Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors

David Mimno, Hanna M. Wallach, Andrew McCallum
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{mimno,wallach,mccallum}@cs.umass.edu

Abstract

Previous work on probabilistic topic models has either focused on models with relatively simple conjugate priors that support Gibbs sampling or models with non-conjugate priors that typically require variational inference. Gibbs sampling is more accurate than variational inference and better supports the construction of composite models. We present a method for Gibbs sampling in non-conjugate logistic normal topic models, and demonstrate it on a new class of topic models with arbitrary graph-structured priors that reflect the complex relationships commonly found in document collections, while retaining simple, robust inference.

1 Introduction

Collections of documents are rarely unstructured: Some documents belong to hierarchies, others may be embedded in a network of citations of hyperlinks, almost all documents are situated in time and space. Not only do these relationships provide information about the documents, they are themselves an important object of study: What areas of nanotechnology are attracting increasing attention? What are people writing about in Canton, OH? Which journals and conferences are influencing each other, and to what degree? In this paper, we present a new class of probabilistic topic models with arbitrary graphs-structured priors that reflect these sorts of complex relationships.

Models for spatial and temporal data often rely on real-valued state space models such as dynamic linear models and, more generally, Gaussian Markov random fields (GMRFs). In contrast, topic models primarily use Dirichlet-multinomial distributions, because they are well-suited to categorical data and lead to simple Gibbs sampling algorithms. The expressiveness of such models is limited, however, due to impoverished covariance structure and non-conjugacy with Gaussian models.

We introduce a new class of probabilistic topic models that capture arbitrary network relationships between documents using Gaussian Markov random fields [13] and logistic normal distributions [1]. Inference in logistic normal topic models has previously been explored by Blei and Lafferty [3, 2], but has been restricted to variational methods due to the non-conjugacy between the multinomial and logistic normal distributions. We overcome this restriction by presenting an efficient algorithm for Gibbs sampling in logistic normal topic models. This algorithm is simple to implement, converges to the true posterior distribution rather than a variational approximation, and can be easily combined with existing tools from the substantial body of literature on real-valued state space models [13].

2 Related Work on Graph-Based Topic Models

There are several recently-introduced topic models that capture relationships between documents by modeling citations and hyperlinks [5, 12, 8, 11]. These models, however, are restricted to di-

rected links between individual documents. They are not capable of representing general, possibly undirected, compatibilities between groups of documents, such as spatial relationships.

Mei et al. [11] present a model that interpolates between a graph and a probabilistic latent semantic analysis. This model does not constitute a coherent generative process. In contrast, the model presented in this paper is fully generative and uses the graph distribution as a Bayesian prior. As a result, it is possible to use Bayesian inference methods rather than expectation maximization.

3 Logistic Normal Topic Models

The logistic normal distribution is a distribution on the simplex, obtained by transforming a random variable drawn from a multivariate Gaussian distribution. A point $\boldsymbol{\theta}$ in the $T-1$ simplex (i.e., a T-dimensional logistic normal random variable) can be generated as follows:

1. Generate a T-dimensional vector of parameters $\boldsymbol{\beta} \in \mathbb{R}^T$ from a T-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. For identifiability it is common practice to set $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ such that β_T is guaranteed to be zero.
2. Transform $\boldsymbol{\beta}$ into $\boldsymbol{\theta}$ using the logistic transform: $\theta_t = \frac{\exp(\beta_t)}{\sum_{t'=1}^{T} \exp(\beta_{t'})} = \frac{\exp(\beta_t)}{1+\sum_{t'=1}^{T-1} \exp(\beta_{t'})}$

A logistic normal distribution can be incorporated into a Dirichlet-based topic model, such as latent Dirichlet allocation, by replacing the Dirichlet priors over the document-specific topic distributions and topic-specific word distributions with logistic normal priors. In this paper, however, we consider only the former scenario, in which the document-specific topic distributions are drawn from a logistic normal, and assume that the topic-specific distributions over words are Dirichlet-distributed. Under such a model, the generative process for a single document (of length N_d) is as follows:

