Implementing the Collapsed Gibbs Sampler for Latent Dirichlet Allocation

Steve Bronder
sab2287@columbia.edu

Antonio Moretti
am4134@columbia.edu

May 17, 2016

Abstract

Latent Dirichlet Allocation is a two stage hierarchical clustering process that is typically fit through MCMC or variational inference. In this note we describe how the collapsed Gibbs sampler is derived and implemented, including the method of integrating out multinomial parameters as an example of Rao-Blackwellization. This allows for all samples to be drawn from simple conditional distributions while also supporting the update of each topic distribution after each word is assigned to a particular topic. We demonstrate our implementation in Fortran by performing posterior inference on a collection of documents from JSTOR. We review a few open problems and discuss heuristics to check rates of convergence.

1 Introduction

Topic models are a family of unsupervised machine learning methods for summarizing and organizing large collections of documents. Topic models aim to mimic the writing process in which an author draws upon a set of topics based on the focus of the narrative. Words represent various topics and provide clues about the underlying themes that comprise the document. Introduced by Blei in 2003, Latent Dirichlet Allocation (LDA) has become a widely popular topic model for text processing due to its simplicity [2]. The same model is derived in 2000 by Pritchard for isolating microbial species using genotype data [12]. Pritchard's motivation is to describe the genome of an individual or species as a mixture of various populations. At its core, LDA is a two level Bayesian heirarchical model for clustering with hidden factors. Formally, a topic is defined as a hidden probability distribution over a fixed vocabulary. Documents are modeled as mixtures of topic distributions, which in turn are modeled as distributions over words. This two stage hierarchical structure supports a soft rather than a hard assignment of each document to multiple topics where words are the only variables observed.

In this note we illustrate how to implement the collapsed Gibbs sampler for LDA. The rest of this document is organized as follows. We introduce the LDA model in Section 2. In Section 3 we discuss the process of learning parameters and the challenge of posterior inference to motivate the MCMC approach. Section 4 describes the collapsed Gibbs sampler for LDA as an application of Rao-Blackwellization including a brief summary of the Hammersley-Clifford theorem. We discuss our implementation and experimental results in Section 5. Section 6 concludes by briefly addressing a few open questions. Mathematical derivations and basic results are provided for completeness along with output and figures in Section 7.

2 LDA Model

In order to state the model we will introduce the notation below. Following Carpenter [3], let K be the number of topics and let M be the number of documents. Let N_m be the number of words in the mth document and let J be the distinct number of words in the corpus. Let $W_{m,n}$ be the observed document term matrix where rows represent documents, columns represent words and the components are the counts of the nth word in the mth document. Let $Z_{m,n} \in 1:K$ be the topic to which the word of $W_{m,n}$ is assigned. Let $\theta_m \in [0,1]^K$ be the topic distribution for the mth document and $\phi_k \in [0,1]^J$ be the word distribution for the kth topic. The hyperparameter $\alpha \in \mathbb{R}^K$ is a vector of prior counts for topics in documents. Similarly $\beta \in \mathbb{R}^J$ is a vector of prior counts for words in a given topic. LDA is defined by the following generative model.

1. For each topic k draw a word distribution:
   (a) $\phi_k \sim Dirichlet(\beta)$
2. For each document m:
   (a) $\theta_m \sim Dirichlet(\alpha)$
   (b) For each of the n words in m:
      i. $z_{m,n} \sim Mult(\theta_m)$
      ii. $w_{m,n} \sim Mult(\phi_{z_{m,n}})$

The joint distribution of the graphical model factorizes as products of the conditional distributions:

$$p(\theta|\alpha)p(z|\theta)p(w|z,\phi)p(\phi|\beta) \tag{1}$$

Define $c_{k,m,j}$ to be the count for the number of times word j is assigned to topic k in document m:

$$c_{k,m,j} = \sum_{n=1}^{Nm} I\Big(z_{m,n} = k \;\&\; w_{m,n} = j\Big)$$

Note that we can marginalize the distribution of counts over any of the three variables k, m and j. For example,

$$c_{k,m,*} = \sum_{j=1}^{J} c_{k,m,j}$$

It will be useful to recast the above using linear algebra. Let NMZ be a matrix of M rows and Z columns whose components denote the number of times document M and topic Z occur together. Similarly, let NZW represent a matrix of Z rows and W columns whose components denote the number of times topic Z and word W interact. Marginalizing the above with respect to W and Z give vectors NM and NZ respectively.

