2-2012

Topic Regression

David Mimno
University of Massachusetts Amherst, david.mimno@gmail.com

TOPIC REGRESSION

A Dissertation Presented

by

DAVID MIMNO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2012

Computer Science

TOPIC REGRESSION

A Dissertation Presented

by

DAVID MIMNO

Approved as to style and content by:

Andrew McCallum, Chair

David Blei, Member

David Jensen, Member

Michael Lavine, Member

David Smith, Member

Lori Clarke, Department Chair
Computer Science

For Jennifer

ACKNOWLEDGMENTS

Charles Sutton told me that graduate school is supposed to be a life changing experience. He was correct. More people have helped me through this process than I can list here, but several deserve particular mention.

I am here because of all of my teachers. Lisa Meeden at Swarthmore introduced me to Computer Science. At UMass, John Staudenmayer, Sridhar Mahadevan and Erik Learned-Miller taught me to reason about uncertainty.

Other graduate students and postdocs at UMass and elsewhere have been a vital part of my education, but the members of the IESL group have been particularly helpful. Aron Culotta, Charles Sutton, Xuerui Wang, Wei Li, Greg Druck, Pallika Kanani, Kedar Bellare, Sameer Singh, Michael Wick, Karl Schultz, Limin Yao, Anton Bakalov and Rachael Shorey created a wonderful environment to learn, practice, and teach. Hanna Wallach was a constant source of inspiration. Our collaboration is my model of an effective research team.

My work for the Perseus Project at Tufts University prepared me for life as a graduate student. Nobody reaches orbit without a big rocket behind them, and Gregory Crane has been a Saturn V.

I chose the members of my committee for their excellence, and I am pleased that they have demanded the same from me. I have known David Smith the longest, since I undertook the difficult task of filling the void he left at the Perseus Project. David Jensen's wisdom has been vital. Michael Lavine arrived, as if summoned, at precisely the time when he made the biggest impact on how I understand statistics and how I explain it to others. David Blei has been an inspiration even before I began graduate school. I depend on his skill and creativity every day.

Andrew McCallum took an enormous risk on an untested student with an odd background and not much math, and has backed me 100% throughout this process. I hope I have repaid his trust.

Owen is almost as old as my graduate career. Nathaniel was born around when I began working on this thesis. Watching them grow and learn has given me perspective on my own growth as a computer scientist. Graduate school is one of many things that I could not have done without Jennifer's constant support and love.

My mother, Nancy, taught me to value creativity, and it has become my most important asset. My father, Gerry, was the first person who told me what a PhD was, and the first to suggest that I might like to get one.

ABSTRACT

TOPIC REGRESSION

FEBRUARY 2012

DAVID MIMNO

B.A., SWARTHMORE COLLEGE

M.Sc., UNIVERSITY OF MASSACHUSETTS, AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Andrew McCallum

Text documents are generally accompanied by non-textual information, such as authors, dates, publication sources, and, increasingly, automatically recognized named entities. Work in text analysis has often involved predicting these non-text values based on text data for tasks such as document classification and author identification. This thesis considers the opposite problem: predicting the textual content of documents based on non-text data. In this work I study several regression-based methods for estimating the influence of specific metadata elements in determining the content of text documents. Such topic regression methods allow users of document collections to test hypotheses about the underlying environments that produced those documents.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

MODELS FOR TEXT ANALYSIS

1.1 Introduction

This thesis is about measuring the association between information *about* documents and the words that appear in those documents. Text documents are typically accompanied by small amounts of structured data — authors, titles, dates, subjects — usually referred to as *metadata*. It is commonly assumed that this succinct block of information is predictive of the contents of the much larger document that it describes. The question I consider is whether there are statistical models that can, first, measure this predictive ability, and second, provide clear, interpretable explanations for these predictions.

Researchers in many fields are increasingly gaining access to large document collections with complicated metadata structures. In this thesis I use collections comprising two decades of New York Times articles, 15 years of US Supreme Court documents, and all the grants funded by the National Institutes of Health for six years, among several others. People are interested in performing experiments using these corpora to answer specific questions. For example, does the media's portrayal of government agencies shift based on the party in power? To what extent does the subject of a supreme court case affect the choice of author for the opinion? Which NIH institutes fund cancer research? My goal is to provide these researchers with tools and methodologies that can accurately measure the association between metadata fields and the words that appear in documents.

In this thesis I define several models for measuring this association. For each model, I also specify an inference method that finds a set of parameter values given a set of training documents. I compare trained models with respect to their ability to predict the contents of new documents given specific metadata values. This evaluation measures the ability of each model to learn patterns of association that underly the observed words in a document collection. The central metric for this evaluation is the marginal probability of held-out

documents under each trained model. Using related but previously unseen documents to measure model performance reduces the risk of overfitting to a specific training set. Due to the vast number of possible documents drawn from a vocabulary of thousands of words, calculating the probability of a set of real, but previously unseen documents under each model is the best known method for comparing how effectively each model has learned the true associations of words and metadata.

The simplest approach to measuring the association between metadata elements and words is to estimate the relationship between each individual metadata element/word pair. This simple approach is problematic for two reasons. The first problem relates to the frequency patterns of natural language. The most frequent words have relatively little specific meaning and are therefore of little interest for researchers. In contrast, more interesting words are often very infrequent and therefore it is difficult to obtain good parameter estimates due to small sample sizes. The second problem lies in interpretation. People want to understand the association between metadata and words, but the large number of potential combinations is overwhelming.

A solution to these problems is to identify a small set of latent semantic dimensions. This projection from a high-dimensional word space into a simpler conceptual space is the goal of statistical topic modeling. Topic models add an intermediate "topic" layer between metadata and words. The latent semantic space summarizes themes and concepts, grouping words that are related and distinguishing between meanings of ambiguous words using context.

The fundamental task of this work is to design a system that quantitatively measures the ability of specific metadata elements to predict the text of documents. With this tool, users can compare the predictive power of different metadata elements or compare the predictive power of the same metadata on different sub-corpora. For example, in a legislative setting, researchers might want to know whether the party in power is a better predictor of the contents of bills than economic indicators (that is, the difference between predictive effects). Alternatively, researchers may wish to measure whether the influence of the party in power changed from one era to another (that is, the effect of the same metadata in different sub-corpora). As an example, Chapter 7 contains a case study of Supreme Court

documents that compares two predictors (authorship and voting patterns) with respect to their predictive ability in different majority sizes.

If we are to trust measurements of predictive ability made by a model, we should choose the strongest predictive model that is computationally tractable for the size of a typical collection. Such a model is most likely to learn the latent association between metadata and text. In Chapter 6, I therefore measure both the predictive ability of models and their associated inference methods, as measured by the probability of held-out test corpora, and also the computational complexity of these inference methods.

1.1.1 Representations

Documents are often described in terms of text and metadata. Text comprises a sequence of words in a natural language. Metadata is a more carefully structured set of named fields that describe the origin and context of the text, such as author, publication date, and publication venue. The value of a particular metadata field is typically drawn from a relatively small set of possible values, although there may be great variation.

Another difference between text and metadata is that metadata is often highly determined by rules governing a given corpus. For example, a document usually has exactly one unique publication date, and there is usually only one way to correctly write that date. In contrast, natural language is more variable. The same idea or fact can be expressed in many different ways, often with no overlap in word use.

For the purposes of this work I represent a document as a set of variables $\{\boldsymbol{x}, \boldsymbol{w}\}$, where $\boldsymbol{x}$ is a vector of features derived from the document metadata and $\boldsymbol{w}$ is a sequence of words from a vocabulary $\mathcal{V}$. A useful method for generating features is to define indicator functions representing whether a particular value appears in a particular field, for example there might be a feature x_{1998} that has the value 1 if a document was published in 1998 and 0 otherwise.

These non-textual data sources provide clues about the underlying environment behind the collection, but a simple analysis of the effect of such metadata may lead to false conclusions due to interactions between different elements. It is easy, for example, to count the occurrence of a given word over time, but if a collection is missing several years of a specific

journal, such analysis may indicate dramatic discontinuities in word prevalence. Factoring in publication venue information could, on the other hand, explain much of this variance.

Words in documents can be thought of as the observable manifestation of "topics": the events, entities, and ideas that the authors intended to describe. Collections generally have a topical focus, implying that the number of these latent topics is small relative to the number of documents. Computational methods for estimating these semantic components of text corpora have been explored for the last 20 years [12]. During this time there has been much work on estimating the relationship between text and specific classes of metadata, such as authors and dates. There has, however, been relatively little work on general methods for estimating the association of arbitrary combinations of non-textual information with the concentration of various topics in the collection. Such analysis can provide overviews identifying the major trends in a collection over time and word-based characterizations of non-text metadata elements.

Previous work in text analysis that has considered text and metadata has generally focused on predicting metadata elements given text. Examples include author identification and spam filtering applications. These tasks seek to estimate the probability of a particular class of metadata elements given the words in a document. The work in this thesis explores the opposite question: predicting the text content of a document given all of its metadata.

Predicting text given metadata is not a new task. Indeed, the basis of naïve Bayes classification is estimating the distribution of words given explanatory variables such as author or spam status. What has so far been difficult is handling more complicated metadata structures. When each document has exactly one metadata value indicating one of a small set of mutually exclusive classes (commonly called a "label"), it is relatively easy to divide the collection into partitions, one for each label, and estimate word distributions. If each document has multiple labels, for example a year, an author, and a publication source, estimation becomes difficult: we can no longer divide the data into partitions for each label as the number of distinct combinations of metadata elements grows exponentially, and we can not estimate the distribution over words for each element because it is not clear how to distinguish the influence of cooccurring elements.

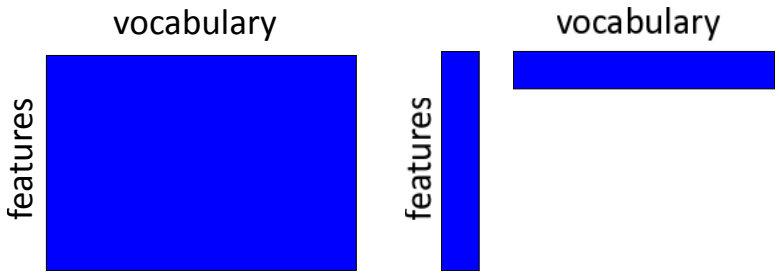

Figure 1.1. Matrix factorization. Representing a matrix of weights for all feature-word pairs (left) leads to large numbers of parameters. A factored representation (right) with a matrix of feature-topic parameters and a matrix of topic-word parameters can be much simpler while representing much of the observed variability in text documents.

Regression models are a well understood and widely used method for evaluating the influence of numerous explanatory variables on some output variable. If the output is real-valued, linear regression is appropriate. Poisson regression is commonly used when the output variable is a non-negative integer. Binary outputs are often modeled using logistic or probit regression. In this work I consider regression models for the case when the output variable consists of text documents.

I consider three Bayesian hierarchical models for disentangling the effects of complicated metadata structures. These include an existing generative model and two novel regression-based methods. In the later chapters, I evaluate the conditions that determine the predictive performance of these models, as measured by the probability of held-out documents, the ability to predict metadata given documents, and the ability to find similar metadata elements. All three methods use latent "topic" representations to simplify the output space. Topic models reduce redundancy and ambiguity, while remaining interpretable to humans. Unfortunately such hidden topic representations require iterative, approximate inference techniques. For comparison, I include results from a non-topic-based model that uses metadata and a topic model that does not use metadata.

Finally, I demonstrate the scalability and practical usefulness of these models in the context of several large, real-world text collections that are characterized by complicated metadata structures, including U.S. Supreme Court decisions, 20 years of the New York Times, the NIH grants database and associated publications, and large collections of academic

papers. I show that the models are effective at identifying large-scale trends, highlighting anomalies and comparing various explanations for observed changes.

1.2 Properties of text collections

Text data is characterized by sparsity and extremely high dimensionality. Technical literature in particular contains very large numbers of distinct word types, on the order of hundreds of thousands, almost all of which occur very rarely. Directly modeling correlations between all possible word types is therefore both computationally infeasible and unlikely to be well estimated as almost all possible word combination events will be extremely rare. At the same time, words are often ambiguous by themselves: "strain" has one meaning for an immunologist and another for a mechanical engineer. Contextual clues are almost always sufficient for human readers to distinguish between meanings, but naïve algorithms may become confused. It is therefore desirable to operate at a more abstract level of meaning than simple vocabulary.

1.3 Topic modeling

Organizing scholarship into categories is an important but challenging problem. Manually curated subject headings have been in development for centuries. Current examples include the Library of Congress subject headings (LCSH) and National Institutes of Health Medical Subject Headings (MeSH). Constructing and maintaining such representations is expensive, fraught with political and philosophical conflicts, and prone to obsolescence, especially in fast moving fields [17].

In the past several decades researchers have developed methods for deriving "topic" clusters from collections of documents. Most of these methods rely on word count vector representations of documents. Latent semantic analysis (LSA) or latent semantic indexing (LSI) [12] use eigenvalue decompositions to find low dimensional subspaces based on word count vectors transformed into real space using term frequency/inverse document frequency (TF-IDF) transformations. LSA components are generally considered uninterpretable [22].

In the past 10 years, work in topic modeling has turned to probabilistic models based on multinomial distributions directly over word counts rather than transformed TF-IDF

vectors. Hofmann introduces probabilistic latent semantic analysis (pLSI) [21], which can be trained using expectation maximization.

More recently, Blei, Ng and Jordan [7] present latent Dirichlet allocation, which recasts pLSI as a Bayesian mixture model with a Dirichlet prior over mixing proportions and a Dirichlet prior over multinomial mixture components. The Bayesian prior distributions allow topics to be inferred using variational EM [7] and Markov chain Monte Carlo methods such as Gibbs sampling [18]. In addition, LDA is a fully generative model that can be used to make inferences about previously unseen documents, a feature that is necessary for some evaluation purposes.

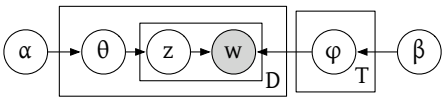

Figure 1.2. Graphical model representation of a simple topic model (LDA).

A Bayesian finite mixture model allows complicated distributions to be composed out of mixtures of simpler distributions. For example, a multi-modal distribution can be modeled as a mixture of simple unimodal distributions. Each data point is assumed to be drawn from a single mixture component, but the variable indicating which component is not observed. The distribution of an observation marginalized over its hidden component is a weighted sum of these component distributions. Inference in such models involves estimating the mixing weights θ and the component distributions.

LDA is a combination of multiple Bayesian mixture models, one per document, that share component distributions but each have a distinct mixing proportion. The component distribution (a "topic" t) is a discrete distribution over words from a vocabulary of size V with probability vector $\boldsymbol{\phi}_t$. Each of these distributions is drawn from a V-dimensional Dirichlet prior. For the purposes of this work I will assume that the prior over these topic-word distributions is symmetric with parameter β: $p(\boldsymbol{\phi}_t) = \frac{\Gamma(\beta)^V}{\Gamma(V\beta)} \prod_w \phi_{tw}^{\beta-1}$. Asymmetric priors, with $\beta_w \neq \beta_{w'}$, are also possible but have been shown by Wallach, Mimno, and McCallum to have little effect [35].

The mixing proportions over components (topics) for each document d are given by a discrete distribution $\boldsymbol{\theta}_d$. The parameters of this distribution are themselves distributed according to a Dirichlet prior with parameter $\boldsymbol{\alpha}$. In contrast to the topic-word distributions, we found that using an asymmetric Dirichlet prior over document-topic distributions had a significant effect on the quality and robustness of the model.

Given these two sets of discrete distributions, the observed words are sampled by first selecting topic indicator variables from $\boldsymbol{\theta}_d$ and then selecting one word each from the topic distributions indexed by the topic indicators. The complete generative model is as follows:

1. For each topic t,
   (a) Draw multinomial $\boldsymbol{\phi}_t \sim \mathcal{D}(\beta)$
2. For each document d,
   (a) Draw multinomial $\boldsymbol{\theta}_d \sim \mathcal{D}(\alpha_1, ..., \alpha_T)$
   (b) For each word i,
      i. Draw $z_i \sim \mathcal{M}(\boldsymbol{\theta}_d)$.
      ii. Draw $w_i \sim \mathcal{M}(\boldsymbol{\phi}_{z_i})$.

This model is specified in terms of categorical observations drawn from discrete probability distributions that are themselves drawn from Dirichlet distributions. The use of a discrete probability distribution adds an additional vector of real-valued parameters that are constrained to be non-negative and sum to one. It is often simpler to work with a simpler model that integrates over this vector.

As a simple example, consider a model such that the random variables $x_1, ..., x_N$ take values in $\{1, ..., K\}$, distributed according to Mult $(\boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is drawn from a symmetric Dirichlet with parameter α. Let $n_k = \sum_{i=1}^{N} I_{x_i=k}$, that is, the number of x_is that have value k. Integrating over a K-dimensional Dirichlet-distributed probability vector $\boldsymbol{\theta}$ with

a sequence of discrete observations results in Dirichlet compound multinomial distribution, also known as a Polya distribution.

$$\int \prod_i \text{Mult}\left(x_i;\boldsymbol{\theta}\right) \times \text{Dir}\left(\boldsymbol{\theta};\alpha\right)\, d\boldsymbol{\theta} = \int \prod_i \theta_{x_i} \times \frac{\Gamma(K\alpha)}{\prod_k \Gamma(\alpha)} \prod_k \theta_k^{\alpha-1}\, d\boldsymbol{\theta} \quad (1.1)$$

$$= \int \frac{\Gamma(K\alpha)}{\prod_k \Gamma(\alpha)} \prod_k \theta_k^{\alpha+n_k-1}\, d\boldsymbol{\theta} \quad (1.2)$$

$$= \frac{\Gamma(K\alpha)}{\Gamma(K\alpha+N)} \prod_k \frac{\Gamma(\alpha+n_k)}{\Gamma(\alpha)} \quad (1.3)$$

Integrating over all multinomials in the LDA model results in the following expression for the probability of the discrete random variables representing observed words and hidden topic indicators $\{\boldsymbol{w}\}_{1:D}, \{\boldsymbol{z}\}_{1:D}$ given Dirichlet parameters.

$$\prod_d \left[\frac{\Gamma(\sum_t \alpha_t)}{\Gamma(\sum_t \alpha_t + n_d)} \prod_t \frac{\Gamma(\alpha_t + n_{t|d})}{\Gamma(\alpha_t)}\right] \prod_t \left[\frac{\Gamma(V\beta)}{\Gamma(V\beta + n_t)} \prod_w \frac{\Gamma(\beta + n_{t|d})}{\Gamma(\beta)}\right] \quad (1.4)$$

The choice of the number of topic components T is an important question, but is not directly addressed in this thesis. I will assume that T is a fixed, known constant.

1.4 Models considered

In order to extend topic models to take into account side information, I consider several models, two of which are novel. For comparison, I include four baseline approaches:

1. Unigram language model. The simplest language model is a single probability distribution over a fixed vocabulary $\mathcal{V}$. This model does not use hidden topics and cannot incorporate metadata. Such a distribution can be estimated from data by adding a small smoothing factor to avoid 0 probabilities for words that do not occur in training data. Let N_v be the number of word tokens of type v that occur in the training documents. Under the unigram model, $P(v) = \frac{N_v+\beta}{\sum_v N_v+\beta|\mathcal{V}|}$. Smaller values of the smoothing parameter β generally lead to greater probability of held-out documents. I use $\beta = 0.01$.

2. Non-topic metadata baseline. The simplest baseline model is to count the proportion of words present in documents with each distinct metadata element. This model

does not identify any latent topic space. Letting $N_{v|m}$ be the number of word tokens of type v occur in documents with metadata element m, $P(v|m) = \frac{N_{v|m}+\beta}{\sum_v N_{v|m}+\beta|\mathcal{V}|}$. For a given set of document-level metadata elements $\mathcal{M}$, I generate a probability distribution over words by taking a uniformly-weighted linear combination of each element-specific language model: $P(v|\mathcal{M}) = |\mathcal{M}|^{-1} \sum_{m \in \mathcal{M}} P(v|m)$.

3. Post-hoc topic baseline. Rather than jointly estimating topic assignments and metadata effects, it is possible to train a topic model with no metadata information and then, holding topic assignments fixed, estimate the effect of metadata elements post hoc. Specifically, I use L-BFGS optimization to train a DMR model (described below) from a fixed topic model. This model cannot support shifting or realigning the topic space given information from non-textual information.

I compare these baselines to three models that identify latent topic spaces conditioned on metadata, which are described more fully in Chapter 3:

1. "Author"-topic model. This model, a generalization of work by Rosen-Zvi, Griffiths, Steyvers and Smyth [33], asserts that each metadata element, for example each author in the original formulation, has a multinomial distribution over topics. Each word token's topic is generated by one of these distributions, but the linkage between tokens and metadata elements is not observed. Inference in this model consists of sampling a topic indicator variable and a metadata assignment indicator variable for each word token. In the case of author metadata, this corresponds to assigning each word in a document to one of the authors listed for that document. There is no particular reason to limit consideration to author metadata, however. In this thesis, I consider extending this model to additional, increasingly fine-grained document metadata elements.

2. Dirichlet-multinomial Regression model. In this model, which I introduced in previous work [30], the inner product between document metadata features and estimated parameters is used to create a document-specific Dirichlet prior over topic-mixing proportions. This prior is then in turn used to generate topic indicator variables.

3. Logistic Normal regression topic model. This model, previously described by Mimno, Wallach and McCallum [31], uses a logistic normal distribution rather than a Dirichlet distribution, as in DMR, to map between sequences of discrete topic indicators to an unconstrained real-valued parameter space. The model then uses arbitrarily structured GMRFs to estimate the probability of document-topic mixture distributions.

1.5 Previous work in topic modeling with metadata

The simplest method of incorporating metadata in generative topic models is to generate both the words and the metadata simultaneously given hidden topic variables. In this type of model, each topic has a distribution over words as in the standard model, as well as a distribution over metadata values. Examples of such "downstream" models include the authorship model of Erosheva, Fienberg and Lafferty [15], the Topics over Time (TOT) model of Wang and McCallum [37], the Group-Topic model of Wang, Mohanty and McCallum [38], the CorrLDA model of Blei and Jordan [5] and the named entity models of Newman, Chemudugunta and Smyth [32].

One of the most flexible members of this family is the supervised latent Dirichlet allocation (sLDA) model of Blei and McAuliffe [6]. sLDA generates metadata such as reviewer ratings by learning the parameters of a generalized linear model (GLM) with an appropriate link function and exponential family dispersion function, which are specified by the modeler, for each type of metadata. It can be shown that the TOT model is an example of sLDA [30].

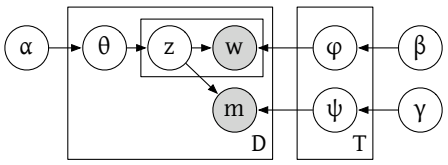

Figure 1.3. Graphical model representation of a "downstream" topic model. Metadata m is generated conditioned on the topic assignment variables $\boldsymbol{z}$ of the document and each topic has some parametric distribution over metadata values.

Another approach involves conditioning on metadata elements such as authors by representing document-topic distributions as mixtures of element-specific distributions. One example of this type of model is the author-topic model of Rosen-Zvi, Griffiths, Steyvers and Smyth [33]. In this model, words are generated by first selecting an author uniformly from an observed author list and then selecting a topic from a distribution over topics that is specific to that author. Given a topic, words are selected as before. This model assumes that each word is generated by one and only one author. Similar models, in which a hidden variable selects one of several multinomials over topics, are presented by Mimno and McCallum [29] for modeling expertise by multiple topical mixtures associated with each individual author, by McCallum, Corrada-Emmanuel, and Wang [26] for authors and recipients of email, and by Dietz, Bickel and Scheffer [14] for inferring the influence of individual references on citing papers. These "upstream" models essentially learn an assignment of the words in each document to one of a set of entities.

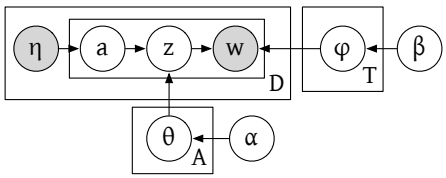

Figure 1.4. An example of an "upstream" topic model (Author-Topic). The observed authors determine a uniform distribution η over authors. Each word is generated by selecting an author, a, then selecting a topic from that author's topic distribution θ_a, and finally selecting a word from that topic's word distribution.

1.6 Contributions of this thesis

This thesis presents models for estimating the relationships between the text and the metadata elements of documents. In particular, in Chapter 3 I present two novel regression-based topic models based on Dirichlet-multinomial regression and on logistic normal regression. I highlight the differences and similarities between these models and the non-topic baselines, with a particular focus on the role of hyperparameters. In Chapter 4, I describe the theoretical difference between these models. In particular, I demonstrate that logistic

normal distributions are more sensitive to variance parameters than Dirichlet distributions in the context of sparse, high-dimensional observations and explain this difference in terms of the symmetry of their error distributions. In Chapter 5 I describe how to add structured, graph-based relationships between regression parameters using Gaussian Markov random field priors. In Chapter 6 I measure the empirical difference between these models and several existing baseline models on real and synthetic data using several measures. These differences are consistent across random initializations and are not significantly affected by rare or extremely infrequent words. Finally, in Chapter 7 I present several applications, in which metadata-enriched models of text produce results predicted by outside information for specific data sets.

I find that the most effective tool for measuring the association of text and metadata is the Dirichlet-multinomial regression (DMR) topic model, because of the following properties.

- DMR shows consistently higher held-out probability than other models. The latent topic space provides a substantial boost in the model's ability to learn underlying regularities in document collections. There is no evidence that this advantage is due to artifacts of hyperparameter settings or to the value of rare "outlier" words.
- DMR is computationally tractable for large collections and scales well to complicated metadata environments.
- As it is based on a regression framework, DMR can accept a wide array of possible input variables, without requiring statistical or computer programming expertise by users.
- DMR provides interpretable results based on associations between metadata and semantically coherent word clusters (topics) rather than individual words.

The result of this thesis is to advocate the use of DMR topic models in any context where large numbers of documents are accompanied by structured metadata, in preference to other methods of measuring the predictive ability of metadata. This tool will allow researchers to determine the relative predictive power of different metadata elements in different corpora,

and to understand the specific details of those associations. More work must be done, however, to make this model more computationally tractable and more accessible to users without computational or statistical backgrounds. In addition, future work will focus on providing users with better indications of the quality and confidence of inferences, both in the semantic coherence of topic distributions and in the significance of the learned regression coefficients linking metadata to topics.

Finally, there are many models in natural language processing that use generative processes based on Dirichlet-multinomial distributions, such as syntactic parsers. One direction for future work is to incorporate regression models into such models in order to give researchers the ability to measure the association not only between metadata and words but also words in their specific syntactic and semantic roles.

CHAPTER 2

APPROXIMATE INFERENCE

All of the topic-based models presented in this thesis involve potentially millions of hidden variables, whose values must be estimated. Due to statistical dependencies, these parameters cannot be learned independently, so exact MAP inference is intractable for all but the most trivial collections. Rather, we rely on approximate inference. Standard approaches to inference in machine learning for data of this scale include Markov chain Monte Carlo and approximate MAP methods, such as variational inference (where parameters are learned within a more tractable family of models) and stochastic EM (where the trainer alternates between sampling over some variables and maximizing the remaining variables).

