2

Navigating the Local Modes of Big Data

The Case of Topic Models[*]

Margaret E. Roberts
University of California, San Diego

Brandon M. Stewart
Princeton University

Dustin Tingley
Harvard University

1 INTRODUCTION

Each day humans generate massive volumes of data in a variety of different forms (Lazer et al., 2009). For example, digitized texts provide a rich source of political content through standard media sources such as newspapers, as well as newer forms of political discourse such as tweets and blog posts. In this chapter we analyze a corpus of 13,246 posts that were written for six political blogs during the course of the 2008 U.S. presidential election. But this is just one small example. An aggregator of nearly every document produced by the U.S. federal government, voxgov.com, has collected more than eight million documents from 2010–2014, including over a million tweets from members of Congress. These data open new possibilities for studies of all aspect of political life from public opinion (Hopkins and King, 2010) to political control (King, Pan, and Roberts, 2013) to political representation (Grimmer, 2013).

The explosion of new sources of political data has been met by the rapid development of new statistical tools for meeting the challenges of analyzing "big data." (National Research Council, 2013; Grimmer and Stewart, 2013; Fan, Han, and Liu, 2014). A prominent example in the field of text analysis is latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan, 2003; Blei, 2012),

[*] The research assistance of Antonio Coppola, under the support of the Harvard BLISS program, was extremely valuable for this chapter. Our thanks to participants in the Harris School Political Economy workshop and Princeton University Political Methodology workshop for reinforcing the need to write this chapter, and to Scott de Marchi, Jetson Leder-Luis, Jimmy Foulds, Padhraic Smyth, and Hanna Wallach for their comments. An R package for estimating the structural topic model is freely available at www.structuraltopicmodel.com.

https://doi.org/10.1017/CBO9781316257340.004

a topic model that uses patterns of word co-occurrences to discover latent themes across documents. Topic models can help us deal with the reality that large data sets of text are also typically unstructured. In this chapter we focus on a particular variant of LDA, the structural topic model (STM) (Roberts et al., 2014), which provides a framework to relate the corpus structure we do have (in the form of document-level metadata) with the inferred topical structure of the model.

Techniques for automated text analysis have been thoroughly reviewed elsewhere (Grimmer and Stewart, 2013). We instead focus on a less often discussed feature of topic models and of latent variable models more broadly: multimodality. That is, the models discussed here give rise to optimization problems that are nonconvex. Thus, unlike workhorse tools such as linear regression, the solution we find can be sensitive to our starting values (in technical parlance, the function we are optimizing has multiple modes). We engage directly with this issue of multimodality, helping the reader understand why it arises and what can be done about it. We provide concrete ways to think about multimodality in topic models, as well as tools for dealing and engaging with it. For example, we enable researchers to ask these questions. How substantively different are the results of different model solutions? Is a "topic," which heuristically can be thought of as a collection of commonly co-occurring words, likely to appear across many solutions of the model? Furthermore, is our key finding between a variable (such as partisan affiliation) and the prevalence of topic usage stable over multiple solutions to the model?

We also discuss initialization strategies for choosing the starting values in a model with multiple modes. Although seldom discussed, these initialization strategies become increasingly important as the size of the data grows and the computational cost of running the model even a single time rises. Starting the algorithm at better starting values not only leads to improved solutions but can also result in dramatically faster convergence.

The outline of this chapter is as follows. In Section 2 we introduce the problem of multimodality and provide several examples of models with multiple modes. In Section 3 we focus on the particular case of topic models and highlight some of the practical problems that can arise in applied research. In Section 4 we introduce a set of tools that allow users to explore the consequences of multimodality in topic models by assessing the stability of findings across multiple runs of the model. In Sections 5 and 6 we discuss procedures for carefully initializing models that may produce better solutions. Finally Section 7 concludes by returning to the constraints and opportunities afforded by big data in light of the statistical tools we have to analyze it.

2 INTRODUCTION TO MULTIMODALITY

Multimodality occurs when the function we are trying to optimize is not globally concave.[1] Thus, when we converge to a solution, we are unsure whether we have converged to a point that is the global maximum or simply a local maximum. In statistical models, the function we are typically trying to maximize is the likelihood function, and when this function is not concave the solution we arrive at can be dependent on our starting values. This issue occurs in many classes of statistical models, but is particularly relevant in those where (1) the data-generating process of the data comes from a mixture of distributions or contains latent variables, which the likelihood then reflects; (2) ridges (essentially flat regions) in the likelihood function appear due to constraints applied to the statistical model; or (3) some parameters are unidentified and therefore multiple solutions exist for the same model. The ability to diagnose and navigate multimodality decreases with the dimension of the parameter space, as visualizing and estimating the likelihood become more difficult in higher dimensions and more complicated models.

Multimodality is particularly prevalent in the context of big data because the same latent variable models that are useful for analyzing largely unstructured data also lead to challenging optimization problems. The models we employ in this setting often involve mixtures of distributions, complicated constraints, and likelihoods that are difficult to visualize because the models contain hundreds, sometimes thousands of parameters. Although simple models from the exponential family with concave likelihoods like regression or lasso (Tibshirani, 1996) still play an important role in big-data applications (Mullainathan, 2014; Belloni, Chernozhukov, and Hansen, 2014), there is an increasing interest in the use of more complex models for discovering latent patterns and structure (National Research Council, 2013). While the latent variable models can bring new insights, they also introduce a complex optimization problem with many modes.

In this section we build up for the reader intuitions about what can lead to multimodality. We first discuss a convex, univariate Gaussian maximum likelihood model that is easily optimized to provide contrast for the nonconvex models we describe later in the section. Then, we extend the univariate Gaussian to a simple mixture of Gaussians and provide an intuition for why mixture models can be multimodal. Last, we connect the simple mixture of Gaussians to topic models and describe how these models, and generally models for big data, contain *latent variables* (variables in the data-generating process that are not observed), which will mean they are more likely to be multimodal.

2.1 Convex Models

To start, we present an example of a convex model in which multimodality is not a problem. A strictly concave function only has (at most) one maximum and has no local maxima. This is convenient for optimization because when the optimization procedure[2] has found a maximum of a concave likelihood function, it has clearly reached the global maximum if only one exists. The natural parameter space for regression models with a stochastic component in the exponential family are convex and therefore are easily optimized (Efron et al., 1978).

We begin with a simple Gaussian (normal) model with mean μ and variance σ.[2,3] In the next section we show how we can generalize this basic setup to a more flexible Gaussian mixture model.

$$Y \sim N(\mu, \sigma^2)$$

The normal distribution is from the exponential family, and therefore the likelihood is concave. This is easy to see by deriving the log-likelihood:

$$\begin{aligned} L(\mu|y) &\propto N(y|\mu,\sigma^2) \\ &= (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i-\mu)^2}{2\sigma^2}\right) \\ \ln L(\mu|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^n y_i^2}{2\sigma^2} + \frac{\sum_{i=1}^n y_i^2}{\sigma^2} + \left(\frac{-n}{2\sigma^2}\right)\mu^2 \end{aligned}$$

If we take the second derivative of the log-likelihood, we get $\frac{-n}{\sigma^2}$. Since n and σ^2 are always positive, the second derivative is always negative.[4] For a fixed σ^2, in a function with only one parameter such as this one, a negative second derivative is sufficient for the likelihood to be convex.[5] As a result, this model is not multimodal. When estimated, the same parameter estimates will be returned regardless of the starting values.[6]

2.2 Mixture Models

Now consider a model where the stochastic component is a combination of Gaussians, instead of one Gaussian with a mean and standard deviation. Imagine a case where the dependent variable could be drawn from one of two different normal distributions. In this data-generating process the Gaussian distribution that the observation is drawn from is first chosen with a particular probability. Then, the value of the dependent variable is drawn from the chosen Gaussian with a particular mean and variance.

For example, say you were trying to model the height of people within a population. Further, you only observed the heights of the people in the population, not any other information about them. You might assume a model where first you draw with 0.5 probability whether the person is male or female. Based on their gender, you would draw the height either from a distribution with a "taller" mean (if the person were male) or from a normal distribution with a "shorter" mean (if the person were female). This is a simple mixture model, because as the data (the heights) would be drawn from a mixture of distributions.

Formally, the data-generating process for this model, a simple Gaussian mixture model, is as follows:

1. Randomly select a distribution d_i with probability $P(d_i) = w_i$, where $\sum w_i = 1$.

2. From the selected distribution, draw $y \sim N(\mu_i, \sigma_i^2)$.

The log-likelihood for this model becomes

$$\ln L(y|\mu_1, \mu 2, \sigma_1^2, \sigma_2^2) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} w_k N(y_n|\mu_k, \sigma_k^2) \right)$$

This model has more parameters to maximize than the normal regression model described in the previous section because (1) the probability of each distribution must be estimated *and* (2) the mean and variance of each distribution must be estimated. Further, the model is considered a *latent variable model* because the latent distribution variables d_i are not observed, but are rather generated as an intermediate step within the data-generating process. Because it is unknown from which distribution each data point comes (the data do not tell us which data points are men and which are women), we cannot solve this problem using the familiar tools of regression. In practice, the maximum likelihood estimate is typically solved using heuristics such as the expectation maximization algorithm (Dempster, Laird, and Rubin, 1977), which alternates between estimating the latent membership variable d_i (the unknown gender in our case) and the parameters of the distribution (the expected height and variance for each gender).[7]

It is easy to see that the estimates of each distribution's parameters will depend on the data points assigned to it and that the estimates of the latent variables will depend on distribution parameters. Because we need one to easily estimate the other, we choose a starting value to initialize our estimator. Unfortunately, different starting values can lead to different final solutions when the optimization method gets stuck in a local maximum. Despite the problems with multimodality, mixture models are often more accurate descriptions for data-generating processes than more traditional regression models, particularly for data that may have quite complicated underlying data-generating processes (e.g., Deb and Trivedi, 2002; DuMouchel, 1999; Fan, Han, and Liu, 2014; Grimmer and Stewart, 2013).

2.3 Latent Dirichlet Allocation

Later in the core sections of this chapter, we address approaches to dealing with multimodality in models of text data. In anticipation of this discussion, we now introduce the latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan, 2003), one of the most popular statistical models of text. We use the intuition from the simple mixture model described in the previous section to provide an intuition for why LDA and similar models are multimodal.

LDA is a mixed membership topic model, meaning that each document is assumed to be a "mixture" of topics. Topics are mathematically described as a probability vector over all V words within a corpus. For example, a topic about summer might place higher probabilities on the words "sun," "vacation," and "summer," and lower probabilities on words such as "cold" or "snow." Each topical vector has a probability assigned to each word within the corpus and therefore is a vector of length V. Topics are typically described by the most probable words for that corpus. The "topic matrix" β contains K (the number of topics estimated from the data) rows of topical vectors, each of length V.

For each document, the data-generating process first decides the number of words within the document N. Then, it draws how much of the document will be in each topic (out of K topics), assigning a probability to each of K topics in the vector θ ($\sum_K \theta = 1$). It then assigns each word within the document to a topic, with probabilities θ. Last, it draws each word for the document from each of the topic probability distributions in β.

More formally, the data-generating process for each document in LDA is as follows:

1. First, the length of the document is chosen from a Poisson, with prior η: $N \sim \text{Poisson}(\eta)$.
2. Next, the proportion of the document in each topic is drawn, with prior α: $\theta \sim \text{Dir}(\alpha)$
3. Last, for each of the N words,
   - A topic for the word is chosen: $z_n \sim \text{Multinomial}(\theta)$.
   - The word is chosen from the topic matrix β, selecting the topic that was chosen z_n: $w_n \sim \text{Multinomial}(\beta^{z_n})$.

The reader should already be able to note that LDA is a more complicated version of the mixture of Gaussians described previously in this section. First, we draw from a distribution that determines the proportion of a document within each topic and the topic assignment for each word. Then, given the topic assignment for each word, we draw the words that we observed within the documents. Although the process is much more complicated, it closely follows the previous section where first we drew a "latent" variable (the distribution (male or female) of the height) and then drew the data (height itself).

Similar to the mixture of Gaussians, optimization of LDA is difficult because of the "latent" parameters that must be drawn before the data is finally drawn. In LDA, these parameters are the proportion of a document in each topic (θ) and the topic assignment for each word (z_n) and are not observed. Similar to the mixture model case, we can optimize the model using a variant of the EM algorithm called variational EM.[8] In the expectation step, we first make a best guess as to the θ and z_n for each individual document, and in the maximization step, we optimize the remaining parameters (in this case β) assuming θ and z_n. We iterate between the expectation and maximization steps until convergence is reached.[9]

This approach maximizes the marginal likelihood (the probability of the data given β and α), which we can use as the objective function for maximizing the model. To get an intuition for the marginal likelihood, first we find the joint distribution of parameters and data:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

To find the probability of the words marginalized over the latent parameters, we integrate over z_n and θ:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) d\theta$$

The marginal likelihood itself is intractable in the case of LDA because of the coupling of β and θ, which leads to a an intractable integration problem. The variational EM approach uses Jensen's inequality to create a lower bound on the marginal likelihood, which we can maximize via coordinate ascent. That is, the algorithm is alternating between updating the content of the topics (β) and the topical makeup of a document (θ). It is this alternating maximization strategy that leads to multiple local optima. If we we could jointly optimize β and θ, we would likely have fewer issues of local modes, but the coupling in the marginal likelihood makes this unfeasible.

