Dynamic Topic Models

David M. Blei BLEI@CS.PRINCETON.EDU
Computer Science Department, Princeton University, Princeton, NJ 08544, USA

John D. Lafferty LAFFERTY@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh PA 15213, USA

Abstract

A family of probabilistic time series models is developed to analyze the time evolution of topics in large document collections. The approach is to use state space models on the natural parameters of the multinomial distributions that represent the topics. Variational approximations based on Kalman filters and nonparametric wavelet regression are developed to carry out approximate posterior inference over the latent topics. In addition to giving quantitative, predictive models of a sequential corpus, dynamic topic models provide a qualitative window into the contents of a large document collection. The models are demonstrated by analyzing the OCR'ed archives of the journal *Science* from 1880 through 2000.

1. Introduction

Managing the explosion of electronic document archives requires new tools for automatically organizing, searching, indexing, and browsing large collections. Recent research in machine learning and statistics has developed new techniques for finding patterns of words in document collections using hierarchical probabilistic models (Blei et al., 2003; McCallum et al., 2004; Rosen-Zvi et al., 2004; Griffiths and Steyvers, 2004; Buntine and Jakulin, 2004; Blei and Lafferty, 2006). These models are called "topic models" because the discovered patterns often reflect the underlying topics which combined to form the documents. Such hierarchical probabilistic models are easily generalized to other kinds of data; for example, topic models have been used to analyze images (Fei-Fei and Perona, 2005; Sivic et al., 2005), biological data (Pritchard et al., 2000), and survey data (Erosheva, 2002).

In an exchangeable topic model, the words of each document are assumed to be independently drawn from a mixture of multinomials. The mixing proportions are randomly drawn for each document; the mixture components, or topics, are shared by all documents. Thus, each document reflects the components with different proportions. These models are a powerful method of dimensionality reduction for large collections of unstructured documents. Moreover, posterior inference at the document level is useful for information retrieval, classification, and topic-directed browsing.

Treating words exchangeably is a simplification that it is consistent with the goal of identifying the semantic themes within each document. For many collections of interest, however, the implicit assumption of exchangeable *documents* is inappropriate. Document collections such as scholarly journals, email, news articles, and search query logs all reflect evolving content. For example, the *Science* article "The Brain of Professor Laborde" may be on the same scientific path as the article "Reshaping the Cortical Motor Map by Unmasking Latent Intracortical Connections," but the study of neuroscience looked much different in 1903 than it did in 1991. The themes in a document collection evolve over time, and it is of interest to explicitly model the dynamics of the underlying topics.

In this paper, we develop a dynamic topic model which captures the evolution of topics in a sequentially organized corpus of documents. We demonstrate its applicability by analyzing over 100 years of OCR'ed articles from the journal *Science*, which was founded in 1880 by Thomas Edison and has been published through the present. Under this model, articles are grouped by year, and each year's articles arise from a set of topics that have evolved from the last year's topics.

In the subsequent sections, we extend classical state space models to specify a statistical model of *topic evolution*. We then develop efficient approximate posterior inference techniques for determining the evolving topics from a sequential collection of documents. Finally, we present qualitative results that demonstrate how dynamic topic models allow the exploration of a large document collection in new

Appearing in *Proceedings of the 23^{rd} International Conference on Machine Learning*, Pittsburgh, PA, 2006.

ways, and quantitative results that demonstrate greater predictive accuracy when compared with static topic models.

2. Dynamic Topic Models

While traditional time series modeling has focused on continuous data, topic models are designed for categorical data. Our approach is to use state space models on the natural parameter space of the underlying topic multinomials, as well as on the natural parameters for the logistic normal distributions used for modeling the document-specific topic proportions.

First, we review the underlying statistical assumptions of a static topic model, such as latent Dirichlet allocation (LDA) (Blei et al., 2003). Let $\beta_{1:K}$ be K topics, each of which is a distribution over a fixed vocabulary. In a static topic model, each document is assumed drawn from the following generative process:

1. Choose topic proportions θ from a distribution over the $(K-1)$-simplex, such as a Dirichlet.
2. For each word:
   (a) Choose a topic assignment $Z \sim \textit{Mult}(\theta)$.
   (b) Choose a word $W \sim \textit{Mult}(\beta_z)$.

This process implicitly assumes that the documents are drawn *exchangeably* from the same set of topics. For many collections, however, the order of the documents reflects an evolving set of topics. In a dynamic topic model, we suppose that the data is divided by time slice, for example by year. We model the documents of each slice with a K-component topic model, where the topics associated with slice t evolve from the topics associated with slice $t-1$.

