \sum_{n=1}^{\infty} (-1)^{n-1} \frac{\left(\sum_t z_{rtx} u_t\right)^n x^n}{n!} \frac{e^{-x}}{x^{1+\sigma}} \mathrm{d}x \mathrm{d}\theta \mathrm{d}\boldsymbol{R}_r \right\} \\ \overset{\text{Integrate out}}{\underset{\text{all } z_{rtx}}{=}} \quad & \left(\frac{\sigma}{\Gamma(1-\sigma)} \right)^{\sum_r K_r} \left(\prod_r M_r^{K_r} \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \end{aligned}$$

$$
\begin{aligned}
& \left(\prod_r \prod_{k:n_{\cdot rk}>0} \frac{\Gamma(n_{\cdot rk}-\sigma)}{\left(1+\sum_t z'_{rtk}u_t\right)^{n_{\cdot rk}-\sigma}}\right)\left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \\
& \prod_r \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)}\left[\sum_{\substack{z'_{rt}\in\{0,1\}\\ \text{for } t=1\cdots T}} \sum_{n=1}^{\infty}(-1)^{n-1}\frac{\left(\sum_{t'} z'_{rt'}u_{t'}\right)^n}{n!}\right.\right. \\
& \qquad \left.\left.\left(\prod_{t'} q_{rt'}^{z'_{rt'}}(1-q_{rt'})^{1-z'_{rt'}}\int_{\mathbb{R}^+} x^{n-\sigma-1}e^{-x}\mathrm{d}x\right)\right]\right\} \\
= \quad & \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r}\left(\prod_r M_r^{K_r}\right)\left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \left(\prod_r \prod_{k:n_{\cdot tk}>0} \frac{\Gamma(n_{\cdot rk}-\sigma)}{\left(1+\sum_t z'_{rtk}u_t\right)^{n_{\cdot rk}-\sigma}}\right)\left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \qquad (7.11) \\
& \prod_r \exp\left\{-M_r\left[\sum_{\substack{z'_{rt}\in\{0,1\}\\ \text{for } t=1\cdots T}}\left(\left(\prod_{t'} q_{rt'}^{z'_{rt'}}(1-q_{rt'})^{1-z'_{rt'}}\right)\left((1+\sum_{t'} z'_{rt'}u_{t'})^{\sigma}-1\right)\right)\right]\right\}
\end{aligned}
$$

where z_{rtx} in (7.10) means a Bernoulli random variable drawn at atom x with parameter q_{rt}. The last equation follows by applying the following result

$$
\begin{aligned}
& \sum_{n=1}^{\infty}(-1)^{n-1}\frac{\lambda^n}{n!}\Gamma(n-\sigma) \\
= \quad & \sum_{n=1}^{\infty}(-1)^{n-1}\lambda^n\frac{\Gamma(n-\sigma)}{n!} \\
= \quad & \frac{1}{\sigma}\left(\sum_{n=1}^{\infty}\frac{(-1)^{n-1}\sigma\Gamma(n-\sigma)}{n!}\lambda^n\right) \\
= \quad & \frac{\Gamma(1-\sigma)}{\sigma}\left(\sum_{n=1}^{\infty}\frac{\sigma(\sigma-1)\cdots(\sigma-n+1)}{n!}\lambda^n\right) \qquad (7.12) \\
= \quad & \frac{\Gamma(1-\sigma)}{\sigma}\left[(1+\lambda)^{\sigma}-1\right] ,
\end{aligned}
$$

where the summation in (7.12) is the Taylor expansion of $(1+\lambda)^{\sigma}-1$. □

■

7.4 Posterior Inference

Without loss of generality, this section shows how to do posterior inference on a specific class of the TNRM – thinned normalized generalize Gamma process (TNGG), with both marginal sampler and slice sampler.

7.4.1 Marginal posterior for thinned normalized generalized Gamma processes

This section proposes a marginal sampler for TNGG based on the marginal posterior (7.6). To achieve full marginalization and sample the topic allocation variables (s_{tl}, g_{tl}), the Bernoulli random variables z_{rtk}'s for the fixed jumps in (7.7) need to be further integrated out. So the terms in the first parenthesis of (7.7) is first augmented by instantiating a set of jump size variables w_{rk}'s distributed as

$$w_{rk} \sim \text{Gamma}\left(n_{\cdot rk} - \sigma, 1 + \sum_t z_{rtk} u_t\right) . \tag{7.13}$$

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 (7.7) 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 simplifies to:

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

where $\tilde{t} = 1$ when $t = 2$; and $\tilde{t} = 2$ when $t = 1$; and similar to previous chapter

$$f_{rk}^{\backslash tl}(x_{tl}) = \frac{\int f(x_{tl}|\theta_{rk}) \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} f(x_{t'l'}|\theta_{rk}) h(\theta_{rk}) \mathrm{d}\theta_{rk}}{\int \prod_{t'l' \neq tl, s_{t'l'}=k, g_{t'l'}=r} f(x_{t'l'}|\theta_{rk}) h(\theta_{rk}) \mathrm{d}\theta_{rk}}$$

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 $(\{u_t\}, \{q_{rt}\}, \sigma)$, the fixed jumps w_{rk} are first instantiated as in (7.13), then the latent Bernoulli variables z_{rtk} for $(k : n_{\cdot rk} > 0)$ can be sampled using the following rule

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

Sampling for other parameters can also be read from the posterior:

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

$$p(u_t | C - u_t) \propto u_t^{N_t - 1} e^{-\left(\sum_r \sum_{k: n_{\cdot rk} > 0} z_{rtk} w_{rk}\right) u_t} e^{-\sum_r M_r \sum_{\mathbf{b}} \left(\prod_{t'} q_{rt'}^{b_{t'}} (1-q_{rt'})^{1-b_{t'}}\right)(1+<\mathbf{u},\mathbf{b}>)^\sigma)} , \tag{7.14}$$

this is log-concave after using a change of variable $v_t = \log(u_t)$. Another possible way for the sampling is to note that the posterior of u_t above is bounded by the first two terms, which is a Gamma distribution. Thus we can first sample u_t from a Gamma distribution: $u_t \sim \text{Gamma}\left(N_t, \sum_r \sum_{k: n_{\cdot rk} > 0} z_{rtk} w_{rk}\right)$, then use a rejection step evaluated on the true posterior (7.14), though the acceptance rate would probably be low.

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

$$p(q_{rt}|C - q_{rt}) \propto q_{rt}^{\sum_{k:n_{\cdot tk}>0} 1(z_{rtk}=1)+a_q-1} \, (1-q_{rt})^{\sum_{k:n_{\cdot tk}>0} 1(z_{rtk}=0)+b_q-1} \tag{7.15}$$

$$e^{-M_r \sum_{\mathbf{b}} \left(\prod_{t'} q_{rt'}^{b_{t'}} (1-q_{rt'})^{1-b_{t'}}\right) ((1+<\mathbf{u},\mathbf{b}>)^{\sigma}-1)} , \tag{7.16}$$

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

$$\text{Beta}\left(\sum_{k:n_{\cdot tk}>0} 1(z_{rtk}=1)+a_q, \sum_{k:n_{\cdot tk}>0} 1(z_{rtk}=0)+b_q\right)$$

proposal distribution and do a rejection step based on the true posterior (7.15).

Sample σ: From (7.6), 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 and Walker, 2011; Favaro and Teh, 2013], a slice sampler for TNGG is developed in the next section which greatly reduces the computational cost.

7.4.2 Posterior inference for the TNGG via slice sampling

This section describes a slice sampler for the thinned normalized generalized Gamma process (TNGG). The idea behind the slice sampler has been described in Chapter 6, which basically is to stochastically truncate the infinite atoms in the model to finite ones so that the model is computationally manageable. Note the derivation here is similar in part to the MNRM, but different in that there are extra Bernoulli random variables z_{rtk}'s to be handled.

Specifically, as in last chapter, let's 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 (7.9), now the joint likelihood of observations and related auxiliary variables becomes

$$\begin{aligned}
& p(\boldsymbol{X}, \boldsymbol{u}, \{\boldsymbol{v_{tl}}\}, \{s_{tl}\}, \{g_{tl}\} | \{\mu_r\}, \{z_{rtk}\}, \{q_{rt}\}) \\
= & \left(\prod_t \prod_l 1\left(w_{g_{tl}s_{tl}} > v_{tl} \right) f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\
& \left(\exp\left\{ -\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk} \right\} \right) \qquad (7.17)
\end{aligned}$$

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

$$\begin{aligned}
& p(\boldsymbol{X}, \boldsymbol{u}, \{\boldsymbol{v_{tl}}\}, \{\mu_r\}, \{s_{tl}\}, \{g_{tl}\} | \{z_{rtk}\}, \{q_{rt}\}) \\
= & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\
& \left(\exp\left\{ -\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk} \right\} \right) \prod_r P(\mathcal{N}_r) \\
\overset{\text{slice at } \mathcal{L}_r}{=} & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\
& \underbrace{\exp\left\{ -\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk} \right\}}_{\text{jumps larger than } \mathcal{L}_r} \\
& \prod_r p(\{(w_{r1}, \theta_{r1})\}, \cdots, \{(w_{rK'_r}, \theta_{rK'_r})\}) \qquad (K'_r \text{ is \# jumps larger than } \mathcal{L}_r) \\
& \underbrace{\prod_r \exp\left\{ -\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_0^{\mathcal{L}_r} \left(1 - \prod_t \left(1 - q_{rt} + q_{rt} e^{-u_t x} \right) \right) \rho'(\mathrm{d}x) \right\}}_{\text{jumps less than } \mathcal{L}_r\text{, according to Theorem 7.2}} \qquad (7.18) \\
\overset{\text{small } \mathcal{L}_r}{=} & \left(\prod_t \prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) f(x_{tl} | \theta_{g_{tl}s_{tl}}) \right) \left(\prod_t \frac{u_t^{N_t - 1}}{\Gamma(N_t)} \right) \\
& \underbrace{\exp\left\{ -\sum_t \sum_r \sum_k z_{rtk} u_t w_{rk} \right\}}_{\text{jumps larger than } \mathcal{L}_r} \\
& \underbrace{\prod_r \left(\frac{\sigma M_r}{\Gamma(1-\sigma)} \right)^{K'_r} \exp\left\{ -\frac{\sigma M_r}{\Gamma(1-\sigma)} \int_{\mathcal{L}_r}^{\infty} \rho'(\mathrm{d}x) \right\} \prod_k w_{rk}^{-1-\sigma} e^{-w_{rk}}}_{p(\{(w_{1k}, \theta_{1k})\}, \{(w_{2k}, \theta_{2k})\}, \cdots, \{(w_{Ik}, \theta_{Ik})\})} \qquad (7.19)
\end{aligned}$$

$$\underbrace{\prod_r \exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)}\int_0^{\mathcal{L}_r}\left((\sum_j q_{rt}u_t)x + O((u_t x)^2)\right)\rho'(\mathrm{d}x)\right\}}_{\text{jumps less than }\mathcal{L}_r} \tag{7.20}$$

$$\begin{aligned}
\approx \quad & \left(\prod_t\prod_l 1(w_{g_{tl}s_{tl}} > v_{tl}) f(x_{tl}|\theta_{g_{tl}s_{tl}})\right)\left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right)\underbrace{\exp\left\{-\sum_t\sum_r\sum_k z_{rtk}u_t w_{rk}\right\}}_{\text{jumps larger than }\mathcal{L}_r} \\
& \underbrace{\prod_r\left(\frac{\sigma M_r}{\Gamma(1-\sigma)}\right)^{K'_r}\exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)}\int_{\mathcal{L}_r}^{\infty}\rho'(\mathrm{d}x)\right\}\prod_k w_{rk}^{-1-\sigma}e^{-w_{rk}}}_{p(\{(w_{1k},\theta_{1k})\},\{(w_{2k},\theta_{2k})\},\cdots,\{(w_{Ik},\theta_{Ik})\})} \\
& \underbrace{\exp\left\{-\sum_r(\sum_t q_{rt}u_t)M_r\frac{\sigma\mathcal{L}_r^{1-\sigma}}{(1-\sigma)\Gamma(1-\sigma)}\right\}}_{\text{jumps less than }\mathcal{L}_r}\ ,
\end{aligned} \tag{7.21}$$