3 Posterior Inference for LDA

The learning task for LDA is to compute the posterior distribution which is the conditional distribution of hidden variables given the observations:

$$P(\theta, z, \phi|w, \alpha, \beta) \propto \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{m=1}^{M} p(\theta_m|\alpha) \prod_{n=1}^{N} p(z_{m,n}|\theta_m)p(w_{m,n}|z_{m,n}, \phi_k) \tag{2}$$

Recall the first two terms are Dirichlet. The third term represents a draw of the topic assignment θ_d for each of the N_m words in the document. The last term represents the likelihood of drawing a word given the joint distribution of observations and hidden variables. It is easy to see that evaluating the above presents computational challenges. Consider for simplicity that we are dealing with one document where the topics ϕ_k are fixed. The per document posterior is then proportional to the following.

$$p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\phi_k) \tag{3}$$

In order to ensure that the above is normalized it is necessary to compute the evidence.

$$\int_\theta p(\theta|\alpha)\prod_{n=1}^{N}\sum_{z=1}^{K} p(z_n|\theta)p(w_n|z_n,\phi_k)d\theta \tag{4}$$

The above is a hypergeometric function shown by Dickey to be intractable [5]. One solution is to move the summation outside the integral to obtain the form of the evidence below:

$$\sum_{z=1}^{K}\int_\theta p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\phi_k)d\theta \tag{5}$$

The simplified form is equivalent to the sum of N^k tractable Dirichlet integrals, however when k is reasonably large it is not practical to compute the solution exactly. Even without or multiple documents, we are forced to compute an exponential number of Dirichlet integrals. Two solutions include the use of approximate inference algorithms such as MCMC via the Gibbs sampler or variational methods such as mean field variational inference [14].

4 The Gibbs Sampler

MCMC methods exploit the fundamental theorem of Markov chains to sample from an intractible distribution such as (2). The MCMC approach is to design an irreducible, aperiodic and positive recurrent chain whose stationary distribution is the posterior distribution of interest. We will use the space of all configurations of hidden variables to represent the state space of a Markov chain. Recall that the Gibbs sampler approach is to iteratively sample from the distribution of one hidden variable at a time conditioned on the observations and the current state of each of the other hidden variables. After a suitable burn-in period, collecting samples will result in a Monte Carlo estimate of the posterior.

A naive sampler would condition on all of the hidden variables, however it is possible to integrate out the multinomial parameters for a faster mixing chain. We summarize the necessary derivations in detail in the Appendix. We note here that by using conditional expectations we can improve the variance of the Monte Carlo estimate.

$$Var(\mathbb{E}[\delta(X)|Y]) \leq Var(\delta(X)) \tag{6}$$

We review some basic properties of the Gibbs sampler and Rao-Blackwellization below.

Definition 4.1. The positivity condition for a joint distribution $f(X_1,\cdots,X_p)$ with marginal densities $f_{X_i}(x_i)$ is satisfied if the following holds: $f_{X_i}(x_i) > 0$ for all $(X_1,\cdots,X_p)$ implies that $f(X_1,\cdots,X_p) > 0$.

Theorem 4.1 (Hammersley-Clifford)**.** *Suppose that* $(X_1, \cdots, X_d)$ *satisfies the positivity condition with joint pdf* $f(X_1, \cdots, X_p)$*. For all* $(Y_1, \cdots Y_p) \in supp(f)$

$$f(X_1, \cdots, X_p) \propto \prod_{i=1}^{P} \frac{f_{X_i}|X_{-i}(X_i|X_1, \cdots X_{i-1}, Y_{i+1}, \cdots Y_p)}{f_{X_i}|X_{-i}(Y_i|X_1, \cdots X_{i-1}, Y_{i+1}, \cdots Y_p)} \tag{7}$$

Proof.

$$f(X_1, \cdots, X_{p-1}, X_p) = f_{X_p}(X_p|X_1, \cdots X_{p-1}) f(X_1, \cdots X_{p-1}) \tag{8}$$

Similarly we have that

$$f(X_1, \cdots, X_{p-1}, Y_p) = f_{X_p}(Y_p|X_1, \cdots X_{p-1}) f(X_1, \cdots X_{p-1}) \tag{9}$$

By equation (5) we can write the following.

$$f(X_1, \cdots, X_p) = f(X_1, \cdots, X_{p-1}) f_{X_p}(X_p|X_1, \cdots X_{p-1}) \tag{10}$$

Rewrite the joint distribtion on the RHS using equation (6).

$$= f(X_1, \cdots, X_{p-1}, Y_p) \frac{f_{X_p}(X_p|X_1, \cdots X_{p-1})}{f_{X_p}(Y_p|X_1, \cdots X_{p-1})} \tag{11}$$

Repeating this process of rewriting the joint in terms of conditionals:

$$= \cdots \tag{12}$$

$$= f(X_1, \cdots, X_p) \frac{f_{X_p}(X_p|Y_1, \cdots Y_p)}{f_{X_p}(Y_p|Y_1, \cdots Y_p)} \cdots \frac{f_{X_p}(X_p|X_1, \cdots X_{p-1})}{f_{X_p}(Y_p|X_1, \cdots X_{p-1})} \tag{13}$$

□

The Hammersley-Clifford theorem specifies when a distribution can be factorized in terms of cliques. In order to make use of Rao-Blackwellization for the Gibbs sampler, we require the joint distribution to be uniquely specified by conditional distributions.

