Crowdsourced Statistics Advice: Topic Modeling stats.stackexchange.com

Paul Hively

M.S. Candidate

Department of Statistics

University of Chicago

May 5, 2017

Advisor: John Lafferty

Abstract

Topic models are hierarchical Bayesian models used to discover latent semantic structure within collections of documents, allowing them to be reduced from millions of words to a few dozen interpretable topics. This paper presents three closely related methods: latent Dirichlet allocation, correlated topic models, and structural topic models. I discuss the estimation challenges associated with topic modeling and compare the three methods by analyzing a collection of 182,308 posts contributed by the general public to the statistics and machine learning community website stats.stackexchange.com.

Table of Contents

Introduction

With the advent of computers and the internet, unstructured text data is generated far more quickly than it can be classified by hand. Automated techniques are needed to explore and organize these collections. This paper focuses on *topic models*, unsupervised generative probabilistic models applying hierarchical Bayesian methods to text documents to discover their underlying structure (Blei & Lafferty, 2009). Topic modeling can be understood as a dimensionality reduction technique (Crain, Zhou, Yang, & Zha, 2012) in which a vocabulary of words is mapped to a much smaller number of *topics*, distinct, interpretable features that capture the original semantic meaning of documents within a *corpus*, or collection.

Since its introduction in 2003, a classic topic modeling technique known as latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) has been extended by correlated topic models (CTM) (Blei & Lafferty, 2007) and structural topic models (STM) (Roberts, Stewart, Tingley, & Airoldi, 2013). The goal of this paper is to explore these three related methods and apply them to a novel dataset, the collection of 92,335 questions and 89,973 answers posted to the statistics and machine learning community website stats.stackexchange.com through December 11, 2016 (Stack Exchange, Inc., 2014).

Terminology and Model Assumptions

The topic models under consideration share several common terms and assumptions, which I detail here. Much of this notation is borrowed from the original LDA paper (Blei, Ng, & Jordan, 2003).

- A *corpus* $\mathcal{D}$ is a collection of D documents, $\mathcal{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_D\}$.
- A *document* $\boldsymbol{w}_d$ is a vector of N_d words, $\boldsymbol{w}_d = \{w_1, w_2, \dots, w_{N_d}\}$, which need not be distinct.
- A *word* w_n is a discrete unit of text from a vocabulary $\mathcal{V}$. Words are separated by defined characters, such as whitespace or punctuation.
- A *vocabulary* $\mathcal{V}$ is the indexed collection of all V unique words in the corpus $\mathcal{D}$.
- A *document-term matrix* T is the $D \times V$ matrix whose elements $t_{i,j}$ count the number of occurrences of word w_j within document $\boldsymbol{w}_i$.
- A *topic* $\boldsymbol{\beta}_k$ is a distribution of words w_n over the vocabulary $\mathcal{V}$. The corpus $\mathcal{D}$ is assumed to contain K distinct topics.
- Topic proportion mixtures $\boldsymbol{\theta}_d$ are the distribution of the K topics $\boldsymbol{\beta}_k$ over document $\boldsymbol{w}_d$.
- Finally, the topic assignments $z_{d,n} \in \{1, \dots, K\}$ are latent variables indicating the topic from which word w_n in document $\boldsymbol{w}_d$ was drawn.

Fundamentally, topic models attempt to estimate the conditional probability a word appears in a document given its topics, i.e. $p(w_n | z_{d,n} = k, \boldsymbol{\beta}_k)$. To achieve this, LDA, CTM, and STM all rely on the bag-of-words assumption: within any given document $\boldsymbol{w}_d$, the words are thought to be *exchangeable*, so that any two documents that contain the same words with the same frequency are equally likely to be observed, no matter the specific order in which the words appear (Blei, Ng, & Jordan, 2003). Mathematically, this means that the joint probability distribution of a set of random variables does not

change under any permutation. If vector $\boldsymbol{p}$ denotes some permutation of the indices $\{1, \dots, N\}$, then exchangeability implies:

$$P(\boldsymbol{w}_d = \{w_1, \dots, w_N\}) = P\left(\boldsymbol{w}_d = \left\{w_{p_1}, \dots, w_{p_N}\right\}\right)$$

LDA is a *generative latent variable model* (Blei & Lafferty, 2009), as are its extensions CTM and STM. Generative models specify the mechanism by which data is produced, and can be used to create synthetic datasets, as well as predict a probability for documents and terms not present in the training set (Blei, Ng, & Jordan, 2003). Latent variables are not directly observed, and serve to reduce the dimensionality of the model (Crain, Zhou, Yang, & Zha, 2012). For example, in my notation the size V vocabulary $\mathcal{V}$ is reduced to the size K set of topics $\boldsymbol{\beta}$ by inferring the latent topics $\boldsymbol{\theta}_d$ and per-word topic assignments $z_{d,n}$ from the words observed in the corpus.

Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) was introduced as an improvement to the earlier Latent Semantic Indexing (LSI) and probabilistic Latent Semantic Indexing (pLSI) methods (Blei, Ng, & Jordan, 2003). Its contribution is the addition of a document-level probability model, which provides a natural way to estimate the probability of previously-unseen documents and reduces the number of parameters that need to be estimated when training the model (Crain, Zhou, Yang, & Zha, 2012).

In LDA, documents are allowed to exhibit one or more topics mixed in various proportions, as specified by $\boldsymbol{\theta}_d$. Each document in a corpus is assumed to be generated by the following process (Blei, Ng, & Jordan, 2003):

1. For each of the K topics, generate a length V multinomial with parameter $\boldsymbol{\beta}_k \sim \mathrm{Dir}(\eta)$. This is the per-word probability vector across the vocabulary $\mathcal{V}$ for topic k.
2. For each of the D documents in corpus $\mathcal{D}$, generate the length K topic proportion distribution as multinomial with parameter $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$. This is the mixture of topics present in each document.
    a. For each of the N_d words within document $\boldsymbol{w}_d$, choose a word topic $z_{d,n} \in \{1, \dots, K\}$, where the probabilities are given by $\mathrm{Multi}(\boldsymbol{\theta}_d)$.
    b. For each of the $z_{d,n}$, choose a word $w_n \sim \mathrm{Multi}\left(\boldsymbol{\beta}_{z_{d,n}}\right)$, the per-word probability vector associated with the current topic.

The Dirichlet distributions used to generate the $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_d$ are typically assumed to have symmetric priors η and α, which induces smoothing over the vocabulary. While other formulations are possible, symmetric priors moderate the estimates of these parameters so that documents containing new vocabulary will have nonzero probability (Wallach, Mimno, & McCallum, 2009).

Finally, documents within the corpus are assumed to be exchangeable, as are the topics within each document. These assumptions allow for a convenient representation of the joint probability function as a mixture distribution due to de Finetti's representation theorem (Blei, Ng, & Jordan, 2003) and lead to the estimation process developed in the next section.

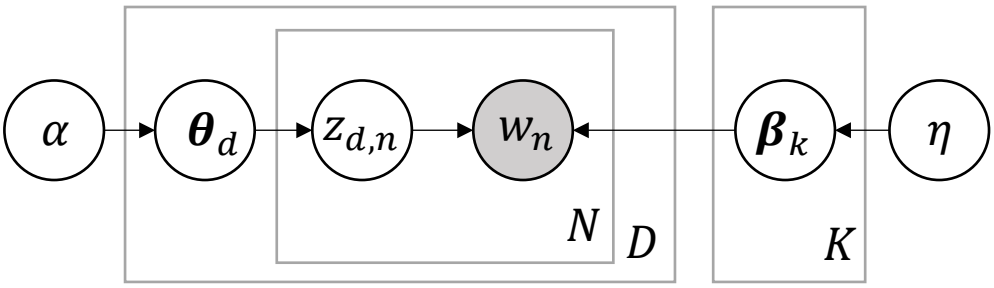

Figure 1: Graphical model for LDA. Shaded nodes are observed, unshaded nodes are hidden, arrows indicate the direction of dependence, and rectangular plates indicate replication. Adapted from Topic Models (Blei & Lafferty, 2009).

LDA Estimation

The likelihood function (Crain, Zhou, Yang, & Zha, 2012) of the corpus $\mathcal{D}$ is given as:

$$\mathcal{L} = \prod_{d=1}^{D} p(\boldsymbol{\theta}_d|\alpha)p(\boldsymbol{\beta}|\eta)\prod_{n=1}^{N} p\left(w_n|z_{d,n},\boldsymbol{\beta}\right)p\left(z_{d,n}|\boldsymbol{\theta}_d\right)$$

Recall that $p(\boldsymbol{\theta}_d|\alpha)$ and $p(\boldsymbol{\beta}|\eta)$ are Dirichlet, given by:

$$\mathrm{Dir}(\boldsymbol{\theta}|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}\prod_i \theta_i^{\alpha_i - 1}$$

And $p\left(w_n\middle|z_{d,n},\boldsymbol{\beta}\right)$ and $p(z_{d,n}|\boldsymbol{\theta}_d)$ are multinomial, which can be written using the gamma function:

$$\mathrm{Multi}(\boldsymbol{z}|\boldsymbol{\theta}) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)}\prod_i \theta_i^{z_i}$$

Finally, consider the LDA model's posterior density. Through a straightforward application of Bayes' rule (Blei & Lafferty, 2009):

$$p\left(\boldsymbol{\theta}, \boldsymbol{z}_{1:D,1:N}, \boldsymbol{\beta}|\boldsymbol{w}_{1:D,1:N}, \alpha, \eta\right) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}_{1:D}, \boldsymbol{\beta}|\boldsymbol{w}_{1:D}, \alpha, \eta)}{\int_{\boldsymbol{\beta}_{1:K}}\int_{\boldsymbol{\theta}_{1:D}}\sum_{\boldsymbol{z}}(\boldsymbol{\theta}, \boldsymbol{z}_{1:D}, \boldsymbol{\beta}|\boldsymbol{w}_{1:D}, \alpha, \eta)}$$

When written in this form, it's easy to see that the Dirichlet distribution is the conjugate prior to the multinomial. However, direct estimation of the parameters of interest is impossible; note that the joint distribution is over both the observed w_n and hidden $z_{d,n}$. This optimization problem is intractable even for individual documents in low dimensions (Crain, Zhou, Yang, & Zha, 2012).

Numerous approximate posterior estimation techniques have been explored in the literature, including several methods based on Gibbs sampling and variational inference (Blei, Ng, & Jordan, 2003; Wei & Croft, 2006; Blei & Lafferty, 2007). I will detail the *variational expectation maximization* (variational EM) procedure suggested for fitting LDA models (Blei, Ng, & Jordan, 2003).

The basic idea behind variational inference is to replace an intractable log likelihood with a simpler distribution. In Figure 1 above, the difficulty in estimating the latent parameters arises from the link between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. By deleting the edge between $\boldsymbol{\theta}_d$ and $z_{d,n}$, we break the dependence between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, enabling the use of an iterative method to find a lower bound on the log likelihood of the true model (Blei, Ng, & Jordan, 2003). Inserting free variational parameters γ and ϕ yields a family of distributions on the latent variables, which can be chosen to minimize the distance between the variational distribution and full posterior.

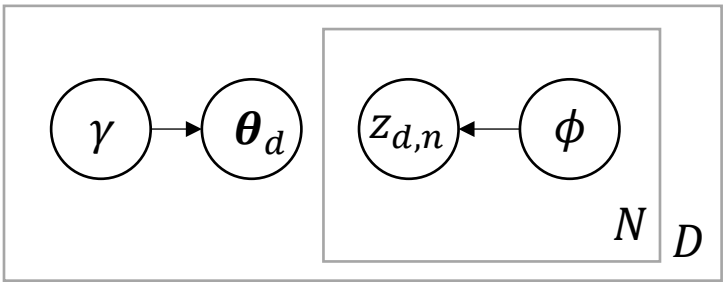

Figure 2: Variational distribution approximating the LDA posterior. Adapted from Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003).

This simplified model takes the following form (Blei, Ng, & Jordan, 2003):

$$q(\boldsymbol{\theta}, \boldsymbol{z}|\gamma, \phi) = q_1(\boldsymbol{\theta}|\gamma) \prod_{n=1}^{N} q_2(z_n|\phi_n)$$

The parameters are estimated to maximize the likelihood of the corpus $\mathcal{D}$ in the following manner (Blei, Ng, & Jordan, 2003; Crain, Zhou, Yang, & Zha, 2012).

$$(\alpha^*, \eta^*, \gamma^*, \phi^*) = \underset{\alpha,\eta}{\operatorname{argmax}} \sum_{d=1}^{D} \underset{\gamma_d,\phi_d}{\operatorname{argmax}} \mathcal{L}(\gamma_d, \phi_d)$$

1. E-step: iterate through each document $\boldsymbol{w}_d, d \in \{1, \dots, D\}$ and set γ^* and ϕ^* to minimize the Kullback-Leibler (KL) divergence between the simplified variational distribution and the posterior:

$$(\gamma^*, \phi^*) = \underset{\gamma,\phi}{\operatorname{argmin}} KL[q(\boldsymbol{\theta}, \boldsymbol{z}|\gamma, \phi) \| p(\boldsymbol{\theta}, \boldsymbol{z}|\alpha, \eta)]$$

2. M-step: set α and η to their maximum likelihood estimates based on the posterior computed in the E-step:

$$(\alpha^*, \eta^*) = \underset{\alpha,\eta}{\operatorname{argmax}} \left[\sum_{d=1}^{D} \log q(\boldsymbol{w}_d|\alpha, \eta) \right]$$

3. Repeat until some convergence criterion is satisfied.

Intuitively, the E-step provides a lower bound for the true log likelihood and the M-step updates the model parameters to optimize the approximate posterior.

