
Markov Topic Models

Chong Wang *

Computer Science Dept.
Princeton University
Princeton, NJ 08540

Bo Thieson, Christopher Meek

Microsoft Research
One Microsoft Way
Redmond, WA 98052

David Blei

Computer Science Dept.
Princeton University
Princeton, NJ 08540

Abstract

We develop *Markov topic models* (MTMs), a novel family of generative probabilistic models that can learn topics simultaneously from multiple corpora, such as papers from different conferences. We apply Gaussian (Markov) random fields to model the correlations of different corpora. MTMs capture both the internal topic structure within each corpus and the relationships between topics across the corpora. We derive an efficient estimation procedure with variational expectation-maximization. We study the performance of our models on a corpus of abstracts from six different computer science conferences. Our analysis reveals qualitative discoveries that are not possible with traditional topic models, and improved quantitative performance over the state of the art.

1 Introduction

Algorithmic tools for analyzing, indexing and managing large collections of online documents are becoming increasingly important. In recent years, algorithms based on *topic models* have become a widely used approach for exploratory and predictive analysis of text. Topic models, such as latent Dirichlet allocation (LDA) (Blei et al. 2003) and the more general discrete component analysis (Buntine 2004), are hierarchical Bayesian models of discrete data that use “topics”, i.e., patterns of word use, to explain an observed document collection. Probabilistic topic models

have been extended and applied to a variety of applications, including collaborative filtering (Marlin 2003), authorship (Rosen-Zvi et al. 2004), computer vision (Fei-Fei and Perona 2005), web blogs (Mei et al. 2006) and information retrieval (Wei and Croft 2006). For a good review, see Griffiths and Steyvers (2006).

Most previous topic models assume that the documents are part of a single corpus, and are exchangeable within it. For many text analysis problems, however, this assumption is not appropriate. For example, papers from different scientific conferences and journals can be viewed as a collection of multiple corpora, related to each other in as much as they discuss similar scientific themes. Articles from newspapers and blogs can also be viewed as multiple corpora, again related to each other in the overlap of their content.

In this paper we study the problem of modeling documents from different corpora, respecting the boundaries of the collections but accounting for and estimating the similarities among their content. Our intuition is that although documents across different corpora should not be assumed exchangeable, they may show different degrees of relationship. As an example, consider papers from multiple computer science conferences. In general, papers from ICML¹ are more likely to be similar to those in NIPS², rather than those in SIGIR³. However, some papers in ICML and SIGIR—specifically those dealing with text processing and information retrieval—can be very similar as well. As the different *topics* can be considered high-level semantic summarizations of a corpus from different aspects, our goal is to discover the relations in the *topic* level among multiple corpora. Thus, the models are able to discover how ICML, SIGIR, and NIPS are correlated, rather than simply saying that ICML and NIPS are more similar.

We introduce *Markov topic models* (MTMs), which use Gaussian Markov random fields (GMRFs) to model the

*Part of this work was done when Chong Wang was an intern at Microsoft Research.

Appearing in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

¹ICML: International Conference of Machine Learning

²NIPS: Neural Information Processing Systems

³SIGIR: International Conference on Research and Development in Information Retrieval.

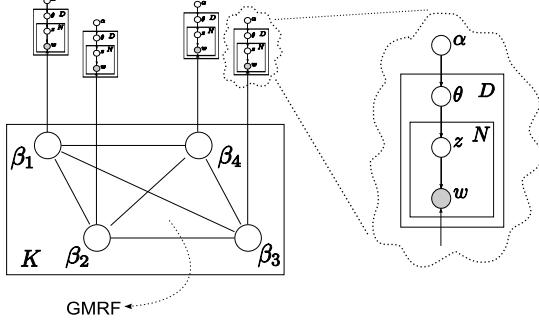


Figure 1: Graphical model for MTMs with multiple corpora. The left part illustrates high-level relations of topics among multiple corpora and the right part illustrates the local LDA model associated with each corpus.

topic relationships among multiple corpora. The models not only capture internal topic structures within each corpus, but also discover the relations between the topics across multiple corpora. Moreover, our approach provides a natural way for smoothing topic parameters using information from multiple collections.

We explain MTMs in detail in Section 2. In Section 3, we present an efficient variational EM algorithm for model learning. In Section 4, we present quantitative and qualitative results on an analysis of the abstracts from different computer science conferences. Our analysis reveals qualitative discoveries that are not possible with traditional topic models, and improved quantitative performance over the state of the art.

2 Markov Topic Models

The class of MTMs is an extension of LDA-based topic models, where we apply a Markovian framework to the topic parameters for different corpora. Figure 1 shows a graphical representation of a Markov topic model with four corpora. The topic parameters β_1, \dots, β_4 are vertices in a Markov random field that governs the relations between corpora, each modeled by an LDA topic model. The standard topic model, for one single corpus, and individual topic models, without any relations between corpora, are both special cases that we will consider in our empirical evaluation.

