Technical Report CISE, University of Florida (2011) 1-12 Submitted 09/12; ID #519

Dirichlet Allocation Using Product Partition Models

Claudio Fuentes CFUENTES@STAT.UFL.EDU
Vikneshwaran Gopal VIKNESH@STAT.UFL.EDU
George Casella CASELLA@STAT.UFL.EDU
Department of Statistics
University of Florida
Gainesville, FL 32611

Clint P. George CGEORGE@CISE.UFL.EDU
Taylor C. Glenn TCG@CISE.UFL.EDU
Joseph N. Wilson JNW@CISE.UFL.EDU
Paul D. Gader PGADER@CISE.UFL.EDU
Department of Computer and Information Science & Engg
University of Florida
Gainesville, FL 32611

Abstract

A popular algorithm for allocation of documents to topics is the Latent Dirichlet Allocation model (LDA), which considers the topics as latent variables, and does not result in a clear assignment of documents to topics. Here we consider a Product Partition Model (PPM) for Dirichlet allocation, which gives more flexibility to the structure of the model, including the topic assignment as a parameter. Moreover, the model does not need to have the number of topics to be pre-specified, and can directly compare models with a variable number of topics, without resorting to ad hoc post processing. We also discuss a Gibbs sampler that facilitates the implementation of the model and the calculation of the required posterior probabilities, along with two illustrative examples. Lastly, we show that not every "collapsed Gibbs sampler" mixes better than the full Gibbs sampler, and previous choices for Dirichlet allocation have not been optimal.

Keywords: Product partition models, topic modeling, Gibbs sampling

1. Introduction

The Latent Dirichlet Allocation model (LDA) introduced by [1], has received a great deal of attention in the recent years as a suitable alternative to work in problems related to topic learning and document classification. Moreover, it is applicable to a wide variety of allocations problems, for example, classifying features of a geographic scenes. Many LDA models have been discussed in detail in the literature, and several authors have proposed different modifications in order to improve and facilitate the implementation of the model. For instance, [2] propose to augment the LDA parameterization by introducing class-label-dependent auxiliary parameters which obtain better results in document classification. [3] combines the usual Gibbs sampler with a variational Bayes technique, based on bounding the log marginal likelihood and optimizing the bound using Gaussian approximation facilitating the implementation of the model. In a similar direction, [4] proposed a collapsed Gibbs

sampler to carry out the calculations intending to improve the algorithm in terms of speed and computational efficiency.

However, all these variations of the LDA models assume that the number of topics is known, which is often unrealistic. In this direction, [5] consider implementing the LDA model for different number of topics and then determine the optimal number of topics based on a model selection criteria. But the problem of introducing the number of topics as a parameter in the model is not easy to solve within the hierarchical structure of the LDA model. To address this problem, we instead propose using a product partition model. These models, first introduced by [6] and [7], and later used by [8] in the context of cluster analysis, provides a more flexible structure allowing the introduction of the number of topics as a parameter and facilitating the calculation of the posterior probabilities of interest.

In Section 2 we first describe the hierarchy of the generative model and present a two-stage Gibbs sampler to estimate the model. In Section 2.3 we compare the performance of the two-stage Gibbs sampler versus the collapsed Gibbs sampler, typically used in the context of LDA models. In Section 3 we show how to learn the number of topics and present some simulation results. Finally, in Section 4, we summarize our main results.

2. Product Partition Models and Gibbs Samplers

The latent Dirichlet allocation model describes one approach to describing collections of discrete data such as a collection of text documents. It is fully developed in [1]. Here we describe a slight modification, that allows us to run a Gibbs sampler in order to estimate the posterior distributions.

2.1 Notation

Using terminology from text collections, suppose that we have D documents in a corpus, and that after removing the stop words, each document has n_d words. Also let V denote the *vocabulary* of the corpus - the total number of unique words. Then each document can be represented by a sequence of n_d binary vectors, each of length V, representing the n_d words in that document. We use the vectors $\boldsymbol{w}_{1,d}, \boldsymbol{w}_{2,d}, \ldots, \boldsymbol{w}_{n_d,d}$ to represent the words. All entries of each vector are zero except for the coordinate indicator $w_{it,d} = 1$, which will tell us exactly which word of the vocabulary is being picked up, with t running from 1 to V. The words are observed, and for each word in document d, let $\boldsymbol{z}_{i,d}$ denote a latent variable, also a binary vector, that identifies the topic that $\boldsymbol{w}_{i,d}$ belongs to. A *topic* can be defined simply as a distribution over the vocabulary of words. Assume we know that there are k topics in the corpus. One goal of Dirichlet Allocation is to identify words that define a topic. The topics themselves are unlabeled, but from the model parameters, they can be identified by the words in the vocabulary that are most likely to appear under that topic. Table 2 contains a summary of the index labels used in this section.