Lemma 4.2. *The transition probability matrix of the Gibbs sampler is*

$$\begin{aligned} K(x^{(t-1)}, x^{(t)}) = f_{X_1|X_{-1}}(x_1|x_1^{t-1}, \cdots x_p^{t-1}) \times f_{X_2|X_{-2}}(x_2^t|x_1^t, x_3^{t-1} \cdots x_p^{t-1}) \\ \times \cdots \times f_{X_p|X_{-p}}(x_p^t|x_1^t, x_{p_1}^t, \cdots x_p^{t-1}) \end{aligned} \tag{14}$$

Proof.

$$P(X^t \in \mathcal{X}|X^{t-1} = x^{t-1}) = \int_{\mathcal{X}} f_{X_t|X_{t-1}}(x^t|x^{t-1}) dx^t \tag{15}$$

$$\begin{aligned} = \int_{\mathcal{X}} f_{X_1|X_{-1}}(x_1|x_1^{t-1}, \cdots x_p^{t-1}) \times f_{X_2|X_{-2}}(x_2^t|x_1^t, x_3^{t-1} \cdots x_p^{t-1}) \\ \times \cdots \times f_{X_p|X_{-p}}(x_p^t|x_1^t, x_{p_1}^t, \cdots x_p^{t-1}) dx^t \end{aligned} \tag{16}$$

□

There are two rules that are required to create an MCMC sampler, one to propose how to move from x to y, and another rule to accept that move as the new position. In the context of the Gibbs sampler, suppose the stationary distribution $X = (X_1, \ldots, X_d)$ is

$$\pi(X_1, \ldots, X_d) = g(X_1, X_2, \ldots, X_d)/\kappa \tag{17}$$

where κ is a normalizing constant. Given the current value $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, for the next move we follow the steps below.

1. Assign $J \sim U(1, 2, \ldots, n)$.
2. Sample X_j with the proposal distribution $\pi(\tilde{x}_J | x_1, \ldots, x_{J-1}, x_{J+1}, \ldots, x_d)$ where $\tilde{x}_J$ is our proposed move from the first rule.

Unlike some MCMC samplers, the Gibbs sampler accepts this next move with probability one. The update equation for the onditional distribution used for Gibbs sampling is given below according to (32) derived in the Appendix:

$$p(z_{a,b} | z_{-(a,b)}, y, \alpha, \beta) = \frac{(\alpha_{z_{a,b}} + c + {}^{-(a,b)}_{z_{a,b},a,*}) \times (\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*}) \big/ (J \times \beta_j + c^{-(a,b)}_{z_{a,b},*,*})}{\sum_{k=1}^{K} (\alpha_{z_{a,b}} + c + {}^{-(a,b)}_{z_{a,b},a,*}) \times (\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*}) \big/ (J \times \beta_j + c^{-(a,b)}_{z_{a,b},*,*})} \tag{18}$$

5 Implementation and Experimental Results

The implementation of the collapsed Gibbs sampler is relatively simple and described in detail in Algorithm 1. High level pseudocode can be found in Heinrich [9]. During initialization, a random topic is assigned for each word of each document and the corresponding counts in the matrices NMZ, NM, NZ and NZW are incremented. The Gibbs sampler itself proceeds as follows. For each iteration, document, and every word in the document, matrices representing the counts by document and topic as well as word and topic are decremented. Then the probability of moving to another topic is selected, and a realization is made from a multinomial distribution as the new topic for that word. The new topic is then added to document and word count matrices. Due to the large amount of looping necessary in the algorithm an implementation was developed in Fortran 95 as higher level languages such as Python or matlab could not satisfy time constraints when working with large matrices[1].

Data was gathered on 200 journal entries submitted to the Journal of the American Statistical Association. Word counts for journal entries were made available by JSTOR's Data for Research program. With five topics selected, the topic model ran over the document term matrix for 30,000 iterations, burning in the first 18,000 iterations. Performing Geweke's [7] diagnostic test on the likelihood, as seen in figure 1, with the first window being at the tenth percentile and the second window at the fiftieth yielded a t-score of 1.504, enough to feel satisfied that the model has converged after this many iterations. Investigating the chain using Heidel's criterion [8] we receive a p-value for the stationarity test of .191 and pass the half-width test, which means we do not reject the null hypothesis of stationarity and have enough samples to accurately estimate the mean.