Correlated Topic Models

One limitation of the latent Dirichlet allocation model is that the per-document topic distributions are generated as $\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$. This implicitly assumes that under LDA, topics within a corpus are nearly independent. Correlated topic models were introduced to extend the LDA framework to allow for correlation between topics, a much more realistic formulation (Blei & Lafferty, 2007). For example, in my dataset a question about linear regression may be more likely to also be about correlation than about random forests.

To allow for correlations between topics, CTM introduces a *logistic normal distribution* prior onto the topic proportion mixtures $\boldsymbol{\theta}_d$ in place of the Dirichlet prior used in LDA. The logistic normal function LN transforms a vector of draws from the normal distribution into a vector of probabilities and takes the following form (Blei & Lafferty, 2007):

$$\mathrm{LN}(\pi_i) = \frac{\exp(\pi_i)}{\sum_j \exp(\pi_j)} \;\; \forall \;\; \pi_i : \boldsymbol{\pi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i \in \{1, \dots, K\}$$

In other words, given a fixed number of topics K, a length K mean vector $\boldsymbol{\mu}$, and a $K \times K$ covariance matrix $\boldsymbol{\Sigma}$, the random draw $\boldsymbol{\pi}$ from the multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is connected to the topic proportions $\boldsymbol{\theta}_d$ for document d via transformation into a length K probability vector by normalization. Besides this change, the generative process for a corpus proceeds as in LDA (Blei & Lafferty, 2007):

1. For each of the K topics, generate a length V vector $\boldsymbol{\beta}_k$ with the constraint that $\sum_i \beta_{k,i} = 1$. This is the per-word probability vector across the vocabulary $\mathcal{V}$ for topic k.
2. For each of the D documents in corpus $\mathcal{D}$, generate the length K topic proportion distribution as multinomial with parameter $\boldsymbol{\theta}_d \sim \mathrm{LN}(\boldsymbol{\pi}_d); \boldsymbol{\pi}_d \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $f(\boldsymbol{\pi}_d)$ as defined above. This is the mixture of topics present in each document, with topic correlations specified by $\boldsymbol{\Sigma}$.
    a. For each of the N_d words within document $\boldsymbol{w}_d$, choose a word topic $z_{d,n} \in \{1, \dots, K\}$, where the probabilities are given by $\mathrm{Multi}(\boldsymbol{\theta}_d)$.
    b. For each of the $z_{d,n}$, choose a word $w_n \sim \mathrm{Multi}\left(\boldsymbol{\beta}_{z_{d,n}}\right)$, the per-word probability vector associated with the current topic.

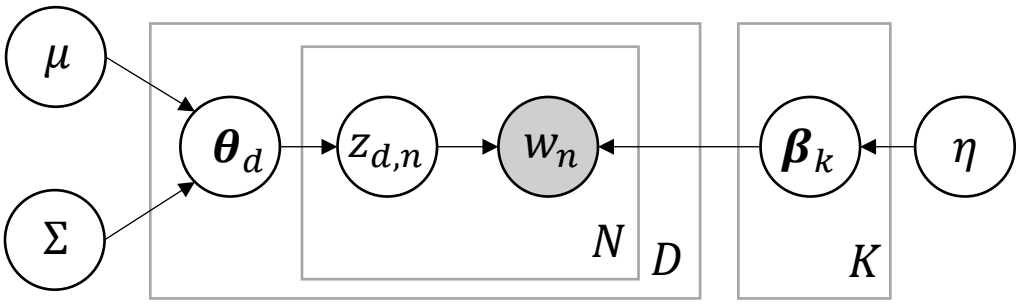

Figure 3: Graphical model for CTM. Adapted from A Correlated Topic Model of Science (Blei & Lafferty, 2007).

CTM Estimation

By Bayes' rule, the posterior of the correlated topic model is (Blei & Lafferty, 2007):

$$p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_n p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}_{1:K})}{\int p(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_n \sum_{z_n=1}^{K} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}_{1:K}) \, d\theta}$$

As before, direct computation is intractable due to the joint distribution over both the observed words w_n and latent topics z_n. There is another difficulty when estimating the parameters of a CTM: since the logistic normal distribution is not conjugate to the multinomial, the integral in the denominator cannot be evaluated analytically. Addressing these issues leads to the following modified variational expectation maximization procedure (Blei & Lafferty, 2007).

First, replace the full model with a simplified version where all variables of interest are independently parametrized so the likelihood can be factored. As before, this is accomplished by breaking the link between $\boldsymbol{\theta}$ and $\boldsymbol{z}$ and inserting free variational parameters on the latent variables (as in Figure 2). Then perform variational EM on this simplified posterior q:

1. E-step: let $\boldsymbol{\lambda}$ stand in as the variational parameter for $\boldsymbol{\mu}$, $\boldsymbol{v}$ for $\boldsymbol{\Sigma}$, and $\boldsymbol{\phi}$ for $\boldsymbol{\beta}$ and minimize the KL divergence $KL(q\|p)$ between the variational distribution q and the true posterior p, where:

$$q(\boldsymbol{\theta}_{1:K}, \boldsymbol{z}_{1:N}|\boldsymbol{\lambda}_{1:K}, \boldsymbol{v}_{1:K}, \boldsymbol{\phi}_{1:N}) = \prod_{k=1}^{K} q_1(\boldsymbol{\theta}_k|\boldsymbol{\lambda}_k, \boldsymbol{v}_k) \prod_{n=1}^{N} q_2(z_n|\boldsymbol{\phi}_n)$$

2. M-step: update $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}$ to their maximum likelihood estimates based on the variational parameters found in the E-step. After taking derivatives (Blei & Lafferty, 2007):

$$\widehat{\boldsymbol{\beta}}_i \propto \sum\nolimits_d \phi_{d,i} \boldsymbol{n}_d$$

$$\widehat{\boldsymbol{\mu}} = \frac{1}{D} \sum\nolimits_d \boldsymbol{\lambda}_d$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{D} \sum\nolimits_d [I\boldsymbol{v}_d + (\boldsymbol{\lambda}_d - \widehat{\boldsymbol{\mu}})(\boldsymbol{\lambda}_d - \widehat{\boldsymbol{\mu}})^T]$$

3. Repeat until some convergence criterion is satisfied.

The overall process is similar to variational EM in LDA. Here, the E-step step finds $\boldsymbol{\lambda}^*, \boldsymbol{v}^*, \boldsymbol{\phi}^*$ minimizing $KL(q\|p)$, and the M-step maximizes the E-step bound with respect to the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}$. However, during estimation an additional variational parameter ζ must be introduced in the E-step due to the nonconjugacy problem. This allows $\boldsymbol{\lambda}$ to be optimized via gradient ascent methods and $\boldsymbol{v}$ via Newton-Raphson iterations; see the original CTM paper (Blei & Lafferty, 2007) for details. Because of the increased number of variational parameters and the need for numerical optimization, variational EM for correlated topic models is more computationally demanding than the analogous procedure in LDA models. In practice, the tradeoff is worthwhile as the increased flexibility of a CTM leads to substantially better predictive performance compared to LDA (Blei & Lafferty, 2007).

Structural Topic Models

In both LDA and CTM, the model structure is estimated solely from word counts and co-occurrences within each document. In practice, documents often have associated *metadata*—corpus-level factors or covariates shared by groups of documents, e.g. author, date, publication source—that affect their topical content. For example, it is natural to suppose that articles published in the latest issue of the *Journal of Machine Learning Research* are more similar to one another than they are to articles published in yesterday's *New York Times*. Structural topic models (Roberts, Stewart, Tingley, & Airoldi, 2013) extend the topic model framework by relaxing the exchangeability assumption of documents within a corpus and introducing a general mechanism to account for metadata when estimating topics.

In contrast to the previously considered topic models, which assume $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are shared globally, STMs formulate these variables as functions of document-level covariates (Roberts, Stewart, Tingley, & Airoldi, 2013). The key contribution made by the STM is the addition of a *logistic normal linear model* prior with associated document covariates $\boldsymbol{X}$ (dimensions $D \times P$) onto the topic proportions $\boldsymbol{\theta}_d$, and factors $\boldsymbol{Y}$ (dimensions $D \times A$) onto the per-topic vocabularies $\boldsymbol{\beta}_k$. This allows both *topical prevalence*, the frequency with which a topic is expected to occur, and *topical content*, the word rate usage within a topic, to vary with changes in $\boldsymbol{X}$ and $\boldsymbol{Y}$ (Roberts, Stewart, & Tingley, 2016).

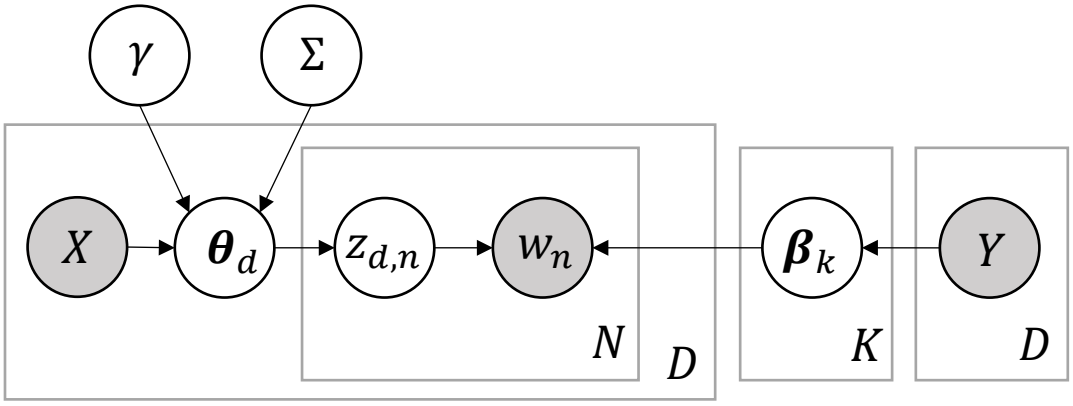

Figure 4: Graphical model for STM. Note that X and Y are observed, while the γ are latent. Adapted from A Model of Text for Experimentation in the Social Sciences (Roberts, Stewart, & Airoldi, 2016).

The model is formulated as follows (Roberts, Stewart, & Airoldi, 2016).

1. For each term in the length V vocabulary $\mathcal{V}$, compute m_v, the baseline rate of word occurrences in the corpus. Define $\boldsymbol{\kappa}^{(t)}$ (dimensions $K \times V$) as the log-transformed deviation from m_v across all the topics, $\boldsymbol{\kappa}^{(c)}$ (dimensions $A \times V$) as the log-transformed deviation from $\boldsymbol{m}_v$ over the levels of factor $\boldsymbol{Y}$, and $\boldsymbol{\kappa}^{(i)}$ (dimensions $A \times K \times V$) the log-transformed interaction effects.
2. Define a set of $\beta^* \propto \exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)}\right)$, where k indexes topics, d documents, and v terms in the vocabulary. This is essentially an additive model on the topic-vocabulary rates (Roberts, Stewart, & Airoldi, 2016). Normalizing these β^* yields the set of probability vectors $\boldsymbol{\beta}_k$:

$$\beta_{k,d,v} = \frac{\exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)}\right)}{\sum_v \exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)}\right)}$$

3. Generate $\boldsymbol{\Gamma}$ as a matrix of column vectors $\boldsymbol{\gamma}_k$, with each $\boldsymbol{\gamma}_k$ a length K multivariate normal vector.

$$\boldsymbol{\gamma}_k \sim \mathcal{N}\left(\mathbf{0}, \sigma_k^2 I_P\right) \ \forall \ k \in \{1, \dots, K\}$$

$$\boldsymbol{\Gamma} = [\gamma_1, \dots, \gamma_k]_{P,K-1}$$

4. Construct the $\boldsymbol{\theta}_d$ as a function of $\boldsymbol{\Gamma}$ and the covariates $\boldsymbol{X}$. Again, LN is the logistic normal function:

$$\boldsymbol{\theta}_d \sim \mathrm{LN}(\boldsymbol{\Gamma}'\boldsymbol{X}_d', \boldsymbol{\Sigma})$$

5. Generate the topics $\boldsymbol{z}$ and words $\boldsymbol{w}$ from multinomial distributions parametrized by the $\boldsymbol{\theta}_d$ and $\boldsymbol{\beta}_k$, as in LDA.

$$\boldsymbol{z}_{d,n} \sim \mathrm{Multi}(\boldsymbol{\theta}_d)$$

$$w_{d,n} \sim \mathrm{Multi}\left(\boldsymbol{\beta}_{z_{d,n}}\right)$$

Conceptually, the priors on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are specified as generalized linear models conditioned on sets of similar documents, as determined by the structure of the observed metadata (Roberts, Stewart, Tingley, & Airoldi, 2013). This allows for partial pooling of parameters during estimation based on the metadata relationships within the corpus. Conversely, if no outside variables are included and the prior means are fixed so that $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \cdots = \boldsymbol{\gamma}_K$, the STM model reduces to a CTM (Roberts, Stewart, & Airoldi, 2016).

