Correlated Topic Models

David M. Blei **John D. Lafferty**
School of Computer Science
Carnegie Mellon University

Abstract

Topic models, such as latent Dirichlet allocation (LDA), have been an effective tool for the statistical analysis of document collections and other discrete data. The LDA model assumes that the words of each document arise from a mixture of *topics*, each of which is a distribution over the vocabulary. A limitation of LDA is the inability to model topic correlation even though, for example, a document about sports is more likely to also be about health than international finance. This limitation stems from the use of the Dirichlet distribution to model the variability among the topic proportions. In this paper we develop the correlated topic model (CTM), where the topic proportions exhibit correlation via the logistic normal distribution [1]. We derive a mean-field variational inference algorithm for approximate posterior inference in this model, which is complicated by the fact that the logistic normal is not conjugate to the multinomial. The CTM gives a better fit than LDA on a collection of OCRed articles from the journal *Science*. Furthermore, the CTM provides a natural way of visualizing and exploring this and other unstructured data sets.

1 Introduction

The availability and use of unstructured historical collections of documents is rapidly growing. As one example, JSTOR (`www.jstor.org`) is a not-for-profit organization that maintains a large online scholarly journal archive obtained by running an optical character recognition (OCR) engine over the original printed journals. JSTOR indexes the resulting text and provides online access to the scanned images of the original content through keyword search. This provides an extremely useful service to the scholarly community, with the collection comprising nearly three million published articles in a variety of fields.

The sheer volume of this unstructured and noisy archive naturally suggests opportunities for the effective use of statistical modeling. For instance, a scholar in a narrow subdiscipline, searching for a particular research article, would certainly be interested to learn that the topic of that article is highly correlated with another topic that the researcher may not have known about and that is not explicitly contained in the article. Alerted to the existence of this new related topic, the researcher could browse the collection in a topic-guided manner to begin to investigate connections to a previously unrecognized body of work. Since the archive comprises millions of articles spanning centuries of scholarly work, automated analysis is essential.

Several statistical models have recently been developed for automatically extracting the

topical structure of large document collections. In technical terms, a topic model is a generative probabilistic model that uses a small number of distributions over a vocabulary to describe a document collection. When fit from data, these distributions often correspond to intuitive notions of topicality. In this work, we build upon the latent Dirichlet allocation (LDA) [3] model. LDA assumes that the words of each document arise from a mixture of topics. The topics are shared by all documents in the collection; the topic proportions are document-specific and randomly drawn from a Dirichlet distribution. LDA allows each document to exhibit multiple topics with different proportions, and it can thus capture the heterogeneity in grouped data which exhibit multiple latent patterns. Recent work has used LDA as a module in more complicated document models [8, 10, 6], and in a variety of settings such as image processing [11], collaborative filtering [7], disability survey data [4], population genetics [9], and the modeling of sequential data and user profiles [5].

Our goal in this paper is to address a limitation of the topic models proposed to date: they fail to directly model correlation between topics. In many—indeed most—text corpora, it is natural to expect that subsets of the underlying latent topics will be highly correlated. In a corpus of scientific articles, for instance, an article about genetics may be likely to also be about health and disease, but unlikely to also be about x-ray astronomy. For the LDA model, this limitation stems from the independence assumptions implicit in the Dirichlet distribution on the topic proportions. Under a Dirichlet, the components of the proportions vector are nearly independent; this leads to the strong and unrealistic modeling assumption that the presence of one topic is not correlated with the presence of another.

In this paper we present the *correlated topic model* (CTM). The CTM uses an alternative, more flexible distribution for the topic proportions that allows for covariance structure among the components. This gives a more realistic model of latent topic structure where the presence of one latent topic may be correlated with the presence of another. In the following sections we develop the technical aspects of this model, and then demonstrate its potential for the applications envisioned above. We fit the model to a portion of the JSTOR archive of the journal *Science*. We demonstrate that the model gives a better fit than LDA, as measured by the accuracy of the predictive distributions over held out documents. Furthermore, we demonstrate qualitatively that the correlated topic model provides a natural way of visualizing and exploring such an unstructured collection of textual data.