3 THE CASE OF TOPIC MODELS

Multimodality occurs in a huge number of statistical models.[10] In the rest of this chapter we focus on unsupervised latent variable models. In practice we use latent variable models to discover low-dimensional latent structure that can explain high-dimensional data. These models have been broadly applied throughout the social sciences to analyze large bodies of texts (Grimmer and Stewart, 2013), discover categories of diseases (Doshi-Velez, Ge, and Kohane, 2014; Ruiz et al., 2014), study human cognition (Tenenbaum et al., 2011), develop ontologies of political events (O'Connor, Stewart and Smith, 2013), build recommendation systems (Lim and Teh, 2007) and reveal the structure of biological and social networks (Airoldi et al., 2009; Hoff, Raftery, and Handcock, 2002). As we have suggested, the flexibility of latent variable models often leads to difficult statistical inference problems, and standard approaches often suffer from highly multimodal solutions.

Statistical topic models are rapidly growing in prominence within political science (Grimmer, 2010*a*; Quinn et al., 2010; Lauderdale and Clark, 2014; Roberts et al., 2014) as well as in other fields (Goldstone et al., 2014; Reich et al., 2015). Here we focus on latent Dirichlet allocation (LDA), which, as discussed in the previous section, models each document as a mixture over topics (Blei, Ng, and Jordan, 2003; Blei, 2012). The *mixed* membership form provides a more flexible representation than the single membership mixture model, but at the cost of an optimization problem with many more local optima.[11]

The posterior of the LDA model cannot be computed in closed form. Two popular approximate inference algorithms are collapsed Gibbs sampling (Griffiths and Steyvers, 2004) and variational inference (Blei, Ng, and Jordan, 2003). In this context, both methods can be seen as a form of alternating maximization; in Gibbs sampling we randomly draw from a single parameter conditional on the others, and in variational inference we update a single parameter averaging over the other parameters with respect to the approximating distribution (Grimmer, 2010*b*). This process of alternating conditional updates, necessitated by the inability to directly integrate over the posterior, leads to a sensitivity to the starting values of the parameters. The myriad solutions that can result from different starting points are well known among computer scientists, but are infrequently discussed.[12]

In fact, we can be more precise about the difficulty of the LDA inference problem by introducing some terminology from theoretical computer science. Nondeterministic polynomial-time-hard (NP-hard) problems are a class of problems that it is strongly suspected cannot be solved in polynomial time.[13] A more complete definition is beyond the scope of this chapter, but the classification conveys a sense of the difficulty of a problem. Maximum likelihood estimation can be shown to be NP-hard even for LDA models with only two topics (Sontag and Roy, 2011; Arora, Ge, and Moitra, 2012). These hardness results suggest not only why local optima are a characteristic of the LDA problem but also why they cannot be easily addressed by changes in the inference algorithm. That is, we can reasonably conjecture from these results that, without additional assumptions to make the problem tractable, it would be impossible to develop a computationally practical, globally optimal inference algorithm for LDA.[14]

How then do we address the practical problem of multimodality in topic models? In this section, we advocate selecting a solution using a broader set of criteria than just the value of the objective function. In the next section we make the argument for looking beyond the objective function when evaluating local modes. We then discuss some specific methods for choosing a single model for analysis. Finally we consider how to assess the stability of the chosen result across many different runs. Throughout we use LDA as a running example, but the arguments are more broadly applicable. In particular we see how they play out in an applied example using the related STM in subsequent sections.

3.1 Evaluating Local Modes

There is a disconnect between the way we evaluate topic models and the way we use them (Blei, 2012). The likelihood function and common evaluation metrics reward models that are predictive of unseen words, but our interest is rarely in predicting the words in a document; instead we want a model that provides a semantically coherent, substantively interesting summary of the documents (Grimmer and Stewart, 2013). This disconnect is not easily remedied; our models and evaluation metrics focus on prediction because it is the most tractable approximation to a human judgment of utility that ultimately must be made on a case-by-case basis. This perspective informs an approach to dealing with multimodality that emphasizes selecting a particular run not solely on the basis of which model yields the highest value of the objective function, but also includes other external assessments of model quality.

If our sole criterion of success were the ability to maximize the objective function, our path would be clear. We would simply generate a large number of candidate solutions by running the model repeatedly with different starting values and then select the one with the highest value. In variational approximations this metric is neatly defined in a single value: the lower bound on the marginal likelihood. We could simply calculate the bound for each model and choose the largest value.

In a general sense, this procedure is both intuitive and well supported theoretically. Not only is the lower bound the objective function we are optimizing but also, as a lower bound on the marginal evidence, it is precisely the quantity commonly used in approaches to Bayesian model selection (Kass and Raftery, 1995; Bishop et al., 2006; Grimmer, 2010*b*). These methods will pick the best model, given the assumptions of the data-generating process, but that may not be the one that is most interesting (Grimmer and Stewart, 2013). While for the purposes of estimating the model we need to rely on our assumptions about the data-generating process, we need not maintain these commitments when making our final selection. This allows us to access a richer set of tools for evaluating model quality.

The implication of this argument is that if we found the global optimum we might not choose to use it. This seems counterintuitive at first, but various forms of the argument have a long tradition in statistics. Consider the argument that we should choose a model on the basis of cross-validation or other forms of held-out prediction. This is the most commonly used evaluation metric for topic models (Wallach et al., 2009; Foulds and Smyth, 2014) and also has a strong tradition in political science (Beck, King, and Zeng, 2000; De Marchi, Gelpi, and Grynaviski, 2004; Ward, Greenhill, and Bakke, 2010). Selecting a model that maximizes a held-out predictive measure implies that we may not choose the model that maximizes the *in-sample* objective function. In settings where forecasting is the primary goal, the ability to predict a held-out sample is the clear gold standard; however, in the case of topic models, prediction is not the only relevant standard.

Implicit in this argument is the claim that the objective function need not directly correspond with human judgment. In human evaluations of topic coherence, selecting model parameters to maximize predictive log-likelihood can actually lead to a mild decrease in assessment of human interpretability (Chang et al., 2009; Lau, Newman, and Baldwin, 2014). Domain expert assessment (Mimno et al., 2011) and alignment to reference concepts

(Chuang et al., 2013) have consistently shown that selecting on the objective function alone does not necessarily yield the same model as human selection.

This is not to say that the objective function is completely useless; we have after all chosen to optimize it. Rather our claim is that among *locally optimal solutions*, model fit statistics provide a weak signal of model quality as judged by human analysts. Due to the nature of the optimization problem we find ourselves having fit a number of candidate models and given that we already have them, it would be wasteful to evaluate them only on the basis of the objective function.

One reaction to this situation would be to improve the objective of the model until it matched a human perception of quality. Unfortunately, this is theoretically impossible across all possible tasks (Grimmer and King, 2011; Wolpert and Macready, 1997). Moreover, the inference problem is already particularly complex, and modifications tend to result in even more intractable models (Mimno et al., 2011).

At the end of the day we trust the objective function enough to optimize it when fitting the model, but not enough to let it be the surrogate for the selection process. Instead, we want to explore the model and its implications, a process that is closely related to the literature on posterior predictive checks (Mimno and Blei, 2011; Blei, 2014; Gelman et al., 2013). In the next section we treat the question of how to choose a particular model for analysis, which we call the reference model. In Section 3.3 we explain how to assess the stability of results across multiple models.

3.2 Finding a Reference Model

Choosing a single reference model for analysis is challenging. The ideal selection criterion is the utility of the model for the analyst, which is an inherently subjective and application-specific assessment (Grimmer and King, 2011; Grimmer and Stewart, 2013). There is an inherent tradeoff in selection criteria between how time intensive the criterion is for the analyst and how closely it approximates the theoretical ideal. In this section we outline methods that span the range of high quality to highly automated.

Manual Review

The most thorough and time-intensive process is a manual review and validation of the model. This entails reading several example documents for each topic and carefully examine the topic-word distributions to verify that the topics are capturing a single well-defined concept. Depending on the number of topics and the length of the documents, this may be a daunting task in itself.

We may also want to consider information beyond the content of the documents themselves. In the social sciences we often have a rich source of additional information in document metadata. Mapping the relations between topics and a document's author (Grimmer, 2010*a*) or date (Quinn et al., 2010) is an important part of understanding if the model is functioning. When an existing typology of the documents is available, we can evaluate how well it corresponds to the inferred topics (Chuang et al., 2013). Ideally we hope that the model will convey some things we already know, allowing us to validate it, while also providing us with some novel insights. The different types of validation criteria have been well developed in the literature for measurement models and content analysis (Quinn et al. 2010; Grimmer and Stewart 2013; Krippendorff 2012).[15]

Manual evaluations of this sort are essentially custom procedures designed specifically for a particular analysis, and they require a large amount of an analyst's time. They are an important and necessary tool for validation of the final model, but are too expensive for evaluation of each candidate model.

Semi-Automated Analysis

A less labor-intensive approach is the human analysis of automated model summaries. The idea is to develop some generic tools for quickly evaluating a model, even if some human intervention is required to make a decision. For topic models we can summarize a topic by looking at the most probable or distinctive words. These word lists can be supplemented by focused reading of documents highly associated with a particular topic. These types of summaries arise naturally from the parameters of the model in the case of LDA, and most latent variable models have some approximate equivalents.

Recent work in information visualization has moved toward the development of automatically generated topic model browsers (Chuang, Manning, and Heer, 2012; Gardner et al., 2010; Chaney and Blei, 2012). Similar approaches have been used to provide browsers that focus on the exploration of covariate effects on word use (O'Connor, 2014). The best of these approaches embody the information visualization mantra of "overview first, zoom and filter, details on demand" (Shneiderman, 1996), which encapsulates the goal of a system that can seamlessly move from high-level model summaries such as word lists all the way down to the document reading experience. Some systems can even incorporate user feedback to allow for an interactive topic modeling experience (Hu et al., 2014). Visualization of topic models is an active area of research that promises to vastly improve the analyst's interaction with the model.

Complete Automated Approaches

The fastest evaluation metrics are those that are completely automated. The most natural metric is the objective function, which is generally either a bound or an approximation to the marginal likelihood (Grimmer, 2010*b*). The default standard within the computer science literature is held-out likelihood, which provides a measure of how predictive the model is for unseen documents

(Wallach et al., 2009; Foulds and Smyth, 2014). Evaluating how well the model predicts new data is appealing in its simplicity, but a predictive model need not be the most semantically interpretable.

Automated metrics can also be useful for narrowing the selection of candidate models that are then evaluated using more labor-intensive approaches. In Roberts et al. (2014) we consider two summary measures: semantic coherence (Mimno et al., 2011), which captures the tendency of a topic's high-probability words to co-occur in the same document, and exclusivity, which captures whether those high-probability words are specific to a single topic. We use these summaries as a coarse filter to focus our attention on a subset of promising candidate models.

Choosing a Balance

This provides only a coarse overview of some of the strategies for choosing a model. Necessarily, model choice is dictated by the particular problem at hand. Once a model is chosen there is always a subjective process of assigning a label to the topic, which implicitly involves arguing that the model representation (a distribution over words) is a good proxy for some theoretical concept represented by the label. Regardless of how the model is chosen, careful validation of the topic to ensure it fits with the theoretical concept is key (Grimmer and Stewart, 2013).

3.3 Assessing Stability

Once we have committed to a particular model and unpacked the publishable findings, we may want to know how stable the finding is across different initializations (i.e., starting values of the optimization algorithm). This serves two distinct purposes: first, we get a sense of how improbable it is that we found the particular local mode we are analyzing, and second, we learn how sensitive the finding is to other arrangements of the parameters.

The first purpose is the most straightforward. We want to build confidence in our readers and in ourselves that we did not stumble across the result completely by chance. The instability across individual runs of LDA has been criticized as unsettling by applied users across fields (Koltcov, Koltsova, and Nikolenko, 2014; Lancichinetti et al., 2014). Understanding how topics map on to the results across runs builds trust in the results (Chuang et al., 2013).

We can also use stability to assess how sensitive our finding is to other configurations of the topics. If a researcher identifies a topic as about "economics," is there some other version of that topic that looks substantially similar but yields contradictory results? These situations can arise when a particular topic or group of topics is of interest, but the model is sensitive to the way the remainder of the topics are allocated. Careful examination of the topic may confirm that it is about "economics," but that it fails to reveal similar content outside the topic that might reasonably be included. Examining the "economics" topic across a large set of models provides a sense of the different representations of the topic supported by the data.

4 SIMILARITY BETWEEN TOPICS ACROSS MODES

In this section we develop tools for assessing the stability of findings of interest across local modes. We start by setting up a running example that uses STM to analyze a corpus of political blogs. We then illustrate several approaches to assessing how similar a pair of topics is to each other. We then show how these metrics can be aggregated to the topic level, to the model level, or across covariates.

The methods we present here serve two related purposes. First, we provide some intuition for the variety of solutions that arise from local modes. Especially for those primarily familiar with globally convex models, this provides a sense of what to expect when using or reading about latent variable models. The methods themselves can also be useful as diagnostics for practictioners. Indeed we show through examples how examination of stability can lead to useful insights about the data and model.

4.1 Political Blogs

To make our discussion concrete we turn to a specific data set: a collection of 13,246 blog posts from American political blogs written during the 2008 presidential election (Eisenstein and Xing, 2010).[16] Six blogs – American Thinker, Digby, Hot Air, Michelle Malkin, Think Progress, and Talking Points Memo – were used to construct the corpus. Each blog is given a rating: liberal or conservative. For each blog post the day of the post is recorded. After stemming and removing a standard list of stopwords and words that appeared in fewer than 1% of the documents, there is left a vocabulary of 2,653 words.

To analyze these texts we use STM (Roberts et al., 2014). STM is a mixed membership topic model in the style of LDA that allows for the inclusion of document-level covariates, in this case rating (liberal/conservative) and time (day of the post). We use the `stm` package in `R` that uses a fast variational EM algorithm. We specify topic prevalence as a function of the partisan rating and a smooth function of time. We estimated the model 685 times, initializing with a short run of LDA (we return to this in Section 5).[17] We note that this set of runs holds a number of things constant, including choices in preprocessing (e.g., stopword removal, stemming) and specification of the model (e.g., the STM prevalence formula, number of topics) that could also lead to differences in model fit.

