Bayesian Analysis (2019) **14**, Number 1, pp. 53–80

Bayesian Analysis of Dynamic Linear Topic Models[*]

Chris Glynn[†], Surya T. Tokdar[‡], Brian Howard[§], and David L. Banks[¶]

Abstract. Discovering temporal evolution of themes from a time-stamped collection of text poses a challenging statistical learning problem. Dynamic topic models offer a probabilistic modeling framework to decompose a corpus of text documents into "topics", i.e., probability distributions over vocabulary terms, while simultaneously learning the temporal dynamics of the relative prevalence of these topics. We extend the dynamic topic model of Blei and Lafferty (2006) by fusing its multinomial factor model on topics with dynamic linear models that account for time trends and seasonality in topic prevalence. A Markov chain Monte Carlo (MCMC) algorithm that utilizes Pólya-Gamma data augmentation is developed for posterior sampling. Conditional independencies in the model and sampling are made explicit, and our MCMC algorithm is parallelized where possible to allow for inference in large corpora. Our model and inference algorithm are validated with multiple synthetic examples, and we consider the applied problem of modeling trends in real estate listings from the housing website Zillow. We demonstrate in synthetic examples that sharing information across documents is critical for accurately estimating document-specific topic proportions. Analysis of the Zillow corpus demonstrates that the method is able to learn seasonal patterns and locally linear trends in topic prevalence.

Keywords: topic model, dynamic linear model, Pólya-Gamma, MCMC.

1 Dynamic text analysis

Text data is ubiquitous. Newspapers, blogs, emails, tweets, and countless other expressions of written language are central to daily communication, as well as formal correspondence. In many cases, the time at which a document is created is an important piece of metadata. When analyzing a corpus of time-stamped documents spanning a considerable amount of time, it is natural to ask whether we can detect and quantify the temporal evolution of its thematic composition. Time-varying detection and quantification of themes may generate hypotheses and lead to insightful answers in social science research.

Thematic analysis of documents is often carried out with probabilistic models where text is summarized as a bag of words. A popular approach is to use topic models (Blei

[*] This research was supported by Zillow and Sciome. The authors thank Kaoru Irie and Ruchir Shah for helpful comments.

[†] Paul College of Business and Economics, University of New Hampshire, Durham, NH, christopher.glynn@unh.edu

[‡] Department of Statistical Science, Duke University, Durham, NC 27708, tokdar@stat.duke.edu

[§] Sciome, LLC., Research Triangle Park, NC, brian.howard@sciome.com

[¶] Department of Statistical Science, Duke University, Durham, NC, banks@stat.duke.edu

https://doi.org/10.1214/18-BA1100

et al., 2003), which are probabilistic descriptions of word frequencies in documents based on a multinomial model with latent factors. Each latent factor, identified as a topic, is an unknown probability vector over the vocabulary. The three primary tasks in a topic model analysis are to (i) learn the topics themselves; (ii) learn the proportional contribution of each topic to each document; and (iii) learn the proportional contribution of each topic to the corpus as a whole.

We develop a novel class of topic models for time indexed corpora by combining the multinomial latent factor model of Blei et al. (2003) with the extremely versatile dynamic linear models (West and Harrison, 1997) widely used in the time series literature. Our dynamic linear topic model (DLTM) greatly expands the dynamic topic model (DTM) of Blei and Lafferty (2006) by allowing the marginal probability of topics to exhibit a rich set of temporal behavior including seasonal patterns and polynomial trends.

The ability of DLTM to model complex dynamics is demonstrated with an analysis of real estate listings from Zillow, spanning 2007–2017. The Zillow corpus calls for a composite dynamic model that allows the marginal probability of topics to generally trend upward (or downward) over time **and** periodically rise and fall with different seasons of the calendar year. Seasonality in the corpus arises as home descriptions emphasize the features most attractive to buyers at the time of year the home is listed for sale. The analysis demonstrates that real estate listings can yield fundamental insights into dynamic housing market behavior.

We adopt a fully Bayesian implementation of DLTM and make equal contributions to: (i) setting up the model and prior for sharp topic identification; and (ii) designing efficient Markov chain Monte Carlo algorithms for reproducible parameter estimation. Our approach helps demystify the often black-box-like appearance of topic models, and it offers a Bayesian modeling platform that can entertain a range of prior beliefs on easy to understand quantities pertaining to the thematic composition of a corpus, such as topic and document overlap. Further, we investigate the implications of these prior choices on the concentration of posterior mass on dynamic topics that minimally overlap in their themes. One particularly important consideration is the interplay between the prior variances of parameters associated with topics and document-specific topic proportions.

The fully Bayesian model platform requires a novel estimation strategy, and we develop an MCMC inferential algorithm for DLTM that seamlessly combines topic models with dynamic linear models. MCMC sampling from the full posterior provides reliable uncertainty quantification on both topic forms and prevalence, and it also enables computation of the widely applicable information criterion (WAIC, Watanabe (2013)), which is practically useful in model selection. These are significant advantages over the variational strategy utilized by Blei and Lafferty (2006) to fit the DTM. We demonstrate that the posterior probability of a topic's prevalence, WAIC, and MCMC reproducibility analyses can be combined to effectively select the number of topics in a corpus (Sections 6.3 and 7), an important open problem in the literature.

Our MCMC algorithm utilizes Pólya-Gamma data augmentation (Polson et al., 2013; Windle et al., 2013) to fuse a multinomial sampling model with Gaussian dynamic linear models. One challenge with Pólya-Gamma data augmentation is that sampling these random variates can be slow for parameter values pertinent to text analysis, making the MCMC algorithm prohibitively slow even for moderately large corpora. An important contribution of the paper is the derivation of a theoretically sound Gaussian approximation to the Pólya-Gamma distribution that allows rapid Pólya-Gamma sampling.

In Section 2, we describe the model; Section 3 constructs and examines the implications of prior distributions for topics and topic proportions; Section 4 details the Markov chain Monte Carlo algorithm for posterior sampling; Section 5 develops an approximate Pólya-Gamma sampler that scales well for text analysis; Section 6 examines the performance of our computational algorithm on a synthetic data set; Section 7 presents a case study with Zillow listings where we demonstrate a dynamic model composition strategy to infer a combination of seasonal patterns and locally linear trends; Section 8 concludes.

2 Model

DLTM is built on a user specified number of topics in the corpus, denoted K. The choice of K, an important open problem in topic models, is discussed in Sections 6.3 and 7. We also require a user specified vocabulary of length V that neither expands nor contracts with time. Each element of the vocabulary is a term, indexed by $v \in \{1, \ldots, V\}$. In situations where the vocabulary may be expected to evolve with time, one can take V to be the union of vocabularies used across time.

At each time point, $t \in \{1, \ldots, T\}$, the corpus contains D_t documents with $d \in \{1, \ldots, D_t\}$ indexing the documents. A document itself, $W_{d,t}$, is a vector where each entry in $W_{d,t}$ corresponds to a word in the document. The entries of document $W_{d,t}$ are denoted by $w_{n,d,t}$, which corresponds to the n^{th} word in the d^{th} document at time t. Each document has its own length of $N_{d,t}$ words, and the words within document $W_{d,t}$ are exchangeable (i.e. the index set $n \in \{1, \ldots, N_{d,t}\}$ can be permuted freely).

Each document is observed at a single time point. Documents themselves do not evolve over time. Only topics and the global topic proportions evolve in time. Despite the static nature of a document, we index documents with t to make clear the membership of each document in a specific time-slice.

2.1 A dynamic latent factor model for documents

One interpretation of DLTM is that it's a dynamic latent factor model for documents. Similar to latent Dirichlet allocation (LDA), DLTM decomposes documents into contributions from latent topics (factors) with document-specific topic proportions (factor loadings). The generative model for a corpus of time indexed documents is described below.

The latent topic associated with the n^{th} word in the d^{th} document at time t is denoted by $z_{n,d,t}$. Conditional on the latent topic variable $z_{n,d,t}$, the word $w_{n,d,t}$ is sampled from a multinomial distribution over the vocabulary,

$$Pr(w_{n,d,t} = v | z_{n,d,t} = k) = \frac{e^{\beta_{k,v,t}}}{\sum_{j=1}^{V} e^{\beta_{k,j,t}}}, \tag{1}$$

where $v \in \{1, \ldots, V\}$. The $\beta_{k,v,t}$ parameter is the natural parameter associated with the v^{th} vocabulary term under the k^{th} topic. Formally, this k^{th} topic is a probability distribution over the V terms in the vocabulary at time t. For the purpose of identifiability, we fix $\beta_{k,V,t} = 0$ for each topic k and time t. Following Blei and Lafferty (2006), the evolution in time of the natural parameter $\beta_{k,v,t}$ is modeled with a random walk:

$$\beta_{k,\cdot,t} = \beta_{k,\cdot,t-1} + \nu_{k,\cdot,t}, \qquad \nu_{k,\cdot,t} \sim N_V(0, \sigma^2 I), \tag{2}$$

$$\beta_{k,\cdot,0} \sim N_V(m_{k,0}, \sigma^2_{k,0} I). \tag{3}$$

The error terms $\nu_{k,\cdot,t}$ are mutually independent: $\nu_{k,\cdot,t} \perp\!\!\!\perp \nu_{k,\cdot,t'}$ for $t \neq t'$, and $\nu_{k,\cdot,t} \perp\!\!\!\perp \nu_{k',\cdot,t}$ for $k \neq k'$. Throughout the paper, when an index is omitted and replaced with $\cdot$, this notation signifies the collection of all elements of the omitted index. As an example, $\beta_{k,\cdot,t} = (\beta_{k,1,t}, \ldots, \beta_{k,V,t})$.