1. Draw a document-specific topic distribution $\boldsymbol{\theta}^{(d)}$ from a logistic normal, as above
2. For each position $n \in \{1, \ldots, N_d\}$
   (a) Draw a topic assignment: $z_n \sim \text{Mult}(\boldsymbol{\theta}^{(d)})$
   (b) Draw a word: $w_n \sim \text{Mult}(\boldsymbol{\phi}^{(z_n)})$

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ define the characteristics of the model: If $\boldsymbol{\Sigma}$ is a diagonal matrix, the model will exhibit the same covariance characteristics (i.e., uncorrelated topics) as latent Dirichlet allocation. Meanwhile, a non-diagonal covariance matrix will result in a correlated topic model [3]. Drawing $\boldsymbol{\mu}$ from a first-order dynamic linear model will give rise to a discrete dynamic topic model [2].

4 Sampling-Based Inference for Logistic Normal Topic Models

Given a corpus $\boldsymbol{w}$ of D documents, the posterior distribution over latent variables—i.e., the topic assignments $\boldsymbol{z}$ and logistic normal parameters $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$—is as follows:

$$P(\boldsymbol{z}, \{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D} \mid \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto P(\boldsymbol{w} \mid \boldsymbol{z}) \prod_{d=1}^{D} \left(\prod_{n=1}^{N_d} \frac{\exp(\beta_{z_n}^{(d)})}{1+\sum_{t=1}^{T-1} \exp(\beta_t^{(d)})} \right) \mathcal{N}(\boldsymbol{\beta}^{(d)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

Due to the non-conjugacy of the logistic normal and multinomial distributions, previous treatments of logistic normal topic models [2, 3] have relied on variational methods when inferring the latent topic assignments $\boldsymbol{z}$ and logistic normal parameters $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$. In this section, we explain how these latent variables can instead be inferred using sampling-based methods, despite non-conjugacy concerns. We take an approach based on blocked Gibbs sampling, in which the topic assignments $\boldsymbol{z}$ and logistic normal parameters $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$ are alternately sampled given all other variables.

4.1 Sampling Topic Assignments

Given a set of logistic normal parameters $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$, the latent topic assignments $\boldsymbol{z}$ can be inferred using an approach similar to that used in sampling-based treatments of latent Dirichlet allocation [6], where each z_n is sequentially re-sampled from its conditional posterior given all other variables.

In latent Dirichlet allocation, both the document-specific topic distributions $\{\boldsymbol{\theta}^{(d)}\}_{d=1}^{D}$ and the topic-specific word distributions $\{\boldsymbol{\phi}^{(t)}\}_{t=1}^{T}$ can be integrated out, due to the conjugacy between the Dirichlet and multinomial distributions. Assuming symmetric Dirichlet priors (with concentration parameters $\alpha, \gamma \in \mathbb{R}^{+}$), the conditional posterior probability for $z_n = t$ is therefore

$$P(z_n = t \mid \boldsymbol{z}_{\setminus n}, \boldsymbol{w}, \boldsymbol{d}, \alpha, \gamma) \propto P(w_n \mid z_n = t, \boldsymbol{z}_{\setminus n}, \boldsymbol{w}_{\setminus n}, \gamma) \left(N_{t|d_n} + \alpha \frac{1}{T} \right), \tag{2}$$

where $\boldsymbol{d}$ is the set of document indices for the entire corpus, T is the number of topics in the model, and the subscript "$\setminus n$" denotes a quantity that excludes data from the n^{th} position in the corpus.

In a logistic normal topic model, however, the document-specific topic distributions $\{\boldsymbol{\theta}^{(d)}\}_{d=1}^{D}$ cannot be integrated out, so the conditional posterior probability for $z_n = t$ is instead

$$P(z_n = t \mid \boldsymbol{z}_{\setminus n}, \boldsymbol{w}, \boldsymbol{d}, \{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}, \gamma) \propto P(w_n \mid z_n = t, \boldsymbol{z}_{\setminus n}, \boldsymbol{w}_{\setminus n}, \gamma) \, \exp(\beta_t^{(d)}). \tag{3}$$

Aside from this difference in posteriors, the algorithm for sampling $\boldsymbol{z}$ in a logistic normal topic model is identical to that used in sampling-based treatments of latent Dirichlet allocation [6].