We briefly define a minimal amount of notation for use in later sections. Let $K = 100$ be the user-selected number of topics, $V = 2{,}653$ be the size of the vocabulary, and $D = 13{,}246$ be the number of documents. Mixed membership topic models, including LDA and STM, can be summarized by two matrices

Topic 18:
law, court, rule, constitut, right, judg, decis, suprem, legal, justic, case, feder, requir, amend, protect, gun, govern, allow, appeal, citizen

Topic 30:
presid, vice, cheney, offic, presidenti, first, execut, dick, decis, leader, role, histori, nation, branch, power, part, govern, order, idea, washington

Topic 48:
global, warm, research, climat, studi, chang, scienc, scientist, gore, caus, human, scientif, earth, emiss, planet, cell, environment, report, water, green

Topic 60:
iran, nuclear, threat, weapon, iranian, program, missil, north, bomb, defens, korea, strike, sanction, intern, build, militari, intellig, capabl, pose, develop

Topic 71:
black, wright, white, race, church, racial, racist, pastor, jeremiah, africanamerican, racism, african, comment, reverend, king, controversi, rev, view, communiti, south

FIGURE 2.1. Five example topics from the reference model. These are given the labels Supreme Court, Cheney, global warming, Iran/N.K. nukes, and Wright, respectively.

of parameters. β is a row-normalized K-by-V matrix of topic-word distributions. The entry $\beta_{k,v}$ can be interpreted as the probability of observing the v-th word in topic k. θ is a row-normalized D-by-K matrix of the document-topic distributions. The entry $\theta_{d,k}$ can be interpreted as the proportion of words in document d that arise from topic k. Both LDA and STM can be framed as a factorization of the row-normalized D-by-V empirical word count matrix W, such that $W \approx \theta\beta$. We use the θ and β matrices to compare the models.

To simplify the resulting discussion, we choose as our reference model the sample maximum of the variational bound. We do not recommend using the sample maximum in general as the selection criteria (for reasons discussed in previous section), but it allows us to proceed more quickly to the comparison of results.

The hundred topics estimated in the model cover a huge range of issues spanning the political dimensions of the 2008 presidential election. We select five topics that illustrate different properties of stability to use as running examples. Figure 2.1 shows the top 20 most probable words for each of the example topics: Supreme Court rulings, Vice President Cheney, global warming research, nuclear weapons issues in Iran and North Korea, and the controversy surrounding Barack Obama's former pastor, Jeremiah Wright.

4.2 Comparing Topics

Our first step is to ask whether there are any differences between the different runs of the model at all. If each run is equivalent up to numerical precision, the question of multimodality would be moot. To answer this question we need a way to measure whether two topics generated across different runs are in fact comparable.

We can compare the similarity of two models by comparing the topic-word distribution β or the document-topic distribution θ. Using β implies that two topics are considered similar if they generate similar observed words. Using θ assesses two topics as similar if they load in the same patterns across the corpus. Although both approaches are useful, β will tend to contract on the true posterior faster than θ, resulting in a less noisy measure. This is because the number of documents will tend to grow faster than the number of unique words in the vocabulary. Before proceeding to pairwise similarity metrics, we need to align topics across runs.

Alignment

Consider a simple case where we have two runs of the model. We first need to establish which two topics from each run to compare. The topic numbers are arbitrary across each run, which on its own is unproblematic, but means that we need to do something additional in order to compare topics to each other across runs. We call the process of deciding which topics to compare the "process of alignment." The alignment itself is determined by some metric of similarity typically on the topic-word distribution. Here we use the inner product between the rows of β.

Given the similarity metric there are at least two reasonable approaches to aligning topics, both of which will yield the same result when the topics are in fact identical up to permutation of the topic numbers. First, we can let each topic in one run of the model choose its favorite in another run of the model, even if that involves a topic being chosen multiple times. We call this process "local alignment" because each topic in the reference model is making a local choice that is independent of the choices of all other topics. A second approach is to choose a one-to-one matching that maximizes the sum of similarities across all the topic pairs. We call this the "global alignment" because each topic's match is contingent on the selection of all other topics. Although this formulation results in a combinatorial optimization problem, it can be solved efficiently using the Hungarian algorithm (Kuhn, 1955).[18] We use global alignment here. The local alignment produced essentially the same relative trends.

Pairwise Similarity

Once we have a candidate alignment we can calculate distance metrics between two topics across model runs. An intuitive measure of distance is the L_1 norm, which is the sum of the absolute value of the difference. It is defined as

$$L_1 = \sum_v \left|\beta^{\text{ref}}_{k,v} - \beta^{\text{cand}}_{k,v}\right|$$

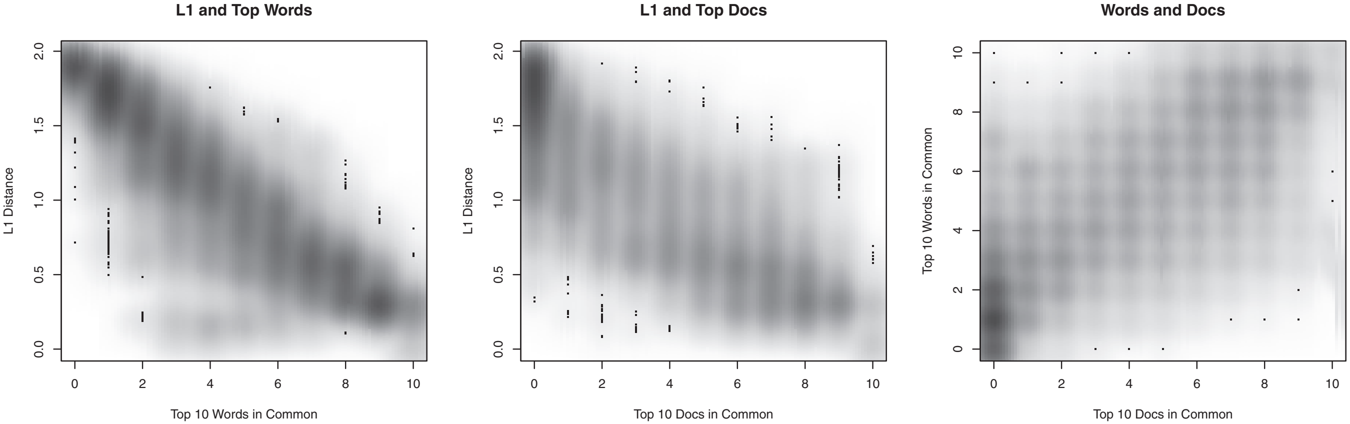


FIGURE 2.2. Relation between three measures of topic similarity across all topics and modes. Plotted surface is a kernel smoothed density estimate.

and has a range: [0,2]. We discuss alternate metrics, but we use L_1 because the result is easy to conceptualize. We discuss the implications of alternative distance metrics in Section 4.5.

We need not constrain ourselves to distance metrics on the parameter space. As an alternative, we compare the number of the top 10 most probable words shared by the reference topic and its match. The result ranges from $\{0, \ldots, 10\}$, indicating the number of words matched.

We can establish the comparable metric for documents. Ranking documents by their use of a particular topic, we can count the overlap in the number of the 10 documents most strongly associated with a topic. This metric ranges from $\{0, \ldots, 10\}$ with 10 indicating complete agreement in the two sets.

Figure 2.2 plots the relations between each of these three metrics across the aligned topics. Each pair of metrics is strongly correlated in the theoretically anticipated direction. Also as expected, the measure based on the documents is somewhat noisier than the corresponding measure based on the words.

This figure also provides us with some insight on the similarities across solutions. Topics range from nearly perfectly aligned to having almost no correspondence. This suggests that there are substantial semantic differences across local modes that could lead to significant differences in interpretation.

4.3 Aggregations

The pairwise similarities shown in Figure 2.2 are useful for contextualizing the full range of topic pairs; however, to make these metrics more interpretable it is helpful to aggregate up to either the model level or the topic level. Aggregation at the model level gives us a sense of how well the local modes approximate the reference model by taking the average over each topic. Aggregation to the topic level gives us information about how stable a given topic in the reference model is across runs.

Model-Level Aggregations

We start with aggregations to the model level. In this case we have a natural summary metric of the complete model: the approximation to the bound on the marginal likelihood.

In Figure 2.3 we plot each of the three similarity metrics on the Y-axis against the approximate bound on the X-axis. The outlier (upper right corner of the first two plots, and lower right of the third) is the reference model, which is, by definition, an exact match for itself. The dashed line marks a natural reference point (5 of 10 words or documents in the left two plots, and an L_1 distance in the middle of the range for the third). The solid line shows the simple linear trend.

The trend between the lower bound and the other three similarity metrics suggests that the objective function can be useful as a coarse measure of

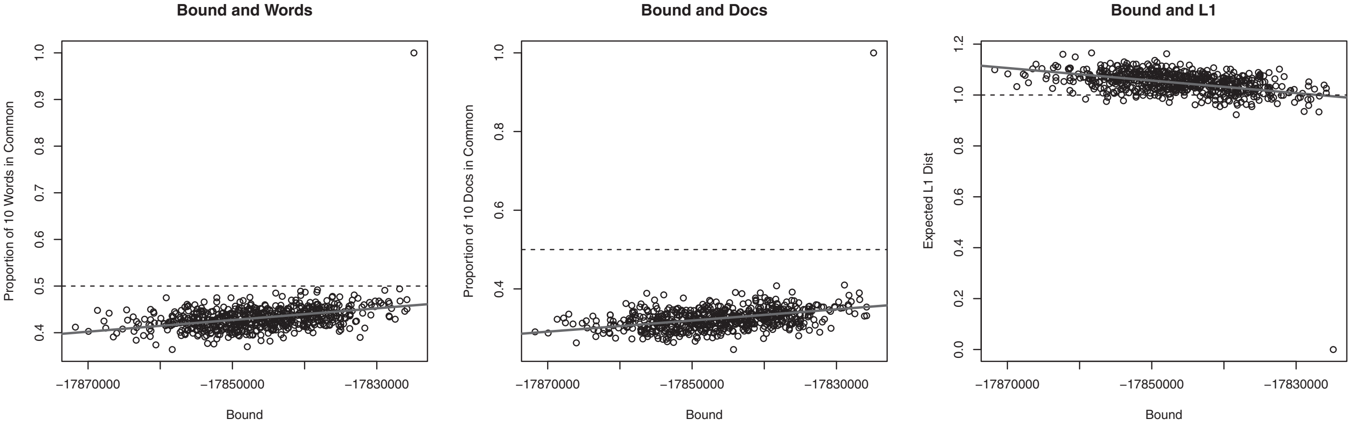


FIGURE 2.3. Comparison between the approximation to the bound on the marginal likelihood (the objective function) with similarity metrics aggregated to the model level.

similarity. That is, as the bound of each of the runs approaches the reference model, all three metrics reveal similarity increasing on average. However, it is only a coarse metric because of the large variance relative to the size of the trend. The high variance around the trend reinforces the observation that, among candidate models with comparable levels of model fit (as measured by the objective function), there is considerable semantic variety in the discovered topics.

Topic-Level Aggregations

Aggregation to the topic level provides us with a measure of how stable a topic within the reference model is across different runs. This helps address the applied situation where a researcher has identified a topic of interest, but wants some understanding of how frequently it occurs across multiple runs of the model.

The distribution over topics is plotted in Figure 2.4 where each topic is represented by the average value of the statistic over the different model runs. The five example topics are each denoted by the dashed lines and a label. In each plot the distribution varies over essentially the full range of the metric, indicating that some topics are extremely stable across all of the runs whereas others are essentially unique to the reference model.

The example topics help explain where some of this variance is coming from. The climate change topic is one of the most stable across all three of the metrics. This reflects the rather specialized language in these blog posts. In a political context, words such as “climate” are very exclusive to a particular topic. These specialized words help pin down the topic, resulting in fewer distinct locally optimal solutions.

One of the least stable topics across runs is the Cheney topic. In the reference model the topic is primarily about Vice President Cheney, whereas other models include broader coverage of the Bush presidency. As an example we chose the local model that is farthest away from the reference model in L_1 distance. In Table 2.1 we compare the two versions of the topic by comparing the topic-specific probabilities of observing 18 terms. These terms define the set of words that have probability of at least 0.01 in one of the two models. We can see that, although both topics discuss Cheney, the local model discusses President Bush using words such as Bush, Bush’s, and George, which have negligible probability under the reference model version of the topic.

Topic-level stability analysis focuses the analyst’s attention on the semantic content covered by a topic. As an analyst, our responsibility is to choose a label for a topic that clearly communicates to the reader what semantic content is included in a topic. We emphasize that an unstable topic is not inferior or less substantively interesting. Depending on the question, a topic that combines discussion of Cheney and the Bush presidency may be more interesting than a topic that just covers the vice president. However, the instability in the topic alerts us that the topic in the reference model is specific to Cheney, with discussion of the Bush presidency being included in a separate topic.