2.2 Generating Hierarchy and the Full Gibbs Sampler

Now let β represent a random $k \times V$ matrix, where each row represents the probability vector over the vocabulary corresponding to one topic. Formally, the data generating hierarchy,

Table 1: Index notation for LDA model.

D documents	$d = 1, \ldots, D$
n_d words in each document	$i = 1, \ldots, n_d$
k topics	$j = 1, \ldots, k$
V unique words in corpus	$t = 1, \ldots, V$

as given in [1], is as follows:

$$\begin{aligned} \beta_j &\sim \text{Dirichlet}(b_1, b_2, \ldots, b_V) \quad \text{for} \quad j = 1, 2, \ldots, k \\ \theta_d &\sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_k) \quad \text{for} \quad d = 1, 2, \ldots, D \\ \text{For each } d, \text{ for } i = 1, 2, \ldots, n_d: & \\ \boldsymbol{z}_{i,d} &\sim \text{Multinomial}(1, \theta_d) \\ \boldsymbol{w}_{i,d} &\sim \text{Multinomial}\left(1, \beta_{j'}\right), \quad \text{where } z_{ij',d} = 1. \end{aligned} \tag{1}$$

Informally, the above display simply says that each document is spread across the topics, with the spread being determined by θ_d. For each word within the document, we pick a random topic first ($\boldsymbol{z}_{i,d}$), and then pick the word according to the distribution (β) determined by that topic. The next step is to write down the conditional distributions for $\boldsymbol{z}_{i,d}$ and $\boldsymbol{w}_{i,d}$ for each document d and simplify them. Ignoring normalizing constants for now, we have

$$\begin{aligned} \Pr(\boldsymbol{z}_d|\theta_d) &= \prod_{i=1}^{n_d}\prod_{j=1}^{k}\theta_{j,d}^{z_{ij,d}} = \prod_{j=1}^{k}\theta_{j,d}^{n_{j,d}}, \quad \text{where } n_{j,d} = \sum_{i=1}^{n_d} z_{ij,d}, \\ \Pr(\boldsymbol{w}_d|\boldsymbol{z}_d, \beta) &= \prod_{i=1}^{n_d}\prod_{t=1}^{V}\prod_{j=1}^{k} = \beta_{jt}^{w_{it,d}z_{ij,d}} = \prod_{t=1}^{V}\prod_{j=1}^{k}\beta_{jt}^{m_{jt,d}}, \quad \text{where } m_{jt,d} = \sum_{i=1}^{n_d} w_{it,d}z_{ij,d}. \end{aligned}$$

In this context, $n_{j,d}$ represents the number of words in document d that are assigned to topic j, and $m_{jt,d}$ represents the number of times that word t in document d is assigned to topic j. The joint density of a corpus can be written as

$$\Pr(\boldsymbol{z}, \theta, \beta, \boldsymbol{w}|\alpha, b) = \left[\prod_{d=1}^{D}\Pr(\boldsymbol{w}_d|\boldsymbol{z}_d, \beta)\Pr(\boldsymbol{z}_d|\theta_d)\Pr(\theta_d|\alpha_1, \ldots, \alpha_k)\right]\Pr(\beta|b_1, \ldots, b_V).$$

Filling in the details, the posterior density of $(\boldsymbol{z}, \theta, \beta)$ is given by

$$\begin{aligned} \Pr(\boldsymbol{z}, \theta, \beta|\boldsymbol{w}, \alpha, b) &\propto \left[\prod_{d=1}^{D}\left(\prod_{t=1}^{V}\prod_{j=1}^{k}\beta_{jt}^{m_{jt,d}}\right)\left(\prod_{j=1}^{k}\theta_{j,d}^{n_{j,d}}\right)\left(\prod_{j=1}^{k}\theta_{j,d}^{\alpha_j-1}\right)\right]\left[\prod_{j=1}^{k}\prod_{t=1}^{V}\beta_{jt}^{b_t-1}\right] \\ &= \left(\prod_{t=1}^{V}\prod_{j=1}^{k}\beta_{jt}^{\sum_{d=1}^{D}m_{jt,d}+b_t-1}\right)\left(\prod_{d=1}^{D}\prod_{j=1}^{k}\theta_{j,d}^{n_{j,d}+\alpha_j-1}\right), \end{aligned} \tag{2}$$

where the second expression follows from straightforward algebra. We can now pick off the full conditionals of a two-stage Gibbs sampler.

