Text Mining for Economics and Finance
Variational Inference: Theory and Applications

Stephen Hansen

Introduction

We have seen that directly computing the posterior for LDA is intractable.

First option is to stochastically approximate posterior by repeatedly sampling from a Markov chain formed by draws from conditional distributions (Gibbs sampling).

We now cover a more recently popularized approach in Bayesian statistics called variational inference.

As with MCMC, many applications, but we focus on LDA for concreteness. Original article used variational approach.

General Idea

Approximate the true posterior distribution with a simpler functional form that depends on a set of variational parameters.

Then optimize the approximate posterior with respect to the variational parameters so that it lies "close to" the true posterior.

The inference problem becomes an optimization problem.

But note that the family of distributions used to approximate the posterior typically does not include the true posterior.

True and Approximate Distributions

Suppose we have observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$ (treat any parameters as fixed for now).

Let $p(\mathbf{x}, \mathbf{z})$ be their joint distribution.

Assume that $p(\mathbf{z} \mid \mathbf{x})$ is intractable to compute, for example because the latent space is too high-dimensional.

Let $q(\mathbf{z})$ be an approximate distribution over the latent variables. It will depend on variational parameters we suppress for now.

Kullback-Leibler Divergence

To measure the closeness of $p(\mathbf{z} \mid \mathbf{x})$ and $q(\mathbf{z})$, we can use the Kullback-Leibler divergence:

$$\mathbb{KL}(p \parallel q) = \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x}) \log \left[\frac{p(\mathbf{z} \mid \mathbf{x})}{q(\mathbf{z})} \right] \qquad \text{(forwards KL)}$$

$$\mathbb{KL}(q \parallel p) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] \qquad \text{(reverse KL)}$$

Forwards KL:

1. "Zero-avoiding"
2. Used in expectation propagation

Reverse KL:

1. "Zero-forcing" → better when multi-modal posterior
2. Used in variational inference

Expressing KL Divergence

We can express reverse KL as:

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x})/p(\mathbf{x})} \right] =$$

$$\underbrace{\log\left[p(\mathbf{x})\right]}_{\text{log evidence}} - \underbrace{\left\{ \sum_{\mathbf{z}} q(\mathbf{z}) \log\left[p(\mathbf{z}, \mathbf{x})\right] - \sum_{\mathbf{z}} q(\mathbf{z}) \log\left[q(\mathbf{z})\right] \right\}}_{\text{evidence lower bound (ELB)}} \geq 0.$$

Expressing KL Divergence

We can express reverse KL as:

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} \right] = \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[\frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x}) / p(\mathbf{x})} \right] =$$

$$\underbrace{\log \left[p(\mathbf{x}) \right]}_{\text{log evidence}} - \underbrace{\left\{ \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[p(\mathbf{z}, \mathbf{x}) \right] - \sum_{\mathbf{z}} q(\mathbf{z}) \log \left[q(\mathbf{z}) \right] \right\}}_{\text{evidence lower bound (ELB)}} \geq 0.$$

$\log [p(\mathbf{x})]$ is hard to compute, but does not depend on $q(\mathbf{z})$.

Minimize KL divergence = maximize ELB, which we can usually compute.

ELB is expected complete data log-likelihood plus entropy of approximating distribution.

Comparison to EM Algorithm

In the EM algorithm, we take the expectation of the complete data log-likelihood with respect to the posterior distribution over $\mathbf{z}$ given fixed parameter values.

We use the true $p(\mathbf{z} \mid \mathbf{x})$ rather than the approximation $q(\mathbf{z})$, so the KL divergence is zero.

The ELB computed using true $p(\mathbf{z} \mid \mathbf{x})$ equals $\log [p(\mathbf{x})]$.

By contrast, with variational inference the ELB is not tight, but we want to make it as tight as possible.

Mean Field Approximation

The space of potential approximating distributions is large, so in practice some restrictions are made.

In mean-field approximation, we assume that q factorizes as

$$q(\mathbf{z}) = \prod_i q_i(z_i).$$

For simplicity, we assume each latent variable is independent, but can also have independent blocks of latent variables.

Independence assumptions implicit in mean field approximation generally not present in true posterior.