In this chapter I discuss methods for approximate inference. I also specify how the choice of method relates to the scientific results of this thesis.

2.1 Distinguishing the effect of model choice and inference method

The contribution of this work has two parts: statistical models and inference methods. A statistical model defines relationships between observed data and unobserved parameters. Such models are usually defined mathematically in terms of a likelihood function. In order to be useful, a model must be accompanied by an inference method, which takes as input an observed data set, and returns as output some setting of the values of the parameters of the model.

Different models have different abilities to represent the underlying regularities of a data set. A univariate Gaussian, for example, cannot represent a bimodal distribution in the way that a mixture of two univariate Gaussians can. This representational power, however, often trades off with complexity of inference. The univariate Gaussian can be estimated in closed form, while a mixture of Gaussians may require iterative inference through EM. The

measurable performance of a model is thus a combination of the representational power of the model and the effectiveness of the inference method.

In complicated models, it is difficult to distinguish the effect on performance of the appropriateness of the model and the quality of inference. In this work I am making comparisons of model and inference pairs. As a result, it is possible that, for example, improved inference methods for the logistic normal regression model might result in improved performance that would match or exceed that of the Dirichlet-multinomial regression model. The findings in this thesis therefore reflect comparisons of models and the chosen inference methods. The differences measured may not reflect the potential value of the models in themselves, if exact inference were available.

2.2 Markov chain Monte Carlo

A commonly used method of inference in statistical topic modeling is Gibbs sampling, which involves iterating through each hidden variable and sampling its value from the conditional distribution of that variable given the current value of all other hidden variables. It can be shown that the stationary distribution of the Markov chain induced by this process is the posterior distribution over the hidden variables given the observed variables and the hyperparameters.

The conditional distribution is given by the following equation.

$$P(z_{di} = t \mid \boldsymbol{w}, \boldsymbol{z}_{\setminus z_{di}}, \alpha, \beta) \quad \propto \quad \frac{n_{t|d} + \alpha_t}{n_d + \sum_t \alpha_t} \frac{n_{w_i|t} + \beta}{n_t + V\beta} \tag{2.1}$$

The performance of a Gibbs sampler is dominated by evaluations of this expression. In previous work I have considered simple methods to reduce this computational burden, particularly for repeated sequential evaluations, in which the summary statistics $n_{w|t}$ and $n_{t|d}$ change slowly [39]. Note that this method is *exact*. The sampling distribution is precisely the same as it would be under Eq. 2.1. The increase in efficiency is not due to an approximation, but rather by caching much of the computation needed to evaluate the normalizing constant for Eq. 2.1 and computing individual terms only when needed during sampling.

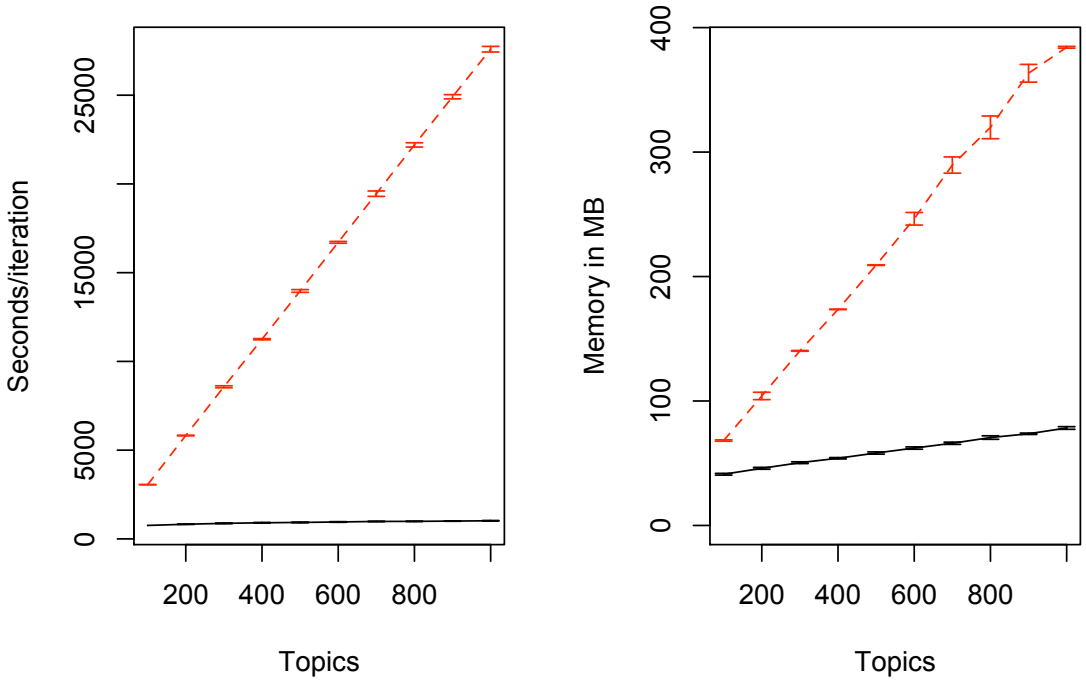

Figure 2.1. Time and memory improvements in Gibbs sampling due to my SparseLDA algorithm and data structure. This plot compares time and space efficiency between standard Gibbs sampling (dashed red lines) and the SparseLDA algorithm and data structure presented in [39] (solid black lines). The corpus is a large collection of New York Times articles. Error bars show the standard deviation over five runs.

One reason that Gibbs samplers may not reach their stationary distribution is that they are sensitive to initial conditions. In practice, starting from many random initializations and sampling for 1000–5000 iterations results in qualitatively very similar models. The same topic-word clusters appear repeatedly, with small variations. Quantitatively, the model log-likelihood of samples also falls within a narrow range, given the number of parameters of the model. Figure 2.2 shows 500 random initializations of an LDA model on the Rexa corpus (see Chapter 6 for details) with $T = 50$. I calculated the log probability of the training data under the current setting of the hidden variables every 10 iterations for a total of 5000 iterations, dividing this value by the number of tokens to get an approximate nats/word value. The plot shows that model log likelihood moves within 500 iterations to a range of probabilities that is relatively narrow compared to the change from the initial log probability.

2.3 Posterior densities vs. point estimates

Standard practice in MCMC involves using multiple samples from a Markov chain to approximate an expectation of some function of interest. In this work I typically use single samples from multiple Markov chains and evaluate the resulting predictive distribution in several ways. As above, I use multiple random initializations and multiple cross-validation folds to test for sensitivity to initial settings of the hidden variables.

This methodology is closer to stochastic MAP than to standard MCMC practice, as it seeks settings of high probability rather than converged samples from the posterior. If we are in fact interested in stochastic MAP to find local optima in the posterior distribution, it makes sense to consider methods that specifically maximize likelihood. A standard method for stochastic optimization that is closely related to Gibbs sampling is iterated conditional modes (ICM) [4]. Rather than sampling from the conditional distribution of over values for each variable given the current values of all other variables, ICM sets the variable to the value with the maximum conditional probability given the current values of all other variables.

As shown previously, Gibbs sampling quickly moves to configurations of hidden variables that have much greater likelihood than random initializations. Running ICM directly from

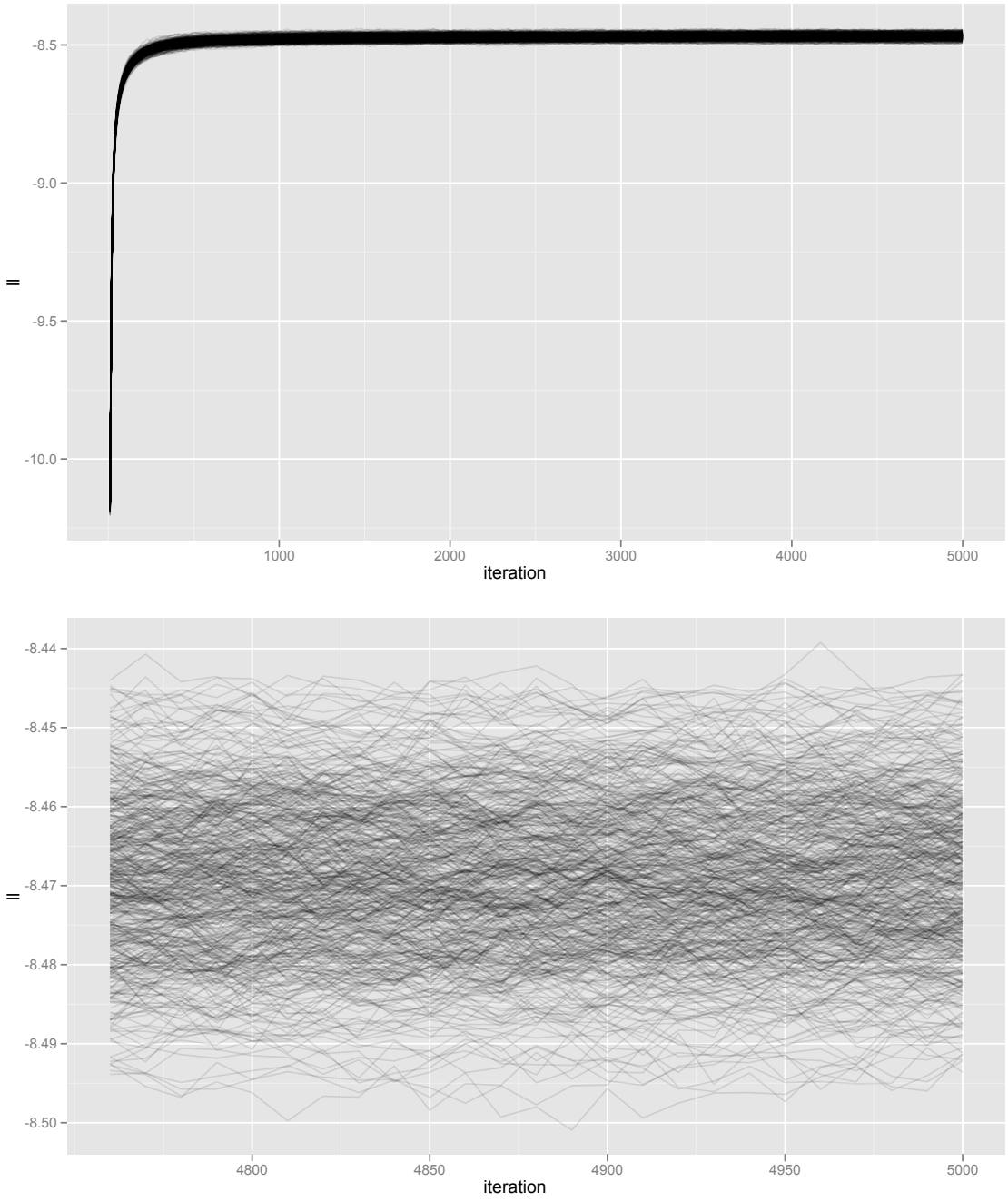

Figure 2.2. Gibbs chains from 500 random initializations. Each chain is run for 5000 iterations, calculating per-token model log likelihood after every 10 iterations. The corpus is Rexa, with $T = 50$. The top figure shows all 5000 iterations, the bottom figure shows just the last 250.

random initializations in LDA produces likelihoods that are significantly worse than Gibbs samples, but ICM initialized from Gibbs samples produces significant increases in model likelihood. In terms of held-out probability, however, there is relatively little improvement in performance. I ran Gibbs sampling with 100 random initializations for 5000 iterations on the Rexa corpus with $T = 50$. I saved samples every 1000 iterations, and, after completion of Gibbs sampling, ran ICM starting from those samples. ICM converged, making no further updates, within 15–20 iterations in all cases.

Figure 2.3 shows results for the 500 Gibbs models and the corresponding 500 ICM models. The average held-out probability is greater for ICM models by 0.0045, a small difference relative to the differences observed between models, and there is substantial overlap between these distributions.

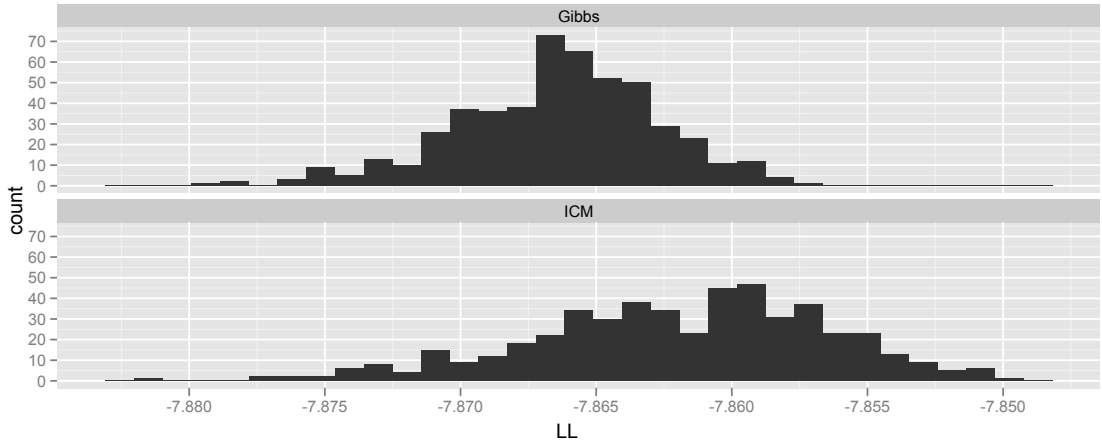

Figure 2.3. Held-out log probability values for the Rexa corpus with $T = 50$. The top figure shows results from 500 models trained by Gibbs sampling (100 chains saving states every 1000 iterations). The bottom shows 500 models trained with iterated conditional modes, initialized from *the same* 500 models. Held-out results are on average better for maximized models, but there is substantial overlap.

CHAPTER 3

MODELS FOR TOPICS CONDITIONED ON METADATA

In this chapter I introduce three extensions to the simple topic model that adapt the topic distribution for each document individually, conditioned on observed metadata for that document. The first model (Author-topic) adds additional per-token hidden variables, while the second two (Dirichlet-multinomial regression and logistic normal regression) add no additional per-token variables but rather use log-linear models to modify the document-level distribution over topics.

3.1 "Author"-topic models

3.1.1 Definition

A simple yet surprisingly powerful method for distinguishing the influences of observed metadata elements is the Author-Topic model of Rosen-Zvi, Griffiths, Steyvers and Smyth [33]. Under this model, each author is modeled as a multinomial distribution over topics. Topics are represented as distributions over the vocabulary, as in a simple topic model. For each document, each word is generated by selecting an author uniformly at random from the set of authors associated with the document and then sampling a topic indicator from that author's distribution over topics and finally a word from that topic distribution. The generative process for this model is as follows.

1. For each author a,
   (a) Draw multinomial $\boldsymbol{\theta}_a \sim \mathcal{D}(\alpha)$
2. For each topic t,
   (a) Draw multinomial $\boldsymbol{\phi}_t \sim \mathcal{D}(\beta)$
3. For each document d,

(a) Let $\mathcal{A}_d$ be the set of authors for document d.

(b) For each word i,

i. Draw $a_i \sim \mathcal{U}(\mathcal{A}_d)$.

ii. Draw $z_i \sim \mathcal{M}(\boldsymbol{\theta}_{a_i})$.

iii. Draw $w_i \sim \mathcal{M}(\boldsymbol{\phi}_{z_i})$.

As with simpler Dirichlet-multinomial topic models, the distributions over topics for each author can be integrated over analytically, leaving a distribution that only depends on hyperparameters α and the hidden author and topic assignment variables $\boldsymbol{a}$ and $\boldsymbol{z}$. The conditional distribution of the pair of variables $\{a_i, z_i\}$ for word i in document d given all other topic and author assignments is

$$P(a_{di} = a, z_{di} = t \mid \boldsymbol{w}, \boldsymbol{a}_{\setminus a_{di}}, \boldsymbol{z}_{\setminus z_{di}}, \alpha, \beta) \quad \propto \quad \frac{1}{|\mathcal{A}_d|} \frac{n_{t|a} + \alpha}{n_a + T\alpha} \frac{n_{w_i|t} + \beta}{n_t + V\beta} \tag{3.1}$$

Characterizations of each author in terms of topics can be derived by averaging over several Gibbs sampling states.

Although the model is defined in terms of authors, in fact any type of categorical metadata can serve the same purpose. Note, however, that real-valued or count-valued covariates are not defined in this model.

3.1.2 Multiple author "personas"

One major assumption of the Author-Topic model is that authors (or other metadata elements) can be well-modeled with a single distribution over topics. Intuitively, many observable characteristics of a document may in fact be themselves combinations of distinct topic mixtures. For example, an author may move from one area to another over time or work on multiple different distinct areas simultaneously.

In previous work [29], which I summarize here, I explored an extension of the Author-Topic model for just this situation, for the purpose of ranking potential reviewers for papers. In the Author-Persona-Topics (APT) model, each observable metadata element (for example an author) has one or more topic distributions. For the purposes of this work I replicated

documents with multiple authors once for each author, so that each document had a single author. For each author I determined a number of personas using a simple heuristic, dividing the author's total number of published papers by 20. This heuristic worked better empirically than non-parametric Dirichlet process methods. I initialized the sampler by randomly assigning each document to one of the author's personas and randomly assigning words to topics. I then alternated, for each document in turn, between sampling topics given the current persona and sampling a new persona for the document given the document's current topic assignments.

This model provides a probability distribution over vectors of words for each author. I ranked potential reviewers for each of a set of held-out papers based on the probability of the document under each author's distribution. Ground truth was provided by a panel of expert annotators, who marked potential reviewers on a four point scale. As a baseline I considered ranking based on the overall distribution of words formed by the author's papers and ranking based on finding the author of the single most similar document in the training corpus. This single-nearest-neighbor ranking had the highest precision at rank 1, but was also extremely good at finding the actual authors of papers. The APT model performed best in precision at rank 10 (that is, the proportion of the top 10 ranked authors that were marked highly relevant to the query document). As reviewer matching is a highly constrained optimization problem, having additional possibilities is valuable.

3.2 Dirichlet-multinomial regression

This model, which I have previously published [30], modifies the generative process of LDA by generating a document-specific Dirichlet prior over topic mixture distributions, derived from a weighted combination of document metadata features. DMR is the first topic model to allow arbitrary combinations of categorical and real-valued metadata features to influence topic mixing proportions.

Guimaraes and Lindrooth [20] use Dirichlet-multinomial regression in economics applications, but do not use a mixture model or any hidden variables. They observe that Dirichlet-multinomial regression falls within the family of overdispersed generalized linear models (OGLMs), and is equivalent to logistic regression in which the output distribution

exhibits extra-multinomial variance. This property is useful because DMR produces unnormalized Dirichlet parameters rather than normalized multinomial parameters. These Dirichlet parameters can then be used as a prior for a Bayesian mixture model.

3.2.1 Definition

For each document d, let $\boldsymbol{x}_d$ be a vector containing features that encode metadata values. For example, if the observed features are indicators for the presence of authors, then $\boldsymbol{x}_d$ would include a 1 in the positions for each author listed on document d, and a 0 otherwise. In addition, to account for the mean value of each topic, we include an intercept term or "default feature" that is always equal to 1.

For each topic t, we also have a vector $\boldsymbol{\lambda}_t$, with length the number of features. Given a feature matrix X, the generative process is:

1. For each topic t,
   (a) Draw $\boldsymbol{\lambda}_t \sim \mathcal{N}(0, \sigma^2 I)$
   (b) Draw $\boldsymbol{\phi}_t \sim \mathcal{D}(\beta)$
2. For each document d,
   (a) For each topic t let $\alpha_{dt} = \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t)$.
   (b) Draw $\boldsymbol{\theta}_d \sim \mathcal{D}(\boldsymbol{\alpha}_d)$.
   (c) For each word i,
      i. Draw $z_i \sim \mathcal{M}(\boldsymbol{\theta}_d)$.
      ii. Draw $w_i \sim \mathcal{M}(\boldsymbol{\phi}_{z_i})$.

The model therefore includes three fixed parameters: σ^2, the variance of the prior on parameter values; β, the Dirichlet prior on the topic-word distributions; and $|T|$, the number of topics.

Integrating over the multinomials $\boldsymbol{\theta}$, we can construct the complete log likelihood for the portion of the model involving the topics $\boldsymbol{z}$:

$$P(\boldsymbol{z}, \boldsymbol{\lambda}) = \prod_d \frac{\Gamma(\sum_t \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t))}{\Gamma(\sum_t \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t) + n_d)} \prod_t \frac{\Gamma(\exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t) + n_{t|d})}{\Gamma(\exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t))} \times \prod_{t,k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\lambda_{tk}^2}{2\sigma^2}\right). \tag{3.2}$$

The derivative of the log of Equation 3.2 with respect to the parameter λ_{tk} for a given topic t and feature k is

$$\frac{\partial \ell}{\partial \lambda_{tk}} = \sum_d x_{dk} \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t) \times \Big(\Psi\big(\sum_t \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t)\big) - \Psi\big(\sum_t \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t) + n_d\big) + \Psi\big(\exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t) + n_{t|d}\big) - \Psi\big(\exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t)\big)\Big) - \frac{\lambda_{tk}}{\sigma^2}. \tag{3.3}$$

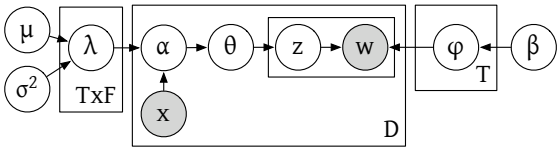

Figure 3.1. The Dirichlet-multinomial Regression (DMR) topic model. Unlike all previous models, the prior distribution over topics, α, is a function of observed document features, and is therefore specific to each distinct combination of metadata feature values.

I trained this model using a stochastic EM sampling scheme, in which I alternate between sampling topic assignments from the current prior distribution conditioned on the observed words and features, and numerically optimizing the parameters $\boldsymbol{\lambda}$ given the topic assignments. This implementation is based on the standard L-BFGS optimizer [23] and Gibbs sampling-based LDA trainer in the Mallet toolkit [27].

3.2.2 Examples

I ran this model on a collection of abstracts from computer science publications selected from the Rexa dataset (see Section 6.2.4). Each document consists of a title and abstract, which I concatenate to create the text output, authors, publication date, and publication venue (journal or conference). In order to provide intuition for the relative values of metadata and topic parameters, Table 3.1 shows examples of estimated parameters for topics related to a single feature, *published in Journal of Machine Learning Research.* Relative to other publication venues in the collection, these results indicate that JMLR publishes more on kernel methods and computational learning theory (VC dimension), and relatively little on information retrieval and human-computer interaction.

Cutting the other way, Table 3.2 shows the estimated parameters for a topic with high probability on the words "reinforcement learning." The top weights are features indicating authors who publish in reinforcement learning and venues (ICML, ECML) that typically publish RL papers. The lowest weights are for linguistics (ACL, COLING) and vision (CVPR, PAMI) venues. The single lowest weight is consistently the intercept term, indicated here as <default>. This magnitude implies that a typical Dirichlet parameter is in the range of $e^{-3} \approx 0.05$, indicating a relatively low precision Dirichlet prior given no additional information. Adding the single feature *published in ICML* leads a Dirichlet parameter of approximately 0.41, with a change in the weight in the sampling distribution equivalent to between a third and a half of a word for this topic.

Table 3.1. Weights on topics for feature *published in Journal of Machine Learning Research.* Multi-word terms are extracted in a post-processing step.

2.27	kernel, kernels, rational kernels, string kernels, fisher kernel
1.74	bounds, vc dimension, bound, upper bound, lower bounds
1.41	reinforcement learning, learning, reinforcement
1.40	blind source separation, source separation, separation, channel
1.37	nearest neighbor, boosting, nearest neighbors, adaboost
	...
-1.12	agent, agents, multi agent, autonomous agents
-1.21	strategies, strategy, adaptation, adaptive, driven
-1.23	retrieval, information retrieval, query, query expansion
-1.36	web, web pages, web page, world wide web, web sites
-1.44	user, users, user interface, interactive, interface

Table 3.2. Weights for the topic *reinforcement learning*. ICML frequently publishes reinforcement learning papers, while COLING focuses on corpus linguistics. <default> is an intercept parameter.

2.99	Sridhar Mahadevan
2.88	ICML
2.56	Kenji Doya
2.45	ECML
2.19	Machine Learning Journal
	...
-1.38	ACL
-1.47	CVPR
-1.54	IEEE Trans. PAMI
-1.64	COLING
-3.76	<default>

3.3 Logistic normal topic models

There are many useful statistical modeling techniques available for regression for data in continuous, real-valued spaces. Text data, however, is fundamentally discrete. Continuous probabilistic models for generating discrete count data are generally constrained to be valid probabilities: multinomial parameters, for example, must be positive and sum to one. In order to introduce the third text regression model, Gaussian Markov random field topic models, it is first necessary to identify a way to bridge the gap between unconstrained continuous variables and discrete spaces. In this chapter, I describe one such method using topic models based on the logistic normal distribution, which is an alternative to the Dirichlet distribution for proportion data [1]. In Chapter 5, after using logistic normal topic models to transition from discrete count data into real-valued spaces, I will focus on GMRF methods for predicting topic distributions given observed metadata.

3.3.1 Definition

The logistic normal distribution is a distribution on the simplex, obtained by transforming a random variable drawn from a multivariate Gaussian distribution. A point $\boldsymbol{\theta}$ in the $T-1$ simplex (i.e., a T-dimensional logistic normal random variable) can be generated as follows:

1. Generate a T-dimensional vector of parameters $\boldsymbol{\beta} \in \mathbb{R}^T$ from a T-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \Sigma)$.

2. Transform $\boldsymbol{\beta}$ into $\boldsymbol{\theta}$ using the logistic transform: $\theta_t = \frac{\exp(\beta_t)}{\sum_{t'=1}^{T} \exp(\beta_{t'})}$

Note that because the proportion is normalized, the distribution is invariant to the addition of a constant to each β_t:

$$\begin{aligned} \frac{\exp(\beta_t + k)}{\sum_{t'=1}^{T} \exp(\beta_{t'} + k)} &= \frac{\exp(\beta_t)\exp(k)}{\sum_{t'=1}^{T} \exp(\beta_{t'})\exp(k)} && (3.4) \\ &= \frac{\exp(k)}{\exp(k)} \frac{\exp(\beta_t)}{\sum_{t'=1}^{T} \exp(\beta_{t'})} && (3.5) \\ &= \frac{\exp(\beta_t)}{\sum_{t'=1}^{T} \exp(\beta_{t'})} && (3.6) \end{aligned}$$

For identifiability it is common practice to set $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ such that β_T (the *reference dimension*) is guaranteed to be zero, so that the transformation becomes $\frac{\exp(\beta_t)}{1+\sum_{t'=1}^{T-1} \exp(\beta_{t'})}$.

3.4 Similarities and differences between models

Each of the topic models and non-topic baselines presented in this thesis can be described using a common framework. Each model consists of two parts:

1. Some number of word distributions over a vocabulary. The prior on these distributions is $\text{Dir}(\beta = 0.01)$ for all models.

2. A distribution over the way that these word distributions combine in a given document for a particular set of metadata elements.

The word-count baseline model has one word distribution for each unique metadata feature. These distributions are combined uniformly over the set of non-zero metadata features for a given document. The unigram language model can be considered a degenerate word-count model that recognizes only one feature, which is present exactly once in all documents.

