Bayesian Analysis (2018) **13**, Number 4, pp. 1065–1093

Nonparametric Bayesian Negative Binomial Factor Analysis

Mingyuan Zhou*

Abstract. A common approach to analyze a covariate-sample count matrix, an element of which represents how many times a covariate appears in a sample, is to factorize it under the Poisson likelihood. We show its limitation in capturing the tendency for a covariate present in a sample to both repeat itself and excite related ones. To address this limitation, we construct negative binomial factor analysis (NBFA) to factorize the matrix under the negative binomial likelihood, and relate it to a Dirichlet-multinomial distribution based mixed-membership model. To support countably infinite factors, we propose the hierarchical gamma-negative binomial process. By exploiting newly proved connections between discrete distributions, we construct two blocked and a collapsed Gibbs sampler that all adaptively truncate their number of factors, and demonstrate that the blocked Gibbs sampler developed under a compound Poisson representation converges fast and has low computational complexity. Example results show that NBFA has a distinct mechanism in adjusting its number of inferred factors according to the sample lengths, and provides clear advantages in parsimonious representation, predictive power, and computational complexity over previously proposed discrete latent variable models, which either completely ignore burstiness, or model only the burstiness of the covariates but not that of the factors.

Keywords: burstiness, count matrix factorization, hierarchical gamma-negative binomial process, parsimonious representation, self- and cross-excitation.

1 Introduction

The need to analyze a covariate-sample count matrix, each of whose elements counts the number of time that a covariate appears in a sample, arises in many different settings, such as text analysis, next-generation sequencing, medical records mining, and consumer behavior studies. The mixed-membership model, widely used for text analysis (Blei et al., 2003) and population genetics (Pritchard et al., 2000), treats each sample as a bag of indices (words), and associates each index with both a covariate that is observed and a subpopulation that is latent. It makes the assumption that there are K latent subpopulations, each of which is characterized by how frequent the covariates are relative to each other within it. Given the total number of indices for a sample, it assigns each index independently to one of the K subpopulations, with a probability proportional to the product of the corresponding covariate's relative frequency in that subpopulation and that subpopulation's relative frequency in the sample. A mixed-membership model constructed in this manner, as shown in Zhou et al. (2012) and Zhou and Carin (2015), can also be connected to Poisson factor analysis (PFA) that factorizes the covariate-

*McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA, mingyuan.zhou@mccombs.utexas.edu

 https://doi.org/10.1214/17-BA1070

sample count matrix, under the Poisson likelihood, into the product of a nonnegative covariate-subpopulation factor loading matrix and a nonnegative subpopulation-sample factor score matrix. Each column of the factor loading matrix encodes the relative frequencies of the covariates in a subpopulation, while that of the factor score matrix encodes the weights of the subpopulations in a sample.

Despite the popularity of both approaches in analyzing the covariate-sample count matrix, they both make restrictive assumptions. Given the relative frequencies of the covariates in subpopulations and the relative frequencies of the subpopulations in samples, the mixed-membership model independently assigns each index to both a covariate and a subpopulation, and hence may not sufficiently capture the tendency for an index to excite the other ones in the same sample to take the same or related covariates. Whereas for PFA, given the factor loading and score matrices, it assumes that the variance and mean are the same for each covariate-sample count, and hence is likely to underestimate the variability of overdispersed counts.

In practice, however, highly overdispersed covariate-sample counts are frequently observed due to self- and cross-excitation of covariate frequencies, that is to say, some covariates are particularly intense and also make other related covariates intense. For example, the tendency for a word present in a document to appear repeatedly is a fundamental phenomenon in natural language that is commonly referred to as word burstiness (Church and Gale, 1995; Madsen et al., 2005; Doyle and Elkan, 2009). If a word is bursty in a document, it is also common for it to excite (stimulate) related words to exhibit burstiness. Without capturing the self- and cross-excitation (stimulation) of covariate frequencies or better modeling the overdispersed covariate-sample counts, the ultimate potential of the mixed-membership model and PFA will be limited no matter how the priors on latent parameters are adjusted. In addition, it could be a waste of computation if the model tries to increase the model capacity to better capture overdispersions that could be simply explained with self- and cross-excitations.

To remove these restrictions, we introduce negative binomial factor analysis (NBFA) to factorize the covariate-sample count matrix, in which we replace the Poisson distributions on which PFA is built, with the negative binomial (NB) distributions. As PFA is closely related to the canonical mixed-membership model built on the multinomial distribution, we show that NBFA is closely related to a Dirichlet-multinomial mixed-membership (DMMM) model that uses the Dirichlet-categorical (Dirichlet-multinomial) rather than categorical (multinomial) distributions to assign an index to both a covariate and a factor (subpopulation). From the viewpoint of count modeling, NBFA improves PFA by better modeling overdispersed counts, while from that of mixed-membership modeling, it improves the canonical mixed-membership model by capturing the burstiness at both the covariate and factor levels (*i.e.*, for topic modeling, it exhibits a rich-get-richer phenomenon at both the word and topic levels). In addition, we will show NBFA could significantly reduce the computation spent on large covariate-sample counts.

Note that with a different likelihood for counts and a different mechanism to generate both the covariate and factor indices, NBFA and the related DMMM model proposed in the paper complement, rather than compete with, PFA (Zhou et al., 2012;

Zhou and Carin, 2015). First, NBFA provides more significant advantages in modeling longer samples, where there is more need to capture both self- and cross-excitation of covariate frequencies. Second, various extensions built on PFA or the multinomial mixed-membership model, such as capturing the correlations between factors (Blei and Lafferty, 2005; Paisley et al., 2012) and learning multilayer deep representations (Ranganath et al., 2015; Gan et al., 2015; Zhou et al., 2016a), could also be applied to extend NBFA. In this paper, we will focus on constructing a nonparametric Bayesian NBFA with a potentially infinite number of factors, and leave a wide variety of potential extensions under this new framework to future research.

To avoid the need of selecting the number of factors K, for PFA and the closely related multinomial mixed-membership model, a number of nonparametric Bayesian priors can be employed to support a potentially infinite number of latent factors, such as the hierarchical Dirichlet process (Teh et al., 2006) and beta-negative binomial process (Zhou et al., 2012; Broderick et al., 2015; Zhou and Carin, 2015). To support countably infinite factors for NBFA, generalizing the gamma-negative binomial process (GNBP) (Zhou and Carin, 2015; Zhou et al., 2016b), we introduce a new nonparametric Bayesian prior: the hierarchical gamma-negative binomial process (hGNBP), where each of the J samples is assigned with a sample-specific GNBP and a globally shared gamma process is mixed with all the J GNBPs. We derive both blocked and collapsed Gibbs sampling for the hGNBP-NBFA, with the number of factors automatically inferred.

The remainder of the paper is organized as follows. In Section 2, we review PFA and the multinomial mixed-membership model. In Section 3, we introduce NBFA and its representation as a DMMM model, and compare them with related models. In Section 4, we propose nonparametric-Bayesian NBFA. In Section 5, we derive both blocked and collapsed Gibbs sampling algorithms. In Section 6, we first make comparisons between different sampling strategies and then compare the performance of various algorithms on real data. We conclude the paper in Section 7. The proofs and Gibbs sampling update equations are provided in the Supplementary Material (Zhou, 2017).

2 Poisson factor analysis and mixed-membership model

2.1 Poisson factor analysis

Let $\mathbf{N}$ be a $V \times J$ covariate-sample count matrix for V covariates among J samples, where n_{vj} is the (v, j) element of $\mathbf{N}$ and gives the number of times that sample j has covariate v. PFA factorizes $\mathbf{N}$ under the Poisson likelihood as

$$\mathbf{N} \sim \text{Pois}(\mathbf{\Phi\Theta}), \tag{1}$$

where "Pois" is the abbreviation for "Poisson," $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K) \in \mathbb{R}_+^{V \times K}$ represents the factor loading matrix, $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J) \in \mathbb{R}_+^{K \times J}$ represents the factor score matrix, and $\mathbb{R}_+ = \{x : x \geq 0\}$, with $\boldsymbol{\phi}_k = (\phi_{1k}, \ldots, \phi_{Vk})^T$ encoding the weights of the V covariates in factor k and $\boldsymbol{\theta}_j = (\theta_{1j}, \ldots, \theta_{Kj})^T$ encoding the popularity of the K factors in sample j. PFA in (1) has an augmented representation as

$$n_{vj} = \sum_{k=1}^{K} n_{vjk},\ n_{vjk} \sim \text{Pois}(\phi_{vk}\theta_{kj}). \tag{2}$$

As in Zhou et al. (2012), it can also be equivalently constructed by first generating n_{vj} and then assigning them to the latent factors using the multinomial distributions as

$$(n_{vj1}, \ldots, n_{vjK}) \,|\, n_{vj} \sim \text{Mult}\left(n_{vj}, \frac{\phi_{v1}\theta_{1j}}{\sum_{k=1}^{K}\phi_{vk}\theta_{kj}}, \ldots, \frac{\phi_{vK}\theta_{Kj}}{\sum_{k=1}^{K}\phi_{vk}\theta_{kj}}\right),$$
$$n_{vj} \sim \text{Pois}\left(\sum_{k=1}^{K}\phi_{vk}\theta_{kj}\right). \tag{3}$$

2.2 Multinomial mixed-membership model

This alternative representation suggests a potential link of PFA to a standard mixed-membership model for text analysis such as probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) and latent Dirichlet allocation (LDA) (Blei et al., 2003). Given the factors $\boldsymbol{\phi}_k$ and factor proportions $\boldsymbol{\theta}_j/\theta_{\cdot j}$, where $\sum_{v=1}^{V}\phi_{vk} = 1$ and $\theta_{\cdot j} := \sum_{k=1}^{K}\theta_{kj}$, a standard procedure is to associate $x_{ji} \in \{1, \ldots, V\}$, the ith index (word) of sample j, with factor (topic) $z_{ji} \in \{1, \ldots, K\}$, and generate a bag of indices $\{x_{j1}, \ldots, x_{jn_j}\}$ as

$$x_{ji} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}),\ z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j/\theta_{\cdot j}), \tag{4}$$

where $n_j = \sum_{v=1}^{V} n_{vj}$ and $n_{vj} = \sum_{i=1}^{n_j}\delta(x_{ji} = v)$. We refer to (4) as the multinomial mixed-membership model. LDA completes the multinomial mixed-membership model by imposing the Dirichlet priors on both $\{\boldsymbol{\phi}_k\}_k$ and $\{\boldsymbol{\theta}_j/\theta_{\cdot j}\}_j$ (Blei et al., 2003).

If in additional to the multinomial mixed-membership model described in (4), one further generates the sample lengths as

$$n_j \sim \text{Pois}(\theta_{\cdot j}), \tag{5}$$

then the joint likelihood of $\boldsymbol{x}_j := (x_{j1}, \ldots, x_{jn_j})$, $\boldsymbol{z}_j := (z_{j1}, \ldots, z_{jn_j})$, and n_j given $\boldsymbol{\Phi}$ and $\boldsymbol{\theta}_j$ can be expressed as

$$P(\boldsymbol{x}_j, \boldsymbol{z}_j, n_j \,|\, \boldsymbol{\Phi}, \boldsymbol{\theta}_j) = \frac{\prod_{v=1}^{V}\prod_{k=1}^{K} n_{vjk}!}{n_j!}\prod_{v=1}^{V}\prod_{k=1}^{K}\text{Pois}(n_{vjk}; \phi_{vk}\theta_{kj}),$$

whose product with the combinatorial coefficient $n_j!/(\prod_{v=1}^{V}\prod_{k=1}^{K} n_{vjk}!)$ becomes the same as the likelihood $P(\{n_{vj}, n_{vj1}, \ldots, n_{vjK}\}_v \,|\, \boldsymbol{\Phi}, \boldsymbol{\theta}_j)$ of (2).

From the viewpoint of PFA, shown in (2), and its alternative representation constituted by (4) and (5), a wide variety of discrete latent variable models, such as nonnegative matrix factorization (NMF) (Lee and Seung, 2001), PLSA, LDA, the gamma-Poisson model of Canny (2004), and the discrete component analysis of Buntine and Jakulin (2006), all have the same mechanism to model the covariate counts that they generate both the covariate and factor indices using the categorical distributions shown in (4); they mainly differ from each other on how the priors on $\boldsymbol{\phi}_k$ and $\boldsymbol{\theta}_j$ (or $\boldsymbol{\theta}_j/\theta_{\cdot j}$) are constructed (Zhou et al., 2012; Zhou and Carin, 2015).