ELB with Mean Field

Consider dependence of ELB just on $q_i(z_i)$:

$$\sum_{\mathbf{z}} q(\mathbf{z}) \log\left[p(\mathbf{z},\mathbf{x})\right] = \sum_{z_i} q_i(z_i)\mathbb{E}_{\mathbf{z}_{-i}}\left(\log\left[p(z_i,\mathbf{z}_{-i},\mathbf{x})\right]\right)$$

$$\begin{aligned}\sum_{\mathbf{z}} q(\mathbf{z}) \log\left[q(\mathbf{z})\right] &= \sum_{\mathbf{z}} \prod_{i=1}^{M} q_i(z_i)\left(\sum_{i=1}^{M} \log[q_i(z_i)]\right) \\ &= \sum_{i=1}^{M} \sum_{z_i} q_i(z_i) \log[q_i(z_i)] \\ &= \sum_{z_i} q_i(z_i) \log\left[q_i(z_i)\right] + \text{constant}\end{aligned}$$

Optimal Update

The optimal distribution is the solution to

$$\max_{q_i(z_i)} \sum_{z_i} q_i(z_i) \left[\mathbb{E}_{\mathbf{z}_{-i}} \left(\log \left[p(z_i, \mathbf{z}_{-i}, \mathbf{x}) \right] \right) - \log \left[q_i(z_i) \right] \right] \text{ s.t. } \sum_{z_i} q_i(z_i) = 1.$$

Optimal $q_i(z_i)$ satisfies

$$\mathbb{E}_{\mathbf{z}_{-i}} \left(\log \left[p(z_i, \mathbf{z}_{-i}, \mathbf{x}) \right] \right) - \log \left[q_i^*(z_i) \right] - 1 + \lambda = 0.$$

so that

$$q_i^*(z_i) \propto \exp \left(\mathbb{E}_{\mathbf{z}_{-i}} \left\{ \log \left[p(z_i, \mathbf{z}_{-i}, \mathbf{x}) \right] \right) \right\} .$$

Inference

The optimal update equation is a function of $q_j(z_j)$ for $j \neq i$.

Coordinate ascent algorithm: update each q_i term holding constant the current values of q_{-i}.

Use optimized q_i to approximate posterior distribution.

Interpretation

Form true conditional posterior $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$ (only need to consider Markov blanket of z_i), then take expectation with respect to approximate distribution over the conditioning variables.

Close relationship to Gibbs sampling:

- In Gibbs sampling, we repeatedly sample values from $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$ to simulate true joint distribution.
- In variational inference, we instead average over $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x})$ rather than take samples.
- Benefit is that analytical averaging “stands in” for collecting many samples.
- But when z_i is strongly correlated with neighboring nodes, averaging distorts the estimated marginal $q_i(z_i)$.

Gibbs Sampling / Variational Inference

Advantages of sampling:

1. Typically easier to derive sampling algorithms
2. More accurate, especially for approximating features of posterior distribution beyond the mode

Advantages of variational inference:

1. Faster, especially when optimized (coordinate ascent not the only algorithm)
2. Deterministic
3. Convergence easy to assess

Variational Bayes

Now suppose we wish to approximate posterior over both latent variables $\mathbf{z}$ and parameters $\boldsymbol{\theta}$ given data $\mathbf{x}$.

Mean field assumption is to approximate $p(\boldsymbol{\theta}, \mathbf{z} \mid \mathbf{x})$ with $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) q_{\mathbf{z}}(\mathbf{z})$, or $q(\boldsymbol{\theta}) \prod_i q_i(z_i)$ given conditional independence of latent variables.

Can implement VBEM algorithm by alternating between updating $q_i(z_i)$ given $q(\boldsymbol{\theta})$ (VB E-step), and updating $q(\boldsymbol{\theta})$ given $q_i(z_i)$ (VB M-step).

Distinction between latent variables and parameters becomes rather artificial, both are treated as unknown quantities and iteratively updated.

