# Augment-and-Conquer Negative Binomial Processes

**Mingyuan Zhou**
Dept. of Electrical and Computer Engineering
Duke University, Durham, NC 27708
mz1@ee.duke.edu

**Lawrence Carin**
Dept. of Electrical and Computer Engineering
Duke University, Durham, NC 27708
lcarin@ee.duke.edu

## Abstract

By developing data augmentation methods unique to the negative binomial (NB) distribution, we unite seemingly disjoint count and mixture models under the NB process framework. We develop fundamental properties of the models and derive efficient Gibbs sampling inference. We show that the gamma-NB process can be reduced to the hierarchical Dirichlet process with normalization, highlighting its unique theoretical, structural and computational advantages. A variety of NB processes with distinct sharing mechanisms are constructed and applied to topic modeling, with connections to existing algorithms, showing the importance of inferring both the NB dispersion and probability parameters.

## 1 Introduction

There has been increasing interest in count modeling using the Poisson process, geometric process [1, 2, 3, 4] and recently the negative binomial (NB) process [5, 6]. Notably, it has been independently shown in [5] and [6] 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 $\boldsymbol{x}_j = \{x_{ji}\}_{i=1,N_j}$. For a territory long occupied by the hierarchical Dirichlet process (HDP) [7] and related models, the inference of which may require substantial bookkeeping and suffer from slow convergence [7], the discovery of the NB process for mixture modeling can be significant. As the seemingly distinct problems of count and mixture modeling are united under the NB process framework, new opportunities emerge for better data fitting, more efficient inference and more flexible model constructions. However, neither [5] nor [6] explore the properties of the NB distribution deep enough to achieve fully tractable closed-form inference. Of particular concern is the NB dispersion parameter, which was simply fixed or empirically set [6], or inferred with a Metropolis-Hastings algorithm [5]. Under these limitations, both papers fail to reveal the connections of the NB process to the HDP, and thus may lead to false assessments on comparing their modeling abilities.

We perform joint count and mixture modeling under the NB process framework, using completely random measures [1, 8, 9] that are simple to construct and amenable for posterior computation. We propose to augment-and-conquer the NB process: by "augmenting" a NB process into both the gamma-Poisson and compound Poisson representations, we "conquer" the unification of count and mixture modeling, the analysis of fundamental model properties, and the derivation of efficient Gibbs sampling inference. We make two additional contributions: 1) we construct a gamma-NB process, analyze its properties and show how its normalization leads to the HDP, highlighting its unique theoretical, structural and computational advantages relative to the HDP. 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; we compare their performance on 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.

### 1.1 Poisson process for count and mixture modeling

Before introducing the NB process, we first illustrate how the seemingly distinct problems of count and mixture modeling can be united under the Poisson process. Denote $\Omega$ 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 $\Omega$, we aim to jointly model the count random variables $\{X_j(A_q)\}$. A natural choice would be to define a Poisson process $X_j \sim \text{PP}(G)$, with a shared completely random measure $G$ on $\Omega$, such that $X_j(A) \sim \text{Pois}\big(G(A)\big)$ for each $A \subset \Omega$. Denote $G(\Omega) = \sum_{q=1}^{Q} G(A_q)$ and $\widetilde{G} = G/G(\Omega)$. Following Lemma 4.1 of [5], the joint distributions of $X_j(\Omega), X_j(A_1), \cdots, X_j(A_Q)$ are equivalent under the following two expressions:

$$X_j(\Omega) = \textstyle\sum_{q=1}^{Q} X_j(A_q), \qquad X_j(A_q) \sim \text{Pois}\big(G(A_q)\big); \tag{1}$$

$$X_j(\Omega) \sim \text{Poisson}(G(\Omega)), \quad [X_j(A_1), \cdots, X_j(A_q)] \sim \text{Mult}\big(X_j(\Omega); \widetilde{G}(A_1), \cdots, \widetilde{G}(A_Q)\big). \tag{2}$$

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 observations into any measurable disjoint partition $\{A_q\}_{1,Q}$ of $\Omega$, conditioning on $X_j(\Omega)$ and the normalized mean measure $\widetilde{G}$.

To complete the model, we may place a gamma process [9] prior on the shared measure as $G \sim \text{GaP}(c, G_0)$, with concentration parameter $c$ and base measure $G_0$, such that $G(A) \sim \text{Gamma}(G_0(A), 1/c)$ for each $A \subset \Omega$, where $G_0$ can be continuous, discrete or a combination of both. Note that $\widetilde{G} = G/G(\Omega)$ now becomes a Dirichlet process (DP) as $\widetilde{G} \sim \text{DP}(\gamma_0, \widetilde{G}_0)$, where $\gamma_0 = G_0(\Omega)$ and $\widetilde{G}_0 = G_0/\gamma_0$. The normalized gamma representation of the DP is discussed in [10, 11, 9] and has been used to construct the group-level DPs for an HDP [12]. The Poisson process has an equal-dispersion assumption for count modeling. As shown in (2), the construction of Poisson processes with a shared gamma process mean measure implies the same mixture proportions across groups, which is essentially the same as the DP when used for mixture modeling when the total counts $\{X_j(\Omega)\}_j$ are not treated as random variables. This motivates us to consider adding an additional layer or using a different distribution other than the Poisson to model the counts. As shown below, the NB distribution is an ideal candidate, not only because it allows overdispersion, but also because it can be augmented into both a gamma-Poisson and a compound Poisson representations.

## 2 Augment-and-Conquer the Negative Binomial Distribution

The NB distribution $m \sim \text{NB}(r, p)$ has the probability mass function (PMF) $f_M(m) = \frac{\Gamma(r+m)}{m!\Gamma(r)}(1-p)^r p^m$. 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 $\sigma^2$) as $r^{-1}$. It has been widely investigated and applied to numerous scientific studies [13, 14, 15]. The NB distribution can be augmented into a gamma-Poisson construction as $m \sim \text{Pois}(\lambda),\ \lambda \sim \text{Gamma}\left(r, p/(1-p)\right)$, where the gamma distribution is parameterized by its shape $r$ and scale $p/(1-p)$. It can also be augmented under a compound Poisson representation [16] as $m = \sum_{t=1}^{l} u_t,\ u_t \sim \text{Log}(p),\ l \sim \text{Pois}(-r\ln(1-p))$, where $u \sim \text{Log}(p)$ is the logarithmic distribution [17] with probability-generating function (PGF) $C_U(z) = \ln(1-pz)/\ln(1-p),\ |z| < p^{-1}$. In a slight abuse of notation, but for added conciseness, in the following discussion we use $m \sim \sum_{t=1}^{l} \text{Log}(p)$ to denote $m = \sum_{t=1}^{l} u_t,\ u_t \sim \text{Log}(p)$.