3 Negative binomial factor analysis and Dirichlet-multinomial mixed-membership model

3.1 Negative binomial factor analysis

To better model overdispersed counts, instead of following PFA to factorize the covariate-sample count matrix under the Poisson likelihood, we propose negative binomial (NB) factor analysis (NBFA) to factorize it under the NB likelihood as

$$\boldsymbol{n}_j \sim \mathrm{NB}(\boldsymbol{\Phi}\boldsymbol{\theta}_j,\, p_j),$$

where $n \sim \mathrm{NB}(r,p)$ denote the NB distribution with probability mass function (PMF) $f_N(n \,|\, r,p) = \frac{\Gamma(n+r)}{n!\Gamma(r)} p^n (1-p)^r$, where $n \in \{0,1,\ldots\}$. NBFA has an augmented representation as

$$n_{vj} = \sum_{k=1}^{K} n_{vjk}, \quad n_{vjk} \sim \mathrm{NB}(\phi_{vk}\theta_{kj},\, p_j). \tag{6}$$

Similar to how the Poisson distribution is related to the multinomial distribution (*e.g.*, Dunson and Herring (2005) and Lemma 4.1 of Zhou et al. (2012)), we reveal how the NB distribution is related to the Dirichlet-multinomial distribution using Theorem 1 shown below, whose proof is provided in Appendix A. As in Mosimann (1962), marginalizing out $\boldsymbol{\theta}$ from $\boldsymbol{z} \sim \prod_{i=1}^{n} \mathrm{Cat}(z_i;\boldsymbol{\theta})$, $\boldsymbol{\theta} \sim \mathrm{Dir}(r_1,\ldots,r_K)$ leads to a Dirichlet-categorical (DirCat) distribution with PMF

$$P(\boldsymbol{z} \,|\, r_1,\ldots,r_K) = \frac{\Gamma(r.)}{\Gamma(n+r.)} \prod_{k=1}^{K} \frac{\Gamma(n_k+r_k)}{\Gamma(r_k)}, \tag{7}$$

where $n_k = \sum_{i=1}^{n} \delta(z_i = k)$, and a Dirichlet-multinomial (DirMult) distribution with PMF $P[(n_1,\ldots,n_K) \,|\, r_1,\ldots,r_K] = \frac{n!}{\prod_{k=1}^{K} n_k!} P(\boldsymbol{z} \,|\, r_1,\ldots,r_K)$.

Theorem 1 (The negative binomial and Dirichlet-multinomial distributions)**.** *Let* $\boldsymbol{x} = (x, x_1, \ldots, x_K)$ *be random variables generated as*

$$x = \sum_{k=1}^{K} x_k, \;\; x_k \sim \mathrm{NB}(r_k, p).$$

Set $r. = \sum_{k=1}^{K} r_k$ *and let* $\boldsymbol{y} = (y, y_1, \ldots, y_K)$ *be random variables generated as*

$$(y_1,\ldots,y_K) \sim \mathrm{DirMult}(y, r_1,\ldots,r_K), \;\; y \sim \mathrm{NB}(r., p).$$

Then the distribution of $\boldsymbol{x}$ *is the same as that of* $\boldsymbol{y}$.

Using Theorem 1, $(n_{vj}, n_{vj1},\ldots,n_{vjK})$ in (6) can be equivalently generated as

$$\begin{aligned} (n_{vj1},\ldots,n_{vjK}) \,|\, n_{vj} &\sim \mathrm{DirMult}\left(n_{vj}, \phi_{v1}\theta_{1j},\ldots,\phi_{vK}\theta_{Kj}\right), \\ n_{vj} &\sim \mathrm{NB}\left(\sum_{k=1}^{K} \phi_{vk}\theta_{kj},\; p_j\right). \end{aligned} \tag{8}$$

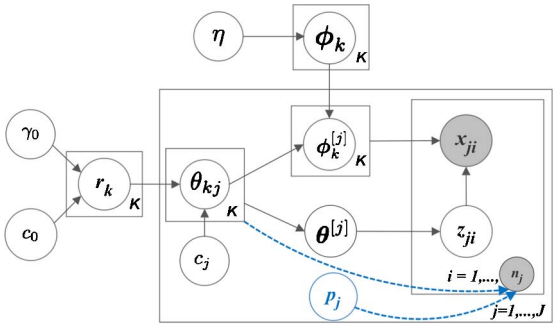

Figure 1: Graphical representations of the Dirichlet-multinomial mixed-membership model (without p_j and the dashed lines) and negative binomial factor analysis (with p_j and the dashed lines).

Clearly, how the factorization under the NB likelihood is related to the Dirichlet-multinomial distribution, as in (6) and (8), mimics how the factorization under the Poisson likelihood is related to the multinomial distribution, as in (2) and (3).

3.2 The Dirichlet-multinomial mixed-membership model

Similar to how we relate PFA in (2) to the multinomial topic model in (4), as in Section 2.1, we may relate NBFA in (6) to a mixed-membership model constructed by replacing the categorical distributions in (4) with the Dirichlet-categorical distributions as

$$\begin{aligned} &x_{ji} \sim \mathrm{Cat}(\boldsymbol{\phi}_{z_{ji}}^{[j]}),\ z_{ji} \sim \mathrm{Cat}(\boldsymbol{\theta}^{[j]}), \\ &\boldsymbol{\phi}_k^{[j]} \sim \mathrm{Dir}(\boldsymbol{\phi}_k \theta_{kj}),\ \boldsymbol{\theta}^{[j]} \sim \mathrm{Dir}(\boldsymbol{\theta}_j), \end{aligned} \tag{9}$$

where $\{\boldsymbol{\phi}_k^{[j]}\}_k$ and $\boldsymbol{\theta}^{[j]}$ represent the factors and factor scores specific for sample j, respectively. A graphical representation of the model, including the settings of the hyper-priors to be discussed later, is shown in Figure 1. Introducing $\boldsymbol{\phi}_k^{[j]}$ into the hierarchical model allows the same set of factors $\{\boldsymbol{\phi}_k\}_k$ to be manifested differently in different samples, whereas introducing $\boldsymbol{\theta}^{[j]}$ allows each sample to have two different representations: the factor scores $\boldsymbol{\theta}^{[j]}$ under the sample-specific factors $\{\boldsymbol{\phi}_k^{[j]}\}_k$, and the factor scores $\boldsymbol{\theta}_j$ under the shared factors $\{\boldsymbol{\phi}_k\}_k$. In addition, under this construction, the variance-to-mean ratio of $\phi_{vk}^{[j]}$ given $\boldsymbol{\phi}_k$ and θ_{kj} becomes $(1-\phi_{vk})/(\theta_{kj}+1)$, which monotonically decreases as the corresponding factor score θ_{kj} increases, allowing the variability of $\boldsymbol{\phi}_k^{[j]}$ in the prior to be controlled by the popularity of $\boldsymbol{\phi}_k$ in the corresponding sample. Moreover, this construction helps simplify the model likelihood and allows the model to be closely related to NBFA, as discussed below.

Explicitly drawing $\{\boldsymbol{\phi}_k^{[j]}\}_k$ for all samples would be computationally prohibitive, especially if the number of samples is large. Fortunately, this operation is totally unnecessary. By marginalizing out $\boldsymbol{\phi}_k^{[j]}$ and $\boldsymbol{\theta}^{[j]}$ in (9), we have

$$\{x_{ji}\}_{i:z_{ji}=k} \sim \text{DirCat}(n_{\cdot jk}, \phi_{1k}\theta_{kj}, \ldots, \phi_{Vk}\theta_{kj}), \quad \boldsymbol{z}_j \sim \text{DirCat}(n_j, \boldsymbol{\theta}_j), \tag{10}$$

where $\boldsymbol{z}_j := (z_1, \ldots, z_{jn_j})$, $n_{vjk} := \sum_{i=1}^{n_j} \delta(x_{ji} = v, z_{ji} = k)$, and $n_{\cdot jk} := \sum_{v=1}^{V} n_{vjk}$. Thus the joint likelihood of $\boldsymbol{x}_j := (x_{j1}, \ldots, x_{jn_j})$ and $\boldsymbol{z}_j$ given $\boldsymbol{\Phi}$, $\boldsymbol{\theta}_j$, and n_j can be expressed as

$$P(\boldsymbol{x}_j, \boldsymbol{z}_j \mid \boldsymbol{\Phi}, \boldsymbol{\theta}_j, n_j) = \frac{\Gamma(\theta_{\cdot j})}{\Gamma(n_j + \theta_{\cdot j})} \prod_{v=1}^{V} \prod_{k=1}^{K} \frac{\Gamma(n_{vjk} + \phi_{vk}\theta_{kj})}{\Gamma(\phi_{vk}\theta_{kj})}. \tag{11}$$

We call the model shown in (9) or (10) as the Dirichlet-multinomial mixed-membership (DMMM) model, whose likelihood given the factors and factor scores is shown in (11).

We introduce the following proposition, with proof provided in Appendix A, to show that one can recover NBFA from the DMMM model by randomizing its sample lengths with the NB distributions, and can reduce NBFA to the DMMM model by conditioning on these lengths. Thus how NBFA and the DMMM are related to each other mimics how PFA and the multinomial mixed-membership model are related to each other.

Proposition 2 (Dirichlet-multinomial mixed-membership (DMMM) modeling and negative binomial factor analysis)**.** *For the DMMM model that generates the covariate and factor indices using (9) or (10), if the sample lengths are randomized as*

$$n_j \sim \text{NB}(\theta_{\cdot j}, p_j), \tag{12}$$

then the joint likelihood of $\boldsymbol{x}_j$, $\boldsymbol{z}_j$, and n_j given $\boldsymbol{\Phi}$, $\boldsymbol{\theta}_j$, and p_j, multiplied by the combinatorial coefficient $n_j!/(\prod_{v=1}^{V}\prod_{k=1}^{K} n_{vjk}!)$, is equal to the likelihood of negative binomial factor analysis (NBFA) described in (6) or (8), expressed as

$$P(\{n_{vj}, n_{vj1}, \ldots, n_{vjK}\}_{v=1,V} \mid \boldsymbol{\Phi}, \boldsymbol{\theta}_j, p_j) = \prod_{v=1}^{V} \prod_{k=1}^{K} \text{NB}(n_{vjk}; \phi_{vk}\theta_{kj}, p_j). \tag{13}$$

The DMMM model could model not only the burstiness of the covariates, but also that of the factors via the Dirichlet-categorical distributions, as further explained below when discussing related models. As far as the conditional posteriors of $\boldsymbol{\phi}_k$ and $\boldsymbol{\theta}_j$ are concerned, the DMMM model with the lengths of its samples randomized via the NB distributions, as shown in (10) and (12), is equivalent to NBFA, as shown in (6). The representational differences, however, lead to different inference strategies, which will be discussed in detail along with their nonparametric Bayesian generalizations.

3.3 Comparisons with related models

Preceding the Dirichlet-multinomial mixed-membership (DMMM) model proposed in this paper, to account for covariate burstiness, Doyle and Elkan (2009) proposed Dirichlet compound multinomial LDA (DCMLDA) that can be expressed as

$$x_{ji} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}^{[j]}), \quad z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j), \quad \boldsymbol{\phi}_k^{[j]} \sim \text{Dir}(\boldsymbol{\phi}_k r_k), \tag{14}$$

where the Dirichlet prior is further imposed on $\boldsymbol{\theta}_j$. Note that the sample-specific factor scores $\{\boldsymbol{\theta}_j\}_j$ are represented under the sample-specific factors $\{\boldsymbol{\phi}_k^{[j]}\}_k$ in DCMLDA, as shown in (14), whereas they are represented under the same set of factors $\{\boldsymbol{\phi}_k\}_k$ in the DMMM model, as shown in (10).

