Beta-Negative Binomial Process and Poisson Factor Analysis

Mingyuan Zhou   **Lauren A. Hannah**[†]   **David B. Dunson**[†]   **Lawrence Carin**
Department of ECE, [†]Department of Statistical Science, Duke University, Durham NC 27708, USA

Abstract

A beta-negative binomial (BNB) process is proposed, leading to a beta-gamma-Poisson process, which may be viewed as a "multi-scoop" generalization of the beta-Bernoulli process. The BNB process is augmented into a beta-gamma-gamma-Poisson hierarchical structure, and applied as a nonparametric Bayesian prior for an infinite Poisson factor analysis model. A finite approximation for the beta process Lévy random measure is constructed for convenient implementation. Efficient MCMC computations are performed with data augmentation and marginalization techniques. Encouraging results are shown on document count matrix factorization.

1 Introduction

Count data appear in many settings. Problems include predicting future demand for medical care based on past use (Cameron et al., 1988; Deb and Trivedi, 1997), species sampling (National Audubon Society) and topic modeling of document corpora (Blei et al., 2003). Poisson and negative binomial distributions are typical choices for univariate and repeated measures count data; however, multivariate extensions incorporating latent variables (latent counts) are under developed. Latent variable models under the Gaussian assumption, such as principal component analysis and factor analysis, are widely used to discover low-dimensional data structure (Lawrence, 2005; Tipping and Bishop, 1999; West, 2003; Zhou et al., 2009). There has been some work on exponential family latent factor models that incorporate Gaussian latent variables (Dunson, 2000, 2003; Moustaki and Knott, 2000; Sammel et al., 1997), but computation tends

to be prohibitive in high-dimensional settings and the Gaussian assumption is restrictive for count data that are discrete and nonnegative, have limited ranges, and often present overdispersion. In this paper we propose a flexible new nonparametric Bayesian prior to address these problems, the beta-negative binomial (BNB) process.

Using completely random measures (Kingman, 1967), Thibaux and Jordan (2007) generalize the beta process defined on $[0, 1] \times \mathbb{R}^+$ by Hjort (1990) to a general product space $[0, 1] \times \Omega$, and define a Bernoulli process on an atomic beta process hazard measure to model binary outcomes. They further show that the beta-Bernoulli process is the underlying de Finetti mixing distribution for the Indian buffet process (IBP) of Griffiths and Ghahramani (2005). To model count variables, we extend the measure space of the beta process to $[0, 1] \times \mathbb{R}^+ \times \Omega$ and introduce a negative binomial process, leading to the BNB process. We show that the BNB process can be augmented into a beta-gamma-Poisson process, and that this process may be interpreted in terms of a "multi-scoop" IBP. Specifically, each "customer" visits an infinite set of dishes on a buffet line, and rather than simply choosing to select certain dishes off the buffet (as in the IBP), the customer may select multiple scoops of each dish, with the number of scoops controlled by a negative binomial distribution with dish-dependent hyperparameters. As discussed below, the use of a negative binomial distribution for modeling the number of scoops is more general than using a Poisson distribution, as one may control both the mean and variance of the counts, and allow overdispersion. This representation is particularly useful for discrete latent variable models, where each latent feature is not simply present or absent, but contributes a distinct count to each observation.

We use the BNB process to construct an infinite discrete latent variable model called Poisson factor analysis (PFA), where an observed count is linked to its latent parameters with a Poisson distribution. To enhance model flexibility, we place a gamma prior on the Poisson rate parameter, leading to a negative bi-

nomial distribution. The BNB process is formulated in a beta-gamma-gamma-Poisson hierarchical structure, with which we construct an infinite PFA model for count matrix factorization. We test PFA with various priors for document count matrix factorization, making connections to previous models; here a latent count assigned to a factor (topic) is the number of times that factor appears in the document.

The contributions of this paper are: 1) an extension of the beta process to a marked space, to produce the beta-negative binomial (BNB) process; 2) efficient inference for the BNB process; and 3) a flexible model for count matrix factorization, which accurately captures topics with diverse characteristics when applied to topic modeling of document corpora.

2 Preliminaries

2.1 Negative Binomial Distribution

The Poisson distribution $X \sim \text{Pois}(\lambda)$ is commonly used for modeling count data. It has the probability mass function $f_X(k) = e^\lambda \lambda^k / k!$, where $k \in \{0, 1, \dots\}$, with both the mean and variance equal to λ. A gamma distribution with shape r and scale $p/(1-p)$ can be placed as a prior on λ to produce a negative binomial (a.k.a, gamma-Poisson) distribution as

$$f_X(k) = \int_0^\infty \text{Pois}(k; \lambda) \text{Gamma}\left(\lambda; r, p/(1-p)\right) d\lambda$$
$$= \frac{\Gamma(r+k)}{k!\Gamma(r)}(1-p)^r p^k \quad (1)$$

where $\Gamma(\cdot)$ denotes the gamma function. Parameterized by $r > 0$ and $p \in (0, 1)$, this distribution $X \sim \text{NB}(r, p)$ has a variance $rp/(1-p)^2$ larger than the mean $rp/(1-p)$, and thus it is usually favored for modeling overdispersed count data. More detailed discussions about the negative binomial and related distributions and the corresponding stochastic processes defined on $\mathbb{R}^+$ can be found in Kozubowski and Podgórski (2009) and Barndorff-Nielsen et al. (2010).

2.2 Lévy Random Measures

Beta-negative binomial processes are created using Lévy random measures. Following Wolpert et al. (2011), for any $\nu^+ \geq 0$ and any probability distribution $\pi(dpd\omega)$ on $\mathbb{R} \times \Omega$, let $K \sim \text{Pois}(\nu^+)$ and $\{(p_k, \omega_k)\}_{1 \leq k \leq K} \overset{iid}{\sim} \pi(dpd\omega)$. Defining $\mathbf{1}_A(\omega_k)$ as being one if $\omega_k \in A$ and zero otherwise, the random measure $\mathcal{L}(A) \equiv \sum_{k=1}^K \mathbf{1}_A(\omega_k)p_k$ assigns independent infinitely divisible random variables $\mathcal{L}(A_i)$ to disjoint Borel sets $A_i \subset \Omega$, with characteristic functions

$$E\left[e^{it\mathcal{L}(A)}\right] = \exp\left\{\int \int_{\mathbb{R} \times A} (e^{itp} - 1)\nu(dpd\omega)\right\} \quad (2)$$

with $\nu(dpd\omega) \equiv \nu^+ \pi(dpd\omega)$. A random signed measure $\mathcal{L}$ satisfying (2) is called a Lévy random measure. More generally, if the Lévy measure $\nu(dpd\omega)$ satisfies

$$\int \int_{\mathbb{R} \times S} (1 \wedge |p|)\nu(dpd\omega) < \infty \quad (3)$$

for each compact $S \subset \Omega$, it need not be finite for the Lévy random measure $\mathcal{L}$ to be well defined; the notation $1 \wedge |p|$ denotes $\min\{1, |p|\}$. A nonnegative Lévy random measure $\mathcal{L}$ satisfying (3) was called a completely random measure (CRM) by Kingman (1967, 1993) and an additive random measure by Çinlar (2011). It was introduced for machine learning by Thibaux and Jordan (2007) and Jordan (2010).