The inference of the NB dispersion parameter $r$ has long been a challenge [13, 18, 19]. In this paper, we first place a gamma prior on it as $r \sim \text{Gamma}(r_1, 1/c_1)$. We then use Lemma 2.1 (below) to infer a latent count $l$ for each $m \sim \text{NB}(r, p)$ conditioning on $m$ and $r$. Since $l \sim \text{Pois}(-r\ln(1-p))$ by construction, we can use the gamma Poisson conjugacy to update $r$. Using Lemma 2.2 (below), we can further infer an augmented latent count $l'$ for each $l$, and then use these latent counts to update $r_1$, assuming $r_1 \sim \text{Gamma}(r_2, 1/c_2)$. Using Lemmas 2.1 and 2.2, we can continue this process repeatedly, suggesting that we may build a NB process to model data that have subgroups within groups. The conditional posterior of the latent count $l$ was first derived by us but was not given an analytical form [20]. Below we explicitly derive the PMF of $l$, shown in (3), and find that it exactly represents the distribution of the random number of tables occupied by $m$ customers in a Chinese restaurant process with concentration parameter $r$ [21, 22, 7]. We denote $l \sim \text{CRT}(m, r)$ as a Chinese restaurant table (CRT) count random variable with such a PMF and as proved in the supplementary material, we can sample it as $l = \sum_{n=1}^{m} b_n,\ b_n \sim \text{Bernoulli}\left(r/(n-1+r)\right)$.

Both the gamma-Poisson and compound-Poisson augmentations of the NB distribution and Lemmas 2.1 and 2.2 are key ingredients of this paper. We will show that these augment-and-concur methods not only unite count and mixture modeling and provide efficient inference, but also, as shown in Section 3, let us examine the posteriors to understand fundamental properties of the NB processes, clearly revealing connections to previous nonparametric Bayesian mixture models.

**Lemma 2.1.** *Denote $s(m,j)$ as Stirling numbers of the first kind [17]. Augment $m \sim \text{NB}(r,p)$ under the compound Poisson representation as $m \sim \sum_{t=1}^{l} \text{Log}(p),\ l \sim \text{Pois}(-r\ln(1-p))$, then the conditional posterior of $l$ has PMF*

$$\Pr(l=j|m,r) = \tfrac{\Gamma(r)}{\Gamma(m+r)}|s(m,j)|r^j,\ \ j=0,1,\cdots,m. \tag{3}$$

*Proof.* Denote $w_j \sim \sum_{t=1}^{j}\text{Log}(p)$, $j=1,\cdots,m$. Since $w_j$ is the summation of $j$ iid $\text{Log}(p)$ random variables, the PGF of $w_j$ becomes $C_{W_j}(z) = C_U^j(z) = [\ln(1-pz)/\ln(1-p)]^j$, $|z| < p^{-1}$. Using the property that $[\ln(1+x)]^j = j!\sum_{n=j}^{\infty}\frac{s(n,j)x^n}{n!}$ [17], we have $\Pr(w_j=m) = C_{W_j}^{(m)}(0)/m! = (-1)^m p^m j! s(m,j)/(m![\ln(1-p)]^j)$. Thus for $0 \le j \le m$, we have $\Pr(L=j|m,r) \propto \Pr(w_j=m)\text{Pois}(j;-r\ln(1-p)) \propto |s(m,j)|r^j$. Denote $S_r(m) = \sum_{j=0}^{m}|s(m,j)|r^j$, we have $S_r(m) = (m-1+r)S_r(m-1) = \cdots = \prod_{n=1}^{m-1}(r+n)S_r(1) = \prod_{n=0}^{m-1}(r+n) = \frac{\Gamma(m+r)}{\Gamma(r)}$. □

**Lemma 2.2.** *Let $m \sim \text{NB}(r,p)$, $r \sim \text{Gamma}(r_1, 1/c_1)$, denote $p' = \frac{-\ln(1-p)}{c_1-\ln(1-p)}$, then $m$ can also be generated from a compound distribution as*

$$m \sim \textstyle\sum_{t=1}^{l}\text{Log}(p),\ l \sim \sum_{t'=1}^{l'}\text{Log}(p'),\ l' \sim\ \text{Pois}(-r_1\ln(1-p')). \tag{4}$$

*Proof.* Augmenting $m$ leads to $m \sim \sum_{t=1}^{l}\text{Log}(p)$, $l \sim \text{Pois}(-r\ln(1-p))$. Marginalizing out $r$ leads to $l \sim \text{NB}(r_1, p')$. Augmenting $l$ using its compound Poisson representation leads to (4). □

## 3 Gamma-Negative Binomial Process

We explore sharing the NB dispersion across groups while the probability parameters are group dependent. We define a NB process $X \sim \text{NBP}(G,p)$ as $X(A) \sim \text{NB}(G(A),p)$ for each $A \subset \Omega$ and construct a gamma-NB process for joint count and mixture modeling as $X_j \sim \text{NBP}(G,p_j),\ G \sim \text{GaP}(c,G_0)$, which can be augmented as a gamma-gamma-Poisson process as

$$X_j \sim \text{PP}(\Lambda_j),\ \ \Lambda_j \sim \text{GaP}((1-p_j)/p_j, G),\ \ G \sim \text{GaP}(c,G_0). \tag{5}$$

In the above $\text{PP}(\cdot)$ and $\text{GaP}(\cdot)$ represent the Poisson and gamma processes, respectively, as defined in Section 1.1. Using Lemma 2.2, the gamma-NB process can also be augmented as

$$X_j \sim \textstyle\sum_{t=1}^{L_j}\text{Log}(p_j),\ L_j \sim \text{PP}(-G\ln(1-p_j)),\ G \sim \text{GaP}(c,G_0); \tag{6}$$

$$L = \textstyle\sum_j L_j \sim \sum_{t=1}^{L'}\text{Log}(p'),\ L' \sim \text{PP}(-G_0\ln(1-p')),\ \ p' = \frac{-\sum_j\ln(1-p_j)}{c-\sum_j\ln(1-p_j)}. \tag{7}$$