Variational Bayes and LDA

We can estimate LDA via Variational Bayes using the mean field approximation

$$p(\Theta, B, \mathbf{z} \mid \mathbf{w}) \approx \prod_{k=1}^{K} q(\beta_k \mid \lambda_k) \prod_{d=1}^{D} \left[q(\theta_d \mid \gamma_d) \prod_{n=1}^{N_d} q(z_{d,n} \mid \phi_{d,n}) \right]$$

where:

- β_k is Dirichlet with parameters λ_k
- θ_d is Dirichlet with parameters γ_d
- $z_{d,n}$ is multinomial with parameters $\phi_{d,n}$

λ_k, γ_d, and $\phi_{d,n}$ are variational parameters we iteratively update according to the mean-field formula above.

Placing variational distributions in the same family as their priors is without loss of generality within exponential family (see Wainwright and Jordan 2008).

Update for γ_d

Recall from Gibbs sampling slides that

$$\theta_d \mid \mathbf{z}_d \sim \mathrm{Dir}\left([\alpha + \sum_n \mathbb{1}(z_{d,n} = k)]_{k=1}^K \right)$$

so

$$\log [p(\theta_d \mid \cdot)] = \sum_k \left[\alpha - 1 + \sum_n \mathbb{1}(z_{d,n} = k) \right] \log (\theta_{d,k}) + \text{constant}.$$

Update for γ_d

Recall from Gibbs sampling slides that

$$\theta_d \mid \mathbf{z}_d \sim \mathrm{Dir}\left([\alpha + \sum_n \mathbb{1}(z_{d,n} = k)]_{k=1}^{K} \right)$$

so

$$\log\left[p(\theta_d \mid \cdot)\right] = \sum_k \left[\alpha - 1 + \sum_n \mathbb{1}(z_{d,n} = k)\right] \log\left(\theta_{d,k}\right) + \text{constant}.$$

Taking expectations (ignoring constant) gives

$$\mathbb{E}_{\mathbf{z}_d}[\,\log\left[p(\theta_d \mid \cdot)\right]\,] = \sum_k \left[\alpha - 1 + \sum_n \phi_{d,n,k}\right] \log\left(\theta_{d,k}\right)$$

so optimal update is

$$\gamma^*_{d,k} = \alpha + \sum_n \phi_{d,n,k}.$$

Update for λ_k

Recall from Gibbs sampling slides that

$$\beta_k \mid \mathbf{z}, \mathbf{w} \sim \text{Dir}\left([\eta + \sum_d \sum_n \mathbb{1}(z_{d,n} = k)\mathbb{1}(w_{d,n} = v)]_{k=1,v=1}^{K,V} \right)$$

Again taking expectations of log probability, optimal update is

$$\lambda^*_{k,v} = \eta + \sum_d \sum_n \phi_{d,n,k} \mathbb{1}(w_{d,n} = v).$$

Updates for both θ_d and β_k very similar to those for Gibbs sampling, but replacing actual with expected counts.

UPDATE FOR $\phi_{d,n}$

From the mean field formula and previous results on Gibbs sampling,

$$\phi_{d,n,k} \propto \exp\left(\mathbb{E}\left[\,\log(\beta_{k,v_{d,n}}\theta_{d,k})\,\right]\right).$$

Result on Dirichlet: $\mathbb{E}[\,\log(\theta_i)\,] = \Psi(\alpha_i) - \Psi(\sum_i \alpha_i)$, so

$$\phi^*_{d,n,k} \propto \exp\left(\Psi\left(\lambda_{k,v_{d,n}}\right) - \Psi\left(\sum_v \lambda_{k,v}\right) + \Psi\left(\gamma_{d,k}\right)\right).$$

(Ψ function is derivative of $\log(\Gamma)$, implemented in many scientific computing packages).

Overall Algorithm

Seed $\phi^1_{d,n,k} = 1/k$. Then at iteration s:

1. For each topic k (or randomly seed if $s = 1$)

$$\lambda^{s+1}_{k,v} = \eta + \sum_d \sum_n \phi^s_{d,n,k} \mathbb{1}(w_{d,n} = v)$$

2. For each document d

   2.1 $\gamma^{s+1}_{d,k} = \alpha + \sum_n \phi^s_{d,n,k}$

   2.2 For each word n in document d

$$\phi^{s+1}_{d,n,k} \propto \exp\left(\Psi\left(\lambda^{s+1}_{k,v_{d,n}}\right) - \Psi\left(\sum_v \lambda^{s+1}_{k,v}\right) + \Psi\left(\gamma^{s+1}_{d,k}\right)\right)$$

3. Check convergence of ELB, if not then proceed to iteration $s + 1$

Model Selection

Another advantage of variational inference over MCMC is the optimized ELB provides an estimate of the log evidence $\log\left[p(\mathbf{x} \mid K)\right]$, which we can use for model selection.