Before describing how an MTM addresses multiple corpora, we describe the standard topic modeling assumptions made for each. We assume that all V corpora cover the same set of W terms (this is accomplished by considering the union of terms across corpora). We will also assume that all corpora contain the same number of topics K . Following Blei et al. (2003), each document in a corpus is represented as a random mixture over latent topics, where each topic is characterized by a Multinomial distribution over the terms. Let $\beta_{v,k,1:W} = [\beta_{v,k,1}, \dots, \beta_{v,k,W}]^T$ be the W -

dimensional vector of parameters for topic k , $1 \leq k \leq K$ in corpus v , $1 \leq v \leq V$.⁴

Given the (marginal) distributions over terms $\beta_{v,1:K,1:W}$ for the K topics at corpus v , the generative process for that corpus is defined by a local LDA model as follows:

For each document d , $1 \leq d \leq D_v$ in corpus v :

1. Draw $\theta_{v,d} \sim \text{Dir}(\alpha_v)$.
2. For each word $w_{v,d,n}$ in the document d :
 - (a) Draw $z_{v,d,n} \sim \text{Mult}(\theta_{v,d})$.
 - (b) Draw $w_{v,d,n} \sim \text{Mult}(\beta_{v,z_{v,d,n},1:W})$.

Note that α_v and $\theta_{v,d}$ are both K -length vectors.⁵

We now turn to the topic distributions, where our goal is to statistically tie these parameters across corpora. The standard representation of a Multinomial distribution is by its mean parameters, with uncertainty about these parameters represented by the conjugate Dirichlet distribution (Griffiths and Steyvers 2006). We instead represent a Multinomial topic-parameter distribution by its *natural parameters* in the exponential family representation, and we model uncertainty about this parameter by a Gaussian (Aitchison 1982). The w 'th mean parameter of the W -dimensional multinomial is denoted π_w . The w 'th natural parameter is the mapping $\beta_w = \log(\pi_w / \pi_W)$, and the reverse mapping is $\pi_w = \exp(\beta_w) / \sum_w \exp(\beta_w)$.

In a MTM, the (marginal) topic parameters associated with local LDA models for different corpora are related, as the graphical structure in Figure 1 suggests. We are therefore considering a huge $V \times K \times W$ dimensional joint Gaussian over all topic parameters in the model with mean \mathbf{m} and precision Δ (The corresponding covariance matrix is $\Sigma = \Delta^{-1}$). That is,

$$\beta_{1:V,1:K,1:W} \sim \mathcal{N}_{V \times K \times W}(\mathbf{m}, \Delta). \quad (1)$$

We apply several constraints to this Gaussian. First, we assume that the per-term parameters across the K topics are mutually independent, as is standard for topic models.

Second, a topic is characterized by the terms with high probabilities in the topic distribution over terms, and different topics will typically focus on different terms. Given a particular topic we tie the mean for a particular term to the same value across corpora. That is $m_{v,k,w} = m_{k,w}$ for

⁴We use subscripts to indicate a particular value for a dimension (e.g. for a corpus, topic, or term) and colon notation (e.g. $1:W$) to denote a range of values. We use various combinations of subscripts and ranges to denote relevant sets of parameters.

⁵We don't write α_v and $\theta_{v,d}$ as $\alpha_{v,1:K}$ and $\theta_{v,d,1:K}$ explicitly, since we don't access $\alpha_{v,1:k}$ and $\theta_{v,d,k}$, $1 \leq k \leq K$, individually in the rest of the paper.

all $v, 1 \leq v \leq V$. This constraint ensures that topics vary smoothly and consistently across corpora.

Third, for simplicity, we assume that topic parameters associated with different terms are mutually independent. In other words, the probability for a particular term in a topic is only directly affected by the probabilities for the same term across corpora. With this additional constraint, the precision matrix $\Delta_{1:V,1:K,1:W}$ is block-diagonal with blocks $\Delta_{1:V,k,w}$, $1 \leq k \leq K$ and $1 \leq w \leq W$.

We further experimented with tying the blocks of precision parameters across words to the same value. That is, $\Delta_{1:V,k,w} = \Delta_{1:V,k}$ for all w . We found that this constraint is too simplistic for the problem at hand. In this case, the precisions associated with the many terms with low probabilities—which in fact do not characterize the topics—overwhelmed the estimation of the tied precisions. Terms with higher topic parameter values are more important to a topic. In order to ensure dominance of the topic parameters for the characteristic terms, we instead scale the tying by the weight of the expected number of observations for each term. That is, the block of precision parameters associated with term w is scaled by the factor

$$g_{k,w} = W \frac{\exp(m_{k,w})}{\sum_w \exp(m_{k,w})}. \quad (2)$$

Note that $\sum_w g_{k,w} = W$. If we set $g_w \equiv 1$, we return to the unscaled model.

