Priors for Random Count Matrices Derived from a Family of Negative Binomial Processes

Mingyuan Zhou, Oscar Hernan Madrid Padilla, and James G. Scott
The University of Texas at Austin

May 25, 2015

Abstract

We define a family of probability distributions for random count matrices with a potentially unbounded number of rows and columns. The three distributions we consider are derived from the gamma-Poisson, gamma-negative binomial, and beta-negative binomial processes, which we refer to generically as a family of negative-binomial processes. Because the models lead to closed-form update equations within the context of a Gibbs sampler, they are natural candidates for nonparametric Bayesian priors over count matrices. A key aspect of our analysis is the recognition that, although the random count matrices within the family are defined by a row-wise construction, their columns can be shown to be independent and identically distributed. This fact is used to derive explicit formulas for drawing all the columns at once. Moreover, by analyzing these matrices' combinatorial structure, we describe how to sequentially construct a column-i.i.d. random count matrix one row at a time, and derive the predictive distribution of a new row count vector with previously unseen features. We describe the similarities and differences between the three priors, and argue that the greater flexibility of the gamma- and beta- negative binomial processes—especially their ability to model over-dispersed, heavy-tailed count data—makes these well suited to a wide variety of real-world applications. As an example of our framework, we construct a naive-Bayes text classifier to categorize a count vector to one of several existing random count matrices of different categories. The classifier supports an unbounded number of features, and unlike most existing methods, it does not require a predefined finite vocabulary to be shared by all the categories, and needs neither feature selection nor parameter tuning. Both the gamma- and beta- negative binomial processes are shown to significantly outperform the gamma-Poisson process when applied to document categorization, with comparable performance to other state-of-the-art supervised text classification algorithms.

The authors are with the Department of Information, Risk, and Operations Management and the Department of Statistics and Data Sciences, the University of Texas at Austin, Austin, TX 78712, USA. *Address for correspondence*: 2110 Speedway Stop B6500, Austin, TX 78712, USA. *Email:* mingyuan.zhou@mccombs.utexas.edu.

1 Introduction

1.1 Models for count matrices

The need to model a random count matrix arises in many settings, from linguistics to marketing to ecology. For example, in text analysis, we often observe a document-term matrix, whose rows record how many times word k appeared in a given document. In a biodiversity study, we may observe a site-species matrix, where each row records the number of times species k was observed at a given site. Similar applications arise in a wide variety of fields; for examples, see Cameron and Trivedi (1998), Chib et al. (1998), Canny (2004), Buntine and Jakulin (2006), Winkelmann (2008), Titsias (2008), and Zhou et al. (2012).

Nonparametric Bayesian analysis provides a natural setting in which to study random matrices, especially those with no natural upper bound on the number of rows or columns. Yet while there is a wide selection of nonparametric Bayesian models for random count vectors and random binary matrices, prior distributions over random count matrices are relatively underdeveloped. Moreover, a major conceptual problem in modeling a random count matrix arises when new rows are added sequentially. For example, as new documents are collected and processed in text analysis, each new document (represented by a new row of the matrix) may contain previously unseen words (features). This requires that new columns be added to the existing count matrix. But it is not obvious how to define the predictive distribution of this new row of a random count matrix, if the row contains previously unseen features. This is especially important in natural language processing, where a common application is to build a naive Bayes model for classifying new documents. Without having a predictive distribution that accounts for new features, one must often use a predetermined vocabulary and simply ignore the previously unseen terms appearing in a new document.

We directly address these issues by investigating a family of nonparametric Bayesian priors for random count matrices constructed from stochastic processes: the gamma-Poisson process, the gamma-negative binomial process (GNBP), and the beta-negative binomial process (BNBP). We show that all these processes lead to random count matrices with independent and identically distributed (i.i.d.) columns, which can be constructed by drawing all the columns at once, or by adding one row at a time. In addition, we show the gamma-Poisson

process, and for special cases of the GNBP and BNBP with common row-wise parameters, the generated random count matrices are exchangeable in both rows and columns.

Our derivation exactly marginalizes out the underlying stochastic processes to arrive at a probability mass function (PMF) for a column-i.i.d. random count matrix. In contrast to existing techniques that take the infinite limit of a finite-dimensional model, this novel procedure allows for the construction and analysis of much more flexible nonparametric priors for random matrices, and highlights certain model properties that are not evident from the finite-model limit. The argument relies upon a novel combinatorial analysis for calculating the number of ways to map a column-i.i.d. random count matrix to a structured random count matrix whose columns are ordered in a certain manner. This is a key step in deriving the predictive distribution of a new random count vector under a random count matrix.

As an application of our proposed framework, we construct a naive-Bayes text classification model. The approach does not require a predefined list of terms (features), and naturally accounts for documents with previously unseen terms. This also implies that random count matrices of different categories can be updated, analyzed, and tested completely in parallel. Moreover, the algorithm requires neither feature selection nor parameter tuning. Following Crammer et al. (2012), the algorithm may also be conveniently extended to an online learning setting. Empirical results suggest that both the proposed GNBP and BNBP models lead to substantially better out-of-sample classification performance, versus both the gamma-Poisson model and the multinomial model with Laplace smoothing. They also clearly outperform the text classification algorithms that first learn lower-dimensional feature vectors for documents and then train a multi-class classifier, and have comparable performance to the state-of-the-art discriminatively trained text classification algorithms, whose features need to be carefully constructed and parameters carefully selected.

1.2 Connections with existing work

Our paper is in the spirit of existing work on nonparametric Bayesian priors for random count vectors and random binary matrices. To model a random count vector, one may use the Chinese restaurant process, or any one of many other stochastic processes characterized by exchangeable partition probability functions (EPPFs) or sample-size dependent EPPFs; see,

for example, Blackwell and MacQueen (1973), Pitman (2006), Lijoi and Prünster (2010), and Zhou and Walker (2014). Likewise, to model a random binary matrix, one may use the Indian buffet process (Griffiths and Ghahramani, 2005, Teh and Gorur, 2009). These well-studied nonparametric Bayesian priors, however, are not directly useful for describing random count matrices. To address this gap, we investigate a family of nonparametric Bayesian priors for random count matrices, each based on a previously proposed stochastic process that has not been thoroughly studied: the gamma-Poisson process (Lo, 1982, Titsias, 2008), the gamma-negative binomial process, or GNBP (Zhou and Carin, 2015); and the beta-negative binomial process, or BNBP (Zhou et al., 2012, Broderick et al., 2015).

All three models can be derived as the marginal distribution of a suitably defined stochastic process with respect to a traditional sampling model for integer-valued counts. This parallels the construction of the models for count vectors or binary matrices mentioned previously. For example, the Chinese restaurant process describes a random count vector as the marginal of the Dirichlet process (Ferguson, 1973) under multinomial sampling. Likewise, the Indian buffet process describes a random binary matrix as the marginal of the beta process (Hjort, 1990) under Bernoulli sampling (Thibaux and Jordan, 2007). Similarly, we present the negative binomial process as the marginal of the gamma process under Poisson sampling, the GNBP as the marginal of the gamma process under negative binomial sampling, and the BNBP as the marginal of the beta process under negative binomial sampling.

The remainder of the paper is organized as follows. After some preliminary definitions and notation, we introduce in Section 2 three distinct nonparametric Bayesian priors for random count matrices. In Section 3, we construct nonparametric Bayesian naive Bayes classifiers to classifier a count vector to one of several existing count matrices and demonstrate their use in document categorization. The details for deriving the random count matrix priors from their underlying hierarchical stochastic processes are provided in the Appendix.

1.3 Notation and preliminaries

Stochastic processes. A gamma process (Ferguson, 1973) $G \sim \Gamma\mathrm{P}(G_0, 1/c)$ on the product space $\mathbb{R}_+ \times \Omega$, where $\mathbb{R}_+ = \{x : x > 0\}$, is defined by two parameters: a finite and continuous base measure G_0 over a complete separable metric space Ω, and a scale

$1/c$, such that $G(A) \sim \text{Gamma}(G_0(A), 1/c)$ for each $A \subset \Omega$. The Lévy measure of the gamma process is $\nu(drd\omega) = r^{-1}e^{-cr}drG_0(d\omega)$. Although the Lévy measure integrates to infinity, $\int_{\mathbb{R}_+\times\Omega} \min\{r, 1\}\nu(drd\omega)$ is finite, and therefore a draw from the gamma process $G \sim \Gamma\text{P}(G_0, 1/c)$ can be represented as the countably infinite sum $G = \sum_{k=1}^{\infty} r_k\delta_{\omega_k},\ \omega_k \sim g_0$, where $\gamma_0 = G_0(\Omega)$ is the mass parameter and $g_0(d\omega) = G_0(d\omega)/\gamma_0$ is the base distribution.

A beta process (Hjort, 1990) $B \sim \text{BP}(c, B_0)$ on the product space $[0, 1]\times\Omega$, is also defined by two parameters: a finite and continuous base measure B_0 over a complete separable metric space Ω, and a concentration parameter $c > 0$. The Lévy measure of the beta process in this paper is defined as

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

As $\int_{[0,1]\times\Omega} \nu(dpd\omega) = \infty$ and $\int_{[0,1]\times\Omega} \min\{p, 1\}\nu(dpd\omega) < \infty$, a draw from $B \sim \text{BP}(c, B_0)$ can be represented as $B = \sum_{k=1}^{\infty} p_k\delta_{\omega_k},\ \omega_k \sim g_0$, where $\gamma_0 = B_0(\Omega)$ is the mass parameter and $g_0(d\omega) = B_0(d\omega)/\gamma_0$ is the base distribution.

Random count matrices. A random count matrix is denoted generically by $\mathbf{N}_J \in \mathbb{Z}^{J\times K_J}$, $\mathbb{Z} = \{0, 1, \ldots\}$, where the J rows of $\mathbf{N}_J$ correspond to the J samples or cases, and the K_J columns to features that have been observed at least once across all rows. Throughout the paper, we will refer to count matrices constructed sequentially by row, for which we require a consistent notation. Suppose that a new case is observed; we use $\mathbf{N}^+_{J+1}$ to refer to the new part introduced to the matrix $\mathbf{N}_J$ by adding row $(J+1)$. Similarly, we use K^+_{J+1} to denote the number of new columns introduced by adding row $(J+1)$, meaning that $K_{J+1} := K_J + K^+_{J+1}$; $\boldsymbol{n}_{:k}$ to indicate the count vector corresponding to column k of the matrix; and $n_{\cdot k} = \sum_{j=1}^{K_J} \boldsymbol{n}_{:k}$ to denote the total number of counts of feature k across all rows. One may think of $\mathbf{N}^+_{J+1}$ as the combination of two submatrices: a row of K_J counts appended below $\mathbf{N}_J$, and then a $(J+1)\times K^+_{J+1}$ submatrix, whose first J rows are entirely zero, and whose K^+_{J+1} columns are inserted into random locations among original columns with their relative orders preserved.

Our convention is that a prior for a random count matrix is named by the stochastic process used to generate each of its rows. In this paper, we study three hierarchical stochastic processes, all in the family of negative binomial processes. Each such stochastic process is defined by the prior for an almost-surely discrete random measure, together with

a sampling model for generating counts. We denote the distribution of such a matrix as $\mathbf{N} \sim \text{ProcessM}(\boldsymbol{\theta})$, where "Process" is the name of the underlying hierarchical stochastic process, "M" stands for matrix, and $\boldsymbol{\theta}$ encodes the parameters of the process.

For example, to construct a gamma-Poisson or negative binomial process random count matrix, $\mathbf{N}_J \sim \text{NBPM}(\gamma_0, c)$, we draw a random measure $G \sim \Gamma\text{P}(G_0, 1/c)$ from a gamma process. Then for each row of the matrix, we independently draw $X_j \mid G \sim \text{PP}(G)$: a Poisson process such that $X_j(A) \sim \text{Pois}[G(A)]$ for all $A \subset \Omega$. As $G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k}$ is atomic, we have $X_j = \sum_{k=1}^{\infty} n_{jk} \delta_{\omega_k},\ n_{jk} \sim \text{Pois}(r_k)$. Although $\{X_j\}_{1,J}$ contains countably many atoms, we will show in later sections that only a finite number of them have nonzero counts. The count matrix $\mathbf{N}_J$ is constructed by organizing all the nonzero column count vectors, $\{\boldsymbol{n}_{:k}\}_{k:n_{\cdot k}>0}$, in an arbitrary order into a random count matrix. Thus the statistical features we care about, such as words or species, are identified with the atoms of the underlying random measure.

Some important distributions. The notation $u \sim \text{Log}(p)$ denotes a random variable having a logarithmic distribution (Quenouille, 1949) with PMF

$$f_U(u \mid p) = \frac{1}{-\ln(1-p)} \frac{p^u}{u} \quad \text{for } u \in \{1, 2, \ldots\}\,.$$

A related distribution, called the sum-logarithmic, is defined as follows. Let $u_t \sim \text{Log}(p)$, and let $n = \sum_{t=1}^{l} u_t$. The marginal distribution of n is a sum-logarithmic distribution (Zhou and Carin, 2015), expressed as $n \sim \text{SumLog}(l, p)$, with PMF

$$f_N(n \mid l, p) = \frac{p^n l!\ |s(n,l)|}{n!\ [-\ln(1-p)]^l}\,,$$

where $|s(n,l)|$ are unsigned Stirling numbers of the first kind. These are related to gamma functions by

$$\frac{\Gamma(n+r)}{\Gamma(r)} = \sum_{l=0}^{n} |s(n,l)| r^l\,. \tag{2}$$

The joint distribution of $n \sim \text{SumLog}(l, p)$ and $l \sim \text{Pois}[-r\ln(1-p)]$ is described as the

Poisson-logarithmic bivariate distribution in Zhou and Carin (2015), with PMF

$$f_{N,L}(n, l \mid r, p) = \frac{|s(n,l)|r^l}{n!} p^n (1-p)^r \,. \tag{3}$$

The marginalization of l from this compound Poisson representation leads to the negative binomial distribution $n \sim \mathrm{NB}(r, p)$, with PMF

$$f_N(n \mid r, p) = \frac{\Gamma(n+r)}{n!\Gamma(r)} p^n (1-p)^r \,.$$

We describe in Appendix D several other useful distributions, including the logarithmic mixed sum-logarithmic (LogLog) distribution, the negative binomial mixed sum-logarithmic distribution, the gamma-negative binomial (GNB) distribution, the beta-negative binomial (BNB) distribution, the digamma distribution, and the logbeta distribution.

