Applications of Latent Variable Models in Modeling Influence and Decision Making

Sean M. Gerrish

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Advisor: David M. Blei

April 2013

Abstract

The past twenty years have seen an avalanche of digital information which is overwhelming people in industry, government, and academics. This avalanche is two-sided: while the past decade has seen an onslaught of digitized records – as governments, publishers, and researchers race to make their records digital, the electronic and software tools for computationally analyzing this data have quickly evolved to face this challenge.

Many of these challenges evolve around recurring patterns, including the presence of text, bits of information about pairs of items, and sequential observations. In this work we present several methods to address these challenges in data analysis which take advantage of these recurring patterns.

We begin with a method for identifying influential documents in a collection which evolves over time. We demonstrate that by encoding our assumptions about influential documents in a statistical model of the changes in textual themes, we are able to provide an alternative bibliometric which provides results consistent with—yet different from—traditional metrics of influence such as citation counts.

We then introduce a model for measuring the relationships between pairs of countries over time. We will demonstrate that this model is able to learn meaningful relationships between countries which is extraordinarily consistent across different human labels.

We next address limitations in existing models of legislative voting. In one extention we predict legislators' votes by using the text of the bills they are voting on combined with individual legislators' past voting behavior. We then introduce a method for inferring these lawmakers' positions on specific issues.

A recurring theme in the methods we present is that by using a small set of statistical primitives, we are able to apply known (or mildly adapted) methods to new problems. Several advances in the past few decades in statistical modeling will make the development and discussion of our models easier, as they will provide both this set of primitives (which can be interchanged easily) and the tools for working with them. As a final contribution, we describe a new method for fitting a statistical model with variational inference, *without* the time investment typically required of practitioners.

Acknowledgements

In this section I outline the various sources of help I have received in completing this thesis.

Funding

This work was supported by grants secured by my advisor from the Office of Naval Research, ONR 175-6343, NSF CAREER 0745520, AFOSR 09NL202, the Alfred P. Sloan foundation, and a grant from Google. I was also supported by a fellowship provided by Princeton to first-year graduate students. I was supported by the Princeton Computer Science department's teaching assignments during my second year.

Various outside organizations have provided funding for travel to conferences for presenting this work. This includes scholarships provided by ICML, NIPS, and the Machine Learning Summer School organized at Cambridge in 2009, all of which have supported me with funding for transportation, lodging, and/or registration. Google, Inc. provided funding for travel to NIPS 2010; and Facebook, Inc. provided funding for travel to ICML 2010. Princeton University's Dean's Fund for Scholarly Travel and Princeton's School of Engineering and Applied Sciences Graduate Travel Funds both provided funds for travel to NIPS 2011.

School

Foremost, I owe my advisor, David Blei, many thanks for his mentorship and support for the past four years. The preponderance of this mentorship has been on research, but Dave's support during the formidable "middle years" also helped me to press through the program. One of the things that stands out with Dave's mentorship is his steadfast insistence that we continue on a project that has been started. This was a healthy counterweight to my "fail fast" philosophy, which, retrospectively, may have led to a very unproductive research career (Dave was also able to differentiate between disillusionment and failure, which helped in this process).

This was exemplified by a diagram Dave pointed to on his board one day. "This is the process of research. This graph plots excitement with your research project as a function of time," he said, pointing to the curve. "You start out excited about your project, but after a while you get sick and tired of it." He pointed to the precipitous dip in the curve. "You'll end up sick of your project at this time – it happens to everybody – but if you keep on pushing forward, then you'll end up with good research." The curve moved back upward, past the dip. This diagram was helpful. It was fairly obvious, and I knew this curve well. But explicitly seeing that other researchers go through the

same thing was helpful nonetheless. (It also helped knowing that Dave did not draw this diagram expressly for me; it was on his board from an earlier meeting he'd had.)

I would also like to thank the various researchers in the field of machine learning and the social sciences who have served as mentors or inspirations, whether they knew it or not. One of these was Leon Bottou, from whom I learned more than I should admit while TAing for him. Leon's deep understanding of mathematics, pleasant manner, and modesty are testaments to his character. I also want to thank Kevin Quinn and Mark Gergen for hosting me for a month at Berkeley to fit several models over the New York Appellate Courts. Finally, my time at JSTOR hosted by Clare Llewellyn, John Burns, Michael Krot, and Ron Snyder was a valuable experience.

I owe many thanks to a number of current and former graduate students in my research lab and our broader research group, who have helped me in this program in various ways. They have served as sounding boards and have served as excellent role models. The graduate students include Jordan Boyd-Graber Ying, Jonathan Chang, Chong Wang, Indraneel Mukherjee, Gungor Polatkan, Lauren Hannah, Matthew Hoffman, Sam Gershman, Melissa Carroll, Berk Kapicioglu, Prem Gopalan, Allison Chaney, and Rajesh Ranganath. The post-doctoral researchers have been equally as helpful; they include David Mimno, John Paisley, and Jeremy Manning.

Committee

I would like to thank my committee, who spent the time to provide me with helpful feedback for this thesis. Their comments were extremely valuable.

- David Blei (reader / advisor)
- Rob Schapire (reader)
- Hanna Wallach (reader)
- Matthew Salganik (non-reader)
- Kosuke Imai (non-reader)

Friends and Family

There are various outside mentors in my life whom I consulted during the "middle years" (circa 2010) to re-orient myself in the graduate program. These included Arash Baratloo, who went through similar periods in graduate school; Jeffrey Oldham, who pointed out the intangible benefits

in a PhD; and Doug Beeferman, who provided an honest and fair point of view from the "other" side. I also want to thank Ricky Wong, who reminded me that the work I am doing is interesting to people outside of Academia (by actually using these tools at his startup) and for providing topics from a biology textbook that I reference in Chapter 2.

I also want to thank those who wrote letters of support for my applications to grad school. These included Dan Rubinstein; who provides incredibly direct, honest, and sound career advice; and who is an excellent role model; Dragomir Radev, whose work and mentorship in computational linguistics and natural language processing pointed my career in its current direction; and Mark Skandera, who mentored me in math reserch as an undergraduate and continues to be a good friend.

My fiancee Sarah was immensely helpful in keeping me sufficiently distracted during this period. Her encouragement and patience was invaluable in keeping me happy during the last few years.

My parents deserve many thanks for their support in the past three decades. Those early years were particularly important, and they set the stage for my interest in science, engineering, and math—let alone stressing the importance of school.

I may have never become interested in computer science if my older brother Josh hadn't shown me things like division, Robot Odyssey and Rocky's Boots, BASIC, fractals, multi-user dungeons, bulletin boards, and countless other geeky things (okay, maybe I would have learned about division in third grade, but it's much cooler if your older brother shows you). Growing up in the wake of Josh was invaluable. His books and articles on logic puzzles, algebra, mathematical proofs, and 3D computer graphics gave me a chance to learn some of these ideas on my own time. Josh also encouraged me to push ahead with my PhD during the times that it would have been easiest to take another path.

My brother Jason has provided me with with many experiences that transcended math and science. Jason's mischief in the first year of my program was probably a good thing for me, as it made me aware that there are important things outside of research that are worth thinking about; this provided a buoyant relief from research. Jason's steadfast encouragement was also important in helping me to continue my PhD.

Finally, I want to thank my little sister Kim. Kim, like Jason, provided a buoyant relief from work. She talked me through my first year of graduate school and inspired me through the remaining years.

This thesis is dedicated to Kimberly Misa Gerrish, 1984-2009.

Contents

Previous Publications

The work here represents expanded versions of the following publications and presentations:

- Sean Gerrish and David Blei. *Predicting legislative roll calls from text.* Proceedings of the 28th International Conference on Machine Learning (ICML), 2011.
- Sean Gerrish and David Blei. *A language-based approach to measuring scholarly impact.* Proceedings of the 27th International Conference on Machine Learning (ICML), 2010.
- Sean Gerrish and David Blei. *How they voted: A model of Issue Voting in the United States Congress.* Proceedings of the Twenty Fifth Annual Conference on Neural Information Processing Systems (NIPS) 2012
- Sean Gerrish and David Blei. *A text-based HMM of foreign-affairs sentiment.* Second Annual Computational Social Science and the Wisdom of Crowds (NIPS Workshop). December 2011.

The current presentation attempts to provide deeper and broader understanding to the reader of the above work.

Variable	Description
Constants	
C	Number of nation-states
D	Number of documents
K	Number of topics
N	Number of words (e.g., in a document)
P	Dimension of a generic latent space
T	Number of discrete time "epochs"
V	Number of words (e.g., in the vocabulary)
Subscripts	
c	Country or nation-state
d	Document
k	Topic
t	Time
u	Person, e.g., a lawmaker
Random variables	
a_d	Document d's polarization
b_d	Document d's popularity
s_d or s_{c_1,c_2}	The sentiment between two nations
v_{ud}	Lawmaker u's vote on item d
$\boldsymbol{w}_d$	A collection of words, as in a document d
x_u	An ideal point for lawmaker u
x_c	Position for country c during interaction
$\bar{x}_c$	Mean position for country c
X	A generic hidden random variable
Y	A generic observed random variable
z_n	$K-$variate topic indicator for term n
$\boldsymbol{z}_u(\boldsymbol{z}_{uk})$	Lawmaker u's position (on issue k)
α	Dirichlet parameter for LDA
$\boldsymbol{\beta}$	Coefficients of words in text regression
$\boldsymbol{\beta}(\boldsymbol{\beta}_t)$	LDA topics (at time t)
$\boldsymbol{\eta}$	Regression coefficient for sLDA
$\boldsymbol{\theta}_d$	Topic mixture for document d
Variational parameters	
γ_d	Variational parameter for document mixture θ_d
$\tilde{\beta}_t$	Variational parameter for topic chain β_t
ϕ_n	Variational parameter for a word's topic indicator z_n
$\tilde{\ell}$	Variational parameter for influence score I
$\tilde{a}_d$	Variational parameter for polarity a_d
$\tilde{b}_d$	Variational parameter for popularity b_d
$\tilde{x}_u$	Variational parameter for lawmaker ideal point x_u
λ_d, κ_d	Variational mean of bill parameters a_d, b_d
$\tau_{d,i}$	Variational mean of lawmaker ideal point x_u

Figure 1: The reader may find the notation in this table a helpful resource in the subsequent chapters.

Chapter 1

Quantitative methods for social research in the digital age

Quantitative social scientists often attempt to understand the behavior of society with numbers and data, and digital records are one of the most useful resources available to them. The digital age has brought to these researchers a deluge of records—particularly in the form of text. This avalanche of data provides more information to these scientists than they have had in the history of mankind.

Researchers are now able to pore over digital copies of all legally binding opinions written by United States Supreme Court Justices, or the text of thousands of bills voted on by members of Congress. Even these numbers are dwarfed by the hundreds of thousands of newspaper articles and blog posts written each day about the events happening in the world. Unfortunately, this flood of information obscures the very insights these researchers aim to discover. Researchers trying to make sense of these collections are subject to the high costs of time spent studying these collections in search of the few key insights.

The goal of this thesis is to describe several new statistical models that are now available for data practitioners and the consumers of that data[1] to better understand society through collections of text documents. I will focus on four high-level research questions that dovetail off one another to illustrate the flexibility and interpretability of latent variable models in large-scale settings.

An implicit premise of this thesis is that patterns are ubiquitous in collections of text documents, and that these patterns can be discovered automatically to describe decisions and behavior of actors

[1] By a (data) practitioner, I am referring to anyone who applies existing methods for data analysis, possibly tweaking or combining these methods to answer specific questions (such as database engineers or lab assistants). This contrasts with fundamental researchers, who research entirely new methods or tools for data analysis. A social scientist may be a researcher in his or her field but a practitioner in the field of data analysis.

in these collections. I will ground this discussion with the development of several specific statistical models, but I will stress throughout this thesis that these methods frequently draw from a suite of common tools which can be used again and again to construct models.

The deluge of information and some statistical tools

Observational social science data – including data about how organizations and the Government work – has become available on a massive scale. The National Archive, which collects information from over 500 federal agencies, has been digitizing its collection of twelve *billion* federal documents (Lazer *et al.*, 2009; National Archives Workshop, 2012; National Archives Press Release, 2012). The problem in handling this data has moved from collecting the data to processing and understanding it. Fortunately for scholars, these data follow recurring patterns which make statistical modeling possible. In this thesis, I will focus on three specific patterns:

Text data. Text data is the low-hanging fruit of most social science research questions. It is ubiquitous because it can—indeed, it must—be easily created, digitized, and stored. It serves as an observation which we can use to better understand the story underlying decisions and politics.

Just as text data is invaluable to researchers, the rate of growth of these text collections is staggering. A single newspaper like the *New York Times* publishes hundreds of thousands of articles each decade. Of the National Archive's collection, billions of its documents are text (National Archives Workshop, 2012; National Archives Press Release, 2012). The rate of growth of sources like the World Wide Web is even more tremendous. As far back as 2008, the Internet was already growing at a rate of several billion webpages per *day* (Google Blog, 2008).

Time-series collections. Many datasets comprise time-series observations. Timestamps are one of the simplest types of metadata to attach to digital collections because they are described by a single scalar and because they are inexpensive and widely available. In spite of its simplicity, the addition of a time variable to statistical models can provide rich insight and a useful historical perspective into collections of documents. It is especially interesting to researchers because it is helpful in framing questions about causation, prediction, and influence.

Relational observations. One of the simplest ways to represent more complicated phenomena is to use the interaction between pairs of items. We will refer to such pairs (and their relationships) as *dyadic*. In later chapters I will use spatial models to represent interactions between lawmakers

and bills (i.e., how congresspersons voted on bills) and between countries (i.e., countries' sentiment toward one another). As we will show, the underlying representation for these cases is very similar.

The role of statistical machine learning

The deluge of information available to researchers means that if researchers want broad coverage over the available sources, they cannot spend long looking over any single document. For example, a graduate student studying patterns governing the relationships between pairs of countries, based on mentions of pairs of countries in the last twenty years of the *New York Times*, would need to spend every day of an entire year, twenty-four hours per day, to code the 300,200 interactions per pair of countries (at two minutes per article). A *computational* treatment is therefore necessary if researchers intend to handle large collections of data.

Statistical representations of these data will be useful because they provide an explicit way to formalize our assumptions. We are fortunate that computers can be programmed to speak in this language, because they are the only means by which we can achieve broad understanding of large collections of text documents.

In this thesis, I will use probabilistic models to encode these statistical assumptions. I will frequently use the paradigm of graphical models (Pearl, 1985) to make our assumptions more clear. Because these statistical methods provide directed summaries, they can serve as optical lenses for researchers to analyze entire collections of documents, which I illustrate as a cartoon in Figure 1.1. Statistical methods enable researchers to describe arbitrarily complex transformations of data with arbitrarily complex lenses. Because of the simplicity of each of these lenses, we will find that a wide array of models can be created by nesting and re-using modules across different applications.

Organization

By the end of this thesis, the reader should have a better understanding of several new models that I have designed for to social scientists. Perhaps more importantly, the reader will be prepared to design his or her own latent-variable model for similar applications.

To this end, I will provide a lower level of detail about latent-variable models in the early chapters of this thesis than normally expected in a doctoral thesis when it may help a reader unfamiliar with this subject to understand the material. I also present some of the most advanced (or uninteresting) math in the appendix to keep the discussion of applications and modeling at the forefront.

I provide preliminary material in Chapter 2, outlining the statistical "primitives" that I will use

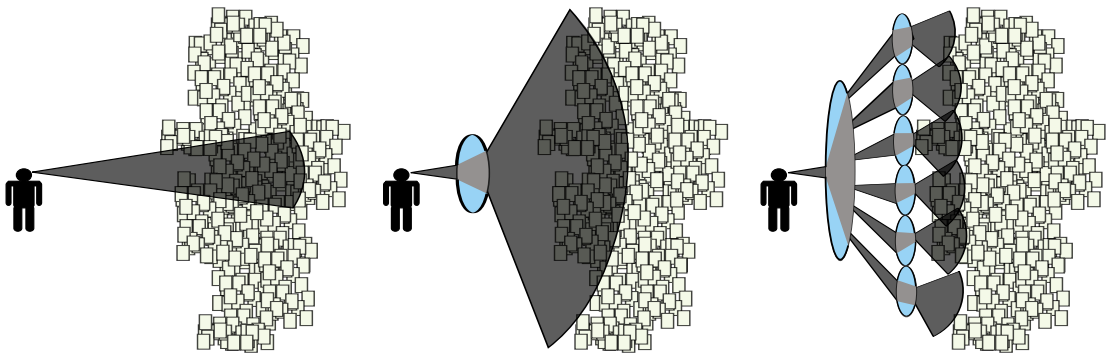


Figure 1.1: A cartoon illustration of the role of statistical models in large-scale data analysis. Left: large data collections are too large to handle without special tools. Center, Right: statistical models serve as lenses which can be nested, adjusted, and custom-designed to glean latent structure from large or complex datasets. Our statistical assumptions define the shape and optical characteristics of these lenses, and fortunately many of these lenses can be re-used.

as building blocks in later chapters: tools for working with text data, time-series data, and dyadic data. This chapter also provides a high-level introduction to the algorithms we will use for Bayesian inference.

Identifying influential documents. After introducing the foundations of this thesis, I will start with high-level social science questions. In Chapter 3, I look at a common challenge in analyzing text collections: that of finding the most important and influential documents in a corpus which has grown over time. This is a challenge in understanding collections of academic articles, legal opinions, email archives, and many other collections. This question even motivated the algorithm behind Larry Page and Sergey Brin's PageRank algorithm, which recursively measures the influence of Webpages, as measured by the hyperlinks between Webpages (Garfield, 1992; Brin and Page, 1998; Garfield, 2002). Unlike Web pages and academic articles, of course, explicit citations or hyperlinks are unavailable, and researchers only see the most basic metadata: documents' timestamps. To this end, I will introduce a model for discovering the most influential documents in such a collection. I will validate this model on a set of several datasets, including several collections of academic articles and a set of opinions written by judges in the New York Appellate Courts system.

Inferring history from a collection of newspaper articles. In Chapter 4 I will zoom in a bit to consider the story within a collection of documents and outline a model to better understand the relationships between pairs of countries over time. I will fit this to a collection of *New York Times* articles and demonstrate that this method discovers a more sophisticated latent story among documents than in Chapter 3. As with the method in Chapter 3, this collection has only the text

of these articles, which I augment with external information such as human labels of sentiment. In this chapter I also incorporate important ideas from the field of dyadic spatial models, which can play a role in modeling various social science phenomena.

Inferring lawmakers' preferences. In Chapters 5 and 6 I will take an even closer look at how documents can be used to better understand how congresspersons vote on bills. I will address two important limitations of ideal point models (the state of the art in spatial voting models) by using the text of bills. One of these limitations is that ideal points cannot be used to predict lawmakers' votes on heldout bills. In Chapter 5, I will introduce several models for predicting votes by lawmakers on previously-unseen bills. I will demonstrate that we can predict lawmakers' votes with high accuracy given their prior voting record and the text of the bills on which they vote.

In Chapter 6 I will address a second shortcoming of ideal point models: the limitation of a one-dimensional latent space. I will do this by using a topic model to identify those issues up for vote in an item of legislation. I will demonstrate that legislators' votes can be better modeled and better understood by describing these lawmakers' positions on different issues.

These models contrast with those in Chapters 3 and 4 in that I ignore documents' timestamps. However, I will use many of the same ideas from these earlier chapters, including mixed-membership models of text and latent-space models, in which I assume that pairs of items interact (in this case a lawmaker and a bill), and that text documents attached to those interactions can provide insight into the interaction.

Additional materials. In Appendix A I discuss details of a new variational inference algorithm which is used in Chapter 6. This Appendix can be treated as a stand-alone contribution of this thesis, making a quantitative rather than a substantive contribution. I save this contribution for the appendix in part to stress my fundamental belief that model development and model implementation can be treated separately (or should be treated as separate whenever possible to enable practitioners to do their magic), and because I believe that this thesis will appeal more broadly if it is not bogged down with mathematical baggage. I provide additional supplementary information for the remaining chapters in Appendix B.

Chapter 2

Preliminary material: quantitative methods

The work in this thesis builds upon the foundations built by decades of research in the development of machine learning. In this chapter, we will describe enough of these foundations for the reader to understand later chapters. Some of this work builds off of general knowledge in the machine learning community; when the foundational work is beyond the scope of this introduction, we will provide references to well-known resources in the community.

In this chapter, we will outline the basic methodology for probabilistic modeling in datasets. We begin by discussing a "data analysis pipeline" so the reader will understand what is meant by phrases like "the data", "fit the model", and "heldout log-likelihood", and where it falls in the overall research pipeline. We then provide basic definitions from the field of probabilistic modeling and illustrate these ideas with models that will be used as building blocks in later chapters.

2.1 Standards and naming conventions

We begin by outlining naming and variable conventions in this work. Random variables and their instantiations are given by Roman or Greek characters; the role of a variable will typically be evident from its context. Multivariate random variables such as vectors are given by boldface, and collections of random variables are sometimes given by uppercase Roman characters. For the reader's convenience, Table 1 provides many of the variables used in this work.

When we refer to a variable, we will sometimes subscript it with multiple indices. For example,

in the next chapter, we will refer to the probability $\beta_{t,k,n}$ of word n in topic k at time t. We sometimes refer to only a subset of these indices. In such cases, we are referring to the appropriately-shaped variable: $\beta_{t,k,n}$ is a scalar; $\beta_{t,k}$ and $\beta_{k,n}$ are vectors; β_t and β_k are matrices; and β is a three-dimensional tensor. In the interest of brevity and clarity, the shape of such variables will be understood from context.

2.2 Latent-variable models, prediction, and exploration

2.2.1 Data analysis pipepline

We will develop the ideas outlined in the last chapter by using the "data analysis pipeline" illustrated ins Figure 2.1.[1] This pipeline, which is driven by specific questions about a set of data, serves as a recipe for answering questions about these data. It will also help to make the contributions of this thesis more explicit. The pipeline has the following steps (this thesis focuses on those steps which are colored blue):

1. **Questions.** One of the first, most critical steps is defining the question at hand. In our case, the questions include "Which articles in a given collection are the most influential?" and "How will the Florida Senator vote on pending legislation?"

2. **Data.** At the same time we are formulating questions, we must also understand which data is available to answer the question at hand. The questions we ask will be informed by the data available to us, and vice-versa.

3. **Modeling assumptions.** Once we have established a set of questions and available data, we define a set of assumptions that will allow us to capture statistics of interest.

   This step arguably allows the practitioner (i.e., someone designing, fitting, and analyzing a model) wide latitude in defining variables of interest. In following chapters, we will spend a lot of time discussing modeling assumptions, and they compose one of the biggest contributions of this work.

4. **Model implementation.** In this stage, the model is defined, and the practitioner must encode these modeling assumptions into an algorithm and run that algorithm. We will variously refer to this stage of the process as *fitting a model*, *performing inference*, and *fitting the pos-*

[1] This pipeline is very much inspired by discussions with David Blei, and I imagine he should get credit for it.

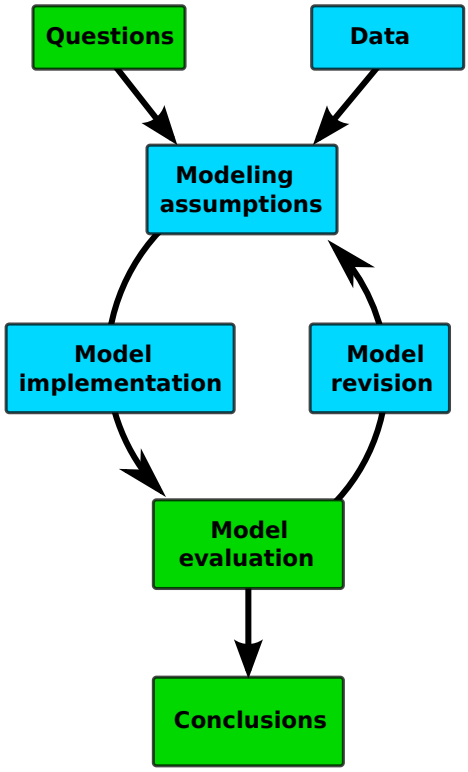


Figure 2.1: A data analysis pipeline. In this work, we make contributions in the areas of **modeling assumptions**, **model implementation**, and **model revision**. We will focus on applications which use text data.

terior. This step also represents a significant contribution of our work, in Chapters 3-6 and Appendix A.

5. **Model evaluation.** The goals of this stage are to evaluate performance of the model and to criticize the model. The criticism may warrant model revision, in which case the modeling assumptions are adjusted, the model is refit, and the model is re-evaluated.

6. **Conclusions.** Finally, the practitioner may draw conclusions from the model. In our case, this leads to two applications: *exploration* and *prediction*. Note that this step may better hang off of *model implementation*.

As alluded to above, this work will focus on modeling assumptions and model implementation. To encode our assumptions, we will use latent variable models.

2.2.2 Latent-variable models

To formalize what we mean by *Modeling assumptions*, we will assume that observed data can be described by a probability distribution. By making this assumption, we will gain several benefits,

which we outline at the end of this section. First, we formalize a *latent variable model*. A latent variable model can be fully specified with

- A set of latent random variables $X_1, \ldots, X_{M_x}$ ($X_{1:M_x}$ for shorthand);
- A set of observed random variables $Y_1, \ldots, Y_{M_y}$ ($Y_{1:M_y}$ for shorthand);
- A joint probability distribution $p(X_{1:M_x}, Y_{1:M_y})$.

As p is a probability distribution, it satisfies

$$\int_{x_{1:M_x}, y_{1:M_y}} p(X_{1:M_x}, Y_{1:M_y}) dx_{1:M_x}, dy_{1:M_y} = 1.$$

For p to be useful, we typically will make distributional assumptions about it. We often describe assumptions about factorization using a *directed graphical model* (sometimes called a Bayesian network) (Pearl, 1985).[2] A graphical model is a directed, acyclic graph $G = (V, E)$ in which vertices $V = 1, \ldots, M = M_x + M_y$ represent random variables and edges connote dependence.

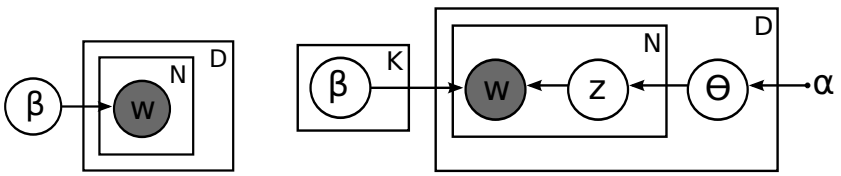
We state this more precisely be defining the "parents" function $\pi_G : \{1, \ldots, M\} \to 2^{\{1,\ldots,M\}}$, which takes the random variable index $m \in \{1, \ldots, M\}$ to its parents $\{i : i \neq m \text{ and } (Z_i, Z_m) \in E\}$. By definition of a graphical model, a probability distribution described by a graphical model G can be factorized as

$$p(Z_1, \ldots, Z_M) = \prod_{i=1,\ldots,M} p_\theta(Z_i | \{Z_j : j \in \pi_G(i)\}). \tag{2.1}$$

Note that there is a many-to-many relationship between graphical models and probability distributions. Each graphical model may describe many different distributions, but all such distributions must be factorizable based on this graphical model. Conversely, each probability distribution can be described by multiple graphical models, but each distribution must factorize according to Equation 2.1 for all of its corresponding graphical models. The language of graphical models makes it possible to succinctly describe many joint probability distributions, and it makes model implementation and inference with these models much easier to discuss formally.

Conventionally, a graphical model G is often drawn as a block-and-arrow diagram, where we write out the graph with each random variable (vertex) drawn as a circle and each edge drawn as an arrow. An additional convention in these diagrams is that boxes, or *plates*, represent replication (with the number of replications shown in a corner of the plate). In the graphical model shown in

[2] Undirected graphical models are also useful. Here we focus on directed graphical models.



Unigram language model Latent Dirichlet allocation

Figure 2.2: Left: graphical model for a unigram language model. Documents $1, \ldots, D$ are treated as *bags of words*, or collections of words w_n. Right: graphical model for Latent Dirichlet Allocation. Circles are random variables, arrows connote dependency, and plates represent replication. The shaded circles represent observed random variables (words in this case).

Figure 2.2 (right), for example, one corresponding factorization (remember that there are many) is

$$\prod_{K} p(\beta_k) \times \prod_{d=1,\ldots,D} p(\theta_d|\alpha) \prod_{n=1,\ldots,N_d} p(z_n|\theta_d)p(w_n|z_n, \beta_{z_n}), \tag{2.2}$$

where we have used the further convention that observed random variables are shaded and hidden random variables are unshaded. Later in this section we will describe this exact model in more detail.

We will sometimes point out conditional independence between groups of random variables. Given random variables Z_1, Z_2, and Z_3, we say that Z_1 is conditionally independent of Z_2 given Z_3 if $p(Z_1, Z_2|Z_3) = p(Z_1|Z_3)p(Z_2|Z_3)$. Conditional independence statements about distributions can be inferred from these distributions' factorizations (or from the graph itself) and become important when one implements a probabilistic model or makes predictions with one. Conditional independence will permeate much of what we discuss in this thesis.

Before proceeding further, it is worth noting the benefits in using latent-variable models. Several of the most compelling motivations are:

1. Flexibility. These models can describe, summarize, and explain a wide variety of phenomena in the physical and social sciences.

2. Embeddability and interpretability. Any quantifiable metric in the dataset can be encoded as a random variable in a probabilistic model. Relationships found within datasets can be likewise encoded explicitly.

3. Modularity. Parts of these models can be re-used across different models. This leads to efficient transfer of resources and common paradigms.

4. Existing toolbox of statistical tools. There is a large and growing body of literature around

how to fit these models, and there are many widely supported packages for fitting these models Bishop (2006). Practitioners no longer need to be experts in statistics to correctly apply many of these tools.

5. Implementation convenience. Latent-variable models provide explicit objective functions. Once a latent-variable model is selected, implementing and fitting it may be a (mostly) solved problem. Over the next couple of decades, increasingly sophisticated and powerful tools will be developed to make general-purpose model-fitting much easier.

The risk with applying latent-variable models is that the credibility and careful deliberation we often associate with statistics lends credence to the results of fitting a model. This may lead researchers to be overconfident in the conclusions they draw from their models, particularly when the model is incorrectly interpreted, when the data is poorly fit by the model, or when the model is poorly defined.

2.2.3 Text as a medium for social science analysis

We first illustrate these ideas in an application of text modeling. As noted in the last chapter, text data is as easy to work with as it is ubiquitous. Importantly, researchers and other practitioners are becoming more proficient with tools for text analysis. Grimmer and Stewart (2012) provide an excellent overview of methods for analyzing text for social scientists; we will summarize several such methods here.

Text data is extremely high-dimensional. A large collection of documents represented by a sequence $\boldsymbol{w}_n$ of words is unweildly for even the most powerful computer. A number of tools have been developed over the past several decades to simply find the *gist* of documents, making it possible to describe collections succinctly and efficiently.

In this work we will use the simplifying assumption that each text document is described by a vector $\boldsymbol{w}_d \in \mathbb{R}^V$ of word counts. This assumption, known as the *bag of words* assumption, removes most of the information in a document (here we use "information" in a very loose sense). At the same time, this assumption still allows us to capture the "gist" of a document very well. One of the simplest bag of words models is the unigram model. In the unigram model, every word is assumed to come from some multinomial distribution β over the vocabulary:

$$p(w_{11}, \ldots, w_{ND}) = p(\beta) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p(w_{n,d}|\beta),$$

where D is the number of documents and N_d is the number of words in document d. We illustrate this model graphically in Figure 2.2 (left). The bag-of-words assumption in particular is illustrated by the model's agnostic treatment of the order between words: these words are fully exchangeable within each document.

Latent Dirichlet Allocation

We will capture the gist of documents using the topic model Latent Dirichlet Allocation (Blei *et al.*, 2003). Latent Dirichlet allocation (LDA) posits a set of K *topics* $\beta_1, \ldots, \beta_K$ to formalize what we mean by the *gist* of a document. LDA describes each document d as a mixture $\boldsymbol{\theta}_d$ of K topics, where $\sum_{k=1}^{K} \boldsymbol{\theta}_{dk} = 1$ and $\boldsymbol{\theta}_{dk} \geq 0$ for all d, k.

Formally, we can represent this using a *generative process* for the creation of documents. The generative process can be interpreted as a recipe for creating the observations—documents, in our case—in a way that fully specifies the joint probability distribution of all random variables:

1. Draw topics $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \sim \text{Dir}(\eta, \ldots, \eta)$.
2. For document $d = 1, \ldots, D$:
   (a) Draw topic mixture $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha, \ldots, \alpha)$.
   (b) For term $n = 1, \ldots, N$:
      i. Draw topic indicator $z_n \sim \text{Mult}(\boldsymbol{\theta}_d)$.
      ii. Draw word $w_n \sim \text{Mult}(\boldsymbol{\beta}_{z_n})$.

The parameter $\alpha > 0$ above is a Dirichlet prior (it is often set by topic model researchers to $1/K$). The distribution $\text{Dir}(\alpha_1, \ldots, \alpha_M)$ refers to the Dirichlet distribution. Its density is given by

$$p(x_1, \ldots, x_M | \alpha_1, \ldots, \alpha_M) = \frac{\Gamma\left(\sum_{i=1}^{M} \alpha_i\right)}{\prod_{i=1}^{M} \Gamma(\alpha_i)} \prod_{i=1}^{M} x_i^{\alpha_i - 1}. \tag{2.3}$$

We illustrate the graphical model for LDA in Figure 2.2. Given the graphical model, we can immediately write the joint distribution of a collection of D documents as

$$p(\boldsymbol{\beta}_k, \boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{W} | \alpha) = \prod_{K} p(\boldsymbol{\beta}_k) \prod_{D} p(\boldsymbol{\theta}_d) \prod_{N} p(z_n | \boldsymbol{\theta}_d) p(w_{n,d} | z_n, \boldsymbol{\beta}_{z_n}), \tag{2.4}$$

where W represents the collection of all random variables $w_{n,d}$, and where $p(\boldsymbol{\beta}_k)$ and $p(\boldsymbol{\theta}_d)$ are understood to be conditioned on η and α respectively. We treat α, η as hyperparameters and omit them so they're not confused with random variables.

Table 2.1: Example topics from Latent Dirichlet Allocation fit to sentences from the the textbook *Biology* by Campbell and Reece. This is a small subset of the 1000 topics. (These topics were provided by Ricky Wong.)

virus	forest	population	dinosaurs
viruses	diversity	growth	dinosaur
viral	plant	rate	birds
host	ectomycorrhizal	age	pterosaurs
phage	fungi	rates	cretaceous
rna	fungal	population growth	bird
genome	treatment	populations	long
infection	emf	life	flight
cell	effects	mortality	feed

In the vast majority of cases, the practitioner observes the words of a set of documents and seeks to learn the topics that describe these documents. Before describing how to fit such a model, however, we point the reader to the four example topics from LDA in Table 2.1. Note that some of these "words" are instead phrases. This can be done by creating a vocabulary of phrases instead of words and describing documents as bags of phrases. We describe how to select phrases for a vocabulary in Appendix B.7.5.

Inference

Of course, we only observe the words $\boldsymbol{Z}$ in a collection of documents, and we are interested in estimating what the topics β and topic mixtures $\boldsymbol{\theta}$ are. We will generally accomplish this with posterior inference, in which we aim to estimate the posterior distribution

$$p(\beta, \boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{W}) = \frac{p(W|\beta, \boldsymbol{\theta}, \boldsymbol{Z})p(\beta, \boldsymbol{\theta}, \boldsymbol{Z})}{p(\boldsymbol{W})}. \tag{2.5}$$

This conditional distribution is impossible to compute efficiently because of the intractable normalizing constant

$$p(\boldsymbol{W}) = \int_{\beta_k} p(\beta_k) \prod_{D} \int_{\boldsymbol{\theta}_d} p(\boldsymbol{\theta}_d|\alpha) \prod_{n=1}^{N} \sum_{K} p(z_{n,d} = k|\boldsymbol{\theta}_d) p(w_n|z_{n,d} = k, \beta_k) d\beta d\boldsymbol{\theta} dz. \tag{2.6}$$

This intractability is common during posterior inference. In Section 2.3.3 we will see details on ways to get around this intractable integral by approximating the posterior Blei *et al.* (2003).

2.2.4 Matrix factorization, latent space models, and multidimensional scaling

Two of the most common primitives in latent-variable models are probabilistic matrix factorization (Salakhutdinov and Mnih, 2008) and multidimensional scaling, which describe relationships between pairs of items, or *dyads*. We first discuss a specific application of matrix factorization called item response theory (IRT), which has been used for decades in political science (Clinton *et al.*, 2004; Martin and Quinn, 2002; Poole and Rosenthal, 1991; Enelow and Hinich, 1984; Albert, 1992).

In IRT, we have two types of objects, and we would like to make predictions about pairs of them. Each of these objects—suppose that they are lawmakers and bills to be concrete—is represented by real-valued random variables: lawmaker $u \in \{u = 1, \ldots, U\}$ has a latent value $X_u \in \mathbb{R}$, and each bill $d \in \{1, \ldots, D\}$ has two latent values $A_d, B_d \in \mathbb{R}$. We make predictions about pairs of them by introducing the likelihood function $p(V_{ud} = 1 | X_u, A_d, B_d) = \sigma(x_u a_d + b_d)$, where $\sigma(s) = \frac{\exp(s)}{1+\exp(s)}$. We illustrate this graphically in Figure 2.3.

In this model, b_d serves as an intercept describing whether the bill is popular or unpopular, independent of the lawmaker voting on it. a_d serves as an indication of how polarizing the bill is, and x_u interacts with a_d to describe the lawmaker's position on bill d. We will look at this model in more detail later.

More formally, we can write $\{V\}_{ud}$ as a matrix of probabilities that boolean random variables (e.g., votes) are true, factorized as

$$\hat{\sigma}\left(\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_U & 1 \end{bmatrix} \begin{bmatrix} a_1 & \cdots & a_D \\ b_1 & \cdots & b_D \end{bmatrix}\right), \tag{2.7}$$

where the matrix operator $\hat{\sigma}(\cdot)$ produces a matrix in which the scalar logistic function $\sigma(s) = \frac{\exp(s)}{1+\exp(s)}$ is applied to each element of its argument.

A wide variety of researchers have used formulations like this for applications such as recommendation and representing the votes of lawmakers (Wang and Blei, 2011; Salakhutdinov and Mnih, 2008; Poole and Rosenthal, 1985, 1991; Clinton *et al.*, 2004). In later chapters, we will use it for models of legislative voting.

Sometimes these pairs of items that interact in dyads are of the same "type", and we wish to model them in the same latent space. Instead of bills and lawmakers interacting, for example, we will consider in Chapter 4 pairs of countries that interact, and we wish for these countries to be

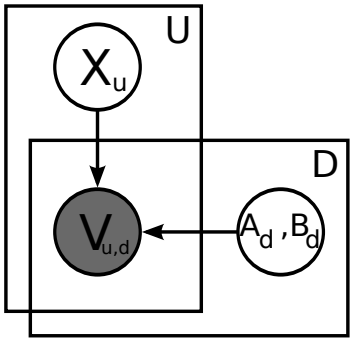


Figure 2.3: Probabilistic matrix factorization. We observe interactions V_{ud} between users represented by X_u and items represented by A_d, B_d.

represented in the same latent, interpretable space. In this case, we will still model each country with a latent position vector $\boldsymbol{x}_u$, and we will model their interaction as above, with

$$p(v|\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma(v|\boldsymbol{x}_i^T \boldsymbol{x}_j, 1),$$

for a suitable distribution σ. We will also frame their relationship using their Euclidean distance in this latent space:

$$p(v|\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma(v|\boldsymbol{\beta} w_{ij} - \log(||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2 + 1), 1),$$

where w_{ij} are observed covariates about the dyad and $\boldsymbol{\beta}$ is hidden along with $\boldsymbol{x}$ (Hoff *et al.*, 2002). We will motivate these expressions and others like them when we develop a model of foreign relations in Chapter 4.

2.2.5 Hidden Markov Models and Kalman Filters

We now turn briefly to abstractions for time-series data. One of the simplest assumptions about a time-series collection is that we have a sequence of observations $Y_1, \ldots, Y_T$ observed at times $t = 1, \ldots, T$. In a *hidden Markov model* (HMM). We assume that these observations can be explained by a hidden set of states $X_1, \ldots, X_T$, which are temporally linked. The model factorizes as

$$p(Y_1, \ldots, Y_T, X_1, \ldots, X_T) = p(X_1)p(Y_1|X_1) \times \prod_{t=2,\ldots,T} p(X_t|X_{t-1})p(Y_t|X_t) \tag{2.8}$$

(see Figure 2.4 for the graphical model). Often the transition distribution $p(X_t|X_{t-1})$ is independent of t (the chain in this case is called *time-homogeneous*). A wide variety of problems can be modeled accurately with a well-selected homogeneous HMM. Importantly, inference in these models is very

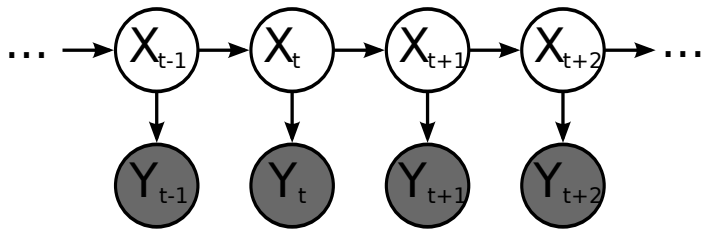


Figure 2.4: A hidden Markov model. Observations $Y_1, \ldots, Y_T$ are observed at discrete times $t = 1, \ldots, T$, and are conditionally independent given the hidden states $X_1, \ldots, X_T$.

efficient because the set of conditional independencies yields a tree: inference can usually be reduced to an application of a forward-backward algorithm, especially when the conditional distribution of each variable given its neighbor is conjugate (even when this is not the case, methods such as Paisley, Gerrish, and Blei (2010) provide approximate ways to perform inference on non-tree graphical models . One of the most famous algorithms for inferring the states of a hidden Markov model is the Kalman filter, which assumes linear (or quadratic) transitions between the states X and Gaussian noise: $p(Y_t|X_t) \propto \mathcal{N}(X_t, \sigma^2)$ for some variance σ^2.

We will use these time-series abstractions in modeling time-series collections of documents. In these collections, the assumption of a hidden, evolving state will allow us to perform inference efficiently while inferring a sequence of states which can be interpreted—for example, we will use this to model themes which evolve over time in Chapter 2 and to infer countries' positions about foreign policy issues in Chapter 3.

We have discussed text and time-series assumptions, which are often seen together in the context of natural language processing. We will not use time-series assumptions at the level of syntactic language modeling. While sequential modeling is useful for many NLP tasks, we will not use them in this work, instead deferring to the bag-of-words assumption described in Section 2.2.3.

2.3 Posterior inference and model evaluation

One of the most fundamental problems in statistical machine learning is that of estimating the values of latent random variables X in a statistical model, given observed random variables Y (i.e., data). In this thesis, we will frequently need to estimate the posterior distribution $p(x|y) = \frac{p(x,y)}{p(y)}$. In this section we outline several common methods for estimating this posterior.

2.3.1 MAP estimation

One of the simplest estimates of the value of a random variable is the *maximum-a-posteriori* (MAP) estimate. The MAP estimate $\hat{X}$ is defined to be the most-likely value of the random variable:

$$\hat{X} = \arg\max_x p(X = x|Y) = \arg\max_x \frac{p(X = x|Y)p(Y)}{p(Y)} = \arg\max_x p(X = x, Y). \tag{2.9}$$

The MAP estimate can typically be found by performing gradient or coordinate ascent on $p(X, Y)$ with respect to X (this is because the normalizer $p(Y)$ is not a function of X). Because MAP estimates can be fast to estimate, they can shorten the development loop described in Section 2.1. The MAP provides a point estimate which is often a good summary of the posterior distribution.