Comparing (9) with (14), it is clear that removing $\boldsymbol{\theta}^{[j]}$ from (9) reduces the DMMM model to DCMLDA in (14). Moreover, if we further let

$$\boldsymbol{\theta}_j \sim \mathrm{Dir}(\boldsymbol{r}), \quad n_j \sim \mathrm{NB}(r_\cdot, p_j), \tag{15}$$

then the joint likelihood of $\boldsymbol{x}_j$, $\boldsymbol{z}_j$, and n_j given $\boldsymbol{\Phi}$, $\boldsymbol{r}$, and p_j can be expressed as

$$P(\boldsymbol{x}_j, \boldsymbol{z}_j, n_j \,|\, \boldsymbol{\Phi}, \boldsymbol{r}, p_j) = \frac{1}{n_j!} \prod_{v=1}^{V} \prod_{k=1}^{K} \frac{\Gamma(n_{vjk} + \phi_{vk} r_k)}{\Gamma(\phi_{vk} r_k)} p_j^{n_{vjk}} (1-p_j)^{\phi_{vk} r_k}, \tag{16}$$

which multiplied by the combinatorial coefficient $n_j!/(\prod_{v=1}^{V}\prod_{k=1}^{K} n_{vjk}!)$ is equal to the likelihood of $\{n_{vj}, n_{vj1}, \ldots, n_{vjK}\}_{v=1,V}$ given $\boldsymbol{\Phi}$, $\boldsymbol{r}$, and p_j in

$$n_{vj} = \sum_{k=1}^{K} n_{vjk}, \;\; n_{vjk} \sim \mathrm{NB}(\phi_{vk} r_k, p_j). \tag{17}$$

Thus, as far as the conditional posteriors of $\{\boldsymbol{\phi}_k\}_k$ and $\{r_k\}_k$ are concerned, DCMLDA constituted by (14)–(15) is equivalent to a special case of NBFA shown in (17), which is the augmented representation of $\boldsymbol{n}_j \sim \mathrm{NB}(\boldsymbol{\Phi}\boldsymbol{r}, p_j)$ that restricts all samples to have the same factor scores $\{r_k\}_k$ under the same set of shared factors $\{\boldsymbol{\phi}_k\}_k$.

Given the factors $\{\boldsymbol{\phi}_k\}_k$ and factor scores $\boldsymbol{\theta}_j$, for the multinomial mixed-membership model in (4), both the covariate and factor indices are independently drawn from the categorical distributions; for DCMLDA in (14), the factor indices but not the covariate indices are independently drawn from the categorical distributions; whereas for the DMMM model in (9), neither the factor indices nor covariate indices are independently drawn from the categorical distributions. Denoting y^{-ji} as the variable y obtained by excluding the contribution of the ith word in sample j, we compare in Table 1 these three different models on their predictive distributions of x_{ji} and z_{ji}.

In comparison to the multinomial mixed-membership model, DCMLDA allows the number of times that a covariate appears in a sample to exhibit the "rich get richer" (*i.e.*, self-excitation) behavior, leading to a better modeling of covariate burstiness, and the DMMM model further models the burstiness of the factor indices and hence encourages not only self-excitation, but also cross-excitation of covariate frequencies. It is clear from Table 1 that DCMLDA models covariate burstiness but not factor burstiness, and the corresponding NBFA restricts all samples to have the same factor scores under the shared factors $\{\boldsymbol{\phi}_k\}_k$. Thus we expect the DMMM model to clearly outperform DCMLDA, as will be confirmed by our experimental results.

Note that Zhou et al. (2016a) have recently extended the single-layer PFA into a multilayer one. For example, a two-layer PFA would further factorize the factor scores

Model	Predictive distribution for x_{ji}	Predictive distribution for z_{ji}
Multinomial mixed-membership	$P(x_{ji}=v\,\vert\,\boldsymbol{\Phi},\boldsymbol{\theta}_j)=\phi_{vz_{ji}}$	$P(z_{ji}=k\,\vert\,\boldsymbol{\theta}_j)=\theta_{kj}/\theta_{\cdot j}$
Dirichlet compound multinomial LDA	$P(x_{ji}=v\,\vert\,\boldsymbol{x}_j^{-ji},z_{ji},\boldsymbol{\Phi},\boldsymbol{r})=\frac{n_{vjz_{ji}}^{-ji}+\phi_{vz_{ji}}r_{z_{ji}}}{n_{\cdot jz_{ji}}^{-ji}+r_{z_{ji}}}$	$P(z_{ji}=k\,\vert\,\boldsymbol{\theta}_j)=\theta_{kj}/\theta_{\cdot j}$
Dirichlet-multinomial mixed-membership	$P(x_{ji}=v\,\vert\,\boldsymbol{x}_j^{-ji},z_{ji},\boldsymbol{\Phi},\boldsymbol{\theta}_j)=\frac{n_{vjz_{ji}}^{-ji}+\phi_{vz_{ji}}\theta_{z_{ji}j}}{n_{\cdot jz_{ji}}^{-ji}+\theta_{z_{ji}j}}$	$P(z_{ji}=k\,\vert\,\boldsymbol{z}_j^{-i},n_j,\boldsymbol{\theta}_j)=\frac{n_{\cdot jk}^{-ji}+\theta_{kj}}{n_j-1+\theta_{\cdot j}}$

Table 1: Comparisons of the predictive distributions for the multinomial mixed-membership model in (4), Dirichlet compound multinomial LDA (DCMLDA) in (14), and Dirichlet-multinomial mixed-membership (DMMM) model in (9).

$\boldsymbol{\theta}_j$ in $\boldsymbol{x}_j\sim\text{Pois}(\boldsymbol{\Phi}\boldsymbol{\theta}_j)$ under the gamma likelihood as $\boldsymbol{\theta}_j\sim\text{Gamma}[\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\theta}}_j,p_j/(1-p_j)]$, which is designed to capture the co-occurrence patterns of the first-layer factors $\boldsymbol{\phi}_k$. With θ_{jk} marginalized out from $n_{\cdot jk}\sim\text{Pois}(\theta_{jk})$, we have $n_{\cdot jk}\sim\text{NB}(\sum_{\tilde{k}}\tilde{\phi}_{k\tilde{k}}\tilde{\theta}_{j\tilde{k}},p_j)$, which can be considered as a NBFA model for the latent factor-sample counts. Thus using our previous analysis on NBFA, this multilayer construction models both the self- and cross-excitations of the first-layer factors and helps capture their bustiness, which could help better explain why adding an additional layer to PFA could often lead to clearly improved modeling performance, as shown in Zhou et al. (2016a). However, due to the choice of the Poisson likelihood at the first layer, a multilayer PFA does not directly capture the bustiness of the covariates. On the other hand, the single-layer NBFA models the self-excitation but not cross-excitation of factors. Therefore, extending the single-layer NBFA into a multi-layer one by exploiting the deep structure of Zhou et al. (2016a) may combine the advantages of both by not only capturing the covariate bustiness, but also better capturing the factor bustiness. That extension is outside of the scope of the paper and we leave it for future study.

4 Hierarchical gamma-negative binomial process negative binomial factor analysis

Let $G\sim\Gamma\text{P}(G_0,1/c_0)$ be a gamma process (Ferguson, 1973) defined on the product space $\mathbb{R}_+\times\Omega$, where $\mathbb{R}_+=\{x:x>0\}$, with two parameters: a finite and continuous base measure G_0 over a complete separable metric space Ω, and a scale $1/c_0$, such that $G(A)\sim\text{Gamma}(G_0(A),1/c_0)$ for each $A\subset\Omega$. The Lévy measure of the gamma process can be expressed as $\nu(drd\boldsymbol{\phi})=r^{-1}e^{-c_0r}drG_0(d\boldsymbol{\phi})$. A draw from $G\sim\Gamma\text{P}(G_0,1/c_0)$

can be represented as the countably infinite sum $G = \sum_{k=1}^{\infty} r_k \delta_{\boldsymbol{\phi}_k}$, $\boldsymbol{\phi}_k \sim g_0$, where $\gamma_0 = G_0(\Omega)$ as the mass parameter and $g_0(d\boldsymbol{\phi}) = G_0(d\boldsymbol{\phi})/\gamma_0$ is the base distribution.

To support countably infinite factors for the DMMM model, we consider a hierarchical gamma-negative binomial process (hGNBP) as

$$X_j \sim \mathrm{NBP}(\Theta_j, p_j),\ \Theta_j \sim \Gamma\mathrm{P}(G, 1/c_j),\ G \sim \Gamma\mathrm{P}(G_0, 1/c_0),$$

where $X \sim \mathrm{NBP}(\Theta, p)$ is a NB process defined such that $X(A_i) \sim \mathrm{NB}(\Theta(A_i), p)$ are independent NB random variables for disjoint partitions A_i of Ω. Given a gamma process random draw $G = \sum_{k=1}^{\infty} r_k \delta_{\boldsymbol{\phi}_k}$, we have

$$\Theta_j = \sum_{k=1}^{\infty} \theta_{kj} \delta_{\boldsymbol{\phi}_k},\ \theta_{kj} \sim \mathrm{Gamma}(r_k, 1/c_j),$$

where $\theta_{kj} := \Theta_j(\boldsymbol{\phi}_k)$ measures the weight of factor k in sample j, and

$$X_j = \sum_{k=1}^{\infty} n_{jk} \delta_{\boldsymbol{\phi}_k},\ n_{jk} \sim \mathrm{NB}(\theta_{kj}, p_j),$$

where $n_j = X_j(\Omega)$ is the length of sample j and $n_{jk} := X_j(\boldsymbol{\phi}_k) = \sum_{i=1}^{n_j} \delta(z_{ji} = k)$ represents the number of times that factor k appears in sample j.

We provide posterior analysis for the proposed hGNBP in Appendix B, where we utilize several additional discrete distributions, including the Chinese restaurant table (CRT), logarithmic, and sum-logarithmic (SumLog) distributions, that will also be used in the following discussion.

4.1 Hierarchical model

We express the hGNBP-DMMM model as

$$\begin{aligned}
&x_{ji} \sim \mathrm{Cat}(\boldsymbol{\phi}_{z_{ji}}^{[j]}),\ z_{ji} \sim \mathrm{Cat}(\boldsymbol{\theta}^{[j]}),\\
&\boldsymbol{\phi}_k^{[j]} \sim \mathrm{Dir}(\boldsymbol{\phi}_k \theta_{kj}),\ \boldsymbol{\theta}^{[j]} \sim \mathrm{Dir}(\boldsymbol{\theta}_j),\\
&\theta_{kj} \sim \mathrm{Gamma}(r_k, 1/c_j),\ c_j \sim \mathrm{Gamma}(e_0, 1/f_0),\\
&n_j \sim \mathrm{NB}(\theta_{\cdot j}, p_j),\ p_j \sim \mathrm{Beta}(a_0, b_0),\ G \sim \Gamma\mathrm{P}(G_0, 1/c_0),
\end{aligned} \tag{18}$$

where the atoms of the gamma process are drawn from a Dirichlet base distribution $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\eta, \ldots, \eta)$. We further let $\gamma_0 \sim \mathrm{Gamma}(a_0, 1/b_0)$ and $c_0 \sim \mathrm{Gamma}(e_0, 1/f_0)$. With Proposition 2, the above model can also be represented as a hGNBP-NBFA as

$$\begin{aligned}
&n_{vj} = \sum_{k=1}^{\infty} n_{vjk},\quad n_{vjk} \sim \mathrm{NB}(\phi_{vk}\theta_{kj}, p_j),\\
&\theta_{kj} \sim \mathrm{Gamma}(r_k, 1/c_j),\ c_j \sim \mathrm{Gamma}(e_0, 1/f_0),\\
&p_j \sim \mathrm{Beta}(a_0, b_0),\ G \sim \Gamma\mathrm{P}(G_0, 1/c_0).
\end{aligned} \tag{19}$$

The DMMM model in (18) and NBFA in (19) have the same conditional posteriors for both the factors $\{\boldsymbol{\phi}_k\}_k$ and factor scores $\{\boldsymbol{\theta}_j\}_j$, but lead to different inference strategies.

To infer $\{\boldsymbol{\phi}_k\}_k$ and $\{\boldsymbol{\theta}_j\}_j$, we first develop both blocked and collapsed Gibbs sampling with (18), and then develop blocked Gibbs sampling with (19), as described below.

5 Inference via Gibbs sampling

We present in this section three different Gibbs sampling algorithms, as summarized in Algorithm 1 in the Supplementary Material.