where (7.19) 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

$$\begin{aligned}
& p(\boldsymbol{X},\boldsymbol{u},\{\boldsymbol{v_{tl}}\},\{w_{rk}\},\{s_{tl}\},\{g_{tl}\}|\sigma,\{M_r\}) \\
\approx \quad & \left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} 1(w_{g_{tl}s_{tl}} > v_{tl}) f(x_{tl}|\theta_{g_{tl}s_{tl}})\right)\left(\prod_t \frac{u_t^{N_t-1}}{\Gamma(N_t)}\right) \\
& \underbrace{\left(\prod_t\prod_r\prod_{k:n_{trk=0}}(1-q_{rt}+q_{rt}e^{-u_t w_{rk}})\prod_{k:n_{trk>0}}e^{-u_t w_{rk}}\right)}_{\text{jumps larger than }\mathcal{L}_r} \\
& \underbrace{\prod_r\left(\frac{\sigma M_r}{\Gamma(1-\sigma)}\right)^{K'_r}\exp\left\{-\frac{\sigma M_r}{\Gamma(1-\sigma)}\int_{\mathcal{L}_r}^{\infty}\rho'(\mathrm{d}x)\right\}\prod_k w_{rk}^{-1-\sigma}e^{-w_{rk}}}_{p(\{(w_{1k},\theta_{1k})\},\{(w_{2k},\theta_{2k})\},\cdots,\{(w_{Ik},\theta_{Ik})\})} \\
& \underbrace{\exp\left\{-\sum_r(\sum_t q_{rt}u_t)M_r\frac{\sigma\mathcal{L}_r^{1-\sigma}}{(1-\sigma)\Gamma(1-\sigma)}\right\}}_{\text{jumps less than }\mathcal{L}_r}
\end{aligned} \tag{7.22}$$

7.4.3 Bound analysis

Note that in the above derivation, to make the inference feasible, a linear approximation for an exponential function in (7.18) is used to make it become (7.21). For the interest of theoretical analysis, this section derives the upper bound and lower bound of the true posterior, which are shown in (7.24) and (7.25) respectively, allowing the integration to be worked out. Actually, the only approximation used is by replacing the term $e^{-u_t w}$ with its linear upper bound and lower bound as in (7.23). 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 [4].

To derive the lower bound and upper bound of the true posterior (7.18), first define the following notation:

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

$$t^r_{max} = \arg\max_{t}\{q_{rt}u_t\}.$$

Also denote the last term in (7.18) as $\tilde{Q}_r(\mathcal{L}_r)$, *i.e.*,

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

Use the following inequality:

$$1-u_t x \le e^{-u_t x} \le 1-\frac{1-e^{-u_t L}}{L}x, \qquad \forall L \ge x. \tag{7.23}$$

the upper bound for $\tilde{Q}_r(\mathcal{L}_r)$ is given as:

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

[4] It is chose as $\mathcal{L}_r = \min\left\{0.001/\max_t\{u_t\}, \min_{(t,l):g_{tl}=r}\{v_{tl}\}\right\}$ in the experiments.

$$= \exp\left\{ -\frac{\sigma M_r}{\Gamma(1-\sigma)} \left(\frac{q_{rt^r_{min}}(1 - e^{-u_{t^r_{min}} \mathcal{L}_r})}{\mathcal{L}_r} \right)^{T/2} \left(2 - q_{rt^r_{min}}(1 - e^{-u_{t^r_{min}} \mathcal{L}_r}) \right)^{T/2} \right.$$
$$\left. \left(\frac{2}{T - 2\sigma} - \frac{2\mathcal{L}_r}{T - 2\sigma + 2} \right) \mathcal{L}_r^{\frac{T}{2} - \sigma} \right\} . \tag{7.24}$$

Similarly, the lower bound is given as:

$$\begin{aligned} & \tilde{Q}_r(\mathcal{L}_r) \geq \\ & \exp\left\{ - \int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)} \left(1 - \prod_t (1 - q_{rt} u_t x) \right) \left(x^{-\sigma-1} - \frac{1 - e^{-\mathcal{L}_r}}{\mathcal{L}_r} x^{-\sigma} \right) \mathrm{d}x \right\} \\ \geq & \exp\left\{ - \int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)} \left(1 - (1 - q_{rt^r_{max}} u_{t^r_{max}} x)^T \right) \left(x^{-\sigma-1} - \frac{1 - e^{-\mathcal{L}_r}}{\mathcal{L}_r} x^{-\sigma} \right) \mathrm{d}x \right\} \\ \geq & \exp\left\{ - \int_0^{\mathcal{L}_r} \frac{\sigma M_r}{\Gamma(1-\sigma)} 2^{T/2} \left(q_{rt^r_{max}} u_{t^r_{max}} \right)^{T/2} x^{T/2} \left(x^{-\sigma-1} - \frac{1 - e^{-\mathcal{L}_r}}{\mathcal{L}_r} x^{-\sigma} \right) \mathrm{d}x \right\} \\ = & \exp\left\{ - \frac{\sigma M_r}{\Gamma(1-\sigma)} \left(q_{rt^r_{max}} u_{t^t_{max}} \right)^{T/2} 2^{T/2} \left(\frac{2}{T - 2\sigma} - \frac{2(1 - e^{-\mathcal{L}_r})}{T - 2\sigma + 2} \right) \mathcal{L}_r^{\frac{T}{2} - \sigma} \right\} . \end{aligned} \tag{7.25}$$

7.4.4 Sampling

Now the sampling is straightforward by inspecting the posterior (7.21). First note 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. The whole set is denoted 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:

$$p(s_{tl} = k, g_{tl} = r | C - \{s_{tl}, g_{tl}\}) \propto 1(w_{rk} > v_{tl}) 1(z_{rtk} = 1) f(x_{tl} | \theta_{g_{tl} s_{tl}}) . \tag{7.26}$$

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

$$v_{tl} | C - v_{tl} \sim \text{Uniform}(0, w_{g_{tl} s_{tl}}) . \tag{7.27}$$

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

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

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

- **Sample w_{rk}'s without observations:** We already know that these w_{rk}'s are Poisson points in a Poisson process, and from Theorem 7.2 we know the mean measure 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 $\nu_r(\mathrm{d}w, \mathrm{d}\theta) = \rho(\mathrm{d}w)H(\mathrm{d}\theta)$ is the Lévy measure of μ_r in region $\mathcal{R}_r$. 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 and Teh, 2013], as it does not require any numerical integrations but only the evaluation of the mean measure $\rho(\mathrm{d}w)$. 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 mean measure $\gamma_x(s)$ 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 $\int_x^\infty \gamma_x(s')\mathrm{d}s' < \infty$ to facilitate the computation. Then the samples from the Poisson process with mean measure $\rho(\mathrm{d}w)$ can be obtained by adaptively thinning some of the instantiated points in the Poisson process with mean measure $\gamma_x(s)$. For TNGG, the following adaptive mean measure is found to be a good one:

$$\gamma_x(s) = \frac{\sigma M_r}{\Gamma(1-\sigma)} \prod_t \left(1 - q_{rt} + q_{rt}e^{-u_t x}\right) e^{-s} x^{-1-\sigma} \tag{7.28}$$

Clearly it satisfies

$$\gamma_x(s) \geq \rho(s), \qquad x \leq s,$$
$$\gamma_x(s) \geq \gamma_{x'}(s), \qquad x' \geq x \; ,$$

Furthermore, we have:

$$\begin{aligned} W_x(s) \;\; &:= \;\; \int_x^s \gamma_x(s')\mathrm{d}s' = \int_x^s \frac{\sigma M_r}{\Gamma(1-\sigma)} \prod_x \left(1 - q_{rt} + q_{rt}e^{-u_t x}\right) e^{-s'} x^{-1-\sigma} \mathrm{d}s' \\ &= \;\; \frac{M_r}{\Gamma(1-\sigma)} \prod_t \left(1 - q_{rt} + q_{rt}e^{-u_t x}\right) x^{-1-\sigma} \left(e^{-x} - e^{-s}\right) \end{aligned} \tag{7.29}$$

So

$$W_x^{-1}(y) = x - \log\left(1 - \frac{\Gamma(1-\sigma)x^{1+\sigma}e^x}{\sigma M_r \prod_t \left(1 - q_{rt} + q_{rt}e^{-u_j x}\right)} y\right) .$$

Finally, the sampling for these w_{ik}'s goes as in Algorithm 1:

Algorithm 1 Simulate inhomogeneous Poisson process with mean measure $\rho(\mathrm{d}s)$ on $[L, \infty]$

```
$N := 0, x := L$;
repeat
  let $e$ be a draw from an Exponential random variable with parameter 1;
  if $e > W_x(\infty)$ then
    terminate;
  else
    set $x' := W_x^{-1}(e)$;
  end if
  with probability $\rho(x')/\gamma_x(x')$ accept sample, and set $N := N + \delta_{x'}$;
  set $x := x'$ and continue to next iteration;
until termination
return $N$ as a draw from the Poisson random measure with mean $\rho$ on $[L, \infty]$.
```

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 (7.8), 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 (7.21) and (7.22):

- **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 $\mathcal{L}_r$, and the integral can be evaluated using numerical evaluation or the incomplete Gamma function described in Theorem 3.15.

- **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:

$$p(q_{rt}|C - q_{rt}) \propto \prod_{k:n_{trk=0}} \left(1 - q_{rt} + q_{rt} e^{-u_t w_{rk}}\right) e^{-\frac{\sigma M_r u_t \mathcal{L}_r^{1-\sigma}}{(1-\sigma)\Gamma(1-\sigma)} q_{rt}} , \qquad (7.30)$$

which is log-concave. Now if we start from the construction, and 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 (7.21), 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 (7.20) 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 and 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 [5], *i.e.*,

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

To sample M, we use the proposal distribution

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

[5] 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}.

the acceptance rate is:

$$\begin{aligned} A &= \min\left(1, \frac{p(M^*, J^*, X)Q(M, J|M^*, J^*)}{p(M, J, X)Q(M^*, J^*|M, J)}\right) \\ &= \min\left(1, \frac{p(M^*, J^*, X)Q(M|M^*)p(J|M)}{p(M, J, X)Q(M^*|M)p(J^*|M^*)}\right) \\ &= \min\left(1, \frac{p(M^*)Q(M|M^*)p(X|M^*, J^*)}{p(M)Q(M^*|M)p(X|M, J)}\right) \qquad (7.31) \end{aligned}$$

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 (7.21), which is unbiased with respective to the random points w_{rk}'s, and also according to the bound analysis in Section 7.4.3, 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, thus this method is not a good choice in term of computational cost. This procedure goes as in Algorithm 2.

Algorithm 2 PMMH sampling for M_r and u_t

```
repeat
    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 Algorithm 1.
    Sample the Bernoulli variables $z_{rtk}$'s
    Use these jumps and $z_{rtk}$'s to evaluate the approximated likelihood (7.21).
    Propose a move
        $M_r^* \sim Q_M(M_r^*|M_r)$,
        $u_t \sim Q_u(u_t^*|u_t)$ , and
        $q_{rt} \sim Q_q(q_{rt}^*|q_{rt})$ .
    Use this state to simulate the jumps larger than $\mathcal{L}_r$ from a Poisson process with Algorithm 1.
    Use these jumps to evaluate the approximated likelihood (7.21).
    Do the accept-reject step using (7.31).
until converged
```

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

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

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

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

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 (7.21) with the current set of parameters $\{\{M_r\}, \{u_t\}, \{q_{rt}\}, \{w_{rk}\}\}$.

The above description completes the sampling procedure for the TNGG.