The word-specific latent topic variable, $z_{n,d,t}$, is sampled from its own multinomial distribution conditional on the set of natural parameters $\eta_{d,\cdot,t}$,

$$Pr(z_{n,d,t} = k | \eta_{d,\cdot,t}) = \frac{e^{\eta_{d,k,t}}}{\sum_{j=1}^{K} e^{\eta_{d,j,t}}}. \tag{4}$$

For the purpose of identifiability, we fix $\eta_{d,K,t} = 0$. Thus far, the model described is identical to that of Blei and Lafferty (2006). Where our model deviates from the DTM is in how we model $\eta_{d,k,t}$. For each $k \in \{1, \ldots, K\}$, we model the vector $\eta_{\cdot,k,t} = \{\eta_{1,k,t}, \eta_{2,k,t}, \ldots, \eta_{D_t,k,t}\}$ with a dynamic linear model (DLM) (West and Harrison, 1997). It is with this DLM that we incorporate periodic and polynomial behavior, covariates, and more broadly, an extensive set of features for temporal dependence in topic proportions:

$$\eta_{\cdot,k,t} = F_{k,t}\alpha_{k,t} + \epsilon_{k,t}, \qquad \epsilon_{k,t} \sim N_{D_t}(0, a^2 I_{D_t}), \tag{5}$$

$$\alpha_{k,t} = G_{k,t}\alpha_{k,t-1} + \xi_{k,t}, \qquad \xi_{k,t} \sim N_p(0, \delta^2 I_p), \tag{6}$$

$$\alpha_{k,0} \sim N(m_{k,0}, C_{k,0}). \tag{7}$$

The error terms are mutually independent: $\epsilon_{k,t} \perp\!\!\!\perp \epsilon_{k',t}$ for $k \neq k'$ and $\epsilon_{k,t} \perp\!\!\!\perp \epsilon_{k,t'}$ for $t \neq t'$. This independence statement implies the K distinct DLMs are mutually independent as well. The integer constant p is the dimension of the underlying state-vector, $\alpha_{k,t}$; $F_{k,t} = (F'_{1,k,t}, \ldots, F'_{D_t,k,t})'$ is a known $D_t \times p$ time-varying design-matrix of document covariates and model component terms corresponding to seasonality, trend, etc; and $G_{k,t}$ is a known $p \times p$ system matrix. Observe that the DTM is a special case of DLTM. The DTM can be recovered by fixing $p = 1$ and $F_{k,t} = G_{k,t} = 1$.

As previously noted, DLTM is a dynamic latent factor model for documents. As with all factor models, identifiability is a challenge. It is possible to find equivalent decompositions of a document by rotating topics (factors) and adjusting the document-specific topic proportions (loadings). We weakly identify the topics and proportional contributions by eliciting prior distributions that favor more distinct topics and documents. These priors will be discussed more in Section 3.

2.2 Likelihood

The likelihood of an entire corpus is computed in (8)–(10) by taking advantage of conditional independencies, which are encoded in the graphical representation of DLTM in Section 1 of the supplement (Glynn et al., 2019). For succinct notation, we let $W_{\cdot,t} = \{W_{1,t}, W_{2,t}, \ldots, W_{D_t,t}\}$ and $W_{\cdot,1:T} = \{W_{\cdot,1}, \ldots, W_{\cdot,T}\}$.

$$p(W_{\cdot,1:T}|Z_{\cdot,1:T}, \alpha_{\cdot,1:T}, \beta_{\cdot,\cdot,1:T}) = \prod_{t=1}^{T} p(W_{\cdot,t}|Z_{\cdot,t}, \alpha_{\cdot,t}, \beta_{\cdot,\cdot,t}) \tag{8}$$

$$= \prod_{t=1}^{T}\prod_{d=1}^{D_t} p(W_{d,t}|Z_{d,t}, \alpha_{\cdot,t}, \beta_{\cdot,\cdot,t}) = \prod_{t=1}^{T}\prod_{d=1}^{D_t}\prod_{n=1}^{N_{d,t}} p(w_{n,d,t}|z_{n,d,t}, \alpha_{\cdot,t}, \beta_{\cdot,\cdot,t}) \tag{9}$$

$$\propto \prod_{t=1}^{T}\prod_{d=1}^{D_t}\prod_{n=1}^{N_{d,t}} \left(\frac{e^{\beta_{z_{n,d,t},1,t}}}{\sum_{j=1}^{V} e^{\beta_{z_{n,d,t},j,t}}}\right)^{\mathbb{1}_{\{w_{n,d,t}=1\}}} \cdots \left(\frac{e^{\beta_{z_{n,d,t},V,t}}}{\sum_{j=1}^{V} e^{\beta_{z_{n,d,t},j,t}}}\right)^{\mathbb{1}_{\{w_{n,d,t}=V\}}} \tag{10}$$

It is useful to examine the likelihood contribution from a specific topic. The objective is to demonstrate that the multinomial likelihood can be reparameterized to one that is proportional to a binomial likelihood if we condition on a specific topic. Proportionality to the binomial likelihood is of interest from a computational perspective, as it makes MCMC simulation with Pólya-Gamma data augmentation (Polson et al., 2013) possible.

If we condition on $z_{n,d,t} = k$, the conditional likelihood is proportional to

$$\ell(\beta_{k,t}|z_{d,n,t} = k) \propto \left(\frac{e^{\beta_{k,1,t}}}{\sum_{j=1}^{V} e^{\beta_{k,j,t}}}\right)^{y_{k,1,t}} \cdots \left(\frac{e^{\beta_{k,V,t}}}{\sum_{j=1}^{V} e^{\beta_{k,j,t}}}\right)^{y_{k,V,t}}, \tag{11}$$

where $y_{k,v,t} = \sum_{d=1}^{D_t}\sum_{n=1}^{N_{d,t}} \mathbb{1}_{\{w_{n,d,t}=v\}}\mathbb{1}_{\{z_{n,d,t}=k\}}$. In (11), $y_{k,v,t}$ is the number of times the vocabulary term v is assigned to topic k across all documents at time t. We reparameterize the likelihood following the strategy of Holmes and Held (2006).

$$\ell(\beta_{k,t}|Z_t = k) \propto \left(\frac{e^{\gamma_{k,v,t}}}{1 + e^{\gamma_{k,v,t}}}\right)^{y_{k,v,t}} \left(\frac{1}{1 + e^{\gamma_{k,v,t}}}\right)^{n^y_{k,t} - y_{k,v,t}} \tag{12}$$

In (12), $\gamma_{k,v,t} = \beta_{k,v,t} - \log \sum_{j \neq v} e^{\beta_{k,j,t}}$ and $n^y_{k,t} = \sum_{j=1}^{V} y_{k,j,t}$ is the total number of words assigned to topic k at time t.

Note that the form of the conditional likelihood in (12) is now proportional to the binomial likelihood. This allows us to proceed with a Gibbs sampling algorithm using Pólya-Gamma data augmentation as outlined in Section 4. For a full derivation of the likelihood conditioning and reparameterization strategy, refer to Section 3 of the supplement.

3 Prior distributions

3.1 Identifiability challenges

Topic models like LDA, the dynamic topic model, and DLTM are not identifiable statistical models. It is possible that multiple different sets of topics and corresponding document-topic proportions result in identical marginal likelihoods for a corpus. Because the DLTM is not identifiable and the likelihood function over topics and topic proportions is multimodal, it is necessary to consider the role that prior choices for $\beta_{k,v,t}$, $\alpha_{k,t}$ and $\eta_{d,k,t}$ play in guiding posterior inference toward modes that are most interpretable and useful to the modeler.

A growing body of literature outlines the identifiability challenge in topic models (see, e.g., Anandkumar et al. (2013); Huang et al. (2016)), proposing various conditions and constraints on document-term and topic-term matrices that guarantee a unique solution to the fundamental non-negative matrix factorization problem. One approach is to "anchor" each topic with a key word that can only occur in that topic and may not appear in any other, leading to a separable non-negative matrix factorization model (Donoho and Stodden, 2004). As noted in Huang et al. (2016), estimation in anchored topic models is performed with either linear programming (Recht et al., 2012; Gillis, 2013) or greedy pursuit algorithms (Gillis, 2014; Gillis and Vavasis, 2014; Kumar et al., 2013; Arora et al., 2013). When the anchor word assumption is not appropriate, it is possible to show that inferences are identifiable if moment conditions of the corpus itself are satisfied (Liu et al., 2012; Anandkumar et al., 2012, 2013; Huang et al., 2016). Without anchor words or moment conditions on the corpus, identifiability is not guaranteed. In the presence of multiple corpus decompositions that result in the same likelihood, prior distributions that favor distinct topics are important. We show in Section 3.2 that the overlap between a pair of topics can be quantified by total variation distance and that the induced prior distribution for pairwise topic overlap is a function of the variance hyperparameter of $\beta_{k,v,0}$.

3.2 Priors for topics

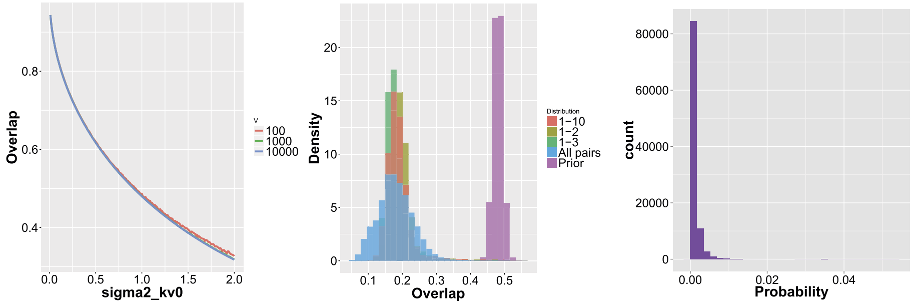
We define a metric to examine the overlap between two topics to guide elicitation of topic priors. The overlap metric is the complement of the total variation (TV) distance between topic k at time s and topic j at time t:

$$\text{Overlap}_{k,j}^{top}(s,t) = 1 - \frac{1}{2}\sum_{v=1}^{V}\left|\frac{e^{\beta_{k,v,s}}}{\sum_{v'=1}^{V} e^{\beta_{k,v',s}}} - \frac{e^{\beta_{j,v,t}}}{\sum_{v'=1}^{V} e^{\beta_{j,v',t}}}\right|. \tag{13}$$

Overlap between two topics is defined on the interval $[0, 1]$. When the overlap between two topics is zero, the topics are completely different. When overlap is one, the topics are exactly the same. In (13), $\frac{e^{\beta_{k,v,s}}}{\sum_{v'=1}^{V} e^{\beta_{k,v',s}}}$ is the marginal prior probability of vocabulary term v in topic k at time s. The variance of $\beta_{k,v,0}$, $\sigma^2_{k,v,0}$ in (3), is the primary driver of expected overlap of topics at time zero. Figure 1(a) illustrates that for two topics, k and j, both at time 0, the expected topic overlap $E[\text{Overlap}_{k,j}^{top}(0,0)]$ decreases as the variance $\sigma^2_{k,v,0}$ increases.