5.1 Blocked Gibbs sampling

As it is impossible to draw all the countably infinite atoms of a gamma process draw, expressed as $G = \sum_{k=1}^{\infty} r_k \delta_{\boldsymbol{\phi}_k}$, for the convenience of implementation, it is common to consider truncating the total number of atoms to be K by choosing a discrete base measure as $G_0 = \sum_{k=1}^{K} \frac{\gamma_0}{K} \delta_{\boldsymbol{\phi}_k}$, under which we have $r_k \sim \text{Gamma}(\gamma_0/K, 1/c_0)$ for $k \in \{1, \ldots, K\}$ (Zhou and Carin, 2015). The finite truncation strategy is also commonly used for Dirichlet process mixture models (Ishwaran and James, 2001; Fox et al., 2011). Although fixing K often works well in practice, it may lead to a considerable waste of computation if K is set to be too large. For nonparametric Bayesian mixture models based on the Dirichlet process (Ferguson, 1973; Escobar and West, 1995) or other more general normalized random measures with independent increments (NRMIs) (Regazzini et al., 2003; Lijoi et al., 2007), one may consider slice sampling to adaptively truncate the number of atoms used in each Markov chain Monte Carlo (MCMC) iteration (Walker, 2007; Papaspiliopoulos and Roberts, 2008). Unlike NRMIs whose atoms' weights have to sum to one and hence are negatively correlated, the weights of the atoms of completely random measures are independent from each other. Exploiting this property, for our models built on completely random measures, we construct a sampling procedure that adaptively truncates the total number of atoms in each iteration.

Note that the conditional posterior of G shown in (B.1) consists of two independent gamma processes: $\mathcal{D} \sim \Gamma\text{P}(\sum_j \widetilde{L}_j, [c_0 - \sum_j \ln(1-\tilde{p}_j)]^{-1})$ and $G_\star \sim \Gamma\text{P}(G_0, [c_0 - \sum_j \ln(1-\tilde{p}_j)]^{-1})$. To approximately represent a draw from $G_\star$ that consists of countably infinite atoms, at the end of each MCMC iteration, we relabel the indices of the atoms with nonzero counts from 1 to $K^+ := \sum_{k=1}^{\infty} \delta(\sum_j X_j(\boldsymbol{\phi}_k) > 0)$; draw $K_\star$ new atoms as

$$\widetilde{G}_\star = \sum_{k=K^++1}^{K^++K_\star} r_k \delta_{\boldsymbol{\phi}_k},\ r_k \sim \text{Gamma}\left(\frac{\gamma_0}{K_\star}, \frac{1}{c_0 - \sum_j \ln(1-\tilde{p}_j)}\right),\ \boldsymbol{\phi}_k \sim \text{Dir}(\eta, \ldots, \eta);$$

and set $K := K^+ + K_\star$ as the total number of atoms to be used for the next iteration.

For the hGNBP DMMM model, we present the update equations for z_{ji} and ℓ_{vjk} below and those for all the other parameters in Appendix C.1.

Sample z_{ji}**.** Using the likelihood in (11), we have

$$P(z_{ji} = k \,|\, x_{ji}, \boldsymbol{z}_j^{-i}, \boldsymbol{\Phi}, \boldsymbol{\theta}_j) \propto n_{x_{ji}jk}^{-ji} + \phi_{x_{ji}k}\theta_{kj}, \quad k \in \{1, \ldots, K\}. \tag{20}$$

Sample ℓ_{vjk}**.** Since $n_{vjk} \sim \mathrm{NB}(\phi_{vk}\theta_{kj}, p_j)$ in the prior, as shown in Proposition 2, we can draw a corresponding latent count ℓ_{vjk} for each n_{vjk} as

$$(\ell_{vjk} \,|\, -) \sim \mathrm{CRT}(n_{vjk}, \phi_{vk}\theta_{kj}), \tag{21}$$

where $\ell_{vjk} = 0$ almost surely if $n_{vjk} = 0$.

5.2 Collapsed Gibbs sampling

Let us denote $\boldsymbol{b} \sim \mathrm{CRP}(n, r)$ as an exchangeable random partition of the set $\{1, \ldots, n\}$, generated by assigning n customers (samples) to ℓ random number of tables (exclusive and nonempty subsets) using a Chinese restaurant process (CRP) (Aldous, 1983) with concentration parameter r. The exchangeable partition probability function of $\boldsymbol{b}$ under the CRP, also known as the Ewens sampling formula (Ewens, 1972; Antoniak, 1974), can be expressed as $P(\boldsymbol{b} \,|\, n, r) = \frac{\Gamma(r) r^{\ell}}{\Gamma(n+r)} \prod_{t=1}^{\ell} \Gamma(n_t)$, where $n_t = \sum_{i=1}^{n} \delta(b_i = t)$ is the size of the tth subset and ℓ is the number of subsets (Pitman, 2006).

One common strategy to improve convergence and mixing for multinomial mixed-membership models is to collapse the factors $\{\boldsymbol{\phi}_k\}_k$ and factor scores $\{\boldsymbol{\theta}_j\}_j$ in the sampler (Griffiths and Steyvers, 2004; Newman et al., 2009). To apply this strategy to the DMMM model, we first need to transform the likelihood in (16) to make it amenable to marginalization. Using an analogy similar to that for the Chinese restaurant franchise of Teh et al. (2006), if we consider z_{ji} as the index of the "dish" that the ith "customer" in the jth "restaurant" takes, then, to make the likelihood in (16) become fully factorized, we may introduce b_{ji} as the index of the table at which this customer is seated. The following proposition, whose proof is provided in Appendix A, reveals how the CRP can be related to the Dirichlet-multinomial and NB distributions, and shows how to introduce auxiliary variables to make the likelihood of $\boldsymbol{z} \sim \mathrm{DirCat}(n, r_1, \ldots, r_K)$, as shown in (7), become fully factorized.

Proposition 3. *Given the sample length n (number of customers) and $\boldsymbol{r} = (r_1, \ldots, r_K)$, the joint distribution of the "table" indices $\boldsymbol{b} = (b_1, \ldots, b_n)$ and "dish" indices $\boldsymbol{z} = (z_1, \ldots, z_n)$ in*

$$\{b_i\}_{i:z_i=k} \sim \mathrm{CRP}(n_k, r_k), \;\; \boldsymbol{z} \sim \mathrm{DirCat}(n, r_1, \ldots, r_K),$$

is the same as that in

$$z_i = s_{b_i}, \;\; s_t \sim \mathrm{Cat}(r_1/r_\cdot, \ldots, r_K/r_\cdot), \;\; \boldsymbol{b} \sim \mathrm{CRP}(n, r_\cdot),$$

with PMF

$$P(\boldsymbol{b}, \boldsymbol{z} \,|\, n, \boldsymbol{r}) = \frac{\Gamma(r_\cdot)}{\Gamma(n + r_\cdot)} \prod_{k=1}^{K} \left\{ r_k^{\ell_k} \prod_{t=1}^{\ell_k} \Gamma(n_{kt}) \right\},$$

where ℓ_k is the number of unique indices in $\{b_i\}_{i:z_i=k}$, $\ell_\cdot = \sum_{k=1}^{K} \ell_k$ is the total number of nonempty tables, $t = 1, \ldots, \ell_\cdot$, and $n_{kt} = \sum_{i=1}^{n} \delta(b_i = t, z_i = k)$ is the number of customers that sit at table t and take dish k.

If we further randomize the sample length as

$$n \sim \mathrm{NB}(r_\cdot, p),$$

then we have the PMF for the joint distribution of $\boldsymbol{b}$, $\boldsymbol{z}$, *and* n *given* $\boldsymbol{r}$ *and* p *in a fully factorized form as*

$$P(\boldsymbol{b},\boldsymbol{z},n\,|\,\boldsymbol{r},p)=\frac{1}{n!}\prod_{k=1}^{K}\left\{r_k^{\ell_k}(1-p)^{r_k}p^{n_k}\prod_{t=1}^{\ell_k}\Gamma(n_{kt})\right\},$$

which, with appropriate combinatorial analysis, can be mapped to the PMF of the joint distribution of $\boldsymbol{n}=(n_1,\ldots,n_k)$, $\boldsymbol{\ell}=(l_1,\ldots,\ell_k)$, *and* n *given* $\boldsymbol{r}$ *and* p *in*

$$n=\sum_{k=1}^{K}n_k,\ \ n_k=\sum_{t=1}^{\ell_k}n_{kt},\ \ n_{kt}\sim\text{Logarithmic}(p),\ \ \ell_k\sim\text{Pois}[-r_k\ln(1-p)].\tag{22}$$

Using (16) and Proposition 3, introducing the auxiliary variables

$$\{b_{ji}\}_{i:\,x_{ji}=v,z_{ji}=k}\sim\text{CRP}(n_{vjk},\phi_{vk}\theta_{kj}),\tag{23}$$

we have the joint likelihood for the DMMM model as

$$P(\boldsymbol{b}_j,\boldsymbol{x}_j,\boldsymbol{z}_j,n_j\,|\,\boldsymbol{\Phi},\boldsymbol{\theta}_j,p_j)=\frac{1}{n_j!}\prod_{v}\prod_{k}\left\{(\phi_{vk}\theta_{kj})^{\ell_{vjk}}p_j^{n_{vjk}}(1-p_j)^{\phi_{vk}\theta_{kj}}\prod_{t=1}^{\ell_{vjk}}\Gamma(n_{vjkt})\right\},\tag{24}$$

where ℓ_{vjk} is the number of unique indices in $\{b_{ji}\}_{i:\,x_{ji}=v,z_{ji}=k}$ and $n_{vjkt}=\sum_{i=1}^{n_j}\delta(x_{ji}=v,z_{ji}=k,b_{ji}=t)$. As in Proposition 3, instead of first assigning the indices to factors using the Dirichlet-multinomial distributions and then assigning the indices with the same factors to tables, we may first assign the indices to tables and then assign the tables to factors. Thus we have the following model

$$\begin{aligned}x_{ji}=w_{jz_{ji}},\ \ z_{ji}=s_{jb_{ji}},\ \ w_{js_{jt}}\sim\text{Cat}(\boldsymbol{\phi}_{s_{jt}}),\ \ s_{jt}\sim\text{Cat}(\boldsymbol{\theta}_j/\theta_{\cdot j}),\\ \boldsymbol{b}_j\sim\text{CRP}(n_j,\theta_{\cdot j}),\ \ n_j\sim\text{NB}(\theta_{\cdot j},p_j),\end{aligned}$$

whose likelihood $P(\boldsymbol{b}_j,\boldsymbol{x}_j,\boldsymbol{z}_j,n_j\,|\,\boldsymbol{\Phi},\boldsymbol{\theta}_j,p_j)$ is the same as the likelihood, as shown in (24), of the DMMM model constituted of (10) and (12) and augmented with (23).

We outline the collapsed Gibbs sampler for the hGNBP-NBFA in Algorithm 1 and provide the derivation and update equations below. This collapsed sampling strategy marginalizes out both the factors $\{\boldsymbol{\phi}_k\}$ and factor scores $\{\boldsymbol{\theta}_j\}_j$, but at the expense of introducing an auxiliary variable b_{ji} for each index x_{ji}.

Marginalizing out $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ from $\prod_j P(\boldsymbol{b}_j,\boldsymbol{x}_j,\boldsymbol{z}_j,n_j\,|\,\boldsymbol{\Phi},\boldsymbol{\theta}_j,p_j)$, we have

$$\begin{aligned}P(\{\boldsymbol{b}_j,\boldsymbol{x}_j,\boldsymbol{z}_j,n_j\}_j\,|\,G,\boldsymbol{p})=e^{r_\star\sum_j\ln(1-\tilde{p}_j)}\left\{\prod_j p_j^{n_j}\frac{\prod_v\prod_k\left(\prod_{t=1}^{\ell_{vjk}}\Gamma(n_{vjkt})\right)}{n_j!}\right\}\\ \times\left\{\prod_{k:\,\ell_{\cdot\cdot k}>0}\frac{\Gamma(V\eta)}{\Gamma(\ell_{\cdot\cdot k}+V\eta)}\prod_{v=1}^{V}\frac{\Gamma(\ell_{v\cdot k}+\eta)}{\Gamma(\eta)}\right\}\end{aligned}$$

$$\times \left\{ \prod_j \prod_{k:\, \ell_{\cdot\cdot k}>0} \frac{\Gamma(r_k+\ell_{\cdot jk})}{\Gamma(r_k)} \frac{c_j^{r_k}}{[c_j-\ln(1-p_j)]^{r_k+\ell_{\cdot jk}}} \right\}, \tag{25}$$

where $\ell_{\cdot\cdot k} := \sum_j \ell_{\cdot jk}$ and $r_\star := \sum_{k:\,\ell_{\cdot\cdot k}=0} r_k$.

