Augment-and-Conquer Negative Binomial Processes

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

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

Abstract

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

1 Introduction

There has been increasing interest in count modeling using the Poisson process, geometric process \cite{1, 2, 3, 4} and recently the negative binomial (NB) process \cite{5, 6}. Notably, it has been independently shown in \cite{5} and \cite{6} that the NB process, originally constructed for count analysis, can be naturally applied for mixture modeling of grouped data \(x_1, \cdots, x_J\), where each group \(x_j = \{x_{ji}\}_{i=1,N_j}\). For a territory long occupied by the hierarchical Dirichlet process (HDP) \cite{7} and related models, the inference of which may require substantial bookkeeping and suffer from slow convergence \cite{7}, the discovery of the NB process for mixture modeling can be significant. As the seemingly distinct problems of count and mixture modeling are united under the NB process framework, new opportunities emerge for better data fitting, more efficient inference and more flexible model constructions. However, neither \cite{5} nor \cite{6} explore the properties of the NB distribution deep enough to achieve fully tractable closed-form inference. Of particular concern is the NB dispersion parameter, which was simply fixed or empirically set \cite{6}, or inferred with a Metropolis-Hastings algorithm \cite{5}. Under these limitations, both papers fail to reveal the connections of the NB process to the HDP, and thus may lead to false assessments on comparing their modeling abilities.

We perform joint count and mixture modeling under the NB process framework, using completely random measures \cite{1, 8, 9} that are simple to construct and amenable for posterior computation. We propose to augment-and-conquer the NB process: by “augmenting” a NB process into both the gamma-Poisson and compound Poisson representations, we “conquer” the unification of count and mixture modeling, the analysis of fundamental model properties, and the derivation of efficient Gibbs sampling inference. We make two additional contributions: 1) we construct a gamma-NB process, analyze its properties and show how its normalization leads to the HDP, highlighting its unique theoretical, structural and computational advantages relative to the HDP. 2) We show that a variety of NB processes can be constructed with distinct model properties, for which the shared random measure can be selected from completely random measures such as the gamma, beta, and beta-Bernoulli processes; we compare their performance on topic modeling, a typical example for mixture modeling of grouped data, and show the importance of inferring both the NB dispersion and probability parameters, which respectively govern the overdispersion level and the variance-to-mean ratio in count modeling.
1.1 Poisson process for count and mixture modeling

Before introducing the NB process, we first illustrate how the seemingly distinct problems of count and mixture modeling can be united under the Poisson process. Denote $\Omega$ as a measure space and for each Borel set $A \subset \Omega$, denote $X_j(A)$ as a count random variable describing the number of observations in $x_j$ that reside within $A$. Given grouped data $x_1, \cdots, x_J$, for any measurable disjoint partition $A_1, \cdots, A_Q$ of $\Omega$, we aim to jointly model the count random variables $\{X_j(A_q)\}$. A natural choice would be to define a Poisson process $X_j \sim \text{PP}(G)$, with a shared completely random measure $G$ on $\Omega$, such that $X_j(A) \sim \text{Pois}(G(A))$ for each $A \subset \Omega$. Denote $G(\Omega) = \sum_{q=1}^Q G(A_q)$ and $\tilde{G} = G/G(\Omega)$. Following Lemma 4.1 of [5], the joint distributions of $X_j(\Omega), X_j(A_1), \cdots, X_j(A_Q)$ are equivalent under the following two expressions:

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

$$X_j(\Omega) \sim \text{Poisson}(G(\Omega)), \quad [X_j(A_1), \cdots, X_j(A_Q)] \sim \text{Mult}(X_j(\Omega); \tilde{G}(A_1), \cdots, \tilde{G}(A_Q)).$$

Thus the Poisson process provides not only a way to generate independent counts from each $A_q$, but also a mechanism for mixture modeling, which allocates the observations into any measurable disjoint partition $\{A_q\}_{1,Q}$ of $\Omega$, conditioning on $X_j(\Omega)$ and the normalized mean measure $\tilde{G}$.

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

2 Augment-and-Conquer the Negative Binomial Distribution

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

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

2
Both the gamma-Poisson and compound-Poisson augmentations of the NB distribution and Lemmas 2.1 and 2.2 are key ingredients of this paper. We will show that these augment-and-concur methods not only unite count and mixture modeling and provide efficient inference, but also, as shown in Section 3, let us examine the posteriors to understand fundamental properties of the NB processes, clearly revealing connections to previous nonparametric Bayesian mixture models.