We can run the above algorithm for different values of K and compare the bound across them.

We need to add a $\log(K!)$ term to the optimized bound to account for multiple modes.

Applications

One of the main advantages LDA is that is provides a basis for more complex probabilistic models for text.

Some of the best known of these drop the Dirichlet prior on term or topic distributions and replace it with the logistic normal, which makes sampling algorithms more difficult to build.

Variational inference used instead, although here we will discuss the structure of the models but not their inference.

Extensions of LDA:

1. Dynamic topic model (Blei and Lafferty 2006).
2. Structural topic model (Roberts et. al. 2016).

Dynamic Topic Model

In LDA, the count of terms within documents are independent and identically distributed given the Dirichlet prior on θ_d.

This rules out all dependencies across texts, which is an unnatural assumption in many social science models (more on this later).

The dynamic topic model introduces time-series dependencies into the data generating process: each time period has a separate topic model, and time periods are linked via smoothly evolving parameters.

These dependencies can be present in the distributions over terms or in the prior distribution from which document-topic distributions are drawn.

Latent Dirichlet Allocation—Original

1. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet($\boldsymbol{\alpha}$).
2. Each word $w_{d,n}$ in document d is generated from a two-step process:
   2.1 Draw topic assignment $z_{d,n}$ from $\boldsymbol{\theta}_d$.
   2.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.

Dynamic Topic Model I

Suppose all documents carry a time stamp t (daily, monthly, quarterly, yearly).

1. Draw $\boldsymbol{\theta}_d$ independently for $d = 1, \ldots, D$ from Dirichlet($\boldsymbol{\alpha}$).
2. Generate parameter vector $b_{t,k} \in \mathbb{R}^V$ where $b_{t,k} \sim \mathcal{N}(b_{t-1,k}, \sigma^2 I_V)$.
3. Form topics according to $\beta_{t,k,v} = \frac{\exp(b_{t,k,v})}{\sum_v \exp(b_{t,k,v})}$.
4. Each word $w_{d,n}$ in document d is generated from a two-step process:
    4.1 Draw topic assignment $z_{d,n}$ from $\boldsymbol{\theta}_d$.

    4.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.

This model captures changes in the words used when discussing topics; topical content of documents remains iid.

Dynamic Topic Model II

Suppose all documents carry a time stamp t (daily, monthly, quarterly, yearly).

1. Generate parameter vector $a_t \in \mathbb{R}^K$ where $a_t \sim \mathcal{N}(a_{t-1}, \delta_1^2 I_K)$.
2. For each document $d = 1, \ldots, D_t$ generate $c_{t,d} \sim \mathcal{N}(a_t, \delta_2^2 I_K)$.
3. Form document-topic distributions according to $\theta_{t,d,k} = \frac{\exp(c_{t,d,k})}{\sum_k \exp(c_{t,d,k})}$.
4. Each word $w_{d,n}$ in document d is generated from a two-step process:
   4.1 Draw topic assignment $z_{d,n}$ from $\boldsymbol{\theta}_d$.
   4.2 Draw $w_{d,n}$ from $\boldsymbol{\beta}_{z_{d,n}}$.

This model captures the evolution of topic content within documents; word distributions induced by topics remains fixed.

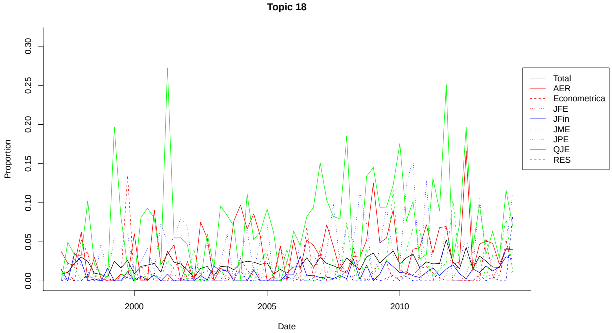
Example with Journal Abstracts

The following results are for a 20-topic version of DTM I on the abstracts of eight top economics journals from 1997:II to 2014:II (thanks to Julian Ashwin for collecting data and estimating):

