Hierarchical Bayesian Modeling of Topics in Time-Stamped Documents

Iulian Pruteanu-Malinici, Lu Ren, John Paisley, Eric Wang and Lawrence Carin

Department of Electrical and Computer Engineering

Duke University

Durham, NC, USA

{ip6,lr,jwp4,ew28,lcarin}@ece.duke.edu

Abstract

We consider the problem of inferring and modeling topics in a sequence of documents with known publication dates. The documents at a given time are each characterized by a topic, and the topics are drawn from a mixture model. The proposed model infers the change in the topic mixture weights as a function of time. The details of this general framework may take different forms, depending on the specifics of the model. For the examples considered here we examine base measures based on independent multinomial-Dirichlet measures for representation of topic-dependent word counts. The form of the hierarchical model allows efficient variational Bayesian (VB) inference, of interest for large-scale problems. We demonstrate results and make comparisons to the model when the dynamic character is removed, and also compare to latent Dirichlet allocation (LDA) and topics over time (TOT). We consider a database of NIPS papers as well as the United States presidential State of the Union addresses from 1790 to 2008.

Index Terms

Hierarchical models, Variational Bayes, Dirichlet Process, Text modeling.

I. Introduction

Topic models attempt to infer sets of words from text data that together form meaningful contextual and semantic relationships. Finding these groups of words, known as topics, allows effective clustering,

searching, sorting, and archiving of a corpus of documents. If we assume the bag-of-words structure, *i.e.*, that words are exchangeable and independent, then there are in general two ways to consider a collection of documents. Factor models such as probabilistic Latent Semantic Indexing (pLSI) [11], Latent Dirichlet Allocation (LDA) [6] and Topics over Time (TOT) [20] assume that each word in a given document is drawn from a mixture model whose components are topics. Other models assume that words in a sentence or even in an overall document are drawn simultaneously from one topic [10], [23]. In [10], the authors propose modeling topics of words as a Markov chain, with successive sentences modeled as being likely to share the same topic. Since topics are hidden, learning and inferring the model are done using tools from hidden Markov models. Whether one draws a topic for every word or considers all words within a sentence/document as being generated by a common topic, documents are represented as counts over the dictionary, and topics are represented as multinomial distributions over the dictionary. This approach to topic representation is convenient, as the Dirichlet distribution is the conjugate prior to the multinomial. However, because the distribution over the dictionary must be normalized, problems can occur if a previously unknown word is encountered, as can often happen when using a trained model on an unknown testing set.

A new factor model has been proposed [7] that represents each integer word count from the term-document matrix as a sample from an independent poisson distribution. This model, called GaP for gamma-poisson, factorizes the sparse term-document matrix into the product of an expected-counts matrix and a theme probability matrix. Note that the GaP model is equivalent to placing a multinomial-Dirichlet implementation over the dictionary, so that one can model both the relative word frequencies and the overall word count. One may use the poisson-gamma characterization as a starting point to building a dynamic topic model by using a closely-related approach to [14]. Using an independent distribution for each word is attractive, as it addresses the problem of adding unknown words to the dictionary. Further, since each word is allowed to evolve independently, this approach leads to a more flexible model than using a traditional multinomial-Dirichlet structure. We build upon this construct in the model presented here.

The main focus of this paper is on development of a hierarchical Bayesian model for characterizing documents with known time stamp. Each document is assumed to have an associated topic, and all documents at a given time are assumed to have topics that are drawn from a mixture model; the mixture weights in this model evolve with time. This framework imposes the idea that documents that appear at similar times are likely to be drawn from similar mixtures of topics. To achieve this goal,

we develop a simplified form of the dynamic hierarchical Dirichlet process (dHDP) [15]. Inference is performed efficiently via a variational Bayesian analysis [2].

Our model differs from other time-evolving topic models [17], [4] in that our topics do not evolve over time; what changes in time are the mixing weights over topics, while the overall set of topics are kept unchanged. Specific topics are typically localized over a period of time, with new dominant topics spawned after other topics diminish in importance (the temporally localized topics may alternatively be viewed as a time evolution of a single topic [4], but such single-topic evolution is not considered here).

The reminder of the paper is organized as follows. We review semi-parametric hierarchical models in Section II. In Section III we introduce the proposed model, and make connections with related models. Sharing properties for the proposed model are addressed in Section IV, and Section V details VB inference. We demonstrate model performance in Section VI, and conclude in Section VII.

II. Review of Semi-Parametric Statistical Modeling

The Dirichlet process (DP) is a semi-parametric measure for development of general mixture models (in principle, in terms of an infinite number of mixture components). Let H be a measure and α is a non-negative real number. A draw from a Dirichlet process parameterized by α and H is denoted $G{\sim}DP(\alpha, H)$. Sethuraman [16] introduced the stick-breaking representation of a DP draw:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*},$$
$$\pi_k = V_k \prod_{l=1}^{k-1}(1-V_l), \qquad V_k \overset{i.i.d.}{\sim} Beta(1,\alpha), \qquad \theta_k^* \overset{i.i.d.}{\sim} H, \tag{1}$$

where $\delta_{\theta_k^*}$ is a point measure concentrated at θ_k^* (each θ_k^* is termed an atom), and $Beta(1,\alpha)$ is a beta distribution with shape parameter α. Note that G is almost surely discrete, with this playing a key role in the utility of DP for clustering. To simplify notation below, an infinite probability vector $\boldsymbol{\pi}$ constructed as above is denote $\boldsymbol{\pi} \sim Stick(\alpha)$.

Suppose the data of interest are divided into different sets, and each data set is termed a "task" for analysis. For clustering of T tasks the DP imposes the belief that when two tasks are associated with the same cluster, all data within the tasks are shared. This may be too restrictive in some applications and has motivated the hierarchical Dirichlet process (HDP) [18]. We denote the data in task t as

$\{\boldsymbol{x}_{t,i}\}_{i=1}^{N_t}$, where N_t is the number of data in the task. The HDP may be represented as

$$\begin{aligned} \boldsymbol{x}_{t,i} &\sim f(\theta_{t,i}); \quad i = 1, 2, ..., N_t; \quad t = 1, 2, ..., T, \\ \theta_{t,i} &\sim G_t; \quad i = 1, 2, ..., N_t; \quad t = 1, 2, ..., T, \\ G_t &\sim DP(\alpha, G); \quad t = 1, 2, ..., T, \\ G &\sim DP(\gamma, H), \end{aligned} \tag{2}$$

where $f(\theta)$ represents the specific parametric model under consideration. Because the task-dependent DPs share the same (discrete) base G, all $\{G_t\}_{t=1}^{T}$ share the same set of mixture atoms, with different mixture weights. The measures $\{G_t\}_{t=1,T}$ are *jointly* drawn from an HDP:

$$\{G_1, \cdots, G_T\} \sim HDP(\alpha, \gamma, H). \tag{3}$$

The HDP assumes the T tasks are exchangeable; however, there are many applications for which it is desirable to remove this exchangeability assumption. Models such as the kernel stick breaking process [9], [1], the generalized product partition model [13], the correlated topic model [5] and the dynamic DP [8] are techniques that impose structure on the dependence of the tasks (removing exchangeability). Some of these models rely on modifying the mixing weights to impose dependence on location [9], [1] or covariate [13], while others impose sequential time dependence on the structure of consecutive tasks (see [8]).

We again consider T tasks, but now index t explicitly denotes the sequential time of data production/collection. To address the sequential nature of the time blocks, [15] imposes a dynamic HDP (dHDP)

$$G_t = w_t D_t + (1 - w_t) G_{t-1}; \quad t = 2, ..., T, \tag{4}$$

where $\{G_1, D_2, \cdots, D_T\} \sim HDP(\alpha, \gamma, H)$. The parameter $w_t \in [0, 1]$ is drawn from a beta distribution $Beta(a_0, b_0)$, and it controls the degree of innovation in G_t relative to G_{t-1}. The DP and HDP are limiting cases of this model:

- when $w_t \to 0$, $G_t \to G_{t-1}$ and there is no innovation, resulting in a common set of mixture weights for all time blocks (DP);
- when $w_t \to 1$, $G_t \to D_t$, where the new innovation distribution D_t controls the sharing mechanism, resulting in each time block having a unique set of mixing weights (HDP).

It is important to restate that dHDP does not assume the mixture components evolve over time, only the mixing weights. The mixture components are shared explicitly across all time blocks. This is fundamentally different from other models that impose temporal dependence through component evolution [4], [17], this allowing a unique and independent set of mixing weights for each block.

III. Semi-Parametric Dynamic Topic Model

A. Model construction

Consider a collection of documents with known time stamps, with time evolving from $t = 1, \cdots, T$. At any particular time we have N_t such independent documents. The total set of documents over all time is represented as $\{\boldsymbol{x}_{t,i}\}_{i=1}^{N_t}{}_{t=1}^{T}$, where $\boldsymbol{x}_{t,i}$ represents a vector of word counts associated with document i at time t. In the form of the model presented here, we are only interested in the number of times a given word is present in a particular document; the set of J unique words in the collection forms a dictionary. Each document is assumed characterized by a single topic, and at time t the topics across all documents are assumed drawn from a mixture model. In the proposed model the mixture weights on the topics are assumed to evolve with time (analogous to as implemented in the dHDP [15] discussed above). The assumption that each document is characterized by a single topic may seem restrictive; however, we observe in Section VI that for our motivating example this assumption is reasonable.

To constitute a model with a time-evolving mixture of topics, we seek a simplified representation of the dHDP. Specifically, the proposed topic model, termed dDTM for dynamic Dirichlet topic model, is represented as

$$\begin{aligned}
x_{t,i}^{j}|z_{t,i} &\sim F(\boldsymbol{\theta}_{z_{t,i}}^{j}); \quad j = 1, ..., J, \\
z_{t,i}|\boldsymbol{\tau}_t &\sim Mult(\boldsymbol{\tau}_t); \quad t = 2, ..., T, \\
z_{1,i}|\boldsymbol{\pi}_1 &\sim Mult(\boldsymbol{\pi}_1), \\
\boldsymbol{\tau}_t &= (1 - w_t)\boldsymbol{\tau}_{t-1} + w_t\boldsymbol{\pi}_t; \quad t = 2, ..., T, \\
\boldsymbol{\pi}_t &\sim Dir(\boldsymbol{\alpha}); \quad t = 1, ..., T, \\
w_t &\sim Beta(c_0, d_0); \quad t = 2, ..., T, \\
\boldsymbol{\theta}_k^j &\sim H^j; \quad j = 1, ..., J; \quad k = 1, ..., K,
\end{aligned} \tag{5}$$

Note that $\boldsymbol{\tau}_1 = \boldsymbol{\pi}_1$. The factorized structure $\boldsymbol{H} = \prod_{j=1}^{J} H^j$ is similar to [7], which allows insertion