1. Given $\boldsymbol{z}$, generate β and θ via

$$\theta_d \sim \text{Dirichlet}(n_{1,d}+\alpha_1,\ldots,n_{k,d}+\alpha_k) \quad \text{for} \quad d=1,2,\ldots,D,$$
$$\beta_j \sim \text{Dirichlet}\left(\sum_{d=1}^{D} m_{j1,d}+b_1,\ldots,\sum_{d=1}^{D} m_{jV,d}+b_V\right) \quad \text{for} \quad j=1,2,\ldots,k.$$

2. Given $\Delta=(\beta,\theta)$, and noting that $n_{j,d}=\sum_{t=1}^{V}\sum_{i=1}^{n_d} z_{ij,d}w_{it,d}=\sum_{t=1}^{V} m_{jt,d}$, the full conditional for $\boldsymbol{z}_d$ can be written

$$\Pr(\boldsymbol{z}_d|\theta_d,\beta,\boldsymbol{w}_d) \propto \left(\prod_{t=1}^{V}\prod_{j=1}^{k}\beta_{jt}^{m_{jt,d}}\right)\left(\prod_{j=1}^{k}\theta_{j,d}^{n_{j,d}}\right)=\prod_{i=1}^{n_d}\prod_{j=1}^{k}\prod_{t=1}^{V}(\beta_{jt}\theta_{j,d})^{z_{ij,d}w_{it,d}}.$$

Thus we can generate $\boldsymbol{z}_{i,d}$ according to

$$\boldsymbol{z}_{i,d} \sim \text{Multinomial}(1,(p_{i1,d},p_{i2,d},\ldots,p_{ik,d})) \quad \text{for} \quad i=1,2,\ldots,n_d,$$
$$\text{where} \quad p_{ij,d} \propto \prod_{t=1}^{V}(\beta_{jt}\theta_{j,d})^{w_{it,d}} \quad \text{for} \quad j=1,2,\ldots,k.$$

As this is a two-stage Gibbs sampler with one of the random variables over a finite state space, we can apply the Duality Principle (see Section 9.2.3 of [6] and [7] for further details) and conclude that the overall chain attains the fastest convergence rate, uniformly ergodicity.

A further point is that since we have a product partition model, we can in fact find the optimal allocation of words to topics using this model. To be explicit, at each iteration of the Gibbs sampler, we can compute the marginal posterior of the sampled $\boldsymbol{z}$ and rank them. To compute the marginal posterior, we simply have to integrate out β and θ from equation (2). Note that, in order to rank the $\boldsymbol{z}$, we do not need to compute the normalizing constant for the marginal posterior distribution of $\boldsymbol{z}$. We only need

$$\begin{aligned}\Pr(\boldsymbol{z}|\boldsymbol{w},\alpha,b) &\propto \int\int \left(\prod_{t=1}^{V}\prod_{j=1}^{k}\beta_{jt}^{\sum_{d=1}^{D} m_{jt,d}+b_t-1}\right)\left(\prod_{d=1}^{D}\prod_{j=1}^{k}\theta_{j,d}^{n_{j,d}+\alpha_j-1}\right)\mathrm{d}\beta\mathrm{d}\theta \\ &= \left[\prod_{j=1}^{k}\prod_{t=1}^{V}\Gamma\left(\sum_{d=1}^{D} m_{jt,d}+b_t\right)\right]\left[\prod_{d=1}^{D}\prod_{j=1}^{k}\Gamma\left(n_{j,d}+\alpha_j\right)\right].\end{aligned} \tag{3}$$

