Northwestern Law

SEARLE CENTER
ON LAW, REGULATION,
AND ECONOMIC GROWTH

WORKFORCE SCIENCE PROJECT

WORKING PAPER SERIES • NO. 15-01

Time-Dependent Topic Analysis with Endogenous and Exogenous Processes

Baiyang Wang
Department of Statistics
Northwestern University

Diego Klabjan
Department of Industrial Engineering and Management Sciences, Northwestern University

WWW.LAW.NORTHWESTERN.EDU/SEARLECENTER

2015

Time-Dependent Topic Analysis with Endogenous and Exogenous Processes

Baiyang Wang*, Diego Klabjan†

March 4th, 2015

Abstract

We consider the problem of modeling time-dependent textual data taking given endogenous and exogenous processes into consideration. Such text documents arise in real world applications, including job advertisements and financial news articles, which are influenced by the fluctuations of the general economy. We propose a hierarchical Bayesian topic model which imposes a dynamic hierarchical structure on the evolution of topics incorporating the effects of exogenous processes, and show that this model can be estimated from Markov chain Monte Carlo sampling methods. We further demonstrate that this model captures the intrinsic relationships between the topic prevalence and the time-dependent factors, and compare its performance with latent Dirichlet allocation (LDA) and the structural topic model (STM). The model is applied to two collections of documents to illustrate its empirical performance: online job advertisements from DirectEmployers Association and journalists' postings on BusinessInsider.com.

1 Introduction

Many organizations nowadays provide portals for job posting and job search, such as glassdoor.com from Glassdoor, indeed.com from Recruit, and my.jobs from DirectEmployers Association. Our work is inspired by data collected from the portal my.jobs, a website where job seekers can apply to the posted job openings through a provided link. The data collected from the website includes user clickstreams (users create accounts on the site) and attributes of the job advertisements, such as their description, location, and posted date.

In this paper, we investigate the relationship between economic fluctuations and the related changes in job advertisements, which can reveal the economic conditions of different time periods. More generally, this question is about the influence of any exogenous process on textual data with temporal dimensions. Given a corpus of text documents with time stamps and a related exogenous process, the problem is to find a relationship between the topics discussed and the exogenous process. This setting is natural in an economic context; for instance, changes in macroeconomic indicators have an impact on government reports and *Wall Street Journal* news articles. We adopt the perspective that the documents are organized into a certain number of topics, and study the impact of the exogenous process on the topic prevalence, i.e. the relative topic proportions.

With the goal of establishing topic dependency on the exogenous process, topic models are particularly suitable. They assume a given number of topics for all the documents and further assume that each word within a document is generated from one topic. The latent Dirichlet allocation (LDA)

*Department of Statistics, Northwestern University, Evanston, IL, USA. Email: baiyang@u.northwestern.edu

†Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA. Email: d-klabjan@northwestern.edu

proposed by Blei et al. (2003) is a model that has significantly contributed to the development of topic models. Since then, a large number of related methods have been proposed, many of which can be found in Blei (2011).

Meanwhile, there has been relatively limited discussion on modeling time-dependent documents when there are relevant simultaneous exogenous processes. Many time-dependent topic models without the exogenous component have been proposed, such as the dynamic topic model in Blei and Lafferty (2009), and the dynamic hierarchical Dirichlet process (dHDP) (Ren et al., 2008). However, to the best of our knowledge, none of these papers incorporate the effect of exogenous processes. On the other hand, the structural topic model (STM) (Roberts et al., 2015) considers the effect of metadata, i.e. the attributes specified for each document, on topic prevalence. This model has been applied to analyze political events (Roberts et al., 2014; Lucas et al., 2015). While STM can be applied for mining time-dependent textual data with exogenous covariates, it does not explicitly consider the time factor, and the model also excludes endogenous topic evolution processes of the time-stamped documents.

Our approach to this problem is to incorporate both endogenous and exogenous processes, assuming that the topic prevalence is affected by both of them. We apply a hierarchical modeling technique partly based on the dHDP in Ren et al. (2008), and show that our model can be derived from a multi-level hierarchical Dirichlet process. We propose a mapping from the endogenous and the exogenous processes to the per-document topic distribution, and adopt a multi-topic approach in order to make our model more flexible, thus reducing perplexity and improving prediction power. Our model has the following contributions: (i) it addresses the question of measuring the influence of exogenous processes on the topics in related documents, (ii) it incorporates both endogenous and exogenous aspects, and (iii) it demonstrates that text mining can also have useful implications in the realm of economics, which, from the authors' perspective, is a relatively new finding.

Section 2 offers a brief review on the topic modeling techniques related to our model. Section 3 develops our hierarchical Bayesian model and describes how to make posterior inferences with a variant of the Markov chain Monte Carlo (MCMC) technique. Section 4 studies the online job advertisements from DirectEmployers Association and journalists' postings in finance on BusinessInsider.com with our proposed method, providing a comparison of performance with the standard LDA and STM. Section 5 suggests possible directions for the future and concludes the paper.

2 Review of Time-Dependent Topic Modeling

We first introduce the standard model of LDA from Blei et al. (2003). Suppose that there is a collection of documents d_i, $i = 1, \ldots, N$ and words $\{x_i^j\}_{j=1}^{J_i}$ within each document d_i indexed by a common dictionary, where N is the number of documents, and J_i is the number of words in d_i. The LDA model is as follows,

$$\boldsymbol{\theta}_i \overset{iid}{\sim} Dir(\boldsymbol{\alpha}),\ \boldsymbol{\phi}_k \overset{iid}{\sim} Dir(\boldsymbol{\beta}),\ z_i^j|\boldsymbol{\theta}_i \overset{iid}{\sim} Cat(\boldsymbol{\theta}_i),\ x_i^j|z_i^j \sim Cat(\boldsymbol{\phi}_{z_i^j}), \tag{1}$$

$i = 1, \ldots, N$, $j = 1, \ldots, J_i$, $k = 1, \ldots, K$, where $\boldsymbol{\theta}_i$ is the per-document topic distribution for d_i, $\boldsymbol{\phi}_k$ is the per-topic word distribution for the k-th topic, z_i^j is the actual topic for the j-th word in d_i, and K is the number of topics. $Dir(\cdot)$ refers to the Dirichlet distribution and $Cat(\cdot)$ refers to the categorical distribution, a special case of the multinomial distribution when $n = 1$.