We recommend that topic modelers specify $\sigma^2_{k,v,0}$ by setting an expected level of overlap appropriate for the corpus under study. In the Zillow corpus, we expect that topics will be composed of common terms (e.g., home, bedroom, bathroom) and terms unique to each topic (e.g., fireplace, deck, stainless steel). As a result, we specify the expected overlap between any two topics to be $\approx \frac{1}{2}$, which implies that $\sigma^2_{k,v,0} \approx 1$ (see Figure 1(a)). Our prior for the natural parameter associated with term v in topic k is then $\beta_{k,v,0} \sim N(0,1)$. By centering the $\beta_{k,v,0}$ prior at zero, we do not favor any particular vocabulary term as being a keyword in the topic. We allow the data to inform which words are keywords, resulting in topics that load on distinct sets of terms. This feature of the prior is demonstrated in Figure 1(b), where the posterior distributions for pairwise overlap between topics concentrate on less overlap than expected under the prior. The prior is sufficiently diffuse – resulting in unique topics – and the data are able to inform the amount of overlap between these well-separated topics.

As a sensitivity check to our choice of $\sigma^2_{k,v,0} = 1$, we increased the prior variance to $\sigma^2_{k,v,0} = 5$ and re-ran our analysis of the Zillow corpus. The posterior topics, document proportions, and dynamic patterns were nearly identical. We advise modelers to begin their analyses by choosing a value for $\sigma^2_{k,v,0}$ that is consistent with their application-specific prior expectation for topic overlap. As a follow-on, it is important for the modeler to verify that the prior is not too concentrated and that the corpus has informed the amount of posterior overlap between topics. Robust results to sensitivity analyses with larger values for $\sigma^2_{k,v,0}$ and demonstrated Bayesian learning in topic overlap will help the modeler determine the most appropriate value of $\sigma^2_{k,v,0}$.



(a) $E[\text{Overlap}^{top}_{k,j}(0,0)]$ (b) Posterior overlap (c) Term prob.

Figure 1: Left: Prior for expected topic overlap between two topics at the same time, $E[\text{Overlap}^{top}_{k,j}(0,0)]$, as a function of hyperparameter $\sigma^2_{k,v,0}$ when $V = 100, 1000, 10000$. Middle: Prior distribution of overlap between two topics at $t = 1$ when $\sigma^2_{k,v,0} = 1$ and posterior distributions of pairwise topic overlap in the Zillow analysis of Section 7 for topics 1 and 2, 1 and 3, 1 and 10, and all pairwise combinations at $t = 1$. Right: Marginal prior probability of an arbitrary vocabulary term appearing in a document, $P(w_{n,d,1} = v | z_{n,d,1} = k)$.

In addition to specifying the time-zero variance hyperparameter, we must also specify the topic innovation variance in (2). We want the expected change in a single topic from t to $t+1$ to be small so that time-varying marginal probabilities of topics (4) are inferred sharply. To specify the innovation variance of the $\beta_{k,v,t}$ process, σ^2, we examine the expected overlap of topic k at t and $t+1$. To model slowly evolving topic forms from $t \rightarrow t+1$, we choose $\sigma^2 = 0.01$. It is important to note that small month-over-month changes in topic form accumulate so that the overlap in topic k at $t = 1$ and $t = T$ is modest. When $\sigma^2 = 0.01$ and the topics evolve for 112 months, the expected overlap between topic k at $t = 1$ and $t = 112$ is ≈ 0.6, allowing the topic to evolve while still preserving elements of its original identity.

Beyond the prior on topic overlap, it is useful to examine the role of $\beta_{k,v,t}$ in the marginal prior probability of observing individual vocabulary terms. To fully assess the uncertainty in the prior distribution for the prevalence of the v^{th} term at time t, $\frac{e^{\beta_{k,v,t}}}{\sum_{j=1}^{V} e^{\beta_{k,j,t}}}$, it is necessary to consider the prior uncertainty for the remaining $V - 1$ terms. Figure 1(c) presents a histogram of samples from the marginal prior distribution for the prevalence of any given vocabulary term v when there are $V = 1000$ terms in the vocabulary (see (1)). Observe that significant prior mass is distributed on the interval $(0, .01]$, which is quite diffuse considering that the uniform distribution of mass over each term in the topic implies a naive term probability of $\frac{1}{V} = 0.001$.

3.3 Priors for document topic proportions

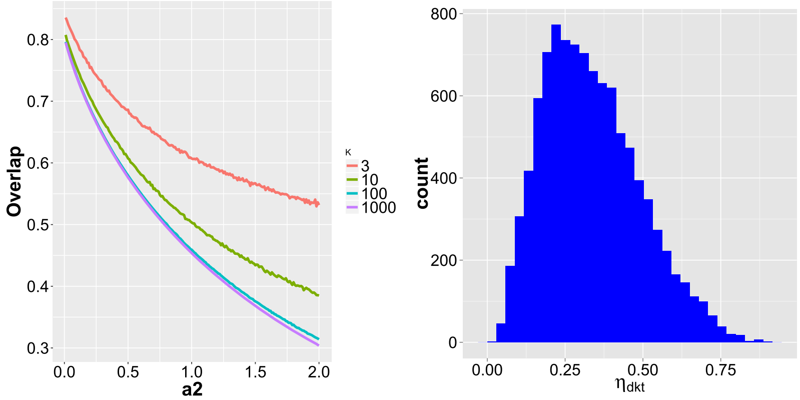
We fix our attention in this subsection to the special case when DTM and DLTM are equivalent (i.e., $F_* = G_* = 1$ in (5) and (6)) to focus the discussion of priors for document-specific topic proportions. As in Section 3.2, we construct a metric for the similarity of documents to guide our choice of hyperparameters a^2 and δ^2 in the DLM model of $\eta\cdot, k, t$. Document overlap is defined as the complement of total variation distance between the vector of topic proportions for two documents d and d' at time t,

$$\text{Overlap}_{d,d'}^{doc} = 1 - \frac{1}{2} \sum_{k=1}^{K} \left| \frac{e^{\eta_{d,k,t}}}{\sum_{j=1}^{K} e^{\eta_{d,j,t}}} - \frac{e^{\eta_{d',k,t}}}{\sum_{j=1}^{K} e^{\eta_{d',j,t}}} \right| . \tag{14}$$

When documents d and d' receive the same proportional contributions from each topic, $\text{Overlap}_{d,d'}^{doc} = 1$. Figure 2(a) illustrates that $E[\text{Overlap}_{d,d'}^{doc}]$ depends on a^2, the variance term in (5), and the number of topics, K. Our strategy is to choose a^2 based on the level of document overlap we expect in the corpus. In the Zillow listings data, we expect that approximately two-thirds of each document will focus on common elements of a home such as descriptions of the kitchen, bedrooms, and bathrooms. We believe one third of each listing will focus on unique features of the home. To achieve an expected overlap of $\approx \frac{2}{3}$ when $K = 10$, we fix $a^2 = 0.25$.

In order to choose δ^2, the variance of the DLM state innovations in (6), we utilize the signal-to-noise ratio (SNR). The SNR may be interpreted as the degree to which the thematic composition of any single document is representative of the corpus as a whole. An important consequence of low SNR is that the proportional contribution of topic k to a document is modeled as a noisy reflection of topic k's prevalence globally in the corpus. In our analysis, we assume that the SNR is $\frac{1}{10}$. In the DLM context, the SNR implies that $\frac{\delta^2}{a^2} = \frac{1}{10}$. Note that lower SNR implies less document overlap. In the synthetic experiments of Section 6, we let $\delta^2 = 0.025$.

The DLM model specification of (5)–(7) is completed with the time-zero prior for $\alpha_{k,0}$. The $\alpha_{k,t}$ state variables govern the expected topic proportion of topic k at time t.



(a) Prior for $E[\text{Overlap}^{doc}_{d,d'}]$ (b) Prior for $P(z_{n,d,1} = k)$

Figure 2: Left: Prior for expected document overlap as a function of variance term a^2. Right: Prior distribution for the proportional contribution of an arbitrary topic to the corpus as a whole when $K = 3$ and $t = 1$.

As in LDA, the expectation of our prior distribution is that documents are an equal mixture of K topics; however, we want our prior to include significant uncertainty regarding *global* topic proportions. To ensure that a wide range of global topic proportions receive prior support, we assume that $\alpha_{k,0} \sim N(0, 0.1)$. Again, simply considering the uncertainty in $\eta_{d,k,1}$ is insufficient for examining the uncertainty on the simplex of document topic proportions. It is necessary to consider the uncertainty in the remaining $K-1$ natural parameters, $\eta_{d,-k,1}$. When $K = 3$, our prior supports document topic proportions ranging from near zero to those exceeding 0.8, as shown in Figure 2(b). The histogram illustrates the marginal distribution of $\frac{e^{\eta_{d,k,t}}}{\sum_{j=1}^{K} e^{\eta_{d,j,t}}}$, the proportional contribution that an arbitrary topic k makes to an arbitrary document d when $K = 3$. The important takeaway is that this prior specification supports a wide range of document topic proportions. Fixing hyperparameters in the manner outlined above facilitates robust learning of latent topics and their dynamic proportional representation in the corpus while still eliciting diffuse prior distributions on those same factors and loadings.

3.4 Interplay between variance components

Whenever specifying the number of topics (K) and vocabulary size (V) in a topic model, we advise analyzing the uncertainty on the probability simplices being modeled before attempting to make inference on topics or document proportions. There is a subtle but important interplay in the variance hyperparameters due to their importance in controlling topic and document overlap. If topic overlap induced by the prior for $\beta_{k,v,t}$ is too high (i.e, if $\sigma^2_{k,v,0}$ is too small), the posterior topic overlap will also be quite high. If K nearly identical topics are estimated, the corresponding estimates of document-specific topic proportions are not meaningful. On the other hand, if document overlap is too high – specifically if a^2 is too small – and documents are modeled as almost certain identical mixtures of K topics, the induced prior belief is that at each time point the corpus contains D_t nearly identical copies of the same document. The result is that the inference procedure learns a single repeated topic – corresponding to the single repeated document – in the corpus. Again, nothing has been learned. The variance components $\sigma^2_{k,v,0}$ and a^2 can be thought of as having a similar role to the concentration parameters in the Dirichlet distributions in LDA. As noted in Wallach et al. (2009), priors have an important effect on the ability of topic models to learn latent structure in documents.