2.3.2 MCMC

We briefly review the key components of Markov Chain Monte Carlo (MCMC) estimation. We will not go into detail about MCMC in this work except to build up to (and draw a contrast with) variational methods, which are introduced in the next section. Readers unfamiliar with MCMC can refer to a standard text such as Bishop (2006).

MCMC methods are often used to inspect a posterior distribution $p(x|y)$. The input to an MCMC sampler is typically an unnormalized probability density $\tilde{p}(x, y) \propto p(x|y)$. [3] Given $\tilde{p}(x, y)$, an MCMC sampler produces a collection of samples from $p(x|y)$. These samples are often used to summarize statistics such as marginal means and variances of $p(x|y)$. They are unbiased and, given enough time, will accurately represent $p(x|y)$.

MCMC methods are used widely, but they have limitations. One of these limitations is runtime: while one may need N *iid* samples from a distribution $p(x|y)$ to estimate its mean and variance, she typically needs many more MCMC samples to estimate these statistics. In some MCMC algorithms, one must select a proposal distribution for sampling; a poorly-chosen proposal distribution can affect runtime, as a Markov chain needs more samples to converge. MCMC algorithms can also suffer from memory bottlenecks, as samples are stored and convergence is measured.

Even when memory is not a bottleneck, the practitioner is often interested in only the marginals of the posterior (as with most mixture-of-Gaussian applications); a large number of discarded samples indicates that there is an inefficiency in the inference pipeline.

[3] For numerical and algebraic convenience, $\tilde{p}(x, y)$ is often specified by $\log \tilde{p}(x, y)$.

2.3.3 Variational inference

Variational methods address some of the shortcomings of MCMC by providing a fast, deterministic alternative to MCMC (Wainwright and Jordan, 2003; Jordan *et al.*, 1999). These algorithms have been successfully applied to many kinds of topic models, where corpus size and vocabulary dimension are large. We review the key ideas of variational inference here for use in later chapters.

Variational methods posit a simplified[4] family of probability distributions, indexed by variational parameters ν, and select the member q_ν of this family that is closest in KL-divergence to the true posterior $p(x|y)$:

$$\arg\min_{\nu} \mathrm{KL}(q_\nu||p) = \arg\min_{\nu} \int_x q_\nu(x) \log \frac{q_\nu(x)}{p(x|y)} dx. \tag{2.10}$$

Finding the optimal variational distribution q_ν is equivalent to optimizing an "evidence lower bound" (ELBO) ($\mathcal{L}_\nu$) on the data likelihood

$$\begin{aligned} \log p(y) \geq & \quad \mathbb{E}_q \left[\log p(x,y) - \log q_\nu(x)\right] & (2.11)\\ = & \quad \mathbb{E}_q \left[\log p(x,y)\right] - H(q_\nu(x)) & (2.12)\\ =: & \quad \mathcal{L}_\nu, & (2.13) \end{aligned}$$

where $H(q_\nu(x))$ is the entropy of that distribution and the slack of the bound is equal to the KL divergence from Equation 2.10.

The family is chosen by the practitioner to make the resulting algorithm tractable and to capture the parameters of interest. A common assumption is that the posterior is fully-factorized into simple marginal distributions; such an assumption is known as *naive mean-field variational inference*. Though simpler, the fitted variational distributions are found to be good proxies for the true posterior (Jordan *et al.*, 1999; Gerrish and Blei, 2011).

For example, a multivariate posterior $p(x|Y), x \in \mathcal{R}^D$ might be represented by the product $q(x_{1:D}) = \prod_D \mathcal{N}(x_d|\mu_d, \sigma_d^2)$ of D Gaussian distributions, and a multinomial posterior might be represented by a Dirichlet distribution (Bishop, 2006). In the case of Latent Dirichlet Allocation, for example, Blei *et al.* (2003) assume that the indicators z_n can be described by a fully-factorized product of multinomial distributions, and they assume that the posterior distribution of topics β and mixture proportions $\boldsymbol{\nu}$ can be represented by a fully-factorized product of Dirichlet distributions.

Once a family is selected, the bound in Equation 2.13 is evaluated symbolically, as a practitioner

[4]Simplified compared to the true posterior.

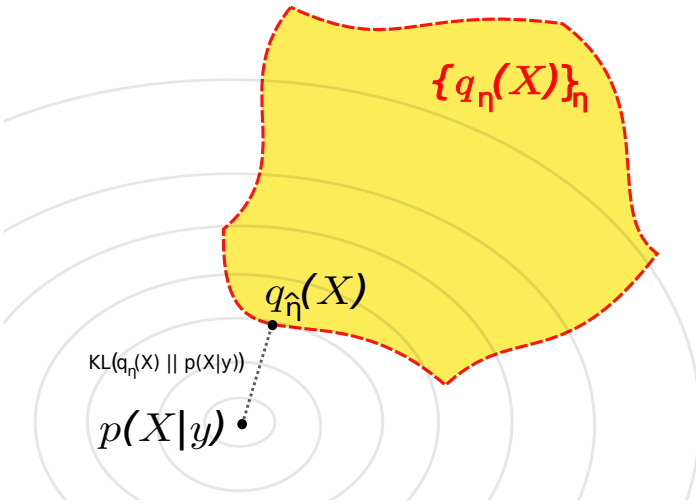


Figure 2.5: Illustration of variational inference. Practitioners define a variational family (shaded yellow region) and find the member of that family $q_{\hat{\eta}}(x)$ which is closest (by KL divergence) to the true posterior.

fully expands $\mathbb{E}_q \left[\log p(x, y) - \log q_\nu(x)\right]$ and (usually) its gradient using pencil, paper, and algebra. As we will show in subsequent chapters, this bound may itself be bounded or approximated with a Taylor approximation such as the delta method (Bickel and Doksum, 2007; Braun and McAuliffe, 2010). These simplifying assumptions – an approximate, fully factorized posterior with further simplifying bounds – make it possible to express the lower bound in terms of the variational parameters ν. The practitioner then uses these bounds and gradients in a coordinate or gradient ascent algorithm. This process, and the role of variational inference in statistical machine learning will become more clear as we develop several algorithms using these methods over the next few chapters.

We have not yet described a limitation of variational inference: with each new set of model assumptions, variational inference requires that the variational lower bound $\mathcal{L}$ be algebraically evaluated, which is a significant time investment by a practitioner for each new model she creates. We will also introduce an alternative method for performing variational inference in Appendix A. This alternative method removes the onus of deriving new variational update equations, making it easy for the practitioner to perform rapid model development on a range of models.

2.3.4 Model evaluation

After a model has been fit with an approach such as variational inference, it is important to evaluate the model (see again the data analysis pipeline in Section 2.2.1). The goal of model evaluation is to evaluate performance of the model and to criticize the model. The criticism may warrant model

revision, in which case the modeling assumptions are adjusted, the model is refit, and the model is re-evaluated.

In general, different practitioners will have different goals in modeling data, so they will have different goals in model evaluation. However, several standard approaches exist.

Likelihood of training data

We first describe one of the simplest metrics of how well a model is fit: its ability to model training data $Y_{\text{obs},1}, \ldots, Y_{\text{obs},N_{\text{obs}}}$. As before, one of the most frequently used metrics for this is the training log-likelihood $\log p(Y_{\text{obs},1}, \ldots, Y_{\text{obs},N_{\text{obs}}})$ of these observations. When these observations are conditionally independent given the observed data (nearly always the case), the log-likelihood can be written $\sum_{n=1}^{N_{\text{obs}}} \log p(Y_{\text{obs},n})$. The downfall of this metric is of course that it does not measure whether a model is overfit. However, it is usually the objective function used to define a stopping criterion when an MAP or MLE estimate is fit. In addition, it can be used to measure the "flexibility" of a model.

Likelihood of heldout observations $Y_1, \ldots, Y_{N_{\textbf{heldout}}}$

A common measure of a model is its ability to represent unseen, heldout observations $Y_{\text{hdt},1}, \ldots, Y_{\text{hdt},N_{\text{hdt}}}$ given a set of "training" observations $Y_{\text{h},1}, \ldots, Y_{\text{h},N_{\text{obs}}}$ One of the most frequently used metrics for this is the log-likelihood

$$\log p(Y_{\text{hdt},1}, \ldots, Y_{\text{hdt},N_{\text{hdt}}} | Y_{\text{obs},1}, \ldots, Y_{\text{obs},N_{\text{obs}}})$$

of these observations, where we condition on the observed data because a "fit" model effectively captures the information in these observations. When these observations are conditionally independent given the observed data (nearly always the case), the log-likelihood can be written $\sum_{n=1}^{N_{\text{hdt}}} \log p(Y_{\text{hdt},n} | Y_{\text{obs},1}, \ldots, Y_{\text{obs},N_{\text{obs}}})$.

When training data is scarce, practitioners often use k-fold cross-validation. Cross-validation requires that the training data be partitioned into K equal-size parts $P_1, \ldots, P_K$, fit K times on each of the $K-1$ subsets which omit exactly one of the partitions, and evaluated on the omitted partition. Such a model has the benefit that all of the data is used for training and evaluation.

Relationship with external data

In this thesis we will use external data sources to validate the results of our model. External data sources are useful because they can confirm a model's assumptions. When a model's inferences do not correspond with a secondary data source, then this can also be useful, because it can be used to inform model development or to confirm that the model can produce new, useful results.[5] We will find this, for example, when we compare influence scores learned from our model in Chapter 3 with citation counts.

2.4 Using these tools to understand influence and decision-making

The ideas outlined in this chapter cover only the tip of the iceberg of tools used by statistical machine learning researchers (and this is only a subset of all machine learning researchers). However, they will serve as important building blocks for future chapters. In the next chapter we will use some of these ideas as we return to the original questions that motivated this thesis: how can we understand patterns of behavior in society using text? How do documents interact with one another, and how can we use them to tell us how people interact with the world?

We will use the tools discussed in the preceding sections to shed light on these questions, and we will do this with exactly the data-analysis recipe that we outlined in Section 2.2.1. This recipe involves defining a question, describing a model to answer that question using data, fitting that model, and drawing inferences using the model.

In the next chapter we will return to a fundamental challenge in managing the huge volumes of text now inundating researchers and companies: how to find the most important and influential documents in a collection. As we will show in the next chapter, a latent-variable treatment of this question will allow us to make our assumptions explicit. This in turn will in turn make the subsequent analysis straightforward.

[5] This was pointed out in a helpful discussion with Matthew Salganik.

Chapter 3

A method for discovering influence in text documents

A fundamental problem in research and industry is that of organizing collections of documents. In many cases this problem can be reduced to identifying those documents which have been the most influential. This is an important and common problem in many fields, including research in academic fields such as political science, history, and science. Influence measurements are used to assess the quality of academic instruments, such as journals, scientists, and universities; as such, they can play a role in decisions surrounding publishing and funding. These measurements are critical for academic researchers: finding and reading the influential articles of a field is central to good research practice.

Measurements of influence are also significant in industry, as regulations such as Sarbanes Oxley require public companies to retain documents. E-discovery is another field in which identifying influential documents is critical. A recent New York Times article cited the need for such tools in industry:

> *"The economic impact will be huge," said Tom Mitchell, chairman of the machine learning department at Carnegie Mellon University in Pittsburgh. "We're at the beginning of a 10-year period where we're going to transition from computers that can't understand language to a point where computers can understand quite a bit about language."* (Markoff, 2011).

The article continues, noting that recent solutions use either keyword-based search methods or take advantage of metadata such as citations or links in emails, which can be helpful when available.

Metadata can be a boon for finding the most influential documents in a collection, but often such metadata is unavailable.

In this chapter, we will describe an approach to identifying influential articles in a collection without the use of metadata like citations. The key assumption of our method is that an influential article will affect how future articles are written and that this effect can be detected by examining the way corpus statistics change over time. We will take advantage of the tools discussed in the last chapter by using them to encode this intuition in a model to measure influence in sequential collections of documents.

Measuring influence with citations

A traditional method of assessing an article's influence is to count the citations to it. The impact factor of a journal, for example, is based on aggregate citation counts (Garfield, 2002). This is intuitive: if more people have cited an article, then more people have read it, and it is likely to have had more impact on its field. Citation counts are used with other types of documents as well. The Pagerank algorithm, for example, uses hyperlinks of web-pages to identify the most influential Web-pages on the Internet, and it was essential to Google's early success in Web search (Brin and Page, 1998). There is a large literature on these and other methods for citation analysis and bibliometrics. See Osareh (1996) for a review.

Though citation counts can be powerful, they can be hard to use in practice. Some collections, such as news stories, blog posts, or legal documents, contain articles that were influential on others but lack explicit citations between them. Other collections, like OCR scans of historical scientific literature, do contain citations, but they are difficult to read in reliable electronic form. Finally, citation counts only capture one kind of influence. All citations from an article are counted equally in an impact factor, when some articles of a bibliography might have influenced the authors more than others.

Using text to measure influence

One possible solution might be to *predict* citation counts, by proposing features and training a regression. Tang and Zhang (2009) and Lokker *et al.* (2008) have used methods like this; successful features include the publishing journal's impact factor, previous citations to last author, key terms, and number of authors (Tang and Zhang, 2009; Lokker *et al.*, 2008). Such research has had measured success: 56% explained variance (Lokker *et al.*, 2008), and 91.5% prediction accuracy (Ibáñez *et al.*,

2009).

However, we seek a model that is applicable to collections for which the notion of citation may not exist. Therefore, predicting citations is an explicit non-goal. Further, work toward predicting citations uses specialized classifiers and restrictive features for narrow application domains; Lokker *et al.* (2008) even note that their results "may not be readily transferable to... basic science articles or journals". They further noted that earlier work in their field of predicting citations to medical journals had only achieved 14% to 20% explained variance (Lokker *et al.*, 2008).

In this chapter we will use a text-based approach to measure influence. We will base our assumptions on a topic model which allows topics to drift over time in a corpus (Blei and Lafferty, 2006). Though our algorithm aims to capture something different from citation, we will validate the inferred influence measurements by comparing them to citation counts.

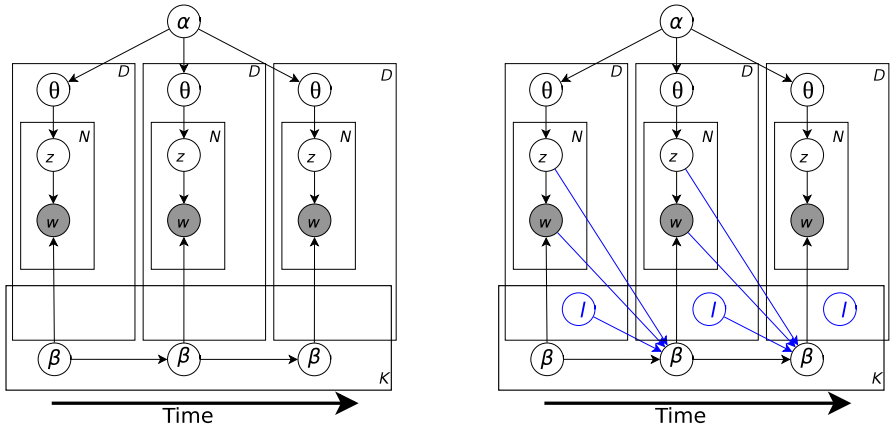
We begin with a discussion of previous work aimed at modeling influential documents. We then describe the Document Influence Model (DIM), our unsupervised model for determining the influence of a document using the changes in language used by documents over time. We follow this with experiments to compare this model with citation counts on three well-known scientific corpora and a collection of legal opinions. We will also provide the reader with an intuition for the model with several real-world examples. With only the language of the articles as input, our algorithm produces a meaningful measure of each document's influence in the corpus.

3.1 The Document Influence Model

In this section we will develop a probabilistic model that captures how past articles exhibit varying influence on future articles. The hypothesis is that an article's influence on the future is corroborated by how the language of its field changes subsequent to its publication. In the model, the influence of each article is encoded as a hidden variable; the posterior distribution of these variables (given the text of documents) reveals the influential articles of the collection.

Past approaches

A number of algorithms link the text of documents to citation counts. This work often models the information in citations by predicting them or modeling them with topics (Nallapati and Cohen, 2008; Chang and Blei, 2009; Dietz *et al.*, 2007; Cohn and Hofmann, 2001) or other semantic tools (McNee *et al.*, 2002; Ibáñez *et al.*, 2009). Other work in this area uses the text of documents along with citations to summarize documents (Qazvinian and Radev, 2008) or to propose new



(a) (b)

Figure 3.1: The Dynamic Topic Model (a) and the the Document Influence Model (b).

bibliometrics: Mann *et al.* (2006) use topic models and citations to map topics over time and define several new bibliometric measurements such as topic Impact Factor, topical diffusion, and topic longevity.

Some work in this area uses the link structure of citation networks to extract higher level structure. Borner *et al.* (2003), for example, have used author and citation networks to understand the evolution of ideas in the history of science.

Dynamic Topics

Our model is based on the dynamic topic model (DTM) (Blei and Lafferty, 2006), a model of sequential corpora that allows language statistics to drift over time. Probabilistic topic models such as LDA (introduced in the last chapter) usually assume that the underlying distribution over words is fixed (Blei *et al.*, 2003; Deerwester *et al.*, 1990; Hofmann, 1999). The DTM introduced a Markov chain of topics (i.e., term distributions) to capture probabilities that drift over the course of the collection. The idea is simple: topics drift in discrete steps over time. At each "epoch", some number of documents are generated based on topics *at that epoch*.

Drifting Topics. We can formalize these assumptions in a statistical model as in Blei and Lafferty (2006). First let V be the number of terms in a vocabulary and consider the natural parameters β_t of a term distribution at time t, where the probability of a word w is given by the soft-max transformation of the unconstrained vector,

$$p(w \mid \beta_t) \propto \exp(\beta_{t,w}). \tag{3.1}$$

The corresponding distribution over terms, i.e., the "topic," is a point on the vocabulary simplex. In the logistic normal Markov chain, this distribution drifts with the stationary Markov process

$$\beta_{t+1} \,|\, \beta_t \sim \mathcal{N}(\beta_t, \sigma^2 I), \tag{3.2}$$

where σ^2 is the transition variance.

Documents generated at time t. Now consider a corpus broken up into discrete epochs $t \in \{1, \ldots, T\}$, with D_t articles at each time t. Let $\mathbf{W}_{t,1:D}$ denote the articles as vectors of word counts, where row $\mathbf{w}_{t,d}$ of $\mathbf{W}_{t,1:D}$ represents the word counts in article d.

At each epoch t, the documents of these articles are drawn independently using the topics described by Equation 3.1. More formally, documents are generated according to the generative process

1. For time $t = 1, \ldots, T$:
   (a) For topics $k = 1, \ldots, K$:
      i. Draw topics $\beta_{k,t} \,|\, \beta_{k,t-1} \sim \mathcal{N}(\beta_{k,t-1}, \sigma^2 I)$
   (b) For document $d = 1, \ldots, D_t$:
      i. Draw topic mixture $\theta_d \sim \text{Dir}(\alpha, \ldots, \alpha)$.
      ii. For position $n = 1, \ldots, N$:
         A. Draw topic indicator $z_n \sim \text{Mult}(\theta_d)$.
         B. Draw the word for the nth term in document d according to Equation 3.1.

We illustrate the graphical model for in Figure 3.1 (a). With this model in hand and a collection of documents, one can then estimate the positions of these topics by computing the posterior distribution of the sequence of topics $\beta_{1:T}$ conditioned on the observed documents. This summarizes the corpus as a smooth trajectory of word frequencies.

The Document Influence Model

We now turn to the original problem: certain ideas are influential in the progression of a field, and we aim to discover what these ideas are (as doing so will allow us to find those documents that are influential). The text of documents will provide a window into these underlying patterns.

In our model, each article is assigned a normally distributed *influence score* ℓ_{td} for this topic, which is a scalar value that describes the influence that the dth article at time t has on the topic. The higher the influence, the more the words of the article affect how the topic drifts.

This is encoded in the time series model. The more influential a document is, the more its words "nudge" the topic's natural parameters at the next time step,

$$\beta_{t+1} \mid \beta_t, (w, \ell)_{t,1:D} \sim \mathcal{N}\left(\beta_t + \exp(-\beta_t) \sum_{d=1}^{D_t} \sum_{n=1}^{N_d} w_{t,d}\ell_{t,d}, \sigma^2 I\right), \tag{3.3}$$

where the exponential $\exp(\beta_t)$ is a vector containing the exponentiated elements of β_t. The words of an article with a high influence will have a higher expected probability in the next epoch; the words of an article with zero influence will not affect the next epoch.

The form of Equation 3.3 is no accident. Specifically, we use this equation to enforce that the increase in words' probability at each time t be proportional to the number of words across the corpus, as well as proportional to the influence $\ell_{t,d}$ of each document.

We illustrate this further by motivating the $\exp(\beta)$ with an appeal to the chain rule of calculus. Writing the unit change $\Delta_t = \sum_d w_{t,d}\ell_{t,d}$ for brevity, we have:

$$\begin{aligned}
\exp(\beta_t) = \exp(\beta_{t-1}) + \Delta_t &\iff 1 = \exp(\beta_{t-1} - \beta_t) + \exp(-\beta_t)\Delta_t \\
&\iff 1 - \exp(-\beta_t)\Delta_t = \exp(\beta_{t-1} - \beta_t) \\
&\iff \log(1 - \exp(-\beta_t)\Delta_t) = \beta_{t-1} - \beta_t \\
&\iff \beta_t = \beta_{t-1} - \log(1 - \exp(-\beta_t)\Delta_t)
\end{aligned} \tag{3.4}$$

When $\exp(-\beta_t)\Delta_t$ is small, we have that $\beta_t \approx \beta_{t-1} + \exp(-\beta_t)\Delta_t$.

We call this model the *document influence model* (DIM). Conditioned on a corpus, the posterior distribution of the topic and influence scores gives a trajectory of term frequencies and a retrospective estimate of the influence of each article. An article whose words can help explain the way the word frequencies change will have a high posterior influence score. We will show in Section 3.3 that this estimate of influence is meaningful.

Multiple topics. Corpora typically contain multiple persistent themes. Accordingly, the full document influence model contains multiple topics, each associated with a time series of distributions. Conditioned on the topics, articles at each time are modeled with latent Dirichlet allocation (LDA).

Each article exhibits the topics with different random proportions θ_d; each word of each article is drawn by choosing a topic assignment from those proportions $z_{d,n}$, and choosing a word from the corresponding topic (Blei *et al.*, 2003).

Modeling multiple topics is important to the influence model because an article might have different impact in the different fields that it discusses. For example, an article about computational genomics may be very important to biology but less important to computer science. We want to discern its influence on each of these topics separately.

As with the DTM, we posit K topic trajectories, and each document of each time point is modeled with LDA. For each document, we now associate an influence score $\ell_{d,k}$ for each topic k. Each of the K topics drifts according to an adapted version of Equation 3.2, where we restrict attention to the influence score for that topic and to the words of each document that were assigned to it,

$$\beta_{k,t+1} \mid \beta_{k,t}, (w, \ell, z)_{t,1:D} \sim \mathcal{N}\left(\beta_{k,t} + \exp(-\beta_{k,t}) \sum_{d=1}^{D_t} \ell_{t,d,k} \sum_n w_{t,d,n} z_{t,d,n,k}, \sigma^2 I\right). \tag{3.5}$$

Here, $z_{t,d,n,k}$ is the indicator that the nth word in the dth document at time t is assigned to topic k. We illustrate the graphical model for this distribution in Figure 3.1 (b).

Although we presented our model in this section with influence spanning one year, we also adapted it to accommodate an "influence envelope", where an article's influence spans W years. This provides a more realistic model of influence (Porter *et al.*, 1988), but it complicates the inference algorithm and may not be necessary, as we note in section 3.3.

To use this model, we analyze a corpus through posterior inference. This reveals a set of K changing topics and influence scores for each article and each topic. The posterior provides a thematic window into the corpus and can help identify which articles most contributed to the development of its themes.

Work with similar goals

It is worth pointing out two pieces of recent research which have similar goals. Leskovec *et al.* (2009) describe a framework for tracking the spread of memes, or ideas, in document collections, and investigate the direction in which ideas tend to percolate. Shaparenko and Joachims (2007) describe a measure of influence by modeling documents as unigram mixtures of earlier documents and use a likelihood ratio test to predict citations between documents. In contrast to this work, the DIM uses dynamic topics to explicitly model the change in *topic* language. Further, we do not attempt to model links between documents, as in Shaparenko and Joachims (2007).

3.2 Inference and parameter estimation

Our computational challenge is to compute the posterior distribution of the latent variables—the sequences of topics and the per-document influence values—conditioned on observed documents in the corpus. As for simpler topic models, this posterior is intractable to compute exactly. We therefore employ variational methods—introduced in Chapter 2—to fit this posterior.

Before proceeding further, we note that this section is particularly dense, in part because variational methods require laborious algebra. We will use variational methods again in Chapter 5, but we introduce a method in Appendix A to mitigate some of the pain of variational inference. We will apply this in Chapter 6.

To apply variational methods, we begin by specifying a variational distribution for the DIM posterior. First, the word assignments z_n and topic proportions θ_d are governed by multinomial parameters ϕ_d and Dirichlet parameters γ_d, as in LDA (Blei *et al.*, 2003); we refer to these distributions as $q(z_n|\phi_n)$ and $q(\theta_d|\gamma_d)$.

The variational distribution for topic trajectories $\{\beta_{k,1}, \ldots, \beta_{k,T}\}$ is described by a linear Gaussian chain. It is governed by parameters $\{\tilde{\beta}_{k,1}, \ldots, \tilde{\beta}_{k,T}\}$, which are interpreted as the "variational observations" of the chain. These induce a sequence of means $\tilde{m}_t$ and variances $\tilde{V}_t$. Blei and Lafferty (2006) call this a "variational Kalman filter."

Finally, the variational distribution of the document influence value $\ell_{d,k}$ is a Gaussian with mean $\tilde{\ell}_{d,k}$ and fixed variance σ_ℓ^2.

In full, the variational distribution is

$$q(\beta, \ell, z, \theta|\tilde{\beta}, \tilde{\ell}, \phi, \gamma) = \prod_{k=1}^{K} q(\beta_{k,1:T}|\tilde{\beta}_{k,1:T}) \prod_{t=1}^{T} \prod_{d=1}^{D_t} q(\theta_{t,d}|\gamma_{t,d}) q(\ell_d|\tilde{\ell}_d) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n}|\phi_{t,d,n}).$$

Using this variational family, our goal is to maximize the Evidence Lower Bound (ELBO) $\mathcal{L}$ on the model evidence of the observed words $\mathbf{W}$:

$$\begin{aligned}
\ln p(\mathbf{W}) \geq & \mathcal{L}(\tilde{\beta}, \phi, \gamma) && (3.6)\\
= & \sum_T \mathbb{E}_q \left[\ln p(\beta_t|\beta_{t-1})\right] + \sum_T \sum_{D_t} \mathbb{E}_q \left[\ln p(\ell_d)\right] + \mathbb{E}_q \left[\ln p(\theta_d|\alpha)\right] && (3.7)\\
& + \sum_T \sum_{D_t} \sum_{N_d} \mathbb{E}_q \left[\ln p(z_n|\theta_d)\right] + \mathbb{E}_q \left[\ln p(w_n|z_n, \beta_t)\right] + H(q). && (3.8)
\end{aligned}$$

Note also that the variational parameters $\tilde{\beta}, \phi$, and γ are implicit in lines 3.7 and 3.8 of the above equation because they parameterize the variational distribution q, and the expectation is taken with

respect to this distribution.

Optimizing the variational bound

This bound is optimized by variational EM, with an update schedule similar to that of Blei and Lafferty (2006):

1. For Topic $k = 1, \ldots, K$:
   (a) Update parameters $\tilde{\beta}_k$.
2. For time $t = 1, \ldots, T$:
   (a) For document $d_{1,t}, \ldots, d_{D_t,t}$:
      i. Update parameters ϕ_d, and γ_d
   (b) Update parameters $\tilde{\ell}_t$ (i.e., update $\tilde{\ell}_d$ as a block for all documents at time t),

where the variational parameters are optimized sequentially in blocks. These updates are repeated until the relative increase in the lower bound is below a threshold (which we specify in the experiments section).

Influence values. In the DIM, changes in a topic's mean parameters are governed by a normal distribution. As a consequence of this choice, updates for the influence parameters $\tilde{\ell}_{t,k}$ solve a linear regression. In this regression, documents' words at time t explain the expected topic drift $\boldsymbol{\Delta_{\beta,t,k}} = (\beta_{t+1,k} - \beta_{t,k})$, where the contributions of each document's words are given by the design matrix $X = \text{Diag}\left(\exp(-\beta_{t,k})\right)(\boldsymbol{W}_{t,k} \circ \phi_{t,k})$. ($\text{Diag}\,(x)$ refers to the matrix having the elements of x on its diagonal, and $\circ$ refers to the element-wise product.)

The parameter updates for document influence $\tilde{\ell}_{t,k}$ are defined, for each time t and each topic k, by the variational normal equation

$$\tilde{\ell}_{t,k} \leftarrow \left(\frac{\sigma^2}{\sigma_d^2}\boldsymbol{I} + \mathbb{E}_q\left[X^T X\right]\right)^{-1} \mathbb{E}_q\left[X^T \boldsymbol{\Delta_{\beta,t,k}}\right]. \tag{3.9}$$

The expectation $\mathbb{E}_q\left[X^T X\right]$ is a matrix with dimension $D_t \times D_t$. Its elements are

$$\mathbb{E}_q\left[X^T X\right]_{d,d'} = \sum_n \exp(-2\tilde{m}_{t,k,n} + 2\tilde{V}_{t,k,n})(w_{t,d,n} w_{t,d',n} \phi_{t,k,d,n} \phi_{t,k,d',n})$$

when $d \neq d'$ and

$$\mathbb{E}_q\left[X^T X\right]_{d,d} = \sum_n \exp(-2\tilde{m}_{t,k,n} + 2\tilde{V}_{t,k,n})(w_{t,d,n}^2 \phi_{t,k,d,n})$$

otherwise. The expectation $\mathbb{E}_q\left[X^T \boldsymbol{\Delta}_{\boldsymbol{\beta},\boldsymbol{t},\boldsymbol{k}}\right]$ is a D_t-dimensional matrix with elements

$$\mathbb{E}_q\left[X^T \boldsymbol{\Delta}_{\boldsymbol{\beta},\boldsymbol{t},\boldsymbol{k}}\right]_d = \sum_n w_{t,d,n}\phi_{t,k,d,n} \times (\tilde{m}_{t+1,k,n} - \tilde{m}_{t,k,n} + \tilde{V}_{t,k,n}/2) \times \exp(-\tilde{m}_{t,k,n} + \tilde{V}_{t,k,n}/2).$$

Topic proportions and topic assignments. Updates for the variational Dirichlet on the topic proportions $\theta_{d,k}$ have a closed-form solution, exactly as in LDA (Blei *et al.*, 2003); we omit details here.

The variational parameter for each word w_n's hidden topic z_n is the multinomial ϕ_n. We solve for $\phi_{n,k}$ by the closed-form updates

$$\begin{aligned}\log(\phi_{n,k}) \leftarrow & \Psi(\gamma_k) + \tilde{m}_{t,k,n} + \frac{1}{\sigma^2} w_t \tilde{\ell}_{d_n,k} \exp(-\tilde{m}_{t,k} + \tilde{V}_{t,k}/2)(\tilde{m}_{t+1,k} - \tilde{m}_{t,k} + \tilde{V}_{t,k}) \\ & - \frac{1}{\sigma^2} w_{t,n}\left[\tilde{\ell}_{d_n,k} \exp(-2\tilde{m}_{t,k} + 2\tilde{V}_{t,k})(\boldsymbol{W}_{t,n,\backslash d_n} \circ \phi_{t,n,k,\backslash d_n})\tilde{\ell}_{t,k,\backslash d_n}\right] \\ & - \frac{1}{\sigma^2} w_{t,n}^2 \exp(-2\tilde{m}_{t,k} + 2\tilde{V}_{t,k})(\tilde{\ell}_{d,n,k}^2 + \sigma_l^2), \end{aligned} \tag{3.10}$$

where Ψ is the digamma function and $\backslash d_n$ refers to the set of all documents *except* d_n. Solving the constrained optimization problem, this update is followed by normalization $\phi_{w,k} \leftarrow \frac{\phi_{w,k}}{\sum_K \phi_{n,k}}$.

3.3 Empirical study

We studied the DIM with four text corpora: three collections of scientific articles and a collection of opinions written by judges in the New York Appellate Court system. For each corpus, we estimated and examined the posterior distributions of its articles' influence.

In this section, we demonstrate that the estimate of an article's influence is robustly correlated to the number of citations it received. While the DIM model is designed for corpora without citations—and, indeed, only the documents' text and dates are used in fitting the model—citations remain an established measure of influence. This study provides validation of the DIM as an exploratory tool of influential articles.

Data

The three scientific corpora we analyzed were the *ACL Anthology*, *The Proceedings of the National Academy of Science*, and the journal *Nature* (we discuss the New York Courts in a later section). For each corpus, we removed short documents, terms that occurred in too few documents, and terms that occurred in too many documents (by thresholds). We also removed terms whose statistics did not vary over the course of the collection, as such terms would not be useful for assessing change in language (a sample of such non-varying terms from *Nature* is "ordinarily", "shake", "centimetre", "traffic", and "themselves"). By applying these filters, we retained the most interesting words from the perspective of a time-series analysis.

ACL Anthology. The *Association for Computational Linguistics Anthology* is a digital collection of publications about computational linguistics and natural language processing (Bird *et al.*, 2008). We analyzed a 50% sample from this anthology, spanning 1964 to 2002. Our sample contains 7,561 articles and 11,763 unique terms after preprocessing. For this corpus we used article citation counts from the *ACL Anthology Network* (Radev *et al.*, 2009).

PNAS. The *Proceedings of the National Academy of Sciences* is a leading, highly-cited, multidisciplinary scientific journal covering biological, physical, and social sciences. We sampled one seventh of the collection, spanning 1914 (when it was founded) to 2004. Our sample contains 12,145 articles and 14,504 distinct terms after preprocessing. We found citations using Google Scholar for 78% of this collection.

Nature. The journal *Nature* is the world's most highly cited interdisciplinary science journal (Thompson Reuters, 2009) with content on a range of scientific fields. We analyzed a 10% sample from this corpus, spanning 1869 (when it was founded) to 2008. Our sample contains 34,418 articles and 6,125 distinct terms after preprocessing. We found citations using Google Scholar for 31% of these documents.

Inference for 10 topics on each corpus above took about 11 hours to converge on a desktop Intel 2.4GHz Core 2 Quad CPU. Our convergence criterion was met when the evidence lower bound increased by no more than 0.01%. For the experiments described below, we set topics' Markov chain variance $\sigma^2 = 0.005$ and $\sigma_d = \sigma_l = 0.0001$. These values were selected to make the topics change at a reasonable, "coherent" rate.

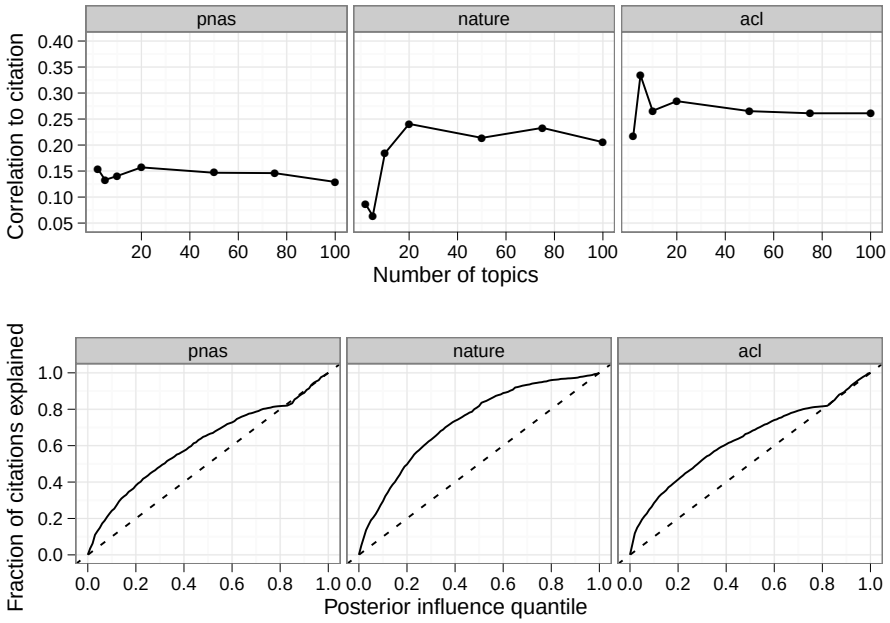


Figure 3.2: Spearman rank correlation between citation counts and posterior influence score, controlling for date (top) and fraction of citations explained by posterior influence (bottom).

Relating posterior influence and citation

We studied the DIM with varying numbers of topics (5, 10, 15, 20, 50, 75, and 100). We measured the relationship between the posterior influence values of each article $\tilde{\ell}_d$ and its citation count c_d.

We first aggregate the influence values across topics. Recall that each document has an influence value $\tilde{\ell}$ for each topic. For the nth word of document d, we compute its expected posterior influence score, with the expectation taken with respect to its (random) topic assignment z. Omitting time indices, this is $\mathrm{E}[z_{d,n} \cdot \tilde{\ell}_d]$. We then sum these values over all words in the document,

$$f(\tilde{\ell}_d) = \sum_{n=1}^{N_d} \mathrm{E}[z_{d,n} \cdot \tilde{\ell}_d]. \tag{3.11}$$

This weights each word by the influence associated with its assigned topic. When we are done with it, $f(\tilde{\ell}_d)$ provides a metric for influence which is topic-independent. (Using the maximum value of influence across topics yielded similar results.)

Figure 3.2 displays the Spearman rank correlation between the aggregated posterior influence score of Equation 3.11 and citation counts. The DIM posterior—which is estimated only from the texts of the articles—has a positive correlation to the number of citations. All of these numbers were found significant up to $p < 10^{-4}$, using permutation tests on the influence scores.

Correlation goes up when we model multiple topics within a corpus. Moving from 2 to 5 topics in the *ACL* corpus increases correlation from 0.25 to 0.37. *Nature* is likewise better with more topics, with a correlation of 0.28 at 20 topics; while *PNAS* performs best near 5 topics, with a correlation of 0.20.

Figure 3.2 also shows the fraction of citations explained by DIM scores: *Nature* documents with the highest 20% of posterior influence, for example, received 56% of citations. The flat regions in *ACL* and *PNAS* are due to aggregate influence scores very close to zero.

Heuristic model. The DIM is a complicated model. To justify its complexity, we describe a simple baseline (the *heuristic*) which captures our intuition with a single topic, is easy to implement, and runs quickly. For this heuristic, we define a word's weight at time t as:

$$w_t := \frac{\text{Frequency of } w \text{ in } [t,t+f]}{\text{Frequency of } w \text{ in } [t-p,t]},$$

for fixed distances f into the future and p into the past. A document's score is the weighted average of its words' weights. This heuristic captures the intuition that influential documents use language adopted by other documents.

The heuristic performed best with large values of its parameters ($f = p = 200$). With these settings, it achieves a correlation of 0.20 for the *ACL*, 0.20 for *PNAS*, and 0.26 for *Nature*. For *Nature*, the model is more correlated with citations than the heuristic for 20, 50, and 75 topics. Correlation is matched for *PNAS*, the model slightly beating the heuristic at 5 topics. *ACL* outperforms the heuristic for all numbers of topics.

Shuffled corpus Though we have eliminated date as a confounder by controlling for it in correlations, there may be other confounders such as document length or topic distribution. We therefore measured the DIM's relationship to citations when dates were randomly shuffled, keeping all documents which share a date together. If non-date confounders exist, then we might see correlation in the shuffled data, marking observed correlation as dubious.

We shuffled dates in the corpora and refit the DIM. We found a *maximum* date-controlled correlation of 0.018 for 29 shuffles of *ACL*; 0.001 for 5 shuffles of *Nature*; and 0.012 for 28 shuffles of *PNAS*. While this shuffled experiment and controlling for date do not entirely preclude confounding, they eliminate many potential confounders.

A closer look

Experiments showing correlation with citations demonstrate consistency with existing bibliometrics. However, the DIM also finds qualitatively different articles than a bibliometric based on citation counts finds. In this section we describe several documents to give the reader an intuition behind the kind of analysis that the DIM provides.

IBM Model 3 The second-most cited article in the *ACL Anthology Network* is *The Mathematics of Statistical Machine Translation: Parameter Estimation* (Brown *et al.*, 1993). It has 450 intra-*ACL* citations and 2,130 total citations listed on Google Scholar. This seminal work describes parameter estimation for five word-based statistical models of machine translation; it provided widely accepted statistical models for word alignment and introduced the well-known "IBM models" for machine translation. The posterior influence score for Brown *et al.* (1993) ranked 6 out of 7,561 articles in a 10-topic model.

This article was most influential in a topic about translation, which had a trend toward "alignment for machine translation." The largest-moving words are shown in Figure 3.3 (left). Upward trends for "alignment", "brown", and "equation" are evident (although it is not clear whether "brown" refers to the author or the corpus).

The Penn Treebank The most-cited article in our subset of the *ACL Anthology Network* is *Building a large annotated corpus of English: the Penn Treebank* (Marcus *et al.*, 1993), with 1,622 *ACL* citations and 2,810 citations on Google Scholar. This article describes the large-scale part-of-speech and syntax tagging of a 4.5-million word corpus. It falls in a topic about part-of-speech tagging and syntax trees; "treebank" had become one of the top words in the topic by 2004.

The DIM assigned a relatively low influence score to this article, ranking it 2,569 out of 7,561 articles. While Marcus *et al.* (1993) introduces a powerful *resource*, most of the article uses conventional language and ideas to detail the annotation of the Penn Treebank. As such, the paper does not discuss paradigm-changing ideas and the model scores it low. We emphasize that this does not undermine the tremendous influence that the Penn Treebank has had on the field of natural language processing. The DIM is not designed to discover this kind of influence.