STM Estimation

The STM posterior is given as (Roberts, Stewart, & Airoldi, 2016):

$$\left[\prod_{d=1}^{D} \mathcal{N}(\boldsymbol{\eta}_d | \boldsymbol{X}_d\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \left(\prod_{n=1}^{N} \mathrm{Multi}\left(z_{n,d} | \boldsymbol{\theta}_d\right) \times \mathrm{Multi}\left(w_n | \boldsymbol{\beta}_{d,z_{d,n}}\right)\right)\right] \times \prod p(\boldsymbol{\kappa}) \times \prod p(\boldsymbol{\Gamma})$$

The coefficients $\boldsymbol{\kappa}$ and $\boldsymbol{\Gamma}$ in the rightmost terms are as defined in steps 1 and 3 of the model description above. Once again, this distribution is intractable due to the coupling of the latent z_n and observed w_n, and as in CTM a further difficulty arises because the logistic normal distribution on the $\boldsymbol{\theta}_d$ is not conjugate to the multinomial. An approximation to variational EM with fast convergence can be implemented by introducing a Laplace approximation (Roberts, Stewart, & Airoldi, 2016), which replaces the intractable likelihood with an easy-to-maximize function. This variational posterior is:

$$\prod_d q_1(\boldsymbol{\eta}_d) q_2(\boldsymbol{z}_d) = \prod_d \mathcal{N}(\boldsymbol{\eta}_d | \boldsymbol{\lambda}_d, \boldsymbol{\upsilon}_d) \, \mathrm{Multi}(\boldsymbol{z}_d | \boldsymbol{\phi}_d)$$

This is maximized in terms of an approximate *evidence lower bound* (ELBO), a function equal to the KL divergence $KL(q\|p)$ up to some constant. After simplification (Roberts, Stewart, & Airoldi, 2016), the log ELBO across the documents and vocabulary is:

$$\mathcal{L}_{\text{ELBO}} = \sum_{d=1}^{D}\left[\left(\sum_{v=1}^{V} w_{d,v}\log(\boldsymbol{\theta}_d\boldsymbol{\beta}_{d,v})\right) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{\lambda}_d - \boldsymbol{\mu}_d)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\lambda}_d - \boldsymbol{\mu}_d) + \frac{1}{2}\log|\boldsymbol{v}_d|\right]$$

Estimation may then proceed as follows (Roberts, Stewart, & Airoldi, 2016).

1. E-step: Solve for the variational parameters $\boldsymbol{\lambda}_d, \boldsymbol{v}_d, \boldsymbol{\phi}_d$. After taking a second-order Taylor expansion and iteratively optimizing, the conditions are:
$$\hat{\boldsymbol{\lambda}}_d = E(\boldsymbol{\eta}_d)$$
$$\hat{\boldsymbol{v}}_d = -\nabla^2 f(E(\boldsymbol{\eta}_d)^{-1})$$
$$\hat{\phi}_{d,n,k} \propto \exp(\lambda_{d,k})\,\beta_{d,k,w_n}$$
2. M-step: update the model parameters $\boldsymbol{\kappa}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}$ with respect to the variational parameters found in the E-step. The topic-word distributions $\boldsymbol{\beta}_d$ are updated by finding the MLE estimates of $\boldsymbol{\kappa}$ based on a multinomial logistic regression of word counts of $\boldsymbol{w}_d$ and the levels of $\boldsymbol{Y}$. Next, each of the components $\boldsymbol{\gamma}_k$ in $\boldsymbol{\Gamma}$ are updated based on a linear regression on the variational parameters $\hat{\boldsymbol{\lambda}}_d$ (see step 3 in the model description above). Finally, $\boldsymbol{\Sigma}$ is updated as:
$$\hat{\boldsymbol{\Sigma}} = \frac{1}{D}\sum_{d=1}^{D}\boldsymbol{v}_d + (\boldsymbol{\lambda}_d - \boldsymbol{X}_d\hat{\boldsymbol{\Gamma}})(\boldsymbol{\lambda}_d - \boldsymbol{X}_d\hat{\boldsymbol{\Gamma}})^T$$
3. Repeat until some convergence criterion is satisfied.

With suitable initialization (Arora, et al., 2013), it has been demonstrated that the Laplace approximation approach to variational EM is quite accurate, and can greatly outperform other estimation methods in time needed to converge (Roberts, Stewart, & Airoldi, 2016; Roberts, Stewart, & Tingley, 2016).

Model Selection

When applying any topic model variant to a dataset, the primary choice a practitioner faces is selecting a suitable K, the number of topics to fit. One general approach is *sensitivity analysis* (Airoldi, et al., 2010), which examines how the macro structure of the topics, represented by $\boldsymbol{\theta}_{1:K}$ and $\boldsymbol{\beta}_{1:K}$, changes as K grows. Intuitively, as K approaches some optimal value, the topic vocabularies and their associated probabilities should stabilize. A more direct method is to compare fitted models by calculating the likelihood, lower bound objective function, or similar quantity of a holdout dataset not used during the estimation process (Wallach, Murray, Salakhutdinov, & Mimno, 2009).

The most straightforward approach is to hold out a subset of documents from the training set, in which case a standard cross-validation approach is easily implemented. Another option is *document completion*, where each document is partitioned into two sets of words $\boldsymbol{w}_d = \{\boldsymbol{w}_d^{(1)}, \boldsymbol{w}_d^{(2)}\}$ and the model is trained on $\boldsymbol{w}_d^{(1)}$ and evaluated on the held out $\boldsymbol{w}_d^{(2)}$. The literature suggests that either method can work well

provided that a suitable evaluation technique is used to estimate the held-out likelihoods (Wallach, Murray, Salakhutdinov, & Mimno, 2009).

Below, I discuss the R packages used in my analysis and a few details of their model implementation and selection routines.

`topicmodels` Package

The `topicmodels` package (Grun & Hornik, 2011) provides an R interface for the variational EM algorithm described in the above section on LDA Estimation (Blei, Ng, & Jordan, 2003). The package's LDA function requires a document-term matrix $\boldsymbol{T}_{D,V}$ and the number of topics K to fit; other settings were left at their defaults. In particular, `topicmodels` randomly initializes the topic proportions and vocabulary; this has serious implications for performance which I discuss below.

The package authors recommend using the held-out document method to choose a suitable K for modeling the corpus. The likelihood for the test data is approximated by the lower bound obtained by VEM estimation (Grun & Hornik, 2011), and *perplexity*, the geometric mean per-word likelihood of a set of test documents $\boldsymbol{w}_{\text{test}}$, is computed as (Grun & Hornik, 2011):

$$\text{Perplexity}(\boldsymbol{w}_{\mathbf{test}}) = \exp\left(-\frac{\ell(\boldsymbol{w}_{\mathbf{test}})}{\sum_{d=1}^{D}\sum_{v=1}^{V} t_{d,v}}\right)$$

Here, $t_{d,v}$ is the element of document-term matrix $\boldsymbol{T}$ counting the number of times vocabulary word $\mathcal{V}_v$ appears in document $\boldsymbol{w}_d$. Under this formulation, the optimal choice of K minimizes expected perplexity.

`stm` Package

Because the structural topic model is an extension of correlated topic models, I used the `stm` R package (Roberts, Stewart, & Tingley, 2016) to estimate both CTM and STM models. The package's STM function requires specification of a list of documents and the corpus vocabulary, which is extracted from the document-term matrix $\boldsymbol{T}_{D,V}$; the number of topics K to fit; topical prevalence covariates $\boldsymbol{X}$ and topical content factors $\boldsymbol{Y}$; and an initialization method.

The `stm` package performs covariate estimation as a distributed multinomial regression (Taddy, 2015), which essentially speeds convergence by factoring the estimation problem into a series of independent Poisson regressions. For initialization, I chose the spectral method, a deterministic method of moments algorithm that initializes the STM by finding *anchor words* that rarely co-occur and placing them in separate topics (Arora, et al., 2013). This algorithm's implementation in `stm` is very fast and in practice leads to good results in fewer iterations than either random initialization or Gibbs sampling (Roberts, Stewart, & Tingley, 2016).

The authors provide a K selection routine based on the document completion approach to likelihood estimation, plus an option to perform residual analysis to check model assumptions (Taddy, 2012).

Analyzing stats.stackexchange.com

StackExchange.com is a user-driven question-and-answer site consisting of dozens of community pages specializing in various subjects. I obtained the dataset used in this analysis from the Stack Exchange data dump website (Stack Exchange, Inc., 2014) and extracted in XML format 92,335 questions and 89,973 answers posted to the statistics and machine learning community page stats.stackexchange.com between February 2, 2009 and December 11, 2016.

Stack Exchange allows for unrestricted use of its data under the Creative Commons Share Alike 3.0 license (Stack Exchange, Inc., 2014; Creative Commons, 2007). All of the data and code used in this analysis is available at https://github.com/phively/uchicago-thesis.

Corpus Creation

In some applications, the documents within a corpus are clearly identifiable: books, articles, reviews, and so on. In this dataset, the collection of posts is split between distinct question and answer types, so I had to take care when creating the corpus. One approach is to treat all 182,308 posts as individual documents. However, since new answers are posted below a single distinct question and all of its earlier answers, this formulation introduces dependence between documents not specified by the topic models under consideration. Consequently, I defined a single document $\boldsymbol{w}_d$ as a question title, the question text, and all threaded answers, resulting in a corpus of 92,335 documents.

Due to the XML format and contents of the posts, this dataset required substantial cleanup before a suitable vocabulary could be extracted. When fitting a topic model, it is standard practice to delete *stop words*, common terms that do not have semantic meaning (Blei & Lafferty, 2009; Crain, Zhou, Yang, & Zha, 2012). In addition to removing a set of 544 common English stop words, I used the R package `xml2` to separate the post text from other data and to remove valid HTML tags, such as `<p>`, `<a href>`, and `<blockquote>`.

Since stats.stackexchange.com is a statistics and machine learning site, many posts contain code snippets and LaTeX markup. While these tokens can be very expressive and would have been interesting to include, because there are multiple ways to write equivalent statements and no clear method to automatically identify synonyms (e.g. different ways of writing the normal distribution), I elected to remove them from the dataset. After examining dozens of examples, it was apparent that in many cases, code and LaTeX markup were either so long and distinctive that they would occur only once or twice in the corpus and be filtered out anyway due to low frequency, or so ubiquitous that they could be treated like stop words (e.g. variables named X).

Stack Exchange implements LaTeX through MathJax, a JavaScript engine that formats text placed between unescaped `$$ $$` and `$ $` characters, or between `\\begin{} \\end{}` tags. I used lookahead and lookbehind regular expressions to recursively identify and strip LaTeX markup from each document. While this method is not perfect, it ran quickly and correctly identified LaTeX in the vast majority of cases. The

ideal method to extract LaTeX from the corpus would be to set up a MathJax installation that flags code for removal rather than parsing it; this remains an opportunity for future improvement.

Finally, to reduce the vocabulary to a manageable size, I removed words appearing fewer than 30 times from the corpus. This reduced the size of the cleaned posts dataset from 20,594,592 total words across a 121,804-term vocabulary to 10,011,781 total words across an 11,905-term vocabulary. The final per-document log word counts were very nearly normally distributed, with little evidence of skewness or excess kurtosis ($\hat{\gamma} = -0.02; \hat{\kappa} = 3.05$). While topic modeling does not require a prior on document-word distributions, it is nevertheless reassuring that the lengths of the observed documents appear to follow an i.i.d. process.

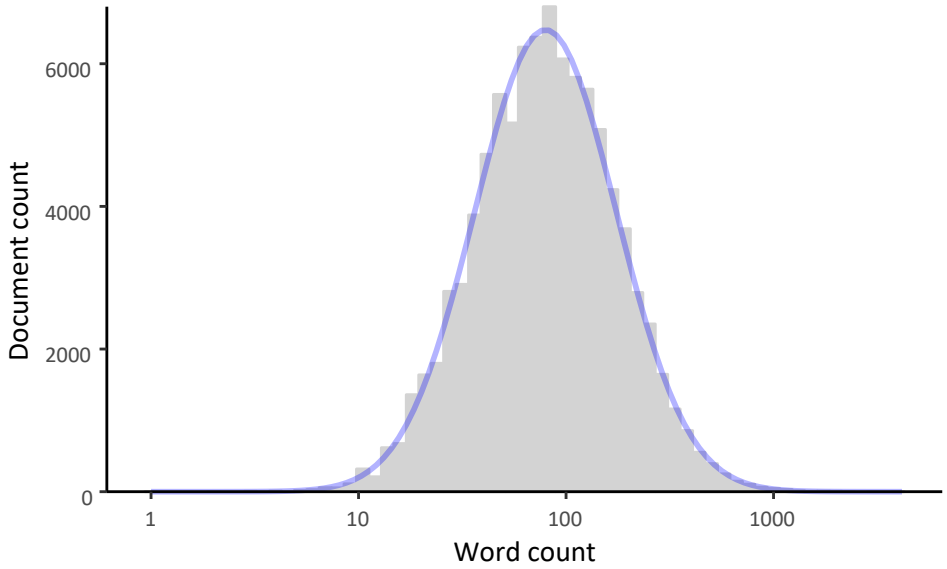

Figure 5: Histogram of per-document word counts with the base-10-exponentiated distribution $exp_{10}\left(N(1.90, 0.34^2)\right)$ *superimposed.*

LDA Results

To select the number of topics K for a latent Dirichlet allocation model, I performed 5-fold cross-validation with different values of $K \in \{10,15, \ldots ,50\}$. As mentioned in the above section on the `topicmodels` package, the LDA implementation I chose uses a random initialization, which unfortunately leads to slow model convergence. The smallest model, $K = 10$, took on average 46 iterations and 1 hour to converge, and the largest model I tested, $K = 50$, took on average 75 iterations and nearly 10.5 hours to converge. As can be seen in Figure 6, convergence time scales linearly with the number of topics K, so fitting increasingly large models quickly becomes impractical.