2 The Correlated Topic Model

The key to the correlated topic model we propose is the logistic normal distribution [1]. The logistic normal is a distribution on the simplex that allows for a general pattern of variability between the components by transforming a multivariate normal random variable. Consider the *natural parameterization* of a K-dimensional multinomial distribution:

$$p(z \mid \eta) = \exp\{\eta^T z - a(\eta)\}. \tag{1}$$

The random variable Z can take on K values; it is represented by a K-vector with one component equal to one to denote a value in $\{1, \ldots, K\}$. The cumulant generating function is:

$$a(\eta) = \log\left(\textstyle\sum_{i=1}^{K} \exp\{\eta_i\}\right). \tag{2}$$

The mapping between the mean parameterization (i.e., the simplex) and the natural parameterization is:

$$\eta_i = \log \theta_i / \theta_K. \tag{3}$$

Notice that this is not the minimal exponential family representation of the multinomial because multiple values of η can yield the same mean parameter.

The logistic normal distribution assumes that η is normally distributed and then mapped to the simplex with the inverse of Eq. 3, that is, $f(\eta_i) = \exp \eta_i / \sum_j \exp \eta_j$. It describes

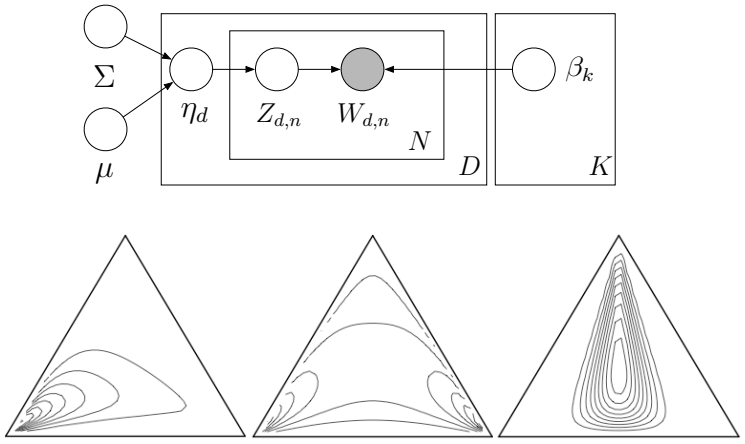

Figure 1: Top: Graphical model representation of the correlated topic model. The logistic normal distribution, used to model the latent topic proportions of a document, can represent correlations between topics that are impossible to capture using a single Dirichlet. Bottom: Example densities of the logistic normal on the 2-simplex. From left: diagonal covariance and nonzero-mean, negative correlation between components 1 and 2, positive correlation between components 1 and 2.

correlations between components of the simplicial random variable through the covariance matrix of the normal distribution. The logistic normal was originally studied in the context of analyzing observed compositional data such as the proportions of minerals in geological samples. In this work, we extend its use to a hierarchical model where it describes the *latent* composition of topics associated with each document.

Let $\{\mu, \Sigma\}$ be a K-dimensional mean and covariance matrix, and let topics $\beta_{1:K}$ be K multinomials over a fixed word vocabulary. The correlated topic model assumes that an N-word document arises from the following generative process:

1. Draw $\eta \mid \{\mu, \Sigma\} \sim \mathcal{N}(\mu, \Sigma)$.
2. For $n \in \{1, \ldots, N\}$:
   (a) Draw topic assignment $Z_n \mid \eta$ from Mult$(f(\eta))$.
   (b) Draw word $W_n \mid \{z_n, \beta_{1:K}\}$ from Mult(β_{z_n}).

This process is identical to the generative process of LDA except that the topic proportions are drawn from a logistic normal rather than a Dirichlet. The model is shown as a directed graphical model in Figure 1.

The CTM is more expressive than LDA. The strong independence assumption imposed by the Dirichlet in LDA is not realistic when analyzing document collections, where one may find strong correlations between topics. The covariance matrix of the logistic normal in the CTM is introduced to model such correlations. In Section 3, we illustrate how the higher order structure given by the covariance can be used as an exploratory tool for better understanding and navigating a large corpus of documents. Moreover, modeling correlation can lead to better predictive distributions. In some settings, such as collaborative filtering, the goal is to predict unseen items conditional on a set of observations. An LDA model will predict words based on the latent topics that the observations suggest, but the CTM has the ability to predict items associated with *additional* topics that are correlated with the conditionally probable topics.

2.1 Posterior inference and parameter estimation

Posterior inference is the central challenge to using the CTM. The posterior distribution of the latent variables conditional on a document, $p(\eta, z_{1:N} \mid w_{1:N})$, is intractable to compute; once conditioned on some observations, the topic assignments $z_{1:N}$ and log proportions η are dependent. We make use of mean-field variational methods to efficiently obtain an approximation of this posterior distribution.