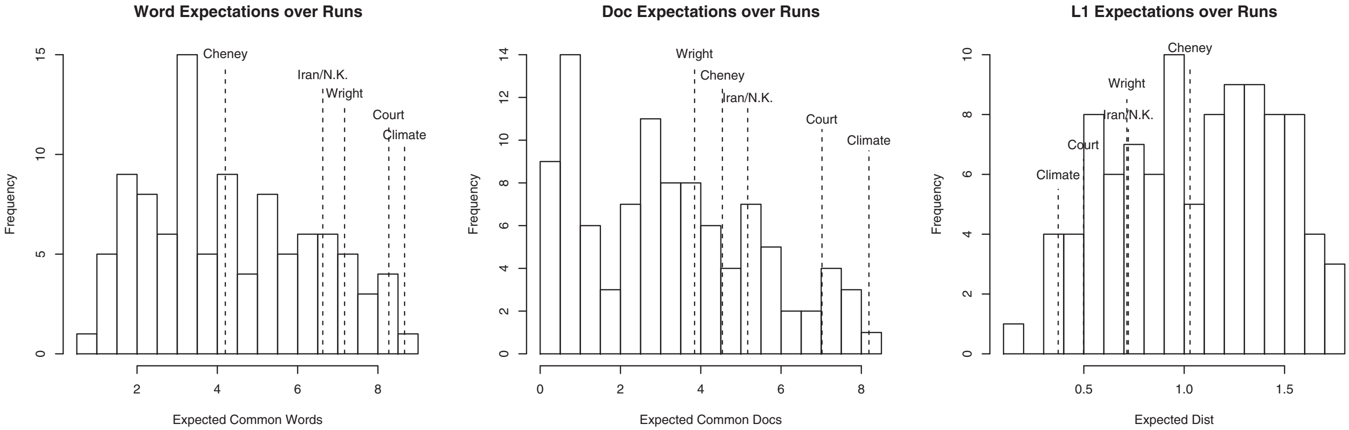


FIGURE 2.4. Aggregation of similarity metrics to the topic level.

TABLE 2.1. *Topic-Specific Probabilities of Observing 18 Words in the Cheney Topic in Both the Reference Model and a Local Solution Far Away from It*

Term	Ref. Model	Local Model
administr	<.0005	0.104
bush	<.0005	0.275
bush'	<.0005	0.0191
cheney	0.0464	0.0279
decis	0.0178	0.0060
dick	0.0195	0.0109
execut	0.0226	0.0022
first	0.0253	0.0001
georg	<.0005	0.0480
histori	0.0104	0.0099
leader	0.0134	<.0005
nation	0.0102	<.0005
offic	0.0414	0.0209
presid	0.5302	0.2868
presidenti	0.0254	0.0003
role	0.0129	0.0001
term	0.0025	0.0130
vice	0.0512	0.0251

Note: Included words have a probability of at least 0.01 under one of the two versions of the topics. The reference model topic is focused primarily on Vice President Cheney, whereas the local mode includes broader coverage of the Bush presidency.

4.4 Covariate Effect Stability

In applied use of STM, we are often interested in the role played by covariates in driving topical prevalence. Indeed this is a principal advantage of the STM framework: it allows for the inclusion of covariate information in the estimation process and facilitates the estimation of covariate effects on the resulting model. In the Poliblog corpus, we can examine the role of partisanship in topical coverage. We start by unpacking the partisanship effects for our example topics in the reference model. We then show how to assess the stability of these findings across other local modes.

Unpacking Covariate Effects

Figure 2.5 plots the expected proportion of topic use in conservative blogs minus the expected proportion of topic use in liberal blogs under the reference model. Thus topics more associated with the conservative blogs appear to the right of zero.

We briefly contextualize the partisan effects in this set of topics. Conservative attention to the Supreme Court topic is primarily driven by the June 2008

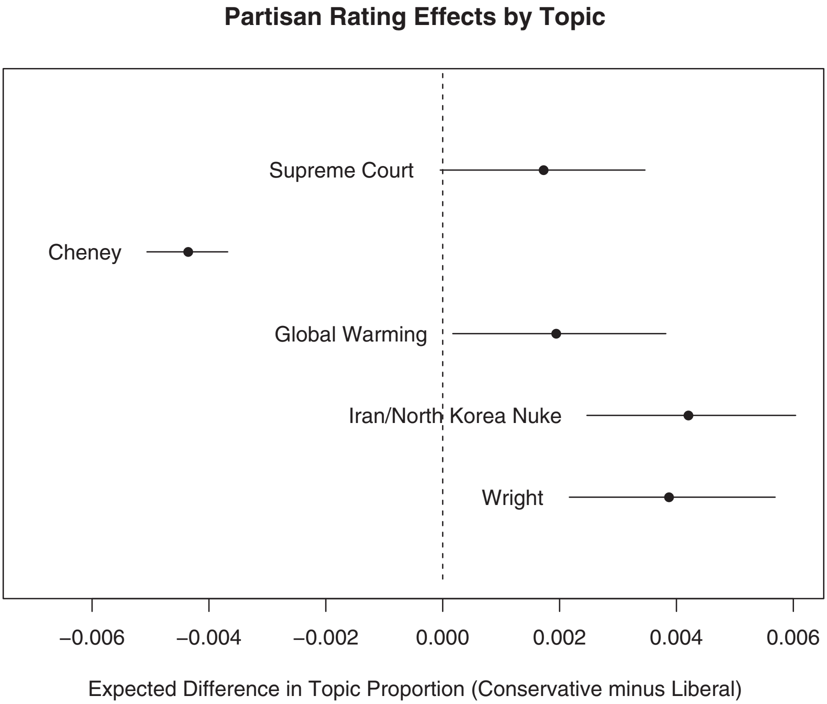
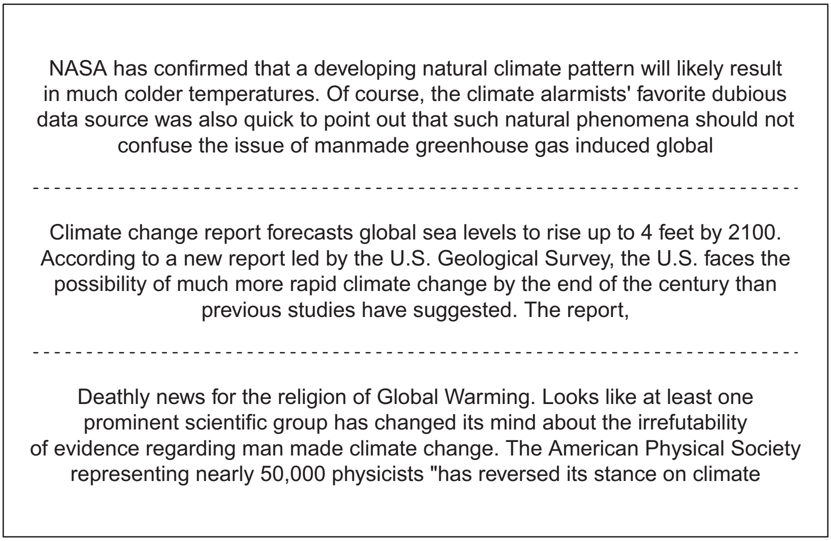


FIGURE 2.5. Differences in topical coverage by rating (controlling for time). Effects to the right of 0 indicate a topic more heavily used by conservatives. Lines indicate 95% confidence intervals using the "global" approximation to measurement uncertainty (Roberts et al., 2014).

Heller v. *District of Columbia* case that Struck down parts of the *Firearms Control Regulations Act of 1975* on Second Amendment grounds. As discussed in the previous section the Cheney topic is primarily about Dick Cheney's legacy on the vice presidency. The coverage is mainly from liberal blogs and is predominantly critical in tone.

The greater conservative attention to global warming is initially surprising given that it is typically a more liberal issue, but it should be remembered that these blogs were posted in 2008, which was before the more recent trend (at time of writing) in liberal assertiveness. We explore this further by examining the posts most associated with this topic. Figure 2.6 shows the first 300 characters of the three posts most associated with the topic. The first and third posts are critical of global warming, whereas the second post describes a report warning against climate change. The first and third are as expected from a conservative blog, and the second is from a Liberal blog.

The Iran and North Korea nuclear weapons topic shows a conservative effect consistent with increased attention to security topics and consistent



NASA has confirmed that a developing natural climate pattern will likely result in much colder temperatures. Of course, the climate alarmists' favorite dubious data source was also quick to point out that such natural phenomena should not confuse the issue of manmade greenhouse gas induced global

Climate change report forecasts global sea levels to rise up to 4 feet by 2100. According to a new report led by the U.S. Geological Survey, the U.S. faces the possibility of much more rapid climate change by the end of the century than previous studies have suggested. The report,

Deathly news for the religion of Global Warming. Looks like at least one prominent scientific group has changed its mind about the irrefutability of evidence regarding man made climate change. The American Physical Society representing nearly 50,000 physicists "has reversed its stance on climate

FIGURE 2.6. The first 300 characters of the three posts most associated with the global warming topic. Posts 1 and 3 come from *American Thinker* and post 2 comes from *Think Progress*.

with conventional views that issue ownership of security is much greater for Republicans. Finally the scandal involving Reverend Jeremeiah Wright, which is critical of then Democratic primary candidate Barack Obama, is more prevalent on conservative blogs.

Stability across Models

How stable are these effects are across other plausible local modes? A simple way to evaluate this stability is to align the topics to the reference model and then calculate the effect for each topic.[19] Although this process produces a distribution over effect sizes, it is important to emphasize the conceptual challenges in interpreting the results. Each model is estimating the effect of the partisan rating, but on a slightly different version of the topic. Thus differences arise for two reasons: the document-topic assignments may be different, but also the topics themselves capture different concepts. The alignment ensures that this concept is the most similar to our reference model (given the alignment method and the similarity metric), but they are not necessarily conceptually identical.

Figure 2.7 plots the distribution of effect sizes. Beginning with the first plot on the top left, we see that the partisan effect for the Supreme Court topic in the reference model has one of the largest observed values across all of the local modes. Not only is the reference model effect out in the tail but also the distribution over effect sizes includes negative as well as positive values. What accounts for this difference? Comparing the most probable words in the

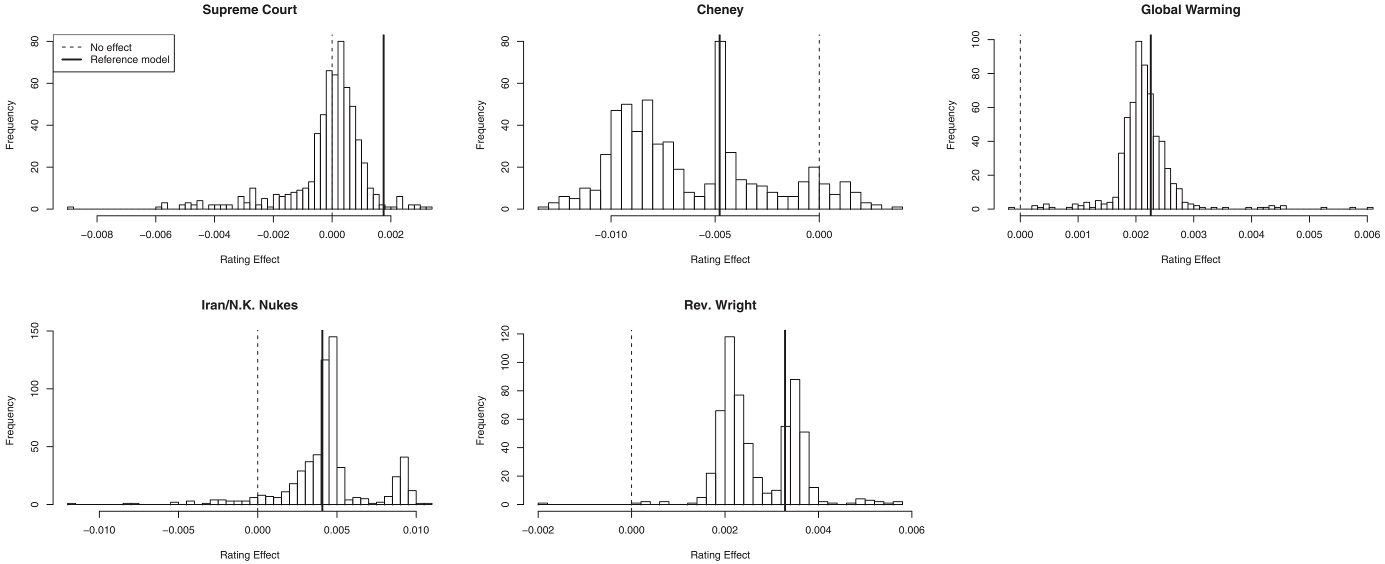


FIGURE 2.7. Distribution of the partisan rating effect across modes for the five example topics. The black solid line shows the effect at the reference mode, and the black dashed line marks an effect size of 0.

reference model with those in an aligned topic for one of the models with a strong liberal effect provides an indication of the differences:

Reference Model: law, court, rule, constitut, right, judg, decis, suprem, legal, justic, case, feder, requir, amend, protect, gun, govern, allow, appeal, citizen

Local Mode: court, tortur, law, justic, legal, rule, judg, suprem, case, interrog, detaine, lawyer, cia, constitut, guantanamo, decis, prison, violat, prosecut, administr

The local mode includes significant discussion of the legal issues surrounding the use of torture and the operation of Guantanamo Bay. By contrast, our reference has a completely separate topic that captures this discussion (top words: tortur, prison, cia, interrog, detaine, use, guantanamo). Thus the fact that the effect size we found is considerably out in the tail of the histogram does not mean that the finding is not valid, but it does suggest that it is very sensitive to the content of the legal cases and the way in which relevant information about legal issues is spread across the other topics.

The second plot in Figure 2.7 shows the Cheney topic. Here we see a distribution with three modes where the reference model sits directly on top of the most typical point. Following the discussion in the previous section, this reflects the difference between having the topic focus exclusively on Vice President Cheney as opposed to including the broader Bush presidency.

The global warming case (third plot) is the most clear-cut, with most of the solutions producing extremely similar effect sizes. This reflects the relatively specialized vocabulary in discussing climate change, which allows the allocation of topics to be less ambiguous across solutions.