Fortunately, fitting huge models was not necessary, since $K = 40$ minimized out-of-sample perplexity for this dataset. I fit several additional LDA models between $K = 35$ and $K = 45$ to verify that this truly represented a minimum, and in Figure 7, we see that the decrease in perplexity with increasing K levels off near $K = 40$; it appears that any $K \geq 38$ would be a reasonable choice.

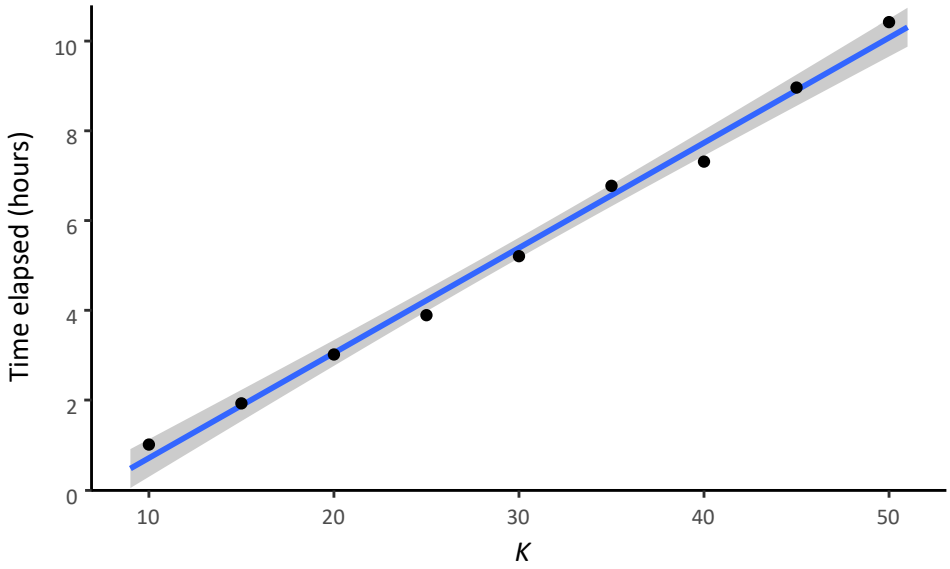

Figure 6: Time elapsed until LDA model convergence in hours as a function of the number of topics K.

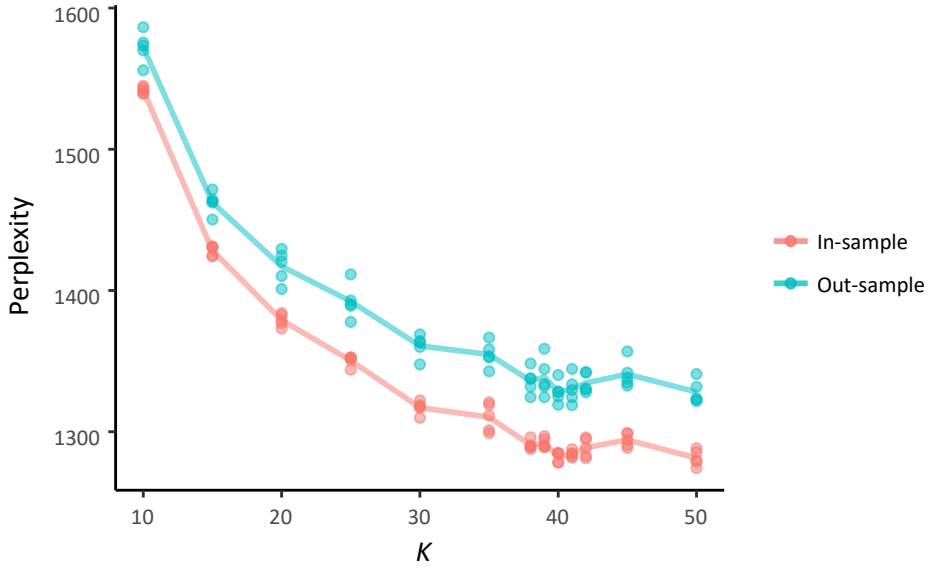

Figure 7: 5-fold cross-validation in-sample and out-sample perplexity results of LDA models. The minimum out-sample perplexity occurs at $K = 40$.

My final LDA model used $K = 40$ topics to fit all documents in the dataset. Cumulatively, the 9 most likely topics, i.e. highest $E(\boldsymbol{\theta}_d)_k$, are expected to account for 35.2% of the observed words. The topic proportions across the entire corpus are given below in Figure 8, and the 15 most probable terms for the most frequent topics are shown in Table 1, with descriptive names suggested for each topic.

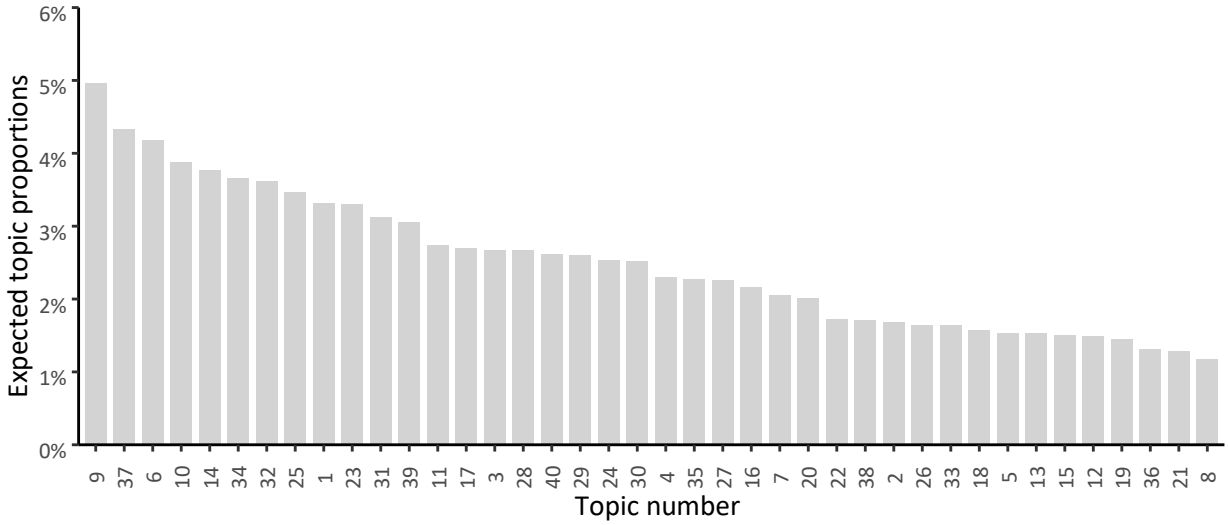

Figure 8: LDA topic proportions across the stats.stackexchange.com dataset, $K = 40$.

9 Distributions	37 Linear Models	6 Commentary on Posts	10 Data	14 Hypothesis Testing	34 Model Selection	32 Coding	25 Prediction	1 Experimental Design
random	variables	question	data	test	model	code	class	group
function	variable	answer	set	hypothesis	models	function	training	groups
distribution	regression	people	values	tests	fit	package	set	anova
variables	logistic	make	dataset	null	regression	values	features	treatment
variable	dependent	good	points	testing	linear	method	validation	control
probability	independent	sense	observations	significance	residuals	results	cross	design
density	model	problem	sets	significant	data	output	classification	subjects
independent	predictors	results	analysis	difference	parameters	bootstrap	feature	effect
conditional	categorical	important	measurements	statistic	fitting	run	data	difference
pdf	predictor	statistical	problem	statistical	fitted	generate	classifier	post
expectation	continuous	case	datasets	values	prediction	simulation	performance	repeated
case	linear	information	make	power	aic	number	classes	experiment
joint	multiple	things	question	reject	predictions	matlab	svm	measures
question	response	point	approach	level	residual	result	selection	analysis
find	binary	questions	observation	means	predicted	problem	dataset	multiple

Table 1: Most frequent terms for each of the 9 most likely LDA topics. See https://github.com/phively/uchicago-thesis/blob/master/results/lda_final_40_top_terms.md for the full list.

CTM Results

Since the STM model reduces to a CTM in the absence of covariates (Roberts, Stewart, & Airoldi, 2016), I used the `stm` package to fit correlated topic models. To find a suitable value of K, I created three size 18,467 $(D/5)$ partially held-out sets of documents using the package's built-in holdout function. Spectral initialization (Arora, et al., 2013) performed wonderfully: convergence was almost always attained in 15-20 iterations, independent of the selection of K, and as shown in Figure 9, the overall estimation process was an order of magnitude faster compared to the `topicmodels` LDA implementation. This allowed the fitting of larger models, so I tried $K \in \{20,30, \dots ,120\}$.

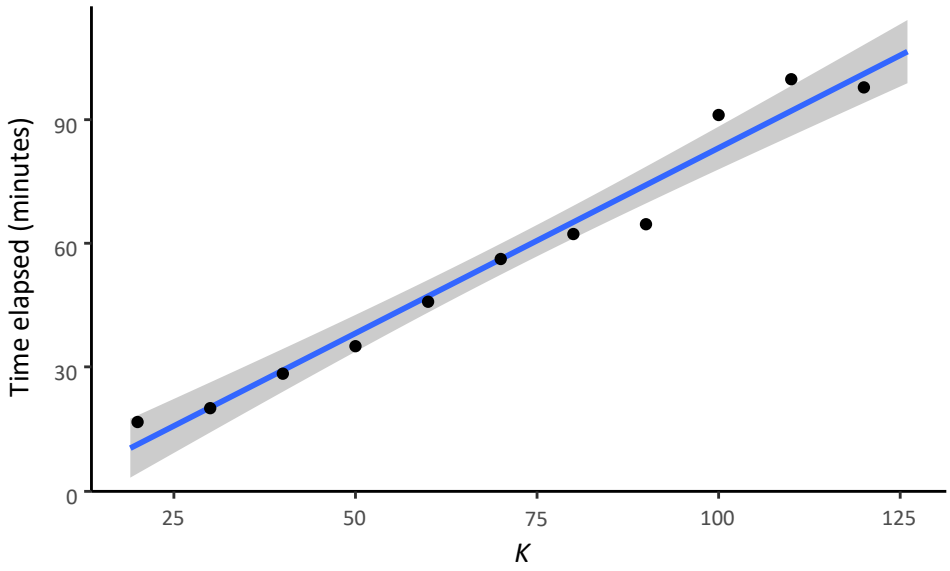

Figure 9: Average time elapsed until CTM model convergence in minutes as a function of the number of topics K.

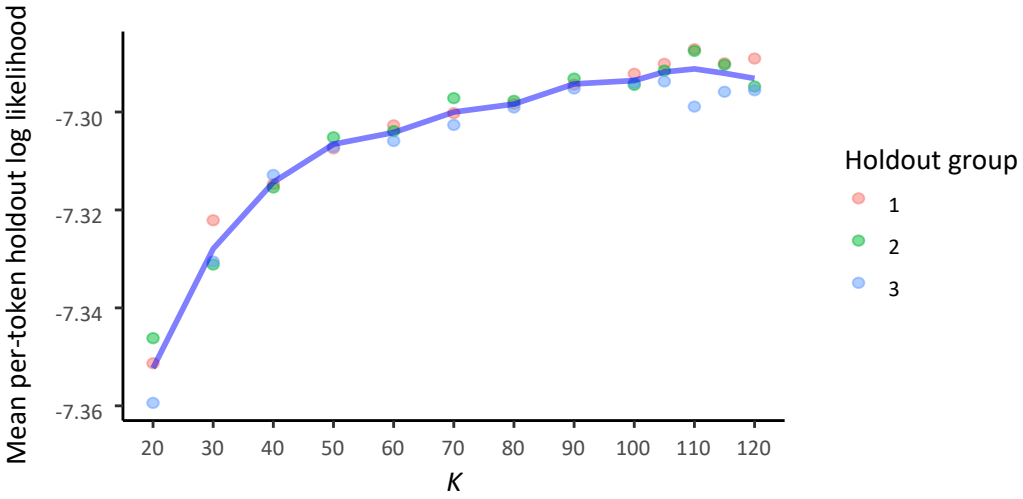

Figure 10: Per-token log likelihood of CTM models on partially held out sets of documents. The mean empirical log likelihood is maximized at $K = 110$.

The results plotted in Figure 10 show that across holdout sets, the mean maximum log likelihood is achieved at $K = 110$, so I chose this value when fitting a final CTM on the entire dataset. Here, the 9 most likely topics are expected to account for 18.9% of the words. $E(\boldsymbol{\theta}_d)_k$ is given for the 40 most likely topics in Figure 11, and the 15 most probable terms for the 9 most common topics are found in Table 2.