In brief, the strategy employed by mean-field variational methods is to form a factorized distribution of the latent variables, parameterized by free variables which are called the variational parameters. These parameters are fit so that the Kullback-Leibler (KL) divergence between the approximate and true posterior is small. For many problems this optimization problem is computationally manageable, while standard methods, such as Markov Chain Monte Carlo, are impractical. The tradeoff is that variational methods do not come with the same theoretical guarantees as simulation methods. See [12] for a modern review of variational methods for statistical inference.

In graphical models composed of conjugate-exponential family pairs and mixtures, the variational inference algorithm can be automatically derived from general principles [2, 13]. In the CTM, however, the logistic normal is *not* conjugate to the multinomial. We will therefore derive a variational inference algorithm by taking into account the special structure and distributions used by our model.

We begin by using Jensen's inequality to bound the log probability of a document:

$$\log p(w_{1:N} \mid \mu, \Sigma, \beta) \;\geq \tag{4}$$
$$\mathrm{E}_q\left[\log p(\eta \mid \mu, \Sigma)\right] + \textstyle\sum_{n=1}^{N} \mathrm{E}_q\left[\log p(z_n \mid \eta)\right] + \mathrm{E}_q\left[\log p(w_n \mid z_n, \beta)\right] + \mathrm{H}\left(q\right),$$

where the expectation is taken with respect to a variational distribution of the latent variables, and $\mathrm{H}\left(q\right)$ denotes the entropy of that distribution. We use a factorized distribution:

$$q(\eta_{1:K}, z_{1:N} \mid \lambda_{1:K}, \nu^2_{1:K}, \phi_{1:N}) = \textstyle\prod_{i=1}^{K} q(\eta_i \mid \lambda_i, \nu_i^2) \prod_{n=1}^{N} q(z_n \mid \phi_n). \tag{5}$$

The variational distributions of the discrete variables $z_{1:N}$ are specified by the K-dimensional multinomial parameters $\phi_{1:N}$. The variational distribution of the continuous variables $\eta_{1:K}$ are K independent univariate Gaussians $\{\lambda_i, \nu_i\}$. Since the variational parameters are fit using a *single* observed document $w_{1:N}$, there is no advantage in introducing a non-diagonal variational covariance matrix.

The nonconjugacy of the logistic normal leads to difficulty in computing the expected log probability of a topic assignment:

$$\mathrm{E}_q\left[\log p(z_n \mid \eta)\right] = \mathrm{E}_q\left[\eta^T z_n\right] - \mathrm{E}_q\left[\log(\textstyle\sum_{i=1}^{K} \exp\{\eta_i\})\right]. \tag{6}$$

To preserve the upper bound on the log probability, we lower bound the negative log normalizer with a Taylor expansion:

$$\mathrm{E}_q\left[\log\left(\textstyle\sum_{i=1}^{K} \exp\{\eta_i\}\right)\right] \leq \zeta^{-1}(\textstyle\sum_{i=1}^{K} \mathrm{E}_q\left[\exp\{\eta_i\}\right]) - 1 + \log(\zeta), \tag{7}$$

where we have introduced a new variational parameter ζ. The expectation $\mathrm{E}_q\left[\exp\{\eta_i\}\right]$ is the mean of a log normal distribution with mean and variance obtained from the variational parameters $\{\lambda_i, \nu_i^2\}$: $\mathrm{E}_q\left[\exp\{\eta_i\}\right] = \exp\{\lambda_i + \nu_i^2/2\}$ for $i \in \{1, \ldots, K\}$.

Given a model $\{\beta_{1:K}, \mu, \Sigma\}$ and a document $w_{1:N}$, the variational inference algorithm optimizes Eq. 4 with respect to the variational parameters $\{\lambda_{1:K}, \nu_{1:K}, \phi_{1:N}, \zeta\}$. We use coordinate ascent, repeatedly optimizing with respect to each parameter while holding the others fixed. In variational inference for LDA, each coordinate can be optimized analytically. However, iterative methods are required for the CTM when optimizing for λ_i and ν_i^2. The details are given in Appendix A.