The standard, non-metadata-enriched topic model has one word distribution for each topic. The method for combining these distributions is through a document-specific discrete

distribution drawn from a shared prior Dir $(\alpha_1, ...\alpha_T)$. The Author-Topic model also has one word distribution per topic, but combines these distributions according to a combination of metadata-specific distributions over topics. The two regression topic models also have the same structure of word distributions, with one per topic. Like AT, the only difference with the simple model is in their method for combining topics given metadata, which is based on a distribution parameterized by an inner product between parameters and observed features.

It is important to note that the fundamental structure of the word distributions (the first component of each model) is identical for all models. Although there are variations in the number of such distributions (exactly one for the unigram model, the number of features for the word-count baseline, and T for all topic models), each distribution is estimated in fundamentally the same way. The probability of a word under such a distribution is proportional to the sum of an observed count and a smoothing parameter $\beta = 0.01$. As a result, in theory no model gets an unfair advantage with regard to words that do not occur in the training set, or occur only rarely. This theoretical observation is experimentally validated in Section 6.5: infrequent words have little to no effect on differences in model performance.

CHAPTER 4

RELATIONSHIP BETWEEN DIRICHLET AND LOGISTIC NORMAL DISTRIBUTIONS

In this chapter I attempt to characterize and explain some of the differences in behavior between the Dirichlet and logistic normal distributions as priors for hierarchical multinomial models. A Dirichlet can be expressed as a logistic transformation of a set of log-gamma distributed random variables. Under this interpretation, a logistic normal with a diagonal covariance matrix can be considered a logistic transformation of a multivariate normal approximation to a set of independent log-gamma random variables.

Ultimately, the decision of whether to use a Dirichlet or logistic normal model depends on the needs of the application. In this thesis I am concerned with distributions over the simplex for use in a regression setting, in which the mean distribution for a particular observation is defined by a linear combination of observed inputs and estimated parameters. The purpose of this section is simply to characterize the behavior of the two distributions in ways that will explain observed phenomena that relate to inference procedures.

Both the logistic normal and Dirichlet distributions provide a distribution over discrete probability distributions. A two-dimensional Dirichlet distribution is a beta distribution. A two-dimensional logistic normal distribution is equivalent to a univariate normal distribution transformed using the inverse logit function $\frac{1}{1+\exp(-x)}$. A beta distribution has two parameters, α_1 and α_2. For simplicity, assume that $\alpha_1 = \alpha_2 = \alpha$, that is, that the beta distribution is *symmetrical*. If α is 1, the distribution of $X \sim \mathcal{B}(\alpha, \alpha)$ is uniform between zero and one. As α increases above 1, random variates drawn from this distribution are increasingly clustered around 0.5. Conversely, as α approaches zero, random variates cluster around zero and one. Similarly, if $Y = \frac{1}{1+\exp(-X)}$ and $X \sim \mathcal{N}(0, \sigma^2)$, the value of Y depends on the ratio between 1 and $\exp(X)$, or equivalently the difference between 0 and X. As the standard deviation σ of the normal distribution increases, this difference tends to increase,

so that Y clusters around zero and one. As σ approaches zero, X becomes increasingly deterministic, the difference between zero and X decreases, and Y tends to cluster around 0.5.

Figure 4.1 shows the relative effect of α and σ^{-1} on the empirical sample variance of both distributions. For each precision k in the powers of ten from 10^{-5} to 10^4, 1000 samples (scalars in the range $(0, 1)$) from $\mathcal{B}(k, k)$ and $\mathcal{LN}(0, k^{-2})$ were drawn and their variance computed. As expected, this variances goes from $1/4$ for small precisions (all values are close to either 1 or 0) and 0 for large precisions (all values are very close to 0.5), but there is a considerable divergence in the middle range. The inverse standard deviation seems to result in a reasonable approximation; the fit with $\sigma^{-2} = \alpha$ is much worse.

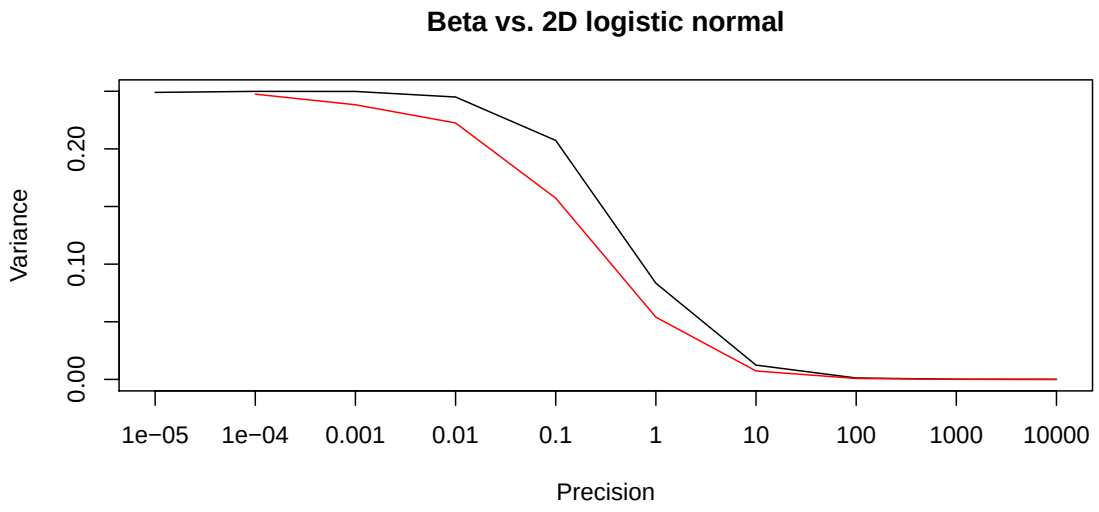

Figure 4.1. Identical values of α and σ^{-1} have similar effects on observed variance. Comparison of the effect of parameter α in a symmetric beta distribution and σ^{-1} in a two-dimensional logistic normal distributions. The top (black) line shows the sample variance of 1000 draws from the beta distribution with each value of α, and the bottom (red) line shows the sample variance of 1000 draws from a logistic normal with the same value for σ^{-1}. There is a rough correspondence between empirical variance of these two distributions, especially in the extreme values.

We can, however, do considerably better. Aitchison and Shen [1] state that the minimum KL divergence logistic normal distribution to a Dirichlet distribution sets $\beta_t = \Psi(\alpha_t) - \Psi(\alpha_T)$ and $\sigma_t^2 = \Psi'(\alpha_t) + \Psi'(\alpha_T)$, where $\Psi(x)$ is the digamma function, $\Psi'(x)$ is trigamma,

and T is the reference dimension. In the two dimensional symmetric case, all values of α_t are the same so the mean remains 0 and the variance is $2\Psi'(\alpha)$. Sample variances from simulations using these parameters are shown in Figure 4.2. Sample variance is slightly higher for the logistic normal in the range around 0.1 to 1, but the overall divergence is much smaller than simply setting the inverse standard deviation to equal α.

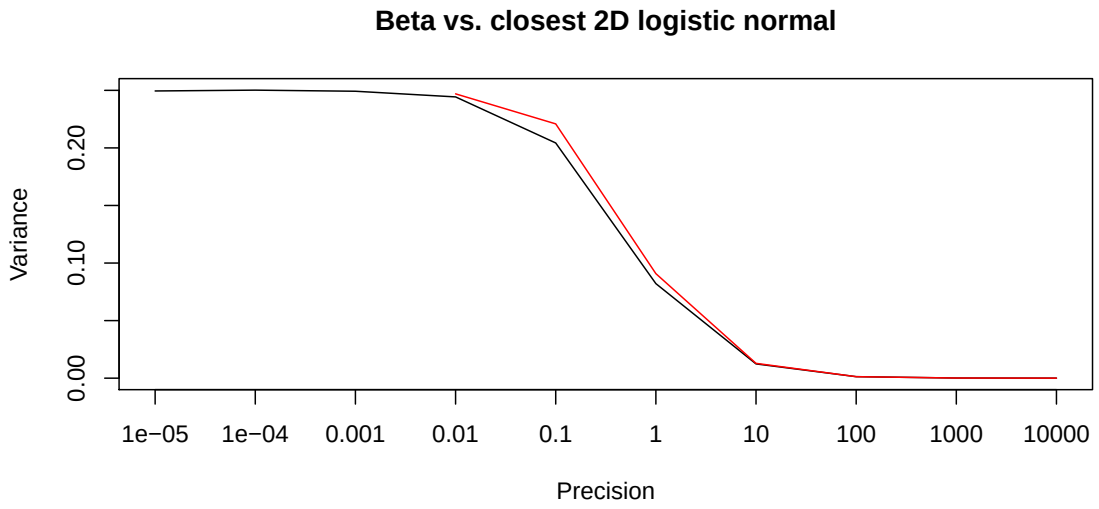

Figure 4.2. Setting σ^{-1} to minimize divergence for values of α results in a close but not perfect match. The same sample variances as in the previous figure, but in this case with the logistic normal variance parameter set to minimize KL divergence with the symmetric beta distribution at each value of α along the horizontal axis. Results are much closer, with the logistic normal having slightly greater variance around 0.1 to 1.

Insight into the significance of the digamma and trigamma functions can be gained by considering the construction of a beta or Dirichlet random variate. Without loss of generality, consider a sample p from a two parameter Dirichlet distribution, that is, Beta (α_1, α_2). To sample from this distribution, we first sample two positive real-valued random variates $X_1 \sim \text{Gamma}\,(\alpha_1, 1)$ and $X_2 \sim \text{Gamma}\,(\alpha_2, 1)$ and then normalize:

$$p = \frac{X_1}{X_1 + X_2} \tag{4.1}$$

Equivalently,

$$
\begin{aligned}
p &= \frac{\exp(\log X_1)}{\exp(\log X_1) + \exp(\log X_2)} && (4.2)\\
&= \frac{\exp(\log X_1 - \log X_2)}{\exp(\log X_1 - \log X_2) + \exp(\log X_2 - \log X_2)} && (4.3)\\
&= \frac{\exp(\log X_1 - \log X_2)}{\exp(\log X_1 - \log X_2) + 1} && (4.4)
\end{aligned}
$$

The last line is the familiar inverse logit function. The transformed variable $Y = \log X$ where $X \sim \text{Gamma}\,(\alpha, 1)$ has the distribution

$$
\begin{aligned}
P(Y \mid \alpha) &= \frac{1}{\Gamma(\alpha)} (\exp Y)^{\alpha-1} \exp(-\exp(Y)) \left| \frac{d}{dY} \exp(Y) \right| && (4.5)\\
&= \frac{1}{\Gamma(\alpha)} (\exp Y)^{\alpha} \exp(-\exp(Y)) && (4.6)\\
&= \frac{1}{\Gamma(\alpha)} \exp\left(\alpha Y - \exp(Y)\right) && (4.7)
\end{aligned}
$$

Frühwirth-Schnatter et al. [16] discuss the distribution of the one-parameter *negative* log-gamma distribution, that is, the variable $-\log X$ where $X \sim \text{Gamma}\,(\alpha, 1)$, which has expectation $-\Psi(\alpha)$ and variance $\Psi'(\alpha)$ and pdf $p(y \mid \alpha) = \Gamma(\alpha)^{-1} \exp(-\alpha y - e^{-y})$. The simple log-gamma distribution in Eqn 4.7 therefore has expectation $\Psi(\alpha)$ (as expectations are linear) and variance $\Psi'(\alpha)$ (by multiplying by -1^2). The log-gamma distribution has support over the real line, so it can be approximated using a univariate Gaussian by matching the mean and variance.

Figure 4.3 shows a comparison of log-gamma and normal densities with values of α in the range 0.01 to 10.0. The normal approximation becomes increasingly tight as α increases, but is quite loose for smaller values. For one thing, the log-gamma is asymmetric: it goes to zero fairly quickly for values greater than its mode, but has a very heavy tail in the negative direction. There is a poor match at 0.01, corresponding to a normal with $\mu = -100$ and $\sigma^2 = 10000$. At $\alpha = 10$, on the other hand, the approximation is quite close.

Figure 4.4 shows a much wider view of the log-gamma distribution at $\alpha = 0.01$ and the associated normal approximation. In this case the normal distribution is a very poor approximation: for Y greater than zero, the $-e^Y$ term in the pdf $\Gamma(\alpha)^{-1} \exp[\alpha Y - \exp(Y)]$ becomes large and negative very quickly, leading to densities near zero and a pdf that becomes close to an exponential distribution reflected around the y-axis.

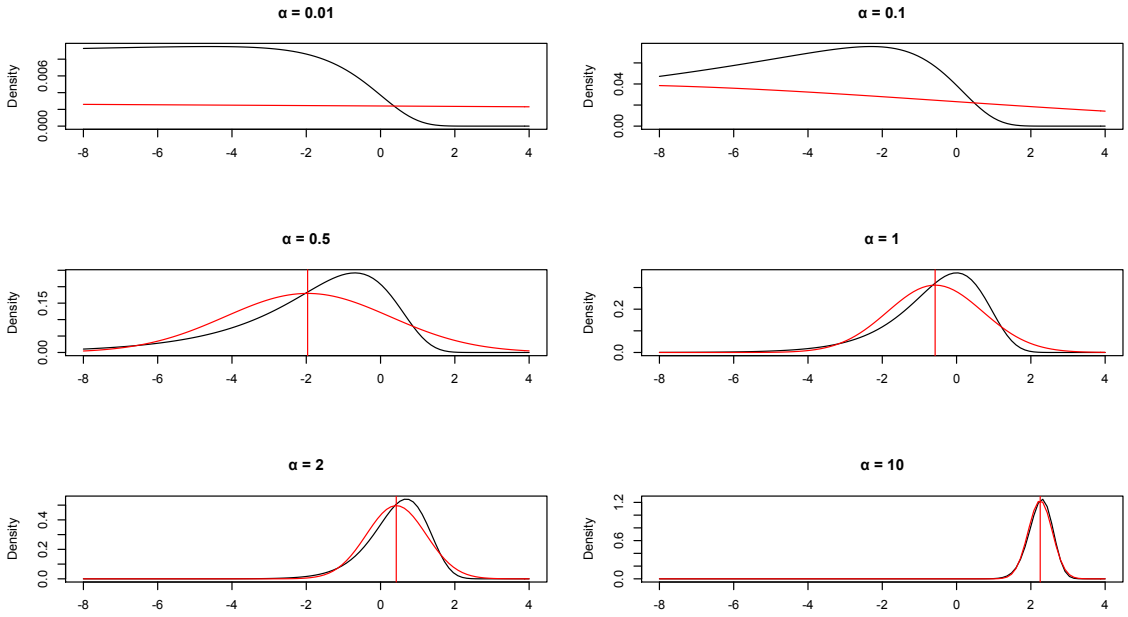

Figure 4.3. Divergence between Dirichlet and LN increases as α decreases. Difference between six log-gamma distributions with $\alpha \in \{0.01, 0.1, 0.5, 1, 2, 10\}$ (black) and the closest Gaussian distribution (red). The means (shown where visible with vertical red lines) are at -100.5, -10.4, -2.0, -0.6, 0.4, and 2.3. Variances are 10000, 101, 4.9, 1.6, 0.64, and 0.10.

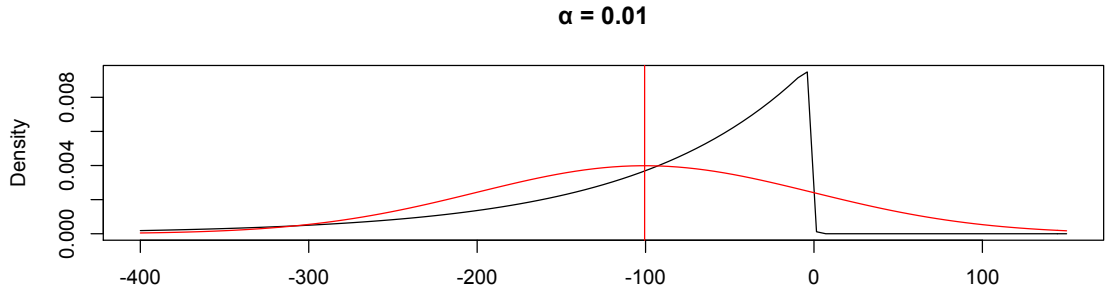

Figure 4.4. Small α leads to an asymmetric shaped log-gamma. A wider view of the log-gamma distribution at $\alpha = 0.01$ and its normal approximation.

4.1 Sparse, high dimensional distributions

The previous analysis has focused for clarity on low-dimensional or even binary distributions. The analysis of text data is characterized by sparse but very high dimensional data. A word-count vector representation of a document, for example, will have as its length the size of the vocabulary (50,000 elements being typical of English corpora), but almost all elements will be zero. Even at the level of topic concepts, we expect that there will be hundreds of distinct topics in any significant collection, but that only a small number will appear in any given document. In addition, there are very few words or topics that appear regularly in many documents: most will occur infrequently, but in high concentrations when they do appear.

To explore the relative effect of parameter settings on sparsity in both distributions, Figure 4.5 shows 20 settings of α between 0.05 and 1.0. For each setting, 1000 samples were drawn from 100-dimensional distributions $\mathcal{D}(\alpha, ..., \alpha)$ and $\mathcal{LN}(0, \Psi'(\alpha))$, which are the most similar distributions by KL divergence. The sparsity of each set of sampled distributions $\boldsymbol{p}$ is measured using four metrics:

1. Shannon entropy ($-\sum_i p_i \log p_i$), which is maximized by a uniform distribution and zero for a deterministic distribution.

2. Number of elements greater than 0.05. A 100-dimensional uniform distribution will have zero elements greater than 0.05 as all elements are 0.01. As the distribution becomes more concentrated on a smaller number of dimensions, this metric will increase, and then decrease as probability mass becomes concentrated on a single dimension.

3. Sum of squared elements. This metric emphasizes elements closer to 1.0. For a uniform distribution the metric will approach zero, while for a deterministic distribution it will be equal to 1.0.

4. The single largest element. The maximum element will be small for a uniform distribution and close to 1.0 for a deterministic distribution.

Figure 4.5 shows that there is substantial divergence between the sparsity of vectors drawn from the Dirichlet and logistic normal distributions using the $\sigma^2 = \Psi'(\alpha_t)$ approxi-

mation. The Dirichlet distribution produces roughly uniform vectors through most of the range of parameters, while the logistic normal produces significantly sparser, more deterministic vectors. The logistic normal is also much more variable in entropy, sum-of-squares, and maximum value.

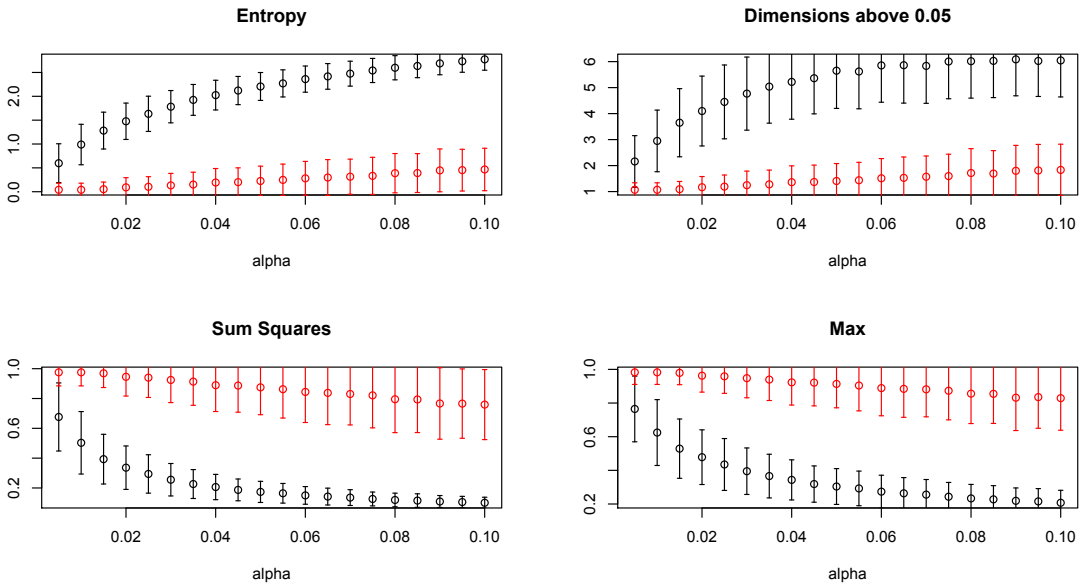

Figure 4.5. LN becomes sparse faster than Dirichlet. Comparison of sparsity metrics for 100-dimensional Dirichlet (black) with specified α_t and 100-dimensional logistic normal (red) with $\mu = 0$ and $\sigma^2 = \Psi'(\alpha_t)$. Although the normal best approximates the log-gamma distribution, the resulting sparsity pattern after the logistic transformation is very different.

The comparison between log-gamma and normal distributions shown in Figure 4.3 offers some insight for this behavior. When α_t is large, the normal approximation is quite good, but for small α_t the log-gamma density is asymmetric and becomes small very quickly above zero, while the normal approximation has no such asymmetry and therefore has significant density for values well above zero. This much heavier tail in the positive direction should result in much greater probability that one of the random variates is very large in comparison to the others. When these variates are exponentiated and normalized, the largest outlier dominates all others, resulting in sparse distributions.

As the Dirichlet parameter α gets smaller, the sparsity of distributions drawn from the Dirichlet increases. As a comparison to Figure 4.5, Figure 4.6 shows the same Dirichlet

parameters compared to logistic normal distributions approximating Dirichlet distributions with parameters that are multiplied by 10, in other words with $\sigma^2 = \Psi'(10\alpha)$. The match is clearly approximate, but suggests a qualitative mapping between parameter ranges. Within these two ranges, the entropy becomes comparable, but the other three metrics show similar curves but wide divergence. The sum of squares and maximum show roughly the same behavior: they are very small at the larger end of the parameter values, but in both cases the logistic normal becomes sparse more quickly. In contrast, the number of dimensions with significant weight (above 0.05) is consistently larger for the Dirichlet, but remains relatively constant at around four for the logistic normal and six for the Dirichlet over the range of parameter values that show increasingly divergent squared sums and maxima, and entropy values that cross over from lower values for the Dirichlet to lower values for the logistic normal.

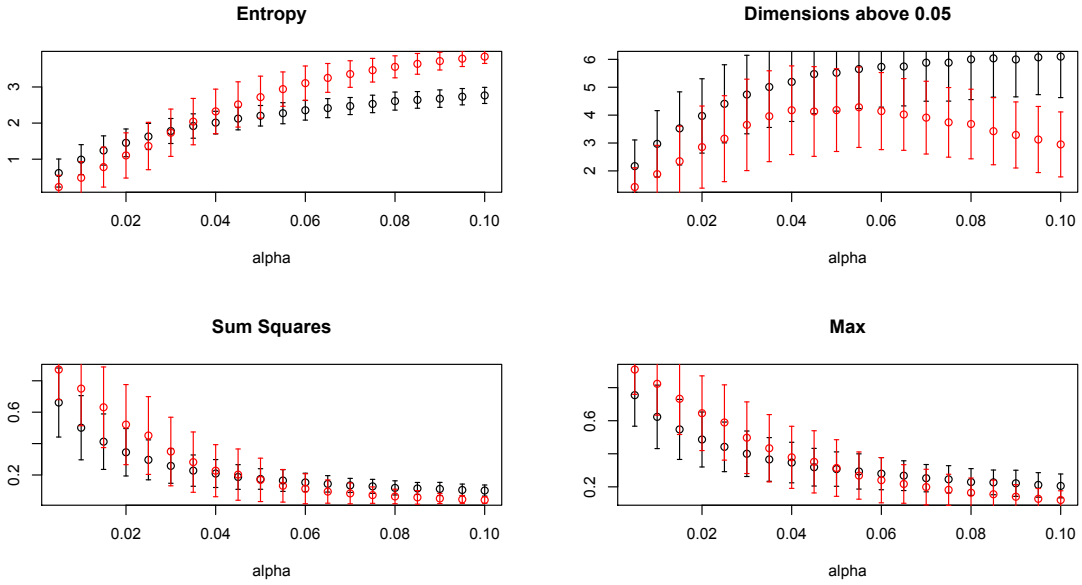

Figure 4.6. Sparsity is closer with $\sigma^2 = \Psi'(10\alpha)$. Comparison of sparsity metrics for 100-dimensional Dirichlet (black) with α_t in the range 0.05 to 0.1 (equal to the values in the previous figure) and 100-dimensional logistic normal (red) with $\mu = 0$ and $\sigma^2 = \Psi'(10\alpha)$, so that each logistic normal approximates a Dirichlet with parameters ten times larger than the Dirichlet it is actually compared to. The logistic normal still converges to a deterministic distribution faster than the Dirichlet, but the two distributions are much closer.

Multiplying α_t by ten is clearly an arbitrary change: I am not aware of any theoretical justification for it and do not claim it to be in any way optimal. Nevertheless, this simple variation produces sparsity patterns that are clearly more closely matched than setting $\sigma^2 = \Psi'(\alpha)$. These results indicate that there may be an optimal method for setting logistic normal variances to match the sparsity patterns of a Dirichlet, but finding the best normal approximation for the log-gamma distribution is not it.

4.2 Logistic-normal-multinomial distributions

The previous sections have explored distributions over values in the $T-1$-dimensional simplex, the space of non-negative vectors that sum to one. Values in the $T-1$-dimensional simplex can be used as the parameters of T-dimensional discrete or multinomial distributions. In the case of the Dirichlet-multinomial hierarchical distribution $P(\boldsymbol{z} \mid \alpha) = \prod_t \theta_t^{n_t} \ \mathcal{D}(\boldsymbol{\theta}; \alpha)$, the posterior distribution over $\boldsymbol{\theta} \mid \boldsymbol{z}$ remains Dirichlet. Unlike the Dirichlet distribution, however, the logistic normal distribution is not conjugate to the multinomial. The posterior distribution of $\boldsymbol{\beta} \mid \boldsymbol{z}$ is

$$
\begin{aligned}
P(\boldsymbol{\beta} \mid \boldsymbol{z}) &\propto \prod_t \left(\frac{\exp(\beta_t)}{\sum_{t'} \exp(\beta_{t'})} \right)^{n_t} \mathcal{N}(\beta_t; \mu, \kappa^{-1}) && (4.8) \\
&\propto \exp\left[\sum_t \left(n_t \beta_t - \frac{\kappa}{2} \beta_t^2 + \kappa \beta_t \mu \right) - N \log\left(\sum_t \exp(\beta_t) \right) \right] && (4.9) \\
&\propto \exp\left[\sum_t \left((n_t + \kappa\mu) \beta_t - \frac{\kappa}{2} \beta_t^2 \right) - N \log\left(\sum_t \exp(\beta_t) \right) \right] && (4.10)
\end{aligned}
$$

where κ is the precision or inverse variance, $n_t = \sum_i I(z_i = t)$ and $N = \sum_t n_t$. This distribution is exponential family with sufficient statistics $\beta_1, ..., \beta_T, \beta_1^2, ...\beta_T^2, \log(\sum_t \exp(\beta_t))$, but is not of any well-known parametric distribution.

It is possible to sample from this posterior distribution using the Metropolis-Hastings algorithm, but the acceptance rate may be low. Although it is not immediately obvious due to the non-conjugacy of the logistic normal and multinomial distributions, it is also possible to use Gibbs sampling by adding auxiliary variables, thus representing the difficult posterior distribution in Eq. 4.10 as a marginalization of a simpler distribution. This method is based on an algorithm for Gibbs sampling of logistic regression parameters presented by

Groenewald and Mokgatlhe [19], which is an extension of Albert and Chib's method for Gibbs sampling in probit regression [3].