4.2 Sampling Logistic Normal Parameters

Given a set of topic assignments $\boldsymbol{z}$, the posterior distribution over $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$ is

$$P(\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D} \mid \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \prod_{d=1}^{D} \left(\prod_{n=1}^{N_d} \frac{\exp(\beta_{z_n}^{(d)})}{1 + \sum_{t=1}^{T-1} \exp(\beta_t^{(d)})} \right) \mathcal{N}(\boldsymbol{\beta}^{(d)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4}$$

As a result, inferring $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$ is equivalent to learning the parameters of a logistic regression model with a Gaussian prior, in which the inputs are the document indices $\boldsymbol{d}$, the labels are the topic assignments $\boldsymbol{z}$, and there are exactly D binary feature functions of the form $f_d(d_n) = \delta(d_n - d)$. Each input d_n will therefore be associated with a vector of D binary feature values—one for each feature function. Clearly, in every such vector, only a single feature will have a non-zero value.

Logistic regression parameters are most commonly inferred using numerical optimization methods, but there are also several Markov chain Monte Carlo methods for Bayesian inference that use auxiliary variables. Holmes and Held [9, 13] use logistically-distributed auxiliary variables, which can be represented as a scale mixture of normals with variances drawn independently from a Kolmogorov-Smirnov distribution. Here, however, we follow the sampling method of Groenewald and Mokgatlhe [7], in which the auxiliary variables are instead drawn from a uniform distribution.

Groenewald and Mokgatlhe's method is best explained by starting with the simplified scenario in which there are two topics and a single document. Since β_2 must be constrained to zero for identifiability, this restriction means that the model has only a single unknown logistic normal parameter.

Under this model, the generative process for topic assignments $\boldsymbol{z}$ is as follows:

1. Generate $\beta_1 \in \mathbb{R}$ from a univariate Gaussian and set β_2 to zero
2. Transform β_1 into $\boldsymbol{\theta}$ using the logistic transform: $\theta_1 = \frac{\exp(\beta_1)}{1+\exp(\beta_1)}$ and $\theta_2 = \frac{1}{1+\exp(\beta_1)}$
3. For each position $n \in \{1, \ldots, N\}$
   (a) Draw a topic assignment: $z_n \sim \text{Mult}(\boldsymbol{\theta})$

In fact, however, θ_1 can be interpreted as the CDF of a logistic distribution,

$$\theta_1 = \frac{\exp(\beta_1)}{1 + \exp(\beta_1)} = \int_{-\infty}^{\beta_1} \frac{\exp(x)}{(1 + \exp(x))^2} \, \mathrm{d}x, \tag{5}$$

as shown in figure 1(a). Using this interpretation, each z_n can be sampled from $\text{Mult}(\boldsymbol{\theta})$ as follows:

Figure 1: The relationship between β_1, z and u in a simple model with a single document and two topics. (a) Given β_1, if $z_n = 1$, u_n (represented by a black dot) must fall below the value of the logistic CDF at β_1. Otherwise, u_n (represented by a white dot) must fall above the value of the logistic CDF at β_1. (b) Given u, β_1 can be anywhere in the interval defined by the maximum black dot and minimum white dot.