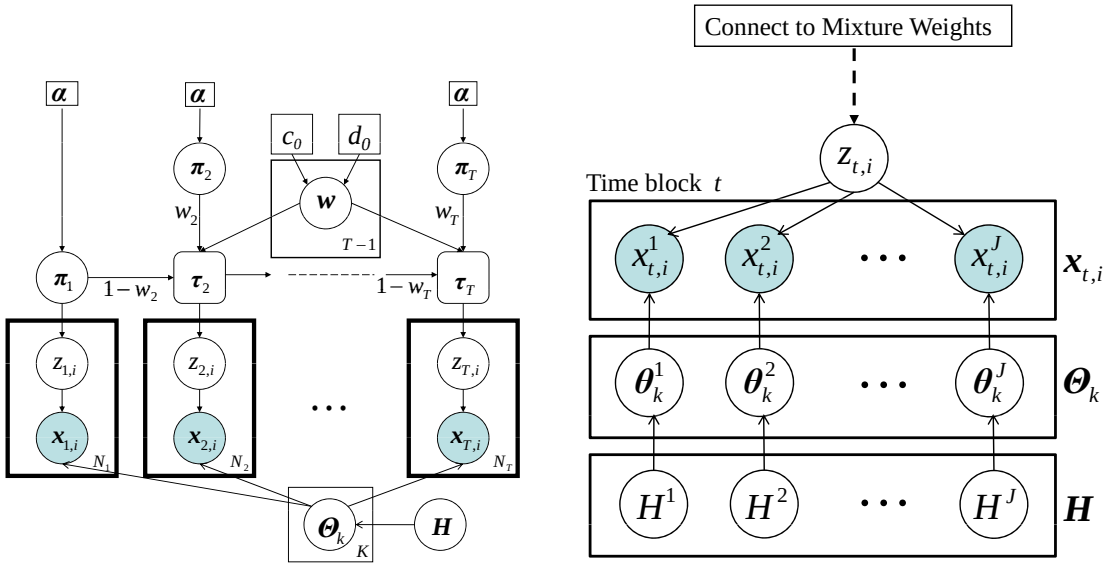


Fig. 1. Dynamic Dirichlet topic model (dDTM). (a) graphical representation of the model, (b) expanded representation of the product measure aspect of the model.

of new words with time.

Although perhaps not apparent at this point, for large K the proposed model is closely related to dHDP; this is analyzed in detail below. The model is represented graphically in Fig. 1(a), and in Fig. 1(b) we illustrate how a single mixture component is drawn, with the parametric model of each dimension drawn independently from its respective prior.

The form of the parametric model $F(\cdot)$ in (5) may vary depending on the application; in the work presented here it corresponds to a multinomial-Dirichlet model. We consider the number of times a word is present in a given document; to do this, $F(\cdot)$ is defined as a multinomial distribution and consequently, to preserve the conjugacy requirements, each H^j is a Dirichlet distribution.

B. Relationship to dHDP

We now make explicit the relationship between dHDP [15] and dDTM represented in (5). Recall that the draws $\{G_1, D_2, \cdots, D_T\} \sim HDP(\alpha, \gamma, H)$ may be constructed as [18]

$$\begin{aligned} G_1 &= \sum_{k=1}^{\infty} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k} \\ D_t &= \sum_{k=1}^{\infty} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k} \quad , \quad t = 2, \cdots, T \\ \boldsymbol{\pi}_t &\sim DP(\alpha, \boldsymbol{v}) \quad , \quad t = 1, \cdots, T \\ \boldsymbol{v} &\sim Stick(\gamma) \\ \boldsymbol{\Theta}_k &\sim \boldsymbol{H} \quad , \quad k = 1, \cdots, \infty \end{aligned} \tag{6}$$

The draw $\boldsymbol{\pi}_t \sim DP(\alpha, \boldsymbol{v})$ may be represented in stick-breaking form, with the kth component of $\boldsymbol{\pi}_t$ constructed as $\pi_{t,k} = \sum_{j=1}^{\infty} w_{t,j} \delta(Y_{t,j} = k)$, with $\boldsymbol{w}_t \sim Stick(\alpha)$, $Y_{t,j} \sim Mult(\boldsymbol{v})$; $\delta(Y_{t,j} = k)$ equals one if $Y_{t,j} = k$, and its zero otherwise. We may also truncate the draw $\boldsymbol{v} \sim Stick(\alpha)$ to K sticks (denoted $\boldsymbol{v}_K \sim Stick_K(\alpha)$), for large K [12]. Using these representations, the overall HDP construction, when truncated to K topics (atoms), may be represented as

$$\begin{aligned} G_1 &= \sum_{k=1}^{K} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k} \\ D_t &= \sum_{k=1}^{K} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k} \quad , \quad t = 2, \cdots, T \\ \pi_{t,k} &= \sum_{j=1}^{\infty} w_{t,j} \delta(Y_{t,j} = k) \quad ; \quad k = 1, \cdots, K; \quad t = 1, \cdots, T \\ \boldsymbol{w}_t &\sim Stick(\alpha) \quad , \quad t = 1, \cdots, T \\ Y_{t,j} &\sim Mult(\boldsymbol{v}_K) \quad ; \quad j = 1, \cdots, \infty \quad ; \quad t = 1, \cdots, T \\ \boldsymbol{v}_K &\sim Stick_K(\gamma) \\ \boldsymbol{\Theta}_k &\sim \boldsymbol{H} \quad , \quad k = 1, \cdots, K \end{aligned} \tag{7}$$

Note that we truncate $Stick(\gamma)$ to K sticks, but do *not* truncate $Stick(\alpha)$. Additionally, $Y_{t,j} \in \{1, \cdots, K\}$, with the particular value of $Y_{t,j}$ depending on which component is selected from the multinomial.

To appreciate the relationship between dHDP and the proposed dDTM, note that (5) corresponds

to drawing atoms/topics at time t from the finite mixture model $G_t = w_t D_t + (1 - w_t)G_{t-1}$, with

$$\begin{aligned} G_1 &= \sum_{k=1}^{K} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k} \\ D_t &= \sum_{k=1}^{K} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k} \quad , \quad t = 2, \cdots, T \\ \boldsymbol{\pi}_t &\sim Dir(\alpha/K, \cdots, \alpha/K) \quad , t = 1, \cdots, T \\ \boldsymbol{\Theta}_k &\sim \boldsymbol{H} \quad , \quad k = 1, \cdots, K \end{aligned} \tag{8}$$

Recall that Sethuraman demonstrated [16] that a draw $\boldsymbol{\pi} \sim Dir(\alpha \boldsymbol{g}_0)$, where $\boldsymbol{g}_0$ is a K-dimensional probability vector and $\alpha > 0$, may be constructed as

$$\begin{aligned} \pi_k &= \sum_{j=1}^{\infty} w_j \delta(Y_j = k) \quad , \quad k = 1, \cdots, K \\ \boldsymbol{w} &\sim Stick(\alpha) \\ Y_j &\sim Mult(\boldsymbol{g}_0) \quad , \quad j = 1, \cdots, \infty \end{aligned} \tag{9}$$

with π_k representing the kth component of $\boldsymbol{\pi}$. Using Sethuraman's stick-breaking representation of the Dirichlet distribution in (8), the proposed dDTM is constructed as

$$\begin{aligned} G_1 &= \sum_{k=1}^{K} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k} \\ D_t &= \sum_{k=1}^{K} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k} \quad , \quad t = 2, \cdots, T \\ \pi_{t,k} &= \sum_{j=1}^{\infty} w_j \delta(Y_{t,j} = k) \quad ; \quad k = 1, \cdots, K; \quad t = 1, \cdots, T \\ \boldsymbol{w}_t &\sim Stick(\alpha) \quad , t = 1, \cdots, T \\ Y_{t,j} &\sim Mult(1/K, \cdots, 1/K) \quad ; \quad j = 1, \cdots, \infty; \quad t = 1, \cdots, T \\ \boldsymbol{\Theta}_k &\sim \boldsymbol{H} \quad , \quad k = 1, \cdots, K \end{aligned} \tag{10}$$

The truncated dHDP model in (4) draws $\{G_1, D_2, \cdots, D_T\}$ from (7), assuming $Stick(\gamma)$ is truncated to K sticks [12]. By contrast, within dDTM the measures $\{G_1, D_2, \cdots, D_T\}$ are drawn from (10). In the former the random variables $Y_{t,j}$ are drawn from $\boldsymbol{v}_K$, which is in turn drawn from the truncated stick-breaking process $Stick_K(\gamma)$; in the latter we simply set $\boldsymbol{v}_K = (1/K, \cdots, 1/K)$ and remove the parameter γ altogether. It is felt that this relatively small change does not significantly

affect the expressibility of the proposed prior. Within the proposed model the weights w_t explicitly impose temporal relationships between the topics (documents at proximate times are more likely to share the same topics).

The above discussion also demonstrates that considering the Dirichlet distribution $Dir(\alpha/K, \cdots, \alpha/K)$ with large K is analogous (but distinct from) a truncated stick-breaking representation. In this sense, the proposed model is non-parametric, in that setting a large K allows the model to infer the proper number of topics from the data, analogous to studies of the truncated stick-breaking representation [12]. Setting a large K ($e.g.$, $K = 50$ in the examples below), does not imply that we believe that there are actually K topics, since from (9) only a relatively small set of components in $\boldsymbol{\pi}_t$ will have appreciable amplitude (the same type of motivation for the stick-breaking view of DP and HDP). As in other non-parametric methods, the proposed model infers a distribution on the proper number of topics, based on the data.

We also emphasize that the stick-breaking representation of a draw from a Dirichlet distribution has been introduced above to make the connection between the proposed model and a truncated representation of dHDP. However, when actually performing inference, it is often simpler to just draw directly from $Dir(\alpha/K, \cdots, \alpha/K)$. However, this issue is revisited in the Conclusions.

C. Limiting cases

In Section II we considered dHDP under limiting cases of w_t, and we do so here for the proposed dDTM in (5). In the limit $w_t \rightarrow 0$, the dDTM parameters are drawn at all time from the same measure $G_1 = \sum_{k=1}^{K} \pi_{1,k} \delta_{\boldsymbol{\Theta}_k}$ with $\boldsymbol{\pi}_1 \sim Dir(\alpha/K, \cdots, \alpha/K)$ and $\boldsymbol{\Theta}_k \sim \boldsymbol{H}$. Therefore, in the limit $K \rightarrow \infty$ and $w_t \rightarrow 0$ the topic-model parameters for dDTM are drawn from $DP(\alpha, \boldsymbol{H})$, as is the case for dHDP when $w_t \rightarrow 0$. Since K is finite in dDTM, the limit $w_t \rightarrow 0$ yields a model similar to LDA [6] (in LDA one performs a point estimate for α, while here α is set).