These three augmentations allow us to derive a sequence of closed-form update equations for inference with the gamma-NB process. Using the gamma Poisson conjugacy on (5), for each $A \subset \Omega$, we have $\Lambda_j(A)|G, X_j, p_j \sim \text{Gamma}\left(G(A)+X_j(A), p_j\right)$, thus the conditional posterior of $\Lambda_j$ is

$$\Lambda_j|G, X_j, p_j \sim \text{GaP}\big(1/p_j, G+X_j\big). \tag{8}$$

Define $T \sim \text{CRTP}(X,G)$ as a CRT process that $T(A) = \sum_{\omega\in A}T(\omega)$, $T(\omega) \sim \text{CRT}(X(\omega), G(\omega))$ for each $A \subset \Omega$. Applying Lemma 2.1 on (6) and (7), we have

$$L_j|X_j, G \sim \text{CRTP}(X_j, G), \quad L'|L, G_0 \sim \text{CRTP}(L, G_0). \tag{9}$$

If $G_0$ is a continuous base measure and $\gamma_0 = G_0(\Omega)$ is finite, we have $G_0(\omega) \to 0\ \forall\,\omega \in \Omega$ and thus

$$L'(\Omega)|L, G_0 = \textstyle\sum_{\omega\in\Omega}\delta(L(\omega)>0) = \sum_{\omega\in\Omega}\delta(\sum_j X_j(\omega)>0) \tag{10}$$

which is equal to $K^+$, the total number of used discrete atoms; if $G_0$ is discrete as $G_0 = \sum_{k=1}^{K}\frac{\gamma_0}{K}\delta_{\omega_k}$, then $L'(\omega_k) = \text{CRT}(L(\omega_k), \frac{\gamma_0}{K}) \ge 1$ if $\sum_j X_j(\omega_k) > 0$, thus $L'(\Omega) \ge K^+$. In either case, let $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$, with the gamma Poisson conjugacy on (6) and (7), we have

$$\gamma_0|\{L'(\Omega), p'\} \sim \text{Gamma}\big(e_0 + L'(\Omega), \tfrac{1}{f_0-\ln(1-p')}\big); \tag{11}$$

$$G|G_0, \{L_j, p_j\} \sim \text{GaP}\big(c - \textstyle\sum_j\ln(1-p_j), G_0 + \sum_j L_j\big). \tag{12}$$

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 $\Lambda_j$ marginalized out, can be expressed as

$$X_{ji}|G, X_j^{-i} \sim \tfrac{\mathbb{E}[\Lambda_j|G,X_j^{-i}]}{\mathbb{E}[\Lambda_j(\Omega)|G,X_j^{-i}]} = \tfrac{G}{G(\Omega)+X_j(\Omega)-1} + \tfrac{X_j^{-i}}{G(\Omega)+X_j(\Omega)-1}. \tag{13}$$

### 3.1 Relationship with the hierarchical Dirichlet process

Using the equivalence between (1) and (2) and normalizing all the gamma processes in (5), denoting $\widetilde{\Lambda}_j = \Lambda_j/\Lambda_j(\Omega)$, $\alpha = G(\Omega)$, $\widetilde{G} = G/\alpha$, $\gamma_0 = G_0(\Omega)$ and $\widetilde{G}_0 = G_0/\gamma_0$, we can re-express (5) as

$$X_{ji} \sim \widetilde{\Lambda}_j,\ \widetilde{\Lambda}_j \sim \text{DP}(\alpha, \widetilde{G}),\ \alpha \sim \text{Gamma}(\gamma_0, 1/c),\ \widetilde{G} \sim \text{DP}(\gamma_0, \widetilde{G}_0) \tag{14}$$

which is an HDP [7]. Thus the normalized gamma-NB process leads to an HDP, yet we cannot return from the HDP to the gamma-NB process without modeling $X_j(\Omega)$ and $\Lambda_j(\Omega)$ 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 $\{A_q\}_{1,Q}$ of $\Omega$; whereas the HDP is not. Practically, the gamma-NB process can exploit conjugacy to achieve analytical conditional posteriors for all latent parameters. The inference of the HDP is a major challenge and it is usually solved through alternative constructions such as the Chinese restaurant franchise (CRF) and stick-breaking representations [7, 23]. In particular, without analytical conditional posteriors, the inference of concentration parameters $\alpha$ and $\gamma_0$ is nontrivial [7, 24] and they are often simply fixed [23]. Under the CRF metaphor $\alpha$ governs the random number of tables occupied by customers in each restaurant independently; further, if the base probability measure $\widetilde{G}_0$ is continuous, $\gamma_0$ governs the random number of dishes selected by tables of all restaurants. One may apply the data augmentation method of [22] to sample $\alpha$ and $\gamma_0$. 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 $K \to \infty$ [11, 7, 24], then using the method of [22] to sample $\gamma_0$ 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 (12) and $G(\Omega)$ plays the role of $\alpha$ in the HDP, which is readily available as

$$G(\Omega)|G_0, \{L_j, p_j\}_{j=1,N} \sim \text{Gamma}\Big(\gamma_0 + \textstyle\sum_j L_j(\Omega), \frac{1}{c - \sum_j \ln(1-p_j)}\Big) \tag{15}$$

and as in (11), regardless of whether the base measure is continuous, the total mass $\gamma_0$ 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 (15) 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 (14), 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 [12] 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 [12] 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 [12].

### 3.2 Augment-and-conquer inference for joint count and mixture modeling

For a finite continuous base measure, the gamma process $G \sim \text{GaP}(c, G_0)$ can also be defined with its Lévy measure on a product space $\mathbb{R}_+ \times \Omega$, expressed as $\nu(drd\omega) = r^{-1}e^{-cr}drG_0(d\omega)$ [9]. Since the Poisson intensity $\nu^+ = \nu(\mathbb{R}_+ \times \Omega) = \infty$ and $\int_{\mathbb{R}_+\times\Omega} r\nu(drd\omega)$ is finite, a draw from this process can be expressed as $G = \sum_{k=1}^{\infty} r_k\delta_{\omega_k},\ (r_k, \omega_k) \sim \pi(drd\omega),\ \pi(drd\omega)\nu^+ \equiv \nu(drd\omega)$ [9]. Here we consider a discrete base measure as $G_0 = \sum_{k=1}^{K} \frac{\gamma_0}{K}\delta_{\omega_k},\ \omega_k \sim g_0(\omega_k)$, 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 gamma-NB process with the discrete base measure as

$$\begin{aligned} \omega_k \sim g_0(\omega_k),\ N_j = \textstyle\sum_{k=1}^{K} n_{jk},\ n_{jk} \sim \text{Pois}(\lambda_{jk}),\ \lambda_{jk} \sim \text{Gamma}(r_k, p_j/(1-p_j)) \\ r_k \sim \text{Gamma}(\gamma_0/K, 1/c),\ p_j \sim \text{Beta}(a_0, b_0),\ \gamma_0 \sim \text{Gamma}(e_0, 1/f_0) \end{aligned} \tag{16}$$