For a K-component model with V terms, let $\beta_{t,k}$ denote the V-vector of natural parameters for topic k in slice t. The usual representation of a multinomial distribution is by its mean parameterization. If we denote the mean parameter of a V-dimensional multinomial by π, the ith component of the *natural parameter* is given by the mapping $\beta_i = \log(\pi_i/\pi_V)$. In typical language modeling applications, Dirichlet distributions are used to model uncertainty about the distributions over words. However, the Dirichlet is not amenable to sequential modeling. Instead, we chain the natural parameters of each topic $\beta_{t,k}$ in a state space model that evolves with Gaussian noise; the simplest version of such a model is

$$\beta_{t,k} \mid \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, \sigma^2 I)\,. \tag{1}$$

Our approach is thus to model sequences of compositional random variables by chaining Gaussian distributions in a dynamic model and mapping the emitted values to the simplex. This is an extension of the logistic normal distribution (Aitchison, 1982) to time-series simplex data (West and Harrison, 1997).

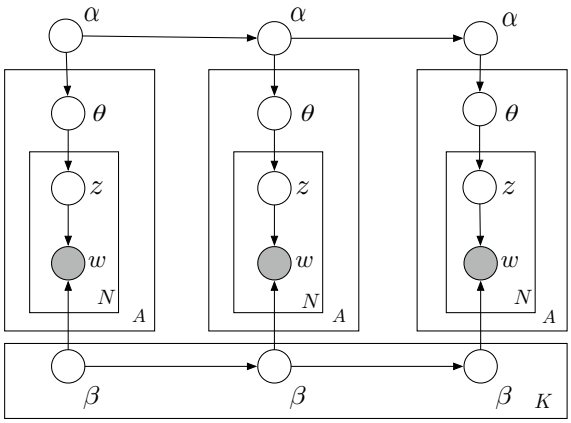

Figure 1. Graphical representation of a dynamic topic model (for three time slices). Each topic's natural parameters $\beta_{t,k}$ evolve over time, together with the mean parameters α_t of the logistic normal distribution for the topic proportions.

In LDA, the document-specific topic proportions θ are drawn from a Dirichlet distribution. In the dynamic topic model, we use a logistic normal with mean α to express uncertainty over proportions. The sequential structure between models is again captured with a simple dynamic model

$$\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)\,. \tag{2}$$

For simplicity, we do not model the dynamics of topic correlation, as was done for static models by Blei and Lafferty (2006).

By chaining together topics and topic proportion distributions, we have sequentially tied a collection of topic models. The generative process for slice t of a sequential corpus is thus as follows:

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
   (b) For each word:
      i. Draw $Z \sim \textit{Mult}(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim \textit{Mult}(\pi(\beta_{t,z}))$.

Note that π maps the multinomial natural parameters to the mean parameters, $\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}$.

The graphical model for this generative process is shown in Figure 1. When the horizontal arrows are removed, breaking the time dynamics, the graphical model reduces to a set of independent topic models. With time dynamics, the kth

topic at slice t has smoothly evolved from the kth topic at slice $t-1$.

For clarity of presentation, we now focus on a model with K dynamic topics evolving as in (1), and where the topic proportion model is fixed at a Dirichlet. The technical issues associated with modeling the topic proportions in a time series as in (2) are essentially the same as those for chaining the topics together.

3. Approximate Inference

Working with time series over the natural parameters enables the use of Gaussian models for the time dynamics; however, due to the nonconjugacy of the Gaussian and multinomial models, posterior inference is intractable. In this section, we present a variational method for approximate posterior inference. We use variational methods as deterministic alternatives to stochastic simulation, in order to handle the large data sets typical of text analysis. While Gibbs sampling has been effectively used for static topic models (Griffiths and Steyvers, 2004), nonconjugacy makes sampling methods more difficult for this dynamic model.

The idea behind variational methods is to optimize the free parameters of a distribution over the latent variables so that the distribution is close in Kullback-Liebler (KL) divergence to the true posterior; this distribution can then be used as a substitute for the true posterior. In the dynamic topic model, the latent variables are the topics $\beta_{t,k}$, mixture proportions $\theta_{t,d}$, and topic indicators $z_{t,d,n}$. The variational distribution reflects the group structure of the latent variables. There are variational parameters for each topic's sequence of multinomial parameters, and variational parameters for each of the document-level latent variables. The approximate variational posterior is

$$\prod_{k=1}^{K} q(\beta_{k,1},\ldots,\beta_{k,T} \mid \hat{\beta}_{k,1},\ldots,\hat{\beta}_{k,T}) \times \prod_{t=1}^{T}\left(\prod_{d=1}^{D_t} q(\theta_{t,d} \mid \gamma_{t,d}) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n} \mid \phi_{t,d,n})\right). \quad (3)$$

In the commonly used mean-field approximation, each latent variable is considered independently of the others. In the variational distribution of $\{\beta_{k,1},\ldots,\beta_{k,T}\}$, however, we retain the sequential structure of the topic by positing a dynamic model with Gaussian "variational observations" $\{\hat{\beta}_{k,1},\ldots,\hat{\beta}_{k,T}\}$. These parameters are fit to minimize the KL divergence between the resulting posterior, which is Gaussian, and the true posterior, which is not Gaussian. (A similar technique for Gaussian processes is described in Snelson and Ghahramani, 2006.)

