Negative Binomial Process Count and Mixture Modeling

Mingyuan Zhou and Lawrence Carin

Abstract

The seemingly disjoint problems of count and mixture modeling are united under the negative binomial (NB) process. We reveal relationships between the Poisson, multinomial, gamma and Dirichlet distributions, and construct a Poisson-logarithmic bivariate count distribution that connects the NB and Chinese restaurant table distributions. Fundamental properties of the models are developed, and we derive efficient Bayesian inference. It is shown that with augmentation and normalization, the NB process and gamma-NB process can be reduced to the Dirichlet process and hierarchical Dirichlet process, respectively. These relationships highlight theoretical, structural and computational advantages of the NB process. A variety of NB processes including the beta-geometric, beta-NB, marked-beta-NB, marked-gamma-NB and zero-inflated-NB processes, with distinct sharing mechanisms, are also constructed. These models are applied to topic modeling, with connections made to existing algorithms under the Poisson factor analysis framework. Example results show the importance of inferring both the NB dispersion and probability parameters, which respectively govern the overdispersion level and variance-to-mean ratio for count modeling.

Index Terms

Beta process, Chinese restaurant process, completely random measures, clustering, count modeling, Dirichlet process, gamma process, hierarchical Dirichlet process, mixed membership modeling, mixture modeling, negative binomial process, Poisson factor analysis, Poisson process, topic modeling.

I. Introduction

Count data appear in many settings, such as modeling the number of insects in regions of interest [1], [2], predicting the number of motor insurance claims [3], [2] and modeling topics of document corpora [4], [5], [6], [7]. There has been increasing interest in count modeling using the Poisson process, geometric process [8], [9], [10], [11], [12] and recently the negative binomial (NB) process [7], [13], [14]. Notably, we have shown in [7] and further demonstrated

M. Zhou and L. Carin are with the Dept. of Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA.

in [14] that the NB process, originally constructed for count analysis, can be naturally applied for mixture modeling of *grouped* data $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_J$, where each group $\boldsymbol{x}_j = \{x_{ji}\}_{i=1,N_j}$.

Mixture modeling infers probability random measures to assign data points into clusters (mixture components), which is of particular interest to statistics and machine learning. Although the number of points assigned to clusters are counts, mixture modeling is not typically considered as a count-modeling problem. Clustering is often addressed under the Dirichlet-multinomial framework, using the Dirichlet process [15], [16], [17], [18], [19] as the prior distribution. With the Dirichlet multinomial conjugacy, Dirichlet process mixture models enjoy tractability because the posterior of the probability measure is still a Dirichlet process. Despite its popularity, it is well-known that the Dirichlet process is inflexible in that a single concentration parameter controls the variability of the mass around the mean [19], [20]; moreover, the inference of the concentration parameter is nontrivial, usually solved with the data augmentation method proposed in [17]. Using probability measures normalized from gamma processes, whose shape and scale parameters can both be adjusted, one may mitigate these disadvantages. However, it is not clear how the parameters of the normalized gamma process can still be inferred under the multinomial likelihood. For mixture modeling of grouped data, the hierarchical Dirichlet process (HDP) [21] has been further proposed to share statistical strength between groups. However, the inference of the HDP is a challenge and it is usually solved under alternative constructions, such as the Chinese restaurant franchise and stick-breaking representations [21], [22], [23].

To construct more expressive mixture models, without losing the tractability of inference, in this paper we consider mixture modeling as a count-modeling problem. Directly modeling the counts assigned to clusters as NB random variables, we perform joint count and mixture modeling via the NB process, using completely random measures [24], [8], [25], [20] that are simple to construct and amenable for posterior computation. We reveal relationships between the Poisson, multinomial, gamma, and Dirichlet distributions and their corresponding stochastic processes, and we connect the NB and Chinese restaurant table distributions under a Poisson-logarithmic bivariate count distribution. We develop data augmentation methods unique to the NB distribution and augment a NB process into both the gamma-Poisson and compound Poisson representations, yielding unification of count and mixture modeling, derivation of fundamental model properties, as well as efficient Bayesian inference using Gibbs sampling.

Compared to the Dirichlet-multinomial framework, the proposed NB process framework provides new opportunities for better data fitting, efficient inference and flexible model constructions. We make four additional contributions: 1) we construct a NB process and a gamma-NB process, analyze their properties and show how they can be reduced to the Dirichlet process and the HDP, respectively, with augmentation and then normalization. We highlight their unique theoretical, structural and computational advantages relative to the Dirichlet-multinomial framework. 2) We show that a variety of NB processes can be constructed with distinct model properties, for which the shared random measure can be selected from completely random measures such as the gamma, beta, and beta-Bernoulli processes. 3) We show NB processes can be related to previously proposed discrete latent variable models under the Poisson factor analysis framework. 4) We apply NB processes to topic modeling, a typical example for mixture modeling of grouped data, and show the importance of inferring both the NB dispersion and probability parameters, which respectively govern the overdispersion level and the variance-to-mean ratio in count modeling.

Parts of the work presented here first appeared in our papers [7], [2], [14]. In this paper, we unify related materials scattered in these three papers and provide significant expansions. New materials include: we construct a Poisson-logarithmic bivariate count distribution that tightly connects the NB, Chinese restaurant table, Poisson and logarithmic distributions, extending the Chinese restaurant process to describe the case that both the number of customers and the number of tables are random variables. We show how to derive closed-form Gibbs sampling for hierarchical NB count models that can share statistical strength in multiple levels. We prove that under certain parameterizations, a Dirichlet process can be scaled with an independent gamma random variable to recover a Gamma process, which is further exploited to connect the NB and Dirichlet processes, and the gamma-NB process and HDP. We show that in Dirichlet process based models, the number of points assigned to a cluster is marginally beta-binomial distributed, distinct from the NB distribution used in NB process based models. In the experiments, we provide a comprehensive comparison of various NB process topic models and related algorithms, and make it clear that key to constructing a successful mixture model is appropriately modeling the distribution of counts, preferably to adjust two parameters for each count to achieve a balanced fit of both the mean and variance.

We mention that beta-NB processes have been independently investigated for mixed membership modeling in [13], with several notable distinctions: we study the properties and inference of the NB distribution in depth and emphasize the importance of learning the NB dispersion parameter, whereas in [13], the NB dispersion parameter is empirically set proportional to the group size N_j in a group dependent manner. We discuss a variety of NB processes, with beta-NB processes independently studied in [7] and [13] as special cases. We show the gamma-Poisson process can be marginalized as a NB process and normalized as a Dirichlet process, thus suitable for mixture modeling but restrictive for mixed membership modeling, as also confirmed by our experimental results; whereas in [13], the gamma-Poisson process is treated parallel to the beta-NB process and considered suitable for mixed membership modeling. We treat the beta-NB process parallel to the gamma-NB process, which can be augmented and then normalized as an HDP; whereas in [13], the beta-NB process is considered less flexible than the HDP, motivating the construction of a hierarchical-beta-NB process.

II. Preliminaries

A. *Completely Random Measures*

Following [20], for any $\nu^+ \geq 0$ and any probability distribution $\pi(dpd\omega)$ on the product space $\mathbb{R}\times\Omega$, let $K \sim \text{Pois}(\nu^+)$ and $\{(p_k,\omega_k)\}_{1,K} \overset{iid}{\sim} \pi(dpd\omega)$. Defining $\mathbf{1}_A(\omega_k)$ as being one if $\omega_k \in A$ and zero otherwise, the random measure $\mathcal{L}(A) \equiv \sum_{k=1}^K \mathbf{1}_A(\omega_k)p_k$ assigns independent infinitely divisible random variables $\mathcal{L}(A_i)$ to disjoint Borel sets $A_i \subset \Omega$, with characteristic functions

$$E\big[e^{it\mathcal{L}(A)}\big] = \exp\Big\{\int\int_{\mathbb{R}\times A}(e^{itp}-1)\nu(dpd\omega)\Big\} \tag{1}$$

with $\nu(dpd\omega) \equiv \nu^+\pi(dpd\omega)$. A random signed measure $\mathcal{L}$ satisfying (1) is called a Lévy random measure. More generally, if the Lévy measure $\nu(dpd\omega)$ satisfies

$$\int\int_{\mathbb{R}\times S}\min\{1,|p|\}\nu(dpd\omega) < \infty \tag{2}$$

for each compact $S \subset \Omega$, the Lévy random measure $\mathcal{L}$ is well defined, even if the Poisson intensity ν^+ is infinite. A nonnegative Lévy random measure $\mathcal{L}$ satisfying (2) was called a completely random measure in [24], [8] and an additive random measure in [26]. It was introduced for machine learning in [27] and [25].

1) Poisson Process: Define a Poisson process $X \sim \text{PP}(G_0)$ on the product space $\mathbb{Z}_+ \times \Omega$, where $\mathbb{Z}_+ = \{0,1,\cdots\}$, with a finite continuous base measure G_0 over Ω, such that $X(A) \sim \text{Pois}(G_0(A))$ for each subset $A \subset \Omega$. The Lévy measure of the Poisson process can be derived from (1) as $\nu(dud\omega) = \delta_1(du)G_0(d\omega)$, where $\delta_1(du)$ is a unit point mass at $u = 1$. If G_0 is discrete (atomic) as $G_0 = \sum_k \lambda_k\delta_{\omega_k}$, then the Poisson process definition is still valid that

$X = \sum_k x_k \delta_{\omega_k},\ x_k \sim \text{Pois}(\lambda_k)$. If G_0 is mixed discrete-continuous, then X is the sum of two independent contributions. As the discrete part is convenient to model, without loss of generality, below we consider the base measure to be continuous and finite.

2) Gamma Process: We define a gamma process [9], [20] $G \sim \text{GaP}(c, G_0)$ on the product space $\mathbb{R}_+ \times \Omega$, where $\mathbb{R}_+ = \{x : x \geq 0\}$, with concentration parameter c and base measure G_0, such that $G(A) \sim \text{Gamma}(G_0(A), 1/c)$ for each subset $A \subset \Omega$, where $\text{Gamma}(\lambda; a, b) = \frac{1}{\Gamma(a)b^a}\lambda^{a-1}e^{-\frac{\lambda}{b}}$ and $\Gamma(\cdot)$ denotes the gamma function. The gamma process is a completely random measure, whose Lévy measure can be derived from (1) as

$$\nu(drd\omega) = r^{-1}e^{-cr}drG_0(d\omega). \tag{3}$$

Since the Poisson intensity $\nu^+ = \nu(R^+ \times \Omega) = \infty$ and $\int\int_{\mathbb{R}_+\times\Omega} r\nu(drd\omega)$ is finite, there are countably infinite points and a draw from the gamma process can be expressed as

$$G = \textstyle\sum_{k=1}^{\infty} r_k \delta_{\omega_k},\ (r_k, \omega_k) \overset{iid}{\sim} \pi(drd\omega),\ \pi(drd\omega)\nu^+ \equiv \nu(drd\omega). \tag{4}$$

3) Beta Process: The beta process was defined by [28] for survival analysis with $\Omega = \mathbb{R}_+$. Thibaux and Jordan [27] generalized the process to an arbitrary measurable space Ω by defining a completely random measure B on the product space $[0, 1] \times \Omega$ with Lévy measure

$$\nu(dpd\omega) = cp^{-1}(1-p)^{c-1}dpB_0(d\omega). \tag{5}$$

Here $c > 0$ is a concentration parameter and B_0 is a base measure over Ω. Since the Poisson intensity $\nu^+ = \nu([0, 1] \times \Omega) = \infty$ and $\int\int_{[0,1]\times\Omega} p\nu(dpd\omega)$ is finite, there are countably infinite points and a draw from the beta process $B \sim \text{BP}(c, B_0)$ can be expressed as

$$B = \textstyle\sum_{k=1}^{\infty} p_k \delta_{\omega_k},\ (p_k, \omega_k) \overset{iid}{\sim} \pi(dpd\omega),\ \pi(dpd\omega)\nu^+ \equiv \nu(dpd\omega). \tag{6}$$

B. Dirichlet Process and Chinese Restaurant Process

1) Dirichlet Process: Denote $\widetilde{G} = G/G(\Omega)$, where $G \sim \text{GaP}(c, G_0)$, then for any measurable disjoint partition $A_1, \cdots, A_Q$ of Ω, we have $\left[\widetilde{G}(A_1), \cdots, \widetilde{G}(A_Q)\right] \sim \text{Dir}\left(\gamma_0\widetilde{G}_0(A_1), \cdots, \gamma_0\widetilde{G}_0(A_Q)\right)$, where $\gamma_0 = G_0(\Omega)$ and $\widetilde{G}_0 = G_0/\gamma_0$. Therefore, with a space invariant concentration parameter, the normalized gamma process $\widetilde{G} = G/G(\Omega)$ is a Dirichlet process [15], [29] with concentration parameter γ_0 and base probability measure $\widetilde{G}_0$, expressed as $\widetilde{G} \sim \text{DP}(\gamma_0, \widetilde{G}_0)$. Unlike the gamma process, the Dirichlet process is no longer a completely random measure as the random variables $\{\widetilde{G}(A_q)\}$ for disjoint sets $\{A_q\}$ are negatively correlated.

2) Chinese Restaurant Process: In a Dirichlet process $\widetilde{G} \sim \text{DP}(\gamma_0, \widetilde{G}_0)$, we assume $X_i \sim \widetilde{G}$; $\{X_i\}$ are independent given $\widetilde{G}$ and hence exchangeable. The predictive distribution of a new data point X_{m+1}, conditioning on $X_1, \cdots, X_m$, with $\widetilde{G}$ marginalized out, can be expressed as

$$X_{m+1}|X_1, \cdots, X_m \sim \mathbb{E}\left[\widetilde{G}\middle| X_1, \cdots, X_m\right] = \textstyle\sum_{k=1}^{K} \frac{n_k}{m+\gamma_0}\delta_{\omega_k} + \frac{\gamma_0}{m+\gamma_0}\widetilde{G}_0 \tag{7}$$

where $\{\omega_k\}_{1,K}$ are discrete atoms in Ω observed in $X_1, \cdots, X_m$ and $n_k = \sum_{i=1}^{m} X_i(\omega_k)$ is the number of data points associated with ω_k. The stochastic process described in (7) is known as the Pólya urn scheme [30] and also the Chinese restaurant process [31], [21], [32].

3) Chinese Restaurant Table Distribution: Under the Chinese restaurant process metaphor, the number of customers (data points) m is assumed to be known whereas the number of nonempty tables (distinct atoms) K is treated as a random variable dependent on m and γ_0. Denote $s(m, l)$ as Stirling numbers of the first kind, it is shown in [16] that the random table count K has PMF

$$\text{Pr}(K = l|m, \gamma_0) = \tfrac{\Gamma(\gamma_0)}{\Gamma(m+\gamma_0)}|s(m, l)|\gamma_0^l, \quad l = 0, 1, \cdots, m. \tag{8}$$

We refer to this distribution as the Chinese restaurant table (CRT) distribution and denote $l \sim \text{CRT}(m, \gamma_0)$ as a CRT random variable. As shown in Appendix A, it can be sampled as $l = \sum_{n=1}^{m} b_n,\ b_n \sim \text{Bernoulli}\left(\gamma_0/(n - 1 + \gamma_0)\right)$ or by iteratively calculating out the PMF under the logarithmic scale. The PMF of the CRT distribution has been used to help infer the concentration parameter γ_0 in Dirichlet processes [17], [21]. Below we explicitly relate the CRT and NB distributions under a Poisson-logarithmic bivariate count distribution.