Consider a very simple model for a single binary event X, in which a single parameter β is sampled from a standard normal, and then X is sampled from a Bernoulli distribution with parameter $p = e^{\beta}/(1 + e^{\beta})$. An alternative representation of the same model is to sample a random variable u from a uniform distribution over the interval $(0, 1)$, $U \sim \mathcal{U}(0, 1)$ and then deterministically set $x = I(U < p)$. Given values of X and β, U is distributed $\mathcal{U}(0, p)$ if $X = 1$ and $\mathcal{U}(p, 1)$ otherwise. If U and X are known but β is unknown, it is possible to infer that if $X = 1$, then we can determine a lower bound on p and therefore β:

$$
\begin{aligned}
e^{\beta}/(1 + e^{\beta}) &> U && (4.11)\\
e^{\beta} &> \frac{U}{1 - U} && (4.12)\\
\beta &> \log \frac{U}{1 - U} && (4.13)
\end{aligned}
$$

The conditional distribution of β is therefore proportional to $\mathcal{N}(\beta; 0, 1)I(\beta > \log \frac{U}{1-U})$, that is, a truncated normal distribution. To estimate the posterior distribution of β, we can therefore alternate between sampling $\beta|U$ and $U|\beta$. Figure 4.7 shows 10,000 Gibbs samples from the posterior distribution of these two variables with $X = 1$. Not surprisingly, estimates vary widely. U varies between zero and one, but tends to be closer to zero. The posterior distribution over β is shifted slightly from the zero-mean prior and has slightly lower variance ($\sigma = 3.0, \bar{s} = 2.17$).

When there is only a single pair $\{X, U\}$, it is only possible to specify either a lower or upper bound on β, depending on whether $X = 1$ or 0, respectively. If instead we have a sequence of binary random variables $X_1, ..., X_N$ all sampled i.i.d. from a Bernoulli distribution with parameter p, we can add an auxiliary variable $U_1, ..., U_N$ alongside each X_i. These multiple pairs $\{X_i, U_i\}$ can be used to get a more accurate estimate of β, particularly if the total number of ones n_1 and the total number of zeros n_0 are both greater than zero. In this case we have both a lower bound and an upper bound on β. In addition, if n_1 is greater than one, then the lower bound on β only depends on the single largest U_i such that $X_i = 1$, and similarly for the case where $n_0 > 1$. Using standard

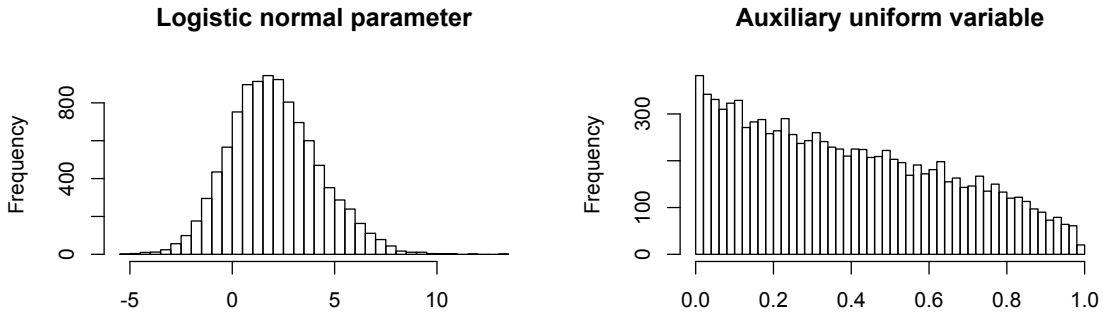

Figure 4.7. After one (positive) observation, the posterior shifts slightly. Histograms of β and U values over 10000 iterations of Gibbs sampling for $X = 1$, $\sigma = 3$. The sample mean of β is 2.04, corresponding to an estimated $p = 0.885$. The sample standard deviation is 2.17, slightly less than the prior.

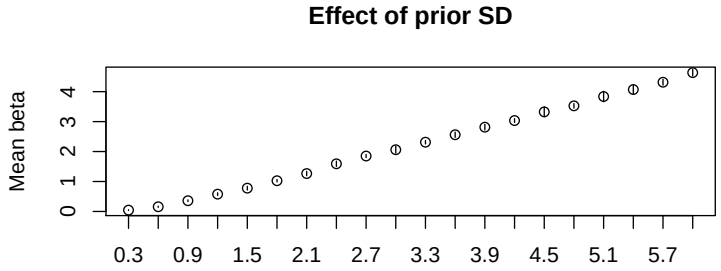

Figure 4.8. With one observation, the prior strongly affects the posterior. Effect of variation in σ on the estimated posterior distribution of β, using the same model as in Figure 4.7. As there is only one (positive) instance, as the prior variance grows the estimate of β increases.

results concerning order statistics, it can be shown that the largest of M uniform random variables is distributed according to Beta $(M, 1)$, while the smallest is distributed according to Beta $(1, M)$. It is therefore unnecessary to sample all N auxiliary variables: Given n_1 and n_0, $\max_{i|X_i=1} U_i$ and $\min_{i|X_i=0} U_i$ (and hence the lower and upper bounds on β) can both be computed directly by drawing a random variable from a beta distribution, scaled and shifted as appropriate. Finally, samples from a beta distribution with either parameter equal to one can be drawn using transformations of uniform random variables [13]: $u^{\frac{1}{M}} \sim$ Beta $(M, 1)$ and $1 - u^{\frac{1}{M}} \sim$ Beta $(1, M)$, where $u \sim \mathcal{U}(0, 1)$.

Figure 4.9 shows samples from the posterior distribution over β and $\max_{i|X_i=1} U_i$ for a larger sample size ($N = 1000, n_1 = 100$) than Figure 4.7. The true parameter is $\beta = \log(0.1/0.9) = -2.20$. After discarding the first 500 of 10000 iterations (shown in the plot as a sequence of values from initial values with $\beta = 0$), the sample mean is -2.18 and the standard deviation is 0.10. Both variables shown explore regions centered around their true values. In the case with a single $X = 1$, the variance of the prior had a significant effect on estimated values of β. Figure 4.10 shows the same experiment run on the larger $N = 1000$ dataset. With this much data, estimated values are expected to be relatively stable with respect to the prior. Indeed, values are tightly centered close to the true value and show little dependence on σ. Finally, Figure 4.11 shows the effect of sample size. Clearly, $N = 1$ is unlikely to provide useful estimates without strong prior knowledge, but $N = 1000$ may be unrealistically large for many applications. The figure shows the same experimental setup as before but with N ranging from 50 to 1000. In each case, n_1 is $0.1N$. Estimated values show little differentiation over this range of sample sizes, although there is a slight increasing trend as the sample size grows. This trend appears to be the result of slower convergence: if we do not remove "burn-in" iterations, $\bar{\beta}$ increases roughly linearly with N.

This sampling scheme leads to a simple physical interpretation. Consider a confined space partitioned into two sections with a flexible barrier. Each section contains some number of particles. Sampling β is equivalent to trying to equalize the pressure in both sections, while sampling the variables $U_1, ..., U_N$ is equivalent to sampling the positions of the particles within each section. If the pressure is unbalanced, particles in one section are more likely to be closer to the barrier and particles in the other section are likely to be

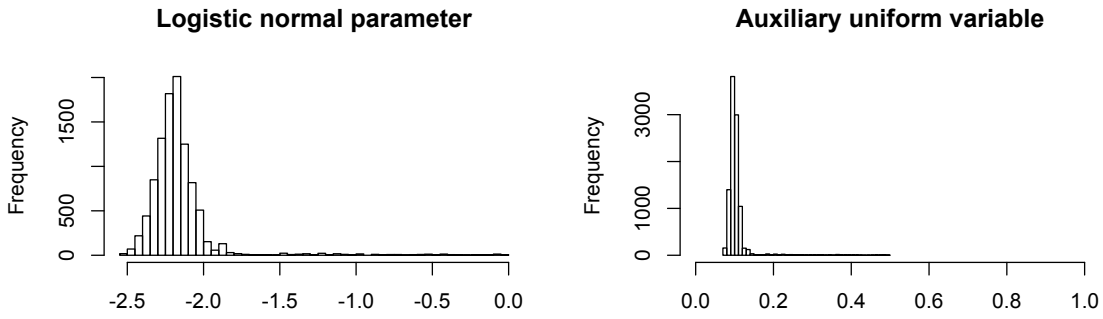

Figure 4.9. Gibbs sampling converging quickly. Histograms of β and $\max_{i|X_i=1} U_i$ values over 10000 iterations of Gibbs sampling for $n_1 = 100, N = 1000$, $\sigma = 3$. After a short burn-in period, the sampler converges to the true parameter value. The sample mean of β is -2.18, corresponding to an estimated $p = 0.107$. The sample standard deviation is 0.10, substantially less than the prior. The maximum uniform random variable U_i such that $X_i = 1$ tends to be tightly concentrated around 0.1.

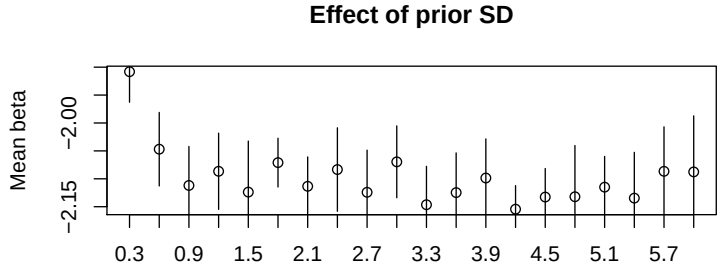

Figure 4.10. With 1000 observations, the prior variance has little effect. Effect of variation in σ on the estimated posterior distribution of β, using the same model as in Figure 4.9 (100 ones out of 1000). There is very little variation in the estimated parameter value even over a wide range of values for σ.

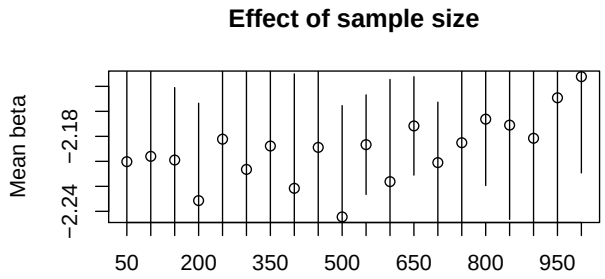

Figure 4.11. Prior variance has little effect at $N \geq 50$. Effect of variation in N on the estimated posterior distribution of β, using the same model as in Figure 4.9, where $n_1 = 0.1N$ and $\sigma = 3$. There is very little variation in the estimated parameter value even over a wide range of values for N, with a slight upturn as sample size increases that appears to be due to slower convergence to the true posterior.

further, increasing the pressure against the barrier from one direction. The gaussian prior corresponds to the initial position and elasticity of the barrier. This metaphor also helps to explain the slower convergence of $\bar{\beta}$ for large values of N, as when the number of particles is large in both sections, the barrier has less "wiggle room" and thus moves less in each iteration.

4.2.1 Point estimates

As a comparison to MCMC inference, it is also possible to consider maximum *a posteriori* estimates. The gradient of the log of Eq. 4.10 is

$$\frac{d\ell}{d\beta_t} = n_t - N\frac{\exp(\beta_t)}{\sum_{t'}\exp(\beta_{t'})} - \kappa(\beta_t - \mu) \tag{4.14}$$

$$= n_t - N\pi_t - \kappa(\beta_t - \mu), \tag{4.15}$$

where π_t represents $\exp(\beta_t)/\sum_{t'}\exp(\beta_{t'}) = P(t \mid \boldsymbol{\beta})$. This expression has the familiar interpretation as the difference between the actual count n_t and the expected count under the current parameter settings $N\pi_t$, plus a regularization term from the Gaussian prior that encourages the difference between the parameter and the prior mean μ, scaled by κ, to be small. This expression cannot be set to zero and analytically solved for β_t, so to find a MAP

estimate the expression must be numerically optimized. The matrix of second derivatives is given by the following equations:

$$\frac{d^2\ell}{d\beta_t^2} = -N\pi_t(1-\pi_t) - \kappa \tag{4.16}$$

$$\frac{d^2\ell}{d\beta_r\beta_s} = N\pi_r\pi_s. \tag{4.17}$$

For a two-dimensional model with a single parameter β, the Newton update with a zero-mean prior is therefore

$$\beta^{new} = \beta - \frac{n_1 - N\pi_1 - \kappa\beta}{-N\pi_1(1-\pi_1) - \kappa}. \tag{4.18}$$

For the earlier case with $n_1 = 100, N = 1000$, $\sigma = 3.0$ ($\kappa = 0.11$), and β initialized to 0, this update converges to $\beta = -2.19$ after three updates and is constant to six decimal places after five.

4.2.2 Multivariate distributions

The Gibbs sampling method described above can be extended to distributions with more than two dimensions by iterating over each dimension t in turn and sampling uniform random variables as if t were one of two binary dimensions, that is, with $n_1 = \sum_i I(X_i = t)$ and $n_0 = \sum_i I(X_i \neq t)$. Solving for the bounds on β_t given $U_{max} = \max_{i|X_i=t} U_i$ and $U_{min} = \min_{i|X_i \neq t} U_i$:

$$\log \frac{\left(\sum_{t' \neq t} \exp \beta_t\right) U_{max}}{1 - U_{max}} < \beta_t < \log \frac{\left(\sum_{t' \neq t} \exp \beta_t\right) U_{min}}{1 - U_{min}}. \tag{4.19}$$

As noted before, in the applications that are the focus of this thesis, count vectors tend to be sparse: the number of non-zero elements of an observation vector is expected to increase more slowly than the number of dimensions. It is therefore important to consider the effect of such sparsity on the accuracy of inference methods.

Identifiability is an important issue. As stated previously, the transformation from a T-dimensional vector $\boldsymbol{\beta}$ in $\mathbb{R}^T$ to a T-dimensional distribution $\boldsymbol{\theta} = \frac{1}{\sum_t e^t} \exp \boldsymbol{\beta}$ in the

simplex $\mathbb{S}^T$ is not unique: adding any scalar k to every element of $\boldsymbol{\beta}$ results in the same $\boldsymbol{\theta}$. Several techniques are commonly used to reduce the dimensionality of $\boldsymbol{\beta}$ to $T-1$ and therefore produce a unique solution. The simplest method is declare one dimension T to be the *reference* dimension and fix its value $\beta_T = 0$. This method is equivalent to the *additive logratio transformation* (ALR) described by Aitchison [2], as the variables become $\beta_t = \log \theta_t/\theta_T$, with $\beta_T = \log \beta_T - \log \beta_T = 0$. In the context of document modeling, having a defined reference dimension is difficult as almost all topics will not occur in any given document. If a particular topic always has parameter zero, then in documents that do not contain that topic the average value of β_t for a topic that does occur must be substantially greater than zero, while in a document that does in fact contain significant numbers of word tokens in the reference topic, the average parameter for all other topics must be substantially less than zero. For example, consider two documents: one that contains the reference topic T and one that does not. In the first case, θ_T must be large, so a value of some other β_t close to zero represents a relatively large probability of topic t occurring. The parameter for some topic r that does not occur must be relatively large and negative. In the second case, θ_T must be small in order to give high probability to the event that no words are assigned to the reference topic, so a value of β_t close to zero would represent a very small probability. If the same topic r also does not exist in the topic, we expect θ_r to be roughly the same as θ_T, and therefore β_r to be close to zero. The prevalence of topic r has not changed between these two documents, but the posterior distribution over β_r for the two documents must be very different.

Another alternative is the *centered logratio transformation* (CLR), in which there is no reference dimension but $\boldsymbol{\beta}$ is constrained to sum to zero, $\beta_t = \log \theta_t - 1/T \sum_{t'} \log \theta_{t'}$: β_t is equal to the log of θ_t divided by the geometric mean of $\boldsymbol{\theta}$. This method does not privilege one particular dimension, but comes at the cost of restricting the prior distribution over $\boldsymbol{\beta}$ to be a $T-1$ dimensional subspace of $\mathbb{R}^T$. The CLR transformation is also problematic in the context of Gibbs sampling, as β_t is deterministic given all other values of $\boldsymbol{\beta}$.

Finally, it is possible to define an over-parameterized model and rely on the Gaussian prior to form a soft centering on the parameter values. Although the likelihood term $P(X \mid \boldsymbol{\beta})$ will not have a unique maximum, the posterior distribution over $\boldsymbol{\beta}|X, \mu, \kappa$ will have

a unique maximum. This method has been successfully used in L_2-regularized multinomial logistic regression in the machine learning community.

4.3 Inference about the mean vector $\boldsymbol{\mu}$ and observation precision κ

The previous section introduces a distribution over count vectors $\boldsymbol{n} = n_1, ..., n_T$:

$$P(\boldsymbol{n}, \boldsymbol{\beta} \mid \boldsymbol{\mu}, \kappa) \quad \propto \quad \prod_t \left(\frac{\exp(\beta_t)}{\sum_{t'} \exp(\beta_{t'})} \right)^{n_t} \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}, \kappa_o^{-1} I_T). \tag{4.20}$$

This expression differs from Eq. 4.10 in that the normal distribution is multivariate and that the *observation precision* κ has been subscripted κ_o for clarity in the following presentation. In order to make meaningful Bayesian inferences about $\boldsymbol{\mu}$ and κ_o, it is necessary to introduce two additional aspects. First, rather than a single count vector $\boldsymbol{n}$, consider D observations, each represented by an observation-specific mean vector $\boldsymbol{\beta}^{(d)}$ and count vector $\boldsymbol{n}^{(d)}$. Second, it is necessary to introduce hyperpriors on the parameters $\boldsymbol{\mu}$ and κ_o. For $\boldsymbol{\mu}$ a reasonable prior is a zero-mean normal with diagonal covariance: $\mathcal{N}(0_T, \kappa_p^{-1} I_T)$. For κ_o, a reasonable prior is Gamma (a, b), with density proportional to $x^{a-1} \exp(-xb)$. Assume for the moment that κ_p, a, and b are fixed, known constants. The full probability of the observations is therefore

$$\begin{aligned} P(\{\boldsymbol{n}\}, \{\boldsymbol{\beta}\} \mid \boldsymbol{\mu}, \kappa) \quad \propto \quad & \prod_d \left[\prod_t \left(\frac{\exp(\beta_t^{(d)})}{\sum_{t'} \exp(\beta_{t'}^{(d)})} \right)^{n_t^{(d)}} \mathcal{N}(\boldsymbol{\beta}^{(d)}; \boldsymbol{\mu}, \kappa^{-1} I_T) \right] \times \\ & \mathcal{N}(\boldsymbol{\mu}; 0_T, \kappa_p^{-1} I_T) \text{ Gamma}(\kappa_o; a, b). \end{aligned} \tag{4.21}$$

The conditional distribution of each μ_t given all other variables is Gaussian and is conditionally independent of the count vectors and (because of the diagonal prior covariance matrix) all other means $\mu_{t' \neq t}$.

$$\boldsymbol{\mu} \mid \{\beta_t\}, \kappa_p, \kappa_o \quad \sim \quad \mathcal{N}\left(\frac{\kappa_o \sum_d \beta_t^{(d)}}{\kappa_p + D\kappa_o}, (\kappa_p + D\kappa_o)^{-1} \right) \tag{4.22}$$

4.3.1 Synthetic data: $\boldsymbol{\mu}$

As before, synthetic data experiments are useful in identifying Figure 4.12 shows synthetic results with $T = 10$ and $D = 100$ over 500 Gibbs samples after a 500 iteration burn-in period. For identifiability, the estimated document mean vectors $\{\boldsymbol{\beta}^{(d)}\}$ were normalized to sum to 0 after each sampling iteration completed. This method leads to estimates close to the true parameters, and fairly stable samples. Overall, the sampler slightly overestimates the absolute magnitude of parameters further away from zero, either positive or negative, but reconstructs their relative positions quite well. The solid line represents the line $y = x$, while the dashed line shows the MLE linear regression line predicting estimated values given true values. Focusing on one particular parameter μ_1, the sampler appears to have converged, staying very close to its overall mean value throughout the sampling run.

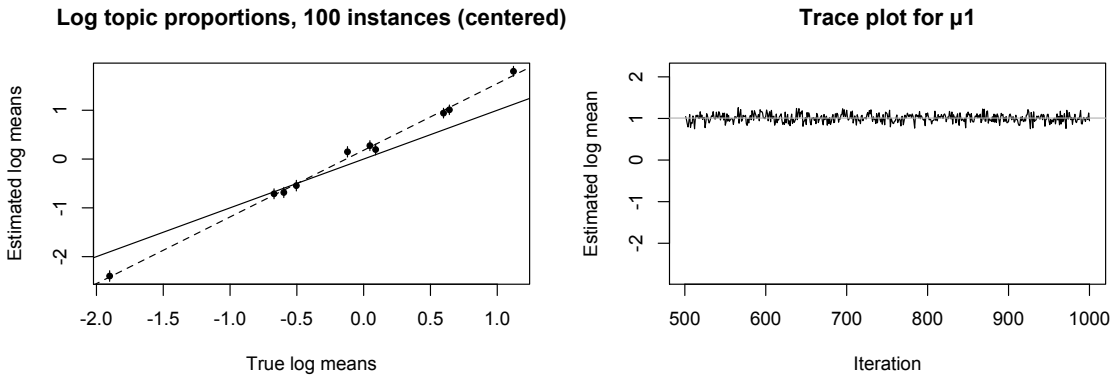

Figure 4.12. Sampled values of μ_0 converge. Comparison of actual and estimated 10-dimensional $\boldsymbol{\mu}$ from a synthetic corpus with 100 observations each with mean vector $\boldsymbol{\beta}^{(d)}$ drawn from a Gaussian with mean vector $\boldsymbol{\mu}$ and $N^{(d)} = 300$ discrete random variables drawn from $\text{logit}^{-1}(\boldsymbol{\beta}^{(d)})$. The solid line is at $y = x$, and the dotted line is the maximum likelihood linear regression line of the estimated parameters given the true parameters. The right plot shows 500 iterations of samples for μ_1, with the overall sample mean shown as a gray line.

4.3.2 Synthetic data: κ

These results were generated with κ_p and κ_o set to their true values, both 1.0. It is next necessary to to consider inference over κ_o. Collecting terms involving κ_o from the overall probability function results in

$$
\begin{aligned}
P(\kappa_o \mid \boldsymbol{\mu}, \{\beta_t\}, \kappa_p) &\propto \prod_d \prod_t \mathcal{N}(\beta_t^{(d)}; \mu_t, \kappa_o^{-1}) \times \text{Gamma}\left(\kappa_o; a, b\right) && (4.23)\\
&\propto \kappa_o^{DT/2} \prod_d \prod_t \exp\left[-\frac{\kappa_o}{2}(\beta_t^{(d)} - \mu_t)^2\right] \times \kappa_o^{a-1} \exp\left[-\kappa_o b\right] && (4.24)\\
&\propto \kappa_o^{DT/2+a-1} \exp\left[-\kappa_o \left(\frac{\sum_{d,t}(\beta_t^{(d)} - \mu_t)^2}{2} + b\right)\right]. && (4.25)
\end{aligned}
$$

This expression is the kernel of a Gamma density, so the posterior becomes

$$
\kappa_o \mid \boldsymbol{\mu}, \{\beta_t\}, \kappa_p \sim \text{Gamma}\left(DT/2 + a, \frac{\sum_{d,t}(\beta_t^{(d)} - \mu_t)^2}{2} + b\right). \quad (4.26)
$$

This distribution has a simple interpretation as the prior shifted by half the number of observations and half the sum of squared divergences from the mean. Unfortunately, this distribution performs poorly in the context of logistic-normal-multinomial models. Specifically, for synthetic data with true $\kappa_o = 1.0$, estimated values of κ_o do not converge to 1.0, and in fact quickly approach zero. The simulation becomes numerically unstable after several hundred iterations. Figure 4.13 shows the estimated values of $\boldsymbol{\beta}^{(d)}$ for one instance with $\boldsymbol{n}^{(d)} = \{3, 12, 12, 44, 9, 0, 4, 0, 215, 1\}$ over 100 Gibbs sampling iterations. The shaded region shows one standard deviation ($1/\sqrt{\kappa_o}$) from 0 for the current value of κ_o at each iteration. The two gray lines represent the two dimensions with $n_t = 0$. These two values have an upper bound but no lower bound, and are essentially draws from the prior. Examining the first 10 iterations shows what is happening: the two values of β_t with no lower bound grow increasingly far from 0, leading to larger and larger squared divergences, which in turn lead to smaller precision κ_o and thus even larger values of β_t in the next iteration. The value of $\{\beta_t : n_t = 0\}$ is essentially a function of κ_o, and the posterior distribution of κ_o is sensitive to outliers in $\boldsymbol{\beta}$, leading to a positive feedback loop.

Sampling values for the observation precision κ_o appears to be difficult and prone to instability. As a result in further examples I treat κ_o as a fixed, known constant. The influence of the value of this constant is explored in the following section.

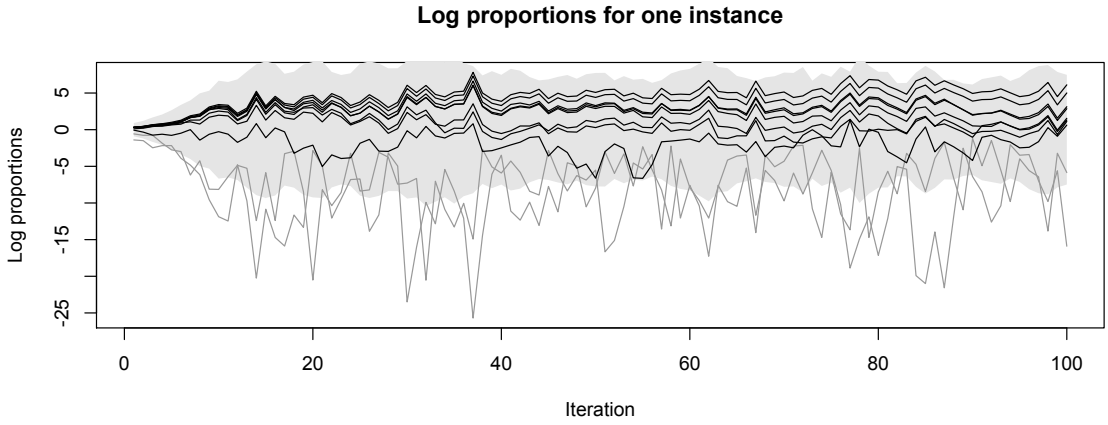

Figure 4.13. Sampling κ_o is unstable. Trace plot of log proportions for one observation with $n_t = 0$ for two dimensions (shown in gray) over 100 iterations of Gibbs sampling. The shaded region shows $1/\sqrt{\kappa_o}$ on each side of zero. Increasingly negative values of β_t for the two dimensions with no lower bound lead to increasingly loose precision, causing κ_o to approach inappropriate values.