Success in 1972 In 1967, The College Science Improvement Program was established to assist predominantly undergraduate institutions. Two years later *Nature* published a short column, which has the highest of our posterior influence in a 20-topic model, out of 34,418 *Nature* articles. No citation

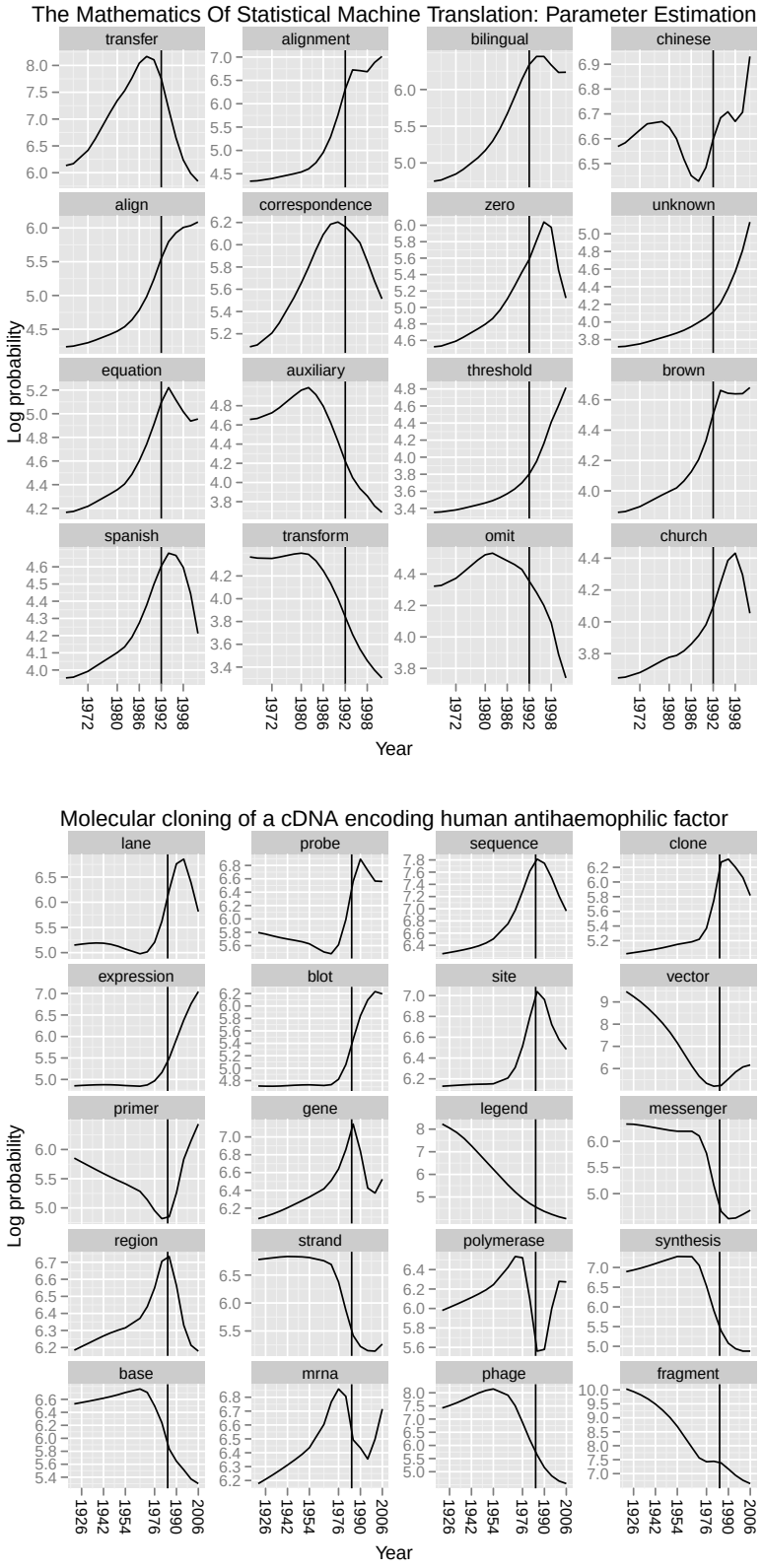


Figure 3.3: Most active words appearing in Brown *et al.* (1993) (left) which have changed the most in a topic about translation. On right are words appearing in Toole *et al.* (1984) in a topic about DNA and genetics. Terms are sorted by increase over 10 years.

information was available about this article in Google Scholar. The column, *How to be Overtaken by Success*, discusses a debate about the "Miller bill", which considers funding for postgraduate education (Nature, 1969). *Overtaken by Success* provides few research resources to researchers, which may explain lack of citation information. Instead, it presciently discusses a paradigm shift in a topic about science, industry, research, and education: "The record of the hearings [on the bill] is not merely an indication of the way the wind is blowing but an important guide to some of the strains which are now accumulating within the system of higher education..."

In 1972, three years after this article's publication, The NSF Authorization Act of 1973 made the NSF explicitly responsible for science education programs *at all levels* (NSF Website, 2010). Where this may have been missed by those using citation counts to study the history of science education, the DIM has provided a metric with which to gauge interest in the article.

Genetics in *Nature* The sixth most influential document by the DIM in a 20-topic model of *Nature* is *Molecular cloning of a cDNA encoding human antihaemophilic factor*, an article describing successful cloning of a human mRNA sequence important in blood clotting (Toole *et al.*, 1984). With 584 citations, this article is among the top 0.2% of these 34,418 documents. The most active words appearing in this article are shown in Figure 3.3 (right). The plot shows some of the document's key words – "expression", "primer", "blot" – become prominent words in the topic.

An application to the New York Appellate Courts.

The New York Appellate Court system hears appeals cases within the state of New York. This court "was established to articulate statewide principles of law in the context of deciding particular lawsuits" (NY CA Website, 2012), acting as a form of "Supreme Court" for the state of New York. Judges who hear these cases make decisions about the cases and write opinions summarizing their reasoning for these decisions. These decisions and opinions are extremely important within the court system because they set precedent for later decisions.

These opinions written by judges are therefore written expressly to be *influential* on later court decisions, and judges' opinions frequently make explicit citations to earlier cases. However, these citations are limited in two respects. First, multiple opinions may exist per case, stating the majority opinion, supporting it in part, or entirely disagreeing with it. Although judges' citations are explicit and well-formatted, their citations do not make this distinction machine-readable, making large-scale analyses difficult without expensive hand-coding. Second, lawmakers may have different reasons for citing opinions; it has been hypothesized by some political methodologists (people who use formal

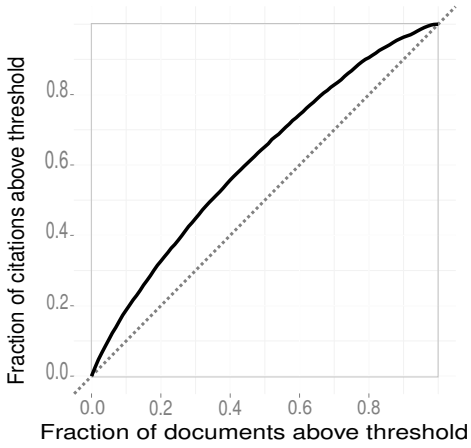


Figure 3.4: Citations explained by influence score in the New York Appellate Courts. Each point on the curve represents a different threshold of the influence score. The x-axis describes threshold on the influence score, and the y-axis describes the fraction of citations for all documents which fall below this threshold.

and quantitative methods to study political science) that researchers do not cite dissenting opinions because dissenting opinions are considered to hold little if any legal sway; citing dissenting opinions is therefore seen as a sign of weakness (Beim correspondence, 2011).

We analyzed this collection, splitting 9,266 appellate court cases into 10,618 distinct opinions, written by judges representing the majority opinion, a concurrence in part (i.e., supporting the majority decision but with a different rationale for reaching that decision), or a dissenting opinion. Our collection contained 13,568 distinct terms after pre-processing. We also scraped citations within this collection and found 37,348 intra-corpus citations.

Based on the analysis of the scientific corpora, we fit a 40-topic model to this collection to discover influential documents. Consistent with the scientific corpora, we measured a Spearman rank-correlation coefficient between posterior influence scores and the logarithm of citation counts at $\rho = 0.24$. We illustrate the fraction of citations explained by documents above different influence thresholds in Figure 3.4. Across all four corpora, the model is consistently correlated with citation counts.

3.4 Conclusions

Traditional bibliometrics like citations are widely used for understanding collections of text documents. Much of the past work for identifying influential documents focuses on measuring or predicting citations for corpora which have citations. In this chapter we described the DIM, which is developed for time-series corpora without bibliometrics. We have demonstrated measured consis-

tency with citations with the model, controlling for confounders like document length. However, the information provided by the model transcends this: the influence score has anecdotally been demonstrated to provide qualitatively different information than citations.

Based only on the changing statistics of the language in a corpus, we computed a measure of influence that is significantly related to observed citation counts. That said, it would be useful to better understand how this metric is qualitatively different from citations and other bibliometrics: expert judgment or usage information obtained from digital libraries might be some avenues. We leave this for future work.

We considered several documents evaluated by the model: Brown *et al.* (1993) and Toole *et al.* (1984), which both had high citations and high posterior influence; and Marcus *et al.* (1993), which had high citations and low posterior influence. These results demonstrate not just that the model is correlated with citations; it also suggests that the model provides qualitatively *different* information than citations.

3.4.1 Avenues for future work

The DIM could be made more realistic and more powerful in many ways. In one variant, individual documents might have their own "windows" of influence. Other improvements may change the way ideas themselves are represented, e.g. as atomic units, or *memes* (Leskovec *et al.*, 2009). Further variants might differently model the flow of ideas, by modeling topics as birth and death processes, using latent force models (Alvarez *et al.*, 2009), or by tracking influence *between* documents, building on the ideas of Shaparenko and Joachims (2007) or Dietz *et al.* (2007).

We also believe that it would be useful to better understand models like the DIM in the context of traditional metrics of influence, such as academic citations, and other metrics of influence, such as usage data. Having a better understanding of when this model and established metrics differ will uncover where our metric may provide new information that is not yet captured by existing statistics.

3.4.2 Next steps

The work presented in this chapter assumes that the collection of documents is described by a set of themes, and that these themes evolve over time. It describes each document using a mixture over themes and a vector describing its influence on each of those themes. This provides a sense of the current of ideas coursing through a collection of documents.

A limitation of this approach is that it provides too broad a view of a corpus: it does not provide explicit detail of the underlying story *within* a collection. This model describes a corpus as a collection of topics, and it describes documents as mixtures of themes and influence weights, but it does not provide any further sense of a story which changes over time.

In the next chapter we will discuss a model to explore some of these shortcomings by explicitly modeling the "story" within a collection of text documents. This approach will use some of the same ideas from this chapter. Again we will assume that a collection of text documents serve as a window into the events within the collection of historical documents, and again we will encode assumptions by explicitly modeling them with latent random variables, linked by a time-series model. However, by modeling the interactions of entities within the collection explicitly, and applying posterior inference, we will learn a story about them.

Chapter 4

A time-series model of foreign affairs: predicting sentiment between nation-states

In this chapter we use the text of newspaper articles to infer a history of the relationships between different nations. An assumption of our work is that the tension between two nations—or a warm and robust relationship between them—is reflected by the language that is used to discuss them. In developing this assumption, we discuss two models designed to infer the relationships between pairs of nations.

Text and latent spaces

The basic unit of analysis in this chapter is paragraphs of text from newspaper articles which discuss pairs of nations. We choose paragraphs because they are small enough to have just one or two concrete ideas but large enough to describe interesting relationships.

We use some of the same ideas presented in the last chapter to model the text of these paragraphs, but we use one of the primitives introduced in Chapter 2 to model relationships between pairs of nations. This allows us to build a history of nations' relationships over time. An advantage of a text-based approach to history is that we can incorporate information from all articles of a given collection with modest computational cost. This means that historians and political scientists can then search and review thousands of historical documents at the push of a button—or identify

forgotten and overlooked incidents in history.

The primitive from Chapter 2 that we use amounts to an assumption that each nation can be summarized by its position in a latent space, so that the sentiment between two nations is determined (up to stochasticity) by the relationship between their positions in this latent space. By making this assumption, we gain two benefits: the ability to interpret these nations' positions, since they provide statistically meaningful summaries of these nations' positions; and the ability to make predictions about the relationships between nations, based on their latent positions. While the last chapter's Document Influence Model allowed us to discover themes which evolved over time and individual documents' influence on these themes, the assumptions we make in this chapter allow us to create a more rich story about the interaction of specific textual entities—nations—over time.

Organization of this chapter

In the next two sections we develop several computational models that link the text of a news source to the relationships between nations.

We begin with a model which infers these relationships by using two sources of labels about the the relationship, or sentiment, between pairs of nations: expert labels and labels assigned by lay paid "workers". To design this model, we develop a set of spatio-temporal assumptions that allow us to describe the sentiment between nations by inspecting their relative positions in this latent space (and, inversely, to interpret their positions based on observed sentiment). We demonstrate that modeling nations in this way allows us to create a history of foreign relations over time. Importantly, we demonstrate that the sentiment inferred from two very different sources of sentiment labels leads to strikingly similar measures of inter-state sentiment.

After developing this supervised model, we invert this question and ask: what sentiment is implied by the text alone of news articles? To answer this question, we describe an unsupervised model of the relationship between nations to qualitatively describe these relationships. We then demonstrate a connection between the unsupervised relationships and the sentiment labels we had used for the supervised model.

4.1 A supervised model of dyadic sentiment

In the last chapter we described a model for identifying influential documents. A defining feature of that model was that it was unsupervised; only after fitting the model could we compare the inferred influence of an article with the number of citations it had received. In this section we will take a

more direct approach, fitting a model with labels *defined* to represent the information that we seek: whether there is a positive or negative relationship between pairs of nations.

In outlining this model, we will provide more detail into the two assumptions made in this chapter: first, that there is a relationship between text and the sentiment between pairs of nations; and second, that we can model the sentiment between nations by representing these nations as vectors in a latent space. After describing these assumptions we adjust the model to extend it to the time-series domain.

4.1.1 Inferring sentiment from text

The first assumption that we make in this chapter is that the relationship between pairs of nations can be described by (at least) a one-dimensional sentiment $s \in \mathbb{R}$, and that when a news source discusses these nations, the author's choice of words $\boldsymbol{w}_d$ reflects the relationship between them. Consider, for example, the relationship between the state of Israel and Palestine in the following passage (emphasis added by me):[1]

> "In Government and opposition circles questions have swirled about how a **Palestinian** truck laden with explosives could have sailed past **Israeli** soldiers stationed at Gaza Strip checkpoints. Some news reports said the vehicle had the required Israeli permits."
> *Failed Truck-Bomb Plot Chills Israel-P.L.O. Autonomy Talks* (Haberman, 2005)

Israel and Palestine have a tense relationship, as suggested by the author's choice of the words "explosives", "questions", and even "required". This relationship is negative, so let's say that the sentiment between them is -3. Now consider the following passage about Egypt and Jordan:

> "The leaders of **Egypt** and **Jordan** too have invested their prestige in the peace plan and would rejoice in private to see Islamic militants crushed." *Middle East Talks are Effort to Aid Peres and Arafat* (Jehl, 1996)

The relationship between Egypt and Jordan—while not fabulous—is certainly more positive, as suggested by words such as "invested", "peace", and "rejoice". Let's say that the sentiment between Egypt and Jordan is 0.5.

The numbers -3 and 0.5 are of course arbitrary, but they convey some sense of the relationship, or "sentiment", between pairs of nations. Our intuition is that the words selected by authors when describing pairs of nations often provide a direct indication of the sentiment between these nations,

[1] The statehood of Palestine is disputed. We considered a collection of of states and territories.

and that we can estimate this sentiment for a new snippet of text (up to a constant factor) with the right model.

To do this, we will use paragraphs of text which mention a pair of nations as the basic unit of analysis in this chapter. We will assign labels to enough of these paragraphs to fit a text-based model, and then we will fit a sentiment model to their text. [2] Paragraphs of text (like the two above) are small enough to contain simple ideas yet large enough to discuss complete ideas—appropriate also for discussing the relationship between pairs of nations.

To relate an author's text to the sentiment between nations, we use a model called text regression (Kogan *et al.*, 2009). In text regression, we model the sentiment s_d in document d using a linear combination of the wordcounts $\boldsymbol{w}_d \in \mathbb{N}^{+V}$ (omitting the names of the nations and major cities) of each article:

$$\begin{aligned} s_d|\boldsymbol{w}_d, \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{w}_d^T\boldsymbol{\beta}, \sigma_W^2) \\ \boldsymbol{\beta} &\sim \mathcal{N}(0, \sigma_\beta^2). \end{aligned} \tag{4.1}$$

For the remainder of this section, we will assume that $\boldsymbol{\beta}$ is observed, so that the s_d is normally distributed with mean $\boldsymbol{w}_d^T\boldsymbol{\beta}$. We describe how to fit $\boldsymbol{\beta}$ with human labels in Section 4.1.5.

A brief comment on notation. Before we describe how to fit this model, we pause to summarize our use of notation. In this chapter, we will use notation flexibly when it is convenient. The typical unit of discussion will be the dth document occurring at time t. The dth document discusses two nations, c_1 and c_2; these define a tuple $(\{c_1, c_2\}, d, t)$ (where the set $\{c_1, c_2\} = \{c_2, c_1\}$). We will generally use d to index documents, t to index time, and c to index a nation. When document d is given, we may refer to its time as t_d (which is unique) or to the two interacting nations as $c_{d,1}, c_{d,2}$ or c_1, c_2. Alternatively, we may refer to the documents in which a nation c appears as $d_{c,1}, \ldots, d_{c,D}$. As another example, we may describe a nation's position $x_{(c_1,d,t)}$ variously as $x_{c_{d,1}}$, $x_{d,1}$, or even x_c if the context is clear. Finally, the sentiment between two nations (when described by a specific document) might be variously described as s_d, s_{c_1,c_2}, $s_{d,t}$, or $s_{c_1,c_2,d,t}$.

4.1.2 Modeling interactions with a latent space

The second assumption we make in this chapter is that each nation can be described by a vector in some p-dimensional latent space, and that the relationship between two nations is determined (up to

[2] We provide more detail about tagging nations in the experiments section.

Description	$\mathcal{F}(\boldsymbol{x}_1, \boldsymbol{x}_2)$	where ...
distance	$-\log(\|\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|\|_2^2 + 1)$	$\boldsymbol{z}_1 = \boldsymbol{x}_{1,2:D}, \boldsymbol{z}_2 = \boldsymbol{x}_{1,2:D}$
inner product	$\boldsymbol{z}_1^T \boldsymbol{z}_2$	$\boldsymbol{z}_1 = \boldsymbol{x}_1,\ \boldsymbol{z}_2 = \boldsymbol{x}_2$
intercept	$y_1 + y_2$	$y_1 = x_{1,1}, y_2 = x_{2,1}$
intercept/inner product	$y_1 + y_2 + \boldsymbol{z}_1^T \boldsymbol{z}_2$	$y_1 = x_{1,1}, y_2 = x_{2,1}$, $\boldsymbol{z}_1 = \boldsymbol{x}_{1,2:D}, \boldsymbol{z}_2 = \boldsymbol{x}_{2,2:D}$
intercept/distance	$y_1 + y_2 - \log(\|\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|\|_2^2 + 1)$	$y_1 = x_{1,1}, y_2 = x_{2,1}$, $\boldsymbol{z}_1 = \boldsymbol{x}_{1,2:D}, \boldsymbol{z}_2 = \boldsymbol{x}_{2,2:D}$

Figure 4.1: Link functions $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. Intercept link functions introduce per-nation intercepts that indicate how prone a nation is to war; distance link functions are based on the distance between nation' vectors; and inner-product link functions represent sentiment as a function of nations' political "orientations". The notation $2 : D$ refers to a collection of indices, so that $x_{2:D}$ is a $D - 1$-dimensional vector.

stochasticity) by the relationship between these nations' vectors. We formalize this assumption by letting each nation c take a position $\bar{x}_{c,0} \in \mathbb{R}^p$. As above, the sentiment of the relationship between these two nations c_1, c_2 is described by the scalar $s_d = s_{c_1,c_2} \in \mathbb{R}$ (we change notation for s as appropriate given the context). This sentiment is determined by the interaction of their positions:

$$\begin{aligned} \boldsymbol{x}_{c_1,d} &\sim \mathcal{N}(\bar{x}_{c_1,0}, \sigma_D^2) \\ \boldsymbol{x}_{c_2,d} &\sim \mathcal{N}(\bar{x}_{c_2,0}, \sigma_D^2) \\ s_d &:= \mathcal{F}(\boldsymbol{x}_{c_1,d}, \boldsymbol{x}_{c_2,d}),, \end{aligned} \tag{4.2}$$

for some suitable function $\mathcal{F} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ (see Table 4.1.2 for examples of $\mathcal{F}$), and where we interpret s_d as the sentiment between c_1 and c_2 as reflected by article d (which appeared at time t_d). We have also introduced the auxiliary random variables x_{c_1} and x_{c_2}, which can be interpreted as the positions these nations take during interaction in an article. We include them for the algebraic convenience that will become evident later.

If $\mathcal{F}$ is continuous and c_1 and c_2 are similar (as measured by the distance between $\bar{x}_{c_1}$ and $\bar{x}_{c_2}$), then c_1 and c_2 will interact with other nations in similar ways. Further, by selecting $\mathcal{F}$ carefully, we can ensure that a poor relationship ($\mathcal{F}(x_{c_1}, x_{c_2}) \ll 0$) between c_1 and c_2 corresponds to intuitive relationships between $\bar{x}_{c_1}$ and $\bar{x}_{c_2}$, such as a large distance.

A spatial model provides us with two benefits. First, it provides interpretability: we can summarize nations' relationships with other nations succinctly with their positions $\bar{x}_c$. Second, this allows us to draw on existing work from multidimensional scaling, which has been used successfully in both political science (Martin and Quinn, 2002; Jackman, 2001) and social network modeling (Hoff *et al.*, 2002; Chang and Blei, 2009). We will empirically validate this model later, but first we extend it to

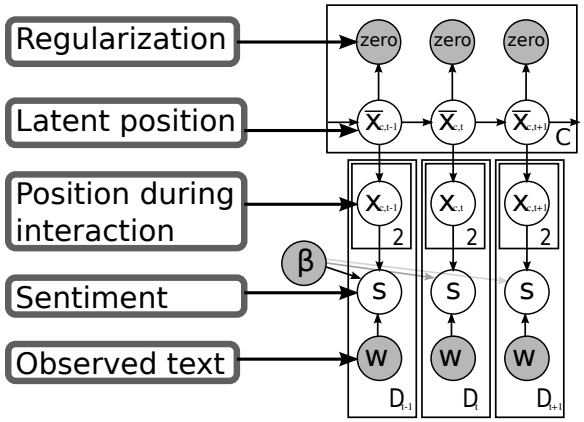


Figure 4.2: A time-series model of nations' interactions. Pseudo-observations of "zero" are added for regularization. Amazon Mechanical Turk labels are used to fit $\boldsymbol{\beta}$, which is used to infer unobserved sentiments.

the time-series domain.

4.1.3 A temporal model of interaction

Foreign relations are not static; nations' alliances and preferences change over time with the evolution of economies, technology, and culture. Therefore we make this a fully temporal model by allowing each nation's mean position (formerly $\bar{x}_c$) to take a position at each time t. We assume that x drifts with the Markov transition

$$\bar{x}_{c,t}|\bar{x}_{c,t-1} \sim \mathcal{N}(\bar{x}_{c,t-1}, \sigma^2_{\text{chain}}), \tag{4.3}$$

as shown in Figure 4.2. At any time t, we may observe the relationship between states c_1 and c_2 in an article d. As before, the distribution of the sentiment between these nations is entirely specified by their positions at this time:

$$\begin{aligned} x_{c_1,d} &\sim \mathcal{N}(\bar{x}_{c_1,t}, \sigma^2_D) \\ x_{c_2,d} &\sim \mathcal{N}(\bar{x}_{c_2,t}, \sigma^2_D) \\ s_d &:= \mathcal{F}(\boldsymbol{x}_{c_1,d}, \boldsymbol{x}_{c_2,d}). \end{aligned} \tag{4.4}$$

We reconcile $p(s_d|\boldsymbol{w}, \boldsymbol{\beta})$ (see Equation 4.1) with Equation 4.4 by recalling that $\boldsymbol{\beta}$ is treated as

constant once it is initially fit. This means that the joint distribution of nations' sentiment is

$$
\begin{aligned}
&p(s_d, \boldsymbol{x}_{c_1,d}, \boldsymbol{x}_{c_2,d} | \boldsymbol{w}_d, \boldsymbol{\beta}, \bar{x}_{c_1,d}, \bar{x}_{c_2,d}) \\
&\qquad \propto p(\mathcal{F}(\boldsymbol{x}_{c_1,d}, \boldsymbol{x}_{c_2,d}) | \boldsymbol{w}_d, \boldsymbol{\beta}) \times p(\boldsymbol{x}_{c_1,d} | \bar{x}_{c_1,t}) \times p(\boldsymbol{x}_{c_2,d} | \bar{x}_{c_2,t}) \\
&\qquad = \mathcal{N}(\mathcal{F}(\boldsymbol{x}_{c_1,d}, \boldsymbol{x}_{c_2,d}) | \boldsymbol{w}_d^T \boldsymbol{\beta}, \sigma_W^2) \times \mathcal{N}(\boldsymbol{x}_{c_1,d} | \bar{x}_{c_1,t}, \sigma_D^2) \times \mathcal{N}(\boldsymbol{x}_{c_2,d} | \bar{x}_{c_2,t}, \sigma_D^2). \qquad (4.5)
\end{aligned}
$$

Regularization and zero-reversion. To complete this model, we add a standard normal prior to the ends of the chain, so that, for all nations c, $p(\bar{x}_{c,0}) = p(\bar{x}_{c,T}) = \mathcal{N}(0, 1)$. We also add an additional regularization term which we call zero-reversion. This term manifests itself as artificial observations of zero coming from the hidden Markov model. In the joint distribution, this is an additional product $\prod_{c=1}^{C} \prod_{t=0}^{T} \mathcal{N}(0 | \bar{x}_{c,t}, \sigma_{\text{zero}}^2)$. Zero-reversion can be motivated anecdotally by noting that, in the absence of news, we can assume that nations tend to have neutral interaction with other nations. We find that for certain link functions $\mathcal{F}$ it improves empirical performance.

Related work

The field of sentiment analysis has received considerable attention in the last couple of decades and is used in a variety of industry fields, ranging from automated trading strategies to restaurant recommendation sites. Models in which individual words are assigned a weight are common; Pang and Lee (2008) provide a review of recent developments in this field. See Taddy (2012) for a model which uses inverse regression on word counts for results which compare favorably with alternatives.

Spatial models such as Item Response Theory (IRT) have been developed over the past century by quantitative social scientists for analyzing behavior. While much of this work has been used to model parliamentary voting behavior, these techniques have also been used to model nations' positions based on their votes in the UN General Assembly. Gartzke et al., for example, use these votes and alliance models to study the nations' affinities (Gartzke, 1998).

These models have been developed for dyadic data more fully in network models such as the latent space model (Hoff *et al.*, 2002; Sarkar and Moore, 2005), in which the probability of a link between two nodes is a function of their latent-space distance. The qualitative relationship of entities' dyadic relationships has been more fully developed with text by the relational topic model, which uses free text to model the relationship between actors in an unsupervised setting (Chang and Blei, 2009).

The areas of sentiment analysis and dyadic models have been combined in recent work focused on content recommendation and unsupervised network discovery. Recommendation systems have

been specialized to items with text for recommending content such as Web content (Agarwal and Chen, 2010) and academic journals (Wang and Blei, 2011); both of these applications used latent Dirichlet allocation for modeling text. Chang and Blei (Chang and Blei, 2009; Chang *et al.*, 2009) have also used unsupervised topic models to discover relationships between entities.

4.1.4 Inference

We fit the *MAP* objective of this probabilistic model. By using a MAP estimate, we will avoid the tedious derivations from the last chapter. Further, the MAP estimate can be interpreted as a form of unregularized variational inference (by letting variance around the posterior estimates go to zero). We optimize the *MAP* objective in this model using an expectation maximization (EM) algorithm.

An expectation maximization algorithm

The MAP solution to this problem can be approximated using an expectation maximization (EM) algorithm because of the way we have specified $p(x_{c,d}|\bar{x}_{c,t_d})$. This makes inference much simpler and allows us to take advantage of a Kalman smoother. Instead of optimizing each variable in the objective, we alternate between optimizing the variables $x_{c,d,t}, s_d$ in an E step and the variable $\bar{x}_{c,t}$ in the M step.

M Step. In the M step, we seek to estimate the mean $\bar{x}_{c,t}|x, \beta, s$ of each nation c's position. Because the Markov blanket of each variable $\bar{x}_{c,t}$ is specified by Gaussian distributions, we have that $\arg\max_{\bar{x}} p(\bar{x}|x) = \mathbb{E}\left[\bar{x}|x\right]$. More generally, this expectation is the optimal value of $\bar{x}$ given the other variables:

$$\arg\max_{\bar{x}} p(s, x, \bar{x}|\boldsymbol{\beta}, \boldsymbol{w}) = \arg\max_{\bar{x}} p(\bar{x}, x) = \arg\max_{\bar{x}} p(\bar{x}|x) = \mathbb{E}\left[\bar{x}|x\right]. \tag{4.6}$$

We can estimate $\bar{x}|s, x, \boldsymbol{\beta}, \boldsymbol{w} = \mathbb{E}\left[\bar{x}|x\right]$ using a variant of the traditional Kalman smoother (Kalman, 1960), where we treat x as observations of the hidden state $\bar{x}$. This step differs from a standard Kalman smoother in that we have no observations on some dates and multiple observations on other dates.

Kalman updates. As with a standard Kalman smoother, the modified Kalman smoother requires a forward filter step and a backward filter step. The forward filter estimates the mean position given

all previous observations:

$$\bar{x}_{\text{forth},c,t}|\bar{x}_{\text{forth},c,t-1}, \{x_{c,d,t-1}\}_d \leftarrow \frac{\bar{x}_{\text{forth},c,t-1}/\sigma^2_{\text{forth},t-1} + \sum_{d=1}^{D_{c,t-1}} x_{c,d,t-1}/\sigma^2_{\text{obs}}}{1/\sigma^2_{\text{forth},t-1} + 1/\sigma^2_{\text{obs}}} \tag{4.7}$$

$$\sigma^2_{\text{forth},t} \leftarrow \frac{1}{1/\sigma^2_{\text{forth},t-1} + D_{c,t-1}/\sigma^2_{\text{obs}}} + \sigma^2_{\text{chain}}, \tag{4.8}$$

where we have used $x_{c,d,t}$ to describe the position of nation c at time t for interaction d and there are D_{ct} documents at time t discussing nation c. We also use initial condition $\bar{x}_{c,0} = 0, \sigma^2_{\text{forth},0} = 10$. The backward step estimates the chain's mean given all current and future observations:

$$\bar{x}_{\text{back},c,t}|\bar{x}_{\text{back},c,t+1}, \{x_{c,d,t}\}_d \leftarrow \frac{\bar{x}_{\text{back},c,t+1}/\sigma^2_{t+1} + \sum_{d=1}^{D_{c,t}} x_{c,d,t}/\sigma^2_{\text{obs}}}{1/\sigma^2_{\text{back},t-1} + 1/\sigma^2_{\text{obs}}}$$

$$\sigma^2_{\text{back},t} \leftarrow \frac{1}{1/(\sigma^2_{\text{back},t+1} + \sigma^2_{\text{chain}}) + D_{c,t}/\sigma^2_{obs}}, \tag{4.9}$$

with initial conditions $\bar{x}_{\text{back},c,T} = 0, \sigma^2_{\text{backward},T} = 10$. The smoothed means—that is, the mean of nations' positions at time t given observations before and after t—are

$$\begin{aligned}\bar{x}_{c,t}|x_{c,t} &= \mathbb{E}\left[x_{c,t}\right]|\bar{x}_{\text{forth},c,t}, \bar{x}_{\text{back},c,t}, \sigma^2_{\text{back}}, \sigma^2_{\text{forth}} \\ &= \frac{\bar{x}_{\text{forth},c,t}/\sigma^2_{\text{forth},t} + \bar{x}_{\text{back},c,t}/\sigma^2_{\text{back},t}}{1/\sigma^2_{\text{forth},t} + 1/\sigma^2_{\text{back},t}}\end{aligned} \tag{4.10}$$

E-Step. In the E-step, our goal is to infer each nation's position $x_{c_{d,1}}|\bar{x}_{c,d,t}, x_{c_{d,2}}, s_d, \boldsymbol{w}_d$ during interaction d given its expected mean $\bar{x}_{c_{d,1},t_d}$ and the text $\boldsymbol{w}_d$ describing this interaction, *and* given the other nation's position for this interaction. Assuming that this nation is indexed by $c_{d,1}$ in each document d, we find these positions by gradient ascent on each interaction:

$$\begin{aligned}x_{c_{d,1},t} &\leftarrow \arg\max_x p(x, x_{c_{d,2},t}, \bar{x}_{c_{d,1},t_d}|\boldsymbol{w}_d, \boldsymbol{\beta}) \\ &= \arg\max_x \mathcal{N}(\mathcal{F}(x, x_{c_{d,2},t})|\boldsymbol{w}_d^T\boldsymbol{\beta}, \sigma^2_W)\mathcal{N}(x|\bar{x}_{c_{d,1},t_d}, \sigma^2_D),\end{aligned} \tag{4.11}$$

For convenience, we iterate between updating $x_{\{c,\cdot\}}$ for all interactions involving nation c and updating $\bar{x}_{c,t}$ with the M step for all times t.

4.1.5 Empirical studies: comparisons with ground truth

We now turn to an experimental analysis of this model. Our goal in this analysis is to demonstrate first that the model captures statistically meaningful patterns in a time-series collection of newspaper

documents and second that it can provide a meaningful view into nations' relationships with one another. We first describe the two distinct label types that we used to define sentiment s_d for this model and summarize the newspaper archive to which we fit this model. We then evaluate the model's ability to infer the relationships between nations and compare results from models inferred with the two different label types.

Parsing the New York Times

We fit and evaluate this model over news articles discussing 245 nations and territories from twenty years of the *New York Times* (NYT). This collection spanned the years 1987 to 2007, a period which included both the Persian Gulf and Iraq wars; the collapse of the Soviet Union; the reunification of Germany; September 11th, 2001; and countless other world events.

Data preparation. We used articles from the Foreign, Business, Financial, and Magazine desks of the newspaper during this period. As noted in Section 4.1.1, we split this collection into paragraphs, which were defined by *Times* editors, and selected the subset of paragraphs which discuss exactly two nations as "documents" d. This resulted in 257,472 paragraphs. We then defined a vocabulary to be those words which satisfied three criteria:

- Appeared at least twenty times,
- Appeared in no more than 40% of documents, and
- Appeared in at least 0.1% of documents.

This resulted in a vocabulary of 5,958 words, mentioned in 40,356 paragraphs. We randomly selected 80% of these paragraphs (32,249) as training examples and used the remaining examples to evaluate our model.

Coding sentiment

We next estimated $\boldsymbol{\beta}$ by fitting ridge regression (i.e., Equation 4.1 with a Gaussian prior on β) on a subset of the training examples. We labeled training examples with information from both inexperienced "workers" and "expert labels", representing vastly different ends of the label spectrum (as we will see, however, they result in strikingly similar predictions).

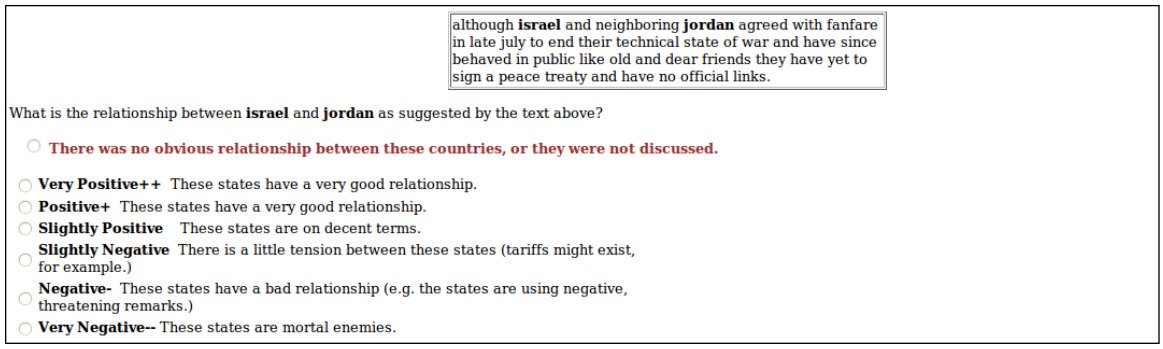


Figure 4.3: A screenshot of a Mechanical Turk labeling task. Sometimes relationships may be complicated; both raters gave this example a score of "slightly positive".

Novice labels: Amazon Mechanical Turk ratings

Amazon Mechanical Turk (AMT) is a crowd-sourcing platform which provides a *requester* with access to thousands of *workers* who perform simple tasks over the Internet. Although the requester can use tests to ensure that workers are high-quality, as well as reject the work of low-quality workers, these workers are very much non-experts.

To fit the model, we asked *Amazon Mechanical Turk* workers to rate the sentiment between two nations mentioned in the text of a paragraph on the scale -5 (mortal enemies), ..., 5 (very good relationship). We illustrate a rating task (as seen by a Mechanical Turk worker) in Figure 4.1.5. Raters were asked to review a random subset of 3607 paragraphs like this from the original collection. Before fitting the model, we manually disqualified eight raters (out of 85) who performed poorly (as measured by inconsistency with other raters).

With all rated paragraphs which were in the training set, we fit the coefficients $\boldsymbol{\beta}$ of the text regression discussed in Section 3.1. This coefficient was then treated as constant in the joint model in Figure 4.2 to allow us to infer sentiment from the words of all 32,249 training paragraphs. This resulted in a regression weight $\boldsymbol{\beta}_w$ for each word w, which we illustrate in Figure 4.4 (left).

Expert labels: Correlates of War

We also used a combined set of expert labels based on the Correlates of War (Sarkees and Warman, 2012) and Issue Correlates of War (Hensel, 2001).

- The *Correlates of War* project "seeks to facilitate the collection, dissemination, and use of accurate and reliable quantitative data in international relations" (CoW Homepage, 2012). The project provides labels describing the relationships between pairs of nations from 1823 to

2003. At-war is a binary relationship (either nations are at war, or they are at peace). We used a list of CoW inter-state wars (version 4.0) from 1823 to 2003 (Sarkees and Warman, 2012).

- The *Issue Correlates of War* project "is a research project that is collecting systematic data on contentious issues in world politics" (ICoW Homepage, 2012), and they provide expert labels on a variety of inter-state conflicts that *do not require militarized conflict*. However, these issue labels do require documented evidence of contention between states; such issues include maritime and territorial disputes (ICoW Homepage, 2012; Hensel, 2001). The Issue Correlates of War are not part of the same project (or produced by the same researchers) as the Correlates of War.

We used these two sets of ratings to label the collection of *New York Times* paragraphs by combining them and treating two nations as having a rating of -5 if they are at war at the time an article was written in the Correlates of War codes and -1 if there was any contentious issue between the nations in the Issue Correlates of War. All other pairs of nations were treated as having a rating of 0.1. These values are somewhat arbitrary (we could have chosen -6.3 for a bad relationship), but they were selected to correspond roughly to the range of the Mechanical Turk labels. Further, they were selected once and kept fixed—changing them during analysis could compromise the statistical power of the results below.

As before, we fit the text regression parameters $\boldsymbol{\beta}$ using these labels on the training set and evaluated nations' ratings on the test dataset. We illustrate the coefficient $\boldsymbol{\beta}$ fit to CoW-labeled paragraphs in Figure 4.4 (right).

Casual vs. expert labels

The CoW represent a data source which is modestly related to Mechanical Turk ratings. In the NYT dataset, CoW ratings and Mechanical Turk ratings were correlated at $\sigma = 0.196$. To illustrate the difference between these ratings, consider the following two examples:

- AMT rating= 1, CoW rating= -5:

  > *As an indication of the dangers the damage occurred in waters where military spokesmen said no mines had been suspected before but where a* ***Saudi*** *officer said today that some 22 were later found.* ***Iraq****i mines widely deployed [sic]* (Cushman, February 1991).

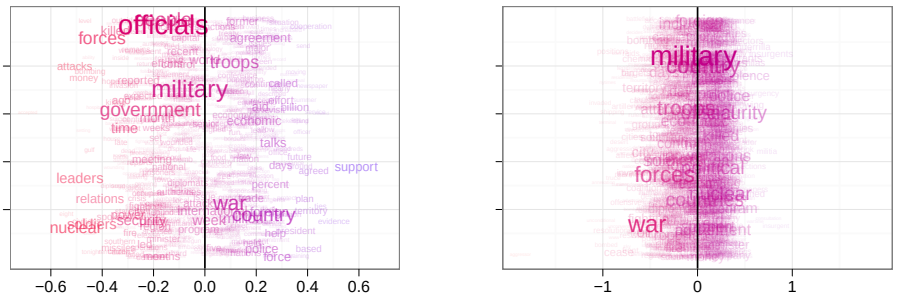


Figure 4.4: Coefficients $\boldsymbol{\beta}_w$ for selected words w fit on text labeled by Amazon Mechanical Turk workers (left) and Correlates of War data (right). Coefficients fit from Mechanical Turk labels are more clearly separated than those fit to Correlates of War labels; this is likely due to explicit positive sentiment in that dataset. The x-axis is $\boldsymbol{\beta}$, and the y-axis is used for display (it corresponds to no variable). Size of each word is proportional to $\sqrt{\text{frequency}}$, and color corresponds to $\boldsymbol{\beta}$.

This example outlines a limitation in our modeling assumptions: a single paragraph is sometimes too small a unit of discussion. Here Mechanical Turk workers likely missed the larger context of the article about the Gulf War (including the article's title, *War in the Gulf: Sea Mines; Allied Ships Hunt Gulf for Iraqi Mines*).

- AMT rating= -5, CoW rating= 0.1:

  > *Not since the grim old days of the cold war have relations between the* ***United States*** *and* ***Russia*** *been quite as problematic as they are this weekend on the eve of president Clinton's visit for celebrations marking the 50th anniversary of the allied victory in Europe in World War II* (Apple, May 1995).

  The second example represents a limitation of both data sources. The two Mechanical Turk ratings of -5 were clearly too strong, as the nations are not at war; but AMT workers likely based their rating in part on the reference to World War II (the instructions provided to MTurk workers suggest that a rating of -3 or -1 would have been more appropriate). In 1995, the United States and Russia were not at war and had no documented territorial conflicts. This means that this sentiment was not reflected in the CoW labels, and their sentiment defaulted to 0.1.

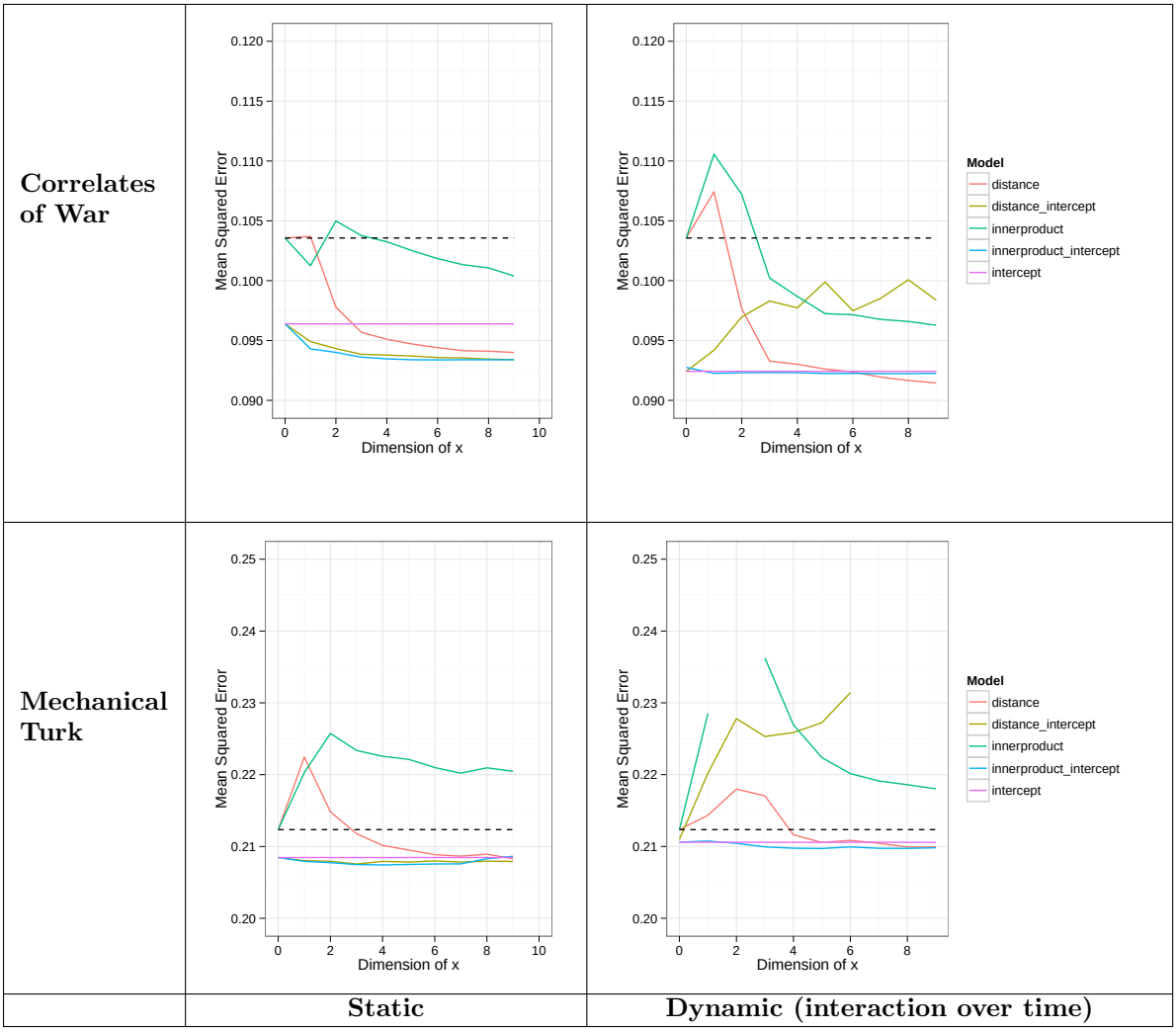


Figure 4.5: The dyadic sentiment model captures text well . Each colored line represents performance of the supervised model on a collection of heldout documents across twenty years of New York Times articles. The black dotted line represents performance based on estimating with the empirical mean of the dataset. An inner-product model with four dimensions (plus intercepts) performs well for most settings. A distance model with many dimensions but no intercepts also performs well across a range of assumptions, performing best with many dimensions.