2.3 Beta Process

The beta process (BP) was defined by Hjort (1990) for survival analysis with $\Omega = \mathbb{R}_+$. Thibaux and Jordan (2007) generalized the process to an arbitrary measurable space Ω by defining a CRM B on a product space $[0, 1] \times \Omega$ with the Lévy measure

$$\nu_{\text{BP}}(dpd\omega) = cp^{-1}(1-p)^{c-1}dpB_0(d\omega). \quad (4)$$

Here $c > 0$ is a concentration parameter (or concentration function if c is a function of ω), B_0 is a continuous finite measure over Ω, called the base measure, and $\alpha = B_0(\Omega)$ is the mass parameter. Since $\nu_{\text{BP}}(dpd\omega)$ integrates to infinity but satisfies (3), a countably infinite number of i.i.d. random points $\{(p_k, \omega_k)\}_{k=1,\infty}$ are obtained from the Poisson process with mean measure ν_{BP} and $\sum_{k=1}^\infty p_k$ is finite, where the atom $\omega_k \in \Omega$ and its weight $p_k \in [0, 1]$. Therefore, we can express a BP draw, $B \sim \text{BP}(c, B_0)$, as $B = \sum_{k=1}^\infty p_k \delta_{\omega_k}$, where δ_{ω_k} is a unit measure at the atom ω_k. If B_0 is discrete (atomic) and of the form $B_0 = \sum_k q_k \delta_{\omega_k}$, then $B = \sum_k p_k \delta_{\omega_k}$ with $p_k \sim \text{Beta}(cq_k, c(1-q_k))$. If B_0 is mixed discrete-continuous, B is the sum of the two independent contributions.

3 The Beta Process and the Negative Binomial Process

Let B be a BP draw as defined in Sec. 2.3, and therefore $B = \sum_{k=1}^\infty p_k \delta_{\omega_k}$. A Bernoulli process $\text{BeP}(B)$ has atoms appearing at the same locations as those of B; it assigns atom ω_k unit mass with probability p_k, and zero mass with probability $1 - p_k$, i.e., Bernoulli. Consequently, each draw from $\text{BeP}(B)$ selects a (finite) subset of the atoms in B. This construction is attractive because the beta distribution is a conjugate prior for the Bernoulli distribution. It has diverse applications including document classification (Thibaux and Jordan, 2007), dictionary learning (Zhou et al., 2012, 2009, 2011) and topic modeling (Li et al., 2011).

The beta distribution is also the conjugate prior for the negative binomial distribution parameter p, which

suggests coupling the beta process with the negative binomial process. Further, for modeling flexibility it is also desirable to place a prior on the negative-binomial parameter r, this motivating a *marked* beta process.

3.1 The Beta-Negative Binomial Process

Recall that a BP draw $B \sim \mathrm{BP}(c, B_0)$ can be considered as a draw from the Poisson process with mean measure ν_{BP} in (4). We can mark a random point (ω_k, p_k) of B with a random variable r_k taking values in $\mathbb{R}^+$, where r_k and $r_{k'}$ are independent for $k \neq k'$. Using the marked Poisson processes theorem (Kingman, 1993), $\{(p_k, r_k, \omega_k)\}_{k=1,\infty}$ can be regarded as random points drawn from a Poisson process in the product space $[0,1] \times \mathbb{R}^+ \times \Omega$, with the Lévy measure

$$\nu_{\mathrm{BP}}^*(dpdrd\omega) = cp^{-1}(1-p)^{c-1}dp R_0(dr) B_0(d\omega) \quad (5)$$

where R_0 is a continuous finite measure over $\mathbb{R}^+$ and $\gamma = R_0(\mathbb{R}^+)$ is the mass parameter. With (5) and using $R_0 B_0$ as the base measure, we construct a marked beta process $B^* \sim \mathrm{BP}(c, R_0 B_0)$, producing

$$B^* = \sum_{k=1}^{\infty} p_k \delta_{(r_k, \omega_k)} \quad (6)$$

where a point $(p_k, r_k, \omega_k) \sim \nu_{\mathrm{BP}}^*(dpdrd\omega)/\nu_{\mathrm{BP}}^+$ contains an atom $(r_k, \omega_k) \in \mathbb{R}^+ \times \Omega$ with weight $p_k \in [0,1]$.

With B^* constituted as in (6), we define the ith draw from a negative binomial process as $X_i \sim \mathrm{NBP}(B^*)$, with

$$X_i = \sum_{k=1}^{\infty} \kappa_{ki} \delta_{\omega_k}, \quad \kappa_{ki} \sim \mathrm{NB}(r_k, p_k). \quad (7)$$

The BP draw B^* in (6) defines a set of parameters $\{(p_k, r_k, \omega_k)\}_{k=1,\infty}$, and (r_k, p_k) are used within the negative binomial distribution to draw a count κ_{ki} for atom ω_k. The $\{(p_k, r_k, \omega_k)\}_{k=1,\infty}$ are shared among all draws $\{X_i\}$, and therefore the atoms $\{\omega_k\}$ are shared; the count associated with a given atom ω_k is a function of index i, represented by κ_{ki}.

In the beta-Bernoulli process (Thibaux and Jordan, 2007), we also yield draws like $\{X_i\}$ above, except that in that case the κ_{ki} is replaced by a one or zero, drawn from a Bernoulli distribution. The replacement of κ_{ki} with a one or zero implies that in the beta-Bernoulli process a given atom ω_k is either used (weight one) or not (weight zero). Since in the proposed beta-negative binomial process the κ_{ki} corresponds to counts of "dish" ω_k, with the number of counts drawn from a negative binomial distribution with parameters (r_k, p_k), we may view the proposed model as a generalization to a "multi-scoop" version of the beta-Bernoulli process. Rather than simply randomly selecting dishes $\{\omega_k\}$ from a buffet, here each X_i may draw *multiple* "scoops" of each ω_k.