where marginally we have $n_{jk} \sim \text{NB}(r_k, p_j)$. Using the equivalence between (1) and (2), we can equivalently express $N_j$ and $n_{jk}$ in the above model as $N_j \sim \text{Pois}\left(\lambda_j\right),\ [n_{j1}, \cdots, n_{jK}] \sim \text{Mult}\left(N_j; \lambda_{j1}/\lambda_j, \cdots, \lambda_{jK}/\lambda_j\right)$, where $\lambda_j = \sum_{k=1}^{K} \lambda_{jk}$. 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 the index

$$z_{ji} \sim \text{Discrete}\left(\lambda_{j1}/\lambda_j, \cdots, \lambda_{jK}/\lambda_j\right) \tag{17}$$

for each $x_{ji}$ and then let $n_{jk} = \sum_{i=1}^{N_j} \delta(z_{ji} = k)$. This provides further insights on how the seemingly disjoint problems of count and mixture modeling are united under the NB process framework. Following (8)-(12), the block Gibbs sampling is straightforward to write as

$$p(\omega_k|-) \propto \prod_{z_{ji}=k} F(x_{ji};\omega_k) g_0(\omega_k), \ \ \Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\lambda_{jk}$$

$$(p_j|-) \sim \text{Beta}\Big(a_0 + N_j, b_0 + \textstyle\sum_k r_k\Big), \ p' = \frac{-\sum_j \ln(1-p_j)}{c-\sum_j \ln(1-p_j)}, \ (l_{jk}|-) \sim \text{CRT}(n_{jk}, r_k)$$

$$(l'_k|-) \sim \text{CRT}(\textstyle\sum_j l_{jk}, \gamma_0/K), \ \ (\gamma_0|-) \sim \text{Gamma}\Big(e_0 + \textstyle\sum_k l'_k, \frac{1}{f_0 - \ln(1-p')}\Big)$$

$$(r_k|-) \sim \text{Gamma}\Big(\gamma_0/K + \textstyle\sum_j l_{jk}, \frac{1}{c-\sum_j \ln(1-p_j)}\Big), \ \ (\lambda_{jk}|-) \sim \text{Gamma}(r_k + n_{jk}, p_j). \quad (18)$$

which has similar computational complexity as that of the direct assignment block Gibbs sampling of the CRF-HDP [7, 24]. If $g_0(\omega)$ is conjugate to the likelihood $F(x;\omega)$, then the posterior $p(\omega|-)$ would be analytical. Note that when $K \to \infty$, we have $(l'_k|-) = \delta(\sum_j l_{jk} > 0) = \delta(\sum_j n_{jk} > 0)$.

Using (1) and (2) and normalizing the gamma distributions, (16) can be re-expressed as

$$z_{ji} \sim \text{Discrete}(\tilde{\boldsymbol{\lambda}}_j), \ \tilde{\boldsymbol{\lambda}}_j \sim \text{Dir}(\alpha\tilde{\boldsymbol{r}}), \ \alpha \sim \text{Gamma}(\gamma_0, 1/c), \ \tilde{\boldsymbol{r}} \sim \text{Dir}(\gamma_0/K, \cdots, \gamma_0/K) \quad (19)$$

which loses the count modeling ability and becomes a finite representation of the HDP, the inference of which is not conjugate and has to be solved under alternative representations [7, 24]. This also implies that by using the Dirichlet process as the foundation, traditional mixture modeling may discard useful count information from the beginning.

## 4 The Negative Binomial Process Family and Related Algorithms

The gamma-NB process shares the NB dispersion across groups. Since the NB distribution has two adjustable parameters, we may explore alternative ideas, with the NB probability measure shared across groups as in [6], or with both the dispersion and probability measures shared as in [5]. These constructions are distinct from both the gamma-NB process and HDP in that $\Lambda_j$ has space dependent scales, and thus its normalization $\widetilde{\Lambda}_j = \Lambda_j/\Lambda_j(\Omega)$ no longer follows a Dirichlet process.

It is natural to let the probability measure be drawn from a beta process [25, 26], which can be defined by its Lévy measure on a product space $[0,1]\times\Omega$ as $\nu(dpd\omega) = cp^{-1}(1-p)^{c-1}dpB_0(d\omega)$. A draw from the beta process $B \sim \text{BP}(c, B_0)$ with concentration parameter $c$ and base measure $B_0$ can be expressed as $B = \sum_{k=1}^{\infty} p_k\delta_{\omega_k}$. A beta-NB process [5, 6] can be constructed by letting $X_j \sim \text{NBP}(r_j, B)$, 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. As in [5], we may also consider a marked-beta-NB[1] process that both the NB probability and dispersion measures are 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 $X_j \sim \text{NBP}(R, B)$ becomes $X_j = \sum_{k=1}^{\infty} n_{jk}\delta_{\omega_k}$, $n_{jk} \sim \text{NB}(r_k, p_k)$. Since the beta and NB processes are conjugate, the posterior of $B$ is tractable, as shown in [5, 6]. 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 $X_j \sim \text{NBP}(RZ_j, p_j)$, where $Z_j \sim \text{BeP}(B)$ is drawn from the Bernoulli process [26] and $(R, B) = \sum_{k=1}^{\infty}(r_k, \pi_k)\delta_{\omega_k}$ is drawn from a marked-beta process, thus $n_{jk} \sim \text{NB}(r_k b_{jk}, p_j)$, $b_{jk} = \text{Bernoulli}(\pi_k)$. This construction can be linked to the model in [27] 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 [28, 29, 30, 31].

### 4.1 Related Algorithms

To show how the NB processes can be diversely constructed and to make connections to previous parametric and nonparametric mixture models, we show in Table 1 a variety of NB processes, which differ on how the dispersion and probability measures are shared. For a deeper understanding on how the counts are modeled, we also show in Table 1 both the VMR and ODL implied by these

[1] We may also consider a beta marked-gamma-NB process, whose performance is found to be very similar.