The variational distribution of the document-level latent variables follows the same form as in Blei et al. (2003). Each proportion vector $\theta_{t,d}$ is endowed with a free Dirichlet parameter $\gamma_{t,d}$, each topic indicator $z_{t,d,n}$ is endowed with a free multinomial parameter $\phi_{t,d,n}$, and optimization proceeds by coordinate ascent. The updates for the document-level variational parameters have a closed form; we use the conjugate gradient method to optimize the topic-level variational observations. The resulting variational approximation for the natural topic parameters $\{\beta_{k,1},\ldots,\beta_{k,T}\}$ incorporates the time dynamics; we describe one approximation based on a Kalman filter, and a second based on wavelet regression.

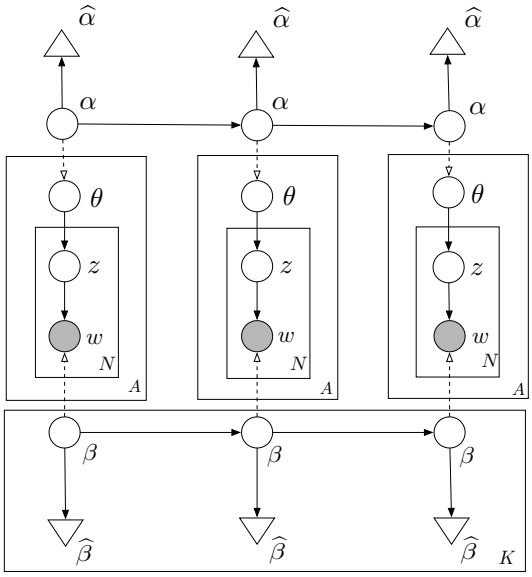

Figure 2. A graphical representation of the variational approximation for the time series topic model of Figure 1. The variational parameters $\hat{\beta}$ and $\hat{\alpha}$ are thought of as the outputs of a Kalman filter, or as observed data in a nonparametric regression setting.

3.1. Variational Kalman Filtering

The view of the variational parameters as outputs is based on the symmetry properties of the Gaussian density, $f_{\mu,\Sigma}(x) = f_{x,\Sigma}(\mu)$, which enables the use of the standard forward-backward calculations for linear state space models. The graphical model and its variational approximation are shown in Figure 2. Here the triangles denote variational parameters; they can be thought of as "hypothetical outputs" of the Kalman filter, to facilitate calculation.

To explain the main idea behind this technique in a simpler setting, consider the model where unigram models β_t (in the natural parameterization) evolve over time. In this model there are no topics and thus no mixing parameters. The calculations are simpler versions of those we need for the more general latent variable models, but exhibit the es-

sential features. Our state space model is

$$\begin{aligned} \beta_t \,|\, \beta_{t-1} &\sim \mathcal{N}(\beta_{t-1}, \sigma^2 I) \\ w_{t,n} \,|\, \beta_t &\sim \textit{Mult}(\pi(\beta_t)) \end{aligned}$$

and we form the variational state space model where

$$\hat{\beta}_t \,|\, \beta_t \sim \mathcal{N}(\beta_t, \hat{\nu}_t^2 I)$$

The variational parameters are $\hat{\beta}_t$ and $\hat{\nu}_t$. Using standard Kalman filter calculations (Kalman, 1960), the forward mean and variance of the variational posterior are given by

$$\begin{aligned} m_t &\equiv \mathbb{E}(\beta_t \,|\, \hat{\beta}_{1:t}) = \\ &\left(\frac{\hat{\nu}_t^2}{V_{t-1} + \sigma^2 + \hat{\nu}_t^2}\right) m_{t-1} + \left(1 - \frac{\hat{\nu}_t^2}{V_{t-1} + \sigma^2 + \hat{\nu}_t^2}\right) \hat{\beta}_t \end{aligned}$$

$$\begin{aligned} V_t &\equiv \mathbb{E}((\beta_t - m_t)^2 \,|\, \hat{\beta}_{1:t}) \\ &= \left(\frac{\hat{\nu}_t^2}{V_{t-1} + \sigma^2 + \hat{\nu}_t^2}\right) (V_{t-1} + \sigma^2) \end{aligned}$$