The Dirichlet process is a class of randomized probability measures and can be applied for semi-parametric modeling of mixture models. Denoting the concentration parameter by α and the mean probability measure by H, a realization G from the Dirichlet process can be written as $G \sim DP(\alpha, H)$. With the stick-breaking notation by Sethuraman (1994), we have

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}, \tag{2}$$

where δ_{θ_k} is a "delta" probability measure with all the probability mass placed at θ_k, $\theta_k \overset{iid}{\sim} H$, $\beta_k = \beta'_k \prod_{i=1}^{k-1}(1-\beta'_i)$, $\beta'_k \overset{iid}{\sim} Beta(1, \alpha)$, $k = 1, 2, \cdots$. More properties of the Dirichlet process can be found in Ferguson (1973).

Teh et al. (2005) proposed a hierarchical Dirichlet process (HDP) which can be applied to text modeling. They assumed the following structure,

$$G \sim DP(\gamma, H),\ G_1, \ldots, G_T | G \overset{iid}{\sim} DP(\alpha, G). \tag{3}$$

This structure was demonstrated to be useful for text modeling, especially when there are clusters within the documents. Given a time structure, the clusters can be set to be the documents within each time period. In this mixture model, the topic is set to be unique for each document, and multiple realizations of words are drawn from each topic. Specifically, we can let $\boldsymbol{\theta}_{t,i} \overset{iid}{\sim} G_t$, $n_{t,i} = f(\boldsymbol{\theta}_{t,i})$, where $\boldsymbol{\theta}_{t,i}$ is the topic distribution for the i-th document in the t-th period, $n_{t,i}$ is a word count statistic, and $f(\theta)$ is a randomized function such as a binomial random variable with n observations and parameter θ.

Ren et al. (2008) further proposed a dynamic HDP which assumes a finite mixture structure where the parameters in every period depend on the last period, thereby taking endogenous factors into consideration. Specifically,

$$G_1 = E_1,\ G_t = w_t E_t + (1 - w_t) E_{t-1},\ t = 2, \ldots, T, \tag{4}$$

where $\{E_t\}_{t=1}^T$ admits an HDP structure. Pruteanu-Malinici et al. (2010) proposed an alternate version of the above problem with truncation of the stick-breaking representation of the Dirichlet processes. The hierarchical Dirichlet structures of these papers are similar to our model; however, we reconsider this problem from another perspective, incorporating exogenous processes and adopting a multi-topic approach, therefore differing from the previous models.

Roberts et al. (2015) proposed the structural topic model (STM) that measures the effect of metadata of each document with the logistic normal distribution. Their model for each document d_i is as follows,

$$\begin{aligned} &\boldsymbol{\theta}_i | (X_i\gamma, \Sigma) \sim LogisticNormal(X_i\gamma, \Sigma),\ p(\boldsymbol{\phi}_{i,k}) = C \cdot \exp(m + \kappa_k + \kappa_{g_i} + \kappa_{kg_i}), \\ &z_i^j | \boldsymbol{\theta}_i \overset{iid}{\sim} Cat(\boldsymbol{\theta}_i),\ x_i^j | z_i^j \sim Cat(\boldsymbol{\phi}_{i,z_i^j}), \end{aligned} \tag{5}$$

where C is a constant, X_i is the metadata matrix, γ is a coefficient vector, Σ is the covariance matrix, $\boldsymbol{\phi}_{i,k}$ is the word distribution for d_i and the k-th topic, m is a baseline log-word distribution, κ_k, κ_{g_i} and κ_{kg_i} are the topic, group, and interaction effects; the rest are defined similarly to LDA. This model is related to our work in that it considers exogenous factors. However, our model treats the time factor explicitly and considers the generation of the document-specific topic distribution $\boldsymbol{\theta}_i$ as a part of the hierarchical Dirichlet process.

3 Model and Algorithm

3.1 A Time-Dependent Hierarchical Bayesian Topic Model

We formulate our problem as follows: we are given a series of time periods $t = 1, \ldots, T$, a set of text documents from each period $d_{t,i}$, $i = 1, \ldots, N_t$, $t = 1, \ldots, T$, and the indices of words $\{x_{t,i}^j\}_{j=1}^{J_{t,i}}$ within each document $d_{t,i}$ from the first one to the last. The words are sourced from a dictionary containing V words in total. Here, we consider N_t to be the number of documents for each time period t, and $J_{t,i}$ to be the word count for each document $d_{t,i}$. We further let E denote the cross-period baseline, i.e. the mean endogenous probability measure for the topic assignment across different time periods. We have the following equations,

$$\begin{cases} E \sim DP(\gamma, H), \\ E_1, \ldots, E_T | (\alpha, E) \overset{iid}{\sim} DP(\alpha, E), \\ w_1, \ldots, w_T \overset{iid}{\sim} Beta(b_1, b_2), \\ G_1 = E_1, G_t = w_t E_t + (1 - w_t) E_{t-1},\ t = 2, \ldots, T, \end{cases} \tag{6}$$

where E_t is the new endogenous probability measure for the topic assignment in each time period t, w_t controls the strength of innovation in the t-th period, and G_t is the baseline, i.e. realized endogenous probability measure for the topic assignment in each time period t, so that $\{G_t\}_{t=1}^T$ corresponds to a dynamic hierarchical Dirichlet process.

Similarly to LDA, we consider the words within each document to be generated from multiple topics, rather than taking the alternative view assuming each document to belong to only one topic. Furthermore, we adopt a one-topic-per-word approach, so that the assignment of topics for each word $x_{t,i}^j$, i.e. the topic distribution, can be ultimately derived from (6).

Therefore, we introduce a random probability measure $\Theta_{t,i}$ for each document $d_{t,i}$, from which all the topics within $d_{t,i}$ are drawn. Specifically, the topic of every word in $d_{t,i}$ is independently and identically distributed from $\Theta_{t,i}$. We assume that for every period t,

$$\Theta_{t,i} \overset{iid}{\sim} DP(a, \mathcal{M}(G_t; \mathbf{y}_t)),\ i = 1, \ldots, N_t, \tag{7}$$

where $\mathbf{y}_t$ is a p-dimensional exogenous covariate, the value of some random process $\{\mathbf{y}_t : t \in [0, T]\}$ or some time series $\{\mathbf{y}_t\}_{t=1}^T$ at time t, on which we elaborate later. Mapping $\mathcal{M}$ is such that the baseline topic distribution for the t-th period G_t and the exogenous covariate $\mathbf{y}_t$ are mapped to the realized mean topic distribution for the t-th period $\mathcal{M}(G_t; \mathbf{y}_t)$. Parameter a controls the topic similarity of all documents within each period.