**Lemma 2.1.** Denote \( s(m,j) \) as Stirling numbers of the first kind [17]. Augment \( m \sim \text{NB}(r,p) \) under the compound Poisson representation as \( m \sim \sum_{t=1}^{l'} \text{Log}(p), \ l \sim \text{Pois}(-r \ln(1-p)) \), then the conditional posterior of \( l \) has PMF

\[
\Pr(l = j | m, r) = \frac{r^j}{(r+m)^j} s(m,j) j! \quad j = 0, 1, \ldots , m.
\]  

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

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

\[
m \sim \sum_{t=1}^{l'} \text{Log}(p), \ l \sim \sum_{t=1}^{l'} \text{Log}(p'), \ l' \sim \text{Pois}(-r_1 \ln(1-p')).
\]  

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

### 3 Gamma-Negative-Binomial Process

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

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

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

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

\[
L = \sum_j L_j \sim \sum_{t=1}^{L'} \text{Log}(p'), \ L' \sim \text{PP}(-G_0 \ln(1-p')), \ p' = -\frac{\ln(1-p)}{c_1-\ln(1-p)}.
\]  

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

\[
\Lambda_j(G, X_j, p_j) \sim \text{GaP}(1/p_j, G + X_j).
\]

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

\[
L_j | X_j, G \sim \text{CRTP}(X_j,G), \ L' | L, G_0 \sim \text{CRTP}(L,G_0).
\]  

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

\[
L(\Omega), G_0 = \sum_{\omega \in \Omega} \delta(\omega) > 0 = \sum_{\omega \in \Omega} \delta(\omega_j(\omega) > 0)
\]

which is equal to \( K^+ \), the total number of used discrete atoms; if \( G_0 \) is discrete as \( G_0 = \sum_{k=1}^{K^+} \delta_{\omega_k} \), then \( L'(\omega_k) = \text{CRT}(L(\omega_k),\frac{K^+}{K^+} \geq 1 \) if \( \sum_j X_j(\omega_k) > 0 \), thus \( L'(\Omega) \geq K^+ \). In either case, let \( \gamma_0 \sim \text{Gamma}(c_0,1/f_0) \), with the gamma Poisson conjugacy on (6) and (7), we have

\[
\gamma_0 | L'(\Omega), p' \sim \text{Gamma}(c_0+L'(\Omega), \frac{1}{f_0-\ln(1-p')}).
\]  

\[
G | G_0, \{L_j,p_j\} \sim \text{GaP}(c - \sum_j \ln(1-p_j), G_0 + \sum_j L_j).
\]  

Since the data \( \{x_{ji}\} \) are exchangeable within group \( j \), the predictive distribution of a point \( X_j \), conditioning on \( X_j^{-1} \neq \emptyset \), is given by

\[
X_j | G, X_j^{-1} \sim \frac{\text{E}[\Lambda_j(G,X_j^{-1})]}{\text{E}[\Lambda_j(G)]} = \frac{G}{G(\Omega) + X_j^{-1}} + \frac{X_j^{-1}}{G(\Omega) + X_j^{-1}}.
\]
3.1 Relationship with the hierarchical Dirichlet process

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

\[
X_{ji} \sim \tilde{\Lambda}_j, \tilde{\Lambda}_j \sim \text{DP}(\alpha, \tilde{G}), \alpha \sim \text{Gamma}(\gamma_0, 1/c), \tilde{G} \sim \text{DP}(\gamma_0, \tilde{G}_0)
\]

which is an HDP [7]. Thus the normalized gamma-NB process leads to an HDP, yet we cannot return from the HDP to the gamma-NB process without modeling \( X_j(\Omega) \) and \( \Lambda_j(\Omega) \) as random variables. Theoretically, they are distinct in that the gamma-NB process is a completely random measure, assigning independent random variables into any disjoint Borel sets \( \{A_q\}_{1,Q} \) of \( \Omega \), whereas the HDP is not. Practically, the gamma-NB process can exploit conjugacy to achieve analytical conditional posteriors for all latent parameters. The inference of the HDP is a major challenge and it is usually solved through alternative constructions such as the Chinese restaurant franchise (CRF) and stick-breaking representations [7, 23]. In particular, without analytical conditional posteriors, the inference of concentration parameters is difficult and it is usually solved through alternative constructions such as the Chinese restaurant franchise (CRF) and stick-breaking representations [7, 23].