The Iran and North Korea topic is a case where, like the Supreme Court topic there is substantial spread across the models. However, in contrast to the Supreme Court topic, the reference model is quite close to the majority of the solutions. Here the largest source of variation is primarily in whether both Iran and North Korea are grouped within the same topic.

Finally, the topic on Reverend Wright shows another case where the reference model is largely consistent with the local modes. There is some distinction between topics that contain coverage of the scandal and those that also contain elements of the positive liberal coverage that followed Barack Obama's speech on the matter ("A More Perfect Union").

These examples highlight the value of local modes for contextualizing the finding in our reference model. By seeing alternative models, such as a Supreme Court topic that focuses on either gun control or the use of torture, we become more attuned to exactly what concepts are included within the model. This in turn allows us to choose labels that more precisely represent the topic's semantic content.

Differences from Alignment

While most of these analyses are insensitive to the method of aligning topics, we do observe significant differences in the covariate effects. Global alignments tend to result in more cases where there are several clusters of effect sizes. Consider for example, the Cheney topic (top-center of Figure 2.7). In the example discussed in Section 4.3 we saw that the matching topic in another model included both discussion of the Bush presidency and Cheney. If the global alignment had assigned that topic to the Bush reference model topic, that would leave it unavailable for the Cheney reference model topic. This tends to manifest in the covariate effect distributions as clusters of certain covariate effect sizes. We still find the global alignment the most useful, however, because it ensures that we are not omitting any topics from the comparison models.

4.5 Additional Comparisons and Related Work

The examples provided here focused on a particular data set with a specific number of topics. Here we briefly discuss findings from additional settings and related work in the literature.

Different Number of Topics

We ran the set of experiments discussed earlier under the same data set with $K = 50$ topics and observed essentially the same patterns and trends reported. Smaller experiments at $K = 20$ revealed higher levels of instability across runs with increased instances of topics that are very poorly aligned. We conjecture that this is primarily a matter of how well the number of topics fit the specific data set, rather than a statement about small numbers of topics in general.[20] If instability was solely a function of the number of topics, we would expect substantially poorer performance in this extreme case. That the instability would be connected to selecting too few topics for a given data set certainly makes intuitive sense, but additional investigation would be necessary to make conclusive statements.

Alternative Distance Measures

In the results discussed earlier, we used two basic measures of distance between the topic-word distributions. We aligned the topics using a dot product measure, and we presented calculations based on L_1 distance. We also performed experiments using a cosine similarity metric (essentially the dot product rescaled by the L_2 norm of the vectors).

The results, depicted in Figure 2.8, show slightly less clear correlations between the similarity metric and the top words and top documents measure. Specifically there are many cases where high cosine similarity topic appears with a comparatively low number of top words or documents in common. Manual examination of topics in these settings demonstrated that this was primarily

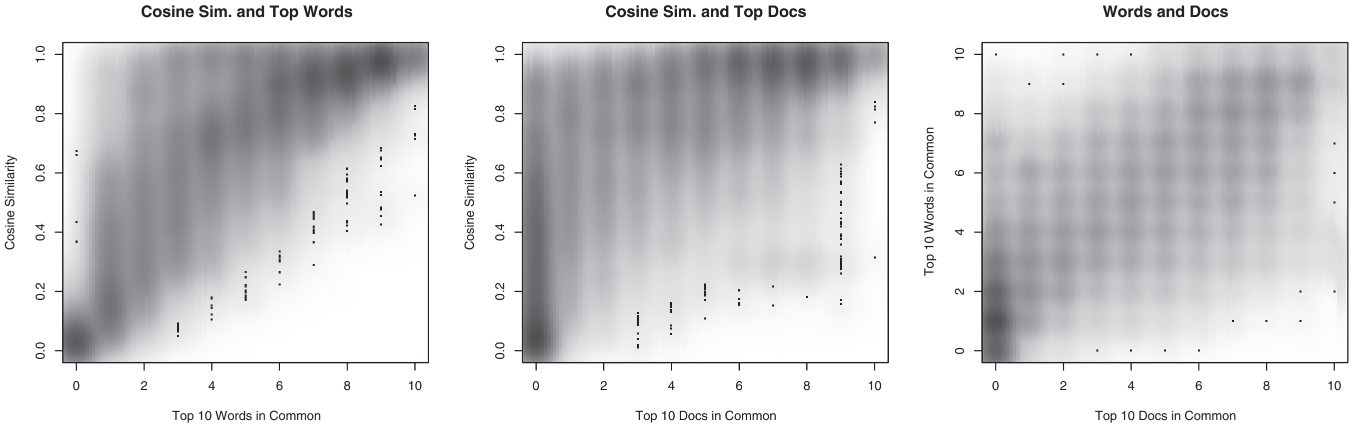


FIGURE 2.8. Comparison of metric based on cosine similarity.

connected with topics where the majority of the probability mass loaded onto fewer than 10 words.[21]

Koltcov, Koltsova, and Nikolenko (2014), in a similar investigation of stability in LDA, guard against the possibility of L_1-style calculations being dominated by the long tail of infrequently occurring words. To guard against this we tested a version where we only calculated the distance over the minimal set of words accounting for 75% of a topic's probability mass within the reference model. The results are substantially the same, but with slightly less noise. We opted to maintain the versions we presented earlier to allow for simpler interpretation.

Alternative Approaches

The similarity metrics described here are automated approximations to semantic similarity. All of the metrics equally penalize deviations from the reference model, regardless of whether they are in the direction of a semantically related word or not. One solution would be to embed words within a vector space such that semantically related words are close together and then calculate differences relative to this space (Mikolov et al., 2013). This has the advantage of more sharply penalizing differences between topics that involve words that are semantically unrelated. However, to perform the word embeddings, we need an extremely large text corpus, which limits the applicability to smaller document settings.[22]

Finally, our focus here has primarily been on estimating similarity across a large number of models. Chuang et al. (2013) focus on comparing two topic models and introduce a rich typology of correspondence between them, including topics that are fused, repeated, junk (unmatched), or resolved (well matched) relative to the reference model. These comparisons require a bit more technical machinery, but can elegantly handle comparisons between a reference and candidate model with different numbers of topics.

This section has presented several approaches to comparing topics across different runs of a model. These methods provide not only a measure of the reference model's stability but also can often give the analyst useful diagnostic information about the contents of the topics. The discussion, however, leaves open the important question of whether there are ways to increase the quality of model runs at the estimation stage. In the next section we discuss approaches to initialization that maximize the quality of the initial runs.

5 INITIALIZATION

When the function we are optimizing is well behaved and globally concave, any starting point will result in the same global solution. Thus initialization of the parameters becomes a trivial detail, possibly chosen to save on computational costs.[23] In the multimodal setting, our initialization influences our final solution. When the computational cost of inference in the model is extremely low, we can simply randomly initialize the parameters and repeat until we have identified the same maximum several times. However, in latent variable models not only may we never encounter a repeat solution but also each solution to the model may be very computationally expensive, a problem that is exacerbated in big-data settings. If fitting a topic model on a million documents takes a week of computational time, rerunning it a thousand different times is not a reasonable strategy. A well-known but little-discussed aspect of statistical optimization is that careful initialization can be an incredibly powerful tool (McLachlan and Peel, 2004; Murphy, 2012).

Before returning to the case of topic models, we consider the simpler case of *k*-means, a central algorithm in the clustering literature closely related to the normal mixture model discussed in Section 2.2. The *k*-means example helps provide some intuition about the role of "smart" initialization. In Section 5.2, we return to the case of topic models and discuss how simpler models such as LDA can be used to initialize more complex models such as STM. In Section 5.3, we provide a simulation study that shows that the LDA-based initialization yields higher values of the approximate evidence lower bound than random initialization.

The initialization approaches we consider in this section are stochastic, and so each time the procedure is repeated we may obtain a different solution. Thus our goal is to initialize such that we produce better solutions in expectation. In special cases such as *k*-means, we may even be able to obtain provable guarantees on the number of trials necessary to come within a certain tolerance of the global solution.

An alternative approach is to explore deterministic approaches to initialization. In Section 6 we outline very recent research that yields deterministic initializations with excellent performance.

5.1 *k*-Means

The *k*-Means algorithm is arguably the central algorithm of the clustering literature. Not only is it important in its own right as a problem in clustering and computational geometry but it is also a common component of larger systems. Because algorithms for *k*-means are extremely fast and easily parallelized, it has widespread applications in big-data settings (Bishop et al., 2006).[24]

k-Means algorithms use an alternating optimization strategy to find a partition of units into *k* distinct clusters such that Euclidean distance between the units and their nearest center is minimized. Finding the optimal partition of units under the *k*-means objective function is a combinatorial optimization problem that is known to be NP-hard (Mahajan, Nimbhorkar, and Varadarajan, 2009). This manifests itself in a tendency of *k*-means algorithms to get stuck in local optima. Nevertheless, it is the most widely used clustering algorithm in practice.

Under the most popular heuristic, cluster centers are chosen randomly from the data points (Lloyd, 1982). Estimation then proceeds by iterating between assigning data points to their closest center and recomputing the location of the cluster center given those points. The result is an incredibly fast procedure, but one that can produce arbitrarily bad partitions relative to the global optimum (Arthur and Vassilvitskii, 2007).

A substantial advance in the literature on the problem came with the development of the *k*-means++ algorithm (Arthur and Vassilvitskii, 2007). The idea is extremely simple: by using a careful seeding of the initial centers we can make probabilistic guarantees on recovery relative to the optimal solution. The seeding strategy is based on selecting the first center uniformly at random from the data points and then choosing subsequent centers at random, but reweighting to prioritize data points that are not near a previously chosen center.

The *k*-means++ algorithm highlights an important general point: carefully considering the initialization procedure can be an important tool for dealing with multimodality in practice. This is an important difference from problems that are globally convex, where starting values are important only for increasing speed or avoiding numerical instability. It is interesting to note that, despite being both simple conceptually and incredibly effective in practice, the *k*-means++ heuristic was not discovered until 25 years after Lloyd's algorithm. Heuristics for solving this problem continue to be an active area of research (Bahmani et al., 2012; Nielsen and Nock, 2014).

5.2 What Makes a Good Initialization?

A good initialization strategy needs to balance the cost of solving for the initial state with the expected improvement in the objective. If the cost of finding the initial values of the parameters is high relative to the model-fitting process, then you might as well use that computational time to randomly restart the original algorithm. Thus the art to initializing a model is finding a procedure that places the model in the right region of the parameter space with as few calculations as possible. The *k*-means++ algorithm is an excellent example of an incredibly low-cost initialization.

In cases where the the model itself is straightforward and the cost of inference rises rapidly with the number of units, a simple but powerful strategy is to run the model on a small subsample of the data. This is generally a good default, particularly in the big-data regime where the computation is costly solely due to scale.

Another steadfast default approach is to initialize a complicated model with a simpler model or algorithm for which inference is easy. The simpler algorithm can often put you into a good region of the parameter space without expending the higher costs of the more complex method. Indeed, this is why the *k*-means algorithm is often used to initialize more complex mixture models (McLachlan and Peel, 2004; Bishop et al., 2006).

In the case of STM, there is a natural simpler model, LDA. Due to the Dirichlet-multinomial conjugacy in LDA we can perform inference using a fast collapsed Gibbs sampler (Griffiths and Steyvers, 2004). They key here is

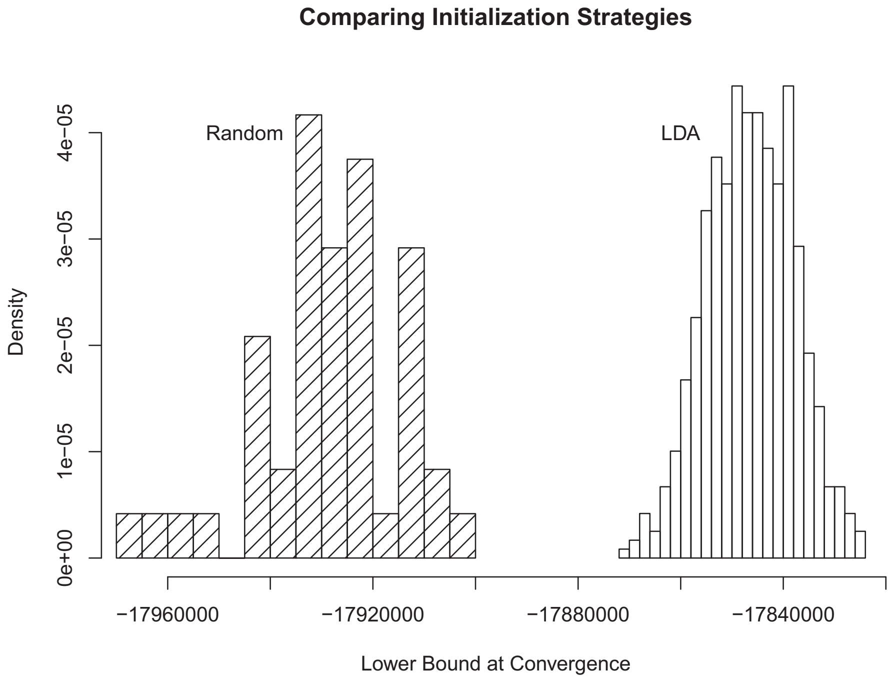


FIGURE 2.9. A comparison of initialization strategies for the $K = 100$ STM models.

that the conjugacy of the model allows for all parameters except the token-level topic latent variables to be integrated out. The result is a very fast sampler that has been heavily optimized (Yao, Mimno, and McCallum, 2009). The cost of inference is linear in the number of individual words (tokens) in the text.[25]

Because LDA is itself multimodal, the result is an initialization that is different each time. Thus like the *k*-means++ algorithm, this approach places STM in a good region of the parameter space, but still allows for variation across runs. The initialization for the LDA algorithm itself is just a random assignment of the tokens, so we do not have a problem of infinite regress.