Alternatively, the model can be reformulated as follows. Note that because there are K topics in our model, the space of the topic probability measures is discrete, so we may as well represent them in terms of K-dimensional probability vectors. Specifically, let $\boldsymbol{\pi}_0$, $\boldsymbol{\pi}$, $\boldsymbol{\pi}_t$, $\boldsymbol{\tau}_t$, and $\boldsymbol{\theta}_{t,i}$ denote the probability vectors of H, E, E_t, G_t, and $\Theta_{t,i}$, $i = 1, \ldots, N_t$, $t = 1, \ldots, T$. Furthermore, let $\mathcal{M}_{\boldsymbol{\eta}}(\boldsymbol{\tau}_t; \mathbf{y}_t)$ denote the probability vector for $\mathcal{M}(G_t; \mathbf{y}_t)$, where $\boldsymbol{\eta}$ is a finite-dimensional parameter for $\mathcal{M}$. From the formal definition of the Dirichlet process, see e.g. Ferguson (1973), we immediately derive the following proposition.

Proposition 1. The following is equivalent to (6) and (7),

$$\begin{cases} \boldsymbol{\pi} \sim Dir(\gamma\boldsymbol{\pi}_0), \\ \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_T | (\alpha, \boldsymbol{\pi}) \overset{iid}{\sim} Dir(\alpha\boldsymbol{\pi}), \\ w_1, \ldots, w_T \overset{iid}{\sim} Beta(b_1, b_2), \\ \boldsymbol{\tau}_1 = \boldsymbol{\pi}_1, \boldsymbol{\tau}_t = w_t\boldsymbol{\pi}_t + (1 - w_t)\boldsymbol{\pi}_{t-1},\ t = 2, \ldots, T, \\ \boldsymbol{\theta}_{t,i} | (a, \boldsymbol{\eta}, \boldsymbol{\tau}_t, \mathbf{y}_t) \overset{iid}{\sim} Dir(a\mathcal{M}_{\boldsymbol{\eta}}(\boldsymbol{\tau}_t; \mathbf{y}_t)), i = 1, \ldots, N_t, t = 1, \ldots, T. \end{cases} \tag{8}$$

□

Formula (8) provides an explicit form for the hierarchical Dirichlet process in our situation. With our multi-topic assumption for word generation within each document, we are able to lay our model in a multi-level hierarchical Dirichlet setting while avoiding the computational burden from applying truncation of stick-breaking processes.

For the word generating process, let $\boldsymbol{\phi}_k$ denote the per-topic word distribution, which is the probability vector for the words listed in the dictionary to be chosen for the k-th topic. Let $\{z_{t,i}^j\}_{j=1}^{J_{t,i}}$ denote the topics for all the words in document $d_{t,i}$, $i = 1, \ldots, N_t$, $t = 1, \ldots, T$. Finally we assume the following

$$\begin{cases} \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K \overset{iid}{\sim} Dir(\boldsymbol{\beta}), \\ z_{t,i}^j | \boldsymbol{\theta}_{t,i} \overset{iid}{\sim} Cat(\boldsymbol{\theta}_{t,i}), \\ x_{t,i}^j | (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K, z_{t,i}^j) \sim Cat(\boldsymbol{\phi}_{z_{t,i}^j}). \end{cases} \tag{9}$$

Now we elaborate on the p-dimensional stochastic process $\{\mathbf{y}_t : t \in [0, T]\}$ with a parameter vector $\boldsymbol{\zeta}$. This exogenous process $\{\mathbf{y}_t : t \in [0, T]\}$ is an underlying circumstance which affects the topic assignments of the documents in the same period. We can find many examples in real data, such as the influence of the general economic condition on the availability of jobs in different categories, as mentioned earlier in the paper.

We discretize this stochastic process into a time series $\{\mathbf{y}_t\}_{t=1}^T$. Let $X = \{x_{t,i}^j\}_{j=1}^{J_{t,i}}; {}_{i=1}^{N_t}; {}_{t=1}^{T}$ denote all the textual data, and let $Y = \{\mathbf{y}_t\}_{t=1}^T$ denote the discretized exogenous process. Then let $D = \{X, Y\}$ denote all the data, and let Δ represent all variables other than D and $\boldsymbol{\zeta}$. We make the assumption that the distribution of Δ conditioning on Y and $\boldsymbol{\zeta}$ only depends on Y, which is a relatively weak condition of independence. The assumption can be represented as

$$p(\Delta | Y, \boldsymbol{\zeta}) = p(\Delta | Y). \tag{10}$$

This also implies that the endogenous process of topic generation does not depend on the internal mechanism of the exogenous process; only the values from each period of the exogenous process suffice. Alternatively, we may only consider a discretized exogenous process $Y = \{\mathbf{y}_t\}_{t=1}^T$ without any continuous component, and all results still hold. From a Bayesian point of view, the discretized exogenous process apparently incorporates AR(p) models, and as a matter of fact, it also incorporates ARMA(p, q) models, as the underlying error terms of ARMA(p, q) can be viewed as parameters of Y, thereby components of $\boldsymbol{\zeta}$.

One question of interest is whether the introduction of the textual data would influence the parameter estimation of the exogenous process. In summary, we have the following proposition.

Proposition 2. Let us assume (8-10) and all priors are independent. Random quantities Δ and $\boldsymbol{\zeta}$ are independent conditioning on the data $D = \{X, Y\}$. Moreover,

$$p(\Delta, \boldsymbol{\zeta}|X, Y) = p(\Delta|X, Y) \cdot p(\boldsymbol{\zeta}|Y). \tag{11}$$

Proof. Note that from (10), $p(\Delta|Y, \boldsymbol{\zeta}) = p(\Delta|Y)$ and from (9), $p(X|Y, \Delta, \boldsymbol{\zeta}) = p(X|Y, \Delta)$. By a straightforward derivation, the equality holds. □

Here we achieve a form of invariance in our model such that the inference of $\boldsymbol{\zeta}$ is not influenced by the contents of the textual data. This has an intuitive explanation as the acquired documents represent only a very small portion of the underlying environment, so their influence on the exogenous parameter $\boldsymbol{\zeta}$ is almost negligible. Therefore, we can treat $\{\mathbf{y}_t\}_{t=1}^T$ as given and focus on (8) and (9), without referring to $\boldsymbol{\zeta}$, in order to simplify our model.

To provide more insights, we present the posterior conditional distribution of the components of Δ as follows. Again assuming (8-10), we obtain the following proposition.