7.5 Experiments

Similar to MNRM in Chapter 6, two settings are designed in the experiments to test the TNGG, a specific class of TNRM: 1) *thinned normalized generalized Gamma process* (TNGG), 2) *hierarchical thinned normalized generalized Gamma process* (HTNGG). The first is simply the thinned construction without hierarchy. Letting G be the base distribution, *e.g.*, Dirichlet distribution in topic modeling while Normal-Wishart in Gaussian mixtures, F the likelihood, *e.g.*, multinomial distribution in topic modeling and Gaussian distribution in Gaussian mixtures, and there are t-time observations in the corpus, the generative process for TNGG is as follows:

$$(\mu_t) \sim \text{TNGG}(\sigma_0, M_0, G, \{q_{rt}\}) \tag{7.32}$$

$$\theta_i^t \sim \mu_t, \qquad x_i^t | \theta_i^t \sim F(\cdot|\theta_i^t) \ , \tag{7.33}$$

where TNGG$(\sigma_0, M_0, G, \{q_{rt}\})$ denotes the dependent NGG constructed via TNGG with index parameter σ, mass parameter M_0, base distribution G and the set of subsampling rates $\{q_{rt}\}$.

The second model HTNGG is constructed similarly as HMNGG in Chapter 6, where the TNRM construction is used to produce a random probability measure (RPM) μ_t for each time-period t; then data in each time period t are generated from an NGG mixture with base-measure μ_t:

$$\begin{aligned} (\mu_t)|\sigma_0, M_0, G, \{q_{rt}\} &\sim \text{TNGG}(\sigma_0, M_0, G, \{q_{rt}\}) \\ \{\mu_{ti}\}|\mu_t &\sim \text{NGG}(\sigma, M, \mu_t) \\ \theta_{ij}^t \sim \mu_{ti}, \quad x_{ij}^t|\theta_{ij}^t &\sim F(\cdot|\theta_{ij}^t) \ , \end{aligned} \tag{7.34}$$

7.5.1 Illustration

First, a demonstration on a thinned Gaussian mixture dataset is given to illustrate the power of TNGG defined above. In this experiment, a mixture of Gaussians dataset is generated with 10 components, each component has a covariance matrix $0.5 \times \mathbf{I}$ and a mean listed in the Table, where $\mathbf{I}$ is the identity matrix. Then a 3-epoch dataset is generated from these Gaussian mixtures. Assume 2 regions are used in this experiment. To generate the dataset, in each epoch t, a uniform random variable is assigned for each (t, r) pair, indicating the probability of choosing a Gaussian component from region r to epoch t. If region r is chosen, one data point is drawn from the corresponding Gaussian component with an added Gaussian noise of covariance $0.5 \times \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This repeats for 70 times to generate 70 points for each epoch, resulting in a 3-epoch dataset, with a total of 210 data points. It is clear that the data in each epoch is a thinned version of the original mixture of Gaussian data, thus fits well with the TNGG setting.

Now TNGG with $\sigma = 0.1$ is then run on this dataset. The hyperparameters for the *Normal-Inverse Wishart* (please refer to definition in Section 3.6) are set to $r = 0.1, \nu = 5, \boldsymbol{m} = [0, 0]^T, \mathbf{S} = 0.1 \times \mathbf{I}$. The number of iterations is set to 200. The result is shown in Figure 7.3, where we see that TNGG successfully recovers the Gaussian components and places it in the 2 regions. More detailed comparison of the true mean and estimated mean is shown in Table 7.1 where it is shown that most of the Gaussian components have been successfully recovered.

7.5.2 On multinomial likelihood synthetic data

This section tests the modeling power of HTNGG on the same synthetic data as for HMNGG in Chapter 6. As a reminder, the data is generated from a hierarchical

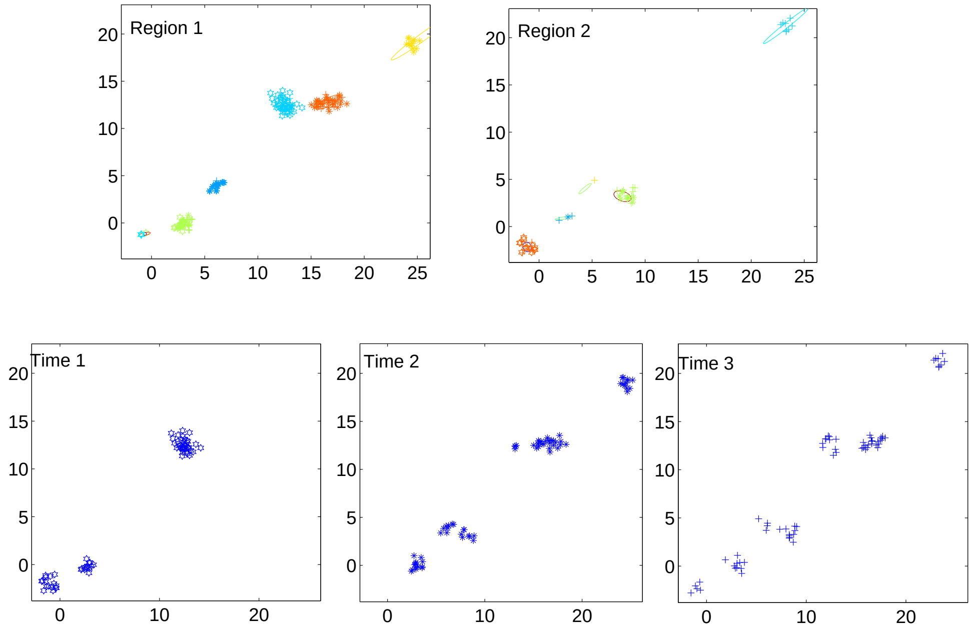


Figure 7.3: TNGG on three mixture of Gaussian synthetic dataset. The top row corresponds to the estimated Gaussian components in the 2 regions, the bottom row corresponds to the 3-epoch dataset.

Pitman-Yor process which contains 3000 data points from 3 groups. In HTNGG, a $\text{Beta}(0.5, 0.5)$ prior is put on the subsampling rate parameters q_{rt}'s. Other parameters and settings are exactly the same as for HMNGG. HTNGG is compared with its non-power-law version– the hierarchical thinned normalized Gamma process (HTNGP), as well as the HMNGG in Chapter 6. Figure 7.4 plots the testing likelihoods along with the number of iterations for all the models. It can be seen that both HMNGG and HTNGG outperform their non-power-law variants HMNGP and HTNGP in terms of predictive likelihoods. Furthermore, HTNGG gets higher likelihoods than HMNGG in this case; this follows from the added flexibility afforded by allowing the thinning of individual atoms[6].

7.5.3 Topic modelling

This section applies the TNGG and HTNGG to topic modeling. For comparison, the same datasets in the experiments of MNRM in Chapter 6 are used, which contains the ICML, TPAMI, Person and NIPS datasets. The parameter setting is also the same as MNRM, where the number of regions is set to be 20 for TNGG, and to the number of years for the HTNGG. The Dirichlet base distribution G is symmetric with concentration parameter of 0.3, $\text{Gamma}(0.1, 0.1)$ prior is placed on all unspecified scalar random variables (*e.g.*, the mass parameter $\{M_j\}$).

[6] We will see that this might not be always true in large real datasets in the following due to the complex posterior structure of the TNRM.

Table 7.1: 10 true and estimated Gaussian means for the dataset. The data from *C_6* and *C_7* seem to mix together due to the noise added, thus they are estimated to be generated only from one component.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10
true	-1.08	3.06	8.21	6.08	12.82	17.36	15.93	12.42	24.51	23.39
	-2.00	-0.03	3.25	3.94	12.06	12.97	12.82	12.98	19.27	21.49
estimated	-1.12	2.90	7.87	6.01	12.43	16.59		12.43	24.58	23.29
	-2.16	-0.15	3.21	3.92	12.43	12.79		12.43	19.74	21.28

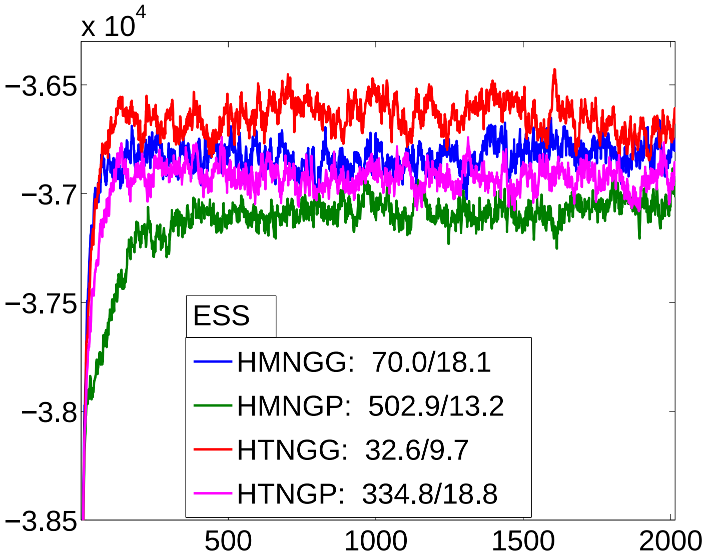


Figure 7.4: HMNGG VS. HMNGP.

In the experiments, 20% of the original data sets is held-out, following the standard dictionary hold-out method (50% of the held-out documents is used to estimate topic probabilities) [Rosen-Zvi et al., 2004]. All the experiments are repeated 10 times with random initializations, mean values and standard deviations of the results are reported. In each run, 2000 cycles are used as burn-in, followed by 1000 cycles to collect samples for the perplexity calculation.

Quantitative comparison for different models Both training and test perplexities are calculated for all the models, including those used in Chapter 6 for comparison. The results are shown in Table 7.2.

In addition to the findings in Chapter 6, it is found that while HTNGG is more flexible than HMNGG, its performances are sightly worse when the datasets becomes large; this is more obvious when comparing TNGG with MNGG. Part of the reason for this is the complex posterior structure for the thinned models, so that the samplers

Table 7.2: Train perplexities and test perplexities for different models on ICML, TPAMI, Person and NIPS datasets.

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

might often be stuck in local optima, resulting in much worse perplexities.

Topic evolution Figure 7.5 is a posterior sample, showing the evolution of 12 randomly selected topics on the NIPS dataset for HTNGG. The figure shows 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, which is defined to be proportional to $q_{rt}w_{rk}$ (see equation 7.26) by integrating out v_{tl} and z_{rtk}. Comparison with the HMNGG in Figure 6.5, 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 Next the performance of the marginal and slice samplers are compared for TNGG and HTNGG. The marginal sampler for TNGG could not handle datasets with more than even 2 times. Instead, we have 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. It is emphasized that this is done only for a comparison with the slice sampler, which can handle more complex datasets. Table 7.3 shows the average effective sample sizes and running times

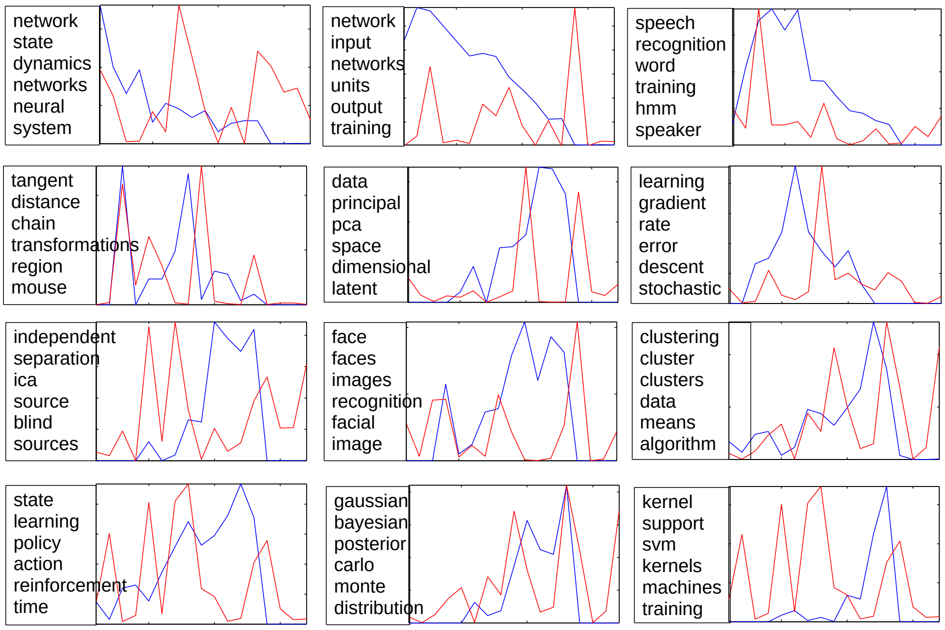


Figure 7.5: Topic evolution on NIPS dataset for 12 randomly chosen topics learned by HTNGG. The two curves give word proportions within each topic (blue) and prediction probabilities (red) for each time. The X axis represents the years from 1988 to 2004 and the Y axis represents the topic proportion and predictive probabilities.

over 5 repeated runs for the two samplers on the original datasets and the 2-time datasets. On the original datasets, the running time of the marginal sampler is more efficient in small datasets (*i.e.*, ICML and TPAMI), while they are comparable in the other datasets. The reason has been analyzed in Chapter 6. In the 2-time datasets, it is observed that the slice sampler obtains larger ESS values than its marginal sampler in 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.