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

For a finite continuous base measure, the gamma process \( G \sim \text{GaP}(c, G_0) \) can also be defined with its Lévy measure on a product space \( \mathbb{R}_+ \times \Omega \), expressed as \( \nu(drd\omega) = r^{-1}e^{-cr}drG_0(d\omega) \) [9]. Since the Poisson intensity \( \nu^+ = \nu(\mathbb{R}_+ \times \Omega) = \infty \) and \( \int_{\mathbb{R}_+ \times \Omega} \nu(drd\omega) \) is finite, a draw from this process can be expressed as \( G = \sum_{k=1}^{\infty} r_k \delta_{\omega_k}, (r_k, \omega_k) \sim \pi(drd\omega), \pi(drd\omega)\nu^+ \equiv \nu(drd\omega) \) [9]. Here we consider a discrete base measure as \( G_0 = \sum_{k=1}^{K} \gamma_k \delta_{\omega_k}, \omega_k \sim g_0(\omega_k) \), then we have \( G = \sum_{k=1}^{K} r_k \delta_{\omega_k}, r_k \sim \text{Gamma}(\gamma_0/K, 1/c), \omega_k \sim g_0(\omega_k) \), which becomes a draw from the gamma process with a continuous base measure as \( K \to \infty \). Let \( x_{ji} \sim F(\omega_{z_{ji}}) \) be observation \( i \) in group \( j \), linked to a mixture component \( \omega_{z_{ji}} \in \Omega \) through a distribution \( F \). Denote \( n_{jk} = \sum_{i=1}^{N_j} \delta(z_{ji} = k) \), we can express the gamma-NB process with the discrete base measure as

\[
\omega_k \sim g_0(\omega_k), N_j = \sum_{k=1}^{K} n_{jk}, n_{jk} \sim \text{Pois}(\lambda_{jk}), \lambda_{jk} \sim \text{Gamma}(r_k, p_j/(1-p_j))
\]

\[
r_k \sim \text{Gamma}(\gamma_0/K, 1/c), p_j \sim \text{Beta}(a_0, b_0), \gamma_0 \sim \text{Gamma}(\epsilon_0, 1/f_0)
\]

where marginally we have \( n_{jk} \sim \text{NB}(r_k, p_j) \). Using the equivalence between (1) and (2), we can equivalently express \( N_j \) and \( n_{jk} \) in the above model as \( N_j \sim \text{Pois}(\lambda_j), [n_{j1}, \cdots, n_{JK}] \sim \text{Mult}(N_j; \lambda_{j1}/\lambda_j, \cdots, \lambda_{JK}/\lambda_j) \), where \( \lambda_j = \sum_{k=1}^{K} \lambda_{jk} \). Since the data \( \{x_{ji}\}_{i=1,N} \) are fully exchangeable, rather than drawing \( [n_{j1}, \cdots, n_{JK}] \) once, we may equivalently draw the index

\[
z_{ji} \sim \text{Discrete}(\lambda_{j1}/\lambda_j, \cdots, \lambda_{JK}/\lambda_j)
\]
for each \( x_{ji} \) and then let \( n_{jk} = \sum_{i=1}^{N_j} \delta(z_{ji} = k) \). This provides further insights on how the seemingly disjoint problems of count and mixture modeling are united under the NB process framework. Following (8)-(12), the block Gibbs sampling is straightforward to write as

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

\[
(p_j | -) \sim \text{Beta} \left( a_0 + N_j, b_0 + \sum_k r_k \right), \quad p' = \frac{1}{c - \sum_j (1 - p_j)} (l_{jk} | -) \sim \text{CRT}(n_{jk}, r_k)
\]

\[
(l'_{k} | -) \sim \text{CRT} \left( \sum_j l_{jk}, \gamma_0 / K \right), \quad (\gamma_0 | -) \sim \text{Gamma} \left( 0 + \sum_k l'_{k}, \frac{1}{\sum_j (1 - p_j)} \right)
\]

\[
(r_{k} | -) \sim \text{Gamma} \left( \gamma_0 / K + \sum_j l_{jk}, \frac{1}{c - \sum_j (1 - p_j)} \right), \quad (\lambda_{jk} | -) \sim \text{Gamma}(r_k + n_{jk}, p_j).
\]