4.4 Logistic normal topic modeling

As described in Chapter 3, the logistic normal distribution can be incorporated into a statistical topic model by replacing the Dirichlet priors over the document-specific topic distribution with a logistic normal prior. The Dirichlet prior over topic-word distributions can be similarly replaced, but for the purposes of this work I assume that the topic-specific distributions over words are Dirichlet-distributed. Under such a model, the generative process for a single document (of length N_d) is as follows:

1. Draw a document-specific topic distribution $\boldsymbol{\theta}^{(d)}$ from a logistic normal, as above
2. For each position $n \in \{1, \ldots, N_d\}$
   (a) Draw a topic assignment: $z_n \sim \text{Mult}\left(\boldsymbol{\theta}^{(d)}\right)$
   (b) Draw a word: $w_n \sim \text{Mult}\left(\boldsymbol{\phi}^{(z_n)}\right)$

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ define the characteristics of the model: If $\boldsymbol{\Sigma}$ is a diagonal matrix, the model will exhibit the same covariance characteristics (i.e., uncorrelated topics) as latent Dirichlet allocation. Meanwhile, a non-diagonal covariance matrix will result in a

correlated topic model [9]. Drawing $\boldsymbol{\mu}$ from a first-order dynamic linear model will give rise to a discrete dynamic topic model [8].

4.5 Gibbs sampling using uniform auxiliary variables

Unlike the Dirichlet distribution, the logistic normal distribution is not conjugate to the multinomial distribution. As a result, it is not immediately clear that Gibbs sampling is tractable in such models. Specifically, the posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{\mu})$ does not follow any easily sampled distribution.

It is, however, possible to introduce a set of auxiliary variables that make Gibbs sampling possible. Groenewald and Mokgatlhe [19] introduce such a data augmentation method for logistic regression, similar to work by Albert and Chib [3] for probit regression. Data augmentation methods represent inconvenient distributions as marginalizations of more tractable distributions. We can then iterate over the individual variables, both the original variables and the introduced auxiliary variables, sampling each from a simple conditional distribution.

The distribution $P(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{\mu})$ is equivalent to the likelihood function for a trivial logistic regression model that consists only of intercept terms, with no covariates. We can therefore use a simplified variant of Groenewald and Mokgatlhe's data augmentation method. Details of the application of this sampling scheme are presented by Mimno, Wallach and McCallum [31].

To evaluate the accuracy of the new sampling algorithm for logistic normal topic models, I compare three models that differ only in their prior distribution over topic mixtures:

- A baseline model in which all documents share the same distribution over topics, θ_0. This model treats all z_n variables as iid samples from θ_0, regardless of document membership, in other words, ignoring the grouping of words into documents. This model is expected to do poorly at predicting which words are likely to occur together, because it does not make use of any cooccurrence data in the training set.

- A Dirichlet-multinomial model (LDA), in which the distribution over topics for each document is drawn from a Dirichlet prior: $\theta \sim \mathcal{D}(\alpha)$.

- A simple logistic normal model, with a single mean $\boldsymbol{\mu}$, shared by all documents, and $\kappa = 1$. Topic assignments and logistic normal parameters for this model were inferred by sampling over auxiliary variables. Table 4.1 shows the held-out likelihood and an example topic for each model from the Rexa collection of computer science papers.

Held-out likelihood was calculated using 10-fold cross-validation by sampling unconditionally from the prior over document-specific topic distributions [36]. Each model used 50 topics. The Dirichlet and logistic normal models perform similarly, and substantially better than the baseline θ_0 model. The Dirichlet and logistic normal models find qualitatively similar topics, related to support vector machines, classification, and regression, which the baseline uniform model contains essentially random words.

$\pi(\theta)$	HOL	Example "classification" topics
$\theta_d = \theta_0$	-876602	web map classification problem based ...
$\theta_d \sim \mathcal{D}$	-846729	vector classification support learning regression ...
$\theta_d \sim \mathcal{LN}$	-846746	classification vector selection support regression ...

Table 4.1. Sampling for logistic normal and Dirichlet-multinomial topic models is equivalent. The baseline θ_0 model performs poorly. Theoretically equivalent models with Dirichlet and logistic normal priors over topic mixtures perform much better, assign similar probability to held-out documents (HOL), and learn qualitatively similar topics.

CHAPTER 5

GRAPHICAL PRIORS FOR TOPIC PROPORTIONS

The goal of this work is to produce models that recover the association between words and covariates derived from non-word metadata, mediated through a latent topic space. When there are large numbers of covariates relative to the number of observations, it may become difficult to estimate the parameters representing the association between covariates and topics individually. In many cases, however, there are additional constraints that allow sharing of statistical strength between parameters. As examples I consider time-series models, in which the value of the parameter for a topic at time t is likely to be similar to the value for the parameter at $t-1$, and multi-level regression models, in which sets of parameters are "grouped" such that parameters within a given group are likely to have similar values.

The two varieties of regression-based models described previously, Dirichlet-multinomial topic models and logistic normal topic models, both learn a matrix of parameters, with one parameter per topic-feature pair. Parameters and document-specific feature vectors are combined to produce mean topic distributions. In both cases there is prior distribution over the values of these parameters. This chapter deals with modifying that prior to take into account relationships between parameters using Gaussian Markov random fields.

5.1 Gaussian Markov random fields

In statistical modeling, graphical models are sets of nodes and edges, in which nodes correspond to random variables and the adjacency structure of the edges defines conditional independence relationships. Any multivariate Gaussian distribution over variables $x_1, ..., x_N$ can be represented as a graphical model with N nodes corresponding to $x_1, ..., x_N$. A graphical model that corresponds to a multivariate Gaussian is known as a Gaussian Markov random field [34]. There is a well known relationship between the graph structure of a

GMRF and the precision matrix $P = \Sigma^{-1}$ of a multivariate Gaussian distribution over the same variables: the entry P_{ij} in the matrix is zero if and only if there does not exist an edge between node i and node j in the graph. In other words, if the entry in the precision matrix for nodes i and j is zero, x_i and x_j are conditionally independent given the remaining variables.

This correspondence between graphical models and multivariate Gaussian distributions is useful because it allows us to use choose either representation as is most convenient for particular purposes, knowing that the model is equivalent. For example, it may be simple to define a temporal or hierarchical model in terms of a graph, but then perform inference in that model using the standard matrix-based representation of a multivariate Gaussian.

To define the relationship between a multivariate Gaussian and a GMRF, we must consider the relationship between each pair of adjacent nodes, which is marginally Gaussian. For example, let x_1, x_2, and x_3, which are random variables in $\mathbb{R}$, be the nodes in a linear chain graph, that is, with edges from x_1 to x_2 and x_2 to x_3. This structure asserts that x_1 is independent of x_3 given x_2, and the joint density $p(x_1, x_2, x_3)$ can be represented as a normalized product of pairwise factors $1/Z f(x_1, x_2) f(x_2, x_3)$. Finally, define these factors to be a Gaussian functions: $\exp(-\frac{\kappa}{2}(x_1 - x_2)^2) \exp(-\frac{\kappa}{2}(x_2 - x_3)^2)$. The additional parameter κ is an inverse variance or precision parameter, representing the strength of the relationship between variables. I will frequently refer to these parameters as *edge weights*. Note that this expression is not a proper density: it depends only on the difference between variables, so we can add an arbitrary scalar to all three variables without changing the density. In order to make such a model proper, we must treat at least one variable as a known constant, which results in a valid conditional distribution for the remaining variables given the constant variable. Adding the arguments of the exponents, expanding the squares, and rearranging terms leads to the expression for a multivariate Gaussian distribution with a precision matrix defined by the adjacency structure of the graph and the edge weights.

As a more realistic example, consider a data set consisting of N real-valued observations that are divided into K groups. Let the number of observations in group k be n_k. We wish to infer the mean of each group. A hierarchical Bayesian model for this data might assert that group parameters $\mu_1 ... \mu_K$ are drawn from a normal prior with (fixed, known) mean μ_p

and precision (i.e. inverse variance) κ_p, and the n_k observations $x_{k1}...x_{kn_k}$ are drawn from a normal distribution with mean μ_k and precision κ_o.

$$\begin{aligned} p(\boldsymbol{\mu} \mid \boldsymbol{x}, \mu_p) \quad \propto \quad & \exp\left[\sum_{k=1}^{K} \frac{-\kappa_p}{2}(\mu_k - \mu_p)^2\right] \\ & \times \exp\left[\sum_{k=1}^{K}\sum_{j=1}^{n_k} \frac{-\kappa_o}{2}(x_{kj} - \mu_k)^2\right] \end{aligned} \tag{5.1}$$

Due to the conjugacy of the normal with itself, we can interpret the term from the prior in the resulting density as pseudo-data, equivalent to another observation from each μ_k, but weighted with κ_p rather than κ_o. As before, the distribution $\boldsymbol{\mu} \mid \boldsymbol{x}, \mu_p$ is the conditional distribution of a multivariate normal distribution with a particular precision matrix $P = \Sigma^{-1}$, which can be derived by expanding the quadratic terms and adding up the coefficients for each term. This square matrix can be partitioned into submatrices by grouping the variables into sets $\boldsymbol{\mu}, \mu_p, \boldsymbol{x}$.

$$\begin{bmatrix} \boldsymbol{\mu} \times \boldsymbol{\mu} & \boldsymbol{\mu} \times \mu_p & \boldsymbol{\mu} \times \boldsymbol{x} \\ \mu_p \times \boldsymbol{\mu} & \mu_p \times \mu_p & \mu_p \times \boldsymbol{x} \\ \boldsymbol{x} \times \boldsymbol{\mu} & \boldsymbol{x} \times \mu_p & \boldsymbol{x} \times \boldsymbol{x} \end{bmatrix}$$

The 1×1 submatrix for the prior mean μ_p contains $\sum_k \kappa_p = K\kappa_p$, as the coefficient for the μ_p^2 term in Eq. 5.1. The $K \times 1$ matrix $P_{\boldsymbol{\mu} \times \mu_p}$ contains $-\kappa_p$ in each element, representing the negative cross terms between μ_p and each μ_i. Note that this value is not $-2\kappa_p$ since $P_{\boldsymbol{\mu} \times \mu_p}$ is the transpose of $P_{\mu_p \times \boldsymbol{\mu}}$. The matrix $P_{\mu_p \times \boldsymbol{x}}$ contains zeros, as the prior mean and the observations are independent conditioned on the parameters $\boldsymbol{\mu}$. The $K \times K$ matrix $P_{\boldsymbol{\mu} \times \boldsymbol{\mu}}$ is zero on the off-diagonals (the means are independent given the prior mean) and $n_k\kappa_o$ on the diagonals. The $K \times N$ matrix $P_{\boldsymbol{\mu} \times \boldsymbol{x}}$ represents the relationship between groups and observations: the first n_1 elements of the first row (one for each observation in the first group) contain $-\kappa_o$, followed by $N - n_1$ zeros. Similarly, the second row contains n_2 non-zero elements in the columns for the observations in the second group, and so forth. Finally, the $N \times N$ matrix $P_{\boldsymbol{x} \times \boldsymbol{x}}$ contains κ_o on the diagonals.

Although this matrix by itself does not represent a proper distribution (each row sums to zero by construction, so the matrix is singular and therefore not invertible) the distribution of any subset of variables conditioned on the remaining variables is well-defined: the rows of the submatrix for those variables will sum to greater than zero. This condition is easily satisfied by fixing values for the hyperparameter μ_p, but also holds for the conditional distributions of $\boldsymbol{x}|\boldsymbol{\mu}$ and $\boldsymbol{\mu}|\boldsymbol{x}, \mu_p$. For example, the distribution of $\boldsymbol{\mu}|\boldsymbol{x}, \mu_p \sim \mathcal{N}(P^{-1}_{\boldsymbol{\mu}\times\boldsymbol{\mu}} P_{\boldsymbol{\mu}\times\mu_p\ \boldsymbol{\mu}\times\boldsymbol{x}} \boldsymbol{x}^\star, P^{-1}_{\boldsymbol{\mu}\times\boldsymbol{\mu}})$, where $\boldsymbol{x}^\star$ is a column vector with μ_p as its first element and $\boldsymbol{x}$ as its remaining elements. Since $P_{\boldsymbol{\mu}\times\boldsymbol{\mu}}$ is diagonal (as each μ_k is conditionally independent of the other group means given μ_p), the mean for $\mu_k = \frac{\kappa_p\mu_p + \kappa_o \sum_{j=1}^{n_k} x_{kj}}{\kappa_p + n_k\kappa_o}$, which is a weighted sum of the observations in group k and the prior.

There are two commonly used extensions to this simple grouped data model. First, we may wish to assert additional dependencies between $\boldsymbol{\mu}$ variables. For example, if groups correspond to years, the value for each year is likely to be similar to the year before and the year after. A first order Markov relationship, in which the difference between adjacent values is normally distributed, with weight κ_t between group parameters involves adding the following term to Eq. 5.1: $\exp\left[\sum_{k=1}^{K-1} \frac{-\kappa_t}{2}(\mu_k - \mu_{k+1})^2\right]$. This term can be encoded in the precision matrix as follows. For each pair of adjacent groups k and $k+1$, we set the off-diagonal entries for $\mu_k, \mu_{k+1} = -\kappa_t$ and add κ_t to the diagonal elements for μ_k and μ_{k+1}. Second order Markov relationships, in which the rate of change in differences between adjacent values is normally distributed, can be handled similarly by adding entries in the precision matrix between each triple of adjacent values. Seasonal models and two-dimensional spatial models are also simple to implement.

A second extension relaxes the requirement that each observation depends on only one value. For example, we may wish to model the effect of a year value μ_y and a separate, non-temporal group value μ_g on an observation x from group g in year y. To represent this dependency, we alter the quadratic term involving x in Eq. 5.1 to $\frac{-\kappa_o}{2}(x - (\mu_y + \mu_g))^2$. Expanding this term, we can update the precision matrix to account for the new dependency. As before, we add κ_o to the diagonal term for x, μ_y and μ_g, and $-\kappa_o$ to the off diagonal elements for the pairs x, μ_y and x, μ_g. The only change is that we have an additional dependency between μ_y and μ_g represented by adding *positive* κ_o to the elements for the

pair μ_y and μ_g. This extension allows regression models to be represented as GMRFs: we expect that if the observations are determined mostly by the year and not by their group membership, μ_y will be large relative to μ_g.

5.2 GMRF topic models

The logistic normal topic model includes multivariate normal priors over topic-metadata parameters and document-topic parameters. Due to the fact that there is a one-to-one relationship between GMRFs and multivariate normal distributions, we can define a graph-based prior for a regression topic model by replacing a diagonal covariance matrix between such parameters with a precision matrix P defined according to the construction given above.

A method for MCMC inference follows immediately from this definition. Any subset of the variables in a GMRF are both marginally multivariate normal distributed and conditionally multivariate normal given the remaining variables. We can sample these values tractably given the precision matrix P, as described in chapter two of Rue and Held [34].

Gibbs sampling in GMRF topic models therefore involves alternating between sampling discrete topic indicator variables for all observed tokens in the corpus, sampling auxiliary variables that make the connection between the discrete word vector space and the continuous topic space, and finally sampling continuous GMRF variables, one set for each topic, either individually as univariate normals or in blocks as multivariate normals.

CHAPTER 6

EVALUATION

6.1 Relationships between metadata and words

This thesis is concerned with methods for evaluating the relationship between sets of metadata elements x and sets of words w. There are four important relationships that can be measured:

1. Prediction of words conditioned on metadata $P(\boldsymbol{w}|\boldsymbol{x})$. This thesis is primarily concerned with this conditional setting, in which metadata $\boldsymbol{x}$ is given, and probability density is only defined over the space of combinations of words.

2. Prediction of metadata conditioned on words $P(\boldsymbol{x}|\boldsymbol{w})$. Given $P(\boldsymbol{w}|\boldsymbol{x})$ and a suitable prior distribution over $P(\boldsymbol{x})$, $P(\boldsymbol{x}|\boldsymbol{w})$ can be estimated using Bayes' rule.

3. Evaluation of the similarity between metadata features. Given two metadata elements x_1 and x_2 and probability distributions $P(\boldsymbol{w}|x_1)$ and $P(\boldsymbol{w}|x_2)$, we can evaluate the relationship between the two elements by defining the similarity of the two probability distributions.

4. Evaluation of the relationship between words. Although the estimation of the conditional distribution of one word given another word $P(w_1|w_2)$ is fundamental to density estimation in document collections and topic modeling in particular, as it does not involve metadata it is tangential to this thesis.

6.2 Predicting words: held-out probability

The first relationship $P(\boldsymbol{w}|\boldsymbol{x})$ can be estimated by calculating the marginal probability of held-out documents given metadata for those documents. This calculation is intractable because the number of possible sequences of topic assignments is exponential in the length of

the document. In this work I use Buntine's sequential left-to-right estimator [11], an extension of Wallach's left-to-right estimator [36]. These methods require a matrix of topic-word probablities Φ and an algorithm for sampling from document-specific conditional probabilities over topics that is essentially identical to a Gibbs sampler. All models are trained with 1000 iterations of Gibbs sampling. Topic distributions are estimated from the saved Gibbs state after training. Using multiple samples would likely improve performance, but due to the the large number of models generated in the course of this thesis I chose to save a single sample for each random initialization.

Table 6.1. Table of models and derived models. The first two models are non-topic baselines. The last four models are designed to show the effect of the topic distributions alone, independent of the distribution over topics.

Name	Metadata?	Topics?	Description
Unigram	no	no	Simple unigram language model. All documents share a single distribution over words.
WC	yes	no	"Word-count" baseline. The word distribution for a document is a linear combination of the feature-specific word distributions for document metadata.
LDA	no	yes	Simple latent Dirichlet allocation topic model.
DMR	yes	yes	Dirichlet-multinomial regression topic model.
LDA+DMR	yes	yes	An LDA model with DMR parameters trained post-hoc.
LN	yes	yes	Logistic normal topic model.
AT	yes	yes	Author-topic model.
LU	no	yes	LDA-uniform. This model uses topic-word distributions trained from LDA, but combines them with a simple symmetric Dirichlet prior rather than a learned, asymmetric prior.
DU	no	yes	DMR-uniform. Like LU — topics from DMR, but without the DMR density over topics.
AU	no	yes	Author-topic-uniform. Author-topic topic distributions with the symmetric Dirichlet from LU.
LNU	no	yes	Logistic-normal-uniform. LN topics with a symmetric Dirichlet prior.

6.2.1 Synthetic data

Before considering whether inference methods for different models are effective on real data, I explore how these inference methods behave when given data that is actually generated by different models. For each of the four metadata-enriched generative processes I have presented (the non-topic "word-count" baseline, the DMR topic model, the logistic normal topic model, and the author-topic model), I created a synthetic corpus of 5000 documents. Each corpus had 20 possible features. For each document I randomly select 1-3 features to activate and then generate a 50-word document (from a vocabulary of size 100) under the appropriate generative process. Statistics for these corpora are shown in Table 6.2. The three topic-based models used 30 topics.

Table 6.2. Synthetic corpora. Each 50-word document contains 1-3 randomly selected features. "Feat./Doc" indicates the number of *non-zero* features per document.

Model	Docs	Vocab	Features	Words/Doc	Feat./Doc	Docs/Feat.
WC	5000	100	20	50 ± 0	2.9 ± 0.4	713.8 ± 28.8
DMR	5000	100	20	50 ± 0	2.9 ± 0.4	713.4 ± 27.6
LN	5000	100	20	50 ± 0	2.9 ± 0.4	712.8 ± 22.5
AT	5000	100	20	50 ± 0	2.9 ± 0.4	712.8 ± 29.8

I then created three test-train splits for each corpus, resulting in 12 training sets. I trained all four metadata-based generative models on all 12 training sets, using three independent random initializations each, for a total of 144 models. The topic-based models have two components: a distribution over topics, and topic-word distributions. To isolate the effect of these two components, I also define four "uniform" models, each based on one of the topic-based models (LDA, DMR, LN, AT). These models keep the topic-word distributions from the real model but replace the distribution over topics for each document with a symmetric Dirichlet prior.

In order to explore the sensitivity of models to the number of topics, I also trained models with $T \in \{20, 30, 40\}$. Results are shown in Tables 6.1–6.4.

In all cases, the unigram model has the worst performance. The AT model has the best performance of all models on the corpus generated from the AT process. Similarly, the DMR model has the best performance of all models for the DMR corpus. The DMR model

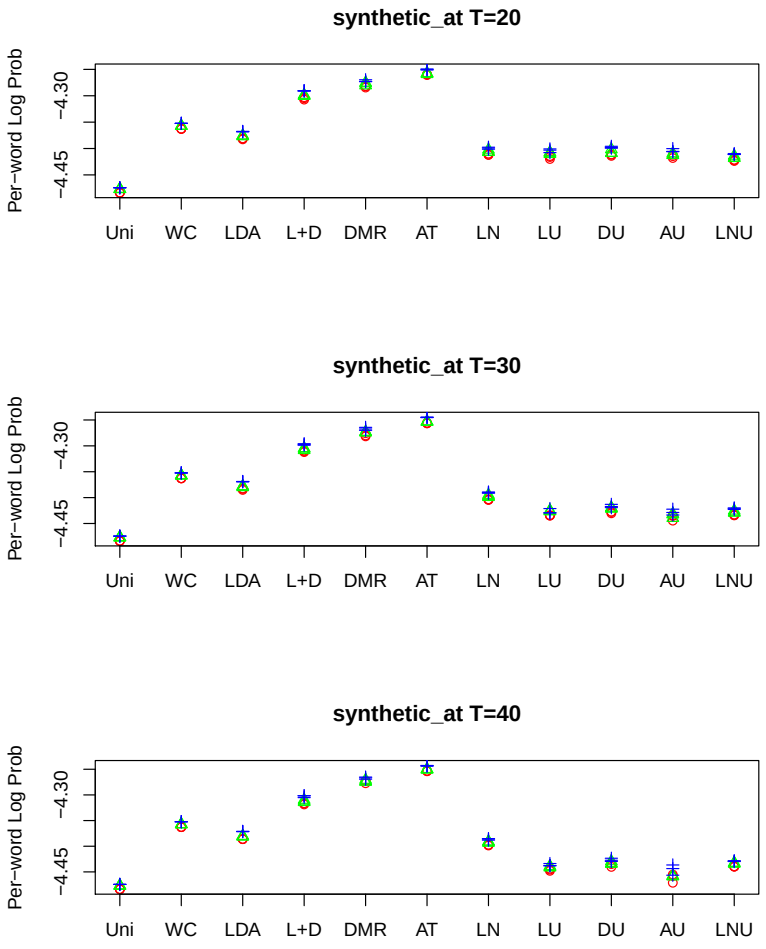

Figure 6.1. Synthetic data from the AT model. Models are a unigram (Uni) language model (no metadata), mixtures of unigram distributions trained with metadata (WC), LDA, LDA plus post-hoc DMR estimation (L+D), DMR, Author-Topic, Logistic normal (LN), and the four topic models evaluated with symmetric Dirichlet priors over topic distributions (LU, DU, AU, LNU). Colors/shapes indicate cross-validation folds, each with three random initializations. The AT model does best on data from its own generative process.

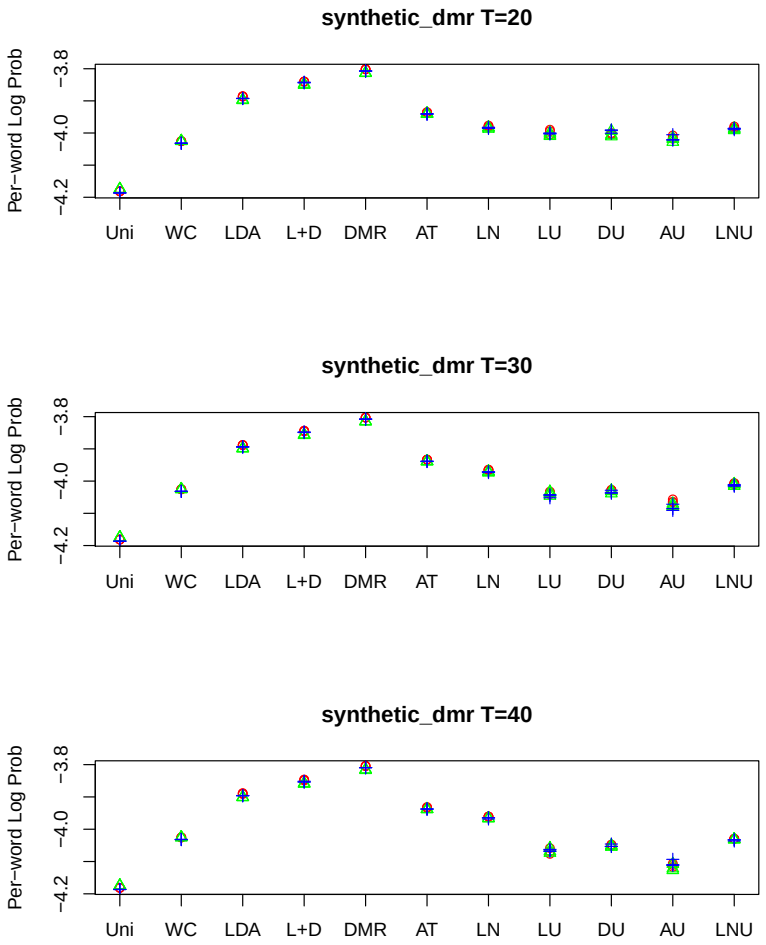

Figure 6.2. Synthetic data from the DMR model. Colors/shapes indicate cross-validation folds, each with three random initializations.

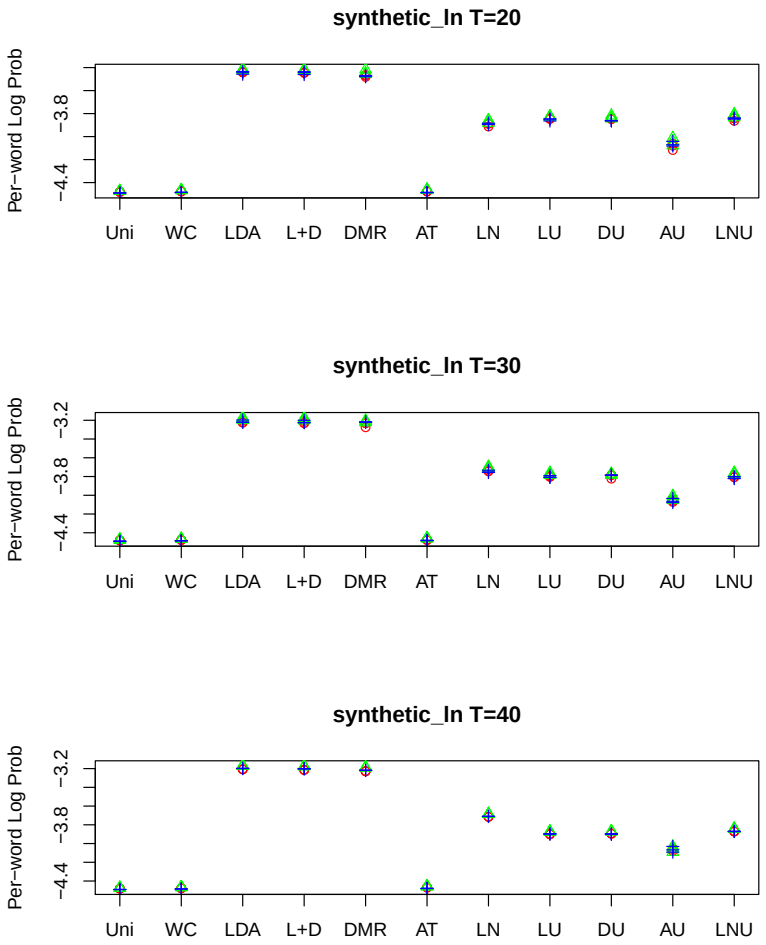

Figure 6.3. Synthetic data from the logistic normal model. Colors/shapes indicate cross-validation folds, each with three random initializations.