7.6 Conclusion

This chapter proposes the *thinned normalized random measure* (TNRM) to address the issue of dense representation of the mixed normalized random measure in the last chapter. The construction involves thinning independent Poisson processes from different region before combining and normalizing them. As for the MNRM, two different MCMC algorithms for posterior inference are developed, a marginal sampler and an approximate slice sampler. However, it seems only the slice sampler is applicable in real applications because of the complexity of the marginal posterior structure. In the experiments of topic modeling, HTNRM shows significantly superior performance compared to related dependent nonparametric models such as

Table 7.3: 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
Models	ESS | Time	ESS | Time
TNGG$_s$	115.2/90.0/4.5|555s	135.7/113.0/11.1|592s
HTNGG$_s$	82.8/80.1/4.7|126s	92.5/105.1/5.4|312s
	Person	NIPS
Models	ESS | Time	ESS | Time
TNGG$_s$	300.6/231.3/3.2|3.3h	223.8/107.7/1.1|1.4h
HTNGG$_s$	184.9/226.3/6.1|4.1h	225.4/210.2/3.4|11.9h
	ICML$_2$	TPAMI$_2$
HTNGG	50.3/46.9/3.0|71s	55.3/58.4/4.3|95s
HTNGG$_s$	94.9/90.9/4.0|76s	116.0/107.8/3.4|106s
	Person$_2$	NIPS$_2$
HTNGG	144.8/170.6/4.2|1.3h	119.1/130.0/2.8|2.3h
HTNGG$_s$	153.2/113.5/2.7|1.1h	176.1/151.0/3.3|1.9h

HDP and SNGP. Interesting future work includes applying the TNRM not just for time-series in topic modeling but also to allow sparsity of probabilities, for instance [Williamson et al., 2010].

Chapter 8

Generalized Dependent Random Probability Measures

8.1 Introduction

In the previous chapters, several methods have been introduced to construct dependent random probability measures (DRPM) including the hierarchical normalized random measure, mixed normalized random measure and thinned normalized random measure. One question raised so far is: is there any way to construct more general DRPMs that are not restricted to the normalized random measure family. The answer is of course positive. This chapter discusses two ways to construct general dependent random probability measures (denoted as $\{\mu_t\}$) beyond this class:

1. Generalize the construction of MNRM and TNRM in Chapter 6 and Chapter 7 by transforming the individual atoms (w_k, θ_k)'s in the Poisson process with appropriate transformation functions, *e.g.*, via the following hierarchical construction:

   $$\begin{aligned} \tilde{\mu}_0 &= \sum_k w_k \delta_{\theta_k}, && \text{a base CRM} \\ \mu_t &\propto \sum_k f_t(w_k) \delta_{T_t(\theta_k)}, && \text{a set of DRPMs} \end{aligned}$$

   where (w_k, θ_k)'s are atoms from the base Poisson process/CRM, $f_t : \mathbb{R}^+ \to \mathbb{R}^+$, $T_t : \Theta \to \Theta$ are measurable functions such that μ_t's are legal random probability measures. Following this idea, Section 8.2 below specifics the forms for theses transformations via a hierarchical construction, which essentially combines MNRM and TNRM to form more general dependent random probability measures. Note the DRPMs following this construction usually do not embody marginal posteriors, thus inference is limited to slice samplers.

2. Follow the idea of Poisson-Kingman processes [Pitman, 2003] by first explicitly mixing the conditional law of the Poisson process with a prior for the total mass $\sum_k w_k$ to form a new RPM, then constructing dependent RPMs following ideas in, for example SNRMs, MNRMs and TNRMs. This forms another generalized

class of DRPMs, we call it *dependent Poisson-Kingman process* (DPKP). This class is attractive in that it not only generalizes the NRM to more general RPMs for constructing *dependent Poisson-Kingman process*, which are beyond the NRM class, but also allows marginal posterior inference algorithms to be developed easily. More detailed construction will be presented in Section 8.3.

In the rest of this chapter, these two classes of DRPMs will be defined, and the corresponding posterior structures and their posterior inference algorithms will also be discussed.

8.2 Thinned-Mixed Normalized Random Measures

This section defines a generalized dependent random probability measure called the *thinned-mixed normalized random measure* (TMNRM). The idea of the TMNRM is to define the transformations f_t above as the combination of the MNRM and TNRM such that it enjoys the flexibility that the atoms of the Poisson process are not only thinned but also weighted. Apart from making the resulting μ_t's sparse by thinning, TMNRM further imposes more variations between different μ_t's by weighting their individual atoms. Note the samples θ_k's are drawn *i.i.d.* from space Θ and independent of w_k's, thus the transformations T_t's are not discussed here for simplicity, *i.e.*, simply set them to be the identity transformation.

Specifically, in TMNRM, assume we have $\#\mathcal{R}$ independent Poisson processes, each is for one *region* and forms a NRM denoted as μ_r. Then the dependent random probability measures μ_t's can be constructed by the following hierarchical construction:

$$\begin{aligned}
\tilde{\mu}_r &= \sum_k w_{rk}\delta_{\theta_{rk}}, && r = 1, 2, \cdots, \#\mathcal{R} \\
q_{0rk} &\sim \text{Gamma}(a_{0q}, b_{0q}), && r = 1, 2, \cdots, \#\mathcal{R}, k = 1, 2, \cdots \\
q_{1rk} &\sim \text{Gamma}(a_{1q}, b_{1q}), && r = 1, 2, \cdots, \#\mathcal{R}, k = 1, 2, \cdots \\
b_{rk} &\sim \text{Beta}(a_b, b_b), && r = 1, 2, \cdots, \#\mathcal{R}, k = 1, 2, \cdots \\
g_{trk} &\sim \text{Gamma}(q_{0rk}, q_{1rk}), && t = 1, 2, \cdots, T, r = 1, 2, \cdots, \#\mathcal{R}, k = 1, 2, \cdots \\
z_{trk} &\sim \text{Bernoulli}(b_{rk}), && t = 1, 2, \cdots, T, r = 1, 2, \cdots, \#\mathcal{R}, k = 1, 2, \cdots \\
\mu_t &= \sum_r \sum_k \frac{g_{trk} z_{trk} w_{rk}}{\sum_{r'} \sum_{k'} g_{tr'k'} z_{tr'k'} w_{r'k'}} \delta_{\theta_{rk}} && t = 1, 2, \cdots, T
\end{aligned}$$

Note the last equation in the above construction combines MNRM and TNRM by associating each atom with a weighting variable g_{trk} and a Bernoulli thinning variable z_{trk}. Also note that by building this hierarchy, the random variables in μ_t generally cannot be integrated out analytically, thus the property that μ_t's are marginally distributed as NRMs is no longer preserved. Although this does not comply the nice theoretical property of the MNRM and TNRM, it obtains a more flexible way of controlling the dependency in the model.

8.2.1 Posterior simulation

Assume in each *region*, $\tilde{\mu}_r's$ are with Lévy measure $\nu_r(\mathrm{d}w, \mathrm{d}\theta)$ defined on space $\mathbb{R}^+ \times \Theta$. In posterior inference, obviously the NRMs μ_r's could not be integrated out because of the coupling of μ_r's and other random variables such as g_{trk}'s and z_{trk}'s. As a result, similar to the TNRM, we need to resort to the slice sampler. Obviously, the variables $\{q_{0rk}\}$, $\{q_{1rk}\}$, $\{b_{rk}\}$, $\{g_{trk}\}$, $\{z_{trk}\}$ have closed form updates given data from their lower level in the hierarchy, *i.e.*, conditioned on g_{trk}'s, q_{1rk} has the following conditional posterior:

$$p(q_{1rk}|\{g_{trk}\}) \sim \text{Gamma}\left(a_{1q}, b_{1a} + \sum_t g_{trk}\right) .$$

Updated formulas for the other variables follow similarly and will be omitted here.

The only difficult task left is to simulate the Poisson points from the underlying Poisson processes. Following the steps in TNRM, for a specific slice level $\mathcal{L}$, there are two kinds of atoms – those with and without observations:

- for the atoms with observations, we can sequentially re-sample $(q_{0rk}, q_{1rk}, b_{rk}, g_{trk}, z_{trk})$ and w_{rk}'s conditioned on other variables, which have analytical formulations thus are easy.
- for those w_{rk}'s without observations, we know that they come from a new completely random measure with a new Lévy measure, say $\nu'(\mathrm{d}w, \mathrm{d}\theta)$. Similar to the TNRM case, these atoms can be sampled by adaptively thinning a unit-rate Poisson process.

To derive the conditional Lévy measure of the CRM, the Poisson process partition calculus framework needs to be used. Following similar steps as in TNRM and after simplification, the conditional Lévy measure is given by

$$\nu'(\mathrm{d}w, \mathrm{d}\theta) = \left(\prod_t \int_0^1 \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \left(b \left(\frac{q_1}{q_1 + u_t w}\right)^{-q_0} + 1 - b\right) \mathrm{d}b\mathrm{d}q_0\mathrm{d}q_1\right) \nu(\mathrm{d}w, \mathrm{d}\theta)$$

The integrations in conditional Lévy measure above deter the adaptive thinning approach [Favaro and Teh, 2013] from working because the Lévy measure is hard to be evaluated. To overcome this difficulty, according to the *Marking Theorem* 2.6 of the Poisson process, we can simply augment the Poisson process from space $\mathbb{R}^+ \times \Theta$ to space $\mathbb{R}^+ \times \Theta \times \mathbb{R}^+ \times \mathbb{R}^+ \times [0, 1]$, where the last three spaces correspond to the spaces where the Gamma random variables q_{0rk}, q_{1rk} and the Beta random variable b_{rk} live on. So now the conditional Lévy measure on this augmented space is simply

$$\begin{aligned}
&\tilde{\nu}(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}q_{0rk}, \mathrm{d}q_{1rk}, \mathrm{d}b_{rk}) \\
&= \prod_t \left(b_{rk} \left(\frac{q_{1rk}}{q_{1rk} + u_t w}\right)^{-q_{0rk}} + 1 - b_{rk}\right) \nu(\mathrm{d}w, \mathrm{d}\theta) G_0(\mathrm{d}q_{0rk}) G_1(\mathrm{d}q_{1rk}) B(\mathrm{d}b_{rk}) ,
\end{aligned}$$

where $G_0(\cdot)$ is the cumulative function of a Gamma distribution with parameters (a_{0q}, b_{0q}), $G_1(\cdot)$ is the cumulative function of a Gamma distribution with parameters (a_{1q}, b_{1q}), $B(\cdot)$ is the cumulative function of a Beta distribution with parameters (a_b, b_b).