Although we have focused here on the case where DLTM and DTM are equivalent, a similar process for eliciting priors in DLTM with polynomial or seasonal patterns should be followed. One point of consideration is the relationship between the different variance components and the choice of model components. Similar marginal distributions for $\eta_{d,k,t}$ can be achieved with different pairs of model components and choices for δ^2. As an example, the random walk model and the locally linear state space model for $\eta_{d,k,t}$ can imply similar marginal distributions if the choice of δ^2 is increased in the random walk specification. While these different models may imply similar marginal distributions, their *predictive* distributions will be quite different. As we show in Section 6.4, explicitly modeling polynomial components improves prediction of document-specific topic proportions when trends do exist. It is for this reason that simply specifying a random walk state space model with high variance δ^2 is not advisable.

4 Markov chain Monte Carlo

The target posterior distribution is

$$p(\alpha_{\cdot,1:T}, \eta_{\cdot,\cdot,1:T}, \beta_{\cdot,\cdot,1:T} | W_{\cdot,1:T}) = \int p(\alpha_{\cdot,1:T}, \eta_{\cdot,\cdot,1:T}, \beta_{\cdot,\cdot,1:T}, Z_{\cdot,1:T} | W_{\cdot,1:T}) dZ_{\cdot,1:T}. \quad (15)$$

Note that we are not interested in inferring $z_{n,d,t}$, and the topic assignment of each word in each document is marginalized out of (15).

The machine learning literature typically utilizes variational methods to fit various forms of topic models (Blei et al., 2003; Blei and Lafferty, 2006; Teh et al., 2007; Hoffman et al., 2010). The appeal of variational procedures is that they scale to meet the demands of massive corpora. Despite their scalability, these methods must be used with caution when fitting dynamic models. In time series analysis, variational methods suffer from a critical shortcoming: parameter uncertainty tends to decrease when the quality of the approximation degrades (Turner and Sahani, 2011). As a result, variational approximations often understate posterior uncertainty. Section 6.3 discusses a common situation in topic modeling when mean-field variational methods are inadequate.

We find that a hierarchical model of document topic proportions yields valuable information about the number of topics necessary to model a corpus. When hierarchical model structure and accurate quantification of parameter uncertainty are imperative, MCMC algorithms are an important part of the topic modelers computational toolbox. MCMC algorithms have been successfully deployed to fit both LDA (Griffiths and Steyvers, 2004) and static logistic Normal topic models (Chen et al., 2013). In this section, we develop an MCMC algorithm for DLTM with Pólya-Gamma data augmentation.

4.1 Forward filtering and backward sampling via data augmentation

Pólya-Gamma data augmentation allows a seamless combination of topic models and dynamic linear models. To make inference on the dynamic processes that govern topics and document topic proportions, we utilize forward filtering and backward sampling (FFBS). FFBS is the predominant MCMC technique for dynamic models (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994). As argued by Chib (1996), forward-backward algorithms are necessary to sample from complex dynamic state-space models. Because Pólya-Gamma data augmentation supports an FFBS algorithm, we favor this data augmentation scheme over alternative MCMC choices such as the adaptive rejection Gibbs sampler of Gilks and Wild (1992).

4.2 Pólya-Gamma Gibbs sampler

We construct a Gibbs sampler by iteratively sampling from the full conditionals derived in Section 6 of the supplement. In order to sample from these full conditionals, we utilize Pólya-Gamma data augmentation (Polson et al., 2013). Chen et al. (2013) introduced the idea of a Pólya-Gamma Gibbs sampler for a static logistic-Normal topic model. We extend this idea to the dynamic setting. In order to make inference for each $\beta_{k,v,t}$, it is necessary to introduce an auxiliary $\zeta_{k,v,t} \sim PG(n^y_{k,t}, 0)$. Additionally, in order to make inference on $\eta_{d,k,t}$, it is necessary to introduce the auxiliary random variable $\omega_{d,k,t} \sim PG(N_{d,t}, 0)$. Note that our target posterior (15) does not include the auxiliary variables. To marginalize out the auxiliary ζ and ω from the posterior, we simply discard them and only store $\alpha_{\cdot,1:T}$,$\eta_{\cdot,\cdot,1:T}$, and $\beta_{\cdot,\cdot,1:T}$. A single MCMC sample from the target posterior is constructed as follows:

1. Sample the natural parameter associated with the v^{th} vocabulary term in the k^{th} topic jointly for all times, $\beta_{k,v,1:T} \ |\beta_{k,-v,1:T}$, $W_{\cdot,1:T}$, $Z_{\cdot,1:T}$, $\zeta_{k,v,1:T}$. This step can be performed independently across topics. The order in which the $\beta_{k,v,1:T}$ are updated is randomly permuted in the index $\{1, \ldots, V\}$ at each MCMC iteration.

2. Sample the auxiliary Pólya-Gamma variable independently for each topic k, vocabulary term v, and time t, $\zeta_{k,v,t} \ |\gamma_{k,v,t}$.

3. Sample the natural parameter for the proportional contribution of topic k in document d at time t, $\eta_{d,k,t} \ |Z_{\cdot,t}, \eta_{d,-k,t}$, $\omega_{d,k,t}, \alpha_{k,t}$. This step can be performed independently across documents d and time t. The order in which the $\eta_{d,k,t}$ are updated is randomly permuted in the index $\{1, \ldots, K\}$ at each MCMC iteration.

4. Sample the auxiliary Pólya-Gamma variable independently for each document d, topic k, and time t, $\omega_{d,k,t}|\eta_{d,\cdot,t}$.

5. Sample the latent state variables for each topic k independently, $\alpha_{k,1:T}|\eta_{\cdot,k,1:T}$.

6. Sample the topic assignment for each word n, document d, and time t, independently $z_{n,d,t} \; |w_{n,d,t}, \; \eta_{d,\cdot,t}, \; \beta_{\cdot,\cdot,t}$.

Several steps in this sampling procedure are easily parallelized. Intuition behind the parallel sampling structure is derived from the graphical model of DLTM in Section 1 of the supplement. In particular, Steps 1 and 5 can be parallelized across topics. Step 3 can be parallelized across documents and time. Step 2 can be parallelized across vocabulary terms, topics, and time. Step 4 can be parallelized across documents, topics, and time. Step 6 can be parallelized across words in a document, documents, and time. Our implementation of this algorithm performs all sampling steps in C++ using the R–C++ interface, Rcpp. It also parallelizes these steps where possible using the mclapply function from the R-package parallel. The code is available at the website http://github.com/G-Lynn/DLTM. Although not implemented in the current version of the code, GPU and distributed computing architectures can be used for faster computation and inference in large corpora.

5 Scalable Pólya-Gamma sampling

A primary computational bottleneck is sampling from the Pólya-Gamma distribution. Each MCMC sample requires drawing $K(VT + \sum_{t=1}^{T} D_t)$ Pólya-Gamma random variables (Steps 2 and 4 in Section 4.2). In the Zillow listings corpus, with more than 11,000 documents spanning 112 months, 912 vocabulary terms, and $K = 10$ topics, each MCMC iteration requires sampling from the Pólya-Gamma distribution 1.1×10^6 times.

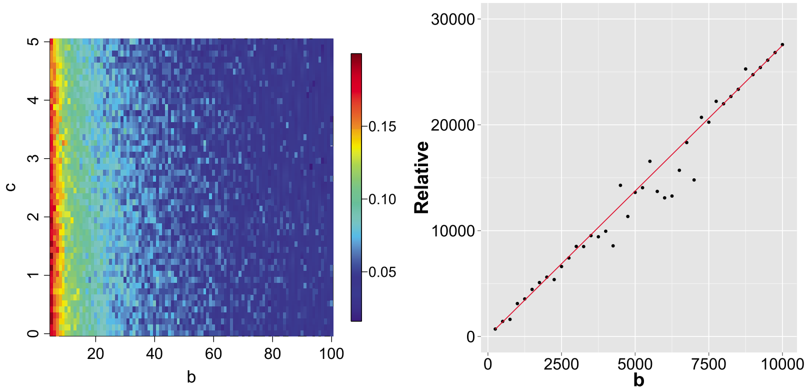
The practical challenge is to rapidly sample from the Pólya-Gamma distribution for a wide range of parameter values. Polson et al. (2013) develop a sampling method based on the work of Devroye (2009), and the BayesLogit R package provides an implementation of this exact sampling algorithm. While the Devroye method works extremely well for Pólya-Gamma sampling when count values are moderate, the practicality of this approach diminishes when counts are high. For the Pólya-Gamma random variates in text analysis, the b parameter is very large.

Section 4.4 of Polson et al. (2013) notes that sampling $\omega \sim PG(b, c)$ when $b \in \mathbb{N}$ is equivalent to the construction $\omega = \sum_{i=1}^{b} \tilde{\omega}_i$, where $\tilde{\omega}_i \sim PG(1, c)$. Because we are focused on counts, the parameter b will always be a natural number. For the Devroye method, the computation time increases with b, as sampling from $PG(b, c)$ requires sampling the b underlying $PG(1, c)$ variables to construct each draw. This additive sampling process imposes a significant limitation on the computational speed.

Several approximate methods for sampling from the Pólya-Gamma distribution exist. To approximately sample from the distribution $PG(b, c)$ when b is large, Chen et al. (2013) propose to sample $z \sim PG(m, c)$ where $m << b$ and then linearly transform z to match the moments of $PG(b, c)$. As discussed in Windle et al. (2014), a saddlepoint approximation is another method for approximate sampling from $PG(b, c)$ when b is large. The saddplepoint approximation constructs a piecewise linear envelope around the true density to approximately sample from $PG(b, c)$.