III. Inference for the Negative Binomial Distribution

The Poisson distribution $m \sim \text{Pois}(\lambda)$ is commonly used to model count data. It has probability mass function (PMF) $f_M(m) = \frac{e^{-\lambda}\lambda^m}{m!},\ m \in \mathbb{Z}_+$, with both the mean and variance equal to λ. Due to heterogeneity (difference between individuals) and contagion (dependence between the occurrence of events), count data are usually overdispersed in that the variance is greater than the mean, making the Poisson assumption restrictive. By placing a gamma distribution prior with shape r and scale $p/(1 - p)$ on λ as $m \sim \text{Pois}(\lambda),\ \lambda \sim \text{Gamma}\left(r, \frac{p}{1-p}\right)$ and marginalizing out λ, a negative binomial (NB) distribution $m \sim \text{NB}(r, p)$ is obtained, with PMF

$$f_M(m) = \textstyle\int_0^\infty \text{Pois}(m; \lambda)\text{Gamma}\left(\lambda; r, \frac{p}{1-p}\right) d\lambda = \frac{\Gamma(r+m)}{m!\Gamma(r)}(1 - p)^r p^m \tag{9}$$

where r is the nonnegative dispersion parameter and p is the probability parameter. Thus the NB distribution is also known as the gamma-Poisson mixture distribution [33]. It has a mean

$\mu = rp/(1-p)$ smaller than the variance $\sigma^2 = rp/(1-p)^2 = \mu + r^{-1}\mu^2$, with the variance-to-mean ratio (VMR) as $(1-p)^{-1}$ and the overdispersion level (ODL, the coefficient of the quadratic term in σ^2) as r^{-1}, and thus it is usually favored over the Poisson distribution for modeling overdispersed counts. As shown in [34], $m \sim \mathrm{NB}(r,p)$ can also be generated from a compound Poisson distribution as

$$m = \textstyle\sum_{t=1}^{l} u_t,\ u_t \overset{iid}{\sim} \mathrm{Log}(p),\ l \sim \mathrm{Pois}(-r\ln(1-p)) \tag{10}$$

where $u \sim \mathrm{Log}(p)$ corresponds to the logarithmic distribution [35], [36] with PMF $f_U(k) = -p^k/[k\ln(1-p)],\quad k \in \{1,2,\dots\}$, and probability-generating function (PGF)

$$C_U(z) = \ln(1-pz)/\ln(1-p), \quad |z| < p^{-1}. \tag{11}$$

In a slight abuse of notation, but for added conciseness, in the following discussion we use $m \sim \sum_{t=1}^{l} \mathrm{Log}(p)$ to denote $m = \sum_{t=1}^{l} u_t,\ u_t \sim \mathrm{Log}(p)$.

The NB distribution has been widely investigated and applied to numerous scientific studies [1], [37], [38], [39]. Although inference of the NB probability parameter p is straightforward, as the beta distribution is its conjugate prior, inference of the NB dispersion parameter r, whose conjugate prior is unknown, has long been a challenge. The maximum likelihood (ML) approach is commonly used to estimate r, however, it only provides a point estimate and does not allow the incorporation of prior information; moreover, the ML estimator of r often lacks robustness and may be severely biased or even fail to converge, especially if the sample size is small [40], [41], [42], [43], [44], [45]. Bayesian approaches are able to model the uncertainty of estimation and incorporate prior information, however, the only available closed-form Bayesian inference for r relies on approximating the ratio of two gamma functions [46].

Lemma III.1. *Augment $m \sim \mathrm{NB}(r,p)$ under the compound Poisson representation as $m \sim \sum_{t=1}^{l} \mathrm{Log}(p),\ l \sim \mathrm{Pois}(-r\ln(1-p))$, then the conditional posterior of l given m and r can be expressed as $l|m,r \sim \mathrm{CRT}(m,r)$.*

Proof: Let $m \sim \mathrm{SumLog}(l,p)$ be the sum-logarithmic distribution as $m \sim \sum_{t=1}^{l} \mathrm{Log}(p)$. Since m is the summation of l iid $\mathrm{Log}(p)$ random variables, its PGF becomes $C_M(z) = C_U^l(z) = [\ln(1-pz)/\ln(1-p)]^l,\ |z| < p^{-1}$. Using the properties that $[\ln(1+x)]^l = l!\sum_{n=l}^{\infty} s(n,l)x^n/n!$ and $|s(m,l)| = (-1)^{m-l}s(m,l)$ [35], we have the PMF of $m \sim \mathrm{SumLog}(l,p)$ as

$$f_M(m|l,p) = \tfrac{C_M^{(m)}(0)}{m!} = \tfrac{p^l l! |s(m,l)|}{m![-\ln(1-p)]^l}. \tag{12}$$

Let $(m,l)\sim\text{PoisLog}(r,p)$ be the Poisson-logarithmic bivariate count distribution that describes the joint distribution of counts m and l as $m\sim\sum_{t=1}^{l}\text{Log}(p),\ l\sim\text{Pois}(-r\ln(1-p))$. Since $f_{M,L}(m,l|r,p)=f_M(m|l,p)f_L(l|r,p)$, we have the PMF of $(m,l)\sim\text{PoisLog}(r,p)$ as

$$f_{M,L}(m,l|r,p)=\tfrac{p^l l!|s(m,l)|}{m![-\ln(1-p)]^l}\tfrac{(-r\ln(1-p))^l e^{r\ln(1-p)}}{l!}=\tfrac{|s(m,l)|r^l}{m!}(1-p)^r p^m. \tag{13}$$

Since $f_{M,L}(m,l|r,p)=f_L(l|m,r)f_M(m|r,p)$, we have

$$f_L(l|m,r)\ \ =\tfrac{f_{M,L}(m,l|r,p)}{f_M(m|r,p)}=\tfrac{|s(m,l)|r^l(1-p)^r p^m}{m!\text{NB}(m;r,p)}=\tfrac{\Gamma(r)}{\Gamma(m+r)}|s(m,l)|r^l$$

which is exactly the same as the PMF of the CRT distribution shown in (8). ■

Corollary III.2. *The Poisson-logarithmic bivariate count distribution with PMF* $f_{M,L}(m,l|r,p)=\frac{|s(m,l)|r^l}{m!}(1-p)^r p^m$ *can be expressed as the product of a CRT and a NB distributions and also the product of a sum-logarithmic and a Poisson distributions as*

$$\text{PoisLog}(m,l;r,p)=\text{CRT}(l;m,r)\text{NB}(m;r,p)=\text{SumLog}(m;l,p)\text{Pois}(l;-r\ln(1-p)). \tag{14}$$

Under the Chinese restaurant process metaphor, the CRT distribution describes the random number of tables occupied by a given number of customers. Using Corollary III.2, we may have the metaphor that the Poisson-Logarithmic bivariate count distribution describes the joint distribution of count random variables m and l under two equivalent circumstances: 1) there are $l\sim\text{Pois}(-r\ln(1-p))$ tables with $m\sim\sum_{t=1}^{l}\text{Log}(p)$ customers; 2) there are $m\sim\text{NB}(r,p)$ customers seated on $l\sim\text{CRT}(m,r)$ tables. Note that according to Corollary A.2 in Appendix A, in a Chinese restaurant with concentration parameter r, around $r\ln\frac{m+r}{r}$ tables would be required to accommodate m customers.

Lemma III.3. *Let* $m\sim\text{NB}(r,p),\ r\sim\text{Gamma}(r_1,1/c_1)$ *represent the gamma-NB distribution, denote* $p'=\frac{-\ln(1-p)}{c_1-\ln(1-p)}$*, then* m *can also be generated from a compound distribution as*

$$m\sim\sum_{t=1}^{l}\text{Log}(p),\ l\sim\sum_{t'=1}^{l'}\text{Log}(p'),\ l'\sim\ \text{Pois}(-r_1\ln(1-p')) \tag{15}$$

which is equivalent in distribution to

$$m\sim\sum_{t=1}^{l}\text{Log}(p),\ l'\sim\ \text{CRT}(l,r_1),\ l\sim\text{NB}(r_1,p'). \tag{16}$$

Proof: We can augment the gamma-NB distribution as $m\sim\sum_{t=1}^{l}\text{Log}(p),\ l\sim\text{Pois}(-r\ln(1-p)),\ r\sim\text{Gamma}(r_1,1/c_1)$. Marginalizing out r leads to $m\sim\sum_{t=1}^{l}\text{Log}(p),\ l\sim\text{NB}\left(r_1,p'\right)$. Augmenting l using its compound Poisson representation leads to (15). Using Corollary III.2, we have that (15) and (16) are equivalent in distribution. ■

Using Corollary III.2, it is evident that to infer the NB dispersion parameter, we can place a gamma prior on it as $r \sim \text{Gamma}(r_1, 1/c_1)$; with the latent count $l \sim \text{CRT}(m, r)$ and the gamma-Poisson conjugacy, we can update r with a gamma posterior. We may further let $r_1 \sim \text{Gamma}(r_2, 1/c_2)$; using Lemma III.3, it is evident that with the latent count $l' \sim \text{CRT}(l, r_1)$, we can also update r_1 with a gamma posterior. Using Corollary III.2 and Lemma III.3, we can continue this process repeatedly, suggesting that for data that have subgroups within groups, we may build a hierarchical model to share statistical strength in multiple levels, with tractable inference. To be more specific, assuming we have counts $\{m_{j1}, \cdots, m_{jN_j}\}_{j=1,J}$ from J data groups; to model their distribution, we construct a hierarchical model as

$$m_{ji} \sim \text{NB}(r_j, p_j),\ p_j \sim \text{Beta}(a_0, b_0),\ r_j \sim \text{Gamma}(r_1, 1/c_1),\ r_1 \sim \text{Gamma}(r_2, 1/c_2).$$

Then Gibbs sampling proceeds as

$$(p_j|-) \sim \text{Beta}\left(a_0 + \textstyle\sum_{i=1}^{N_j} m_{ji}, b_0 + N_j r_j\right),\ p'_j = \tfrac{-N_j \ln(1-p_j)}{c_1 - N_j \ln(1-p_j)}$$

$$(l_{ji}|-) \sim \text{CRT}(m_{ji}, r_j),\ (l'_j|-) \sim \text{CRT}\left(\textstyle\sum_{i=1}^{N_j} l_{ji}, r_1\right)$$

$$r_1 \sim \text{Gamma}\left(r_2 + \textstyle\sum_{j=1}^{J} l'_j, \tfrac{1}{c_2 - \sum_{j=1}^{J} \ln(1-p'_j)}\right),\ r_j \sim \text{Gamma}\left(r_1 + \textstyle\sum_{i=1}^{N_j} l_{ji}, \tfrac{1}{c_1 - N_j \ln(1-p_j)}\right).$$

The conditional posterior of the latent count l was first derived by us in [2] and its analytical form as the CRT distribution was first discovered by us in [14]. Here we provide a more comprehensive study to reveal connections between various discrete distributions. These connections are key ingredients of this paper, which not only allow us to unite count and mixture modeling and derive efficient inference, but also, as shown in Sections IV and V, let us examine the posteriors to understand fundamental properties of the NB processes, clearly revealing connections to previous nonparametric Bayesian mixture models.

IV. Negative Binomial Process Joint Count and Mixture Modeling

A. *Poisson Process for Joint Count and Mixture Modeling*

The Poisson distribution is commonly used for count modeling [47] and the multinomial distribution is usually considered for mixture modeling, and their conjugate priors are the gamma and Dirichlet distributions, respectively. To unite count modeling and mixture modeling, and to derive efficient inference, we show below the relationships between the Poisson and multinomial random variables and the gamma and Dirichlet random variables.

Lemma IV.1. *Suppose that $x_1, \dots, x_K$ are independent Poisson random variables with*

$$x_k \sim \mathrm{Pois}(\lambda_k), \quad x = \textstyle\sum_{k=1}^K x_k.$$

Set $\lambda = \sum_{k=1}^K \lambda_k$; let $(y, y_1, \dots, y_K)$ be random variables such that

$$y \sim \mathrm{Pois}(\lambda), \ (y_1, \dots, y_k)|y \sim \mathrm{Mult}\left(y; \tfrac{\lambda_1}{\lambda}, \dots, \tfrac{\lambda_K}{\lambda}\right).$$

Then the distribution of $\boldsymbol{x} = (x, x_1, \dots, x_K)$ is the same as the distribution of $\boldsymbol{y} = (y, y_1, \dots, y_K)$ [7].

Corollary IV.2. *Let $X \sim \mathrm{PP}(G)$ be a Poisson process defined on a completely random measure G such that $X(A) \sim \mathrm{Pois}(G(A))$ for each $A \subset \Omega$. Define $Y \sim \mathrm{MP}(Y(\Omega), \frac{G}{G(\Omega)})$ as a multinomial process, with total count $Y(\Omega)$ and base probability measure $\frac{G}{G(\Omega)}$, such that $(Y(A_1), \cdots, Y(A_Q)) \sim \mathrm{Mult}\Big(Y(\Omega); \frac{G(A_1)}{G(\Omega)}, \cdots, \frac{G(A_Q)}{G(\Omega)}\Big)$ for any disjoint partition $\{A_q\}_{1,Q}$ of Ω; let $Y(\Omega) \sim \mathrm{Pois}(G(\Omega))$. Since $X(A)$ and $Y(A)$ have the same Poisson distribution for each $A \subset \Omega$, X and Y are equivalent in distribution.*

Lemma IV.3. *Suppose that random variables y and $(y_1, \dots, y_K)$ are independent with $y \sim \mathrm{Gamma}(\gamma, 1/c), \quad (y_1, \dots, y_K) \sim \mathrm{Dir}(\gamma p_1, \cdots, \gamma p_K)$, where $\sum_{k=1}^K p_k = 1$. Let $x_k = y y_k$, then $\{x_k\}_{1,K}$ are independent gamma random variables with $x_k \sim \mathrm{Gamma}(\gamma p_k, 1/c)$.*

Proof: Since $x_K = y(1 - \sum_{k=1}^{K-1} y_k)$ and $\left|\frac{\partial(y_1, \cdots, y_{K-1}, y)}{\partial(x_1, \cdots, x_K)}\right| = y^{-K+1}$, we have $f_{X_1, \cdots, X_K}(x_1, \cdots, x_K) = f_{Y_1, \cdots, Y_{K-1}}(y_1, \cdots, y_{K-1}) f_Y(y) y^{-K+1} = \prod_{k=1}^K \mathrm{Gamma}(x_k; \gamma p_k, 1/c)$. ■

Corollary IV.4. *If the gamma random variable α and the Dirichlet process $\widetilde{G}$ are independent with $\alpha \sim \mathrm{Gamma}(\gamma_0, 1/c), \quad \widetilde{G} \sim \mathrm{DP}(\gamma_0, \widetilde{G}_0)$, where $\gamma_0 = G_0(\Omega)$ and $\widetilde{G}_0 = G_0/\gamma_0$, then the product $G = \alpha\widetilde{G}$ is a gamma process with $G \sim \mathrm{GaP}(c, G_0)$.*

Using Corollary IV.2, we illustrate how the seemingly distinct problems of count and mixture modeling can be united under the Poisson process. Denote Ω as a measure space and for each Borel set $A \subset \Omega$, denote $X_j(A)$ as a count random variable describing the number of observations in $\boldsymbol{x}_j$ that reside within A. Given grouped data $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_J$, for any measurable disjoint partition $A_1, \cdots, A_Q$ of Ω, we aim to jointly model the count random variables $\{X_j(A_q)\}$. A natural choice would be to define a Poisson process $X_j \sim \mathrm{PP}(G)$, with a shared completely random measure G on Ω, such that $X_j(A) \sim \mathrm{Pois}\big(G(A)\big)$ for each $A \subset \Omega$ and $G(\Omega) = \sum_{q=1}^Q G(A_q)$. Following Corollary IV.2, letting $X_j \sim \mathrm{PP}(G)$ is equivalent to letting

$$X_j \sim \mathrm{MP}(X_j(\Omega), \widetilde{G}), \quad X_j(\Omega) \sim \mathrm{Pois}(G(\Omega)) \tag{17}$$

where $\widetilde{G} = G/G(\Omega)$. Thus the Poisson process provides not only a way to generate independent counts from each A_q, but also a mechanism for mixture modeling, which allocates the $X_j(\Omega)$ observations into any measurable disjoint partition $\{A_q\}_{1,Q}$ of Ω, conditioning on the normalized mean measure $\widetilde{G}$. A distinction is that in most clustering models the number of observations $X_j(\Omega)$ is assumed given and $X_j(A) \sim \text{Binomial}(X_j(\Omega), \widetilde{G}(A))$, whereas here $X_j(\Omega)$ is modeled as a Poisson random variable and $X_j(A) \sim \text{Poisson}(G(A))$.