Now the adaptive thinning approach to simulate the Poisson process is applicable. Following the thinning procedure of TNRM, obviously we have the adaptively upper bound $\tilde{\nu}'(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}q_{0rk}, \mathrm{d}q_{1rk}, \mathrm{d}b_{rk})$ [1] of $\tilde{\nu}(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}q_{0rk}, \mathrm{d}q_{1rk}, \mathrm{d}b_{rk})$, then we do the thinning as follows:

- Sample q_{0rk}, q_{1rk}, b_{rk} from the corresponding Gamma and Betta distributions respectively as follows

$$\begin{aligned} q_{0rk} &\sim \mathrm{Gamma}(a_{0q}, b_{0q}) \\ q_{1rk} &\sim \mathrm{Gamma}(a_{1q}, b_{1q}) \\ b_{rk} &\sim \mathrm{Beta}(a_b, b_b) \end{aligned}$$

- Use the sampled values q_{0rk}, q_{1rk}, b_{rk} to evaluate formulas related to $\tilde{\nu}'(\cdots)$ and $\tilde{\nu}(\cdots)$, and do the thinning as in TNRM.

8.3 Dependent Poisson-Kingman Processes[2]

This section describes the second way to construct more general DRPMs by first extending the NRM to a more general class of RPM called Poisson-Kingman process (PKP), then building dependent RPMs based on the corresponding PKPs. As a result, the dependent random probability measures constructed also are not NRM distributed. In the following, how to construct the general Poisson-Kingman process from the NRM will be first introduced.

8.3.1 Poisson-Kingman processes

Generally speaking, Poisson-Kingman processes are generalizations of normalized random measures by *tilting* the total masses with some appropriate functions [Pitman, 2003]. Formally, let $\nu(\mathrm{d}w, \mathrm{d}\theta)$ be a measure on the space $\mathcal{W} \times \Theta$. Denote $P_\nu(\cdot)$ as the law of a Poisson random measure $\mathcal{N}(\mathrm{d}w, \mathrm{d}\theta)$ on this space with mean measure $\nu()$, and denote $E_\nu[\cdot]$ as the expectation over the Poisson random measure ν. Let $f : \mathbb{R}^+ \to [0, 1]$ be a probability distribution on $\mathbb{R}^+$. The idea of Poisson-Kingman process is to define a "*tilted*" point process by *tilting* the total mass $Z = \sum_k w_k$ by the function/distribution f such that the new law of the Poisson random measure now becomes [Pitman, 2003]

$$P_{f,\nu}(\mathrm{d}\mathcal{N}) = \int_{\mathbb{R}^+} P_v(\mathrm{d}\mathcal{N} | Z = t) f(Z) \mathrm{d}Z = \frac{1}{K} \int_{\mathbb{R}^+} P_v(\mathrm{d}\mathcal{N}) f(Z) \mathrm{d}Z \,, \tag{8.1}$$

[1] We can use the same upper bound on $\nu(\mathrm{d}w, \mathrm{d}\theta)$ as in the TNGG case.

[2] This section is based on personal communication with Vinayak Rao and Yee Whye Teh.

where K is the normalization factor to ensure (8.1) to be a valid probability distribution, and the subscripts (f, ν) means the distribution law depends on f and ν.

Usually f could take many specific forms. For a conjugate class of the NRM, please refer to [Lau, 2013]. Here to facilitate calculations, a general class is to define f via an augmented form as [Favaro et al., 2013b]:

$$f(Z) \propto \tilde{L}(\tau) e^{-\tau Z} , \tag{8.2}$$

where τ is an auxiliary nonnegative random variable, $\tilde{L} : \mathbb{R}^+ \to \mathbb{R}$ is a measurable function on $\mathbb{R}$. Now the *tilted* law $P_{\tilde{L},\nu}$ becomes:

$$P_{\tilde{L},\nu}(\mathrm{d}\mathcal{N}) = \frac{\int_{\mathbb{R}^+} \tilde{L}(\tau) e^{-\tau Z} \mathrm{d}\tau}{\zeta_L} P_\nu(\mathrm{d}\mathcal{N}) , \tag{8.3}$$

where $\zeta_{\tilde{L}} = E_\nu[\int_{\mathbb{R}^+} \tilde{L}(\tau) e^{-\tau Z} \mathrm{d}\tau]$ is the normalization constant, $\tilde{L}(\tau)$ can be any measurable functions such that $\zeta_{\tilde{L}}$ is finite. It is easy to see the joint density of τ and $\mathcal{N}$ becomes:

$$p_{\tau,\mathcal{N}}(\mathrm{d}\tau, \mathrm{d}\mathcal{N}) = \frac{1}{\zeta_{\tilde{L}}} \tilde{L}(\tau) e^{-\tau Z} \mathrm{d}\tau P_v(\mathrm{d}\mathcal{N}) \tag{8.4}$$

It can also be seen that the marginal density of τ is

$$P(\tau) \propto \tilde{L}(\tau) \phi(\tau t), \quad \text{where} \tag{8.5}$$

$$\phi(\tau Z) = E_\nu[\exp(-\tau Z)] = \exp\left(-\int_{\mathcal{W}\times\Theta} (1 - e^{-\tau w}) \nu(\mathrm{d}w, \mathrm{d}) \right) \tag{8.6}$$

Clearly, from the gamma identity, the marginal density of $\mathcal{N}$ corresponds to $P_{\tilde{L},\nu}$. The following result is from [Rao, 2013], which corresponds to the conditional mean measure of the Poisson process:

Theorem 8.1. *Conditioned on τ, $\mathcal{N}$ is a Poisson process with mean measure*

$$\exp(-\tau w) \nu(\mathrm{d}w, \mathrm{d}\theta) .$$

Proof. The proof follows the technique of James [2002]. Denote by $P(\mathrm{d}\mathcal{N}|\tau)$ the conditional law of $\mathcal{N}$. We have:

$$P(\mathrm{d}\mathcal{N}|\tau) = \frac{1}{K} \exp(-\tau Z) P(\mathrm{d}\mathcal{N})$$

where K is the normalization constant. Denote $\mathcal{M}$ as the space of bounded finite measures, then since $Z = \int_{\mathcal{W}\times\Theta} w \mathcal{N}(\mathrm{d}w, \mathrm{d}\theta)$, based on Cambell's theorem 2.10 we have

$$K = \int_{\mathcal{M}} \exp(-\tau Z) P(\mathrm{d}\mathcal{N})$$

$$= \exp\left(-\int_{\mathcal{W}\times\Theta}(1-e^{-\tau w})\nu(\mathrm{d}w,\mathrm{d}\theta)\right)$$

For some nonnegative function $g(\theta, w)$, we calculate the characteristic functional of g with respect to the law $P(\mathrm{d}\mathcal{N}|\tau)$:

$$\begin{aligned}
\phi_\tau(g) &= \int_{\mathcal{M}} \exp\left(-\int_{\mathcal{W}\times\Theta} g(\theta,w)\mathcal{N}(\mathrm{d}w,\mathrm{d}\theta)\right) P(\mathrm{d}\mathcal{N}|\tau) \\
&= \int_{\mathcal{M}} \exp\left(-\int_{\mathcal{W}\times\Theta} (g(\theta,w)+\tau h(w))\mathcal{N}(\mathrm{d}w,\mathrm{d}\theta)\right) \frac{1}{K} P(\mathrm{d}\mathcal{N}) \\
&= \frac{1}{K}\exp\left(-\int_{\mathcal{W}\times\Theta}(1-e^{-(g(\theta,w)+\tau h(w))})\nu(\mathrm{d}w,\mathrm{d}\theta)\right) \\
&= \exp\left(-\int_{\mathcal{W}\times\Theta}(1-e^{-(g(\theta,w))})\exp(-\tau h(w))\nu(\mathrm{d}w,\mathrm{d}\theta)\right) \qquad (8.7)
\end{aligned}$$

Clearly from (8.7) and according to the unity of the characteristic functional, we conclude that $P(\mathrm{d}\mathcal{N}|\tau)$ is the law of a Poisson process with mean measure as

$$\tilde{\nu}(\mathrm{d}w,\mathrm{d}\theta) = \exp(-\tau h(w))\nu(\mathrm{d}w,\mathrm{d}\theta)\ .$$

□

Theorem 8.1 allows us to sample from a general Poisson-Kingman process as follows:

- First sample τ based on (8.4).
- Conditioned on τ, sample a Poisson process $\mathcal{N}$ from $P_\nu(\mathrm{d}\mathcal{N}|\tau)$.
- Construct a completely random from this Poisson process and normalize it.

Thus, we call the random measure tilted by f in (8.2) the *tilted*-PK(ν) process, and its normalized version the *normalized tilted*-PK(ν) process. It can be shown that the familiar two-parameter Poisson-Dirichlet process (Pitman-Yor process) is a specific class of the *tilted*-PK(ν) process.

Pitman-Yor processes When ν corresponds to the Lévy measure of the σ-stable subordinator, *e.g.*, $\nu(\mathrm{d}w,\mathrm{d}\theta) \propto w^{-1-\sigma}\mathrm{d}wH(\mathrm{d}\theta)$ where $0 < \sigma < 1$, H a probability measure on Θ, and $\tilde{L}(\tau)$ takes the form

$$\tilde{L}(\tau) = \tau^{b-1}\ ,$$

where b can take values to ensure $\zeta_{\tilde{L}}$ defined above is finite, then it is called a *polynomially tilted* Poisson process, which corresponds to the Pitman-Yor process [Pitman, 2003].

8.3.2 Dependent Poisson-Kingman processes

Given the definition of the generalized Poisson-Kingman process in the last section, it is straightforward to construct dependent Poisson-Kingman processes (DPKP) by simply applying the ideas of, for example the *spatial normalized random measure, mixed normalized random measure* or *thinned normalized random measure*. In the following a specific class called *spatial Pitman-Yor process* (SPYP) is introduced, which combines the idea of spatial normalized random measures and Poisson-Kingman process. Other dependency models can be constructed similarly thus the details will be omitted.

To define the SPYP, it is assumed that there is a Poisson process defined on the augmented product space $\mathbb{R}^+ \times \Theta \times \mathcal{R}$. Each *time* corresponds to a RPM, which is associated with several regions on the *region space* $\mathcal{R}$. To define a Pitman-Yor process from this Poisson process, an auxiliary variable τ is introduced as above. Now conditioned on τ and after applying Theorem 8.1, we get the conditional Lévy measure of the CRM constructed from the Poisson process to be (note the conditional Lévy measure of the PYP is a *polynomially tilting* stable subordinator):

$$\exp(-\tau w)\nu(\mathrm{d}w, \mathrm{d}\theta, \mathrm{d}r) \triangleq w^{-1-\sigma} e^{-\tau w} \mathrm{d}w Q(\mathrm{d}r) H(\mathrm{d}\theta) ,$$

where Q and H are measures on space $\mathcal{R}$ and Θ, respectively. Now we can introduce spacial structures into the Poisson process as in the case of SNRM, *e.g.*, each element $r \in \mathcal{R}$ corresponds to one region and is associated with a Poisson process. For more flexible modeling, we further allow each region to have its own τ parameter, written as τ_r. By integrating over the region $\mathcal{R}_r$, we get the conditional Lévy measure in region $\mathcal{R}_r$ as

$$\nu_r(\mathrm{d}w, \mathrm{d}\theta) = \frac{\sigma Q_r}{\Gamma(1-\sigma)} w^{-1-a} e^{-\tau_r w} \mathrm{d}w H(\mathrm{d}\theta) ,$$

where $Q_r = Q(\mathcal{R}_r)$. Denote the completely random measure with Lévy measure $\nu_r(\mathrm{d}w, \mathrm{d}\theta)$ in region $\mathcal{R}_r$ as G_r^*, then the joint distribution of $(\boldsymbol{X}, \boldsymbol{u}, \{G_r^*\}, \{\tau_r\})$ can then be written as

$$p(\boldsymbol{X}, \boldsymbol{u}, \{G_r^*\}, \{\tau_r\}\}) \propto p(\{\tau_r\}) p(\boldsymbol{X}, \boldsymbol{u}, |\{G_r^*\}, \{\tau_r\}\}) .$$

The above representation of the joint distribution is useful in deriving the posterior of the SPYP model. Using the same notation as in Chapter 6, employing priors $p(\{\tau_r\})$ for $\{\tau_r\}$ and integrating out all G_r^*'s with the Poisson process partition calculus (Theorem 2.12), the posterior is given by

$$\begin{aligned} & p(\boldsymbol{X}, \boldsymbol{u}, \{\tau_r\} | \sigma, \{b_r\}) = \mathbb{E}\left[p(\{\tau_r\}) p(\{G_r^*\} | \{\tau_r\}) p(\boldsymbol{X}, \boldsymbol{u} | \{\mu_r\}, \{\tau_r\})\right] \\ \propto & \left(\frac{\sigma}{\Gamma(1-\sigma)}\right)^{\sum_r K_r} \left(\prod_r \frac{\sigma}{\Gamma\left(\frac{b_r}{\sigma}\right)} \tau_i^{b_r - 1} p(\tau_r) Q_r^{K_r}\right) \left(\prod_{r=1}^{I} \prod_{k=1}^{K_r} \frac{\Gamma(n_{\cdot rk} - \sigma)}{\left(1 + \sum_{t : \tilde{R}_r \in \mathcal{R}_t} u_t\right)^{n_{\cdot rk} - \sigma}}\right) \end{aligned}$$

$$\left(\prod_t \frac{u_t^{N_t-1}}{N_t!}\right)\left(\prod_r e^{-M_r Q_r\left(\left(\tau_r+\sum_{t:\tilde{R}_r\in\mathcal{R}_t} u_t\right)^{\sigma}-\tau_r\right)}\right)\left(\prod_{t=1}^{T}\prod_{l=1}^{L_t} f(x_{tl}|\theta_{g_{tl}s_{tl}})\right) \tag{8.8}$$

Now the posterior inference can be done by iteratively sampling the related random variables based on the posterior (again denote the whole set of random variables as C):

Sampling (s_{tl}, g_{tl}), u_t and σ: These are similar to the spatial normalized generalized Gamma process described in Section 6.2.3 of Chapter 6.