Another possible approximation suggested by Windle et al. (2014), though not explored in the technical report, is based on the Central Limit Theorem. To approximate the sampling of a $PG(b, c)$ draw when $b \in \mathbb{N}$, we rely on the additive construction of a $PG(b, c)$ random variable from $\omega_i \sim PG(1, c)$ variates. The Central Limit Theorem provides that $\sqrt{b}((\frac{1}{b}\sum_{i=1}^{b}\tilde{\omega}_i) - E[\tilde{\omega}_i]) \overset{d}{\Rightarrow} N(0, Var(\tilde{\omega}_i))$, which suggests that $\omega = \sum_{i=1}^{b}\tilde{\omega}_i \overset{d}{\approx} N(bE[\tilde{\omega}_i], bVar(\tilde{\omega}_i))$ for large values of b. The mean and variance of the approximating Normal distribution are the appropriately scaled mean and variance of $\tilde{\omega}_i \sim PG(1, c)$.

The advantage of the approximation developed here is that sampling from a Pólya-Gamma distribution is not necessary at all. Rather than sampling the b underlying $PG(1, c)$ distributions, it is possible to generate an approximate draw from a $PG(b, c)$ distribution with a single draw from an approximating Gaussian. This removes the problem of additivity altogether for sufficiently large values of b. The remainder of this section examines the reliability of the Gaussian approximation for different choices of b and c. To consider the quality of the approximation for different parameter choices, we estimate the total variation distance between the Gaussian approximation and the kernel density estimate from a large sample of Pólya-Gamma draws. Section 2 of the supplement presents density overlays of the Gaussian approximation and the Pólya-Gamma distribution for several instances.



(a) TV distance

(b) Sampling time

Figure 3: Left: TV distance between true distribution and approximation over a grid of different choices for b and c. Right: Relative sampling time for the Devroye sampling method in BayesLogit compared to the Gaussian approximation for 1000 draws at different b values. For each value of b, $c_i \sim N(0, 1)$ for $i = 1, \ldots, 1000$. Red line is the linear relationship $2.75 \times b$.

Figure 3(a) presents the TV distance between the approximation and the Pólya-Gamma distribution as b and c vary jointly. Observe that for low values of b, the approximation is unreliable. Also note that for sufficiently large values of b, the profile of the TV distance is uniform across c. Figure 3(a) provides evidence that the Gaussian approximation works well for a wide range of choices for both b and c. In our MCMC simulation, if $b < 20$, we utilize exact Pólya-Gamma sampling techniques to minimize approximation error. This approach balances the tradeoff between computational speed and accuracy of the approximation.

To examine the scalability of the approximate sampling algorithm, the relative sampling times for the Gaussian approximation and Devroye method are compared as a function of increasing b using the *benchmark* function in R. Figure 3(b) demonstrates that relative sampling time increases linearly with b. The relative computation time from the approximate sampling algorithms of Chen et al. (2013) and Windle et al. (2013) are on the same order of magnitude as the Gaussian approximation. The advantage of the Gaussian approximation is that it has theoretical guarantees from the Central Limit Theorem.

The reliability of the approximation and the reduction in computation time significantly increase with b. The Gaussian approximation is scalable and safe when counts are large, which they are in most corpora. Further, the approximation method makes Pólya-Gamma data augmentation a practical simulation tool in text analysis, where it was previously infeasible.

6 Simulation study

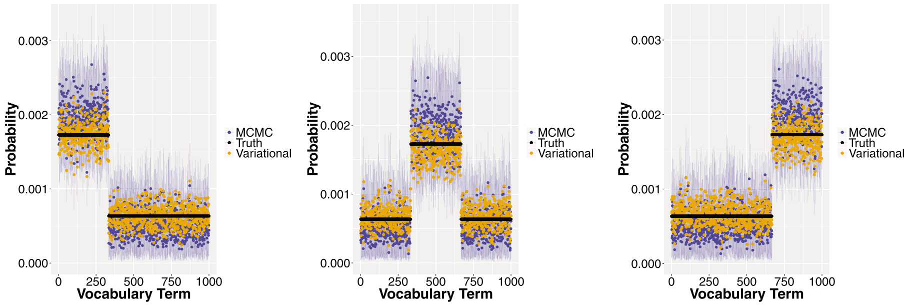
We conducted a simulation study to validate and examine the reproducibility of our MCMC algorithm. We constructed a synthetic data set with $K = 3$ topics and a vocabulary with $V = 1000$ terms. The objective of the simulation study is to benchmark the computational method in recovering a known truth as compared to existing variational strategies for inference in the DTM (Gerrish and Blei, 2011). In Sections 6.1–6.3, the differences between DLTM and DTM are due to the differences in computational strategies, as we focus on the special case when DLTM and DTM are equivalent (discussed at the end of Section 2.1). In Section 6.4, the underlying dynamic models are different (e.g., random walk for DTM and dynamic linear model for DLTM), but MCMC is used to fit all models.

6.1 Synthetic data

The synthetic data set was constructed by sampling a random number of documents at $T = 5$ different time points. The number of documents at each time point was sampled from a Poisson distribution with mean of 1000. Each document was endowed with a random number of words, which was sampled from a Poisson distribution with mean 150.

The proportional contribution of the three topics to each document was generated by sampling from the DTM data generating model for document proportions with $m_{k,0} = 0$, $C_{k,0} = 0.025$, $\delta^2 = 0.001$, and $a^2 = 0.5$. Setting $\delta^2 = 0.001$ makes it likely that there is no overall trend in topic proportions. Setting $a^2 = 0.5$ ensures heterogeneity in documents.

The three topics were constructed so that three disjoint subsets of vocabulary terms would occur with high probability in each topic. This is visually demonstrated by the black lines in Figures 4(a)–4(c). The first topic places high probability on vocabulary



(a) Topic 1

(b) Topic 2

(c) Topic 3

Figure 4: Posterior means for probabilities of v^{th} term for each topic. The blue dots represent the posterior mean of the topic probability for each vocabulary term, as estimated by the MCMC algorithm. The light blue verticals associated with each blue dot represent the 95% posterior credible interval for the probability. The orange dots represent the mean estimate from the variational Kalman filter of Gerrish and Blei (2011), which is publicly available at the Blei Lab GitHub repository.

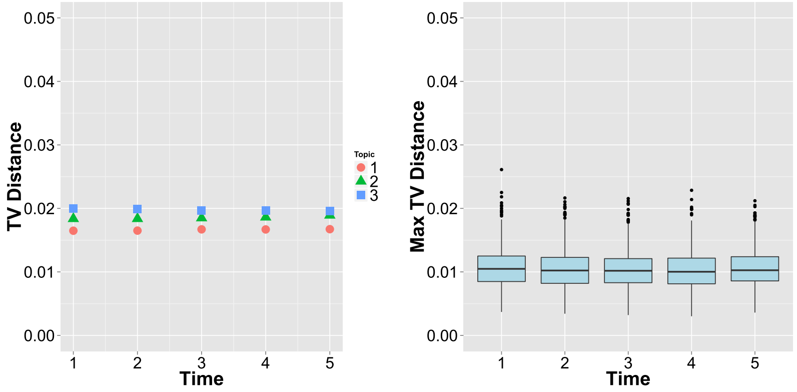
terms 1 through 333. The second topic places high probability on terms 334–667. The third topic places high probability on terms 668–1000. The true topics were allowed to evolve after $t = 1$ with an innovation variance of $\sigma^2 = .01$.

Both the MCMC algorithm and variational Kalman filter of Blei and Lafferty (2006) learn the simulated topics in the corpus. Figure 4 demonstrates that the posterior means for the probability of the v^{th} term in each topic correspond well to the true probability for both the MCMC and variational methods. Our MCMC computation of the full posterior distribution for each topic is emphasized with the light blue vertical lines in the figure. Importantly, the 95% posterior credible intervals from the MCMC simulation always include the true term probability. Further comparison of the computational methods is conducted in Section 4 of the supplement.

6.2 Reproducibility and MCMC convergence

The second objective of the simulation study is to examine the reproducibility and convergence of our MCMC algorithm. We ran five different MCMC simulations where the parameters in each run were initialized by sampling from the data generating model with high variance, providing overly disperse starting points. Each MCMC simulation was run for 600,000 iterations, and the chain was thinned by recording every 100^{th} sample, resulting in 6,000 MCMC samples. After thinning, we discarded 2,000 of the remaining samples as a burn-in period. On an 8 core workstation, this took approximately one day.

Topic inference is reproducible, as Figure 5(a) shows that the maximum TV distance between topics learned across the five simulations is small. The figure was constructed



(a) Topic TV

(b) Document % TV

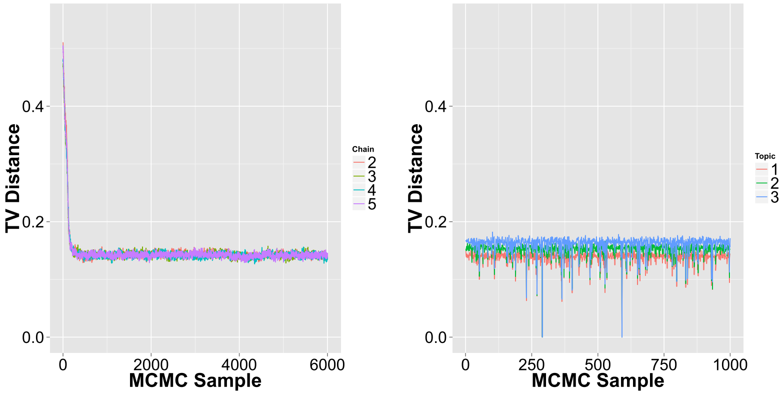
Figure 5: Left: Maximum TV distance between posterior means of topics from chains 2–5 and chain 1 for each topic. Right: Boxplot of maximum TV distances between posterior means of document topic proportions from chains 2–5 and chain 1 for all documents.

by computing TV distances between the posterior mean topics from chains two through five and chain one and then taking the maximum. The takeaway is that we learn the same topics across different MCMC simulations with very different starting points.