2 Nonparametric Priors for Random Count Matrices

In this section, we introduce three nonparametric Bayesian priors for random count matrices; for the gamma-Poisson process, we describe in detail its PMF, row- and column-wise construction, and some other basic properties; and for the GNBP and BNBP, we present their PMFs and defer other details to the Appendix. We then describe the predictive distribution of a new row count vector under a random count matrix, and highlight some important differences among the three priors. Although results here are quoted without proof, and the detailed construction is deferred to the Appendix, the basic manner of argument in each case is similar. Our goal is to marginalize out the infinite-dimensional random measure to obtain the unconditional PMF of the random count matrix $\mathbf{N}_J \in \mathbb{Z}^{J \times K_J}$, where $\mathbb{Z} = \{0, 1, \ldots\}$. We are able to do so by separating the absolutely continuous and discrete components of the underlying random measure, and applying a result for Poisson processes known as the Palm formula (e.g. Daley and Vere-Jones, 1988, James, 2002), together with combinatorics. This is a very general approach, which can also be employed to derive the PMF of the Indian buffet process random binary matrix using the beta-Bernoulli process.

2.1 The gamma-Poisson or negative binomial process

Let $\mathbf{N}_J \sim \mathrm{NBPM}(\gamma_0, c)$ denote a gamma-Poisson or negative binomial process (NBP) random count matrix, parameterized by a mass parameter γ_0 and a concentration parameter c. This prior arises from marginalizing out the gamma process $G \sim \Gamma\mathrm{P}(G_0, 1/c)$ from J conditionally independent Poisson process draws $X_j \mid G \sim \mathrm{PP}(G)$, with the rows of $\mathbf{N}_J$ corresponding to the X_j's and the columns of $\mathbf{N}_J$ corresponding to the atoms with at least one nonzero count.

2.1.1 Conditional likelihood

As $\{X_j\}_{1,J}$ are i.i.d. given G, they are exchangeable according to de Fennetti's theorem. With a draw from the gamma-Poisson process expressed as $X_j = \sum_{k=1}^{\infty} n_{jk}\delta_{\omega_k},\ n_{jk} \sim \mathrm{Pois}(r_k)$, where $r_k = G(\omega_k)$ is the weight of the atom ω_k of the gamma process $G \sim \Gamma\mathrm{P}(G_0, 1/c)$, we may write the likelihood of $\{X_j\}_{1,J}$, given G, as

$$p(\{X_j\}_{1,J} \mid G) = \prod_{k=1}^{\infty} \frac{r_k^{n_{\cdot k}}}{\prod_{j=1}^{J} n_{jk}!} e^{-Jr_k} = \left\{ \prod_{k:n_{\cdot k}>0} \frac{r_k^{n_{\cdot k}}}{\prod_{j=1}^{J} n_{jk}!} e^{-Jr_k} \right\} \cdot \left\{ \prod_{k:n_{\cdot k}=0} e^{-Jr_k} \right\},$$

where $n_{\cdot k} = \sum_{j=1}^{J} n_{jk}$. Let $\mathcal{D}_J = \{\omega_k\}_{k:n_{\cdot k}>0}$ denote the set of all observed atoms with nonzero counts, and let $K_J = |\mathcal{D}_J|$. Our goal is to marginalize out the random measure G to obtain the unconditional PMF of the random count matrix $\mathbf{N}_J \in \mathbb{Z}^{J\times K_J}$, where $\mathbb{Z} = \{0, 1, \ldots\}$, and to show that this "feature count" matrix is row-column exchangeable. The rows correspond to the X_j's, and the K_J columns represent those atoms in Ω with at least one nonzero count across the X_j's. Representing the infinite dimensional X_j's as a finite random matrix brings interesting combinatorial questions that need to be carefully addressed.

Fix an arbitrary labeling of the indices of the atoms in $\mathcal{D}_J$ from 1 to K_J. We now appeal to the definition of a gamma process and rewrite the conditional likelihood of $\{X_j\}_{1,J}$ as

$$p(\{X_j\}_{1,J} \mid G) = e^{-JG(\Omega\backslash\mathcal{D}_J)} \prod_{k=1}^{K_J} \frac{r_k^{n_{\cdot k}} e^{-Jr_k}}{\prod_{j=1}^{J} n_{jk}!}, \tag{4}$$

where $G(\Omega\backslash\mathcal{D}_J) := \sum_{k:n_k=0} r_k$ is the total mass of the rest of the (absolutely continuous) space. The idea is to first marginalize out G from (4) to obtain the marginal distribution

$p(\{X_j\}_{1,J} \mid \gamma_0, c)$, whose derivation using the Palm formula is provided in the Appendix, and then use combinatorial argument to find the marginal distribution of the random count matrix $\mathbf{N}_J$ organized from $\{X_j\}_{1,J}$.

2.1.2 Marginal distribution and combinatorial analysis

One of our main results is that the PMF of $\mathbf{N}_J \sim \text{NBPM}(\gamma_0, c)$, with J rows and a random K_J number of columns, is

$$f(\mathbf{N}_J \mid \gamma_0, c) = \frac{p(\{X_j\}_{1,J} \mid \gamma_0, c)}{K_J!} = \frac{\gamma_0^{K_J} \exp\left[-\gamma_0 \ln(\frac{J+c}{c})\right]}{K_J!} \prod_{k=1}^{K_J} \frac{\frac{\Gamma(n_{\cdot k})}{(J+c)^{n_{\cdot k}}}}{\prod_{j=1}^{J} n_{jk}!}, \tag{5}$$

where the unordered column vectors $\{\boldsymbol{n}_{:k}\}_{1,K_J}$ of the count matrix $\mathbf{N}_J$ represent a draw from the underlying stochastic process, and the normalization constant of $1/K_J!$ arises from the fact that the mapping from a realization of $\{X_j\}_{1,J}$ to $\mathbf{N}_J$ is one-to-many, with $K_J!$ distinct column orderings.

By construction, the rows of a NBP random count matrix are exchangeable. Moreover, one may verify by direct calculation that a NBP random count matrix with PMF (5) can be generated column by column as i.i.d. count vectors:

$$\begin{aligned} &\boldsymbol{n}_{:k} \sim \text{Multinomial}(n_{\cdot k}, 1/J, \ldots, 1/J), \\ &n_{\cdot k} \sim \text{Log}[J/(J+c)], \\ &K_J \sim \text{Pois}\left\{\gamma_0 \left[\ln(J+c) - \ln(c)\right]\right\}. \end{aligned} \tag{6}$$

It is clear from (6) that the columns of $\mathbf{N}_J$ are independent multivariate count vectors, which all follow the same logarithmic-multinomial (mixture) distribution. Thus the NBP random count matrix $\mathbf{N}_J$ is row-column exchangeable (see, e.g. Hoover, 1982, Aldous, 1985, Orbanz and Roy, 2014, for a general treatment of row-column exchangeable matrices).

Now consider the row-wise sequential construction of the NBP random matrix, recalling that $\mathbf{N}_{J+1}^{+}$ represents the "new" part of the matrix added by the new row. With the prior

on $\mathbf{N}_J \in \mathbb{Z}^{J\times K_J}$ well defined, one may construct $\mathbf{N}_J$ in a sequential manner as

$$f(\mathbf{N}_J \mid \boldsymbol{\theta}) = f(\mathbf{N}_1 \mid \boldsymbol{\theta}) \frac{f(\mathbf{N}_2 \mid \boldsymbol{\theta})}{f(\mathbf{N}_1 \mid \boldsymbol{\theta})} \cdots \frac{f(\mathbf{N}_J \mid \boldsymbol{\theta})}{f(\mathbf{N}_{J-1} \mid \boldsymbol{\theta})},$$

where $\boldsymbol{\theta} := \{\gamma_0, c\}$ and $p(\mathbf{N}^+_{j+1} | \mathbf{N}_j, \boldsymbol{\theta}) := f(\mathbf{N}_{j+1} | \boldsymbol{\theta}) / f(\mathbf{N}_j | \boldsymbol{\theta})$ is the prediction rule to add the new part brought by row $(j+1)$ into the matrix $\mathbf{N}_j$. Direct calculations using (6) yield the following form for this prediction rule, expressed in terms of familiar PMFs:

$$\begin{aligned} p(\mathbf{N}^+_{J+1} \mid \mathbf{N}_J, \boldsymbol{\theta}) = \frac{K_J! K^+_{J+1}!}{K_{J+1}!} \prod_{k=1}^{K_J} \mathrm{NB}\left(n_{(J+1)k}; n_{\cdot k}, \frac{1}{J+c+1}\right) \\ \times \prod_{k=K_J+1}^{K_{J+1}} \mathrm{Log}\left(n_{(J+1)k}; \frac{1}{J+c+1}\right) \\ \times \mathrm{Pois}\left\{K^+_{J+1}; \gamma_0 \left[\ln(J+c+1) - \ln(J+c)\right]\right\}. \end{aligned} \tag{7}$$

This formula says that to add a new row to $\mathbf{N}_J \in \mathbb{Z}^{J\times K_J}$, we first draw count $\mathrm{NB}[n_{\cdot k}, 1/(J+c+1)]$ at each existing column. We then draw K^+_{J+1} new columns as $K^+_{J+1} \sim \mathrm{Pois}\{\gamma_0[\ln(J+c+1) - \ln(J+c)]\}$. Finally, each entry in the new columns has a $\mathrm{Log}[1/(J+c+1)]$ distributed random count; crucially, new columns brought by the new row must have positive counts.

The normalizing constant $(K_J!\ K^+_{J+1}!)/K_{J+1}!$ in (7) plays a key role in our combinatorial analysis, and will appear again in both the gamma- and beta- negative binomial processes. It emerges directly from the calculations, and can also be interpreted in the following way. After drawing K^+_{J+1} new columns, we must insert them into the original K_J columns while keeping the relative orders of both the original and new columns unchanged. This is a one-to-many mapping, with the number of such order-preserving insertions given by the binomial coefficient. For example, if the original $\mathbf{N}_J$ has two columns and the new row $J+1$ introduces two more columns, then we construct $\mathbf{N}_{J+1}$ by rearranging the two old columns 1 and 2 and the two new columns iii and iv in one of $\binom{4}{2} = 6$ possible ways: (1 2 iii iv), (1 iii 2 iv), (iii 1 2 iv), (1 iii iv 2), (iii 1 iv 2), and (iii iv 1 2), where (1 2 iii iv) represents the construction appending the new columns to the right of the original matrix.

It is instructive to compare (6), which generates a NBP random matrix by drawing all its columns at once, with (7), which generates an identically distributed random matrix one

row at a time. The matrix generated with (6) has i.i.d. columns. The matrix generated with (7) adds K_{J+1}^+ new columns when it adds the $(J+1)$th row, and if the newly added columns are inserted into random locations among original columns with their relative orders preserved, then we arrive at an identically distributed column-i.i.d. random count matrix. If the newly added columns are inserted in a particular way, then the distribution of the generated random matrix would be different up to a multinomial coefficient. For example, if we generate row vectors $\boldsymbol{n}_j$ from $j = 1$ to $j = J$ and each time we append the new columns to the right of the original matrix, then this ordered matrix $\widetilde{\mathbf{N}}_J$ will appear with probability

$$f(\widetilde{\mathbf{N}}_J \mid \boldsymbol{\theta}) = f(\mathbf{N}_1 \mid \boldsymbol{\theta}) \prod_{j=1}^{J-1} p(\mathbf{N}_{j+1}^+ \mid \mathbf{N}_j, \boldsymbol{\theta}) \frac{K_{j+1}!}{K_j! K_{j+1}^+!} = \binom{K_J}{K_1^+, \ldots, K_J^+} f(\mathbf{N}_J \mid \boldsymbol{\theta}). \tag{8}$$

Shown in the first row of Figure 1 are three NBP random count matrices simulated in this manner. We note that the gamma-Poisson process is related to the model of Lo (1982), as well as the model of Titsias (2008), which can be considered as a special case of the NBP with the concentration parameter c fixed at one.

2.1.3 Inference for parameters

Although the marginal likelihood alone is not amenable to posterior analysis, the NBP parameters can be conveniently inferred using both the conditional and marginal likelihoods. To complete the model, we let $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$ and $c \sim \text{Gamma}(c_0, 1/d_0)$. With (4), (5) and $G(\Omega) := G(\Omega \backslash \mathcal{D}_J) + \sum_{k=1}^{K_J} r_k$, we sample the parameters in closed form as

$$\begin{aligned}
&(\gamma_0 \mid -) \sim \text{Gamma}\bigg(e_0 + K_J, \frac{1}{f_0 - \ln(\frac{c}{c+J})}\bigg), \\
&(r_k \mid -) \sim \text{Gamma}\big(n_{\cdot k}, 1/(c+J)\big), \\
&\{G(\Omega \backslash \mathcal{D}_J) \mid -\} \sim \text{Gamma}\big(\gamma_0, 1/(c+J)\big), \\
&(c \mid -) \sim \text{Gamma}\big(c_0 + \gamma_0, 1/[d_0 + G(\Omega)]\big).
\end{aligned} \tag{9}$$

Similar strategies will be used to infer the parameters of the other two stochastic processes. Having closed-form update equations for parameter inference via Gibbs sampling is a unique

feature shared by all the nonparametric Bayesian priors proposed in this paper.