5.3 The Effects of Initialization

Unlike the case of the *k*-means++ algorithm, we cannot make theoretical guarantees on the quality of LDA as a method for initializing STM.[26] This naturally leads us to ask about how it performs as an initialization in practice. To investigate this issue we compared the objective function values in the 685 model runs initialized with LDA to a set of 50 runs initialized from random starting values.[27] Figure 2.9 plots the resulting distributions over the final level of the objective function.

These substantial gains come at a very low computational cost courtesy of the efficient Gibbs sampler in the lda package (Chang, 2012). The initialization process takes only a few seconds to complete 50 iterations of the 2.6 million tokens in the Poliblog data. Indeed this is why initializing with LDA is the current default method in the stm package in R. Furthermore, not only do the LDA initialized models performed uniformly better but they also converged significantly more quickly. Most of the LDA models took between 60 to 120 iterations to converge, whereas the randomly initialized versions took close to 200 iterations. Interestingly, we were not able to increase the average quality by running the sampler longer, suggesting that without considerable further effort this may be close to the optimal strategy for this type of initialization.

6 GLOBAL SOLUTIONS

In the previous sections we discussed how nonconvex models can lead to inference algorithms that exhibit multimodality. For the important case of topic models we provided a series of tools both for exploring a set of local modes and for improving the average quality of our solutions through careful initialization. These approaches work well in settings where it is feasible to run the model many times. However, in the truly big-data setting, every single optimization of the model may be so costly that we want to strictly limit the number of times we run the model.

In this section we introduce recent innovations in theoretical computer science that allow for global optimization of nonconvex models using *spectral learning*. As we show, these algorithms introduce additional assumptions into the model to achieve tractable inference with provable guarantees of recovering the globally optimal parameters. Following the logic of Section 5, we use an algorithm for LDA as an initialization to the STM. Our results suggest that this hybrid strategy can be a useful technique for tackling big-data problems.

We remind the reader that these techniques are very much "on the frontier," and so the substantive implications for applied projects have not been charted out, something that is beyond the scope of this chapter. Furthermore, we emphasize that these initialization strategies do not "solve" the multimodality problem. These techniques do not yield a correct answer, and even though they do very well at maximizing the approximate evidence lower bound, this does not mean that the solution is optimal with respect to other criteria (as discussed earlier). The types of robustness exercises discussed earlier should continue to be an important part of the research process. Nevertheless, we find that these deterministic initialization procedures are a promising contribution to the topic modeling toolkit.

6.1 Introduction to Spectral Learning

When we define an inference procedure we would like to be able to prove that the algorithm will converge to the global optimum. For the types of problems that we discuss here, we generally settle for heuristics, such as expectation-maximization, which has provable convergence to a local optimum (Dempster, Laird, and Rubin, 1977), or MCMC algorithms, which have no finite sample guarantees but will asymptotically recover the posterior (Robert and Casella, 2004). In practice both approaches get stuck in local optima.

Here we describe a class of spectral learning algorithms for estimating the parameters of latent variable models while retaining guarantees of globally optimal convergence.[28] The key insight is that by using matrix (or array) decomposition techniques we can recover the parameters from low-order moments of the data. This approach relies on a method of moments inferential framework, as opposed to the likelihood-based framework we have adopted thus far (Pearson, 1894; King, 1989; Anandkumar, Ge, Hsu, Kakade and Telgarsky, 2014). In models with certain structures this can lead to procedures with provable theoretical guarantees of recovering the true parameters, as well as algorithms that are naturally scalable.

Spectral algorithms have been applied to a wide array of models: Gaussian mixture models (Hsu and Kakade, 2013), hidden Markov models (Anandkumar, Hsu, and Kakade, 2012), latent tree models (Song, Xing, and Parikh, 2011), community detection on a graph (Anandkumar, Ge, Hsu and Kakade, 2014), dictionary learning (Arora, Ge, and Moitra, 2014), and many others (Anandkumar, Ge, Hsu, Kakade and Telgarsky, 2014). Of particular interest for our purposes is the development of spectral approaches to estimating topic models (Arora, Ge, and Moitra, 2012; Anandkumar, Liu, Hsu, Foster, and Kakade, 2012). There are two basic approaches to spectral learning in LDA that differ in their assumptions and methods. For clarity we focus on a simple and scalable algorithm developed in Arora, Ge, Halpern, et al. (2013).

The discussion of these methods is unavoidably more technical than the previous material. However, the common theme is straightforward: we are making stronger assumptions about the model in order to obtain an algorithm that does not suffer from problems of local modes. Importantly for our case we use the spectral algorithm as an initialization, rather than as a procedure to fit the model. In doing so we weaken our reliance on the assumptions in the spectral algorithm while still achieving its desirable properties. In this sense the spectral learning algorithms are complementary to the likelihood-based approach we have considered here (Anandkumar, Ge, Hsu, Kakade and Telgarsky, 2014).

6.2 An Algorithm for LDA

Here we briefly describe the intuition behind the inference algorithm of Arora, Ge, Halpern, et al. (2013) that uses a non-negative matrix factorization (NMF)[29] to recover the model parameters from the word co-occurrence matrix, as we show later, to separate the β parameter (the topic distributions) from the data. The main input to the algorithm is a matrix of word-word co-occurrences that is of size V-by-V where V is the number of the words in the vocabulary. Normalizing this matrix so all entries sum to 1, we get the matrix Q. If we assume that Q is constructed from an infinite number of documents, then it is the second-order moment matrix, and the element $Q_{i,j}$ has the interpretation as the probability of observing word i and word j in the same document. We can write the Q matrix in terms of the model parameters as

$$\begin{aligned} Q &= \mathbb{E}\left[\beta^T\theta^T\theta\beta\right] && (1) \\ &= \beta^T\mathbb{E}\left[\theta^T\theta\right]\beta, && (2) \end{aligned}$$

where the second line follows by treating the parameters as fixed but unknown. Arora, Ge, Halpern, et al. (2013) show that we can recover β^T from the rest of the parameters using a non-negative matrix factorization.

The NMF problem is also NP-hard in general (Vavasis, 2009) and suffers from the same local mode problems as LDA in practice (Gillis, 2014). However recent work by Arora, Ge, Kannan, and Moitra (2012) showed that we can provably compute the NMF for the class of matrices that satisfy the *separability* condition (Donoho and Stodden, 2003). In this context, separability assumes that for each topic there is at least one word, called an anchor word, that is assigned only to that topic. The anchor word for topic k does not need to be in every document about topic k, but if a document contains the anchor word, we know that it is at least partially about topic k. Separability implies that all non-anchor word rows of the Q matrix can be recovered as a convex combination of the anchor rows (Arora, Ge, Halpern, et al., 2013). Thus if we can identify the anchors, we can solve for β using convex optimization methods.

Thus the algorithm of Arora, Ge, Halpern, et al. (2013) proceeds in two parts. First we identify the anchors, and then given the anchors we uncover the model parameters β. Crucially these steps do not need to be iterated and are not sensitive to the starting values of the algorithm. There are many different approaches to these two steps that differ in computational complexity and robustness to noise (Kumar, Sindhwani, and Kambadur, 2012; Recht et al., 2012; Gillis and Luce, 2014; Ding, Rohban, Ishwar, and Saligrama, 2013).[30]

Advantages

The main advantage of the Arora, Ge, Halpern, et al. (2013) algorithm is that we can give theoretical guarantees that it will recover the optimal parameters (given the model and separability assumption). In practice this means that we completely sidestep the multimodality concerns described in this chapter. The second crucial advantage is that the method is extremely scalable. Note that Q is V-by-V, and thus the algorithm does not increase in complexity with the number of documents. This means that, for a fixed vocabulary size, the cost of doing inference on a million documents is essentially the same as inference for a hundred. This is an incredibly useful property for the big-data setting. Many of the algorithms cited earlier for other models are similarly scalable.[31]

Disadvantages

Naturally there are practical drawbacks to spectral algorithms. Because we are substituting the observed sample moments for the population moments, spectral methods require a lot of data to perform well. In experiments on synthetic data reported in Arora, Ge, Halpern, et al. (2013), spectral methods only approach the accuracy of Gibbs sampling at around 40,000 documents. This is particularly troubling because as the power-law distribution of natural language ensures that we will need an incredibly large number of documents to estimate co-occurrences of highly infrequent words. In practice this is addressed by filtering out low-frequency words before performing anchor selection.

The second major concern is that spectral methods lean more heavily on the model assumptions, which can lead to somewhat less interpretable models in real data (Nguyen, Hu, and Boyd-Graber, 2014). Finally, as a practical matter the spectral method only recovers the topic word distributions β so additional methods are still required to infer the document-topic proportions. These can be obtained by a single pass of Gibbs sampling or variational inference (Roberts et al., 2014).

6.3 Spectral Learning as Initialization

Here we apply the Arora, Ge, Halpern, et al. (2013) algorithm as an initialization for the structural topic model. Using the spectral method as an initialization weakens our reliance on the assumptions of the methods. For example, our initialization will have anchor words, but once we begin variational inference of STM, those anchor words are free to move some of their probability mass onto other topics. Thus we simply use the spectral algorithm to place us into an optimal region of the space. Because the spectral method is deterministic, we also only need to run the model once.

We apply the algorithm as an initialization for the same 100-topic model of the Poliblog corpus used previously. Note that the approximately 13,000-document corpus is smaller than previous findings would suggest are necessary to match the quality of Gibbs sampling.

Figure 2.10 shows the results of the model with the spectral initialization. Not only is the result dramatically better with respect to the lower bound than the random and LDA initializations but also the model converged considerably faster as well.[32] Because our focus here is on introducing this class of algorithms, we do not go through the process of reinterpreting the 100-topic model.

6.4 Future Directions

Spectral algorithms are a very active area of current research. Here we focused on a particular algorithm that leverages non-negative matrix factorization under a separability assumption. There have been several algorithmic

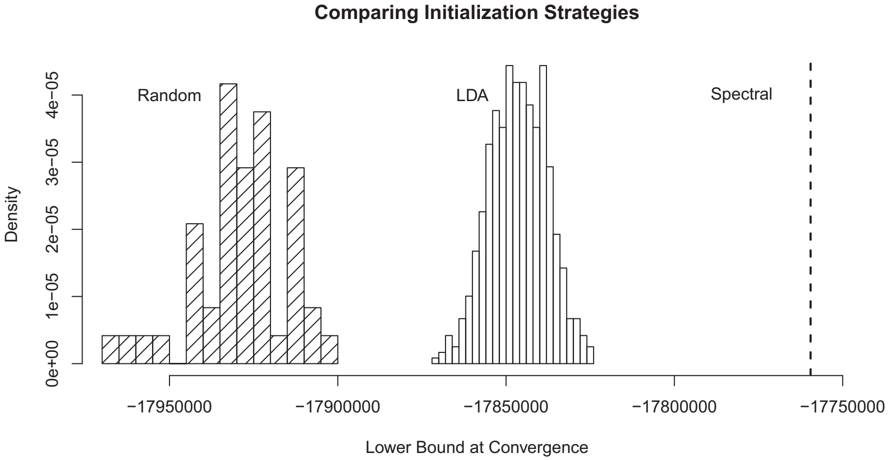


FIGURE 2.10. A comparison of the spectral initialization strategy to random and LDA for the $K = 100$ STM models. The dashed line denotes the result of the spectral initialized solution.

improvements since Arora, Ge, Kannan, and Moitra (2012) introduced the anchor-based method (Recht et al., 2012; Kumar, Sindhwani, and Kambadur, 2012; Ding, Rohban, Ishwar, and Saligrama, 2013; Gillis and Luce, 2014; Gillis, 2014; Zhou, Bilmes, and Guestrin, 2014). There has also been substantial work applying the approach to other problem domains (Arora, Ge, Moitra, and Sachdeva, 2012; Arora, Ge, and Moitra, 2014; Arora, Bhaskara, Ge, and Ma, 2014; Zhou, Bilmes, and Guestrin, 2014).

A separate line of work uses higher order moments of the data along with tools for array (tensor) decomposition (Anandkumar, Ge, Hsu, Kakade and Telgarsky, 2014). These methods have also resulted in algorithms for an incredibly rich set of applications and models. Importantly we can also use this framework to develop algorithms for LDA with provable global convergence guarantees (Anandkumar, Liu, Hsu, Foster, and Kakade, 2012; Anandkumar, Hsu, Javanmard, and Kakade, 2013).[33] This work differs in both the assumptions and methods used. Crucially the tensor method of moments approach uses the third moments of the data, which may require an even higher sample size to accurately estimate.[34]

7 CONCLUSION

Alongside rapid increases in data and processing power have come the development and deployment of a range of new data analysis tools. All of these tools enable new insights and new ways of looking at data that even a decade ago would have been difficult. In this chapter, we focus on the problem of multimodality that affects many of these tools, with specific attention to topic models for textual data. The purpose of this chapter has been to convey an understanding of where this multimodality comes from and then engage in a sustained discussion about what to do about multimodality from an applied perspective when analyzing text data.

Any modeling approach requires transparency about both process and guiding principles. The topic models we focus on in this chapter are no different in this respect from more traditional statistical tools. Even in traditional general linear models, there is always the choice of model specification in both variables and functional form. Although multimodality brings new issues to the table, the responsibility of the researcher to carefully validate the chosen model is fundamentally the same. This is true regardless of whether the choice between competing models arises due to a nonconvex latent variable model or due to the selection of an important model-tuning parameter in a globally convex problem. Thus even if multimodality is an unfamiliar problem, social scientists can draw on the same set of best practices that they employ throughout their research.