- The Quarterly Journal of Economics
- Journal of Political Economy
- American Economic Review
- Econometrica
- Journal of Financial Economics
- Journal of Finance
- Review of Economic Studies
- Journal of Monetary Economics

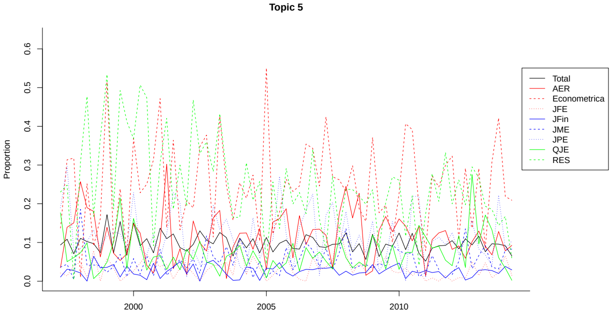
Example with Journal Abstracts

$$\begin{pmatrix}\text{school}\\ \text{educ}\\ \text{network}\\ \text{program}\\ \text{student}\\ \text{famili}\\ \text{score}\\ \text{year}\\ \text{black}\\ \text{class}\end{pmatrix} \to \begin{pmatrix}\text{school}\\ \text{educ}\\ \text{network}\\ \text{famili}\\ \text{student}\\ \text{program}\\ \text{colleg}\\ \text{class}\\ \text{group}\\ \text{year}\end{pmatrix} \to \begin{pmatrix}\text{school}\\ \text{educ}\\ \text{famili}\\ \text{group}\\ \text{student}\\ \text{network}\\ \text{program}\\ \text{black}\\ \text{score}\\ \text{white}\end{pmatrix} \to \begin{pmatrix}\text{school}\\ \text{student}\\ \text{network}\\ \text{group}\\ \text{famili}\\ \text{educ}\\ \text{social}\\ \text{peer}\\ \text{score}\\ \text{year}\end{pmatrix} \to \begin{pmatrix}\text{school}\\ \text{group}\\ \text{network}\\ \text{student}\\ \text{peer}\\ \text{social}\\ \text{educ}\\ \text{year}\\ \text{score}\\ \text{famili}\end{pmatrix}$$



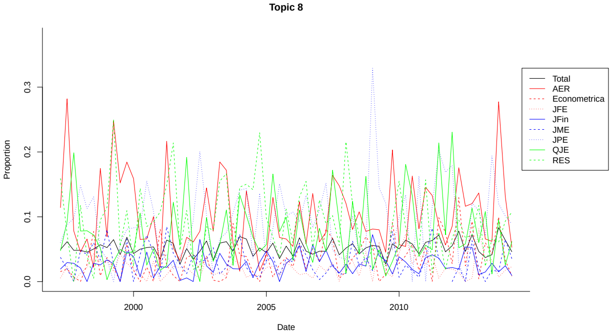
Example with Journal Abstracts

$$\begin{pmatrix}\text{equilibrium}\\ \text{agent}\\ \text{inform}\\ \text{contract}\\ \text{game}\\ \text{auction}\\ \text{optim}\\ \text{may}\\ \text{equilibria}\\ \text{effici}\end{pmatrix} \rightarrow \begin{pmatrix}\text{inform}\\ \text{equilibrium}\\ \text{agent}\\ \text{game}\\ \text{contract}\\ \text{optim}\\ \text{auction}\\ \text{equilibria}\\ \text{effici}\\ \text{player}\end{pmatrix} \rightarrow \begin{pmatrix}\text{agent}\\ \text{equilibrium}\\ \text{inform}\\ \text{game}\\ \text{optim}\\ \text{contract}\\ \text{auction}\\ \text{equilibria}\\ \text{effici}\\ \text{privat}\end{pmatrix} \rightarrow \begin{pmatrix}\text{inform}\\ \text{equilibrium}\\ \text{agent}\\ \text{game}\\ \text{contract}\\ \text{optim}\\ \text{effici}\\ \text{auction}\\ \text{set}\\ \text{player}\end{pmatrix} \rightarrow \begin{pmatrix}\text{inform}\\ \text{agent}\\ \text{equilibrium}\\ \text{game}\\ \text{privat}\\ \text{optim}\\ \text{contract}\\ \text{equilibria}\\ \text{player}\\ \text{effici}\end{pmatrix}$$