with initial conditions specified by fixed m_0 and V_0. The backward recursion then calculates the marginal mean and variance of β_t given $\hat{\beta}_{1:T}$ as

$$\begin{aligned} \widetilde{m}_{t-1} &\equiv \mathbb{E}(\beta_{t-1} \,|\, \hat{\beta}_{1:T}) = \\ &\left(\frac{\sigma^2}{V_{t-1} + \sigma^2}\right) m_{t-1} + \left(1 - \frac{\sigma^2}{V_{t-1} + \sigma^2}\right) \widetilde{m}_t \end{aligned}$$

$$\begin{aligned} \widetilde{V}_{t-1} &\equiv \mathbb{E}((\beta_{t-1} - \widetilde{m}_{t-1})^2 \,|\, \hat{\beta}_{1:T}) \\ &= V_{t-1} + \left(\frac{V_{t-1}}{V_{t-1} + \sigma^2}\right)^2 \left(\widetilde{V}_t - (V_{t-1} + \sigma^2)\right) \end{aligned}$$

with initial conditions $\widetilde{m}_T = m_T$ and $\widetilde{V}_T = V_T$. We approximate the posterior $p(\beta_{1:T} \,|\, \mathbf{w}_{1:T})$ using the state space posterior $q(\beta_{1:T} \,|\, \hat{\beta}_{1:T})$. From Jensen's inequality, the log-likelihood is bounded from below as

$$\begin{aligned} &\log p(d_{1:T}) \geq \qquad (4) \\ &\int q(\beta_{1:T} \,|\, \hat{\beta}_{1:T}) \log \left(\frac{p(\beta_{1:T})\, p(d_{1:T} \,|\, \beta_{1:T})}{q(\beta_{1:T} \,|\, \hat{\beta}_{1:T})}\right) d\beta_{1:T} \\ &= \mathbb{E}_q \log p(\beta_{1:T}) + \sum_{t=1}^{T} \mathbb{E}_q \log p(d_t \,|\, \beta_t) + H(q) \end{aligned}$$

Details of optimizing this bound are given in an appendix.

3.2. Variational Wavelet Regression

The variational Kalman filter can be replaced with variational wavelet regression; for a readable introduction standard wavelet methods, see Wasserman (2006). We rescale time so it is between 0 and 1. For 128 years of *Science* we take $n = 2^J$ and $J = 7$. To be consistent with our earlier notation, we assume that

$$\hat{\beta}_t = \widetilde{m}_t + \hat{\nu}\epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$. Our variational wavelet regression algorithm estimates $\{\hat{\beta}_t\}$, which we view as observed data, just as in the Kalman filter method, as well as the noise level $\hat{\nu}$.

For concreteness, we illustrate the technique using the Haar wavelet basis; Daubechies wavelets are used in our actual examples. The model is then

$$\hat{\beta}_t = \alpha\phi(x_t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j - 1} D_{jk}\psi_{jk}(x_t)$$

where $x_t = t/n$, $\phi(x) = 1$ for $0 \leq x \leq 1$,

$$\psi(x) = \begin{cases} -1 & \text{if } 0 \leq x \leq \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} < x \leq 1 \end{cases}$$

and $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$. Our variational estimate for the posterior mean becomes

$$\widetilde{m}_t = \hat{\alpha}\phi(x_t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j - 1} \hat{D}_{jk}\psi_{jk}(x_t).$$

where $\hat{\alpha} = n^{-1} \sum_{t=1}^{n} \hat{\beta}_t$, and $\hat{D}_{jk}$ are obtained by thresholding the coefficients

$$Z_{jk} = \frac{1}{n} \sum_{t=1}^{n} \hat{\beta}_t \psi_{jk}(x_t).$$

To estimate $\hat{\beta}_t$ we use gradient ascent, as for the Kalman filter approximation, requiring the derivatives $\partial\widetilde{m}_t / \partial\hat{\beta}_t$. If soft thresholding is used, then we have that

$$\frac{\partial \widetilde{m}_t}{\partial \hat{\beta}_s} = \frac{\partial \hat{\alpha}}{\partial \hat{\beta}_s}\, \phi(x_t) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j - 1} \frac{\partial \hat{D}_{jk}}{\partial \hat{\beta}_s}\, \psi_{jk}(x_t).$$

with $\partial\hat{\alpha}/\partial\hat{\beta}_s = n^{-1}$ and

$$\partial \hat{D}_{jk} / \partial \hat{\beta}_s = \begin{cases} \frac{1}{n}\psi_{jk}(x_s) & \text{if } |Z_{jk}| > \lambda \\ 0 & \text{otherwise.} \end{cases}$$

Note also that $|Z_{jk}| > \lambda$ if and only if $|\hat{D}_{jk}| > 0$. These derivatives can be computed using off-the-shelf software for the wavelet transform in any of the standard wavelet bases.