Putting these three constraints together, the parameterization of the distribution in (1) simplifies dramatically. The distribution can now be represented by K independent $V \times W$ -dimensional block-diagonal Gaussians with a V dimensional block for each term w . Each block defines the distribution for a term in a specific topic across corpora, and is constrained as follows,

$$\beta_{1:V,1:K,1:W} \sim \prod_{k=1}^K \prod_{w=1}^W \mathcal{N}(m_{k,w} \mathbf{1}_{1:V}, g_{k,w} \Delta_{1:V,k}), \quad (3)$$

where $\mathbf{1}_{1:V}$ denotes a V dimensional vector of ones.

Finally, in a Markov topic model, structural relations between corpora may restrict the distributions in (3) further. For example, the corpora could be local news stories and one could have reason to believe that topic parameters evolve smoothly through a geo-spatial structure of neighboring locations. The structure in this way defines the Markov properties that the distribution for $\beta_{1:V,k,w}$ has to obey, i.e., a GMRF. Alternatively to the a priori decided structural constraints one could also choose to learn a structure via model selection methods, e.g., Meinshausen and Bühlmann (2006).

In some modeling situations, we would like multiple corpora to share a set of common “background” topics. Background topics can be modeled as isolated topics in the

GMRF representation. Notice that if all topics in the model are background topics, the model simplifies to a standard LDA model (with logistic normal smoothing of the topic parameters). The generative process of MTMs with B shared background topics is a simple extension of basic MTMs. To generate a document, we follow the same procedure, as described in this section, except that we will now for each corpus consider $K + B$ topics instead of just the K corpus specific topics.

3 Inference and Estimation

In this section, we present the approximate inference and parameter estimation for MTMs. The models are learned by the variational EM algorithm, which are described in the following two sections.

3.1 Approximate Inference: E-step

The E-step computes the posterior distribution of the latent topic structure conditioned on the observed documents, and the current values for the GMRF parameterization of the topic distributions (defined by $\mathbf{m}_{1:K,1:W}$ and $\Delta_{1:V,1:K}$). In a MTM, the latent topic structure comprises the per-document topic proportions at each corpus $\theta_{v,d}$, the per-word topic assignments at each corpus $z_{v,d,n}$, and the K Markov structures of topics $\beta_{1:V,k,1:W}$. The true posterior is not tractable. We appeal to an approximation.

We derive an efficient variational approximation for MTMs. The main idea behind variational methods is to posit a simple family of distributions over the latent variables, indexed by free *variational parameters*, and to find the member of that family which is closest in Kullback-Leibler divergence to the true posterior. Good overviews of this methodology can be found in Jordan et al. (1999) and Wainwright and Jordan (2003). The fully-factorized variational distribution over the latent variables is:

$$q(\beta, z, \theta | \hat{\beta}, \phi, \gamma) = \prod_{k=1}^K \prod_{w=1}^W q(\beta_{1:V,k,w} | \hat{\beta}_{1:V,k,w}) \times \prod_{v=1}^V \prod_{d=1}^{D_v} \left(q(\theta_{v,d} | \gamma_{v,d}) \prod_{n=1}^{N_{v,d}} q(z_{v,d,n} | \phi_{v,d,n}) \right). \quad (4)$$

The free variational parameters are the Dirichlets $\gamma_{v,d}$ for the per-document topic proportions, the multinomials $\phi_{v,d,n}$ for each word’s topic assignment, and the variational parameters $\hat{\beta}_{1:V,k,w}$ for $\beta_{1:V,k,w}$. The updates for document-level variational parameters $\theta_{v,d}$ and $z_{v,d,n}$ follow similar forms of those in Blei et al. (2003), where the difference is that we replace the topic distribution parameters with their variational expectations.

We now turn to variational inference for the topic distributions. For clarity of presentation, we focus on a model with only one topic and assume that each corpus has only one document. These calculations are simpler versions of those we need for the full model, but exhibit the essential features of the algorithm. Generalization to the full model is straightforward.

In this case, we only need to compute $q(\beta|\hat{\beta})$. Note that we drop some indices to make the following part easier to follow. Specifically, we don't need subscript k and d anymore. Further simplifying notation, we use Δ ($\hat{\Delta}$) to represent $\Delta_{1:V}$ ($\hat{\Delta}_{1:V}$).

We use the following variational posterior, for term w ,

$$q(\beta_{1:V,w} | \hat{\beta}_{1:V,w}) = \varphi_V(\hat{\mathbf{m}}_{1:V,w}, \hat{\Delta}), \quad (5)$$

where $\hat{\beta}_{1:V,w} = \{\hat{\mathbf{m}}_{1:V,w}, \hat{\Delta}\}$ and $\varphi_V(\hat{\mathbf{m}}_{1:V,w}, \hat{\Delta})$ indicates the Gaussian density with mean $\hat{\mathbf{m}}_{1:V,w}$ and precision $\hat{\Delta}$. Unlike m_w , which is the same for all corpora, $\hat{\mathbf{m}}_{1:V,w}$ are free parameters to fit. $\hat{\Delta}$ is chosen to be the same for all terms, which is required for the numerical stability. We will see $\hat{\Delta}$ is able to preserve the structure of GMRF if Δ represents a non-dense GMRF.