We also learn the same document topic proportions across the different simulations. The boxplot in 5 shows that the maximum TV distance between document-specific topic proportions learned across different simulations is small for all documents. For each document, we compute the TV distance between the posterior means for topic proportions from chains two through five and the posterior mean for topic proportions from chain one. We then take the maximum TV distance for each document. When taken together, Figures 5(a) and 5(b) confirm reproducible inference across MCMC runs.

While inferences across simulations are nearly identical, the question remains: have the chains converged? To investigate convergence, we explore (i) the TV distance between MCMC samples of topic k in chain one and topic k in chains two through five; and (ii) the TV distance between the q^{th} MCMC sample of topic k and the r^{th} MCMC sample of topic k in the same chain. Across and within chain comparison of topics provides insight into how efficiently the simulations are exploring the posterior. Essentially, we adapt MCMC convergence ideas from Gelman and Rubin (1992) for topic modeling. Rather than consider across chain and within chain variance, we consider across chain and within chain TV distance between topics.

Figure 6(a) illustrates that topic reproducibility across simulations is not caused by the chains simply mirroring each other as they move through parameter space. The TV distance between MCMC samples for topic one across the simulations is at least 0.15, approximately ten times larger than the max TV distance computed from the topic



(a) Across Chain TV

(b) Within Chain 1 TV

Figure 6: Left: Across Chain TV Distance to initialization 1 for topic 1. Right: Within Chain TV for each topic in initialization 1. The TV distances are between 1000 randomly selected pairs of post-burn-in posterior samples of topics.

means (Figure 5(a)). At each MCMC iteration, chains two through five are in different places than chain one.

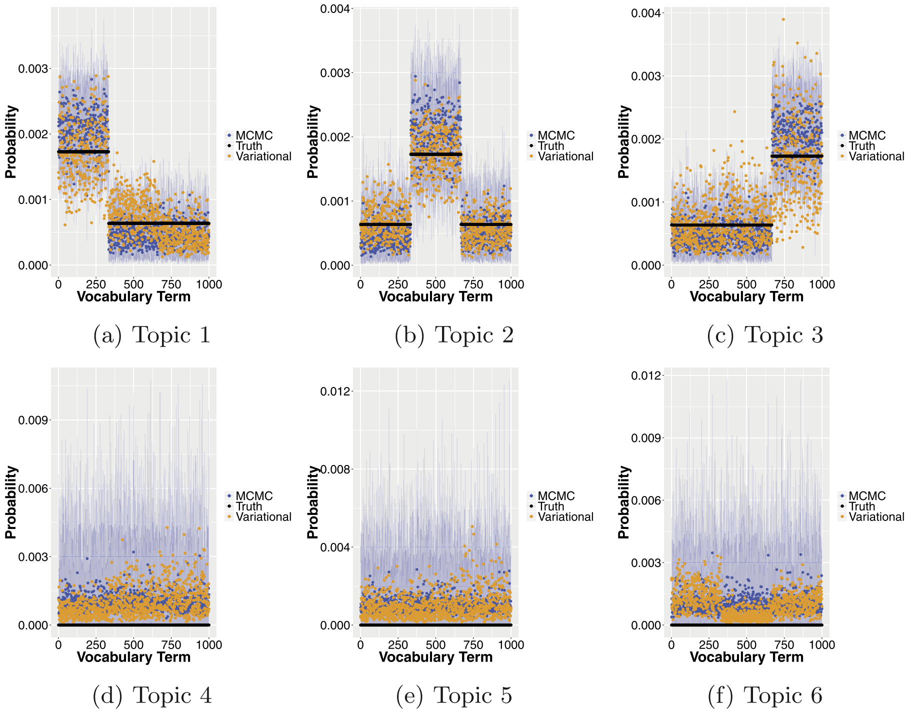
The Markov chain is also efficiently exploring the posterior. Figure 6(b) illustrates that the TV distance between different MCMC samples of the same topic in the same chain is between 0.15–0.2, again considerably larger than the values in Figure 5(a). The takeaway is that topic reproducibility across simulations (Figure 5(a)) is due to the convergence of the chains to the stationary distribution and not a fluke artifact of the MCMC simulations.

6.3 Misspecification of K

One of the critical questions for topic modeling is how to specify the number of topics in a corpus. Rarely is a true number of topics known to the researcher. For this reason, behavior of the model and computational algorithm when the number of topics is misspecified is of practical importance. In this section, we let $K = 6$ and repeat the analysis of our synthetic data.

Figure 7 presents the posterior means for Topics 1–3 and 4–6, respectively. Topics 1, 2 and 3 correspond to the three topics that generated the synthetic corpus. Topics 4, 5, and 6 are extraneous. The extremely wide uncertainty intervals for each vocabulary term in topics 4–6 suggest that there is no meaningful thematic differentiation among these topics. The uncertainty intervals are valuable for identifying that a topic lacks a distinct theme.

The variational DTM also recovers the three original topics. The advantage of DLTM is in estimating the document-specific topic proportions. By sharing information across



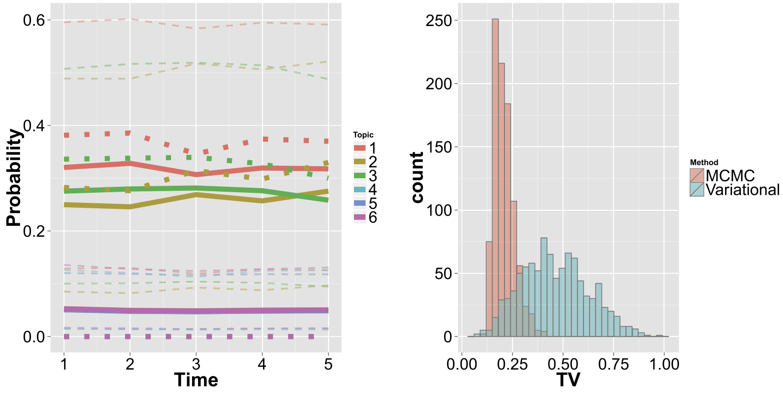
(a) Topic 1 (b) Topic 2 (c) Topic 3

(d) Topic 4 (e) Topic 5 (f) Topic 6

Figure 7: Posterior mean for v^{th} term in Topics 1–6.

documents with a hierarchical model for document topic proportions, DLTM infers which topics are unnecessary to model the corpus as a whole. This is evidenced by the inferred global proportions of topics 4, 5, and 6 in Figure 8(a), which concentrate on low probabilities. The lack of thematic differentiation and the low marginal probabilities indicate that topics 4, 5, and 6 are extraneous. Taken together, Figures 7 and 8 demonstrate that even if the researcher misspecifies the number of topics to be larger than the true number, DLTM will still identify the three topics actually present in the data.

The MCMC algorithm significantly outperforms the variational approximation when estimating document topic proportions (Figure 8(b)), an advantage of identifying that topics four through six are globally unnecessary. Because the variational algorithm approximates each document topic proportion with an independent Dirichlet distribution, the hierarchical structure between documents is not adequately preserved and extraneous topics in the corpus as a whole cannot be identified. The MCMC algorithm is able to more effectively share information across documents and identify extraneous topics, a form of shrinkage that improves estimation of document topic proportions in the corpus. The mean (median) TV distance between the estimated document-specific topic proportions and the truth is 0.20 (0.20) for the MCMC method of estimation. For the variational method of estimation, the mean (median) TV distance between the estimated document-specific topic proportions and the truth is 0.46 (0.44). The important



(a) Topic Proportions

(b) TV distances

Figure 8: Left: Marginal probability of topics over time. Right: TV distances between estimated document-specific topic proportions and true topic proportions for all documents at time $t = 5$.

point is that borrowing information across documents is essential when the number of topics is misspecified.

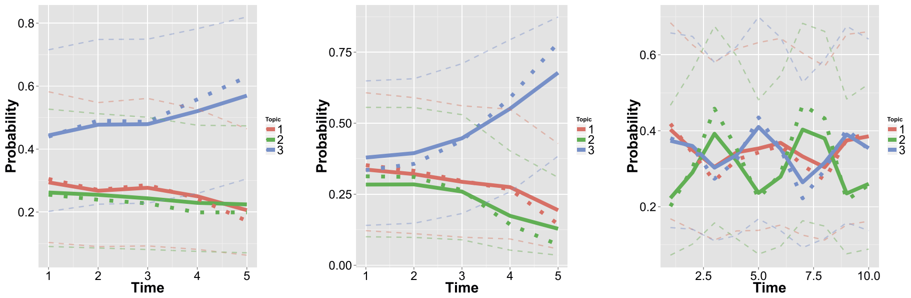
6.4 Linear, quadratic, and harmonic trends

In this section, we report simulation examples with linear, quadratic, and harmonic trends. We maintain the same topics as in previous examples but allow their marginal probabilities in the corpus to exhibit more complex dynamic behavior. In the linear case, $F_{k,t} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{bmatrix}$. The system matrix is $G_{k,t} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$. The harmonic case can be simulated by maintaining the $F_{k,t}$ matrix as in the linear case and replacing the system matrix with $G_{k,t} = \left[\begin{smallmatrix} \cos(\omega) & \sin(\omega) \\ -\sin(\omega) & \cos(\omega) \end{smallmatrix}\right]$. We let $\omega = \frac{\pi}{2}$ in our example. To simulate quadratic trends,

$$F_{k,t} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad G_{k,t} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

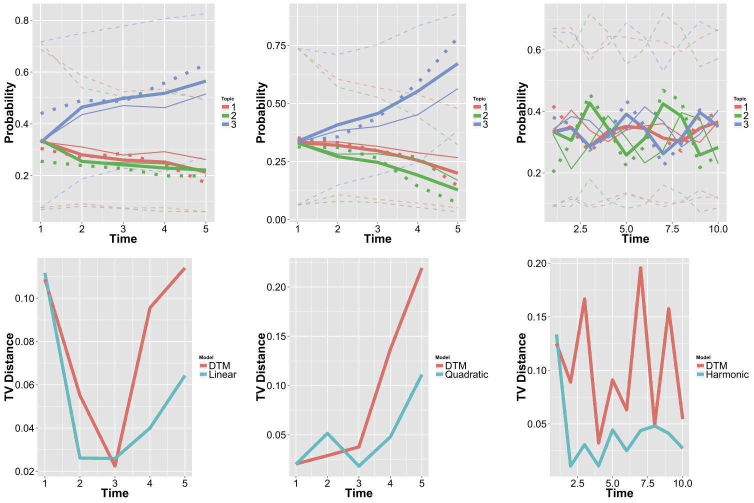
The topics recovered from these analyses and comparison with their variational counterparts are presented in Section 5 of the supplement. DLTM clearly recovers linear, quadratic, and seasonal patterns in the proportional contribution of each topic to the corpus (see Figure 9). Inference on such dynamics represents a significant advance over existing dynamic topic models.