2.2 The gamma-negative binomial process

Let $\mathbf{N}_J \sim \text{GNBPM}(\gamma_0, c, p_1, \ldots, p_J)$ denote a gamma-negative binomial process (GNBP) random count matrix, parameterized by a mass parameter γ_0, a concentration parameter c, and J row-specific probability parameters $\{p_j\}_{1,J}$. This random count matrix is the direct outcome of marginalizing out the gamma process $G \sim \Gamma\text{P}(G_0, 1/c)$, with data augmentation, from J conditionally independent negative binomial process draws $X_j \mid G \sim \text{NBP}(G, p_j)$, which are defined such that $X_j(A) \sim \text{NB}\left(G(A), p_j\right)$ for each $A \subset \Omega$.

As directly marginalizing out the gamma process under negative binomial sampling is difficult, our construction is based on the compound-Poisson representation of the negative binomial, described in Section 1.3. Specifically, consider the joint distribution of $\mathbf{N}_J$ and a latent count matrix $\mathbf{L}_J$, whose dimension and locations of nonzero counts are the same as those of $\mathbf{N}_J$. These two matrices parallel the scalar n and l given in the joint PMF of the Poisson-logarithmic distribution (3). This joint distribution is defined as

$$f(\mathbf{N}_J, \mathbf{L}_J \mid \boldsymbol{\theta}) = \frac{\gamma_0^{K_J} \exp\left[-\gamma_0 \ln(\frac{c+q_\cdot}{c})\right]}{K_J!} \prod_{k=1}^{K_J} \frac{\Gamma(l_{\cdot k})}{(c+q_\cdot)^{l_{\cdot k}}} \left(\prod_{j=1}^{J} \frac{|s(n_{jk}, l_{jk})| p_j^{n_{jk}}}{n_{jk}!} \right), \qquad (10)$$

where $\boldsymbol{\theta} := \{\gamma_0, c, p_1, \ldots, p_J\}$, $q_j := -\ln(1-p_j)$ and $q_\cdot := \sum_{j=1}^{J} q_j$. The detailed derivation is in the Appendix.

Similar to the analysis in Section 2.1 for the NBP, we show in the Appendix that the GNBP random count matrix can be constructed by either drawing its i.i.d. columns at once or adding one row at a time, and it has closed-form Gibbs sampling update equations for model parameters. Different from the NBP random count matrix that is row-column exchangeable, the GNBP random count matrix no longer maintains row exchangeability if its row-wise probability parameters p_j are set differently for different rows.

Shown in the second row of Figure 1 are three sequentially constructed GNBP random count matrices, with the new columns introduced by each row appended to the right of the matrix. Similar to the combinatorial arguments that lead to (8), this particularly structured

matrix and its auxiliary matrix appear with probability $\binom{K_J}{K_1^+,\ldots,K_J^+} f(\mathbf{N}_J, \mathbf{L}_J \,|\, \boldsymbol{\theta})$.

2.3 The beta-negative binomial process

Let $\mathbf{N}_J \sim \mathrm{BNBPM}(\gamma_0, c, r_1, \ldots, r_J)$ denote a beta-negative binomial process (BNBP) random count matrix, parameterized by a mass parameter γ_0, a concentration parameter c, and J row-specific dispersion parameters $\{r_j\}_{1,J}$, whose PMF is defined as

$$f(\mathbf{N}_J \mid \boldsymbol{\theta}) = \frac{\gamma_0^{K_J} \exp\left\{-\gamma_0 \left[\psi(c + r_\cdot) - \psi(c)\right]\right\}}{K_J!} \prod_{k=1}^{K_J} \frac{\Gamma(n_{\cdot k})\Gamma(c + r_\cdot)}{\Gamma(c + n_{\cdot k} + r_\cdot)} \prod_{j=1}^{J} \frac{\Gamma(n_{jk} + r_j)}{n_{jk}!\Gamma(r_j)}, \quad (11)$$

where $\boldsymbol{\theta} := \{\gamma_0, c, r_1, \ldots, r_J\}$. The PMF is the direct outcome of marginalizing out the beta process $B \sim \mathrm{BP}(c, B_0)$ from J conditionally independent negative binomial process draws $X_j \,|\, B \sim \mathrm{NBP}(r_j, B)$, which are defined such that $X_j(A) = \sum_{k:\omega_k \in A} n_{jk},\ n_{jk} \sim \mathrm{NB}(r_j, p_k)$ for each $A \subset \Omega$, where $p_k = B(\omega_k)$ is the weight of atom k. The detailed derivation is provided in the Appendix.

Similar to the analysis in Section 2.1 for the NBP, we show in the Appendix that the BNBP random count matrix can be constructed by either drawing its i.i.d. columns at once or adding one row at a time using an "ice cream" buffet process, and it has closed-form Gibbs sampling update equations for all model parameters except for the concentration parameter c. The BNBP random count matrix no longer maintains row exchangeability if its row-wise dispersion parameters r_j are set differently for different rows.

Shown in the last row of Figure 1 are three sequentially constructed BNBP random count matrices, with the new columns introduced by each row appended to the right of the matrix. Similar to the combinatorial arguments that lead to (8), this particularly structured matrix appears with probability $\binom{K_J}{K_1^+,\ldots,K_J^+} f(\mathbf{N}_J \,|\, \boldsymbol{\theta})$.

2.4 The predictive distribution of a new row count vector

It is critical to note that the prediction rule $p(\mathbf{N}_{J+1}^+ \,|\, \mathbf{N}_J, \boldsymbol{\theta})$ of the NBP shown in (7) is for sequentially constructing a column-i.i.d. random count matrix, but it is not the predictive distribution for a new row count vector. The $1 \times K_J$ submatrix of $\mathbf{N}_{J+1}^+$ orders its column

in the same way as $\mathbf{N}_J$ does, and the $(J+1) \times K^+_{J+1}$ submatrix of $\mathbf{N}^+_{J+1}$ also maintains a certain order of its columns; however, the indexing of these K^+_{J+1} columns are in fact arbitrarily chosen from $K^+_{J+1}!$ possible permutations. Therefore, the predictive distribution of a row vector $\boldsymbol{n}_{J+1}$ that brings K^+_{J+1} new columns shall be

$$p(\boldsymbol{n}_{J+1} \mid \mathbf{N}_J, \boldsymbol{\theta}) = \frac{p(\mathbf{N}^+_{J+1} \mid \mathbf{N}_J, \boldsymbol{\theta})}{K^+_{J+1}!} \tag{12}$$

$$= \frac{K_J!}{K_{J+1}!} \frac{\frac{K_{J+1}!}{K_J! K^+_{J+1}!} f(\mathbf{N}_{J+1} \mid \boldsymbol{\theta})}{f(\mathbf{N}_J \mid \boldsymbol{\theta})} \,. \tag{13}$$

The normalizing constant $1/K^+_{J+1}!$ in (12) arises because a realization of $\mathbf{N}^+_{J+1}$ to $\boldsymbol{n}_{J+1}$ is one-to-many, with $K^+_{J+1}!$ distinct orderings of these new columns brought by the $(J+1)$th row. Our experimental results show that omitting this normalizing term may significantly deteriorate the out-of-sample prediction performance.

An equivalent representation in (13) shows that one may first consider the distribution of a matrix constructed by appending the new columns brought by $\boldsymbol{n}_{J+1}$ to the right of $\mathbf{N}_J$, which is $\frac{K_{J+1}!}{K_J! K^+_{J+1}!} f(\mathbf{N}_{J+1} \,|\, \boldsymbol{\theta})$, and then apply the Bayes' rule to derive the conditional distribution of this particularly ordered $\boldsymbol{n}_{J+1}$ given $\mathbf{N}_J$. The normalizing constant $K_J!/K_{J+1}!$ in (13) can be interpreted in the following way. We need to insert the K^+_{J+1} new columns one by one into the original matrix. The first, second, ..., and last new columns can choose from $K_J + 1$, $K_J + 2$, ..., and $K_J + K^+_{J+1}$ possible locations, respectively, thus there are $\prod_{i=1}^{K^+_{J+1}} (K_J + i)! = K_{J+1}!/K_J!$ ways to insert the K^+_{J+1} new columns into the original ordered K_J columns, which is again a one-to-many mapping. The same combinatorial analysis applies to both the GNBP and BNBP. For the GNBP, to compute the predictive likelihood of $\boldsymbol{n}_{J+1}$, one will need to take extra care as the computation involves $\mathbf{L}_J$, an auxiliary random count matrix that is not directly observable. In Section 3, we will discuss in detail how to compute the predictive likelihood via Monte Carlo integration.

2.5 Comparison

In the appendix, we provide further details on the construction of random count matrices from the negative binomial process, as well as those derived from two further generaliza-

Table 1: Comparison of the prediction rules of the NBP, GNBP, and BNBP random count matrices.

Model	Number of new columns K_{J+1}^{+}	Counts in existing columns	Counts in new columns
NBP	$\text{Pois}\left\{\gamma_0[\ln(J+c+1)-\ln(J+c)]\right\}$	$\text{NB}\left[n_{\cdot k}, 1/(J+c+1)\right]$	$\text{Log}\left[1/(J+c+1)\right]$
GNBP	$\text{Pois}\left\{\gamma_0\left[\ln(c+q_{\cdot}+q_{J+1})-\ln(c+q_{\cdot})\right]\right\}$	$\text{GNB}\left(l_{\cdot k}, c+q_{\cdot}, p_{J+1}\right)$	$\text{LogLog}\left(c+q_{\cdot}, p_{J+1}\right)$
BNBP	$\text{Pois}\left\{\gamma_0\left[\psi(c+r_{\cdot}+r_{J+1})-\psi(c+r_{\cdot})\right]\right\}$	$\text{BNB}(r_{J+1}, n_{\cdot k}, c+r_{\cdot})$	$\text{Digam}(r_{J+1}, c+r_{\cdot})$

tions: the gamma-negative binomial process (GNBP) and the beta-negative binomial process (BNBP). While the PMFs for all three proposed nonparametric priors are complicated, their relationship and differences become evident once we show that they all govern random count matrices with a Poisson-distributed number of i.i.d. columns. Table 1 shows the differences among the three priors' row-wise sequential construction, and the following list shows the variance-mean relationship for each prior for the counts at existing columns. Together, these provide additional insights on how the priors differ from each other.

$$\text{NBP:}\quad \text{Var}[n_{(J+1)k}] = \mathbb{E}[n_{(J+1)k}] + \frac{\mathbb{E}^2[n_{(J+1)k}]}{n_{\cdot k}} \tag{14}$$

$$\text{GNBP:}\quad \text{Var}[n_{(J+1)k}] = \frac{\mathbb{E}[n_{(J+1)k}]}{1-p_{J+1}} + \frac{\mathbb{E}^2[n_{(J+1)k}]}{l_{\cdot k}} \tag{15}$$

$$\text{BNBP:}\quad \text{Var}[n_{(J+1)k}] = \frac{\mathbb{E}[n_{(J+1)k}]}{\frac{c+r_{\cdot}}{n_{\cdot k}+c+r_{\cdot}-1}} + \frac{\mathbb{E}^2[n_{(J+1)k}]}{\frac{n_{\cdot k}(c+r_{\cdot}-2)}{n_{\cdot k}+c+r_{\cdot}-1}} \tag{16}$$

The NBP can be used to generate a row-column exchangeable random count matrix with a potentially unbounded number of columns. However, as shown in (6), to model the total count of a column $n_{\cdot k}$, the NBP uses the logarithmic distribution, which has only one free parameter, always has the mode at one, and monotonically decreases. In addition, each column sum $n_{\cdot k}$ is assigned to the J rows with a multinomial distribution that has a uniform probability vector $(1/J, \ldots, 1/J)$. Furthermore, as shown in Table 1, for out-of-sample prediction, it models counts at existing columns using $\text{NB}\left[n_{(J+1)k}; n_{\cdot k}, 1/(J+c+1)\right]$, whose variance-mean relationship (14) may be restrictive in modeling highly overdispersed counts. Finally, the expected number of new columns brought by a row, equal to $\gamma_0 \ln[1+1/(J+c)]$,