1. Draw a vertical line through the x-axis of the CDF at β_1, as in figure 1(a)
2. Sample an auxiliary variable $u_n \sim \mathrm{U}(0, 1)$
3. Plot u_n on the vertical line through β_1
4. If u_n lies below the curve, let $z_n = 1$; otherwise let $z_n = 2$

The same approach (i.e., the introduction of a single uniformly-distributed auxiliary variable u_n for each topic assignment z_n) can also be used to *infer* β_1 from z, given an initial value of β_1:

1. Draw a vertical line through the x-axis of the CDF at the current β_1
2. For each position $n \in \{1, \ldots, N\}$
   (a) Draw $u_n \sim \begin{cases} \mathrm{U}(0, \frac{\exp(\beta_1)}{1+\exp(\beta_1)}) & \text{if } z_n = 1 \\ \mathrm{U}(\frac{\exp(\beta_1)}{1+\exp(\beta_1)}, 1) & \text{if } z_n = 2 \end{cases}$
   (b) Plot u_n on the vertical line through β_1, as in figure 1(a)

Having generated an auxiliary variable u_n for each z_n, sampling β_1 given z simply corresponds to shifting the vertical line left or right, within the region bounded as follows:

$$\max_{n|z_n=1} \log \frac{u_n}{1-u_n} < \beta_1 < \min_{n|z_n=2} \log \frac{u_n}{1-u_n}. \tag{6}$$

When selecting a new location for the vertical line (i.e., a new value for β_1) possible locations must be weighted according to the prior $\mathcal{N}(\beta_1; \mu_1, \sigma_1^2)$. This is equivalent to sampling from a normal distribution multiplied by an indicator function that "zeroes out" the entire real line except for the segment defined by the lower and upper bounds. Once a new β_1 has been selected, the auxiliary variables can be discarded and the process repeated until sufficient samples have been generated.

The efficiency of this algorithm can be improved by noting that $\log \frac{u_n}{1-u_n}$ is monotonic in u_n. Finding $\max_{n|z_n=1} \log \frac{u_n}{1-u_n}$ is therefore equivalent to finding the maximum of $N_1 = \sum_{n=1}^{N} \delta(z_n - 1)$ uniform random variables, while finding $\min_{n|z_n=2} \log \frac{u_n}{1-u_n}$ is equivalent to finding the minimum of $N_2 = \sum_{n=1}^{N} \delta(z_n - 2)$ uniform random variables. Using standard results from order statistics, the largest of M uniform random variables is distributed according to Beta $(M, 1)$, while the smallest is distributed according to Beta $(1, M)$. It is therefore unnecessary to sample all N auxiliary variables: Given N_1 and N_2, $\max_{n|z_n=1} u_n$ and $\min_{n|z_n=2} u_n$ (and hence the lower and upper bounds on β_1) can both be computed directly by drawing a random variable from a beta distribution, scaled and shifted as appropriate. Finally, the efficiency can be further improved by noting that samples from a beta distribution with either parameter equal to one can be drawn using transformations of uniform random variables [4]: $u^{\frac{1}{M}} \sim$ Beta $(M, 1)$ and $1 - u^{\frac{1}{M}} \sim$ Beta $(1, M)$, where $u \sim \mathrm{U}(0, 1)$.

If the above scenario is extended to include T topics, then $T - 1$ logistic normal parameters must be sampled: $\{\beta_t\}_{t=1}^{T-1}$. For each topic t, the corresponding parameter β_t can be sampled using an approach similar to that described above, by noting that $\theta_t = \frac{\exp(\beta_t)}{1+\sum_{t'=1}^{T-1} \exp(\beta_{t'})}$ can be interpreted as the CDF of a logistic distribution. A new value for β_t can therefore be sampled as follows:

1. Draw a vertical line through the x-axis of the CDF at the current β_t

2. For each position $n \in \{1, \ldots, N\}$

   (a) Draw $u_n \sim \begin{cases} \mathrm{U}\,(0, \frac{\exp(\beta_t)}{1+\sum_{t'=1}^{T-1} \exp(\beta_{t'})}) & \text{if } z_n = t \\ \mathrm{U}\,(\frac{\exp(\beta_t)}{1+\sum_{t'=1}^{T-1} \exp(\beta_{t'})}, 1) & \text{if } z_n \neq t \end{cases}$

   (b) Plot u_n on the vertical line through β_t, as in figure 1(a)