An important practical contribution of this chapter is that it extends the set of tools available to scholars using topic models in applied research. While we have focused on STM, many of the procedures we use are helpful for a broader class of latent variable models. For instance, the approaches to aligning topics and calculating stability across runs can all be applied directly to the broader class of statistical topic models and with minor modifications to most latent variable models.

We see great potential for the analysis of "big" data in the social sciences, but rather than focus on the data we have taken a more methodological focus. We think this has important implications not only for methodological development but also could structure the types of questions we ask and the types of data sets we seek to build. Methodologically, we think that there will be important advances in areas such as optimal initialization strategies, which will be especially important as our data sets grow in size. From an applied perspective, users will be unlikely to want to wait for extended periods of time to get even a single set of results. Advances in computational power need to be matched with smart ways to leverage that power. From a research design perspective, we think more focus should be put on bringing greater structure to so-called unstructured data. In the STM we focus on the inclusion of metadata for modeling and hypothesis testing, but this is only one possible use. Can more direct supervision help us with issues of multimodality? Of course, in the end, big data will be at its best when there is active dialogue between those who pose the big question and those who might provide the big answers.

Notes

1. In this chapter, we refer to convex optimization problems and convex models as those where the likelihood is globally concave and therefore has one maximum, instead of a globally convex likelihood with one minimum. Our main interest, however, is in the number of modes the likelihood has.

2. There exist a large number of optimization procedures for finding optima of a particular function; see Boyd and Vandenberghe (2009) for a review.
3. This model is equivalent to a normal linear regression that only models the intercept; that is, without regressors.
4. See King (1989) for a more in-depth discussion of this example.
5. For multidimensional likelihoods, if the Hessian is positive definite, the model will be strictly convex (only has one optimum); if it is positive semi-definite, it will be convex (two points may share a optimum on the same plane.)
6. Other normal linear regression models that are sometimes used in big data applications include lasso (Tibshirani, 1996).
7. Although see additional strategies for the lower dimensional case in Kalai, Moitra, and Valiant (2012).
8. Variational inference provides an approximation to the posterior distribution that falls within a tractable parametric family, unlike EM, which provides a point estimate of the model parameters. Here we simplify some of the differences between these approaches by referring to variational inference as optimizing the "model parameters" rather than the parameters of the approximating posterior. For more information, see Jordan et al. (1998); Grimmer (2010*b*); and Bishop et al. (2006).
9. The posterior distribution of LDA can also be estimated using Gibbs sampling; see Griffiths and Steyvers (2004) for more information.
10. For example, neural network models (Cochocki and Unbehauen, 1993), which allow for layered combinations of the model matrix, are extremely useful for modeling more complex data-generating processes (Beck, King, and Zeng, 2000). However, they too often suffer from extremely multimodal likelihoods, and rarely is the global maximum found (Bishop et al., 2006; De Marchi, Gelpi, and Grynaviski, 2004). Additional examples include Bayesian nonparametric processes (Teh et al., 2006; Griffiths and Tenenbaum, 2004), hidden Markov models (Rabiner and Juang, 1986; Park, 2012), switching time series models (Hamilton, 1989), and seemingly unrelated regression models (Srivastava and Giles, 1987; Drton and Richardson, 2004), to name a few. The item response (IRT) model (Hambleton, 1991), popular in political science (Poole and Rosenthal, 1997), is unidentified because solutions that are rotations of each other can exist for the same set of data (Poole and Rosenthal, 1997; Rivers, 2003). To estimate the model, a few parameters must first be pinned down before the rest of the parameters can be known. In essence, there are multiple and sometimes equally likely solutions to the same problem. While different from multimodality in the previous examples, "multiple solutions" of an unidentified likelihood can also be classified under models with likelihoods that have multiple modes.
11. LDA, and mixture models more generally, have $K!$ substantively identical modes arising from posterior invariance to label switching (i.e., permutation of the order of the topics). This type of multimodality is only a nuisance because each of the modes will yield the same inferences in an applied setting.
12. For example, Blei (2012) provides an excellent overview of LDA and related models, but does not mention the issue of local optima at all. The original paper introducing LDA mentions local optima only in passing to warn against degenerate initializations (Blei, Ng, and Jordan, 2003). Notable exceptions to this trend are Koltcov, Koltsova, and Nikolenko (2014) and Lancichinetti et al. (2014), which investigate the stability more directly, as do our the efforts in this chapter.

13. That is, if $P \neq NP$ then this is the case. However, there is no formal proof that $P \neq NP$.
14. The exact connection between *NP*-hard complexity and local modes is difficult to concisely state. Not all convex problems can be provably solved in polynomial time (de Klerk and Pasechnik, 2002). However it is sufficient for the argument here to establish that the hardness results imply that there is something inherently difficult about the nature of the problem, which makes it unlikely that a computationally practical algorithm with global convergence properties exists without adding assumptions.
15. Quinn et al. (2010) present five types of validity for topic models: external, semantic, discriminant, predictive, and hypothesis.
16. The CMU Poliblog corpus is available at http://sailing.cs.cmu.edu/socialmedia/blog2008.html, and documentation on the blogs is available at http://www.sailing.cs.cmu.edu/socialmedia/blog2008.pdf. A sample of 5,000 posts is also available in the `stm` package.
17. Each model is run to convergence (a relative change of less than 10^{-5} in the objective).
18. The Hungarian algorithm is a polynomial time algorithm for solving the linear sum assignment problem. Given a K by K matrix, where entry i, j gives the cost of matching row i to columns j, the Hungarian algorithm finds the optimal assignment of rows to columns such that the cost is minimized. The Hungarian algorithm guarantees that this can be solved in $O(K^3)$ time (Papadimitriou and Steiglitz, 1998). We use the implementation in the `clue` package in `R` (Hornik, 2005).
19. This is similar to the permutation test methodology developed in Roberts et al. (2014). In Roberts et al. (2014) we are interested in testing whether our finding on the effect of the binary treatment indicator is driven by including it as a topic prevalence covariate (that is, are we at risk of baking in our conclusion?). We randomly permute the treatment indicator across documents and rerun the model. In each case we calculate the *largest* treatment effect observed within the data across *all topics* and compare this distribution to the observed level. If we were baking in the conclusion, the model would discover large treatment effects even though the treatment indicator had been randomly assigned. In practice the observed effect is substantially larger than the randomly permuted data sets, suggesting that the model is working as expected. Here we are *aligning the topics* first and then comparing effect sizes across model runs.
20. In Roberts et al. (2014) we examined a small open-ended survey response data set with $K = 3$ and found results to be extremely stable even under a more demanding permutation test.
21. Chuang et al. (2013) presented a number of different distance metrics (e.g., testing KL divergence, cosine metric, and Spearman rank coefficient) against human judgments of similarity. They find that the cosine metric most directly matches human judgment and that it could even be further improved using a rescaled dot product measure that they introduced. The strong findings for the cosine metric provide an interesting contrast to Figure 2.8 and suggest that it may perform better in other circumstances.
22. An alternate strategy is to cast the notion of distance between topics entirely in the realm of human judgments. This is essentially the approach of Grimmer and

King (2011), which offers experimental protocols for evaluating similarity between topics.

23. We mean well behaved because in practice even globally convex problems can be sensitive to starting values due to practical issues in numerical optimization.
24. By easily parallelized, we mean that it can be easily fit into the Map-Reduce paradigm (Dean and Ghemawat, 2008). The algorithm is still serial in the iterations, but the expensive calculations within each iteration can be performed in parallel.
25. Also crucially the collapsed sampler mixes dramatically faster than an uncollapsed version (Carpenter, 2010; Asuncion Jr., 2011). By integrating out the topic-word distribution β we are implicitly updating the global parameters every time we take a new sample at the document level. As a result we only need a few passes through the data to reach a good region of the parameter space.
26. Such a theoretical analysis is likely possible under a certain set of assumptions, but would lead to a lengthy and technical digression here.
27. Specifically we initialize topic-word distributions with random draws from a Dirichlet distribution and set the document-topic proportion prior mean to zero. This is the commonly used initialization procedure in many variational algorithms for LDA.
28. Spectral methods derive their name from the use of tools from linear algebra that are connected to the spectral theorem. Here we use an inclusive definition of spectral learning that includes methods using a variety of matrix and array decomposition techniques beyond the canonical singular value decomposition.
29. NMF is similar to a singular value decomposition except that all elements of the decomposition are constrained to be non-negative.
30. Anchor selection methods use either a sparse regression framework (Recht et al., 2012) or appeal to geometric properties of the anchors (Kumar, Sindhwani, and Kambadur, 2012). See Gillis (2014) for a summary of these approaches. For our experiments here, we focus the approach defined in Arora, Ge, Halpern, et al. (2013), which falls into the geometric properties camp. They use a combinatorial search based on a modified Gram Schmidt orthogonalization process for the anchor selection. Parameter recovery then uses an exponentiated gradient descent algorithm (Kivinen and Warmuth, 1997) with an L_2 norm loss.
31. A good example is the mixed membership stochastic blockmodel, which is, loosely speaking, LDA for community detection on a network (Airoldi et al., 2009). Huang et al. (Forthcoming) give a spectral algorithm that learns hundreds of communities in a network of millions of nodes in under 10 minutes.
32. It took 25 iterations to converge after the spectral initialization, compared to 60 iterations for LDA initialization and close to 200 iterations for random initialization.
33. Technically the work in Anandkumar, Liu, Hsu, Foster, and Kakade (2012) uses an approach called excess correlation analysis, which involves two singular value decompositions on the second and third moments of the data. The approach based on the tensor method of moments strategy is described in Anandkumar, Ge, Hsu, Kakade and Telgarsky (2014) and applies to a wider class of models. We group them together here because they emerged from the same research group and use similar techniques.
34. An excellent discussion of differing assumptions of spectral methods is given in Ding, Ishwar, Rohban, and Saligrama (2013).

References

Airoldi, Edoardo M., David M. Blei, Stephen E. Fienberg, and Eric P. Xing. 2009. "Mixed membership stochastic blockmodels." In *Advances in Neural Information Processing Systems*. pp. 33–40.

Anandkumar, Animashree, Rong Ge, Daniel Hsu, and Sham M. Kakade. 2014. "A tensor approach to learning mixed membership community models." *Journal of Machine Learning Research* 15:2239–2312. http://jmlr.org/papers/v15/anandkumar14a.html.

Anandkumar, Animashree, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. 2014. "Tensor decompositions for learning latent variable models." *Journal of Machine Learning Research* 15:2773–2832. http://jmlr.org/papers/v15/anandkumar14b.html.

Anandkumar, Animashree, Yi-kai Liu, Daniel J. Hsu, Dean P. Foster, and Sham M. Kakade. 2012. "A spectral algorithm for latent Dirichlet allocation." In *Advances in neural information processing systems*. pp. 917–925.

Anandkumar, Animashree, Daniel Hsu, Adel Javanmard, and Sham Kakade. 2013. "Learning linear bayesian networks with latent variables." In *Proceedings of the 30th International Conference on Machine Learning*. pp. 249–257.

Anandkumar, Animashree, Daniel Hsu, and Sham M. Kakade. 2012. "A method of moments for mixture models and hidden Markov models." *arXiv preprint arXiv:1203.0683*.

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. "A practical algorithm for topic modeling with provable guarantees." In *Proceedings of the 30th International Conference on Machine Learning*. pp. 280–288.

Arora, Sanjeev, Rong Ge, Ravindran Kannan, and Ankur Moitra. 2012. "Computing a nonnegative matrix factorization–provably." In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*. ACM. pp. 145–162.

Arora, Sanjeev, Aditya Bhaskara, Rong Ge, and Tengyu Ma. 2014. "Provable bounds for learning some deep representations." In *Proceedings of the 31st International Conference on Machine Learning*. 32. pp. 584–592.

Arora, Sanjeev, Rong Ge, and Ankur Moitra. 2012. "Learning topic models – going beyond SVD." In *Foundations of Computer Science (FOCS), 2012 IEEE. 53rd Annual Symposium on*. IEEE. pp. 1–10.

Arora, Sanjeev, Rong Ge, and Ankur Moitra. 2014. "New algorithms for learning incoherent and overcomplete dictionaries." In *Proceedings of The 27th Conference on Learning Theory*. 35. pp. 779–806.

Arora, Sanjeev, Rong Ge, Ankur Moitra, and Sushant Sachdeva. 2012. "Provable ICA with unknown Gaussian noise, with implications for Gaussian mixtures and autoencoders." In *Advances in Neural Information Processing Systems*. pp. 2375–2383.

Arthur, David and Sergei Vassilvitskii. 2007. "k-means++: The advantages of careful seeding." In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. pp. 1027–1035.

Asuncion Jr., Arthur Uy. 2011. "Distributed and accelerated inference algorithms for probabilistic graphical models." Technical report. California State University at Long Beach.

Bahmani, Bahman, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. 2012. "Scalable k-means++." *Proceedings of the VLDB Endowment* 5(7):622–633.

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving quantitative studies of international conflict: A conjecture." *American Political Science Review* 94:21–36.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-dimensional methods and inference on structural and treatment effects." *Journal of Economic Perspectives* 28(2):29–50.

Bishop, Christopher M., et al. 2006. *Pattern recognition and machine learning*. Vol. 1. Springer: New York.

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55(4):77–84.

Blei, David M. 2014. "Build, compute, critique, repeat: Data analysis with latent variable models." *Annual Review of Statistics and Its Application* 1:203–232.

Blei, David M, Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning Research* 3:993–1022.

Boyd, Stephen and Lieven Vandenberghe. 2009. *Convex optimization*. Cambridge University Press.

Carpenter, Bob. 2010. "Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling." Technical report. LingPipe.

Chaney, Allison June-Barlow and David M. Blei. 2012. "Visualizing topic models." In *International Conference on Web and Social Media*.