B. Gamma-Poisson Process and Dirichlet Process

To complete the Poisson process, we may place a gamma process prior on G as

$$X_j \sim \text{PP}(G),\ j = 1, \cdots, J,\ G \sim \text{GaP}(J(1-p)/p, G_0). \tag{18}$$

Here the base measures of the Poisson process (PP) and gamma process (GaP) are not restricted to be continuous. Marginalizing out G of the gamma-Poisson process leads to a NB process $X = \sum_{j=1}^{J} X_j \sim \text{NBP}(G_0, p)$, in which $X(A) \sim \text{NB}(G_0(A), p)$ for each $A \subset \Omega$. We comment here that when $J > 1$, i.e., there is more than one data group, one need avoid the mistake of marginalizing out G in $X_j \sim \text{PP}(G),\ G \sim \text{GaP}(c, G_0)$ as $X_j \sim \text{NBP}(G_0, 1/(c+1))$. The gamma-Poisson process has also been discussed in [9], [10], [11], [12] for count modeling. Here we show that it can be represented as a NB process, leading to fully tractable closed-form Bayesian inference, and we demonstrate that it can be natrually applied for mixture modeling.

Define $L \sim \text{CRTP}(X, G_0)$ as a CRT process that for each $A \subset \Omega$, $L(A) = \sum_{\omega \in A} L(\omega),\ L(\omega) \sim \text{CRT}(X(\omega), G_0(\omega))$. Under the Chinese restaurant process metaphor, $X(A)$ and $L(A)$ represent the customer count and table count, respectively, observed in each $A \subset \Omega$. Using Corollary III.2, their joint distribution is the same for: 1) first drawing $L(A) \sim \text{Pois}(-G_0(A)\ln(1-p))$ tables and then assigning $\text{Log}(p)$ customers to each table, with $X(A)$ total number of customers; 2) first drawing $X(A) \sim \text{NB}(G_0(A), p)$ customers and then assigning them into $L(A) \sim \sum_{\omega \in A} \text{CRT}(X(\omega), G_0(\omega))$ tables. Therefore, the NB process augmented under the compound Poisson representation as $X \sim \sum_{t=1}^{L} \text{Log}(p),\ L \sim \text{PP}(-G_0 \ln(1-p))$ is equivalent in distribution to $L \sim \text{CRTP}(X, G_0),\ X \sim \text{NBP}(G_0, p)$. These equivelent reprsentations allow us to derive closed-form Bayesian inference for the NB process.

If we impose a gamma prior $\text{Gamma}(e_0, 1/f_0)$ on $\gamma_0 = G_0(\Omega)$ and a beta prior $\text{Beta}(a_0, 1/b_0)$ on p, using the conjugacy between the gamma and Poisson and the beta and NB distributions, we have the conditional posteriors as

$$G|X,p,G_0 \sim \text{GaP}(J/p, G_0+X), \quad (p|X,G) \sim \text{Beta}(a_0+X(\Omega), b_0+\gamma_0)$$
$$L|X,G_0 \sim \text{CRTP}(X,G_0), \quad (\gamma_0|L,p) \sim \text{Gamma}\left(e_0+L(\Omega), \tfrac{1}{f_0-\ln(1-p)}\right). \tag{19}$$

If the base measure G_0 is continuous, then $G_0(\omega) \to 0$ and we have $L(\omega) \sim \text{CRT}(X(\omega), G_0(\omega)) = \delta(X(\omega)>0)$ and thus $L(\Omega) = \sum_{\omega\in\Omega} \delta(X(\omega)>0)$, i.e., the number of tables is equal to K^+, the number of observed discrete atoms. The gamma-Poisson process is also well defined with a discrete base measure as $G_0 = \sum_{k=1}^K \frac{\gamma_0}{K}\delta_{\omega_k}$, which becomes continuous only if $K\to\infty$. With such a base measure, we have $L = \sum_{k=1}^K l_k\delta_{\omega_k},\ l_k \sim \text{CRT}(X(\omega_k), \gamma_0/K)$; it becomes possible that $l_k > 1$ if $X(\omega_k) > 1$, which means $L(\Omega) \geq K^+$. Thus when G_0 is discrete, using the number of observed atoms K^+ instead of the the number of tables $L(\Omega)$ to update the mass parameter γ_0 may lead to a biased estimate, especially if K is not sufficiently large.

Based on Corollaries IV.2 and IV.4, the gamma-Poisson process is equivalent to

$$X_j \sim \text{MP}(X_j(\Omega), \widetilde{G}),\ \widetilde{G} \sim \text{DP}(\gamma_0, \widetilde{G}_0),\ X_j(\Omega) \sim \text{Pois}(\alpha),\ \alpha \sim \text{Gamma}(\gamma_0, p/(J(1-p))) \tag{20}$$

where $G = \alpha\widetilde{G}$ and $G_0 = \gamma_0\widetilde{G}_0$. Thus without modeling $X_j(\Omega)$ and $G(\Omega) = \alpha$ as random variables, the gamma-Poisson process becomes the Dirichlet process, which is widely used for mixture modeling [15], [17], [29], [19]. Note that for the Dirichlet process, no analytical forms are available for the conditional posterior of γ_0 when $\widetilde{G}_0$ is continuous [17] and no rigorous inference for γ_0 is available when $\widetilde{G}_0$ is discrete. Whereas for the proposed gamma-Poisson process augmented from the NB process, as shown in (19), the conditional posteriors are analytic regardless of whether the base measure G_0 is continuous or discrete.

C. Block Gibbs Sampling for the Negative Binomial Process

For a finite continuous base measure, a draw from the gamma process $G \sim \text{GaP}(c, G_0)$ can be expressed as an infinite sum as in (4). Here we consider a discrete base measure as $G_0 = \sum_{k=1}^K \frac{\gamma_0}{K}\delta_{\omega_k}$, then we have $G = \sum_{k=1}^K r_k\delta_{\omega_k},\ r_k \sim \text{Gamma}(\gamma_0/K, 1/c),\ \omega_k \sim g_0(\omega_k)$, which becomes a draw from the gamma process with a continuous base measure as $K\to\infty$. Let $x_{ji} \sim F(\omega_{z_{ji}})$ be observation i in group j, linked to a mixture component $\omega_{z_{ji}} \in \Omega$ through a distribution F. Denote $n_{jk} = \sum_{i=1}^{N_j}\delta(z_{ji}=k)$, we can express the NB process as

$$x_{ji} \sim F(\omega_{z_{ji}}),\ \omega_k \sim g_0(\omega_k),\ N_j = \textstyle\sum_{k=1}^K n_{jk},\ n_{jk} \sim \text{Pois}(r_k)$$
$$r_k \sim \text{Gamma}(\gamma_0/K, p/(J(1-p))),\ p \sim \text{Beta}(a_0,b_0),\ \gamma_0 \sim \text{Gamma}(e_0, 1/f_0) \tag{21}$$

where marginally we have $n_k = \sum_{j=1}^J n_{jk} \sim \mathrm{NB}(\gamma_0/K, p)$. Note that if $J > 1$, one need avoid marginalizing out r_k in $n_{jk} \sim \mathrm{Pois}(r_k),\ r_k \sim \mathrm{Gamma}(\gamma_0/K, 1)$ as $n_{jk} \sim \mathrm{NB}(\gamma_0/K, 1/2)$. Denote $r = \sum_{k=1}^K r_k$, using Lemma IV.1, we can equivalently express N_j and n_{jk} in (21) as

$$N_j \sim \mathrm{Pois}\left(r\right),\ \left(n_{j1}, \cdots, n_{jK}\right) \sim \mathrm{Mult}\left(N_j; r_1/r, \cdots, r_K/r\right). \tag{22}$$

Since the data $\{x_{ji}\}_{i=1,N_j}$ are fully exchangeable, rather than drawing $(n_{j1}, \cdots, n_{jK})$ once, we may equivalently draw index z_{ji} for each x_{ji} and then calculate n_{jk} as

$$z_{ji} \sim \mathrm{Discrete}\left(r_1/r, \cdots, r_K/r\right), \quad n_{jk} = \textstyle\sum_{i=1}^{N_j} \delta(z_{ji} = k). \tag{23}$$

This provides further insights on how the seemingly disjoint problems of count and mixture modeling are united under the NB process framework. Following (19), the block Gibbs sampling is straightforward to write as

$$\begin{aligned}
&\mathrm{Pr}(z_{ji} = k|-) \propto F(x_{ji}; \omega_k) r_k, \quad (p|-) \sim \mathrm{Beta}\left(a_0 + \textstyle\sum_{j=1}^J N_j, b_0 + \gamma_0\right) \\
&(l_k|-) \sim \mathrm{CRT}(n_k, \gamma_0/K), \quad (\gamma_0|-) \sim \mathrm{Gamma}\left(e_0 + \textstyle\sum_{k=1}^K l_k, \frac{1}{f_0 - \ln(1-p)}\right) \\
&(r_k|-) \sim \mathrm{Gamma}\left(\gamma_0/K + n_k, p/J\right), \quad p(\omega_k|-) \propto \textstyle\prod_{z_{ji}=k} F(x_{ji}; \omega_k) g_0(\omega_k).
\end{aligned} \tag{24}$$

If $g_0(\omega)$ is conjugate to the likelihood $F(x; \omega)$, then the conditional posterior of ω would be analytic. Note that when $K \to \infty$, we have $(l_k|-) = \delta(n_k > 0)$ and then $\sum_{k=1}^K l_k = K^+$.

Without modeling N_j and r as random variables, we can re-express (21) as

$$x_{ji} \sim F(\omega_{z_{ji}}),\ z_{ji} \sim \mathrm{Discrete}(\tilde{\boldsymbol{r}}),\ \tilde{\boldsymbol{r}} \sim \mathrm{Dir}(\gamma_0/K, \cdots, \gamma_0/K),\ \gamma_0 \sim \mathrm{Gamma}(e_0, 1/f_0) \tag{25}$$

which loses the count modeling ability and becomes a finite representation of the Dirichlet process [15], [29]. The conditional posterior of $\tilde{\boldsymbol{r}}$ is analytic, whereas γ_0 can be sampled as in [17] when $K \to \infty$. This also implies that by using the Dirichlet process as the foundation, traditional mixture modeling may discard useful count information from the beginning.

The Poisson process has an equal-dispersion assumption for count modeling. For mixture modeling of grouped data, the gamma-Poisson process augmented from the NB process might be too restrictive in that, as shown in (20), it implies the same mixture proportions across groups. This motivates us to consider adding an additional layer into the gamma-Poisson process or using a different distribution other than the Poisson to model the counts for grouped data. As shown in Section III, the NB distribution is an ideal candidate, not only because it allows overdispersion, but also because it can be equivalently augmented into a gamma-Poisson and a compound Poisson representations; moreover, it can be used together with the CRT distribution to form a Poisson-logarithmic bivariate distribution to jointly model the counts of customers and tables.

V. Count and Mixture Modeling of Grouped Data

For joint count and mixture modeling of grouped data, we explore sharing the NB dispersion while the probability parameters are group dependent. We construct a gamma-NB process as

$$X_j \sim \text{NBP}(G, p_j),\ G \sim \text{GaP}(c, G_0). \tag{26}$$

Note that we may also let $X_j \sim \text{NBP}(\alpha_j G, p_j)$ and place a gamma prior on α_j to increase model flexibility, whose inference will be slightly more complicated and thus omitted here for brevity. The gamma-NB process can be augmented as a gamma-gamma-Poisson process as

$$X_j \sim \text{PP}(\Theta_j), \quad \Theta_j \sim \text{GaP}((1-p_j)/p_j, G), \quad G \sim \text{GaP}(c, G_0). \tag{27}$$

This construction introduces gamma random measures Θ_j based on G, which are essential to construct group-level probability measures $\widetilde{\Theta}_j$ to assign observations to mixture components. The gamma-NB process can also be augmented under the compound Poisson representation as

$$X_j \sim \textstyle\sum_{t=1}^{L_j} \text{Log}(p_j),\ L_j \sim \text{PP}(-G\ln(1-p_j)),\ G \sim \text{GaP}(c, G_0) \tag{28}$$

which, using Corollary III.2, is equivalent in distribution to

$$L_j \sim \text{CRTP}(X_j, G),\ X_j \sim \text{NBP}(G, p_j),\ G \sim \text{GaP}(c, G_0). \tag{29}$$

Using Lemma III.3 and Corollary III.2, we further have two equivalent augmentations as

$$L \sim \textstyle\sum_{t=1}^{L'} \text{Log}(p'),\ L' \sim \text{PP}(-G_0\ln(1-p')), \quad p' = \frac{-\sum_j \ln(1-p_j)}{c - \sum_j \ln(1-p_j)}; \tag{30}$$

$$L' \sim \text{CRTP}(L, G_0),\ L \sim \text{NBP}(G_0, p') \tag{31}$$

where $L = \sum_j L_j$. These augmentations allow us to derive a sequence of closed-form update equations for inference with the gamma-NB process, as described below.