Quantitative results

Inference and Prediction

We next turn to an empirical validation of the model laid out so far in this chapter. After fitting $\boldsymbol{\beta}$ to each set of labels on a subset of training documents, we estimated the MAP solution $\bar{x}, x, s|\boldsymbol{\beta}, \boldsymbol{w}$ using the entire training set described in Section 4.1.4.

For two nations c_1 and c_2 mentioned together at time t, we predict their sentiment to be $\tilde{s}_{c_1,c_2} = \mathcal{F}(\bar{x}_{c_1,t}, \bar{x}_{c_2,t})$ and calculate the mean-squared error between that prediction and their predicted sentiment $\boldsymbol{\beta}^T \boldsymbol{w}_d$ under text regression. We made the latter choice so we could analyze the text regression part of the model separately from the latent-space assumption (analyzing them together would make it difficult to discern the effect of each model).

Text regression. The text-regression model for CoW labels predicted heldout labels with MSE 0.98, compared with 1.02 if we estimate using the empirical mean $\bar{s} = -0.21$ of training examples. The text-regression model for AMT predicted heldout labels with MSE of 6.37, compared with 6.78 under the empirical mean.

While these errors are very large compared to the variance of the sentiment label, the scale of these errors is a result of the small number of training examples, the large number of features (1998 in each case), and the sparsity of these word-features. Still, we find that the coefficients $\boldsymbol{\beta}_{\text{CoW}}, \boldsymbol{\beta}_{\text{AMT}}$ learned from the respective CoW and AMT labels are subjectively intuitive. We illustrate the coefficients fit with these labels in Figure 4.4. The coefficients $\boldsymbol{\beta}_{\text{CoW}}$ and $\boldsymbol{\beta}_{\text{AMT}}$ are correlated at $\sigma = 0.18$.

Static latent space. With the text model in place, we next turn to evaluating the latent-space assumption. To do this, we hold fixed the coefficients $\boldsymbol{\beta}_{\text{CoW}}, \boldsymbol{\beta}_{\text{AMT}}$. This makes the mean $\boldsymbol{\beta}^T \boldsymbol{w}_d$ of sentiment s_d available to the latent-space models.

We first check the assumptions described in Section 4.1.2, which model nations' pairwise sentiment but do not assume that they change over time. We predict the sentiment between nations interacting in document d to be $\bar{s}_d = \mathcal{F}(\bar{x}_{c_1}, \bar{x}_{c_2})$ and evaluated the latent-space assumption based on its ability to reproduce predictions from the text-based sentiment model $s_d = \boldsymbol{w}_d^T \boldsymbol{\beta}$.

We evaluated this model for the five link functions $\mathcal{F}(\boldsymbol{x}_{c_1}, \boldsymbol{x}_{c_2})$ summarized in Table 4.1.2 and for a range of dimensions $p = \dim(\boldsymbol{z}) = 1, \ldots, 9$. We report the MSE for this range of experiments in Figure 4.1.5 and compare these models with a baseline model, which uses the empirical mean of the ratings.

We find that the the `inner product` assumption $\bar{z}_{c_1}^T \bar{z}_{c_2}$ alone is poor because it provides no natural way to model nations which are in frequent conflict with others. When the `inner product` link function and the `distance` link function are endowed with the intercepts y_{c_1}, y_{c_2}, their performance improves substantially: they consistently represent inter-nations' sentiment better than other models, with the `intercept/inner product` model consistently outperforming `intercept/distance` for most values of the latent-space dimension p. Based on the `intercept/inner product` model, the space of political sentiment appears to have dimension four or five. This is consistent between both label types.

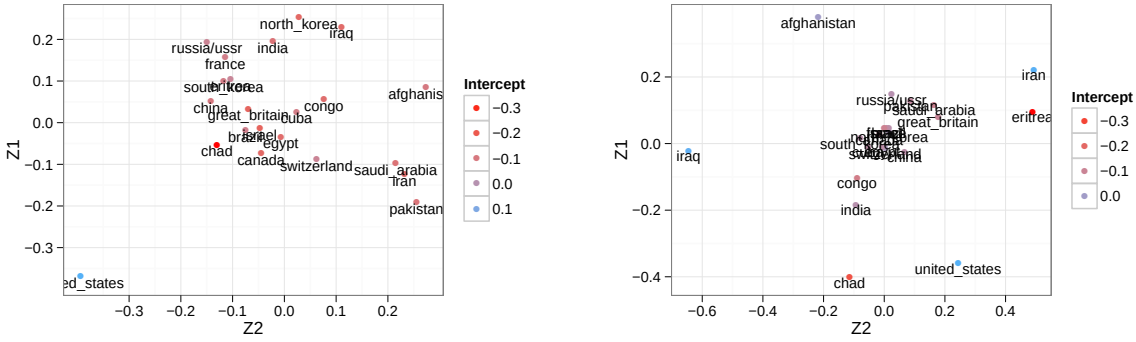
The improvement of these intercept models over their counterparts appears to be largely because intercepts enable these models to explain how conflict-prone a nation is. At the same time, they can use $\bar{z}$ to explain how each nation interacts with others; both `intercept/inner product` and `intercept/distance` outperform `intercept` for most values of p. Interestingly, the `distance` link function is able to model data well as p grows large without an indication that the model overfits (we only measured this up to 9 dimensions).

The benefit in adding a time-series assumption. We can add more flexibility to this model – and an ability to model much more interesting behavior—by extending it to the time-series domain as described in Section 4.1.3. Under this assumption, we allow $\bar{x}_c$ to drift over time for each nation c. Again we fit the model to a range of latent-space dimensions $p = 1 \dots 9$. We illustrate these results in Figure 4.1.5.

The `inner product` model again performed poorly, often worse than the baseline model. Adding an intercept term harms performance for the `distance` model. The time-series assumption overall improved performance for correlates of war and harmed performance for Mechanical Turk labels.

We note that the time-series models performed better than the static model for the CoW labels but *not* for the Mechanical Turk labels. One possible explanation is that the formal relationships between nations – as accurately represented by expert labels – is indeed changing over time; while the lay relationships between these nations – as determined by lay interpretations of nations' relationships – remains more static over time.

Improvement due to zero-reversion regularization A further explanation for the decrease in performance for the time-series models (compared to the static model) is sensitivity to parameters. The static models have one parameter for each link function: the prior of nations' positions σ_c^2. In the dynamic model, we must set the priors over nations' positions $\sigma_{c,d}^2$ for each interaction, chain



Mechanical Turk Correlates of War

Figure 4.6: Positions of selected nations according to the static issue-adjusted model for articles labeled with Amazon Mechanical Turk (left) and Correlates of War (right). Nations' positions were inferred with the intercept / distance model, with distance dimension p=2. Intercepts are illustrated by color.

variance σ^2_{chain} , and zero-reversion variance σ^2_p. We selected chain variance $\sigma^2_{\text{chain}} = 0.0001$ and zero-reversion variance $\sigma^2_p = 1$ and 0.01 by grid search for these models at 3 dimensions and report results above based on the setting which worked best for each model. Setting σ^2_p had a substantial impact on model performance for the inner product models.

A closer look

What relationships between nations does this model infer? Because the relationships between nations are treated as functions of their positions $\boldsymbol{x} \in \mathbb{R}^p$, we can interpret these nations' positions $\boldsymbol{x}$ as summaries of nations' geopolitical orientations. We illustrate the positions of selected nations in Figure 4.6.

With both CoW and AMT labels, the relationships between nations can be inferred from the distance between their positions. In Figure 4.6, the United States stands out from a cluster of other nations, with Iraq, Iran, and Afghanistan—nations with which the U.S. has been at odds in the past twenty years—furthest away.

Correlates of War and Mechanical Turk labels provide different patterns of inter-nation sentiment. Nations' positions under CoW tend to be very clustered, with a few outliers, while their positions under AMT labels are more uniformly distributed. However, the two datasets provide extraordinarily consistent measures of nations' relationships.

To measure the consistency of these two models, we measured the Spearman rank correlation

coefficient

$$\underset{c_1,c_2 \in C, c_1 \neq c_2}{\text{Correlation}} (d_{\text{AMT}}(c_1, c_2), d_{\text{CoW}}(c_1, c_2))$$

between all pairs of nations c_1, c_2 in the set C of nations. The two-dimensional `intercept/distance` models have a Spearman rank correlation coefficient of $\sigma = 0.900$. Of course, these $\binom{|C|}{2}$ distances are far from independent, and a single outlier in each model could skew the correlation. To mitigate any such effect, we also measured the average correlation coefficient

$$\frac{1}{|C|} \sum_{c_1 \in C} \underset{c_2 \in C, c_2 \neq c_1}{\text{Correlation}} (d_{\text{AMT}}(c_i, c_2), d_{\text{CoW}}(c_i, c_2)),$$

which was *even higher*, at $\sigma = 0.901$. Recall that this is higher than the correlation coefficient between the original labels ($\sigma = 0.196$) – an effect possible because these models remove noise.

Under the second metric of correlation, most per-nation correlations were very high: over 90% of nations had correlation coefficient higher than 0.86. One of the most-differently-represented nations in this collection under the two different label types was Iran, which accounted for 7% of documents; the per-Iran correlation coefficient $\underset{c \in C, c \neq \text{Iran}}{\text{Cor}} (d_{\text{AMT}}(\text{Iran}, c), d_{\text{CoW}}(\text{Iran}, c)$ was 0.65 (higher only than Eritrea, which was 0.62 but accounted for 0.2% of documents).

Mutual sentiment with the United States and differences between CoW and AMT model fits. We illustrate mutual sentiment with the United States for a selection of these nations over time in Figure 4.1.5. To estimate the sentiment in these plots, we fit the `intercept/distance` model with $\dim(\boldsymbol{z}) = 2$. We summarize major events for two of these nations below.

- **Ukraine** was emancipated in 1991 with the dissolution of the Soviet Union. The U.S. has given Ukraine over $4.1 billion in aid, targeted to "promote political, security, and economic reform and to address urgent social and humanitarian needs" (State Department, 2012b). In return, Ukraine has been an active member of the UN and has assisted the NATO allies with defense aid in Kosovo (1999), Afghanistan (2011), Iraq, the Middle East, and Africa. Ukraine adopted its first post-Soviet constitution June 28, 1996, the same year taking part in the Olympics for the first time as an independent nation (the Olympics were hosted in the U.S. that year). At the same time, Ukraine has been taking active steps in eliminating the nuclear weapons program it inherited, permanently closing the last operating reactor at the Chernobyl site in 2000 (State Department, 2012b).

Ukraine's sentiment with Iraq, as inferred from the AMT model, was at its lowest in January 1993, January 1998, and again in April 2006. Its CoW sentiment with Iraq was at its lowest in May 2006, April 2003, and February 1991 (technically before its independence, during the Persian Gulf War). Its relationship with the U.S. was much stronger than with Iraq, peaking in 1996 (AMT) and June 2002 (CoW), when it supported the U.S. invasion of Iraq.

- **Iran** has had a poor relationship with the United States since the U.S. Embassy seizure in 1981. Between 1987 and 1988, U.S. and Iranian forces clashed in the Persian Gulf. (CIA Factbook, 2012). Transfers of power have since then increased political tension, with the election of a reformist president in 1997 and a reformist legislature in 2000, followed by conservative re-elections starting in 2003 and continuing through 2004. Hardliner President Mahmud Ahmadinejad was inaugurated in August 2005 and re-elected in 2009 (CIA Factbook, 2012).

  Ahmadinejad's rule has been met with increasing pressure from the United Nations. The Council has made successive resolutions imposing sanctions on Iran in 2006, 2007, 2008, and 2010 (State Department, 2012a).

  The Mechanical Turk sentiment between Iran and the U.S. has clearly dropped in the lead-up to Ahmadinejad's election (see again Figure 4.1.5), but this contrasts with the Correlates of War sentiment, which was lowest in 1988, when AMT sentiment was not as low.

Both of these low periods with Iran are clearly periods of bad relationships between these nations, but why did one model pick up sentiment in one case and not the other? This could be explained in part because the tension picked up by the CoW labels was unilateral, while the tension picked up in the later period did not fall under the dictum of CoW labels: the U.S. and Iran were neither at war nor having a territorial dispute. Instead, the U.S., as a member of the U.N., has supported Iran sanctions.

4.2 A comparison with unsupervised relationship mining

The preceding approach has limitations, of course. First, sentiment labels measure only one kind of interaction: whether nations are at war or peace. In reality, relationships between nations may be characterized in many ways, some of which are independent of the $[war, peace]$ dimension. For example, the relationship between nations may be characterized by trade in goods, or by the exchange of culture and ideas. Another limitation to a supervised sentiment model is that labels of the sentiment between nations may be unavailable or limited, or (as we saw before), the labels may be

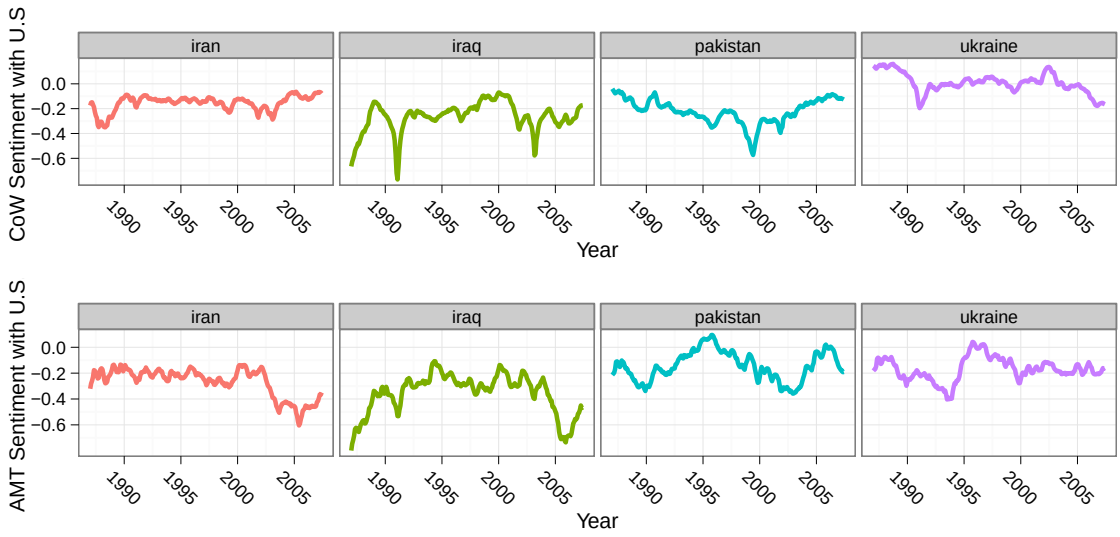


Figure 4.7: Selected nations' relationships with the United States over time. Each line in the plot above represents a specific nation's relationship with the United states inferred with the intercept/distance link function, with a two-dimensional distance space, using CoW labels (top) and AMT labels (bottom). Sentiment between all nations and either Iran or Pakistan was least consistent between CoW and AMT. Ukraine was the most consistently represented with these labels.

noisy.

In this section we will briefly compare the results of the previous model with the results of an unsupervised model. Because the unsupervised model is preliminary, we leave details of it in Appendix B.4. The unsupervised sentiment model uses the same latent-space assumption that we introduced in Sections 4.1.2 and 4.1.3. The curious reader can refer to the appendix for a fuller description of these assumptions.

Topics

A key assumption behind the unsupervised sentiment model is that each document can be described by a mixture of four topics. Two of these topics correspond to the two nations discussed in the paragraph (there are C of these topics, one for each nation). A third topic is a "background" topic, and the fourth topic is one of two sentiment topics. This sentiment topic is linked to a distance-based latent-space model exactly as in the last section, but with a binary sentiment indicator instead of a real-valued sentiment.

By fitting the unsupervised model, we learn which words are most-likely in each of these topics. With these assumptions, we inferred a set of topics using the same NYT corpus for the supervised sentiment model. Table 4.8 lists the most-likely words from a sample of topics fit to these twenty

years of articles.

State-specific topics $\beta_{C,\cdot}$. The state-specific topics describe words used when one of these nations is mentioned in text. Many of these topics are intuitive: "oil" shows up in the Iran topic, and "drug" is the top word in the Mexico topic. As these nations are mentioned in a major U.S. news source, the topics are sometimes biased toward ideas specific to the U.S. relationship with these nations (for example, "border" and "traffickers" in the Mexico topic). The U.S. topic contains phrases specific to policy and leadership.

While these topics are intuitive, they serve little role in analyzing this collection. From a modeling perspective, they serve as a "sponge" to explain away words commonly used to describe a nation, especially when those words might otherwise be interpreted to refer to a specific relationship.

An economics/military dichotomy. The sentiment topic, on the other hand, appears to demonstrate that one of the most prominent directions of variance in the text of paragraphs corresponds to the sentiment that we have been measuring.

Again using the convention that $\kappa_d = 1$ indicates *negative* sentiment between nations, the negative-sentiment topic $\beta_{S,1}$ matches our intuition: it contains words typically associated with conflict: *military, officials, soldiers, killed, troops*, and *police* are among the top words. On the other hand, the words most likely in the supplementary topic $\beta_{S,0}$ are associated more with economics: *million, percent, people, billion, oil*, and *officials*.

Are these words the same words that tend to be associated with expert labels of sentiment? To quantify this, we used coefficients from the text regression fit to Mechanical Turk ratings in the last section. Among the top 12 words in this topic (shown in Table 4.8), we estimated the average coefficient learned in the supervised sentiment model. The average coefficient for these terms was -0.225, which is less than the mean 0.007 of the entire vocabulary ($p < 0.02$ by a 2-sample t-test). The mean of the per-word sentiment β for this collection of words was at the $20th$ percentile of words in the vocabulary.

This contrasts with the top words in either the complementary topic $\beta_{S,0}$ or the background topic β_B. The top words in these topics had respective mean sentiments β of $-0.11, -0.08$. Neither of these was statistically noteworthy ($p = 0.15, 0.16$).

Background topic (β_B)
war
political
officials
country
people
military
international
peace
confirmed
week
following
government

Economics topic ($\beta_{S,0}$) **vs.**	**Military topic** ($\beta_{S,1}$)
million	military
percent	officials
people	soldiers
billion	killed
oil	troops
officials	police
country	forces
money	people
aid	attack
government	border
companies	near
military	air

Pakistan ($\beta_{C,\text{Pakistan}}$)	**Mexico** ($\beta_{C,\text{Mexico}}$)	**Israel** ($\beta_{C,\text{Israel}}$)
nuclear	drug	peace
weapons	officials	territories
military	border	occupied
officials	law	talks
terrorism	enforcement	officials
border	traffickers	negotiations
war	agents	agreement
government	police	state
aid	authorities	settlement
support	trade	security

United States ($\beta_{C,\text{United States}}$)	**Iran** ($\beta_{C,\text{Iran}}$)	**China** ($\beta_{C,\text{China}}$)
officials	nuclear	rights
military	war	human
official	program	trade
policy	weapons	relations
political	officials	officials
support	arms	nuclear
government	oil	visit
meeting	hostages	political
leaders	gulf	democracy
administration	uranium	economic

Figure 4.8: Per-nation topics ($\beta_{C,\cdot}$), a background topics ($\beta_{B,0}$), and the two interaction topics ($\beta_{S,0}, \beta_{S,1}$).

4.3 Conclusions

In this chapter we took a closer look at the story within a collection of documents. To do this, we reviewed a model for representing the relationship between countries, and we saw that this model provides an empirically meaningful benefit over simpler baselines. We also demonstrated that the predicted sentiment between pairs of countries with two entirely different sets of labels was strikingly similar. We finally demonstrated that an unsupervised model can produce a sentiment dimension aligned with our conception of inter-nation sentiment.

The set of assumptions we used in this chapter provide a broad view of global politics. Unfortunately it provides no sense for the internal factors motivating the positions countries take within the latent space. In the following chapter we will zoom in to take a closer look at how politicians within a country—the United States in particular—make decisions. To do this, we will use the text of the bills on which they are voting to better understand the positions they take. By using the text of bills, we will also overcome some limitations of a traditional model of how lawmakers vote.

We will continue to see two of the primitives discussed in this chapter. The traditional model of how lawmakers vote is in fact very much like the latent-space model we described in this chapter, and lawmakers' positions within this latent space are widely disseminated statistics. Second, we will continue to see that tools for text analysis – both mixed-membership models and text regression – can provide meaningful extensions of this model.

Chapter 5

Predicting Legislative Votes with Text Models

In the United States, as in many Western democracies, laws are made by committees of lawmakers. A defining characteristic of these committees is that each member casts a vote indicating whether she supports or rejects the proposed legislation. Legislative behavior centers around these votes, and it is a common goal of quantitative political science to characterize patterns of lawmakers' behavior with these votes. Voting behavior exhibits enough of a regularity that simple statistical models easily capture the broad political structure of legislative bodies.

One of these models is the *ideal point model*, a mainstay in quantitative political science for analyzing votes (Clinton *et al.*, 2004). It posits a latent "political space" along the real line and assumes each lawmaker has a position in that space; bills take a position in a related latent space (look ahead to Figure 5.2 for an intuition of these positions). A lawmaker's probability of voting `Yea` on pending legislation is then characterized by her position on this real line and parameters specific to that legislation.

Just as we saw with the last chapter's spatial models, ideal point models can be used to interpret lawmakers' positions on the political spectrum and to represent votes meaningfully.[1] However, ideal point models have certain limitations. One important limitation of these models is that they are not *predictive* models: while they can be used to model the bills that have been voted on, they cannot be used to predict lawmakers' votes on new bills. (A second limitation of these models is that lawmakers do not fit neatly into the assumptions made by such models. We address this limitation

[1] The interpretation of a lawmaker's latent position and a bill's position in the same space are slightly more nuanced than the last chapter. We clarify this relationship in the next section.

in the next chapter.) In this chapter we will extend the ideal point model so that we can make predictions about how lawmakers will vote on bills before these bills have seen a single vote. We will do this by using the text of bills to make this prediction.

Using text to predict future votes

These limitations dovetail with increasing access to both public records and tools for algorithmic text analysis. In the past decade, the text of congressional bills and other government records has become readily available to the broad public and research scientists. Websites like the Library of Congress's `thomas.loc.gov` release this information to the public, and sites like `www.govtrack.us` collect this information, synthesize it, and make it available for researchers and the public to better understand both the content and behavior around legislative decision-making (Govtrack website, 2010).

Just as text has become more available in this field in digitized formats, tools for text analysis have matured. Tools which were once available only to computational linguistics are becoming more familiar to political methodologists (Zimmer and Stewart, 2012). Topic models have evolved from vector-space models such as latent semantic analysis (Deerwester *et al.*, 1990) into probabilistic topic models (Hofmann, 1999; Blei *et al.*, 2003), which can be used as modules in more sophisticated statistical models.

In the next two chapters, we will take advantage of this broader availability of digitized text collections and tools for text analysis to address the above shortcomings of ideal point models. We begin this chapter by reviewing ideal point models (Poole and Rosenthal, 1985, 1991; Jackman, 2001; Martin and Quinn, 2002; Clinton *et al.*, 2004). After describing ideal point models, we will describe how to combine ideal point models with the models of text used in Chapters 3 and 4, including topic models (Blei *et al.*, 2003) and text regression (Kogan *et al.*, 2009), to enable us to predict votes on previously-unseen bills. Through this chapter and the next, the abstraction enabled by latent variable models will enable us to address these shortcomings of ideal point models with intuitive solutions.

5.1 The ideal point model

U.S. lawmakers' votes are captured during *roll call votes*, public records of lawmakers' votes on pending legislation. We can represent these votes as a matrix, with lawmakers in the rows and proposed legislation in the columns. We illustrate a sample of roll call votes for the United States

Senate in Figure 5.1.

Ideal points

Roll-call votes like this are often modeled with ideal point models. Ideal point models are based on item response theory, a statistical theory that models how members of a population judge a set of items. Loosely, an ideal point model assumes that each lawmaker u is described by a latent position $x_u \in \mathbb{R}$ summarizing her political preferences. A lawmaker's (stochastic) voting behavior is characterized by the relationship between her position in this space and the bill's position (Poole and Rosenthal, 1985, 1991; Jackman, 2001; Martin and Quinn, 2002; Clinton *et al.*, 2004).

In fact, we can motivate ideal points with explicit behavioral assumptions. Following the treatment in Clinton *et al.* (2004), we assume that a proposed item of legislation d would, if passed, move the current state of the world from the status quo $\boldsymbol{\zeta}_d \in \mathbb{R}^p$ to a new location $\boldsymbol{\psi}_d \in \mathbb{R}^p$. Lawmaker u observes the utility of each of these positions based on her ideal point $\boldsymbol{x}_u \in \mathbb{R}^p$ with noisy quadratic loss $||\boldsymbol{\zeta}_d - \boldsymbol{x}_u||^2 + \varepsilon_1$ and $||\boldsymbol{\psi}_d - \boldsymbol{x}_u||^2 + \varepsilon_2$, where $\varepsilon_1, \varepsilon_2$ follow an extreme value distribution. She will cast a vote toward whichever outcome maximizes her utility. These positions therefore represent each lawmaker's ideal "state of the world" (where passage of a bill moves this state of the world). For this reason, lawmakers' positions x_u are often called their *ideal points*.

Reparameterizing, we can write the probability $p(v_{ud}|\boldsymbol{X}_u, \boldsymbol{\zeta}_d, \boldsymbol{\psi}_d)$ of an affirmative vote with the probit or logistic function (Clinton *et al.*, 2004). Setting $\boldsymbol{b}_d = 2(\boldsymbol{\zeta}_d - \boldsymbol{\psi}_d)$ and $a_d = (\boldsymbol{\psi}_d^T \boldsymbol{\psi}_d - \boldsymbol{\zeta}_d^T \boldsymbol{\zeta}_d)$, we have

$$p(v_{ud} = \textbf{Yea}|\boldsymbol{a}_d, b_d, \boldsymbol{x}_u) = \sigma(\boldsymbol{x}_u^T \boldsymbol{a}_d + b_d), \tag{5.1}$$

Example roll call votes

Lawmaker	Item of legislation				
Bill	S. 3930	H.R. 5631	H.R. 6061	H.R. 5682	S. 3711
Mitch McConnell (R)	Yea	Yea	Yea	Yea	Yea
Olympia Snowe (R)		Yea	Yea	Yea	Nay
John McCain (R)	Yea	Yea	Yea	Yea	Yea
Patrick Leahy (D)	Nay	Yea	Nay	Nay	Nay
Paul Sarbanes (D)	Nay	Yea	Nay	Yea	Nay
Debbie Stabenow (D)	Yea	Yea	Yea	Yea	Yea

Figure 5.1: A sample roll-call matrix illustrating lawmakers' votes on items of legislation. These votes are from the Senate in the 109th Congress (2005-2006). The party of each Senator – (D)emocrat or (R)epublican – is provided in parentheses. The matrix of roll calls is sometimes incomplete (see Snowe's vote on S. 3930, for example).

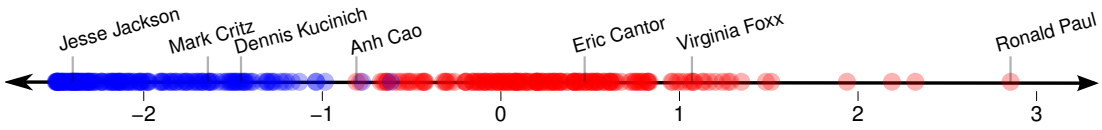


Figure 5.2: Example one-dimensional ideal points from the 111th House of Representatives. Ideal points represent lawmakers' voting preferences. Democrats are blue and Republicans are red.

where $\sigma(s)$ is the logistic function $\frac{\exp(s)}{1+\exp(s)}$. [2] Legislation d can therefore be fully characterized by specifying its *polarity* $\boldsymbol{a}_d$ and its *popularity* b_d. [3] When the popularity of a bill b_d is high, nearly everyone votes "Yea" on bill d; when the popularity is low, nearly everyone votes "Nay". When the popularity is near zero, the probability that a lawmaker votes "Yea" depends on how her ideal point x_u interacts with bill polarity a_d. We will make the common assumption that the latent variables a_d, b_d, and x_u have standard normal priors (Clinton *et al.*, 2004).

Given a matrix of votes, we use posterior inference to estimate the ideal point of each lawmaker, which reveal their intuitive political preferences. Figure 5.2 illustrates that ideal points fit to the U.S. House of Representatives from 2009-2010 clearly separate lawmakers by their political party. In U.S. politics, these inferred positions correspond to the commonly-known political spectrum: right-wing lawmakers are at one extreme, and left-wing lawmakers are at the other.

5.2 A model for predicting votes with the text of new bills

In this section, we extend ideal point models to use the text of bills to estimate a bill's polarity and popularity. This gives a new way of exploring and analyzing the government record and, further, gives a useful predictor of government. While traditional methods can only fill in missing votes, we develop tools that can predict how lawmakers will vote on a new bill. We will study the predictive accuracy of votes on new bills, where we use a spatial voting model as a "cold" prediction mechanism.

We will describe several models that connect the voting patterns of lawmakers to the original text of bills. One of these models embeds the statistical assumptions of *supervised topic modeling* (Blei and McAuliffe, 2008) into the ideal point model, where the locations of the bills are predicted from the latent topics in their texts. This model—the ideal point topic model—can predict complete votes on pending bills and provides a new way of exploring how legislative language is correlated with political support. The other models predict inferred ideal points using different forms of regression on phrase counts.

[2] The probability σ is sometimes taken to be probit; this amounts to $\varepsilon_1, \varepsilon_2$ taking on the Normal distribution.

[3] Popularity is also called *difficulty*, and polarity is called *discrimination*, in the context of educational testing applications of this model (Clinton *et al.*, 2004). We move away from these terms in favor of more appropriate terms for this application.

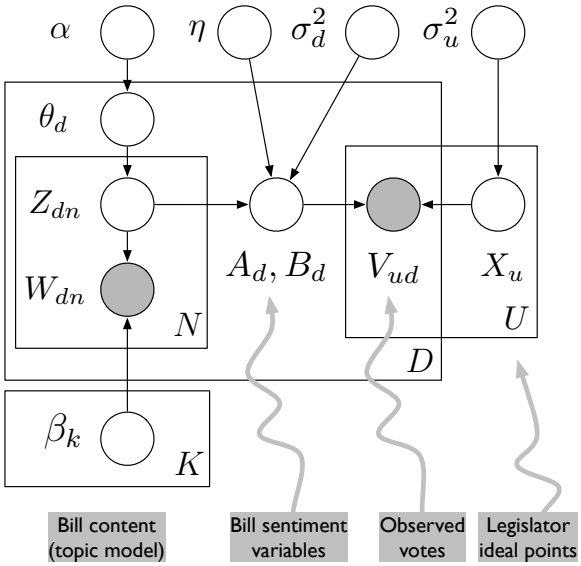


Figure 5.3: The ideal point topic model. Priors over the multinomials θ_d and β are both symmetric Dirichlet distributions.

In the following sections, we review the details of ideal point estimation and develop several models for predicting votes from legislative text. We derive an approximate posterior inference algorithm for ideal point models based on variational methods and analyze six Congresses (12 years) of legislative data from the United States Congress. Given a legislative history, these models can accurately predict votes on future legislation. One of these models, the ideal point topic model, can help summarize and visualize the political landscape of a government body based both on the voting patterns of its members and the language of its issues.

We now develop models relating the text of a bill to the variables a_d and b_d. Associating text to bill variables has a predictive advantage because new bills can be situated in the space of ideal points. It also has an interpretive advantage because language becomes associated with political sentiment.

Modeling ideal points with text regression. We developed two predictive ideal-point models which use text regression (Kogan *et al.*, 2009). For these, we first fit an ideal-point model to a training set of bills and all lawmakers using the variational algorithm described in Section 5.1. We then fit Lasso regression[4] (`LARS`)[5] and Ridge regression (`L2`) to these bills' parameters a_d, b_d using a vector of their n-gram[6] counts $\boldsymbol{w}_d$ as covariates.

Modeling ideal points with supervised topics. The text regression models link individual words or phrases to bill sentiment. In this section, we connect textual *themes* with bill sentiment.

[4]Implemented in the "penalized" package for R
[5]implemented with the "lars" package for R
[6]See Section 5.3 for details.

We refer to this model as an ideal point topic model (IPTM).

To model themes, we use the assumptions of supervised Latent Dirichlet Allocation (sLDA) (Blei and McAuliffe, 2008). As in Latent Dirichlet Allocation (Blei *et al.*, 2003), each bill is represented as a mixture of latent topics θ_d, where each of K topics β_k is a multinomial probability distribution over terms. For the n^{th} term of bill d, we draw topic z_{dn} from Mult(θ_d), and then draw word w_{dn} from the topic β_{z_n}.

Like sLDA, the ideal point topic model further assumes each bill d is attached to a response variable. In this case, the response variable is the 2-component vector of bill variables (a_d, b_d). The distribution of the response is a linear model whose covariates are the empirical distribution of the topics $\boldsymbol{z}_d$ for the bill,

$$\begin{aligned} a_d \sim & \qquad \mathcal{N}(\boldsymbol{\eta}_a^\top \bar{z}_d, \sigma_d^2) \\ b_d \sim & \qquad \mathcal{N}(\boldsymbol{\eta}_b^\top \bar{z}_d, \sigma_d^2), \end{aligned}$$

where $\bar{z}_d = (1/N) \sum_n z_{dn}$. This setting is more complex than the original sLDA model: the response variables are *hidden*—they are not observed directly, but are used downstream in the voting model.

Finally, we add a Gaussian prior to $\boldsymbol{\eta}$. The full model is represented as a graphical model in Figure 6.2.

The only observed variables in the model are the bill texts and votes. Our goal in fitting this model is to uncover the posterior

$$p(a_d, b_d, x_u, \boldsymbol{\eta}, \beta, z, \theta | \boldsymbol{W}, \boldsymbol{V}), \tag{5.2}$$

which can then be used in exploratory or predictive tasks. Conditioned on these variables, our analysis proceeds with the posterior distribution of the ideal points, polarities and popularities, topics, and coefficients. Computing the posterior exactly is intractable, so we use variational inference to approximate it. We describe this in further detail in Section 6.2.

This posterior allows us to explore the connection between language and political tone. For example, the coefficients $\boldsymbol{\eta}$ are a direct connection between bills' topics and the political tone of these bills. Examples of this are provided in Section 5.3. The topics β, learned from both text and votes, provide a lexical window into legislative issues. The parameters $\boldsymbol{\eta}, \beta$ together also allow us to predict votes using the text of new bills; Section 6.2 provides detail about this.

Multimodal solutions and identification

Note that a fit of the ideal point model has multiple modes. In one mode, Democrats tend to have positive ideal points, while Republicans are negative; in another, Republicans are positive, while Democrats are negative. To keep fits of the different models identifiable, several researchers have applied nonzero priors over specific lawmakers to encourage the model to prefer one of these modes (Jackman, 2001; Clinton *et al.*, 2004; Martin and Quinn, 2002).

In the study in Section 5.3, we anchor four lawmakers with strong priors ($\sigma_d = 10^{-3}$) at ideal points ± 4. We select two congresspersons from each chamber and two from each party: Kennedy (S-Dem) and Waxman (H-Dem) are centered at +4 and Enzi (S-Rep) and Donald Young (H-Rep) are centered at -4.[7] We selected these Senators for consistency with previous work such as Clinton *et al.* (2004). We selected the Representatives because they have held long offices in the House. Without these sharp priors, the model still discovers ideal points which cleanly separate political parties but may converge on "opposite" modes in different fits. With the priors, we obtain consistent ideal points at the expense of predictive performance.

Related work

Ideal point models, a form of spatial voting model, have roots as far back as the 1920s (Enelow and Hinich, 1984). They are fit by both frequentist (Poole and Rosenthal, 1985; Heckman and Snyder, 1996) and Bayesian methods (Jackman, 2001; Martin and Quinn, 2002; Clinton *et al.*, 2004), have been embedded in a time series (Martin and Quinn, 2002; Wang *et al.*, 2010), and have been developed for higher dimensional political spaces (Jackman, 2001; Heckman and Snyder, 1996).

Topic models have been applied to Senate speeches, such as to discern "the substantive structure of the rhetorical [legislative] agenda" (Quinn *et al.*, 2006). They have also been used with legislative speeches to gauge lawmakers' sentiment toward legislation using roll-calls (Thomas *et al.*, 2006). Modeling sentiment in text is more generally discussed in the field of sentiment analysis; see Pang and Lee (2008) for a review.

The ideal point topic model relates closely to user-recommendation models based on matrix factorization (Salakhutdinov and Mnih, 2008). Matrix factorization methods for recommendation are akin to large-scale spatial behavior models (though usually with no "popularity" term, which acts as an intercept). Many of these matrix factorization models for user recommendation do not provide a method of predicting one user's item preference without other users' preferences on the

[7] This value was selected to be large yet not completely out of the ordinary.

same item.

Two works stand out as closely related to this work. One of these is fLDA, which models binary or continuous ratings with user affinity to topics (Agarwal and Chen, 2010). Another is Wang *et al.* (2010), who describe a similar application by combining topic models and matrix completion. Their work also draws on ideal point models, models transitions over time, and is designed to learn the dimensionality of the latent factors. Under the generative assumptions of their model, bills and matrix cells (e.g., votes) are conditioned on a shared mixture; in our model, votes are conditioned on words' topics.

Posterior estimation for the ideal point topic model

Computing the posterior in Equation 5.2 is intractable. Posterior inference for traditional Bayesian ideal point models is traditionally implemented with MCMC methods such as Gibbs sampling (Johnson and Albert, 1999; Jackman, 2001; Martin and Quinn, 2002; Clinton *et al.*, 2004). However, in the ideal point topic model, fast Gibbs samplers are unavailable because the conditionals needed are not analytically computable; an MCMC strategy would require a more complicated sampling scheme. We therefore use an alternative algorithm—which can be applied to both the standard ideal point model and the ideal point topic model—which uses variational methods (Jordan *et al.*, 1999).

Recall from Chapters 2 and 3 that variational inference requires specification of a variational distribution which will serve as a proxy for the true posterior distribution. Word assignments z_{dn} and topic proportions are governed by multinomial parameters $\boldsymbol{\phi}_d$ and Dirichlet parameters $\boldsymbol{\gamma}_d$, as in LDA (Blei *et al.*, 2003). The variational distribution for lawmakers' ideal points x_u; bills' parameters a_d, b_d; and coefficients $\boldsymbol{\eta}$ are Gaussian with respective means τ_u, $\tilde{a}_d$, $\tilde{b}$, $\hat{\eta}$ and variances σ^2_τ, $\sigma^2_{\tilde{a}}$, $\sigma^2_{\tilde{b}}$, and σ^2_η. The variational distribution is

$$q(\tau,\sigma_\tau,\tilde{a},\sigma_{\tilde{a}},\phi,\theta) = \prod_u q(x_u|\tau_u,\sigma^2_\tau) \prod_D q(a_d,b_d|\tilde{a}_d,\sigma^2_{\tilde{a}}) \prod_D q(\theta_d|\gamma_d) \prod_{N_d} p(z_n|\phi_n) q(\boldsymbol{\eta}|\hat{\eta},\sigma_{\hat{\eta}}). \quad (5.3)$$

Inference proceeds by minimizing the KL between the variational posterior (Equation 5.3) and the true posterior (Equation 5.2), which is equivalent to maximizing a lower bound on the marginal probability of the observations. Coordinate ascent only works for some of the random variables, but we must use gradient ascent on a_d, b_d, and x_u. We give further details of the variational inference algorithm in Appendix B.5.

Prediction After they are fit to lawmakers' votes and bill text, the variational parameters τ,

$\hat{\eta}$, and β can be used to estimate the vote of each lawmaker on a *new* bill d using its text. To predict whether lawmaker u votes `yea` on d, the per-word parameters ϕ_n of d are estimated using the topics β. Once ϕ has been estimated, the probability of a `yea` vote is given by $p(v_{ud} = \texttt{yea}) = \sigma(\tau_u(\bar{\phi_d}\hat{\eta}_b) + \bar{\phi_d}\hat{\eta}_a)$ [8], where $\bar{\phi_d}$ is $\frac{1}{N_d}\sum_{N_d}\phi_n$. In practice, we fit $\hat{\eta}$ with no regularization after the model has converged. This gives slightly better results which are more robust to parameter selection.

5.3 An empirical analysis

Analyzing the U.S. House and Senate

We studied the performance of these models on 12 years of data from the United States House of Representatives and Senate. We first demonstrate how the ideal point topic model can be used to explore legislative data; then we evaluate the models' generalization performance in predicting votes from bill texts.

We collected roll-call votes for Congressional sessions 106 through 111 (January 1997 to January 2011). We used votes about bills and resolutions, and only votes regarding the legislation as a whole (as opposed to, e.g., amendments of the legislation). We downloaded the data from `www.govtrack.us`, an independent Website which provides comprehensive legislative information to the public. Our collection contains 4,447 bills, 1,269 unique lawmakers, and 1,837,033 `yea` or `Nay` roll-call votes.

To select the vocabulary, we lemmatized (i.e., normalized the forms of) words in the bills with Treetagger (Schmid, 1994). Then we retained a vocabulary of statistically significant n-grams ($1 \leq n \leq 5$) using likelihood ratios. These n-grams were treated as terms.[9] We removed n-grams occurring in fewer than 0.2% of all bills and more than 15% of bills. We also removed an n-gram if it accounted for more than 0.2% of all tokens or fewer than than 0.001% of all tokens. After this process, our vocabulary contained 4,743 unique n-grams.

We used the anchor lawmakers described in Section 5.1. We ran variational inference until the change in increase in the objective function was less than 0.01%.