Proposition 3. The conditional posteriors for $\Delta_1 = \{\boldsymbol{\pi}, \boldsymbol{\pi}_t, w_t, a, \boldsymbol{\eta}\}_{t=1}^T$, $\{\boldsymbol{\theta}_{t,i}\}_{i=1}^{N_t}$; $_{t=1}^{T}$, $\{\boldsymbol{\phi}_k\}_{k=1}^K$, $\{z_{t,i}^j\}_{j=1}^{J_{t,i}}$; $_{i=1}^{N_t}$; $_{t=1}^{T}$ given D are as follows

$$\begin{cases} p(\Delta_1|(\Delta\backslash\Delta_1, D)) = C \cdot \left\{ \prod_{t=1}^{T} \left[\prod_{i=1}^{N_t} p_{Dir}(\boldsymbol{\theta}_{t,i}; a\mathcal{M}_{\boldsymbol{\eta}}(\boldsymbol{\tau}_t; \mathbf{y}_t)) \right] p_{Dir}(\boldsymbol{\pi}_t; \alpha\boldsymbol{\pi}) \right\} p_0(\Delta_1\backslash\{\boldsymbol{\pi}_t\}), \\ \boldsymbol{\theta}_{t,i}|(\Delta\backslash\boldsymbol{\theta}_{t,i}, D) \sim Dir\left(\left(\sum_j 1_{z_{t,i}^j=1}, \ldots, \sum_j 1_{z_{t,i}^j=K} \right) + a\mathcal{M}_{\boldsymbol{\eta}}(\boldsymbol{\tau}_t; \mathbf{y}_t) \right), \\ \boldsymbol{\phi}_k|(\Delta\backslash\boldsymbol{\phi}_k, D) \sim Dir\left(\left(\sum_{t,i,j} 1_{z_{t,i}^j=k,\, x_{t,i}^j=1}, \ldots, \sum_{t,i,j} 1_{z_{t,i}^j=k,\, x_{t,i}^j=V} \right) + \boldsymbol{\beta} \right), \\ z_{t,i}^j|(\Delta\backslash z_{t,i}^j, D) \sim Cat\left(\phi_1^{x_{t,i}^j}\theta_{t,i}^1, \ldots, \phi_K^{x_{t,i}^j}\theta_{t,i}^K \right). \end{cases} \tag{12}$$

Here C is a constant, $p_{Dir}(\cdot;\cdot)$ denotes the probability density function of the Dirichlet distribution with some parameter, and $p_0(\cdot)$ denotes joint prior of the included variables.

Proof. Note that $p(\Delta|D)$, the joint posterior for Δ, is proportional to

$$\left\{ \prod_{t=1}^{T} \left[\prod_{i=1}^{N_t} \left(\prod_{j=1}^{J_{t,i}} \theta_{t,i}^{z_{t,i}^j} \phi_{z_{t,i}^j}^{x_{t,i}^j} \right) p_{Dir}(\boldsymbol{\theta}_{t,i}; a\mathcal{M}_{\boldsymbol{\eta}}(\boldsymbol{\tau}_t; \mathbf{y}_t)) \right] p_{Dir}(\boldsymbol{\pi}_t; \alpha\boldsymbol{\pi}) \right\} p_0(\{\boldsymbol{\phi}_k\} \cup \Delta_1\backslash\{\boldsymbol{\pi}_t\}).$$

The proposition immediately follows. □

In summary, this is a hierarchical Bayesian topic model which incorporates the endogenous structure of topic assignments over different periods by applying a dynamic hierarchical Dirichlet process and introduces the exogenous structure by assuming a mapping from both processes to a topic distribution. We have also shown that our model is conditional on the whole exogenous process. Estimation of the posterior of the model is described in the next section.

3.2 *Sampling the posterior: An MCMC Approach*

Direct estimation of the Bayesian posterior is often intractable since the closed-form expression, if it exists, can be difficult to integrate and thus, many approaches to approximate the posterior have been proposed. Monte Carlo methods, which draw a large number of samples from the posterior as its approximation, are particularly helpful. In this paper, we adopt the Markov chain Monte Carlo (MCMC) approach which constructs samples from a Markov chain and is asymptotically exact. Below we provide the Metropolis-Hastings-within-Gibbs sampling approach tailored to our situation, a variant of the general MCMC approach.

We intend to sample from the posterior $\Delta|D$ using the conditional distributions in (12), and re-parametrize the parameter $\Delta_1 = \{\boldsymbol{\pi}, \boldsymbol{\pi}_t, w_t, a, \boldsymbol{\eta}\}_{t=1}^T$ as $\Delta_1 = (\delta_1, \ldots, \delta_s)$, Δ as $(\delta_1, \ldots, \delta_t)$, $s < t$. We set the initial value $\Delta^{(0)}$ to Δ, and carry out an iterative algorithm representing the variables at the r-th iteration with superscript (r). At the r-th iteration, we first update sequentially $\delta_1, \ldots, \delta_s$. We let

$$\delta_i^* \sim q(\cdot|\delta_i^{(r-1)}),\ \delta_i^{(r)} = \begin{cases} \delta_i^* & \text{with probability } P, \\ \delta_i^{(r-1)} & \text{with probability } 1-P, \end{cases} \tag{13}$$

where $P = \min\left\{p\left(\delta_i^*\middle|\delta_1^{(r-1)}, \ldots, \delta_{i-1}^{(r-1)}, \delta_{i+1}^{(r)}, \ldots, \delta_t^{(r)}\right) q\left(\delta_i^{(r-1)}\middle|\delta_i^*\right) \Big/ p\left(\delta_i^{(r-1)}\middle|\delta_1^{(r-1)}, \ldots, \delta_{i-1}^{(r-1)}, \delta_{i+1}^{(r)}, \ldots, \delta_t^{(r)}\right) q\left(\delta_i^*\middle|\delta_i^{(r-1)}\right), 1\right\}$, and $q(\cdot|\cdot)$ is a known conditional distribution with positive density over the range of δ_i, $i = 1, \ldots, s$. Note that when P is low, we update each δ_i, $i = 1, \ldots, s$, for a fixed number of times in each iteration.