A. Model Properties

Let $p_j \sim \text{Beta}(a_0, b_0)$, using the beta NB conjugacy, we have

$$(p_j|-) \sim \text{Beta}\left(a_0 + X_j(\Omega), b_0 + G(\Omega)\right). \tag{32}$$

Using (29) and (31), we have

$$L_j|X_j, G \sim \text{CRTP}(X_j, G), \qquad L'|L, G_0 \sim \text{CRTP}(L, G_0). \tag{33}$$

If G_0 is continuous and finite, we have $G_0(\omega) \rightarrow 0 \ \forall \ \omega \in \Omega$ and thus $L'(\Omega)|L, G_0 = \sum_{\omega \in \Omega} \delta(L(\omega) > 0) = \sum_{\omega \in \Omega} \delta(\sum_j X_j(\omega) > 0) = K^+$; if G_0 is discrete as $G_0 = \sum_{k=1}^{K} \frac{\gamma_0}{K} \delta_{\omega_k}$, then $L'(\omega_k) = \text{CRT}(L(\omega_k), \frac{\gamma_0}{K}) \geq 1$ if $\sum_j X_j(\omega_k) > 0$, thus $L'(\Omega) \geq K^+$. In either case, let $\gamma_0 = G_0(\Omega) \sim \text{Gamma}(e_0, 1/f_0)$, with the gamma Poisson conjugacy on (28) and (30), we have

$$\gamma_0 | \{L'(\Omega), p'\} \sim \text{Gamma}\left(e_0 + L'(\Omega), \tfrac{1}{f_0 - \ln(1-p')}\right); \tag{34}$$

$$G | G_0, \{L_j, p_j\} \sim \text{GaP}\left(c - \textstyle\sum_j \ln(1-p_j), G_0 + L\right). \tag{35}$$

Using the gamma Poisson conjugacy on (27), we have

$$\Theta_j | G, X_j, p_j \sim \text{GaP}\left(1/p_j, G + X_j\right). \tag{36}$$

Since the data $\{x_{ji}\}_i$ are exchangeable within group j, the predictive distribution of a point X_{ji}, conditioning on $X_j^{-i} = \{X_{jn}\}_{n:n \neq i}$ and G, with Θ_j marginalized out, can be expressed as

$$X_{ji} | G, X_j^{-i} \sim \tfrac{\mathbb{E}[\Theta_j | G, X_j^{-i}]}{\mathbb{E}[\Theta_j(\Omega) | G, X_j^{-i}]} = \tfrac{G}{G(\Omega) + X_j(\Omega) - 1} + \tfrac{X_j^{-i}}{G(\Omega) + X_j(\Omega) - 1}. \tag{37}$$

B. Relationship with the Hierarchical Dirichlet Process

Based on Corollaries IV.2 and IV.4, we can equivalently express (27) as

$$X_j(\Omega) \sim \text{Pois}(\theta_j), \ \theta_j \sim \text{Gamma}(\alpha, p_j/(1-p_j)) \tag{38}$$

$$X_j \sim \text{MP}(X_j(\Omega), \widetilde{\Theta}_j), \ \widetilde{\Theta}_j \sim \text{DP}(\alpha, \widetilde{G}), \ \alpha \sim \text{Gamma}(\gamma_0, 1/c), \ \widetilde{G} \sim \text{DP}(\gamma_0, \widetilde{G}_0) \tag{39}$$

where $\Theta_j = \theta_j \widetilde{\Theta}_j$, $G = \alpha \widetilde{G}$ and $G_0 = \gamma_0 \widetilde{G}_0$. Without modeling $X_j(\Omega)$ and θ_j as random variables, (39) becomes an HDP [21]. Thus the augmented and then normalized gamma-NB process leads to an HDP. However, we cannot return from the HDP to the gamma-NB process without modeling $X_j(\Omega)$ and θ_j as random variables. Theoretically, they are distinct in that the gamma-NB process is a completely random measure, assigning independent random variables into any disjoint Borel sets $\{A_q\}_{1,Q}$ in Ω, and the count $X_j(A)$ has the distribution as $X_j(A) \sim \text{NB}(G(A), p_j)$; by contrast, due to normalization, the HDP is not, and marginally

$$X_j(A) \sim \text{Beta-Binomial}\big(X_j(\Omega), \alpha \widetilde{G}(A), \alpha(1 - \widetilde{G}(A))\big). \tag{40}$$

Practically, the gamma-NB process can exploit conjugacy to achieve analytical conditional posteriors for all latent parameters. The inference of the HDP is a challenge and it is usually solved through alternative constructions such as the Chinese restaurant franchise (CRF) and stick-breaking representations [21], [23]. In particular, without analytical conditional posteriors, the inference of concentration parameters α and γ_0 is nontrivial [21], [22] and they are often simply fixed [23]. Under the CRF metaphor α governs the random number of tables occupied by customers in each restaurant independently; further, if the base probability measure $\widetilde{G}_0$ is continuous, γ_0 governs the random number of dishes selected by tables of all restaurants. One may apply the data augmentation method of [17] to sample α and γ_0. However, if $\widetilde{G}_0$ is discrete as $\widetilde{G}_0 = \sum_{k=1}^K \frac{1}{K}\delta_{\omega_k}$, which is of practical value and becomes a continuous base measure as $K \to \infty$ [29], [21], [22], then using the method of [17] to sample γ_0 is only approximately correct, which may result in a biased estimate in practice, especially if K is not large enough.

By contrast, in the gamma-NB process, the shared gamma process G can be analytically updated with (35) and $G(\Omega)$ plays the role of α in the HDP, which is readily available as

$$G(\Omega)|G_0, \{L_j, p_j\}_{j=1,N} \sim \text{Gamma}\Big(\gamma_0 + \textstyle\sum_j L_j(\Omega), \frac{1}{c-\sum_j \ln(1-p_j)}\Big) \qquad (41)$$

and as in (34), regardless of whether the base measure is continuous, the total mass γ_0 has an analytical gamma posterior whose shape parameter is governed by $L'(\Omega)$, with $L'(\Omega) = K^+$ if G_0 is continuous and finite and $L'(\Omega) \geq K^+$ if $G_0 = \sum_{k=1}^K \frac{\gamma_0}{K}\delta_{\omega_k}$. Equation (41) also intuitively shows how the NB probability parameters $\{p_j\}$ govern the variations among $\{\widetilde{\Lambda}_j\}$ in the gamma-NB process. In the HDP, p_j is not explicitly modeled, and since its value becomes irrelevant when taking the normalized constructions in (39), it is usually treated as a nuisance parameter and perceived as $p_j = 0.5$ when needed for interpretation purpose. Fixing $p_j = 0.5$ is also considered in [48] to construct an HDP, whose group-level DPs are normalized from gamma processes with the scale parameters as $\frac{p_j}{1-p_j} = 1$; it is also shown in [48] that improved performance can be obtained for topic modeling by learning the scale parameters with a log Gaussian process prior. However, no analytical conditional posteriors are provided and Gibbs sampling is not considered as a viable option [48].

C. Block Gibbs Sampling for the Gamma-Negative Binomial Process

As with the NB process described in Section IV, with a discrete base measure as $G_0 = \sum_{k=1}^K \frac{\gamma_0}{K}\delta_{\omega_k}$, we can express the gamma-NB process as

$$x_{ji} \sim F(\omega_{z_{ji}}),\ \omega_k \sim g_0(\omega_k),\ N_j = \textstyle\sum_{k=1}^K n_{jk},\ n_{jk} \sim \text{Pois}(\theta_{jk}),\ \theta_{jk} \sim \text{Gamma}(r_k, p_j/(1-p_j))$$
$$r_k \sim \text{Gamma}(\gamma_0/K, 1/c),\ p_j \sim \text{Beta}(a_0, b_0),\ \gamma_0 \sim \text{Gamma}(e_0, 1/f_0) \qquad (42)$$

where marginally we have $n_{jk} \sim \text{NB}(r_k, p_j)$. Following Section V-A, the block Gibbs sampling for (42) is straightforward to write as

$$\begin{aligned}
\Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\theta_{jk}, \quad (l_{jk}|-) \sim \text{CRT}(n_{jk}, r_k), \quad (l'_k|-) \sim \text{CRT}\left(\textstyle\sum_j l_{jk}, \gamma_0/K\right) \\
(p_j|-) \sim \text{Beta}\left(a_0 + N_j, b_0 + \textstyle\sum_k r_k\right), \; p' = \tfrac{-\sum_j \ln(1-p_j)}{c-\sum_j \ln(1-p_j)} \\
(\gamma_0|-) \sim \text{Gamma}\left(e_0 + \textstyle\sum_k l'_k, \tfrac{1}{f_0 - \ln(1-p')}\right) \\
(r_k|-) \sim \text{Gamma}\left(\gamma_0/K + \textstyle\sum_j l_{jk}, \tfrac{1}{c-\sum_j \ln(1-p_j)}\right) \\
(\theta_{jk}|-) \sim \text{Gamma}(r_k + n_{jk}, p_j), \quad p(\omega_k|-) \propto \textstyle\prod_{z_{ji}=k} F(x_{ji};\omega_k) g_0(\omega_k)
\end{aligned} \tag{43}$$

which has similar computational complexity as that of the direct assignment block Gibbs sampling of the CRF-HDP [21], [22]. Note that when $K \to \infty$, we have $(l'_k|-) = \delta(\sum_j l_{jk} > 0) = \delta(\sum_j n_{jk} > 0)$ and thus $\sum_k l'_k = K^+$.

Without treating N_j and θ_j as random variables, we can re-express (42) as

$$z_{ji} \sim \text{Discrete}(\tilde{\boldsymbol{\theta}}_j), \; \tilde{\boldsymbol{\theta}}_j \sim \text{Dir}(\alpha\tilde{\boldsymbol{r}}), \; \alpha \sim \text{Gamma}(\gamma_0, 1/c), \; \tilde{\boldsymbol{r}} \sim \text{Dir}(\gamma_0/K, \cdots, \gamma_0/K) \tag{44}$$

which becomes a finite representation of the HDP, the inference of which is usually solved under the Chinese restaurant franchise [21], [22] or stick-breaking representations [23].

VI. The Negative Binomial Process Family

The gamma-NB process shares the NB dispersion across groups while the NB probability parameters are group dependent. Since the NB distribution has two adjustable parameters, we may explore alternative ideas, with the NB probability measure shared across groups as in [13], or with both the dispersion and probability measures shared as in [7]. These constructions are distinct from both the gamma-NB process and HDP in that Θ_j has space dependent scales, and thus its normalization $\widetilde{\Theta}_j = \Theta_j/\Theta_j(\Omega)$, which is still a probability measure, no longer follows a Dirichlet process.

It is natural to let the NB probability measure be drawn from the beta process [28], [27]. A beta-NB process [7], [13] can be constructed by letting $X_j \sim \text{NBP}(r_j, B), \; B \sim \text{BP}(c, B_0)$, with a random draw expressed as $X_j = \sum_{k=1}^{\infty} n_{jk}\delta_{\omega_k}, \; n_{jk} \sim \text{NB}(r_j, p_k)$. Under this construction, the NB probability measure is shared and the NB dispersion parameters are group dependent. Note that if r_j are fixed as one, then the beta-NB process reduces to the beta-geometric process, related to the one for count modeling discussed in [11]; if r_j are empirically set to some other values, then the beta-NB process reduces to the one proposed in [13]. These simplifications are not considered in the paper, as they are often overly restrictive. As in [7], we may also consider a marked-beta-NB process, with both the NB probability and dispersion measures shared, in which each point of the beta process is marked with an independent gamma random variable. Thus a draw from the marked-beta process becomes $(R, B) = \sum_{k=1}^{\infty}(r_k, p_k)\delta_{\omega_k}$, and the NB process $X_j \sim$ NBP(R, B) becomes $X_j = \sum_{k=1}^{\infty} n_{jk}\delta_{\omega_k},\ n_{jk} \sim$ NB(r_k, p_k). Since the beta and NB distributions are conjugate, the posterior of B is tractable, as shown in [7], [13]. Similar to the marked-beta-NB process, we may also consider a beta marked-gamma-NB process, whose performance is found to be similar. If it is believed that there are excessive number of zeros, governed by a process other than the NB process, we may introduce a zero inflated NB process as $X_j \sim$ NBP(RZ_j, p_j), where $Z_j \sim$ BeP(B) is drawn from the Bernoulli process [27] and $(R, B) = \sum_{k=1}^{\infty}(r_k, \pi_k)\delta_{\omega_k}$ is drawn from a marked-beta process, thus $n_{jk} \sim$ NB$(r_k b_{jk}, p_j),\ b_{jk} =$ Bernoulli(π_k). This construction can be linked to the model in [49] with appropriate normalization, with advantages that there is no need to fix $p_j = 0.5$ and the inference is fully tractable. The zero inflated construction can also be linked to models for real valued data using the Indian buffet process (IBP) or beta-Bernoulli process spike-and-slab prior [50], [51], [52], [53], [54], [55], [56], [57]. More details on the NB process family can be found in Appendix B.

VII. Negative Binomial Process Topic Modeling and Poisson Factor Analysis

We consider topic modeling (mixed membership modeling) of a document corpus, a special case of mixture modeling of grouped data, where the words of the jth document $x_{j1}, \cdots, x_{jN_j}$ constitutes a group $\boldsymbol{x}_j$ (N_j words in document j), each word x_{ji} is an exchangeable group member indexed by v_{ji} in a vocabulary with V unique terms. The likelihood $F(x_{ji}; \boldsymbol{\phi}_k)$ is simply $\phi_{v_{ji}k}$, the probability of word x_{ji} under topic (atom/factor) $\boldsymbol{\phi}_k \in \mathbb{R}^V$, with $\sum_{v=1}^{V} \phi_{vk} = 1$. We refer to NB process mixture modeling of grouped words $\{\boldsymbol{x}_j\}_{1,J}$ as NB process topic modeling.

Denote $n_{vjk} = \sum_{i=1}^{N_j} \delta(z_{ji} = k, v_{ji} = v)$, $n_{jk} = \sum_v n_{vjk}$, $n_{v\cdot k} = \sum_j n_{vjk}$ and $n_{\cdot k} = \sum_j n_{jk}$, and for modeling convenience, we place Dirichlet priors on topics $\boldsymbol{\phi}_k \sim$ Dir$(\eta, \cdots, \eta)$, then for block Gibbs sampling of the gamma-NB process in (43) with K atoms, we have

$$\text{Pr}(z_{ji} = k|-) = \frac{\phi_{v_{ji}k}\theta_{jk}}{\sum_{k=1}^{K} \phi_{v_{ji}k}\theta_{jk}} \tag{45}$$

$$(\boldsymbol{\phi}_k|-) \sim \text{Dir}\left(\eta + n_{1\cdot k}, \cdots, \eta + n_{V\cdot k}\right) \tag{46}$$

which would be the same for the other NB processes, since the gamma-NB process differs from them on how the gamma priors of θ_{jk} and consequently the NB priors of n_{jk} are constituted. For example, marginalizing out θ_{jk}, we have $n_{jk} \sim$ NB(r_k, p_j) for the gamma-NB process,

$n_{jk} \sim \text{NB}(r_j, p_k)$ for the beta-NB process, $n_{jk} \sim \text{NB}(r_k, p_k)$ for both the marked-beta-NB and marked-gamma-NB processes, and $n_{jk} \sim \text{NB}(r_k b_{jk}, p_j)$ for the zero-inflated-NB process.

Since in topic modeling the majority of computation is spent updating z_{ji}, $\boldsymbol{\phi}_k$ and θ_{jk}, the proposed Bayesian nonparametric models pay a small amount of additional computation, relative to parametric ones such as latent Dirichlet allocation (LDA) [5], for updating other parameters.