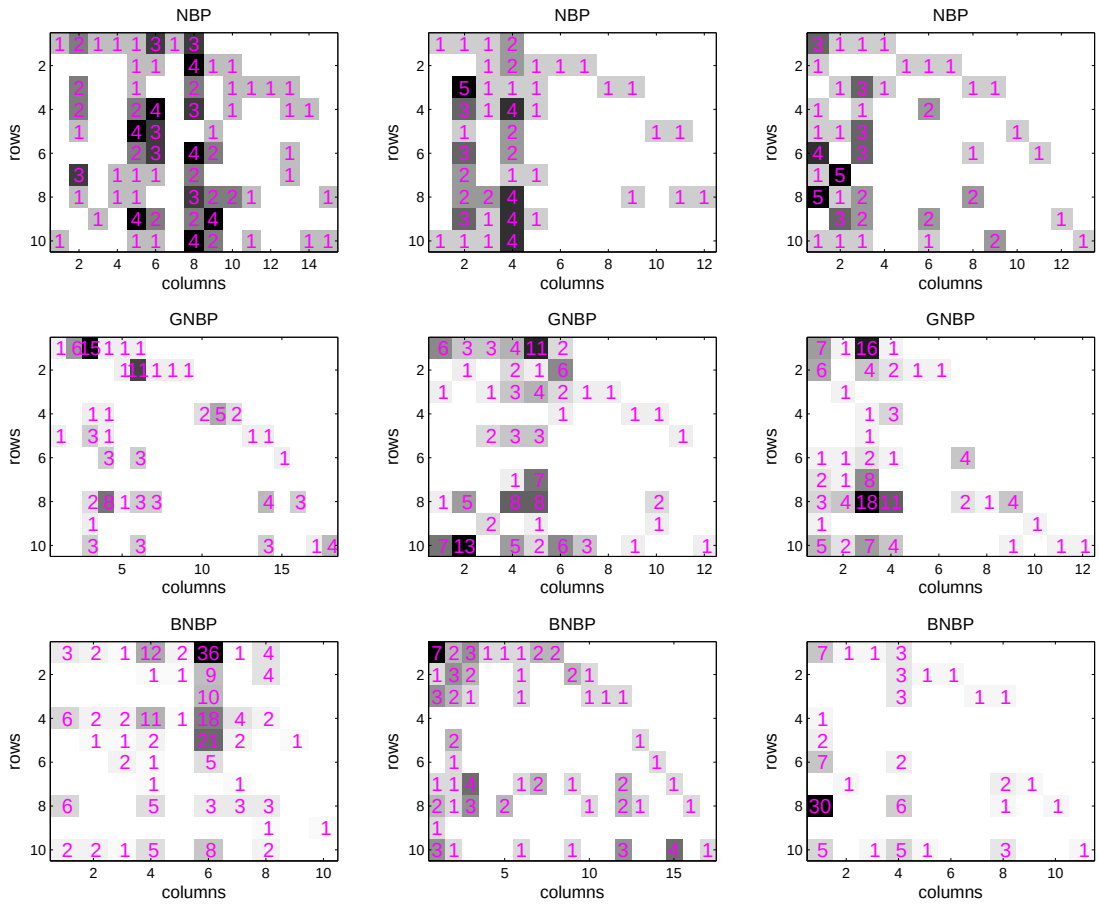

Figure 1: Sequentially constructed negative binomial process (NBP), gamma-negative binomial process (GNBP), and beta-negative binomial process (BNBP) random count matrices (the blank cells indicate zero counts). The ten rows of each matrix are added one by one, with the new columns introduced by each row appended to the right of the matrix. To make the expected total count of a random matrix as 100 and the expected number of columns approximately as 12, the parameters are set as $\gamma_0 = 5$ and $c = 0.5$ for the NBP, set as $c = 1$, $\gamma_0 = 4.79$, and $\sum_j \frac{p_j}{1-p_j} = 20.88$ for the GNBP, and set as $c = 2$, $\gamma_0 = 4.31$, and $\sum_j r_j = 23.20$ for the BNBP. The randomized row wise parameters $[p_1/(1-p_1), \ldots, p_J/(1-p_J)]^T$ and $(r_1, \ldots, r_J)^T$ are generated via $\mathrm{Dir}(1, \ldots, 1) \sum_j \frac{p_j}{1-p_j}$ and $\mathrm{Dir}(1, \ldots, 1) \sum_j r_j$, respectively.

monotonically decreases. These constraints limit the potential use of the NBP model.

Both the GNBP and BNBP relax these constraints in their own unique ways. Examining the sequential construction of the GNBP helps us understand the advantages of the GNBP over the NBP. As shown in Table 1, to model the likelihood of a new row count vector, one may find that the GNBP employs the three-parameter GNB instead of the two-parameter negative binomial distribution to model the count at an existing column, and employs the

two-parameter LogLog instead of the logarithmic distribution to model the count at a new column. As the GNB random variable $n_{(J+1)k} \sim \text{GNB}\left(l_{\cdot k}, c + q_{\cdot}, p_{J+1}\right)$ can be generated as $n_{(J+1)k} \sim \text{NB}(r_{(J+1)k}, p_{J+1})$, $r_{(J+1)k} \sim \text{Gamma}\left[l_{\cdot k}, 1/(c + q_{\cdot})\right]$, using the laws of total expectation and total variance, we express $\text{Var}[n_{(J+1)k}]$ in terms of $\mathbb{E}[n_{(J+1)k}]$ in (15). Since $p_{J+1} < 1$ and $l_{\cdot k} \leq n_{\cdot k}$, the GNBP can model much more overdispersed counts than the NBP. Moreover, the GNBP allows each row count vector to have its own probability parameter, allowing finer control on the expected number of new columns brought by a new row, which is $\gamma_0 \ln[1+q_{J+1}/(c + q_{\cdot})]$. The NBP random count matrix is row-column exchangeable, whereas the GNBP random count matrix is column exchangeable, but not row exchangeable if the row-wise probability parameters p_j are fixed at different values.

As shown in Table 1, to model the likelihood of a new row count vector, one may find that the BNBP employs the three-parameter BNB instead of the two-parameter negative binomial distribution to model the count at an existing column, and employs the two-parameter digamma instead of the logarithmic distribution to model the count at a new column. Note that the BNB random variable $n_{(J+1)k} \sim \text{BNB}(r_{J+1}, n_{\cdot k}, c+r_{\cdot})$ can be generated as $n_{(J+1)k} \sim \text{NB}(r_{J+1}, p_{(J+1)k})$, $p_{(J+1)k} \sim \text{Beta}\left(n_{\cdot k}, c + r_{\cdot}\right)$, using the laws of total expectation and total variance, for $c+r_{\cdot} > 2$, we express $\text{Var}[n_{(J+1)k}]$ in terms of $\mathbb{E}[n_{(J+1)k}]$ in (16). As $\frac{c+r_{\cdot}}{n_{\cdot k}+c+r_{\cdot}-1} \leq 1$ and $\frac{n_{\cdot k}(c+r_{\cdot}-2)}{n_{\cdot k}+c+r_{\cdot}-1} < n_{\cdot k}$ for $c+r_{\cdot} > 2$, the BNBP can also model much more overdispersed counts than the NBP. Moreover, the BNBP allows each row count vector to have its own dispersion parameter, allowing finer control on the expected number of new columns brought by a row, which is $\gamma_0[\psi(c + r_{\cdot} + r_{J+1}) - \psi(c + r_{\cdot})]$; the NBP random count matrix is row-column exchangeable, whereas the BNBP random count matrix is column exchangeable, but not row exchangeable if the row-wise dispersion parameters r_j are different.

The variance-mean relationships expressed by (14)-(16) show that the GNBP and BNBP can model much more overdispersed counts than the NBP. This fact is borne out by the simulated random count matrices in Figure 1, which provide some intuition for the practical differences among the models. The parameters for the three priors have been chosen so that each random matrix has the same expected total count. Yet the counts in the NBP random count matrices have small dynamic ranges, whereas the counts in both the GNBP and BNBP matrices can contain values that are significantly above the average.

2.6 Parameter inference

An appealing feature of all three negative binomial process random count matrix priors is that their parameters can be inferred with closed-form Gibbs sampling update equations, by exploiting both the conditional and marginal distributions, together with the data augmentation and marginalization techniques unique to the negative binomial distribution. Parameter inference for the NBP is provided in Section 2.1.3. The details of parameter inference for both the GNBP and BNBP are provided in the Appendix.

3 Negative Binomial Process Naive Bayes Classifiers

3.1 Background

Given a random count matrix, finding the predictive distribution of a row count vector, which may bring additional columns, involves interesting and challenging combinatory arguments that have been throughly addressed in this paper. With these combinatorial structures carefully analyzed, we are ready to construct a NBP, a GNBP, and a BNBP naive Bayes classifiers. We do so as follows. First, for each category, the training row count vectors are summarized as a random count matrix $\mathbf{N}_J$, each column of which must contain at least one nonzero count (i.e. columns with all zeros are excluded). Second, Gibbs sampling is used to infer the parameters $\boldsymbol{\theta}$ that generate $\mathbf{N}_J$. To represent the posterior of $\boldsymbol{\theta}$, S MCMC samples $\{\boldsymbol{\theta}^{[s]}\}_{1,S}$ are collected. For the GNBP, a posterior MCMC sample $\mathbf{L}_J^{[s]}$ for the auxiliary random matrix is also collected when $\boldsymbol{\theta}^{[s]}$ is collected. Finally, to test a row count vector $\boldsymbol{n}_{J+1}$, its predictive likelihood given $\mathbf{N}_J$ is calculated via Monte Carlo integration using

$$p(\boldsymbol{n}_{J+1} \mid \mathbf{N}_J) = \frac{1}{S}\sum_{s=1}^{S} \frac{p(\mathbf{N}_{J+1}^{+} \mid \mathbf{N}_J, \boldsymbol{\theta}^{[s]})}{K_{J+1}^{+}!} \tag{17}$$

for both the NBP and BNBP, and using

$$p(\boldsymbol{n}_{J+1} \mid \mathbf{N}_J) = \frac{1}{S}\sum_{s=1}^{S} \frac{p(\mathbf{N}_{J+1}^{+} \mid \mathbf{N}_J, \mathbf{L}_J^{[s]}, \boldsymbol{\theta}^{[s]})}{K_{J+1}^{+}!} \tag{18}$$

for the GNBP. Although a larger S shall lead to a more accurate calculation of the predictive likelihood, the computational complexity for testing is a linear function of S. It is therefore of practical importance to find out how the value of S impacts the performance of the proposed nonparametric Bayesian naive classifiers. Below we consider experiments on document categorization, for which we will show that $S = 1$ performs essentially just as well as selecting a much larger S in terms of the categorization accuracy.

3.2 Experiment settings

We consider the example of categorizing the 18,774 documents of the 20 newsgroup dataset[1], where each bag-of-words document is represented as a word count vector under a vocabulary of size $V = 61{,}188$. We also consider the TDT2 corpus[2] (NIST Topic Detection and Tracking corpus): with the documents appearing in two or more categories removed, this subset of TDT2 consists of 9,394 documents from the largest 30 categories, with a vocabulary of size $V = 36{,}771$; this dataset was used to compare document clustering algorithms in Cai et al. (2005). We train all three negative binomial processes using 10%, 20%, ..., or 80% of the documents in each newsgroup of the 20 newsgroup dataset, and in each category of the TDT2 corpus. We then test on the remaining documents. We report our results based on five random training/testing partitions.

To make comparison to other commonly used text categorization algorithms, we also consider a default setting for the 20 newsgroup dataset: using the first 11,269 documents for training and the other 7,505 documents collected at later times for testing. For this setting, we reports our results based on five independent runs with random initializations. This allows us to compare our performance to many other papers that have proposed text classification algorithms and benchmarked their method using this same split of the 20-newsgroup dataset.

For the ith newsgroup/category with $J^{(i)}$ training documents, we construct a document-term count matrix $\mathbf{N}^{(i)}_{J^{(i)}} \in \mathbb{Z}^{J^{(i)} \times K_{J^{(i)}}}$, whose element $n^{(i)}_{jk}$ represents the number of times term k appearing in document j. Since only the terms present in the training documents of the ith category are considered, the column indices of $\mathbf{N}^{(i)}_{J^{(i)}}$ correspond to the terms that

[1] http://qwone.com/~jason/20Newsgroups/

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html

appear at least once in training. We use $x^{(i)}$ to denote that x is a parameter inferred from $\mathbf{N}^{(i)}_{J^{(i)}}$. Note that the column indices of $\mathbf{N}^{(i)}_{J^{(i)}}$ can be arbitrarily ordered, which affects neither training nor out-of-sample prediction as long as their corresponding features are recorded.

We collect S MCMC samples of model parameters and auxiliary variables to compute the predictive likelihood for a new row count vector. In this paper, we run S independent Markov chains and collect the 2500th sample of each chain. Note that one may also consider collecting S samples at a certain interval from a single Markov chain after the burn-in period. We consider non-informative hyper-parameters as $a_0 = b_0 = \ldots = f_0 = 0.001$. For the BNBP, we set $c_0 = d_0 = 1$. The document-term training count matrix of the ith newsgroup is modeled as $\mathbf{N}^{(i)}_{J^{(i)}} \sim \mathrm{NBPM}(\gamma_0^{(i)}, c^{(i)})$, $\mathbf{N}^{(i)}_{J^{(i)}} \sim \mathrm{GNBPM}\big(\gamma_0^{(i)}, c^{(i)}, p_1^{(i)}, \ldots, p_{J^{(i)}}^{(i)}\big)$, and $\mathbf{N}^{(i)}_{J^{(i)}} \sim \mathrm{BNBPM}\big(\gamma_0^{(i)}, c^{(i)}, r_1^{(i)}, \ldots, r_{J^{(i)}}^{(i)}\big)$ under the three priors respectively.

Note that we are facing typical "small n and large p" problems as the number of rows of a document-term count matrix is typically much smaller than the number of columns. For example, the first newsgroup of the 20 newsgroup dataset contains 798 documents with 12,665 unique words, which is summarized as a 798×12665 count matrix; and the 30th category of the TNT2 subset contains 52 documents with 2904 unique words, which is summarized as a 52×2904 count matrix. As the number of unique terms in a category might be significantly smaller than the vocabulary size of the whole corpus, our approach for both training and testing could be much faster than the approach that considers all the terms in the vocabulary of the corpus. In addition, our approach provides a principled, model-based way to handle terms that appear in a testing document but not in the training documents. By contrast, many traditional approaches have to discard these terms not present in training.