Once an auxiliary variable u_n has been generated for each z_n, sampling a new value for β_t corresponds to shifting the vertical line left or right within a bounded region, under the constraint that potential locations must be weighted according to $\mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The bounds on β_t are:

$$\max_{n|z_n=t} \log \frac{C\,u_n}{1-u_n} < \beta_t < \min_{n|z_n \neq t} \log \frac{C\,u_n}{1-u_n}, \tag{7}$$

where

$$C = 1 + \sum_{t'=1}^{T-1} \exp(\beta_{t'})(1 - \delta(t' - t)). \tag{8}$$

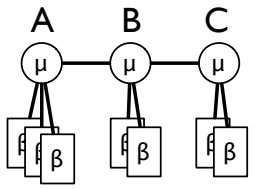
Since C is a constant, these bounds are monotonic in u_n. Consequently, finding $\max_{n|z_n=t} \log \frac{C\,u_n}{1-u_n}$ is equivalent to finding the maximum of $N_t = \sum_{n=1}^{N} \delta(z_n - t)$ uniform random variables, while finding $\min_{n|z_n \neq t} \log \frac{C\,u_n}{1-u_n}$ is equivalent to finding the minimum of $N_{\neg t} = \sum_{n=1}^{N} (1 - \delta(z_n - t))$ uniform random variables. Again, standard results from order statistics mean that it is not necessary to sample an auxiliary variable for each z_n: Given N_t and $N_{\neg t}$, $\max_{n|z_n=t} u_n$ and $\min_{n|z_n \neq t} u_n$ (and hence the bounds on β_t) can be computed directly by drawing random variables from beta distributions. An important special case is when $N_t = 0$. In this situation, there is no lower bound on β_t, so the prior $\mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is responsible for ensuring that β_t stays within a reasonable range.

Finally, Groenewald and Mokgatlhe's method can be extended to sample the logistic normal parameters for D documents $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^{D}$ by restricting the topic assignments used to sample each parameter $\beta_t^{(d)}$ to only those assignments z_n that belong to document d (i.e., those for which $d_n = d$).

5 Gaussian MRF Priors for Topic Models

Topic-based document similarities often reflect underlying relationships between documents that can be expressed using some kind of graph structure. For example, documents written in 2005 are more likely to have similar topic distributions to each other, and, to a lesser extent, to documents written in 2004 and 2006, than to documents written in other years. Similarly, it is more likely that documents that are in the same or nearby geographical areas will have similar topic distributions than documents that are geographically dispersed. These kinds of relationships and their effects on document contents can be modeled using a graph-based prior over document-specific topic distributions.

A Gaussian Markov random field (GMRF) [13] defines a multivariate Gaussian distribution $P(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ over a set of real-valued variables $\boldsymbol{x}$. Letting $\mathcal{G}$ be an undirected graphical model over $\boldsymbol{x}$, then $P(\boldsymbol{x})$ can be decomposed such that each x_n is normally distributed conditioned on its neighbors. Furthermore, there is a direct correspondence between the graph structure of $\mathcal{G}$ and the precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ of $P(\boldsymbol{x})$: Element Q_{mn} is zero if and only if variables x_m and x_n are conditionally independent—i.e., there is no edge in $\mathcal{G}$ directly connecting x_m and x_n. If $\boldsymbol{Q}$ is diagonal, all variables in $\boldsymbol{x}$ are independent. If $\boldsymbol{Q}$ is tridiagonal, then $\mathcal{G}$ is a linear chain. More general graphs, such as irregular lattices, correspond to more complicated precision matrices. However, Rue and Held [13] show that as long as $\boldsymbol{Q}$ is sparse, inference in GMRFs remains tractable.