Explicitly incorporating trend components is important for prediction as well as inference. Figure 10 presents the one-step-ahead predictions for topic proportions from DLTM models with structured dynamic components as compared to the baseline ran-



(a) Linear Trend (b) Quadratic Trend (c) Harmonic Trend

Figure 9: Marginal probability of topics over time. The solid lines are the posterior means. The light dashed lines are the 95% posterior credible intervals. The thick dotted lines are the true trends.



(a) Linear Trend (b) Quadratic Trend (c) Harmonic Trend

Figure 10: First row: One-step-ahead predictions for a DLTM with trend (thick solid line) and a random walk baseline (thin solid line) compared with the truth (thick dashed line). Light dashed lines are the 95% posterior credible intervals for the DLTM estimate. Second row: prediction errors for the DLTM with trend and random walk baseline, labeled DTM.

dom walk case. The one-step-ahead predictions are generated in the forward filtering step of the MCMC. Predictions from models with linear and quadratic components outperform their random walk counterpart as trends become more established. This outperformance is significant by the final time point. The outperformance of predictions from a harmonic model is significant throughout.

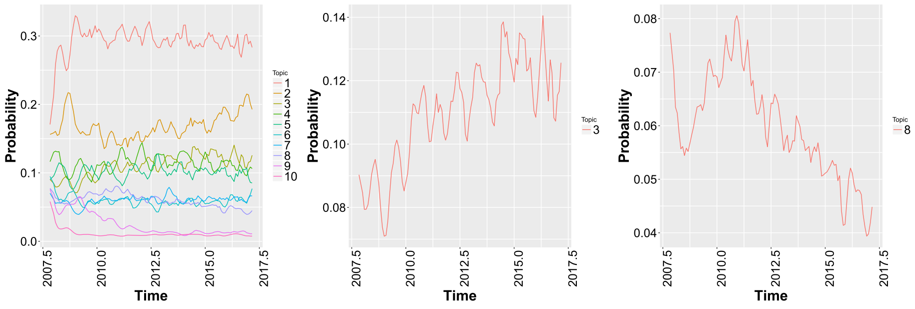
7 Zillow listings

We analyzed real estate listings from the housing website Zillow to demonstrate the utility of DLTM in a practical application. The Zillow corpus contains more than 11,000 home listings in Seattle, WA from 2007–2017. The vocabulary contains 912 distinct terms. An important aspect of real estate listings is that they emphasize features of a home that are most appealing to potential buyers at the time of listing (e.g., fireplaces in winter months, outdoor spaces in summer months).

We model seasonality and potential local linear trends in topics with a composite dynamic model for each $\eta_{k,t}$. Specifically, for each topic k, we define

$$F_{k,t} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad G_{k,t} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\omega) & \sin(\omega) \\ 0 & 0 & -\sin(\omega) & \cos(\omega) \end{bmatrix}.$$

The frequency of the seasonal patterns is set to be annual and $\omega = \frac{2\pi}{12}$. Though we fix the periodicity of seasonal patterns, the months at which topics peak (i.e. the phase shift) and the magnitude of the seasonal swings (amplitude) are learned from the data. The composite dynamic model for $\eta_{k,1:T}$ allows each inferred topic to exhibit a unique combination of locally linear and seasonal dynamics. DLTM shrinks the effect of some features toward zero when they are not needed. This is evidenced by the posterior means of global proportions for $K = 10$ topics presented in Figure 11(a). Topics one through three illustrate elements of both locally linear growth and seasonal patterns at different epochs. Topics four and five are characterized primarily by their seasonal patterns, which peak at different months. Topics six and seven exhibit little dynamic behavior at all, while topics eight through ten decay towards zero posterior probability. There are two important takeaways from Figures 11(a)–11(c). First, DLTM is able to learn complicated seasonal patterns and locally linear growth / decay in global topic proportions. Second, the data inform which components of the composite DLTM are needed.

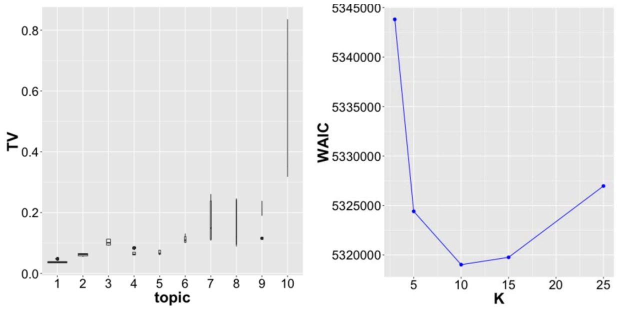


(a) All topics (b) Topic 3 (c) Topic 8

Figure 11: Inferred dynamics of topic prevalence in the Zillow corpus with $K = 10$ topics.

As observed in Section 6.3, the near zero posterior probability of topics nine and ten raises the question of whether these are extraneous topics not necessary to model the corpus. Running the MCMC simulation multiple times provides evidence that the ninth topic is genuinely needed, whereas the tenth is unnecessary. Figure 12(a) presents boxplots of TV distance between the same topics inferred from 8 different MCMC runs. The width of each boxplot is proportional to the February 2017 posterior proportion of the topic in the corpus as a whole. The first topic, which is the most prevalent in the corpus and has the widest boxplot, is nearly identical across the 8 different runs. The TV distances slightly increase as the topics become less prevalent. Notice that, although the ninth topic has low posterior probability, MCMC simulations repeatedly identify that same topic in the data. The inferred tenth topic, on the other hand, is quite different from one MCMC run to the next. Although both topics nine and ten have quite low posterior probability (see Figure 11(a)), our MCMC simulations are able to robustly infer topic nine but not ten. This suggests that the inability to replicate topic ten is not due to the low posterior probability but is rather a function of its superfluous nature.

We computed the WAIC measure of Watanabe (2013) to quantify model fit as a function of K to further examine the number of topics required. Figure 12(b) illustrates that when computed for $K = 3, 5, 10, 15$, and 25, the minimum WAIC is realized at $K = 10$. We believe that Figures 11(a), 12(a), and 12(b) illustrate a practical model selection strategy for the number of topics in a corpus.



(a) Reproducibility

(b) WAIC

Figure 12: Left: Boxplot of TV distance across multiple MCMC simulations for each topic in February 2017. Right: WAIC of models with different numbers of K topics.

Keywords of inferred topics provide important context to the dynamics of topic proportions in Figure 11(a). The ten vocabulary terms in each topic with highest posterior mean probability in February 2017 are presented in Table 1. The topics correspond to those presented in Figure 11(a) so that the first topic is the most prevalent in the corpus and the ninth topic is next to last in terms of prevalence.

The first topic describes the basic common elements of a home, including bedrooms, dining rooms, kitchens, and living rooms. It makes sense that this is the most prevalent topic in the corpus. Fireplace, the 11^{th} most likely term, and garage, the 12^{th}, help describe some of the seasonal patterns in the data.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
room	parks	home	master	new	views	basement	ceilings	garden
bedroom	home	maintained	kitchen	newer	lake	charm	doors	trees
bath	located	entertaining	tiled	updated	deck	finishes	original	newer
floors	seattle	spaces	floors	windows	sound	craftsman	details	ravenna
living	shops	open	suite	roof	home	updated	windows	uw
large	close	design	bath	painted	beach	green	leaded	village
kitchen	lot	light	walk	gas	mountains	porch	built	trail
level	neighborhood	modern	appliances	hardwoods	rainier	ballard	french	gilman
main	access	floors	counter	floors	washington	heart	wood	mature
dining	light	living	remodeled	kitchen	olympic	block	glass	burke

Table 1: The 10 vocabulary terms with the highest posterior mean probability in topics one through nine (February 2017). Each column is its own topic, and the words are ordered so that the first word has the highest posterior mean probability.

The second topic gets at the location of a home and nearby amenities such as parks, shops, and neighborhoods. Though not presented in the table, the terms downtown (11^{th} most likely), restaurant (12) and school (13) are also clearly identified with a home's location. Terms light (10) and rail (18) combine to form the name of Seattle's public transit system, the light rail. Observe in Figure 11(a) that the second topic peaks in September of 2008. The posterior probability rapidly falls thereafter, bottoming out in May 2010, and then steadily climbs through 2017. September 2008 marks the beginning of the global financial crisis, and we believe the dynamics of topic two, which identifies with a home's location and downtown Seattle, track the boom, bust, and recovery of the Seattle housing market from 2007–2017.

Topic three focuses on modern design elements of a home and outdoor spaces. Observe in Figure 11(b) that the prevalence of this topic has increased in recent years. Key words include spaces (term 4), open (5), design (6), light (7), and modern (8). Garden (12), outdoor (13), deck (16), patio (17), and landscaping (19) round out the top 20. Figure 11(b) also illustrates the seasonal nature of the topic's prevalence, which peaks in June or July and is at its lowest probability in December and January. As expected, outdoor spaces of a home are most prominently featured in summer listings. Composite seasonal and linear dynamics are evident in both topics two and three. The takeaway is that DLTM is able to learn complicated dynamics in the themes of a corpus.

Recall that topics four and five exhibit seasonal fluctuations that peak at different months and have different amplitudes. Topic four loads on luxury items such as master suite and master bath. Stainless (term 12) describes appliances (8). Granite (13) describes counter (9). Topic five identifies with remodeled homes, renovations, and home maintenance.

Topic six describes a home's view, referencing Seattle-specific attractions such as Lake (term 2) Washington (9) and Puget (14) Sound (4), the Olympic (10) Mountains (6), and Mt. Rainier (8). Figure 11(a) illustrates that the global proportion of topic six is near constant over time. It makes sense that a home's view is always included in a listing.