Continuing from above, we then update the values of $\{\boldsymbol{\theta}_{t,i}\}_{i=1}^{N_t}{}_{t=1}^{T}$, $\{\boldsymbol{\phi}_k\}_{k=1}^K$, $\{z_{t,i}^j\}_{j=1}^{J_{t,i}}{}_{i=1}^{N_t}{}_{t=1}^{T}$ from the conditional posterior of each variable,

$$\begin{cases} \boldsymbol{\theta}_{t,i}^{(r)} \sim Dir\left(\left(\sum_j 1_{z_{t,i}^{j(r-1)}=1}, \ldots, \sum_j 1_{z_{t,i}^{j(r-1)}=K}\right) + a^{(r)}\mathcal{M}_{\boldsymbol{\eta}}^{(r)}\left(\boldsymbol{\tau}_t^{(r)}; \mathbf{y}_t\right)\right), \\ \boldsymbol{\phi}_k^{(r)} \sim Dir\left(\left(\sum_{t,i,j} 1_{z_{t,i}^{j(r-1)}=k,\, x_{t,i}^j=1}, \ldots, \sum_{t,i,j} 1_{z_{t,i}^{j(r-1)}=k,\, x_{t,i}^j=V}\right) + \boldsymbol{\beta}\right), \\ z_{t,i}^{j(r)} \sim Cat\left(\phi_1^{x_{t,i}^j(r)}\theta_{t,i}^{1(r)}, \ldots, \phi_K^{x_{t,i}^j(r)}\theta_{t,i}^{K(r)}\right). \end{cases} \tag{14}$$

This completes our sampling approach. We apply this approach as both the Gibbs sampler and the Metropolis-Hastings algorithm cannot be directly applied to (12). Our approach can be viewed as an approximation of the Gibbs sampler replacing each Gibbs update by one or multiple univariate Metropolis-Hastings updates. Thus we construct a Markov chain $\Delta^{(0)} \to \Delta^{(1)} \to \cdots$ from sampling, which we can regard as approximate samples from the posterior of $\Delta|D$.

To establish the theoretical properties of the Markov chain $\Delta^{(0)} \to \Delta^{(1)} \to \cdots$, we let all priors have positive density over the range of each parameter, and let $\mathcal{M}_{\boldsymbol{\eta}}$ in (12) be positive and measurable. From Roberts and Rosenthal (2006), we immediately know that the Markov chain $\Delta^{(0)} \to \Delta^{(1)} \to \cdots$ is μ-irreducible and aperiodic, μ being the probability measure of the posterior distribution $\Delta|D$. Following Theorem 12 in Roberts and Rosenthal (2006), we know that the Markov chain $\Delta^{(0)} \to \Delta^{(1)} \to \cdots$ is Harris recurrent as a deterministic-scan Metropolis-Hastings-within-Gibbs sampler (see definitions in the paper). Furthermore, from Theorem 6.51 and Theorem 6.63 in Robert and Casella (2004), we derive the following.

Proposition 4. For the Markov chain $\Delta^{(0)} \to \Delta^{(1)} \to \cdots$ constructed from (13) and (14),

(a) $\|p_{\Delta^{(n)}}(\cdot) - p_{\Delta|D}(\cdot)\|_\infty \xrightarrow{a.s.} 0$ as $n \to \infty$ for any $\Delta^{(0)}$;

(b) $\frac{1}{n}\sum_{i=1}^{n} f(\Delta^{(i)}) \xrightarrow{a.s.} E[f(\Delta)|D]$ as $n \to \infty$, for every $f \in L_1(\mu)$. □

4 Case Studies

The proposed model is demonstrated on the two data sets: (1) online job advertisements from my.jobs from February to September in 2014, and (2) journalists' postings from January 3rd, 2014 to December 31st, 2014 in the "Finance" section in BusinessInsider.com, an American business and technology news website. Our algorithm has been implemented in Java, and we compare our proposed model with LDA and STM.

4.1 Model Specification

We denote our proposed model as "EeLDA." For the online job advertisements, we initialize the hyperparameters of EeLDA as follows: $\alpha = K^2$, $\boldsymbol{\beta} = (1, \ldots, 1)$, $\gamma = K$, $\boldsymbol{\pi}_0 = (1/K, \ldots, 1/K)$, $b_1 = b_2 = 1$. We set the priors of a and $\boldsymbol{\eta}$ to be flat. We carry out the Metropolis-Hastings-within-Gibbs algorithm as described in Section 3.2, and run 10,000 iterations of the Markov chain with 2,000 burn-in samples. The number of topics is set to $K = 20$ for both data sets. For LDA, we let $\boldsymbol{\alpha} = (1, \ldots, 1)$, $\boldsymbol{\beta} = (1, \ldots, 1)$. For STM, we apply the default settings in the R package `stm`. We perform data cleaning, remove the stopwords, stem the documents, and keep most frequent 2,000 words in each study, so that $V = 2{,}000$. We do this as we have found that for both studies, introducing more rare words makes the results more unstable and less interpretable.

For the journalists' postings, every setting is the same with the above paragraph except that the hyperparameter $\boldsymbol{\beta}$ is set to $(0.1, \ldots, 0.1)$ for EeLDA and LDA. We note that when $\boldsymbol{\beta}$ increases, the topics are assumed to be more alike, and *vice versa*. We let $\boldsymbol{\beta}$ be small as we suppose that the topics in the articles should be more distinguishable from each other than using a flat prior.

For the form of $\mathcal{M}_{\boldsymbol{\eta}}$, we choose the following

$$\mathcal{M}_{\boldsymbol{\eta}}(\boldsymbol{\tau}_t; \mathbf{y}_t) = \boldsymbol{\tau}_t + \boldsymbol{\eta}\mathbf{y}_t, \ \mathbf{1}^T\boldsymbol{\eta} = \mathbf{0}, \tag{15}$$

so that $\boldsymbol{\eta}$ represents the "co-movement propensity" of the topics, i.e. the propensity of change in the topic proportions with regard to changes in $\{\mathbf{y}_t\}$.

Moreover, we recommend that the hyperparameter $\boldsymbol{\alpha}$ for LDA be no less than $\mathbf{1}$. We find that since not every topic necessarily appears in each document, $\boldsymbol{\alpha}$ being close to $\mathbf{0}$ can lead LDA into degeneration. For EeLDA, the same would happen if α goes to 0. To remedy this situation, we assume that there are κ units of "latent" topics equally assigned to each topic and document. Specifically, the second equation in (12) can be rewritten as

$$\boldsymbol{\theta}_{t,i}|(\Delta \backslash \boldsymbol{\theta}_{t,i}, D) \sim Dir\left(\left(\sum_j 1_{z_{t,i}^j=1}, \ldots, \sum_j 1_{z_{t,i}^j=K}\right) + a\mathcal{M}_{\boldsymbol{\eta}}(\boldsymbol{\tau}_t; \mathbf{y}_t) + \kappa\mathbf{1}\right). \tag{16}$$

Everything else remains the same. Note that LDA with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ can be identified as LDA with $\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\alpha}_0$ units of latent topics. We set $\kappa = 1$ and apply (15) and (16) in both studies.