3.2 Model Properties

Assume we already observe $\{X_i\}_{i=1,n}$. Since the beta and negative binomial processes are conjugate, the conditional posterior of p_k at an observed point of discontinuity (r_k, ω_k) is

$$p_k \sim \mathrm{Beta}(m_{nk}, c + nr_k) \quad (8)$$

where $m_{nk} = \sum_{i=1}^{n} \kappa_{ki}$ and $\kappa_{ki} = X_i\{\omega_k\}$. The posterior of the Lévy measure of the continuous part can be expressed as

$$\nu_{\mathrm{BP},n}^*(dpdrd\omega) = cp^{-1}(1-p)^{c_n-1}dp R_0(dr) B_0(d\omega) \quad (9)$$

where $c_n = c + nr$ is a concentration function. With (8) and (9), the posterior $B^*|\{X_i\}_{1,n}$ is defined, and following the notation in Kim (1999); Miller (2011); Thibaux (2008); Thibaux and Jordan (2007), it can be expressed as

$$B^*|\{X_i\}_{1,n} \sim \mathrm{BP}\left(c_n, \frac{c}{c_n}R_0 B_0 + \frac{1}{c_n}\sum_k m_{nk}\delta_{(r_k,\omega_k)}\right) \quad (10)$$

where

$$c_n = \begin{cases} c + m_{nk} + nr_k, & \text{if } (r, \omega) = (r_k, \omega_k) \in \mathcal{D} \\ c + nr, & \text{if } (r, \omega) \in (\mathbb{R}^+ \times \Omega) \backslash \mathcal{D} \end{cases} \quad (11)$$

where $\mathcal{D} = \{(r_k, \omega_k)\}_k$ is the discrete space including all the points of discontinuity observed so far.

Thibaux and Jordan (2007) showed that the IBP (Griffiths and Ghahramani, 2005) can be generated from the beta-Bernoulli process by marginalizing out the draw from the beta process. In the IBP metaphor, customer $n + 1$ walks through a buffet line, trying a dish tasted by previous customers with probability proportional to the popularity of that dish among previous customers; additionally, this customer tries $K_{n+1} \sim \mathrm{Pois}(\frac{c\alpha}{c+n})$ new dishes. By placing a count distribution on each observation, a BNB process prior naturally leads to a "multiple-scoop" generalization of the original IBP (i.e., a msIBP). That is, each customer now takes a number of scoops of each selected dish while walking through the buffet line.

Unlike the IBP, however, the msIBP predictive posterior distribution of $X_{n+1}|\{X_i\}_{i=1}^n$ does not have a closed form solution for a general measure R_0. This occurs because R_0 is not conjugate to the negative binomial distribution. We can analytically compute the distribution of the number of new dishes sampled. To find the number of new dishes sampled, note that an atom $\{(p_k, r_k, \omega_k)\}$ produces a dish that is sampled with probability $1 - (1 - p_k)^{r_k}$. Using the Poisson process decomposition theorem, we see that the number

of sampled dishes has distribution $\text{Pois}(\nu_{\text{msIBP}(n)}^{+})$,

$$\nu_{\text{msIBP}(n)}^{+}$$
$$= \alpha \int_{0}^{\infty} \int_{0}^{1} c(1-(1-p)^{r})p^{-1}(1-p)^{c+nr-1}dpR_{0}(dr).$$

When $R_{0} = \delta_{1}$, that is $r = 1$ and $\gamma = 1$, then the number of new dishes sampled has distribution $\text{Pois}\left(\alpha\frac{c}{c+n}\right)$. Note that for choices of R_{0} where $\nu_{\text{msIBP}(0)}^{+}$ is finite, $\nu_{\text{msIBP}(n)}^{+} \to 0$ as $n \to \infty$, and hence the number of new dishes goes to 0. Note that for the special case $R_{0} = \delta_{1}$, the negative binomial process becomes the geometric process, and (10) reduces to the posterior form discussed in Thibaux (2008); this simplification is not considered in the experiments, as it is often overly restrictive.

3.3 Finite Approximations for Beta Process

Since $\nu_{\text{BP}}^{+} = \nu_{\text{BP}}^{*}([0,1] \times \mathbb{R}^{+} \times \Omega) = \infty$, a BP generates countably infinite random points. For efficient computation, it is desirable to construct a finite Lévy measure which retains the random points of BP with non-negligible weights and whose infinite limit converges to ν_{BP}^{*}. We propose such a finite Lévy measure as

$$\nu_{\epsilon\text{BP}}^{*}(dpdrd\omega) = cp^{c\epsilon-1}(1-p)^{c(1-\epsilon)-1}dpR_{0}(dr)B_{0}(d\omega) \tag{12}$$

where $\epsilon > 0$ is a small constant, and we have

$$\nu_{\epsilon\text{BP}}^{+} = \nu_{\epsilon\text{BP}}^{*}([0,1] \times \mathbb{R}^{+} \times \Omega) = c\gamma\alpha\text{B}(c\epsilon, c(1-\epsilon)) \tag{13}$$

where $\text{B}(c\epsilon, c(1-\epsilon))$ is the beta function and it approaches infinity as $\epsilon \to 0$. Since

$$\lim_{\epsilon\to 0} \frac{\nu_{\epsilon\text{BP}}^{*}(dpdrd\omega)}{\nu_{\text{BP}}^{*}(dpdrd\omega)} = \lim_{\epsilon\to 0}\left(\frac{p}{1-p}\right)^{c\epsilon} = 1$$

we can conclude that $\nu_{\epsilon\text{BP}}^{*}(dpdrd\omega)$ approaches $\nu_{\text{BP}}^{*}(dpdrd\omega)$ as $\epsilon \to 0$.

Using $\nu_{\epsilon\text{BP}}^{*}(dpdrd\omega)$ as the finite approximation, a draw from ϵBP $B_{\epsilon}^{*} \sim \epsilon\text{BP}(c, B_{0})$ can be expressed as

$$B_{\epsilon}^{*} = \sum_{k=1}^{K} p_{k}\delta_{(r_{k},\omega_{k})}, \quad K \sim \text{Pois}(\nu_{\epsilon\text{BP}}^{+})$$

where $\{(p_{k}, r_{k}, \omega_{k})\}_{k=1}^{K} \overset{iid}{\sim} \pi(dpdrd\omega) \equiv \nu_{\epsilon\text{BP}}^{*}(dpdrd\omega)/\nu_{\epsilon\text{BP}}^{+}$ and we have $\pi(dp) = \text{Beta}(p; c\epsilon, c(1-\epsilon))dp$, $\pi(dr) = R_{0}(dr)/\gamma$ and $\pi(d\omega) = B_{0}(d\omega)/\alpha$. A reversible jump MCMC algorithm can be used to sample the varying dimensional parameter K. We can also simply choose a small enough ϵ and set $K = \mathbb{E}[\text{Pois}(\nu_{\epsilon\text{BP}}^{+})] = \nu_{\epsilon\text{BP}}^{+}$.