Sample z_{ji} ***and*** b_{ji}**.** Using the likelihood in (25), with $(K^+)^{-ji}$ representing the number of active atoms without considering z_{ji}, we have

$$P(z_{ji}=k, b_{ji}=t \,|\, x_{ji}, \boldsymbol{z}^{-ji}, \boldsymbol{b}^{-ji}, G)$$
$$\propto \begin{cases} n^{-ji}_{x_{ji}jkt}, & \text{if } k \le (K^+)^{-ji}, t \le \ell^{-ji}_{x_{ji}jk}; \\ \frac{\ell^{-ji}_{x_{ji}\cdot k}+\eta}{\ell^{-ji}_{\cdot\cdot k}+V\eta} \frac{r_k+\ell^{-ji}_{\cdot jk}}{c_j-\ln(1-p_j)}, & \text{if } k \le (K^+)^{-ji}, t = \ell^{-ji}_{x_{ji}jk}+1; \\ \frac{1}{V}\frac{r_\star}{c_j-\ln(1-p_j)}, & \text{if } k = (K^+)^{-ji}+1, t=1; \end{cases}$$

and if $k = (K^+)^{-ji}+1$ happens, similar to the direct assignment sampler for the hierarchical Dirichlet process (Teh et al., 2006), we draw $\beta \sim \text{Beta}(1,\gamma_0)$ and then let $r_k = \beta r_\star$ and $r_\star = (1-\beta) r_\star$. This is based on the stick-breaking representation of the Dirichlet process, $\widetilde{G}_\star \sim \text{DP}(\gamma_0, G_0/\gamma_0)$, whose product with an independent random variable $r_\star \sim (\gamma_0, [c_0 - \sum_j \ln(1-\tilde{p}_j)]^{-1})$ recovers the gamma process $G_\star \sim \Gamma\text{P}(G_0, [c_0 - \sum_j \ln(1-\tilde{p}_j)]^{-1})$. Note that instead of drawing both z_{ji} and b_{ji} at the same time, one may first draw z_{ji} and then draw b_{ji} given z_{ji}, or first draw b_{ji} and then draw z_{ji} given b_{ji}. We describe how to sample the other parameters in Appendix C.2.

5.3 Blocked Gibbs sampling under compound Poisson representation

Examining both the blocked and collapsed Gibbs samplers presented Sections 5.1 and 5.2, respectively, and their sampling steps shown in Algorithm 1, one may notice that to obtain n_{vjk} in each iteration, one has to go through all individual indices x_{ji}, for each of which the sampling of z_{ji} from a multinomial distribution takes $O(K)$ computation. However, as it is the ℓ_{vjk} but not n_{vjk} that are required for sampling all the other model parameters, one may naturally wonder whether the step of sampling z_{ji} to obtain n_{vjk} can be skipped. To answer that question, we first introduce the following theorem, whose proof is provided in Appendix A.

Theorem 4. *Conditioning on n and $\boldsymbol{r}$, with $\{n_k\}_k$ marginalized out, the distribution of $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ in*

$$\ell_k \,|\, n_k \sim \text{CRT}(n_k, r_k), \;\; \boldsymbol{n} \,|\, n \sim \text{DirMult}(n, r_1, \ldots, r_K)$$

is the same as that in

$$\boldsymbol{\ell} \,|\, \ell_\cdot \sim \text{Mult}(\ell_\cdot, r_1/r_\cdot, \ldots, r_K/r_\cdot), \;\; \ell_\cdot \,|\, n \sim \text{CRT}(n, r_\cdot),$$

with PMF

$$P(\boldsymbol{\ell} \,|\, n, r_1, \ldots, r_K) = \frac{\ell_\cdot!}{\prod_{k=1}^K \ell_k!} \frac{\Gamma(r_\cdot)}{\Gamma(n + r_\cdot)} |s(n, \ell_\cdot)| \prod_{k=1}^K r_k^{\ell_k}.$$

Rather than representing NBFA in (19) as the DMMM model in (18), we may directly consider its compound Poisson representation as

$$n_{vj} = \sum_{t=1}^{\ell_{vj}} n_{vjt},\ n_{vjt} \sim \text{Logarithmic}(p_j),\ \ell_{vj} \sim \text{Pois}\left[-\sum_k \phi_{vk}\theta_{kj} \ln(1 - p_j)\right]. \quad (26)$$

Under this representation, we may first infer ℓ_{vj} for each n_{vj} and then factorize the latent count matrix $\{\ell_{vj}\}_{v,j}$ under the Poisson likelihood, as described below.

Rather than first sampling z_{ji} (and hence n_{vjk}) using (20) and then sampling ℓ_{vjk} using (21), with Theorem 4 and the compound Poisson representation in (26), we can skip sampling z_{ji} and directly sample ℓ_{vjk} as

$$(\ell_{vj} \,|\, -) \sim \text{CRT}\left(n_{vj}, \sum_k \phi_{vk}\theta_{kj}\right), \quad (27)$$

$$[(\ell_{vj1}, \ldots, \ell_{vjK}) \,|\, -] \sim \text{Mult}\left(\ell_{vj}, \frac{\phi_{v1}\theta_{1j}}{\sum_k \phi_{vk}\theta_{kj}}, \ldots, \frac{\phi_{vK}\theta_{kj}}{\sum_k \phi_{vk}\theta_{kj}}\right). \quad (28)$$

All the other model parameters can be sampled in the same way as they are sampled in the regular blocked Gibbs sampler, with (C.1)–(C.7).

Note that $\ell_{vj} = n_{vj}$ a.s. if $n_{vj} \in \{0, 1\}$, $\ell_{vj} \le n_{vj}$ a.s. if $n_j \ge 2$, and

$$\mathbb{E}[\ell_{vj} \,|\, n_{vj}, \phi_{vk}\theta_{kj}] = \left(\textstyle\sum_k \phi_{vk}\theta_{kj}\right) \left[\psi(n_{vj} + \phi_{vk}\theta_{kj}) - \psi(\phi_{vk}\theta_{kj})\right],$$

where $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$ is the digamma function. Thus $\mathbb{E}[\ell_{vj}] \approx \left(\sum_k \phi_{vk}\theta_{kj}\right) \ln(n_{vj})$ when n_{vj} is large. Clearly, this new sampling strategy significantly impacts n_{vj} that are large. In comparison to the regular blocked Gibbs sampler described in detail in Section 5.1, the compound Poisson based blocked Gibbs sampler essentially replaces (20)–(21) of the regular one with (27)–(28), which can be readily justified by Theorem 4. Instead of directly assigning the n_{vj} covariate indices to the K latent factors, now we only need to assign them to $O[\ln(n_{vj} + 1)]$ tables and directly assign these tables to latent factors. As assigning an index to a table can be accomplished by just a single Bernoulli random draw, this new sampling procedure reduces the computational complexity for sampling all ℓ_{vjk} from $O(n_{\cdot\cdot}K)$ to $O[\sum_v \sum_j \ln(n_{vj} + 1)K]$, which could lead to a considerable saving in computation for long samples where large counts n_{vj} are abundant.

This new sampling algorithm not only is less expensive in computation, but also may converge much faster as there is no need to worry about the dependencies between the MCMC samples for the factor indices z_{ji}, which are not used at all under the compound Poisson representation. Note that the collapsed Gibbs sampler in Section 5.2 removes the need to sample $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ at the expense of having to sample z_{ji} and augment a b_{ij} for each z_{ij}. We will show in Appendix F that maintaining $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ but removing the need to sample z_{ji} leads to a more effective sampler.

5.4 Model comparison

We also consider NBFA based on the GNBP, in which each of the J samples is assigned with a sample-specific NB process and a globally shared gamma process is mixed with all the J NB processes. Given its connection to DCMLDA, as revealed in Section 3.3, we call this nonparametric Bayesian model as the GNBP-DCMLDA, which is shown to be restrictive in that although each sample has its own factor scores under the corresponding sample-specific factors, all the samples are enforced to have the same factor scores under the same set of globally shared factors. By contrast, by modeling not only the burstiness of the covariates, but also that of the factors, the hGNBP-NBFA provides sample-specific factor scores under the same set of shared factors, making it suitable for extracting low-dimensional latent representations for high-dimensional covariate count vectors.

We describe in detail in Appendix D how to use the gamma-negative binomial process (GNBP) as a nonparametric Bayesian prior for both PFA and DCMLDA. In the prior, for PFA, we have $n_{vj} \sim \text{Pois}(\sum_k \phi_{vk}\theta_{kj})$, whereas for NBFA, we have $n_{vj} \sim \text{NB}(\sum_k \phi_{vk}\theta_{kj}, p_j)$, which can be augmented as

$$n_{vj} \sim \text{Pois}(\lambda_{vj}),\ \lambda_{vj} \sim \text{Gamma}\left[\textstyle\sum_k \phi_{vk}\theta_{kj}, p_j/(1-p_j)\right].$$

Thus we have $(\lambda_{vj} \,|\, n_{vj}, \boldsymbol{\Phi}, \boldsymbol{\theta}_j, p_j) \sim \text{Gamma}(n_{vj} + \sum_k \phi_{vk}\theta_{kj},\ p_j)$ for NBFA. Similarly, we have $(\lambda_{vj} \,|-) \sim \text{Gamma}(n_{vj} + \sum_k \phi_{vk} r_k,\ p_j)$ for the GNBP-DCMLDA. To better understand the similarities and differences between the GNBP-PFA, GNBP-DCMLDA, and hGNBP-NBFA, in Table 2 we compare their Poisson rates of n_{vj}, estimated with the factors and factor scores in a single MCMC sample, and several other important model properties.

To estimate the latent Poisson rates for each count n_{vj} and hence the smoothed normalized covariate frequencies, it is clear from the second row of Table 2 that PFA (the multinomial mixed-membership model) solely relies on the inferred factors $\{\boldsymbol{\phi}_k\}$ and factor scores $\{\theta_{kj}\}$, DCMLDA adds a sample-invariant smoothing parameter, calculated as $\sum_v \phi_{vk} r_k$, into the observed count n_{vj} and weights that sum by a sample-specific probability parameter p_j, whereas NBFA (the DMMM model) adds a sample-specific smoothing parameter, calculated as $\sum_v \phi_{vk}\theta_{kj}$, into the observed count n_{vj} and weights that sum by p_j. Thus PFA represents an extreme that the observed counts are used to infer the factors and factor scores but are not used to directly estimate the Poisson rates; DCMLDA represents another extreme that the covariate frequencies in all samples are indiscriminately smoothed by the same set of smoothing parameters; whereas NBFA combines the observed counts with the inferred sample-specific smoothing parameters. This unique working mechanism also makes NBFA have reduced hyper-parameter sensitivity, as will be demonstrated with experiments.