Sampling τ_r: τ_r has posterior proportional to

$$p(\tau_r|C-\tau_r) \propto \tau_r^{b_r-1} e^{-M_r Q_r\left((\tau_r+\sum_{t:\tilde{R}_r\in\mathcal{R}_t} \mu_t)^{\sigma}-\tau_r^{\sigma}\right)} p(\tau_r) \ .$$

Employing a Gamma prior for τ_r and using a change of variable $\gamma_r = \tau_r^{\sigma}$, then it can be shown that $p(\tau_r|C-\tau_r)$ is log-concave.

Sample b_r: b_r is the concentration parameter in the two-parameter Poisson-Dirichlet process. The conditional distribution for b_r, with prior $p(b_r)$ is

$$p(b_r|C-b_r) \propto \frac{\tau_r^{b_r-1}}{\Gamma\left(\frac{b_r}{\sigma}\right)} p(b_r) \ .$$

If we use a Gamma prior for b_r, then $p(b_r|C-b_r)$ is log-concave thus can be easily sampled

We can see that the sampling procedure for the SPYP follows similarly as the SNRM except for some extra random variables, thus the code for SNGG can be easily adapted to the SPYP model.

8.4 Conclusion

Based on previous chapters, this chapter discusses possible generalizations of dependent normalized random measures. This chapter considers two possible ways: 1) via proper transformations on the atoms of the Poisson process before using it to construct dependent normalized random measures. The transformations is usually defined via hierarchical constructions, for example, the thinned-mixed normalized random measure defined in Section 8.2. 2) via dependency operators on a larger family of random probability measures called the Poisson-Kingman process. This class is attractive partially because it includes some familiar Bayesian nonparametric priors such as the well known Pitman-Yor process [Pitman, 2003]. Furthermore, posterior inference can be derived easily with this construction. Though feasible for posterior inference, the generalized dependent random probability measures have not been empirically tested because of the lack of motivation in real applications for now, leaving an interesting direction for future work.

Chapter 9

Conclusion and Future Work

9.1 Conclusion

Bayesian nonparametrics, as an extension of finite dimensional Bayesian models, has gained increasing attention in modern machine learning due to its flexibility in adapting model complexity with data sizes. Among those Bayesian nonparametric priors, the random probability measure, a generalization of the Dirichlet process, plays an important role in modeling probability vectors, *e.g.*, topic distribution vectors and topic-word distribution vectors in topic modeling, friendship distribution vectors in social network modeling, *etc*.

As is known, the Dirichlet process [Ferguson, 1973] is restricted in modeling because it is incapable of dealing with long tailed distributions (power-law distributions). In this thesis, a generalized random probability family called normalized random measure (NRM), is introduced based on the theory of Poisson processes. The NRM is first introduced by Regazzini et al. [2003] from the statistical community; this thesis constitutes the first systematic study on the NRM by extending the concept and reformulating it to deal with machine learning problems. Then several dependent Bayesian nonparametric models are built based on the NRM to deal with different dependency structures. These includes the hierarchical normalized random measures (HNRM) for hierarchical modeling, dependent hierarchical normalized random measures (DHNRM) for hierarchical and Markovian modeling, mixed normalized random measures (MNRM) for general dependency modeling, and thinned normalized random measures (TNRM) for general and sparsity modeling. Finally some generalized dependent random probability measures including the mixed-thinned normalized random measures and the dependent Poisson-Kingman process.

Although there has been some related work by using the normalized random measure for dependency modeling from the statistical community, *e.g.*, the time dependent stick-breaking process [Griffin and Steel, 2009], the Ornstein-Uhlenbeck-like time varying stochastic process [Griffin, 2011], the nested Dirichlet process [Rodriguez et al., 2008], some other simple dependent NRM models such as [Griffin et al., 2013; A. Lijoi and B. Nipoti and I. Prunster, 2013a,b; F. Leisen and A. Lijoi and D. Spano', 2013], their foci were on the analysis of their dependency structures, and their posterior structures and inference are usually complicated. Instead, the models proposed in the thesis focus more on the computational side but still have very nice distributional properties. The major contributions of these models are described in more details below.

The hierarchical normalized random measure, mimicking the construction of the hierarchical Dirichlet process [Teh et al., 2006], augments the modeling ability to deal with long tailed distributions. A related model that exhibits similar function is the hierarchical Pitman-Yor process [Teh, 2006b,c; Du et al., 2010; Sato and Nakagawa, 2010; Chen et al., 2011, 2014a]. The Pitman-Yor process is closely related to the normalized random measure via, for example [Pitman and Yor, 1997] or Corollary 3.8, and it is an instance of the general Poisson-Kingman process as discussed in Chapter 8, thus we expect the modeling ability of these two models to be similar, though the hierarchical Pitman-Yor process seems to have some more attractive distributional properties such as the closed form stick-break construction. To sum up, the chapter introduces an efficient MCMC sampling algorithm for the general HNRM, which is shown to be comparable to the HDP in running time while obtains more modeling flexibilities.

The dependent hierarchical normalized random measure proposed in Chapter 5 makes a first attempt to extend the HNRM with Markovian dependent operators for dynamic topic modeling, which are adapted from [Lin et al., 2010]. These operators allow topics to die, to be born and to vary at any time, which well fits the dynamic topic model scenario. Furthermore, by employing tools from Poisson process theory, the dependencies can be quantitatively worked out. The model extends the dependent Dirichlet process of Lin et al. [2010] by generalizing it to general NRMs, thus obtaining more flexible models. Similar to [Lin et al., 2010], one deficiency of the model is that the posterior inference relies on some approximations for efficiency. However, the approximations do not seem to impair the model performance too much, and is acceptable in practice. How to design an efficient and exact algorithm is an interesting future work.

The thesis then continues contributing by proposing a more theoretically clean dependency model called the mixed normalized random measure. The construction of the MNRM is fairy simple by first weighting the Poisson process in disjoint regions and then doing a superposition. This generalizes existing work on spatial normalized Gamma process [Rao and Teh, 2009] and a simple dependent NRM model [Griffin et al., 2013]. Though simple in construction, it endows the ability of flexible dependency modeling, *e.g.*, empirically it is better than the HNRM. Furthermore, it has nice distributional properties. For example, marginally each of the dependent NRMs can be shown to belong to the same class of the NRM, *e.g.*, marginally Dirichlet process distributed. Note also that the posterior inference algorithm for the model can be easily adapted to all the models within the NRM family, which is different from existing work. Furthermore, the construction of the MNRM can be easily adapted to do hierarchical model as well as the Markovian dependency modeling.

To introduce sparsity in the model, the thinned normalized random measure is proposed in Chapter 7. Rather than weighting the whole Poisson process as in the MNRM, TNRM chooses individual atoms independently. The price for this flexibility, of course, is that the resulting posterior structure is complicated. This hinders the development of an efficient marginal sampler. Thanks to recent advances in the MCMC theory, an efficient slice sampler can be developed for the TNRM. The slice sampler developed for the TNRM overcomes the problem of complicated coupling of random variables by simulating the underlying Poisson processes. Conditioned on these Poisson processes, the sampling becomes much easier and can deal with relative large datasets as well. Note that existing work on the "thinning" idea include [Foti et al., 2013; Lin et al., 2010; Lin and Fisher, 2012]. However, [Foti et al., 2013] is only restricted in the Dirichlet process case and posterior inference relies on a truncated approximation of the DP; while [Lin et al., 2010; Lin and Fisher, 2012] fail to recognize the true posterior structure of the thinning modeling thus have incorrect samplers. The TNRM constitutes the first work on correctly dealing with the thinning operation on the NRM family.

Finally, more general dependent random probability measures (DRPM) are discussed in Chapter 8. This chapter introduces two ways of constructing more general DRPMs: by transforming individual atoms of the Poisson process and by introducing dependencies on a more general class of random probability measures called Poisson-Kingman processes. The ideas of inference for the models are also discussed but experiments are omitted because of the lack of enough motivations in real applications. However, we believe these more general DRPMs will have great potential implications in the future.

9.2 Future Work

There are several possible avenues for future work:

More applications for the DNRM framework Though the theory for the dependent normalized random measures has been well developed in the thesis, applications are still limited in literature. Except for the topic modeling and Gaussian mixture applications discussed in the thesis, other applications we are aware of include the cosmic microwave background radiation modeling and motorcycle crash simulation study [Foti and Williamson, 2012], and stock exchange analysis [Griffin, 2011]. More applications are expected within the DNRM framework, for example, the Bayesian sparsity modeling for regression [Griffin and Brown, 2013] with the TNRM.

Alternatives for other nonparametric Bayesian priors Given the flexible construction of DNRM from Poisson processes (or completely random measures), it is interesting to further study the relationships between the DNRM and other nonparametric Bayesian priors. An interesting Bayesian nonparametric prior in recent machine learning is the the prior for exchangeable random graph built on the theory of exchangeable random arrays [Hoover, 1979; Aldous, 1981; Orbanz and Roy, 2014]. A concrete example is the Gaussian process random function network model of [Lloyd et al., 2012]. A recent progress by extending the completely random measure framework for this task is done by Caron and Fox [2014]. It will be interesting to further extend the techniques developed in this thesis for such dependency modeling. Some other nonparametric Bayesian priors related to the completely random measure include the prior for ranked data [Caron and Teh, 2012] and the prior for bipartite graphs [Caron, 2012], which extends the Indian buffet process by allowing power-law distributional property in the number of features for dishes for each customer. A different treatment with similar goal appears in [Williamson et al., 2013] by restricting the domain of distributions. Note that the techniques used in the thesis, *e.g.*, the thinning operator, can also be adapted to the completely random measure to construct alternatives for the IBP, which is an interesting future work.

Large scale DNRM learning A recent innovation in machine learning is large scale Bayesian inference, which has received increasing attention due to high demand in real applications. Some representative models include the large scale distributed learning for the latent Dirichlet allocation topic model [Smola and Narayanamurthy, 2010] and some supervised topic models such as the max-margin topic model [Zhu et al., 2013] and the Logistic-Normal topic model [Chen et al., 2013c]. Moreover, some recently developed large scaled distributed systems such as [Ho et al., 2013; Li et al., 2013] further popularize this topic for future research.