A finite approximation by restricting $p \in [\epsilon, 1]$ (Wolpert et al., 2011) and stick breaking representations (Paisley et al., 2010; Teh et al., 2007) may also be employed to approximate the infinite model.

4 Poisson Factor Analysis

Given $K \leq \infty$ and a count matrix $\mathbf{X} \in \mathbb{R}^{P \times N}$ with P terms and N samples, discrete latent variable models assume that the entries of $\mathbf{X}$ can be explained as a sum of smaller counts, each produced by a hidden factor, or in the case of topic modeling, a hidden topic. We can factorize $\mathbf{X}$ under the Poisson likelihood as

$$\mathbf{X} = \text{Pois}(\mathbf{\Phi}\mathbf{\Theta}) \tag{14}$$

where $\mathbf{\Phi} \in \mathbb{R}^{P \times K}$ is the factor loading matrix, each column of which is a factor encoding the relative importance of each term; $\mathbf{\Theta} \in \mathbb{R}^{K \times N}$ is the factor score matrix, each column of which encodes the relative importance of each atom in a sample. This is called Poisson factor analysis (PFA).

We can augment (14) as

$$x_{pi} = \sum_{k=1}^{K} x_{pik}, \quad x_{pik} \sim \text{Pois}(\phi_{pk}\theta_{ki}) \tag{15}$$

which is also used in Dunson and Herring (2005) for a discrete latent variable model. This form is useful for inferring ϕ_{pk} and θ_{ki}. As proved in Lemma 4.1 (below), we can have another equivalent augmentation as

$$x_{pi} \sim \text{Pois}\left(\sum_{k=1}^{K} \phi_{pk}\theta_{ki}\right), \quad \zeta_{pik} = \frac{\phi_{pk}\theta_{ki}}{\sum_{k=1}^{K}\phi_{pk}\theta_{ki}}$$
$$[x_{pi1}, \cdots, x_{piK}] \sim \text{Mult}(x_{pi}; \zeta_{pi1}, \cdots, \zeta_{piK}) \tag{16}$$

which assigns x_{pi} into the K latent factors. Both augmentations in (15) and (16) are critical to derive efficient inferences, which were not fully exploited by related algorithms (Buntine and Jakulin, 2006; Canny, 2004; Lee and Seung, 2000; Williamson et al., 2010).

Lemma 4.1. *Suppose that* $x_{1}, \ldots, x_{K}$ *are independent random variables with* $x_{k} \sim \text{Pois}(\lambda_{k})$ *and* $x = \sum_{k=1}^{K} x_{k}$. *Set* $\lambda = \sum_{k=1}^{K} \lambda_{k}$; *let* $(y, y_{1}, \ldots, y_{K})$ *be random variables such that*

$$y \sim \text{Pois}(\lambda), \quad (y_{1}, \ldots, y_{k})|y \sim \text{Mult}\left(y; \frac{\lambda_{1}}{\lambda}, \ldots, \frac{\lambda_{K}}{\lambda}\right).$$

Then the distribution of $\boldsymbol{x} = (x, x_{1}, \ldots, x_{K})$ *is the same as the distribution of* $\boldsymbol{y} = (y, y_{1}, \ldots, y_{K})$.

Proof. For $t = [t_{0}, \cdots, t_{K}] \in \mathbb{R}^{K+1}$ and compare the characteristic functions (CF) of $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\mathbb{E}\left[e^{it^{T}\boldsymbol{x}}\right] = \prod_{k=1}^{K}\mathbb{E}\left[e^{it(t_{0}+t_{k})x_{k}}\right] = e^{\sum_{k=1}^{K}\lambda_{k}e^{i(t_{0}+t_{k})-1}};$$

$$\mathbb{E}\left[e^{it^{T}\boldsymbol{y}}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{it^{T}\boldsymbol{y}}\Big|y\right]\right] = \mathbb{E}\left[\left(\sum_{k=1}^{K}\frac{\lambda_{k}}{\lambda}e^{i(t_{0}+t_{k})}\right)^{y}\right]$$
$$= e^{\sum_{k=1}^{K}\lambda_{k}e^{i(t_{0}+t_{k})-1}}.$$

Since the CF uniquely characterizes the distribution, the distributions are the same. $\square$

4.1 Beta-Gamma-Gamma-Poisson Model

Notice that for a sample, the total counts are usually observed, and it is the counts assigned to each factor that are often latent and need to be inferred. However, the BNB process in Section 3.1 does not tell us how the total counts are assigned to the latent factors. Recalling (1), the negative binomial distribution is a gamma-Poisson mixture distribution. Therefore, we can equivalently represent $X_i \sim \text{NBP}(B^*)$ as $X_i \sim \mathcal{P}(\mathcal{T}(B^*))$, where $\mathcal{T}(B^*)$ is a gamma process defined on B^* and $\mathcal{P}(\mathcal{T}(B^*))$ is a Poisson process defined on $\mathcal{T}(B^*)$. Thus at atom ω_k, the associated count $\kappa_{ki} \sim \text{NB}(r_k, p_k)$ can be expressed as

$$\kappa_{ki} \sim \text{Pois}(\theta_{ki}), \quad \theta_{ki} \sim \text{Gamma}(r_k, p_k/(1-p_k)) \quad (17)$$

where θ_{ki} is the weight of X_i at atom ω_k in the gamma process $\mathcal{T}(B^*)$. If we further place a gamma prior on r_k, then $(p_k, r_k, \theta_{ki}, \kappa_{ki})$ naturally forms a beta-gamma-gamma-Poisson hierarchical structure, and we call the BNB process PFA formulated under this structure the $\beta\gamma\Gamma$-PFA. This kind of hierarchical structure is useful for sharing statistical strength between different groups of data, and efficient inference is obtained by exploiting conjugacies in both the beta-negative binomial and gamma-Poisson constructions.