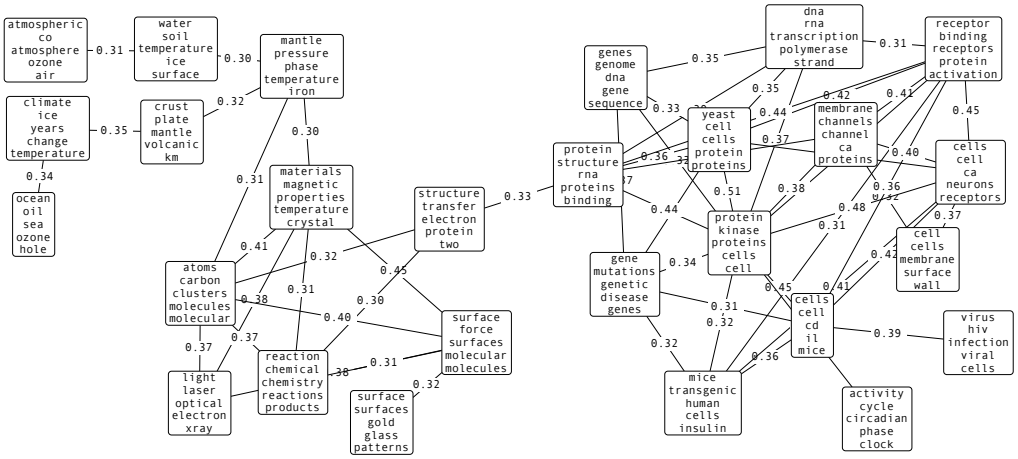

Figure 2: A portion of the topic graph learned from 15,744 OCR articles from *Science*. Each node represents a topic, and is labeled with the five most probable words from its distribution; edges are labeled with the correlation between topics.

Given a collection of documents, we carry out parameter estimation in the correlated topic model by attempting to maximize the likelihood of a corpus of documents as a function of the topics $\beta_{1:K}$ and the multivariate Gaussian (μ, Σ). We use variational expectation-maximization (EM), where we maximize the bound on the log probability of a collection given by summing Eq. 4 over the documents.

In the E-step, we maximize the bound with respect to the variational parameters by performing variational inference for each document. In the M-step, we maximize the bound with respect to the model parameters. This is maximum likelihood estimation of the topics and multivariate Gaussian using expected sufficient statistics, where the expectation is taken with respect to the variational distributions computed in the E-step. The E-step and M-step are repeated until the bound on the likelihood converges. In the experiments reported below, we run variational inference until the relative change in the probability bound of Eq. 4 is less than 0.0001%, and run variational EM until the relative change in the likelihood bound is less than 0.001%.

3 Examples and Empirical Results: Modeling *Science*

In order to test and illustrate the correlated topic model, we estimated a 100-topic CTM on 15,744 *Science* articles spanning 1971 to 1998. We constructed a graph of the latent topics and the connections among them by examining the most probable words from each topic and the between-topic correlations. Part of this graph is illustrated in Figure 2. In this subgraph, there are three densely connected collections of topics: material science, geology, and cell biology. Furthermore, an estimated CTM can be used to explore otherwise unstructured observed documents. In Figure 4, we list articles which are assigned to the cognitive science topic and articles which are assigned to both the cognitive science and visual neuroscience topics. The interested reader is invited to visit `http://www.cs.cmu.edu/~lemur/science/` to interactively explore this model, including the topics, their connections, and the articles that exhibit them.

We compared the CTM to LDA by fitting a smaller collection of articles to models of varying numbers of topics. This collection contains the 1,452 documents from 1960; we

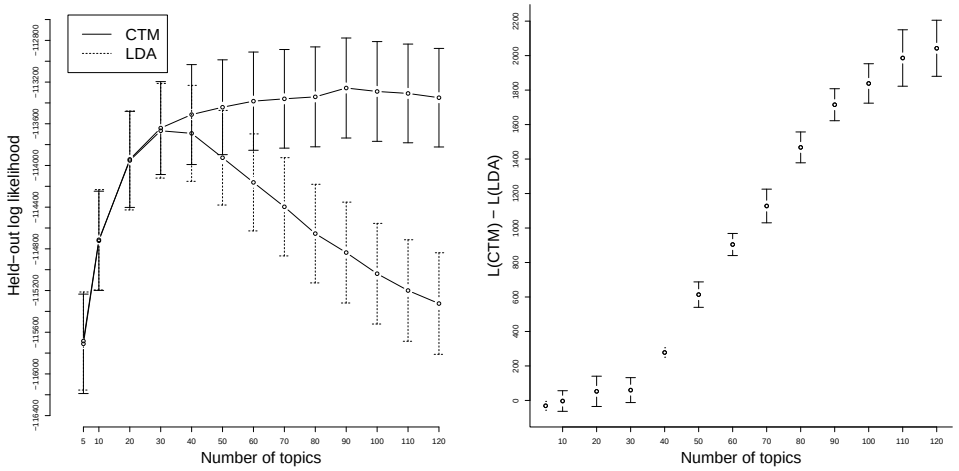

Figure 3: (L) The average held-out probability; CTM supports more topics than LDA. See figure at right for the standard error of the difference. (R) The log odds ratio of the held-out probability. Positive numbers indicate a better fit by the correlated topic model.