In the limit $w_t \rightarrow 1$, at time t the dDTM model parameters are drawn from $G_t = \sum_{k=1}^{K} \pi_{t,k} \delta_{\boldsymbol{\Theta}_k}$, again with $\boldsymbol{\Theta}_k \sim \boldsymbol{H}$, and with each $\boldsymbol{\pi}_t \overset{iid}{\sim} Dir(\alpha/K, \cdots, \alpha/K)$. Thus the $\{G_t\}_{t=1,T}$ all share the same atoms (topics), with distinct t-dependent probability weights $\boldsymbol{\pi}_t$. The dHDP model has a similar limit when $w_t \rightarrow 1$, with the weights drawn $\boldsymbol{\pi}_t \overset{iid}{\sim} DP(\alpha, \boldsymbol{v})$ for $\boldsymbol{v} \sim Stick(\gamma)$. In both cases the atoms/topics are shared across all time, with different mixture weights. The dHDP arguably allows for more modeling flexibility, through the parameter γ, while dDTM yields a simpler model with very similar structural form.

IV. Model Properties

To examine properties of the model in (5), we consider the discrete indicator's space $I = \{1, 2, ..., K\}$ with $k \in I$ indicating one of the K mixing components of the model. Therefore, we can write

$$\begin{aligned} \boldsymbol{\tau}_t(I)|\boldsymbol{\tau}_{t-1}, w_t &= (1-w_t)\boldsymbol{\tau}_{t-1}(I) + w_t\boldsymbol{\pi}_t(I) \\ &= \boldsymbol{\tau}_{t-1}(I) + \Delta_t(I), \end{aligned} \tag{11}$$

where $\Delta_t(I) = w_t(\boldsymbol{\pi}_t(I) - \boldsymbol{\tau}_{t-1}(I))$ is the random deviation from $\boldsymbol{\tau}_{t-1}(I)$ to $\boldsymbol{\tau}_t(I)$.

Theorem 1. The mean and the variance of the random deviation Δ_t are controlled by the innovating weight w_t and model parameter $\boldsymbol{\alpha} = [\alpha/K, ..., \alpha/K]$:

$$\begin{aligned} E\{\Delta_t(I)|\boldsymbol{\tau}_{t-1}, w_t, \boldsymbol{\alpha}\} &= w_t(E(\boldsymbol{\pi}_t(I)) - \boldsymbol{\tau}_{t-1}(I)) \\ &= w_t([\frac{1}{K}, ..., \frac{1}{K}] - \boldsymbol{\tau}_{t-1}(I)), \end{aligned} \tag{12}$$

$$V\{\Delta_t(I)|\boldsymbol{\tau}_{t-1}, w_t, \boldsymbol{\alpha}\} = \frac{w_t^2}{K}[\frac{1}{\alpha+1}(1-\frac{1}{K}), ..., \frac{1}{\alpha+1}(1-\frac{1}{K})], \tag{13}$$

where we observe two limiting cases:

- when $w_t \to 0$, $\boldsymbol{\tau}_t = \boldsymbol{\tau}_{t-1}$.
- when $\boldsymbol{\tau}_{t-1} \to [\frac{1}{K}, ..., \frac{1}{K}]$, $E\{\boldsymbol{\tau}_t(I)|\boldsymbol{\tau}_{t-1}, w_t\} = \boldsymbol{\tau}_{t-1}(I)$.

Theorem 2. The correlation coefficient between two adjacent distributions $\boldsymbol{\tau}_{t-1}$ and $\boldsymbol{\tau}_t$ for $t = 2, ..., T$ is

$$\begin{aligned} Corr(\tau_{t-1,k}, \tau_{t,k}) &= \frac{E\{\tau_{t-1,k}\tau_{t,k}\} - E\{\tau_{t-1,k}\}E\{\tau_{t,k}\}}{\sqrt{V\{\tau_{t-1,k}\}V\{\tau_{t,k}\}}} \\ &= (1-w_t)\sqrt{\frac{\sum_{l=1}^{t-1} w_l^2 \prod_{q=l+1}^{t-1}(1-w_q)^2}{\sum_{l=1}^{t} w_l^2 \prod_{q=l+1}^{t}(1-w_q)^2}}, \end{aligned} \tag{14}$$

for any $k \in I$. The proofs of these theorems are provided in Appendix A.

To compare the similarity of two adjacent tasks/documents, the two theorems yield insights through the mean and variance of the random deviation and the correlation coefficient which can be estimated from (14), using the posterior expectation of $\boldsymbol{w}$. Although dDTM represents a simplification of the dHDP framework [15], the sharing properties are similar. The proofs to both theorems are summarized in Appendix A.

V. Variational Bayes Inference

To motivate the theory of variational inference, we first recognize that the equality

$$\int_{\boldsymbol{O}} Q(\boldsymbol{O}) \ln \frac{Q(\boldsymbol{O})}{P(\boldsymbol{O}|\boldsymbol{X})P(\boldsymbol{X})} d\boldsymbol{O} = \int_{\boldsymbol{O}} Q(\boldsymbol{O}) \ln \frac{Q(\boldsymbol{O})}{P(\boldsymbol{X}|\boldsymbol{O})P(\boldsymbol{O})} d\boldsymbol{O}, \tag{15}$$

can be rewritten as

$$\ln P(\boldsymbol{X}) = \mathcal{L}(Q) + KL(Q||P), \tag{16}$$

where $\boldsymbol{O}$ represents the model latent parameters $\boldsymbol{O} = \{\{\boldsymbol{\Theta}_k\}_{k=1}^K, \boldsymbol{z}, \boldsymbol{d}, \{\boldsymbol{\pi_t}\}_{t=1}^T, \boldsymbol{w}\}$, $\boldsymbol{X}$ the observed data, $Q(\boldsymbol{O})$ some yet to be determined approximating density and

$$\mathcal{L}(Q) = \int_{\boldsymbol{O}} Q(\boldsymbol{O}) \ln \frac{P(\boldsymbol{X}|\boldsymbol{O})P(\boldsymbol{O})}{Q(\boldsymbol{O})} d\boldsymbol{O}, \quad KL(Q||P) = \int_{\boldsymbol{O}} Q(\boldsymbol{O}) \ln \frac{Q(\boldsymbol{O})}{P(\boldsymbol{O}|\boldsymbol{X})} d\boldsymbol{O}. \tag{17}$$

For inference purposes, instead of drawing $z_{t,i} \sim Mult(\boldsymbol{\tau}_t)$, we use an extra variable $d_{t,i}$ indicating the task/document we are drawing the mixing weights $\boldsymbol{\tau}_{d_{t,i}}$ from; for each document-dependent $\boldsymbol{x}_{t,i}$ we first draw the task indicator variable $d_{t,i}$ from a stick-breaking construction and then the corresponding topic indicator $z_{t,i}$ as follows:

$$z_{t,i} \sim Mult(\boldsymbol{\pi}_{d_{t,i}}), \quad d_{t,i} \sim Mult(\boldsymbol{V}),$$

$$V_q = w_q \textstyle\prod_{l=1}^{q-1}(1 - w_l), \quad w_q \sim Beta(1, d_0), \tag{18}$$

where $Beta(1, d_0)$ corresponds to $Beta(c_0, d_0)$ in (5), with $c_0 = 1$.

Therefore, the joint distribution of the indicator variables $\boldsymbol{d}$ and $\boldsymbol{z}$ can be written as follows:

$$\begin{aligned} p(d_{t,i} = v, z_{t,i} = k|\boldsymbol{\pi}_t, \boldsymbol{\Theta}_k, \boldsymbol{x}_{t,i}) &\propto (\prod_{i=1}^{N_t} p(\boldsymbol{x}_{t,i}|\boldsymbol{\Theta}_{z_{t,i}=k}, \boldsymbol{\pi}_t)) p(z_{t,i} = k|d_{t,i}, \boldsymbol{\pi}_t, \boldsymbol{\Theta}_k) p(d_{t,i} = v) \\ &= (\prod_{i=1}^{N_t} \prod_{j=1}^{J} Mult(\boldsymbol{\theta}^j_{z_{t,i}=k}) \pi_{t,k} w_v \prod_{l=1}^{v-1}(1 - w_l), \end{aligned} \tag{19}$$

where N_t is the total number of documents in block t, and $x^j_{t,i}$ corresponds to word j in $\boldsymbol{x}_{t,i}$.

Our desire is to best approximate the true posterior $P(\{\boldsymbol{\Theta}_k\}_{k=1}^K, \boldsymbol{z}, \boldsymbol{d}, \{\boldsymbol{\pi_t}\}_{t=1}^T, \boldsymbol{w}|\boldsymbol{X})$ by minimizing $KL(Q||P)$, and this is accomplished by maximizing $\mathcal{L}(Q)$. In doing so, we assume that $Q(\boldsymbol{O})$

can be factorized, meaning

$$Q(\boldsymbol{O}) = Q(\{\boldsymbol{\Theta}_k\}_{k=1}^K, \boldsymbol{z}, \boldsymbol{d}, \{\boldsymbol{\pi_t}\}_{t=1}^T, \boldsymbol{w}) = Q(\{\boldsymbol{\Theta}_k\}_{k=1}^K)Q(\boldsymbol{z})Q(\boldsymbol{d})Q(\{\boldsymbol{\pi_t}\}_{t=1}^T)Q(\boldsymbol{w}). \tag{20}$$

A general method for writing inference for conjugate-exponential Bayesian networks, as outlined in [22], is as follows: for a given node in a graph, write out the posterior as though everything were known, take the natural algorithm, the expectation with respect to all unknown parameters and exponentiate the result. Since it requires computational resources comparable to the expectation-maximization (EM) algorithm, variational inference is fast relative to Markov chain Monte Carlo (MCMC) [15] methods (based on empirical studies for this particular application, and depending on what level of convergence MCMC is run to).

A. VB-E step

For the VB-E step, we calculate the variational expectation with respect to all unknown model parameters $\boldsymbol{\Theta}_k$, $\boldsymbol{\pi}_t$ and w_t. The variational equations of the model parameters $\boldsymbol{\Theta}_k$, $\boldsymbol{\pi}_t$ and w_t are shown below; their derivation is summarized in Appendix C. The analysis yields

$$\begin{aligned}
\widetilde{\boldsymbol{\Theta}_k} &= exp[\sum_{j=1}^{J}\sum_{m=1}^{M}(\sum_{T_k}\langle x_{t,i}^j\rangle + \beta/M - 1)\ln(\theta_k^{j,m})], \\
\widetilde{\boldsymbol{\pi_t}} &= exp[\sum_{i=1}^{N}\sum_{k=1}^{K}(\langle \ln \pi_{t,k}\rangle + \langle \ln w_t\rangle + \sum_{q=1}^{t-1}\langle \ln(1-w_q)\rangle) + \sum_{k=1}^{K}(\alpha/K - 1)\langle \ln \pi_{t,k}\rangle], \\
\widetilde{w_t} &= exp[\sum_{i=1}^{N_t}\sum_{k=1}^{K}(\langle \ln w_t\rangle + \sum_{q=1}^{t-1}\langle \ln(1-w_q)\rangle) + (d_0 - 1)\langle \ln(1-w_t)\rangle],
\end{aligned} \tag{21}$$

where

$$\langle \ln \pi_{t,k}\rangle = \psi(\langle n_{t,k}\rangle + \alpha) - \psi(N_t + 1),$$

$$\langle x_{t,i}^j\rangle = x_{t,i}^j p(x_{t,i}^j | z_{t,i} = k),$$

$$\langle \ln w_v\rangle = \psi(1 + N_v) - \psi(1 + d_0 + \sum_{l=v}^{T} N_l),$$

$$\langle \ln(1-w_l)\rangle \;=\; \psi(d_0+\sum_{m=l+1}^{T} N_m)-\psi(1+d_0+\sum_{m=l}^{T} N_m), \tag{22}$$

with $\psi(\cdot)$ the digamma function, $\langle n_{t,k}\rangle$ the number of words sharing topic k in block t, $\boldsymbol{\beta}=[\beta/M,...,\beta/M]$ the Dirichet hyper-parameters for the priors on the words distribution, and $m\in\{1,2.,...,M\}$ a possible outcome of the multinomial distributions on the word counts.