Using ϵBP $\nu^*_{\epsilon\text{BP}}(dpdrd\phi)$ on $[0,1] \times \mathbb{R}^+ \times \mathbb{R}^P$ as the base measure for the negative binomial process, we construct the ϵBNB process and apply it as the nonparametric Bayesian prior for PFA. We have $K \sim \text{Pois}(\nu^+_{\epsilon\text{BP}}) = \text{Pois}(c\gamma\alpha\text{B}(c\epsilon, c(1-\epsilon)))$, and thus $K \leq \infty$ with the equivalence obtained at $\epsilon = 0$. Using the beta-gamma-gamma-Poisson construction, we have

$$x_{pi} = \sum_{k=1}^{K} x_{pik}, \quad x_{pik} \sim \text{Pois}(\phi_{pk}\theta_{ki}) \quad (18)$$

$$\phi_k \sim \text{Dir}(a_\phi, \cdots, a_\phi) \quad (19)$$

$$\theta_{ki} \sim \text{Gamma}\left(r_k, \frac{p_k}{1-p_k}\right) \quad (20)$$

$$r_k \sim \text{Gamma}(c_0 r_0, 1/c_0) \quad (21)$$

$$p_k \sim \text{Beta}(c\epsilon, c(1-\epsilon)). \quad (22)$$

4.2 MCMC Inference

Denote $x_{\cdot ik} = \sum_{p=1}^{P} x_{pik}$, $x_{p\cdot k} = \sum_{i=1}^{N} x_{pik}$, $x_{\cdot\cdot k} = \sum_{p=1}^{P}\sum_{i=1}^{N} x_{pik}$ and $x_{\cdot\cdot\cdot} = \sum_{p=1}^{P}\sum_{k=1}^{K} x_{pik}$. The size of K is upper bounded by $\nu^+_{\epsilon\text{BP}} = c\gamma\alpha\text{B}(c\epsilon, c(1-\epsilon))$.

Sampling x_{pik}. Use (16).

Sampling ϕ_k. Exploiting the relationships between the Poisson and multinomial distributions and using $\sum_{p=1}^{P} \phi_{pk} = 1$, one can show that $p([x_{1ik}, \cdots, x_{Pik}]|-) \sim \text{Mult}(x_{\cdot ik}; \phi_k)$, thus we have

$$p(\phi_k|-) \sim \text{Dir}(a_\phi + x_{1\cdot k}, \cdots, a_\phi + x_{P\cdot k}). \quad (23)$$

Sampling p_k. Marginalizing ϕ_k and θ_{ki} out, $x_{\cdot ik} \sim \text{NB}(r_k, p_k)$, $p_k \sim \text{Beta}(c\epsilon, c(1-\epsilon))$, thus

$$p(p_k|-) \sim \text{Beta}(c\epsilon + x_{\cdot\cdot k}, c(1-\epsilon) + Nr_k). \quad (24)$$

Sampling r_k. It can be shown that $p(r_k|-) \propto \text{Gamma}(r_k; c_0 r_0, 1/c_0) \prod_{i=1}^{N} \text{NB}(x_{\cdot ik}; r_k, p_k)$, thus

$$p(r_k|-) \sim \text{Gamma}\left(c_0 r_0, \frac{1}{c_0 - N\log(1-p_k)}\right) \quad (25)$$

if $x_{\cdot\cdot k} = 0$. If $x_{\cdot\cdot k} \neq 0$, we prove in Lemma 4.2 that $g(r_k) = \log p(r_k|-)$ is strictly concave if $c_0 r_0 \geq 1$, then we can use Newton's method to find an estimate as

$$\tilde{r}_k = r_k^t - \frac{p'(r_k^t|-)}{p''(r_k^t|-)}$$

which can be used to construct a proposal in a Metropolis-Hastings (MH) algorithm as

$$r'_k \sim \mathcal{N}\left(\tilde{r}_k, \mu\sqrt{\tilde{r}_k}\right) \quad (26)$$

where μ is an adjustable stepsize. Note that the adaptive rejection sampling in Gilks and Wild (1992) may also be used to sample r_k.

Sampling θ_{ki}. Using (18) and (20), we have

$$p(\theta_{ki}|-) \sim \text{Gamma}(r_k + x_{\cdot ik}, p_k). \quad (27)$$

Lemma 4.2. *If $x_{\cdot\cdot k} \neq 0$, then for any $c_0 r_0 \geq 1$, $g(r_k) = \log p(r_k|-)$ is strictly concave.*

Proof. Since $g'(r_k) = (c_0 r_0 - 1)/r_k - c_0 + N\log(1 - p_k) - N\psi(r_k) + \sum_{i=1}^{N} \psi(r_k + x_{\cdot ik})$ and $g''(r_k) = -(c_0 r_0 - 1)/r_k^2 - N\psi_1(r_k) + \sum_{i=1}^{N} \psi_1(r_k + x_{\cdot ik})$, where $\psi(x)$ is the diagmma function and $\psi_1(x)$ is the trigamma function which is strictly decreasing for $x > 0$, if $c_0 r_0 \geq 1$ and $x_{\cdot\cdot k} \neq 0$, we have $g''(r_k) < 0$, and thus $g(r_k)$ is strictly concave and has a unique maximum. $\square$

5 Related Discrete Latent Variable Models

The hierarchical form of the $\beta\gamma\Gamma$-PFA model shown in (18)-(22) can be modified in various ways to connect to previous discrete latent variable models. For example, we can let $\{\theta_{ki}\}_{i=1,N} \equiv g_k$ and $g_k \sim \text{Gamma}(g_0/K, 1)$, resulting in the infinite gamma-Poisson feature model in Titsias (2008) as $K \to \infty$. Thibaux (2008) showed that it can also be derived from a gamma-Poisson process. Although this is a nonparametric model supporting an infinite number of features, requiring $\{\theta_{ki}\}_{i=1,N} \equiv g_k$ may be too restrictive. We mention that Broderick et al. (2012) have independently investigated beta-negative binomial processes for mixture and admixture models.

Table 1: Algorithms related to PFA under various prior settings. The gamma scale and shape parameters and the sparsity of the factor scores are controlled by p_k, r_k and z_{ki}, respectively.

PFA priors	Infer p_k	Infer r_k	Infer z_{ki}	Infer K	Related algorithms
Γ	$\times$	$\times$	$\times$	$\times$	NMF
Dir	$\times$	$\times$	$\times$	$\times$	LDA
$\beta\Gamma$	$\checkmark$	$\times$	$\times$	$\checkmark$	GaP
S$\gamma\Gamma$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	FTM
$\beta\gamma\Gamma$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	

Before examining the details, defined by p_k and r_k in (20), z_{ki} in (34) and the subset of the K factors needed to represent the data are inferred, we summarize in Table 1 the connections between related algorithms and PFA with various priors, including non-negative matrix factorization (NMF) (Lee and Seung, 2000), latent Dirichlet allocation (LDA) (Blei et al., 2003), gamma-Poisson (GaP) (Canny, 2004) and the focused topic model (FTM) (Williamson et al., 2010). Note that the gamma scale parameters $p_k/(1-p_k)$ and $p_{k'}/(1-p_{k'})$ in (20) are generally different for $k \neq k'$ in $\beta\gamma\Gamma$-PFA, and thus the normalized factor score $\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i / \sum_{k=1}^{K} \theta_{ki}$ does not follow a Dirichlet distribution. This is a characteristic that distinguishes $\beta\gamma\Gamma$-PFA from models with a global gamma scale parameter, where the normalized factor scores follow Dirichlet distributions.