Apart from distributed methods, large scale learning can also be done via stochastic variational methods [Hoffman et al., 2013] or its Gibbs sampling alternative via stochastic gradient Langevin dynamics [Welling and Teh, 2011]. These methods make learning scalable by considering a subset of the dataset each time for parameter updating. Recent progress on improving the stochastic gradient Langevin dynamics is to consider its natural gradient instead of the raw gradient [Patterson and Teh, 2013].

One challenge of the above methods is that they are currently only be able to deal with finite Bayesian models, so extending them to Bayesian nonparametric models such as the dependent normalized random measure is an important and interesting future work.

Bibliography

A. Lijoi and B. Nipoti and I. Prunster, 2013a. Bayesian inference with dependent normalized completely random measures. *To appear in Bernoulli*, (2013). (cited on pages 79 and 163)

A. Lijoi and B. Nipoti and I. Prunster, 2013b. Dependent mixture models: clustering and borrowing information. *To appear in Computational Statistics and Data Analysis*, (2013). (cited on page 163)

Accardi, L., 2001. *De Finetti theorem*. Springer. (cited on page 63)

Ahmed, A. and Xing, E., 2010. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the Twenty-sixth Conference Annual Conference on Uncertainty in Artificial Intelligence*, 411–418. (cited on pages 85 and 93)

Aldous, D. J., 1981. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11, 4 (1981), 581–598. (cited on page 165)

Aldous, D. J., 1985. Exchangeability and related topics. *École d'Été St Flour 1983*, (1985), 1–198. (cited on page 6)

Andrieu, C. and Roberts, G. O., 2009. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37, 2 (2009), 697–725. (cited on page 145)

Arbel, J.; Mengersen, K.; and Rousseau, J., 2014. Bayesian nonparametric dependent model for the study of diversity for species data. *submitted to the Annals of Applied Statistics*, (2014). (cited on page 14)

Arratia, R.; Barbour, A. D.; and Tavaré, S., 2003. *Logarithmic combinatorial structures: a probabilistic approach*. EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich. ISBN 3-03719-000-0. (cited on page 10)

Bache, K. and Lichman, M., 2013. UCI machine learning repository. http://archive.ics.uci.edu/ml. (cited on page 55)

Bartlett, N.; Pfau, D.; and Wood, F., 2010. Forgetting counts: constant memory inference for a dependent hierarchical Pitman-Yor process. In *International conference on machine learning*. (cited on page 86)

Bertoin, J., 2006. *Random fragmentation and coagulation processes*. Cambridge University Press, Cambridge, vol.102 of Cambridge Studies in Advanced Mathematics. (cited on page 29)

Blei, D. and Lafferty, J., 2006. Dynamic topic models. In *International conference on machine learning*. (cited on page 92)

Blei, D. M.; Ng, A. Y.; and Jordan, M. I., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 (2003), 993–1022. (cited on pages 2, 9, 61, 71, and 72)

Brix, A., 1999. Generalized Gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31 (1999), 929–953. (cited on page 38)

Broderick, T.; Jordan, M. I.; and Pitman, J., 2012. Beta processes, stick-breaking, and Power laws. *Bayesian Analysis*, 7 (2012), 439–476. (cited on page 130)

Buntine, W. and Hutter, M., 2012. A Bayesian view of the Poisson-Dirichlet process. Technical Report arXiv:1007.0296, NICTA and ANU, Australia. http://arxiv.org/abs/1007.0296. (cited on pages 4, 44, and 45)

Caron, F., 2012. Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*. (cited on page 166)

Caron, F. and Fox, E. B., 2014. Bayesian nonparametric models of sparse and exchangeable random graphs. Technical report, University of Oxford and University of Washington. (cited on page 166)

Caron, F. and Teh, Y. W., 2012. Bayesian nonparametric models for ranked data. In *Advances in Neural Information Processing Systems*. (cited on page 166)

Çinlar, E., 2010. *Probability and stochastics*. Springer. (cited on pages 3, 7, 19, 20, 21, 22, 23, 24, 26, 30, 36, and 101)

Chen, C.; Buntine, W.; and Ding, N., 2012a. Theory of dependent hierarchical normalized random measures. Technical Report arXiv:1205.4159, ANU and NICTA, Australia. http://arxiv.org/abs/1205.4159. (cited on pages 15 and 25)

Chen, C.; Buntine, W.; Ding, N.; Xie, L.; and Du, L., 2014a. Differential topic models. *IEEE Transactions on Pattern Recognition and Machine Intelligence (accept with minor revision)*, (2014). (cited on pages 15 and 164)

Chen, C.; Ding, N.; and Buntine, W., 2012b. Dependent hierarchical normalized random measures for dynamic topic modeling. In *International conference on machine learning (ICML)*. (cited on pages 71 and 130)

Chen, C.; Du, L.; and Buntine, W., 2011. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 296–311. (cited on pages 15, 45, and 164)

Chen, C.; Rao, V.; Buntine, W.; and Teh, Y. W., 2013a. Dependent normalized random measures. In *International conference on machine learning*. (cited on page 15)

Chen, C.; Rao, V.; Buntine, W.; and Teh, Y. W., 2013b. Supplementary material for dependent normalized random measures. Technical report, Supplements. (cited on pages 15, 130, and 133)

Chen, C.; Zhu, J.; and Zhang, X., 2014b. Robust bayesian max-margin clustering. In *Advances in Neural Information Processing Systems*. (cited on page 15)

Chen, J.; Zhu, J.; Wang, Z.; Zheng, X.; and Zhang, B., 2013c. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*. (cited on page 166)

Daley, D. J. and Vere-Jones, D., 1998. *An introduction to the theory of point processes*. Springer, New York. (cited on pages 22 and 29)

De Iorio, M.; Müller, P.; Rosner, G. L.; and MacEachern, S. N., 2004. An ANOVS model for dependent random measures. *Journal of the American Statistical Association*, 99, 465 (2004), 205–215. (cited on page 14)

Ding, N.; Chen, C.; and Vishwanathan, S. V. N., 2014. *t*-conditional random fields. *To be submitted to Neurocomputing*, (2014). (cited on page 15)

Du, L., 2012. *Non-parametric Bayesian Methods for Structured Topic Models*. PhD thesis, the Australian National University, Canberra, Australia. (cited on page 61)

Du, L.; Buntine, W.; and Jin, H., 2010. A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Machine Learning*, 81 (2010), 5–19. (cited on pages 122 and 164)

Du, L.; Buntine, W.; Jin, H.; and Chen, C., 2012. Sequential latent Dirichlet allocation. *Knowledge and Inforamtion Systems*, 31, 3 (2012), 475–503. (cited on page 15)

Du, L.; Ren, L.; Dunson, D. B.; and Carin, L., 2009. A Bayesian model for simultaneous image clustering, annotation and object segmentation. In *Advances in Neural Information Processing Systems*. (cited on page 66)

Eisenstein, J.; Ahmed, A.; and Xing, E., 2011. Sparse additive generative models of text. In *International conference on machine learning*. (cited on page 73)

F. Leisen and A. Lijoi and D. Spano', 2013. A vector of Dirichlet processes. *Electronic Journal of Statistics*, 7 (2013), 62–90. (cited on page 163)

Favaro, S.; Lijoi, A.; and Prünster, I., 2012. On the stick-breaking representation of normalized inverse Gaussian priors. *Biometrika*, 99, 3 (2012), 663–674. (cited on page 68)

Favaro, S.; Lijoi, A.; and Prünster, I., 2013a. Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability*, 23, 5 (2013), 1721–1754. (cited on pages 45 and 114)

Favaro, S.; Nipoti, B.; and Teh, Y. W., 2013b. A note on exponentially tilted stable distributions. *Submitted*, (2013). (cited on page 159)

Favaro, S. and Teh, Y. W., 2013. MCMC for normalized random measure mixture models. *Statistical Science*, 28, 3 (2013), 335–359. (cited on pages 38, 41, 55, 110, 118, 138, 143, and 157)

Ferguson, T. and Klass, M., 1972. A representation of independent increment processes without Gaussian component. *The Annals of Mathematical Statistics*, 43, 5 (1972), 1634–1643. (cited on pages 34 and 36)

Ferguson, T. S., 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 2 (1973), 209–230. (cited on pages 3, 13, 38, 62, 63, and 163)

Foti, N. J.; Futoma, J.; Rockmore, D.; and Williamson, S. A., 2013. A unifying representation for a class of dependent random measures. In *International Conference on Artificial Intelligence and Statistics*. (cited on pages 130 and 165)

Foti, N. J. and Williamson, S. A., 2012. Slice sampling normalized kernel-weighted completely random measure mixture models. In *Advances in Neural Information Processing Systems*. (cited on page 165)

Franti, P. and Virmajoki, O., 2006. Iterative shrinking method for clustering problems. *Pattern Recognition*, 39, 5 (2006), 761–765. (cited on page 55)

Gilks, W. Adaptive rejection sampling. Technical report, Department of Statistics, University of Leeds, UK. (cited on page 73)

Gilks, W. R. and Wild, P., 1992. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41, 2 (1992), 337–348. (cited on pages 51, 70, and 111)

Globerson, A.; Chechik, G.; Pereira, F.; and Tishby, N., 2007. Euclidean Embedding of Co-occurrence Data. *Journal of machine learning research*, 8 (2007), 2265–2295. (cited on pages 91 and 122)

Goldwater, S.; Griffiths, T.; and Johnson, M., 2006. Interpolating between types and tokens by estimating Power-law generators. In *Advances in Neural Information Processing Systems 18*, 459–466. (cited on pages 10 and 85)

Griffin, J.; Kolossiatis, M.; and Steel, M. F. J., 2013. Comparing distributions using dependent normalized random measure mixtures. *Journal of the Royal Statistical Society, Series B*, 75 (2013), 499–529. (cited on pages 96, 163, and 164)

GRIFFIN, J. AND WALKER, S., 2011. Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics*, 20, 1 (2011), 241–259. (cited on pages 33, 41, 51, 52, 53, 54, 57, 59, 89, and 138)

GRIFFIN, J. E., 2011. The Ornstein–Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference. *Journal of Statistical Planning and Inference*, 141 (2011), 3648–3664. (cited on pages 163 and 165)

GRIFFIN, J. E. AND BROWN, P. J., 2013. Some priors for sparse regression modelling. *Bayesian Analysis*, 8, 3 (2013), 691–702. (cited on page 165)

GRIFFIN, J. E. AND STEEL, M. F. J., 2009. Time-dependent stick-breaking processes. *working paper*, (2009). (cited on page 163)

GRIFFITHS, T. L. AND GHAHRAMANI, Z., 2011. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12 (2011), 1185–1224. (cited on pages 5, 11, and 130)

HO, Q.; CIPAR, J.; CUI, H.; LEE, S.; KIM, J.; GIBBONS, P. B.; GIBSON, G. A.; GANGER, G.; AND XING, E. P., 2013. More effective distributed ML via a stale synchronous parallel parameter server. In *Advances in Neural Information Processing Systems*. (cited on page 166)

HOFFMAN, M.; BLEI, D.; PAISLEY, J.; AND WANG, C., 2013. Stochastic variational inference. *Journal of Machine Learning Research*, 14 (2013), 1303–1347. (cited on page 166)

HOOVER, D. N., 1979. Relations on probability spaces and arrays of random variables. Technical report, Institute for Advanced Study, Princeton. (cited on page 165)

ISHWARAN, H. AND JAMES, L., 2001. Gibbs sampling methods for stick breaking priors. *Journal of the American Statistical Association*, 96 (2001), 161–173. (cited on page 4)

ISHWARAN, H. AND JAMES, L. F., 2003. Generalized weighted chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13 (2003), 2003. (cited on page 4)

JAMES, L., 2003. Bayesian calculus for Gamma processes with applications to semiparametric intensity models. *Sankhya: The Indian Journal of Statistics*, 65, 1 (2003), 179–206. (cited on page 14)

JAMES, L.; LIJOI, A.; AND PRUNSTER, I., 2006. Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, 33 (2006), 105–120. (cited on pages 41 and 85)

JAMES, L.; LIJOI, A.; AND PRUNSTER, I., 2009. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36 (2009), 76–97. (cited on pages 33, 39, 41, 42, 43, 46, 52, 58, 59, 68, 83, 90, 91, 100, 118, and 132)