Recall that, for simplicity, we assume each corpus has only one document. Let \mathbf{w}_v be the observed document for corpus v . With the variational posterior distributions (5) in hand, we turn to the details of posterior inference. Equivalent to minimizing KL is tightening the bound on the likelihood of the observations given by Jensen's inequality (Jordan et al. 1999),

$$\begin{aligned} \log p(\mathbf{w}_{1:V} | \mathbf{m}, \Delta) &\geq \mathbb{E}_q[\log p(\mathbf{w}_{1:V} | \beta)] + \mathbb{E}_q[\log p(\beta | \mathbf{m}, \Delta)] + H(q) \\ &= \mathcal{L}(\hat{\mathbf{m}}, \hat{\Delta}; \mathbf{m}, \Delta), \end{aligned} \quad (6)$$

where $H(q)$ is the entropy of the variational distribution. Now we expand the right side of Equation 6 term by term,

$$\begin{aligned} \mathbb{E}_q[\log p(\mathbf{w}_{1:V} | \beta)] &= \sum_v \mathbb{E}_q[\log p(\mathbf{w}_v | \beta_{v,1:W})] \\ &= \sum_v \sum_w n_{v,w} \mathbb{E}_q \left[\beta_{v,w} - \log \sum_w \exp(\beta_{v,w}) \right] \\ &\geq \sum_v \sum_w n_{v,w} \hat{m}_{v,w} \\ &\quad - \sum_v n_v \left(\log \sum_w \exp(\hat{m}_{v,w}) + \hat{\Sigma}_{v,v} \right), \end{aligned} \quad (7)$$

where the count of term w in document \mathbf{w}_v is $n_{v,w}$, $n_v = \sum_w n_{v,w}$ and $\hat{\Sigma}_{v,v}$ is the entry (v, v) in matrix $\hat{\Sigma} = \hat{\Delta}^{-1}$. The last inequality comes from Jensen's inequality.

$$\mathbb{E}_q[\log p(\beta | \mathbf{m}, \Delta)] = \sum_w \mathbb{E}_q[\log p(\beta_{1:V,w} | \mathbf{m}, \Delta)],$$

where

$$\begin{aligned} \mathbb{E}_q[\log p(\beta_{1:V,w} | \mathbf{m}, \Delta)] &= \\ &= -\frac{V}{2} \log 2\pi + \frac{V}{2} \log g_w + \frac{1}{2} \log |\Delta| - \frac{g_w}{2} \text{Tr}(\Delta \hat{\Sigma}) \\ &\quad - \frac{g_w}{2} (\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1})^T \Delta (\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1}). \end{aligned} \quad (8)$$

$$H(q) = \frac{VW}{2} \log 2\pi - \frac{W}{2} \log |\hat{\Delta}| + \frac{VW}{2}. \quad (9)$$

Now we proceed to compute the required derivatives for $\hat{\Delta}$ and $\hat{\mathbf{m}}_{1:V,w}$. Now, we isolate the terms that contain $\hat{\Delta}$,

$$\begin{aligned} \mathcal{L}_{[\hat{\Delta}]} &= -\frac{1}{2} \sum_v n_v \hat{\Sigma}_{v,v} - \frac{W}{2} \log |\hat{\Delta}| - \frac{W}{2} \text{Tr}(\Delta \hat{\Sigma}) \\ &= -\frac{W}{2} \left(\log |\hat{\Delta}| + \text{Tr}(\Delta + \text{diag}(\mathbf{n})/W) \right), \end{aligned} \quad (10)$$

where we have used $\sum_w g_w = W$ in the first "=". and $\mathbf{n} = [n_1, n_2, \dots, n_V]$. The optimal value of $\hat{\Delta}$ is obtained by:

$$\hat{\Delta} = \Delta + \text{diag}(\mathbf{n})/W, \quad (11)$$

where we use the following Equation 12:

$$\log |\mathbf{X}| + \text{Tr}(\mathbf{X}^{-1} \mathbf{A}) \geq \log |\mathbf{A}| + d, \quad (12)$$

where both \mathbf{X} and \mathbf{A} are $d \times d$ positive definitive matrixes and the equality holds if and only if $\mathbf{X} = \mathbf{A}$. Equation 11 means that to obtain $\hat{\Delta}$, one only needs to add a diagonal matrix $\text{diag}(\mathbf{n})/W$ to Δ . Then if Δ is sparse, $\hat{\Delta}$ preserves the sparsity. Recall that n_v is the counts of all terms in the corpus v . Then if n_v becomes larger ($\hat{\Delta}_{v,v}$ becomes larger and $\hat{\Sigma}_{v,v}$ becomes smaller), the marginal variational distribution of $q(\beta_{v,w})$ tends to peak at $\hat{m}_{v,w}$.