2.3 Comparison with the Collapsed Gibbs Sampler

For clarity, we investigate the relationship between the full Gibbs sampler and the collapsed Gibbs sampler in a special case of (1), using only one document, excluding the variables β and $\boldsymbol{w}$, and taking $n_d=2$. The model becomes

$$\begin{aligned}\theta &\sim \text{Dirichlet}(\alpha_1,\alpha_2,\ldots,\alpha_V) \\ \boldsymbol{z}_{i,} &\sim \text{Multinomial}(1,\theta), \quad \text{for } i=1,2,\end{aligned} \tag{4}$$

and note the following three samplers:

$$\begin{array}{ll} (a) & \boldsymbol{z}_1, \boldsymbol{z}_2|\theta, \quad \theta|\boldsymbol{z}_1, \boldsymbol{z}_2, \\ (b) & \boldsymbol{z}_1|\theta, \quad \theta|\boldsymbol{z}_1, \text{ or } \boldsymbol{z}_2|\theta, \quad \theta|\boldsymbol{z}_2, \\ (c) & \boldsymbol{z}_1|\boldsymbol{z}_2, \quad \boldsymbol{z}_2|z_1. \end{array} \tag{5}$$

Sampler (a) is the full Gibbs sampler of Section 2.2, and (b) and (c) are collapsed versions. The samplers in (b) are the subject of [9] and [10], who note that (b) is better than (a) in convergence rate (with the original theory in [11]). However, this theory does not cover Sampler (c), which is the collapsed Gibbs sampler of [5], [3], and [4]. Moreover, we now show that Sampler (c) is inferior to Sampler (a).

We compare the samplers through their lag-1 autocovariance. To be specific, run the collapsed Gibbs transition in the order $\boldsymbol{z} \to \boldsymbol{z}'$, where $\boldsymbol{z} = (z_1, z_2)$. Let $h(\boldsymbol{z})$ be any function with $\mathrm{E}h(\boldsymbol{z}) = 0$ and $\mathrm{Var}h(\boldsymbol{z}) < \infty$. The lag-1 covariance is given by

$$\mathrm{Cov}(h(\boldsymbol{z}), h(\boldsymbol{z}')) = \int_{\mathcal{Z}} \int_{\mathcal{Z}'} h(\boldsymbol{z}) h(\boldsymbol{z}') f(\boldsymbol{z}'|\boldsymbol{z}) f(\boldsymbol{z}) d\boldsymbol{z} d\boldsymbol{z}', \tag{6}$$

where $f(\boldsymbol{z}'|\boldsymbol{z})$ is the conditional distribution of $\boldsymbol{z}'$ given $\boldsymbol{z}$ (the Markov chain transition kernel) and $f(\cdot)$ is the stationary distribution of the chain. (We use f generically to denote a function, the roles should be clear from the arguments of the function.) Note that the better Markov chain will have a smaller lag-1 autocovariance, as the the chain will be closer to independence. In fact, for a two-stage Gibbs sampler the lag-1 autocorrelation is the convergence rate. For more about lag-1 autocovariance domination see [12].

We have the following theorem.

Theorem 1 *For the model* (4) *and samplers* (5)*, let* $\mathrm{Cov}_c(h(\boldsymbol{z}), h(\boldsymbol{z}'))$ *denote the lag-1 autocovariance under the collapsed Gibbs sampler* (c)*, and* $\mathrm{Cov}_f(h(\boldsymbol{z}), h(\boldsymbol{z}'))$ *that under the full Gibbs sampler* (a)*. The for any* $h(\boldsymbol{z})$ *satisfying* $\mathrm{E}h(\boldsymbol{z}) = 0$ *and* $\mathrm{Var}h(\boldsymbol{z}) < \infty$*,*

$$|\mathrm{Cov}_f(h(\boldsymbol{z}), h(\boldsymbol{z}'))| \leq |\mathrm{Cov}_c(h(\boldsymbol{z}), h(\boldsymbol{z}'))|.$$

Proof In (6), we can use the calculus of probability to write the density of the collapsed Gibbs sampler as

$$f(\boldsymbol{z}'|\boldsymbol{z}) f(\boldsymbol{z}) = \int_{\Theta} f(\boldsymbol{z}', \boldsymbol{z}|\theta) f(\theta) d\theta = \int_{\Theta} f(\boldsymbol{z}'|\theta) f(\boldsymbol{z}|\theta) f(\theta) d\theta,$$

where the last equality follows from the fact that in model (4), as in model (1), $\boldsymbol{z}$ and $\boldsymbol{z}'$ are conditionally independent given θ. Thus

$$\mathrm{Cov}_c(h(\boldsymbol{z}), h(\boldsymbol{z}')) = \int_{\Theta} \left[\int_{\mathcal{Z}} h(\boldsymbol{z}) f(\boldsymbol{z}|\theta) d\boldsymbol{z} \right]^2 f(\theta) d\theta = \mathrm{Var}[\mathrm{E}(h(\boldsymbol{z})|\theta)].$$