Table one shows the top 10 words for each topic. We can see that each topic tends to follow a pattern in the statistical literature. For instance, topic three is associated with words about

[1]The code and Python notebook associated with this research can be found at `https://github.com/stevo15025/topicmodelpy`

Table 1: The top 10 words for each topic

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	inconsistent	hildreth	adapted	zin	presumably
2	ext	affects	snedecor	harvey	stop
3	algebra	ath	exposition	pooled	paired
4	advantages	lag	gets	eliminate	indicators
5	benjamin	requiring	averaging	diagram	exception
6	thompson	heuristic	bls	singular	computable
7	continues	classified	mcleod	load	locally
8	sectors	alternatively	anything	lhs	markets
9	width	cal	marvin	carlin	empirically
10	originally	lot	payments	friedman	adler

approximation while topic four requires linear algebra and singular value decomposition. We can also see topic five may hold words associated with econometrics.

Tables two through five in the appendix show the articles that have the highest probability of being in each topic. Note how topic one seems to be comprised of topics on sampling procedures and estimating significance. One thing we see in all topics is how frequent back matter, letters to the editor, and book reviews matter. This may be due to these papers in particular journals summing up large breaths of what was popular at that particular time in statistical literature. Also note that some articles such as "Exact Distribution of the Sum of Independent Identically Distributed Discrete Random Variables" appear multiple times. This is due to the probabilistic nature of the topic model algorithm. As each article has a certain probability of being in a topic, some articles will be general enough to fit in one or more topics.

6 Conclusion and Future Work

One important area of research that has not seen much focus is dealing with memory issues when study convergence properties for these large models. At every iteration of a topic model you receive back two matrices, one of size documents by topics and another by topics and words. If you have one thousand documents and thirty thousand words it is very easily to run into memory out of bounds errors as only 10 iterations can take up a significant portion of memory. Implementing chunking methods as well as online models for the convergence procedure, for instance an online ARMA model, to reduce memory consumption would allow researchers to investigate convergence properties of these models much more clearly.

Another open question about MCMC methods concerns whether or not theoretical bounds can be constructed that address how long to run the chain. Rates of convergence between the transition kernel and the stationary distribution $K^n(x, y) \to \pi$ are usually measured by total variation distance:

$$\|K_x^n - \pi\|_{TV} = \frac{1}{2} \sum_y |K^n(x, y) - \pi(y)| = \max_{A \subseteq \mathcal{X}} |K^n(x, A) - \pi(A)| \tag{19}$$

A natural question is whether or not we can derive eigenvalue bounds on rates of convergence by using a reversible chain Gibbs sampler. Note that the Gibbs sampler implemented here

is not reversible. That is, $\pi(x)K(x,y) \neq \pi(y)K(y,x)$. Liu (1995) describes the following reversible chain:

1. Sample an index j uniformly form the distribution on $\{1, \cdots, p\}$
2. Sample $X_j^{(t)} \sim f_{X_j|X_{-j}}(\cdot|X_1^{(t-1)}, \cdots, X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)}, \cdot X_p^{(t-1)})$ setting $X_i^{(t)} = X_i^{(t-1)}$ $\forall i \neq j$.

For the reversible case, $Kg(x) = \sum g(y)K(x,y)$ and so the Markov kernel is a self adjoint operator $\langle Kg, h\rangle = \langle g, Kh\rangle$. We can then factorize the transition kernel using a spectral decomposition. Let ψ_i be the eigenvectors and β_i be the corresponding eigenvalues such that $K\psi_i = \beta\psi_i$ for $0 \leq i \leq |\mathcal{X}| - 1$:

$$K(x,y) = \pi(y) \sum_{i=0}^{|\mathcal{X}|-1} \beta_i \psi_i(x)\psi_i(y) \qquad K^n(x,y) = \pi(y) \sum_{i=0}^{|\mathcal{X}|-1} \beta_i^n \psi_i(x)\psi_i(y) \tag{20}$$

Diaconis shows that we can derive an eigenvalue bound using the Cauchy-Schwartz inequality:

$$4\|K_x^n - \pi\|_{TV}^2 \leq \sum_y \frac{(K^n(x,y) - \pi(y))^2}{\pi(y)} = \sum_{i=0}^{|\mathcal{X}|-1} \beta_i^{2n} \psi_i^2(x) \tag{21}$$

Several other heuristics exist for MCMC convergence rates and NLP models. Two examples the perplexity score which is a function of entropy and quantifies how well a distribution fits a sample $2^{H(p)} = 2^{-\sum p(x) log p(x)}$, and the Gelman-Rubin test. In future work we wish to explore various criteria for model convergence.

One approach that is not discussed here is that of variational inference. The idea here is to select a family of distributions over the hidden variables and to find corresponding parameters that make the variational parameters as close as possible to the posterior. This often involves minimizing the KL divergence between the two distributions. Naturally there is a nice parallel and intuitive analogy with the EM algorithm. A good discussion of this approach is given by Hoffman [10]. LDA has also been expanded in an attempt to automatically learn topic heirarchies to model increasingly complex phenomena. Several approaches have been proposed that draw on the theory of the distribution of random partitions known as Chinese restuarant processes. For an interesting discussion of the above the reader is referred to the following work by Paisley [11] and Blei [1].