Exploring topics and bills

In this section, we examine a fit of the ideal point topic model for all the bills and votes of a session. This demonstrates the model's use as an exploratory tool of political data. For this analysis, we used

[8] The estimate $\mathbb{E}_q\left[\sigma(x_u(\bar{z_d}\eta_b) + \bar{z_d}\eta_a)\right]$ can be more theoretically justified, but results from the two estimates are (in practice) identical.

[9] When one n-gram subsumes another, we chose to observe the longer of the two

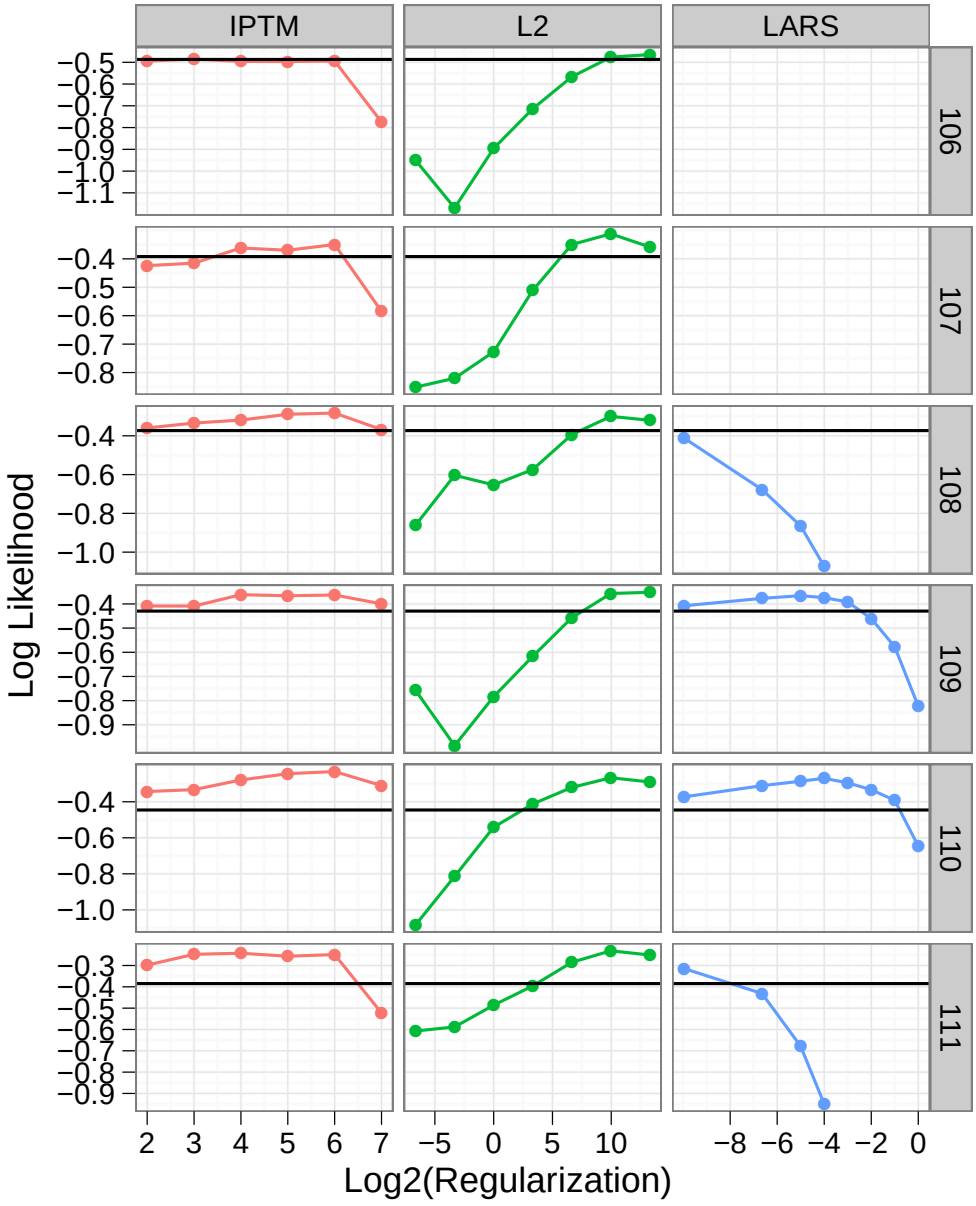


Figure 5.4: Vote log likelihood on heldout votes. Models are shown by color for different regularizations (x axis), for Congresses 106 to 111. For LARS and L2, the regularization is the complexity parameter; for the ITPM, the regularization is the the number of topics. The *yea* baseline is the horizontal black line. LARS is below the fold for 106-107. The ideal point topic model performs with less variance across its regularization parameter.

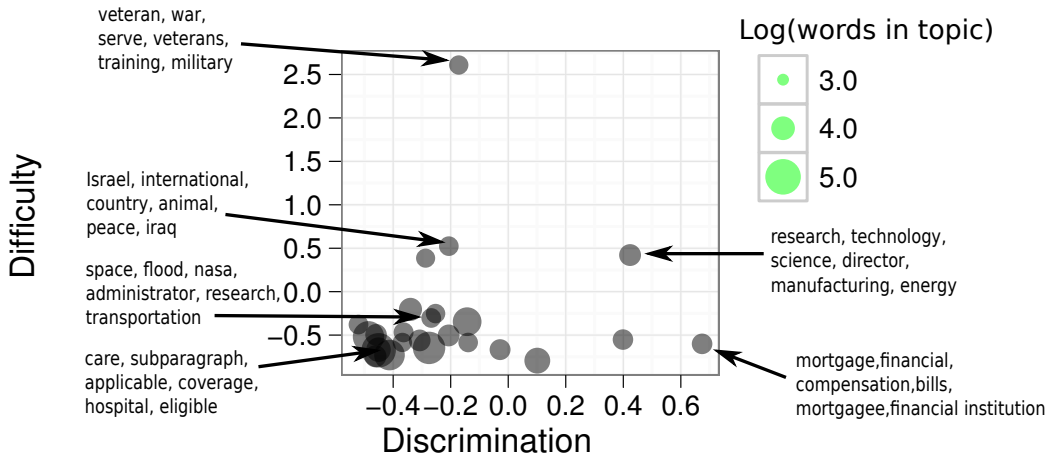


Figure 5.5: Topics can be visualized in the same latent political space as lawmakers and bills. This plot shows selected topics by coefficients $\hat{\eta}$, for a 64-topic model ($\hat{\eta}$s are normalized by mean and variance). Two topics (*people, month, recognize, ...* and *clause, motion, chair, ...*) with popularity 4.68 and polarity 7.4 (respectively) are not shown.

dispersion $\sigma_d = \sigma_u = 1.0$ and 64 topics. We focus on the 111^{th} session (January 2009 to January 2011).

Exploring topics with $\hat{\eta}$. As noted in Section 5.1, the coefficients $\hat{\eta}$ relate each topic's weight in a bill with the bill's popularity and polarity parameters. Figure 5.5 shows some example topics and their corresponding coefficients $\hat{\eta}$. Below we describe some of these topics in more detail and connect them to the data.

One popular topic in the 111th Congress focused on national recognition: *people, month, recognize, history, week, woman.* In contrast, the *least*-supported topic was more procedural, frequently appearing in bills under consideration or with many amendments (*clause, motion, chair, print, offer, read*). In this case, such legislation is sometimes summarily rejected before further consideration; the language of amendments is a signal that legislation is contentious.

While these topics often explained overwhelming support or rejection of legislation, much legislation was considerably more partisan.

Health Care. One contentious topic was about qualification for public health care: *care, subparagraph, applicable, coverage, hospital, eligible.* This topic was among the most-Democratic 10% of topics, in large part because it helped to explain the *Patient Protection and Affordable Care Act*, i.e. the "Health Care Bill" of 2009. Although this 906-page bill was barely passed: of the 311 Democrats voting on it, 276 voted in favor; of the 217 Republicans voting on it, none voted in favor. The model was moderately accurate on this bill: it correctly predicted 93.8% of votes. The two other topics highly expressed in this bill were about different aspects of public health, including

one about government health options (medicare and social security) and one about health insurance coverage; both were slightly Democratic.

NASA Authorization. Another contentious topic was about spaceflight: *space, flood, NASA, administrator, research, transportation.* This topic was expressed in one of the most-poorly predicted bills of the 111$^{\text{th}}$ Congress. This bill, the *NASA Authorization Act of 2010*, was a "compromise between the Obama administration, which wants... a commercial space industry in which private companies would transport astronauts, and House lawmakers, who wanted... one government-owned rocket" (Herszenhorn, 2010). In the house vote (a Senate record was not kept), of 249 Democrats voting on the bill, 185 voted in favor; of the 173 Republicans, 119 voted in favor. Because this bill had mixed but nonpartisan support, the model could not represent it well, with only 72% of votes correctly predicted.

Checking the ideal points

We can also use the in-sample fit to assess the quality of the ideal points of the lawmakers. In classical ideal point modeling, this is done via in-sample accuracy: How well does the model explain the observed votes?

The average per-lawmaker accuracy in the in-sample fit was 96% (only 10% of lawmakers had accuracy lower than 90%). As expected, accuracy increases with more votes ($\rho = 0.51$). Among lawmakers with over 100 votes, only two stand out. Donald Young (713 votes; accuracy 0.83) had a pre-defined ideal point (see Section 5.1). Ron Paul, a Republican in the 111$^{\text{th}}$ Congress, was also poorly predicted (761 votes; accuracy 0.84). Paul is known for his Libertarian beliefs, even having run for President for the Libertarian party in 1988.

The poor prediction of Paul points to a limitation of the 1-dimensional ideal-point model, which can only capture the two main parties, instead of a limitation of the supervised prediction: fitting votes to the classical ideal point model (ignoring bill text), Paul's in-sample accuracy was consistently poor across sessions. We will address this limitation in the next chapter by incorporating a bill's issues in the prediction task.

Predicting votes from text

Prediction on heldout bills. We measured predictive accuracy and log likelihood for these models under a variety of regularization settings (`LARS` is parameterized by $0 < f \leq 1$, `L2` is parameterized by regression coefficient $\Lambda \geq 0$, and `IPTM` is parameterized by topics K).

We also devised two baselines for comparison with the three models described so far. The first of these provides a lower bound: assume all votes are `yea`. Because the majority (85%) of votes in our corpus were `yea` votes, this presents a more reasonable overall baseline than random guessing (at 50%). We call this model the `yea` model. The second baseline fit a logistic regression trained for members of each party (with a separate one for mixed or independent lawmakers), with terms as covariates. This baseline (implemented with the R `glm` library) used too much memory to use more than 800 terms and therefore led to results worse than the `yea` baseline. We believe that a better baseline could be used.

For each 2-year period (called a Congress), the bills were partitioned into 6 folds. For each model, we iteratively (1) remove a fold, (2) fit the model to the remaining folds (by Congress), and (3) form predictions on the bills in the removed fold. Across folds, we thus obtain a complete data set of held-out votes.

Across all sessions, the `yea` baseline predicts votes correctly 85% of the time. The ideal point topic model is better, correctly predicting 89% of votes with 64 topics (this means that 62,000 more votes are correctly predicted). Overall performance for `L2` was best for $\Lambda = 1000$ (90%), and `LARS` was best at $f = 0.01$ (82%). While the ideal point topic model had lower accuracy than `L2`, its log-likelihood was nearly the same. These results are summarized in Figure 5.3, and further details are in Appendix B.6.

Sequential prediction. Our final study examined the performance of these models on predicting future votes from past votes. To do this, we fit a 64-topic `IPTM` and `L2` predictive models on the first $3, 6, 9, \ldots, 21$ months of a Congress.[10] We then tested these each of these fits on the following three months of unseen votes. The ideal-point topic model correctly predicted 87.0% of votes, and `L2` correctly predicted 88.1% of votes; their log-likelihood was identical.

With these models, one could predict 31,000 to 55,000 votes above the baseline, *based only on the text of the bills*. The simpler of the two models, `L2`, performs better at prediction.

5.4 Conclusions and limitations of these models

We have developed several models associating the text of legislation to lawmakers' voting patterns. These models provide a way of exploring large collections of legislative data and predicting the votes of new bills. The text-regression models and the ideal point topic model have incorporated bill texts into the simplest kind of ideal point model of roll call data.

[10] A bug prevented LARS from completing in most runs of this setting

Though we were motivated by (and focused on) political science data, we note that these models are among several (e.g., Agarwal and Chen (2010)) that can be applied in a variety of collaborative filtering settings. They provide a way to model a collection of users and their decisions about collections of textual items.

One of the central advantages of latent-variable models is their modularity. Because we have modeled the text of legislation as a vector of topics (or a vector of word counts), it is straightforward to incorporate other elements of the legislative process, such as speech transcripts (Quinn *et al.*, 2006; Thomas *et al.*, 2006) or bill sponsor, into this model's supervision. This could improve both the predictive power and exploratory capabilities of the ideal point topic model. The modularity of latent-variable models allows us to swap in modeling assumptions for each of these types of data.

However, even optimal features for prediction would be limited by the power of the downstream model for lawmakers' votes on bills. Here we have studied multiple topics with a one-dimensional political space. As noted in Section 5.3, this is a predictive bottleneck. (The "true" number of dimensions is debatable—Heckman and Snyder (1996) argued that there are at least 6 statistically significant dimensions in roll-call data, while Jackman (2001) barely found more than one.) One solution is to increase the dimension of the lawmaker and bill variables or use a mixture model as in Wang *et al.* (2010), which can increase the strength of the model at the expense of interpretability.

An alternative solution is to model individual lawmakers' affinities to issues, using ideas explored by Agarwal and Chen (2010) and Wang and Blei (2011) for matching users with text content. We will use these ideas in the following chapter, where we explicitly model lawmakers' positions on a variety of issues. This will allow us to represent lawmakers' votes better than an ideal point model while providing an interpretable window into individual lawmakers' voting behavior.

Chapter 6

Lawmakers' issue preferences in the U.S. Congress

In the last chapter we introduced several models for predicting lawmakers' votes on previously-unseen bills. One limitation of these models—and one-dimensional ideal point models in general—is that they were designed around a restrictive latent space: lawmakers are described by a single number, and the predictive performance of these models is bottlenecked by lawmakers for whom a single number is not sufficient.

Indeed, there are some votes that the traditional ideal point given in Equation 5.1 fails to capture. For example, Ronald Paul, Republican representative from Texas, and Dennis Kucinich, Democratic representative from Ohio, are poorly modeled by ideal points because they diverge from the left-right spectrum on issues like foreign policy. Because some lawmakers deviate from their party on certain substantive issues, their positions on these issues are not captured by ideal point models.

In this chapter we will develop the *issue-adjusted ideal point model*, a latent variable model of roll-call data that accounts for the contents of the bills that lawmakers are voting on. The idea is that each lawmaker has both a general position and a sparse set of position adjustments, one for each issue. The votes on a bill depend on a lawmaker's position, *adjusted* for the bill's content. The text of the bill encodes the issues it discusses. Our model can be used as an exploratory tool for identifying exceptional voting patterns of individual lawmakers, and it provides a richer description of lawmakers' voting behavior than the models traditionally used in political science.

In the following sections, we develop our model and describe an approximate posterior inference algorithm based on variational methods. We will again analyze six Congresses (12 years) of legislative

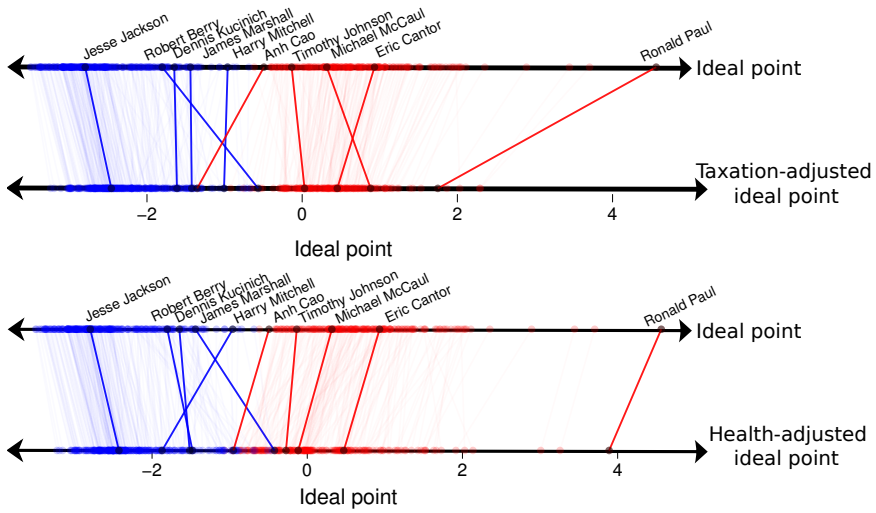


Figure 6.1: In a traditional ideal point model, lawmakers' ideal points are static (top line of each figure). In the issue-adjusted ideal point model, lawmakers' ideal points change when they vote on certain issues, such as *Taxation* (top) and *Health* (bottom).

data from the United States Congress. We finally show that our model gives a better fit to legislative data and provides an interesting exploratory tool for analyzing legislative behavior.

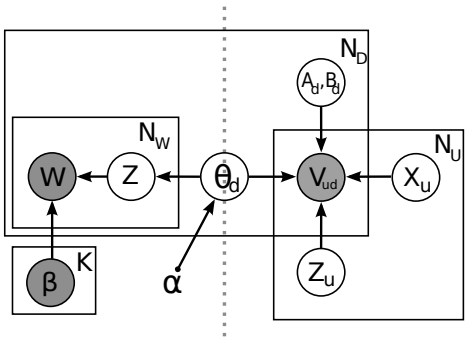
An additional contribution of this chapter is that we will also motivate an alternative algorithm for variational inference (which we fully describe in Appendix A) that will allow practitioners to iterate more quickly with their modeling assumptions.

6.1 A model of exceptional voting patterns

A one-dimensional ideal point model fit to the House of Representatives from 2009-2010 correctly models 98% of all lawmakers' votes on training data. But it only captures 83.3% of Baron Hill's (D-IN) votes and 80.0% of Ronald Paul's (R-TX) votes. Why is this?

The ideal point model assumes that lawmakers are ordered. Each bill d, described by polarization a_d and popularity b_d, splits them at a *cut point* $-\frac{b_d}{a_d}$. Lawmakers to one side of the cut point are more likely to support the bill, and lawmakers to the other side are likely to reject it. For lawmakers like Paul and Hill, this assumption is too strong because their voting behavior does not fit neatly into a single ordering. Their location among the other lawmakers changes with different bills.

These lawmakers do not change their positions randomly. They vote consistently within individual areas of policy, such as financial regulation and education. Paul consistently votes against United States involvement in foreign military engagements, a position that contrasts with other Republicans. Democratic representatives from New York are more likely to hold conservative positions



The issue-adjusted ideal point model

Terrorism		Commemorations	
terrorist	york	nation	serve
September	terrorist attack	people	percent
attack	Hezbollah	life	community
nation	national guard	world	family

Education		Transportation	
student	history	transportation	land
school	nation	minor	coast guard
university	child	print	substitute
charter school	college	tax	nature

Labeled topics

Figure 6.2: Left: the issue-adjusted ideal point model, which models votes v_{ud} from lawmakers and legislative items. Classic item response theory models votes v using x_u and a_d, b_d. For our work, documents' issue vectors $\boldsymbol{\theta}$ were estimated fit with a topic model (left of dashed line) using bills' words w and labeled topics β. Expected issue vectors $\mathbb{E}_q[\boldsymbol{\theta}|\boldsymbol{w}]$ are then treated as constants in the issue model (right of dashed line). Right: Top words from topics fit using labeled LDA (Ramage *et al.*, 2009).

on financial services regulation, even though they vote Democratically on social issues.

We refer to voting behavior like this as *issue voting*. An *issue* is any federal policy area, such as "financial regulation," "foreign policy," "civil liberties," or "education," on which lawmakers are expected to take positions. Lawmakers' positions on these issues will often diverge from their traditional left/right stances. The model we develop in this chapter captures this intuition, which we illustrated in Figure 6.1. Charles Djou is more similar to Republicans on *Taxation* (right) and more similar to Democrats on *Health* (left), while Ronald Paul is more Republican-leaning on *Health* and less extreme on *Taxation*.

6.1.1 Issue-adjusted ideal points

Suppose that there are K issues in the political landscape. We will use the words $\boldsymbol{w}_d$ of each bill d to code it with a mixture $\boldsymbol{\theta}_d$ of issues, where each element θ_{dk} corresponds to an issue; the components of $\boldsymbol{\theta}_d$ are positive and sum to one. (These vectors will come from a topic model, which we describe below.) In our proposed model, each lawmaker is also associated with a K-vector $\boldsymbol{z}_u \in \mathbb{R}^K$, which describes how her ideal point changes for bills about each issue.

We use these variables in a model based on the traditional ideal point model of Equation 5.1. As above, x_u is the ideal point for lawmaker u and a_d, b_d are the polarity and popularity of bill d. In our model, votes are modeled with a logistic regression

$$p(v_{ud}|a_d, b_d, z_u, x_u, \boldsymbol{w}_d) = \sigma\left((\boldsymbol{z}_u^\top \mathbb{E}_q[\boldsymbol{\theta}_d|\boldsymbol{w}_d] + x_u)a_d + b_d\right), \tag{6.1}$$

where we use an estimate $\mathbb{E}_q[\boldsymbol{\theta}_d|\boldsymbol{w}_d]$ of the bill's issue vector from its words $\boldsymbol{w}_d$ as described below.

We put standard normal priors on the ideal points, polarity, and difficulty variables. We use Laplace priors for $\boldsymbol{z}_u$: $p(z_{uk} \mid \lambda_1) \propto \exp(-\lambda_1||z_{uk}||_1)$. This enforces a sparse penalty with MAP inference and a "nearly-sparse" penalty with Bayesian inference. See Figure 6.2 (left) for the graphical model.

To better understand this model, assume that bill d is only about *Finance*. This means that $\boldsymbol{\theta}_d$ has a one in the *Finance* dimension and zero everywhere else. With a classic ideal point model, a lawmaker u's ideal point, x_u, gives his position on each issue, including *Finance*. With the issue-adjusted ideal point model, his *effective ideal point* for *Finance*, $x_u + z_{u,\text{Finance}}$, gives his position on *Finance*. The adjustment $z_{u,\text{Finance}}$ affects how lawmaker u feels about *Finance* alone. When $z_{u,k} = 0$ for all u, k, this model becomes the classic ideal point model.

This model lets us inspect lawmakers' overall voting patterns by issue. Given a collection of votes and a coding of bills to issues, posterior estimates of the ideal points and per-issue adjustments give us a window into voting behavior that is not available to classic ideal point models.

6.1.2 Using Labeled LDA to associate bills with issues.

Equation 6.1 adjusts a lawmaker's ideal point by using the conditional expectation of a bill's thematic labels $\boldsymbol{\theta}_d$ given its words $\boldsymbol{w}_d$. We estimate this vector using labeled latent Dirichlet allocation (LDA) (Ramage *et al.*, 2009).

Labeled LDA is a topic model, a bag-of-words model that assumes a set of themes for the collection of bills and that each bill exhibits a mixture of those themes. The themes, called topics, are distributions over a fixed vocabulary. In unsupervised LDA (Blei *et al.*, 2003) they are learned from the data. In labeled LDA, they are defined by using an existing tagging scheme. Each tag is associated with a topic; its distribution is found by taking the empirical distribution of words for documents assigned to that tag.[1] This gives interpretable names (the tags) to the topics.

We used tags provided by the Congressional Research Service (Congressional Research Service, 2011), which provides subject codes for all bills passing through Congress. These subject codes describe the bills using phrases which correspond to traditional issues, such as *Civil rights* and *National security*. Each bill may cover multiple issues, so multiple codes may apply to each bill. (Many bills have more than twenty labels.) We used the 74 most-frequent issue labels. Table 6.2 (right) illustrates the top words from several of these labeled topics.[2] We fit the issue vectors

[1] Ramage *et al.* (2009) explore more sophisticated approaches, but we found this simplified version to work well.

[2] After defining topics, we performed two iterations of unsupervised LDA with variational inference to smooth the

$\mathbb{E}[\boldsymbol{\theta}_d|\boldsymbol{w}_d]$ as a preprocessing step. In the issue-adjusted ideal point model (Equation 6.1), $\mathbb{E}[\boldsymbol{\theta}_d]$ was treated as observed when estimating the posterior distribution $p(x_u, a_d, b_d, \boldsymbol{z}_d|\mathbb{E}[\boldsymbol{\theta}_d|\boldsymbol{w}_d], v_{ud})$. We summarize all 74 issue labels in Appendix B.7.4

Related Work

Item response theory has been used for decades in political science (Clinton *et al.*, 2004; Martin and Quinn, 2002; Poole and Rosenthal, 1985); see Enelow and Hinich for a historical perspective (Enelow and Hinich, 1984) and Albert for Bayesian treatments of the model (Albert, 1992). Some political scientists have used higher-dimensional ideal points, where each lawmaker is attached to a vector of ideal points $\boldsymbol{x}_u \in \mathbb{R}^K$ and each bill polarization $\boldsymbol{a}_d$ takes the same dimension K (Heckman and Snyder, 1996). The probability of a lawmaker voting "Yea" is $\sigma(\boldsymbol{x}_u^T \boldsymbol{a}_d + b_d)$. The principle component of ideal points explains most of the variance and explains party affiliation. However, other dimensions are not attached to issues, and interpreting beyond the principal component is painstaking (Jackman, 2001).

Note that our goal in this chapter is fundamentally different than it was in the last chapter. The last chapter describes how to predict votes on bills which had not yet received any votes. Those models fit a_d and b_d using supervised topics, but the underlying voting model is one-dimensional: it cannot model individual votes better than a one-dimensional ideal point model. Along the same lines, Wang et al. created a Bayesian nonparametric model of votes and text over time (Wang *et al.*, 2010). Predicting votes on new bills is a non-goal in this chapter, in contrast to these related works (which do not model individuals' affinity toward issues).

The issue-adjusted model is conceptually more similar to recent models for content recommendation. Specifically, Wang and Blei (2011) describe a method to recommend academic articles to individuals, and Agarwal and Chen (2010) propose *fLDA* to match users to Web content (Agarwal and Chen, 2010). Agarwal et al. learn a separate user-item offset y_{ud} and a user-topic affinity which interacts with $\mathbb{E}_q[\boldsymbol{\theta}_d|\boldsymbol{w}_d]$. Wang and Blei (2011) fit a linear regression, again learning a user-topic affinity. Our model differs in its introduction of the polarity a_d: lawmakers take a position z_{uk} on issue k which only creates an affinity toward k if the bill leans the correct way. Finally, we have an explicit goal of interpretability.

word counts.

6.2 Inference for the adjusted ideal point model

With a way to map bills to issues, we turn to fitting lawmakers' issue adjustments $\boldsymbol{z}_u$. We estimate issue adjustments $\boldsymbol{z}_u$ by using the observed votes v and bills' issues $\boldsymbol{\theta}_d$ with the posterior distribution $p(x, \boldsymbol{z}, a, b|v, \boldsymbol{\theta})$.

Bayesian ideal point models are usually fit with Gibbs sampling (Johnson and Albert, 1999; Jackman, 2001; Martin and Quinn, 2002; Clinton *et al.*, 2004). However, fast Gibbs samplers are unavailable for our model because the conditionals needed are not analytically computable. We therefore estimated the posterior with variational Bayes.

Recall that in variational Bayes, we posit a family of distributions $\{q_{\boldsymbol{\eta}}\}$ over the latent variables that is likely to contain a distribution similar to the true posterior (Jordan *et al.*, 1999) and select η to minimize the KL divergence between the variational and true posteriors. In the ideal point topic model, we let $\{q_{\boldsymbol{\eta}}\}$ be the family of fully factorized distributions

$$q(x, \boldsymbol{z}, a, b|\boldsymbol{\eta}) = \prod_U \mathcal{N}(x_u|\tilde{x}_u, \sigma^2_{x_u})\mathcal{N}(\boldsymbol{z}_u|\tilde{z}_u, \lambda_{z_u}) \prod_D \mathcal{N}(a_d|\tilde{a}_d, \sigma^2_{a_d})\mathcal{N}(b_d|\tilde{b}_d, \sigma^2_{b_d}), \tag{6.2}$$

where above we parameterize our variational posterior with $\boldsymbol{\eta} = \{(\tilde{x}_u, \sigma_x), (\tilde{z}_u, \sigma_{\boldsymbol{z}_u}), (\tilde{a}, \sigma_a), (\tilde{b}, \sigma_b)\}$. Above we assumed full factorization to make inference tractable. Though simpler than the true posterior, fitted variational distributions can be excellent proxies for it. The similarity between ideal points fit with variational inference and MCMC has been demonstrated in particular (Gerrish and Blei, 2011).

As seen in Chapters 2, 3 and 5, variational inference usually proceeds by optimizing $\mathcal{L}_{\boldsymbol{\eta}} = \mathbb{E}_{q_{\boldsymbol{\eta}}}[\log p(x, \boldsymbol{z}, a, b, v, \boldsymbol{\theta})] - \mathbb{E}_{q_{\boldsymbol{\eta}}}[\log q_{\boldsymbol{\eta}}(x, \boldsymbol{z}, a, b)]$, with gradient or coordinate ascent. Optimizing this bound is challenging when the expectation is not analytical, which makes computing the exact gradient $\nabla_{\boldsymbol{\eta}}\mathcal{L}_{\boldsymbol{\eta}}$ more difficult. In this chapter we will take a different approach, by optimizing this bound with stochastic gradient ascent (Robbins and Monro, 1951; Bottou and LeCun, 2004), approximating the gradient with samples from $q_{\boldsymbol{\eta}}$:

$$\nabla_{\boldsymbol{\eta}}\mathcal{L}_{\boldsymbol{\eta}} \approx \frac{1}{M} \sum_{y_m \sim q_{\boldsymbol{\eta}}} \frac{\partial q_{\boldsymbol{\eta}}}{\partial \boldsymbol{\eta}} (\log p(y_m, v, \boldsymbol{\theta}) - \log q_{\boldsymbol{\eta}}(y_m)), \tag{6.3}$$

where $y_m = (x_m, \boldsymbol{z}_m, a_m, b_m)$ is a sample from $q_{\boldsymbol{\eta}}$. The algorithm proceeds by following this stochastic gradient with decreasing step size; we provide much more complete details of this algorithm, along with an empirical analysis of it, in Appendix A.

6.3 Understanding twelve years of U.S. congressional votes

In this section we will summarize the data on which we fit the issue-adjusted ideal point model and the methods we used to fit the model. In the subsequent sections we will fit models to this data to evaluate these models' performance on votes from this period and provide a qualitative look at U.S. lawmakers' issue preferences. We begin this section with a closer look at votes in the U.S. Congress from 1999-2010.

The United States Congress from 1999-2010

We studied U.S. Senate and House of Representative roll-call votes from 1999 to 2010. This period spanned Congresses 106 to 111, the majority of which Republican President George W. Bush held office. Bush's inauguration and the attacks of September 11th, 2001 marked the first quarter of this period, followed by the wars in Iraq and Afghanistan. Democrats gained a significant share of seats from 2007 to 2010, taking the majority from Republicans in both the House and the Senate. Democratic President Barack Obama was inaugurated in January 2009.

Not all votes in the U.S. Congress are recorded during roll-calls. Some bills are simply passed when no lawmaker objects to an anonymous vote and a voice vote is unambiguous. We ignored votes on such bills. Bills with roll-call votes, which are explicitly recorded, are more interesting, because some lawmaker wanted an explicit record of votes on the bill. Such records are useful for demonstrating lawmakers' (and lawmakers' opponents') positions on issues. Roll calls serve as an incontrovertible record for any lawmaker who wants such a record.

We downloaded both roll-call tables and bills from `www.govtrack.us`, a nonpartisan website which provides records of U.S. Congressional voting (Govtrack website, 2010). Not all bills were available in text form, but we had over one hundred for each Congress. Votes on bills without text were discarded. We provide a summary of statistics for our datasets in these Congresses in Table 6.3.

We fit both models to two-year periods in the House and (separately) to two-year periods in the Senate. Some bills received votes in both the House and Senate; in those cases, the issue-adjusted model's treatment of the bill in the House was completely independent of its treatment by the model in the Senate.

Vocabulary

To fit the labeled topic model to each bill, we represented each bill as a vector of phrase counts (the *vocabulary*). This "bag of phrases" is similar to the "bag of words" assumption commonly used in

Figure 6.3: Roll-call data sets used in the experiments. These counts include votes in both the House and Senate. The number of lawmakers within each House and Senate varies by congress because there was some turnover within each Congress. In addition, some lawmakers never voted on legislation in our experiments (recall, we used legislation for which both text was available and for which the roll-call was recorded).

Statistics for the U.S. Senate

Congress	Years	Lawmakers	Bills	Votes
106	1999-2000	81	101	7,612
107	2001-2002	78	76	5,547
108	2003-2004	101	83	7,830
109	2005-2006	102	74	7,071
110	2007-2008	103	97	9,019
111	2009-2010	110	62	5,936

Statistics for the U.S. House of Representatives

Congress	Years	Lawmakers	Bills	Votes
106	1999-2000	437	345	142,623
107	2001-2002	61	360	18,449
108	2003-2004	440	490	200,154
109	2005-2006	441	458	187,067
110	2007-2008	449	705	287,645
111	2009-2010	446	810	330,956

natural language processing. This vocabulary omitted content-free phrases such as "and", "when", and "the" (known as stop words) and awkward, non-informative phrases such as "and the". The full vocabulary consisted of 5,000 n-grams. We provide further details of vocabulary selection in Appendix B.7.5. We used these words to algorithmically define topics and assign issue weights to bills as described in Section 6.1.2.

Identification

We discussed in Section 5.2 the ways in which the ideal point model is under-specified. The issue-adjusted ideal point model has similar identification nuances. We address this by flipping ideal points (and bill polarities) if necessary to make Republicans positive and Democrats negative. As with the ideal point model, this does not affect model performance.

Traditional ideal points vs. issue-adjusted ideal points

The issue-adjusted ideal point model in Equation 6.1 is a generalization of the classic ideal point model (they are the same when $z_{uk} = 0$ for all u, k). The goal of this section is to empirically justify this increased complexity with a comparison of issue-adjusted ideal points and traditional ideal points. We first give a qualitative discussion of these differences and follow this with quantitative validation of the issue-adjusted model.

Examples: adjusting for issues

We give a side-by-side comparison of traditional ideal points x_u and issue-adjusted ideal points ($x_u + \boldsymbol{z}_u^T \boldsymbol{\theta}$) for the ten most-improved bills of Congress 111 (2009-2010) in Table 6.4. For each bill, the top row shows the ideal points of lawmakers who voted "Yea" on the bill and the bottom row shows lawmakers who voted "Nay". The top and bottom rows are a partition of votes rather than separate treatments of the same votes. On these bills, "Yea" and "Nay" votes fall to the correct sides of the split more often when lawmakers' issue-adjusted ideal points are used instead of their traditional ideal points.

A comparison of issue-adjusted ideal points x_u and traditional ideal points

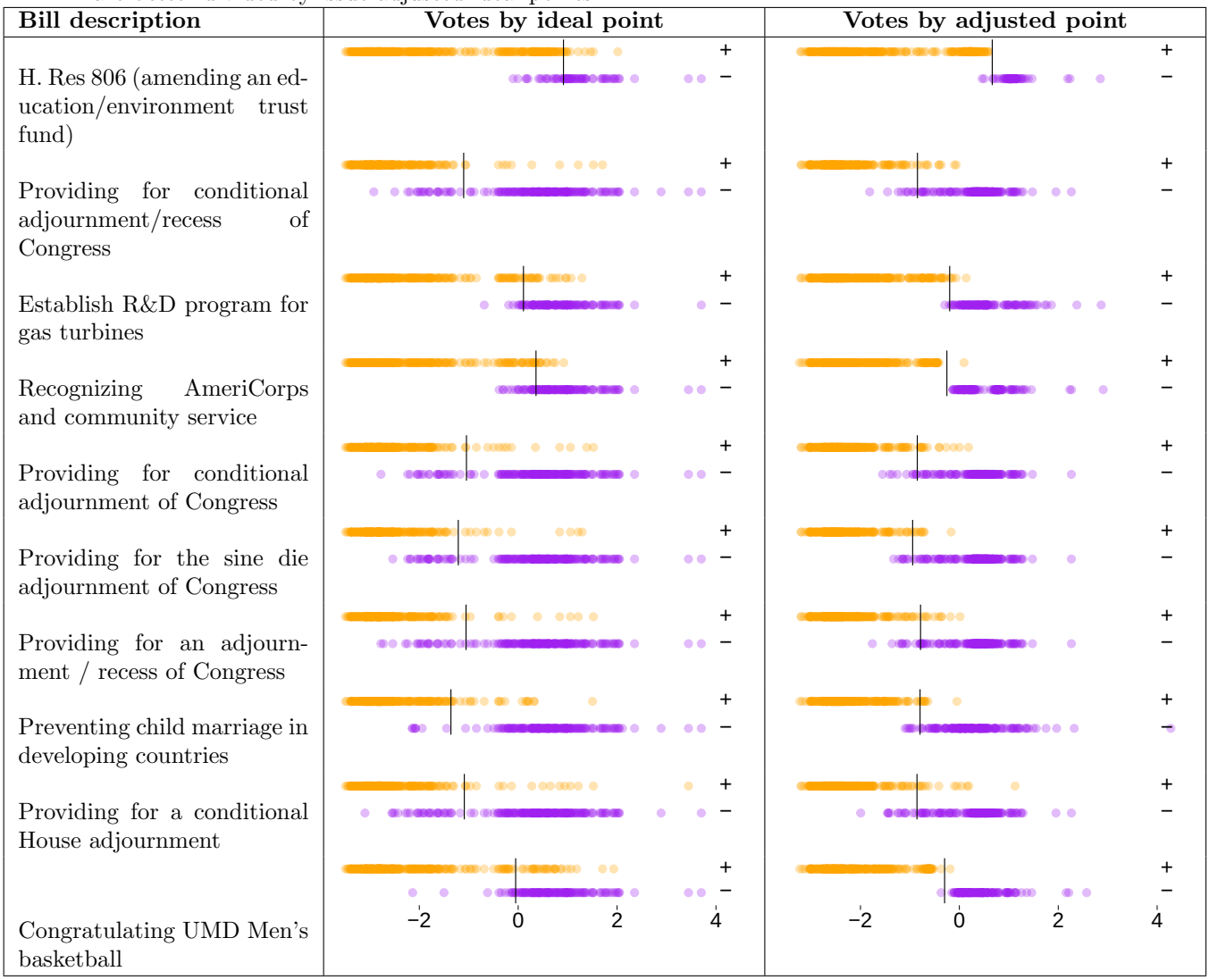
The traditional ideal point model (Equation 5.1) uses one parameter per lawmaker—x_u—to explain all of her voting behavior. In contrast, the issue-adjusted model (Equation 6.1) uses x_u along with seventy-four other parameters—one per issue—to describe each lawmaker. How does x_u under these two models differ? We fit ideal points to the 111th House (2009 to 2010) and issue-adjusted ideal points $\tilde{x}_u$ to the same period ($\lambda = 1$) and compare these ideal points in Figure 6.5

In this figure we use an alternative to a scatterplot called a *parallel plot*. In a parallel plot (which we will use several more times in this chapter), we plot the two variables we wish to compare along parallel axes and draw line segments connecting two points when they represent the same variable under different treatments. In Figure 6.5, the top axis axis represents a lawmaker's ideal point x_u under traditional IRT, and the bottom "treatment" axis represents his ideal point x_u under the issue-adjusted model. Here and later we will use the convention that the bottom row represents a special treatment. When it is helpful, we use darker line segments for those items which change the most under treatment.[3]

In the parallel plot in Figure 6.5, the traditional ideal point model's $\tilde{x}_u$ and the issue-adjusted model's un-adjusted ideal points $\tilde{x}_u$ are similar – their correlation coefficient is 0.998. The most noteworthy change is that lawmakers appear more partisan under the traditional ideal point model — enough that Democrats are completely separated from Republicans — while issue-adjusted ideal points provide a softer split. This is not surprising, because the issue-adjusted model is able to use lawmakers' adjustments $\boldsymbol{z}_u$ to more than make up for this difference. For the same reason, issue-adjusted ideal points are slightly less extreme than classic ideal points.

[3]Specifically, we fit a linear model to predict the bottom row from the top row and color line segments with opacity proportional to the squared residual of this pair. We specified opacity in *ggplot* for *R* with the alpha parameter.

Figure 6.4: Issue-adjusted ideal points can explain votes better than standard ideal points. The x-axis of each small plot shows ideal point or issue-adjusted ideal point for a lawmaker. Each bill's indifference point $-\frac{b_d}{a_d}$ is shown as a vertical line. Positive votes (orange) and negative votes (purple) are better-divided by issue-adjusted ideal points.



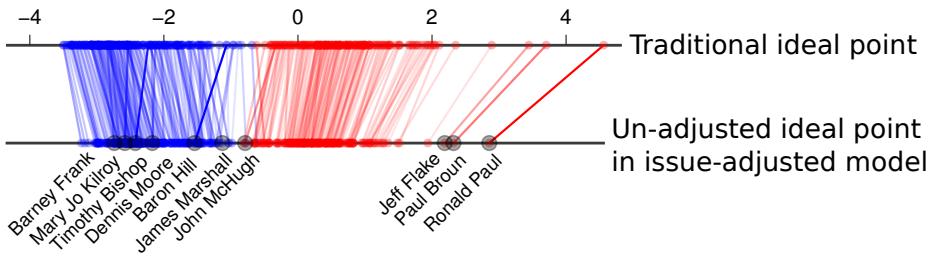


Figure 6.5: Classic issue-adjusted ideal points x_u (bottom row) separate lawmakers by party better than un-adjusted ideal points x_u from the issue-adjusted model (top row). Republicans are colored red, and Democrats are blue. These ideal points were estimated in the 111th House of Representatives. The line connecting ideal points from each model has opacity proportional to the squared residuals in a linear model fit to predict issue-adjusted ideal points from ideal points.

Evaluation of the predictive distribution

The issue-adjusted model contains the ideal point model as the special case $z_{uk} = 0, \forall u, k$. Does this greater expressivity—74 extra random variables per lawmaker—model meaningful patterns? We answer this question by comparing the issue-adjusted ideal point model with two alternatives:

1. A variational ideal point model (Equation 5.1), which treats lawmakers with the single variate x_u.

2. A permutation test. The goal of this test is to attribute any improvement over traditional ideal points to the issues assigned to bills. In this test, we randomly permute topic vectors' document labels: $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D) \mapsto (\boldsymbol{\theta}_{\pi_i(1)} \dots \boldsymbol{\theta}_{\pi_i(D)})$, for five random permutations $\pi_1, \dots, \pi_5$. This permutation test removes information contained in the matching from bills and topic mixtures. At the same time, the empirical distribution over topic mixtures $\boldsymbol{\theta}_{dk}$ stays the same, and each bill is still matched to a topic mixture with $\sum_k \boldsymbol{\theta}_{dk} = 1$. This is important because it any improvement we see over traditional ideal points is due to the bills' topics, not due to spurious factors (such as the change in dimension).

Sensitivity to λ. The main parameter in the issue-adjusted model is the regularization λ, which is shared for all issue adjustments. We report the effect of different λ by fitting the issue-adjusted model to the 109th Congress (1999-2000) of the House and Senate for a range $\lambda = 0.0001, \dots, 1000$ of regularizations. We performed 6-fold cross-validation, holding out one sixth of votes in each fold, and calculated average log-likelihood $\sum_{v_{ud} \in V_{\text{heldout}}} \log p(v_{ud} | \tilde{x}_u, \tilde{z}_u, \tilde{a}_d, \tilde{b}_d)$ for votes V_{heldout} in the heldout set. Following the algorithm described in Section 6.2, we began with $M = 21$ samples to estimate the approximate gradient (Equation 6.3) and scaled it by 1.2 each time the ELBO $\mathcal{L}$ dropped below a threshold, until it was 500. We also fixed variance $\sigma_x^2, \sigma_z^2, \sigma_a^2, \sigma_b^2 = \exp(-5)$. We summarize these results in Table 6.6.