6 Example results

We apply the proposed models to factorize covariate-sample count matrices, each column of which is represented as a V dimensional covariate-frequency count vector, where V is the number of unique covariates. We set the hyper-parameters as $a_0 = b_0 = 0.01$ and

	GNBP-PFA (multinomial mixed-membership model)	GNBP-DCMLDA	hGNBP-NBFA (Dirichlet-multinomial mixed-membership model)
Estimated Poisson rate of n_{vj} given the factors and factor scores	$\sum_k \phi_{vk}\theta_{kj}$	$(n_{vj} + \sum_k \phi_{vk} r_k)p_j$	$(n_{vj} + \sum_k \phi_{vk}\theta_{kj})p_j$
Factor analysis	$\boldsymbol{n}_j \sim \text{Pois}(\boldsymbol{\Phi}\boldsymbol{\theta}_j)$	$\boldsymbol{n}_j \sim \text{NB}(\boldsymbol{\Phi}\boldsymbol{r}, p_j)$	$\boldsymbol{n}_j \sim \text{NB}(\boldsymbol{\Phi}\boldsymbol{\theta}_j, p_j)$
Mixed-membership modeling	$z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j/\theta_{\cdot j})$, $x_{ji} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}})$	$\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{r})$, $z_{ji} \sim \text{Cat}(\boldsymbol{\theta}_j)$, $\boldsymbol{\phi}_k^{[j]} \sim \text{Dir}(\boldsymbol{\phi}_k r_k)$, $x_{ji} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}^{[j]})$	$\boldsymbol{\theta}^{[j]} \sim \text{Dir}(\boldsymbol{\theta}_j)$, $z_{ji} \sim \text{Cat}(\boldsymbol{\theta}^{[j]})$, $\boldsymbol{\phi}_k^{[j]} \sim \text{Dir}(\boldsymbol{\phi}_k \theta_{kj})$, $x_{ji} \sim \text{Cat}(\boldsymbol{\phi}_{z_{ji}}^{[j]})$
Distribution of $X_j = \sum_{k=1}^{\infty} n_{\cdot jk}\delta_{\boldsymbol{\phi}_k}$ given G	$X_j \,\vert\, G, p_j \sim \text{NBP}(G, p_j)$	$X_j \,\vert\, G, p_j \sim \text{NBP}(G, p_j)$	$X_j \,\vert\, G, c_j, p_j \sim \text{GNBP}(G, c_j, p_j)$
Variance-to-mean ratio of $n_{\cdot jk}$ given c_j and p_j	$\frac{1}{1-p_j}$	$\frac{1}{1-p_j}$	$\frac{1}{1-p_j} + \frac{p_j}{c_j(1-p_j)^2}$

Table 2: Comparisons of the GNBP-PFA, GNBP-DCMLDA, and hGNBP-NBFA.

$e_0 = f_0 = 1$. We consider the Journal of the ACM (JACM, http://www.cs.princeton.edu/~blei/downloads/), Psychological Review (PsyReview, http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm), and NIPS Conference Papers Vols0-12 (NIPS12, http://www.cs.nyu.edu/~roweis/data.html) datasets, choosing covariates that occur in five or more samples. In addition, we consider the 20newsgroups dataset (http://qwone.com/~jason/20Newsgroups/), consisting of 18,774 samples from 20 different categories. It is partitioned into a training set of 11,269 samples and a testing set of 7,505 ones that were collected at later times. We remove a standard list of stopwords and covariates that appear less than five times. As summarized in Table 3 of Appendix G, for the PysReview and JACM datasets, each of whose sample corresponds to the abstract of a research paper, the average sample lengths are only about 56 and 127, respectively. By contrast, a NIPS12 sample that includes the words of all sections of a research paper is in average more than ten times longer. By varying the percentage of covariate indices randomly selected from each sample for training, we construct a set of covariate-sample matrices with a large variation on the average lengths of samples, which will be used to help make comparison between different models. Depending on applications, we either treat the Dirichlet smoothing parameter η as a tuning parameter, or sample it via data augmentation, as described in Appendix E.

To learn the factors in all the following experiments, we use the compound Poisson representation based blocked Gibbs sampler for both the hGNBP-NBFA and GNBP-NBFA, and use collapsed Gibbs sampling for the GNBP-PFA. We compare different samplers for the hGNBP-NBFA and provide justifications for choosing these samplers in Appendix F.

6.1 Prediction of heldout covariate indices

Experimental settings

We randomly choose a certain percentage of the covariate indices in each sample as training, and use the remaining ones to calculate heldout perplexity. As shown in Zhou and Carin (2015), the GNBP-PFA performs similarly to the hierarchical Dirichlet process LDA of Teh et al. (2006) and outperforms a wide array of discrete latent variable models, thus we choose it for comparison. To demonstrate the importance of modeling both the burstiness of the covariates and that of the factors, we also make comparison to the GNBP-DCMLDA that considers only covariate burstiness. Since the inferred number of factors and hence the performance often depends on the Dirichlet smoothing parameter η, we set η as 0.005, 0.02, 0.01, 0.05, 0.10, 0.25, or 0.50. We vary both the training percentage and η to examine how the average sample length and the value of η influence the behaviors of the GNBP-PFA, GNBP-DCMLDA, and hGNBP-NBFA and impact their performance relative to each other.

For all three algorithms, we initialize the number of factors as $K = 400$ and consider 5000 Gibbs sampling iterations, with the first 2500 samples discarded and every sample per five iterations collected afterwards. For each collected sample, for the GNBP-PFA, we draw the factors $(\boldsymbol{\phi}_k \,|\, -) \sim \mathrm{Dir}(\eta + n_{1\cdot k}, \ldots, \eta + n_{V\cdot k})$ and factor scores $(\theta_{kj} \,|\, -) \sim \mathrm{Gamma}(n_{jk} + r_k, p_j)$ for $k \in \{1, \ldots, K^+ + K_\star\}$, where we let $n_{v\cdot k} = 0$ for all $k > K^+$; for the GNBP-DCMLDA, we draw the factors $(\boldsymbol{\phi}_k \,|\, -) \sim \mathrm{Dir}(\eta + \ell_{1\cdot k}, \ldots, \eta + \ell_{V\cdot k})$ and the weights $(r_k \,|\, -) \sim \mathrm{Gamma}(\ell_{\cdot\cdot k}, 1/[c_0 - \sum_j \ln(1 - p_j)])$, where we let $\ell_{v\cdot k} = 0$ and $\ell_{\cdot\cdot k} = \gamma_0/K_*$ for all $k > K^+$; and for the hGNBP-NBFA, we draw the factors $(\boldsymbol{\phi}_k \,|\, -) \sim \mathrm{Dir}(\eta + \ell_{1\cdot k}, \ldots, \eta + \ell_{V\cdot k})$ and factor scores $(\theta_{kj} \,|\, -) \sim \mathrm{Gamma}[r_k + \ell_{\cdot jk}, 1/(c_j - \ln(1 - p_j))]$, where we let $\ell_{v\cdot k} = 0$ and $r_k = \gamma_0/K_\star$ for all $k > K^+$. We set $K_* = 20$ for all three algorithms.

We compute the heldout perplexity as

$$\exp\left(-\frac{1}{m_{\cdot\cdot}^{\mathrm{test}}} \sum_v \sum_j m_{vj}^{\mathrm{test}} \ln \frac{\sum_s \lambda_{vj}^{(s)}}{\sum_s \sum_{v'} \lambda_{v'j}^{(s)}}\right),$$

where $s \in \{1, \ldots, S\}$ is the index of a collected MCMC sample, m_{vj}^{test} is the number of test covariate indices at covariate v in sample j, $m_{\cdot\cdot}^{\mathrm{test}} = \sum_v \sum_j m_{vj}^{\mathrm{test}}$, and $\lambda_{vj}^{(s)}$ are computed using the equations shown in the second row of Table 2, *e.g.*, we have $\lambda_{vj}^{(s)} = (n_{vj} + \sum_{k=1}^{K^+ + K_\star} \phi_{vk}^{(s)} \theta_{kj}^{(s)}) p_j^{(s)}$ for the hGNBP-NBFA. For each unique combination of η and the training percentage, the results are averaged over five random training/testing

partitions. The evaluation method is similar to those used in Wallach et al. (2009b), Paisley et al. (2012), and Zhou and Carin (2015). All algorithms are coded in MATLAB, with the steps of sampling factor and table indices coded in C to optimize speed. We terminate a trial and omit the results for that particular setting if it takes a single core of an Intel Xeon 3.3 GHz CPU more than 24 hours to finish 5000 iterations. The code will be made available in the author's website for reproducible research.

We first consider the NIPS12 dataset, whose average sample length is about 1323, and present its results in Figures 2–5. We also consider both the PsyReview and JACM datasets, whose average sample lengths are about 56 and 127, respectively, and provide related plots in Appendix G.

General observations

For multinomial mixed-membership models, generally speaking, the smaller the Dirichlet smoothing parameter η is, the more sparse and specific the inferred factors are encouraged to be, and the larger the number of inferred factors using a nonparametric Bayesian mixed-membership modeling prior, such as the hierarchical Dirichlet process and the gamma- and beta-negative binomial processes (Paisley et al., 2012; Zhou et al., 2012). As shown in Figures 2(a)–(e), for the hGNBP-NBFA, a nonparametric Bayesian DMMM model, we observe a relationship between the number of inferred factors and η similar to that for the GNBP-PFA, a nonparametric Bayesian multinomial mixed-membership model.

In comparison to multinomial mixed-membership models such as the GNBP-PFA, what make the hGNBP-NBFA different and desirable are: 1) its parsimonious representation that uses fewer factors to achieve better heldout prediction, as shown in Figures 3(a)–(e); 2) ts distinct mechanism in adjusting its number of inferred factors according to the lengths of samples, as shown in Figures 4(a)–(f); 3) its significantly lower computational complexity for a covariate-sample matrix with long samples (large column sums), with the differences becoming increasingly more significant as the average sample length increases, as shown in Figures 3(f)–(j) and 5(a)–(f); 4) its ability to achieve the same predictive power with significantly less time, as shown in Figures 5(g)–(l); 5) and its overall better predictive performance both under various values of η while controlling the sample lengths, as shown in Figures 2(f)–(j), and under various sample lengths while controlling η, as shown in Figures 4(g)–(l).

Detailed discussions

Distinct behavior and parsimonious representation. When fixing η but gradually increasing the average sample length, the number of factors inferred by a nonparametric Bayesian multinomial mixed-membership model such as the GNBP-PFA often increases at a near-constant rate, as shown with the blue curves in Figure 4(a)–(f). The GNBP-DCMLDA, which models covariate burtiness, behaves similarly in the number of inferred factors, as shown with the red curves in Figure 4(a)–(f). Under the same setting, the number of inferred factors by the hGNBP-NBFA often first increases at a similar near-constant rate when the average sample length is short, however, it starts decreasing

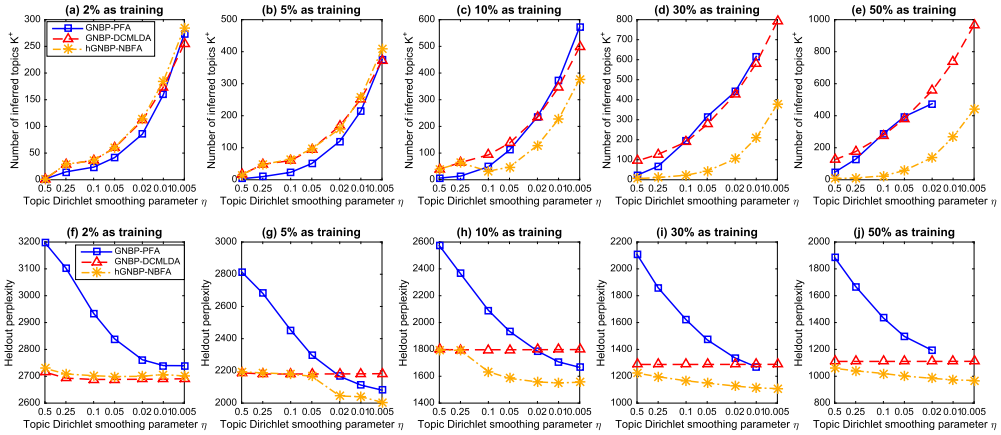

Figure 2: Comparisons of the GNBP-PFA (multinomial mixed-membership model), GNBP-DCMLDA, and hGNBP-NBFA (Dirichlet-multinomial mixed-membership model) on (a)–(e) the posterior means of the number of inferred factors K^+ and (f)–(j) heldout perplexity, both as a function of the Dirichlet smoothing parameter η for the NIPS12 dataset. The values of η are plot in the logarithmic scale from large to small. In both rows, the plots from left to right are obtained using 2%, 5%, 10%, 30%, and 50% of the covariate indices for training, respectively. All plots are based on five independent random trials. The error bars are not shown as variations across different trials are small. Some results for the GNBP-PFA are missing as they took more than 24 hours to run 5000 Gibbs sampling iterations on a 3.3 GHz CPU and hence were terminated before completion.