References

[1] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, page 2003. MIT Press, 2004.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[3] Bob Carpenter. Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling. Technical report, LingPipe, 2010.

[4] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, November 2008.

[5] James M. Dickey. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78(383):628–637, 1983.

[6] Edward I. George George Casella. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[7] John Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *IN BAYESIAN STATISTICS*, pages 169–193. University Press, 1992.

[8] Philip Heidelberger and Peter D. Welch. Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144, 1983.

[9] Gregor Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[10] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013.

[11] John William Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. Nested hierarchical dirichlet processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):256–270, 2015.

[12] J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000a.

[13] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[14] Yee W. Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1353–1360. MIT Press, 2007.

7 Appendix

7.1 Integrating out Multinomial Parameters

$$\int\int p(\theta|\alpha)p(z|\theta)p(w|z,\phi)p(\phi|\beta)d\theta d\phi \tag{22}$$

The joint distribution factorizes based on dependencies specified by the graphical model.

$$= \int p(\theta|\alpha)p(z|\theta)d\theta \int p(w|z,\phi)p(\phi|\beta)d\phi \tag{23}$$

Take the product over documents and topics and replace the conditional probabilities with the Multinomial and Dirichlet distributions.

$$= \prod_{m=1}^{M} \int \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta^{\alpha_k - 1} \prod_{n=1}^{Nm} \theta^{c_{k,m,*}} d\theta \prod_{k=1}^{K} \int \frac{\Gamma(\sum \beta_j)}{\prod \Gamma(\beta_j)} \prod_{j=1}^{J} \phi^{\beta_j - 1} \prod_{m=1}^{M} \prod_{n=1}^{Nm} \phi^{c_{k,*,j}} d\phi \tag{24}$$

Rewrite the above exploiting the congucacy between prior and posterior.

$$= \prod_{m=1}^{M} \int \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta^{\alpha_k + c_{k,m,*} - 1} d\theta \prod_{k=1}^{K} \int \frac{\Gamma(\sum \beta_j)}{\prod \Gamma(\beta_j)} \prod_{j=1}^{J} \phi^{\beta_j + c_{k,*,j} - 1} d\phi \tag{25}$$

Note that both integrals are unnormalized Dirichlets. Multiply by the normalization constant and its reciprocal to cancel both integrals as they evaluate to 1 by definition.

$$= \prod_{m=1}^{M} \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \frac{\prod \Gamma(\alpha_k + c_{k,m,*})}{\Gamma(\sum \alpha_k + c_{k,m,*})} \prod_{k=1}^{K} \frac{\Gamma(\sum \beta_j)}{\prod \Gamma(\beta_j)} \frac{\prod \Gamma(\beta_j + c_{k,*,j})}{\Gamma(\sum \beta_j + c_{k,*,j})} \tag{26}$$

Drop terms that do not depend on the distribution over counts $c_{k,m,j}$

$$\propto \prod_{m=1}^{M} \frac{\prod \Gamma(\alpha_k + c_{k,m,*})}{\Gamma(\sum \alpha_k + c_{k,m,*})} \prod_{k=1}^{K} \frac{\prod \Gamma(\beta_j + c_{k,*,j})}{\Gamma(\sum \beta_j + c_{k,*,j})} \tag{27}$$

Split the product over documents and topics including/excluding the (a, b)th component.

$$= \prod_{m \neq a} \frac{\prod \Gamma(\alpha_k + c_{k,m,*})}{\Gamma(\sum \alpha_k + c_{k,m,*})} \frac{\prod \Gamma(\alpha_k + c_{k,a,*})}{\Gamma(\sum \alpha_k + c_{k,a,*})} \times \prod_{k=1}^{k} \frac{\prod_{j \neq y_{a,b}} \Gamma(\beta_j + c_{k,*,j}) \Gamma(\beta_{y_{ab}} + c_{k,*,y_{a,b}})}{\Gamma(\sum \beta_j + c_{k,*,j})} \tag{28}$$

Drop terms that do not depend on current position (a,b)

$$\propto \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + c_{k,a,*})}{\Gamma(\sum_{k=1}^{K} \alpha_k + c_{k,a,*})} \prod_{k=1}^{K} \frac{\Gamma(\beta_{y_{ab}} + c_{k,*,y_{ab}})}{\Gamma(\sum \beta_j + c_{k,*,j})} \tag{29}$$