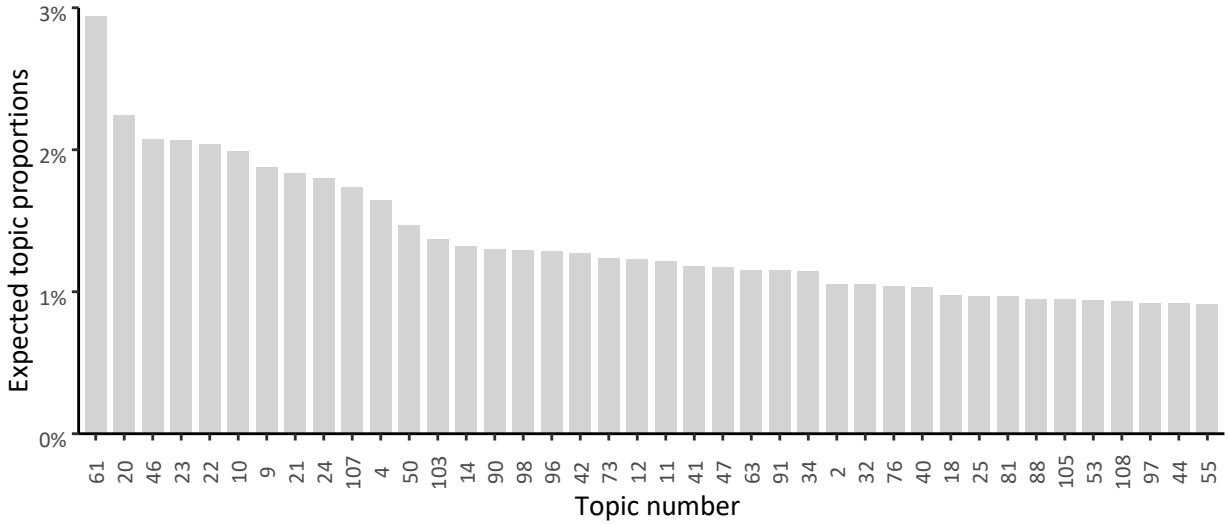

Figure 11: Top 40 CTM topic proportions across the stats.stackexchange.com dataset, $K = 110$.

61 Best Practice	20 Data	46 Model Selection	23 Research	22 Commentary on Posts	10 Variable Selection	9 Statistical Theory	21 ANOVA	24 Reference Materials
sense	data	model	methods	question	dependent	expectation	wilcoxon	book
makes	sets	models	approaches	answers	variables	proof	mann	books
reason	set	modelling	method	answer	variable	moments	whitney	introduction
cases	raw	modeling	paper	asked	explanatory	theorem	test	journal
issue	collected	building	literature	questions	vif	notation	kruskal	chapter
situation	synthetic	fit	approach	answered	multicollinearity	prove	wallis	edition
things	consists	aicc	technique	correct	collinearity	derivation	paired	introductory
matter	kind	saturated	techniques	comment	independent	finite	tests	author
common	generated	full	papers	comments	collinear	inequality	signed	science
reasons	apply	fitting	recommend	incorrect	enter	denote	difference	john
issues	working	compare	references	answering	vifs	moment	compare	modern
situations	gathered	simpler	popular	feel	dependant	measurable	comparing	article
practice	advice	complex	advantages	wrong	remove	summation	unpaired	press
reasonable	datas	fits	helpful	hope	include	limit	differences	springer
generally	collection	build	aware	edit	individually	expectations	significantly	authors

Table 2: Most probable and characteristic terms for each of the 9 most likely CTM topics. See https://github.com/phively/uchicago-thesis/blob/master/results/ctm_final_110_top_terms.md for the full list.

Finally, the correlations between topics induced by the covariance matrix $\boldsymbol{\Sigma}$ were estimated from the posterior document topic proportions $\boldsymbol{\theta}_d$ and visualized using the Fruchterman-Reingold force-directed layout algorithm provided by the `igraph` package. To enforce sparsity and reduce the number of edges, small correlations $|\hat{\rho}_{i,j}| < 0.05$ were truncated to 0. Figure 12 illustrates a part of the resulting undirected graph, which reveals thematic structure among the topics. For instance, in the top left of the graph, topics 76 (group treatment), 21 (Wilcoxon Mann), 50 (results statistically), and 108 (size small) all tend to appear together. This could be interpreted as suggesting the existence of a constellation of documents about testing for between-group differences given a small sample size.

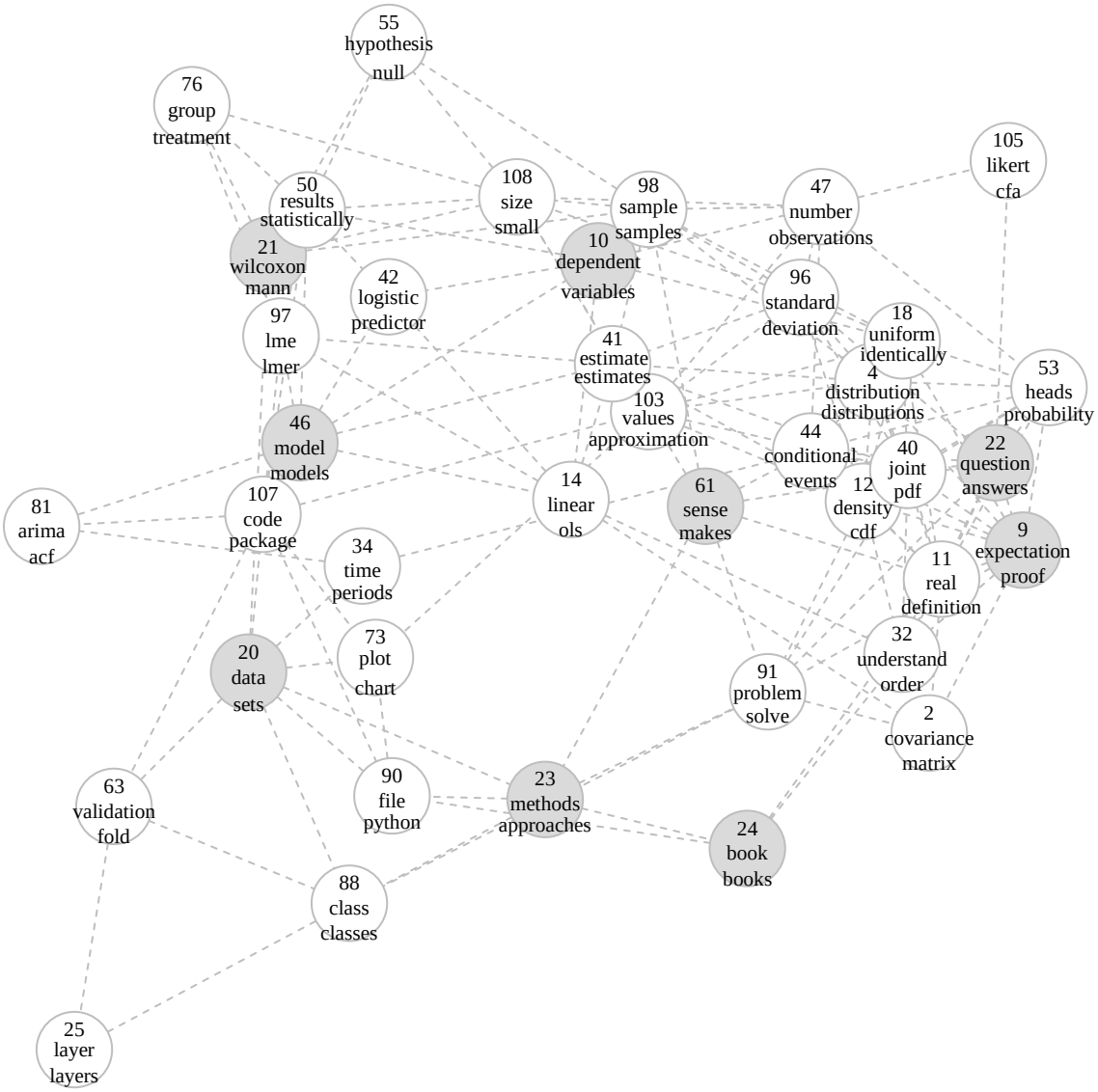

Figure 12: CTM topic correlations of the 40 most likely topics across the stats.stackexchange.com dataset. The topic number $k \in \{1, \ldots, 110\}$ and two characteristic terms are shown for each topic, and nodes for the 9 most likely topics are shaded gray.

STM Results

Before fitting a structural topic model to the stats.stackexchange.com dataset, I explored the fields associated with each question to identify any metadata that could potentially be included as covariates or factors. Candidate fields included creation date and last activity date as measures of recency; view count, answer count, and score as measures of popularity based respectively on the number of page views, recorded answers, and upvotes a question had received; and tags, which allow users to associate up to five keywords with a given question.

Statistic	CreationDate	LastActivityDate	ViewCount	AnswerCount	Score
Minimum	2009-02-02	2010-07-19	1	0	-22
First Quartile	2013-08-22	2014-02-11	46	0	0
Median	2015-02-04	2015-06-02	120	1	1
Mean	2014-10-08	2015-01-25	981.63	0.97	2.43
Third Quartile	2016-02-25	2016-04-21	440	1	3
Maximum	2016-12-11	2016-12-11	340000	140	460

Table 3: Metadata summary statistics for the stats.stackexchange.com dataset.

Of the continuous metadata, creation date and score are the most interesting. I prefer creation date to last activity date because creation date is a fixed point in time, and score gives a sense of community sentiment not captured by view or answer counts. These were included as covariates $\boldsymbol{X}$ affecting the topical prevalence of $\boldsymbol{\theta}_d$, since it's reasonable to expect that topic prevalence can change over time, and that some topics are more popular than others. I transformed creation date into the time in months since the first post was made, and included both variables in the model as B-splines with 10 degrees of freedom.

The dataset contains nearly 1,500 distinct tags, practically a vocabulary in its own right, though most tags are used only once. It would be interesting to see how user-added tags compare to the latent topics discovered by the STM and include a selection of tags as both topical factors $\boldsymbol{X}$ and content factors $\boldsymbol{Y}$. However, due to the formulation of the model, the introduction of each new factor level necessitates the estimation of additional topic-vocabulary terms $\boldsymbol{\beta}_{1:K}$, plus any necessary interaction terms; this quickly becomes computationally prohibitive. As a result, I looked at the 10 most common tags and included only the one I found most interesting: `homework/self-study`, which is associated with 4,139 questions.

The observed data in the STM was therefore formulated as follows:

$$\boldsymbol{X} = \left[f(\text{Score}_d) \quad g(\text{CreationDate}_d) \quad \mathbf{1}\{\text{Tag}_{\text{HW},d}\} \right]_{D,3}$$

$$\boldsymbol{Y} = \left[\mathbf{1}\{\text{Tag}_{\text{HW},d}\} \right]_{D,1}$$

$$\boldsymbol{w}_{1:D} = \{w_1, \ldots, w_{N_d}\}$$

As in the previous CTM results section, the number of topics K was chosen by creating three size 18,467 partially held-out sets of documents and identifying the K producing the mean maximum log likelihood. STM convergence as a function of K was faster than LDA but substantially slower than CTM, so with the knowledge that $K = 110$ was the optimal choice above, I tried models with $K \in \{60,70,\dots,120\}$.

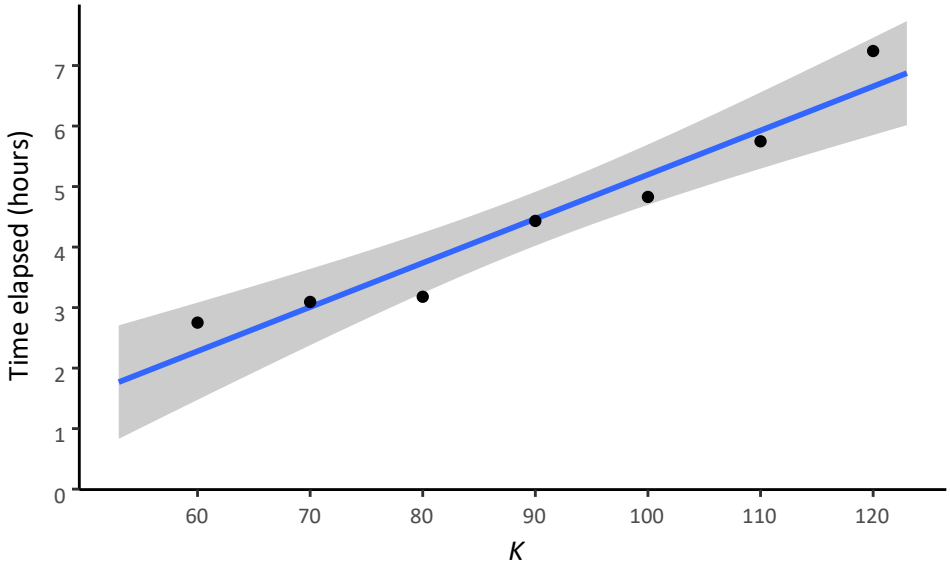

Figure 13: Average time elapsed until STM model convergence in hours as a function of the number of topics K.