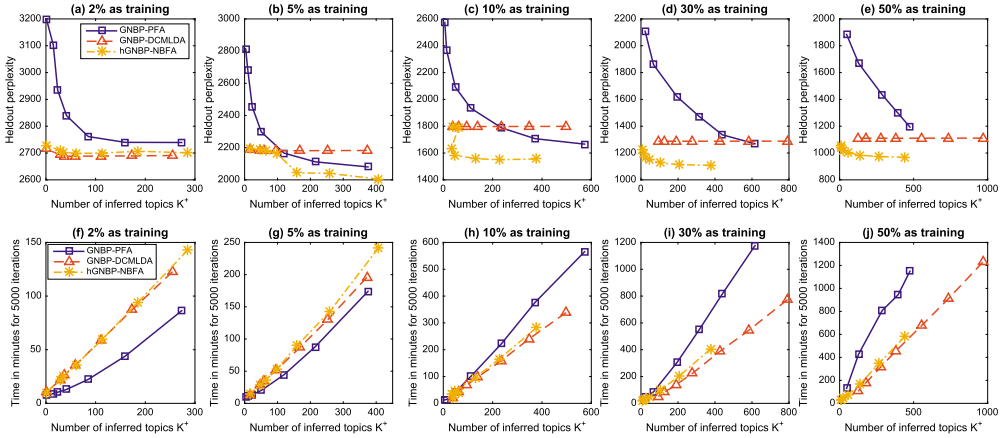

Figure 3: Using the same results shown in Figure 2, we plot (a)–(e) the obtained heldout perplexity and (f)–(j) the number of minutes to finish 5000 Gibbs sampling iterations, both as a function of the number of inferred factors K^+.

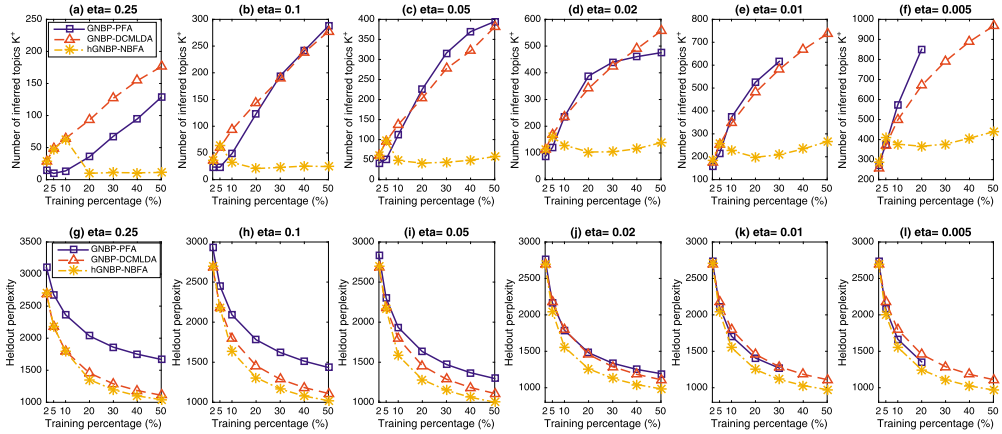

Figure 4: Comparisons of the GNBP-PFA (multinomial mixed-membership model), GNBP-DCMLDA, and hGNBP-NBFA (Dirichlet-multinomial mixed-membership model) on (a)–(f) the posterior means of the number of inferred factors K^+ and (g)–(l) heldout perplexity, both as a function of the percentage of covariate indices used for training for the NIPS12 dataset. In both rows, the plots from left to right are obtained with $\eta = 0.25, 0.1, 0.05, 0.02, 0.01$, and 0.005, respectively. Other specifications are the same as those of Figure 2.

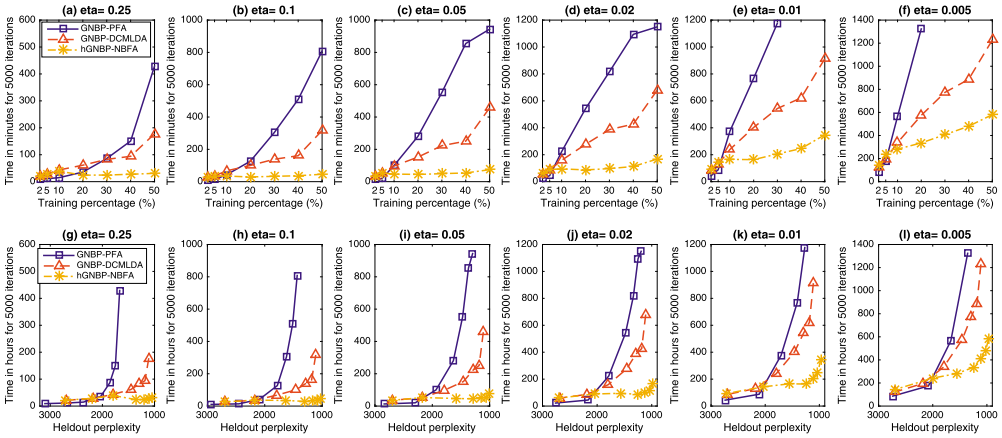

Figure 5: Using the results shown in Figure 4, we plot the number of minutes to finish 5000 Gibbs sampling iterations both (a)–(f) as a function of the percentage of covariates used for training and (g)–(l) as a function of heldout perplexity.

once the average sample length becomes sufficiently long, and eventually turns around and increases, but at a much lower rate, as the average sample length further increases, as shown with the yellow curves in Figure 4(a)–(f). This distinct behavior implies that

by exploiting its ability to model both covariate and factor burstiness, the hGNBP-NBFA could have a parsimonious representation of a dataset with long samples. By contrast, a nonparametric Bayesian multinomial mixed-membership model, such as the GNBP-PFA, models neither covariate nor factor burstiness. Consequently, it has to increase its latent dimension at a near-constant rate as a function of the average sample length, in order to adequately capture self- and cross-excitation of covariate frequencies, which are often more prominent in longer samples. It is clear that by decreasing η and hence increasing the number of inferred factors, the GNBP-PFA can gradually approach and eventually outperform the GNBP-DCMLDA, but still clearly underperform the hGNBP-NBFA in most cases, even if using many more factors and consequently significantly more computation.

Combining factorization and the modeling of burstinss. As shown in Figure 2(f), when the training percentage is as small as 2%, the GNBP-DCMLDA, which combines the observed counts n_{vj} and the inferred sample-invariant smoothing parameters $\sum_k \phi_{vk} r_k$ to estimate the Poisson rates (and hence the smoothed normalized covariate frequencies), achieves the best predictive performance (lowest heldout perplexity); the hGNBP-NBFA tries to improve DCMLDA by combining the observed counts and document-specific smoothing parameters $\sum_k \phi_{vk}\theta_{kj}$, and the GNBP-PFA only relies on $\sum_k \phi_{vk}\theta_{kj}$, yielding slightly and significantly worse performance, respectively, at this relatively extreme setting. This suggests that when the observed counts are too small, using factorization may not provide any advantages than simply smoothing the raw covariate counts with sample-invariant smoothing parameters.

As the training percentage increases, all three algorithms quickly improve their performance, as shown in Figures 2(g)–(j). Given a training percentage that is sufficiently large, e.g., 10% for this dataset (*i.e.*, the average training sample length is about 132), all three algorithms tend to increase their numbers of inferred factors K^+ as η decreases, although the hGNBP-NBFA usually has a lower increasing rate. They differ from each other significantly, however, on how the performance improves as the inferred number factors increases, as shown in Figures 3(a)–(e): for the GNBP-DCMLDA, as it relies on $\sum_k \phi_{vk} r_k$ to smooth the observed counts, its predictive power is almost invariant to the change of η and its number of factors; for the GNBP-PFA, by decreasing η and hence increasing its number of inferred factors, it can approach and eventually outperform DCMLDA; whereas for the hGNBP-NBFA, it follows DCMLDA closely when η is large or the lengths of samples are short, but often reduces its rate of increase for the number of inferred factors as η decreases and quickly lowers its perplexity as K^+ increases, as long as η is sufficiently small or the samples are sufficiently long. Thus in general, the hGNBP-NBFA provides the lowest perplexity using the least number of inferred factors.

Note that when the lengths of the training samples are short, setting η to be large will make the factors $\boldsymbol{\phi}_k$ of NBFA become over-smoothed and hence NBFA becomes essentially the same as DCMLDA. As η decreases given the same average sample length, or as the average sample length increases given the same η, the factorization of NBFA with sample-dependent factor scores gradually take effect to improve the estimation of

the Poisson rates and hence the smoothed normalized covariate frequencies for each sample. Overall, by combining the factorization, as used in PFA, the modeling of covariate burstiness, as used in DCMLDA, and the modeling of factor burstiness, unique to NBFA, the hGNBP-NBFA captures both self- and cross-excitation of covariate frequencies and achieves the best predictive performance with the most parsimonious representation as long as the average sample length is not too short and the value of η is not set too large to overly smooth the factors.

Significantly lower computation for sufficiently long samples. For the GNBP-PFA, the collapsed Gibbs sampler samples all the factor indices with a computational complexity of $O(n_{\cdot\cdot}K^+)$, whereas for the hGNBP-NBFA, the corresponding computation has a complexity of $O[\sum_v \sum_j \ln(n_{vj}+1)K^+]$ and sampling $\{\phi_k\}_k$ and $\{\boldsymbol{\theta}_j\}_j$ adds an additional computation of $O(VK^+ + NK^+)$. Thus the computation for the hGNBP-NBFA not only is often lower given the same K^+ for a dataset consisting of sufficiently long samples, but also becomes much lower because the inferred K^+ is often much smaller when the sample lengths are sufficiently long. For example, as shown in Figure 3(i), when 30% of the covariate indices in each sample are used for training, which means the average training sample length is about 397, the time for the GNBP-PFA to finish 5000 Gibbs sampling iteration on a 3.3 GHz CPU is about double that for the hGNBP-NBFA when their inferred numbers of factors are similar to each other; and when 20% of the covariate indices in each sample are used for training (*i.e.*, the average training sample length is around 265), in comparison to the hGNBP-NBFA, the GNBP-PFA takes about three times more minutes when $\eta = 0.1$, as shown in Figures 5(b), and four times more minutes when $\eta = 0.01$, as shown in Figures 5(e). Overall, for a dataset whose samples are not too short to exhibit self- and cross-excitation of covariate frequencies, the hGNBP-NBFA often takes the least time to finish the computation while controlling the value of η and average sample length, has lower computation given the same inferred number of factors K^+, and achieves a low perplexity with significantly less computation.

6.2 Unsupervised feature learning for classification

To further verify the advantages of NBFA that models both self- and cross-excitation of covariate frequencies, we use the proposed models to extract low-dimensional feature vectors from high-dimensional covariate-frequency count vectors of the 20newsgroups dataset, and then examine how well the unsupervisedly extracted feature vector of a test sample can be used to correctly classify it to one of the 20 news groups. As the classification accuracy often strongly depends on the dimension of the feature vectors, we truncate the total number of factors at $K = 25$, 50, 100, 200, 400, 600, 800, or 1000. Correspondingly, we slightly modify the gamma process based nonparametric Bayesian models by choosing a discrete base measure for the gamma process as $G_0 = \sum_{k=1}^{K} \frac{\gamma_0}{K}\delta_{\boldsymbol{\phi}_k}$, $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\eta, \ldots, \eta)$. Thus in the prior we now have $r_k = G(\boldsymbol{\phi}_k) \sim \mathrm{Gamma}(\gamma_0/K, 1/c_0)$ and consequently the Gibbs sampling update equations for $\{r_k\}_k$ and γ_0 will also slightly change. We omit these details for brevity, and refer to Zhou and Carin (2015) on how the same type of finite truncation is used in inference for nonparametric Bayesian models.

For this application, we fix the truncation level K but impose the non-informative Gamma$(0.01, 1/0.01)$ prior on the Dirichlet smoothing parameter η, letting it be inferred from the data using (E.4). The same as before, we consider collapsed Gibbs sampling for the GNBP-PFA and the compound Poisson representation based blocked Gibbs sampler for the hGNBP-NBFA, with the main difference in that a fixed instead of an adaptive truncation is now used for inference. We do not consider the GNBP-DCMLDA here since it does not provide sample specific feature vectors under the same set of shared factors. Note that although we fix K, if K is set to be large enough, not necessarily all factors would be used and hence a truncated model still preserves its ability to infer the number of active factors $K^+ \leq K$; whereas if K is set to be small, a truncated model may lose its ability to infer K^+, but it still maintains asymmetric priors (Wallach et al., 2009a) on the factor scores.