Topic 9, which decays in prevalence after February 2009, describes the area of Seattle near the University of Washington (UW). This area includes the neighborhood Ravenna, University Village mall, and the Burke Gilman Trail, a biking and walking trail that runs along UW and extends through Ravenna. Topic ten was deemed extraneous.

8 Discussion

In this paper, we advance the topic model literature by fusing multinomial models for text documents with the power of dynamic linear models. DLTM offers a mathematically principled framework for modeling complex dynamic behavior in corpora. The applied analysis of Zillow listings demonstrates that (i) real corpora exhibit rich temporal structure; (ii) DLTM is practically useful in modeling a wide range of dynamic features; and (iii) the themes extracted from the corpus can yield meaningful insights into dynamic patterns in real estate markets.

Perhaps the biggest open question for topic modeling is choosing the number of latent topics. A scalable model-based approach to inferring the number of topics remains elusive. While the hierarchical Dirichlet process of Teh et al. (2006) allows the data to inform the number of latent topics in the corpus, the DTM, LDA, and DLTM all require the modeler to specify K. We demonstrated in the Zillow case study that the posterior probability of a topic, WAIC, and a reproducibility analysis can be effectively utilized to select the number of topics.

A significant contribution of our work is a procedure for specifying a diverse set of prior beliefs on interpretable model features. These prior beliefs have important implications for Bayesian learning from a corpus. Learning well-separated topics and distinct document proportions requires careful prior consideration of the interplay between topic overlap and document similarity.

The DLTM model class requires intensive computation for inference and prediction. We demonstrate that Pólya-Gamma data augmentation is a method that allows for fully Bayesian posterior inference in the DLTM. While our Gibbs sampling algorithm is an admittedly slower method of inference than the variational Kalman filter of Blei and Lafferty (2006), our primary goal was not to compete with existing methods on speed and scalability. Rather, our aim was to develop an inference algorithm for an extended model class that allows inference of complicated temporal structure in text. The DLTM, combined with the Pólya-Gamma MCMC strategy, is a significant advance in this regard.

Although inference of complicated dynamics in massive corpora isn't directly addressed by the present work, we developed our computational algorithm with an eye toward scalability. Since our Gaussian approximation to the Pólya-Gamma random variable does not slow down as the length and number of documents increases, our data-augmentation and MCMC scheme will continue to work as more documents are added to the corpus. Parallel computation offers a promising path for scaling most steps in the current computational algorithm to handle corpora with a huge number of documents.

One step in the sampling procedure that does not scale particularly well with increasingly large corpora is sampling the $\beta_{k,v,t}$ natural parameter (step 1 in Section 4.2). As the size of the vocabulary significantly increases, the effective size of the MCMC sample decreases. Although a vocabulary of one-thousand terms is not a hard upper limit on the algorithm, computation with much larger vocabularies requires far more MCMC iterations for sufficient mixing. We find in synthetic data experiments that increasing the vocabulary from one-thousand to ten-thousand terms reduces the effective sample size by approximately 70%.

We note that the Pólya-Gamma stick-breaking method of Linderman et al. (2015) can offer further computational speed up by facilitating joint updates of the $\beta_{k,v,t}$ parameters associated with topics as opposed to the sequential updates performed in our computing. However, stick-breaking constructions of a probability vector naturally favor a stochastic ordering of the vector elements, while the multinomial logistic construction adopted here maintains exchangeability. The latter is indeed desirable from a modeling perspective when no pre-specified ordering of the vocabulary terms is available. A stick-breaking construction that maintains exchangeability and also offers tight control on prior variance appears to present significant mathematical difficulties that have not been considered in the present work.

Supplementary Material

Supplement to "Bayesian Analysis of Dynamic Linear Topic Models" (DOI: 10.1214/18-BA1100SUPP; .pdf).

References

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Liu, Y.-K. (2012). "A spectral algorithm for latent dirichlet allocation." In *Advances in Neural Information Processing Systems*, 917–925. 58

Anandkumar, A., Hsu, D. J., Janzamin, M., and Kakade, S. M. (2013). "When are overcomplete topic models identifiable? uniqueness of tensor tucker decompositions with structured sparsity." In *Advances in Neural Information Processing Systems*, 1986–1994. 58

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013). "A practical algorithm for topic modeling with provable guarantees." In *International Conference on Machine Learning*, 280–288. 58

Blei, D. M. and Lafferty, J. D. (2006). "Dynamic topic models." In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 113–120. New York, NY, USA: ACM. URL `http://doi.acm.org/10.1145/1143844.1143859` 53, 54, 56, 62, 67, 77

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3: 993–1022. URL `http://dl.acm.org/citation.cfm?id=944919.944937` 53, 54, 62

Carter, C. K. and Kohn, R. (1994). "On Gibbs sampling for state space models." *Biometrika*, 81(3): 541–553. MR1311096. doi: `https://doi.org/10.1093/biomet/81.3.541`. 63

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). "Scalable inference for logistic-Normal topic models." In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, 2445–2453. Curran Associates, Inc. 62, 63, 64, 66

Chib, S. (1996). "Calculating posterior distributions and modal estimates in Markov mixture models." *Journal of Econometrics*, 75: 79–97. MR1414504. doi: `https://doi.org/10.1016/0304-4076(95)01770-4`. 63

Devroye, L. (2009). "On exact simulation algorithms for some distributions related to Jacobi theta functions." *Statistics & Probability Letters*, 79(21): 2251–2259. URL http://www.sciencedirect.com/science/article/pii/S0167715209002867 MR2591982. doi: https://doi.org/10.1016/j.spl.2009.07.028. 64

Donoho, D. and Stodden, V. (2004). "When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?" In Thrun, S., Saul, L. K., and Schölkopf, B. (eds.), *Advances in Neural Information Processing Systems 16*, 1141–1148. MIT Press. 58

Frühwirth-Schnatter, S. (1994). "Data augmentation and dynamic linear models." *Journal of Time Series Analysis*, 15(2): 183–202. MR1263889. doi: https://doi.org/10.1111/j.1467-9892.1994.tb00184.x. 63

Gelman, A. and Rubin, D. B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7(4): 457–472. URL http://dx.doi.org/10.1214/ss/1177011136 68

Gerrish, S. and Blei, D. (2011). "DTM." https://github.com/blei-lab/dtm. 66, 67

Gilks, W. R. and Wild, P. (1992). "Adaptive rejection sampling for Gibbs sampling." *Applied Statistics*, 337–348. 63

Gillis, N. (2013). "Robustness Analysis of Hottopixx, a Linear Programming Model for Factoring Nonnegative Matrices." *SIAM Journal on Matrix Analysis and Applications*, 34(3): 1189–1212. MR3085119. doi: https://doi.org/10.1137/120900629. 58

Gillis, N. (2014). "Successive Nonnegative Projection Algorithm for Robust Nonnegative Blind Source Separation." *SIAM Journal on Imaging Sciences*, 7(2): 1420–1450. MR3224576. doi: https://doi.org/10.1137/130946782. 58

Gillis, N. and Vavasis, S. A. (2014). "Fast and robust recursive algorithmsfor separable nonnegative matrix factorization." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4): 698–714. 58

Glynn, C., Tokdar, S. T., Howard, B., and Banks, D. L. (2019). "Supplement to "Bayesian Analysis of Dynamic Linear Topic Models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1100SUPP. 57

Griffiths, T. L. and Steyvers, M. (2004). "Finding scientific topics." *Proceedings of the National Academy of Sciences*, 101(Suppl. 1): 5228–5235. 62

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). "Online learning for latent dirichlet allocation." In *Advances in Neural Information Processing Systems*, 856–864. 62

Holmes, C. C. and Held, L. (2006). "Bayesian auxiliary variable models for binary and multinomial regression." *Bayesian Analysis*, 1: 145–168. MR2227368. doi: https://doi.org/10.1214/06-BA105. 57

Huang, K., Fu, X., and Sidiropoulos, N. D. (2016). "Anchor-free correlated topic modeling: Identifiability and algorithm." In *Advances in Neural Information Processing Systems*, 1786–1794. MR3562709. doi: https://doi.org/10.1109/TSP.2016.2601291. 58

Kumar, A., Sindhwani, V., and Kambadur, P. (2013). "Fast conical hull algorithms for near-separable non-negative matrix factorization." In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 231–239. 58

Linderman, S., Johnson, M., and Adams, R. P. (2015). "Dependent multinomial models made easy: Stick-breaking with the Pólya-Gamma augmentation." In *Advances in Neural Information Processing Systems*, 3456–3464. 77

Liu, Y.-K., Anandkumar, A., Foster, D. P., Hsu, D., and Kakade, S. M. (2012). "Two SVDs Suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation." In *Neural Information Processing Systems (NIPS)*. 58

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Polya-Gamma latent variables." *Journal of the American Statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 54, 57, 63, 64

Recht, B., Re, C., Tropp, J., and Bittorf, V. (2012). "Factoring nonnegative matrices with linear programs." In *Advances in Neural Information Processing Systems*, 1214–1222. 58

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 77

Teh, Y. W., Newman, D., and Welling, M. (2007). "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation." In *Advances in Neural Information Processing Systems*, 1353–1360. 62

Turner, R. E. and Sahani, M. (2011). "Two problems with variational expectation maximisation for time-series models." In Barber, D., Cemgil, T., and Chiappa, S. (eds.), *Bayesian Time Series Models*, chapter 5, 109–130. Cambridge University Press. MR2894235. 62

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). "Rethinking LDA: Why priors matter." In *Advances in Neural Information Processing Systems 22*, 1973–1981. Curran Associates, Inc. URL http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.pdf 62

Watanabe, S. (2013). "A widely applicable Bayesian information criterion." *Journal of Machine Learning Research*, 14(Mar): 867–897. MR3049492. 54, 74

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Modeling*. New York, NY: Springer-Verlag, second edition. MR1482232. 54, 56

Windle, J., Carvalho, C. M., Scott, J. G., and Sun, L. (2013). "Efficient data augmentation in dynamic models for binary and count data." *ArXiv e-prints*. 54, 66

Windle, J., Polson, N. G., and Scott, J. G. (2014). "Sampling Polya-Gamma random variates: alternate and approximate techniques." *ArXiv e-prints*. 64, 65