B. VB-M step

Updating the variational posteriors in the VB-M step is performed by updating the sufficient statistics of the model parameters, obtained from the VB-E step. The analysis yields

$$\begin{aligned} Q(\boldsymbol{\Theta}_k) &= \prod_{j=1}^{J} Dir(\beta/M+\langle p_{k,1}\rangle,...,\beta/M+\langle p_{k,M}\rangle),\\ Q(\boldsymbol{\pi}_t) &= Dir(\alpha/K+\langle n_{t,1}\rangle,...,\alpha/K+\langle n_{t,K}\rangle),\\ Q(w_t) &= Beta(1+\langle m_t\rangle, d_0+\sum_{b=1}^{t-1}\langle m_b\rangle), \end{aligned} \tag{23}$$

where $\langle m_t\rangle=\sum_{k=1}^{K}\langle n_{t,k}\rangle$ and $\langle p_{k,m}\rangle$ is the number of words with outcome m in topic k.

VI. Experimental Results

The proposed model is demonstrated on two data sets, each corresponding to a sequence of documents with known time dependence: (i) the NIPS data set [10] containing publications from the NIPS conferences between 1987 and 1999 and (ii) every United States presidential State of the Union Address from 1790-2008.

As comparisons to the dDTM model developed here, we consider LDA [6] and TOT [20], and dDTM with innovation weights set as $\{w_t\}_{t=2}^{T}=1$ (termed DTM). For the dDTM framework, we initialized the hyper-parameters as follows: the parameter $\alpha=1$, $c_0=1$, $d_0=2$, and Dirichlet distributions with uniform parameters $\boldsymbol{\beta}=[\frac{1}{M},...,\frac{1}{M}]$ as priors on the words distribution; the integer M defines the number of possible outcomes concerning the occurrence of a given word in a document, and this is detailed below for the particular examples. We ran VB until the relative change in the marginal likelihood bound [3] was less than 0.01%. For the LDA and TOT model initializations, we used exchangeable Dirichlet distributions as priors on word probabilities and initialized the Dirichlet hyper-parameters for the topic mixing weights with $\boldsymbol{\alpha}=[\frac{1}{K},...,\frac{1}{K}]$. The truncation level was set to

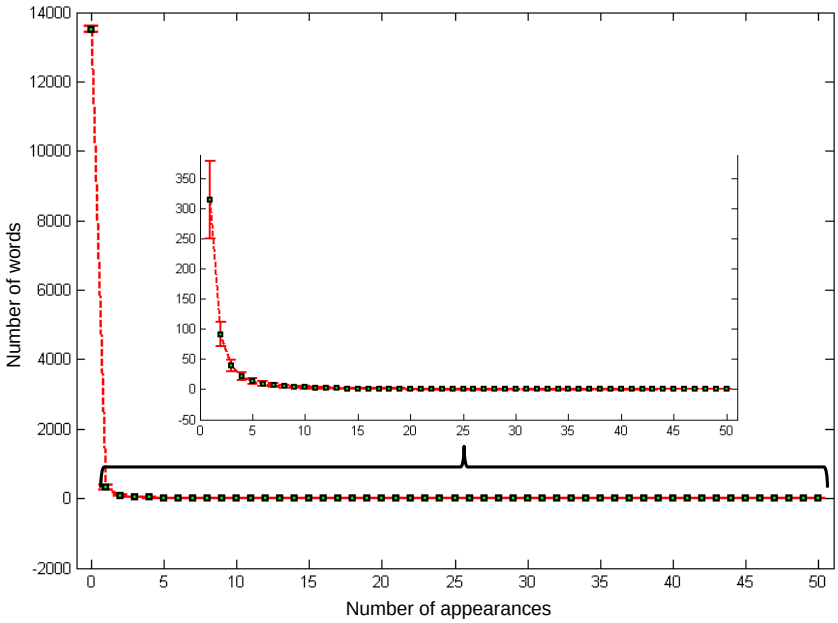


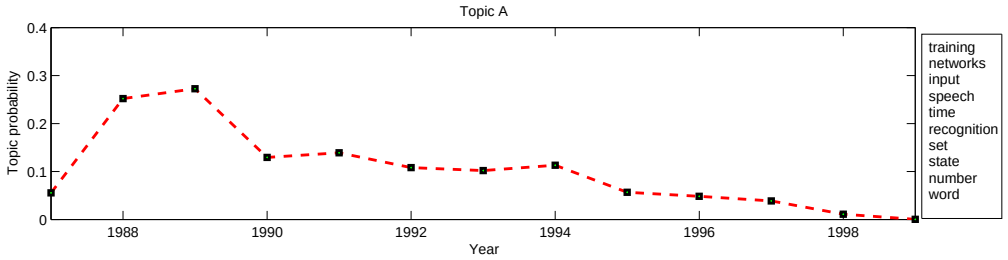
Fig. 2. Histogram of the rate of word appearances in the NIPS data set; the horizontal axis represents the number of times a given word appears in one document, and the vertical axis quantifies the number of times such words occurred across all documents. For example, in an average document, there will be 95 words that appear twice. From this we note that most words rarely occur more than five times in a given document.

$K = 50$ topics in all four models. For the reasons discussed in Section III, the dDTM are expected to be insensitive to the setting of K, as long as it is "large enough"; we also performed studies of the below data for $K = 75$ and $K = 100$, with very similar results manifested.

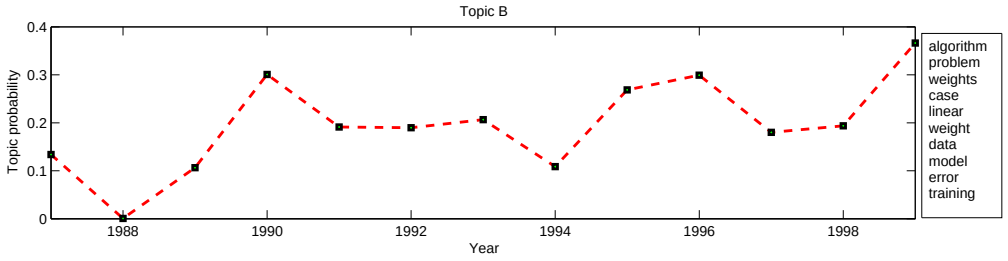
A. NIPS Data Set

The NIPS (Neural Information Processing Systems) data set comprises $1,740$ publications. The total number of unique words was $J = 13,649$. The observation vector $\boldsymbol{x}_{t,i}$ corresponds to the frequency of all words in paper i of the NIPS proceedings from year t. We set the total number of outcomes of the multinomial distributions to $M = 5$; $m = 1$ corresponds to a word occurring zero times, $m = 2$ corresponds to a word occurring once or twice, $m = 3$ corresponds to a word occurring between three-five times, $m = 4$ corresponds to a word occurring between six-ten times, and $m = 5$ corresponds to a word occurring more than ten times in a publication. This decomposition was defined based on examining a histogram of the rate with which any given word appeared in a given publication (see Fig. 2).

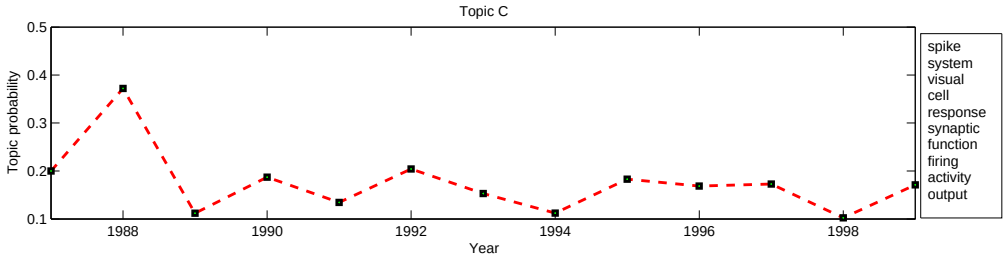
We first estimated the dDTM posterior distributions over the entire set of topics; the time evolution of the posterior dDTM probabilities for four representative topics and their ten most probable words,



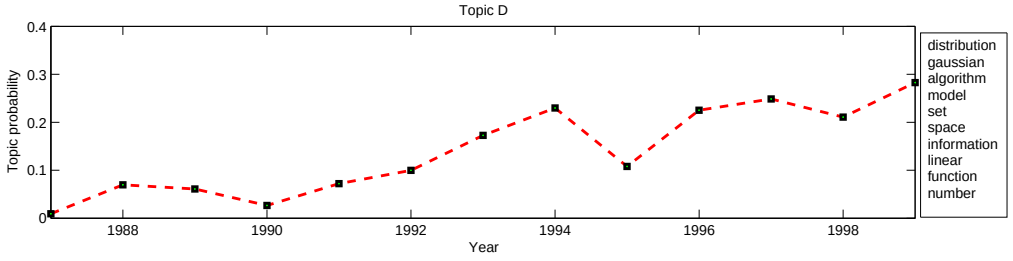
(a)



(b)



(c)



(d)

Fig. 3. Posterior topic probabilities distribution and most probable words for NIPS data set, as computed by the dDTM model.

as computed via the posterior updates of words distributions within topics, are shown in Fig. 3; we ran the algorithm 20 times (with different randomly selected initializations) and chose the VB realization with the highest lower bound.

We then selected the years when the four topics represented above reached their highest probability of being drawn and identified associated publications; as we can see in Table I, for a given topic,

there is a strong dependency between the most probable words and associated publications, with this proving to be a useful method of searching for papers based on a topic name or topic identifying words. These representative results are interpreted as follows: Topics A and C appear to be related to neural networks and speech processing, which appear to have a diminishing importance with time. By contrast, Topics B and D appear to be related to more statistical approaches, which have an increasing importance with time. The specific topic label is artificially given; it corresponds to one indicator variable in the VB solution.

In our next experiment, we quantitatively compared the dDTM, LDA, TOT and DTM models by computing the *perplexity* of a held-out test set; perplexity [6] is a popular measure used in language modeling, reflecting the difficulty of predicting new unseen documents after learning the model from a training data set. The perplexity results considered here are not the typical held-out at random type, but real prediction where we are using the past to build a model for the future; a lower perplexity score indicates better model performance. The perplexity for a test set of N_{test} documents is defined to be

$$P = exp\{-\frac{\sum_{i=1}^{N_{test}} \ln p(\boldsymbol{x}_{test,i})}{\sum_{i=1}^{N_{test}} n_{test,i}}\}, \tag{24}$$

where $\boldsymbol{x}_{test,i}$ represents the document i in the test set and $n_{test,i}$ is the number of words in document $\boldsymbol{x}_{test,i}$.