For the Gibbs full sampler we have

$$\mathrm{Cov}_f(h(\boldsymbol{z}), h(\boldsymbol{z}')) \quad = \quad \int_{\mathcal{Z}} \int_{\mathcal{Z}'} \int_{\Theta} \int_{\Theta'} h(\boldsymbol{z}) h(\boldsymbol{z}') f(\boldsymbol{z}', \theta'|\boldsymbol{z}, \theta) f(\boldsymbol{z}, \theta) d\boldsymbol{z} d\boldsymbol{z}' d\theta d\theta'.$$

Now write

$$f(\boldsymbol{z}', \theta' | \boldsymbol{z}, \theta) f(\boldsymbol{z}, \theta) = f(\boldsymbol{z}', \theta', \boldsymbol{z}, \theta) = f(\boldsymbol{z}' | \theta' \boldsymbol{z}, \theta) f(\theta', \boldsymbol{z}, \theta),$$

where we have factored the joint density into the conditional of $\boldsymbol{z}'$ times the marginal of $(\theta', \boldsymbol{z}, \theta)$. We can write this last marginal as $f(\theta', \boldsymbol{z}, \theta) = f(\boldsymbol{z}|\theta', \theta) f(\theta', \theta) = f(\boldsymbol{z}|\theta') f(\theta', \theta)$, as $\boldsymbol{z}$ is conditionally independent of θ given θ'. We thus have

$$f(\boldsymbol{z}', \theta' | \boldsymbol{z}, \theta) f(\boldsymbol{z}, \theta) = f(\boldsymbol{z}' | \theta' \boldsymbol{z}, \theta) f(\boldsymbol{z}|\theta') f(\theta', \theta) = f(\boldsymbol{z}'|\theta') f(\boldsymbol{z}|\theta) f(\theta', \theta),$$

where the last step again follows from conditional independence, that $\boldsymbol{z}'$ is conditionally independent of $(\boldsymbol{z}, \theta)$ given θ'. Putting this all together yields

$$\begin{aligned} \mathrm{Cov}_f(h(\boldsymbol{z}), h(\boldsymbol{z}')) &= \int_\Theta \int_{\Theta'} \left[\int_{\mathcal{Z}} h(\boldsymbol{z}') f(\boldsymbol{z}'|\theta') d\boldsymbol{z}' \right] \left[\int_{\mathcal{Z}} h(\boldsymbol{z}) f(\boldsymbol{z}|\theta) d\boldsymbol{z} \right] f(\theta', \theta) d\theta d\theta' \\ &= \mathrm{Cov}\left[\mathrm{E}(h(\boldsymbol{z}|\theta'), \mathrm{E}(h(\boldsymbol{z}|\theta) \right]. \end{aligned}$$

And, finally, recalling that the Cauchy-Schwarz inequality states that for any random variables X and Y, $|\mathrm{Cov}(X, Y)| \leq [\mathrm{Var}(X)\mathrm{Var}(Y)]^{1/2}$, we have

$$|\mathrm{Cov}_f(h(\boldsymbol{z}), h(\boldsymbol{z}'))| = |\mathrm{Cov}\left[\mathrm{E}(h(\boldsymbol{z}|\theta'), \mathrm{E}(h(\boldsymbol{z}|\theta) \right]| \leq \mathrm{Var}[\mathrm{E}(h(\boldsymbol{z})|\theta)] = |\mathrm{Cov}_c(h(\boldsymbol{z}), h(\boldsymbol{z}'))|,$$

showing that the collapsed sampler (c) has larger lag-1 autocovariances than the full Gibbs sampler (a). ■

Thus the full Gibbs sampler dominates the collapsed Gibbs sampler in the covariance ordering. This implies that the full Gibbs sampler is closer to independent sampling, and is expected to have smaller Monte Carlo variances. The following example gives some idea of the extent of the improvement, showing that the dominance can result in a reduction in Monte Carlo error of up to 50%.