Numerical approaches, such as L-BFGS (Liu and Nocedal 1989), can be used to estimate $\hat{\mathbf{m}}_{1:V,w}$. After isolating the terms that contain $\hat{\mathbf{m}}_{1:V,w}$, we have

$$\begin{aligned} \mathcal{L}_{[\hat{\mathbf{m}}_{1:V,w}]} &= \sum_v \left(n_{v,w} \hat{m}_{v,w} - n_v \log \sum_w \exp(\hat{m}_{v,w}) \right) \\ &\quad - \frac{g_w}{2} (\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1})^T \Delta (\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1}). \end{aligned} \quad (13)$$

By taking the derivative w.r.t. $\hat{\mathbf{m}}_{1:V,w}$, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{m}}_{1:V,w}} &= \mathbf{n}_{1:V,w} - \zeta_{1:V,w} \\ &\quad - g_w \Delta (\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1}), \end{aligned} \quad (14)$$

where

$$\zeta_{v,w} = n_v \frac{\exp(\hat{m}_{v,w})}{\sum_w \exp(\hat{m}_{v,w})}. \quad (15)$$

3.2 Parameter estimation: M-step

Parameter estimation is done in the M-step that maximizes the lower bound of the log likelihood of the data obtained by the variational approximation in section 3.1. In other words, variational E-step computes the variational posterior $q(\beta, z, \theta)$ given the current settings of model parameter $\mathbf{m}_{1:K,1:W}$ and $\Delta_{1:V,1:K}$. Then M-step finds the maximum likelihood estimate of these model parameters. The variational EM runs alternatively between two steps until the lower bound converges.

Recall that we consider a single topic model here. Let $\Sigma = \Delta^{-1}$. First, isolating the terms that contain Δ from 6, we have

$$\begin{aligned} \mathcal{L}[\Delta] &= \frac{W}{2} \left(\log |\Delta| - \text{Tr}(\Delta \hat{\Delta}^{-1}) - \text{Tr}(\Delta M) \right) \\ &= \frac{W}{2} \left(\log |\Delta| - \text{Tr}(\Delta (\hat{\Delta}^{-1} + M)) \right) \\ &= -\frac{W}{2} \left(\log |\Sigma| + \text{Tr}(\Sigma^{-1} (\hat{\Delta}^{-1} + M)) \right), \end{aligned} \quad (16)$$

where

$$M = \frac{1}{W} \sum_w g_w (\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1})(\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1})^T. \quad (17)$$

Applying the Equation 12 to Equation 16, we obtain the optimal value of Δ as:

$$\Delta^{-1} = \Sigma = \hat{\Delta}^{-1} + M. \quad (18)$$

Clearly, M is a weighted combination by the relative importance of terms, g_w . According to the form of $\hat{\Delta}$ in equation 11, Δ is somehow determined by M and the counts of all terms (or expected counts for K topic models) for each corpus.

Second, isolating the terms that contain \mathbf{m} from 6, we have

$$\mathcal{L}[\mathbf{m}] = \frac{V}{2} \sum_w \log g_w - \frac{1}{2} \sum_w g_w f_w, \quad (19)$$

where

$$f_w = (\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1})^T \Delta (\hat{\mathbf{m}}_{1:V,w} - m_w \mathbf{1}). \quad (20)$$

To derive the derivative of m_w , we first compute

$$\frac{\partial g_{w'}}{\partial m_w} = \begin{cases} g_w(1 - g_w/W) & \text{if } w' = w \\ -g_w g_{w'}/W & \text{otherwise} \end{cases}$$

By taking the derivative w.r.t. m_w , we have

$$\begin{aligned} \frac{\partial \mathcal{L}[\mathbf{m}]}{\partial m_w} &= \frac{V}{2} (1 - g_w) \\ &\quad - \frac{g_w}{2} \left(f_w + f'_w - \frac{1}{W} \sum_w g_w f_w \right), \end{aligned} \quad (21)$$

Algorithm 1 IPF algorithm for Δ

Input: S, \mathcal{C} and initial guess Δ_0

Output: the optimal Δ_{opt}

repeat

for $a \in \mathcal{C}$ **do**

$\Delta_{aa} \leftarrow S_{aa}^{-1} + \Delta_{aa^c} \Delta_{a^c a^c}^{-1} \Delta_{a^c a}$

end for

until converge

where $f'_w = \partial f_w / \partial m_w$, a linear function of m_w .