(a)

4	-1		-1	-1	-1				
-1	4	-1				-1	-1		
	-1	3						-1	-1
-1			1						
-1				1					
-1					1				
	-1					1			
	-1						1		
		-1						1	
		-1							1

(b)

Figure 2: Figure 2(a) shows a linear chain backbone graph with three topic mean parameters $\boldsymbol{\mu}^{(A)}$, $\boldsymbol{\mu}^{(B)}$ and $\boldsymbol{\mu}^{(C)}$, which in turn determine three, two and two sets of logistic normal document parameters, respectively. Figure 2(b) shows the precision matrix $\boldsymbol{Q}$ for an intrinsic GMRF associated with this graph. Each of the first three variables ($\boldsymbol{\mu}^{(A)}$, $\boldsymbol{\mu}^{(B)}$ and $\boldsymbol{\mu}^{(C)}$) depends on its neighboring mean variables (e.g., $\boldsymbol{\mu}^{(A)}$ depends on $\boldsymbol{\mu}^{(B)}$ but not on $\boldsymbol{\mu}^{(C)}$) and the logistic normal variables for all documents belonging to that group (e.g., $\boldsymbol{\mu}^{(A)}$ depends on $\{\boldsymbol{\beta}^{(d)}\}_{d\in A}$. The logistic normal variables are conditionally independent of each other given the mean variables. Note that the diagonals represent the negative sum of the off-diagonals for each row.

Given an undirected graph $\mathcal{G}$ over $\boldsymbol{x}$ it is possible to define $P(\boldsymbol{x})$ such that the *difference* in values between each pair of adjacent variables $x_m \sim x_n$ is Gaussian with some precision κ:

$$P(\boldsymbol{x}) \propto \prod_{x_m \sim x_n} \exp\left(-\frac{\kappa}{2}(x_m - x_n)^2\right) \tag{9}$$

$$= \exp\left(-\frac{\kappa}{2}\sum_{x_m \sim x_n}(x_m - x_n)^2\right) \tag{10}$$

$$= \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}\right). \tag{11}$$

$P(\boldsymbol{x})$ is therefore a multivariate Gaussian with precision matrix $\boldsymbol{Q}$, whose elements are $Q_{nn} = \kappa \deg(x_n)$ and $Q_{mn} = -\kappa$ for $m \neq n$. This matrix is singular and does not define a proper distribution, as $P(x)$ is invariant to the addition of a constant to each x_n; equivalently, this model makes no claims about the *absolute* value of $\boldsymbol{x}$, only about relative differences between the variables. Rue and Held [13] refer to this form of GMRF as an intrinsic GMRF, or IGMRF, of first-order.

In the previous section, we introduced a blocked Gibbs sampling algorithm for inferring $\boldsymbol{z}$ and $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^D$ in a logistic normal topic model with a single prior mean $\boldsymbol{\mu}$. Using the IGMRF framework, we now extend this algorithm to account for *groups* of documents that share a prior mean.

As an example, figure 2(a) depicts three groups of documents: Group A contains three documents, while groups B and C each contain two. Rather than a single prior mean $\boldsymbol{\mu}$, there are *three* prior means: $\boldsymbol{\mu}^{(A)}$, $\boldsymbol{\mu}^{(B)}$ and $\boldsymbol{\mu}^{(C)}$. Furthermore, the conditional independences between the means $\{\boldsymbol{\mu}^{(g)}\}_{g\in\{A,B,C\}}$ and logistic normal parameters $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^D$ are expressed via the graph structure.

The example in figure 2(a) is best explained by starting with the simplified scenario in which there are only two topics. Since each $\beta_2^{(d)}$ must be constrained to zero for identifiability, this restriction means that each document has only a single unknown logistic normal parameter. Similarly, $\mu_1^{(g)} \in \mathbb{R}$ and $\mu_2^{(g)} = 0$ for each group $g \in \{A, B, C\}$. As a result, the model has ten parameters: three means $\{\mu_1^{(g)}\}_{g\in\{A,B,C\}}$—one for each group—and seven logistic normal parameters $\{\beta_1^{(d)}\}_{d=1}^7$—one for each document. Using a first order IGMRF, it is possible to define a joint distribution over $\{\mu_1^{(g)}\}_{g\in\{A,B,C\}}$ and $\{\beta_1^{(d)}\}_{d=1}^7$, parameterized by the precision matrix shown in figure 2(b). The three-by-three block in the top left, denoted by $\boldsymbol{Q}_{\mu\mu}$, contains the degrees of $\{\mu_1^{(g)}\}_{g\in\{A,B,C\}}$ along the diagonal and the graph structure between them (indicated by "-1"s) in the off-diagonal elements. The top right and bottom left blocks, denoted by $\boldsymbol{Q}_{\mu\beta}$ and $\boldsymbol{Q}_{\beta\mu}$, represent the connections between