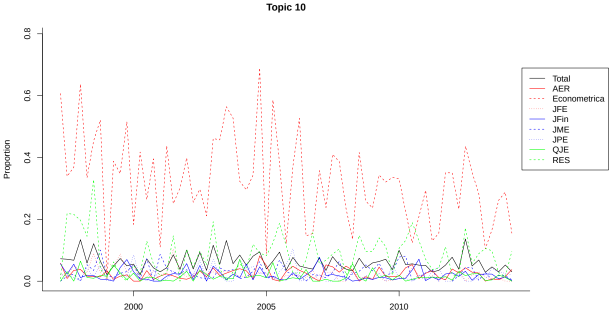
Example with Journal Abstracts

$$\begin{pmatrix}\text{cost}\\ \text{product}\\ \text{competit}\\ \text{incent}\\ \text{effici}\\ \text{contract}\\ \text{demand}\\ \text{may}\\ \text{effort}\\ \text{welfar}\end{pmatrix} \rightarrow \begin{pmatrix}\text{product}\\ \text{cost}\\ \text{competit}\\ \text{good}\\ \text{industri}\\ \text{demand}\\ \text{profit}\\ \text{qualiti}\\ \text{effort}\\ \text{consum}\end{pmatrix} \rightarrow \begin{pmatrix}\text{product}\\ \text{cost}\\ \text{competit}\\ \text{consum}\\ \text{good}\\ \text{industri}\\ \text{demand}\\ \text{qualiti}\\ \text{entri}\\ \text{welfar}\end{pmatrix} \rightarrow \begin{pmatrix}\text{cost}\\ \text{product}\\ \text{competit}\\ \text{consum}\\ \text{industri}\\ \text{demand}\\ \text{entri}\\ \text{welfar}\\ \text{qualiti}\\ \text{good}\end{pmatrix} \rightarrow \begin{pmatrix}\text{product}\\ \text{cost}\\ \text{competit}\\ \text{good}\\ \text{demand}\\ \text{consum}\\ \text{entri}\\ \text{qualiti}\\ \text{export}\\ \text{industri}\end{pmatrix}$$



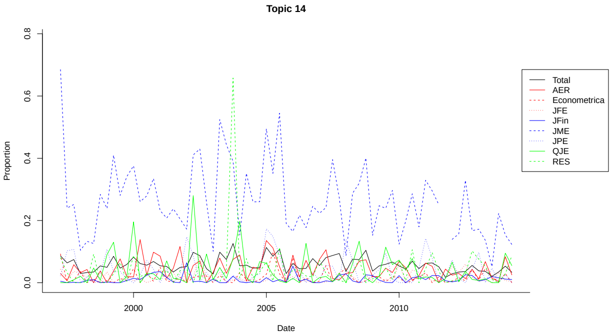
Example with Journal Abstracts

$$\begin{pmatrix}\text{estim}\\ \text{test}\\ \text{distribut}\\ \text{paramet}\\ \text{condit}\\ \text{asymptot}\\ \text{method}\\ \text{sampl}\\ \text{function}\\ \text{regress}\end{pmatrix} \rightarrow \begin{pmatrix}\text{estim}\\ \text{distribut}\\ \text{function}\\ \text{test}\\ \text{paramet}\\ \text{method}\\ \text{asymptot}\\ \text{condit}\\ \text{sampl}\\ \text{consist}\end{pmatrix} \rightarrow \begin{pmatrix}\text{estim}\\ \text{method}\\ \text{distribut}\\ \text{function}\\ \text{condit}\\ \text{test}\\ \text{paramet}\\ \text{asymptot}\\ \text{variabl}\\ \text{consist}\end{pmatrix} \rightarrow \begin{pmatrix}\text{estim}\\ \text{distribut}\\ \text{function}\\ \text{paramet}\\ \text{condit}\\ \text{method}\\ \text{identif}\\ \text{test}\\ \text{asymptot}\\ \text{variabl}\end{pmatrix} \rightarrow \begin{pmatrix}\text{estim}\\ \text{method}\\ \text{distribut}\\ \text{condit}\\ \text{test}\\ \text{paramet}\\ \text{function}\\ \text{identif}\\ \text{asymptot}\\ \text{propos}\end{pmatrix}$$