Sample results of running this and the Kalman variational algorithm to approximate a unigram model are given in Figure 3. Both variational approximations smooth out the

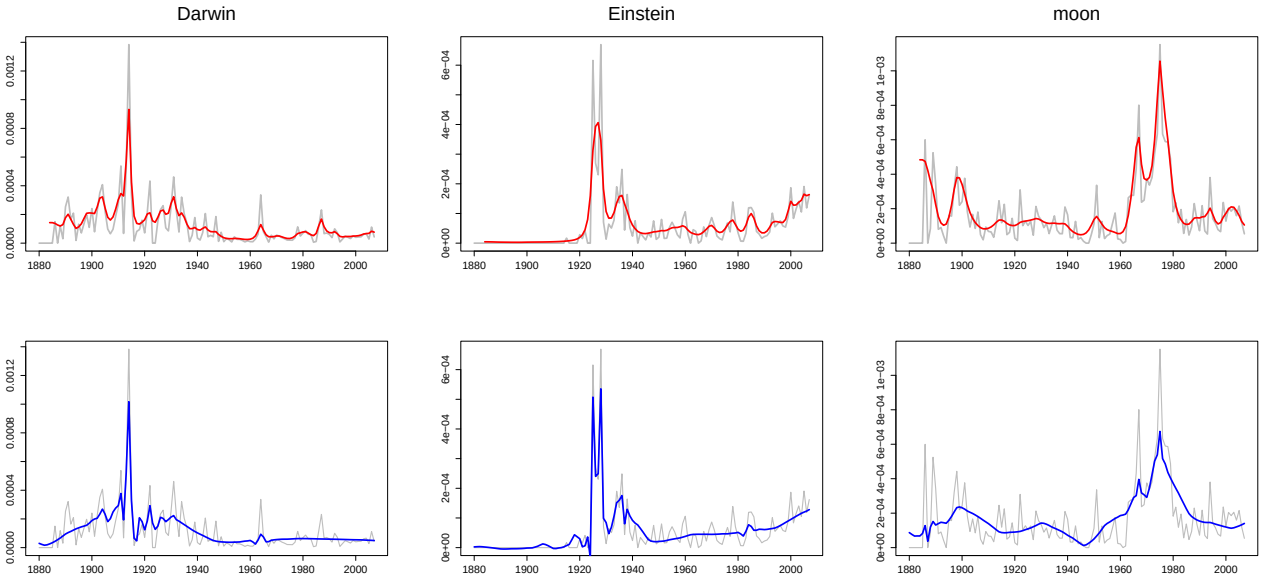

Figure 3. Comparison of the Kalman filter (top) and wavelet regression (bottom) variational approximations to a unigram model. The variational approximations (red and blue curves) smooth out the local fluctuations in the unigram counts (gray curves) of the words shown, while preserving the sharp peaks that may indicate a significant change of content in the journal. The wavelet regression is able to "superresolve" the double spikes in the occurrence of Einstein in the 1920s. (The spike in the occurrence of Darwin near 1910 may be associated with the centennial of Darwin's birth in 1809.)

local fluctuations in the unigram counts, while preserving the sharp peaks that may indicate a significant change of content in the journal. While the fit is similar to that obtained using standard wavelet regression to the (normalized) counts, the estimates are obtained by minimizing the KL divergence as in standard variational approximations.

In the dynamic topic model of Section 2, the algorithms are essentially the same as those described above. However, rather than fitting the observations from true observed counts, we fit them from expected counts under the document-level variational distributions in (3).

4. Analysis of *Science*

We analyzed a subset of 30,000 articles from *Science*, 250 from each of the 120 years between 1881 and 1999. Our data were collected by JSTOR (`www.jstor.org`), a not-for-profit organization that maintains an online scholarly 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.

Our corpus is made up of approximately 7.5 million words. We pruned the vocabulary by stemming each term to its root, removing function terms, and removing terms that occurred fewer than 25 times. The total vocabulary size is 15,955. To explore the corpus and its themes, we estimated a 20-component dynamic topic model. Posterior inference took approximately 4 hours on a 1.5GHZ PowerPC Macintosh laptop. Two of the resulting topics are illustrated in Figure 4, showing the top several words from those topics in each decade, according to the posterior mean number of occurrences as estimated using the Kalman filter variational approximation. Also shown are example articles which exhibit those topics through the decades. As illustrated, the model captures different scientific themes, and can be used to inspect trends of word usage within them.