A. Poisson Factor Analysis

Note that under the bag-of-words representation (the ordering of words in a document is not considered), without losing information, we can form $\{\boldsymbol{x}_j\}_{1,J}$ as a term-document count matrix $\mathbf{M} \in \mathbb{R}^{V\times J}$, where m_{vj} counts the number of times term v appears in document j. Given $K \leq \infty$ and a count matrix $\mathbf{M}$, discrete latent variable models assume that each entry m_{vj} can be explained as a sum of smaller counts, each produced by one of the K hidden factors, or in the case of topic modeling, a hidden topic. We can factorize $\mathbf{M}$ under the Poisson likelihood as

$$\mathbf{M} = \text{Pois}(\mathbf{\Phi}\mathbf{\Theta}) \tag{47}$$

where $\mathbf{\Phi} \in \mathbb{R}^{V\times K}$ is the factor loading matrix, each column of which is an atom encoding the relative importance of each term; $\mathbf{\Theta} \in \mathbb{R}^{K\times J}$ is the factor score matrix, each column of which encodes the relative importance of each atom in a sample. This is called Poisson Factor Analysis (PFA). We may augment $m_{vj} \sim \text{Pois}(\sum_{k=1}^K \phi_{vk}\theta_{jk})$ as

$$m_{vj} = \textstyle\sum_{k=1}^K n_{vjk},\ n_{vjk} \sim \text{Pois}(\phi_{vk}\theta_{jk}). \tag{48}$$

and if $\sum_{v=1}^V \phi_{vk} = 1$, we have $n_{jk} \sim \text{Pois}(\theta_{jk})$, and with Lemma IV.1, we also have

$$(n_{vj1}, \cdots, n_{vjK}|-) \sim \text{Mult}\left(m_{vj}; \tfrac{\phi_{v1}\theta_{j1}}{\sum_{k=1}^K \phi_{vk}\theta_{jk}}, \cdots, \tfrac{\phi_{vK}\theta_{jK}}{\sum_{k=1}^K \phi_{vk}\theta_{jk}}\right) \tag{49}$$

$$(n_{v\cdot 1}, \cdots, n_{v\cdot K}|-) \sim \text{Mult}(n_{\cdot k}; \boldsymbol{\phi}_k), \quad (n_{j1}, \cdots, n_{jK}|-) \sim \text{Mult}\left(N_j; \boldsymbol{\theta}_j\right) \tag{50}$$

where (49) would lead to (45) under the assumption that the words $\{x_{ji}\}_i$ are exchangeable and (50) would lead to (46) if $\boldsymbol{\phi}_k \sim \text{Dir}(\eta, \cdots, \eta)$. Thus the NB process topic modeling can be considered as factorization of the term-document count matrix under the Poisson likelihood as $\mathbf{M} \sim \text{Pois}(\mathbf{\Phi}\mathbf{\Theta})$, with the requirement that $\sum_{v=1}^V \phi_{vk} = 1$ (implying $n_{jk} \sim \text{Pois}(\theta_{jk})$).

B. Related Discrete Latent Variable Models

We show below that previously proposed discrete latent variable models can be connected under the PFA framework, with the differences mainly on how the priors of ϕ_{vk} and θ_{jk} are constituted and how the inferences are implemented.

1) Latent Dirichlet Allocation: We can construct a Dirichlet-PFA (Dir-PFA) by imposing Dirichlet priors on both $\boldsymbol{\phi}_k = (\phi_{1k}, \cdots, \phi_{Vk})^T$ and $\boldsymbol{\theta}_j = (\theta_{j1}, \cdots, \theta_{jK})^T$ as

$$\boldsymbol{\phi}_k \sim \text{Dir}(\eta, \cdots, \eta), \quad \boldsymbol{\theta}_j \sim \text{Dir}\left(\alpha/K, \cdots, \alpha/K\right). \tag{51}$$

Sampling z_{ji} with (45), which is the same as sampling n_{vjk} with (49), and using (50) with the Dirichlet multinomial conjugacy, we have

$$(\boldsymbol{\phi}_k|-) \sim \text{Dir}\left(\eta + n_{1\cdot k}, \cdots, \eta + n_{V\cdot k}\right), \quad (\boldsymbol{\theta}_j|-) \sim \text{Dir}\left(\alpha/K + n_{j1}, \cdots, \alpha/K + n_{jK}\right). \tag{52}$$

Using variational Bayes (VB) inference [58], [59], we can approximate the posterior distribution with the product of $Q_{(n_{vj1},\cdots,n_{vjK})} = \text{Mult}\left(m_{vj}; \tilde{\zeta}_{vj1}, \cdots, \tilde{\zeta}_{vjK}\right)$, $Q_{\boldsymbol{\phi}_k} = \text{Dir}\left(\tilde{a}_{\phi 1k}, \cdots, \tilde{a}_{\phi Vk}\right)$ and $Q_{\boldsymbol{\theta}_j} = \text{Dir}\left(\tilde{a}_{\theta j1}, \cdots, \tilde{a}_{\theta jK}\right)$ for $v = 1, \cdots, V$, $j = 1, \cdots, J$ and $k = 1, \cdots, K$, where

$$\tilde{\zeta}_{vjk} = \frac{\exp\left(\langle \ln \phi_{vk}\rangle + \langle \ln \theta_{jk}\rangle\right)}{\sum_{k'=1}^{K} \exp\left(\langle \ln \phi_{vk'}\rangle + \langle \ln \theta_{jk'}\rangle\right)} \tag{53}$$

$$\tilde{a}_{\phi vk} = \eta + \langle n_{v\cdot k}\rangle, \quad \tilde{a}_{\theta jk} = \alpha/K + \langle n_{jk}\rangle; \tag{54}$$

these moments are calculated as $\langle \ln \phi_{vk}\rangle = \psi(\tilde{a}_{\phi vk}) - \psi\left(\sum_{v'=1}^{V} \tilde{a}_{\phi v'k}\right)$, $\langle \ln \theta_{jk}\rangle = \psi(\tilde{a}_{\theta jk}) - \psi\left(\sum_{k'=1}^{K} \tilde{a}_{\theta jk'}\right)$ and $\langle n_{vjk}\rangle = m_{vj}\tilde{\zeta}_{vjk}$, where $\psi(x)$ is the diagmma function. Therefore, Dir-PFA and LDA [5], [60] have the same block Gibbs sampling and VB inference. It may appear that Dir-PFA should differ from LDA via the Poisson distribution; however, imposing Dirichlet priors on both factor loadings and scores makes it essentially loose that distinction.

2) Nonnegative Matrix Factorization and a Gamma-Poisson Factor Model: We can construct a Gamma-PFA (Γ-PFA) by imposing gamma priors on both ϕ_{vk} and θ_{jk} as

$$\phi_{vk} \sim \text{Gamma}(a_\phi, 1/b_\phi), \;\; \theta_{jk} \sim \text{Gamma}(a_\theta, g_k/a_\theta). \tag{55}$$

Note that if we set $b_\phi = 0$, $a_\phi = a_\theta = 1$ and $g_k = \infty$, then we are imposing no priors on ϕ_{vk} and θ_{jk}, and a maximum a posterior (MAP) estimate of Γ-PFA would become an ML estimate of PFA. Using (48) and (55), one can show that

$$(\phi_{vk}|-) \sim \text{Gamma}(a_\phi + n_{v\cdot k}, 1/(b_\phi + \theta_{\cdot k})) \tag{56}$$

$$(\theta_{jk}|-) \sim \text{Gamma}(a_\theta + n_{jk}, 1/(a_\theta/g_k + \phi_{\cdot k})) \tag{57}$$

where $\theta_{\cdot k} = \sum_{j=1}^{J} \theta_{jk}$, $\phi_{\cdot k} = \sum_{v=1}^{V} \phi_{vk}$. If $a_\phi \geq 1$ and $a_\theta \geq 1$, using (49), (56) and (57), we can substitute $\mathbb{E}[n_{vjk}|\phi_{vk}, \theta_{jk}] = \frac{m_{vj}\phi_{vk}\theta_{jk}}{\sum_{k=1}^{K}\phi_{vk}\theta_{jk}}$ into the modes of ϕ_{pk} and θ_{ki}, leading to a MAP Expectation-Maximization (MAP-EM) algorithm as

$$\phi_{vk} = \phi_{vk}\frac{\frac{a_\phi - 1}{\phi_{vk}} + \sum_{j=1}^{J}\frac{m_{vj}\theta_{jk}}{\sum_{k=1}^{K}\phi_{vk}\theta_{jk}}}{b_\phi + \theta_{k\cdot}}, \quad \theta_{jk} = \theta_{jk}\frac{\frac{a_\theta - 1}{\theta_{jk}} + \sum_{v=1}^{V}\frac{m_{vj}\phi_{vk}}{\sum_{k=1}^{K}\phi_{vk}\theta_{jk}}}{a_\theta/g_k + \phi_{\cdot k}}. \tag{58}$$

If we set $b_\phi = 0$, $a_\phi = a_\theta = 1$ and $g_k = \infty$, the MAP-EM algorithm reduces to the ML-EM algorithm, which is found to be the same as that of non-negative matrix factorization (NMF) with an objective function of minimizing the KL divergence $D_{KL}(\mathbf{M}||\mathbf{\Phi\Theta})$ [61]. If we set $b_\phi = 0$ and $a_\phi = 1$, then the update equations in (58) are the same as those of the gamma-Poisson (Gap) model of [6], in which setting $a_\theta = 1.1$ and estimating g_k with $g_k = \mathbb{E}[\theta_{jk}]$ are suggested. Since all latent variables are in the exponential family with conjugate update, following the VB inference for Dir-PFA in Section VII-B1, we can conveniently derive the VB inference for Γ-PFA, omitted here for brevity. Note that the inference for the basic gamma-Poisson model and its variations have also been discussed in detail in [62], [63]. Here we show using Lemma IV.1, the derivations of the ML-EM, MAP-EM, Gibbs sampling and VB inference are all straightforward. The NMF has been widely studied and applied to numerous applications, such as image processing and music analysis [61], [64]. Showing its connections to NB process topic modeling, under the Poisson factor analysis framework, may allow us to extend the proposed nonparametric Bayesian techniques to these broad applications.

C. Negative Binomial Process Topic Modeling

From the point view of PFA, a NB process topic model factorizes the count matrix under the constraints that each factor sums to one and the factor scores are gamma distributed random variables, and consequently, the number of words assigned to a topic (factor/atom) follows a NB distribution. Depending on how the NB distributions are parameterized, as shown in Table I, we can construct a variety of NB process topic models, which can also be connected to previous parametric and nonparametric topic models. For a deeper understanding on how the counts are modeled, we also show in Table I both the VMR and ODL implied by these settings.

We consider eight differently constructed NB processes in Table I: (*i*) The NB process described in (21) is used for topic modeling. It improves over the count-modeling gamma-Poisson process discussed in [10], [11] in that it unites mixture modeling and has closed-form Bayesian inference. Although this is a nonparametric model supporting an infinite number of topics, requiring $\{\theta_{jk}\}_{j=1,J} \equiv r_k$ may be too restrictive. (*ii*) Related to LDA [5] and Dir-PFA [7], the NB-LDA is also a parametric topic model that requires tuning the number of topics. However, it uses a document dependent r_j and p_j to automatically learn the smoothing of the gamma distributed topic weights, and it lets $r_j \sim \text{Gamma}(\gamma_0, 1/c)$, $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$ to share statistical strength between documents, with closed-form Gibbs sampling. Thus even the most

TABLE I

A variety of negative binomial processes are constructed with distinct sharing mechanisms, reflected with which parameters from θ_{jk}, r_k, r_j, p_k, p_j and π_k (b_{jk}) are inferred (indicated by a check-mark ✓), and the implied VMR and ODL for counts $\{n_{jk}\}_{j,k}$. They are applied for topic modeling of a document corpus, a typical example of mixture modeling of grouped data. Related algorithms are shown in the last column.

Algorithms	θ_{jk}	r_k	r_j	p_k	p_j	π_k	VMR	ODL	Related Algorithms
NB	$\theta_{jk} \equiv r_k$	✓					$(1-p)^{-1}$	r_k^{-1}	Gamma-Poisson [10], [11]
NB-LDA	✓		✓		✓		$(1-p_j)^{-1}$	r_j^{-1}	LDA [5], Dir-PFA [7]
NB-HDP	✓	✓			0.5		2	r_k^{-1}	HDP[21], DILN-HDP [48]
NB-FTM	✓	✓			0.5	✓	2	$(r_k)^{-1}b_{jk}$	FTM [49], S$\gamma\Gamma$-PFA [7]
Beta-Geometric	✓		1	✓			$(1-p_k)^{-1}$	1	Beta-Geometric [11], BNBP [7], [13]
Beta-NB	✓		✓	✓			$(1-p_k)^{-1}$	r_j^{-1}	BNBP [7], [13]
Gamma-NB	✓	✓			✓		$(1-p_j)^{-1}$	r_k^{-1}	CRF-HDP [21], [22]
Marked-Beta-NB	✓	✓		✓			$(1-p_k)^{-1}$	r_k^{-1}	BNBP [7]

basic parametric LDA topic model can be improved under the NB count modeling framework. (*iii*) The NB-HDP model is related to the HDP [21], and since p_j is an irrelevant parameter in the HDP due to normalization, we set it in the NB-HDP as 0.5, the usually perceived value before normalization. The NB-HDP model is comparable to the DILN-HDP [48] that constructs the group-level DPs with normalized gamma processes, whose scale parameters are also set as one. (*iv*) The NB-FTM model introduces an additional beta-Bernoulli process under the NB process framework to explicitly model zero counts. It is the same as the sparse-gamma-gamma-PFA (S$\gamma\Gamma$-PFA) in [7] and is comparable to the focused topic model (FTM) [49], which is constructed from the IBP compound Dirichlet process. The Zero-Inflated-NB process improves over these approaches by allowing p_j to be inferred, which generally yields better data fitting. (*v*) The Gamma-NB process explores sharing the NB dispersion measure across groups, and it improves over the NB-HDP by allowing the learning of p_j. It reduces to the HDP [21] by normalizing both the group-level and the shared gamma processes. (*vi*) The Beta-Geometric process explores the idea that the probability measure is shared across groups, which is related to the one proposed for count modeling in [11]. It is restrictive that the NB dispersion parameters are fixed as one. (*vii*) The Beta-NB process explores sharing the NB probability measure across groups, and it improves over the Beta-Geometric process and the beta negative binomial process (BNBP) proposed in [13], allowing inference of r_j. (*viii*) The Marked-Beta-NB process is comparable to the BNBP proposed in [7], with the distinction that it allows analytical update of r_k. The constructions and inference of various NB processes and related algorithms in Table I all follow the formulas in (42) and (43), respectively, with additional details presented in Appendix B.

Note that as analyzed in Section VII, NB process topic models can also be considered as factor analysis of the term-document count matrix under the Poisson likelihood, with $\boldsymbol{\phi}_k$ as the kth factor that sums to one and θ_{jk} as the factor score of the jth document on the kth factor, which can be further linked to nonnegative matrix factorization [61] and a gamma Poisson factor model [6]. If except for proportions $\tilde{\boldsymbol{\lambda}}_j$ and $\tilde{r}$, the absolute values, e.g., θ_{jk}, r_k and p_k, are also of interest, then the NB processes based joint count and mixture models would be more appropriate than the Dirichlet process and the HDP based mixture models.

VIII. Example Results

Motivated by Table I, we consider topic modeling using a variety of NB processes, which differ on how the NB dispersion and probability parameters of the latent counts $\{n_{jk}\}_{j,k}$ are learned and consequently how the VMR and ODL are modeled. We compare them with LDA [5], [65] and CRF-HDP [21], [22], in which the latent count n_{jk} is marginally distributed as

$$n_{jk} \sim \text{Beta-Binomial}(N_j, \alpha\tilde{r}_k, \alpha(1-\tilde{r}_k)) \tag{59}$$

with $\tilde{r}_k$ fixed as $1/K$ in LDA and learned from the data in CRF-HDP. For fair comparison, they are all implemented with block Gibbs sampling using a discrete base measure with K atoms, and for the first fifty iterations, the Gamma-NB process with $r_k \equiv 50/K$ and $p_j \equiv 0.5$ is used for initialization. We consider 2500 Gibbs sampling iterations and collect the last 1500 samples.