used a vocabulary of 5,612 words after pruning common function words and terms which occur once in the collection. We split the data into ten groups; for each group we computed the log probability of the held-out data given a model estimated from the remaining groups. A better model of the document collection will assign higher probability to the held out group. To avoid comparing bounds, we used importance sampling to compute the log probability of a document where the fitted variational distribution is the proposal.

Figure 3 illustrates the average held out log probability for each model and the average difference between them. The CTM provides a better fit than LDA and supports more topics; the likelihood for LDA peaks near 30 topics while the likelihood for the CTM peaks close to 90 topics. The means and standard errors of the *difference* in log-likelihood of the models is shown at right; this indicates that the CTM always gives a better fit.

Another quantitative evaluation of the relative strengths of LDA and the CTM is how well the models predict the remaining words after observing a portion of the document. Suppose we observe words $w_{1:P}$ from a document and are interested in which model provides a better predictive distribution $p(w \mid w_{1:P})$ of the remaining words. To compare these distributions, we use *perplexity*, which can be thought of as the effective number of equally likely words according to the model. Mathematically, the perplexity of a word distribution is defined as the inverse of the geometric per-word average of the probability of the observations. Note that lower numbers denote more predictive power.

The plot in Figure 4 compares LDA and the CTM in terms of predictive perplexity. When only a small number of words have been observed, the uncertainty about the remaining words under the CTM is much less than under LDA—the perplexity is reduced by nearly 200 words, or roughly 10%. The reason is that after seeing a few words in one topic, the CTM uses topic correlation to infer that words in a related topic may also be probable. In contrast, LDA cannot predict the remaining words as well until a large portion of the document as been observed so that all of its topics are represented.

Top Articles with
{brain, memory, learning}

(1) Distributed Neural Network Underlying Musical Sight-Reading and Keyboard Performance
(2) The Primate Hippocampal Formation: Evidence for a Time-Limited Role in Memory Storage
(3) Separate Neural Bases of Two Fundamental Memory Processes in the Temporal Lobe
(4) A Neostriatal Habit Learning System in Humans
(5) The Mental Representation of Hand Movement after Parietal Cortex Damage

Top Articles with
{brain, memory, learning} and {neurons, visual, cell}

(1) Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs
(2) Visual Instruction of the Neural Map of Auditory Space in the Developing Optic Tectum
(3) Visually Evoked Oscillations of Membrane Potential in Cells of Cat Visual Cortex
(4) Corticofugal Modulation of Time-Domain Processing of Biosonar Information in Bats
(5) A Map of Visual Space Induced in Primary Auditory Cortex

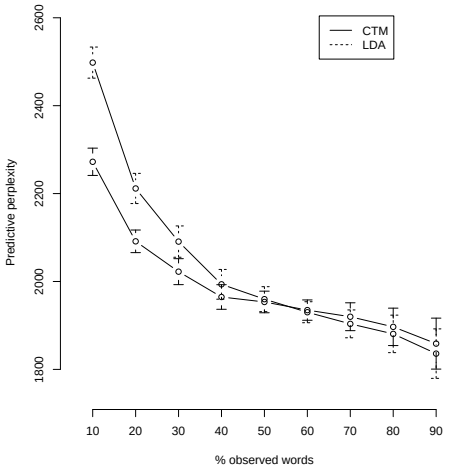

Figure 4: (Left) Exploring a collection through its topics. (Right) Predictive perplexity for partially observed held-out documents from the 1960 *Science* corpus.

References

[1] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44(2):139–177, 1982.

[2] C. Bishop, D. Spiegelhalter, and J. Winn. VIBES: A variational inference engine for Bayesian networks. In *NIPS 15*, pages 777–784. Cambridge, MA, 2003.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[4] E. Erosheva. *Grade of membership and latent structure models with application to disability survey data*. PhD thesis, Carnegie Mellon University, Department of Statistics, 2002.