3.3 Training and posterior predictive checking

We train the NBP, GNBP, and BNBP with the document-term word count matrix $\mathbf{N} \in \mathbb{Z}^{52\times 2904}$ that summarizes all the 52 documents in the 30th category of the TDT2 subset. We then run 2500 MCMC iterations and collect the last 1500 samples to infer the posterior means of the parameters in $\mathbf{N} \sim \mathrm{NBPM}(\gamma_0, c)$, $\mathbf{N} \sim \mathrm{GNBPM}(\gamma_0, c, p_1, \ldots, p_{52})$, and $\mathbf{N} \sim \mathrm{BNBPM}(\gamma_0, c, r_1, \ldots, r_{52})$. Using the corresponding parameters learned from the training count matrix, we regenerate a NBP, a GBNP, and a BNBP random count matrix as an

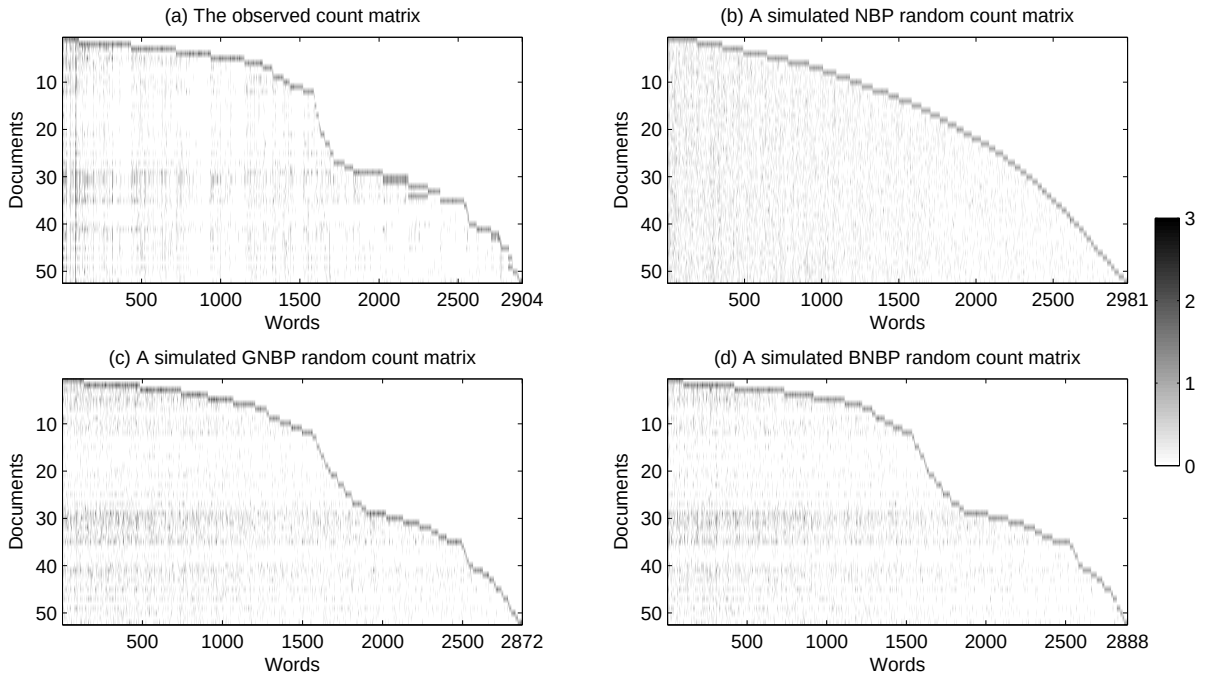

Figure 2: The parameters of the negative binomial processes are inferred using (a) the observed document-term count matrix. These parameters are used to simulate (b) a NBP random count matrix, (c) a GNBP random count matrix, and (d) a BNBP random count matrix. These matrices are visualized by arranging the new columns brought by each new row to the right of the original matrix. The counts larger than 3 are displayed as 3.

informal posterior predictive check on the model. The observed count matrix is shown in Figure 2 (a), and the three simulated random count matrices are shown in Figure 2 (b)-(d). These matrices are displayed by arranging the new columns brought by a new row to the right of the original matrix.

It is clear that the NBP is restrictive, in that the generated random matrix looks the least similar to the observed count matrix. This is unsurprising, as the NBP has a limited ability to model highly overdispersed counts, does not model row-heterogeneity, and can barely adjust the number of new columns brought by a row. On the other hand, both the generated GNBP and BNBP random count matrices resemble the original count matrix much more closely. This is expected, since both priors use heavy-tailed count distribution to model highly overdispersed counts, and have row-wise probability or dispersion parameters to model row-heterogeneity and to control the number of new columns brought by each row. Note that the observed matrix has 2904 columns, but each of the generated random count

matrices has a different (random) number of columns. This is because there are one-to-one correspondences between their row indices, but not their column indices.

3.4 Out-of-sample prediction and categorization for count vectors

For out-of-sample prediction on a new row vector, we first compute that vector's likelihood under different categories' training count matrices. We then use these likelihoods in a naive-Bayes classifier to categorize the new vector. For example, for testing row count vector $\boldsymbol{n}_{j'}$ under category i, we will first match the column indices (features) of this row count vector to those of the training count matrix $\mathbf{N}^{(i)}_{J^{(i)}}$; each feature that belongs to one of the $K^{(i)}_{J^{(i)}}$ features of $\mathbf{N}^{(i)}_{J^{(i)}}$ but not present in $\boldsymbol{n}_{j'}$ will be assigned a zero count; and the $K^{+(i)}_{j'}$ features that are present in vector j' but not in $\mathbf{N}^{(i)}_{J^{(i)}}$ will be treated as new features brought by vector j' to to $\mathbf{N}^{(i)}_{J^{(i)}}$. For the the GNBP, we first find an estimate of $p^{(i)}_{j'}$ as $p^{(i)}_{j'} = (a_0 + n^{(i)}_{j'\cdot})/[a_0 + b_0 + n^{(i)}_{j'\cdot} + G^{(i)}(\Omega)]$. For the BNBP, we first find an expectation-maximization estimate of $r_{j'}$ by running the updates

$$
\begin{aligned}
l^{(i)}_{j'k} &= r^{(i)}_{j'}\big[\psi(r^{(i)}_{j'} + n^{(i)}_{j'k}) - \psi(r^{(i)}_{j'})\big], \\
r^{(i)}_{j'} &= \frac{a_0 - 1 + l^{(i)}_{j'\cdot}}{b_0 + p^{(i)}_{*} - \sum_{k=1}^{K_{J^{(i)}}} \ln(1 - p^{(i)}_{k})}
\end{aligned}
$$

iteratively for 20 iterations, where for a testing row vector with all zeros, we let $l^{(i)}_{j'\cdot} = 1$. Given the column sums of $\mathbf{N}^{(i)}$ and the inferred model parameters (together with auxiliary variables for the GBNB), the predictive likelihoods of a new row count vector are calculated using (17) for both the NBP and BNBP and with (18) for the GNBP.

Note that when the predictive distributions are used to calculate the likelihoods, the models are not constrained under a predetermined vocabulary. But if we are given a vocabulary of size V that includes all the important terms, exploiting that information might further improve the performance. Thus to test document j', we also consider using

$$
p(\boldsymbol{n}_{j'} \mid \mathbf{N}^{(i)}_{J^{(i)}}, \boldsymbol{\theta}^{(i)}) = \prod_{v=1}^{V} \mathrm{NB}\left[n_{j'v}; n^{(i)}_{\cdot v} + \gamma^{(i)}_0/V, 1/(J^{(i)} + c^{(i)} + 1)\right] \tag{19}
$$

as the likelihood for the NBP, using

$$p(\boldsymbol{n}_{j'} \mid \mathbf{N}^{(i)}_{J^{(i)}}, \mathbf{L}^{(i)}_{J}, \boldsymbol{\theta}^{(i)}) = \prod_{v=1}^{V} \mathrm{GNB}\left(n_{j'v}; l^{(i)}_{\cdot v} + \gamma^{(i)}_0/V, c^{(i)} + q^{(i)}_{\cdot}, p^{(i)}_{j'}\right) \tag{20}$$

as the likelihood for the GNBP, and using

$$p(\boldsymbol{n}_{j'} \mid \mathbf{N}^{(i)}_{J^{(i)}}, \boldsymbol{\theta}^{(i)}) = \prod_{v=1}^{V} \mathrm{BNB}\left(n_{j'v}; r^{(i)}_{j'}, n^{(i)}_{\cdot v} + \gamma^{(i)}_0/V, c^{(i)} + r^{(i)}_{\cdot}\right) \tag{21}$$

as the likelihood for the BNBP. Note that for this testing procedure we also compute $p(\boldsymbol{n}_{j'} \mid \mathbf{N}^{(i)}_{J})$ using Monte Carlo integration based on S posterior MCMC samples. In contrast to its truly nonparametric Bayesian counterpart with an infinite vocabulary, this testing procedure is expected to have higher computational complexity, but may produce better out-of-sample prediction if the predetermined finite vocabulary fits the testing documents well. Below we show the results produced by both testing procedures.

For comparison, we consider the multinomial naive Bayes classifier with Laplace smoothing (McCallum and Nigam, 1998, Manning et al., 2008), where a test document j' has the likelihood under newsgroup i as

$$\prod_{v=1}^{V} \left(\frac{n^{(i)}_{\cdot v} + 1}{\sum_{v=1}^{V} (n^{(i)}_{\cdot v} + 1)} \right)^{n_{j'v}}. \tag{22}$$

The results of some other commonly used text classification algorithms will also be included as benchmarks. Note that all these classifiers require the same predefined finite vocabulary for both training and testing. Thus any new terms in a testing document that are not listed in that vocabulary must be discarded.

3.5 Example results

We first consider choosing $S = 10$ in (17) and (18) to compute the predictive likelihood $p(\boldsymbol{n}_{j'} \mid \mathbf{N}^{(i)}_{J})$ for test document j'. Assuming a uniform prior for all the C categories, we assign document j' to category i with probability

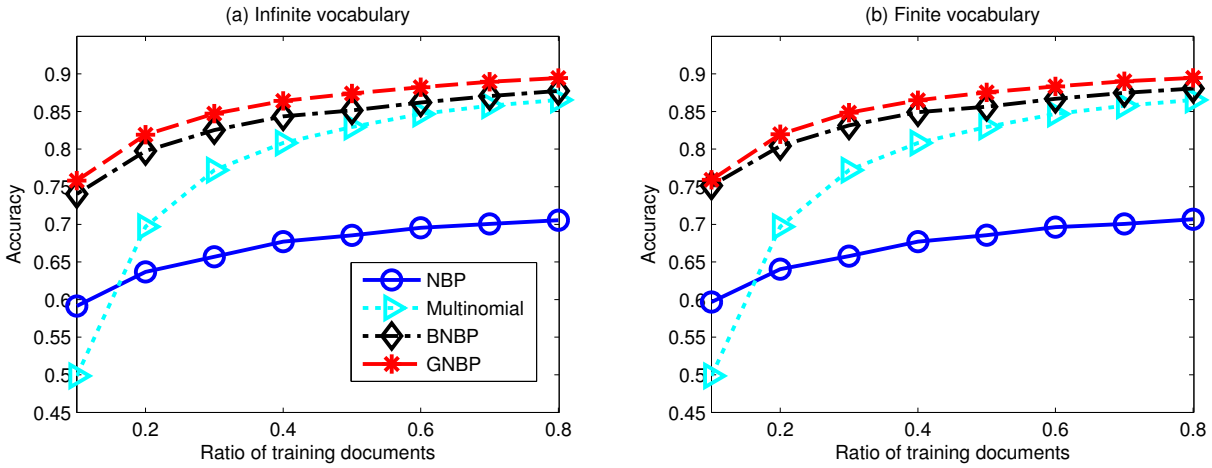

Figure 3: Document categorization results on the 20 Newsgroup dataset with (a) an unconstrained vocabulary that can grow to infinite, and (b) a predetermined finite vocabulary of size $V = 61{,}188$, using the negative binomial process (NBP), gamma-negative binomial process (GNBP), and beta-negative binomial process (BNBP). The results of the multinomial naive Bayes classifier using Laplace smoothing are included for comparison.

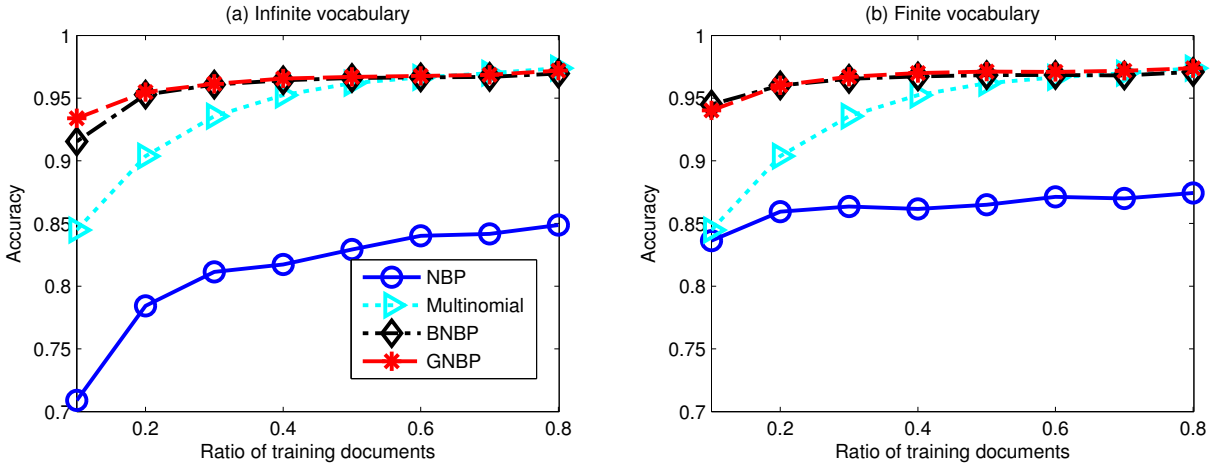

Figure 4: Analogous plots to Figures 3 (a) and (b) for the TDT2 dataset. The predetermined finite vocabulary has the size of $V = 36{,}771$.

$$\frac{p(\boldsymbol{n}_{j'} \mid \mathbf{N}_J^{(i)})}{\sum_{i=1}^{C} p(\boldsymbol{n}_{j'} \mid \mathbf{N}_J^{(i)})} \tag{23}$$

and categorize document j' to the category under which its word count vector $\boldsymbol{n}_{j'}$ has the highest probability. As shown in Figures 3 and 4, the NBP has the worst categorization accuracy. Both the BNBP and GNBP clearly outperform the NBP and the multinomial naive-Bayes classifier with Laplace smoothing, especially when the number of training documents

is small. Both for fitting the training count matrix and making out-of-sample prediction, the NBP is the most restrictive, as it has only two free parameters γ_0 and c. In addition to these two parameters, the GNBP (BNBP) has a probability (dispersion) parameter for each row count vector. Moreover, as both the GNB and BNB distributions are mixed negative-binomial distributions, they have heavier tails that may help model the burstiness of words in documents (Church and Gale, 1995, Madsen et al., 2005, Clinchant and Gaussier, 2008).