For both the hGNBP-NBFA and GNBP-PFA, we consider 2000 Gibbs sampling iterations on the 11,269 training samples of the 20newsgroups dataset, and retain the weights $\{r_k\}_{1,K}$ and the posterior means of $\{\boldsymbol{\phi}_k\}_{1,K}$ as factors, according to the last MCMC sample, for testing. With these K inferred factors and weights, we further apply 1000 blocked Gibbs sampling iterations for both models and collect the last 500 MCMC samples to estimate the posterior mean of the feature usage proportion vector $\boldsymbol{\theta}_j / \theta_{\cdot j}$, for every sample in both the training and testing sets. Denote $\bar{\boldsymbol{\theta}}_j \in \mathbb{R}^K$ as the estimated feature vector for sample j. We use the L_2 regularized logistic regression provided by the LIBLINEAR (https://www.csie.ntu.edu.tw/∼cjlin/liblinear/) package (Fan et al., 2008) to train a linear classifier on all $\bar{\boldsymbol{\theta}}_j$ in the training set and use it to classify each $\bar{\boldsymbol{\theta}}_j$ in the test set to one of the 20 news groups; the regularization parameter C of the classifier five-fold cross validated on the training set from $(2^{-10}, 2^{-9}, \ldots, 2^{15})$.

We first consider distinguishing between the *alt.atheism* and *talk.religion.misc* news groups, and between the *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware* news groups. For each binary classification task, we remove a standard list of stop words and only consider the covariates that appear at least five times in both newsgroups combined, and report the classification accuracies based on twelve independent runs with random initializations. For the first binary classification task, we have 856 training documents, with 6509 unique terms and about 116K words, while for the second one, we have 1162 training documents, with 4737 unique terms and about 91K words.

As shown in Figures 6(a) and 6(b), NBFA clearly outperforms PFA for both binary classification tasks in that it in general provides higher classification accuracies on testing samples while controlling the truncation level K (*i.e.*, the dimension of the extracted feature vectors). It is also interesting to examine how the inferred Dirichlet smoothing parameter η changes as the truncation level K increases, as shown in Figures 6(d) and 6(e). It appears that the inferred η's and active factors K^+'s could be fitted with a decreasing straight line in the logarithmic scale, except for the tails that seem slightly concave up, for both NBFA and PFA. When the truncation level K is not sufficiently large, the inferred η of NBFA is usually smaller than that of PFA given the same K. This may be explained by examining (E.4), where $\ell_{vjk} \leq n_{vjk}$ a.s. and the differences could be significant for large n_{vjk}. Note when using the raw word counts as the features, the classification accuracies of logistic regression are 79.4% and 86.7% for the first and

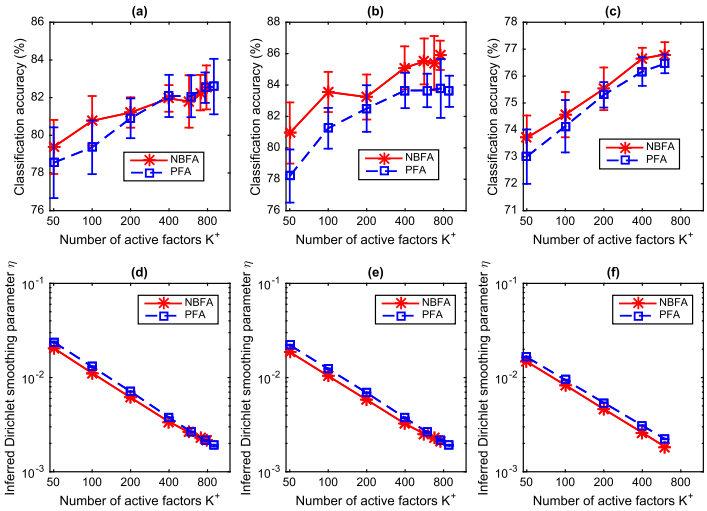

Figure 6: Comparison between negative binomial factor analysis (NBFA) and Poisson factor analysis (PFA) on three different classification tasks, with the number of factors K fixed and the Dirichlet smoothing parameter η inferred from the data. For the *alt.atheism* versus *talk.religion.misc* binary classification task, based on twelve independent random trials, we plot (a) the classification accuracy and (d) the inferred η, both as a function of the number active topics $K^+ \leq K$, where $K \in \{50, 100, 200, 400, 600, 800, 1000\}$. (d) and (e): Analogous plots to (a) and (b) for the *comp.sys.ibm.pc.hardware* versus *comp.sys.mac.hardware* binary classification task, with $K \in \{50, 100, 200, 400, 600, 800, 1000\}$. (c) and (f): Analogous plots to (a) and (b) for the 20newsgroups multi-class classification task, with $K \in \{50, 100, 200, 400, 600\}$.

second binary classification tasks, respectively, and when using the normalized term frequencies as the features, these are 80.8% and 88.0%, respectively.

In addition to these two binary classification tasks, we consider multi-class classification on the 20newsgroups dataset. After removing stopwords and terms (covariates) that appear less than five times, we obtain 33,420 unique terms and about 1.4 million words for training, as summarized in Table 3. We use all 11,269 training documents to infer the factors and factor scores, and mimic the same testing procedure used for binary classification to extract low-dimensional feature vectors, with which each testing sample is classified to one of the 20 news groups using the same L_2 regularized logistic regression. Note the classification accuracies of logistic regression with the raw counts or normalized term frequencies as features are 78.0% and 79.4%, respectively. As shown in Figure 6(f), NBFA generally outperforms PFA in terms of classification accuracies given the same feature dimensions, consistent with our observations for both binary classification tasks. We also observe similar relationship between the K^+ and inferred η as we do in both binary classification tasks.

7 Conclusions

Negative binomial factor analysis (NBFA) is proposed to factorize the covariate-sample count matrix under the NB likelihood. Its equivalent representation as the Dirichlet multinomial mixed-membership model reveals its distinctions from previously proposed discrete latent variable models. The hierarchical gamma-negative binomial process (hGNBP) is further proposed to support NBFA with countably infinite factors, and a compound Poisson representation based blocked Gibbs sampler is shown to converge fast and have low computational complexity. By capturing both self- and cross-excitation of covariate frequencies and by smoothing the observed counts with both sample and covariate specific rates obtained through factorization under the NB likelihood, the hGNBP-NBFA not only infers a parsimonious representation of a covariate-sample count matrix, but also achieves state-of-the-art predictive performance at low computational cost. In addition, the latent feature vectors inferred under the hGNBP-NBFA are better suited for classification than those inferred by the GNBP-PFA. It is of interest to investigate a wide variety of extensions built on Poisson factor analysis under this new modeling framework.

Supplementary Material

Nonparametric Bayesian Negative Binomial Factor Analysis: Supplementary Material (DOI: 10.1214/17-BA1070SUPP; .pdf).

References

Aldous, D. (1983). "Exchangeability and related topics." In *Ecole d'Ete de Probabilities de Saint-Flour XIII*, 1–198. Springer. MR0883646. doi: https://doi.org/10.1007/BFb0099421. 1076

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *Annals of Statistics*, 2(6): 1152–1174. MR0365969. 1076

Blei, D. and Lafferty, J. (2005). "Correlated Topic Models." In *NIPS*, 147–154. 1067

Blei, D., Ng, A., and Jordan, M. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3: 993–1022. 1065, 1068

Broderick, T., Mackey, L., Paisley, J., and Jordan, M. I. (2015). "Combinatorial Clustering and the Beta Negative Binomial Process." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1067

Buntine, W. and Jakulin, A. (2006). "Discrete Component Analysis." In *Subspace, Latent Structure and Feature Selection Techniques*. Springer-Verlag. 1068

Canny, J. (2004). "GaP: a factor model for discrete data." In *SIGIR*. MR1969388. doi: https://doi.org/10.1016/S0012-365X(02)00881-6. 1068

Church, K. W. and Gale, W. A. (1995). "Poisson mixtures." *Natural Language Engineering*. 1066

Doyle, G. and Elkan, C. (2009). "Accounting for burstiness in topic models." In *ICML*. 1066, 1071

Dunson, D. B. and Herring, A. H. (2005). "Bayesian latent variable models for mixed discrete outcomes." *Biostatistics*, 6(1): 11–25. MR2490545. doi: https://doi.org/10.1016/S0169-7161(05)25025-3. 1069

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*. MR1340510. 1075

Ewens, W. J. (1972). *Theoretical Population Biology*, 3(1): 87–112. MR0325177. 1076

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). "LIBLINEAR: A Library for Large Linear Classification." *Journal of Machine Learning Research*, 1871–1874. 1088

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1(2): 209–230. MR0350949. 1073, 1075

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). "A sticky HDP-HMM with application to speaker diarization." *Annals of Applied Statistics*. MR2840185. doi: https://doi.org/10.1214/10-AOAS395. 1075

Gan, Z., Chen, C., Henao, R., Carlson, D., and Carin, L. (2015). "Scalable Deep Poisson Factor Analysis for Topic Modeling." In *ICML*. 1067

Griffiths, T. L. and Steyvers, M. (2004). "Finding Scientific Topics." *PNAS*. 1076

Hofmann, T. (1999). "Probabilistic Latent Semantic Analysis." In *UAI*. 1068

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453). MR1952729. doi: https://doi.org/10.1198/016214501750332758. 1075

Lee, D. D. and Seung, H. S. (2001). "Algorithms for Non-negative Matrix Factorization." In *NIPS*. 1068

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Controlling the reinforcement in Bayesian non-parametric mixture models." *Journal of the Royal Statistical Society: Series B*, 69(4): 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 1075

Madsen, R. E., Kauchak, D., and Elkan, C. (2005). "Modeling word burstiness using the Dirichlet distribution." In *ICML*. 1066

Mosimann, J. E. (1962). "On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions." *Biometrika*, 65–82. MR0143299. 1069

Newman, D., Asuncion, A., Smyth, P., and Welling, M. (2009). "Distributed algorithms for topic models." *Journal of Machine Learning Research*. MR2540777. 1076

Paisley, J., Wang, C., and Blei, D. M. (2012). "The Discrete Infinite Logistic Nor-

mal Distribution." *Bayesian Analysis*. MR3000022. doi: https://doi.org/10.1214/12-BA734. 1067, 1083

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 1075

Pitman, J. (2006). *Combinatorial stochastic processes*. Lecture Notes in Mathematics. Springer-Verlag. MR2245368. 1076

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). "Inference of population structure using multilocus genotype data." *Genetics*, 155(2): 945–959. 1065

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. (2015). "Deep exponential families." In *AISTATS*. 1067

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments." *Annals of Statistics*, 31(2): 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 1075

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101: 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 1067, 1076, 1078, 1082

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics Simulation and Computation*. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 1075

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009a). "Rethinking LDA: Why priors matter." In *NIPS*. 1088

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009b). "Evaluation Methods for Topic Models." In *ICML*. 1083

Zhou, M. (2017). "Nonparametric Bayesian Negative Binomial Factor Analysis: Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1070SUPP. 1067

Zhou, M. and Carin, L. (2015). "Negative Binomial Process Count and Mixture Modeling." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 307–320. 1065, 1066, 1067, 1068, 1075, 1082, 1083, 1087

Zhou, M., Cong, Y., and Chen, B. (2016a). "Augmentable Gamma Belief Networks." *JMLR*, 17(163): 1–44. MR3555054. 1067, 1072, 1073

Zhou, M., Hannah, L., Dunson, D., and Carin, L. (2012). "Beta-Negative Binomial Process and Poisson Factor Analysis." In *AISTATS*, 1462–1471. 1065, 1066, 1067, 1068, 1069, 1083

Zhou, M., Padilla, O. H. M., and Scott, J. G. (2016b). "Priors for Random Count Matrices Derived from a Family of Negative Binomial Processes." *Journal of the American*

Statistical Association, 111(515): 1144–1156. MR3561938. doi: https://doi.org/10.1080/01621459.2015.1075407. 1067

Acknowledgments

The author would like to thank the editor-in-chief, editor, associate editor, and two anonymous referees for their invaluable comments and suggestions, which have helped improve the paper substantially.