Example 1 *A typical approach to LDA models is to run a "collapsed" Gibbs sampler* (c)*. To be precise,* θ *and* β *are analytically integrated out of the joint posterior in equation (2), and an* n*-stage Gibbs sampler is run to obtain samples from* $\Pr(\boldsymbol{z}|\boldsymbol{w}, \alpha, b)$*. This approach is detailed in [6]. When each sample of* $\boldsymbol{z}$ *is obtained, a Rao-Blackwellized estimate of* θ *and* β *is computed. Here we present an empirical comparison of the two Markov chains based on a (synthetic) data set.*

We ran 20 *chains, each with* 500 *iterations. For each chain we calculate the estimates of* θ*, and get their standard deviation between the* 20 *chains, which estimates the true Monte Carlo error. We take* $n = 8$ *to be the dimension of* $\boldsymbol{z}$*, and assume that there are* $k = 5$ *topics. The results are in Table 2, where we see that the full Gibbs sampler has a smaller Monte Carlo error than the collapsed Gibbs sampler.*

3. Learning the Number of Topics

To allow variable k, we start by putting a subscript on k, and now write k_d, as has been a typical approach. Also, both β and θ have dimension V, the maximum number of topics.

Table 2: Estimates of θ and standard errors for $n = 8$ and $k = 5$, based on 20 Markov chains, each with 500 iterations.

True $\theta = (.2, .2, .2, .2, .2)$

	θ					Std. Dev
Full Gibbs	0.1998	0.2017	0.2042	0.1965	0.1979	0.0525
Collapsed Gibbs	0.1940	0.2035	0.1949	0.2120	0.1956	0.0819

True $\theta = (0.3333, 0.3333, 0.1111, 0.1111, 0.1111)$

	θ					Std. Dev
Full Gibbs	0.3320	0.3351	0.1126	0.1077	0.1126	0.0269
Collapsed Gibbs	0.3334	0.3367	0.1114	0.1101	0.1084	0.0381

The generating hierarchy (1) is modified to

$$\begin{aligned} \beta_j &\sim \text{Dirichlet}(b_1, b_2, \ldots, b_V) \quad \text{for} \quad j = 1, 2, \ldots, V \\ \theta_d &\sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_V) \quad \text{for} \quad d = 1, 2, \ldots, D \\ \text{For each } d, \text{ for } i = 1, 2, \ldots, n_d: & \qquad (7) \\ \boldsymbol{z}_{i,d} &\sim \text{Multinomial}(1, \theta_d) \\ \boldsymbol{w}_{i,d} &\sim \text{Multinomial}\left(1, \beta_{j'}\right), \quad \text{where } z_{ij',d} = 1, \end{aligned}$$

where the dimension of the β and θ have been extended to V. The joint posterior calculation in (2) remains the same except that $\sum_{j=1}^{k}$ becomes $\sum_{j=1}^{V}$. We run the two-stage Gibbs sampler in the following way.

1. Given $\boldsymbol{z}$, generate β and θ via

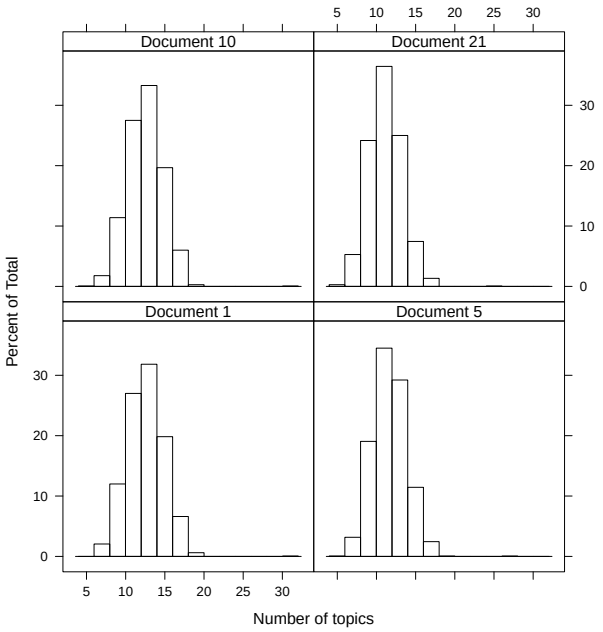
$$\theta_d \sim \text{Dirichlet}(n_{1,d} + \alpha_1, \ldots, n_{k_d,d} + \alpha_{k_d}, \underbrace{1, \ldots, 1}_{V - k_d \text{ times}}) \quad \text{for} \quad d = 1, 2, \ldots, D$$

$$\beta_j \sim \text{Dirichlet}\left(\sum_{d=1}^{D} m_{j1,d} + b_1, \ldots, \sum_{d=1}^{D} m_{jV,d} + b_V\right) \quad \text{for} \quad j = 1, 2, \ldots, V$$