For the 20 newsgroup dataset, with the 7,505 documents collected at later times used for testing, our NBP, BNBP, and GNBP with an infinite vocabulary and $S = 10$ achieve categorization accuracies of 61.9%, 78.7%, and 80.9%, respectively. With a finite vocabulary they achieve accuracies of 61.7%, 79.1%, and 80.9%, respectively. Despite the simplicity of the model, this performance meets or exceeds that of other competing methods, which we briefly describe. The multinomial naive Bayes classifier with Laplace smoothing achieves an accuracy of 78.1%. Lan et al. (2009) consider a range of reweighted term-frequency features in a k-nearest neighbors (kNN) classifier. Under an optimal choice of k and set of features, they achievs an accuracy of 69.1%. The same authors report that a support vector machine (SVM) classifier achieves an accuracy of 80.8%. Larochelle et al. (2012) use restricted Boltzmann machine for classification, with an optimized training strategy and cross-validated model parameters. They report an accuracy of 76.2% using binary features for the 5000 most frequent words. The accuracy increases to 79.1% when using binary features for the 25247 most frequent words, but the algorithm is too computationally intensive to include more word features.

We also note that text categorization performance significantly deteriorates if one trains a multi-class classifier on the lower-dimensional features extracted using unsupervised feature learning algorithms, such as latent Dirichlet allocation (LDA) (Blei et al., 2003) or the deep Boltzmann machine (Srivastava et al., 2013). As shown in Srivastava et al. (2013), even with tuned parameters, neither LDA nor deep Boltzmann machines combined with a multinomial logistic regression classifier can achieve an accuracy above 70% on this data set. It is also shown in Zhu et al. (2012) that LDA plus an SVM classifier fails to achieve an accuracy above 65%. The performance of LDA could be improved by using a supervised training strategy (Blei and Mcauliffe, 2008). However, as shown in Zhu et al. (2012), the maximum-

entropy discrimination LDA (MedLDA), a state-of-the-art supervised LDA algorithm, still does not achieve an accuracy above 80%, despite the fact that the number of topics and model parameters are carefully tuned through cross validation and complex inference and heavy computations are employed to learn the latent features. Both the BNBP and GNBP naive classifiers, while being tuning-free and fast and simple to train using the raw counts, compare favorably to the state-of-the-art text classification algorithms that often rely on heavy computation and carefully selected features and parameters.

Note that for the proposed naive Bayes classifiers, a larger S usually leads to a more accurate computation of the predictive likelihood via Monte Carlo integration, but may not necessarily lead to a clear gain in accuracy for document categorization. This is confirmed by examining the experimental results with S set as small as one (i.e. a single MCMC sample) on both the 20 newsgroup and TDT2 datasets, which are found to be very similar to the results with $S = 10$ that are shown in Figures 3 and 4. This is not surprising since it is not the absolute magnitude of the category-specific predictive likelihoods, only their relative rankings, that determine the categorization accuracy.

To further elaborate on this point, we consider the CNAE-9 dataset[3] of Ciarelli and Oliveira (2009), which contains 1080 documents of free text business descriptions of Brazilian companies divided into nine categories, with a vocabulary size of $V = 856$; and we randomly select 20% of documents from each category as training, and calculate each test document's predictive probabilities under the nine categories, using the GNBP naive Bayes classifier with $S = 1000$ samples, each of which is the 2500th MCMC sample of an independent Markov chain. As shown in Figure 5 (a), in most cases, there is a little ambiguity on which category a test document should be assigned to. Hence letting $S = 1000$ or $S = 1$ make little practical difference in terms of categorization accuracy. In Figure 5 (b), from the left to right, we show the boxplot of 1000 accuracies produced by 1000 independent runs of the same testing procedure, each of which is calculated with $S = 1$ MCMC sample; the boxplot of 250 accuracies with $S = 4$; the boxplot of 100 accuracies with $S = 10$; and the boxplot of 20 accuracies with $S = 50$. It is clear from Figure 5 (b) that the larger the S is, the less the categorization accuracy varies, which is expected as the error of Monte Carlo

[3]https://archive.ics.uci.edu/ml/datasets/CNAE-9

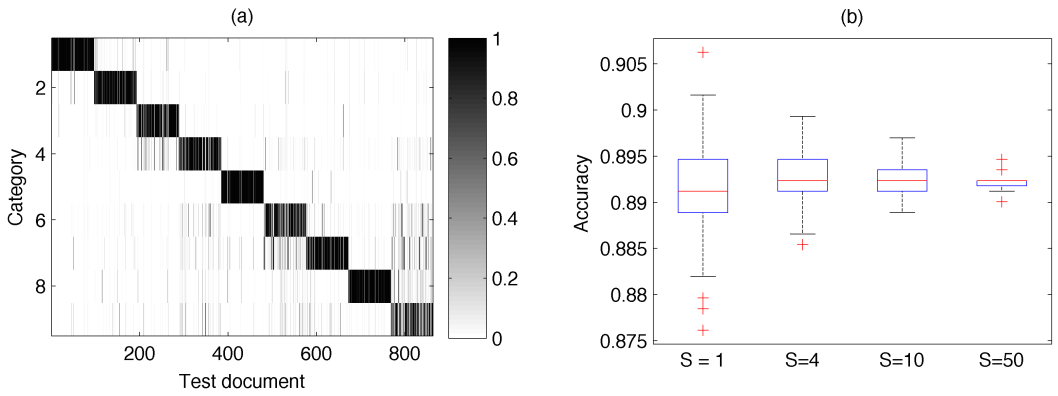

Figure 5: (a) The predicted probabilities of the test documents under different categories for the CNAE-9 dataset, using the GNBP nonparametric Bayesian naive Bayes classifier with 20% of the documents of each of the nine categories used for training. Each column shows the estimated probabilities across all nine categories for a singe document. Because the test documents from left to right were arranged from small to large according to their class labels, the dark diagonal band shows that most documents were placed with high posterior probability into the correct class. (b) Monte Carlo variability of document categorization accuracies under different settings of S, the number of MCMC samples used in computing the predictive likelihood. The boxplots show the variability of categorization accuracy when using $S = 1$, $S = 4$, $S = 10$, and $S = 50$ MCMC samples. While the variability is clearly higher with fewer samples, there is no evident bias for using a small S, and the actual scale of the variability (standard error $< 1\%$) is quite modest.

integration decreases with $\sqrt{N}$. However, there is no substantial improvement for the mean of the accuracies as S increases. Even with $S = 1$, the worst categorization accuracy is not too far from its mean. Therefore, in practice one may simply choose a small S to compute the predictive likelihoods for the purpose of document categorization.

As opposed to the conventional multinomial naive-Bayes classifier that estimates the probability of each word in the vocabulary by normalizing the word counts, the proposed negative binomial processes provide new methods that directly analyze the raw counts and take into account the total length of a document. Moreover, there is no need to predetermine the vocabulary, as new features not present in the training data have been taken care of by the nonparametric Bayesian predictive distributions of the negative binomial processes that are discussed in Section 2.4.

4 Conclusions

This paper fills a gap in the nonparametric Bayesian literature, deriving a family of probability mass functions for random count matrices by exploiting the gamma-Poisson, gamma-negative binomial, and beta-negative binomial processes. The resulting random count matrices have a random number of i.i.d. columns, and their parameters can be inferred with closed-form update equations. Any random count matrix in this family can be constructed by generating all its i.i.d. columns at once, or by adding one row at a time. Our results also allow us to define the predictive distribution of an infinite-dimensional random count vector under any of the proposed priors, leading to three nonparametric Bayesian naive Bayes classifiers for count vectors. The proposed classifiers, which directly operate on the raw counts and require no parameter tuning, alleviate the need to predetermine a shared finite vocabulary, and can account for features not present in the training data. Example results on document categorization show that the proposed gamma-negative binomial process and beta-negative binomial process clearly outperform both the negative binomial process and the multinomial naive Bayes classifier with Laplace smoothing, and have comparable performance to other state-of-the-art discriminatively-trained text classification algorithms. We are currently extending the techniques developed here to construct nonparametric Bayesian priors for a random count matrix, which has an unbounded number of columns and each row of which sums to a fixed integer; this extension can be used to construct nonparametric Bayesian discrete latent variable models, whose feature usages are represented with infinite random count matrices that are not directly observable.

Acknowledgements

The authors would like to thank the Associate Editor and two anonymous referees for their invaluable comments and suggestions that help significantly improve the paper.

References

D. Aldous. Exchangeability and related topics. *École d'ete de probabilités de Saint-Flour XIII-1983*, pages 1–198, 1985.

J. Bertoin. *Random fragmentation and coagulation processes*, volume 102. Cambridge University Press, 2006.

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

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

D. M. Blei and J. D. Mcauliffe. Supervised topic models. In *NIPS*, 2008.

T. Broderick, L. Mackey, J. Paisley, and M. I. Jordan. Combinatorial clustering and the beta negative binomial process. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2015.

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

D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE Trans. Knowledge and Data Engineering*, 2005.

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

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

F. Caron, Y. W. Teh, and B. T. Murphy. Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data. *Annal of Applied Statistics*, 2014.

S. Chib, E. Greenberg, and R. Winkelmann. Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics*, 1998.

K. W. Church and W. A. Gale. Poisson mixtures. *Natural Language Engineering*, 1995.

P. M. Ciarelli and E. Oliveira. Agglomeration and elimination of terms for dimensionality reduction. In *International Conference on Intelligent Systems Design and Applications*, 2009.

S. Clinchant and E. Gaussier. The BNB distribution for text modeling. In *Advances in Information Retrieval*. 2008.

K. Crammer, M. Dredze, and F. Pereira. Confidence-weighted linear classification for text categorization. *Journal of Machine Learning Research*, 13(1):1891–1926, 2012.

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes*, volume 2. Springer, 1988.

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

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

C. Heaukulani and D. M. Roy. The combinatorial structure of beta negative binomial processes. *arXiv:1401.0062*, 2013.

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

D. N. Hoover. Row-column exchangeability and a general model for exchangeability. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*. 1982.

L. F. James. Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093*, 2002.

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

M. Lan, C. L. Tan, J. Su, and Y. Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(4): 721–735, 2009.

H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio. Learning algorithms for the classification restricted Boltzmann machine. *Journal of Machine Learning Research*, 13(1):643–669, 2012.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian nonparametrics*. Cambridge University Press, 2010.

A. Y. Lo. Bayesian nonparametric statistical inference for Poisson point processes. *Zeitschrift fur*, pages 55–66, 1982.

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

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *AAAI/ICML- 98 Workshop on Learning for Text Categorization*, 1998.

J. E. Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, pages 65–82, 1962.

P. Orbanz and D. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

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

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

M. S. Ridout. Generating random numbers from a distribution specified by its Laplace transform. *Statistics and Computing*, pages 439–450, 2009.

C. Ritter and M. A. Tanner. Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, 1992.

M. Sibuya. Generalized hypergeometric, digamma and trigamma distributions. *Annals of the Institute of Statistical Mathematics*, pages 373–390, 1979.

N. Srivastava, R. Salakhutdinov, and G. Hinton. Modeling documents with a deep Boltzmann machine. In *Uncertainty in Artificial Intelligence*, 2013.

Y. W. Teh and D. Gorur. Indian buffet processes with power-law behavior. In *NIPS*, 2009.

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

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

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

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

M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2015.

M. Zhou and S. G. Walker. Sample size dependent species models. *arXiv:1410.3155*, 2014.

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

J. Zhu, A. Ahmed, and E. P. Xing. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(1):2237–2278, 2012.

A The Negative Binomial Process: Details

A.1 Negative binomial process random count matrix

To generate a random count matrix, we construct a gamma-Poisson process as

$$X_j \sim \mathrm{PP}(G),\ G \sim \Gamma\mathrm{P}(G_0, 1/c). \tag{24}$$

Zhou and Carin (2015) derives the marginal distribution of $X = \sum_{j=1}^{J} X_j$ and calls it as the negative binomial process (NBP), a draw from which is represented as an exchangeable random count *vector*. We do not consider that simplification in this paper and consequently our definition of the NBP, a draw from which is represented as a row-column exchangeable random count *matrix*, differs from the one in Zhou and Carin (2015).