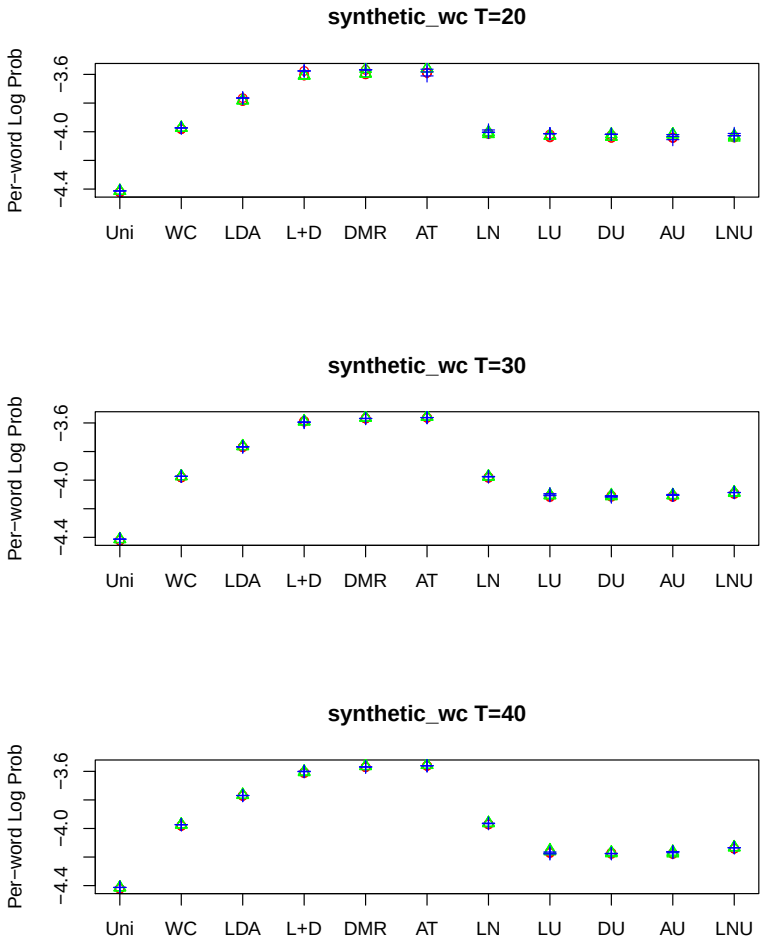

Figure 6.4. Synthetic data from the non-topic "word-count" baseline model. Colors/shapes indicate cross-validation folds, each with three random initializations.

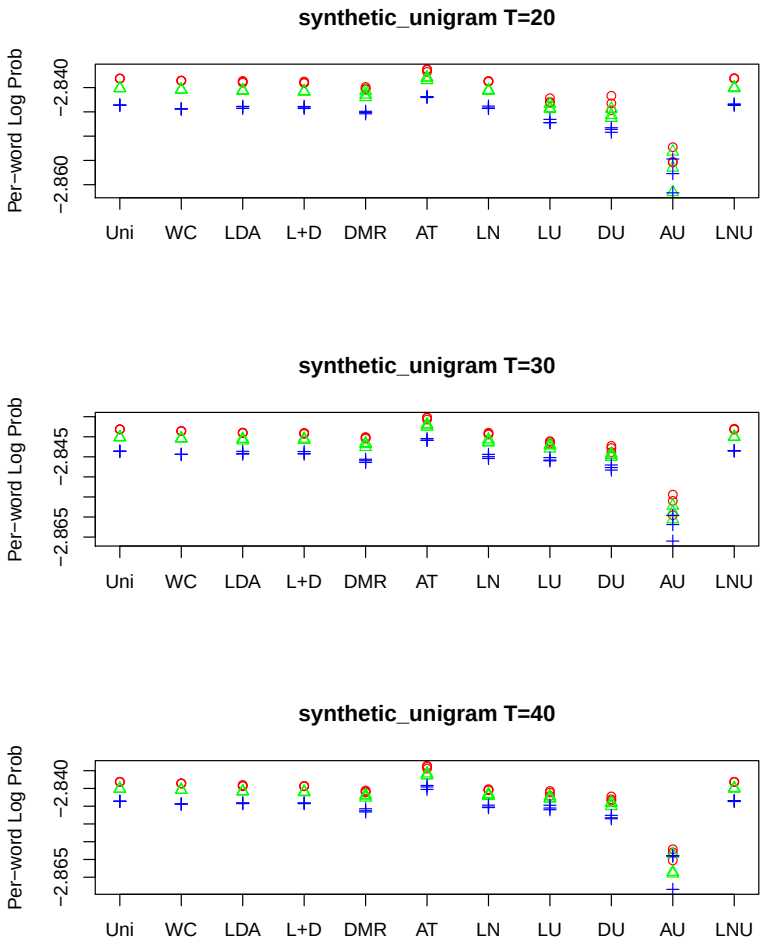

Figure 6.5. Synthetic data from the simple unigram baseline model. Colors/shapes indicate cross-validation folds, each with three random initializations. Predictor variables are purely random and have no relationship to words. All models are essentially indistinguishable, with a range of values much smaller than for all other corpora.

also, however, shows the best held-out probability for the WC corpus and (along with LDA and LDA+DMR) on the LN corpus.

It is important to consider why LN does not perform well on data generated by the LN generative process. As stated in Chapter 2, I cannot directly measure the quality of a model without considering the quality of inference. There is evidence that this phenomenon is due to less effective inference for the LN model.

First we must consider the data set itself. The synthetic LN data appears to have significantly different properties than the AT and DMR corpora. Specifically, the WC baseline and the AT model do as poorly as a simple unigram model, while the LDA and DMR models perform equivalently. These results indicate that the metadata is not informative for this corpus. The generative process for topic distributions in both LN and DMR involve a systematic component (an inner product between features and parameters) and a random component (normally distributed error in the case of LN, log gamma distributed error in the case of DMR). The lack of difference between natural "pairs" such as unigram/WC and LDA/DMR suggests that the systematic component may be being overwhelmed by the random component in the LN generative process. LDA and DMR are able to maintain high performance in this situation, while AT and WC fail. In other words, the data is a good fit for topic models, but the metadata is not informative.

Second we can consider the relative quality of different portions of the learned LN model. As stated in Section 3.4, each model has two components: a set of word distributions and a method for combining those distributions given metadata. The results for the uniform topic models (LU, DU, AU, LNU) helps distinguish the effect of these two components. AU performs poorly compared to the other uniform models. At least some of the poor performance of AT is therefore due to finding poor word distributions, which is reasonable since AT does not distinguish topic distributions at the document level, only through metadata. The full AT model actually performs worse than its word distributions alone, indicating that the non-informative metadata is actively confusing the model. The performance of LN falls between the uniform topic models and the Dirichlet-based models (LDA, LDA+DMR, DMR). The topic distributions of LN by themselves (LNU) are as good as those of LDA

and DMR, so the difference is likely to be due to the ability of LN to adapt to the topic distribution of the individual documents.

In the WC corpus, it is surprising that the WC model has lower predictive ability than the topic models. There are several possible explanations for this effect. First, it is possible that the topic models simply always return better held-out probability scores than WC. To show that this is not the case, Figure 6.5 shows predictive performance for a corpus such that the content of every document is drawn from a single shared multinomial distribution, and with predictive values that have no effect on word choice. In that case, all models perform very similarly, with AU the only small exception. A second possibility is that the WC generative process produces data that is in fact well matched with topic model assumptions: clusters of words frequently appear together. The difference in predictive ability on the WC corpus may therefore come from the method used to train the WC model rather than the model itself. By simply counting the number of words associated with metadata features, the model does not learn to distinguish the precise effect of specific features. In contrast, the topic models can learn specific distributions over words that happen to be associated with particular metadata features.

The "uniform" versions of the four topic models produce comparable performance in most cases, indicating that the difference in held-out probability between the full models is not due to finding better topic dimensions. The exception is the uniform version of the AT model, which does relatively poorly in the DMR and LN corpora, especially at $T = 40$.

6.2.2 Possible problems with logistic normal models

There is a substantial difference in observed performance between logistic normal and DMR topic models. Although they are theoretically similar, as shown in Chapter 4, logistic normal models diverge significantly from Dirichlet-based models when observations become sparse. The observed difference in performance is therefore not surprising, but in order to rule out other conditions, I also tried several modifications to the inference procedure:

DMR training alternates between sampling topic assignments and maximizing (rather than sampling) topic parameters. It did not appear to make a difference whether LN parameters were sampled or maximized (as in DMR).

The held-out probability results shown are based on parameters averaged over several saved samples. For DMR, I evaluate using the values of parameters at the end of training. Results for LN were slightly worse when using the sampled state for topic-metadata parameters at the last iteration than for averaging over several states.

6.2.3 Research administration: NIH grants

I now consider real document collections, beginning with a set of biomedical research grant abstracts. The National Institutes of Health funds a significant portion of all biomedical research. It is an extremely complicated organization, consisting of many institutes that operate with varying degrees of independence. Funding decisions are made through a complicated interaction between institutes, programs, and study sections. Understanding this ecosystem is vital both for grant applicants, who need to know where to send a proposal, and the funding agency itself, which needs to watch for duplication of work and opportunities for collaboration.

In this section I consider a database of NIH grant applications that were accepted for funding in 2009. Each grant is annotated with the following metadata elements:

- the institute that funded the grant
- an activity code (distinguishing, for example, research or training programs)
- an ID (PCC code) for the program officer
- the funding opportunity announcement (FOA) title
- the study section that reviewed the proposal
- the integrated review group (IRG) that contains the study section
- reporting categories from a controlled vocabulary

Descriptive statistics for this corpus are shown in Table 7.1. This corpus illustrates the case where there are large numbers of metadata features for each document, most of which are rare.

Table 6.3. NIH: Many features per document. Abstracts are fairly short, with many features. Most features are rare, so variance in documents per feature is high.

Docs	Vocab	Features	Words/Doc	Feat./Doc	Docs/Feat.
43362	41440	2848	164.2 ± 46.5	10.7 ± 3.4	163.6 ± 838.0

I repeated the held-out probability experiments on the NIH corpus. Results for three test/train splits, each with three random initializations, are shown in Figure 6.6. Values are very consistent within models, so points are not distinguished. The relative performances of the different models are similar to the results from the Rexa data set. DMR and LDA with DMR post hoc perform best overall, with a slight advantage for the iterative model. AT outperforms the word-count baseline, but is not competitive with LDA.

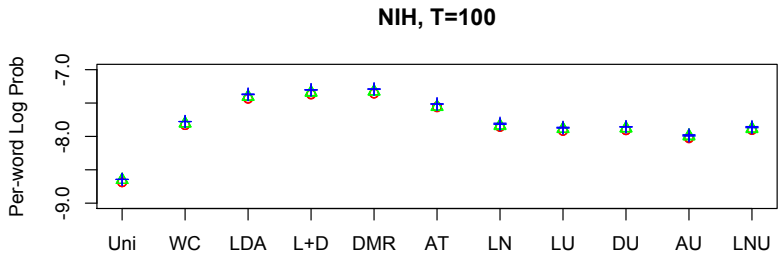

Figure 6.6. DMR has the best held-out performance. Results are very consistent across test sets and random initializations. This corpus has a challenging metadata environment with many, often sparsely represented elements. The corpus is particularly difficult for logistic normal models, which perform worse than the WC baseline.

6.2.4 Scientific publications: the Rexa corpus

I next consider a corpus of scientific publications from journals and conferences in artificial intelligence. All papers are taken from the Rexa database.[1] Corpus statistics are shown in Table 6.4. Metadata features are simpler in the Rexa corpus than in the NIH corpus. This corpus is representative of the case where most documents are assigned to a small number of overlapping categories, each of which is well-represented in the corpus. Each

[1] `http://rexa.info`

document is annotated with a publication venue (journal or conference) and a publication year. I encoded these values as indicator functions, with one function per year and one per venue. In addition, I added a single "intercept" or "default" feature that is always set to 1.

Table 6.4. Rexa: Many documents per feature. Many documents are titles alone, so documents tend to be short. All documents have exactly three non-zero features each (indicators for year, venue, and intercept). Most features are common, but with high variance.

Docs	Vocab	Features	Words/Doc	Feat./Doc	Docs/Feat.
45802	32512	51	23.1 ± 41	3 ± 0	2,694.2 ± 6,294.8

I created three 90%/10% test-train splits for cross validation to reduce the effect of the specific held-out documents. For each test-train split I ran each model with three random initializations. The models evaluated include LDA with optimized $\boldsymbol{\alpha}$ hyperparameters, DMR, LDA with one round of post-hoc DMR training, Author-Topic, and a logistic normal model. As stated above, held-out probability estimation has two inputs: fixed topic-word distributions and a distribution over topic distributions, which may or may not depend on metadata. In order to distinguish the contribution of the topic-word distributions from the contribution of the distribution over topics, I also evaluated the four topic models using their topic-word distributions combined with a symmetric Dirichlet prior ($\alpha_t = 0.1\ \forall t$). Finally, I also evaluated two non-topic baselines, one with a single probability distribution over words (Unigram) and another that is a mixture of metadata-specific empirical distributions (word-count).

Results are shown in Figure 6.7 for models with $T \in \{50, 75\}$. Typical per-word log-likelihoods ($\log p(D)/|D|$, related to the geometric mean) are around -7.0, which corresponds to a probability of 1 in 1000, which is reasonably good given that the vocabulary size is in the tens of thousands. The unigram language model, essentially a single distribution over words, has the worst performance. LDA, with no access to metadata, does better than WC, indicating that the regularities identified in the latent topic space are helpful in making predictions. Of the metadata-enriched topic models, the two variants of DMR perform best. Iteratively alternating between optimizing DMR parameters and resampling

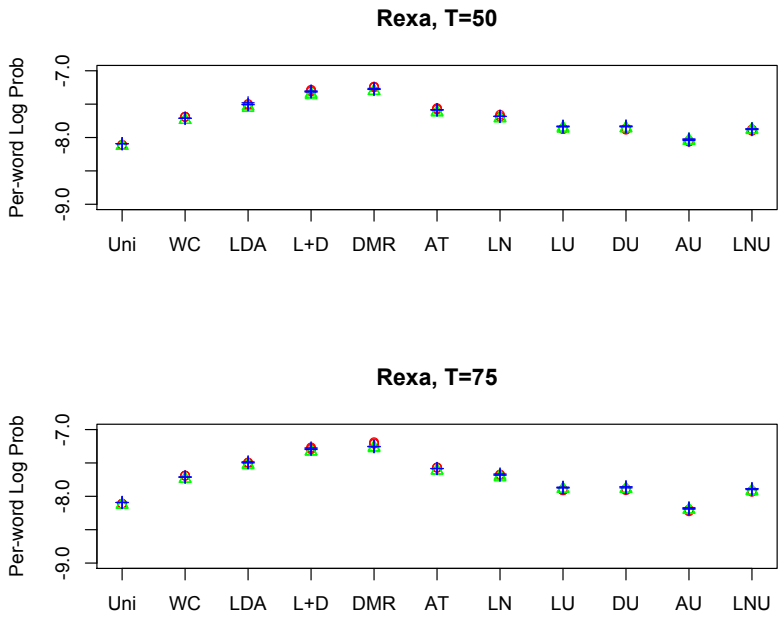

Figure 6.7. DMR has the best held-out performance. Models are a unigram (Uni) language model (no metadata), mixtures of unigram distributions trained with metadata (WC), LDA, LDA plus post-hoc DMR estimation (L+D), DMR, Author-Topic, Logistic normal (LN), and the four topic models evaluated with symmetric Dirichlet priors over topic distributions (LU, DU, AU, LNU). Colors/shapes indicate cross-validation folds, each with three random initializations.

topics performs best overall, but a single round of post-hoc is sufficient to account for much of the gap between DMR and LDA.

In this corpus, both AT and LN perform poorly. The evaluations of topics alone (with a common symmetric Dirichlet prior over topic proportions) show that these models are doing badly for different reasons. LDA and DMR topics alone (LU, DU) have similar held-out probabilities, indicating that most of the benefit of DMR comes from better estimation of topic distributions given metadata, and that LDA benefits substantially from optimized hyperparameters. Logistic normal topics alone are slightly worse than LDA topics alone, indicating that most of LN's lack of performance is due to difficulty in estimating topic distributions, most likely as a result of a more complicated sampling procedure that involves realization of the document-level topic distribution. AT topics alone, on the other hand, produce much poorer results than LDA, indicating that this model is not finding a good latent topic space. The characteristics of this corpus, with many documents per metadata element, are not well-suited for Author-Topic, which cannot take advantage of document-specific contextual patterns.

6.3 Predicting metadata from text

Evaluating the second relationship $P(\boldsymbol{x}|\boldsymbol{w})$ is challenging for the models studied, because they are all defined conditionally on metadata. As a result, none of these models has an explicit density over combinations of metadata features. This property is an advantage for many applications, because it allows us to use high-dimensional feature spaces with complex, non-independent interactions (e.g. conjunction features) without increasing the complexity of the model.

One simple way to evaluate $P(\boldsymbol{x}|\boldsymbol{w})$ is to estimate the probability of a held-out document $\boldsymbol{w}$ under each possible metadata configuration such that exactly one feature is "on". For example, if there are three binary metadata features x_1, x_2, and x_3, this process would involve calculating $P(\boldsymbol{w}|x_1 = 1, x_2 = 0, x_3 = 0)$, $P(\boldsymbol{w}|x_1 = 0, x_2 = 1, x_3 = 0)$, and $P(\boldsymbol{w}|x_1 = 0, x_2 = 0, x_3 = 1)$. I can then rank features x_1, x_2, and x_3 by their probability of generating $\boldsymbol{w}$ and compare this ranking to the actual features that are non-zero for the held-out document. This evaluation avoids searching over an exponential number

of metadata configurations to find the maximum likelihood solution, but at the cost of penalizing features that by themselves do not have high likelihood but might combine with other features to obtain high likelihood.

An alternative to considering each feature individually would be to perform an $A^\star$ or stochastic search for feature combinations, but such procedures are beyond the scope of this thesis. Depending on the complexity of the metadata features, such a search may not be likely to find optimal combinations. In addition, estimating held-out probability is time-consuming, especially when multiplied over many models, data sets, cross-validation folds, and random initializations. If these models are to be used as a method for inferring sets of metadata features, more work needs to be done to make such search and evaluation of held-out probability more efficient. Rather, the results presented in this section measure the ability of models to identify the "meaning" (with respect to words or topics) of specific metadata features.

A natural evaluation is to rank all features for a held-out document and record the rank of the true features for that document. A good predictive model should put the true features at low ranks. As noted above, I did not consider sets of features in combination, only individual features. As a result, we are asking each feature to "explain" all the words in a document when it may in fact only account for a small proportion of those words. It is therefore possible that a single "incorrect" feature could legitimately have higher likelihood than all "correct" features taken individually, for example if the true features are very specific and the incorrect feature is very general.

In order to make this evaluation more challenging, I created partly synthetic feature sets by adding noise features to the true features. These "distractor" features are either 1 or 0 with some probability p. I varied the number of distractors and the probability p by corpus. Estimating a word or topic distribution for such random features will lead to distributions very close to the overall distribution of the corpus.

The ability of models to predict true features and ignore distractor features varies considerably between models and between data sets. In general, the WC baseline does a good job of ranking correct features, but also ranks distractors towards the beginning. The post-

hoc DMR model (LDA+DMR) does a good job as well, with slightly worse performance on real features but better discrimination for noise features.

For a first example, I evaluate the ability of different models to rank features for previously unseen documents in a subset of a larger collection of opinions from the U.S. Supreme Court. This corpus will be discussed in more detail in Section 7.1. Each "document" is a paragraph from a Supreme Court opinion. For this example I use a very simple metadata structure: there is one non-zero feature per document, which is an indicator for the case ("docket") that the paragraph is drawn from. These case IDs are chosen to strongly indicate a specific set of themes, i.e. the subject of the case. Details are given in Table 6.5.

Table 6.5. Supreme Court with `case` features. Each document is a paragraph from a Supreme Court decision. The single metadata feature is a case ID (docket number).

Docs	Vocab	Features	Words/Doc	Feat./Doc	Docs/Feat.
12317	10330	76	39.6 ± 30.6	1 ± 0	162.1 ± 140.3

Table 6.6 shows feature ranking results for this corpus. The first column shows precision at rank 1, or the proportion of held-out documents for which the first ranked feature is a real (non-distractor) feature of the document. The second column shows the mean and standard deviation of the rank of the true features (lower numbers are better). The third column shows the same statistics for the distractor features (higher numbers are better). The final column shows the difference between the mean rank of the real features and the mean rank of the distractor features. In this corpus the WC, AT and DMR models have similar performance for real features, but differ in their handling of distractor features. The WC model has the least ability to distinguish between real and noise features. The AT model has the best discriminative ability: it correctly selects true features and ignores noise features. DMR+LDA has a worse average rank than DMR for true features but better average rank for noise features. LN shows poor performance overall, but appears to distinguish between real and noise features.

For the next example, I use a slightly more complicated metadata set from the Rexa database, where features are indicators for year and publication venue (Table 6.7). The AT

Table 6.6. Ranked feature results for 76 Supreme Court cases (divided into paragraphs) with case number as the single true predictor, plus 3 noise features. In this corpus, with a small number of highly meaningful predictors, the AT model is able to distinguish true features from distractors.

Model	P@1	Feat. Rank	Distr. Rank	Diff.
WC	74.1%	2.8 ± 6.6	4.0 ± 3.9	1.3
AT	75.4%	2.8 ± 7.4	21.0 ± 20.0	18.2
LDA+DMR	59.1%	4.1 ± 9.1	16.7 ± 14.6	12.6
DMR	71.0%	2.8 ± 6.8	8.9 ± 7.5	6.1
LN	14.7%	20.9 ± 20.7	38.0 ± 22.6	17.1

model does much more poorly, with distractor features consistently ranked *ahead* of true features. LDA+DMR again performs well.

Table 6.7. Ranked feature results for Rexa, with 51 real features and 3 noise features.

Model	P@1	Feat. Rank	Distr. Rank	Diff.
WC	40.7%	11.4 ± 10.1	13.0 ± 8.0	1.6
AT	18.8%	18.0 ± 11.9	9.0 ± 6.8	-9.0
LDA+DMR	41.3%	15.8 ± 14.6	20.6 ± 8.6	4.7
DMR	39.7%	15.6 ± 13.3	16.7 ± 6.8	1.2
LN	19.1%	22.1 ± 15.7	26.3 ± 14.7	4.2

Finally, I produced feature rankings for the NIH grants corpus, which has a much more complicated metadata structure than that for the Supreme Court and Rexa corpora. Results are shown in Table 6.8. This collection has two orders of magnitude more features than the previous two examples, and many more non-zero features per document. WC has significantly better performance on true features than any of the topic models, but also consistently ranks noise features along with true features. AT has surprisingly poor results for true features, much worse than random chance.[2] The one positive observation about the performance of AT in this case is that it consistently (i.e. with low variance) ranks distractor features lower than all real features. DMR, although not competitive with WC in ranking true features, has relatively strong ability to distinguish between real and noise features while still ranking a true feature at the top one in nine times.

[2]I have reviewed this result, and I have not been able to detect a bug.

Table 6.8. Ranked feature results for NIH grants with 2848 real features and 10 noise features. WC has the best performance for true features, but is fooled by noise features. AT performance appears worse than random. DMR shows the best differentiation between real and noise features.

Model	P@1	Feat. Rank	Distr. Rank	Diff.
WC	66.5%	40.2 ± 103.0	39.4 ± 56.6	-0.8
AT	0.0%	2508.9 ± 492.9	2734.9 ± 59.6	226.0
LDA+DMR	8.1%	1074.4 ± 842.3	1363.9 ± 831.5	289.6
DMR	12.8%	964.9 ± 839.2	1390.3 ± 817.0	425.4
LN	0.7%	1249.4 ± 831.6	1379.1 ± 827.8	129.6

These results illustrate the effect of different types of metadata elements on the different models. When there is a single active feature per document that strongly indicates the content of a document, as in the `case` features for the Supreme Court data, AT does well. When there are multiple features that have a more diffuse relationship to document contents, such as the `year` and `publication venue` features in the Rexa corpus, AT begins to break down. One difference between AT and DMR is that AT represents metadata using multinomial distributions over topics, where DMR uses parameters that are combined into a Dirichlet prior. The multinomial distribution has less flexibility to accommodate variance in observations than the Dirichlet. The AT model, therefore, cannot assign high probability to documents that do not closely match the existing topic distribution for a given metadata element, while the Dirichlet distribution can provide a "softer" preference for some topics over others.

6.4 Similarity of clusterings

In this section I consider whether the different methods of accounting for the effect of metadata on topic mixing proportions result in the discovery of different latent components (ie topics). If the choice of LDA, Author-Topic, DMR or LN makes no difference, topics inferred from different random initializations of the same method should be as similar to each other as they are to topics inferred from other methods.

I compared topic assignment vectors $\boldsymbol{z}$ between eight models: LDA, Author-Topic, DMR, and a logistic normal model with $\kappa_o \in \{0.1, 0.2, 0.4, 0.8, 1.6\}$. As the value of topic indicators is not meaningful across models, I used a method for comparing clusterings,

Meilă's variation of information (VI) metric [28]. I ran five random initializations of each model on the Rexa dataset with $T = 100$ and saved the Gibbs sampling state after 1000 iterations. I then computed a matrix of pairwise VI distances. I projected this matrix into two-dimensional space using multidimensional scaling[3]. The resulting embedding is shown in Figure 6.8.

Figure 6.8. LN topics move away from LDA and DMR as κ_o increases. Multidimensional scaling of clustering distances between five random initializations of each of eight models: LDA, green inverted triangles; DMR, blue diamonds; AT, red triangles; LN, circles shading from black to light gray, with $\kappa_o \in \{0.1, 0.2, 0.4, 0.8, 1.6\}$ respectively. Distances between points are related to variation of information distance: the axes themselves are not meaningful.

All models produce clusters of topic assignment vectors that are closer to each other than any other model's topic assignments. LN with $\kappa_o = 0.1$ is closest to DMR and LDA. AT is far from either set of topic assignments, and by far the most consistent across random initializations. These results indicate that the choice of model has a significant and consistent effect on the assignment of words to topics.

6.5 Effect of infrequent words

In some probabilistic models, rare events can have a disproportionate effect on overall model log likelihood. For example, if data sampled from a heavy-tailed Student t distribution with $df = 1$ is evaluated under a standard normal log likelihood function, the overall log probability of the data set is likely to be dominated by one or two extreme data points.

[3] using the R command `cmdscale`.

The variance of estimates of log probability of such samples can consequently be large, even for large sample sizes.

Held-out probability in topic models is essentially a sum of log-probabilities of individual words (in fact the estimator I use is the sum of log probabilities *conditioned* on the previous words in a document). It is well-known that words follow a heavy-tailed "Zipfian" distribution, such that the probability of most words in the vocabulary is small, but the probability that an individual word token is rare is in fact rather high. It is therefore important to consider two questions:

1. Are the values of held-out probability and similarity of clusterings that I report earlier in this chapter dominated by rare words?

2. Are the relative performance rankings of different models affected by rare words?

The answer to both questions appears to be no. I first consider differences in held-out likelihood results and then move on to the similarity of clusterings based on variation of information.