To validate the dynamic topic model quantitatively, we consider the task of predicting the next year of *Science* given all the articles from the previous years. We compare the predictive power of three 20-topic models: the dynamic topic model estimated from all of the previous years, a static topic model estimated from all of the previous years, and a static topic model estimated from the single previous year. All the models are estimated to the same convergence criterion. The topic model estimated from all the previous data and dynamic topic model are initialized at the same point.

The dynamic topic model performs well; it always assigns higher likelihood to the next year's articles than the other two models (Figure 5). It is interesting that the predictive power of each of the models declines over the years. We can tentatively attribute this to an increase in the rate of specialization in scientific language.

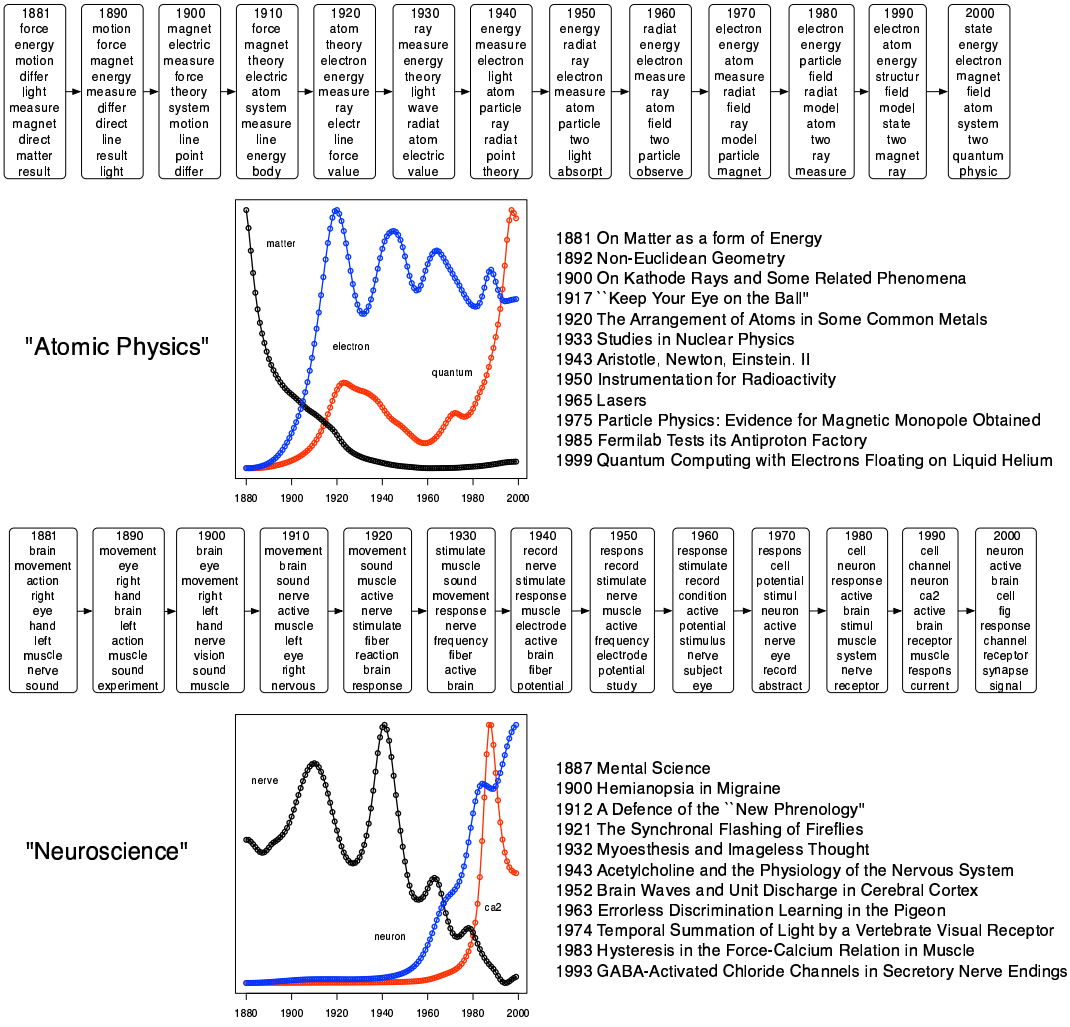

Figure 4. Examples from the posterior analysis of a 20-topic dynamic model estimated from the *Science* corpus. For two topics, we illustrate: (a) the top ten words from the inferred posterior distribution at ten year lags (b) the posterior estimate of the frequency as a function of year of several words from the same two topics (c) example articles throughout the collection which exhibit these topics. Note that the plots are scaled to give an idea of the shape of the trajectory of the words' posterior probability (i.e., comparisons across words are not meaningful).