The conditional likelihood in (4) can be re-written as

$$p(\{X_j\}_{1,J} \mid G) = e^{-JG(\Omega)} \prod_{k=1}^{K_J} \sum_{k'=1}^{\infty} \frac{r_{k'}^{n_{\cdot k'}}}{\prod_{j=1}^{J} n_{jk'}!} \delta(\omega_{k'} = \omega_k)\,.$$

Applying the Palm formula (Daley and Vere-Jones, 1988, James, 2002, Bertoin, 2006, Caron et al., 2014) to the expectation $\mathbb{E}_G[p(\{X_j\}_{1,J} \mid G)]$, we have

$$\begin{aligned}
&\mathbb{E}_G[p(\{X_j\}_{1,J} \mid G)] = \mathbb{E}\left[e^{-JG(\Omega)} \prod_{k=1}^{K_J} \sum_{k'=1}^{\infty} \frac{r_{k'}^{n_{\cdot k'}}}{\prod_{j=1}^{J} n_{jk'}!} \delta(\omega_{k'} = \omega_k)\right] \\
&= \int_{\mathbb{R}_+ \times \Omega} \frac{r_1^{n_{\cdot 1}}}{\prod_{j=1}^{J} n_{j1}!} e^{-Jr_1} \nu(dr_1 d\omega_1) \mathbb{E}\left[e^{-JG(\Omega\setminus\{\omega_1\})} \prod_{k=2}^{K_J} \sum_{k'=1}^{\infty} \frac{r_{k'}^{n_{\cdot k'}}}{\prod_{j=1}^{J} n_{jk'}!} \delta(\omega_{k'} = \omega_k)\right] \\
&= \ldots \\
&= \left\{\prod_{k=1}^{K_J} \int_{\mathbb{R}_+ \times \Omega} \frac{r_k^{n_{\cdot k}}}{\prod_{j=1}^{J} n_{jk}!} e^{-Jr_k} \nu(dr_k d\omega_k)\right\} \cdot \left\{\mathbb{E}_G\left[e^{-JG(\Omega\setminus\mathcal{D}_J)}\right]\right\}.
\end{aligned}$$

Directly calculation with $\int_{\mathbb{R}_+ \times \Omega} r^n e^{-Jr} \nu(dr d\omega) = \gamma_0 (J + c)^{-n} \Gamma(n)$ and $\mathbb{E}_G[e^{-JG(\Omega\setminus\mathcal{D}_J)}] = (1 + J/c)^{-\gamma_0}$ leads to

$$p(\{X_j\}_{1,J} \mid \gamma_0, c) = \mathbb{E}_G[p(\{X_j\}_{1,J} \mid G)] = \gamma_0^{K_J} e^{-\gamma_0 \ln(\frac{J+c}{c})} \prod_{k=1}^{K_J} \frac{\frac{\Gamma(n_{\cdot k})}{(J+c)^{n_{\cdot k}}}}{\prod_{j=1}^{J} n_{jk}!}.$$

B Gamma-Negative Binomial Process: Details

B.1 GNBP random count matrix

Given the gamma process $G \sim \Gamma\mathrm{P}(G_0, 1/c)$, we define $X \mid G \sim \mathrm{NBP}(G, p)$ as a negative binomial process such that $X(A) \sim \mathrm{NB}(G(A), p)$ for each $A \subset \Omega$. Replacing the Poisson processes in (24) with the negative binomial processes defined in this way yields a gamma-negative binomial process (GNBP):

$$X_j \sim \mathrm{NBP}(G, p_j)\,, \quad G \sim \Gamma\mathrm{P}(G_0, 1/c)\,.$$

With a draw from the gamma process $G \sim \Gamma\mathrm{P}(G_0, 1/c)$ expressed as $G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k}$, a draw from $X_j \mid G \sim \mathrm{NBP}(G, p_j)$ can be expressed as $X_j = \sum_{k=1}^{\infty} n_{jk} \delta_{\omega_k},\ n_{jk} \sim \mathrm{NB}(r_k, p_j)$. The GNBP employs row-specific probability parameters p_j to model row heterogeneity, and hence X_j are conditionally independent but not identically distributed if p_j at different rows are set differently. Note that the GNBP is previously proposed in Zhou and Carin (2015), which focuses on finding the conditional posterior of G, without considering the marginalization of G.

The GNBP hierarchical construction is conceptually simple, but to obtain a random count matrix, we have to marginalize out the gamma process $G \sim \Gamma\mathrm{P}(G_0, 1/c)$. As it is difficult to directly marginalize G out of the conditional likelihood of the observed J rows as

$$p(\{X_j\}_{1,J} \mid G, \boldsymbol{p}) = \prod_{k=1}^{\infty} \prod_{j=1}^{J} \frac{\Gamma(n_{jk} + r_k)}{n_{jk}! \Gamma(r_k)} p_j^{n_{jk}} (1 - p_j)^{r_k},$$

where $\boldsymbol{p} := (p_1, \ldots, p_J)$, we first augment each $n_{jk} \sim \mathrm{NB}(r_k, p_j)$ under its compound Poisson representation as $n_{jk} \sim \mathrm{SumLog}(l_{jk}, p_j),\ l_{jk} \sim \mathrm{Pois}(r_k q_j)$.

Define $X \sim \mathrm{SumLogP}(L, p)$ as a sum-logarithmic process such that $X(A) \sim \mathrm{SumLog}(L(A), p)$ for each $A \subset \Omega$. With $X_j \sim \mathrm{NBP}(G, p_j)$ augmented as $X_j \sim \mathrm{SumLogP}(L_j, p_j),\ L_j \sim$

$\mathrm{PP}(q_j G)$, we may express the joint likelihood of X_j and L_j as

$$p(\{X_j, L_j\}_{1,J} \mid G, \boldsymbol{p}) = \prod_{j=1}^{J} \prod_{k=1}^{\infty} \frac{|s(n_{jk}, l_{jk})| r_k^{l_{jk}}}{n_{jk}!} p_j^{n_{jk}} (1 - p_j)^{r_k},$$

With $l_{\cdot k} := \sum_{j=1}^{J} l_{jk}$, similar to the analysis in Section A, we can reexpress the likelihood as

$$p(\{X_j, L_j\}_{1,J} \mid G, \boldsymbol{p}) = e^{-q_\cdot G(\Omega \setminus \mathcal{D})} \prod_{k=1}^{K_J} r_k^{l_{\cdot k}} e^{-q_\cdot r_k} \left(\prod_{j=1}^{J} \frac{|s(n_{jk}, l_{jk})| p_j^{n_{jk}}}{n_{jk}!} \right). \tag{25}$$

Similar to the analysis in Section A.1, with G marginalized out as $p(\{X_j, L_j\}_{1,J} \mid \gamma_0, c, \boldsymbol{p}) = \mathbb{E}_G[p(\{X_j, L_j\}_{1,J} \mid G, \boldsymbol{p})]$, we obtain the GNBP random matrix prior in (10) using

$$f(\mathbf{N}_J, \mathbf{L}_J \mid \gamma_0, c, \boldsymbol{p}) = \frac{p(\{X_j, L_j\}_{1,J} \mid \gamma_0, c, \boldsymbol{p})}{K_J!}. \tag{26}$$

Although not obvious, one may verify that (10) defines the PMF of a compound random count matrix, which can be generated via

$$\begin{aligned}
&n_{jk} \sim \mathrm{SumLog}(l_{jk}, p_j), \\
&(l_{1k}, \ldots, l_{Jk}) \sim \mathrm{Mult}(l_{\cdot k}, q_1/q_\cdot, \ldots, q_J/q_\cdot), \\
&l_{\cdot k} \sim \mathrm{Log}[q_\cdot/(c + q_\cdot)], \\
&K_J \sim \mathrm{Pois}\{\gamma_0[\ln(c + q_\cdot) - \ln(c)]\}.
\end{aligned} \tag{27}$$

Let $\sigma(1), \ldots, \sigma(J)$ denote a random permutation of the column indices. If p_j are set differently for different rows, then $\mathrm{Mult}(l_{\cdot k}, q_{\sigma(1)}/q_\cdot, \ldots, q_{\sigma(J)}/q_\cdot) \overset{d}{\neq} \mathrm{Mult}(l_{\cdot k}, q_1/q_\cdot, \ldots, q_J/q_\cdot)$ and hence the introduced random count matrix no longer maintains row exchangeability.

Comparing (27) with (6), one may identify several key differences between the GNBP and NBP random count matrices. First, one may increase p_j to encourage the jth row to have larger counts than the others. Second, both n_{jk} and the column sum $n_{\cdot k}$ are generated from compound distributions. In fact, if we let $p_j \equiv 1 - e^{-1}$, then the matrix $\{l_{jk}\}_{jk}$ in (27) is exactly a NBP random count matrix, and the GNBP builds its random matrix using $n_{jk} \sim \mathrm{SumLog}(l_{jk}, p_j)$.

The sequential construction of a GNBP random count matrix can be intuitively explained as drawing dishes, drawing tables at each dish, and then drawing customers at each table. Similar to the definition of $\mathbf{N}^+_{J+1}$, we let $\mathbf{L}^+_{J+1}$ represent the new row and columns added to $\mathbf{L}_J$. Using (10), following the analysis in Section 2.1, one may show with direct calculation that

$$\begin{aligned} p(\mathbf{N}^+_{J+1}, \mathbf{L}^+_{J+1} \mid \mathbf{N}_J, \mathbf{L}_J, \boldsymbol{\theta}) = \frac{K_J! K^+_{J+1}!}{K_{J+1}!} \prod_{k=1}^{K_{J+1}} \mathrm{SumLog}\left(l_{(J+1)k}, p_{J+1}\right) \\ \times \prod_{k=1}^{K_J} \mathrm{NB}\left(l_{(J+1)k}; l_{\cdot k}, \frac{q_{J+1}}{c + q_{\cdot} + q_{J+1}}\right) \\ \times \prod_{k=K_J+1}^{K_{J+1}} \mathrm{Log}\left(l_{(J+1)k}; \frac{q_{J+1}}{c + q_{\cdot} + q_{J+1}}\right) \\ \times \mathrm{Pois}\left\{K^+_{J+1}; \gamma_0 \left[\ln(c + q_{\cdot} + q_{J+1}) - \ln(c + q_{\cdot})\right]\right\}. \end{aligned} \tag{28}$$

Thus to add a new row, we first draw $\mathrm{NB}[l_{\cdot k}, q_{J+1}/(c + q_{\cdot} + q_{J+1})]$ tables at existing columns (dishes); we then draw $K^+_{J+1} \sim \mathrm{Pois}\{\gamma_0[\ln(c+q_{\cdot}+q_{J+1}) - \ln(c+q_{\cdot})]\}$ new dishes, each of which is associated with $\mathrm{Log}[q_{J+1}/(c + q_{\cdot} + q_{J+1})]$ tables; we further draw $\mathrm{Log}(p_{J+1})$ customers at each table and aggregate the counts across the tables of the same dish as $n_{(J+1)k} = \sum_{t=1}^{l_{(J+1)k}} n_{(J+1)kt}$; and in the final step, we insert the K^+_{J+1} new columns into the K_J original columns without reordering, which again is a one to $K_{J+1}!/\left(K_J!\, K^+_{J+1}!\right)$ mapping. We emphasize that the number of tables (customers) for a new dish, which follows a logarithmic (sum-logarithmic) distribution, must be at least one; the implication is that there are infinite many dishes that have not yet been ordered by any of the tables seated by existing customers. The sequential construction provides a convenient way to construct a GNBP random count matrix one row at a time.

With the latent counts $l_{(J+1)k}$ marginalized out, one may show that the predictive distribution for $\mathbf{N}^+_{J+1}$, given $\mathbf{N}_J$ and $\mathbf{L}_J$, can be expressed in terms of the Poisson, LogLog and

GNB distributions as

$$p(\mathbf{N}^+_{J+1} \mid \mathbf{N}_J, \mathbf{L}_J, \boldsymbol{\theta}) = \frac{K_J! K^+_{J+1}!}{K_{J+1}!} \prod_{k=1}^{K_J} \mathrm{GNB}\left(n_{(J+1)k}; l_{\cdot k}, c+q_{\cdot}, p_{J+1}\right) \times \prod_{k=K_J+1}^{K_{J+1}} \mathrm{LogLog}\left(n_{(J+1)k}; c+q_{\cdot}, p_{J+1}\right) \times \mathrm{Pois}\left\{K^+_{J+1}; \gamma_0 \left[\ln(c+q_{\cdot}+q_{J+1}) - \ln(c+q_{\cdot})\right]\right\}, \quad (29)$$

where $n \sim \mathrm{LogLog}(c,p)$ represents a logarithmic mixed sum-logarithmic distribution defined on positive integers and $n \sim \mathrm{GNB}(l,c,p)$ represents a gamma mixed negative binomial distribution defined on $\mathbb{Z}$, whose PMFs are shown in Appendix D.

B.2 Inference for parameters

Both the GNB and LogLog distributions have complicated PMFs involving Stirling numbers of the first kind and it seems difficult to infer their parameters. Fortunately, using the likelihoods (25) and (10) and the data augmentation techniques developed for the negative binomial distribution (Zhou and Carin, 2015), we are able to derive closed-form conditional posteriors for the GNBP. To complete the model, we let $\gamma_0 \sim \mathrm{Gamma}(e_0, 1/f_0)$, $p_j \sim \mathrm{Beta}(a_0, b_0)$ and $c \sim \mathrm{Gamma}(c_0, 1/d_0)$. We sample the model parameters as

$$\begin{aligned}
&(\gamma_0|-) \sim \mathrm{Gamma}\left(e_0 + K_J, \frac{1}{f_0 - \ln(\frac{c}{c+q_{\cdot}})}\right),\\
&(l_{jk}|-) = \sum_{t=1}^{n_{jk}} u_t,\ u_t \sim \mathrm{Bernoulli}\left(\frac{r_k}{r_k + t - 1}\right),\\
&(r_k|-) \sim \mathrm{Gamma}\big(l_{\cdot k}, 1/(c+q_{\cdot})\big),\\
&\{G(\Omega \backslash \mathcal{D}_J)|-\} \sim \mathrm{Gamma}\big(\gamma_0, 1/(c+q_{\cdot})\big),\\
&(p_j|-) \sim \mathrm{Beta}\big(a_0 + m_j, b_0 + G(\Omega)\big),\\
&(c|-) \sim \mathrm{Gamma}\big(c_0 + \gamma_0, 1/[d_0 + G(\Omega)]\big).
\end{aligned} \quad (30)$$

C Beta-Negative Binomial Process: Details

C.1 BNBP random count matrix

The GNBP generalizes the NBP by replacing the Poisson process in (24) using a negative binomial process and shares the negative binomial dispersion parameters across rows. Exploiting an alternative strategy that shares the negative binomial probability parameters across rows, we construct a BNBP as

$$X_j \sim \mathrm{NBP}(r_j, B),\ B \sim \mathrm{BP}(c, B_0),$$

where $p_k = B(\omega_k)$ is the weight of the atom ω_k of the beta process $B \sim \mathrm{BP}(c, B_0)$, and $X_j \mid B \sim \mathrm{NBP}(r_j, B)$ is a negative binomial process such that $X_j(A) = \sum_{k:\omega_k \in A} n_{jk},\ n_{jk} \sim \mathrm{NB}(r_j, p_k)$ for each $A \subset \Omega$.