Example with Journal Abstracts

$$\begin{pmatrix} \text{polici} \\ \text{monetari} \\ \text{rate} \\ \text{inflat} \\ \text{exchang} \\ \text{money} \\ \text{currenc} \\ \text{real} \\ \text{chang} \\ \text{nomin} \end{pmatrix} \rightarrow \begin{pmatrix} \text{polici} \\ \text{rate} \\ \text{monetari} \\ \text{inflat} \\ \text{exchang} \\ \text{interest} \\ \text{money} \\ \text{real} \\ \text{scienc} \\ \text{currenc} \end{pmatrix} \rightarrow \begin{pmatrix} \text{polici} \\ \text{monetari} \\ \text{inflat} \\ \text{rate} \\ \text{money} \\ \text{exchang} \\ \text{shock} \\ \text{rule} \\ \text{interest} \\ \text{real} \end{pmatrix} \rightarrow \begin{pmatrix} \text{polici} \\ \text{monetari} \\ \text{inflat} \\ \text{rate} \\ \text{shock} \\ \text{interest} \\ \text{real} \\ \text{nomin} \\ \text{optim} \\ \text{exchang} \end{pmatrix} \rightarrow \begin{pmatrix} \text{polici} \\ \text{rate} \\ \text{monetari} \\ \text{exchang} \\ \text{inflat} \\ \text{nomin} \\ \text{money} \\ \text{shock} \\ \text{interest} \\ \text{real} \end{pmatrix}$$



Structural Topic Model

A typical application of topic modeling in the social sciences first estimates LDA, then uses estimates of θ_d as the dependent variable in an regression on covariates to test whether different types of documents have different content.

This is contradictory because documents are assumed to be generated by a statistical process that we subsequently reject.

The structural topic model (STM) of Roberts et. al. (2016) explicitly introduces covariates into a topic model, and allows one to estimate the impact of document-level covariates on topic content and prevalence as part of the topic model itself.

There is a user-friendly R package stm for implementing it and additional information on `www.structuraltopicmodel.com`.

Topic Prevalence vs. Content

The process for generating individual words is the same as for plain LDA conditional on the $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_d$ terms.

However both objects can depend on potentially different sets of document-level covariates:

1. Topic Prevalence. Each document has P attributes $\mathbf{x}_d$ that affect the likelihood of discussing topic k.
2. Topic Content. Each document has A attributes $\mathbf{v}_d$ that affect the likelihood of discussing term v overall, and of discussing it within topic k.

The generation of the $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_d$ terms is via multinomial logistic regression, similar to the dynamic topic model.

Topic Prevalence Model

For each topic, draw $\gamma_k \sim \mathcal{N}_P(\mathbf{0}, \sigma_k^2 \mathbf{I}_P)$.

$\gamma_{k,p}$ is a coefficient that determines the effect of covariate p on the use of topic k. Its prior shrinks it towards 0 but does not induce sparsity.

$\boldsymbol{\mu}_d \in \mathbb{R}^{K-1}$ is the mean of the logistic normal from which $\boldsymbol{\theta}_d$ is drawn, where $\mu_{d,k} = \sum_p \gamma_{k,p} x_{d,p}$.

First draw $\boldsymbol{\eta}_d \sim \mathcal{N}_{K-1}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma})$, then form $\theta_{d,k} = \frac{\exp(\eta_{d,k})}{\sum_k \exp(\eta_{d,k})}$ where $\eta_{d,K}$ is normalized to 0.

When the covariates are constants, this becomes the correlated topic model (Blei and Lafferty 2007). The $\boldsymbol{\Sigma}$ matrix captures that some topics may be more or less likely to be discussed jointly.