In our experiment the role of a document is played by a publication; the perplexity results correspond to a real prediction scenario, where we are using the past to build a model for the future: we held out all the publications from one year for test purposes and trained the models on all the publications from all the years prior to the testing year; as testing years we considered the last five years between 1995 and 1999.

The perplexity for the LDA and TOT models was computed as in [6]; for the dDTM and DTM models it was computed as follows:

$$\begin{aligned} P_{dDTM} &= exp\{-\frac{\sum_{i=1}^{N_{test}} \ln(\sum_{j=1}^{J} \sum_{z_n} \sum_{t=1}^{T} p(x_{test,i}^{j}|z_{test,i}, \boldsymbol{\Theta}) p(z_{test,i}|\boldsymbol{\tau}) p(\boldsymbol{\tau}_{d_{test,i}=t}|\boldsymbol{\alpha}) \frac{d_0}{1+d_0} w_t \prod_{l>t}(1-w_l))}{\sum_{i=1}^{N_{test}} n_{test,i}}\}, \\ P_{DTM} &= exp\{-\frac{\sum_{i=1}^{N_{test}} \ln(\sum_{j=1}^{J} \sum_{z_n} \sum_{t=1}^{T} p(x_{test,i}^{j}|z_{test,i}, \boldsymbol{\Theta}) p(z_{test,i}|\boldsymbol{\tau}) p(\boldsymbol{\tau}_{d_{test,i}=t}|\boldsymbol{\alpha})}{\sum_{i=1}^{N_{test}} n_{test,i}}\}, \end{aligned} \tag{25}$$

where z is the topic indicator, i is the publication index, d is the block/year indicator, T is the total number of training years, and d_0 is the hyper-parameter of the beta prior distributions $Beta(1, d_0)$ on the innovating weights $\{w_t\}_{t=2}^{T}$. The perplexity computation for the dDTM model is provided in

TABLE I
REPRESENTATIVE TOPICS FROM THE NIPS DATABASE, WITH THEIR MOST PROBABLE WORDS AND ASSOCIATED PUBLICATIONS.

Topic A (year 1989)	training networks input speech time recognition set state number word	'A Continuous Speech Recognition System Embedding MLP into HMM' 'Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters' 'Speaker Independent Speech Recognition with Neural Networks and Speech Knowledge' 'The Cocktail Party Problem: Speech/Data Signal Separation Comparison between Back propagation and SONN‘
Topic B (year 1999)	algorithm problem weights case linear weight data model error training	'Model Selection for Support Vector Machines' 'Uniqueness of the SVM Solution' 'Differentiating Functions of the Jacobian with Respect to the Weights' 'Transductive Inference for Estimating Values of Functions'
Topic C (year 1988)	spike system visual cell response synaptic function firing activity output	'Models of Ocular Dominance Column Formation: Analytical and Computational Results' 'Modeling the Olfactory Bulbs Coupled Nonlinear Oscillators' 'A Model for Resolution Enhancement (Hyperacuity) in Sensory Representation' 'A Computationally Robust Anatomical Model for Retinal Directional Selectivity'
Topic D (year 1999)	distribution gaussian algorithm model set space information linear function number	'Local Probability Propagation for Factor Analysis‘ 'Algorithms for Independent Components Analysis and Higher Order Statistics‘ 'Correctness of Belief Propagation in Gaussian Graphical Models of Arbitrary Topology‘ 'Data Visualization and Feature Selection: New Algorithms for Nongaussian Data'

Appendix B.

Figure 4 shows the mean value and standard deviation of the perplexity of dDTM, LDA TOT and DTM models with $K = 50$ topics; we ran 20 VB realizations for the dDTM, LDA and DTM and 20 MCMC realizations (with 1000 iterations each) for the TOT model. We see that the dDTM model slightly outperforms the other models, with the LDA and TOT better than the DTM. The improved performance of dDTM model is due to the time evolving structure; the order of publications plays an important role in predicting new documents, through the innovation weight probability $\boldsymbol{w}$, as can be seen in (25).

While the NIPS database is widely used for topic modeling, the relatively small number of years it entails mitigates interesting analysis of the ability of dDTM to model the time-evolving properties of documents. This motivates the next example, which corresponds to a yearly database extending over 200 years.

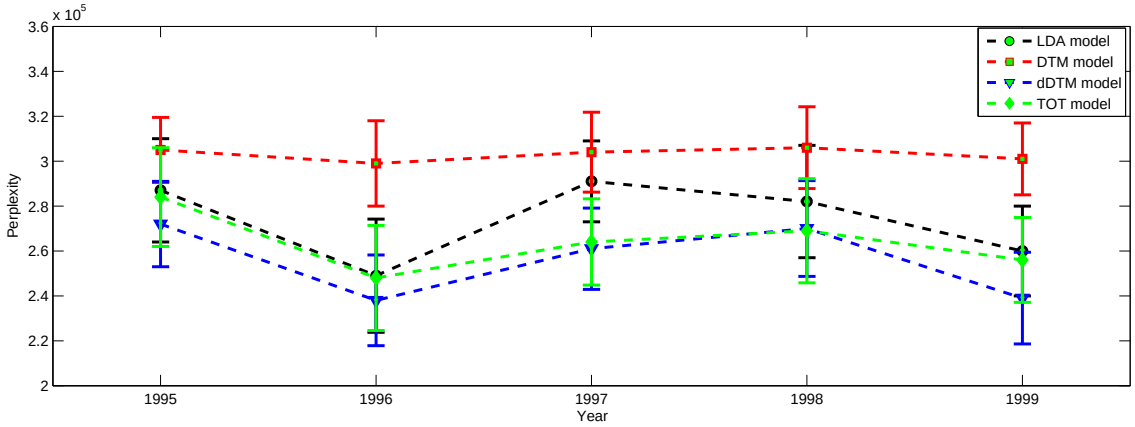


Fig. 4. Perplexity results on the NIPS data set for dDTM, LDA, TOT and DTM: mean value and standard deviation.

B. State of the Union Data Set

The State of the Union data set comprised 20,431 paragraphs, each with a time stamp from 1790 to 2008. The observation vector $\boldsymbol{x}_{t,i}$ corresponds to the frequency of all words in paragraph i of the State of the Union from year t. In this (motivating) example, "document" i for year t corresponds to paragraph i from the State of the Union for year t. Therefore, the model assumes the State of the Union is represented by a mixture of topics, and within dDTM the mixture weights evolve with time.

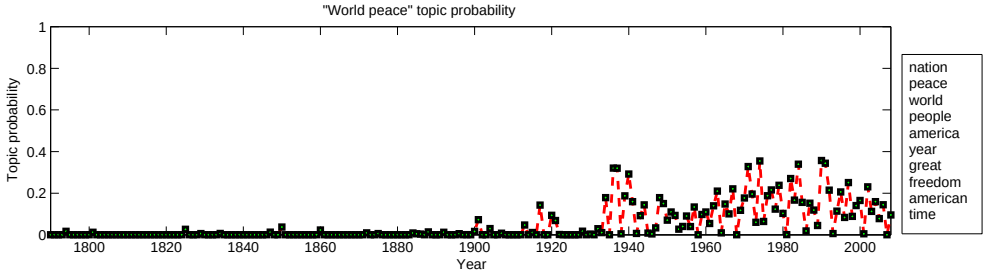
After removing common stop words by referencing a common list which can be found at $http://www.dcs.gla.ac.uk/idom/ir\_resources/linguistic\_utils/stop\_words$, and applying the Porter stemming algorithm [19], the total number of unique words was $J = 747$. In the rare years where two state of the union addresses were given, the address given by the outgoing president was used. Similar to the NIPS data set, each paragraph was represented as a datum, a vector of word counts over the dictionary. However, to match the data structure, we set the number of possible outcomes as $M = 2$, indicating whether a given word was present ($m = 1$) or not ($m = 2$) in a given paragraph. This structure corresponds to a binomial-beta representation of the words distribution, a special case of the multinomial-Dirichlet model used in the NIPS experiment.

For our first experiment we estimated the posterior distributions over the entire set of topics, for each of the three models mentioned above. Results for the dDTM model are shown in Fig. 5 for the time evolution of the posterior dDTM probabilities for five important topics in American history: 'American civil war', 'world peace', 'health care', 'U. S. Navy' and 'income tax'; similar to the NIPS experiment, we ran the algorithm 20 times (with random initialization) and chose the

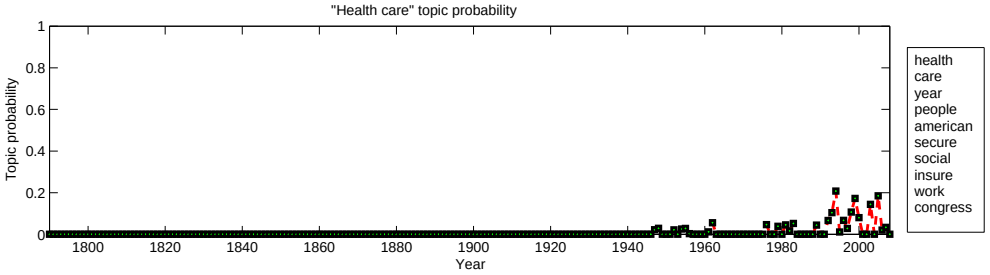
VB realization with the highest lower bound. The topic distributions preserve sharp peaks in time indicating significant information content at particular historical time points. It is important to mention that we have (artificially) named the topics based on their ten most probable words. The corresponding most probable words are shown in the right hand side of each plot. In comparison, the dDTM seems to perform better than LDA, TOT and DTM: 'American civil war' and 'health care' are topics that were not found by LDA, TOT or DTM. The better performance of the dDTM model can be explained by the sharing properties that exist between adjacent blocks, properties controlled by the innovation weight $\boldsymbol{w}$. Figures 6, 7 and 8 show topic distributions and their associated ten most probable words for the LDA, TOT and DTM models, respectively.

Concerning the interpretation of these results, we note that the US was not a world power until after World War II, consistent with Fig. 5(a). National health care in the US became a political issue in the early and mid 1990s, and continues such to this day. The US Navy was an important defense issue from the earliest days of the country, particularly in wars with Britain and Spain. With the advent of aircraft, the importance of the navy diminished, while still remaining important today. Concerning Fig. 5(d) on taxation, the first federal laws on federal (national) income tax were adopted by Congress in 1861 and 1862, and the Sixteenth Amendment to the US Constitution (1913) also addressed federal taxation. The heavy importance of this topic around 1920 is attributed to World War I, with this becoming an important issue/topic thereafter (concerning the appropriate tax rate). The US Civil War, which had a heavy focus on "state rights" was of course in the 1860-1865 period, with state rights being a topic of some focus sporadically thereafter.