Algorithm	HOL	Example "classification" topics
GibbsUniform	-876602	web map classification problem based
GibbsLDA	-846729	vector classification support learning regression
GibbsLogistic	-846746	classification vector selection support regression

Table 1: Held-out likelihood ("HOL") and example topics. GibbsLDA and GibbsLogistic produce similar results, while the baseline GibbsUniform algorithm exhibits much worse performance.

the groups and the documents, while the bottom right block, denoted by $\boldsymbol{Q}_{\beta\beta}$, represents the fact that the logistic normal parameters $\{\beta_1^{(d)}\}_{d=1}^7$ are conditionally independent given the means.

The logistic normal topic model discussed in section 4 is equivalent to a first order IGMRF topic model with a single group (and hence a single mean $\boldsymbol{\mu}$). Consequently, the sampling algorithm in section 4.2 can be used without modification to sample the logistic normal parameters $\{\boldsymbol{\beta}_1^{(d)}\}_{d=1}^D$ of a single-group IGMRF topic model. The only modification necessary for a multi-group IGMRF topic model is to ensure that the correct group-specific mean is used when sampling each $\beta_1^{(d)}$.

To complete the Gibbs sampling framework for a first order IGMRF topic model, the group-specific means, e.g., $\{\mu_1^{(g)}\}_{g\in\{A,B,C\}}$, must also be sampled, given all other variables. The posterior distribution over the means is given by the conditional distribution of a multivariate Gaussian, e.g.,

$$P(\{\mu_1^{(g)}\}_{g\in\{A,B,C\}} \mid \{\beta_1^{(d)}\}_{d=1}^7) = \mathcal{N}\left(\boldsymbol{Q}_{\mu\mu}^{-1}\,\boldsymbol{Q}_{\mu\beta}\,\{\beta_1^{(d)}\}_{d=1}^7, \boldsymbol{Q}_{\mu\mu}^{-1}\right). \tag{12}$$

Drawing a sample from this distribution requires inverting $\boldsymbol{Q}_{\mu\mu}$. Fortunately, this non-singular submatrix grows only with the number of groups, rather than the size of the entire document collection.

To extend the above scenario to include T topics, the graph structure must be replicated for each topic $t = 1, \ldots, T-1$. Inference of the logistic normal parameters $\{\boldsymbol{\beta}^{(d)}\}_{d=1}^D$ can then be performed as in section 4.2, while inference of the means, e.g., $\{\mu_1^{(g)}\}_{g\in\{A,B,C\}}$, can proceed as above. Since $\boldsymbol{Q}_{\mu\mu}$ depends only on the graph structure (which is identical for all topics) and κ, matrix inversion can be performed once and $\boldsymbol{Q}_{\mu\mu}^{-1}$ shared between topics, provided κ is the same for each topic.

Finally, it is possible to give the precision parameter κ a conjugate gamma prior and sample its value [13]. It is also possible to to define different precision parameters for different edges in the graph.

6 Experimental Results

In this section, we present two sets of results. The first set demonstrates that the sampling method in section 4 is as accurate as collapsed Gibbs sampling for Dirichlet-based topic models. Meanwhile, the second set highlights the effects of the graph-based prior on unseen document prediction.