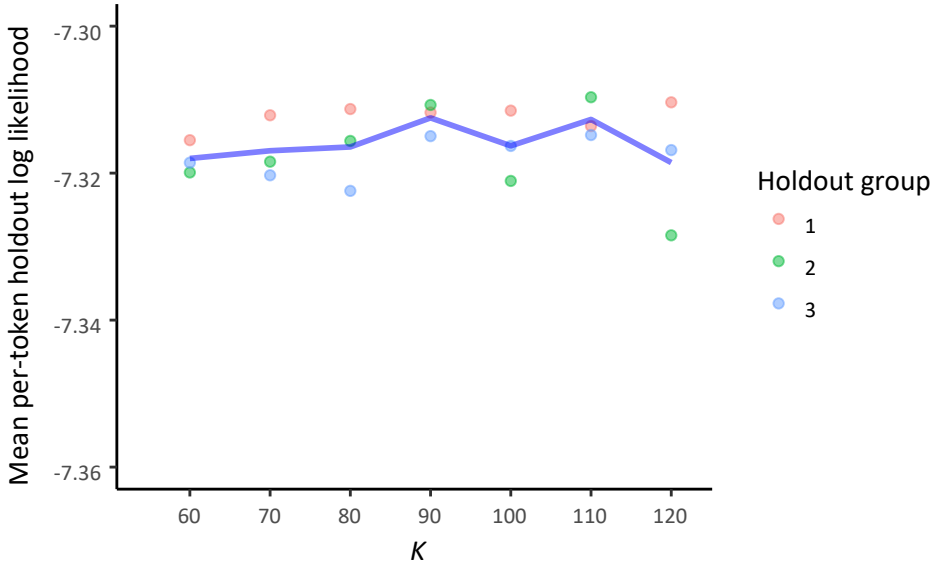

Figure 14: Per-token log likelihood of STM models on partially held out sets of documents. The mean empirical log likelihood is maximized at $K = 90$.

From these results, the mean maximum log likelihood and minimum variance are both achieved at $K = 90$, so I chose this value for my final model. We do see that as a result of the additional parameters

estimated by the STM there is considerably more variability for a fixed value of K than was present when fitting the CTMs (compare Figure 14 to Figure 10). Computational resources permitting, since the spectral initialization algorithm is deterministic a full 5- or 10-fold cross-validation scheme would help identify a suitable K for STMs. For $K = 90$, the 9 most likely topics are expected to account for 21.6% of the words, and as before topic proportions and top vocabulary for the most likely topics are given below.

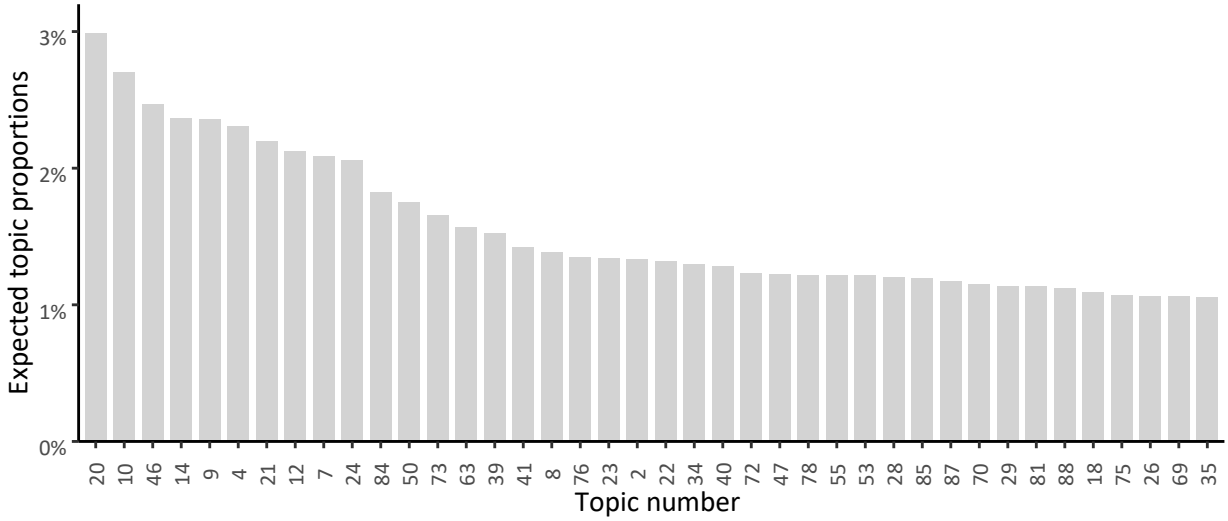

Figure 15: Top 40 STM topic proportions across the stats.stackexchange.com dataset, $K = 90$.

20 Data	10 Variable Selection	46 Model Selection	14 Linear Models	9 Statistical Theory	4 Distributions	21 ANOVA	12 Calculus	7 Work
imputation	dependent	aic	linear	proof	distribution	wilcoxon	cdf	tools
missing	categorical	bic	ols	expectation	distributions	paired	integral	open
data	variable	model	regression	definition	kurtosis	mann	formula	tool
datasets	explanatory	fit	regressions	prove	weibull	whitney	closed	skills
dataset	variables	models	predictor	inequality	normal	smirnov	approximation	science
sets	ordinal	fits	nonlinear	finite	uniform	test	cumulative	phd
imputed	multicollinearity	fitting	coefficients	theorem	tail	permutation	derivative	programs
impute	collinearity	aicc	regressors	expectations	pareto	signed	derivation	programming
set	nominal	fitted	linearity	holds	skewness	kolmogorov	delta	projects
handle	vif	reml	predictors	measurable	exponential	parametric	integrate	fields
missingness	ivs	nested	regress	notation	cauchy	kruskal	derive	project
synthetic	continuous	goodness	regressor	limit	bimodal	wallis	expression	field
complete	dichotomous	deviance	quadratic	converges	empirical	compare	formulas	job
nas	independent	akaike	ordinary	infinity	shape	difference	equation	academic
frame	numeric	saturated	regressing	moments	lognormal	comparing	numerically	thesis

Table 4: Most probable and characteristic terms for each of the 9 most likely STM topics. See https://github.com/phively/uchicago-thesis/blob/master/results/stm_final_90_top_terms.md for the full list.

For between-topic correlations, I used the same settings as in the CTM network graphs to make the two models as comparable as possible. In particular, I truncated all small correlations $|\hat{\rho}_{i,j}| < 0.05$ to 0. Note that the most common topics—20, 10, 46, and 14—are all located in the middle of the graph and appear to have high network centrality.

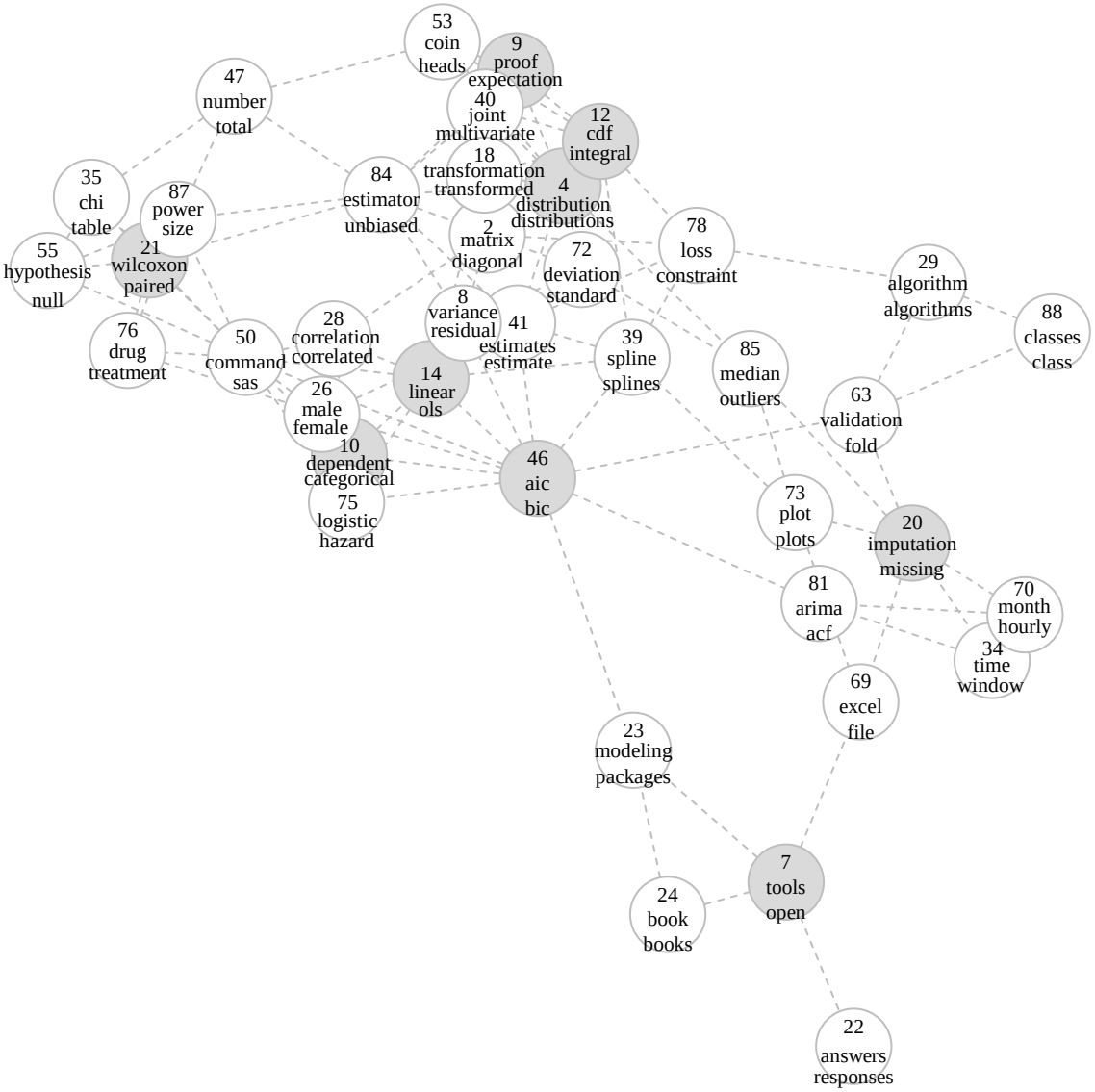

Figure 16: STM topic correlations of the 40 most likely topics across the stats.stackexchange.com dataset. The topic number $k \in \{1, \ldots, 90\}$ and two characteristic terms are shown for each topic, and nodes for the 9 most likely topics are shaded gray.

The primary reason to fit an STM is to investigate whether observed document metadata is associated with corpus structure (Roberts, Stewart, Tingley, & Airoldi, 2013). To that end, I fit several of the most interesting topics to each of the covariates. In Figure 17 below, 5 of the 9 most likely topics are associated with higher scores and 3 with lower scores, while the last demonstrates no association. Across all 90 topics, topic 9 (proof expectation) was associated with the lowest scores and topic 10 (dependent categorical) the highest scores. I will focus on these two topics when exploring the other metadata.

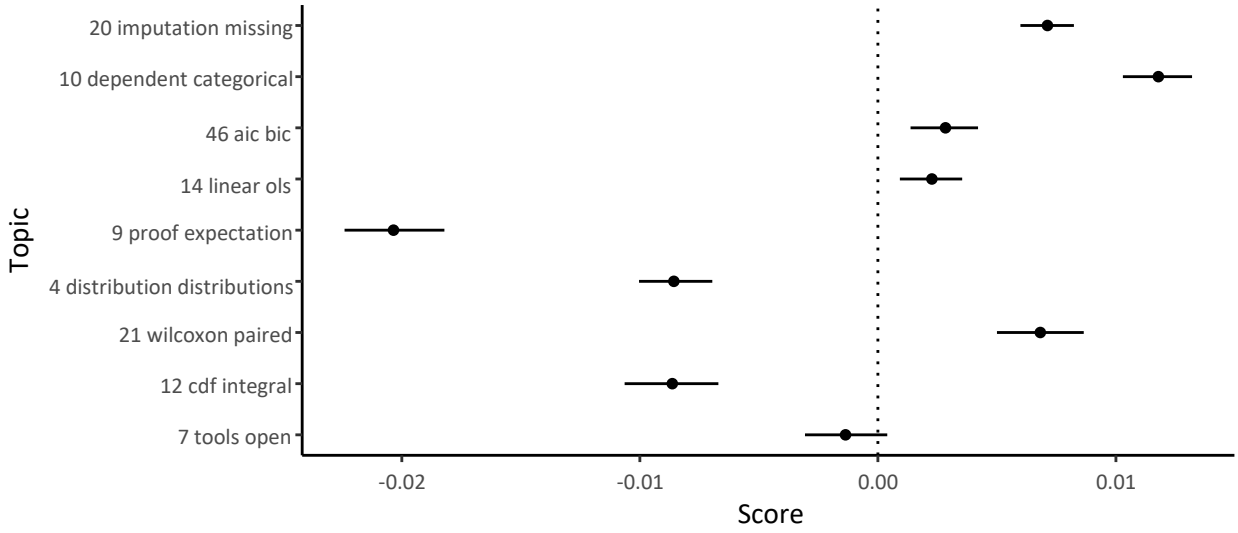

Figure 17: STM topic associations with question score for each of the 9 most likely topics.