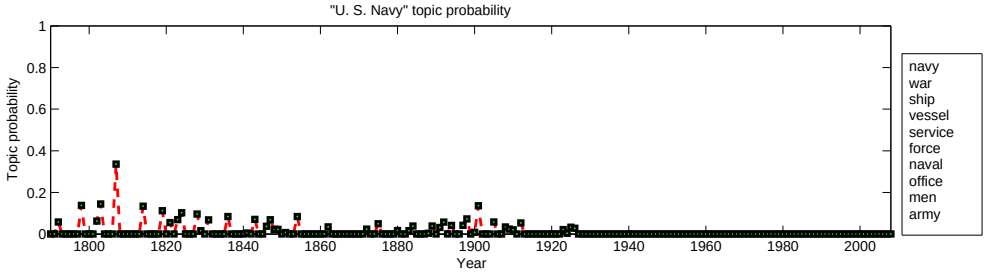
Another advantage of dDTM over LDA, TOT and DTM is that it allows us to analyze the dynamic evolution of topic mixing weights through innovation probabilities. For that, using the dDTM model, we examined the innovation weight probability $\boldsymbol{w}$, for each year from 1790 to 2008. Table II shows the years when the mean innovation probability was greater than 0.8, the year-period description and the name of the associated president. As observed during those years, important political events are well identified by dDTM. For each of the innovating years shown in Table II we also estimated the 'most innovative' words with respect to their previous year. For example, we were interested in finding the words that caused innovation during year 1829. For that, we first calculated the distribution of the words within one year, by integrating out the topics; we then estimated the Kullback-Leibler (KL) divergence between the probabilities of a given word belonging to two consecutive years, 1828 and 1829. The higher the KL distance is for a given word, the more innovation it produces; the ten



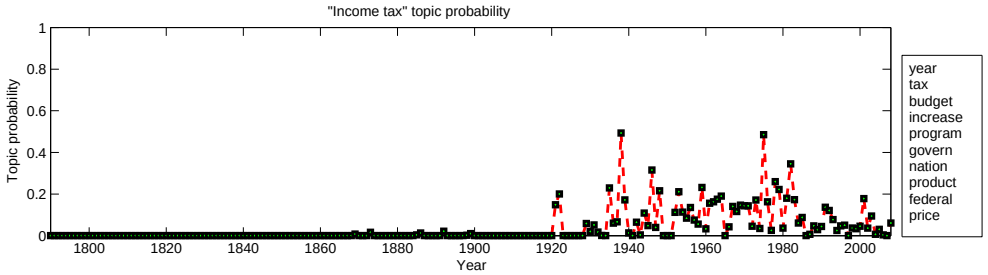
(a)



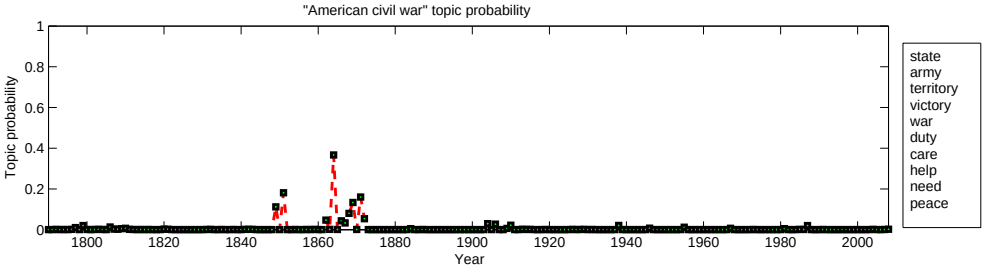
(b)



(c)



(d)



(e)

Fig. 5. dDTM model - topic probabilities distribution and most probable words for State of the Union data set. (a) World peace, (b) health care, (c) U.S. Navy, (d) income tax, (e) Civil War.

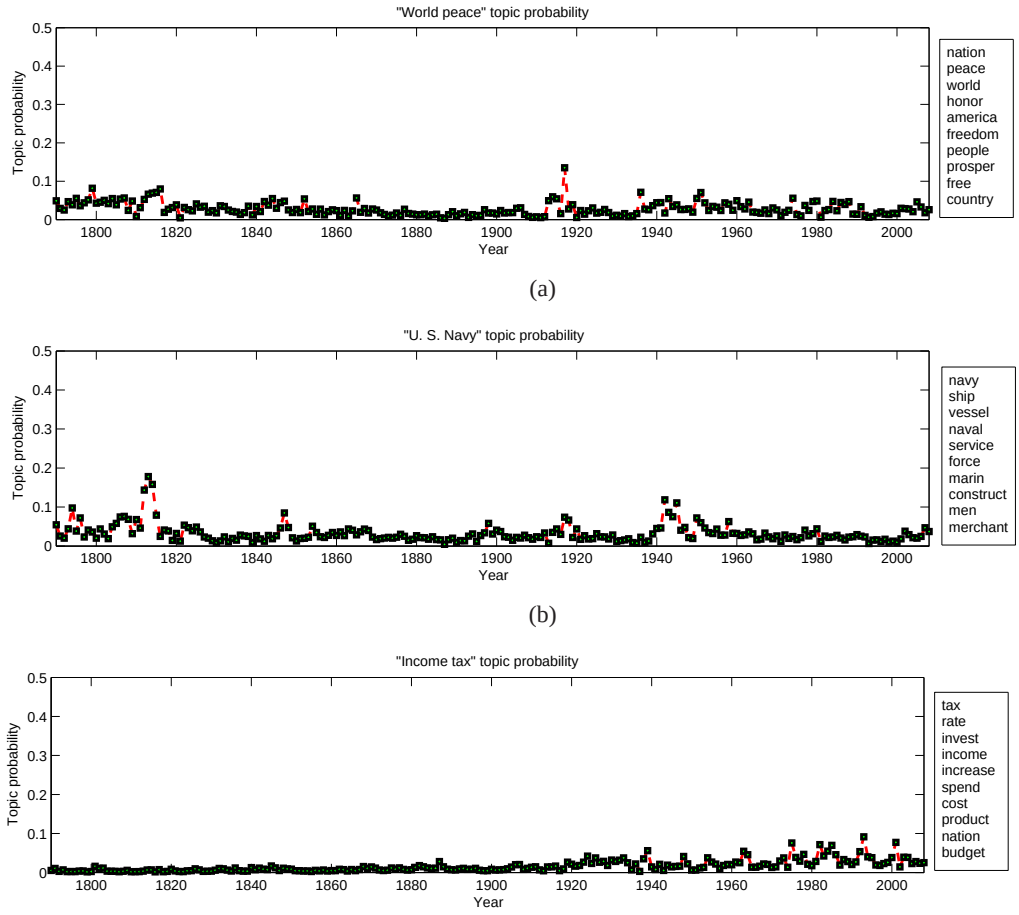


Fig. 6. LDA model - topic probabilities distribution and most probable words. (a) World peace, (b) U.S. Navy, (c) income tax.

most innovative words for each of the years of interest are shown in Table III.

The results in Table II ideally (if dDTM works properly) correspond to periods of significant change in the US. Concerning interpretation of these results, President Jackson was the first non-patrician US president, and he brought about significant change ($e.g.$, he ended the national banking system in the US). The Civil War, World War I, World War II, Vietnam and the end of the Cold War were all significant changes of "topics" within the US. Ronald Reagan also brought a level of conservative government to the US which was a significant change. These key periods, as inferred automatically via dDTM, seem to be in good agreement with historical events in the US.

We also analyzed the ability of dDTM to group paragraphs into topics. We chose two distinguishing years in American history, 1861 (during the American Civil War) and 2002 (post terrorist attacks) and show the most probable three topics as computed via the VB posterior updates and their associated

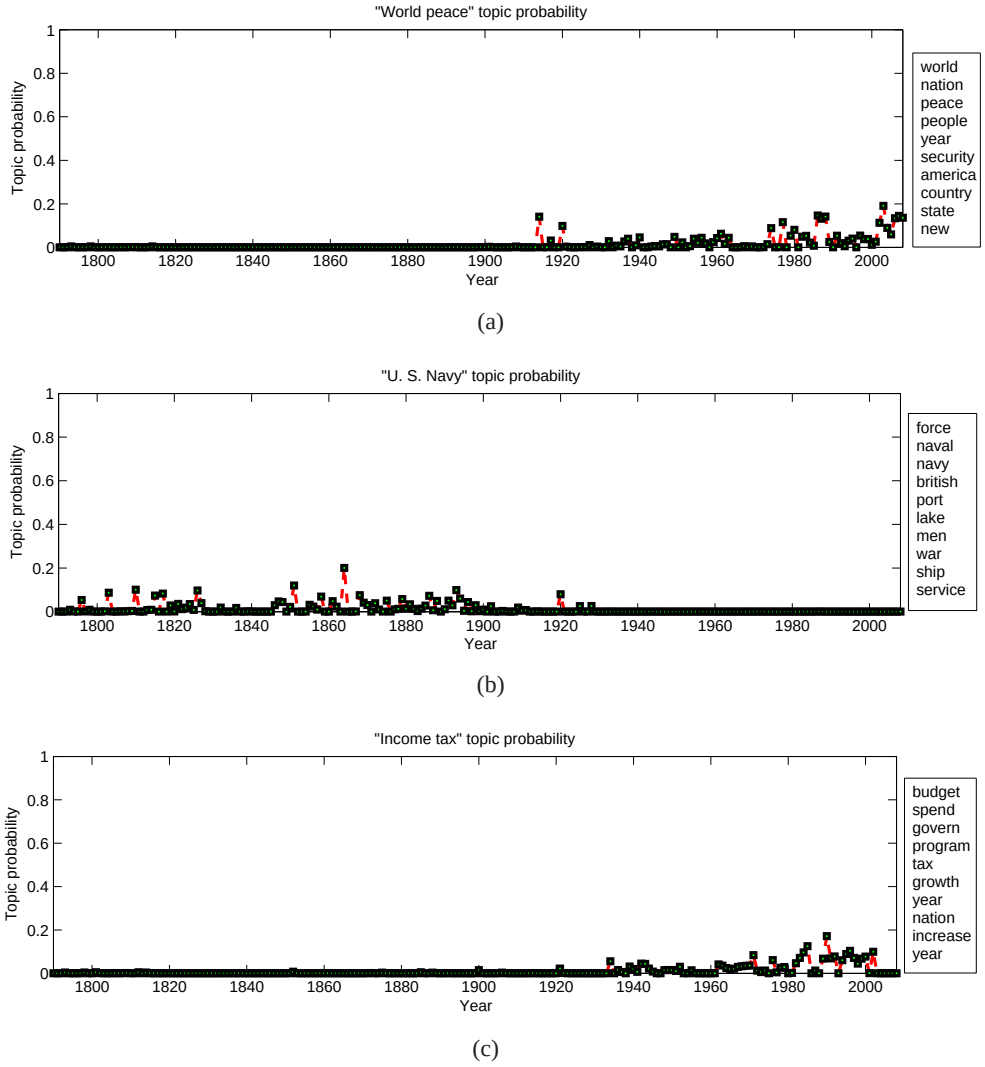


Fig. 7. TOT model - topic probabilities distribution and most probable words. (a) World peace, (b) U.S. Navy, (c) income tax.

paragraphs (see Tables IV and V). In 1861 the three major topics were ‘political situation’, ‘finances’ and ‘army’, whereas in 2002 the topics were ‘terrorism’, ‘national budget’ and ‘overall progress of the country’. In both cases, the algorithm automatically clusters the paragraphs using what appears to be an accurate topic representation.