5.1 Nonnegative Matrix Factorization and a Gamma-Poisson Factor Model

We can modify the $\beta\gamma\Gamma$-PFA into a Γ-PFA by letting

$$\phi_{pk} \sim \text{Gamma}(a_\phi, 1/b_\phi) \tag{28}$$

$$\theta_{ki} \sim \text{Gamma}(a_\theta, g_k/a_\theta). \tag{29}$$

Using (18), (28) and (29), one can show that

$$p(\phi_{pk}|-) \sim \text{Gamma}(a_\phi + x_{p \cdot k}, 1/(b_\phi + \theta_{k \cdot})) \tag{30}$$

$$p(\theta_{ki}|-) \sim \text{Gamma}(a_\theta + x_{\cdot ik}, 1/(a_\theta/g_k + \phi_{\cdot k})) \tag{31}$$

where $\theta_{k\cdot} = \sum_{i=1}^{N} \theta_{ki}$ and $\phi_{\cdot k} = \sum_{p=1}^{P} \phi_{pk}$. If $a_\phi \geq 1$ and $a_\theta \geq 1$, using (16), (30) and (31), we can substitute $\mathbb{E}[x_{pik}]$ into the modes of ϕ_{pk} and θ_{ki}, leading to an Expectation-Maximization (EM) algorithm as

$$\phi_{pk} = \phi_{pk} \frac{\frac{a_\phi - 1}{\phi_{pk}} + \sum_{i=1}^{N} \frac{x_{pi}\theta_{ki}}{\sum_{k=1}^{K} \phi_{pk}\theta_{ki}}}{b_\phi + \theta_{k\cdot}} \tag{32}$$

$$\theta_{ki} = \theta_{ki} \frac{\frac{a_\theta - 1}{\theta_{ki}} + \sum_{p=1}^{P} \frac{x_{pi}\phi_{pk}}{\sum_{k=1}^{K} \phi_{pk}\theta_{ki}}}{a_\theta/g_k + \phi_{\cdot k}}. \tag{33}$$

If we set $b_\phi = 0$, $a_\phi = a_\theta = 1$ and $g_k = \infty$, then (32) and (33) are the same as those of non-negative matrix factorization (NMF) with an objective function of minimizing the KL divergence $D_{KL}(\mathbf{X}||\boldsymbol{\Phi}\boldsymbol{\Theta})$ (Lee and Seung, 2000). If we set $b_\phi = 0$ and $a_\phi = 1$, then (32) and (33) are the same as those of the gamma-Poisson (GaP) model of Canny (2004), in which setting $a_\theta = 1.1$ and estimating g_k with $g_k = \mathbb{E}[\theta_{ki}]$ are suggested. Therefore, as summarized in Table 1, NMF is a special cases of the Γ-PFA, which itself can be considered as a special case of the $\beta\gamma\Gamma$-PFA with fixed r_k and p_k. If we impose a Dirichlet prior on $\boldsymbol{\phi}_k$, the GaP can be consider as a special case of $\beta\Gamma$-PFA, which itself is a special case of the $\beta\gamma\Gamma$-PFA with a fixed r_k.

5.2 Latent Dirichlet Allocation

We can modify the $\beta\gamma\Gamma$-PFA into a Dirichlet PFA (Dir-PFA) by changing the prior of $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_i \sim \text{Dir}(a_\theta, \cdots, a_\theta)$. Similar to the derivation of $p(\boldsymbol{\phi}_k|-)$ in (23), one can show $p([x_{\cdot i1}, \cdots, x_{\cdot iK}]|-) \sim \text{Mult}(x_{\cdot i \cdot}; \boldsymbol{\theta}_i)$ and thus $p(\boldsymbol{\theta}_i|-) \sim \text{Dir}(a_\theta + x_{\cdot i1}, \cdots, a_\theta + x_{\cdot iK})$. Dir-PFA and LDA (Blei et al., 2003; Hoffman et al., 2010) have the same block Gibbs sampling and variational Bayes inference equations (not shown here for brevity). It may appear that Dir-PFA should differ from LDA via the Poisson distribution; however, imposing Dirichlet priors on both factor loadings and scores makes it essentially lose that distinction.

5.3 Focused Topic Model

We can construct a sparse $\gamma\Gamma$-PFA (S$\gamma\Gamma$-PFA) with the beta-Bernoulli process prior by letting $\theta_{ki} = z_{ki}s_{ki}$ and

$$s_{ki} \sim \text{Gamma}(r_k, p_k/(1-p_k)), \ r_k \sim \text{Gamma}(r_0, 1)$$

$$z_{ki} \sim \text{Bernoulli}(\pi_k), \ \pi_k \sim \text{Beta}(c\epsilon, c(1-\epsilon)). \tag{34}$$

If we fix $p_k = 0.5$, it can be shown that under the PFA framework, conditioning on z_{ki}, we have

$$x_{\cdot ik} \sim \text{NB}(z_{ki}r_k, 0.5) \tag{35}$$

$$x_{\cdot i \cdot} \sim \text{NB}\left(\sum_{k=1}^{K} z_{ki}r_k, 0.5\right) \tag{36}$$

$$\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i \bigg/ \sum_{k=1}^{K} \theta_{ki} \sim \text{Dir}(z_{1i}r_1, \cdots, z_{Ki}r_K) \tag{37}$$

$$[x_{pi1}, \cdots, x_{piK}] \sim \text{Mult}\left(x_{pi \cdot}; \tilde{\zeta}_{pi1}, \cdots, \tilde{\zeta}_{piK}\right) \tag{38}$$

where $\tilde{\zeta}_{pik} = (\phi_{pk}\tilde{\theta}_{ki})/\sum_{k=1}^{K} \phi_{pk}\tilde{\theta}_{ki}$. Therefore, S$\gamma\Gamma$-PFA has almost the same MCMC inference as the focused topic model (FTM) using the IBP compound Dirichlet priors (Williamson et al., 2010). Note that (35) and (36) are actually used in the FTM to infer r_k and z_{ki} (Williamson et al., 2010) without giving explicit explanations under the multinomial-Dirichlet construction, however, we show that both equations

naturally arise under SγΓ-PFA under the constraint that $p_k = 0.5$. In this sense, SγΓ-PFA provides justifications for the inference in Williamson et al. (2010).