2. Given $\Delta = (\beta, \theta)$, and noting that $n_{j,d} = \sum_{t=1}^{V} \sum_{i=1}^{n_d} z_{ij,d} w_{it,d} = \sum_{t=1}^{V} m_{jt,d}$, the full conditional for z_d is

$$\begin{aligned} \Pr(\boldsymbol{z}_d | \theta_d, \beta, \boldsymbol{w}_d) &\propto \left(\prod_{t=1}^{V} \prod_{j=1}^{V} \beta_{jt}^{m_{jt,d}}\right) \left(\prod_{j=1}^{V} \theta_{j,d}^{n_{j,d}}\right) \\ &= \left(\prod_{i=1}^{n_d} \prod_{t=1}^{V} \prod_{j=1}^{V} \beta_{jt}^{z_{ij,d} w_{it,d}}\right) \left(\prod_{i=1}^{n_d} \prod_{t=1}^{V} \prod_{j=1}^{V} \theta_{j,d}^{z_{ij,d} w_{it,d}}\right) \\ &= \prod_{i=1}^{n_d} \prod_{j=1}^{V} \prod_{t=1}^{V} (\beta_{jt} \theta_{j,d})^{z_{ij,d} w_{it,d}} \end{aligned}$$

Figure 1: The sampler was run with a "spiky" $\alpha = (30, 30, 30, 30, 0.1, 0.1, \ldots)$ in order to favor a smaller number of topics. The histograms demonstrate that the sampler is able to pick out the true number of topics present.



Thus we can generate $\boldsymbol{z}_{i,d}$ according to

$$\boldsymbol{z}_{i,d} \sim \text{Multinomial}(1, (p_{i1,d}, p_{i2,d}, \ldots, p_{iV,d})) \quad \text{for} \quad i = 1, 2, \ldots, n_d$$

$$\text{where} \quad p_{ij,d} \propto \prod_{t=1}^{V} (\beta_{jt}\theta_{j,d})^{w_{it,d}} \quad \text{for} \quad j = 1, 2, \ldots, V$$

3. We then calculate, for each d,

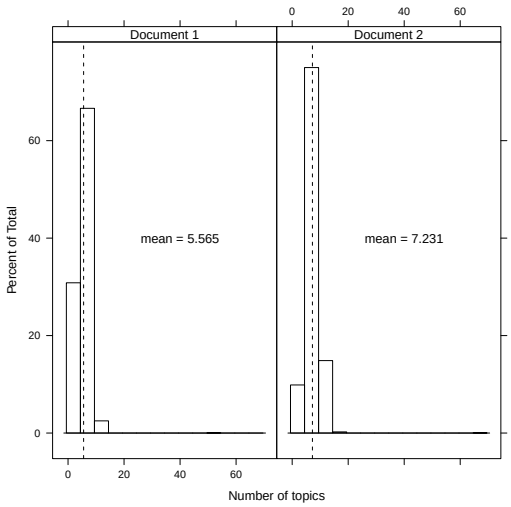
$$n_{j,d} = \sum_{i=1}^{n_d} z_{ij,d}, j = 1, \ldots V.$$

Some of these $n_{j,d}$ will be zero, and the k_d nonzero ones define the current number of topics.

3.1 Simulation Studies

Here, on two datasets, we demonstrate how the Gibbs sampler learns the number of topics. One is a synthetic dataset with the true number of topics known. The other is a small corpus of documents consisting of Associated Press articles.

Figure 2: Results from the variable k Gibbs sampler for the documents in Section 3.1.2. The Gibbs sampler of this section was run for 3000 iterations, and produced the following histograms for the number of topics in each document.



3.1.1 Synthetic Dataset

For this section, we simulated a dataset from the hierarchy specified in (1) with 10 topics, a vocabulary size of 50 words and 25 documents. Figure 1 contains histograms of the number of topics selected for 4 of the 25 documents.