The variational implementation generalized well for the entire range, representing votes best in the range $1 \leq \lambda \leq 10$. Log-likelihood dropped modestly for $\lambda < 1$. In the worst case, log-likelihood was -0.159 in the House (this corresponds with 96% heldout accuracy) and -0.242 in the Senate (93% heldout accuracy).

Performance across all sessions. We fit the issue-adjusted model to both the House and Senate for Congresses 106 to 110 (1999-2010) with $\lambda = 1$. For comparison we also fit an ideal point model to each of these congresses and fit an issue-adjusted model to each congress with topics' document

Figure 6.6: Average log-likelihood of heldout votes by regularization λ. Log-likelihood was averaged across folds using six-fold cross validation for Congress 109 (2005-2006). The variational distribution represented votes with higher heldout log-likelihood than traditional ideal points for $1 \leq \lambda \leq 10$. In a model fit with permuted issue labels (Perm. Issue), heldout likelihood of votes was worse than traditional ideal points for all regularizations λ.

Model	**Senate**							
λ	1e-4	1e-3	1e-2	1e-1	1	10	100	1000
Ideal	-0.188	-0.189	-0.189	-0.189	-0.189	-0.190	-0.189	-0.189
Issue (LDA)	-0.191	-0.191	-0.188	-0.186	-0.188	-0.189	-0.189	0.198
Perm. Issue	-0.242	-0.245	-0.231	-0.221	-0.204	-0.208	-0.208	-0.208
Model	**House**							
λ	1e-4	1e-3	1e-2	1e-1	1	10	100	1000
Ideal	-0.119	-0.119	-0.119	-0.119	-0.120	-0.119	-0.119	-0.119
Issue (LDA)	-0.159	-0.159	-0.158	-0.139	-0.118	-0.119	-0.119	0.119
Perm. Issue	-0.191	-0.192	-0.189	-0.161	-0.122	-0.120	-0.120	-0.120

Figure 6.7: Average log-likelihood of heldout votes across all sessions for the House and Senate. Log-likelihood was averaged across folds using six-fold cross validation for Congresses 106 to 111 (1999-2010) with regularization $\lambda = 1$. The variational distribution had higher heldout log-likelihood for all congresses in both chambers than the the ideal point model and the issue-adjusted distribution fit from permuted data.

Model	**Senate**					
Congress	106	107	108	109	110	111
Ideal	-0.209	-0.209	-0.182	-0.189	-0.206	-0.182
Issue (LDA)	**-0.208**	**-0.209**	**-0.181**	**-0.188**	**-0.205**	**-0.180**
Issue (label)	**-0.208**	**-0.209**	-0.182	-0.189	-0.206	-0.181
Perm. Issue	-0.210	-0.210	-0.183	-0.203	-0.211	-0.186
	House					
Ideal	-0.168	-0.154	-0.096	-0.120	-0.090	-0.077
Issue (LDA)	**-0.167**	**-0.151**	-0.095	-0.118	-0.089	-0.076
Issue (label)	**-0.167**	**-0.151**	**-0.094**	**-0.117**	**-0.088**	**-0.075**
Perm. Issue	-0.167	-0.155	-0.096	-0.122	-0.090	-0.077

labels permuted $(\boldsymbol{\theta}_{\pi(1)}, \ldots, \boldsymbol{\theta}_{\pi(1)})$. We summarize these results in Table 6.7. In all chambers in both Congresses, the issue-adjusted model represents heldout votes with higher log-likelihood than an ideal point model. Further, every permutation represented votes with lower log-likelihood than the issue-adjusted model. In most cases they were also lower than an ideal point model.

Human labels vs. inferred text-based labels. The issue-adjusted model assumes a fixed issue vector $\boldsymbol{\theta}_d$ for each bill. We described a method in Section 6.1.2 for inferring this issue vector based on the text of bills using labeled LDA; this method uses the original Congressional Research Service (CRS) labels. What happens if we skip this preprocessing step and just use the original CRS labels? We checked this by converting the original CRS issue labels into a K-vector of issues. For each document d having issue labels $J \subset K$, each issue $\boldsymbol{\theta}_{dk}$ was assigned a weight of $1/|J|$ if $k \in J$ and zero if $k \notin J$. We fit an issue-adjusted model using these with CRS labels and performed six-fold cross validation as described above and illustrate predictive performance in Table 6.7 in the "Issue (label)" row.

Across Congresses, the predictive benefit in using text-based issue vectors over labels provided by the CRS is negligible. However, we see at least two benefits in using text-based labels. First, they provide a defensible way to distribute weight to each issue: an issue should receive less than $1/|J|$ weight if it is mentioned only in passing in a bill. Second, this method allows us to fit issue vectors to the 107 bills which were missing CRS labels.

Changes in bills' parameters. Bills' polarity a_d and popularity b_d are similar under both the traditional ideal point model and the issue-adjusted model. We illustrate bills' parameters in these two models in Figure 6.8 and note two exceptions. First, procedural bills stand out from other bills in becoming more popular overall. In Figure 6.8, procedural bills have been separated from traditional ideal points. We attribute the difference in procedural bills' parameters to *procedural cartel theory*, which we describe further in Section 6.3.1. Second, the remaining bills have become less popular but more polarized under the issue-adjusted model. This is because the model depends more on lawmakers' positions to explain votes, because it has many more dimensions with which it can describe each lawmaker.

Sparsity of $\tilde{z}_{uk}$. The variational estimates $\tilde{z}_{uk}$ of issue adjustments were not sparse, although a high mass of these issue adjustments was concentrated around zero: twenty-nine percent of issue adjustments were within $[-0.01, 0.01]$, and eighty-seven percent of issue adjustments were within

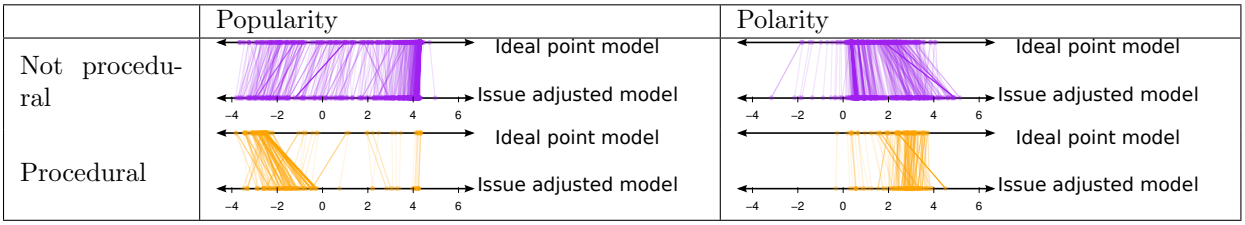


Figure 6.8: Procedural bills are more popular under the issue-adjusted voting model. Top: popularity b_d of procedural bills under the issue-adjusted voting model is greater than with traditional ideal points. Bottom: consistent with Cox and Poole (2002) and *procedural cartel theory*, the polarity of procedural bills is generally more extreme than that of non-procedural bills. However, issue adjustments lead to increased polarity (i.e., certainty) among non-procedural votes as well. The procedural issues include *Congressional reporting requirements*, *Government operations and politics*, *House of Representatives*, *House rules and procedure*, *Legislative rules and procedure*, and *Congress*.

$[-0.1, 0.1]$. We illustrate the distribution of lawmakers' offsets for selected issues in Figure 6.10 and describe them further in Section 6.3.1.

6.3.1 Issues and Lawmakers

We illustrate Representatives' issue adjustments for the issues *Finance* and the procedural issue *Congressional sessions* in Figure 6.11 for the 111th House. These adjustments illustrate the way in which the issue-adjusted voting model allows us to better understand how lawmakers feel about specific issues, but they do not tell us which issues were well-fit by the model, or whether these issue adjustments were systemic (i.e., predictable using lawmakers' ideal points) or even statistically significant.

The goal of this section is to address these concerns by providing a qualitative look at lawmakers' issue preferences such as those in Figure 6.11. We begin by identifying those issues which were best- and worst-represented by the issue-adjusted model. We then look at when lawmakers' issue adjustments can be explained by party affiliation and discuss how to control for these systemic biases to identify lawmakers who transcend party lines. We finally describe a theory explaining why certain lawmakers have such different preferences on procedural issues like *Congressional sessions* than substantive issues like as *Finance*.

Issues improved by issue adjustment

Those issues which tended to move lawmakers the most (by standard deviation of $\hat{z}_k$) also tended to give issue-adjusted ideal points an edge over traditional ideal points. We measure the performance

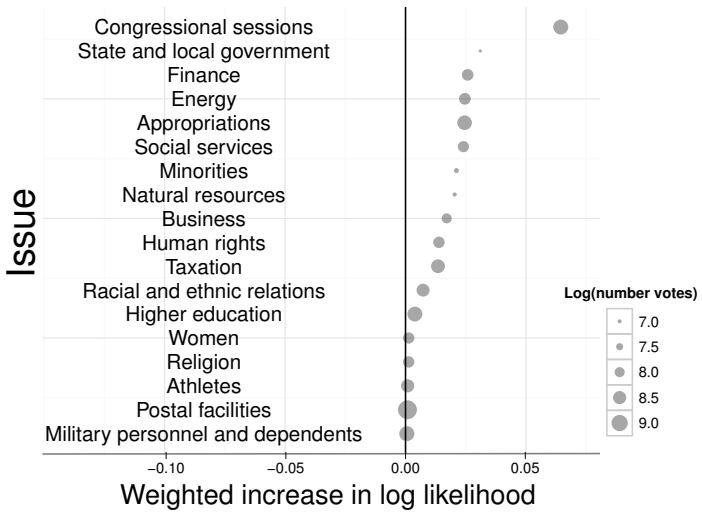


Figure 6.9: Log-likelihood increases when using adjusted ideal points most for procedural and strategic votes and less for issues frequently discussed during elections. Imp_k is shown on the x-axis, while issues are spread on the y-axis for display. The size of each issue k is proportional to the logarithm of the weighted sum $\sum_{v_{ud}} \boldsymbol{\theta}_{dk}$ of votes about the issue.

improvement for any issue by first taking the issue-adjusted log likelihood of each vote

$$J_{ud} = 1_{\{v_{ud}=\text{yea}\}} p - \log(1 + \exp(p)), \tag{6.4}$$

where $1_{\dots}$ is an indicator function and $p = (x_u + \boldsymbol{z}_u^T \boldsymbol{\theta}_d) a_d + b_d$ is the log-odds of a vote under the issue adjusted voting model. We also measure the corresponding log-likelihood I_{ud} under the ideal point model, using $p = x_u a_d + b_d$ with an ideal point model. The improvement of issue k is then the sum of the improvement in log-likelihood, weighted by issue k:

$$\text{Imp}_k = \frac{\sum_{v_{ud}} \boldsymbol{\theta}_{d_v k} (J_{ud} - I_{ud})}{\sum_{v_{ud}} \boldsymbol{\theta}_{d_v k}}. \tag{6.5}$$

A high value of Imp_k indicates that issue k is associated with an increase in log-likelihood, while a low value is associated with a decrease in log-likelihood.

We illustrate Imp_k for a selection of issues in Figure 6.9. All issues increased log-likelihood; those associated with the greatest increase tended to be related to procedural votes. For example, *Women*, *Religion*, and *Military personnel* issues are nearly unaffected by lawmakers' offsets. These improvements Imp_k were correlated with the standard deviation of residual offsets $\hat{z}_k$ ($\sigma_{\text{cor}} = 0.68$), but not with coefficients β_k ($\sigma_{\text{cor}} = 0.05$), indicating that issue offsets, and not ideal points, explain most of the improvement.

Issues associated with worse predictions. We also note several poorly-fit issues. We evaluated issues by taking the number of incorrectly-fit votes under the issue-adjusted model *minus* the number of incorrectly-fit votes under a traditional ideal point model. We call this the number of "new mis-predicted votes" for each issue. Those issues which had the most "new mis-predicted votes" also had the most "new correctly-predicted votes", which is largely because votes on these issues are simply hard to predict. For example, *Athletics* was one of the issues which saw the most most newly-mispredicted votes. *Postal Facilities* and *Military Personnel* were other examples.

Bills which expressed many issues were also less-well fit. The bill which decreased the most by log-likelihood of its votes from the ideal point model in the 111th House was the *Consolidated Land, Energy, and Aquatic Resources Act of 2010* (H.R. 3534). This bill had substantial weight in five issues, with most in *Public lands and natural resources*, *Energy*, and *Land transfers*, but its placement in many issues appears to have harmed its performance. This effect was common, and it suggests that methods which represent bills with fewer issues (such as unsupervised topics) may perform better, at the expense of interpretability.

Understanding lawmakers' voting amidst party bias

Many lawmakers' issue adjustments can be explained by party affiliation (hence, their ideal point). We illustrate the distribution across lawmakers of $\tilde{z}_{uk}$ for selected issues k in Figure 6.10. This figure shows this distribution for the four issues with the greatest variance in $\tilde{z}_{uk}$ across lawmakers and the four issues with the least variance across lawmakers. Note the systematic bias in Democrats' and Republicans' issue preferences: they become *more partisan* on certain issues, particularly procedural ones.

Controlling for ideal points. A typical Republican tends to have a Republican offset on taxation, but this surprises nobody. Instead, we are more interested in understanding when this Republican lawmaker *deviates* from behavior suggested by her ideal point. We can shed light on this systemic issue bias by explicitly controlling for it. To do this, we fit a regression for each issue k to explain away the effect of a lawmaker's ideal point $\boldsymbol{x}_u$ on her offset $\boldsymbol{z}_{uk}$:

$$\boldsymbol{z}_k = \beta_k \boldsymbol{X} + \boldsymbol{\varepsilon},$$

where $\beta_k \in \mathbb{R}$. Instead of evaluating a lawmaker's observed offsets, we use her residual $\hat{z}_{uk} = \boldsymbol{z}_{uk} - \beta_k \boldsymbol{x}_u$, which we call the corrected issue adjustment. By doing this, we can evaluate lawmakers

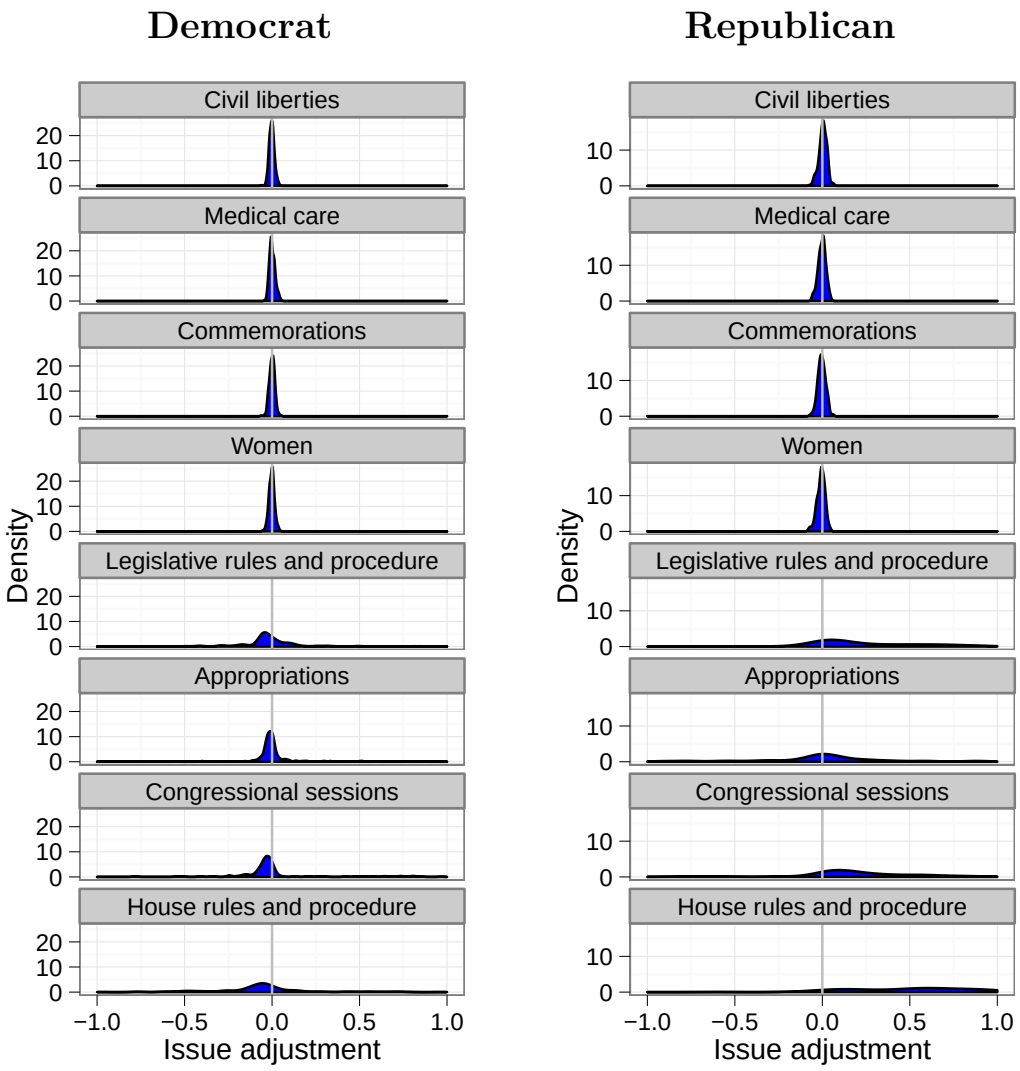


Figure 6.10: Histogram of issue adjustments for selected issues. Democrats are in the left column, and Republicans are in the right column. Both Democrats and Republicans tend to have small issue adjustments for traditional issues. Their issue adjustments differ substantially for procedural issues. A more-dispersed distribution of issue adjustments does not mean that these lawmakers tend to feel differently from one another about these issues. Instead, it means that lawmakers deviate from their ideal points more.

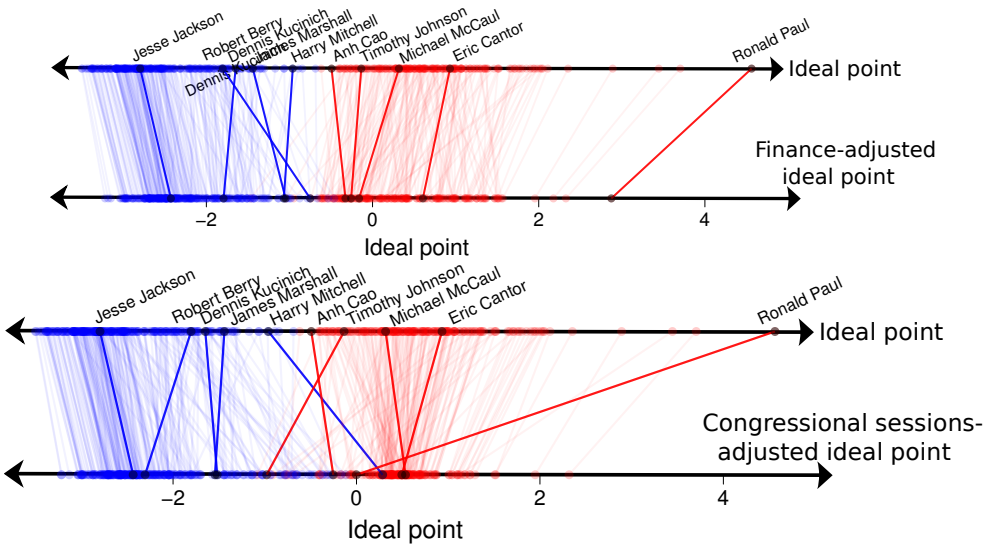


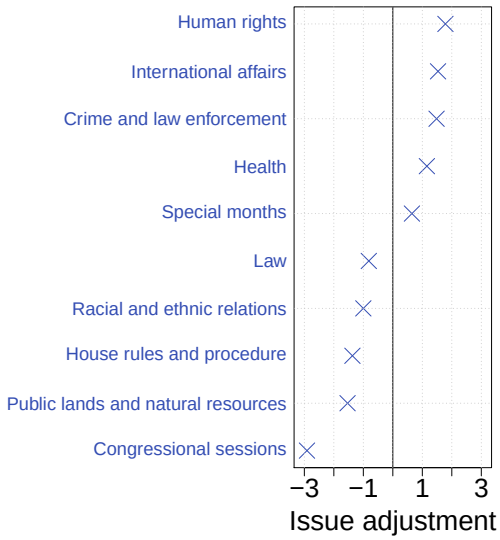
Figure 6.11: Ideal points x_u and issue-adjusted ideal points $x_u + z_{uk}$ from the 111th House for the substantive issue *Finance* and the procedural issue *Congressional Sessions*. Democrats are blue and Republicans are red. Votes about *Finance* and *Congressional Sessions* were better fit using issue-adjusted ideal points. For procedural votes such as *Congressional sessions*, lawmakers become more polarized by political party, behavior predicted by procedural cartel theory (Cox and McCubbins, 1993).

in the context of other lawmakers who share the same ideal points: a positive offset $\hat{z}_{uk}$ for a Democrat means she tends to vote more liberally about issue k than others with the same ideal point (most of whom are Democrats).[4]

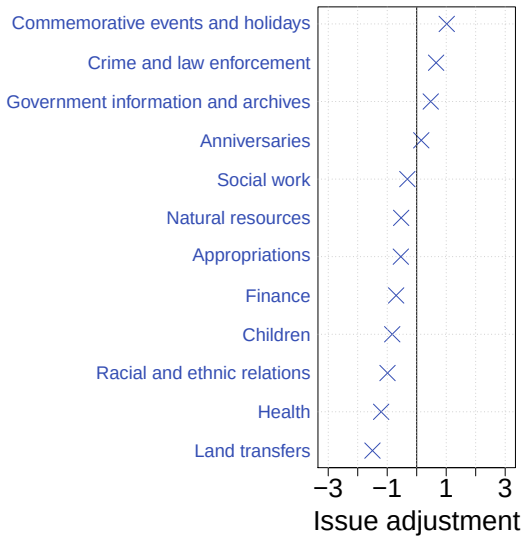
Most issues had only a moderate relationship to ideal points. *House rules and procedure* was the most-correlated with ideal points, moving the adjusted ideal point $\beta_k = 0.26$ right for every unit increase in ideal point. *Public land and natural resources* and *Taxation* followed at a distance, moving an ideal point 0.04 and 0.025 respectively with each unit increase in ideal point. *Health*, on the other hand, moved lawmakers $\beta_k = 0.04$ left for every unit increase in ideal point. The issues *Women*, *Religion*, and *Military personnel* were nearly unaffected by lawmakers' offsets.

Finding exceptional issue-adjustments. We next use these corrected issue adjustments to identify lawmakers' exceptional issue preferences. To identify adjustments which are significant, we turn again to the same nonparametric check described in the last section: permute issue vectors' document labels, i.e. $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D) \mapsto (\boldsymbol{\theta}_{\pi_i(1)} \ldots \boldsymbol{\theta}_{\pi_i(D)})$, and refit lawmakers' adjustments using both the original issue vectors and permuted issue vectors, for permutations $\pi_1, \ldots, \pi_{20}$. By mixing up the matching between issue vectors and bills, this serves to separate issue adjustments that might arise accidentally from noise in the data from issue adjustments that arise from the observed data.

[4] We also fit a model with this regression explicitly encoded. That model performed slightly worse in experiments on heldout data.



Ronald Paul (House Republican)



Donald Young (House Republican)

Figure 6.12: Significant issue adjustments for exceptional senators in Congress 111. Each illustrated issue is significant to $p < 0.05$ by a permutation test.

We then compare a corrected issue adjustment $\hat{z}_{uk}$'s absolute value with corrected issue adjustments estimated with permuted issue vectors $\boldsymbol{\theta}_{\pi_i(d)k}$.

This provides a nonparametric method for finding issue adjustments which are more extreme than expected by chance: an extreme issue adjustment has a greater absolute value than all of its permuted counterparts. We use these to discuss several unique lawmakers.

Extreme lawmakers. Using corrected issue adjustments, we identified several of the most-unique lawmakers. We focused this analysis on votes from 2009-2010, the most recent full session of Congress, using $\lambda = 1$. We fit the variational model to all votes in the House and computed lawmakers' corrected issue adjustments $\hat{z}_{uk}$, which are conditioned on their ideal points as described in Section 6.3.1. Figure 6.12 illustrates those issue preferences for two lawmakers from this Congress which significant under 20 permutation replications ($p < 0.05$).

- **Ron Paul.** We return to Ron Paul, one of the most unique House Republicans, and a lawmaker who first motivated this analysis. Paul's offsets were very extreme; he tended to vote more conservatively than expected on *health*, *human rights* and *international affairs*. He voted more liberally on social issues such as *racial and ethnic relations*, and broke with behavior expected under a procedural cartel (congressional sessions). The issue-adjusted training accuracy of Paul's votes increased from 83.8% to 87.9% with issue offsets, placing him among the two

most-improved lawmakers with this model.

The issue-adjusted improvement Imp_K (Equation 6.5) when restricted to Paul's votes indicate significant improvement in *international affairs* and *East Asia* (he tends votes against U.S. involvement in foreign countries); *congressional sessions*; *human rights*; and *special months* (he tends to vote against recognition of months as special holidays).

- **Donald Young.** One of the most exceptional lawmakers in the 111th House was Donald Young, Alaska Republican. Young stood out most in a topic used frequently in House bills about naming local landmarks. In many cases, Young voted against the majority of his party (and the House in general) on a series of largely symbolic bills and resolutions. For example, in the *commemorative events and holidays* topic, Young voted (with only two other Republicans and against the majority of the House) *not to* commend "the members of the Agri-business Development Teams of the National Guard and the National Guard Bureau for their efforts... to modernize agriculture practices and increase food production in war-torn countries."

  Young's divergent symbolic voting was also evident in a series of votes against naming various landmarks—such as post offices—in a topic about such symbolic votes. Yet Donald Young's ideal point is -0.35, which is not particularly distinctive (see Figure 5.2): using the ideal point alone, we would not recognize his unique voting behavior.

Procedural Cartels

Above we briefly noted that Democrats and Republicans become *more partisan* on procedural issues. Lawmakers' more partisan voting on procedural issues can be explained by theories about partisan strategy in the House. In this section we summarize a theory underlying this behavior and note several ways in which it is supported by issue adjustments.

The sharp contrast in voting patterns between procedural votes and substantive votes has been noted and studied over the past century (Jr., 1965; Cox and McCubbins, 1993; Cox and Poole, 2002). Cox and McCubbins (1993) provide a summary of this behavior: "parties in the House – especially the majority party – are a species of 'legislative cartel' [which usurp the power] to make rules governing the structure and process of legislation." A defining assumption made by Cox and McCubbins (2005) is that the majority party delegates an agenda-setting monopoly to senior partners in the party, who set the procedural agenda in the House. As a result, the cartel ensures that senior members hold agenda-setting seats (such as committee chairs) while rank-and-file members of the party support agenda-setting decisions.

This *procedural cartel theory* has withstood tests in which metrics of polarity were found to be *greater* on procedural votes than substantive votes (Cox and McCubbins, 1993; Cox and Poole, 2002; Cox and McCubbins, 2005). We note that issue adjustments support this theory in several ways. First, lawmakers' systematic bias for procedural issues was illustrated and discussed in Section 6.3.1 (see Figure 6.10): Democrats systematically lean left on procedural issues, while Republicans systematically lean right. Importantly, this discrepancy is more pronounced among procedural issues than substantive ones. Second, lawmakers' positions on procedural issues are more partisan than expected under the underlying un-adjusted ideal points (see Section 6.3.1 and Figure 6.11). Finally, more extreme polarity and improved prediction on procedural votes (see Section 6.3 and Figure 6.8) indicate that that issue adjustments for procedural votes are associated with *more extreme* party affiliation – also observed by Cox and Poole (2002).

Conclusions

In the last two chapters we took a closer look at the decision-making process in the U.S. by addressing two important shortcomings of ideal point models: their inability to predict votes on previously-unseen bills and their inability to represent lawmakers with nontrivial voting behavior. In this chapter we addressed the latter limitation by developing and studying the issue-adjusted ideal point model, a model designed to tease apart lawmakers' preferences from their general political positions. This is a model of roll-call data that captures how lawmakers vary, issue by issue, and it gives a new way to explore legislative data. On a large data set of legislative history, we demonstrated that it is able to represent votes better than a classic ideal point model and illustrated its use as an exploratory tool.

This work could be extended in several ways. One of the most natural ways is to incorporate lawmakers' stated positions on issues – which may differ from how they actually vote on these issues; in preliminary analyses, we have found little correlation to external sources. We might also study lawmakers' activities outside of voting to understand their issue positions. For example, lawmakers' fund-raising by industry area might (or might not) be useful in predicting their positions on different issues. Additional work includes modeling how lawmakers' positions on issues change over time, by incorporating time-series assumptions as in Martin and Quinn (2002).

The ideal point model introduced in the last chapter represents one of the simplest models of dyadic data. By framing it as a latent variable model as in Clinton *et al.* (2004), we were able to use it as a modular piece of larger latent-variable models. By combining it with other modules

for analyzing text data, including Latent Dirichlet Allocation and text regression, we were able to address two important shortcomings of ideal points. As in chapters 3 and 4, this allowed us to explicitly state nontrivial assumptions in interesting ways.

Chapter 7

Conclusions

In the past five chapters we introduced several models to address important problems in social science research. We accomplished this by framing our assumptions as latent variable models, using data sources to estimate the values of the latent random variables, and empirically validating these models. Throughout, we repeatedly made use of a small set of statistical primitives.

We introduced the reader to these primitives in Chapter 2. Having been developed over the past century, these ideas were sufficient to provide the majority of the scaffolding for our assumptions. The most important of these assumptions were models for text analysis, including topic models and text regression, which have seen huge advances in the past two decades. We also used hidden Markov models for time-series applications and latent spatial assumptions to model interaction between pairs of items.

In Chapter 3 we explored the problem of finding influential documents in text corpora which have evolved over time. This problem affects a wide range of fields, with concrete motivations in both academia and industry. We introduced a model which uses the change in language to find documents which use language that becomes more popular over time in a field. We then fit this model to four corpora: three scientific journals and a corpus of legal opinions. We consistently found a correlation between our measure of influence and unseen citation counts for these corpora, and we explored several anecdotal examples within these collections.

We then used some of the same time-series assumptions to build a recent history of the sentiment between nations in Chapter 4. To do this, we encoded assumptions about how nations interact into a dyadic model of the sentiment between two nations. We defined the sentiment between nations using hand-labeled codes from both experts and novices. Upon fitting this model, we discovered

that the sentiment between nations predicted by this model was extraordinarily correlated between a model fit with expert labels and a model fit with novice labels.

In Chapters 5 and 6, we used primitives from text analysis to improve the ideal point model, a well-known dyadic model of legislative voting. Using models of text, we first constructed models to enable us to predict lawmakers' votes on unseen documents. We demonstrated that such models allow us to accurately predict lawmakers' votes on unseen bills.

We then made ideal point models more interpretable by extending it to incorporate lawmakers' positions on different issues. We then used supervised topic models to assign interpretable labels to bills and fit lawmakers' positions. We demonstrated that, in doing this, we were able to improve the ideal point model's representation of heldout votes while providing an interpretable description of each lawmaker.

7.1 Latent variable models for understanding the social sciences

The core tools we used to build the models in these chapters were outlined in the introductory materials chapter. By framing our questions as latent-variable models, we were able to use a handful of "statistical primitives" to encode our assumptions, make predictions, and interpret hidden random variables. These primitives were tools for text data, including bag-of-words models like latent Dirichlet allocation (Blei and Lafferty, 2006) and text regression (Kogan *et al.*, 2009). The time-series primitive we used was that of a hidden Markov model. Finally, we modeled interactions between pairs of items – whether they were a pair of warring nations or lawmakers and bills – using simple distributions over pairs of variables with well-defined distributions.

Our ability to use these primitives has been made possible by several recent advances in the past few decades. These include the wider availability of documents in digital form – including the text of millions of academic articles on sites like JSTOR (`www.jstor.org`), millions of newspaper articles like the New York Times, and billions of government records such as those on independent sites like Govtrack (`www.govtrack.us`) and government sites like the National Archives (`www.archives.gov`).

Just as these corpora have become more widely available, the tools for gleaning information from them have continued to improve. These include first Judea Pearl's abstraction of graphical models, which now allow us to piece together latent variable models as easily as children build cars from Legos^{TM}(Pearl, 1985). Since graphical models have become mainstream, old primitives such

as HMMs have been ported to this paradigm, while new primitives such as LDA and text regression have been developed within this paradigm.

At the same time, the statistical tools for fitting these models has continued to improve, as tools such as variational inference improve model runtime and tools like stochastic variational optimization (which I introduce in Appendix A) decrease model development time. Finally, Moore's law now enables us to fit these models on larger and larger collections of data, as the speed and memory of researchers' workstations allows us to process millions of observations per minute.

7.2 Future work

The conclusion section of each chapter in this thesis describes future research directions for the work in that chapter. In this section I outline high-level work I see ahead for this research community around the themes discussed in this thesis.

The explosion of text data and tools for working with text suggest that fundamental research will continue around model-building, primitive development, and posterior inference. As this happens, analysis and model-building will become easier for the casual "practitioner". As we researchers better understand how to abstract away details about inference without compromising the quality of posterior estimation and without compromising data integrity, it will become our responsibility to develop tools for data practitioners.

The primitives I have recapitulated throughout this thesis are only a handful of the primitives available to practitioners. I have focused on these primitives because they are simple enough to recur frequently while still retaining enough power to provide meaningful utility to practitioners. There exist a variety of other primitives for researchers working with text, including alternative models of documents and alternative models for time-series analysis. While books such as Bishop (2006) describe many of these abstractions from a machine learning perspective, I expect many of these abstractions to receive explicit attention in future resources for practitioners in fields outside of machine learning.

Appendix A

Optimizing the variational bound stochastically

Estimating an arbitrary probability distribution $p(x)$ is a fundamental problem in statistical modeling. This problem arises in posterior inference, for example, where we seek to estimate a conditional $p(x|y)$ of latent variables x given observations y. There are two main classes of solutions—Markov chain Monte Carlo (MCMC) (Bishop, 2006) and variational methods (Jordan *et al.*, 1999). In MCMC, we define a Markov chain whose stationary distribution is the target distribution. We run the chain to try to collect independent samples from the stationary distribution, and then use them to form an approximation. In variational inference, we posit a parameterized family of distributions $q_\theta(x)$ and find the member of that family that is "closest" to the true posterior. This turns the problem of inference into one of optimization.

Deriving and implementing a variational inference algorithm can be painstaking, as evidenced in the tedious update equations in Appendix B. It involves defining the variational family, forming an objective function, taking derivatives with respect to the variational parameters, and running an optimization algorithm. In this appendix, we present an alternative algorithm for variational inference. Our algorithm circumvents many of the challenges to using variational inference by optimizing the variational objective function stochastically.

To do this, we form the derivative of the variational objective as an expectation with respect to the variational distribution. We then sample from that distribution to obtain realizations of a stochastic gradient. Our algorithm is a "black box" algorithm in that it only requires that we evaluate the joint likelihood $p(x, y)$ of the hidden and observed variables (up to a constant factor), the

variational likelihood $q(x)$ (up to a constant factor), and the derivative of $\log q(x)$. (Note that this derivative can be reused across variational inference problems.) Unlike other automated approaches to variational inference (Winn and Bishop, 2001), we have no other restrictions on the model or variational family, e.g., that the hidden variables come in conjugate pairs or that the variational distributions are in the exponential family.

There have been several recent algorithms that are similar in spirit to ours. Both Carbonetto *et al.* (2009) and Graves (2011) perform variational inference by taking samples from the variational posterior to estimate a gradient. Carbonetto et al. assume that the variational distribution comes from the exponential family (Carbonetto *et al.*, 2009). Graves (2011) approximates the first-order gradient for fitting a neural network.

Our work significantly expands on this research. We make weaker assumptions than Carbonetto *et al.* (2009) on the forms of p and $q_\theta(x)$. Our posterior $p(x|y)$ and $q(x)$ must be well-behaved—the KL-divergence between $p(x|y)$ and $q(x)$ must exist, $\log q_\theta(x)$ must be differentiable almost everywhere, and q_θ must have finite variance—but is otherwise unrestricted. Our method can be used for a wider variety of statistical models, with benefits over both MCMC and traditional variational inference.

A.1 Stochastic optimization of the variational objective

We begin this section by reviewing variational inference for approximating posterior distributions. We then derive our algorithm for optimizing the variational objective with stochastic optimization. We discuss an illustrative example and describe several extensions to the algorithm.

A.1.1 Variational inference

Variational methods are a fast, deterministic alternative to MCMC for approximate inference (Wainwright and Jordan, 2003; Jordan *et al.*, 1999). Variational methods posit a parameterized family of distributions $q_\theta(x)$ and try to find the member (i.e., the setting of variational parameters θ) that is closest in KL-divergence to the posterior $p(x|y)$,

$$\arg\min_\theta \mathrm{KL}(q_\theta||p) = \arg\min_\theta \int q_\theta(x) \log \frac{q_\theta(x)}{p(x|y)} dx. \tag{A.1}$$

We select the family to make this optimization problem tractable. A commonly chosen family is the mean-field family, where the variational distribution is fully factorized. For example, if x is a

collection of real-values that are dependent in $p(x|y)$ then the mean-field distribution might be a product $\prod_K \mathcal{N}(\mu_k, \sigma_k^2)$ of independent Gaussian distributions.

Optimizing Equation A.1 is equivalent to optimizing the "evidence lower bound" (ELBO) $\mathcal{L}_\theta$ Bishop (2006):

$$\log p(y) \geq \mathbb{E}_q \left[\log p(x, y) - \log q_\theta(x)\right] =: \mathcal{L}_\theta, \tag{A.2}$$

where the slack of the bound is equal to the KL divergence from Equation A.1. Typical variational inference algorithms optimize this bound by coordinate ascent. This requires evaluating $\mathbb{E}_q \left[\log p(x, y) - \log q_\theta(x)\right]$ and its gradient with respect to θ. If the variational distribution is not conjugate to the joint distribution $p(x, y)$, the expectation $\mathbb{E}_q \left[\log p(x, y)\right]$ will not be analytically tractable. We may then need to perform further bounds or approximations (Jaakkola and Jordan, 2000; Jordan *et al.*, 1999; Bickel and Doksum, 2007; Braun and McAuliffe, 2010).

This procedure makes variational methods challenging for two reasons. First, they require a steep learning curve and careful attention to detail to derive the coordinate updates. Second, this process must be repeated each time the model $p(x, y)$ changes form. Deriving the variational algorithm becomes a bottleneck when we seek rapid model development.

A.1.2 An algorithm for stochastic optimization of the variational objective

We now describe an alternative method for optimizing the ELBO $\mathcal{L}$. We form a noisy estimate of the gradient using Monte-Carlo integration (Graves, 2011; Wei and Tanner, 1990; Carbonetto *et al.*, 2009), and follow it with stochastic optimization (Robbins and Monro, 1951). This avoids difficult derivations; we need only evaluate $\log p(x, y)$, $q_\theta(x)$, and $\nabla \log q_\theta(x)$.

We now show that the gradient of Equation A.2 can be written as an expectation. We first exchange integration and differentiation[1], and apply the chain the rule,

$$\begin{aligned} \nabla \mathcal{L}_\theta &= \nabla \Big[\int q_\theta(x)(\log p(x, y) - \log q_\theta(x)) dx \Big] \\ &= \int \nabla \Big[q_\theta(x)(\log p(x, y) - \log q_\theta(x)) \Big] dx \\ &= \int \nabla q_\theta(x)(\log p(x, y) - \log q_\theta(x) - 1) dx. \end{aligned} \tag{A.3}$$

[1]This assumes the support of q_θ is not a function of θ, and that $\log q_\theta(x)$ and $\nabla \log q_\theta(x)$ are continuous with respect to θ.

We can write this as an expectation by using the identity $q_\theta(x)\nabla \log q_\theta(x) = \nabla q_\theta(x)$,

$$\nabla \mathcal{L}_\theta = \mathbb{E}_q \left[\nabla \log q_\theta(x) \left(\log p(x, y) - \log q_\theta(x) - 1\right)\right] \tag{A.4}$$

Now we use Monte Carlo integration to form an unbiased estimate of the gradient at $\theta = \theta_0$. We obtain M samples from the variational distribution $q_{\theta_0}(x)$, $\{x_1, \ldots, x_M\}$ and approximate,

$$\nabla \mathcal{L}_\theta \approx \frac{1}{M} \sum_{m=1}^{M} \nabla \log q_\theta(x_m)\Big|_{\theta_0} (\log p(x_m, y) - \log q_{\theta_0}(x_m) - C). \tag{A.5}$$

Note we replaced the one in Equation A.4 with a constant C. This follows because $\mathbb{E}_q \left[\nabla \log q_\theta(x)\right] = 0$. For now we will assume C equals zero, but see Section A.1.2 for how to improve performance by adjusting this constant.

Related estimates of similar gradients have been studied in recent work (Carbonetto *et al.*, 2009; Graves, 2011) and in the context of expectation maximization (Wei and Tanner, 1990).

The quality of this estimate depends on the sample size M. A small number of samples leads to a fast but crude approximation, while a large number of samples will be slower but more accurate. We will explain in Section A.1.2 how to decrease the variance of this approximation by using batches of carefully selected, non-*iid* samples and provide an experiment to explore the effect of sample size.

With regard to the model, the gradient estimate in Equation A.5 only requires we can evaluate the joint distribution. This means that variational inference can take the form of a "black box": we do not need to compute expectations of $p(x, y)$ or gradients of $\mathcal{L}_\theta$ with respect to q_θ or θ. The other requirements—that we can sample from the variational distribution $q_\theta(x)$ and evaluate its log and gradient of its log—are usually easy. (And, if not, they can be worked out once and then placed in reference for use in many variational algorithms.) We give concrete examples of the gradient of the log for several types of distributions in Section A.1.2, Section A.2.4 and in the supplementary materials.

Stochastic optimization. We can now embed this approximation in a stochastic optimization algorithm. In this algorithm, we proceed with a sequence of estimates $q_{\theta_0}, q_{\theta_1}, \ldots$ of the variational distribution. On the nth iteration, we use Monte-Carlo samples from the previous distribution $q_{\theta_{n-1}}$ to stochastically estimate the gradient to find the next distribution:

$$\theta_n \leftarrow \theta_{n-1} + \frac{\eta}{n^k} \tilde{\nabla}_\theta \mathcal{L}_\theta\Big|_{\theta_{n-1}}, \tag{A.6}$$

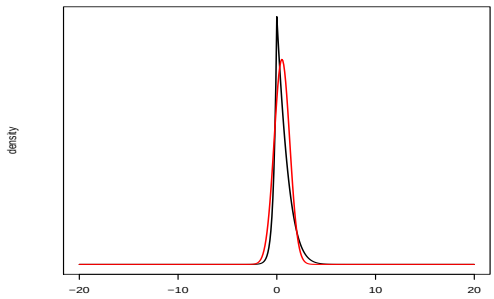


Figure A.1: A non-conjugate posterior density (dashed) and a Gaussian variational approximation (solid).

where $\eta > 0$ is a learning rate parameter and $k \in (0.5, 1.0]$ (Carbonetto *et al.*, 2009). Importantly, the expectation of the stochastic gradient, $\mathbb{E}_{\mathcal{D}}\left[\frac{\partial f(x_n,\theta)}{\partial\theta}\right]$, is the gradient of the objective (up to a constant factor). We call this method first-order Stochastic Variational Optimization (SVO) and summarize it in Algorithm 1.