We consider the Psychological Review[1] corpus, restricting the vocabulary to terms that occur in five or more documents. The corpus includes 1281 abstracts from 1967 to 2003, with $V = 2566$ and 71,279 total word counts. We randomly select 20%, 40%, 60% or 80% of the words from each document to learn a document dependent probability for each term v and calculate the per-word perplexity on the held-out words as

$$\text{Perplexity} = \exp\left(-\tfrac{1}{y_{\cdot\cdot}}\textstyle\sum_{j=1}^{J}\sum_{v=1}^{V} y_{jv}\log f_{jv}\right),\ \ f_{jv} = \frac{\sum_{s=1}^{S}\sum_{k=1}^{K}\phi_{vk}^{(s)}\theta_{jk}^{(s)}}{\sum_{s=1}^{S}\sum_{v=1}^{V}\sum_{k=1}^{K}\phi_{vk}^{(s)}\theta_{jk}^{(s)}} \tag{60}$$

[1] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

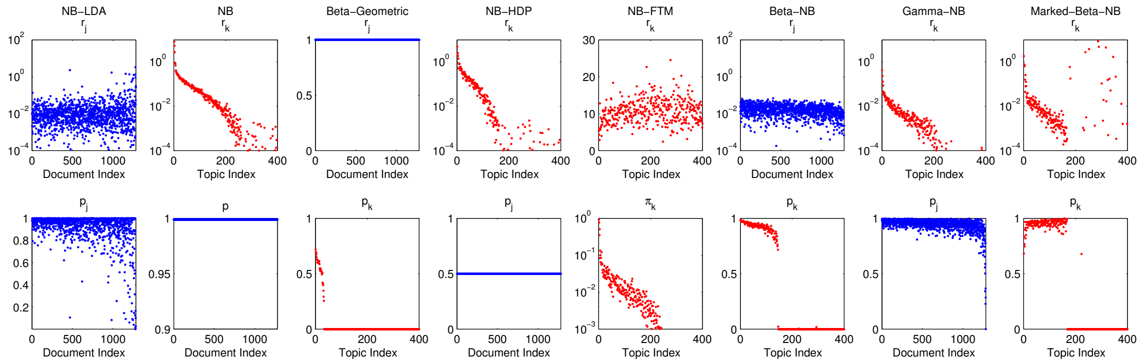

Fig. 1. Distinct sharing mechanisms and model properties are evident between various NB process topic models, by comparing their inferred NB dispersion and probability parameters. Note that the transition between active and non-active topics is very sharp when p_k is used and much smoother when r_k is used. Both the documents and topics are ordered in a decreasing order based on the number of words associated with each of them. These results are based on the last Gibbs sampling iteration, on the Psychological Review corpus with 80% of the words in each document used as training. The values are shown in either linear or log scales for convenient visualization.

where y_{jv} is the number of words held out at term v in document j, $y_{\cdot\cdot} = \sum_{j=1}^{J}\sum_{v=1}^{V} y_{jv}$ is the total number of held-out words, and $s = 1, \cdots, S$ are the indices of collected samples. Note that the per-word perplexity is equal to V if $f_{jv} = 1/V$, thus it should be no greater than V for a functional topic model. The final results are averaged from five random training/testing partitions. The performance measure is the same as in [7] and also similar to those used in [66], [67], [23]. Note that the perplexity per held-out word is a fair metric to compare topic models. However, when the actual Poisson rates or NB distribution parameters for counts instead of the mixture proportions are of interest, a NB process based joint count and mixture model would be more appropriate than a Dirichlet process or an HDP based mixture model.

We show in Fig. 1 the NB dispersion and probability parameters learned by various NB process topic models listed in Table I, revealing distinct sharing mechanisms and model properties. In Fig. 2 we compare the per-held-out-word prediction performance of various algorithms. We set the parameters as $c = 1$, $\eta = 0.05$ and $a_0 = b_0 = e_0 = f_0 = 0.01$. For LDA and NB-LDA, we search K for optimal performance and for the others, we set $K = 400$ as an upper-bound. All the other NB process topic models are nonparametric Bayesian algorithms that can automatically learn the number of active topics K^+ for a given corpus. When $\theta_{jk} \equiv r_k$ is used, as in the NB process, different documents are imposed to have the same topic weights, leading to the worst held-out-prediction performance.

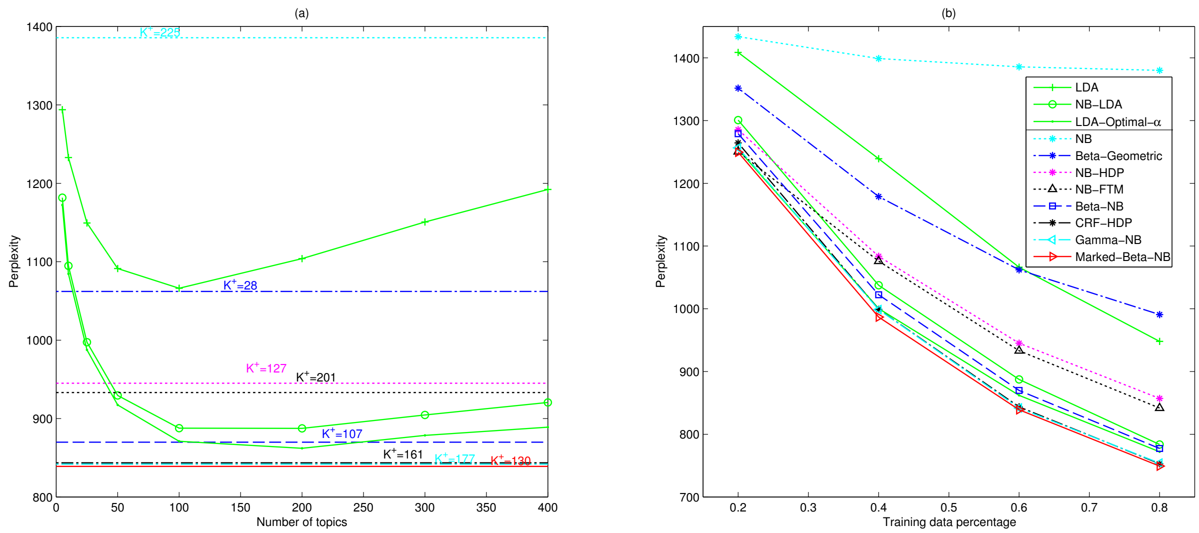

Fig. 2. Comparison of per-word perplexity on held out words between various algorithms listed in Table I on the Psychological Review corpus. LDA-Optimal-α refers to an LDA algorithm whose topic proportion Dirichlet concentration parameter α is optimized based on the results of the CRF-HDP on the same dataset. (a) With 60% of the words in each document used for training, the performance varies as a function of K in both LDA and NB-LDA, which are parametric models, whereas the NB, Beta-Geometric, NB-HDP, NB-FTM, Beta-NB, CRF-HDP, Gamma-NB and Marked-Beta-NB all infer the number of active topics, which are 225, 28, 127, 201, 107, 161, 177 and 130, respectively, according to the last Gibbs sampling iteration. (b) Per-word perplexities of various algorithms as a function of the percentage of words in each document used for training. The results of LDA and NB-LDA are shown with the best settings of K under each training/testing partition. Nonparametric Bayesian algorithms listed in Table I are ranked in the legend from top to bottom according to their overall performance.

With a symmetric Dirichlet prior Dir$(\alpha/K, \cdots, \alpha/K)$ placed on the topic proportion for each document, the parametric LDA is found to be sensitive to both the number of topics K and the value of the concentration parameter α. We consider $\alpha = 50$, following the suggestion of the topic model toolbox[1] provided for [65]; we also consider an optimized value as $\alpha = 2.5$, based on the results of the CRF-HDP on the same dataset. As shown in Fig. 2, when the number of training words is small, with optimized K and α, the parametric LDA can approach the performance of the nonparametric CRF-HDP; as the number of training words increases, the advantage of learning $\tilde{r}_k$ in the CRF-HDP than fixing $\tilde{r}_k = 1/K$ in LDA becomes clearer. The concentration parameter α is important for both LDA and CRF-HDP since it controls the VMR of the count n_{jk}, which is equal to $(1 - \tilde{r}_k)(\alpha + N_j)/(\alpha + 1)$ based on (59). Thus fixing α may lead to significantly under- or overestimated variations and then degraded performance, as evident by comparing the performance of LDA with $\alpha = 50$ and LDA-Optima-α in Fig. 2.

When (r_j, p_j) is used, as in NB-LDA, different documents are weakly coupled with $r_j \sim$

Gamma$(\gamma_0, 1/c)$, and the modeling results in Fig. 1 show that a typical document in this corpus usually has a small r_j and a large p_j, thus a large ODL and a large VMR, indicating highly overdispersed counts on its topic usage. NB-LDA is a parametric topic model that requires tuning the number of topics K. It improves over LDA in that it only has to tune K, whereas LDA has to tune both K and α. With an appropriate K, the parametric NB-LDA may outperform the nonparametric NB-HDP and NB-FTM as the training data percentage increases, showing that even by learning both the NB dispersion and probability parameters r_j and p_j in a document dependent manner, we may get better data fitting than using nonparametric models that share the NB dispersion parameters r_k across documents, but fix the NB probability parameters.

When (r_j, p_k) is used to model the latent counts $\{n_{jk}\}_{j,k}$, as in the Beta-NB process, the transition between active and non-active topics is very sharp that p_k is either far from zero or almost zero. That is because p_k controls the mean as $\mathbb{E}[\sum_j n_{jk}] = p_k/(1-p_k)\sum_j r_j$ and the VMR as $(1-p_k)^{-1}$ on topic k, thus a popular topic must also have large p_k and thus large overdispersion measured by the VMR; since the counts $\{n_{jk}\}_j$ are usually overdispersed, particularly true in this corpus, a small p_k indicating an small mean and small overdispersion is not favored by the model and thus is rarely observed.

The Beta-Geometric process is a special case of the Beta-NB process that $r_j \equiv 1$, which is more than ten times larger than the values inferred by the Beta-NB process on this corpus, as shown in Fig. 1; therefore, to fit the mean $\mathbb{E}[\sum_j n_{jk}] = Jp_k/(1-p_k)$, it has to use a substantially underestimated p_k, leading to severely underestimated variations and thus degraded performance, as confirmed by comparing the curves of the Beta-Geometric and Beta-NB processes in Fig. 2.

When (r_k, p_j) is used, as in the Gamma-NB process, the transition is much smoother that r_k gradually decreases. The reason is that r_k controls the mean as $\mathbb{E}[\sum_j n_{jk}] = r_k \sum_j p_j/(1-p_j)$ and the ODL r_k^{-1} on topic k, thus popular topics must also have large r_k and thus small overdispersion measured by the ODL, and unpopular topics are modeled with small r_k and thus large overdispersion, allowing rarely and lightly used topics. Therefore, we can expect that (r_k, p_j) would allow more topics than (r_j, p_k), as confirmed in Fig. 2 (a) that the Gamma-NB process learns 177 active topics, significantly more than the 107 ones of the Beta-NB process. With these analysis, we can conclude that the mean and the amount of overdispersion (measure by the VMR or ODL) for the usage of topic k is positively correlated under (r_j, p_k) and negatively correlated under (r_k, p_j).

The NB-HDP is a special case of the Gamma-NB process that $p_j \equiv 0.5$. From a mixture modeling viewpoint, fixing $p_j = 0.5$ is a natural choice as p_j becomes irrelevant after normalization. However, from a count modeling viewpoint, this would make a restrictive assumption that each count vector $\{n_{jk}\}_{k=1,K}$ has the same VMR of 2. It is also interesting to examine (41), which can be viewed as the concentration parameter α in the HDP, allowing the adjustment of p_j would allow a more flexible model assumption on the amount of variations between the topic proportions, and thus potentially better data fitting.

The CRF-HDP and Gamma-NB process have very similar performance on predicting held-out words, although they have distinct assumption on count modeling: n_{jk} is modeled as a NB distribution in the Gamma-NB process while it is modeled as a beta-binomial distribution in the CRF-HDP. The Gamma-NB process adjust both r_k and p_j to fit the NB distribution, whereas the CRF-HDP learns both α and $\tilde{r}_k$ to fit the beta-binomial distribution. The concentration parameter α controls the VMR of the count n_{jk} as shown in (59), and we find through experiments that prefixing its value may substantially degrade the performance of the CRF-HDP, thus this option is not considered in the paper and we exploit the CRF metaphor to update α as in [21], [22].

When (r_k, π_k) is used, as in the NB-FTM model, our results show that we usually have a small π_k and a large r_k, indicating topic k is sparsely used across the documents but once it is used, the amount of variation on usage is small. This modeling properties might be helpful when there are excessive number of zeros which might not be well modeled by the NB process alone. In our experiments, we find the more direct approaches of using p_k or p_j generally yield better results, but this might not be the case when excessive number of zeros are better explained with the underlying beta-Bernoulli processes, e.g., when the training words are scarce, the NB-HDP can approach the performance of the Marked-Beta-NB process.

When (r_k, p_k) is used, as in the Marked-Beta-NB process, more diverse combinations of mean and overdispersion would be allowed as both r_k and p_k are now responsible for the mean $\mathbb{E}[\sum_j n_{jk}] = J r_k p_k/(1-p_k)$. For example, there could be not only large mean and small overdispersion (large r_k and small p_k), indicating a popular topic frequently used by most of the documents, but also large mean and large overdispersion (small r_k and large p_k), indicating a topic heavily used in a relatively small percentage of documents. Thus (r_k, p_k) may combine the advantages of using only r_k or p_k to model topic k, as confirmed by the superior performance of the Marked-Beta-NB over the Beta-NB and Gamma-NB processes.

IX. Conclusions

We propose a variety of negative binomial (NB) processes for count modeling, which can be naturally applied for the seemingly disjoint problem of mixture modeling. The proposed NB processes are completely random measures, which assign independent random variables to disjoint Borel sets of the measure space, as opposed to the Dirichlet process and the hierarchical Dirichlet process (HDP), whose measures on disjoint Borel sets are negatively correlated. We reveal connections between various distributions and discover unique data augmentation methods for the NB distribution, with which we are able to unite count and mixture modeling, analyze fundamental model properties, and derive efficient Bayesian inference using Gibbs sampling. We demonstrate that the NB process and the gamma-NB process can be normalized to produce the Dirichlet process and the HDP, respectively. We show in detail the theoretical, structural and computational advantages of the NB process. We examine the distinct sharing mechanisms and model properties of various NB processes, with connections made to existing discrete latent variable models under the Poisson factor analysis framework. Experimental results on topic modeling show the importance of modeling both the NB dispersion and probability parameters, which respectively govern the overdispersion level and variance-to-mean ratio for count modeling.

References

[1] C. I. Bliss and R. A. Fisher. Fitting the negative binomial distribution to biological data. *Biometrics*, 1953.

[2] M. Zhou, L. Li, D. Dunson, and L. Carin. Lognormal and gamma mixed negative binomial regression. In *ICML*, 2012.

[3] C. Dean, J. F. Lawless, and G. E. Willmot. A mixed Poisson-inverse-Gaussian regression model. *Canadian Journal of Statistics*, 1989.

[4] T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, 1999.

[5] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

[6] J. Canny. Gap: a factor model for discrete data. In *SIGIR*, 2004.

[7] M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, 2012.

[8] J. F. C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

[9] R. L. Wolpert and K. Ickstadt. Poisson/gamma random field models for spatial statistics. *Biometrika*, 1998.

[10] M. K. Titsias. The infinite gamma-Poisson feature model. In *NIPS*, 2008.

[11] R. J. Thibaux. *Nonparametric Bayesian Models for Machine Learning*. PhD thesis, UC Berkeley, 2008.

[12] K. T. Miller. *Bayesian Nonparametric Latent Feature Models*. PhD thesis, UC Berkeley, 2011.

[13] T. Broderick, L. Mackey, J. Paisley, and M. I. Jordan. Combinatorial clustering and the beta negative binomial process. *arXiv:1111.1802v3*, 2012.