I found that most topic proportions are very stable through time. In Figure 18, the high-scoring topic 10 is nearly flat, with an expected proportion near 0.02 throughout the 8-year period. On the other hand, the low-scoring topic 9 has greatly increased in prevalence since 2009, and there is perhaps a hint of seasonality from late 2015 on, with slightly lower topic prevalence during the summer months. Topic 69 (excel file) is included for comparison and as a point of interest: its prevalence peaked in 2010 and has declined ever since.

It's worth remarking on the wide confidence bands around each topic proportion in 2009, and their rapid narrowing starting in 2010. Studying the site's post history, it appears that a handful of questions were asked in 2009 but the site did not fully launch until the second half of 2010. Data from before July 2010 should therefore be treated with care.

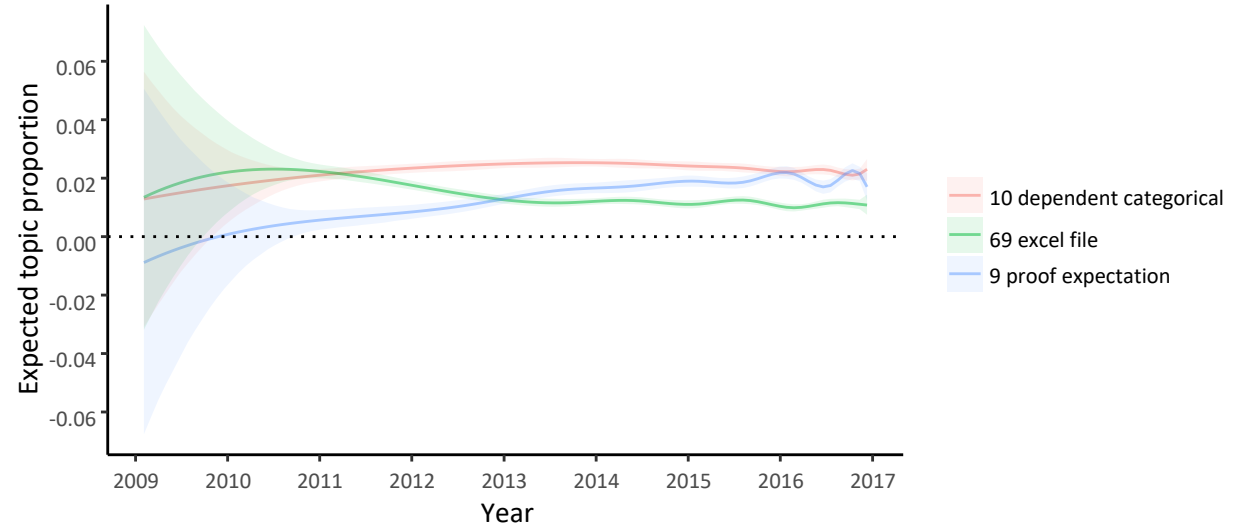

Figure 18: STM expected topic proportions over time. Across the corpus, most topic proportions appear nearly flat like topic 10; the increasing and decreasing trends seen in topics 9 and 69 are atypical.

Since topic 9 seems to involve statistical theory and may have lower topic prevalence during the summer months, it seems plausible that a homework effect is in play. Figure 19 and Figure 20 show the expected prevalence of topics 9 and 10 split between questions that are or aren't tagged homework; topic 9 is indeed substantially more likely to occur given the homework tag, while topic 10 is less likely to be present given the same tag. Finally, Figure 21 provides a visualization of the differences in vocabulary between the topic 9 questions tagged homework and those that aren't. While it is incorrect to read too much into these exploratory results, all together these findings do suggest that homework-tagged posts are less likely to receive high scores than others.

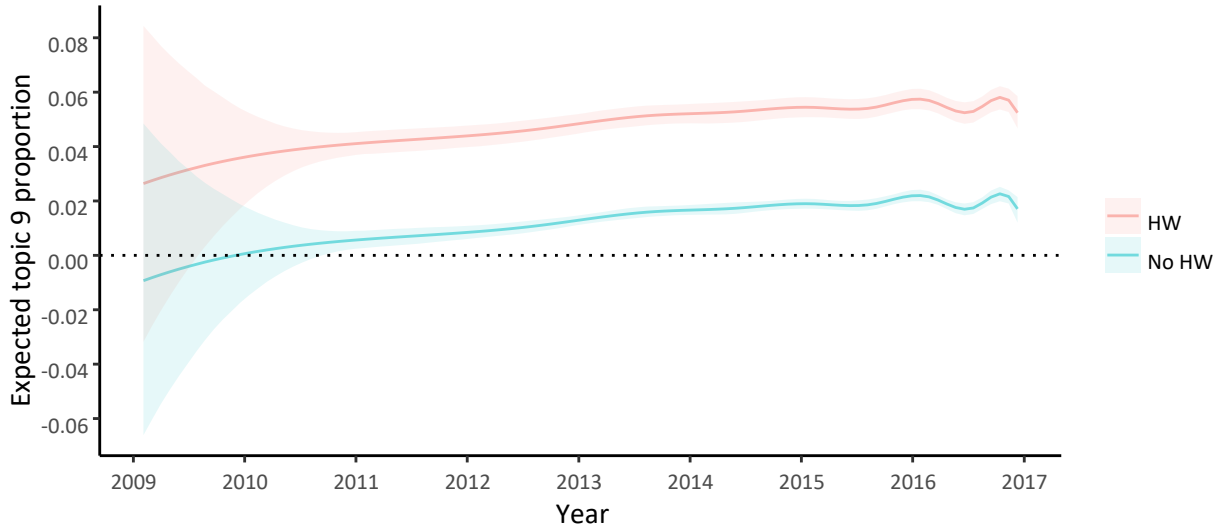

Figure 19: STM topic 9 expected proportions over time for posts with and without the homework tag.

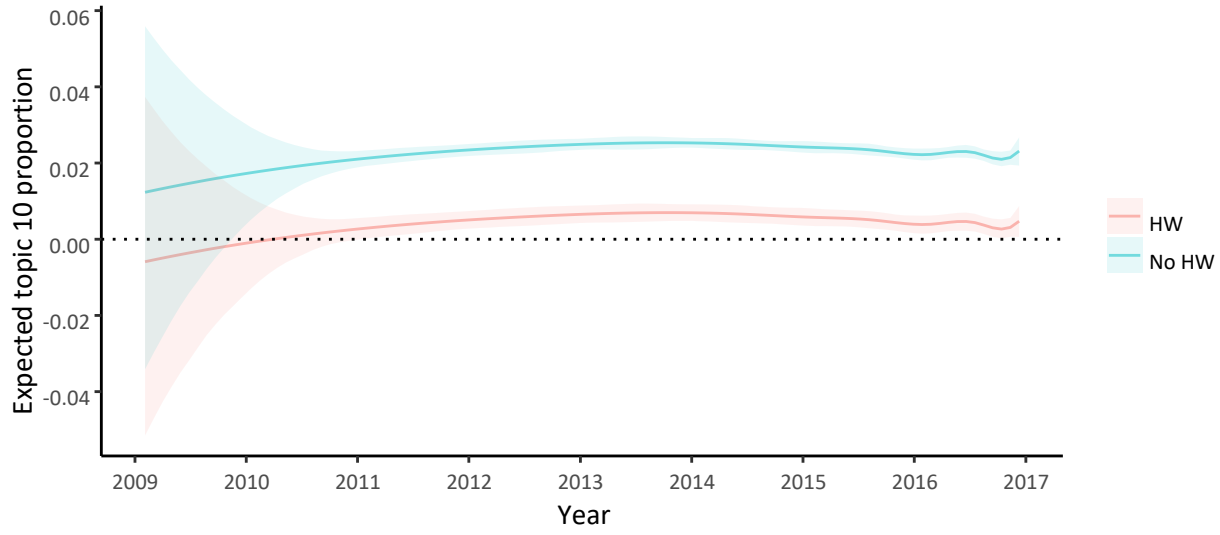

Figure 20: STM topic 10 expected proportions over time for posts with and without the homework tag.

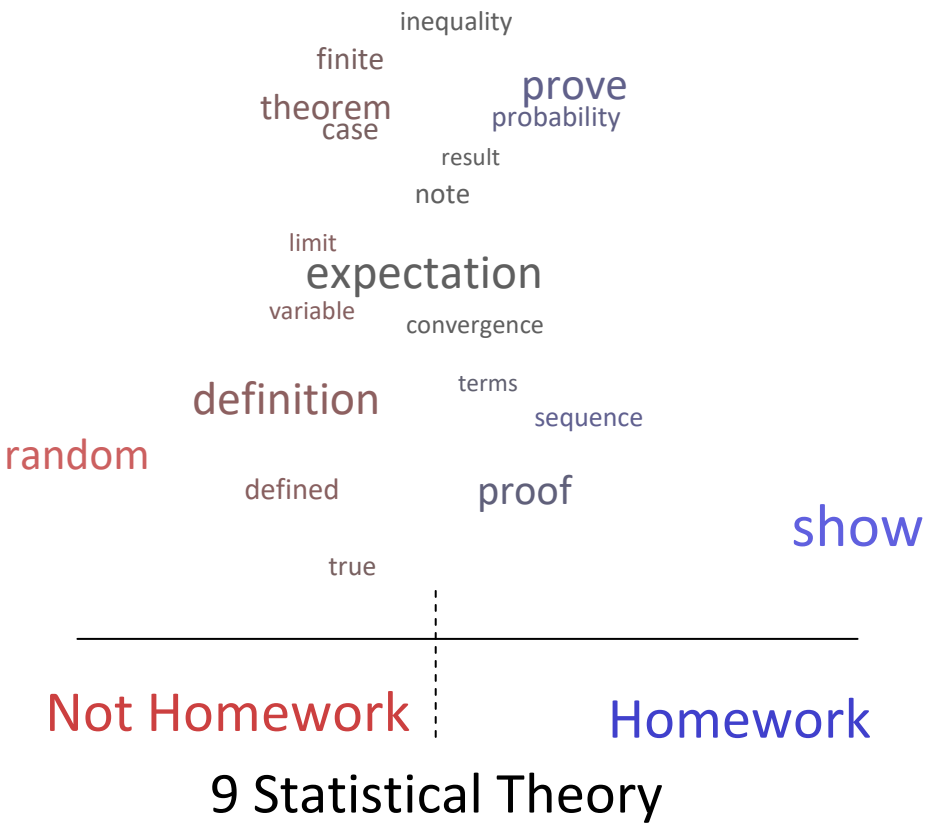

Figure 21: Per-term contrasts between posts tagged or not tagged as homework for STM topic 9. Larger font size indicates a term occurs more frequently within the corpus.

Model Comparison

In examining the results for each of the LDA, CTM, and STM models, one interesting observation is that CTM and STM support more topics than LDA. For LDA, I found $K = 40$, compared to $K = 110$ for CTM and $K = 90$ for STM. The ratio of CTM to LDA topics is nearly 3 to 1, which is similar to that found in previous comparisons of LDA and CTM models (Blei & Lafferty, 2007). This reflects the greater flexibility and better fit yielded by the CTM's logistic normal prior.

The reason for the different topic counts between the CTM and STM models is less clear. While one interpretation is that STM suffers compared to CTM due to the greater number of parameters that need to be estimated, it's important to acknowledge that a more exhaustive cross-validation approach may have yielded a different choice of K. On the other hand, if the STM model is correct and metadata effects really are involved in the document generation process, its model structure should be preferred over the other methods. Previous simulation studies have looked at how well various topic model formulations can recover covariate effects on document content, and demonstrate that the STM is substantially better at estimating the effect's direction and magnitude than a 2-step process involving fitting an LDA model and then performing a post-hoc analysis on the resulting topics (Roberts, et al., 2014).

Examining the top LDA topics in Table 1, CTM topics in Table 2, and STM topics in Table 4, there appear to be two sets of similar topics shared across all three models which I have designated "Data" and "Model Selection" respectively. Table 5 shows the most common vocabulary for these sets of topics side by side.

LDA		CTM		STM	
10 Data	**34 Model Selection**	**20 Data**	**46 Model Selection**	**20 Data**	**46 Model Selection**
data	model	data	model	imputation	aic
set	models	sets	models	missing	bic
values	fit	set	modelling	data	model
dataset	regression	raw	modeling	datasets	fit
points	linear	collected	building	dataset	models
observations	residuals	synthetic	fit	sets	fits
sets	data	consists	aicc	imputed	fitting
analysis	parameters	kind	saturated	impute	aicc
measurements	fitting	generated	full	set	fitted
problem	fitted	apply	fitting	handle	reml
datasets	prediction	working	compare	missingness	nested
make	aic	gathered	simpler	synthetic	goodness

Table 5: Comparing the data and model selection topics across the LDA, CTM, and STM models.

There is substantial vocabulary overlap between the three models across these two topics. To better understand their similarity and differences it's instructive to examine a few of the documents most associated with these topics in each model.