A standard measure of the rareness of words, used in information retrieval, is the *inverse document frequency* (IDF) of a term. This measure is defined as

$$IDF(w) \quad = \quad \log\left(\frac{D}{DF(w)}\right) \tag{6.1}$$

where D is the number of documents in a corpus and $DF(w)$ is the number of documents that contain at least one instance of word w. A word that occurs in 5% of all documents (relatively frequent) will have IDF $\log(20) = 3.00$. A rare word that occurs in one out of 10,000 documents will have IDF 9.24.

6.5.1 Held-out likelihood

I used Buntine's variant of Wallach's left-to-right estimator to evaluate the marginal probability of each token in the held-out documents given all previous tokens in a given document. This process results in a marginal log probability for each token in the test set. The probability of each token depends on its order in the document, but I hold the order fixed across different models.

I ordered each word in a test set from the Rexa corpus according to its IDF, from the most common words to the rarest. I then calculated the cumulative log probability for each model at each IDF value by summing over the marginal token probabilities of each token. Figure 6.9 shows the cumulative log likelihood of a held-out test set under five models. The steepest decline is between IDFs of roughly 3.5 and 6, corresponding to words that occur in between 3% (1200 documents) and 0.2% (100 documents). Words with IDF less than or equal to 6.0 make up only 4.6% of the vocabulary — these words are infrequent, but they are not rare. The most infrequent words, at IDF values greater than 8.0, contribute relatively little to the overall log probability of the held-out test set. In particular, they do not change the relative position of the models.

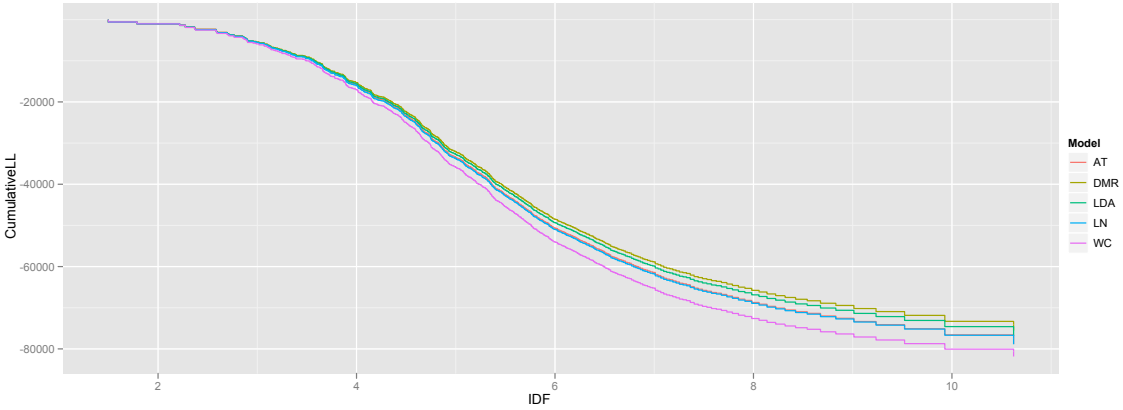

Figure 6.9. Rexa: Held-out likelihood is dominated by moderately frequent words. Each line shows the cumulative marginal log probability of the model as words are added in order of increasing IDF. The AT model closely tracks the LN model in this plot, and is thus not easily visible.

6.5.2 Similarity of clusterings

The previous section indicates that there is a significant difference between the latent topic spaces found by different models. I found that even when models are initialized randomly, after a fixed number of iterations of Gibbs sampling the resulting pattern of topic assignments is more similar to other topic assignment patterns from the same model than to those of any other model. Random initializations from, for example, LDA, consistently lead to topic assignment patterns that are more similar to other LDA models than they are

to topic assignment patterns from DMR or AT. Again, we would like to know whether this difference is the result of rare words. It is possible that if we ignore the most infrequent words, differences between models will disappear.

As in previous sections, I calculate the similarity of clusterings — that is, assignments of a fixed set of tokens to some set of topics — by measuring variation of information. In order to measure the effect of rare words, I evaluate variation of information for varying IDF cutoffs. For example, at a cutoff of 3.0, I remove all words with IDF greater than 3.0 from both clusterings and measure the distance between the topic assignment patterns of the remaining words.

Results for the Rexa corpus are shown in Figure 6.10. The plot shows the variation of information between a single LDA model with $T = 50$ and four other models (AT, DMR, LN, and another random initialization of LDA) at five different IDF cutoffs. These IDF values $\{3.0, 5.0, 7.0, 9.0, 11.0\}$ correspond to 0.06%, 1.55%, 8.62%, 26.74% and 100% of the vocabulary, respectively. There is only one case where restricting the vocabulary changes the ordering of models: AT matches the other LDA model and is closer than DMR or LN only for the IDF ≤ 3.0 cutoff, which considers a tiny fraction of the vocabulary. This result indicates that the allocation of the most frequent words in the corpus — words occurring in more than 5% of all documents — is similar in LDA and AT, but the allocation of all other words is substantially different.

6.6 Computational cost

The models evaluated in this chapter have substantially different computational cost, which must be taken into account when considering the empirical differences in performance demonstrated above. The following is a discussion of the relative costs of each model as a function of the size of the corpus and the quantity of metadata.

Experimental results showing 10 random initializations of each topic model on two corpora are shown in Figures 6.11 and 6.12. Results for each model are very consistent across runs, so I have added random jitter in the x-axis to show all points clearly.

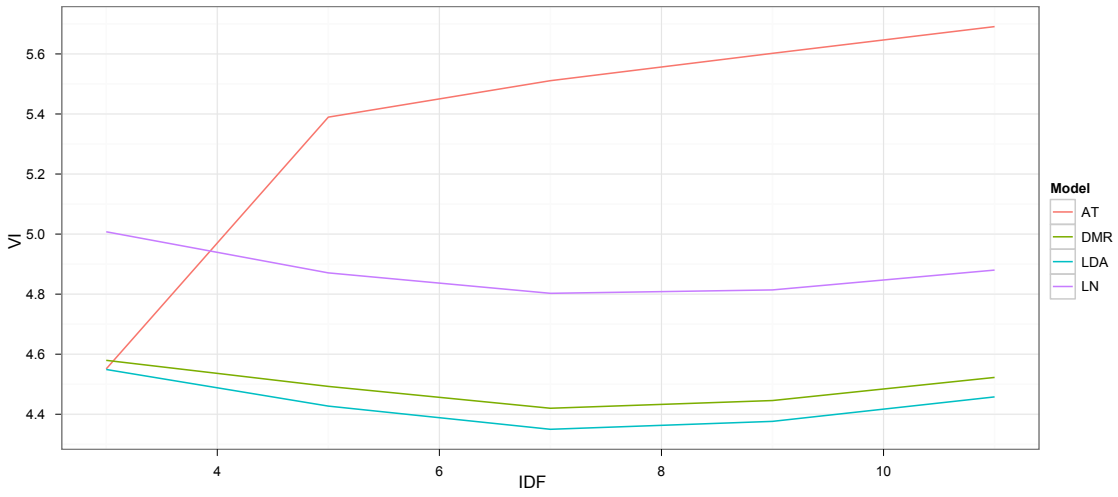

Figure 6.10. Rexa: Differences between models are consistent for all but the most frequent words. These plots show the variation of information between an LDA model and four other models, including another LDA model.

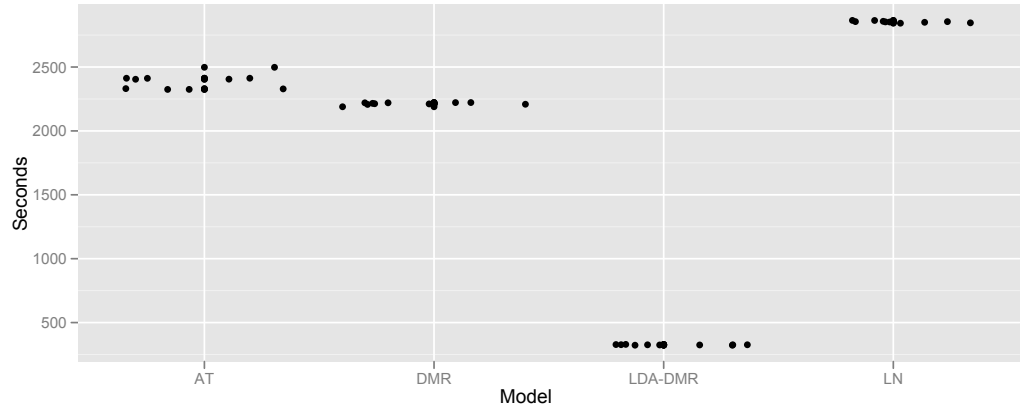

Figure 6.11. Rexa: Models are well-matched when there are few features per document. The y-axis shows the number of seconds needed to perform 1000 iterations of sampling. The x-axis includes random jitter to reduce overplotting. Each document has three features (publication venue, publication year, and a constant "intercept" parameter representing a background distribution.

6.6.1 Word-count baseline.

Estimating parameters for this model involves iterating over all training documents once to collect the frequency that words appear with each metadata element, weighted by the number of non-zero features for each document. This model is by far the most computationally efficient method considered.

6.6.2 LDA+DMR.

This model involves training a standard topic model and then performing a single numerical optimization step to learn metadata-topic parameters. The cost is therefore the cost of sampling an LDA model with T topics and then optimizing parameters. As a result of my work in efficient sampling, the first step can be accomplished in time roughly logarithmic in T and linear in the number of word tokens in the corpus [39]. This sampling process usually dominates the time taken for a single optimization of DMR parameters.

6.6.3 DMR.

The full DMR topic model differs from the LDA+DMR model in that it performs many rounds of numerical optimization on metadata-topic parameters rather than just one. The difference, however, is not strictly linear in the number of optimization rounds, as the time taken for L-BFGS to converge decreases after each optimization round, as the parameters start off closer to their maximum. On the other hand, sampling topic assignments is slightly slower for than for pure LDA: as every document has its own Dirichlet prior, certain data structures must be recomputed from scratch for every document, rather than being lazily updated for the small number of topics that actually occur in each document.

6.6.4 LN.

LN is similar to DMR in that it alternates between sampling topic indicators and updating metadata-topic parameters. The LN sampling scheme, however, is more complicated than in Dirichlet-based topic models, resulting in roughly twice the computational time.

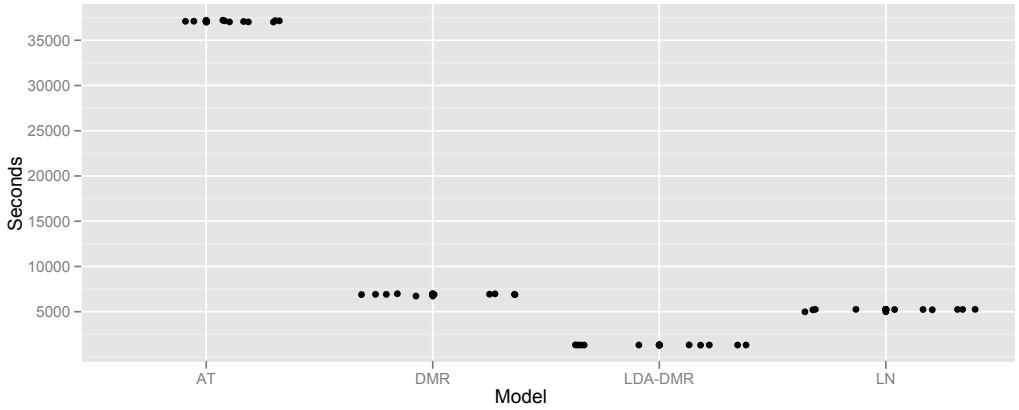

Figure 6.12. NIH: AT performs poorly when there are many features per document. The NIH corpus includes on average 10 features per document, resulting in substantially poorer performance for AT.

6.6.5 Author-topic.

Unlike LN and DMR, the author-topic model must sample both a topic and a metadata assignment for every word token. Thus, rather than an additive cost (number of tokens $\times$ number of topics + number of metadata elements $\times$ number of topics), the cost becomes multiplicative (number of tokens $\times$ number of topics $\times$ number of metadata elements). In addition, sparse computational methods [39] are more difficult to implement than in LDA. This multiplicative scaling leads to significantly longer computational times for corpora with many non-zero features per document. As shown in Figure 6.12, the number of features available for sampling has a strong effect on the efficiency of sampling.

CHAPTER 7

APPLICATIONS

In the previous chapter I focused on evaluations that are independent of the content of the collection. In this chapter I examine several case studies, in which outside knowledge suggests ways that specific metadata elements should interact with text. The observed effects of metadata-enriched topic models follow these predictions.

7.1 Predicting authorship of Supreme Court opinions

The U.S. Supreme court decides roughly 100 cases per year. For each decision, a justice is selected to write the majority opinion. The content of this opinion determines how law is interpreted, and has an effect beyond the simple binary vote on a case. The choice of author is therefore significant. This selection is made by the chief justice if he is in the majority, and otherwise by the senior associate justice in the majority. Political scientists studying the U.S. Supreme Court have compiled substantial evidence through interviews and statistical methods that the author of an opinion is not chosen randomly [24, 25]. Maltzman [24] identifies several factors that contribute to the decision to select an author for a decision. Some of these factors are always relevant, and do not depend, for example, on the "closeness" of the case. These include each justice's current workload and the expertise of a given justice in the specific area of a case. Other factors are only salient in particular situations. If the case is "significant" (Maltzman defines this condition heuristically as a case reported on the front page of the New York Times), authorships are frequently given to ideological allies of the chief. When a case is close (with a majority size of five or six), there is a preference for more ideologically distant justices, in which case the authorship may be offered to solidify a slim majority. However, Bonneau et al. have also suggested that there is *not* a clear link between the "median" justice and authorship decisions [10].

I constructed a corpus of supreme court opinions by downloading documents from the Legal Information Institute (LII) at Cornell.[1] This collection covers the last 20 years of supreme court decisions, and includes opinions, concurrences, and dissents. Each case is identified by a docket number that includes the last two digits of the term during which the case was filed followed by a case number. For example, the case UNION PACIFIC RAILROAD COMPANY v. REGAL-BELOIT CORPORATION et al., filed in 2008, has docket 08-1554. The HTML-formatted opinions contain authorship information and links to cited cases.

The LII corpus includes only documents. In order to gather voting records I downloaded the "justice-centered" spreadsheet from the Supreme Court Database (SCDB) at Washington University, St. Louis.[2] This database contains records of all cases going back to the 1940s. It includes the vote of each justice, the chief justice (either Rehnquist or Roberts for cases since 1990), and information about the geographic origin of the case and about the authorship of opinions, concurrences and dissents.

Table 7.1. Supreme Court: `disagreement` features. Each opinion is broken up into paragraphs. Most decisions have large majorities: only 3.3 justices on average disagree with the chief justice.

Docs	Vocab	Features	Words/Doc	Feat./Doc	Docs/Feat.
79457	24432	21	41.6 ± 29.7	3.3 ± 1.7	12,330.6 ± 9,738.2

Now consider the task of of predicting the author of an opinion, given the set of justices that voted with the majority. The work in political science summarized above suggests that a measure of ideological distance should have a significant predictive power *in close cases*, but not for unanimous cases, while a measure of expertise in the particular areas of a case should have good predictive power regardless of the size of the majority. In this section I consider the predictive effect of two sets of metadata derived from Supreme Court authorship and voting data that attempt to measure ideological distance and expertise, respectively.

[1] http://www.law.cornell.edu/supct/

[2] http://scdb.wustl.edu/

1. `Disagreement` **features**. For these features, I use one indicator variable to represent the fact that an associate justice votes against the chief justice on a case. For example, if Justice Scalia votes against Chief Justice Rehnquist, I set the feature `AGAINST_Scalia_Rehnquist` = 1, and 0 otherwise. This type of variable does not depend on the polarity of the vote, so it is independent of whether the chief justice is in the majority or the minority. This set of features is designed to measure the particular topics on which the chief justice and a given associate justice disagree.

2. `Expertise` **features**. These features indicate that a given associate justice is the author of the opinion for a case.

I created 19 test-train splits by holding out cases based on their year of filing (the first two digits of the docket ID). I then trained three random initializations of each topic model (AT, DMR, LN, LDA) on each of the 19 training sets using just disagreement features and then again using just expertise features, for a total of 456 models.

For both sets of features, each individual feature corresponds to exactly one associate justice. I can therefore use the methodology presented in Section 6.3 to rank features for held-out documents and transform that ranking of features into a ranking of associate justices. For each document $\boldsymbol{w}_d$ and each feature f, I calculate $score(\boldsymbol{w}_d, f) = \log P(\boldsymbol{w}_d | f = 1)$ and rank features in descending order by score for each document. For example, I can evaluate the probability of words in a case from the Rehnquist court given the feature `AGAINST_Scalia_Rehnquist` = 1, with all other features set to 0, to estimate the probability of the words given that Scalia (hypothetically) voted against Rehnquist on this case. As I am interested in authorships assigned by the chief justice, I dropped cases where the chief justice was not in the majority and therefore did not assign the opinion, and where the chief justice wrote the opinion himself. The set of relevant `disagreement` features for a given case are those between an associate justice that voted in the majority and the chief justice for the case.

Note that sorting by score is not equal to sorting by the probability $P(f = 1 | \boldsymbol{w}_d)$. We can use Bayes rule to construct this probability from $P(\boldsymbol{w}_d | f = 1)$ by calculating the prior probability $P(f = 1)$. For the `disagreement` features, this means calculating the

probability that a given associate justice disagrees with a given chief justice. I calculated this probability by counting the number of times each justice voted against each chief justice as a proportion of all cases on which both justices voted. Adding the log of these probabilities did not affect rankings.

In the case of the `disagreement` features, my goal is to guess which of the justices that voted with the majority was most likely to have opposed the chief justice's preferred outcome. Ranking justices by the probability of their disagreement feature given a document should provide an estimate of the probability that that justice would have disagreed with the chief justice on the particular issues of the case.

For each case, I ranked justices according to their score under the `disagreement` model, and recorded the rank of the true author. If the model is a good predictor, the true author should appear at a low rank (1 or 2). If the model has no predictive effect, true authors should be uniformly distributed among ranks. Results with $T = 150$ are shown in Table 7.1. Each row represents one model, and each column represents a majority size, from 5 (close cases) to 9 (unanimous cases). For each row, the points in the histograms represent a single case: the rank of the author within the justices voting with the majority. Each column represents the same authorships for the same set of cases, but ranked by the different models.

The only new model is "prior", which ranks justices by their overall probability of disagreeing with the chief justice for a held-out case. This ranking does not take into account any word information from the cases.

If we consider unanimous cases (the rightmost column), we find that for most models except the prior model there is no clear trend: the rank of the real author is uniformly distributed from 1 to 8 (the chief justice is not included). In contrast, for most models in the close cases (the leftmost column) there is a consistent downward slope from rank 1 to rank 4, particularly for DMR, AT, and LN. The prior distribution does not by itself predict authorship well in any context.[3]

[3]It is, however, by far the best predictor considered of whether an individual justice will in fact vote against the chief justice

I assume that the probability of disagreeing with the chief justice on a particular case is a good proxy for ideological distance. This probability appears to be a good predictor of authorship in close cases, but not in unanimous cases. This result is consistent with Maltzman's findings: the ideological distance between an associate justice and the chief is a factor when a majority must be solidified, but is not significant otherwise.

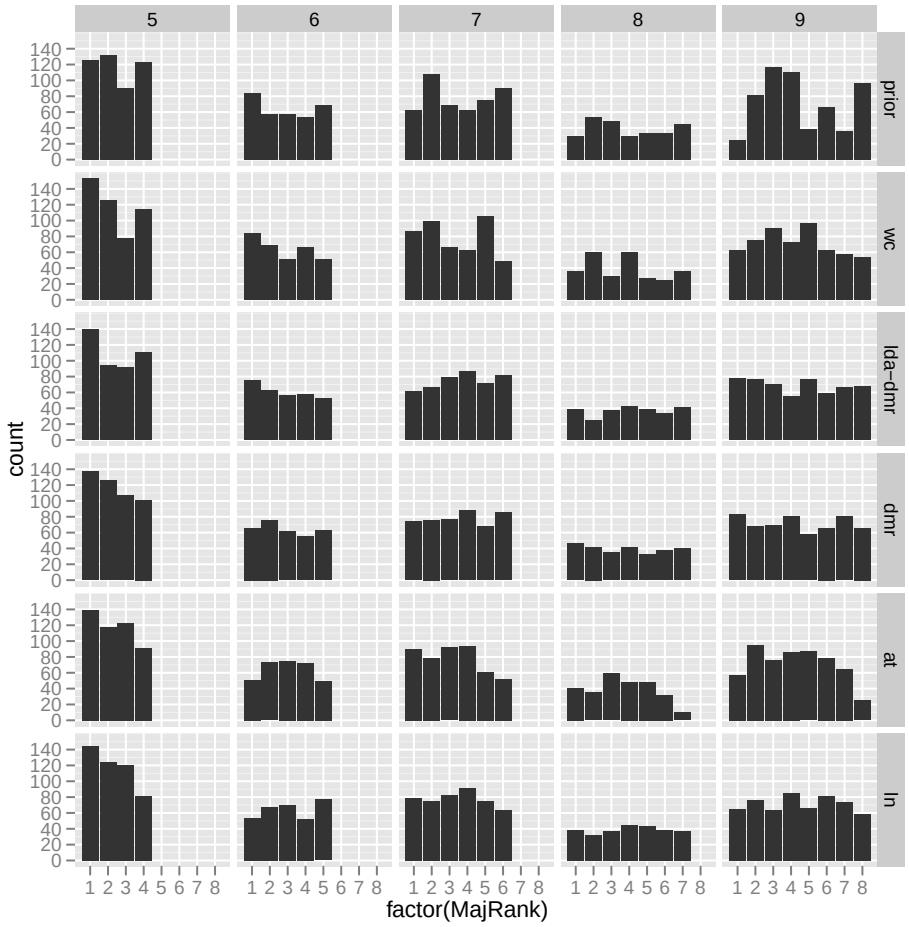

Figure 7.1. The ability of `disagreement` features to predict authorships. Columns correspond to models. Columns of plots correspond to the size of the majority (close 5-4 decisions on the left, unanimous 9-0 decisions on the right). In most word-based models (all but `prior`), the probability that a justice will disagree with the chief justice is a better predictor of the fact that that justice will write the opinion `only for close decisions`.

In comparison, results for *expertise* features are shown in Table 7.2. In this case, authorship of similar documents in the training set is a good indicator of authorship in held-out documents both for close cases and unanimous cases.

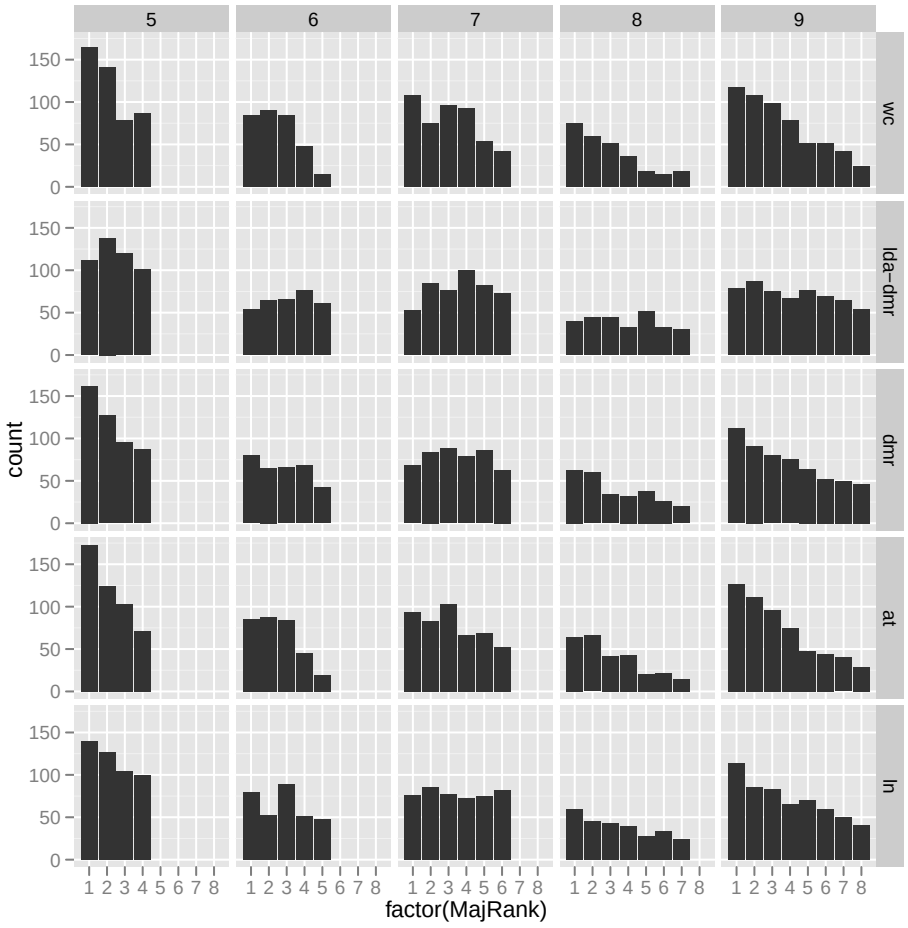

Figure 7.2. The ability of `expertise` features to predict authorships. As in the previous figure, columns correspond to models and rows correspond to majority sizes. Features indicating authorship of documents in training cases show a strong ability to predict authorship of held-out cases in both close cases (majority size 5) and unanimous cases (majority size 9).

7.2 News data: New York Times

My next example corpus consists of 20 years of New York Times articles, from 1987 to the first six months of 2007. These articles are available in XML format from the Linguistic

Data Consortium. For each article, the collection includes the lead paragraph and the full text, as well as substantial metadata, such as the publication date, desk of origin, section and page number, and indexing service annotations such as people, locations, descriptions, and taxonomic classifiers. Considering only lead paragraphs, this collection contains 1.85 million articles and more than 150 million word tokens.

In order to explore potential topic variation in the NYT collection, I ran a simple topic model using no non-textual information. Figures 7.3 and 7.4 show the proportions of two topics ($T = 1000$) relative to the total number of tokens in each month. The first topic, with highest probability words *Series Yankees Sox Red World* appears to relate to baseball, particularly the world series and the Yankees/Red Sox rivalry. As expected, the topic shows a strong cyclical pattern over time, peaking in October every year and then crashing to a low baseline level, with the exception of 1994. A similar anomaly in 1994 is present in the second topic, characterized by high probability words *players League owners*. In fact, the 1994 world series was cancelled due to a players strike. The time series in Figure 7.3 might therefore be well modeled using a combination of a cyclical component with period one year and a first or second order markov chain that might account for the sudden dip in 1994. Much of the variation in this topic might also be well explained by the presence of indexing service annotations.

Cyclic patterns appear to exist at multiple time scales in this collection, from multi-year cycles such as elections and sports championships (world cup soccer is highly regular, the Olympic games less so due to the shift in winter games in 1994), to cycles within years (spring and fall collections, quarterly earnings reports) and from week to week (movie reviews, science articles).