[14] M. Zhou and L. Carin. Augment-and-conquer negative binomial processes. In *NIPS*, 2012.

[15] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1973.

[16] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 1974.

[17] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *JASA*, 1995.

[18] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 2000.

[19] Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

[20] R. L. Wolpert, M. A. Clyde, and C. Tu. Stochastic expansions using continuous dictionaries: Lévy Adaptive Regression Kernels. *Annals of Statistics*, 2011.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *JASA*, 2006.

[22] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Developing a tempered HDP-HMM for systems with state persistence. *MIT LIDS, TR #2777*, 2007.

[23] C. Wang, J. Paisley, and D. M. Blei. Online variational inference for the hierarchical Dirichlet process. In *AISTATS*, 2011.

[24] J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 1967.

[25] M. I. Jordan. Hierarchical models, nested models and completely random measures. In M.-H. Chen, D. Dey, P. Mueller, D. Sun, and K. Ye, editors, *Frontiers of Statistical Decision Making and Bayesian Analysis: in Honor of James O. Berger*. New York: Springer, 2010.

[26] E. Çinlar. *Probability and Stochastics*. Springer, New York, 2011.

[27] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *AISTATS*, 2007.

[28] N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, 1990.

[29] H. Ishwaran and M. Zarepour. Exact and approximate sum-representations for the Dirichlet process. *Can. J. Statist.*, 2002.

[30] D. Blackwell and J. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1973.

[31] D. Aldous. Exchangeability and related topics. In *Ecole d'Ete de Probabilities de Saint-Flour XIII 1983*, pages 1–198. Springer.

[32] J. Pitman. *Combinatorial stochastic processes*. Lecture Notes in Mathematics. Springer-Verlag, 2006.

[33] M. Greenwood and G. U. Yule. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, 1920.

[34] M. H. Quenouille. A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics*, 1949.

[35] N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate Discrete Distributions*. John Wiley & Sons, 2005.

[36] O. E. Barndorff-Nielsen, D. G. Pollard, and N. Shephard. Integer-valued Lévy processes and low latency financial econometrics. *Preprint*, 2010.

[37] A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*. Cambridge, UK, 1998.

[38] R. Winkelmann. *Econometric Analysis of Count Data*. Springer, Berlin, 5th edition, 2008.

[39] M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 2008.

[40] E. P. Pieters, C. E. Gates, J. H. Matis, and W. L. Sterling. Small sample comparison of different estimators of negative binomial parameters. *Biometrics*, 1977.

[41] L. J. Willson, J. L. Folks, and J. H. Young. Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter k. *Biometrics*, 1984.

[42] J. F. Lawless. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 1987.

[43] W. W. Piegorsch. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, 1990.

[44] K. Saha and S. Paul. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 2005.

[45] J. O. Lloyd-Smith. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS ONE*, 2007.

[46] E. T. Bradlow, B. G. S. Hardie, and P. S. Fader. Bayesian inference for the negative binomial distribution via polynomial expansions. *Journal of Computational and Graphical Statistics*, 2002.

[47] D. B. Dunson and A. H. Herring. Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 2005.

[48] J. Paisley, C. Wang, and D. M. Blei. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 2012.

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

[50] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2005.

[51] D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Independent Component Analysis and Signal Separation*, 2007.

[52] J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *ICML*, 2009.

[53] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric Bayesian dictionary learning for sparse image representations. In *NIPS*, 2009.

[54] M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin. Nonparametric Bayesian matrix completion. In *IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2010.

[55] M. Zhou, H. Yang, G. Sapiro, D. Dunson, and L. Carin. Dependent hierarchical beta process for image interpolation and denoising. In *AISTATS*, 2011.

[56] L. Li, M. Zhou, G. Sapiro, and L. Carin. On the integration of topic modeling and dictionary learning. In *ICML*, 2011.

[57] M. Zhou, H. Chen, J. Paisley, L. Ren, L. Li, Z. Xing, D. Dunson, G. Sapiro, and L. Carin. Nonparametric Bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE TIP*, 2012.

[58] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[59] C. M. Bishop and M. E. Tipping. Variational relevance vector machines. In *UAI*, 2000.

[60] D. Blei M. Hoffman and F. Bach. Online learning for latent Dirichlet allocation. In *NIPS*, 2010.

[61] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2000.

[62] W. Buntine and A. Jakulin. Discrete component analysis. In *Subspace, Latent Structure and Feature Selection Techniques*. Springer-Verlag, 2006.

[63] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Intell. Neuroscience*, 2009.

[64] C. Févotte, N. Bertin, and J. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural Comput.*, 2009.

[65] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 2004.

[66] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *UAI*, 2009.

[67] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *ICML*, 2009.

[68] Y. Kim. Nonparametric Bayesian estimators for counting processes. *Annals of Statistics*, 1999.

Appendix A

Chinese Restaurant Table Distribution

Lemma A.1. *A CRT random variable* $l \sim \mathrm{CRT}(m, r)$ *with PMF* $f_L(l|m,r) = \frac{\Gamma(r)}{\Gamma(m+r)}|s(m,l)|r^l,\ \ l = 0, 1, \cdots, m$ *can be generated as*

$$l = \sum_{n=1}^{m} b_n,\ b_n \sim \text{Bernoulli}\left(\frac{r}{n-1+r}\right). \tag{61}$$

Proof: Since l is the summation of independent Bernoulli random variables, its PGF becomes $C_L(z) = \prod_{n=1}^{m}\left(\frac{n-1}{n-1+r} + \frac{r}{n-1+r}z\right) = \frac{\Gamma(r)}{\Gamma(m+r)}\sum_{k=0}^{m}|s(m,k)|(rz)^k$. Thus we have $f_L(l|m,r) = \frac{C_L^{(l)}(0)}{l!} = \frac{\Gamma(r)}{\Gamma(m+r)}|s(m,l)|r^l,\quad l = 0, 1, \cdots, m.$ ■

Corollary A.2. *If* $f_L(l|m,r) = \frac{\Gamma(r)}{\Gamma(m+r)}|s(m,l)|r^l,\quad l = 0, 1, \cdots, m$*, i.e.* $l \sim \mathrm{CRT}(m, r)$*, then*

$$\mathbb{E}[l|m,r] = \sum_{n=1}^{m}\frac{r}{n-1+r},\quad \text{Var}[l|m,r] = \sum_{n=1}^{m}\frac{(n-1)r}{(n-1+r)^2} \tag{62}$$

and approximately we have the mean and variance as

$$\hat{\mu}_l = \int_1^{m+1}\tfrac{r}{x-1+r}dx = r\ln\tfrac{m+r}{r},\qquad \hat{\sigma^2}_l = \int_1^{m+1}\tfrac{(x-1)r}{(x-1+r)^2}dx = r\ln\tfrac{m+r}{r} - \tfrac{(m+1)r}{m+1+r}. \tag{63}$$

Although $l \sim \mathrm{CRT}(m, r)$ can be generated as the summation of independent Bernoulli random variables, it may be desirable to directly calculate out its PMF in some case. However, it is numerically instable to recursively calculate the unsigned Stirling numbers of the first kind $|s(m,l)|$ based on $|s(m,l)| = (m-1)|s(m-1,l)| + |s(m-1,l-1)|$, as $|s(m,l)|$ would rapidly reach the maximum value allowed by a finite precision machine as m increases. Denote $P_r(m,l) = \frac{\Gamma(r)}{\Gamma(m+r)}|s(m,l)|r^l$, then $\mathbf{P}_r$ is a probability matrix as a function of r, each row of which sums to one, with $P_r(0,0) = 1$, $P_r(m,0) = 0$ if $m > 0$ and $P_r(m,l) = 0$ if $l > m$. We propose to calculate $P_r(m,l)$ under the logarithmic scale based on

$$\ln P_r(m,l) = \ln P_1(m,l) + l\ln(r) + \ln\Gamma(r) - \ln\Gamma(m+r) + \ln\Gamma(m+1) \tag{64}$$

where $\ln P_1(m,l)$ is iteratively calculated with $\ln P_1(m,1) = \ln\frac{m-1}{m} + \ln P_1(m-1,1)$, $\ln P_1(m,l) = \ln\frac{m-1}{m} + \ln P_1(m-1,l) + \ln(1 + \exp(\ln P_1(m-1,l-1) - \ln P_1(m-1,l) - \ln(m-1)))$ for $2 \le l \le m-1$, and $\ln P_1(m,m) = \ln P_1(m-1,m-1) - \ln m$. This approach is found to be numerically stable, but it requires calculating and storing the matrix $\mathbf{P}_1$, which would be time and memory consuming when m is large.

APPENDIX B

MODEL AND INFERENCE FOR NEGATIVE BINOMIAL PROCESS TOPIC MODELS

A. *CRF-HDP*

The CRF-HDP model [7, 26] is constructed as

$$\begin{aligned} & x_{ji} \sim F(\boldsymbol{\phi}_{z_{ji}}), \quad \boldsymbol{\phi}_k \sim \text{Dir}(\eta, \cdots, \eta), \quad z_{ji} \sim \text{Discrete}(\tilde{\boldsymbol{\lambda}}_j) \\ & \tilde{\boldsymbol{\lambda}}_j \sim \text{Dir}(\alpha \tilde{\boldsymbol{r}}), \quad \alpha \sim \text{Gamma}(a_0, 1/b_0), \quad \tilde{\boldsymbol{r}} \sim \text{Dir}(\gamma_0/K, \cdots, \gamma_0/K). \end{aligned} \tag{65}$$

Under the CRF metaphor, denote n_{jk} as the number of customers eating dish k in restaurant j and l_{jk} as the number of tables serving dish k in restaurant j, the direct assignment block Gibbs sampling can be expressed as

$$\begin{aligned} & \Pr(z_{ji} = k|-) \propto \phi_{v_{ji}k} \tilde{\lambda}_{jk} \\ & (l_{jk}|-) \sim \text{CRT}(n_{jk}, \alpha \tilde{r}_k), \quad w_j \sim \text{Beta}(\alpha + 1, N_j), \quad s_j \sim \text{Bernoulli}\left(\frac{N_j}{N_j + \alpha}\right) \\ & \alpha \sim \text{Gamma}\left(a_0 + \sum_{j=1}^J \sum_{k=1}^K l_{jk} - \sum_{j=1}^J s_j, \frac{1}{b_0 - \sum_j \ln w_j}\right) \\ & (\tilde{\boldsymbol{r}}|-) \sim \text{Dir}\left(\gamma_0/K + \sum_{j=1}^J l_{j1}, \cdots, \gamma_0/K + \sum_{j=1}^J l_{jK}\right) \\ & (\tilde{\boldsymbol{\lambda}}_j|-) \sim \text{Dir}\left(\alpha \tilde{r}_1 + n_{j1}, \cdots, \alpha \tilde{r}_K + n_{jK}\right) \\ & (\boldsymbol{\phi}_k|-) \sim \text{Dir}\left(\eta + n_{1\cdot k}, \cdots, \eta + n_{V\cdot k}\right). \end{aligned} \tag{66}$$

When $K \to \infty$, the concentration parameter γ_0 can be sampled as

$$\begin{aligned} & w_0 \sim \text{Beta}\left(\gamma_0 + 1, \sum_{j=1}^J \sum_{k=1}^\infty l_{jk}\right), \quad \pi_0 = \frac{e_0 + K^+ - 1}{(f_0 - \ln w_0) \sum_{j=1}^J \sum_{k=1}^\infty l_{jk}} \\ & \gamma_0 \sim \pi_0 \text{Gamma}\left(e_0 + K^+, \frac{1}{f_0 - \ln w_0}\right) + (1 - \pi_0) \text{Gamma}\left(e_0 + K^+ - 1, \frac{1}{f_0 - \ln w_0}\right) \end{aligned} \tag{67}$$

where K^+ is the number of used atoms. Since it is infeasible in practice to let $K \to \infty$, directly using this method to sample γ_0 is only approximately correct, which may result in a biased estimate especially if K is not set large enough. Thus in the experiments, we do not sample γ_0 and fix it as one. Note that for implementation convenience, it is also common to fix the concentration parameter α as one [25]. We find through experiments that learning this parameter usually results in obviously lower per-word perplexity for held out words, thus we allow the learning of α using the data augmentation method proposed in [7], which is modified from the one proposed in [24].

B. NB-LDA

The NB-LDA model is constructed as

$$\begin{aligned}
&x_{ji} \sim F(\boldsymbol{\phi}_{z_{ji}}), \quad \boldsymbol{\phi}_k \sim \text{Dir}(\eta, \cdots, \eta) \\
&N_j = \sum_{k=1}^{K} n_{jk}, \quad n_{jk} \sim \text{Pois}(\theta_{jk}), \quad \theta_{jk} \sim \text{Gamma}(r_j, p_j/(1-p_j)) \\
&r_j \sim \text{Gamma}(\gamma_0, 1/c), \quad p_j \sim \text{Beta}(a_0, b_0), \quad \gamma_0 \sim \text{Gamma}(e_0, 1/f_0) \qquad (68)
\end{aligned}$$

Note that letting $r_j \sim \text{Gamma}(\gamma_0, 1/c)$, $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$ allows different documents to share statistical strength for inferring their NB dispersion parameters.

The block Gibbs sampling can be expressed as

$$\begin{aligned}
&\Pr(z_{ji} = k|-) \propto \phi_{v_{ji}k}\theta_{jk} \\
&(p_j|-) \sim \text{Beta}\left(a_0 + N_j, b_0 + Kr_j\right), \quad p'_j = \frac{-K\ln(1-p_j)}{c - K\ln(1-p_j)} \\
&(l_{jk}|-) \sim \text{ CRT}(n_{jk}, r_j), \quad l'_j \sim \text{CRT}(\sum_{k=1}^{K} l_{jk}, \gamma_0), \quad \gamma_0 \sim \text{Gamma}\left(e_0 + \sum_{j=1}^{J} l'_j, \frac{1}{f_0 - \sum_{j=1}^{J}\ln(1-p'_j)}\right) \\
&(r_j|-) \sim \text{Gamma}\left(\gamma_0 + \sum_{k=1}^{K} l_{jk}, \frac{1}{c - K\ln(1-p_j)}\right), \quad (\theta_{jk}|-) \sim \text{Gamma}(r_j + n_{jk}, p_j) \\
&(\boldsymbol{\phi}_k|-) \sim \text{Dir}\left(\eta + n_{1\cdot k}, \cdots, \eta + n_{V\cdot k}\right). \qquad (69)
\end{aligned}$$

C. NB-HDP

The NB-HDP model is a special case of the Gamma-NB process model with $p_j = 0.5$. The hierarchical model and inference for the Gamma-NB process are shown in (42) and (43) of the main paper, respectively.