What if Δ is sparse? If Δ is sparse, i.e. Δ represents a non-dense GMRF, it becomes difficult to obtain an analytical solution like equation 18. We then choose to use iterative proportional fitting (IPF) (Ruschendorf 1995). We outline the procedure as follows. Let $S = \hat{\Delta}^{-1} + M$. Recall that $\mathcal{L}[\Delta]$ can be written as

$$\begin{aligned} \mathcal{L}[\Delta] &= \frac{W}{2} \left(\log |\Delta| - \text{Tr}(\Delta (\hat{\Delta}^{-1} + M)) \right) \\ &= \frac{W}{2} \log |\Delta| - \frac{W}{2} \text{Tr}(\Delta S). \end{aligned} \quad (22)$$

Viewing S as the sufficient statistics for the Gaussian distribution $\mathcal{N}(\mathbf{0}, \Delta)$, this optimization falls in the IPF framework. Let G be the graph that Δ represents and \mathcal{C} be the collections cliques of G . For $a \in \mathcal{C}$, a^c (complement of a) contains all the other vertices in G . Define $\Delta_{ab} = \{\Delta_{i,j}\}_{(i,j) \in a \times b}$, $a, b \in \mathcal{C}$ and $S_{ab} = \{S_{i,j}\}_{(i,j) \in a \times b}$, $a, b \in \mathcal{C}$. Algorithm 1 computes the optimal Δ for equation 22.

4 Experimental Results

In this section, we demonstrate the use of MTMs on a multi-corpora dataset constructed from several international conferences held in the last few years. We report predictive perplexity, compared to LDA models, and interesting topical patterns. The Dirichlet parameter α is fixed to be a symmetric prior (2.0) for every model and we use a dense GMRF in the MTM.

4.1 Multi-corpora Dataset

We analyzed the abstracts from six international conferences: CIKM⁶, ICML, KDD⁷, NIPS, SIGIR and WWW⁸. These conferences were held between year 2005 and year 2008. The publications from these conferences cover a wide range of topics related to information processing. For example, CIKM mainly covers “databases”, “information

⁶ACM Conference on Information and Knowledge Management.

⁷ACM International Conference on Knowledge Discovery & Data Mining.

⁸International World Wide Web Conference.

CONF.	YEARS	#DOCS	#WORDS	AVG.WORDS
CIKM	05-07	410	27609	67.3
ICML	06-08	447	28419	63.6
KDD	06-08	374	29179	78.0
NIPS	07-08	355	25031	70.5
SIGIR	06-08	573	34607	60.4
WWW	07-08	439	27718	63.1
TOTAL	05-08	2598	172563	66.4

Table 1: Information about the multi-corpora dataset. The vocabulary size is 3733. Year: the years when the conferences were held; #Docs: the total number documents (abstracts of papers or posters); #Words: the total number of words; Avg.Words: the average number of words in a document.

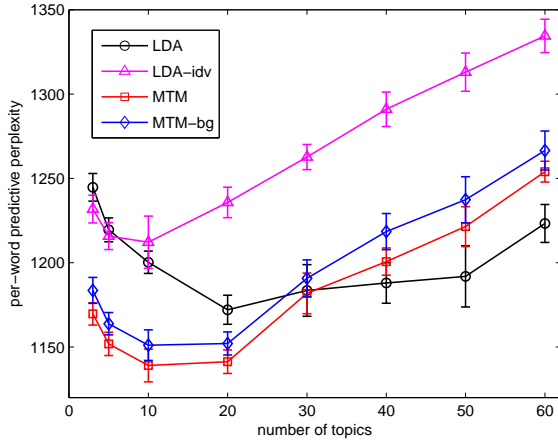


Figure 2: Per-word predictive perplexity comparison. MTM and MTM-bg achieve their best performances when the K is around 10, while LDA achieves its best performance when K is around 20. MTM gives the *lowest* predictive perplexity around $K = 10$.

retrieval” and “knowledge management”, while SIGIR focuses on all aspects of “information retrieval”. WWW covers all aspects of World Wide Web, also including “web information retrieval”. We expect that these conference are correlated in some sense. For example, artificial intelligence and machine learning techniques are studied and used in these areas, but in many different ways.

Abstracts from the same conference form a corpus. After pruning the vocabulary by removing the functional terms and the terms that occurred less than 5 times or in less than 3 documents, the entire dataset contains 170K words split among the 6 corpora. The vocabulary size is 3733. Table 1 shows some statistical information of these corpora.