To evaluate the new sampling algorithm for logistic normal topic models, we compared three models: The first is a simple logistic normal model, referred to as "GibbsLogistic", with a single mean $\boldsymbol{\mu}$, shared by all documents, and $\kappa = 1$. Topic assignments and logistic normal parameters for this model were inferred using the algorithm in section 4. The second model is a standard Dirichlet-based topic model, referred to as "GibbsLDA", with optimized document-topic hyperparameters. The algorithms for sampling latent topic assignments $\boldsymbol{z}$ in GibbsLogistic and GibbsLDA differ only in the conditional posterior distribution for each assignment z_n, as indicated by equations 3 and 2. Finally, "GibbsUniform" is a baseline model with uniform document-specific topic distributions.

Table 1 shows the held-out likelihood and an example topic for each model, obtained using a corpus of approximately 50,000 abstracts from computer science papers[1]. Held-out likelihood was calculated using 10-fold cross-validation by sampling unconditionally from the prior over document-specific topic distributions [10]. Each model used 50 topics. Topic assignments and parameters (if any) were inferred using 1000 sampling iterations. The results for GibbsLDA and GibbsLogistic are almost indistinguishable, while the results for GibbsUniform are, predictably, nearly random. The sampling algorithm in section 4 is therefore an effective inference method for topic models.

[1] http://rexa.info

	CVPR	NIPS	2001	1990
Single	-694911	-1121797	-724802	-67779
Years	-691880	-1123727	-724381	-68019
Venues	-673732	-1112114	-700662	-64817
Years and Venues	-690893	-1118951	-721598	-67745

Table 2: Held-out likelihoods for different document groups under various graph topologies.

To determine the effects of using a graph-based prior over document-specific topic distributions, we compared four graph structures over group means for the corpus described above: a single mean (equivalent to LDA; no edges), a linear chain over publication years (21 edges), an irregular graph based on venue-to-venue citation patterns (85 edges), and a "spatio-temporal" graph with a mean for each year–venue pair, such that the mean for "NIPS 2005" is connected to the means for "NIPS 2004", "NIPS 2006" and "ICML 2005" (2458 edges). For each graph structure, two precision parameters were inferred: an "evolution" precision for edges between group means, and an "observation" precision for edges connecting the means to document-specific logistic normal parameters.

For each model, the held-out likelihood was computed using a subset of the corpus, e.g., all papers from a particular venue or year. The remaining papers were used to infer the group-specific means. Note that if the held-out data and graph structure are aligned such that all documents from a particular group belong to the held-out set, then the mean for that group is inferred using only information from adjacent means. The held-out likelihood therefore provides a measure of the extent to which the inferred mean accurately represents the document-specific topic distributions for that group.

Table 2 shows the held-out likelihood results for papers from CVPR and NIPS, as well as papers published in 2001 and 1990. The venue graph consistently gives rise to the best held-out likelihood, even when the held-out data consist of documents from a particular year. These results indicate that venue is generally a better predictor of topic than year. The year–venue graph exhibits relatively poor performance compared to the venue graph. It is possible that this difference is due to the fact that there is only a single "evolution" precision parameter in each model. Inferring different "evolution" precisions for year–year edges and venue–venue edges may result in improved performance.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by Lockheed Martin through prime contract #FA8650-06-C-7605 from the Air Force Office of Scientific Research, in part by DoD contract #HM1582-06-1-2013, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010, and AFRL #FA8750-07-D-0185. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, 1986.

[2] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, 2006.

[3] D. M. Blei and J. D. Lafferty. A correlated topic model of *Science*. *AAS*, 1(1):17–35, 2007.

[4] L. Devroye. Random variate generation in one line of code. In *Winter Simulation Conference*, 1996.

[5] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML*, 2007.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[7] P. C. Groenewald and L. Mokgatlhe. Bayesian computation for logistic regression. *Computational Statistics and Data Analysis*, 48:857–868, 2005.

[8] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Latent topic models for hypertext. In *UAI*, 2008.

[9] C. C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

[10] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, 2006.

[11] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, 2008.

[12] R. Nallapati and W. Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. In *ICWSM*, 2008.

[13] H. Rue and L. Held. *Gaussian Markov Random Fields*. Chapman & Hall/CRC, 2005.