Posts most associated with the LDA data topic:

1. 244815 Scaling large range values
2. 143414 Combining Standard Deviation (combined measurement uncertainty)
3. 175945 PCA on bimodal data: how to standardize the data?
4. 249959 Normalization of data with lots of zeros

Posts most associated with the CTM data topic:

1. 79511 How to combine two data sets (level-1 and level-2 data sets) for multilevel analysis in R
2. 58333 What method can be used to test if three or more categorical sample data sets are from the same distribution?
3. 123621 Can I use K-S test to distinguish distributions?
4. 52010 Choosing a better data-set

Posts most associated with the STM data topic:

1. 104392 Multiple Imputation: Before or after case exclusion?
2. 181399 How to combine multiply imputed datasets created with MICE from different cohorts?
3. 104904 How to replace the missing data from AMELIA results
4. 228463 perform quality check for imputed data with MICE in R

While it's clear that all of these posts do deal with data in some manner, there are clearly themes within each model that my simple label of "data" did not capture. For instance, the LDA data topic appears to concern data standardization, while the CTM data topic addresses methods for comparing datasets, and the STM data topic deals with missing data. Notably, the focus of each topic is distinctive enough that none of the top 4 posts are shared between any of the models.

On the other hand, the model selection topics are much more homogeneous between the three models. Posts most associated with the LDA model selection topic:

1. 198365 Higher SIC and lower S.E. of residuals
2. 221016 AIC (QIC) - comparision of models?
3. 206808 Model comparison with chi-square difference test and BIC
4. 94292 Comparing nested, non-linear models

Posts most associated with the CTM model selection topic:

1. 94292 Comparing nested, non-linear models
2. 194601 Comparison between different models: hierarchical models and model with dummy
3. 76893 Comparing linear models
4. 174018 Extract goodness of fit from linear mixed effects models

Posts most associated with the STM model selection topic:

1. 194601 Comparison between different models: hierarchical models and model with dummy
2. 127353 What model should I use for this?
3. 221016 AIC (QIC) - comparision of models?
4. 76893 Comparing linear models

This time, all of the topics clearly address model selection, with several of the same posts flagged as exemplars by multiple models. The different results for the two topics examined above raise questions about interpretability. Surprisingly, it has been found that topics identified by models with higher out-of-sample predictive performance can be less interpretable than those identified by models with lower performance. Humans are worse at identifying *intruder words* inserted into the topics of CTMs than into those of simpler models, such as LDA or pLSI, and are also worse at identifying *intruder topics* incorrectly associated with documents, despite CTM's superior predictive performance as measured by the log likelihood of holdout data (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009). This suggests that the inter-topic correlations allowed by CTM and STM improve model fit but also confound human intuition.

We can further assess differences between the three models by examining how they assign topic proportions to a single document. As of December 11, 2017, with a score of 462, the most popular stats.stackexchange.com question of all time was post number 2691—Making sense of principal component analysis, eigenvectors & eigenvalues—which seeks to "explain these concepts to a layman or grandma." The below figures show the expected posterior topic proportions for each model.

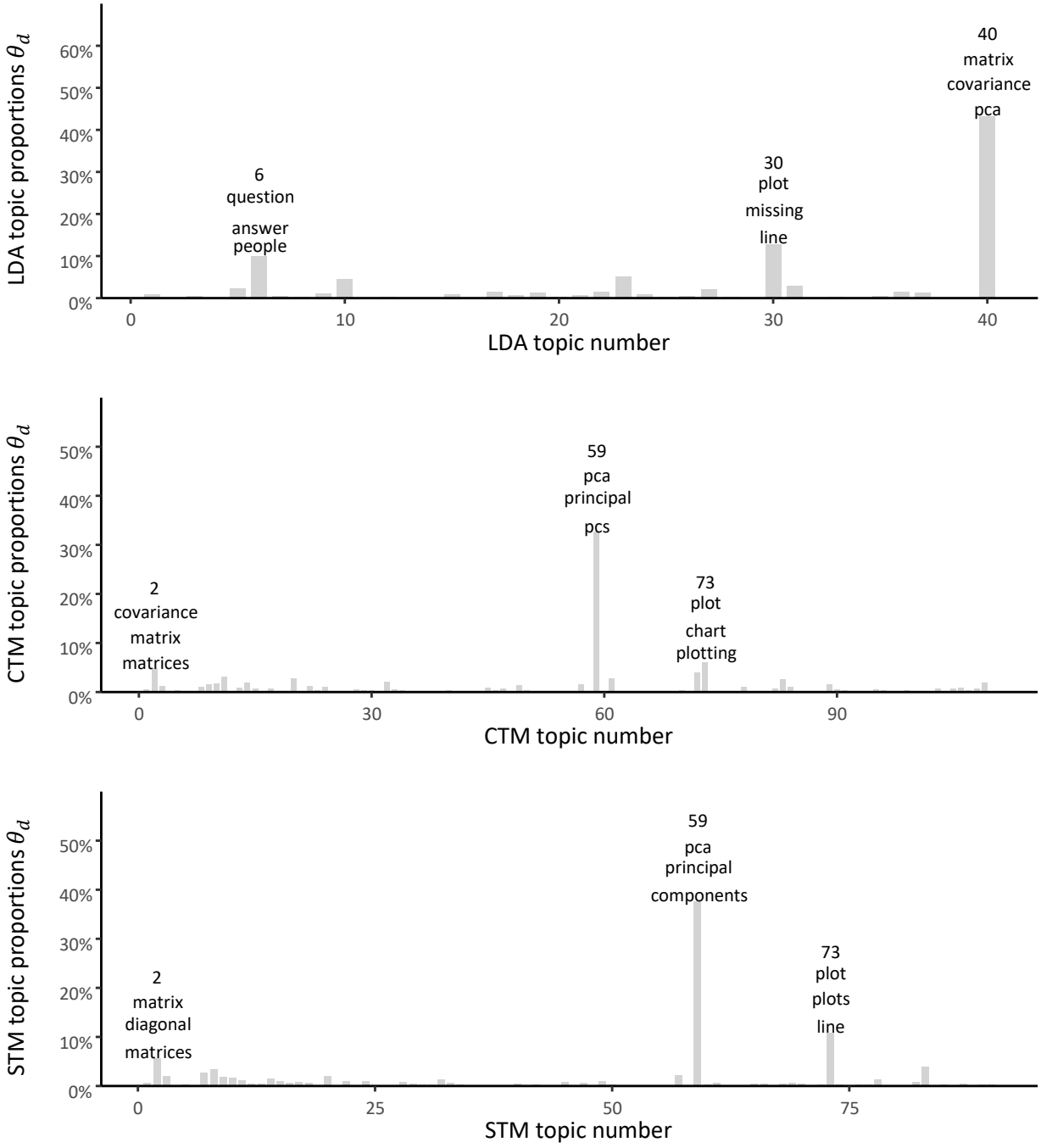

Figure 22: LDA, CTM, and STM posterior topic probabilities for post 2691. For each model, the 3 topics with highest expected posterior probability are numbered and labeled with their 3 most likely terms.

The most likely topics are very similar: each model ranks PCA as first by a large margin, and data visualization is second. For CTM and STM, the third most likely topic is matrices; LDA appears to have combined matrices and PCA into topic 40, and the third topic is difficult to interpret. Also note that corresponding CTM and STM topics actually share the same index k. This occurs because both models were fit on the full dataset and the spectral initialization used by the `stm` package is deterministic, identifying the same anchor words and sorting them into the same initial topics each run.

We can also explore differences in how the three models fit the corpus by examining documents that are expected to be similar. Expected Hellinger distance $E\left[d\left(\theta_i, \theta_j\right)\right] = E\left[\sum_k\left(\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}}\right)^2\right]$ provides a measure of document similarity by finding the distance between the posterior topic proportions of documents i and j (Blei & Lafferty, 2007).

These posts are most similar to 2691 according to the LDA model (with Hellinger distance in parentheses):

1. 48244 PCA vector interpretation (0.169)
2. 10225 Extracting multiple columns from a matrix in R (0.188)
3. 164054 Is there a typo in this paper on Slow feature analysis? (0.197)

These posts are most similar according to the CTM model:

1. 232941 How to interpret ggbiplot() visualization of PCA in R? (0.121)
2. 137430 PCA and diagonalization of the covariance matrix (0.128)
3. 234619 CCA Interpretation (0.136)

And these posts are most similar according to the STM model:

1. 234619 CCA Interpretation (0.138)
2. 130882 Why are principal components in PCA (eigenvectors of the covariance matrix) mutually orthogonal? (0.148)
3. 232941 How to interpret ggbiplot() visualization of PCA in R? (0.156)

As before, the CTM and STM model give very similar results: 2 of the 3 documents identified as closest to post 2691 are shared. LDA is quite dissimilar, likely due to the smaller number of topics K used by the model. Recall from Figure 22 that LDA seemed to combine PCA and matrices into a single topic; looking at minimum Hellinger distances here, all but the first most similar document appear to focus on matrices rather than PCA.

As a point of interest, stats.stackexchange.com has its own post similarity algorithm. The top 3 questions flagged as related to post 2691 are:

1. 31510 Correlating variables with eigenvalues in principal components analysis
2. 44229 How do I convert the variances explained by a principal component into an eigenvalue?
3. 88537 Principal Components, Canonical Correlation and Eigenvalue problems

Note that each of these linked posts include the words “principal components” and “eigenvalue,” which suggests that the website simply matches posts with similar titles. By capturing the semantic meaning behind questions, topic model similarity scores suggest a richer set of related documents than simple keyword matches can provide.

Discussion

Having investigated latent Dirichlet allocation, correlated topic models, and structural topic models, which method is best? As usual, the choice of model must depend on the problem under investigation.

While latent Dirichlet allocation does not have the flexibility of the other methods, its assumption of a conjugate Dirichlet prior is a great strength during the model estimation process, and given a suitable initialization algorithm (Arora, et al., 2013) LDA should be able to outperform any other method. Additionally, for certain applications, LDA’s greater interpretability could make it a better choice than more complex models (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009).

The correlated topic model implementation in the `stm` package is very fast, so the CTM is an obvious choice over LDA if improved predictive performance is desired, and particularly if topic correlations are of interest. The estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$ provides a useful way to understand the relationships between topics and can be interpreted as an undirected graph, giving rise to network visualizations that can assist in navigating a corpus (Blei & Lafferty, 2007).

Finally, structural topic models provide a systematic way to understand the effects of covariates on topic prevalence and content. By incorporating metadata into the model, STMs relax the exchangeability assumptions of LDA and offer improved effect estimation compared to post-hoc tests (Roberts, et al., 2014; Roberts, Stewart, & Airoldi, 2016), providing an interesting tool for researchers studying relationships between language usage and other variables.

References

Airoldi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C., Love, T., & Shringarpure, S. (2010). Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences, 107*(49). doi:0.1073/pnas.1013452107

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., . . . Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. *Proceedings of the 30th international conference on machine learning, 28*(2), 280-288. Retrieved from http://proceedings.mlr.press/

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics, 1*(1), 17-35. doi:10.1214/07-AOAS114

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In V. Kumar (Ed.), *Text mining classification, clustering, and applications* (pp. 71-93). Boca Raton, FL: Chapman & Hall/CRC.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022. Retrieved from http://jmlr.csail.mit.edu/

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: how humans interpret topic models. *Proceedings of the 22nd International Conference on Neural Information Processing Systems, 22*, 288-296. Retrieved from https://papers.nips.cc/

Crain, S. P., Zhou, K., Yang, S., & Zha, H. (2012). Dimensionality reduction and topic modeling: from latent semantic indexing to latent Dirichlet allocation and beyond. In C. Aggarwal, & C. Zhai (Eds.), *Mining Text Data* (pp. 129-161). New York, NY: Springer-Verlag. doi:10.1007/978-1-4614-3223-4_5

Creative Commons. (2007). *Attribution-ShareAlike 3.0 Unported CC BY-SA 3.0.* Retrieved from creativecommons.org: https://creativecommons.org/licenses/by-sa/3.0/

Grun, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software, 40*(13). Retrieved from http://www.jstatsoft.org/

Roberts, M., Stewart, B., & Airoldi, E. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association, 111*(515), 988-1003. doi:10.1080/01621459.2016.1141684

Roberts, M., Stewart, B., & Tingley, D. (2016). stm: R package for structural topic models. Working paper. Retrieved from https://cran.r-project.org/

Roberts, M., Stewart, B., Tingley, D., & Airoldi, E. (2013). The structural topic model and applied social science. *Neural Information Processing Society.* Retrieved from https://scholar.princeton.edu/bstewart/

Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., . . . Rand, D. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064-1082. doi:10.1111/ajps.12103

Stack Exchange, Inc. (2014). *Stack Exchange Data Dump published December 15, 2016.* Retrieved February 24, 2017, from Internet Archive: https://archive.org/details/stackexchange

Taddy, M. (2012). On estimation and selection for topic models. *Journal of Machine Learning Research, 22*, 1184-1193. Retrieved from http://jmlr.csail.mit.edu/

Taddy, M. (2015). Distributed multinomial regression. *The Annals of Applied Statistics, 9*(3), 1394-1414. doi:10.1214/15-AOAS831

Wallach, H. M., Mimno, D., & McCallum, A. (2009). Rethinking LDA: why priors matter. *Advances in Neural Information Processing Systems, 22*. Retrieved from https://papers.nips.cc/

Wallach, H., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th International Conference on Machine Learning*, 1105-1112. doi:10.1145/1553374.1553515

Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. *SIGIR '06 Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 178-185). doi:10.1145/1148170.1148204