Split the product over topics including/excluding the $c^{a,b}$th component

$$\propto \frac{\prod_{k\neq z_{ab}} \Gamma(\alpha_k + c_{k,a,*}^{-(a,b)}) \times \Gamma(\alpha_{z_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)})}{\Gamma(1 + \sum \alpha_k + c_{k,a,*}^{-(a,b)})} \times \prod_{k\neq z_{a,b}}^{K} \frac{\Gamma(\beta_{y_{a,b}} + c_{k,*,y_{a,b}}^{-(a,b)})}{\Gamma(\sum \beta_j + c_{k,*,j})} \frac{\Gamma(\beta_{y_{a,b}} + c_{z_{a,b},*,y_{a,b}}^{-(a,b)} + 1)}{\Gamma(1 + \sum \beta_j + c_{z_{a,b},*,j}^{-(a,b)})} \tag{30}$$

Use the fact that that $\Gamma(x+1) = x\Gamma(x)$

$$= \frac{\prod_{k\neq z_{a,b}} \Gamma(\alpha_k + c_{k,a,*}^{-(a,b)}) \Gamma(\alpha_{z_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)})(\alpha_{z_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)})}{\Gamma(1 + \sum \alpha_k + c_{k,a,*}^{-(a,b)})} \times \prod_{k\neq z_{a,b}} \frac{\Gamma(\beta_{y_{a,b}} + c_{k,*,y_{a,b}}^{-(a,b)})}{\Gamma(\sum \beta_j + c_{k,*,j})} \frac{\Gamma(\beta_{y_{a,b}} + c_{z_{a,b},*,y_{a,b}}^{-(a,b)})}{\Gamma(\sum \beta_j + c_{z_{a,b},*,j}^{-(a,b)})} \frac{(\beta_{y_{a,b}} + c_{z_{a,b},*,y_{a,b}}^{-(a,b)})}{\sum(\beta_j + c_{z_{a,b},*,j}^{-(a,b)})} \tag{31}$$

Combine the Γ terms back into the product:

$$= \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + c_{k,a,*}^{-(a,b)})(\alpha_{z_{a,b}} + c_{z_{a,b},*,y_{a,b}}^{-(a,b)})}{\Gamma(1 + \sum \alpha_k + c_{k,a,*}^{-(a,b)})} \times \prod_{k=1}^{K} \frac{\Gamma(\beta_{y_{a,b}} + c_{k,*,y_{a,b}}^{-(a,b)})}{\Gamma(\sum \beta_j + c_{z_{a,b},*,j}^{-(a,b)})} \frac{(\beta_{y_{a,b}} + c_{z_{a,b},*,y_{a,b}}^{-(a,b)})}{\sum(\beta_j + c_{z_{a,b},*,j}^{-(a,b)})} \tag{32}$$

Drop the topic denominator and general product terms which are constant:

$$\propto \frac{(\alpha_{z_{a,b}} + c +_{z_{a,b},a,*}^{-(a,b)}) \times (\beta_{y_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)})}{\sum(\beta_j + c_{z_{a,b},*,j}^{-(a,b)})} = \frac{(\alpha_{z_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)}) \times (\beta_{y_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)})}{\sum(\beta_j) + c_{z_{a,b},*,*}^{-(a,b)}} \tag{33}$$

The conditional distribution used for Gibbs sampling is given by normalizing the result above:

$$p(z_{a,b}|z_{-(a,b)}, y, \alpha, \beta) = \frac{(\alpha_{z_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)}) \times (\beta_{y_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)}) / (J \times \beta_j + c_{z_{a,b},*,*}^{-(a,b)})}{\sum_{k=1}^{K} (\alpha_{z_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)}) \times (\beta_{y_{a,b}} + c_{z_{a,b},a,*}^{-(a,b)}) / (J \times \beta_j + c_{z_{a,b},*,*}^{-(a,b)})} \tag{34}$$

7.2 The Gamma, Beta and Dirichlet Redux

Euler would introduce the useful function below.

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt = 2 \int_0^\infty u^{2x-1} e^{-u^2} du$$

Substituting $t = u^2$ and evaluating $x = 1/2$ gives the Euler-Poisson integral.

$$\Gamma(1/2) = 2 \int_0^\infty e^{-u^2} du = \sqrt{\pi}$$

Using integration by parts it is easy to see $\Gamma(x+1) = x\Gamma(x)$ and thus $\Gamma(x+1) = x!$ when x is an integer.

$$\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx$$

$$\Gamma(n+1) = \int_0^\infty e^{-x}x^{n}dx$$

Let $u = x^n, du = nx^{n-1}$, $dv = e^{-x}$ and $v = -e^{-x}$

$$= -x^n e^{-x}\Big|_0^\infty - \int_0^\infty -e^{-x}nx^{n-1}dx$$

$$= n\int_0^\infty e^{-x}x^{n-1}dx = n\Gamma(n)$$

Euler found that by using the Gamma function we can define the Beta function:

$$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = B(x,y) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

The beta is constructed using Gamma functions of two variables. The Gamma function can be used to derive the Beta. Perform the substitution $t = sin^2(\theta)$:

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

$$= \int_0^{\pi/2} sin(t)^{2x-1}cos(t)^{2y-1}dt = B(y,x)$$

$$\Gamma(x)\Gamma(y) = 4\int_0^\infty u^{2x-1}e^{-u^2}du \int_0^\infty v^{2y-1}e^{-v^2}dv$$

$$= 4\int_0^\infty\int_0^\infty e^{-(u^2+v^2)}u^{2x-1}v^{2y-1}dudv$$

Let $u = r\ cos(\theta)$ and $v = r\ sin(\theta)$:

$$
\begin{aligned}
&= 4 \int_0^\infty \int_0^{\pi/2} e^{-r^2} r^{2(x+y)-1} cos^{2x-1}(\theta) sin^{2y-1}(\theta) dr d\theta \\
&= 2 \int_0^\infty r^{2(x+y)-1} e^{-r^2} dr \times 2 \int_0^{\pi/2} cos^{2x-1}(\theta) sin^{2y-1}(\theta) d\theta \\
\Gamma(x)\Gamma(y) &= \Gamma(x+y) B(y,x) \\
\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} &= B(y,x)
\end{aligned}
$$

□

Recall that the Dirichlet is constructed from Gamma functions of d variables:

$$p(x|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{d} x_i^{\alpha_i - 1}$$

The Dirichlet distribution is the conjugate prior to the Multinomial. Note the similar functional form.

$$
\begin{aligned}
&posterior \propto likelihood \times prior \\
&= \frac{\Gamma(n+1)}{\prod \Gamma(x_i+1)} \prod_{i=1}^{K} p_i^{x_i} \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod x_i^{\alpha_i - 1} \\
&\propto \prod_{y_i} \prod_{j=1}^{K} p_j^{y_i^{(j)}} \prod_{j=1}^{K} p_j^{\alpha_j - 1} \\
&= \prod_{j=1}^{K} p_j^{\alpha_j - 1 + \sum y_i^{(j)}} \\
&= Dir(a_j + \sum y_i)
\end{aligned}
$$

The posterior is again a Dirichlet random variable whose new parameters are increased by the sum of the multinomial parameters.

Algorithm 1 Topic Model Procedure

```
                                                          ▷ Define Variables
integer :: m,n, top_size, Z, gg, kk, i, j, ll, nn
integer, dimension(n,m) :: matrix
integer, dimension(ntopics,n) :: NZW
integer, dimension(ntopics,m) :: NZM
integer, dimension(ntopics) :: NZ
integer, dimension(m) :: NM
integer, intent(in) :: ntopics, max_iter
integer, dimension(top_size) :: topics
integer, dimension(ntopics) :: genZ
real, dimension(ntopics) :: p_z
real,intent(in) :: alpha, beta
procedure COLLAPSED GIBBS SAMPLER

    for i in 1:iterations do
        kk = 1

        for j in 1:len(Documents) do

            for ll in 1:len(words) do
                if matrix(ll, j) == 0 then Next

                for nn in 1:matrix(ll,j) do
                    Z = topics(kk)
                                    ▷ Decrement DocumentTopic and WordTopic matrices
                    NZM(Z,j) = NZM(Z,j) - 1
                    NM(j) = NM(j) - 1
                    NZW(Z,ll) = NZW(Z,ll) - 1
                    NZ(Z) = NZ(Z) - 1

                    vocab = count which(matrix(:,j) != 0)

                    p_z = (NZM(:,j) + alpha) × (NZW(ll,:) + beta) / (NZ + vsize × beta)

                    p_z = p_z / sum(p_z)
                                    ▷ Generate Realization from Multinomial Distribution
                    call GENMUL(1,p_z,ntopics, genZ)
                    Z = which(genZ == 1)
                    topics(kk) = Z
                    kk = kk + 1
                                    ▷ Increment DocumentTopic and WordTopic matrices
                    NZM(Z,j) = NZM(Z,j) + 1
                    NM(j) = NM(j) + 1
                    NZW(Z,ll) = NZW(Z,ll) + 1
                    NZ(Z) = NZ(Z) + 1




    Phi = NZM / sum(NZM,by='cols')
    Theta = NZW / sum(NZW,by='rows')
```