4.2 My.jobs: Online Job Advertisements

The number of online job advertisements on my.jobs from February to September in 2014 amounts to $17{,}147{,}357$ in total, and the number of advertisements each day varies greatly. Therefore, we gather a stratified sample of $44{,}660$ advertisements with a roughly equal number of samples for each day, so that we have sampled 0.26% of all the documents in total. The training data set consists of $40{,}449$ advertisements, and the testing data set consists of $4{,}211$ advertisements (9.4% of the sample). The discrepancy between 9.4% and targeted 10% is caused by removal of advertisements in Spanish after splitting. For the exogenous variable $\{\mathbf{y}_t\}_{t=1}^{T}$, we use the standardized Consumer Price Index from February to September in 2014, so that $p = 1$, and $T = 8$.

We use perplexity to compare the difference of the prediction power of different methods. The perplexity for N_{test} held-out documents given the training data D is defined as

$$perp = \exp\left\{-\frac{\sum_{i=1}^{N_{test}} \log p(d_{test,i}|D)}{\sum_{i=1}^{N_{test}} n_{test,i}}\right\} \tag{17}$$

where $d_{test,i}$ represents the i-th held-out document, and $n_{test,i}$ is the number of words in $d_{test,i}$. We expect the perplexity to be small when a model performs well, since this means that under the estimated model, the probability of a word in the testing documents being written *a priori* is large. Figure 1 shows the different perplexity of the three models.

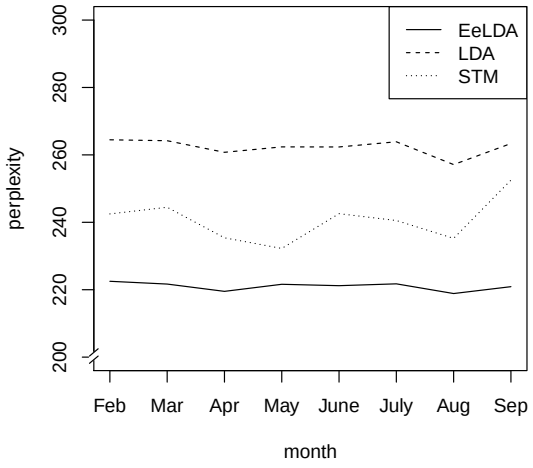

Figure 1: Perplexity results for the job advertisements from February to September in 2014.

Figure 1 implies that EeLDA better predicts the words in the new documents in terms of perplexity. This is due to the fact that the introduction of the endogenous and exogenous processes allows us to make more accurate inference on the topic distributions of the documents in a given period of time. We also observe that the perplexity is more stable against time for EeLDA.

The actual topics are presented in Figure 2. The x-axis and y-axis in Figure 2 are the relative co-movement propensities $\nu = \eta/\pi$ for all the topics, i.e. the percent change in the topic proportion

given one unit change in the exogenous covariate, and the cross-period baseline topic proportions π. Here π and η denote the related component of $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$ for each topic. Table 1 lists the 10 highest probability words sorted by their probabilities from high to low inside the five topics with highest relative co-movement propensities ν.

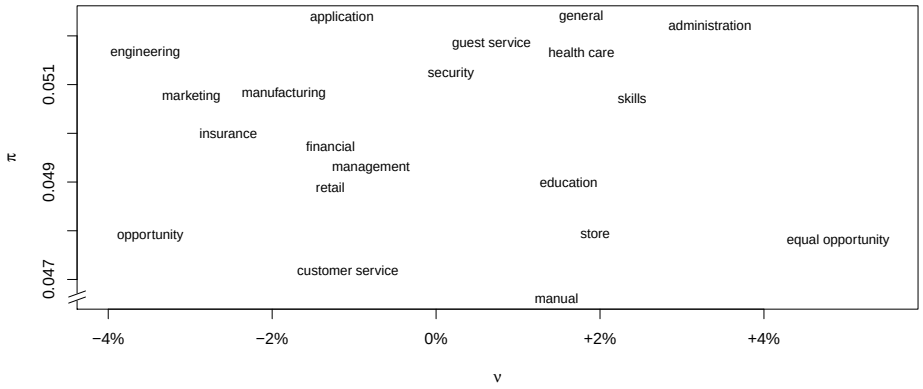

Figure 2: The topics for the job advertisements, the cross-period baseline proportions π, and the relative co-movement propensities ν from EeLDA.

equal opportunity	employment status equal opportunity disabled veteran protected national origin job
administration	procedures policies reports maintain ensure appropriate assigned duties time information
skills	skills ability work experience communication excellent strong knowledge required written
store	customer products job work store perform required friendly department order
health care	care health nursing patient medical clinical provide hospital practice current

Table 1: The 10 highest probability words sorted by probability inside the five topics with highest ν.

A number of facts can be inferred from Figure 2. The topics with positive ν are those that have a positive correlation with the growth of the CPI in 2014. We can observe that several of them are supported by the U.S. government spending, namely "equal opportunity," "administration," "health care," and "education." This confirms that there is a causal relationship between the increase in government spending and the increase in the number of jobs in these categories, and the former was also an underlying factor in the growth of the CPI in 2014.

On the other hand, many observers of the U.S. labor market pointed out that while there was a boom in the number of jobs in 2014, the job market created more lower-paid jobs than high-paid ones. From the results of our model, we can also observe that with the growth of CPI in 2014, lower-paid job categories, such as equal opportunity jobs, tend to move in the same direction. Meanwhile, some traditional higher-paid ones, such as engineering and marketing, do not. Therefore the suggestions from our analysis of the job advertisements agree with a number of news articles on the U.S. labor market in 2014; for instance, see Lowrey (2014) and Coy (2014).

We also conduct sensitivity analyses of EeLDA below. We are mainly concerned with the hyper-parameters we introduce into EeLDA in addition to the original structure of LDA, namely α and κ, which are not the parameters of any prior distribution. Other hyperparameters we introduced constitute flat priors which are trivial. Among the results, the co-movement propensities ν are the most useful for comparison. The plots of π and ν for different values of α and κ are in Figure 3.