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

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

\[
z_{ji} \sim \text{Discrete}(\tilde{\lambda}_j), \quad \tilde{\lambda}_j \sim \text{Dir}(\alpha \tilde{r}), \quad \alpha \sim \text{Gamma}(\gamma_0, 1/c), \quad \tilde{r} \sim \text{Dir}(\gamma_0 / K, \ldots, \gamma_0 / K)
\]

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

4 The Negative Binomial Process Family and Related Algorithms

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

It is natural to let the probability measure be drawn from a beta process [25, 26], which can be defined by its Lévy measure on a product space \( [0, 1] \times \Omega \) as \( \nu(d\alpha d\omega) = cp^{-1}(1 - p)^{-1} d\alpha d\omega \). A draw from the beta process \( B \sim \text{BP}(c, B_0) \) with concentration parameter \( c \) and base measure \( B_0 \) can be expressed as \( B = \sum_{k=1}^{\infty} p_k \delta_{\omega_k} \). A beta-NB process [5, 6] can be constructed by letting \( X_j \sim \text{NB}(r_j, B) \), with a random draw expressed as \( X_j = \sum_{k=1}^{\infty} n_{jk} \delta_{\omega_k}, \quad n_{jk} \sim \text{NB}(r_{jk}, p_k) \). Under this construction, the NB probability measure is shared and the NB dispersion parameters are group dependent. As in [5], we may also consider a marked-beta-NB\(^1\) process that both the NB probability and dispersion measures are shared, in which each point of the beta process is marked with an independent gamma random variable. Thus a draw from the marked-beta process becomes \( (R, B) = \sum_{k=1}^{\infty} (r_k, p_k) \delta_{\omega_k} \), and the NB process \( X_j \sim \text{NB}(R, B) \) becomes \( X_j = \sum_{k=1}^{\infty} n_{jk} \delta_{\omega_k}, \quad n_{jk} \sim \text{NB}(r_{jk}, p_k) \). Since the beta and NB processes are conjugate, the posterior of \( B \) is tractable, as shown in [5, 6]. If it is believed that there are excessive number of zeros, governed by a process other than the NB process, we may introduce a zero inflated NB process as \( X_j \sim \text{NB}(ZR_{jk}, p_j) \), where \( Z_{jk} \sim \text{BeP}(B) \) is drawn from the Bernoulli process [26] and \( (R, B) = \sum_{k=1}^{\infty} (r_k, p_k) \delta_{\omega_k} \) is drawn from a marked-beta process, thus \( n_{jk} \sim \text{NB}(r_{jk}, p_j) \). This construction can be linked to the model in [27] with appropriate normalization, with advantages that there is no need to fix \( p_j = 0.5 \) and the inference is fully tractable.

4.1 Related Algorithms

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

\(^1\)We may also consider a beta marked-gamma-NB process, whose performance is found to be very similar.
Table 1: A variety of negative binomial processes are constructed with distinct sharing mechanisms, reflected with which parameters from \( r_k, r_j, p_k, p_j \) and \( \pi_k \) (\( b_{jk} \)) are inferred (indicated by a check-mark \( ✓ \)), and the implied VMR and ODL for counts \( \{n_{jk}\}_{j,k} \). They are applied for topic modeling of a document corpus, a typical example of mixture modeling of grouped data. Related algorithms are shown in the last column.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>( r_k )</th>
<th>( r_j )</th>
<th>( p_k )</th>
<th>( p_j )</th>
<th>VMR</th>
<th>ODL</th>
<th>Related Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB-LDA</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>((1 - p_j)^{-1})</td>
<td>((1 - p_k)^{-1})</td>
<td>LDA [32], Dir-PFA [5]</td>
</tr>
<tr>
<td>NB-HDP</td>
<td>✓</td>
<td>✓</td>
<td>0.5</td>
<td></td>
<td>( r_j^{-1} )</td>
<td>( r_k^{-1} )</td>
<td>HDP[7], DILN-HDP [12]</td>
</tr>
<tr>
<td>NB-FTM</td>
<td>✓</td>
<td>✓</td>
<td>0.5</td>
<td>✓</td>
<td>((1 - p_j)^{-1})</td>
<td>((1 - p_k)^{-1})</td>
<td>FTM [27], S-T-G-PFA [5]</td>
</tr>
<tr>
<td>Beta-NB</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>((1 - p_j)^{-1})</td>
<td>( p_j^{-1} )</td>
<td>BNBP [5], BNBP [6]</td>
</tr>
<tr>
<td>Gamma-NB</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>((1 - p_j)^{-1})</td>
<td>( p_j^{-1} )</td>
<td>CRF-HDP [7, 24]</td>
</tr>
<tr>
<td>Marked-Beta-NB</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>((1 - p_k)^{-1})</td>
<td>( p_k^{-1} )</td>
<td>BNBP [5]</td>
</tr>
</tbody>
</table>