James, L. F., 2002. Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. Technical report, http://arXiv.org/abs/math/0205093. (cited on pages 30, 32, 38, and 159)

James, L. F., 2005. Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics*, 33, 4 (2005), 1771–1799. (cited on pages 30, 35, 38, 40, and 41)

James, L. F., 2010. Coag-Frag duality for a class of stable Poisson-Kingman mixtures. Technical report, the Hong Kong University of Science and Technology, arxiv:1008.2420. (cited on page 78)

James, L. F., 2013. Stick-breaking PG(α, ζ)-Generalized Gamma processes. Technical report, the Hong Kong University of Science and Technology, arXiv:1308.6570. (cited on pages 39, 68, and 78)

Johnson, M. Pitman-Yor adaptor grammar sampler. Technical Report http://web.science.mq.edu.au/ mjohnson/Software.htm, Department of Computing, Faculty of Science, Macquarie University, Australia. (cited on page 72)

Johnson, M.; Griffiths, T.; and Goldwater, S., 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19*, 641–648. (cited on page 85)

Kingman, J., 1967. Completely random measures. *Pacific Journal of Mathematics*, 21, 1 (1967), 59–78. (cited on pages 19, 34, 35, and 79)

Kingman, J., 1975. Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37, 1 (1975), 1–22. (cited on page 38)

Kingman, J., 1993. *Poisson Processes*. Oxford University Press. (cited on pages 13, 14, 19, 22, 23, 24, 26, and 35)

Kolokoltsov, A. N., 2011. *Markov processes, semigroups and generators*. (cited on page 13)

Lau, J. W., 2013. A conjugate class of random probability measures based on tilting and with its posterior analysis. *Bernoulli*, 19, 5B (2013), 2590–2626. (cited on page 159)

Li, M.; Zhou, L.; Yang, Z.; Li, A.; Xia, F.; Andersen, D. G.; and Smola, A., 2013. Parameter server for distributed machine learning. In *NIPS workshop on Big Learning*. (cited on page 166)

Lijoi, A. and Bernardo, N., 2014. A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association*, 109, 506 (2014), 802–814. (cited on page 14)

Lijoi, A.; Mena, R.; and Prunster, I., 2007. Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of Royal Statistical Society B*, 69, 4 (2007), 715–740. (cited on page 39)

Lijoi, A. and Prunster, I., 2010. *Models beyond the Dirichlet process*. Cambridge University Press. (cited on page 38)

Lim, K. W.; Chen, C.; and Buntine, W., 2013. Twitter-network topic model: a full bayesian treatment for social network and text modeling. In *NIPS workshop on Topic Modeling*. (cited on page 15)

Lin, D. and Fisher, J., 2012. Coupling nonparametric mixtures via latent Dirichlet processes. In *Advances in Neural Information Processing Systems*. (cited on pages 130, 133, and 165)

Lin, D.; Grimson, E.; and Fisher, J., 2010. Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in Neural Information Processing Systems*. (cited on pages 25, 79, 85, 130, 133, 164, and 165)

Lloyd, J. R.; Orbanz, P.; Ghahramani, Z.; and Roy, D. M., 2012. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems*. (cited on page 165)

MacEachern, S. N., 1999. Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*. (cited on page 14)

MacEachern, S. N., 2000. Dependent dirichlet processes. Technical report, Ohio State University. (cited on page 14)

MacEachern, S. N., 2001. Decision theoretic aspects of dependent nonparametric processes. *Bayesian Methods with Application Science, Policy and Official Statistics*, (2001), 551–560. (cited on page 14)

MacEachern, S. N.; Athanasios, K.; and Gelfand, A., 2001. Spatial nonparametric bayesian models. *Proceedings of the 2001 Joint Statistical Meetings*, 3 (2001). (cited on page 14)

Mochihashi, D. and Sumita, E., 2008. The infinite Markov model. In *Advances in Neural Information Processing Systems 20*, 1017–1024. (cited on page 66)

Murphy, K. P., 2007. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, UCB. (cited on page 53)

Neal, R. M., 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9 (2000), 249–265. (cited on page 110)

Neal, R. M., 2003. Slice sampling. *The Annals of Statistics*, 31, 3 (2003), 705–767. (cited on pages 51, 70, 111, 119, and 145)

Orbanz, P. and Buhmann, J. M., 2007. Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 77 (2007), 25–45. (cited on page 66)

Orbanz, P. and Roy, D. M., 2014. Bayesian models of graphs, arrays and other exchangeable random structures. Technical report, Columbia University and Cambridge University. (cited on page 165)

Otis, G.; Hatcher, E.; and McCandless, M., 2009. Lucene in action. In *Manning Publications*, 475. (cited on page 122)

Patterson, S. and Teh, Y. W., 2013. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*. (cited on page 166)

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, 11 (1901), 559–572. (cited on page 11)

Perman, M.; Pitman, J.; and Yor, M., 1992. Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92, 1 (1992), 21–39. (cited on pages 3 and 38)

Pitman, J., 1995. Exchangeable and partially exchangeable random partitions. *Probability Theory Related Fields*, 102 (1995), 145–158. (cited on page 64)

Pitman, J., 1996. Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, 245–267. (cited on page 4)

Pitman, J., 2003. Poisson-Kingman partitions. Technical Report No. 625, Dep. Statist., UCB. (cited on pages 38, 39, 155, 158, 160, and 162)

Pitman, J., 2006. *Combinatorial stochastic processes*, vol. 1875 of *Lecture Notes in Mathematics*, x+256. Springer-Verlag. (cited on page 4)

Pitman, J. and Yor, M., 1997. The two-parameter Poisson-Diriclet distribution derived from a stable subordinator. *Annals of Probability*, 25, 2 (1997), 855–900. (cited on pages 3, 39, 44, and 164)

Rao, V., 2012. *Markov Chain Monte Carlo for Continuous-time Discrete-state Systems*. PhD thesis, University College London, London, Britian. (cited on pages 7, 13, and 24)

Rao, V., 2013. Spatial normalized random measures. Technical report, Unpublished Note. (cited on pages 15 and 159)

Rao, V. and Teh, Y. W., 2009. Spatial normalized Gamma processes. In *Advances in Neural Information Processing Systems*. (cited on pages 17, 105, 106, 111, 112, 114, 121, 128, and 164)

RASMUSSEN, C., 2000. The infinite Gaussian mixture model. In *Advances in information processing systems 12*, 554–560. (cited on page 120)

RASMUSSEN, C. E. AND WILLIAMS, C. K. I., 2006. *Gaussian Processes for Machine Learning*. The MIT Press. (cited on pages 5, 10, and 11)

REGAZZINI, E.; LIJOI, A.; AND PRUNSTER, I., 2003. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31, 2 (2003), 560–585. (cited on pages 37, 38, and 163)

REN, L.; DUNSON, D.; AND CARIN, L., 2008. The dynamic hierarchical Dirichlet process. In *Proceedings of the 25th international conference on Machine learning*, 824–831. (cited on page 85)

RODRIGUEZ, A.; DUNSON, D. B.; ; AND GELFAND, A. E., 2008. The nested Dirichlet process. *Journal of the American Statistical Association*, 103 (2008), 1149–1151. (cited on page 163)

ROSEN-ZVI, M.; GRIFFITHS, T.; STEYVERS, M.; AND SMYTH, P., 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in Artificial Intelligence*, 487–494. (cited on pages 123 and 150)

ROY, D. AND TEH, Y. W., 2009. The Mondrian process. In *Advances in Neural Information Processing Systems*. (cited on pages 6, 7, and 11)

RUBIN, T.; CHAMBERS, A.; SMYTH, P.; AND STEYVERS, M., 2011. Statistical topic models for multi-label document classification. Technical Report arXiv:1107.2462v2, University of California, Irvine, ASA. (cited on page 85)

SATO, I. AND NAKAGAWA, H., 2010. Topic models with Power-law using Pitman-Yor process. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. (cited on page 164)

SATO, K., 1990. *Levy Processes and Infinitely Divisible Distributions*. Cambridge University Press. (cited on page 63)

SERFOSO, R., 1999. *Introduction to Stochastic Networks*. Springer, New York. (cited on page 29)

SETHURAMAN, J., 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4 (1994), 639–650. (cited on pages 3 and 65)

SLIVNYAK, I., 1962. Some properties of stationary flows of homogeneous random events. *Theory of Probability and Its Applications*, (1962). (cited on page 29)

SMOLA, A. AND NARAYANAMURTHY, S., 2010. An architecture for parallel topic models. In *International Conference on Very Large Data Bases*. (cited on page 166)

Sudderth, E.; Torralba, A.; Freeman, W.; and Willsky, A., 2005. Describing visual scenes using transformed Dirichlet processes. In *Advances in Neural Information Processing Systems*. (cited on page 66)

Sudderth, E. B. and Jordan, M. I., 2008. Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems*. (cited on pages 85 and 86)

Teh, Y., 2006a. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 985–992. (cited on pages 10, 44, 66, and 85)

Teh, Y. and Gorur, D., 2009. Indian buffet processes with Power-law behavior. In *Advances in Neural Information Processing Systems*. (cited on page 130)

Teh, Y. W., 2004. Nonparametric Bayesian mixture models–release 2.1. Technical report, UCL, UK, http://www.stats.ox.ac.uk/ teh/research/npbayes/npbayes-r21.tgz. (cited on page 92)

Teh, Y. W., 2006b. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore. (cited on pages 10, 44, 66, and 164)

Teh, Y. W., 2006c. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 985–992. http://www.aclweb.org/anthology/P/P06/P06-1124. (cited on page 164)

Teh, Y. W., 2007. Exponential families: Gaussian, Gaussian-Gamma, Gaussian-Wishart, Multinomial. Technical report, University College London, UK. (cited on pages 53 and 54)

Teh, Y. W.; Blundell, C.; and Elliott, L. T., 2011. Modelling genetic variations with fragmentation-coagulation processes. In *Advances in Neural Information Processing Systems*. (cited on pages 7, 12, and 78)

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M., 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101, 476 (2006), 1566–1581. (cited on pages 9, 17, 45, 62, 65, 66, 67, 68, 70, 72, 78, 88, 90, 91, 93, 122, and 164)

Thibaux, R. and Jordan, M. I., 2007. Hierarchical Beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*. (cited on page 5)

Vinh, N. X.; Epps, J.; and Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, , 11 (2010), 2837–2854. (cited on page 55)

Wallach, H.; Murray, I.; Salakhutdinov, R.; and Mimno, D., 2009. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning*, 672–679. (cited on page 72)

Welling, M. and Teh, Y. W., 2011. Bayesian learning via stochastic gradient Langevin dynamics. In *International conference on machine learning*. (cited on page 166)

Williamson, S. A.; MacEachern, S. N.; and Xing, E. P., 2013. Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems*. (cited on page 166)

Williamson, S. A.; Wang, C.; Heller, K. A.; and Blei, D., 2010. The IBP compound Dirichlet process and its application to focused topic modeling. In *International conference on machine learning (ICML)*. (cited on pages 130 and 153)

Wood, F.; Archambeau, C.; Gasthaus, J.; James, L. F.; and Teh, Y. W., 2009. A stochastic memoizer for sequence data. In *Proceedings of the International Conference on Machine Learning*. (cited on page 66)

Xu, Z.; Tresp, V.; Yu, K.; and Kriegel, H.-P., 2006. Infinite hidden relational models. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, 544–551. (cited on page 66)

Yano, T.; Cohen, W.; and Smith, N., 2009. Predicting response to political blog posts with topic models. In *Proc. of the NAACL-HLT*. (cited on page 91)

Zhang, J.; Song, Y.; Zhang, C.; and Liu, S., 2010. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. (cited on page 85)

Zhou, M.; Chen, H.; Paisley, J.; Ren, L.; Sapiro, G.; and Carin, L., 2009. Nonparametric Bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems*. (cited on page 11)

Zhu, J.; Zheng, X.; Zhou, L.; and Zhang, B., 2013. Scalable inference in max-margin supervised topic models. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. (cited on page 166)