Figure 1: Likelihood Over Sampled Iterations

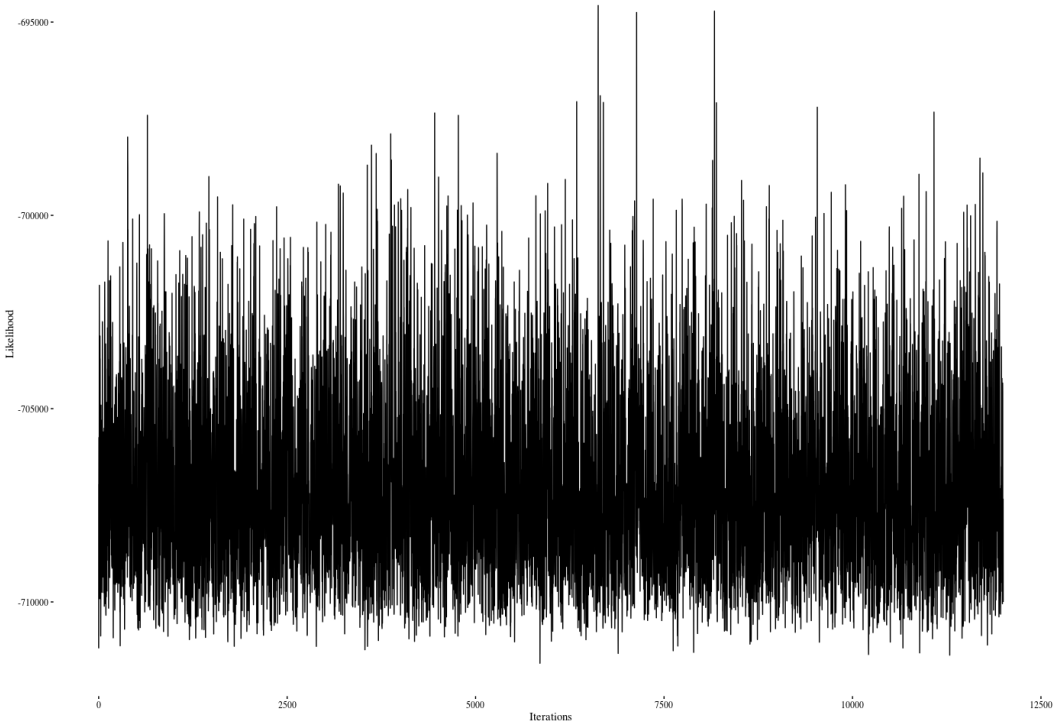

Table 2: Articles in Topic One

	Topic One
1	A Test for Significance in a Unique Sample
2	Volume Information
3	Letter to the Editor
4	Selective Service's Medical Statistics Program
5	Analysis of Coarsely Grouped Data from the Lognormal Distribution
6	Sub-Balanced Data and the Mixed Analysis of Variance
7	On Variance Estimation With Imputed Survey Data: Rejoinder
8	Exact Distribution of the Sum of Independent Identically Distributed Discrete Random Variables
9	Back Matter
10	Note on Estimating Significance of a Number of Samples

Table 3: Articles in Topic Two

	Topic Two
1	Bayesian Approach to Life Testing and Reliability Estimation
2	Inadmissibility of the Usual Scale Estimate for a Shifted Exponential Distribution
3	[List of Book Reviews]
4	Progress of Work in the Census Bureau
5	Nomographs for ... Differences Between the Frequencies of Events in Two Contrasted Series or Groups
6	Estimating Population Size with Exponential Failure
7	Missing Observations in Multivariate Statistics. III: Large Sample Analysis of Simple Linear Regression
8	Self-Calibrating Priors Do Not Exist: Comment
9	Wage Rates and Per Capita Productivity
10	A Method of Appraising Short-Term Forecasts

Table 4: Articles in Topic Three

	Topic Three
1	Committee on Nominations
2	Back Matter
3	Miscellaneous Notes
4	On Non-Negative Quadratic Unbiased Estimation of Variance Components
5	Corrections: Comment on "Semi-Parametric Nonlinear Mixed-Effects Models and Their Application"
6	Notes About Authors
7	[List of Book Reviews]
8	Self-Calibrating Priors Do Not Exist: Comment
9	The Revised Index of the Volume of Trade
10	Calculation of Chi-Square for Complex Contingency Tables

Table 5: Articles in Topic Four

	Topic Four
1	Exact Distribution of the Sum of Independent Identically Distributed Discrete Random Variables
2	Notes About Authors
3	A Stochastic Analysis of the Size Distribution of Firms
4	Bayesian Approach to Life Testing and Reliability Estimation
5	Index of Book Reviewers
6	Note on Estimating Significance of a Number of Samples
7	Diversity as a Concept and its Measurement: Rejoinder
8	[List of Book Reviews]
9	Committee on Nominations
10	A Two-Stage Minimax Procedure for Selecting the Normal Population with the Smallest Variance

Table 6: Articles in Topic Five

	Topic 5
1	Exact Distribution of the Sum of Independent Identically Distributed Discrete Random Variables
2	[List of Book Reviews]
3	A Stochastic Analysis of the Size Distribution of Firms
4	Committee on Nominations
5	Notes About Authors
6	Letter to the Editor
7	A Test for Extreme Value Domain of Attraction
8	Estimates of Bounded Relative Error for the Ratio of Variances of Normal Distributions
9	Progress of Work in the Census Bureau
10	Continuous Sequential Testing of a Poisson Process to Minimize the Bayes Risk