To show the dynamic structure of dDTM, we selected 2002 as a reference year and its two years before and after as years where topic transition could be manifested. For each of the five years, we estimated the most probable topic and identified its associated paragraphs. As we can see in Table VI, a topic transition is manifested during this time interval: if in 2000, the main topic was ‘economy’, in the following years attention is paid to ‘education’, ‘terrorism’, ‘economy’ again and ‘war in Iraq’,

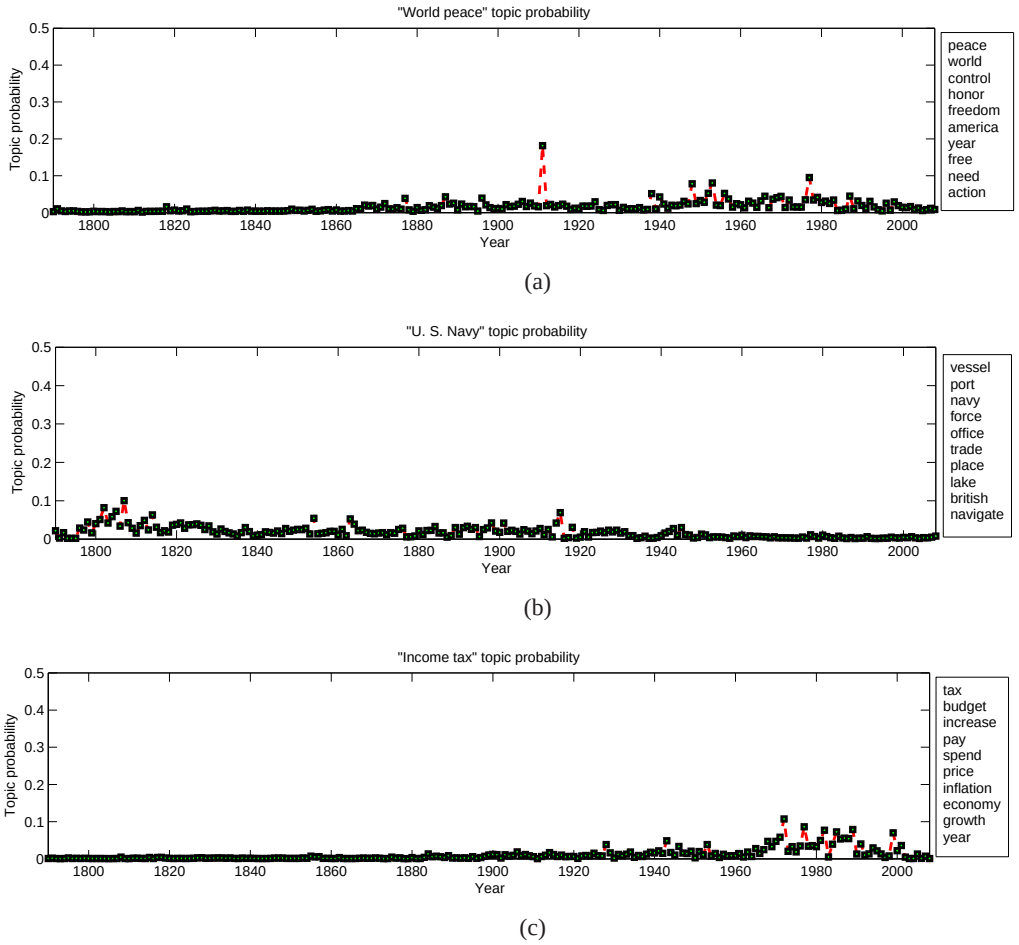


Fig. 8. DTM model - topic probabilities distribution and most probable words. (a) World peace, (b) U.S. Navy, (c) income tax.

TABLE II

YEARS WITH THE MEAN INNOVATION WEIGHT PROBABILITY GREATER THAN 0.8 IN THE DDTM MODEL, YEAR-PERIOD DESCRIPTION AND THE ASSOCIATED PRESIDENT.

Year	Mean innovation weight probability	Period description	President
1829	0.87	Pres. A. Jackson's era	A. Jackson
1831	0.84		
1861	0.82	Civil war	A. Lincoln
1909	0.81	Industrialization	W. H. Taft
1919	0.85	Post "world war I" era	W. Wilson
1938	0.84	Roosevelt's economical recovery	F. D. Roosevelt
1939	0.82	Second world war	F. D. Roosevelt
1965	0.8	Vietnam's war	L. B. Johnson
1981	0.81	R. Reagan's promised economic revival and the recession	R. Reagan
1982	0.89		
1990	0.82	The end of the "cold war"	G. H. W. Bush

TABLE III
MOST INNOVATIVE WORDS IN THE YEARS WITH THE MEAN INNOVATION WEIGHT PROBABILITY GREATER THAN 0.8

1829	1831	1861	1909	1919	1938	1939	1965	1981	1982	1990
Indian	Treaty	State	Unit	Legislation	Federal	Peace	War	Federal	Spend	World
Law	Unit	Right	Treaty	Army	Tax	Freedom	World	Public	Budget	Peace
Tribe	Claim	War	British	Labor	Budget	America	Help	Program	Agriculture	War
Report	Convent	Increase	Report	Navy	Billion	War	Social	Budget	Senate	Free
Secretary	Prosper	Total	Negotiate	Peace	Deficit	Cut	Nation	Union	Let	Cut
Service	Nation	Power	Spain	District	Increase	God	Care	Increase	Know	Union
Work	Report	June	People	Federal	Fiscal	Spend	Million	Health	House	Strength
Constitute	Negotiate	Total	Subject	Ship	Income	Budget	Parent	Legislation	Tax	Cooper
Construct	Minister	Year	Session	Regulation	Rate	Percent	Peace	Tax	People	Rate
Navy	People	Service	Claim	Law	Spend	Army	Drug	America	Represent	Great

TABLE IV
PARAGRAPH CLUSTERING ANALYSIS FOR YEAR 1861: TOP THREE MOST PROBABLE TOPICS AND THEIR ASSOCIATED PARAGRAPHS.

Topic 40	Topic 41	Topic 22
Nations thus tempted to interfere are not always able to resist the counsels of seeming expediency and ungenerous ambition although measures adopted under such influences seldom fail to be unfortunate and injurious to those adopting them.	The revenue from all sources including loans for the financial year ending on the 10th of June was and the expenditures for the same period including payments on account of the public debt were leaving a balance in the Treasury on the 1st of July of	I respectfully refer to the report of the Secretary of War for information respecting the numerical strength of the Army and for recommendations having in view an increase of its efficiency and the wellbeing of the various branches of the service intrusted to his care.
It is not my purpose to review our discussions with foreign states because whatever might be their wishes or dispositions the integrity of our country and the stability of our Government mainly depend not upon them but on the loyalty virtue patriotism and intelligence of the people.	For the first quarter of the financial year ending on the 30th of September the receipts from all sources including the balance of the 1st of July were and the expenses leaving a balance on the 1st of October of	The large addition to the Regular Army in connection with the defection that has so considerably diminished the number of its officers gives peculiar importance to his recommendation for increasing the corps of cadets to the greatest capacity of the Military Academy.
Some treaties designed chiefly for the interests of commerce and having no grave political importance have been negotiated and will be submitted to the Senate for their consideration.	The revenue from all sources during the fiscal year ending June including the annual permanent appropriation of for the transportation of free mail matter was being about 12 per cent less than the revenue for	It is gratifying to know that the patriotism of the people has proved equal to the occasion and that the number of troops tendered greatly exceeds the force which Congress authorized me to call into the field.

respectively. The terrorist attacks on the World Trade Center and on the Pentagon occurred in 2001, manifesting the clear change in the important “topics”.

Finally, we again compared dDTM, LDA, TOT and DTM models by computing their perplexities; in this case, the role of a document was represented by a paragraph and, similar to the NIPS experiment, we considered the task of real prediction, by holding out all the paragraphs from one year for test purposes and training the models on all the paragraphs from all the years prior to the testing year; as testing years we considered the ending year of each decade from 1901 to 2000.

Figure 9 shows the mean perplexity of dDTM, LDA, TOT and DTM models with $K = 50$ topics and 10 testing years. We ran 20 VB realizations for the dDTM, LDA and DTM and 20 MCMC realizations (with 1000 iterations each) for the TOT model; the standard deviation values are included as well. We see that the dDTM model consistently performs better than the other models. We also observe that LDA and TOT slightly outperform the DTM model due to the Dirichlet distribution approximations

TABLE V

PARAGRAPH CLUSTERING ANALYSIS FOR YEAR 2002: TOP THREE MOST PROBABLE TOPICS AND THEIR ASSOCIATED PARAGRAPHS.

Topic 19	Topic 2	Topic 39
America has a window of opportunity to extend and secure our present peace by promoting a distinctly American internationalism. We will work with our allies and friends to be a force for good and a champion of freedom. We will work for free markets free trade.	Government cannot be replaced by charities or volunteers. Government should not fund religious activities. But our Nation should support the good works of these good people who are helping their neighbors in need. So I propose allowing all taxpayers whether they itemize or not to deduct their charitable contributions. Estimates show this could encourage as much as one billion a year in new charitable giving money that will save and change lives.	Together we are changing the tone in the Nation's Capital. And this spirit of respect and cooperation is vital because in the end we will be judged not only by what we say or how we say it we will be judged by what were able to accomplish.
Our Nation also needs a clear strategy to confront the threats of this century threats that are more widespread and less certain. They range from terrorists who threaten with bombs to tyrants in rogue nations intent upon developing weapons of mass destruction .To protect our own people our allies and friends we must develop and we must deploy effective missile defenses.	I propose we make a major investment in conservation by fully funding the Land and Water Conservation Fund and our national parks. As good stewards we must leave them better than we found them. So I propose to provide one billion over ten years for the upkeep of these national treasures.	The last time I visited the Capitol I came to take an oath on the steps of this building. I pledged to honor our Constitution and laws and I asked you to join me in setting a tone of civility and respect in Washington. I hope America is noticing the difference because we're making progress.
Yet the cause of freedom rests on more than our ability to defend ourselves and our allies. Freedom is exported every day as we ship goods and products that improve the lives of millions of people. Free trade brings greater political and personal freedom. Each of the previous five Presidents has had the ability to negotiate far reaching trade agreements.	The budget adopts a hopeful new approach to help the poor and the disadvantaged. We must encourage and support the work of charities and faith based and community groups that offer help and love one person at a time. These groups are working in every neighborhood in America to fight homelessness and addiction and domestic violence to provide a hot meal or a mentor or a safe haven for our children. Government should welcome these groups to apply for funds not discriminate against them.	Neither picture is complete in and of itself. Tonight I challenge and invite Congress to work with me to use the resources of one picture to repaint the other to direct the advantages of our time to solve the problems of our people. Some of these resources will come from Government, some but not all.