6 Example Results and Discussions

We consider the JACM[1] and PsyRev[2] datasets, restricting the vocabulary to terms that occur in five or more documents in each corpus. The JACM includes 536 abstracts of the Journal of the ACM from 1987 to 2004, with 1,539 unique terms and 68,055 total word counts; the PsyRev includes 1281 abstracts from Psychological Review from 1967 to 2003, with 2,566 unique terms and 71,279 total word counts. As a comparison, the stopwords are kept in JACM and removed in PsyRev. We obtain similar results on other document corpora such as the NIPS corpus[3]. We focus on these two datasets for detailed comparison.

For each corpus, we randomly select 80% of the words from each document to form a training matrix $\mathbf{T}$, holding out the remaining 20% to form a testing matrix $\mathbf{Y} = \mathbf{X} - \mathbf{T}$. We factorize $\mathbf{T}$ as $\mathbf{T} \sim \text{Pois}(\mathbf{\Phi\Theta})$ and calculate the held-out per-word perplexity as

$$\exp\left(-\frac{1}{y_{..}}\sum_{p=1}^{P}\sum_{i=1}^{N} y_{pi} \log \frac{\sum_{s=1}^{S}\sum_{k=1}^{K}\phi_{pk}^s\theta_{ki}^s}{\sum_{s=1}^{S}\sum_{p=1}^{P}\sum_{k=1}^{K}\phi_{pk}^s\theta_{ki}^s}\right)$$

where S is the total number of collected samples, $y_{..} = \sum_{p=1}^{P}\sum_{i=1}^{N} y_{pi}$ and $y_{pi} = \mathbf{Y}(p,i)$. The final results are based on the average of five random training/testing partitions. We consider 2500 MCMC iterations, with the first 1000 samples discarded and every sample per five iterations collected afterwards. The performance measure is similar to those used in Asuncion et al. (2009); Wallach et al. (2009); Wang et al. (2011).

As discussed in Sec. 5 and shown in Table 1, NMF (Lee and Seung, 2000) is a special case of Γ-PFA; GaP (Canny, 2004) is a special case of βΓ-PFA; and in terms of inference, Dir-PFA is equivalent to LDA (Blei et al., 2003; Hoffman et al., 2010); and SγΓ-PFA is closely related to FTM (Williamson et al., 2010). Therefore, we are able to compare all these algorithms with βγΓ-PFA under the same PFA framework, all with MCMC inference.

We set the priors of Γ-PFA as $a_\phi = a_\theta = 1.01$, $b_\phi = 10^{-6}$, $g_k = 10^6$ and the prior of βΓ-PFA as $r_k = 1.1$; these settings closely follow those that lead to the EM algorithms of NMF and GaP, respectively. We find that Γ- and βΓ-PFAs in these forms generally yield better prediction performance than their EM

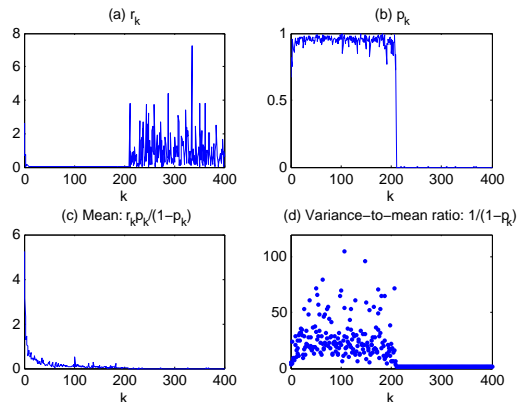Figure 1: Inferred r_k, p_k, mean $r_k p_k/(1-p_k)$ and variance-to-mean ratio $1/(1-p_k)$ for each factor by βγΓ-PFA with $a_\phi = 0.05$. The factors are shown in decreasing order based on the total number of word counts assigned to them. Of the $K_{\max} = 400$ possible factors, there are 209 active factors assigned nonzero counts. The results are obtained on the training count matrix of the PsyRev corpus based on the last MCMC iteration.

counterparts, thus we report the results of Γ- and βΓ-PFAs under these prior settings. We set the prior of Dir-PFA as $a_\theta = 50/K$, following the suggestion of the topic model toolbox[2] provided for Griffiths and Steyvers (2004). The parameters of βγΓ-PFA are set as $c = c_0 = r_0 = \gamma = \alpha = 1$. In this case, $\nu_{\epsilon\text{BP}}^+ = \pi/\sin(\pi\epsilon) \approx \epsilon^{-1}$ for a small ϵ, thus we preset a large upper-bound $K_{\max}$ and let $\epsilon = 1/K_{\max}$. The parameters of SγΓ-PFA are set as $c = r_0 = 1$ and $\epsilon = 1/K_{\max}$. The stepsize in (26) is initialized as $\mu = 0.01$ and is adaptively adjusted to maintain an acceptance rate between 25% and 50%. For all the algorithms, $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are preset with random values. Under the above settings, it costs about 1.5 seconds per iteration for βγΓ-PFA on the PsyRev corpus using a 2.67 GHz PC.

Figure 1 shows the inferred r_k and p_k, and the inferred mean $r_k p_k/(1-p_k)$ and variance-to-mean ratio (VMR) $1/(1-p_k)$ for each latent factor using the βγΓ-PFA algorithm on the PsyRev corpus. Of the $K = 400$ possible factors, there are 209 active factors assigned nonzero counts. There is a sharp transition between the active and nonactive factors for the values of r_k and p_k. The reason is that for nonactive factors, p_k and r_k are drawn from (24) and (25), respectively, and thus p_k with mean $c\epsilon/(c+Nr_k)$ is close to zero and r_k is approximately drawn from its prior Gamma$(c_0 r_0, 1/c_0)$; for active factors, the model adjusts the negative binomial distribution with both p_k and r_k to fit the data, and r_k would be close to zero and p_k would be close to one for an active factor with a small mean and a large VMR, as is often the case for both corpora considered.

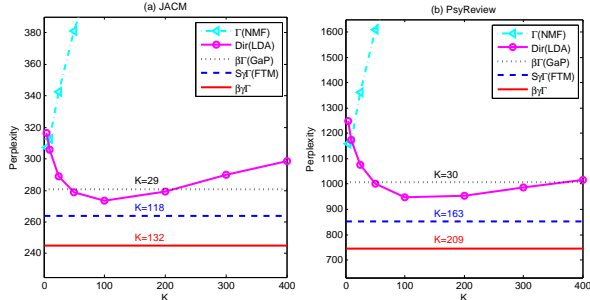We find that the first few dominant factors correspond to common topics popular both across and in-

Figure 2: Per-word perplexities on the test count matrix for (a) JACM and (b) PsyRev with $a_\phi = 0.05$. The results of Γ- and Dir-PFAs are a function of K. $\beta\Gamma$-, $S\gamma\Gamma$- and $\beta\gamma\Gamma$-PFAs all automatically infer the number of active factors, with the number at the the last MCMC iteration shown on top of the corresponding lines.