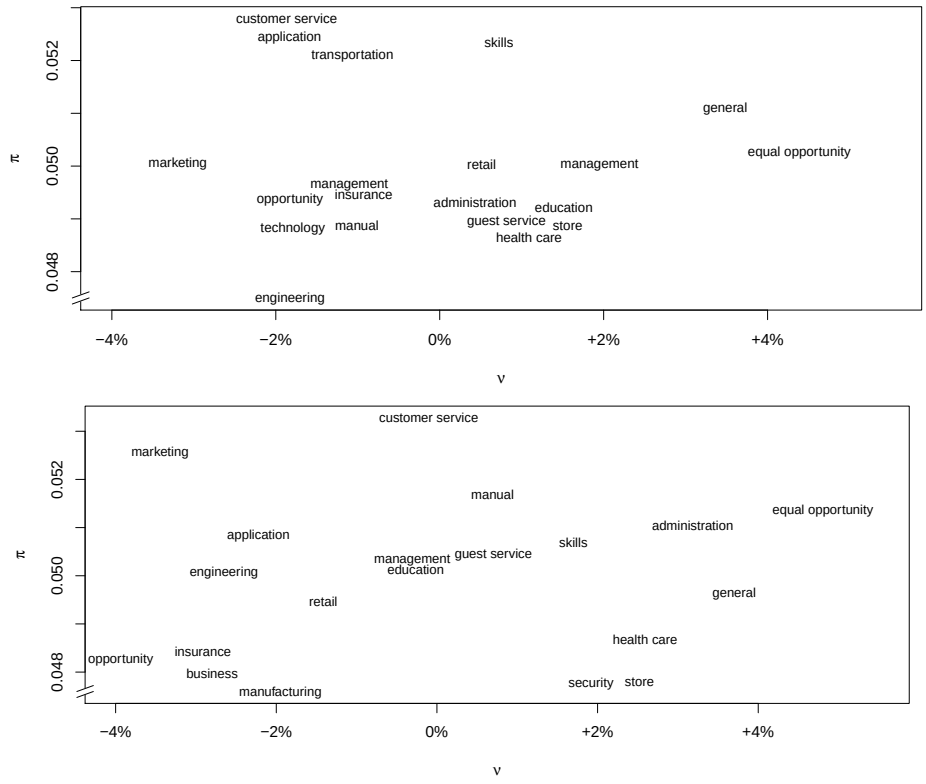

Figure 3: The topics for the job advertisements, the cross-period baseline proportions π, and the relative co-movement propensities ν given $\alpha = 2K^2$ (top); $\kappa = 2$ (bottom) from EeLDA; other parameters unchanged.

Topic models are often sensitive to changes in hyperparameters; for instance, changing V, K, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ can make results much less interpretable for LDA and EeLDA in both studies ($\boldsymbol{\alpha}$ only exists in LDA). For STM, we have even observed sensitivity with regard to the initialization hyperparameters and convergence criterion in both studies. The actual comparisons are omitted for brevity.

However, in Figure 3, we observe that the topics produced from these settings are generally alike. Although there are changes in both π and ν, the orders of ν of the topics are generally similar. In other words, they provide us with the same conclusion that the increase in employment in 2014 is largely driven by government spending and reflected in lower-paying sectors. In this sense, these results have shown that EeLDA is quite insensitive towards changes in hyperparameters in terms of the panoramic view of the labor market. They have also asserted the validity of EeLDA, since different hyperparameter settings result in the same conclusion, and in accordance with the general public opinion. These results also hold for the second study, again omitted here for brevity.

As a rule of thumb, it is recommended that for EeLDA, $\alpha = K^2$ so that the prior coefficient of variation for each component of $\{\boldsymbol{\pi}_t\}_{t=1}^T$ is approximately $1/\sqrt{K}$ to encourage clustering; $\kappa = 1$ intuitively; K is between 10 and 20; V is approximately 2000 (K and V should be decreased for smaller data sets); $\boldsymbol{\beta} = \mathbf{1}$ or moderately less given evidence of highly distinguishable topics; other hyperparameters should be initialized as in Section 4.1 so that they form flat priors from intuition.

4.3 *BusinessInsider.com: Financial News Articles*

We consider all contributions in the "Finance" section of BusinessInsider.com on all trading days in 2014. There are $15{,}659$ articles in total, which are divided into a training data set containing $12{,}527$ articles and a testing data set containing $3{,}132$ articles (20% of all articles). We increase the proportion of testing documents to 20% because we anticipate that the corpus of financial articles is highly predictable. We apply the daily price of the Chicago Board Options Exchange Market Volatility Index (VIX) as the exogenous process, measuring the volatility of the U.S. financial market. The other settings are the same as those in Section 4.2. We provide an analysis of the perplexity of different models in Figure 4. The lines are smoothed by LOESS with a span of 0.1, as there are large fluctuations in perplexity from day to day.

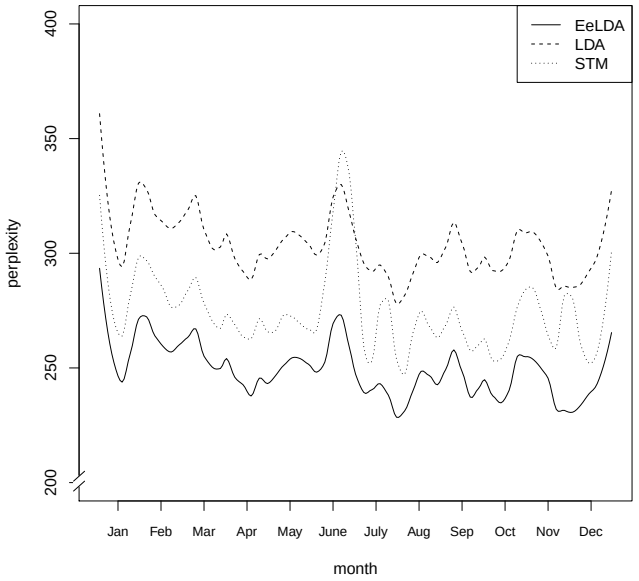

Figure 4: Perplexity results for contributions in the "Finance" section in businessinsider.com in 2014.

Again we observe that EeLDA generates the lowest perplexity for the testing documents, therefore improving the fitting of the topic model. We present the topics from our analysis and their relative co-movement propensities and average proportions in Figure 5, which is similar to Figure 2 in Section 4.2. We also list the 10 highest probability words inside the five topics with highest ν, sorted by their probabilities in Table 2.

energy	price oil production energy gas supply cost crude industrial cut
Europe/Russia	Russia government country president tax Ukraine politics European minister national
stock market	market stock price time Morgan Stanley sell note earnings chart average
currency	bank debt financial credit central currency dollar crisis government ECB
daily news	week trade day close free morning dollar Friday Thursday Monday

Table 2: The 10 highest probability words sorted by probability inside the five topics with highest ν.