Motivated by Table 1, we consider topic modeling using a variety of NB processes, which differ on which parameters are learned and consequently how the VMR and ODL of the latent counts \( \{n_{jk}\}_{j,k} \) are modeled. We compare them with LDA [32, 35] and CRF-HDP [7, 24]. For fair comparison, they are all implemented with block Gibbs sampling using a discrete base measure with \( K \) atoms, and for the first fifty iterations, the Gamma-NB process with \( r_k = 50/K \) and \( p_j = 0.5 \) is used for initialization. For LDA and NB-LDA, we search \( K \) for optimal performance and for the other models, we set \( K = 400 \) as an upper-bound. We set the parameters as \( c = 1, \eta = 0.05 \) and \( a_0 = b_0 = c_0 = f_0 = 0.01 \). For LDA, we set the topic proportion Dirichlet smoothing parameter as \( 50/K \), following the topic model toolbox provided for [35]. We consider 2500 Gibbs sampling iterations, with the last 1500 samples collected. Under the NB processes, each word \( x_{ji} \) would...
be assigned to a topic $k$ based on both $F(x_{ji}; \omega_k)$ and the topic weights $\{\lambda_{jk}\}_{k=1}^K$; each topic is drawn from a Dirichlet base measure as $\omega_k \sim \text{Dir}(\eta, \cdots, \eta) \in \mathbb{R}^V$, where $V$ is the number of unique terms in the vocabulary and $\eta$ is a smoothing parameter. Let $v_{ji}$ denote the location of word $x_{ji}$ in the vocabulary, then we have $(\omega_k | -) \sim \text{Dir}(\eta + \sum_j \sum_i \delta(z_{ji} = k), v_{ji} = 1), \cdots, \eta + \sum_j \sum_i \delta(z_{ji} = k), v_{ji} = V)$. We consider the Psychological Review corpus, restricting the vocabulary to terms that occur in five or more documents. The corpus includes 1281 abstracts from 1967 to 2003, with 2,566 unique terms and 71,279 total word counts. We randomly select 20%, 40%, 60% or 80% of the words from each document to learn a document dependent probability for each term $v$ as $f_{ji} = \sum_{s=1}^S \sum_{k=1}^K \omega_{vk} \lambda_{jk}^{(s)} / \sum_{s=1}^S \sum_{k=1}^K \sum_{v\in\mathcal{S}} \omega_{vk} \lambda_{jk}^{(s)}$, where $\omega_{vk}$ is the probability of term $v$ in topic $k$ and $S$ is the total number of collected samples. We use $\{f_{ji}\}_{j,v}$ to calculate the per-word perplexity on the held-out words as in [5]. The final results are averaged from five random training/testing partitions. Note that the perplexity per test word is the fair metric to compare topic models. However, when the actual Poisson rates or distribution parameters for counts instead of the mixture proportions are of interest, it is obvious that a NB process based joint count and mixture model would be more appropriate than an HDP based mixture model.

Figure 1 compares the performance of various algorithms. The Marked-Beta-NB process has the best performance, closely followed by the Gamma-NB process, CRF-HDP and Beta-NB process. With an appropriate $K$, the parametric NB-LDA may outperform the nonparametric NB-HDP and NB-FTM as the training data percentage increases, somewhat unexpected but very intuitive results, showing that even by learning both the NB dispersion and probability parameters $r_j$ and $p_j$ in a document dependent manner, we may get better data fitting than using nonparametric models that share the NB dispersion parameters $r_k$ across documents, but fix the NB probability parameters.