TABLE VI

DYNAMIC STRUCTURE ANALYSIS FOR YEARS 2000-2004: MOST PROBABLE TOPIC AND ASSOCIATED PARAGRAPHS.

Year 2000 (topic 37)	Year 2001 (topic 12)	Year 2002 (topic 19)	Year 2003 (topic 37)	Year 2004 (topic 34)
We begin the new century with over one million new jobs; the fastest economic growth in more than ten years; the lowest unemployment rates in years; the lowest poverty rates in years; the lowest African American and Hispanic unemployment rates on record. America will achieve the longest period of economic growth in our entire history. We have built a new economy.	A budget's impact is counted in dollars but measured in lives. Excellent schools quality health care a secure retirement a cleaner environment a stronger defense, these are all important needs and we fund them. The highest percentage increase in our budget should go to our children's education. Education is my top priority and by supporting this budget you'll make it yours as well.	America has a window of opportunity to extend and secure our present peace by promoting a distinctly American internationalism. We will work with our allies and friends to be a force for good and a champion of freedom. We will work for free markets free trade.	To lift the standards of our public schools we achieved historic education reform which must now be carried out in every school and in every classroom so that every child in America can read and learn and succeed in life. To protect our country we reorganized our Government and created the Department of Homeland Security which is mobilizing against the threats of a new era. To bring our economy out of recession we delivered the largest tax relief in a generation.	We have faced serious challenges together and now we face a choice. We can go forward with confidence and resolve or we can turn back to the dangerous illusion that terrorists are not plotting and outlaw regimes are no threat to us. We can press on with economic growth and reforms in education and Medicare or we can turn back to old policies and old divisions.
Our economic revolution has been matched by a revival of the American spirit crime down by percent to its lowest level in years teen births down years in a row adoptions up by percent welfare rolls cut in half to their lowest levels in years.	When it comes to our schools dollars alone do not always make the difference. Funding is important and so is reform. So we must tie funding to higher standards and accountability for results.	Our Nation also needs a clear strategy to confront the threats of this century threats that are more widespread and less certain. They range from terrorists who threaten with bombs to tyrants in rogue nations intent upon developing weapons of mass destruction. To protect our own people our allies and friends we must develop and we must deploy effective missile defenses.	Our first goal is clear we must have an economy that grows fast enough to employ every man and woman who seeks a job. After recession terrorist attacks corporate scandals and stock market declines our economy is recovering. Yet its not growing fast enough or strongly enough. With unemployment rising our Nation needs more small businesses to open more companies to invest and expand more employers to put up the sign that says Help Wanted.	Having broken the Baathist regime we face a remnant of violent Saddam supporters. Men who ran away from our troops in battle are now dispersed and attack from the shadows. These killers joined by foreign terrorists are a serious continuing danger. Yet we're making progress against them. The once all powerful ruler of Iraq was found in a hole and now sits in a prison cell. The top officials of the former regime we have captured or killed. Our forces are on the offensive leading over patrols a day and conducting an average of raids a week.
Eight years ago it was not so clear to most Americans there would be much to celebrate in the year. Then our Nation was gripped by economic distress social decline political gridlock. The title of a bestselling book asked America What Went Wrong.	Schools will be given a reasonable chance to improve and the support to do so. Yet if they don't if they continue to fail we must give parents and students different options a better public school a private school tutoring or a charter school. In the end every child in a bad situation must be given a better choice because when it comes to our children failure is simply not an option.	Yet the cause of freedom rests on more than our ability to defend ourselves and our allies. Freedom is exported every day as we ship goods and products that improve the lives of millions of people. Free trade brings greater political and personal freedom. Each of the previous five Presidents has had the ability to negotiate far reaching trade agreements.	A growing economy and a focus on essential priorities will be crucial to the future of Social Security. As we continue to work together to keep Social Security sound and reliable we must offer younger workers a chance to invest in retirement accounts that they will control and they will own.	As democracy takes hold in Iraq the enemies of freedom will do all in their power to spread violence and fear. They are trying to shake the will of our country and our friends but the United States of America will never be intimidated by thugs and assassins. The killers will fail and the Iraqi people will live in freedom.

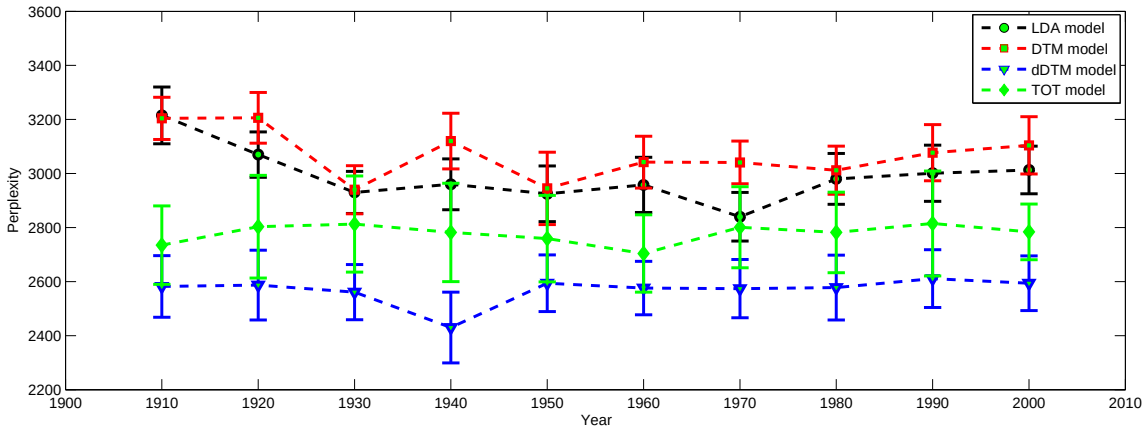


Fig. 9. Perplexity results on the United States presidential State of the Union Address for dDTM, LDA, TOT and DTM: mean value and standard deviation, estimated from 20 randomly initialized VB realizations.

made in the DTM model.

Concerning computational costs, all code was run in Matlab^{TM} on a PC with Intel 2.33GHz processor. For the NIPS data dDTM, LDA, TOT and DTM required (for each VB and MCMC runs) 4 hours and 16 minutes, 3 hours and 22 minutes, 10 hours and 31 minutes, and 3 hours and 45 minutes, respectively. For the State of the Union data these respective times were 25, 22, 104 and 23 minutes. These times are meant to give relative computational costs; none of the software was optimized.

VII. Conclusions

We have developed a novel topic model, the truncated dynamic HDP, or dDTM, to analyze topics associated with documents with known time stamps. The new model allows simple variational Bayesian (VB) inference, yielding fast computation times. The algorithm has been demonstrated on a large database, the US State of the Unions for a 220 year period, and the results seem to be able to highlight significant events in the US history (although it should be emphasized that the authors are not historians, and much further testing and evaluation is required). The algorithm is able to identify important historical topics, as well as periods of time over which significant changes in topics are realized. The model compares favorably with LDA, TOT and a simplified form of dDTM (for which time dependence is ignored).

Concerning future research, other approaches that might be considered for approximate inferences include collapsed sampling [21]. It would be interesting to analyze how these different inferences

influence the overall performance of the model. In order to capture semantics evolution with time, one may consider a similar dynamic model for topics themselves. This could be accomplished by allowing the words distributions change in time; for identifiability, constraints could be used so that the majority of words in a topic, and their associated frequencies, remain constant across time. In addition, the evolution of the model occurred in only one dimension (time). There may be problems for which documents may be collected at different geographical locations, for example from different cities across the world. In this case one may have spatial proximity as well as temporal proximity to consider, when considering inter-document relationships. It is of interest to extend the dynamic structure from one dimension to perhaps a graphical structure, where the nodes of the graph may represent space and time.

We also note there may be general interest within topic-model research in representing a draw from a Dirichlet process in the form in (9). While this increases the complexity of the analysis, it has the significant advantage of allowing one to place a Gamma prior on α and perform full VB inference (we no longer have to set α). As discussed in Section III, α plays an important role in defining the number of expected topics per document (since it controls the number of important mixture weights). One may place a separate prior on the distinct α associated with each document, so that the number of important topics per document may change. The complication with doing this, rather than just directly drawing from $Dir(\alpha/K, \cdots, \alpha/K)$ is that one must now perform inference on many more parameters (on the sticks of the stick-breaking representation). In some applications such added complexity will be warranted by a desire to infer α in a full VB analysis.

ACKNOWLEDGEMENTS

The authors wish to thank Prof. Andrew McCallum and Xuerui Wang of the University of Massachusetts, for allowing us to use their "Topics over time" software, and for several helpful discussions. The authors also thank the reviewers for their comments, which have significantly improved the paper.

REFERENCES

[1] Q. An, E. Wang, I. Shterev, L. Carin., and D. B. Dunson. Hierarchical kernel stick-breaking process for multi-task image analysis. *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[2] M. J. Beal. Variational algorithms for approximate bayesian inference. *Gatsby Computational Neuroscience Unit, Ph.D. thesis, University College London*, 2003.

[3] D. M. Blei and M. I. Jordan. Variational methods for the dirichlet process. *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[4] D. M. Blei and J. D. Lafferty. Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006.

[5] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

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

[7] J. F. Canny and T. L. Rattenbury. A dynamic topic model for document segmentation. *Technical Report, Department of Electrical Engineering and Computer Sciences, University of California at Berkeley*, 2006.

[8] D. B. Dunson. Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, 7:551–568, 2006.

[9] D. B. Dunson and J.-H. Park. Kernel stick-breaking processes. *Biometrika*, 2007.

[10] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Hidden topic markov models. *Artificial Intelligence and Statistics*, 2007.

[11] T. Hofmann. Probabilistic latent semantic analysis. *Proceedings of Uncertainty in Artificial Intelligence*, 1999.

[12] J. Ishwaran and L. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161174, 2001.

[13] J.-H. Park and D. B. Dunson. Bayesian generalized product partition model. 2006.

[14] M. L. Pennell and D. B. Dunson. Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics*, 6:1044–1052, 2006.

[15] L. Ren, D. B. Dunson, and L. Carin. The dynamic hierarchical dirichlet process. *International Conference on Machine Learning*, 2008.

[16] J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

[17] N. Srebro and S. Roweis. Time-varying topic models using dependent dirichlet processes. *Technical Report, Department of Computer Science, University of Toronto*, 2005.

[18] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1582, 2005.

[19] C. J. van Rijsbergen, S. E. Robertson, and M. F. Porter. Information retrieval. *Butterworths, London, 2nd edition*, 6:111–143, 1979.

[20] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. *The Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, 2006.

[21] M. Welling, I. Porteous, and E. Bart. Infinite state bayes-nets for structured domains. *Proceedings of the International Conference on Neural Information Processing Systems*, 2007.

[22] J. Winn and C. M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

[23] J. Zhang, Z. Ghahramani, and Y. Yang. A probabilistic model for online document clustering with application to novelty detection. *Proceedings of Neural Information Processing Systems*, 2004.