Topic Content Model

$$\beta_{d,k,v} = \frac{\exp(m_v + \kappa'_{k,v} + \kappa''_{\mathbf{y}_d,v} + \kappa'''_{\mathbf{y}_d,k,v})}{\sum_v \exp(m_v + \kappa'_{k,v} + \kappa''_{\mathbf{y}_d,v} + \kappa'''_{\mathbf{y}_d,k,v})}$$

1. m_v is the baseline log-transformed rate of term v. Effect of covariates will be to generate deviations from these baseline frequencies.
2. κ' terms capture propensity of term v to appear in topic k across all documents.
3. κ'' terms capture propensity of covariates to generate term v.
4. κ''' terms capture propensity of covariates to generate term v in topic k.

Laplace priors placed on all κ terms to promote sparsity in the estimates.

STM vs LDA

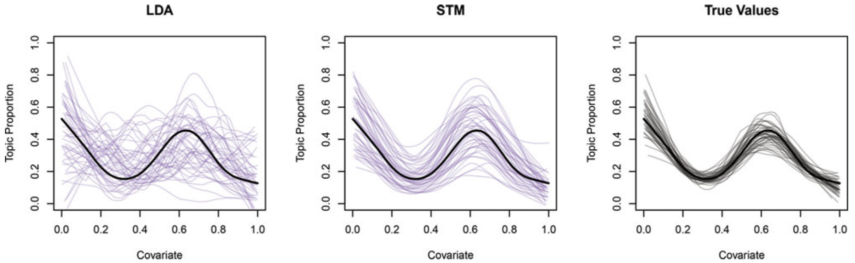


Figure 2. Plot of fitted covariate-topic relationships from 50 simulated datasets using LDA and the proposed structural topic model of text. The third panel shows the estimated relationship using the true values of the topic and thus only reflects sampling variability in the data-generating process.

Example

Gadarian and Albertson (2014) conduct an experiment in which they cue a treatment group to worry about immigration and a control group to think about immigration, and record open-ended responses.

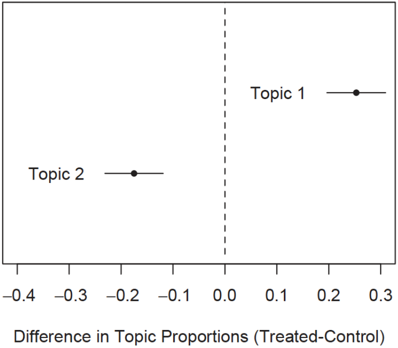
Estimate three-topic structural topic model on the corpus of survey responses.

Covariates: (i) political party affiliation; (ii) treatment dummy; (iii) interaction effect.

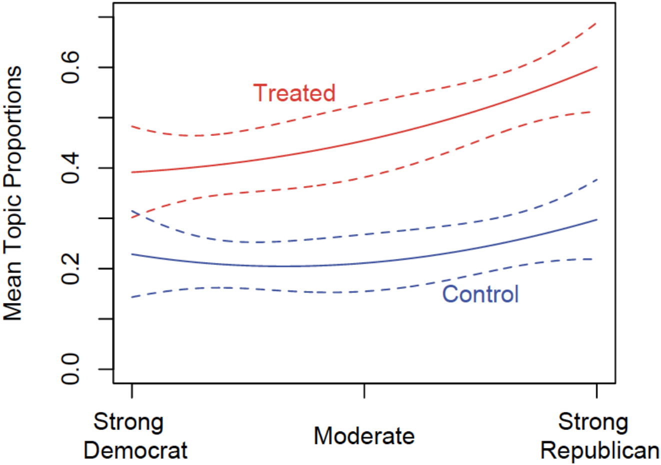
Topic Output

Topic 1:
illeg, job, immigr, tax, pai, american, care, welfar, crime, system, secur, social, cost, health, servic, school, languag, take, us, free

Topic 2:
immigr, illeg, legal, border, need, worri, mexico, think, countri, law, mexican, make, america, worker, those, american, fine, concern, long, fenc



Effect of Covariate



Conclusion

Variational inference provides a fast, deterministic means of approximating the posterior distribution in complex models.

Particularly relevant for processing massive corpora (see code and demo in gensim package in Python).

Also used in topic models where the logistic normal replaces the Dirichlet as the prior distribution for multinomial probabilities.

Within economics, MCMC is the dominant approach to posterior approximation, and variational inference is hardly known.