4.2 Quantitative: Predictive Perplexity

In our quantitative evaluation, we compare the following models: a standard LDA model over all corpora (LDA),

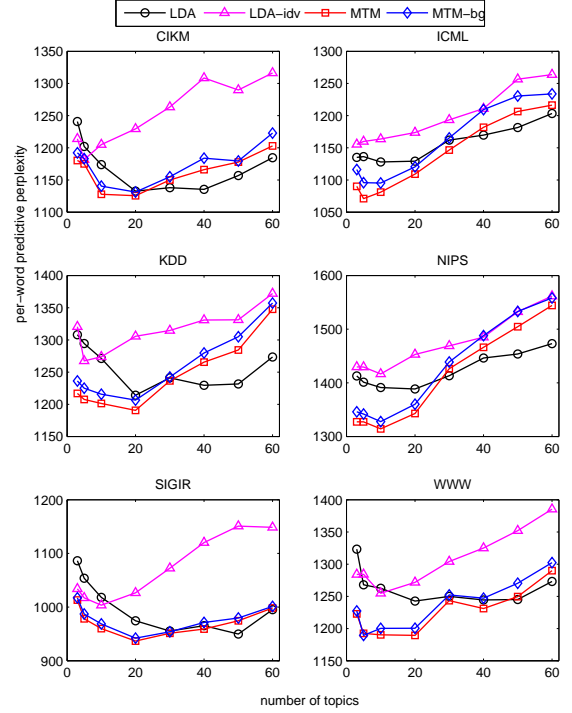


Figure 3: Per-word predictive perplexity comparison for each corpus (Standard errors are not shown). As we can see, MTM generally gives the best performance for all the corpora.

individual LDA models for each corpus (LDA-idv)⁹, the basic MTM (MTM), and an extension of the MTM with one background topic (MTM-bg). We use 5-fold cross validation for the evaluation. In each fold, 80% of the documents from each of the six conferences are chosen as the training set and the remaining 20% is used as the testing set. We compute the per-word predictive perplexity over a test dataset D_{test} as our test criterion. This perplexity is defined as

$$\text{perplexity}_{pw} = \exp \left\{ - \frac{\sum_{d \in D_{test}} \log p(\mathbf{w}_d | \beta)}{\sum_{d \in D_{test}} N_d} \right\}, \quad (23)$$

where β denotes all the estimated topic parameters in a model.

For LDA, we use variational inference to approximate $\log p(\mathbf{w}_d)$ with a lower bound (Blei et al. 2003). The situation is slightly different for the local LDA models in the MTM and MTM-bg. For these local models, we in fact learn variational posterior distributions for the topic parameters—see Section 3.2—and we instead use the mean values as the estimated parameterization. To be clear, for corpus v , the topic parameter for the k th-topic is estimated by $\tilde{\beta}_{v,k,w} \approx \exp(\hat{m}_{v,k,w}) / \sum_w \exp(\hat{m}_{v,k,w})$. The perplexity computation now proceeds as for a standard LDA

⁹We achieve this by removing all the edges in GMRF in the MTM.

model, except that we pick the estimated parameterization according to the corpus of each document.

We studied the performance of the models for a wide range of numbers of topics: $K = 3, 5, 10, 20, 30, 40, 50, 60$. Figure 2 shows the overall performance and figure 3 shows the performance over each corpus. (Note that lower perplexity is better.) We see that MTM and MTM-bg achieve the best perplexity around $K = 10$, and LDA achieves its best perplexity around $K = 20$. Most importantly, modeling interrelated local corpora with MTM and MTM-bg outperforms standard LDA and the individual LDA models, with MTM achieving the overall lowest predictive perplexity for this application.

All three models begin to overfit after $K = 20$. As K increases, the overfitting effect of MTM and MTM-bg is worse than for LDA. There is a natural explanation to this fact. In MTM and MTM-bg, each corpus (modeled by a local LDA model) has K topics, and these topics are tied to the topics from other corpora. Therefore, the “effective” number of topics for MTM or MTM-bg is larger than K , though smaller than KV . For each individual corpus, from figure 2, we can see similar results. (Note that for difference corpora, the numbers of topics for the best performance are not the same. How to discover the right number of topics for each corpus under the MTM framework is a question for future work.)

Observe that MTM-bg always has higher perplexity results than MTM, indicating that the background topic is not of use in this data. We do not expect this finding to carry over to different types of documents, but rather attribute it to the fact that we have been analyzing abstracts, where the writing style is quite constrained. In abstracts, people are only allowed to use a few concise sentences to outline the whole paper, and these sentences must therefore be very relevant to the main content. It is unlikely that all abstracts would share the same background information.