In order to account for cyclical variation, I trained a logistic normal topic model using a combination of a first order dynamic linear model and a seasonal model. These represent, first, a linear chain time series with one covariate for each month of the 20 year period covered by the corpus, and second, a set of independent covariates, one for each month. Thus the mean log topic probability for March 2003 is modeled as a parameter for *March 2003* plus a parameter for *March*. For identifiability, the parameter for January is restricted

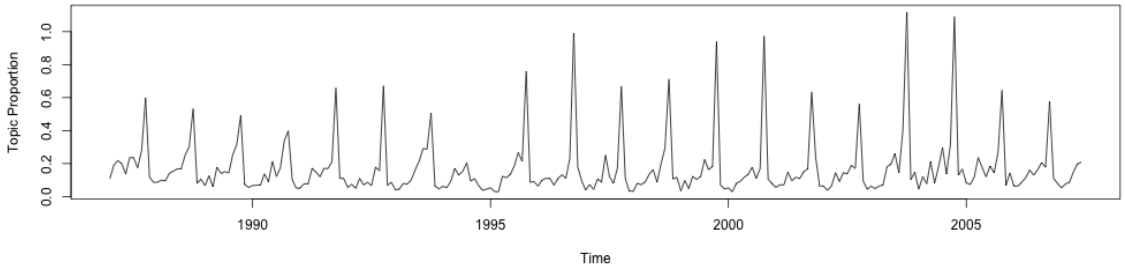

Figure 7.3. Plot of relative prominence over time for one of 1000 topics on the New York Times corpus, with highest probability words *Series Yankees Sox Red World League game Boston team games baseball Mets Game series won Clemens Braves Yankee teams.*

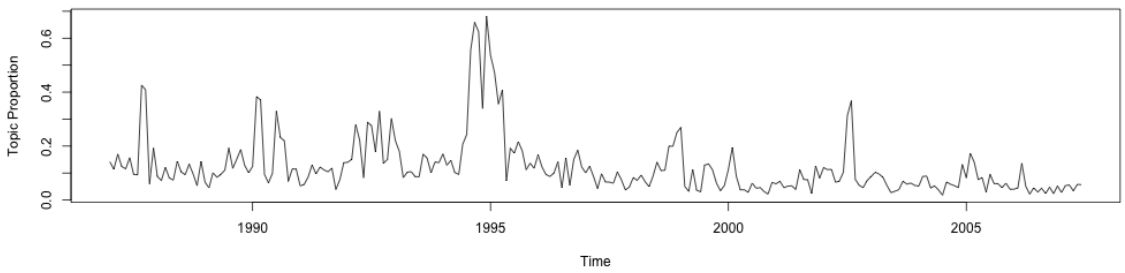

Figure 7.4. Plot of relative prominence over time for one of 1000 topics on the New York Times corpus, with highest probability words *players League owners league baseball union commissioner Baseball Association labor Commissioner Football major teams Selig agreement strike team bargaining.*

to 0. In order to emphasize topics with cyclical structure, I selected the subset of articles from the sports desk. Statistics for this corpus are shown in Table 7.3

Table 7.2. NYT: common features and rare features. Article leads are short with relatively high variance. The small variance in features per document results from dropping the parameter for *January*. The large variance in documents per feature reflects the combination of frequent month features and rare day-to-day features.

Docs	Vocab	Features	Words/Doc	Feat./Doc	Docs/Feat.
174991	131744	258	39 ± 57.5	2.9 ± 0.3	1,974.8 ± 11,169.9

Estimated values for these parameters for two topics are shown in Figures 7.5 and 7.6. The 11 month-specific parameters for the seasonal model are shown on top with the linear chain model underneath. As we have multiple samples from the approximate posterior over parameters, the gray line in the time series shows one standard error. Standard errors are small for the month-specific parameters. These results demonstrate the ability of the model to filter out seasonal variation from changes in the underlying distribution. The most clear anomaly in the time series for the baseball topic in Figure 7.5 is the downward spike in 1994 when the World Series was cancelled. Less clear patterns include a general increase in the proportion of coverage devoted to baseball after 1996, when the New York Yankees began seriously competing in postseason baseball, and a general decrease in the proportion of coverage of college basketball beginning in the 1990's.

7.3 Political speeches: State of the Union addresses

For the next example application, I consider the problem of long documents. Topic models are often applied to short, topically focused documents such as abstracts. Many longer documents, however, range over a variety of topics, often with little indication in the text. Here, we consider the state of the union addresses delivered by US presidents, which typically discuss a sequence of issues but contain no explicit section headings. Modeling each speech as a single document may not provide enough focus to learn meaningful topic-specific distributions over words. However, modeling paragraphs independently may not provide enough context to the model. I therefore model each paragraph p as a point in topic space $\boldsymbol{\beta}^{(p)}$ in a logistic normal topic model, and define a graph connecting each paragraph's

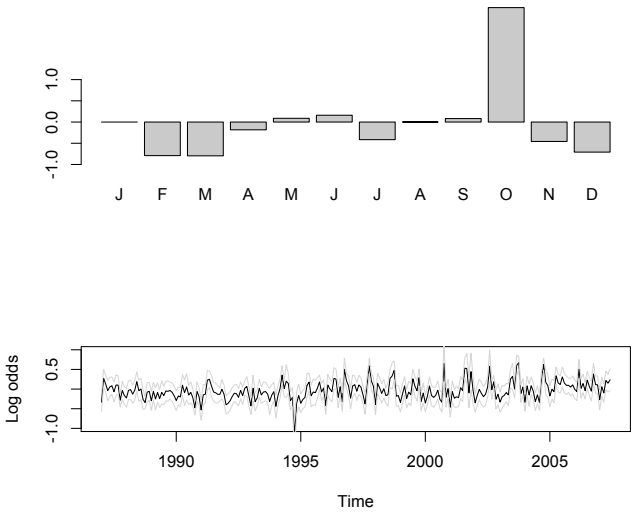

Figure 7.5. The world series happens in October. This topic, which relates to the baseball World Series, is low during the Winter, increases as baseball season starts in April, and peaks in October. After accounting for seasonal variation, the filtered time-series shows clearly both the baseball strike in 1994 and increased coverage during the early 2000's.

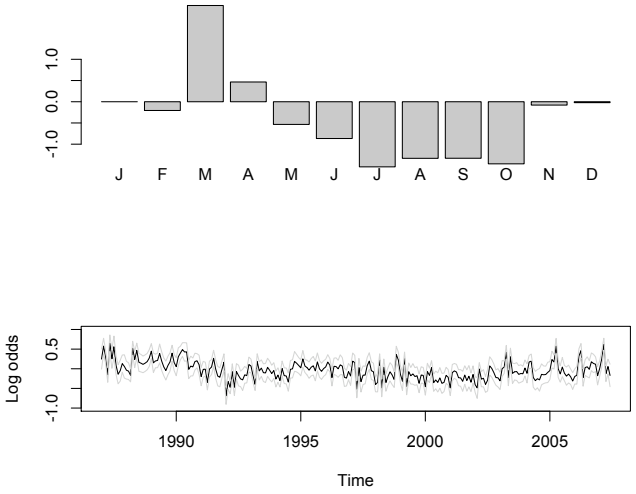

Figure 7.6. March Madness. This topic, which relates to college basketball, peaks during the March NCAA tournament, declines during the summer, and begins again with the college basketball season in November.

parameters to those of the previous and next paragraphs in order to construct a GMRF prior on topic proportions.

Table 7.3. SotU: unique features. Each paragraph has its own topic distribution that depends on the year, the previous paragraph, and the subsequent paragraph.

Docs	Vocab	Features	Words/Doc	Feat./Doc	Docs/Feat.
19254	10035	19254	27.7 ± 26.5	1 ± 0	1 ± 0

It is also useful to measure the prevalence of topics over time, as well as the degree of topic change from year to year. Previous work in topic-based analysis of presidential addresses [37] uses a beta distribution for each topic, which can only support a limited number of patterns (constant, gradual decline or increase; single narrow spikes). A more flexible model that allows large discontinuities and multiple modes is therefore preferable. I augment the graph structure previously proposed for speeches by adding additional edges between every paragraph from year y and a single mean vector $\boldsymbol{\mu}_y$ for that year's speech, as well as edges between $\boldsymbol{\mu}_y$ and $\boldsymbol{\mu}_{y+1}$.

Figure 7.8 shows the exponentiated parameters estimated for three topics, relating to Native Americans, wars, and energy. Figure 7.8 shows the posterior distribution over precisions between speeches from 1950 to 2009. Vertical lines indicate transitions between presidents. We can make several observations from these results. The transition between Kennedy and Johnson (1963, both democrats) was larger than that between Eisenhower and Kennedy (1960, republican to democrat) and between Johnson and Nixon (1968, democrat to republican). The topics during the Vietnam era were generally very stable. Clinton's presidency was also relatively stable except for 1994 (when democrats lost congress to republicans). George W. Bush's first speech was very different from Clinton's last (2000), but Bush's administration was generally much more turbulent than Clinton's. The variance between Bush's last speech (2008) and Obama's first was the largest in this time frame.

7.3.1 Model effect on trend analysis

An important goal in analyzing collections that span many years, particularly in scientific literature, is to understand trends over time. Policymakers, for example, need to follow the development of new fields and track the progress of more mature areas in order to make

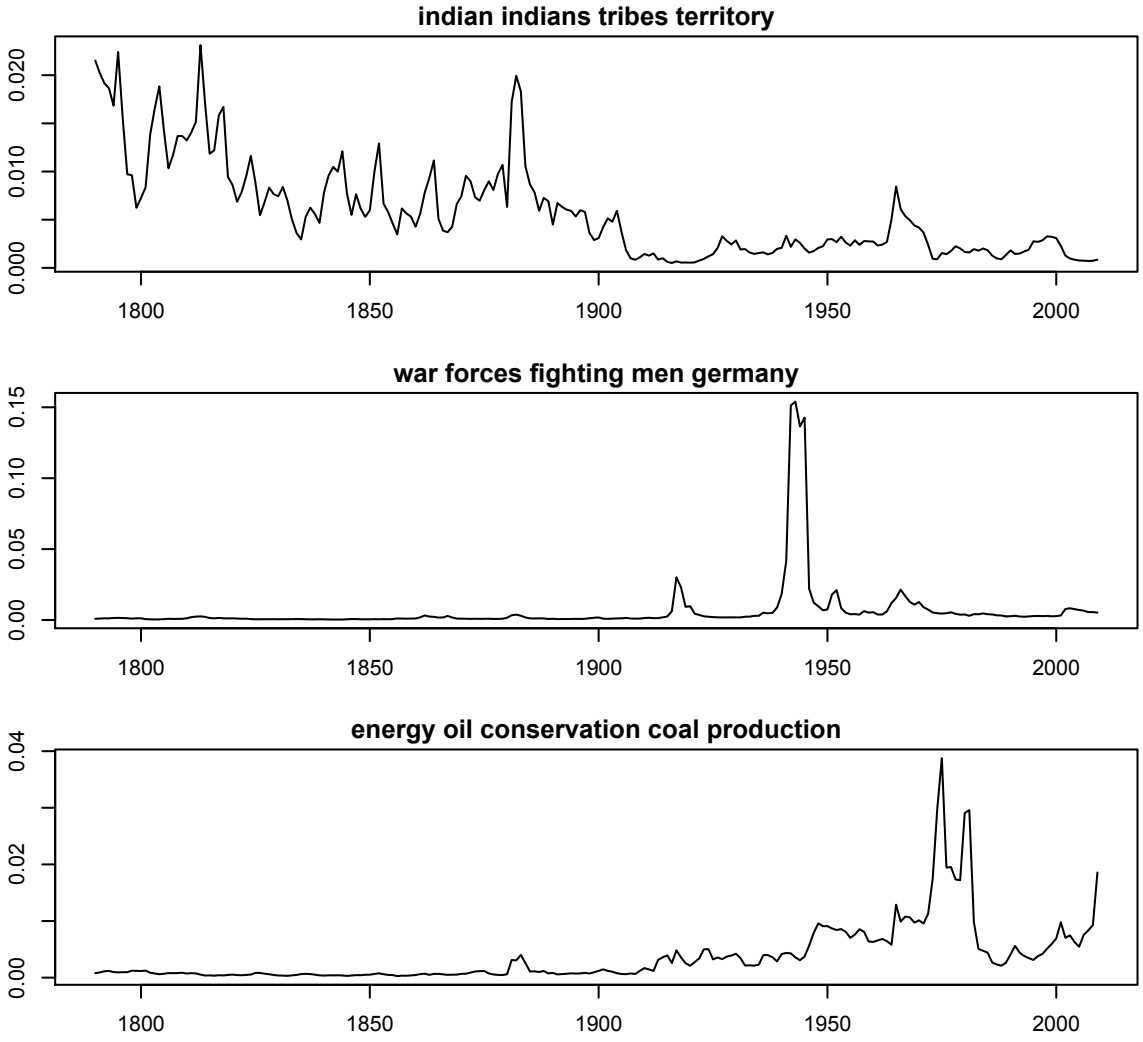

Figure 7.7. Exponentiated topic parameters for three topics.

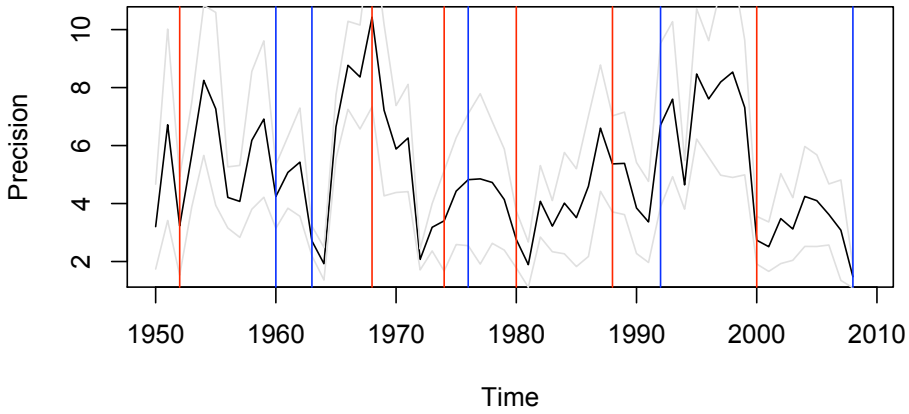

Figure 7.8. Year-to-year precisions between state of the union addresses. Larger values indicate more topic similarity between speeches. Gray lines are one SD, vertical lines indicate transitions between presidents.

strategic funding decisions. Evaluating temporal trends from data, however, may lead to incorrect conclusion if care is not taken to account for the conditions under which a corpus was collected. Journals may publish irregularly or be missing for particular periods, leading to apparent dramatic changes in topic frequency.

Figure 7.9 shows mean topic proportions over time for one topic from each of three models trained on the Rexa corpus with $T = 50$. The three topics were selected by finding the topic in each model that had highest probability for the terms *problems* and *search*. Because topic models are stochastic clustering models, it is difficult to compare topics across models. To show that these topics are comparable, I show the most probable words and phrases[4] for each, along with the number of word tokens shared by each pair of topics in Table 7.4.

The first model is a simple topic model trained with no access to metadata. The observed time series is highly variable, but a great deal of this variation is in fact due to the fact that this topic is very important in the conference IJCAI, which takes place every two years.

[4] "phrases" are identified by searching for sequences of two or more consecutive words assigned to the same topic, and reporting the most frequent multi-word sequences for each topic.

Table 7.4. Comparable topics across models. The third topic includes constraint satisfaction, but also *heuristic search* at slightly lower rank than the other topics. The number of tokens assigned to the topic at one Gibbs sampling state is on the diagonal, tokens shared by each pair of topics are shown in off-diagonals. 5997 word tokens are shared by all three topics. As a comparison, three randomly selected topics of similar size have 8 words in common.

search, problems, heuristic search, algorithms, problem, algorithm, local search, search space, heuristic, optimal, space, solving, problem solving, heuristics, np hard	16947	9286	7595
search, problems, heuristic search, problem, algorithms, heuristic, local search, problem solving, solving, heuristics, search space, local, finding, algorithm, scheduling		20235	9224
search, problems, constraint satisfaction, constraint, problem, problem solving, algorithms, constraints, solving, local search, arc consistency, ..., heuristic search			19526

The second plot is from a logistic normal topic model with features for year of publication. The plot shows, at each year, the exponentiated parameter for this topic divided by the sum of the exponentiated parameters for all topics in that year. This plot is significantly smoother, but still shows spikes in odd years (1997, 1999, 2005). The third plot is from a logistic normal model with year of publication and publication venue metadata. The plot represents the same quantity as the previous plot: the exponentiated parameter for the year, normalized against all other topics. Although this topic has only 500 fewer tokens than the previous topic, its overall mean year-to-year proportion is significantly less: this topic is very localized to a small number of venues, so most of its presence is accounted for by the venue-specific parameters.

These three plots show different stories. Without access to metadata, the topic appears prominent, accounting for nearly 3% of all tokens in the corpus, but chaotic, with no clear temporal trend. Adding a dynamic linear model to smooth across time results in a similar but less variable picture. Finally, adding publication-venue information results in a slightly different trend, and a greatly different average proportion.

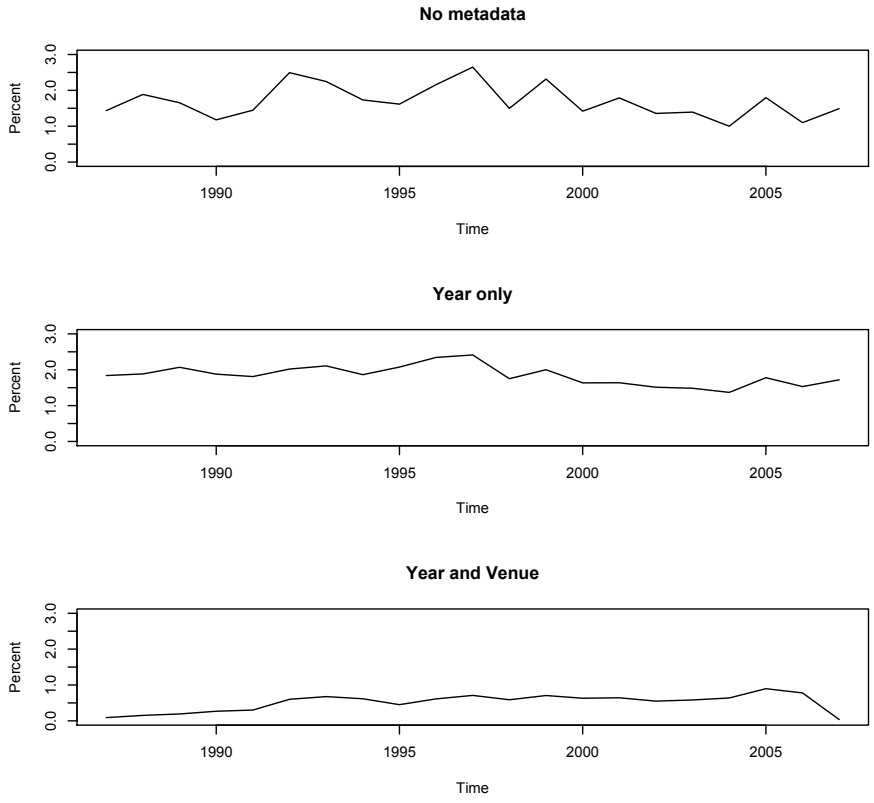

Figure 7.9. Accounting for publication venue explains an apparent trend. These plots show three views of the same corpus. The first (left) shows a time-series plot for a topic related to heuristic search in artificial intelligence, derived from a simple topic model that had no access to document-level metadata. The second plot (top right) shows parameters for the most similar topic from a GMRF topic model with a second-order dynamic linear model over time. The third plot shows parameters for the most similar topic from a second-order DLM with the addition of indicator features for publication venue.

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

The goal of this work is to discover methods for estimating the association between text and metadata. A good model should identify the patterns underlying text collections and use those patterns to predict

In this thesis I have explored several statistical models for measuring the association of words with non-textual metadata. My major findings are as follows:

- Latent topic spaces provide an efficient and effective tool as an intermediate representation between observed words and metadata as measured by predictive probability.
- Although theoretically similar, there is a significant empirical difference in performance between logistic normal and Dirichlet-based topic regression models. This difference is likely to be the result of more effective inference techniques available for the DMR models due to the conjugacy of the Dirichlet and multinomial distributions.
- DMR topic models provide good performance relative to Author-Topic models across a wide variety of metadata environments, from very sparse, high-dimensional settings to models with a handful of well-represented metadata elements. In cases where metadata is not informative, they are able to match closely the performance of standard topic models.
- Although there are consistent differences in the latent topic spaces found by different metadata-based models, when computation is at a premium, DMR models can be effectively approximated using simple post hoc estimation technique.

The focus of this thesis has been on understanding the difference between possible models for measuring the association between text and metadata. The DMR model appears to be

the simplest and most reliable method, while also demonstrating the strongest ability to learn patterns with predictive ability.

DMR topic models provide a simple, efficient tool for users to compare the predictive ability of different metadata elements, and to explore whether that association changes between sub-corpora. The next step in supporting such experiments is to formulate a standard methodology for using the tool of DMR and held-out probability, assessing model fit and making recommendations to practitioners about when observed differences are significant.

This work is about modeling the choices people make in writing documents, at the level of themes or topics. DMR topic models use metadata to inform distributions over the choice of very simple linguistic models: that is, unigram distributions over words, but the DMR framework applies to any model that contains a distribution over some finite choice of discrete possibilities. When people write documents, however, they make many more choices than simply a mixture of topics. Syntactic structures and the implied semantic roles of nouns within sentences provide important information as well, and natural language processing is starting to be able to model these phenomena effectively. Researchers may be interested, for example, in whether particular groups of authors use a particular class of nouns more in active or passive settings. This difference could be modeled by adding a regression model that changes the expected distribution over particular choices within a syntactical parsing or semantic role labeling model. Given the success of DMR in improving the ability of topic models to learn the latent relationships between text and metadata, and in providing users with the ability to *measure* those relationships, it is likely that other interesting characteristics of text documents would be measurable with more powerful regression models.

BIBLIOGRAPHY

[1] Aitchison, J., and Shen, S.M. Logistic-normal distributions: Some properties and uses. *Biometrika 67*, 2 (Aug. 1980), 261–272.

[2] Aitchison, John. *The Statistical Analysis of Compositional Data*. Chapman & Hall, 1986.

[3] Albert, James H., and Chib, Siddhartha. Bayesian analysis of binary and polychotomous response data. *JASA*, 422 (1993).

[4] Besag, Julian. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological) 48*, 3 (1986), 259–302.

[5] Blei, David, and Jordan, Michael. Modeling annotated data. In *SIGIR* (2003).

[6] Blei, David, and McAuliffe, Jon D. Supervised topic models. In *NIPS* (2007).

[7] Blei, David, Ng, Andrew, and Jordan, Michael. Latent Dirichlet allocation. *Journal of Machine Learning Research 3* (January 2003), 993–1022.

[8] Blei, David M., and Lafferty, John D. Dynamic topic models. In *ICML* (2006).

[9] Blei, David M., and Lafferty, John D. A correlated topic model of *Science*. *AAS 1*, 1 (2007), 17–35.

[10] Bonneau, Chris, Hammond, Thomas, Maltzman, Forrest, and Wahlbeck, Paul. Agenda control, the median justice, and the majority opinion on the u.s. supreme court. *American Journal of Political Science 51* (October 2007), 890–90.

[11] Buntine, Wray L. Estimating likelihoods for topic models. In *Asian Conference on Machine Learning* (2009).

[12] Deerwester, Scott, Dumais, Susan T., Furnas, George W., Landauer, Thomas K., and Harshman, Richard. Indexing by latent semantic analysis. *JASIS 41*, 6 (1990), 391–407.

[13] Devroye, Luc. Random variate generation in one line of code. In *Winter Simulation Conference* (1996).

[14] Dietz, Laura, Bickel, Steffen, and Scheffer, Tobias. Unsupervised prediction of citation influences. In *ICML* (2007).

[15] Erosheva, Elena, Fienberg, Stephen, and Lafferty, John. Mixed membership models of scientific publications. *PNAS 101*, Suppl. 1 (2004), 5220–5227.

[16] Frühwirth-Schnatter, Sylvia, Frühwirth, Rudolf, Held, Leonhard, and Rue, Havard. Improved auxiliary mixture sampling for hierarchical models of non-gaussian data. Tech. Rep. 2008-34, IFAS Research Paper Series, Sep 2008.

[17] Funk, Mark E., Reid, Carolyn Anne, and McGoogan, Leon S. Indexing consistency in MEDLINE. *Bull. Med. Libr. Assoc. 71*, 2 (April 1983).

[18] Griffiths, Thomas L., and Steyvers, Mark. Finding scientific topics. *PNAS 101*, suppl. 1 (2004), 5228–5235.

[19] Groenewald, Pieter C.N., and Mokgatlhe, Lucky. Bayesian computation for logistic regression. *Computational Statistics and Data Analysis 48* (2005), 857–868.

[20] Guimaraes, Paulo, and Lindrooth, Richard. Dirichlet-multinomial regression. Econometrics 0509001, EconWPA, Sept. 2005.

[21] Hofmann, Thomas. Probilistic latent semantic analysis. In *UAI* (1999).

[22] Landauer, Thomas K., Laham, Darrell, and Derr, Marcia. From paragraph to graph: Latent semantic analysis for information visualization. *PNAS 101*, Suppl 1 (2004), 5214–5219.

[23] Liu, D.C., and Nocedal, J. On the limited memory method for large scale optimization. *Mathematical Programming B 45*, 3 (1989), 503–528.

[24] Maltzman, Forrest, II, James F. Spriggs, and Wahlbeck, Paul J. *Crafting law on the Supreme Court: the collegial game*. Cambridge University Press, 2000.

[25] Maltzman, Forrest, and Wahlbeck, Paul J. Opinion assignment on the rehnquist court. *Judicature 89* (November/December 2005), 121–126.

[26] McCallum, Andrew, Corrada-Emmanuel, Andrés, and Wang, Xuerui. Topic and role discovery in social networks. In *IJCAI* (2005).

[27] McCallum, Andrew Kachites. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[28] Meilă, Marina. Comparing clusterings by the variation of information. In *COLT* (2003).

[29] Mimno, David, and McCallum, Andrew. Expertise modeling for matching papers with reviewers. In *KDD* (2007).

[30] Mimno, David, and McCallum, Andrew. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *UAI* (2008).

[31] Mimno, David, Wallach, Hanna, and McCallum, Andrew. Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs* (2008).

[32] Newman, David, Chemudugunta, Chaitanya, and Smyth, Padhraic. Statistical entity-topic models. In *KDD* (2006).

[33] Rosen-Zvi, Michal, Griffiths, Tom, Steyvers, Mark, and Smyth, Padhraic. The author-topic model for authors and documents. In *UAI* (2004).

[34] Rue, Havard, and Held, Leonhard. *Gaussian Markov Random Fields*. Chapman & Hall/CRC, 2005.

[35] Wallach, Hanna, Mimno, David, and McCallum, Andrew. Rethinking LDA: Why priors matter. In *NIPS* (2009).

[36] Wallach, Hanna, Murray, Iain, Salakhutdinov, Ruslan, and Mimno, David. Evaluation methods for topic models. In *ICML* (2009).

[37] Wang, Xuerui, and McCallum, Andrew. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD* (2006).

[38] Wang, Xuerui, Mohanty, Natasha, and McCallum, Andrew. Group and topic discovery from relations and their attributes. In *NIPS* (2005).

[39] Yao, Limin, Mimno, David, and McCallum, Andrew. Efficient methods for topic model inference on streaming document collections. In *KDD* (2009).