Table 1: A variety of negative binomial processes are constructed with distinct sharing mechanisms, reflected with which parameters from $r_k$, $r_j$, $p_k$, $p_j$ and $\pi_k$ ($b_{jk}$) 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 | $r_k$ | $r_j$ | $p_k$ | $p_j$ | $\pi_k$ | VMR | ODL | Related Algorithms |
|---|---|---|---|---|---|---|---|---|
| NB-LDA | | ✓ | | ✓ | | $(1-p_j)^{-1}$ | $r_j^{-1}$ | LDA [32], Dir-PFA [5] |
| NB-HDP | ✓ | | | 0.5 | | 2 | $r_k^{-1}$ | HDP[7], DILN-HDP [12] |
| NB-FTM | ✓ | | | 0.5 | ✓ | 2 | $(r_k)^{-1}b_{jk}$ | FTM [27], S$\gamma\Gamma$-PFA [5] |
| Beta-NB | | ✓ | ✓ | | | $(1-p_k)^{-1}$ | $r_j^{-1}$ | BNBP [5], BNBP [6] |
| Gamma-NB | ✓ | | | ✓ | | $(1-p_j)^{-1}$ | $r_k^{-1}$ | CRF-HDP [7, 24] |
| Marked-Beta-NB | ✓ | | ✓ | | | $(1-p_k)^{-1}$ | $r_k^{-1}$ | BNBP [5] |

settings. We consider topic modeling of a document corpus, a typical example of mixture modeling of grouped data, where each a-bag-of-words document constitutes a group, each word is an exchangeable group member, and $F(x_{ji};\omega_k)$ is simply the probability of word $x_{ji}$ in topic $\omega_k$.

We consider six differently constructed NB processes in Table 1: (*i*) Related to latent Dirichlet allocation (LDA) [32] and Dirichlet Poisson factor analysis (Dir-PFA) [5], 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 inference. Thus even the most basic parametric LDA topic model can be improved under the NB count modeling framework. (*ii*) The NB-HDP model is related to the HDP [7], 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 [12] that constructs the group-level DPs with normalized gamma processes, whose scale parameters are also set as one. (*iii*) 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 [5] and is comparable to the focused topic model (FTM) [27], which is constructed from the IBP compound DP. Nevertheless, they apply about the same likelihoods and priors for inference. The Zero-Inflated-NB process improves over them by allowing $p_j$ to be inferred, which generally yields better data fitting. (*iv*) The Gamma-NB process explores the idea that the dispersion measure is shared across groups, and it improves over the NB-HDP by allowing the learning of $p_j$. It reduces to the HDP [7] by normalizing both the group-level and the shared gamma processes. (*v*) The Beta-NB process explores sharing the probability measure across groups, and it improves over the beta negative binomial process (BNBP) proposed in [6], allowing inference of $r_j$. (*vi*) The Marked-Beta-NB process is comparable to the BNBP proposed in [5], with the distinction that it allows analytical update of $r_k$. The constructions and inference of various NB processes and related algorithms in Table 1 all follow the formulas in (16) and (18), respectively, with additional details presented in the supplementary material.

Note that as shown in [5], NB process topic models can also be considered as factor analysis of the term-document count matrix under the Poisson likelihood, with $\omega_k$ as the $k$th factor loading that sums to one and $\lambda_{jk}$ as the factor score, which can be further linked to nonnegative matrix factorization [33] and a gamma Poisson factor model [34]. If except for proportions $\tilde{\boldsymbol{\lambda}}_j$ and $\tilde{r}$, the absolute values, e.g., $\lambda_{jk}$, $r_k$ and $p_k$, are also of interest, then the NB process based joint count and mixture models would apparently be more appropriate than the HDP based mixture models.

## 5 Example Results

Motivated by Table 1, we consider topic modeling using a variety of NB processes, which differ on which parameters are learned and consequently how the VMR and ODL of the latent counts $\{n_{jk}\}_{j,k}$ are modeled. We compare them with LDA [32, 35] and CRF-HDP [7, 24]. 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. For LDA and NB-LDA, we search $K$ for optimal performance and for the other models, we set $K = 400$ as an upper-bound. We set the parameters as $c = 1$, $\eta = 0.05$ and $a_0 = b_0 = e_0 = f_0 = 0.01$. For LDA, we set the topic proportion Dirichlet smoothing parameter as $50/K$, following the topic model toolbox[2] provided for [35]. We consider 2500 Gibbs sampling iterations, with the last 1500 samples collected. Under the NB processes, each word $x_{ji}$ would

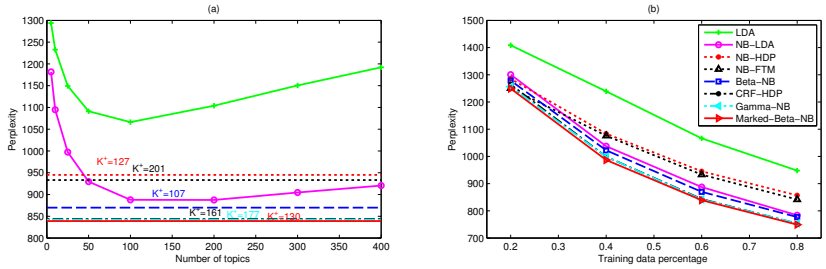


Figure 1: Comparison of per-word perplexities on the held-out words between various algorithms. (a) With 60% of the words in each document used for training, the performance varies as a function of $K$ in both LDA and NB-LDA, which are parametric models, whereas the NB-HDP, NB-FTM, Beta-NB, CRF-HDP, Gamma-NB and Marked-Beta-NB all infer the number of active topics, which are 127, 201, 107, 161, 177 and 130, respectively, according to the last Gibbs sampling iteration. (b) Per-word perplexities of various models as a function of the percentage of words in each document used for training. The results of the LDA and NB-LDA are shown with the best settings of $K$ under each training/testing partition.

be assigned to a topic $k$ based on both $F(x_{ji};\omega_k)$ and the topic weights $\{\lambda_{jk}\}_{k=1,K}$; each topic is drawn from a Dirichlet base measure as $\omega_k \sim \text{Dir}(\eta,\cdots,\eta) \in \mathbb{R}^V$, where $V$ is the number of unique terms in the vocabulary and $\eta$ is a smoothing parameter. Let $v_{ji}$ denote the location of word $x_{ji}$ in the vocabulary, then we have $(\omega_k|-) \sim \text{Dir}\big(\eta+\sum_j\sum_i \delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_j\sum_i\delta(z_{ji}=k,v_{ji}=V)\big)$. We consider the Psychological Review[2] corpus, restricting the vocabulary to terms that occur in five or more documents. The corpus includes 1281 abstracts from 1967 to 2003, with 2,566 unique terms 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 $v$ as $f_{jv}=\sum_{s=1}^S\sum_{k=1}^K \omega_{vk}^{(s)}\lambda_{jk}^{(s)}/\sum_{s=1}^S\sum_{v=1}^V\sum_{k=1}^K\omega_{vk}^{(s)}\lambda_{jk}^{(s)}$, where $\omega_{vk}$ is the probability of term $v$ in topic $k$ and $S$ is the total number of collected samples. We use $\{f_{jv}\}_{j,v}$ to calculate the per-word perplexity on the held-out words as in [5]. The final results are averaged from five random training/testing partitions. Note that the perplexity per test word is the fair metric to compare topic models. However, when the actual Poisson rates or distribution parameters for counts instead of the mixture proportions are of interest, it is obvious that a NB process based joint count and mixture model would be more appropriate than an HDP based mixture model.

