Augment-and-Conquer Negative Binomial Processes

Mingyuan Zhou Lawrence Carin

Department of Electrical & Computer Engineering

Duke University, Durham, NC 27708, USA

Introduction

- We develop data augmentation methods unique to the negative binomial (NB) distribution.
- We perform joint count and mixture modeling under the NB process, using completely random measures (Poisson process, gamma process, beta process, NB process) 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.
- 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.

Poisson Process for Count & Mixture Modeling

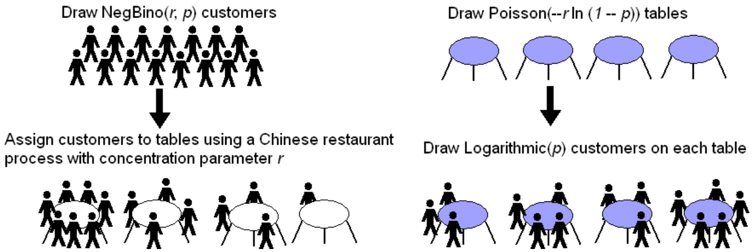
- Mixture modeling infers probability random measures to assign data points into clusters (mixture components).
- Since the number of points assigned to clusters are counts, we consider mixture modeling as a count-modeling problem.
- The NB distribution, parameterized by a dispersion parameter and a probability parameter, is used to model overdispersed counts.
- The NB distribution is a gamma-Poisson mixture distribution (for hierarchical construction), and can also be represented as a compound Poisson distribution (for inference tractability).
- The Poisson process provides not only a way to generate independent counts from each partition of the space, but also a mechanism for mixture modeling, which allocates the observations into any measurable disjoint partition of space, conditioning on the total count and the normalized mean measure.

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

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

Poisson-Logarithmic bivariate count distribution

The joint distribution of the customer count and table count are equivalent:



Augment-and-Conquer the Negative Binomial Distribution

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

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

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

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

Gamma-Negative Binomial Process

Gamma-Negative Binomial Process

$$X_j\sim \mathrm{NBP}(G,p_j), \ G\sim \ \mathrm{GaP}(c,G_0)$$

$$X_j\sim \mathrm{PP}(\Lambda_j), \ \ \Lambda_j\sim\mathrm{GaP}((1-p_j)/p_j,G), \ \ G\sim\mathrm{GaP}(c,G_0)$$

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

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

Posterior Analysis

$$\Lambda_j|G,X_j,p_j\sim\mathrm{GaP}(1/p_j,G+X_j) \qquad (p_j|-)\sim\mathrm{Beta}\left(a_0+X_j(\Omega),b_0+G(\Omega)\right)$$

$$L_j|X_j,G\sim\mathrm{CRTP}(X_j,G), \quad L'|L,G_0\sim\mathrm{CRTP}(L,G_0)$$

$$\gamma_0|\{L'(\Omega),p'\}\sim\mathrm{Gamma}\left(e_0+L'(\Omega),\frac{1}{f_0-\ln(1-p')}\right)$$

$$G|G_0,\{L_j,p_j\}\sim\mathrm{GaP}\left(c-\sum_j\ln(1-p_j),G_0+\sum_j L_j\right)$$

Predictive Distribution

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

Related to the Hierarchical Dirichlet Process

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

The NB Process Family and Related Algorithms

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 π_k (b_{jk}) are inferred (indicated by a check-mark ✓), and the implied VMR and ODL for counts $\{n_{jk}\}_{j,k}$. They are applied for topic modeling of a document corpus, a typical example of mixture modeling of grouped data. Related algorithms are shown in the last column.

Algorithms	r_k	r_j	p_k	p_j	π_k	VMR	ODL	Related Algorithms
NB-LDA		✓		✓		$(1-p_j)^{-1}$	r_j^{-1}	LDA [32], Dir-PFA [5]
NB-HDP	✓			0.5		2	r_k^{-1}	HDP[7], DILN-HDP [12]
NB-FTM	✓			0.5	✓	2	$(r_k)^{-1}b_{jk}$	FTM [27], S$\gamma\Gamma$-PFA [5]
Beta-NB		✓	✓			$(1-p_k)^{-1}$	r_j^{-1}	BNBP [5], BNBP [6]
Gamma-NB	✓			✓		$(1-p_j)^{-1}$	r_k^{-1}	CRF-HDP [7, 24]
Marked-Beta-NB	✓		✓			$(1-p_k)^{-1}$	r_k^{-1}	BNBP [5]

Beta-NB $n_{jk}\sim\mathrm{NB}(r_j,p_k)$ **NB-FTM** $n_{jk}\sim\mathrm{NB}(r_kb_{jk},p_j)$

Gamma-NB $n_{jk}\sim\mathrm{NB}(r_k,p_j)$ **Marked-Beta-NB** $n_{jk}\sim\mathrm{NB}(r_k,p_k)$

Topic Modeling (Mixture Modeling of Grouped Words)

Comparison of Per-Word Perplexities

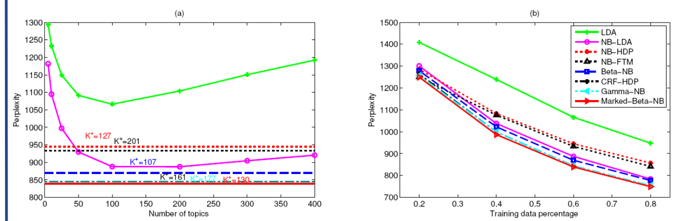


Figure 1: Comparison of per-word perplexities on the held-out words between various algorithms. (a) With 60% of the words in each document used for training, the performance varies as a function of K in both LDA and NB-LDA, which are parametric models, whereas the NB-HDP, NB-FTM, Beta-NB, CRF-HDP, Gamma-NB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

Inferred NB Process Model Parameters

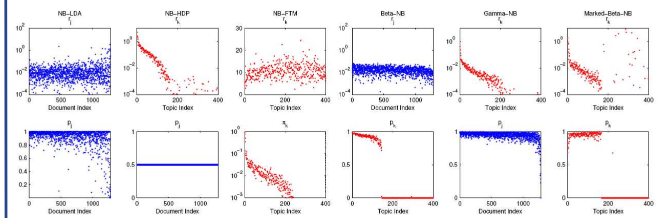


Figure 2: Distinct sharing mechanisms and model properties are evident between various NB processes, by comparing their inferred parameters. Note that the transition between active and non-active topics is very sharp when p_k is used and much smoother when r_k is used. Both the documents and topics are ordered in a decreasing order based on the number of words associated with each of them. These results are based on the last Gibbs sampling iteration. The values are shown in either linear or log scales for convenient visualization.