Convergence. We apply these updates until a predefined convergence criterion is met (we give details in the next section). In most stochastic optimization settings, the sampling distribution $\mathcal{D}$ is stationary, and many theoretical results demonstrate when and how stochastic optimization converges to an optimum of the objective with this assumption (Bottou and LeCun, 2004; Robbins and Monro, 1951). We violate this assumption because the distribution q_θ changes in each iteration. Still, we find that this method reliably converges in practice.

Example: Gaussian variational marginal

In the next section we will describe ways to improve this gradient ascent algorithm (Equation A.6), but first we illustrate this method by estimating an "unknown" posterior with a Gaussian variational posterior.

We let $p(x, y)$ be the joint likelihood. In this example, $p(x, y)$ is a synthetic distribution: a unimodal mixture of two Gaussians, $\mathcal{N}(x|5.1, 12)$ (with component weight 0.5) and $\mathcal{N}(x|5, 3)$ (with component weight 1). We illustrate this distribution in Figure A.1. We will make only the joint likelihood $\log p(x, y)$ of this posterior available to SVO (note that y is a dummy variable in this example).

We initialize the Gaussian variational posterior to the distribution $q_{\mu_1,\sigma_1^2}(x) = \mathcal{N}(x|\mu_1, \sigma_1^2)$, with $\mu_1 = 0$ and $\sigma_1^2 = 4$. We proceed by drawing samples $x_1, \ldots, x_{15} \sim N(x|\mu_1, \sigma_1^2)$ and calculating, for each sample, the gradients

$$\frac{\partial \log q_{\mu,\sigma_1^2}(x_m)}{\partial\mu}\Big|_{\mu_{n1}} = \frac{\partial}{\partial\mu}\frac{-(x_m-\mu)^2}{2\sigma_1^2}\Big|_{\mu_1} = \frac{x_m-\mu_1}{\sigma_1^2}, \tag{A.7}$$

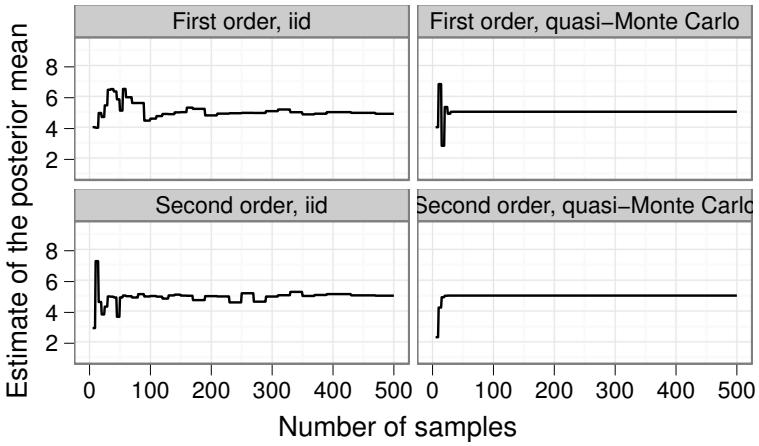


Figure A.2: A comparison of our algorithm using first-order vs. second-order updates (top vs. bottom); and estimating gradients using *iid* samples from q vs. quasi-Monte Carlo samples from q (left vs. right).

using the Gaussian density q_{μ_1,σ_1^2}. We estimate a gradient of the objective by combining these samples (Equation A.5)

$$\tilde{\nabla}_\mu \mathcal{L}_{\mu,\sigma_1^2}\Big|_{\mu_1} = \frac{1}{M}\sum_M \frac{x_m - \mu_1}{\sigma_1^2}\Big(\log p(x_m, y) - \log q_{\mu_1,\sigma_2^2}(x_m)\Big)$$

and finally update the mean μ (Equation A.6):

$$\mu_2 \leftarrow \mu_1 + \frac{\eta_\mu}{1^k}\tilde{\nabla}_\mu \mathcal{L}_{\mu,\sigma_1^2}\Big|_{\mu_1}. \tag{A.8}$$

The update for variance is similar, but we optimize $\nu = \log\sigma^2$ instead of σ^2:[2]

$$\tilde{\nabla}_\nu \mathcal{L}_{\mu_1,\nu}\Big|_{\nu_1} \approx \frac{1}{M}\sum_{m=1}^{M}\Bigg[\left(\frac{(x_m-\mu_1)^2}{\exp(\nu_1)} - 1\right) \times \left(\log p(x_m, y) - \log q_{\mu_1,\nu_1}(x_m)\right)\Bigg]. \tag{A.9}$$

The variance is then updated with Equation A.6: $\nu_2 \leftarrow \nu_1 + \frac{\eta_\nu}{1^k}\tilde{\nabla}_\nu \mathcal{L}_{\mu_1,\nu}\big|_{\nu_1}$.

Testing convergence. We repeated this process for iterations $n = 1, 2, \ldots$ until convergence. The variational estimate of the mean by the total number of samples is shown in Figure A.2 (top-left corner). To test convergence, we estimated the evidence lower bound at each iteration,

$$\mathcal{L}_n = \frac{1}{M}\sum_M \left(\log p(x_{nm}, y) - \log q_{\theta_n}(x_{nm})\right).$$

[2] σ^2 must be strictly positive, so ν is a more natural choice for stochastic updates).

We performed these updates until the exponential moving average $\Delta_{\text{est},n} \leftarrow 0.8\Delta_{\text{est},n-1} + 0.2(\mathcal{L}_n - \mathcal{L}_{n-1})$ of the ELBO dropped below one. Any time this happened, we scaled the number M of samples by a factor of 1.2. When the moving average dropped below one and $M > 500$, we stopped the algorithm.

Note that the functional form of $p(x, y)$ was never used in these updates: only the form of $q(x)$ was used. A variety of other variational posteriors are used in practice. We provide these gradients for Dirichlet and multinomial posteriors, which are conjugate to multinomial and indicator random variables respectively, in the supplementary materials. To fit these variational distributions, we need to compute $\frac{\partial \log q_\theta}{\partial \theta}$ as in Equations A.7 and A.9; the other steps are mechanical.

Improving performance

The algorithm described above lays the foundation for our approach. We now make several adjustments to complete the algorithm. These adjustments revolve around (1) improving samples used to estimate the gradient, which we can do because we have intimate knowledge of q_{θ_n}, and (2) improving step sizes with second-order updates.

Minibatch sampling

The stochastic gradients in Equation A.5 were estimated with "minibatches" of M *iid* samples from q. As Figure A.2 (top left) shows, the first-order estimate may need many samples to reach satisfactory convergence, a common observation in stochastic optimization.

One key insight for our algorithm is that we have more control over samples because we have perfect knowledge of q_θ. This contrasts with many stochastic optimization methods, in which samples may be drawn *iid* from an unknown distribution $\mathcal{D}$. By carefully selecting minibatches with non-*iid* samples, we can decrease the variance of our estimate of the ELBO $\frac{\partial \mathcal{L}_\theta}{\partial \theta}$. Quasi-Monte Carlo methods such as the Latin hypercube design have been developed for exactly this purpose (Tang, 1993; Owen, 1998; Niederreiter, 1992).

To sample values from a univariate variational distribution q, we select M equidistant points from the uniform distribution and pass these points through the inverse CDF of q.[3] To sample from multivariate distributions $\Pi_D q_d$, we select M samples from each of the D distributions, randomly permute samples from each distribution, and group them into M D-variate samples. We increase the number of samples as the algorithm converges as described in the experiments section (Wei

[3]For a truly unbiased minibatch sample from q, these points could be jittered with uniform noise within each interval.

$n \leftarrow 1$

while not converged **do**

Draw samples $x_{n1}, \ldots, x_{nM} \sim q_{\theta_{n-1}}$ using quasi-Monte Carlo sampling.

Compute $\frac{\partial \log q_{\theta_n}(x_{nm})}{\partial \theta}\big|_{\theta_{n-1}}$.

Estimate $\frac{\partial \mathcal{L}_\theta}{\partial \theta}\big|_{\theta_{n-1}}$ using Equation A.5.

Update θ, using Equation A.6:

$$\theta_n \leftarrow \theta_{n-1} - \frac{\eta}{n^k} \frac{\partial \mathcal{L}_\theta}{\partial \theta}$$

end while

$n \leftarrow 1$

while not converged **do**

Draw samples $x_{n1}, \ldots, x_{nM} \sim q_{\theta_{n-1}}$ using quasi-Monte Carlo sampling.

Compute $\frac{\partial \log q_{\theta_n}(x_{nm})}{\partial \theta}\big|_{\theta_{n-1}}$.

Compute $\frac{\partial^2 \log q_{\theta_n}(x_{nm})}{\partial \theta^2}\big|_{\theta_{n-1}}$.

Estimate $\frac{\partial \mathcal{L}_\theta}{\partial \theta}\big|_{\theta_{n-1}}$ using Equation A.5.

Estimate $\frac{\partial^2 \mathcal{L}_\theta}{\partial \theta^2}\big|_{\theta_{n-1}}$ using Equation A.10.

Update θ, using Equation A.11:

$$\theta_n \leftarrow \theta_{n-1} - \left(\frac{\partial^2 \mathcal{L}_\theta}{\partial \theta^2}\right)^{-1} \frac{\partial \mathcal{L}_\theta}{\partial \theta}.$$

$n \leftarrow n + 1$

end while

Figure A.3: First-order SVO (left) and second-order SVO (right). In each, we begin with a variational distribution $q_{\theta_0}(x)$ and joint likelihood $p(x, y)$.

and Tanner, 1990). Figure A.2 (top right) illustrates the effect of quasi-Monte Carlo sampling on convergence of SVO.

Numerical estimates with these samples can be vectorized, which can speed up computation significantly.[4] This use of samples contrasts with standard MCMC methods, which require sequential, dependent samples from a given random variable. When MCMC does not require sequential samples (e.g., updates to variables which are conditionally independent), SVO does not require sequential samples.

Second-order updates

We also note that the step size parameters η and k have a large impact on convergence to an optimal solution: they must be carefully tuned in both stochastic optimization and our algorithm. We circumvent the challenge of selecting step size with second-order updates, which are sometimes used in stochastic optimization (Robbins and Monro, 1951; Bottou and LeCun, 2004) and were used by Carbonetto *et al.* (2009) and Wei and Tanner (1990). To derive the second-order updates, we make a Taylor approximation of the variational objective $\mathcal{L}_\theta$ (Equation A.2) around the current estimate θ_0:

$$\mathcal{L}_\theta \approx \mathcal{L}_{\theta_0} + (\frac{\partial \mathcal{L}_\theta}{\partial \theta}\Big|_{\theta_0})^T \Delta_\theta + \Delta_\theta^T (\frac{\delta^2 \mathcal{L}_\theta}{\delta \theta \delta \theta^T}\Big|_{\theta_0}) \Delta_\theta,$$

where $\Delta_\theta = \theta - \theta_0$. This approximation becomes exact as θ_0 approaches the optimal solution.

[4] Vectorization uses software libraries such as BLAS and hardware such as GPUs to use samples more efficiently.

In addition to estimating the gradient $\frac{\partial \mathcal{L}_\theta}{\partial \theta}\Big|_{\theta_0}$, we also estimate the curvature $\frac{\partial^2 \mathcal{L}_\theta}{\partial \theta^2}\Big|_{\theta_0}$ empirically with samples:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_\theta}{\partial \theta^2} \approx \frac{1}{M}\sum_M \Bigg(&\left(\frac{\partial \log q_\theta(x_{nm})}{\partial \theta}\Big|_{\theta_0}\right)^2 \\ &\times (\log p(x_{nm}, y) - \log q_{\theta_0}(x_{nm}) - 1) \\ + &\left(\frac{\partial^2 \log q_\theta(x_{nm})}{\partial \theta^2}\Big|_{\theta_0}\right)(\log p(x_{nm}, y) - \log q_{\theta_0}(x_{nm}))\Bigg).\end{aligned} \tag{A.10}$$

The estimate of the optimum is then

$$\theta \leftarrow \theta_0 - \left(\frac{\partial^2 \mathcal{L}_\theta}{\partial \theta^2}\Big|_{\theta_0}\right)^{-1} \frac{\partial \mathcal{L}_\theta}{\partial \theta}\Big|_{\theta_0}. \tag{A.11}$$

This algorithm is summarized in Figure A.3 (right), and it can be used instead of the first-order algorithm (Figure A.3 (left)), just as in stochastic optimization (Bottou and LeCun, 2004). The results of applying this algorithm to the synthetic dataset described in the last section is shown in Figure A.2 (bottom two panels). Second order methods can help to avoid both high variance in a posterior estimate and poor learning rates.

We can approximate both the gradient and curvature arbitrarily well by increasing the number of samples M, provided that q and p are well-behaved. This turns the problem into approximate Newton-Raphson optimization, which means that this approach can converge more reliably than stochastic optimization by using more (and better) samples during the final updates. It also means that a tunable learning rate is no longer necessary.

Decreasing variance of the gradient

In Section A.1.2, we noted that C can be set arbitrarily without changing $\mathbb{E}_q[\tilde{\nabla}\mathcal{L}_\theta]$. However, changing C does affect the variance of $\tilde{\nabla}\mathcal{L}_\theta$, and we can set it to minimize this variance. Specifically, we set

$$C = \frac{\sum_M (\frac{\partial \log q(x_m)}{\partial})^2 (\log p(x_m, y) - \log q(x_m))}{\sum_M (\frac{\partial \log q(x_m)}{\partial})^2}. \tag{A.12}$$

This minimizes the variance of the estimate of the gradient; we find that it works well in practice. If the gradient of $\log p(x, y)$ is available to the user, he may also use this to improve the estimated gradient, as described in §A.2.

A.1.3 Multivariate distributions

Most interesting latent-variable models are multivariate, so we now describe our algorithm in the multivariate setting. With traditional variational inference, we update the posterior estimate q_θ of each hidden random variable x_i successively, given the current estimate of the remaining variables' distributions. This update is typically accomplished by gradient or coordinate ascent. In these cases, the distributions of hidden random variables are usually represented by their expectation under the variational distribution.

SVO optimizes the objective similarly: the distribution of each hidden random variable x_i is updated by holding the distributions of the remaining hidden variables fixed. To represent the distributions of variables in x_i's Markov blanket, SVO uses samples from their variational posteriors.

Related work

Stochastic optimization. SVO differs from traditional stochastic optimization in several important ways. We draw a contrast from methods which optimize a variational lower bound with *iid* training examples (Hoffman *et al.*, 2010) from an unknown distribution; we optimize the probability distribution with respect to which we are taking an expectation. Further, the samples we use to optimize this bound are drawn from this distribution. This is not the case for stochastic optimization in general. We address a specific problem using ideas from stochastic optimization, making improvements for the specific problem at hand. Many of these improvements do not apply in the general stochastic optimization setting.

Stochastic sampling with variational inference. Carbonetto *et al.* (2009) used stochastic optimization in an approach conceptually very similar to ours. They sample from the variational posterior and use importance sampling along with second-order updates to estimate a similar gradient. They further assume that the family of variational distributions includes an unbiased estimate of the true posterior, and that both the variational posterior and true posterior come from the same exponential family.

We make weaker assumptions on the forms of p and $q_\theta(x)$. Our posterior $p(x|y)$ and $q(x)$ must be well-behaved: the KL-divergence between $p(x|y)$ and $q(x)$ must exist, and it must be approximable with Monte-Carlo methods. We further require that (1) $\log q_\theta(x)$ be differentiable almost everywhere and (2) q_θ have finite variance.

Carbonetto *et al.* (2009) used importance sampling to approximate a gradient and require learn-

ing rates to be carefully set. We address both of these by using the sampling methods discussed in Section A.1.2.

Wei and Tanner (1990) use Monte-Carlo sampling to perform the E-step of EM using a finite-sum approximation of an integral. While they explicitly outline the gradient and Hessian of the expectation, they never use these values.

A.2 Empirical study

In this section we studied SVO in the synthetic toy example of Section A.1.2, Bayesian logistic regression, probit ideal point models, and the switching Kalman filter. We compared SVO to MCMC, classical variational inference, and an "oracle" sampler (when one was available).

A.2.1 Univariate examples

We return to the toy example presented in Section A.1.2 and compare our estimates of the posterior mean for second-order SVO with two sampling methods. As before, we assume that the synthetic dataset has a posterior distribution that is a logistic distribution with mean $\nu = 5$ and scale $\gamma = 2$, illustrated in Figure A.1. We make only $\log p(x, y)$ available to SVO.

Second-order SVO. We used second-order SVO (Figure A.3) with quasi-Monte Carlo samples. We assessed convergence using the method described in Section A.1.2, tracking an exponential moving average of the ELBO $\mathcal{L}$ and doubling sample size each time the moving average was low. We illustrate SVO's estimate of the mean as a function of the number of samples (and evaluations of the joint) in Figure A.5.

MCMC estimate. We compare this estimate with a Metropolis Hastings (MH) sampler, a "typical" sampler for such a problem. This sampler used a standard normal proposal distribution. We assumed a burn-in period of 100 samples. For $n \geq 101$, we plot the mean of samples $101, \ldots, n$ in Figure A.5. SVO approaches the posterior mean much more quickly than the MH estimate.

Oracle sample The above comparison with a specific MCMC sampling method depends on our choice of MCMC algorithm and parameters such as the proposal distribution. Therefore we also compare with an oracle sampler, which provides error bounds on the best possible *iid* sampling algorithm (most standard MCMC algorithms produce samples which, when thinned, are treated as *iid*). An oracle sampler is able to draw *iid* samples from p(x, y) to estimate the mean. For each sample size M, we explicitly calculate the 95% standard error confidence intervals of an estimated mean from M samples. We plot these error bars around the true mean in Figure A.5. Even with a

Probit Item Response Theory			
Metric	**SVO**	**Variational**	**Gibbs**
Heldout LL	-0.181	-0.214	-0.214
Time	27 sec.	5 sec.	122 sec.
"True" MSE	0.048	0.031	0.001

Switching Kalman filter (well log data)		
Metric	**SVO**	**Gibbs**
Heldout MSE	3.6e6	3.5e6
Time	92 sec.	104 sec.
"True" MSE	2.2e6	2.4e6

Figure A.4: Experimental results comparing SVO and MCMC estimates. We show lawmaker posteriors in the probit IRT model (left) and observation means from a change point model (right). In each table we illustrate runtime, log-likelihood (LL) or mean-squared error (MSE) on heldout observations. We also estimate MSE against the "True" posterior means, estimated using long Gibbs runs (500K and 50K samples for left and right respectively).

perfect sampler, an estimate of the mean takes much longer to converge than univariate SVO.

A.2.2 Probit regression and ideal points

We next studied SVO for approximating a complex posterior in a large high-dimensional model.

We fit a matrix of U.S. lawmakers' votes using Item Response Theory (IRT), a class of models frequently used in political science (Poole and Rosenthal, 1991; Martin and Quinn, 2002; Albert, 1992). IRT is used to position each lawmaker l in a latent space with positions $x_l \in \mathbb{R}$; these positions are often studied by political scientists to understand the lawmakers' political preferences. Lawmakers' positions interact with latent bill variables $a_d, b_d \in \mathbb{R}$; all latent variables take a standard normal prior. The probability of lawmaker l voting "Yes" on bill d is then given by $p(v_{ld} = \text{Yes}) = \text{probit}(x_l a_d + b_d)$ (Clinton *et al.*, 2004).

Experiments. Political methodologists usually implement these models with MCMC methods (Albert, 1992) (a variational implementation was introduced by Gerrish and Blei (2011), although that used the logistic response). We fit these models with MCMC, traditional variational Bayes, and SVO. We chose fully-factorized Gaussian posterior distributions.

We can use an auxiliary random variable to yield a fast Gibbs sampler and a variational algorithm (Armagan and Zaretzki, 2011) (this is not possible with a logistic response). We fit the posterior with these algorithms as well as with second-order SVO.

Results. We estimated the means of these random variables for 68 Senate bills, 95 senators, and 5,145 votes during the years 2009-2010 (this was 219 dimensions). We fit these models and compared the estimated means $\bar{x}, \bar{a}, \bar{b}$ of ideal points and bill variables; we summarize the results in Figure A.4. MCMC was the slowest, while traditional variational inference was the fastest. The

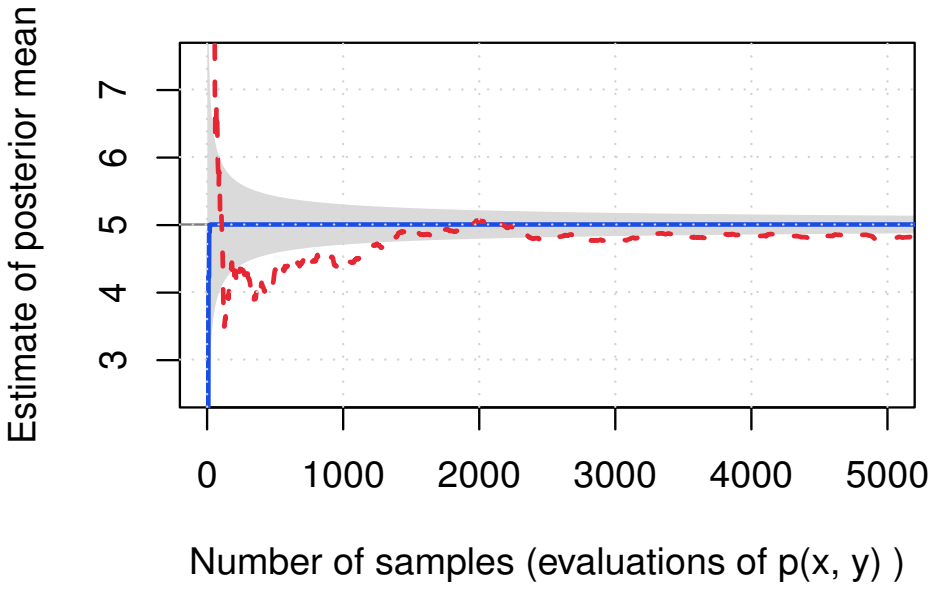


Figure A.5: figure

SVO can converge quickly to a univariate non-conjugate posterior $p(x|y)$. Solid blue: the estimated mean of a variational posterior against the number of samples (and evaluations of the joint) using second-order SVO. Dashed red: estimated mean of the posterior using a Metropolis-Hastings sampler. Shaded: 95% confidence intervals of the mean estimate from an oracle sampler.

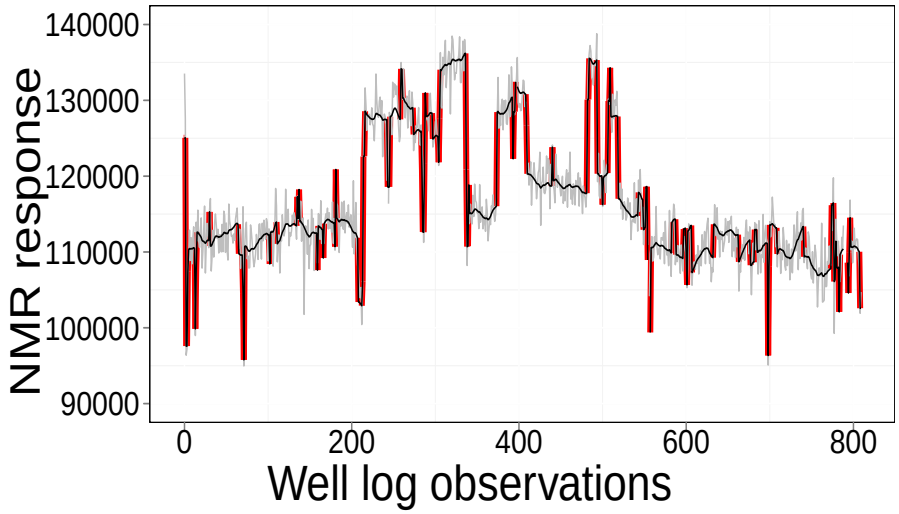


Figure A.6: Well-log data (grey) fit with a variational switching Kalman filter. The inferred means $\bar{\mu}_t$ of the filter are shown in black. Each timestamp also has an associated variational change point $\bar{c}_t$ which indicates the probability that the filter is making a large transition. Transitions at change points with mean $\bar{c}_t > \frac{1}{2}$ are marked in red.

latter is not surprising because variational Bayes uses coordinate ascent, while SVO uses slower Newton-Raphson updates on coordinates. (Further, the variational algorithm takes advantage of the derivatives of the ELBO, which we do not need to derive for SVO.)

We estimated the "true" posterior mean using 500,000 MCMC samples and found that the means estimated with both variational Bayes and MCMC samplers were closer to the "true" mean than means estimated with SVO. However, we also assessed the posteriors by their predictive distribution, using six-fold cross validation to measure log-likelihood on held-out votes. SVO formed much better predictions than the other two algorithms. (The difference between SVO and the variational Bayes estimates is explained by the auxiliary variable.)

A.2.3 Switching Kalman filter

We next illustrate this method in the task of identifying change points—positions of large changes—in a time-series dataset. To this end, we assume a series of real-valued observations $y_{1:T}$ arising from underlying means $\mu_{1:T}$. These means transition with low variance but occasionally make a large transition. These changes are characterized by random switch variates, $c_t \in \{0, 1\}$, which may indicate a large transition ($c_t = 1$) with low probability. This distribution—a switching Kalman filter (Murphy, 1998)—has the density

$$p(c_{1:T}, \mu_{1:T}, y_{1:T}) = p(\mu_1)p(\mu_T) \prod_{t=2}^{T} p(c_t)p(y_t|\mu_t)p(\mu_t|\mu_{t-1}), \tag{A.13}$$

with Gaussian observation density $p(y_t|\mu_{t-1})$ and Gaussian transition probabilities (with variance depending on c_t). $p(c_i)$ is the probability of a change point, with $p(c_i) = 0.001$. While the conditional distributions are conjugate (enabling fast Gibbs samplers), there is no analytical solution to describe the posterior distribution, so variational approximations are sometimes used (Ghahramani and Hinton, 1996; Murphy, 1998). (The derivation of the variational inference algorithm in Ghahramani and Hinton (1996) was 2.5 pages.)

Experiments. We implemented this model using both a Gibbs sampler and SVO. We used the fully-factorized posterior $\prod_T q(c_t|\bar{c}_t)q(\mu_t|\bar{\mu}_t)$ of Bernoulli and Gaussian variational distributions. We fit this model to a set of 809 measurements taken during the drilling of a well using nuclear magnetic resonance (NMR). The well log data were "used to interpret the geophysical structure of the rock surrounding the well" (Adams and MacKay, 2007) and have been studied previously using change point models (Ruanaidh and Fitzgerald, 1996; Adams and MacKay, 2007). We illustrate these data (along with SVO posterior means) in Figure A.6. We fixed the variances and π by a-priori estimation

for the well data before fitting any models.

Results. We summarize these results in Figure A.4 and in Figure A.6. We first observe that SVO takes nearly as long to converge as a 1500-sample (after 500 burn-in) Gibbs run. Why is this? In this specific case, the Gibbs sampler is very high-quality, drawing "oracle" examples with high probability once it has burned in. SVO's time performance suffers because it wastes effort updating variables that are highly dependent.

We compared these posterior estimates with posterior means from a 49,500-sample Gibbs run, which we treat as ground truth. In contrast to the IRT experiment, the SVO fit estimated better posterior means than the 1500-sample Gibbs estimate, and SVO estimated a predictive distribution which is no better than the Gibbs estimate. This is surprising but may be because the Gibbs sampler had not converged. Although the estimated means μ_t of these distributions were similar, the variational distribution discovered nearly three times as many active change points (i.e., $\bar{c}_t > 0.5$) as either Gibbs posteriors, illustrating the inherent bias in variational methods.

A.2.4 Alternative variational distributions: Laplace variational posterior

We have discussed variational Gaussian and multinomial posteriors, which are both commonly used in variational inference. But SVO opens the door to many kinds of variational distributions, as all we require is to sample from them and compute the gradient of their logs. In §A.4, we report a study fitting L_1-regularized logistic regression with a multivariate Laplace variational posterior on two standard datasets. This factorized Laplace posterior had the density $q_{\boldsymbol{\mu},\kappa}(\boldsymbol{\beta}) = \propto \exp(-\exp(\kappa)\sum_I |\beta_i - \mu_i|)$, with free variational parameters $\mu_1, \ldots, \mu_d$ and κ. This leads to a "fat-tailed" posterior which estimates posterior means which are closer to the prior mean. Importantly, these posterior distributions yielded higher held-out log-likelihood performance This and similar alternative posteriors are interesting avenues for future work.

A.3 Discussion

We described stochastic variational optimization, a generic method for variational inference that does not require taking gradients of the evidence lower bound. SVO uses stochastic optimization, taking advantage of second-order updates and quasi-Monte Carlo sampling to improve this optimization. The main benefit of SVO is that it is independent of the functional form of $p(x, y)$. With a cache of sampling methods and gradients of variational distributions, we can us SVO to rapidly build and fit many kinds of models. We demonstrated that SVO provides a good fit to the variational objective,

often forming superior predictive distributions to competing algorithms.

Appendix B

Supplementary materials

B.1 Derivation of update equations for the Document Influence Model

In this section, we describe the evidence lower bound and expand its terms to derive the variational updates for Chapter 3. The evidence is given by the following formula:

$$\mathcal{L}(q) = \log p(d_{1:T}) \tag{B.1}$$

$$\geq \int q(\beta, l, \theta, z|\tilde{\beta}, \tilde{l}, \gamma, \phi) \log \left(\frac{p(\beta, l, \theta, z) p(d|\beta, l, \theta, z)}{q(\beta, l, \theta, z|\tilde{\beta}, \tilde{l}, \gamma, \phi)} \right) d_{\beta_{1:T}} \tag{B.2}$$

$$= \mathbb{E}_q \left[\log \prod_T \prod_K p(l_{T,k}) \right] \tag{B.3}$$

$$+ \mathbb{E}_q \left[\log \prod_T \prod_{D_t} \prod_{N_{d_t}} p(z_n|\theta_{d_t}) \right] \tag{B.4}$$

$$+ \mathbb{E}_q \left[\log \prod_{t=1}^{T} \prod_K p(\beta_{t,k}|\beta_{t-f,k}) \right] \tag{B.5}$$

$$+ \mathbb{E}_q \left[\log \prod_T \prod_{D_t} \prod_{N_{d_t}} p(w_n|z_n) \right] \tag{B.6}$$

$$+ H(q) \tag{B.7}$$

$$+ \ldots, \tag{B.8}$$

where we have left out some terms (B.8) which are not relevant to this model's derivation. To maximize this lower bound, we find locally optimal values for the parameters $\phi, \tilde{\beta}, \tilde{l}$, and γ numerically

through the variational updates described below.

To derive these updates, we expand each term symbolically and find the gradient of the evidence lower bound with respect to each parameter. We then solve for the optimal value of the parameter if possible or perform gradient ascent on the parameter of interest.

We can expand B.3 as:

$$\begin{aligned}\mathbb{E}_q\left[\log\prod_T\prod_K p(l_{T,k})\right] &= \sum_T\sum_{D_t}\sum_K \mathbb{E}_q\left[-\frac{l_{d_t,k}^2}{2\sigma_d^2}-\frac{1}{2}(\log 2\pi+\log\sigma_d^2)\right] \\ &= \sum_T\sum_{D_t}\sum_K -\frac{1}{2\sigma_d^2}(\tilde{l}_{d_t,k}^2+\sigma_\ell^2)-\frac{1}{2}(\log 2\pi+\log\sigma_d^2)\end{aligned} \tag{B.9}$$

Equation B.4 can be expanded as demonstrated in the original LDA algorithm (Blei *et al.*, 2003):

$$\begin{aligned}\mathbb{E}_q\left[\log\prod_T\prod_{D_t}\prod_{N_{d_t}} p(z_n|\theta_{d_t})\right] &= \sum_T\sum_{D_t}\sum_{N_{d_t}}\mathbb{E}_q\left[\log p(\mathbf{z}_{d_t}|\theta_{d_t})\right] \\ &= \sum_N\sum_K \phi_{n,k}\left(\Psi(\gamma_i)-\Psi(\sum_{j=1}^K\gamma_j)\right)\end{aligned} \tag{B.10}$$

Finally, we expand B.5, first defining convenience functions g_w and h:

$$
\begin{aligned}
g_w(s) :=& (\mathbf{W}_{s,k} \circ \phi_{s,k})_w \tilde{l}_{s,k} \qquad \text{(B.11)}\\
h(s,q) :=& \left((\mathbf{W}_{s,k} \circ \phi_{s,k}) l_{s,k}\right)^T \Lambda_{\exp(-2\tilde{m}_{q,k}+2\tilde{V}_{q,k})} \left((\mathbf{W}_{s,k} \circ \phi_{s,k}) l_{s,k}\right)^T \\
&+ \exp(-2\tilde{m}_{q,k} + 2\tilde{V}_{q,k})^T (\mathbf{W}_{s,k} \circ \mathbf{W}_{s,k} \circ (\phi_{s,k} - \phi_{s,k} \circ \phi_{s,k}))(\tilde{l}_{s,k} \circ \tilde{l}_{s,k} + \vec{\sigma^2_{\ell}}_{D_s}) \\
&+ \exp(-2\tilde{m}_{q,k} + 2\tilde{V}_{q,k})^T (\mathbf{W}_{s,k} \circ \mathbf{W}_{s,k} \circ \phi_{s,k} \circ \phi_{s,k}) \vec{\sigma^2_{\ell}}_{D_s}.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_q \left[\log \prod_{t=1}^{T} \prod_{K} p(\beta_{t,k} | \beta_{t-1,k}) \right] \\
&\quad = \sum_{t=1}^{T} \sum_{K} \sum_{W} -\frac{1}{2\sigma^2} \mathbb{E}_q \left[\beta^2_{t,k,w} + \beta^2_{t-1,k,w} \right] \\
&\qquad + \frac{1}{\sigma^2} \mathbb{E}_q \left[\beta_{t,k,w} \beta_{t-1,k,w} \right] \\
&\qquad - \frac{1}{\sigma^2} \mathbb{E}_q \left[(\beta_{t-1,k,w} - \beta_{t,k,w}) \circ \exp(-\beta_{t-1,k,w}) (\mathbf{W}_{t-1,w} \circ [z_w]_k) l_{t-1,k} \right] \\
&\qquad + \frac{1}{\sigma^2} \mathbb{E}_q \left[\exp(-2\beta_{t-1,k,w}) \left((\mathbf{W}_{t-1,k,w} \circ [z_w]_k) l_{t-2,k} \right)^2 \right] \\
&\qquad - \frac{VT}{2} (\log \sigma^2 + \log 2\pi) \\
&\quad = - \frac{VT}{2} (\log \sigma^2 + \log 2\pi) \\
&\qquad - \frac{1}{\sigma^2} \sum_{t=1}^{T} \mathrm{Tr}(\tilde{V}_t) + \frac{1}{2\sigma^2} \left(\mathrm{Tr}(\tilde{V}_0) - \mathrm{Tr}(\tilde{V}_T) \right) \\
&\qquad - \frac{1}{2\sigma^2} (\tilde{m}_t - \tilde{m}_{t-1})^2 \\
&\qquad - \frac{1}{2\sigma^2} (\tilde{m}_{t,k} + \tilde{V}_{t-1,k} - \tilde{m}_{t-1,k})_w \exp(-\tilde{m}_{t-1,k} + \tilde{V}_{t-1,k}/2)_w \sum_{i=0}^{t} r(i) g_w(t-i) \qquad \text{(B.12)} \\
&\qquad - \frac{1}{2\sigma^2} \sum_{i=0}^{t} r(i) h(t-i, t-1)
\end{aligned}
$$

Above, $\circ$ refers to the Hadamard element-wise product and $\Lambda_{\vec{x}}$ refers to a diagonal matrix having the elements of $\vec{x}$ on its diagonal. At the line indicated by Equation B.12, we have also used the fact that $\mathbb{E}_q \left[\beta_t \exp(-\beta_t) \right] = (\tilde{m} - \tilde{V}) \exp(-\tilde{m} + \tilde{V}/2)$. Finally, we use the notation $r(s)$ to represent the envelope of influence over time. $r(s)$ satisfies $r(s) > 0$ for $s = 1, \ldots, T$ and $\sum_{s=1}^{T} r(s) = 1$.

B.1.1 Update equations

We update θ as in the DTM. The updates for $\tilde{\beta}$ and ϕ are different in the Document Influence Model, and the document weights $\tilde{l}$ must also be updated. As shown in Equation 3.9, the document weights are updated with a regression. We determine this regression by collecting terms with $\tilde{l}$, taking the

derivative, and setting equal to zero.

To find the updates for ϕ, we gather all terms from the evidence lower bound containing ϕ and form the Lagrangian to enforce the constraint $\sum_{j=1}^{K} \phi_{n,j} = 1$:

$$g_{w,k}(s) := (W_{s,k} \circ \phi_{s,k})\tilde{l}_{s,k}$$

$$\begin{aligned}
L[\phi] = \sum_{N}\sum_{K}\Bigg(&\phi_{n,k}\left(-\log\phi_{n,k} + (\Psi(\gamma_k) - \Psi(\sum_{j=1}^{K}\gamma_j)) + \tilde{m}_{n,k}\right) \\
&+ \lambda_n(\sum_{j=1}^{K}\phi_{n,j} - 1) \\
&+ \frac{1}{\sigma^2}\sum_{i=t}^{T-1} r(i-t)\exp(-\tilde{m}_i + \tilde{V}_i/2)(\tilde{m}_{i+1} - \tilde{m}_i + \tilde{V}_i)\phi_{n_d,k}w_{i,n}l_{d,k} \\
&- \frac{1}{\sigma^2}\sum_{i=t}^{T-1} r(i-t)\exp(-2\tilde{m}_i + 2\tilde{V}_i)\phi_{n,k}w_{i,n}l_{d,n}\sum_{j=0\ldots t, j\neq i}\Big((W_j \circ \phi_{j,k})\tilde{l}_{j,k,d_n}r(i-j)\Big) \\
&- \frac{1}{\sigma^2}\sum_{i=t}^{T-1} r(i-t)^2\exp(-2\tilde{m}_i + 2\tilde{V}_i)\phi_{n,k}w_{i,n}(\phi_{\setminus d_n,k} \circ W^2_{\setminus d_n})(\tilde{l}^2_{\setminus d_n}) \\
&- \frac{1}{\sigma^2}\sum_{i=t}^{T-1} r(i-t)^2\exp(-2\tilde{m}_i + 2\tilde{V}_i)\phi_{n,k}w^2_{i,n}(\tilde{l}^2_{d_n} + \sigma^2_l) \\
&- \frac{1}{\sigma^2}\sum_{i=t}^{T-1} r(i-t)^2\exp(-2\tilde{m}_i + 2\tilde{V}_i)\phi_{n,k}w^2_{i,n}(\sigma^2_{\ell\setminus d_n}) \\
&- \frac{1}{\sigma^2}\sum_{i=t}^{T-1} r(i-t)^2\exp(-2\tilde{m}_i + 2\tilde{V}_i)(1-\phi_{n,k})w^2_{i,n}(\tilde{l}^2_{d_n} + \sigma^2_{\ell\, d_n})
\end{aligned}$$

Next, take the derivative with respect to $\phi_{n,i}$:

$$
\begin{aligned}
\frac{\partial L}{\partial \phi_{n,k}} =& \sum_N \sum_K \Bigg(- \log \phi_{n,k} + \Psi(\gamma_k) - \Psi(\sum_{j=1}^{K} \gamma_j) + \tilde{m}_{n,k} + \lambda_n \\
&+ \frac{1}{\sigma^2} w_n l_{d_n} \sum_{i=t}^{T-1} r(i-t) \exp(-\tilde{m}_i + \tilde{V}_i/2)(\tilde{m}_{i+1} - \tilde{m}_i + \tilde{V}_i) \\
&- \frac{1}{\sigma^2} w_n l_{d_n} \sum_{i=t}^{T-1} r(i-t) \exp(-2\tilde{m}_i + 2\tilde{V}_i) \sum_{j=0\ldots i} \Big((W_j \circ \phi_{j,k})_{\setminus D_n} \tilde{l}_{j,k \setminus D_n} r(i-j) \Big) \\
&- \frac{1}{\sigma^2} w_n \sum_{i=t}^{T-1} r(i-t)^2 \exp(-2\tilde{m}_i + 2\tilde{V}_i)(\phi_{D_n,\setminus w,k} \circ W_{D_n,\setminus w})(\tilde{l}^2_{D_n}) \\
&- \frac{1}{\sigma^2} \sum_{i=t}^{T-1} r(i-t)^2 \exp(-2\tilde{m}_i + 2\tilde{V}_i) w_n^2 (\tilde{l}^2_{d_n} + \sigma_l^2) \\
=& \sum_N \sum_K \Bigg(- \log \phi_{n,k} + \Psi(\gamma_k) - \Psi(\sum_{j=1}^{K} \gamma_j) + \tilde{m}_{n,k} + \lambda_n \\
&+ \frac{1}{\sigma^2} \sum_{i=t}^{T-1} r(i-t) \exp(-\tilde{m}_i + \tilde{V}_i/2)(\tilde{m}_{i+1} - \tilde{m}_i + \tilde{V}_i) w_{i,k} l_{d,n} \\
&- \frac{1}{\sigma^2} w_{t,n} l_{d,n} \sum_{i=t}^{T-1} r(i-t) \exp(-2\tilde{m}_i + 2\tilde{V}_i) \sum_{j=0\ldots i, j \neq t} \Big((W_j \circ \phi_{j,k}) \tilde{l}_{j,k,d_n} r(i-j) \Big) \\
&- \frac{1}{\sigma^2} (\phi^{\text{last}}_{D_n,k} \circ W_{D_n})(\tilde{l}^2_{D_n}) \sum_{i=t}^{T-1} r(i-t)^2 \exp(-2\tilde{m}_i + 2\tilde{V}_i) \\
&- \frac{1}{\sigma^2} w_n^2 (\tilde{l}^2_{d_n} + \sigma_l^2) \sum_{i=t}^{T-1} r(i-t)^2 \exp(-2\tilde{m}_i + 2\tilde{V}_i) \\
&+ \frac{1}{\sigma^2} \phi^{\text{last}}_{n,k} w_n^2 (\tilde{l}^2_{d_n}) \sum_{i=t}^{T-1} r(i-t)^2 \exp(-2\tilde{m}_i + 2\tilde{V}_i)
\end{aligned}
\tag{B.13}
$$

$$
\begin{aligned}
=&\sum_N \sum_K \Bigg(-\log \phi_{n,k} + \Psi(\gamma_k) - \Psi(\sum_{j=1}^{K} \gamma_j) + \tilde{m}_{n,k} + \lambda_n \\
&+ \frac{1}{\sigma^2} w_n l_{D_n} \sum_{i=t}^{T-1} r(i-t) \exp(-\tilde{m}_i + \tilde{V}_i/2)(\tilde{m}_{i+1} - \tilde{m}_i + \tilde{V}_i) \\
&- \frac{1}{\sigma^2} \sum_{i=t}^{T-1} r(i-t) \exp(-2\tilde{m}_i + 2\tilde{V}_i) w_n l_{d,n} \sum_{j=0\ldots i} \Big((W_j \circ \phi_{j,k}^{\text{last}}) \tilde{l}_{j,k,d_n} r(i-t) \Big) \\
&+ \frac{1}{\sigma^2} (w_n^2 (\tilde{l}_{D_n}^2 \phi_{n,k}^{\text{last}} - (\tilde{l}_{D_n}^2 + \sigma_{\ell\, D_n}^2)) \sum_{i=t}^{T-1} r(i-t) \exp(-2\tilde{m}_i + 2\tilde{V}_i),
\end{aligned} \tag{B.14}
$$

.

$$
\begin{aligned}
=&\sum_N \sum_K \Bigg(-\log \phi_{n,k} + \Psi(\gamma_k) - \Psi(\sum_{j=1}^{K} \gamma_j) + \tilde{m}_{n,k} + \lambda_n \\
&+ \frac{1}{\sigma^2} \mathbf{w}_n l_{D_n} \sum_{i=t}^{T-1} r(i-t) \exp(-\tilde{m}_i + \tilde{V}_i/2)(\tilde{m}_{i+1} - \tilde{m}_i + \tilde{V}_i) \\
&- \frac{1}{\sigma^2} w_n l_{D_n} \sum_{i=t}^{T-1} r(i-t) \exp(-2\tilde{m}_i + 2\tilde{V}_i) \sum_{j=0\ldots i} \Big((W_j \circ \phi_{j,k}^{\text{last}}) \tilde{l}_{j,k,d_n} r(i-j) \Big) \\
&+ \frac{1}{\sigma^2} w_{t,n}^2 (\phi_{n,k}^{\text{last}} \tilde{l}_{d_n}^2 - \tilde{l}_{d_n}^2 - \sigma_l^2) \sum_{i=t}^{T-1} r(i-t)^2 \exp(-2\tilde{m}_i + 2\tilde{V}_i),
\end{aligned} \tag{B.15}
$$

where we have introduced ϕ^{last}, which is the last known value of ϕ. Therefore the update equation can be found by solving for ϕ:

$$
\begin{aligned}
\log(\phi) \leftarrow &\Psi(\gamma_k) - \Psi(\sum_{j=1}^{K} \gamma_j) + \tilde{m}_{n,k} + \lambda_n \\
&+ \frac{1}{\sigma^2} \mathbf{w}_{t,k} l_{d_n} \sum_{i=t}^{T-1} r(i-t) \exp(-\tilde{m}_i + \tilde{V}_i/2)(\tilde{m}_{i+1} - \tilde{m}_i + \tilde{V}_i) \\
&- \frac{1}{\sigma^2} w_{t,n} l_{d_n} \sum_{i=t}^{T-1} r(i-t) \exp(-2\tilde{m}_i + 2\tilde{V}_i) \sum_{j=0\ldots i} \Big((W_j \circ \phi_{j,k}^{\text{last}}) \tilde{l}_{j,k,d_n} r(i-j) \Big) \\
&+ \frac{1}{\sigma^2} w_{t,n}^2 (\phi_{n,k}^{\text{last}} \tilde{l}_{d_n}^2 - \tilde{l}_{d_n}^2 - \sigma_l^2) \sum_{i=t}^{T-1} r(i-t)^2 \exp(-2\tilde{m}_i + 2\tilde{V}_i),
\end{aligned} \tag{B.16}
$$

The update for $\tilde{\beta}$ can be found by collecting terms containing $\tilde{\beta}$ from Equation B.1. We then

maximize with respect to $\tilde{\beta}$, again using two new helper functions g and h:

$$
\begin{aligned}
g(s) :=& \mathbb{E}_q\left[\exp(-\beta_{s,k,w})(\mathbf{W}_{s,k,w} \circ \mathbf{z}_{s,k,w})\ell_{s,k}\right] \\
&= \exp(-\tilde{m}_{s,k,w} + \tilde{V}_{s,k,w}/2)(\mathbf{W}_{s,k,w} \circ \phi_{s,k,w})\tilde{l}_{s,k} \\
h(s) :=& \left(\exp(-2\tilde{m}_{s,k} + 2\tilde{V}_{s,k}) + \exp(-2\tilde{m}_{s,k} + \tilde{V}_{s,k})\right) \\
&\times \Big(\left((\mathbf{W}_{s,k,w} \circ \phi_{s,k,w})l_{s,k}\right)^2 \\
&+ (\mathbf{W}_{s,k,w} \circ \mathbf{W}_{s,k,w} \circ (\phi_{s,k,w} - \phi_{s,k,w} \circ \phi_{s,k,w}))(\tilde{l}_{s,k} \circ \tilde{l}_{s,k} + \vec{\sigma^2_{\ell}}_{D_s}) \\
&+ (\mathbf{W}_{s,k,w} \circ \mathbf{W}_{s,k,w} \circ \phi_{s,k,w} \circ \phi_{s,k,w})\vec{\sigma^2_{\ell}}\Big) \\
\frac{\partial \mathcal{L}}{\partial \tilde{\beta}_{sw}} =& -\frac{1}{\sigma^2}\sum_{t=1}^{T}\left(\tilde{m}_{tw} - \tilde{m}_{t-1,w} - \sum_{i=0}^{t-1} r(i)g(t-i-1)\right) \\
&\times \left(\frac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}} - \frac{\partial \tilde{m}_{t-1,w}}{\partial \tilde{\beta}_{sw}} + \sum_{i=0}^{t-1} r(i)g(t-i-1)\frac{\partial \tilde{m}_{t-i-1,w}}{\partial \tilde{\beta}_{sw}}\right) \\
&+ \sum_{T}\left(n_{tw} - n_t\zeta^{-1}\exp(\hat{m}_{\beta_{tw}} + \frac{\tilde{V}_{tw}}{2})\right)\frac{\partial \tilde{m}_t}{\partial \tilde{\beta}_{sw}} \\
&+ \frac{1}{\sigma^2}\sum_{t=1}^{T}\sum_{i=0}^{t-1}\frac{\partial \tilde{m}_{t-i-1}}{\partial \tilde{\beta}_{sw}} r(i)^2\left(h(t-i-1) - g(t-i-1)^2\right) \\
&+ \frac{1}{\sigma^2}\sum_{t=0}^{T-1}\frac{\partial \tilde{m}_t}{\partial \tilde{\beta}_{sw}} r(0)g(t)\tilde{V}_{t,k}.
\end{aligned}
$$

B.1.2 Topic trajectories.