3.1.2 Associated Press Dataset

The corpus used in this section only consists of the following 2 documents:

1. At least 15 people died and 25,000 residents of Surat town were evacuated after torrential rains flooded the west Indian town, United News of India reported Friday. UNI said the victims lived in slum localities which were the worst affected in the deluge Thursday night. It said 25,000 slum residents lost their homes in the floods and were moved to relief camps. Surat in Gujarat state is 560 miles southwest of New Delhi. At least 50 people have drowned or died in collapsed houses across India since the monsoon broke this month, newspapers reported. Worst affected are northeastern Assam state and eastern Bihar state.

2. A Navy anti-submarine helicopter crashed while preparing to land on a frigate in the North Arabian Sea and its three crewmen were presumed dead, officials announced Friday. The SH-2F helicopter was returning to the USS Barbey at the end of a dawn flight and crashed on approach, said Ken Mitchell, spokesman for North Island Naval Air Station. The Barbey is based at San Diego. The crash occurred about 7 p.m. PST Thursday, Mitchell said. Lost and presumed dead were Lt. Cmdr. Gerald C. Pelz, 37, of Coronado, Calif., Lt. j.g. Gerald T.

Table 3: The more frequent words in document 1 are assigned to topic 2 rather than topic 2, and vice versa for the frequent words in document 2. This suggests that in this purely instructional example, the topics are defined by the documents.

	Topic 1	Topic 2	Neither
"died"	6	9	
"india"	6	9	
"slum"	6	9	
"state"	7	8	
"surat"	5	9	1
"barbey"	11	4	
"helicopter"	8	7	
"uss"	12	3	
"north"	7	7	1
"gerald"	10	5	

Ramsdell, age unknown, of Ridgewood, N.J., and the anti-submarine warfare operator, Petty Officer 3rd Class William E. Martinie, 24, of Peoria, Ill. Helicopters from the aircraft carrier USS Nimitz, 70 miles away, and boats from the Barbey and USS California unsuccessfully searched for survivors. The craft was part of Helicopter Anti-Submarine Squadron Light 33.

The variable k gibbs sampler of this section was run for 3000 iterations, producing the histograms in Figure 2 for the number of topics in each document. Using the formula in equation (3), we can rank the sampled $\boldsymbol{z}$'s. In Table 3, we display the top 15 topic assignments in the course of running the Gibbs sampler, for 5 words from document 1 and 5 words from document 2.

4. Discussion

The product partition model has proven to be extremely effective in cluster algorithms, and here we see that it is equally effective in topic modeling. One big advantage of the model is the ability to include the number of topics in the search algorithm, eliminating the need for ad hoc post-processing with criteria such as BIC (Bayesian Information Criterion). We also saw that the collapsed Gibbs sampler is not necessarily an improvement over the full Gibbs sampler; although it may require fewer calculations, it does not mix as well and thus has a larger Monte Carlo error.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (Grants # 0712799, # MMS-1028314, and # DMS-1105127) and the Army Research Office

(Grant # W911NF-08-10410). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, Army Research Laboratory, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. The authors would like to thank R. Harmon, R. Weaver, P.Howard, and T. Donzelli for their support of this work.

References

[1] D.M. Blei, Ng A.Y., and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] S. Lacoste-Julien, F. Sha, and M.I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems*, 21, 2008.

[3] Y.W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent dirichlet allocation. *Advances in neural information processing systems*, 19, 2007.

[4] I. Porteous, D. Newman, A. Ihler, A. Asuncion, and M. Smyth, P.and Welling. Fast collapsed Gibbs sampling for latent dirichlet allocation. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[5] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proc. Nat. Acad. Sci. U. S.*, 101:5228 – 5235, 2004.

[6] J. A. Hartigan. Partition models. *Comm. Statist.Theory Meth.*, 19:27452756, 1990.

[7] D. Barry and J.A. Hartigan. Product partition models for change point problems. *Ann. Statist.*, 20:260–279, 1992.

[8] F.A. Quintana and P.L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Series B*, 65:557–574, 2003.

[9] Jun S. Liu. The collapsed Gibbs sampler in bayesian computations with applications to a gene regulationproblem. *Journal of the American Statistical Association*, 89:958–966, 1994.

[10] David A. van Dyk and Taeyoung Park. Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103:790–796, 2008.

[11] Jun S. Liu, Wing Hung Wong, and Augustine Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81:27–40, 1994.

[12] A. Mira. Ordering and improving the performance of Monte Carlo markov chains. *Statistical Science*, 16:340–350, 2001.