Figure 2 shows the learned model parameters by various algorithms under the NB process framework, revealing distinct sharing mechanisms and model properties. When $(r_j, p_j)$ is used, as in the NB-LDA, different documents are weakly coupled with $r_j \sim \text{Gamma}(\gamma_0, 1/c)$, and the modeling results show that a typical document in this corpus usually has a small $r_j$ and a large $p_j$, thus a large ODL and a large VMR, indicating highly overdispersed counts on its topic usage. When $(r_j, p_k)$ is used to model the latent counts $\{n_{jk}\}_{j,k}$, as in the Beta-NB process, the transition between active and non-active topics is very sharp that $p_k$ is either close to one or close to zero. That is because $p_k$ controls the mean as $\mathbb{E}[\sum_j n_{jk}] = p_k/(1 - p_k) \sum_j r_j$ and the VMR as $(1 - p_k)^{-1}$ on topic $k$, thus a popular topic must also have large $p_k$ and thus large overdispersion measured by the VMR; since the counts $\{n_{jk}\}_{j}$ are usually overdispersed, particularly true in this corpus, a middle range $p_k$ indicating an appreciable mean and small overdispersion is not favored by the model and thus is rarely observed. When $(r_k, p_j)$ is used, as in the Gamma NB-process, the transition is much smoother that $r_k$ gradually decreases. The reason is that $r_k$ controls the mean as $\mathbb{E}[\sum_j n_{jk}] = r_k \sum_j p_j / (1 - p_j)$ and the ODL $r_k^{-1}$ on topic $k$, thus popular topics must also have large $r_k$ and thus small overdispersion measured by the ODL, and unpopular topics are modeled with small $r_k$ and thus large overdispersion, allowing rarely and lightly used topics. Therefore, we can expect that $(r_k, p_j)$ would allow...
more topics than \((r_j, p_k)\), as confirmed in Figure 1 (a) that the Gamma-NB process learns 177 active topics, significantly more than the 107 ones of the Beta-NB process. With these analysis, we can conclude that the mean and the amount of overdispersion (measure by the VMR or ODL) for the usage of topic \(k\) is positively correlated under \((r_j, p_k)\) and negatively correlated under \((r_k, p_k)\).

When \((r_k, p_k)\) is used, as in the Marked-Beta-NB process, more diverse combinations of mean and overdispersion would be allowed as both \(r_k\) and \(p_k\) are now responsible for the mean \(E[\sum_j n_{jk}] = J r_k p_k / (1 - p_k)\). For example, there could be not only large mean and small overdispersion (large \(r_k\) and small \(p_k\)), but also large mean and large overdispersion (small \(r_k\) and large \(p_k\)). Thus \((r_k, p_k)\) may combine the advantages of using only \(r_k\) or \(p_k\) to model topic \(k\), as confirmed by the superior performance of the Marked-Beta-NB over the Beta-NB and Gamma-NB processes. When \((r_k, \pi_k)\) is used, as in the NB-FTM model, our results show that we usually have a small \(\pi_k\) and a large \(r_k\), indicating topic \(k\) is sparsely used across the documents but once it is used, the amount of variation on usage is small. This modeling properties might be helpful when there are excessive number of zeros which might not be well modeled by the NB process alone. In our experiments, we find the more direct approaches of using \(p_k\) or \(p_j\) generally yield better results, but this might not be the case when excessive number of zeros are better explained with the underlying beta-Bernoulli or IBP processes, e.g., when the training words are scarce.

It is also interesting to compare the Gamma-NB and NB-HDP. From a mixture-modeling viewpoint, fixing \(p_j = 0.5\) is natural as \(p_j\) becomes irrelevant after normalization. However, from a count modeling viewpoint, this would make a restrictive assumption that each count vector \(\{n_{jk}\}_{k=1,K}\) has the same VMR of 2, and the experimental results in Figure 1 confirm the importance of learning \(p_j\) together with \(r_k\). It is also interesting to examine (15), which can be viewed as the concentration parameter \(\alpha\) in the HDP, allowing the adjustment of \(p_j\) would allow a more flexible model assumption on the amount of variations between the topic proportions, and thus potentially better data fitting.

6 Conclusions

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

Acknowledgments

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