[5] M. Girolami and A. Kaban. Simplicial mixtures of Markov chains: Distributed modelling of dynamic user profiles. In *NIPS 16*, pages 9–16, 2004.

[6] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, 2005.

[7] B. Marlin. Collaborative filtering: A machine learning perspective. Master's thesis, University of Toronto, 2004.

[8] A. McCallum, A. Corrada-Emmanuel, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks. 2004.

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

[10] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smith. In *UAI '04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494.

[11] J. Sivic, B. Rusell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. Technical report, CSAIL, MIT, 2005.

[12] M. Wainwright and M. Jordan. A variational principle for graphical models. In *New Directions in Statistical Signal Processing*, chapter 11. MIT Press, 2005.

[13] E. Xing, M. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of UAI*, 2003.

A Variational Inference

We describe a coordinate ascent optimization algorithm for the likelihood bound in Eq. 4 with respect to the variational parameters.

The first term of Eq. 4 is:

$$\mathrm{E}_q\left[\log p(\eta \mid \mu, \Sigma)\right] = (1/2)\log|\Sigma^{-1}| - (K/2)\log 2\pi - (1/2)\mathrm{E}_q\left[(\eta-\mu)^T\Sigma^{-1}(\eta-\mu)\right], \tag{8}$$

where

$$\mathrm{E}_q\left[(\eta-\mu)^T\Sigma^{-1}(\eta-\mu)\right] = \nu^{2T}\mathrm{diag}(\Sigma^{-1}) + (\lambda-\mu)^T\Sigma^{-1}(\lambda-\mu). \tag{9}$$

The second term of Eq. 4, using the additional bound in Eq. 7, is:

$$\mathrm{E}_q\left[\log p(z_n \mid \eta)\right] = \textstyle\sum_{i=1}^K \lambda_i\phi_{n,i} - \zeta^{-1}\left(\sum_{i=1}^K \exp\{\lambda_i + \nu_i^2/2\}\right) + 1 - \log\zeta. \tag{10}$$

The third term of Eq. 4 is:

$$\mathrm{E}_q\left[\log p(w_n \mid z_n, \beta)\right] = \textstyle\sum_{i=1}^K \phi_{n,i}\log\beta_{i,w_n}. \tag{11}$$

Finally, the fourth term is the entropy of the variational distribution:

$$\textstyle\sum_{i=1}^K \frac{1}{2}(\log\nu_i^2 + \log 2\pi + 1) - \sum_{n=1}^N\sum_{i=1}^k \phi_{n,i}\log\phi_{n,i}. \tag{12}$$

We maximize the bound in Eq. 4 with respect to the variational parameters $\lambda_{1:K}$, $\nu_{1:K}$, $\phi_{1:N}$, and ζ. We use a coordinate ascent algorithm, iteratively maximizing the bound with respect to each parameter.

First, we maximize Eq. 4 with respect to ζ, using the second bound in Eq. 7. The derivative with respect to ζ is:

$$f'(\zeta) = N\left(\zeta^{-2}\left(\textstyle\sum_{i=1}^K \exp\{\lambda_i + \nu_i^2/2\}\right) - \zeta^{-1}\right), \tag{13}$$

which has a maximum at:

$$\hat{\zeta} = \textstyle\sum_{i=1}^K \exp\{\lambda_i + \nu_i^2/2\}. \tag{14}$$

Second, we maximize with respect to ϕ_n. This yields a maximum at:

$$\hat{\phi}_{n,i} \propto \exp\{\lambda_i\}\beta_{i,w_n}, \quad i \in \{1,\ldots,K\}. \tag{15}$$

Third, we maximize with respect to λ_i. Eq. 4 is not amenable to analytic maximization. We use the conjugate gradient algorithm with derivative

$$dL/d\lambda = -\Sigma^{-1}(\lambda-\mu) + \textstyle\sum_{n=1}^N \phi_{n,1:K} - (N/\zeta)\exp\{\lambda + \nu^2/2\} \tag{16}$$

Finally, we maximize with respect to ν_i^2. Again, there is no analytic solution. We use Newton's method for each coordinate, constrained such that $\nu_i > 0$:

$$dL/d\nu_i^2 = -\Sigma_{ii}^{-1}/2 - N/2\zeta\exp\{\lambda + \nu_i^2/2\} + 1/(2\nu_i^2). \tag{17}$$

Iterating between these optimizations defines a coordinate ascent algorithm on Eq. 4.