Figure 1 compares the performance of various algorithms. The Marked-Beta-NB process has the best performance, closely followed by the Gamma-NB process, CRF-HDP and Beta-NB process. With an appropriate $K$, the parametric NB-LDA may outperform the nonparametric NB-HDP and NB-FTM as the training data percentage increases, somewhat unexpected but very intuitive results, 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.

Figure 2 shows the learned model parameters by various algorithms under the NB process framework, revealing distinct sharing mechanisms and model properties. When $(r_j,p_j)$ is used, as in the NB-LDA, different documents are weakly coupled with $r_j \sim \text{Gamma}(\gamma_0,1/c)$, and the modeling results 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. 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 close to one or close to 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 middle range $p_k$ indicating an appreciable mean and small overdispersion is not favored by the model and thus is rarely observed. 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

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

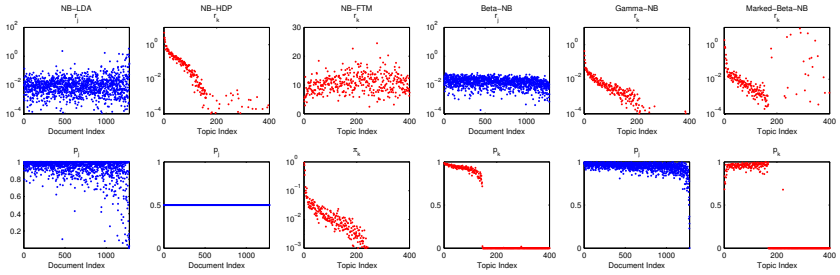


Figure 2: Distinct sharing mechanisms and model properties are evident between various NB processes, by comparing their inferred parameters. Note that the transition between active and non-active topics is very sharp when $p_k$ is used and much smoother when $r_k$ is used. Both the documents and topics 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. The values are shown in either linear or log scales for convenient visualization.

more topics than $(r_j, p_k)$, as confirmed in Figure 1 (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 $(r_k, p_j)$.

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 $\mathbb{E}[\sum_j n_{jk}] = Jr_kp_k/(1-p_k)$. For example, there could be not only large mean and small overdispersion (large $r_k$ and small $p_k$), but also large mean and large overdispersion (small $r_k$ and large $p_k$). 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. When $(r_k, \pi_k)$ is used, as in the NB-FTM model, our results show that we usually have a small $\pi_k$ 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 or IBP processes, e.g., when the training words are scarce.

It is also interesting to compare the Gamma-NB and NB-HDP. From a mixture-modeling viewpoint, fixing $p_j = 0.5$ is natural 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, and the experimental results in Figure 1 confirm the importance of learning $p_j$ together with $r_k$. It is also interesting to examine (15), which can be viewed as the concentration parameter $\alpha$ 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.

## 6 Conclusions

We propose a variety of negative binomial (NB) processes to jointly model counts across groups, which can be naturally applied for mixture modeling of grouped data. The proposed NB processes are completely random measures that they assign independent random variables to disjoint Borel sets of the measure space, as opposed to the hierarchical Dirichlet process (HDP) whose measures on disjoint Borel sets are negatively correlated. We discover augment-and-conquer inference methods that by "augmenting" a NB process into both the gamma-Poisson and compound Poisson representations, we are able to "conquer" the unification of count and mixture modeling, the analysis of fundamental model properties and the derivation of efficient Gibbs sampling inference. We demonstrate that the gamma-NB process, which shares the NB dispersion measure across groups, can be normalized to produce the HDP and we show in detail its theoretical, structural and computational advantages over the HDP. We examine the distinct sharing mechanisms and model properties of various NB processes, with connections to existing algorithms, with experimental results on topic modeling showing the importance of modeling both the NB dispersion and probability parameters.

### Acknowledgments

The research reported here was supported by ARO, DOE, NGA, and ONR, and by DARPA under the MSEE and HIST programs.

## References


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

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

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

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

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

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

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

[8] M. I. Jordan. Hierarchical models, nested models and completely random measures. 2010.

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

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

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

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

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

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

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

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

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

[18] S. J. Clark and J. N. Perry. Estimation of the negative binomial parameter $\kappa$ by maximum quasi-likelihood. *Biometrics*, 1989.

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

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

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

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

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

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

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

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

[27] 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.

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

[29] 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.

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

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

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

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

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

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

# A Generating a CRT random variable

**Lemma A.1.** *A CRT random variable $l \sim \text{CRT}(m, r)$ can be generated with the summation of independent Bernoulli random variables as*

$$l = \sum_{n=1}^{m} b_n,\ b_n \sim \text{Bernoulli}\left(\frac{r}{n-1+r}\right). \tag{20}$$

*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 $f_L(l|m,r) = \frac{C_L^{(l)}(0)}{l!} = \frac{\Gamma(r)}{\Gamma(m+r)}|s(m,l)|r^l,\ \ l = 0, 1, \cdots, m.$ □

# B Dir-PFA and LDA

The Dirichlet Poisson factor analysis (Dir-PFA) model [5] is constructed as

$$\begin{aligned} &x_{ji} \sim F(\omega_{z_{ji}}),\ \ \omega_k \sim \text{Dir}(\eta, \cdots, \eta) \\ &N_j = \sum_{k=1}^{K} n_{jk},\ \ n_{jk} \sim \text{Pois}(\tilde{\lambda}_{jk}),\ \ \tilde{\boldsymbol{\lambda}}_j \sim \text{Dir}(50/K, \cdots, 50/K) \end{aligned} \tag{21}$$

where $\eta$ is the Dirichlet smoothing parameter for the topic's distribution over the vocabulary, $n_{jk} = \sum_{i=1}^{N_j}\delta(z_{ji}=k)$, and the data likelihood $F(x_{ji};\omega_k)$ in topic modeling is $\omega_{v_{ji}k}$, the probability of the $i$th word in $j$th document under topic $\omega_k$.