With $\boldsymbol{r} := (r_1, \ldots, r_J)$, similar to the analysis in Section B, the likelihood of the BNBP can be expressed as

$$p(\{X\}_{1,J} \mid B, \boldsymbol{r}) = e^{-p_* r_\cdot} \prod_{k=1}^{K_J} p_k^{n_{\cdot k}} (1-p_k)^{r_\cdot} \prod_{j=1}^{J} \frac{\Gamma(n_{jk} + r_j)}{n_{jk}! \Gamma(r_j)}, \tag{31}$$

where p_* denotes the sum over all the atoms in the absolutely continuous space $\Omega \backslash \mathcal{D}_J$ as

$$p_* := -\textstyle\sum_{k:n_{\cdot k}=0} \ln(1-p_k)$$

and $r_\cdot := \sum_{j=1}^{J} r_j$. Using the Lévy-Khintchine theorem and (1), the Laplace transform of p_* can be expressed as

$$\begin{aligned}
\mathbb{E}[e^{-s p_*}] &= \exp\left\{ \int_{[0,1]\times\Omega} \left[(1-p)^s - 1\right] \nu(dp d\omega) \right\} \\
&= \exp\left[-\gamma_0 \sum_{i=0}^{\infty} \left(\frac{1}{c+i} - \frac{1}{c+i+s} \right) \right] \\
&= \exp\left\{ -\gamma_0 \left[\psi(c+s) - \psi(c) \right] \right\},
\end{aligned}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function; we define such a random variable as the logbeta random variable

$$p_* \sim \text{logBeta}(\gamma_0, c),$$

whose mean and variance are $\mathbb{E}[p_*] = \gamma_0\psi_1(c)$ and $\text{Var}[p_*] = -\gamma_0\psi_2(c)$, respectively, where $\psi_n(x) = \frac{d^n\psi(x)}{dx^n}$.

As before, one may verify with direct calculation that (11) defines the PMF of a column-i.i.d. random count matrix $\mathbf{N}_J \in \mathbb{Z}^{J\times K_J}$, which can be generated via

$$\begin{aligned}
&\boldsymbol{n}_{:k} \sim \text{DirMult}(n_{\cdot k}, r_1, \ldots, r_J),\\
&n_{\cdot k} \sim \text{Digam}(r_\cdot, c),\\
&K_J \sim \text{Pois}\big\{\gamma_0 \left[\psi(c + r_\cdot) - \psi(c)\right]\big\},
\end{aligned} \tag{32}$$

where the PMFs of both the Dirichlet-multinomial (DirMult) and digamma distributions are shown in the Appendix. Note that if r_j are set differently for different rows, then $\text{DirMult}(n_{\cdot k}, r_{\sigma(1)}, \ldots, r_{\sigma(J)}) \overset{d}{\neq} \text{DirMult}(n_{\cdot k}, r_1, \ldots, r_J)$ and hence the corresponding random count matrix no longer maintains row exchangeability.

The sequential construction of a BNBP random count matrix can be intuitively understood as an "ice cream" buffet process (ICBP). Using (11), similar to the analysis in Section 2.1, we have

$$\begin{aligned}
p(\mathbf{N}^+_{J+1} \mid \mathbf{N}_J) = &\frac{K_J! K^+_{J+1}!}{K_{J+1}!} \prod_{k=1}^{K_J} \text{BNB}(n_{(J+1)k}; r_{J+1}, n_{\cdot k}, c + r_\cdot)\\
&\times \prod_{k=K_J+1}^{K_{J+1}} \text{Digam}(n_{(J+1)k}; r_{J+1}, c + r_\cdot)\\
&\times \text{Pois}\left\{K^+_{J+1}; \gamma_0 \left[\psi(c + r_\cdot + r_{J+1}) - \psi(c + r_\cdot)\right]\right\},
\end{aligned} \tag{33}$$

where the PMF for the beta-negative binomial (BNB) distribution is shown in Appendix D. Thus to add a row to $\mathbf{N}_J \in \mathbb{Z}^{J\times K_J}$, customer $J+1$ takes $n_{(J+1)k} \sim \text{BNB}(r_{J+1}, n_{\cdot k}, c + r_\cdot)$ number of scoops at an existing ice cream (column); the customer further selects $K^+_{J+1} \sim \text{Pois}\left\{\gamma_0 \left[\psi(c + r_\cdot + r_{J+1}) - \psi(c + r_\cdot)\right]\right\}$ new ice creams out of the buffet line and takes $n_{(J+1)k} \sim$

Digam$(r_{J+1}, c + r.)$ number of scoops at each new ice cream. Thus the ICBP can also be considered as a "multiple-scoop" Indian buffet process, an analogy used in Zhou et al. (2012). Note that when $r_j \equiv 1$, we have $K^+_{J+1} \sim \text{Pois}[\gamma_0/(c + J)]$, confirming the derivation about the number of new dishes (ice creams) in Section 3.2 of Zhou et al. (2012)[4], which, however, provides no descriptions about the distributions of the number of scoops at existing and new ice creams. We emphasize that the number of scoops at a new ice cream, which follows a digamma distribution, must be at least one; the implication is that there are infinite many ice creams in the buffet line that have not yet been scooped by any of the existing customers. Similar to the GNBP random count matrix, the BNBP random count matrix is column exchangeable, but not row exchangeable if the row-specific dispersion parameters r_j are fixed at different values.

A related marked BNBP of Zhou et al. (2012), Zhou and Carin (2012) attaches an independent negative binomial dispersion parameter r_k for each atom of the beta process, and infers its values under a finite approximation of the beta process; another related BNBP of Broderick et al. (2015) uses a single dispersion parameter r and sets its value empirically. None of these papers, however, marginalize out the beta process to define a prior on column-i.i.d. random count matrices, a challenge tackled in this paper.

Independently of our work, Heaukulani and Roy (2013) also describe the marginalization of the beta process from the negative binomial process, where the obtained BNBP is called the negative binomial Indian buffet process. Although the idea of marginalizing out the beta process is shared by both papers, the techniques and combinatorial arguments used are quite different. Their paper focuses on a special case of the BNBP where a single dispersion parameter r is used for all the X_j's. Our model allows row-specific dispersion parameters r_j, develops an efficient inference scheme for all model parameters, derives the predictive distribution of a new row count vector under a BNBP random count matrix, and also situates the BNBP in the larger family of count-matrix priors derived from negative-binomial processes.

[4]Due to different parameterization of the Lévy measure, the beta process mass parameter γ_0 in this paper can be considered as $\gamma_0 c$ in Thibaux and Jordan (2007) and Zhou et al. (2012).

C.2 Inference for parameters

For all the atoms in the absolutely continuous part of the space, $\Omega \backslash \mathcal{D}_J$, we have that

$$(\nu(dpd\omega) \mid -) = p^{-1}(1-p)^{c+r_\cdot-1} dp B_0(d\omega)\,.$$

Thus the Laplace transform of $(p_*|-)$ can be expressed as

$$\mathbb{E}[e^{-s(p_*|-)}] = \exp\left\{-\gamma_0\left[\psi(c+r_\cdot+s)-\psi(c+r_\cdot)\right]\right\},$$

and hence we have $(p_*|-) \sim \text{logBeta}(\gamma_0, c+r_\cdot)$. With its Laplace transform, we sample $(p_*|-)$ using the method proposed in Ridout (2009). To complete the model, we let $\gamma_0 \sim \text{Gamma}(e_0, 1/f_0)$, $r_j \sim \text{Gamma}(a_0, b_0)$ and $c \sim \text{Gamma}(c_0, 1/d_0)$. Using both the conditional likelihood (31) and the marginal likelihood (11), and the data augmentation techniques developed in Zhou and Carin (2015), we sample the model parameters as

$$\begin{aligned}
&(\gamma_0|-) \sim \text{Gamma}\bigg(e_0 + K_J, \frac{1}{f_0 + \psi(c+r_\cdot) - \psi(c)}\bigg),\\
&(p_k|-) \sim \text{Beta}(n_{\cdot k}, c+r_\cdot), \quad (p_*|-) \sim \text{logBeta}(\gamma_0, c+r_\cdot),\\
&(l_{jk}|-) = \sum_{t=1}^{n_{jk}} u_t,\ u_t \sim \text{Bernoulli}\bigg(\frac{r_j}{r_j+t-1}\bigg),\\
&(r_j|-) \sim \text{Gamma}\bigg(a_0 + l_{j\cdot}, \frac{1}{b_0 + p_* - \sum_{k=1}^{K_J}\ln(1-p_k)}\bigg).
\end{aligned} \tag{34}$$

The only parameter that does not have an analytic conditional posterior is the concentration parameter c. Since using Campbell's theorem (Kingman, 1993), we have $\mathbb{E}[\sum_k p_k] = \int_{[0,1]\times\Omega} p\nu(dpd\omega) = \gamma_0/c$, to sample c, we use

$$Q(c') = \text{Gamma}\bigg(c_0 + \gamma_0, \frac{1}{d_0 + p_* + \sum_{k=1}^{K_J} p_k}\bigg) \tag{35}$$

as the proposal distribution in an independence chain Metropolis-Hastings sampling step. One may also sample c using a griddy-Gibbs sampler (Ritter and Tanner, 1992).

D Some useful distributions

Direct calculation shows that the logarithmic mixed sum-logarithmic (LogLog) distribution, expressed as $n \sim \text{SumLog}(l,p),\ l \sim \text{Log}\left(\frac{-\ln(1-p)}{c-\ln(1-p)}\right)$, has PMF

$$f_N(n|c,p) = \frac{\sum_{l=1}^{n} \frac{|s(n,l)|p^n}{n!} \frac{\Gamma(l)}{[c-\ln(1-p)]^l}}{\ln[c-\ln(1-p)] - \ln(c)}$$

for $n \in \{1,2,\ldots\}$; and the negative binomial mixed sum-logarithmic distribution, expressed as $n \sim \text{SumLog}(l,p),\ l \sim \text{NB}\left(e, \frac{-\ln(1-p)}{c-\ln(1-p)}\right)$, has PMF

$$f_N(n|e,c,p) = \sum_{l=0}^{n} \frac{c^e p^n |s(n,l)|}{\Gamma(e)n!} \frac{\Gamma(e+l)}{[c-\ln(1-p)]^{e+l}}$$

for $n \in \{0,1,\ldots\}$. The iterative calculation of $|s(n,l)|/n!$ under the logarithmic scale is described in Appendix E. Using (2), one may show that the negative binomial mixed sum-logarithmic distribution shown above is equivalent to a gamma mixed negative binomial (GNB) distribution, generated by $n \sim \text{NB}(r,p),\ r \sim \text{Gamma}(e,1/c)$. Note that $n \sim \text{LogLog}(c,p)$ is the limit of $n \sim \text{GNB}(e,c,p)$ as $e \to 0$, conditioning on $n > 0$, thus it can be considered as a truncated GNB distribution.

The Dirichlet-multinomial (DirMult) distribution (Mosimann, 1962, Madsen et al., 2005) is a Dirichlet mixed multinomial distribution, with PMF

$$\text{DirMult}(\boldsymbol{n}_{:k} \mid n_{\cdot k}, \boldsymbol{r}) = \frac{n_{\cdot k}!}{\prod_{j=1}^{J} n_{kj}!} \frac{\Gamma(r_\cdot)}{\Gamma(n_{\cdot k}+r_\cdot)} \prod_{j=1}^{J} \frac{\Gamma(n_{kj}+r_j)}{\Gamma(r_j)},$$

and the digamma distribution (Sibuya, 1979) has PMF

$$\text{Digam}(n \mid r,c) = \frac{1}{\psi(c+r)-\psi(c)} \frac{\Gamma(r+n)\Gamma(c+r)}{n\Gamma(c+n+r)\Gamma(r)}, \tag{36}$$

where $n = 1,2,\ldots$. Since the beta-negative binomial (BNB) distribution has PMF

$$f_N(n \mid r,e,c) = \int_0^1 \text{NB}(n;r,p)\text{Beta}(p;e,c)dp = \frac{\Gamma(r+n)}{n!\Gamma(r)} \frac{\Gamma(c+r)\Gamma(e+n)\Gamma(e+c)}{\Gamma(e+c+r+n)\Gamma(e)\Gamma(c)},$$

one may show that conditioning on $n > 0$, $n \sim \text{BNB}(r, e, c)$ becomes $n \sim \text{Digam}(r, c)$ as $e \to 0$. Thus the digamma distribution can be considered as a truncated BNB distribution.

Since the Laplace transform of the logbeta random variable $p_* \sim \text{logBeta}(\gamma_0, c)$ can be reexpressed as

$$\mathbb{E}[e^{-sp_*}] = \prod_{i=0}^{\infty} \exp\left\{\frac{\gamma_0}{c+i}\left[\left(1+\frac{s}{c+i}\right)^{-1} - 1\right]\right\},$$

we can generate $p_* \sim \text{logBeta}(\gamma_0, c)$ as an infinite sum of independent compound Poisson random variables as

$$p_* = \sum_{i=0}^{\infty} \lambda_i,\ \lambda_i = \sum_{t=1}^{u_i} \lambda_{it},\ u_i \sim \text{Pois}\left(\frac{\gamma_0}{c+i}\right),\ \lambda_{it} \sim \text{Gamma}\left(1, \frac{1}{c+i}\right). \tag{37}$$

E Calculating Stirling Numbers of the First Kind

The unsigned Stirling numbers of the first kind $|s(n,l)|$ appear in the predictive distribution for the GNBP. It is numerically unstable to recursively calculate $|s(n,l)|$ based on $|s(n,l)| = (n-1)|s(n-1,l)| + |s(n-1,l-1)|$, as $|s(n,l)|$ would rapidly reach the maximum value allowed by a finite precision machine as n increases. Denoting

$$g(n,l) = \ln(|s(n,l)|) - \ln(n!),$$

we iteratively calculate $g(n,l)$ with $g(n,1) = \ln(n-1) - \ln(n) + \ln g(n-1,1)$, $g(n,n) = g(n-1,n-1) - \ln n$, and

$$g(n,l) = \ln\frac{n-1}{n} + g(n-1,l) + \ln\left\{1 + \exp[g(n-1,l-1) - g(n-1,l) - \ln(n-1)]\right\}$$

for $2 \le l \le n-1$. This approach is found to be numerically stable.