D. NB-FTM

The NB-FTM model is a special case of zero-inflated NB process with $p_j = 0.5$, which is constructed as

$$\begin{aligned}
& x_{ji} \sim F(\boldsymbol{\phi}_{z_{ji}}), \quad \boldsymbol{\phi}_k \sim \mathrm{Dir}(\eta, \cdots, \eta) \\
& N_j = \sum_{k=1}^{K} n_{jk}, \quad n_{jk} \sim \mathrm{Pois}(\theta_{jk}) \\
& \theta_{jk} \sim \mathrm{Gamma}(r_k b_{jk}, 0.5/(1-0.5)) \\
& \ r_k \sim \mathrm{Gamma}(\gamma_0, 1/c), \ \gamma_0 \sim \mathrm{Gamma}(e_0, 1/f_0) \\
& b_{jk} \sim \mathrm{Bernoulli}(\pi_k), \quad \pi_k \sim \mathrm{Beta}(c/K, c(1-1/K)).
\end{aligned} \tag{70}$$

The block Gibbs sampling can be expressed as

$$\begin{aligned}
& \mathrm{Pr}(z_{ji} = k|-) \propto \phi_{v_{ji}k}\theta_{jk} \\
& b_{jk} \sim \delta(n_{jk} = 0)\mathrm{Bernoulli}\left(\frac{\pi_k(1-0.5)^{r_k}}{\pi_k(1-0.5)^{r_k} + (1-\pi_k)}\right) + \delta(n_{jk} > 0) \\
& \pi_k \sim \mathrm{Beta}\left(c/K + \sum_{j=1}^{J} b_{jk}, c(1-1/K) + J - \sum_{j=1}^{J} b_{jk}\right), \ p'_k = \frac{-\sum_j b_{jk}\ln(1-0.5)}{c - \sum_j b_{jk}\ln(1-0.5)} \\
& (l_{jk}|-) \sim \mathrm{CRT}(n_{jk}, r_k b_{jk}), \ (l'_k|-) \sim \mathrm{CRT}\left(\sum_{j=1}^{J} l_{jk}, \gamma_0\right) \\
& (\gamma_0|-) \sim \mathrm{Gamma}\left(e_0 + \sum_{k=1}^{K} l'_k, \frac{1}{f_0 - \sum_{k=1}^{K}\ln(1-p'_k)}\right) \\
& (r_k|-) \sim \mathrm{Gamma}\left(\gamma_0 + \sum_{j=1}^{J} l_{jk}, \frac{1}{c - \sum_{j=1}^{J} b_{jk}\ln(1-0.5)}\right) \\
& (\theta_{jk}|-) \sim \mathrm{Gamma}(r_k b_{jk} + n_{jk}, 0.5) \\
& (\boldsymbol{\phi}_k|-) \sim \mathrm{Dir}\left(\eta + n_{1\cdot k}, \cdots, \eta + n_{V\cdot k}\right).
\end{aligned} \tag{71}$$

E. Beta-Negative Binomial Process

We consider a beta-NB process that the NB probability measure is shared and drawn from a beta process while the NB dispersion parameters are group dependent. As in Section II-A3, a draw from the beta process $B \sim \mathrm{BP}(c, B_0)$ can be expressed as $B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k}$, thus a beta-NB process can be constructed as $X_j \sim \mathrm{NBP}(r_j, B)$, with a random draw expressed as

$$X_j = \sum_{k=1}^{\infty} n_{jk}\delta_{\omega_k},\ n_{jk} \sim \mathrm{NB}(r_j, p_k). \tag{72}$$

1) Posterior Analysis: Assume we already observe $\{X_j\}_{1,J}$ and a set of discrete atoms $\mathcal{D} = \{\omega_k\}_{1,K}$. Since the beta and NB distributions are conjugate, at an observed discrete atom $\omega_k \in \mathcal{D}$, with $p_k = B(\omega_k)$ and $n_{jk} = X_j(\omega_k)$, we have $p_k|\{r_j, X_j\}_{1,J} \sim \mathrm{Beta}\left(\sum_{j=1}^J n_{jk}, c + \sum_{j=1}^J r_j\right)$. For the continuous part $\Omega\backslash\mathcal{D}$, the Lévy measure can be expressed as $\nu(dpd\omega)|\{r_j, X_j\}_{1,J} = cp^{-1}(1-p)^{c+\sum_{j=1}^J r_j - 1}dpB_0(d\omega)$. Following the notation in [68], [27], [11], we have the posterior of the beta process as

$$B|\{r_j, X_j\}_{1,J} \sim \mathrm{BP}\left(c + \sum_{j=1}^J r_j, \frac{c}{c+\sum_{j=1}^J r_j}B_0 + \frac{1}{c+\sum_{j=1}^J r_j}\sum_{k=1}^K\sum_{j=1}^J n_{jk}\delta_{\omega_k}\right). \tag{73}$$

Placing a gamma prior $\mathrm{Gamma}(c_0, 1/d_0)$ on r_j, we have

$$l_{jk}|r_j, X_j \sim \mathrm{CRT}(n_{jk}, r_j),\ r_j|\{l_{jk}\}_k, B \sim \mathrm{Gamma}\left(c_0 + \sum_{k=1}^K l_{jk}, \frac{1}{d_0 - \sum_{k=1}^K \ln(1-p_k)}\right). \tag{74}$$

Note that if r_j are fixed as one, then the beta-NB process reduces to the beta-geometric process discussed in [11], and if r_j are empirically set to some other values, then the beta-NB process reduces to the one proposed in [13]. These simplifications are not considered in the paper, as they are often overly restrictive.

With a discrete base measure $B_0 = \sum_{k=1}^K \frac{1}{K}\delta_{\phi_k}$, the beta-NB process topic model is constructed as

$$\begin{aligned} &x_{ji} \sim F(\boldsymbol{\phi}_{z_{ji}}),\quad \boldsymbol{\phi}_k \sim \mathrm{Dir}(\eta, \cdots, \eta) \\ &N_j = \sum_{k=1}^K n_{jk},\quad n_{jk} \sim \mathrm{Pois}(\theta_{jk}),\quad \theta_{jk} \sim \mathrm{Gamma}(r_j, p_k/(1-p_k)) \\ &\ r_j \sim \mathrm{Gamma}(e_0, 1/f_0),\quad p_k \sim \mathrm{Beta}(c/K, c(1-K)) \end{aligned} \tag{75}$$

The block Gibbs sampling can be expressed as

$$\begin{aligned}
&\Pr(z_{ji} = k|-) \propto \phi_{v_{ji}k}\theta_{jk} \\
&(p_k|-) \sim \text{Beta}\left(c/K + \sum_{j=1}^{J} n_{jk}, c(1-1/K) + \sum_{j=1}^{J} r_j\right), \quad l_{jk} \sim \text{CRT}(n_{jk}, r_j) \\
&(r_j|-) \sim \text{Gamma}\left(e_0 + \sum_{k=1}^{K} l_{jk}, \frac{1}{f_0 - \sum_{k=1}^{K}\ln(1-p_k)}\right) \\
&(\theta_{jk}|-) \sim \text{Gamma}(r_j + n_{jk}, p_k) \\
&(\boldsymbol{\phi}_k|-) \sim \text{Dir}\left(\eta + n_{1\cdot k}, \cdots, \eta + n_{V\cdot k}\right).
\end{aligned} \tag{76}$$

F. Marked-Beta-Negative Binomial Process

We may also consider a marked-beta-NB process that both the probability and dispersion measures are shared, in which each random point (ω_k, p_k) of the beta process is marked with an independent gamma random variable r_k taking values in $\mathbb{R}_+$. Using the marked Poisson process theorem [8], we may regard $(R, B) = \sum_{k=1}^{\infty}(r_k, p_k)\delta_{\omega_k}$ as a random draw from a marked beta process defined in the product space $[0,1] \times \mathbb{R}_+ \times \Omega$, with Lévy measure

$$\nu(dpdrd\omega) = cp^{-1}(1-p)^{c-1}dpR_0(dr)B_0(d\omega) \tag{77}$$

where R_0 is a continuous finite measure over $\mathbb{R}_+$. A marked-beta-NB process can be constructed by letting $X_j \sim \text{NBP}(R, B)$, with a random draw expressed as

$$X_j = \sum_{k=1}^{\infty} n_{jk}\delta_{\omega_k}, \;\; n_{jk} \sim \text{NB}(r_k, p_k). \tag{78}$$

1) Posterior Analysis: At an observed discrete atom $\omega_k \in \mathcal{D}$, with $r_k = R(\omega_k)$, we have $p_k|R, \{X_j\}_{1,J} \sim \text{Beta}\left(\sum_{j=1}^{J} n_{jk}, c + Jr_k\right)$. For the continuous part $\Omega \backslash \mathcal{D}$, with $r = R(\omega)$ for $\omega \in \Omega\backslash\mathcal{D}$, we have $\nu(dpd\omega)|R, \{X_j\}_{1,J} = cp^{-1}(1-p)^{c+Jr-1}dpB_0(d\omega)$. Thus the posterior of B can be expressed as

$$B|R, \{X_j\}_{1,J} \sim \text{BP}\left(c_J, \tfrac{c}{c_J}B_0 + \tfrac{1}{c_J}\sum_{k=1}^{K}\sum_{j=1}^{J} n_{jk}\delta_{\omega_k}\right) \tag{79}$$

where c_J is the concentration function as $c_J(\omega) = c + JR(\omega) + \sum_{j=1}^{J} X_j(\omega)$. Let $R_0(dr)/R_0(\mathbb{R}_+) = \text{Gamma}(r; e_0, 1/f_0)dr$, then for $\omega_k \in \mathcal{D}$, we have

$$l_{jk}|R, X_j \sim \text{CRT}(n_{jk}, r_k), \;\; r_k|\{l_{jk}\}_{j=1,J}, B \sim \text{Gamma}\left(e_0 + \sum_{j=1}^{J} l_{jk}, \tfrac{1}{f_0 - J\ln(1-p_k)}\right) \tag{80}$$

and for $\omega \in \Omega \backslash \mathcal{D}$, the posterior of $r = R(\omega)$ is the same as the prior $r \sim \text{Gamma}(e_0, 1/f_0)$.

With a discrete base measure $B_0 = \sum_{k=1}^{K} \frac{1}{K} \delta_{\boldsymbol{\phi}_k}$, the Marked-Beta-NB process topic model is constructed as

$$\begin{aligned}
& x_{ji} \sim F(\boldsymbol{\phi}_{z_{ji}}), \quad \boldsymbol{\phi}_k \sim \text{Dir}(\eta, \cdots, \eta) \\
& N_j = \sum_{k=1}^{K} n_{jk}, \quad n_{jk} \sim \text{Pois}(\theta_{jk}), \quad \theta_{jk} \sim \text{Gamma}(r_k, p_k/(1-p_k)) \\
& r_k \sim \text{Gamma}(e_0, 1/f_0), \quad p_k \sim \text{Beta}(c/K, c(1-K))
\end{aligned} \tag{81}$$

The block Gibbs sampling can be expressed as

$$\begin{aligned}
& \Pr(z_{ji} = k | -) \propto \phi_{v_{ji}k} \theta_{jk} \\
& p_k \sim \text{Beta}\left(c/K + \sum_{j=1}^{J} n_{jk}, c(1-1/K) + J r_k\right), \quad l_{jk} \sim \text{CRT}(n_{jk}, r_k) \\
& (r_k | -) \sim \text{Gamma}\left(e_0 + \sum_{j=1}^{J} l_{jk}, \frac{1}{f_0 - J \ln(1-p_k)}\right) \\
& (\theta_{jk} | -) \sim \text{Gamma}(r_k + n_{jk}, p_k) \\
& (\boldsymbol{\phi}_k | -) \sim \text{Dir}\left(\eta + n_{1 \cdot k}, \cdots, \eta + n_{V \cdot k}\right).
\end{aligned} \tag{82}$$

G. Marked-Gamma-Negative Binomial Process

We may also consider a marked-gamma-NB process that each random point (r_k, ω_k) of the gamma process is marked with an independent beta random variable p_k taking values in $[0,1]$. We may regard $(G, P) = \sum_{k=1}^{\infty} (r_k, p_k) \delta_{\omega_k}$ as a random draw from a marked gamma process defined in the product space $\mathbb{R}_+ \times [0,1] \times \Omega$, with Lévy measure

$$\nu(drdpd\omega) = r^{-1} e^{-cr} dr P_0(dp) G_0(d\omega) \tag{83}$$

where P_0 is a continuous finite measure over $[0,1]$.

1) Posterior Analysis: At an observed discrete atom $\omega_k \in \mathcal{D}$, we have

$$l_{jk} | G, X_j \sim \text{CRT}(n_{jk}, r_k), \;\; r_k | \{l_{jk}\}_{j=1,J}, P \sim \text{Gamma}\left(\sum_{j=1}^{J} l_{jk}, \frac{1}{c - J \ln(1-p_k)}\right) \tag{84}$$

where $r_k = G(\omega_k)$ and $p_k = P(\omega_k)$. For the continuous part $\Omega \backslash \mathcal{D}$, with $p = P(\omega)$ for $\omega \in \Omega \backslash \mathcal{D}$, the Lévy measure of G can be expressed as $\nu(drd\omega) | P, \{X_j\}_{1,J} = r^{-1} e^{-(c - J \ln(1-p))r} dr G_0(d\omega)$. Thus the posterior of G can be expressed as

$$G|P, \{X_j\}_{1,J} \sim \text{GaP}\left(c_J, G_0 + \textstyle\sum_{k=1}^{K}\sum_{j=1}^{J} l_{jk}\delta_{\omega_k}\right) \tag{85}$$

where c_J is the concentration function as $c_J(\omega) = c - J\ln(1-P(\omega))$. Let $P_0(dp)/P_0([0,1]) =$ Beta$(p; a_0, b_0)dp$, then for $\omega_k \in \mathcal{D}$, we have

$$p_k|R, \{X_j\}_{1,J} \sim \text{Beta}\left(a_0 + \textstyle\sum_{j=1}^{J} n_{jk}, b_0 + Jr_k\right) \tag{86}$$

and for $\omega \in \Omega\backslash\mathcal{D}$, the posterior of $p = P(\omega)$ is the same as the prior $p \sim$ Beta(a_0, b_0).

With a discrete base measure $G_0 = \sum_{k=1}^{K} \frac{\gamma_0}{K}\delta_{\phi_k}$, the Marked-Gamma-NB process topic model is constructed as

$$\begin{aligned}
&x_{ji} \sim F(\boldsymbol{\phi}_{z_{ji}}), \quad \boldsymbol{\phi}_k \sim \text{Dir}(\eta, \cdots, \eta)\\
&N_j = \sum_{k=1}^{K} n_{jk}, \quad n_{jk} \sim \text{Pois}(\theta_{jk}),\ \theta_{jk} \sim \text{Gamma}(r_k, p_k/(1-p_k))\\
&r_k \sim \text{Gamma}(\gamma_0/K, 1/c), \quad p_k \sim \text{Beta}(a_0, b_0), \quad \gamma_0 \sim \text{Gamma}(e_0, 1/f_0).
\end{aligned} \tag{87}$$

The block Gibbs sampling can be expressed as

$$\begin{aligned}
&\Pr(z_{ji} = k|-) \propto \phi_{v_{ji}k}\theta_{jk}\\
&p_k \sim \text{Beta}\left(a_0 + \sum_{j=1}^{J} n_{jk}, b_0 + Jr_k\right), \quad p_k' = \frac{-J\ln(1-p_k)}{c - J\ln(1-p_k)}\\
&l_{jk} \sim \text{CRT}(n_{jk}, r_k),\ l_k' \sim \text{CRT}(\sum_{j=1}^{J} l_{jk}, \gamma_0/K),\ \gamma_0 \sim \text{Gamma}\left(e_0 + \sum_{k=1}^{K} l_k', \frac{1}{f_0 - \sum_{k=1}^{K}\ln(1-p_k')/K}\right)\\
&(r_k|-) \sim \text{Gamma}\left(\gamma_0/K + \sum_{j=1}^{J} l_{jk}, \frac{1}{c - J\ln(1-p_k)}\right), \quad (\theta_{jk}|-) \sim \text{Gamma}(r_k + n_{jk}, p_k)\\
&(\boldsymbol{\phi}_k|-) \sim \text{Dir}\left(\eta + n_{1\cdot k}, \cdots, \eta + n_{V\cdot k}\right).
\end{aligned} \tag{88}$$