Chang, Jonathan. 2012. *lda: Collapsed Gibbs sampling methods for topic models*. R package version 1.3.2. http://CRAN.R-project.org/package=lda.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. "Reading tea leaves: How humans interpret topic models." In *Advances in Neural Information Processing Systems*. pp. 288–296.

Chuang, Jason, Christopher D. Manning, and Jeffrey Heer. 2012. "Termite: Visualization techniques for assessing textual topic models." In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM. pp. 74–77.

Chuang, Jason, Sonal Gupta, Christopher Manning, and Jeffrey Heer. 2013. "Topic model diagnostics: Assessing domain relevance via topical alignment." In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. pp. 612–620.

Cochocki, A. and Rolf Unbehauen. 1993. *Neural networks for optimization and signal processing*. John Wiley & Sons.

Dean, Jeffrey and Sanjay Ghemawat. 2008. "MapReduce: Simplified data processing on large clusters." *Communications of the ACM* 51(1):107–113.

Deb, Partha and Pravin K. Trivedi. 2002. "The structure of demand for health care: Latent class versus two-part models." *Journal of Health Economics* 21(4):601–625.

de Klerk, Etienne and Dmitrii V. Pasechnik. 2002. "Approximation of the stability number of a graph via copositive programming." *SIAM Journal on Optimization* 12(4):875–892.

De Marchi, Scott, Christopher Gelpi, and Jeffrey D. Grynaviski. 2004. "Untangling neural nets." *American Political Science Review* 98(2):371–378.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*. pp. 1–38.

Ding, Weicong, Prakash Ishwar, Mohammad H. Rohban, and Venkatesh Saligrama. 2013. "Necessary and sufficient conditions for novel word detection in separable topic models." *arXiv preprint arXiv:1310.7994*.

Ding, Weicong, Mohammad H. Rohban, Prakash Ishwar, and Venkatesh Saligrama. 2013. "Topic discovery through data dependent and random projections." *arXiv preprint arXiv:1303.3664*.

Donoho, David and Victoria Stodden. 2003. "When does non-negative matrix factorization give a correct decomposition into parts?" In *Advances in Neural Information Processing Systems*. pp. 1141–1148.

Doshi-Velez, Finale, Yaorong Ge, and Isaac Kohane. 2014. "Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis." *Pediatrics* 133(1):e54–e63.

Drton, Mathias and Thomas S. Richardson. 2004. "Multimodality of the likelihood in the bivariate seemingly unrelated regressions model." *Biometrika* 91(2): 383–392.

DuMouchel, William. 1999. "Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system." *American Statistician* 53(3):177–190.

Efron, Bradley et al. 1978. "The geometry of exponential families." *Annals of Statistics* 6(2):362–376.

Eisenstein, Jacob and Eric Xing. 2010. "The CMU 2008 Political Blog Corpus."

Fan, Jianqing, Fang Han, and Han Liu. 2014. "Challenges of Big Data analysis." *National Science Review* 1:293–324.

Foulds, J. R. and P. Smyth. 2014. "Annealing paths for the evaluation of topic models." In *Proceedings of the Thirtieth Conference Conference on Uncertainty in Artificial Intelligence*.

Gardner, Matthew J., Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. 2010. "The topic browser: An interactive tool for browsing topic models." In *NIPS Workshop on Challenges of Data Visualization*.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian data analysis*. CRC press.

Gillis, Nicolas. 2014. "The why and how of nonnegative matrix factorization." In *Regularization, Optimization, Kernels, and Support Vector Machines*. J.A.K. Suykens, M. Signoretto and A. Argyriou (eds), Chapman & Hall/CRC, Machine Learning and Pattern Recognition Series. pp. 257–291.

Gillis, Nicolas, and Robert, Luce. 2014. "Robust near-separable nonnegative matrix factorization using linear optimization." *Journal of Machine Learning Research* 15 (Apr). pp. 1249–1280.

Goldstone, Andrew and Ted Underwood, et al. 2014. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." New Literary History 45, no. 3:359–384.

Griffiths, D.M.B.T.L. and M.I.J.J.B. Tenenbaum. 2004. "Hierarchical topic models and the nested Chinese restaurant process." *Advances in Neural Information Processing Systems* 16:17.

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

Grimmer, Justin. 2010*a*. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18(1):1–35.

Grimmer, Justin. 2010*b*. "An introduction to Bayesian inference via variational approximations." *Political Analysis* 19(1):32–47.

Grimmer, Justin. 2013. *Representational style in Congress: What legislators say and why it matters*. Cambridge University Press.

Grimmer, Justin and Gary King. 2011. “General purpose computer-assisted clustering and conceptualization.” *Proceedings of the National Academy of Sciences* 108(7):2643–2650.

Grimmer, Justin and Brandon M. Stewart. 2013. “Text as data: The promise and pitfalls of automatic content analysis methods for political texts.” *Political Analysis* 21(3): 267–297.

Hambleton, Ronald K. 1991. *Fundamentals of item response theory*. Vol. 2. Sage publications.

Hamilton, James D. 1989. “A new approach to the economic analysis of nonstationary time series and the business cycle.” *Econometrica: Journal of the Econometric Society* 57(2): 357–384.

Hoff, Peter D., Adrian E. Raftery, and Mark S. Handcock. 2002. “Latent space approaches to social network analysis.” *Journal of the American Statistical Association* 97(460):1090–1098.

Hopkins, Daniel J. and Gary King. 2010. “A method of automated nonparametric content analysis for social science.” *American Journal of Political Science* 54(1):229–247.

Hornik, Kurt. 2005. “A clue for cluster ensembles.” *Journal of Statistical Software* 14(12).

Hsu, Daniel and Sham M. Kakade. 2013. “Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*. ACM. pp. 11–20.

Hu, Yuening, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. “Interactive topic modeling.” *Machine Learning* 95(3):423–469.

Huang, Furong, U.N. Niranjan, M. Hakeem, and Animashree Anandkumar. Forthcoming. “Online tensor methods for learning latent variable models.” *Journal of Machine Learning Research*. arXiv:1309.0787.

Jordan, Michael I., Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1998. *An introduction to variational methods for graphical models*. Springer.

Kalai, Adam Tauman, Ankur Moitra, and Gregory Valiant. 2012. “Disentangling Gaussians.” *Communications of the ACM* 55(2):113–120.

Kass, Robert E. and Adrian E. Raftery. 1995. “Bayes factors.” *Journal of the American Statistical Association* 90(430):773–795.

King, Gary. 1989. *Unifying political methodology*. Cambridge University Press.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. “How censorship in China allows government criticism but silences collective expression.” *American Political Science Review* 107(02):326–343.

Kivinen, Jyrki and Manfred K. Warmuth. 1997. “Exponentiated gradient versus gradient descent for linear predictors.” *Information and Computation* 132(1):1–63.

Koltcov, Sergei, Olessia Koltsova, and Sergey Nikolenko. 2014. “Latent Dirichlet allocation: Stability and applications to studies of user-generated content.” In *Proceedings of the 2014 ACM Conference on Web Science*. ACM. pp. 161–165.

Krippendorff, Klaus. 2012. *Content analysis: An introduction to its methodology*. Sage.

Kuhn, Harold W. 1955. “The Hungarian method for the assignment problem.” *Naval Research Logistics Quarterly* 2(1–2):83–97.

Kumar, Abhishek, Vikas Sindhwani, and Prabhanjan Kambadur. 2012. “Fast conical hull algorithms for near-separable non-negative matrix factorization.” *arXiv preprint arXiv:1210.1190*.

Lancichinetti, Andrea, M. Irmak Sirer, Jane X. Wang, Daniel Acuna, Konrad Körding, and Luís A. Nunes Amaral. 2014. "A high-reproducibility and high-accuracy method for automated topic classification." *arXiv preprint arXiv:1402.0422*.

Lau, Jey Han, David Newman, and Timothy Baldwin. 2014. "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality." In *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Lauderdale, Benjamin E. and Tom S. Clark. 2014. "Scaling politically meaningful dimensions using texts and votes." *American Journal of Political Science* 58(3):754–771.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. "Life in the network: The coming age of computational social science." *Science* 323(5915):721.

Lim, Yew Jin and Yee Whye Teh. 2007. "Variational Bayesian approach to movie rating prediction." In *Proceedings of KDD Cup and Workshop*. Vol. 7. Citeseer. pp. 15–21.

Lloyd, Stuart. 1982. "Least squares quantization in PCM." *IEEE Transactions on Information Theory* 28(2):129–137.

Mahajan, Meena, Prajakta Nimbhorkar, and Kasturi Varadarajan. 2009. "The planar k-means problem is NP-hard." In *WALCOM: Algorithms and Computation*. Springer. pp. 274–285.

McLachlan, Geoffrey and David Peel. 2004. *Finite mixture models*. John Wiley & Sons.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." In *Advances in Neural Information Processing Systems*. pp. 3111–3119.

Mimno, David and David Blei. 2011. "Bayesian checking for topic models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. pp. 227–237.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing semantic coherence in topic models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. pp. 262–272.

Mullainathan, Sendhil. 2014. "What big data means for social science." Behavioral and Experimental Seminar.

Murphy, Kevin P. 2012. *Machine learning: A probabilistic perspective*. MIT press.

National Research. Council. 2013. *Frontiers in massive data analysis*. National Academies Press.

Nguyen, Thang, Yuening Hu, and Jordan, Boyd-Graber. 2014. "Anchors regularized: Adding robustness and extensibility to scalable topic-modeling algorithms." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics. pp. 359–369. http://www.aclweb.org/anthology/P14-1034.

Nielsen, Frank and Richard Nock. 2014. "Further heuristics for k-means: The merge-and-split heuristic and the (k,l)-means." *arXiv preprint arXiv:1406.6314*.

O'Connor, Brendan. 2014. "MiTextExplorer: Linked brushing and mutual information for exploratory text data analysis." In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Association for Computational Linguistics. pp. 1–13.

O'Connor, Brendan, Brandon M. Stewart, and Noah A. Smith. 2013. "Learning to extract international relations from political context." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics. pp. 1094–1104. http://www.aclweb.org/anthology/P13-1108.

Papadimitriou, Christos H. and Kenneth Steiglitz. 1998. *Combinatorial optimization: Algorithms and complexity*. Courier Dover Publications.

Park, Jong Hee. 2012. "A unified method for dynamic and cross-sectional heterogeneity: Introducing hidden Markov panel models." *American Journal of Political Science* 56(4):1040–1054.

Pearson, Karl. 1894. "Contributions to the mathematical theory of evolution." *Philosophical Transactions of the Royal Society of London. A*. pp. 71–110.

Poole, Keith T. and Howard Rosenthal. 1997. *Congress: A political-economic history of roll call voting*. Oxford University Press.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Rabiner, Lawrence and Biing-Hwang Juang. 1986. "An introduction to hidden Markov models." *ASSP Magazine, IEEE* 3(1):4–16.

Recht, Ben, Christopher Re, Joel Tropp, and Victor Bittorf. 2012. "Factoring nonnegative matrices with linear programs." In *Advances in Neural Information Processing Systems*. pp. 1214–1222.

Reich, Justin, Dustin Tingley, Jetson Leder-Luis, Margaret E. Roberts, and Brandon M. Stewart. 2015. Computer Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses. *Journal of Learning Analytics*. 2(1):156–184.

Rivers, Douglas. 2003. "Identification of multidimensional spatial voting models." Typescript. Stanford University.

Robert, Christian P. and George Casella. 2004. *Monte Carlo statistical methods*. Vol. 319. Springer: New York.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science*. 58(4):1064–1082.

Ruiz, Francisco J. R., Isabel Valera, Carlos Blanco, and Fernando Perez-Cruz. 2014. "Bayesian nonparametric comorbidity analysis of psychiatric disorders." *Journal of Machine Learning Research* 15:1215–1247. http://jmlr.org/papers/v15/ruiz14a.html.

Shneiderman, Ben. 1996. "The eyes have it: A task by data type taxonomy for information visualizations." In *Proceedings of the IEEE Symposium on Visual Languages*. pp. 336–343.

Song, Le, Eric P. Xing, and Ankur P. Parikh. 2011. "A spectral algorithm for latent tree graphical models." In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. pp. 1065–1072.

Sontag, David and Dan Roy. 2011. "Complexity of inference in latent Dirichlet allocation." In *Advances in Neural Information Processing Systems*. pp. 1008–1016.

Srivastava, Virendera K. and David E. A. Giles. 1987. *Seemingly unrelated regression equations models: Estimation and inference*. Vol. 80. CRC Press.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. "Hierarchical Dirichlet processes." *Journal of the American Statistical Association* 101(476).

Tenenbaum, Joshua B., Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. “How to grow a mind: Statistics, structure, and abstraction.” *Science* 331(6022):1279–1285.

Tibshirani, Robert. 1996. “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Vavasis, Stephen A. 2009. “On the complexity of nonnegative matrix factorization.” *SIAM Journal on Optimization* 20(3):1364–1377.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. “Evaluation methods for topic models.” In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. pp. 1105–1112.

Ward, Michael D., Brian D., Greenhill, and Kristin M. Bakke. 2010. “The perils of policy by p-value: Predicting civil conflicts.” *Journal of Peace Research* 47(4):363–375.

Wolpert, David H. and William G. Macready. 1997. “No free lunch theorems for optimization.” *IEEE Transactions on Evolutionary Computation* 1(1):67–82.

Yao, Limin, David Mimno, and Andrew McCallum. 2009. “Efficient methods for topic model inference on streaming document collections.” In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. pp. 937–946.

Zhou, Tianyi, Jeff Bilmes, and Carlos Guestrin. 2014. “Divide-and-conquer learning by anchoring a conical hull.” *arXiv preprint arXiv:1406.5752*.