The Dir-PFA has the same block Gibbs sampling as LDA [34], expressed as

$$\begin{aligned} &\Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\tilde{\lambda}_{jk} \\ &(\omega_k|-) \sim \text{Dir}\left(\eta + \sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k, v_{ji}=1), \cdots, \eta + \sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k, v_{ji}=V)\right) \\ &(\tilde{\boldsymbol{\lambda}}_j|-) \sim \text{Dir}\left(50/K + n_{j1}, \cdots, 50/K + n_{jK}\right). \end{aligned} \tag{22}$$

# C CRF-HDP

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

$$\begin{aligned} &x_{ji} \sim F(\omega_{z_{ji}}),\ \ \omega_k \sim \text{Dir}(\eta, \cdots, \eta),\ \ z_{ji} \sim \text{Discrete}(\tilde{\boldsymbol{\lambda}}_j) \\ &\tilde{\boldsymbol{\lambda}}_j \sim \text{Dir}(\alpha\tilde{\boldsymbol{r}}),\ \ \alpha \sim \text{Gamma}(a_0, 1/b_0),\ \ \tilde{\boldsymbol{r}} \sim \text{Dir}(\gamma_0/K, \cdots, \gamma_0/K). \end{aligned} \tag{23}$$

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

$$
\begin{aligned}
&\Pr(z_{ji}=k|-) \propto F(x_{ji};\omega_k)\tilde{\lambda}_{jk}\\
&(l_{jk}|-) \sim \text{CRT}(n_{jk},\alpha\tilde{r}_k),\ \ w_j \sim \text{Beta}(\alpha+1,N_j),\ \ s_j \sim \text{Bernoulli}\left(\frac{N_j}{N_j+\alpha}\right)\\
&\alpha \sim \text{Gamma}\left(a_0+\sum_{j=1}^{J}\sum_{k=1}^{K} l_{jk}-\sum_{j=1}^{J} s_j, \frac{1}{b_0-\sum_j \ln w_j}\right)\\
&(\tilde{\boldsymbol{r}}|-) \sim \text{Dir}\left(\gamma_0/K+\sum_{j=1}^{J} l_{j1},\cdots,\gamma_0/K+\sum_{j=1}^{J} l_{jK}\right)\\
&(\tilde{\boldsymbol{\lambda}}_j|-) \sim \text{Dir}\left(\alpha\tilde{r}_1+n_{j1},\cdots,\alpha\tilde{r}_K+n_{jK}\right)\\
&(\omega_k|-) \sim \text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right). \quad (24)
\end{aligned}
$$

When $K\to\infty$, the concentration parameter $\gamma_0$ can be sampled as

$$
\begin{aligned}
&w_0 \sim \text{Beta}\left(\gamma_0+1,\sum_{j=1}^{J}\sum_{k=1}^{\infty} l_{jk}\right),\ \ \pi_0=\frac{e_0+K^+-1}{e_0+K^+-1+(f_0-\ln w_0)\sum_{j=1}^{J}\sum_{k=1}^{\infty} l_{jk}}\\
&\gamma_0 \sim \pi_0\text{Gamma}\left(e_0+K^+,\frac{1}{f_0-\ln w_0}\right)+(1-\pi_0)\text{Gamma}\left(e_0+K^+-1,\frac{1}{f_0-\ln w_0}\right) \quad (25)
\end{aligned}
$$

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 $\gamma_0$ is only approximately correct, which may result in a biased estimate especially if $K$ is not set large enough. Thus in the experiments, $\gamma_0$ is not sampled and is fixed as one. Note that for implementation convenience, it is also common to fix the concentration parameter $\alpha$ 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 $\alpha$ using the data augmentation method proposed in [7], which is modified from the one proposed in [24].

## D NB-LDA

The NB-LDA model is constructed as

$$
\begin{aligned}
&x_{ji} \sim F(\omega_{z_{ji}}),\ \ \omega_k \sim \text{Dir}(\eta,\cdots,\eta)\\
&N_j=\sum_{k=1}^{K} n_{jk},\ \ n_{jk} \sim \text{Pois}(\lambda_{jk}),\ \ \lambda_{jk} \sim \text{Gamma}(r_j,p_j/(1-p_j))\\
&r_j \sim \text{Gamma}(\gamma_0,1/c),\ \ p_j \sim \text{Beta}(a_0,b_0),\ \ \gamma_0 \sim \text{Gamma}(e_0,1/f_0) \quad (26)
\end{aligned}
$$

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

$$
\begin{aligned}
&\Pr(z_{ji}=k|-)\propto F(x_{ji};\omega_k)\lambda_{jk}\\
&(p_j|-)\sim \text{Beta}\left(a_0+N_j,b_0+Kr_j\right),\ \ p'_j=\frac{-K\ln(1-p_j)}{c-K\ln(1-p_j)}\\
&(l_{jk}|-)\sim\ \text{CRT}(n_{jk},r_j),\ \ l'_j\sim\text{CRT}(\sum_{k=1}^{K}l_{jk},\gamma_0),\ \ \gamma_0\sim\text{Gamma}\left(e_0+\sum_{j=1}^{J}l'_j,\frac{1}{f_0-\sum_{j=1}^{J}\ln(1-p'_j)}\right)\\
&(r_j|-)\sim\text{Gamma}\left(\gamma_0+\sum_{k=1}^{K}l_{jk},\frac{1}{c-K\ln(1-p_j)}\right),\ \ (\lambda_{jk}|-)\sim\text{Gamma}(r_j+n_{jk},p_j)\\
&(\omega_k|-)\sim\text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}
\tag{27}
$$

## E 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 (16) and (18) of the main paper, respectively.

## F NB-FTM

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

$$
\begin{aligned}
&x_{ji}\sim F(\omega_{z_{ji}}),\ \ \omega_k\sim\text{Dir}(\eta,\cdots,\eta)\\
&N_j=\sum_{k=1}^{K}n_{jk},\ \ n_{jk}\sim\text{Pois}(\lambda_{jk})\\
&\lambda_{jk}\sim\text{Gamma}(r_kb_{jk},0.5/(1-0.5))\\
&r_k\sim\text{Gamma}(\gamma_0,1/c),\ \gamma_0\sim\text{Gamma}(e_0,1/f_0)\\
&b_{jk}\sim\text{Bernoulli}(\pi_k),\ \ \pi_k\sim\text{Beta}(c/K,c(1-1/K)).
\end{aligned}
\tag{28}
$$