4.3 Qualitative: Topic Pattern Discovery

The analysis in this section is based on the 10-topic MTM of the previous section. In Figure 4, we visualize the correlation coefficients (scaled) for two topics using the covariance matrixes from the variational posterior distributions. The whiter the square is, the more correlated the two conferences are on this topic. Figure 4(a) and 4(b) correspond to Table 2 and Table 3, where we visualize the topics using top 12 terms due to the limited space. In Figure 4(a), the topic is about *clustering*, where almost all the conferences have “clustering, data, similarity” in top 12 terms. However, different conferences may have different aspect on this *clustering* topic. Among these, for example, we see that ICML and NIPS are highly correlated, they also share “graph, kernels, spectral”, while CIKM and WWW are also quite correlated on “pattern, mining”. Another example is

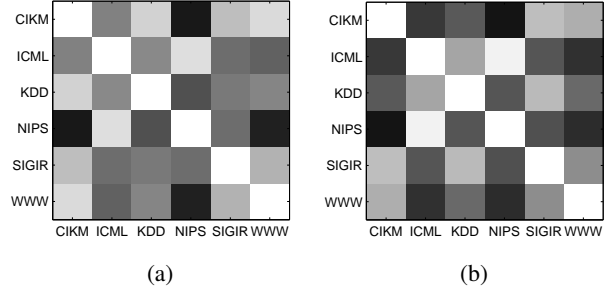


Figure 4: Correlation coefficient analysis (rescaled). (a) The correlation coefficient analysis of the topics in Table 2. (b) The correlation coefficient analysis of the topics in Table 3.

shown in Table 3, the topic is about *learning & classification*. ICML and NIPS are mainly on the theoretical side (NIPS also has image classification papers though), while CIKM, SIGIR and WWW are on the application side. KDD seems right in the middle.

5 Related Work

Previous work, including *dynamic topic models* (DTM) (Blei and Lafferty 2006) and *continuous time dynamic topic models* (cDTM) (Wang et al. 2008), has studied the problem of topic evolution when time information is available. If documents from the same time period are considered as a corpus, then DTM and cDTM are within the framework of MTM by designing a precision matrix that only allows dependence along the time line.

Several topic models have considered meta information, such as times, locations or authors, in estimating topics (Wang and McCallum 2006; Mei et al. 2006, 2008; Rosen-Zvi et al. 2004). In principle, corpus assignment can be considered a type of meta information. However, all of these previous models assume a single set of global and independent topics. Methods such as these do not provide a mechanism for modeling topic relations among multiple corpora, as we have developed here for MTM.

6 Conclusions

In this paper, we developed MTMs for simultaneously modeling multiple corpora. Across corpora, MTMs use GMRFs to model the correlations between their topics. These models not only capture the internal topic structures within one corpus, but also discover the relationships of the topics across many.

While we examined MTMs in the context of LDA-based models, we emphasize that the MTM framework can be integrated into many other topic models. The inference and estimation procedures provide a general way of incorporating multiple corpora into topic analysis. In future work, we plan to study other datasets, e.g., local news articles,

topic: <i>clustering</i>					
CIKM	ICML	KDD	NIPS	SIGIR	WWW
clustering	clustering	clustering	clustering	clustering	spam
data	graph	data	graph	semantic	clustering
similarity	data	mining	similarity	similarity	similarity
algorithms	kernels	patterns	data	filtering	mining
algorithm	constraints	algorithm	cluster	based	detection
patterns	relational	frequent	clusters	document	algorithms
time	based	algorithms	algorithms	cluster	extraction
mining	similarity	clusters	matching	information	based
method	pairwise	set	spectral	spam	data
set	cluster	cluster	kernels	clusters	web
series	spectral	graph	shape	algorithm	patterns
based	algorithms	pattern	set	items	existing

Table 2: The corresponding topic visualization of Figure 4(a).

topic: <i>learning & classification</i>					
CIKM	ICML	KDD	NIPS	SIGIR	WWW
classification	learning	model	learning	classification	learning
learning	model	data	model	text	models
text	data	classification	data	image	topic
features	models	models	models	features	images
training	algorithm	learning	image	learning	classification
models	bayesian	labels	inference	labeled	image
classifier	approach	training	bayesian	data	text
model	using	labeling	structure	training	topics
image	structure	labeled	features	using	approach
approach	semi-supervised	algorithm	classification	classifier	method
categorization	markov	text	using	algorithm	features
based	multiple	multiple	images	segmentation	framework

Table 3: The corresponding topic visualization of Figure 4(b).

and explore other possible representations of relationships between topics.

Acknowledgments David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520, and grants from Google and Microsoft.

References

- J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44(2):139–177, 1982.
- D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, 2006.
- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. ISSN 1533-7928.
- W. Buntine. Applying discrete PCA in data analysis. In *UAI*. AUAI Press, 2004.
- L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.
- T. Griffiths and M. Steyvers. Probabilistic topic models. In *Latent Semantic Analysis: A Road to Meaning*, 2006.
- M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.
- D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(3):503–528, 1989.
- B. Marlin. Modeling user rating profiles for collaborative filtering. In *NIPS*. MIT Press, 2003.
- Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*. ACM, 2006.
- Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, 2008.
- N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.
- M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI*, 2004.
- L. Ruschendorf. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, 23(4):1160–1174, 1995.
- M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. Technical Report 649, UC Berkeley, Dept. of Statistics, 2003.
- C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *UAI*, 2008.
- X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD*, 2006.
- X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR*, New York, NY, USA, 2006. ACM.