5. Discussion

We have developed sequential topic models for discrete data by using Gaussian time series on the natural parameters of the multinomial topics and logistic normal topic proportion models. We derived variational inference algorithms that exploit existing techniques for sequential data; we demonstrated a novel use of Kalman filters and wavelet regression as variational approximations. Dynamic topic models can give a more accurate predictive model, and also offer new ways of browsing large, unstructured document collections.

There are many ways that the work described here can be extended. One direction is to use more sophisticated state space models. We have demonstrated the use of a simple Gaussian model, but it would be natural to include a drift term in a more sophisticated autoregressive model to explicitly capture the rise and fall in popularity of a topic, or in the use of specific terms. Another variant would allow for heteroscedastic time series.

Perhaps the most promising extension to the methods presented here is to incorporate a model of how new topics in the collection appear or disappear over time, rather than assuming a fixed number of topics. One possibility is to use a simple Galton-Watson or birth-death process for the topic population. While the analysis of birth-death or branching processes often centers on extinction probabilities, here a goal would be to find documents that may be responsible for spawning new themes in a collection.

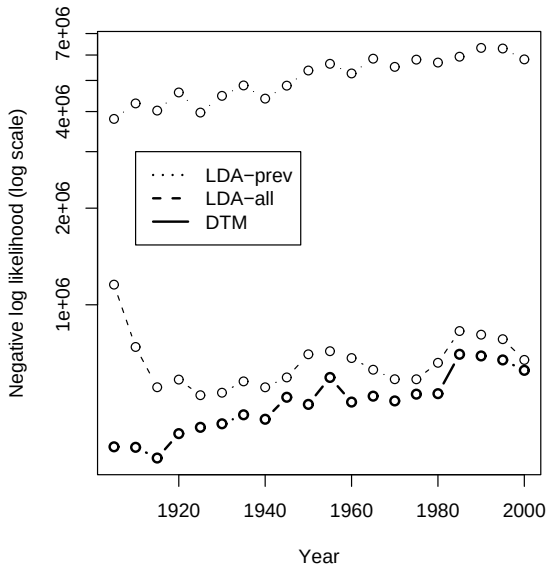

Figure 5. This figure illustrates the performance of using dynamic topic models and static topic models for prediction. For each year between 1900 and 2000 (at 5 year increments), we estimated three models on the articles through that year. We then computed the variational bound on the negative log likelihood of next year's articles under the resulting model (lower numbers are better). DTM is the dynamic topic model; LDA-prev is a static topic model estimated on just the previous year's articles; LDA-all is a static topic model estimated on all the previous articles.

Acknowledgments

This research was supported in part by NSF grants IIS-0312814 and IIS-0427206, the DARPA CALO project, and a grant from Google.

References

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

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

Blei, D. M. and Lafferty, J. D. (2006). Correlated topic models. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA.

Buntine, W. and Jakulin, A. (2004). Applying discrete PCA in data analysis. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 59–66. AUAI Press.

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

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. *IEEE Computer Vision and Pattern Recognition*.

Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science*, 101:5228–5235.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transaction of the AMSE: Journal of Basic Engineering*, 82:35–45.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2004). The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and academic email. Technical report, University of Massachusetts, Amherst.

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

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smith, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494. AUAI Press.

Sivic, J., Rusell, B., Efros, A., Zisserman, A., and Freeman, W. (2005). Discovering objects and their location in images. In *International Conference on Computer Vision (ICCV 2005)*.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, Cambridge, MA. MIT Press.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.

A. Derivation of Variational Algorithm

In this appendix we give some details of the variational algorithm outlined in Section 3.1, which calculates a distribution $q(\beta_{1:T} \mid \hat{\beta}_{1:T})$ to maximize the lower bound on

$\log p(d_{1:T})$. The first term of the righthand side of (5) is

$$\begin{aligned}\sum_{t=1}^{T} \mathbb{E}_q \log p(\beta_t \mid \beta_{t-1}) &= -\frac{VT}{2}\left(\log \sigma^2 + \log 2\pi\right) \\ &\quad -\frac{1}{2\sigma^2}\sum_{t=1}^{T} \mathbb{E}_q (\beta_t - \beta_{t-1})^T (\beta_t - \beta_{t-1}) \\ &= -\frac{VT}{2}\left(\log \sigma^2 + \log 2\pi\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{T} \|\widetilde{m}_t - \widetilde{m}_{t-1}\|^2 \\ &\quad -\frac{1}{\sigma^2}\sum_{t=1}^{T} \mathrm{Tr}\left(\widetilde{V}_t\right) + \frac{1}{2\sigma^2}\left(\mathrm{Tr}\left(\widetilde{V}_0\right) - \mathrm{Tr}\left(\widetilde{V}_T\right)\right)\end{aligned}$$

using the Gaussian quadratic form identity

$$\begin{aligned}&\mathbb{E}_{m,V}(x-\mu)^T\Sigma^{-1}(x-\mu) = \\ &\quad (m-\mu)^T\Sigma^{-1}(m-\mu) + \mathrm{Tr}\left(\Sigma^{-1}V\right).\end{aligned}$$