side documents. For example, the first two most dominant topics are characterized by "proposed, evidence, data, discussed, experiments" and "model, models, effects, response, predictions" in PysRev where the stop words are removed; the first two dominant topics in JACM, where the stop words are kept, are characterized by "the, of, a, is, in" and "we, of, in, and, the". These top factors generally have large means and small VMRs. The remaining topics have a diverse range of means and VMRs. A PsyRev factor with prominent words "masking, visual, stimulus, metacontrast" and a JACM factor with prominent words "local, consistency, finite, constraint" are example topics with large mean and large VMR. A PsyRev factor "rivalry, binocular, monocular, existence" and a JACM factor "search, binary, tree, nodes" are example topics with small mean and large VMR. Therefore, the $\beta\gamma\Gamma$-PFA captures topics with distinct characteristics by adjusting the negative binomial parameters r_k and p_k, and the characteristics of these inferred parameters may assist in factor/topic interpretation. Note that conventional topic models are susceptible to stop words, in that they may produce topics that are not readily interpretable if stop words are not removed (Blei et al., 2010). Our results show that when stop words are present, $\beta\gamma\Gamma$-PFA usually absorbs them into a few dominant topics with large mean and small VMR and the remaining topics are easily interpretable.

Figure 2 shows the performance comparison for both corpora with the factor loading (topic) Dirichlet prior set as $a_\phi = 0.05$. In both Γ- and Dir-PFAs, the number of factors K is a tuning parameter, and as K increases, Γ-PFA quickly overfits the training data and Dir-PFA shows signs of ovefitting around $K = 100$. In $\beta\Gamma$-, $S\gamma\Gamma$- and $\beta\gamma\Gamma$-PFAs, the number of factors is upper bounded by $K_{\max} = 400$ and an appropriate K is automatically inferred. As shown in Fig. 2, $\beta\gamma\Gamma$-PFA produces the smallest held out perplexity, followed by $S\gamma\Gamma$- and $\beta\Gamma$-PFAs, and their inferred sizes of K at the
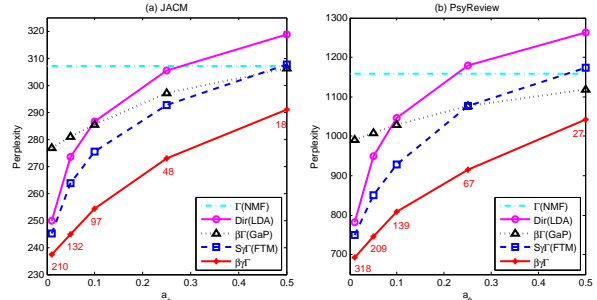

Figure 3: Per-word perplexities on the test count matrix of (a) JACM and (b) PsyRev as a function of the factor loading (topic) Dirichlet prior $a_\phi \in \{0.01, 0.05, 0.1, 0.25, 0.5\}$. The results of Γ- and Dir-PFAs are shown with the best settings of K under each a_ϕ. The number of active factors under each a_ϕ are all automatically inferred by $\beta\Gamma$-, $S\gamma\Gamma$- and $\beta\gamma\Gamma$-PFAs. The number of active factors inferred by $\beta\gamma\Gamma$-PFA at the last MCMC interaction are shown under the corresponding points.

last MCMC iteration are 132, 118 and 29 for JACM and 209, 163 and 30 for PsyRev, respectively.

Figure 3 shows the performance of these algorithms as a function of a_ϕ. For Γ-PFA, a_ϕ is fixed. For Dir-, $\beta\Gamma$-, $S\gamma\Gamma$- and $\beta\gamma\Gamma$-PFAs, a_ϕ influences the inferred sizes of K and the accuracies of held-out predictions. We find that a smaller a_ϕ generally supports a larger K, with better held-out prediction. However, if a_ϕ is too small it leads to overly specialized factor loadings (topics), that concentrate only on few terms. As shown in 3, $\beta\gamma\Gamma$-PFA yields the best results under each a_ϕ and it automatically infers the sizes of K as a function of a_ϕ.

7 Conclusions

A beta-negative binomial (BNB) process, which leads to a beta-gamma-Poisson process, is proposed for modeling multivariate count data. The BNB process is augmented into a beta-gamma-gamma-Poisson hierarchical structure and applied as a nonparametric Bayesian prior for Poisson factor analysis (PFA), an infinite discrete latent variable model. A finite approximation to the beta process Lévy random measure is proposed for convenient implementation. Efficient MCMC inference is performed by exploiting the relationships between the beta, gamma, Poisson, negative binomial, multinomial and Dirichlet distributions. Connections to previous models are revealed with detailed analysis. Model properties are discussed, and example results are presented on document count matrix factorization. Results demonstrate that by modeling latent factors with negative binomial distributions whose mean and variance are both learned, the proposed $\beta\gamma\Gamma$-PFA is well suited for topic modeling, defined quantitatively via perplexity calculations and more subjectively by capturing both common and specific aspects of a document corpus.

References

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

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

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of ACM*, 2010.

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

T. Broderick, L. Mackey, J. Paisley, and M. I. Jordan. Combinatorial clustering and the beta negative binomial process. *arXiv:1111.1802v2*, 2012.

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

A. C. Cameron, P. K. Trivedi, F. Milne, and J. Piggott. A microeconometric model of the demand for health care and health insurance in Australia. *The Review of Economic Studies*, 1988.

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

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

P. Deb and P. Trivedi. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 1997.

D. B. Dunson. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society Series B*, 2000.

D. B. Dunson. Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association*, 2003.

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

W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 1992.

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

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

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 1990.

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

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

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

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

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

T. J. Kozubowski and K. Podgórski. Distributional properties of the negative binomial Lévy process. *Probability and Mathematical Statisitcs*, 2009.

N. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 2005.

D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. In *NIPS*, 2000.

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

K. T. Miller. *Bayesian Nonparametric Latent Feature Models*. PhD thesis, University of California at Berkeley, 2011.

I. Moustaki and M. Knott. Generalized latent trait models. *Psychometrika*, 2000.

National Audubon Society. The Christmas Bird Count Historical Results. URL `http://www.audubon.org/bird/cbc`. Accessed 10/4/2011.

J. Paisley, A. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin. A stick-breaking construction of the beta process. In *ICML*, 2010.

M. D. Sammel, L. M. Ryan, and J. M. Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society Series B*, 1997.

Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *AISTATS*, 2007.

R. J. Thibaux. *Nonparametric Bayesian Models for Machine Learning*. PhD thesis, University of California at Berkeley, 2008.

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

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B*, 1999.

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

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

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

M. West. Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, 2003.

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.

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

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 Trans. Image Processing*, 2012.

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

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