The block Gibbs sampling can be expressed as

$$
\begin{aligned}
&\Pr(z_{ji}=k|-)\propto F(x_{ji};\omega_k)\lambda_{jk}\\
&b_{jk}\sim\delta(n_{jk}=0)\text{Bernoulli}\left(\frac{\pi_k(1-0.5)^{r_k}}{\pi_k(1-0.5)^{r_k}+(1-\pi_k)}\right)+\delta(n_{jk}>0)\\
&\pi_k\sim\text{Beta}\Big(c/K+\sum_{j=1}^{J}b_{jk},c(1-1/K)+J-\sum_{j=1}^{J}b_{jk}\Big),\ p'_k=\frac{-\sum_j b_{jk}\ln(1-0.5)}{c-\sum_j b_{jk}\ln(1-0.5)}\\
&(l_{jk}|-)\sim\text{CRT}(n_{jk},r_k b_{jk}),\ (l'_k|-)\sim\text{CRT}\left(\sum_{j=1}^{J}l_{jk},\gamma_0\right)\\
&(\gamma_0|-)\sim\text{Gamma}\left(e_0+\sum_{k=1}^{K}l'_k,\frac{1}{f_0-\sum_{k=1}^{K}\ln(1-p'_k)}\right)\\
&(r_k|-)\sim\text{Gamma}\left(\gamma_0+\sum_{j=1}^{J}l_{jk},\frac{1}{c-\sum_{j=1}^{J}b_{jk}\ln(1-0.5)}\right)\\
&(\lambda_{jk}|-)\sim\text{Gamma}(r_k b_{jk}+n_{jk},0.5)\\
&(\omega_k|-)\sim\text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}
\tag{29}
$$

## G Beta-NB

The beta-NB process model is constructed as

$$
\begin{aligned}
&x_{ji}\sim F(\omega_{z_{ji}}),\ \ \omega_k\sim\text{Dir}(\eta,\cdots,\eta)\\
&N_j=\sum_{k=1}^{K}n_{jk},\ \ n_{jk}\sim\text{Pois}(\lambda_{jk}),\ \ \lambda_{jk}\sim\text{Gamma}(r_j,p_k/(1-p_k))\\
&r_j\sim\text{Gamma}(e_0,1/f_0),\ \ p_k\sim\text{Beta}(c/K,c(1-K))
\end{aligned}
\tag{30}
$$

The block Gibbs sampling can be expressed as

$$
\begin{aligned}
&\Pr(z_{ji}=k|-)\propto F(x_{ji};\omega_k)\lambda_{jk}\\
&(p_k|-)\sim\text{Beta}\left(c/K+\sum_{j=1}^{J}n_{jk},c(1-1/K)+\sum_{j=1}^{J}r_j\right),\ \ l_{jk}\sim\text{CRT}(n_{jk},r_j)\\
&(r_j|-)\sim\text{Gamma}\left(e_0+\sum_{k=1}^{K}l_{jk},\frac{1}{f_0-\sum_{k=1}^{K}\ln(1-p_k)}\right)\\
&(\lambda_{jk}|-)\sim\text{Gamma}(r_j+n_{jk},p_k)\\
&(\omega_k|-)\sim\text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}
\tag{31}
$$

## H Marked-Beta-NB

The Marked-Beta-NB process model is constructed as

$$
\begin{aligned}
&x_{ji}\sim F(\omega_{z_{ji}}),\ \ \omega_k\sim\text{Dir}(\eta,\cdots,\eta)\\
&N_j=\sum_{k=1}^{K}n_{jk},\ \ n_{jk}\sim\text{Pois}(\lambda_{jk}),\ \ \lambda_{jk}\sim\text{Gamma}(r_k,p_k/(1-p_k))\\
&r_k\sim\text{Gamma}(e_0,1/f_0),\ \ p_k\sim\text{Beta}(c/K,c(1-K))
\end{aligned}
\tag{32}
$$

The block Gibbs sampling can be expressed as

$$\begin{aligned}
&\Pr(z_{ji}=k|-)\propto F(x_{ji};\omega_k)\lambda_{jk}\\
&p_k\sim \text{Beta}\left(c/K+\sum_{j=1}^{J}n_{jk}, c(1-1/K)+Jr_k\right),\ \ l_{jk}\sim \text{CRT}(n_{jk},r_k)\\
&(r_k|-)\sim \text{Gamma}\left(e_0+\sum_{j=1}^{J}l_{jk}, \frac{1}{f_0-J\ln(1-p_k)}\right)\\
&(\lambda_{jk}|-)\sim \text{Gamma}(r_k+n_{jk},p_k)\\
&(\omega_k|-)\sim \text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}\tag{33}$$

# I Marked-Gamma-NB

The Marked-Gamma-NB process model is constructed as

$$\begin{aligned}
&x_{ji}\sim F(\omega_{z_{ji}}),\ \ \omega_k\sim \text{Dir}(\eta,\cdots,\eta)\\
&N_j=\sum_{k=1}^{K}n_{jk},\ \ n_{jk}\sim \text{Pois}(\lambda_{jk}),\ \lambda_{jk}\sim \text{Gamma}(r_k,p_k/(1-p_k))\\
&r_k\sim \text{Gamma}(\gamma_0/K,1/c),\ \ p_k\sim \text{Beta}(a_0,b_0),\ \ \gamma_0\sim \text{Gamma}(e_0,1/f_0).
\end{aligned}\tag{34}$$

The block Gibbs sampling can be expressed as

$$\begin{aligned}
&\Pr(z_{ji}=k|-)\propto F(x_{ji};\omega_k)\lambda_{jk}\\
&p_k\sim \text{Beta}\left(a_0+\sum_{j=1}^{J}n_{jk}, b_0+Jr_k\right),\ \ p'_k=\frac{-J\ln(1-p_k)}{c-J\ln(1-p_k)}\\
&l_{jk}\sim \text{CRT}(n_{jk},r_k),\ l'_k\sim \text{CRT}(\sum_{j=1}^{J}l_{jk},\gamma_0/K),\ \gamma_0\sim \text{Gamma}\left(e_0+\sum_{k=1}^{K}l'_k, \frac{1}{f_0-\sum_{k=1}^{K}\ln(1-p'_k)/K}\right)\\
&(r_k|-)\sim \text{Gamma}\left(\gamma_0/K+\sum_{j=1}^{J}l_{jk}, \frac{1}{c-J\ln(1-p_k)}\right),\ \ (\lambda_{jk}|-)\sim \text{Gamma}(r_k+n_{jk},p_k)\\
&(\omega_k|-)\sim \text{Dir}\left(\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=1),\cdots,\eta+\sum_{j=1}^{J}\sum_{i=1}^{N_j}\delta(z_{ji}=k,v_{ji}=V)\right).
\end{aligned}\tag{35}$$