The second term of (5) is

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}_q \log p(d_t \mid \beta_t) = \\ &\quad \sum_{t=1}^{T}\sum_{w} n_{tw}\, \mathbb{E}_q\left(\beta_{tw} - \log \sum_{w} \exp(\beta_{tw})\right) \\ &\geq \sum_{t=1}^{T}\sum_{w} n_{tw}\, \widetilde{m}_{tw} - n_t\, \hat{\zeta}_t^{-1} \sum_{w} \exp(\widetilde{m}_{tw} + \widetilde{V}_{tw}/2) \\ &\quad + \sum_{t=1}^{T} n_t - n_t \log \hat{\zeta}_t\end{aligned}$$

where $n_t = \sum_w n_{tw}$, introducing additional variational parameters $\hat{\zeta}_{1:T}$. The third term of (5) is the entropy

$$\begin{aligned}H(q) &= \sum_{t=1}^{T}\left(\frac{1}{2}\log|\widetilde{V}_t| + \frac{T}{2}\log 2\pi\right) \\ &= \frac{1}{2}\sum_{t=1}^{T}\sum_{w}\log \widetilde{V}_{tw} + \frac{TV}{2}\log 2\pi.\end{aligned}$$

To maximize the lower bound as a function of the variational parameters we use a conjugate gradient algorithm. First, we maximize with respect to $\hat{\zeta}$; the derivative is

$$\frac{\partial \ell}{\partial \hat{\zeta}_t} = \frac{n_t}{\hat{\zeta}_t^2}\sum_{w}\exp(\widetilde{m}_{tw} + \widetilde{V}_{tw}/2) - \frac{n_t}{\hat{\zeta}_t}.$$

Setting to zero and solving for $\hat{\zeta}_t$ gives

$$\hat{\zeta}_t = \sum_{w}\exp(\widetilde{m}_{tw} + \widetilde{V}_{tw}/2).$$

Next, we maximize with respect to $\hat{\beta}_s$:

$$\begin{aligned}&\frac{\partial \ell(\hat{\beta}, \hat{\nu})}{\partial \hat{\beta}_{sw}} = \\ &\quad -\frac{1}{\sigma^2}\sum_{t=1}^{T}(\widetilde{m}_{tw} - \widetilde{m}_{t-1,w})\left(\frac{\partial \widetilde{m}_{tw}}{\partial \hat{\beta}_{sw}} - \frac{\partial \widetilde{m}_{t-1,w}}{\partial \hat{\beta}_{sw}}\right) \\ &\quad + \sum_{t=1}^{T}\left(n_{tw} - n_t\hat{\zeta}_t^{-1}\exp(\widetilde{m}_{tw} + \widetilde{V}_{tw}/2)\right)\frac{\partial \widetilde{m}_{tw}}{\partial \hat{\beta}_{sw}}.\end{aligned}$$

The forward-backward equations for $\widetilde{m}_t$ can be used to derive a recurrence for $\partial \widetilde{m}_t / \partial \hat{\beta}_s$. The forward recurrence is

$$\begin{aligned}\frac{\partial m_t}{\partial \hat{\beta}_s} &= \left(\frac{\hat{\nu}_t^2}{v_{t-1} + \sigma^2 + \hat{\nu}_t^2}\right)\frac{\partial m_{t-1}}{\partial \hat{\beta}_s} + \\ &\quad \left(1 - \frac{\hat{\nu}_t^2}{v_{t-1} + \sigma^2 + \hat{\nu}_t^2}\right)\delta_{s,t}\,,\end{aligned}$$

with the initial condition $\partial m_0 / \partial \hat{\beta}_s = 0$. The backward recurrence is then

$$\begin{aligned}\frac{\partial \widetilde{m}_{t-1}}{\partial \hat{\beta}_s} &= \left(\frac{\sigma^2}{V_{t-1} + \sigma^2}\right)\frac{\partial m_{t-1}}{\partial \hat{\beta}_s} + \\ &\quad \left(1 - \frac{\sigma^2}{V_{t-1} + \sigma^2}\right)\frac{\partial \widetilde{m}_t}{\partial \hat{\beta}_s},\end{aligned}$$

with the initial condition $\partial \widetilde{m}_T / \partial \hat{\beta}_s = \partial m_T / \partial \hat{\beta}_s$.