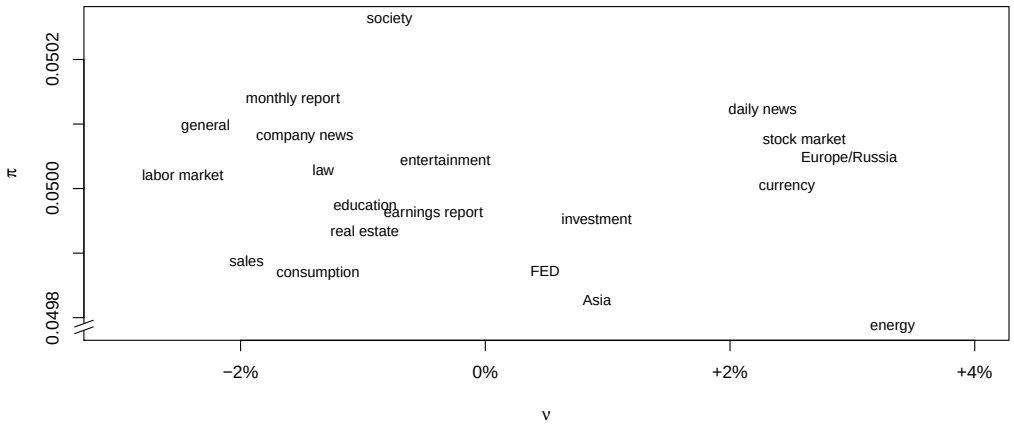

Figure 5: The topics for contributions in the "Finance" section, the cross-period baseline proportions π, and the relative co-movement propensities ν from EeLDA.

The topics that are strongly positively correlated with the VIX include short-term news, such as stock market news and announcements from central banks, as in the topics "stock market" and "currency." From our analysis, the drop in oil price and the instability in Russia and Ukraine were also major causes of fluctuations in the stock market in 2014. On the other hand, we observe that news about longer-term economic trends is not positively correlated with the VIX, such as "company" and "labor market." We suggest that EeLDA can be used for finding topics that are major contributors to changes in an exogenous process during a period of time.

We also observe that EeLDA is able to extract certain topics highly correlated with an exogenous process. For instance, the "energy" and "Europe/Russia" topics significantly contribute to VIX and are successfully extracted. To assert our observation, we make a closer comparison of the 20 highest probability words sorted by probability in the closest energy-related topic and the closest Europe-related topic from LDA, STM and EeLDA. The results are presented in Table 3.

LDA	energy/Russia	oil price Russia energy Ukraine gas country production supply crude production cut Moscow region since fall world natural western military
STM	energy	price oil energy production gas supply crude cost industrial demand cut commodity market fuel solar project natural global plants mine
STM	Russia/conflict	Russian Ukraine country region south India international Moscow western war president military flight United world European force eastern Africa minister
EeLDA	energy	price oil production energy gas supply cost crude industrial cut demand project fall commodity natural power food global lower fuel
EeLDA	Europe/Russia	Russia government country president tax Ukraine politics Europe minister national budget officials leaders support economy Union prime vote foreign independence

Table 3: A comparison of the 20 highest probability words in the closest energy-related topic and Europe-related topic.

For LDA there is only one combined topic. For STM and EeLDA, the "energy" topics are generally alike. The "Europe/Russia" topic from EeLDA tends to be much more Europe-related and captures the Scotland Independence Vote compared to the "Russia/conflict" topic from STM. These findings assert our view that EeLDA improves the structure of the topic model and makes it more time-dependent.

5 Conclusion

We have developed a time-dependent topic model which analyzes temporal text documents with known exogenous processes. The new model takes both endogenous and exogenous processes into account, and applies Markov chain Monte Carlo sampling for calibration. We have demonstrated that this model better fits temporal documents in terms of perplexity, and extracts well information from job advertisements and financial news articles. We suggest that a possible direction for the future could be analyzing the contents of temporal documents so that they could predict the trends of related exogenous processes.

Acknowledgements

This research was conducted in collaboration with the Workforce Science Project of the Searle Center for Law, Regulation and Economic Growth at Northwestern University. We are indebted to Deborah Weiss, Director, Workforce Science Project, for introducing us to the subject of workforce and providing guidance. We are also very grateful for the help and data from DirectEmployers Association.

Reference

D. M. Blei. (2011). Introduction to Probabilistic Topic Models. *Communications of the ACM*, 55:77-84.

D. M. Blei, and J. Lafferty. (2009). Topic Models. In *Text Mining: Theory and Applications*, Taylor and Francis, London, UK.

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

P. Coy. (2014). America's Low-Paying Recovery: More Jobs Than Ever, Worse Wages. *Bloomberg Business*. Retrieved from http://www.bloomberg.com/bw/articles/2014-08-11/report-new-jobs-in-u-dot-s-dot-offer-lower-wages-than-before-recession

T. Ferguson. (1973). Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics*, 1:209-230.

A. Lowrey. (2014). Recovery Has Created Far More Low-Wage Jobs Than Better-Paid Ones. *The New York Times*. Retrieved from http://www.nytimes.com/2014/04/28/business/economy/recovery-has-created-far-more-low-wage-jobs-than-better-paid-ones.html

C. Lucas, R. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley. (2015). Computer Assisted Text Analysis for Comparative Politics. *Political Analysis*, forthcoming.

I. Pruteanu-Malinici, L. Ren, J. Paisley, E. Wang, and L. Carin. (2010). Hierarchical Bayesian Modeling of Topics in Time-Stamped Documents, *IEEE Trans. Pattern Analysis Machine Intelligence*, 32:996-1011.

L. Ren, D. B. Dunson, and L. Carin. (2008). The Dynamic Hierarchical Dirichlet Process. *International Conference on Machine Learning*, Helsinki, Finland.

C. P. Robert, and G. Casella. (2004). *Monte Carlo Statistical Methods*, 2nd ed., Springer, New York.

G. O. Roberts, and J. S. Rosenthal. (2006). Harris Recurrence of Metropolis-Within-Gibbs and Trans-Dimensional Markov Chains. *The Annals of Applied Probability*, 16:2123-2139.

M. E. Roberts, B. M. Stewart, and E. M. Airoldi. (2015). *A Model of Text for Experimentation in the Social Sciences*. Working paper. Retrieved from http://scholar.harvard.edu/files/bstewart/files/stm.pdf

M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, D. G. Rand. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58:1064-1082.

J. Sethuraman. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4:639-650.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. (2005). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101:1566-1582.