The variational update for $\tilde{\beta}$ is similar to that in Blei and Lafferty (2006). For each topic, we update the variational Kalman "observations" by applying gradient ascent:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \tilde{\beta}_{sw}} =& -\frac{1}{\sigma^2}\sum_{t=1}^{T}(\tilde{m}_{tw} - \tilde{m}_{t-1,w} - G_{t-1,w})\left(\frac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}} - \frac{\partial \tilde{m}_{t-1,w}}{\partial \tilde{\beta}_{sw}} + G_{t-1,w}\frac{\partial \tilde{m}_{t-1,w}}{\partial \tilde{\beta}_{sw}}\right) \\
&+ \sum_{T}\left(N_{w,t} - N_t\zeta_t^{-1}\exp(\hat{m}_{\beta_{tw}} + \frac{\tilde{V}_{tw}}{2})\right)\frac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}} \\
&+ \frac{1}{\sigma^2}\sum_{t=1}^{T}\frac{\partial \tilde{m}_{t-1,w}}{\partial \tilde{\beta}_{sw}}\left(H_{t-1,w} - G_{t-1,w}^2\right) + \frac{1}{\sigma^2}\sum_{t=0}^{T-1}\frac{\partial \tilde{m}_{tw}}{\partial \tilde{\beta}_{sw}} G_{tw}\tilde{V}_{tw},
\end{aligned}
$$

where

$$
\begin{aligned}
G_{sn} &= \mathbb{E}_q\left[\exp(-\beta_{s,k,n})(\mathbf{W}_{s,k,n} \circ z_{s,k,n})\ell_{s,k}\right] \\
H_{sn} &= \mathbb{E}_q\big[\exp(-2\beta_{s,k,n})\left((\mathbf{W}_{s,k,n} \circ z_{s,k,n})\ell_{s,k}\right)^2\big].
\end{aligned}
$$

Note also that we have added the additional variational parameter ζ_t and the term $\frac{\partial \tilde{m}_{tn}}{\partial \tilde{\beta}_{sn}}$, which are both described in Blei and Lafferty (2006). The former can be updated once per iteration with $\zeta_t \leftarrow \sum_w \exp(\tilde{m}_{t,n} + \tilde{V}_{t,n}/2)$. The latter can be derived from the variational Kalman filter updates (see Appendix B.1 and Blei and Lafferty (2006)).

B.2 A parallel implementation of the model

The algorithm described in Section 6.2 takes approximately 11 hours on a modern desktop computer[1], for about 30,000 documents. For a larger dataset—such as all scientific articles in *Nature*, *Science*, and *PNAS* combined—this naïve implementation takes considerably longer to complete, and it requires too much memory to fit on a traditional desktop computer.[2]

In this section, we describe a parallel implementation for this model. As with the standard algorithm, the parallel algorithm optimizes the evidence lower bound by local coordinate ascent. Here, however, many of these steps are made in parallel. While most of this algorithm involves simply scheduling these updates across many computers, we also describe below how to handle an update that cannot be distributed without modification.

In this section, we will refer to a single computer as a processor. We will differentiate between the roles a processor may take by referring to a "master", which coordinates the entire algorithm, and the "workers", which perform lower-level computing. The master launches workers, checks when they are complete, and monitors model convergence. Each worker performs updates for a partition of the entire collection of random variables.

Algorithm overview

With both the parallel implementation and the standard implementation, we initialize the model with LDA topics. We therefore first fit LDA in parallel. The parallel implementation of LDA distributes the work of the E step among the many workers during each iteration. The LDA M step for each iteration—which simply aggregates sufficient statistics—is then run on the master.

Following this initialization with LDA topics, the DIM model is fit. This is driven by a single master program which alternates between two steps: a *topic* M-step and a *document* E-step.

[1] This was a 2.2GHz, 1MB cache, Dual core AMD Opteron 275 processor

[2] Circa 2009.

The parallel topics M-step

In the topic step of the original algorithm, our goal is to re-fit topics $\tilde{\beta}_k|\phi_k, \tilde{\ell}_k, \mathbf{w}$ by adjusting their variational observations. Topic chains $\tilde{\beta}_k$ are conditionally independent given the documents' variational parameters, so the parallel algorithm performs the same operations as the original algorithm, but in parallel. We simply split the work among W distinct workers.

The parallel documents E-step

In the documents E-step of the algorithm, our goal is to re-fit each document's parameters $\gamma, \phi|\tilde{\ell}, \tilde{\beta}, \mathbf{w}$ and $\tilde{\ell}|\phi, \tilde{\beta}, \mathbf{w}$ using the topics estimated in the M-step. In this step the master partitions the entire collection of documents into time-contiguous chunks and assigns each chunk to a worker. The algorithm can update γ and ϕ using the same operations as the original algorithm because the Markov blanket of γ contains variables from a single timestamp and because each ϕ is conditionally independent given the remaining variational parameters. Therefore we fit these using alternating updates of ϕ_d and γ_d as in LDA.

Parallel update dampening

The update for $\tilde{\ell}|\phi, \tilde{\beta}, \mathbf{w}$ is a bit trickier because the influence of documents at time t is not conditionally independent of the influence of documents at a different time $s \neq t$. This means that we cannot simply estimate the optimal influence of documents independently in W different workers because it is not guaranteed to improve the variational objective. In a worst-case scenario, updating $\tilde{\ell}_t$ in parallel for all times $t = 1, \ldots, T$ might result in over-estimating $\tilde{\ell}_t$, over-explaining influence.

We address this with **parallel update dampening**. In parallel update dampening, we use the fact that documents have been partitioned into W sets, and workers $\omega = 1, \ldots, W$ each manage one of those sets. In parallel update dampening:

1. Each worker calculates the optimal variational influence $\tilde{\ell}_{k,\omega}$ for its managed documents, given all other (unmanaged) documents. Each worker then has a list of all influence scores, this list comprising its unmanaged documents' scores (which take the old values) and its managed documents' scores (which are assigned new values).

2. After each worker has run and saved its scores, a master then calculates the average of scores in these lists.

Importantly, this process maintains the requirement that the variational lower bound never decrease. In the first step, each of the "local" estimates has not decreased the variational bound. The variational lower bound is concave in $\tilde{\ell}_k$, so the average of these estimates does not decrease the variational bound. Therefore, both the first and second steps guarantee that the variational objective never decrease.

Instead of taking the global average in a second stage, this can be implemented in each worker by taking the estimate of the optimal solution $\tilde{\ell}_{d,k}^{\text{worker}}$, and dampening it with the current estimate, $\tilde{\ell}_{d,k}^{\text{old}}$:

$$\tilde{\ell}_{d,k}^{\text{new}} \leftarrow \frac{W-1}{W}\tilde{\ell}_{d,k}^{\text{old}} + \frac{1}{W}\tilde{\ell}_{d,k}^{\text{worker}},$$

or, equivalently,

$$\tilde{\ell}_{d,k}^{\text{new}} \leftarrow \tilde{\ell}_{d,k}^{\text{old}} + \frac{1}{W}(\tilde{\ell}_{d,k}^{\text{worker}} - \tilde{\ell}_{d,k}^{\text{old}}).$$

We stress again that we can only guarantee that parallel update dampening increases the variational objective because the objective is concave in the parameter $\tilde{\ell}$.

B.3 Notes on the unsupervised sentiment model

B.4 Additional notes on unsupervised sentiment analysis

In this section we describe a model for inferring relationships between countries in an unsupervised fashion. This model is based on the model in the last section, but it requires no explicit labels of the relationship between pairs of countries. Instead it infers a qualitative relationship between countries – a relationship which we can attempt to interpret post-hoc. The significance of this approach is that it infers a relationship between countries based more on the discussion of these countries than explicit labels. Particularly, if there is a relationship which has been overlooked by historians, then we might be able to learn it.

In the remainder of this section we will outline a probabilistic model for inferring sentiment between pairs of countries. We will outline the key assumptions of this model – first, a language model inspired by the *Networks Uncovered by Bayesian Inference* model (Chang *et al.*, 2009); and second, a spatial model of dyadic relationships. We will then describe inference for this model, and finally provide an empirical analysis of this model.

This section necessarily represents a very cursory look at unsupervised sentiment analysis. Because there are many parts to the model, we focus on the `intercept/distance` link function defined

in Table 4.1.2. As our goal is to qualitatively observe the inferred sentiment topic, we will focus on that and skip a rigorous analysis of this model's performance.

B.4.1 A model of unsupervised foreign relations

A key variable in this model is that each document has a sentiment parameter κ_d. This become important when we link this sentiment model to text. Intuitively, if two countries are far apart in the latent space at time t, we expect that κ is more likely to be 1 when they interact. Otherwise, κ is more likely to be 0. As we develop the language model, we will use this random variable to decide which topic is used to describe the pair of countries.

Binary Relational language model

We incorporate text using a mixed-membership language model similar to LDA. Recall that in LDA, each word comes from a specific topic. In our model, which we dub the *binary relational language model*, we assume that the words describing a pair of countries come from topics about those countries.

A mixture of four topics. To be concrete, consider a document discussing Iran and the United States. We assume that each word in this document will serve one of four roles:

1. It discusses the U.S. only,
2. It discusses Iran only,
3. It discusses the relationship between the U.S. and Iran.
4. It is a "filler" word, providing little contribution to the discussion.

The first two roles for a word are self-explanatory. The relationship in (3) above could be any type of relationship – the goal of this section is of course to discover the relationships in a collection of documents about these countries. The "filler" words in (4) above are those words found in any document – stopwords, for example – that are unrelated to either country or the relationship between them.

We therefore keep (N_c+2+1) topics—topics $\beta_{C,1}, \ldots, \beta_{C,N_c}$ for each of the N_c countries, exactly two sentiment topics $\beta_{S,0}, \beta_{S,1}$, and a single, global background topic β_{B0} (Chemudugunta *et al.*, 2006). We assume, as in LDA, that a document about the United States and Iran is a mixture of topics; in contrast to LDA, however, we constrain this document's topics to be exactly the four

topics enumerated above: $\beta_{C,\text{Iran}}$, $\beta_{C,\text{United States}}$, $\beta_{B,0}$, and either $\beta_{S,0}$ or $\beta_{S,1}$ (we describe below how to make the choice between $\beta_{S,0}$ and $\beta_{S,1}$). A document about Hungary and Germany, in contrast, would be a mixture of the topics $\beta_{C,\text{Germany}}$, $\beta_{C,\text{Hungary}}$, $\beta_{B,0}$, and either $\beta_{S,0}$ or $\beta_{S,1}$.

Once these topics are fixed for a document, the language model proceeds as with LDA for each word: each word in the document comes from one of four topics, with probability for topic k proportional to the topic mixture $\mathbb{E}[\theta]_k$. We illustrate this model graphically in Figure B.1. Note that we keep the topic mixture θ global instead of local to each document because the topics are already very constrained.

Determining the sentiment topic: connecting dyadic sentiment and text

Up to now the dyadic sentiment model and the language model have been developed independently. We connect the two models by using the binary sentiment parameter κ_d to index the sentiment topic for a document: document d takes topic β_{S,κ_d} for its sentiment topic.[3] In other words, if two countries are far apart in the latent space, then when they interact in document d, this interaction is likely to be negative (i.e., $\kappa_d = 1$, and the language used to describe their relationship will come from topic $\beta_{S,1}$. If they were instead close together in this latent space, the language used to describe their relationship would come from topic $\beta_{S,0}$.

We can now specify the generative model of a document language, given the sentiment κ_d for each interaction between countries. We begin by specifying the global topics.

1. First, draw topics:
   (a) For nation $c = 1, \ldots, C$:
      - Draw topic $\beta_{\text{C},c} \sim \text{Dir}(1, \ldots, 1)$.
   (b) Draw background topic $\beta_{\text{B},0} \sim \text{Dir}(1, \ldots, 1)$.
   (c) Draw positive-interaction topic $\beta_{\text{S},0} \sim \text{Dir}(1, \ldots, 1)$
   (d) Draw negative-interaction topic $\beta_{\text{S},1} \sim \text{Dir}(1, \ldots, 1)$
2. Next, draw the global topic mixture $\theta \sim \text{Dir}(1, 1, 1, 1)$.
3. Finally, draw documents.

   For document $d = 1, \ldots, D$, each representing interactions between pairs of countries $c_{d,1}, c_{d,2}$:

[3] We also make a small adjustment to ensure that the model converges to a reasonable mode. There are two main components of this model: a language model and a sentiment model. We introduce a parameter $\nu \sim \mathbb{N}(\nu, \nu\nu\nu)$ and per-document parameters $\nu_d \sim \mathbb{R}(\nu, 0.001)$ and define the binary sentiment $\kappa_d \sim \sigma(s_d \nu_d)$.

(a) Draw sentiment index $\kappa_d \sim \sigma(s_d)$

(b) For word $n = 1, \ldots, N_d$:

- Draw $z_n \sim \text{Mult}(\theta_d)$.
- Switch(z_n):
  - If $z_n = (1, 0, 0, 0)$, draw $w_n \sim \beta_{\text{C},c_{d,1}}$.
  - If $z_n = (0, 1, 0, 0)$, draw $w_n \sim \beta_{\text{C},c_{d,2}}$.
  - If $z_n = (0, 0, 1, 0)$, draw $w_n \sim \beta_{\text{B},0}$.
  - If $z_n = (0, 0, 0, 1)$, draw $w_n \sim \beta_{\text{S},\kappa_d}$.

We illustrate the combined model in Figure B.1.

Related work

The binary relational language model is founded on ideas discussed by several recent models. Chang *et al.* (2009) developed a model to describe the relationships between "entities" (e.g., countries) with a similar assumption of entity-specific and relationship-specific topics. In Chang *et al.* (2009)'s *Networks Uncovered by Bayesian Inference* (Nubbi) model, each entity had its own entity-specific topic, which was active when that country is discussed. An additional mixture of topics was then used to describe the relationship between countries. Nubbi was then be used to infer relationships between countries that have been tagged in a collection of text documents.

Nubbi inferred relationships between countries by finding similar topic weights between documents. In contrast, we use sentiment to select between topics, with an "upstream" model in which actors are embedded in a latent space. This idea of merging topics at different levels of a hierarchy has also been explored by Chemudugunta *et al.* (2006). Neither of these approaches included a switch variable for selecting between topics.

As noted in the last section, the idea of associating language with sentiment has been explored in considerable detail lately. Some of the most successful supervised approaches handle this with regression methods such as text regression (Kogan *et al.*, 2009). Supervised topic models (Blei and McAuliffe, 2008) offer a fully probabilistic generative model of documents which have an attached label. A key assumption behind supervised topics is that the model can learn topics that capture the underlying sentiment. Supervised topic models do this by assuming that the distribution of documents' sentiment parameters s_d are fully specified given their words' topic indices $\boldsymbol{z}_d$ and regression coefficients $p(s_d|\boldsymbol{z}_d, \boldsymbol{\eta})$. This requires that $p(s_d|\boldsymbol{w}_d, \boldsymbol{\eta}, \boldsymbol{\beta})$ they are fully specified given the text of

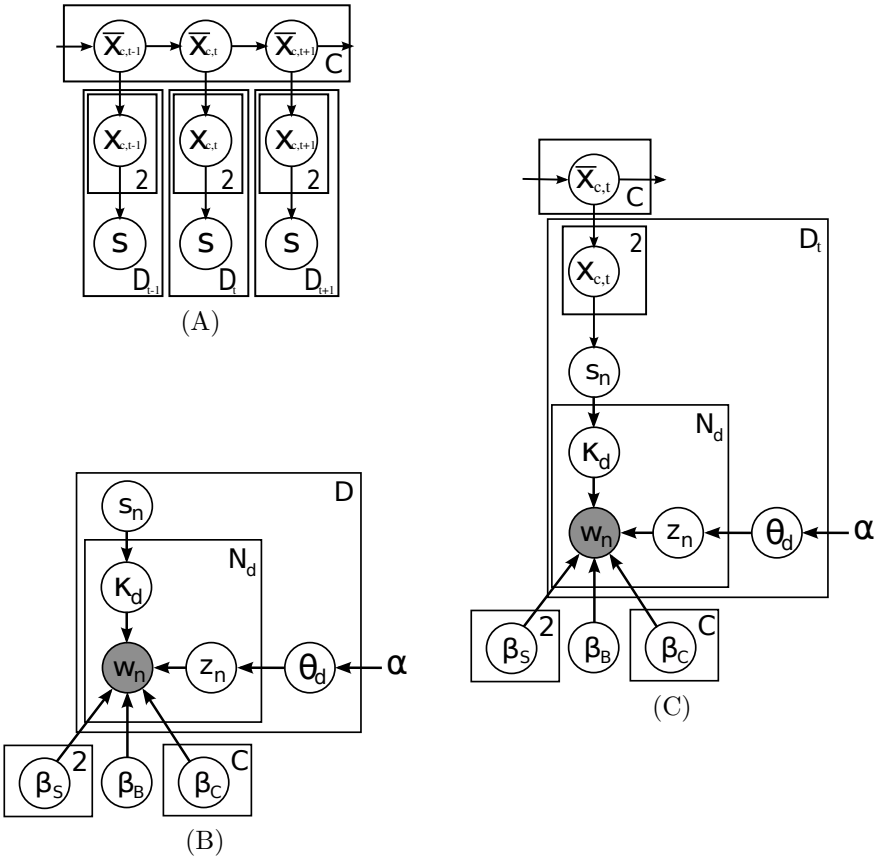


Figure B.1: The dynamic sentiment model (A), a binary mask mixed-membership language model (B), and the full unsupervised foreign relations model (C) (which is a combination of (A) and (C). In (B) and (C), we assign each country its own topic $\beta_{C,\cdot}$. Interactions between countries are characterized by sentiment s_d, which is reflected in the sentiment topic β_{S,κ_d}. The background topic β_B is provided to "soak up" background noise.

documents and η. This means that the topics learned by a standard LDA algorithm will differ from those learned by a supervised LDA algorithm, because they adjust to explain documents' sentiment.

The unsupervised sentiment model is similar to supervised LDA in that the topics adjust as the underlying sentiment parameter s_d differs. In contrast to supervised topics, we assume an inverted conditional independence: words of two documents are conditionally independent given the document's and other model parameters: $p(\boldsymbol{w}_d|s_d, \beta)$, while supervised LDA assumes that sentiment is conditionally independent given words and regression coefficients $\boldsymbol{\eta}$.

B.4.2 Inference

As before, we only observe a collection $\{(\boldsymbol{w}_d, c_{d,1}, c_{d,2})\}_{d \in D}$ of interactions between countries. Each of these interactions takes the form of a vector of wordcounts $\boldsymbol{w}_d$ and a pair of countries interacting. To perform an empirical analysis with this model, we must estimate the latent positions of countries

and the latent topics associated with documents. These are described by the hidden random variables $\bar{x}_c, \theta$, and β. We accomplish this with posterior inference, which will provide us with an estimate of the distribution $p(\bar{x}_c, \theta, \beta | \{(\boldsymbol{w}_d, c_{d,1}, c_{d,2})\}_{d \in D})$.

We fit this model with *maximum a posteriori* (MAP) inference, which has the benefit of a simpler derivation than variational inference. As the reader may recall, the MAP estimate is

$$\begin{aligned}\hat{x}_c, \hat{\theta}, \hat{\beta} &= \arg\max_{\bar{x}_c, \theta, \beta} p(\bar{x}_c, \theta, \beta | \{(\boldsymbol{w}_d, c_{d,1}, c_{d,2})\}_{d \in D}) \\ &= \arg\max_{\bar{x}_c, \theta, \beta} p(\bar{x}_c, \theta, \beta, \{(\boldsymbol{w}_d, c_{d,1}, c_{d,2})\}_{d \in D}). \end{aligned} \tag{B.17}$$

Deriving the algorithm for MAP estimation requires expanding the full likelihood objective, lower-bounding this objective, and maximixing the lower bound with respect to the parameters. We have designed the model in such a way that updates can be performed by a combination of exact coordinate ascent on each parameter (or its expectation), with the exception of countries' mean position $\hat{x}_{c_{d,1}, c_{d,2}}$ during interactions.

The lower bound on the likelihood uses the expectations $\mathbb{E}[\kappa_d]$ and $\mathbb{E}[z_n]$. This means that each interaction is manifested as a mixture $\mathbb{E}[\kappa_d]$ of sentiment, and the observed words are treated as mixtures $\mathbb{E}[\boldsymbol{z}_d]$ of topics. We estimate countries' mean positions $\bar{\hat{x}}$ using a Kalman filter (Kalman, 1960) as in the last section. This inference step is exactly as in the last section.

Countries' per-interaction positions $x_{c_{d,1}, c_{d,2}}$. As in the last section, we infer countries' positions during an interaction by gradient ascent on the objective with respect to their positions $x_{c_{d,1}, c_{d,2}}$.

Estimating topics $\beta_{C,\cdot}, \beta_{S,\cdot}$, and β_B. The update for topics is similar to that in LDA. In both cases, we aggregate the sufficient statistics and normalize during an M-step. We also use Laplace smoothing by adding pseudo counts of 0.1 to these statistics.

Estimating $\mathbb{E}[\kappa]$ and $\mathbb{E}[z_n]$. During inference, we compute the expectations $\mathbb{E}[\kappa_d]$ and $\mathbb{E}[z_n]$, to perform EM. The goal of performing EM is to optimize the bound

$$\log p(\boldsymbol{w}_d | \boldsymbol{\beta}, s_d) \geq \log \mathbb{E}_q \left[\frac{q(\kappa_d, \boldsymbol{z}_d)}{q(\kappa_d, \boldsymbol{z}_d)} p(\boldsymbol{w}_d | \boldsymbol{\kappa}_d, \boldsymbol{z}_d, \boldsymbol{\beta}, s_d) \right] \tag{B.18}$$

$$\geq \mathbb{E}_q \left[q(\kappa_d, \boldsymbol{z}_d) \log \frac{p(\boldsymbol{w}_d | \boldsymbol{\kappa}_d, \boldsymbol{z}_d, \boldsymbol{\beta}, s_d)}{q(\kappa_d, \boldsymbol{z}_d)} \right] \tag{B.19}$$

$$= \mathbb{E}_q \left[p(\boldsymbol{w}_d | \boldsymbol{\kappa}_d, \boldsymbol{z}_d, \boldsymbol{\beta}, s_d) \right] - H(q(\kappa_d, \boldsymbol{z}_d)) \tag{B.20}$$

on the likelihood of documents, where we specify $q(\boldsymbol{\kappa}_d, \boldsymbol{z}_d)$ to be the factorized distribution $q(\boldsymbol{\kappa}_d)q(\boldsymbol{z}_d)$ and write the expectations $q(\boldsymbol{\kappa}_d = 1) = \mathbb{E}_q[\boldsymbol{\kappa}_d]$, $q(\boldsymbol{z}_{dn} = 1) = \mathbb{E}_q[(\boldsymbol{z}_{d,n})]$.

As an aside, note the similarity between Equation B.20 and the variational objective (Equation 2.13). MAP inference using EM can be interpreted as variational inference, in which we use point estimates for many of the random variables and distributions to represent the remaining variables.

Letting S_0 and S_1 index the sentiment word-topic distributions, and letting S index the sentiment topic in the topic indicators z, and recalling that the indicator z_n describes word w_n, this update is:

$$\begin{aligned}
\kappa_{d,0} &\propto \sum_{n=1}^{N_d} \beta_{S_0,w_n} \mathbb{E}[z_{n,S}] \\
\kappa_{d,1} &\propto \exp(s_d) \sum_{n=1}^{N_d} \beta_{S_1,w_n} \mathbb{E}[z_{n,S}] \\
\mathbb{E}[\kappa_{d,i}] &= \frac{\kappa_{d,m}}{\sum_k \kappa_{d,m}}
\end{aligned} \tag{B.21}$$

The update for $\mathbb{E}[z_n]$ is similar. Again letting $S(S_0, S_1)$ refer to the sentiment topic indices, and describing the remaining indices with C_1, C_2, B, we have:

$$\begin{aligned}
z_{n,S} &\propto \mathbb{E}[\theta_S] \left(\beta_{S_0,w_n} \mathbb{E}[\kappa_{d_z,0}] + \beta_{S_1,w_n} \mathbb{E}[\kappa_{d_z,1}]\right) \\
z_{n,k_{c1}} &\propto \mathbb{E}[\theta_{C_1}] \beta_{C_1,w_n} \\
z_{n,k_{c2}} &\propto \mathbb{E}[\theta_{C_2}] \beta_{C_2,w_n} \\
z_{n,k_b} &\propto \mathbb{E}[\theta_B] \beta_{B,w_n} \\
\mathbb{E}[z_{n,i}] &= \frac{z_{n,i}}{\sum_k z_{n,k}}
\end{aligned} \tag{B.22}$$

The update for $\mathbb{E}[\theta_k]$ is similar to κ_{dk}, but we use sufficient statistics from all documents:

$$\begin{aligned}
\theta_k &\propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} \mathbb{E}[z_{n,k}] \\
\mathbb{E}[\theta_k] &= \frac{\theta_k}{\sum_m \theta_m}
\end{aligned} \tag{B.23}$$

B.4.3 Empirical analysis

In this section we perform a very cursory empirical discussion of this model. For this analysis, we used the same *New York Times* (NYT) articles described in the last section. The dimension of the latent space was $p = 2$.

B.5 Derivation of update equations for the Ideal Point Topic Model

Variational inference for the ideal point topic model

Inference for the ideal point topic model requires variational updates (see Jordan *et al.* (1999) for more details about variational inference). Minimizing the KL between the variational distribution and the true posterior is equivalent to maximizing the following lower bound on the model evidence (called the "evidence lower bound", or ELBO):

$$\begin{aligned}
\log p(\boldsymbol{W}, \boldsymbol{V}) = &\int p(\boldsymbol{W}, \boldsymbol{V} | \beta, \boldsymbol{\eta}, I, X, z, \theta) p(\beta, \boldsymbol{\eta}, I, X, z, \theta) \\
&\geq \mathbb{E}_q \left[\sum_D \sum_N \log p(w_n | z_n, \beta) + \log p(z_n | \theta_d) \right] \\
&\quad + \mathbb{E}_q \left[\sum_D \log p(A_d, B_d | z_{d,1:n}, \boldsymbol{\eta}) + \log p(\boldsymbol{\eta}) \right] \\
&\quad + \mathbb{E}_q \left[\sum_U \log p(x_u) + \sum_D \log p(v_{ud} | x_u, A_d, B_d) \right] \\
&\quad + \mathbb{E}_q \left[\sum_D \log p(\theta_d | \alpha) \right] + H(q) \\
&\quad =: \mathcal{L}(\hat{\eta}, \tilde{a}, \tau, \phi, \gamma),
\end{aligned} \tag{B.24}$$

where the expectations are taken with respect to the variational distribution q. This bound is optimized by block coordinate ascent. We repeatedly optimize each variational parameter until the relative increase in the lower bound is below a specified threshold.

One important detail in this equation is that $\mathbb{E}_q \left[\log p(v_{ud} | x_u, A_d, B_d) \right]$ is not available in closed form under the variational distribution. We approximate the expectation in Equation B.24 by applying the second-order multivariate Delta method Bickel and Doksum (2007), also applied to the logit distribution in Chang and Blei (2009); Braun and McAuliffe (2010). This Taylor approximation no longer guarantees that our objective is a lower bound; however, Braun and McAuliffe (2010) have

found it to work better than a first-order approximation (which does maintain the lower bound).

We now turn to the coordinate updates.

Updates for $\boldsymbol{\eta}$ The variational update for $\hat{\eta}$ can be found by collecting terms in the evidence lower bound, taking the derivative with respect to $\hat{\eta}$, setting this to zero, and solving for $\hat{\eta}$. Letting $\boldsymbol{\kappa}_{\text{disc}}$ be a bill's discrimination parameters, we have the the exact update for the vector $\hat{\eta}_{\text{disc}}$:

$$\hat{\eta}_{\text{disc}} \leftarrow \left(\mathbb{E}_q\left[\bar{Z}^T\bar{Z}\right] + \frac{\sigma_d^2}{\sigma_\eta^2}\right)^{-1} \mathbb{E}_q\left[\bar{Z}\right]^T \boldsymbol{\kappa}_{\text{disc}}.$$

The update for $\hat{\eta}_{\text{diff}}$, controlling a bill's difficulty parameter $\boldsymbol{\kappa}_{\text{diff}}$, is analogous.

Updates for β, ϕ, and γ The updates for β and γ are exactly as in LDA Blei *et al.* (2003), and the update for ϕ is exactly as in sLDA Blei and McAuliffe (2008); we omit details here.

Updates for κ_d and τ_u We cannot solve for κ and τ exactly, so they must be optimized via gradient ascent. For bill d, the gradient with respect to κ is

$$\begin{aligned}\nabla_{\kappa_{d,i}}\mathcal{L}(\kappa_{d,i}) = &\sum_D -\frac{\kappa_{d,i} - \eta_i\overline{\phi}}{\sigma_d^2} + \sum_{v\in V(u)} 1_v\tilde{x}_{u_v,i} - \tilde{x}_{u_v,i}\rho_{ud} \\ &- \sum_{v\in V(d)} \frac{1}{2}\Big(\left(\sigma_\kappa^2(\tilde{x}_{u_v}^T\tilde{x}_{u_v}) + \sigma_{\tilde{x}}^2(\kappa_d^T\kappa_d)\right) \\ &\qquad\qquad \times \tilde{x}_{u_v,i}\left(\rho_{ud} - 2\rho_{ud}^2 + 2\rho_{ud}^3\right)\Big) \\ &- \sum_{v\in V(d)} \frac{1}{2}\sigma_{\tilde{x}}^2\Big(\kappa_{d,i}\circ\left(\rho_{ud} - \rho_{ud}^2\right)\Big),\end{aligned}$$

where $\rho_{ud} = \frac{\exp(\tau_u^T\kappa_d - a_d)}{\exp(\tau_u^T\kappa_d - a_d)+1}$ and 1_v is an indicator describing whether vote v was a `yea`-vote.

To optimize this, we apply second-order gradient ascent to the sum $\sum_d \frac{\partial\mathcal{L}}{\partial\kappa_d}$, repeating the updates

$$\kappa_d^n = \kappa_d^{n-1} - \frac{1000}{1000 + n^{0.6}} H^{-1}\left(\nabla_{\kappa_d}\mathcal{L}(\kappa_d)\right)$$

until convergence. In implementation, we constructed the Hessian H numerically by evaluating the above gradient with coordinates perturbed by 10^{-5}. For the data we used, this was sufficiently fast; if a bill has enough votes, an alternative implementation might use more frequent updates and fewer iterations through the votes.

The gradient for the user-ideal parameter τ_u is nearly identical to that for κ:

$$\begin{aligned}\nabla\tau_{u,i}\mathcal{L}(\tau_{u,i}) = &\sum_U -\frac{\tau_{u,i}}{\sigma_u^2} + \sum_{v\in V(u)} 1_v\kappa_{d_v,i} - \kappa_{d_v,i}\rho_{ud} \\ &- \sum_{v\in V(u)} \frac{1}{2}\Big(\big(\sigma_\tau^2(\kappa_{d_v}^T\kappa_{d_v}) + \sigma_\kappa^2(\tau_u^T\tau_u)\big) \\ &\qquad \times \kappa_{d_v,i}\left(\rho_{ud} - 2\rho_{ud}^2 + 2\rho_{ud}^3\right)\Big) \\ &- \sum_{v\in V(u)} \frac{1}{2}\sigma_\kappa^2\Big(\tau_{u,i} \circ \left(\rho_{ud} - \rho_{ud}^2\right)\Big).\end{aligned}$$

Again, we update this via second-order gradient ascent.

Updates for σ_κ and $\sigma_{\tilde{x}}$. Once per iteration, we update the the variances σ_κ and $\sigma_{\tilde{x}}$. As with $\hat{\eta}$, these updates are exact:

$$\begin{aligned}\sigma_\kappa^2 &\leftarrow \frac{ND}{\sum_{D,v\in V(d)} \tau_u^T\tau_u(\rho_{u_vd} - \rho_{u_vd}^2)_n + ND/\sigma_d^2} \\ \sigma_\tau^2 &\leftarrow \frac{NU}{\sum_{U,v\in V(u)} \kappa_d^T\kappa_d(\rho_{ud_v} - \rho_{ud_v}^2)_n + NU/\sigma_u^2},\end{aligned}$$

where above we have U users, D bills, and an N-dimensional ideal-point model.

B Implementation details

We provided details of a variational implementation of the ideal point topic model. Here we describe several modifications to improve this algorithm.

Second order updates. Note that the second-order updates for κ and τ may violate the convexity assumption. To mitigate this, and to prevent the parameters from diverging for large σ_d or σ_u, we add a constant to each element of the diagonal Levenberg (1944). We add a sufficiently large constant to guarantee that all 1×1 and 2×2 principal minors have positive determinant (this is necessary but not sufficient to guarantee that H is positive definite). We have observed that H only requires this adjustment for early model iterations.

Identifiability. In the modeling section, we discussed using nonzero priors for certain legislators to make the posterior identifiable. These priors may not be sufficient to guarantee that the model

Model	Regularization	Accuracy	Log Likelihood	Expected Correct Probability
lars	0.001	0.819	**-0.855**	0.792
lars	0.01	**0.822**	-0.984	**0.793**
lars	0.03125	0.817	-1.091	0.792
lars	0.0625	0.807	-1.214	0.787
lars	0.125	0.799	-1.337	0.781
lars	0.25	0.786	-1.479	0.770
lars	0.5	0.770	-1.640	0.755
lars	1	0.735	-1.903	0.723
l2	0.01	0.815	-0.914	0.793
l2	0.1	0.832	-0.794	0.811
l2	1	0.850	-0.636	0.829
l2	10	0.876	-0.498	0.853
l2	100	0.891	-0.371	0.866
l2	1000	**0.897**	**-0.302**	**0.868**
l2	10000	0.873	-0.324	0.841
iptm	4	0.871	-0.370	0.849
iptm	8	0.869	-0.348	0.845
iptm	16	0.883	-0.321	0.858
iptm	32	0.883	-0.314	0.856
iptm	64	**0.887**	**-0.306**	**0.858**
iptm	128	0.873	-0.456	0.845
yea		0.853	-0.417	0.749

Figure B.2: Prediction metrics for heldout prediction experiments.

Model	Accuracy	Log Likelihood	Expected Correct Probability
l2	0.881	-0.346	0.852
iptm	0.870	-0.346	0.824
yea	0.851	-0.422	0.746

Figure B.3: Prediction metrics for time-series prediction experiments.

finds specific modes. To encourage the model to converge to the desired optimum, we allow the first two iterations of this model one extra dimension for the ideal point. We believe this "blessing of dimensionality" allows the model to rotate ideal points toward the desired mode.

Annealing. We set the model parameters y for σ_d^2 to 1.0 before the first iteration and update it with $y \leftarrow y^{0.9}(\sigma_d^2)^{0.1}$ in a form of "variational annealing". We apply the same annealing to σ_u.

B.6 Experimental Results

The experimental results for cross-fold validation are presented in Figure B.3. Top performers by various metrics are highlighted in bold.

We also display ideal points for all Senators (Figure B.5) and all legislators (Senators and House representatives) (Figure B.4) in the fit of the 111th Congress.

B.7 Additional notes for the Issue-Adjusted Ideal Point Model

B.7.1 Sparsity

Issue adjustments $\boldsymbol{z}_u$ ranged widely, moving some lawmakers significantly. The variational estimates were not sparse, although a high mass was concentrated around 0. Twenty-nine percent of issue adjustments were within $[-0.01, 0.01]$, and eighty-seven percent of issue adjustments were within $[-0.1, 0.1]$.

B.7.2 Hyperparameter settings

The most obvious parameter in the issue voting model is the regularization term λ. The Bayesian treatment described in the Inference section of *How they Vote* demonstrated considerable robustness to overfitting at the expense of precision. With $\lambda = 0.001$, for example, issue adjustments z_{uk} remained on the order of single digits, while the MAP estimate yielded adjustment estimates over 100.

We recommend a modest value of $1 < \theta < 10$. At this value, the model outperforms ideal points in validation experiments on both the House and Senate while maintaining stability in the two-stage model.

B.7.3 Implementation

When performing the second-order updates described in the Inference section, we skipped variable updates when the estimated Hessian was not positive definite (this disappeared when sample sizes grew large enough). We also limited step sizes to 0.1 (another possible reason for smaller coefficients).

B.7.4 Issue labels

In the empirical analysis, we used issue labels obtained from the Congressional Research Service. There were $5,861$ labels, ranging from *World Wide Web* to *Age*. We only used issue labels which were applied to at least twenty five bills in the 12 years under consideration. This filter resulted in seventy-four labels which correspond fairly well to political issues. These issues, and the number of documents each label was applied to, is given in Table B.1.

Table B.1: Issue labels and the number of documents with each label (as assigned by the Congressional Research Service) for Congresses 106 to 111 (1999 to 2010).

Issue label	Bills
Women	25
Military history	25
Civil rights	25
Government buildings; facilities; and property	26
Terrorism	26
Energy	26
Crime and law enforcement	27
Congressional sessions	27
East Asia	28
Appropriations	28
Business	29
Congressional reporting requirements	30
Congressional oversight	30
Special weeks	31
Social services	31
Health	33
Special days	33
California	33
Social work; volunteer service; charitable organizations	33
State and local government	34
Civil liberties	35
Government information and archives	35
Presidents	35
Government employees	35
Executive departments	35
Racial and ethnic relations	36
Sports and recreation	36
Labor	36
Special months	39
Children	40
Veterans	40
Human rights	41
Finance	41
Religion	42
Politics and government	43
Minorities	44
Public lands and natural resources	44

Issue label	Bills
Europe	44
Military personnel and dependents	44
Taxation	47
Government operations and politics	47
Postal facilities	47
Medicine	48
Transportation	48
Emergency management	48
Sports	52
Families	53
Medical care	54
Athletes	56
Land transfers	56
Armed forces and national security	56
Natural resources	58
Law	60
History	61
Names	62
Criminal justice	62
Communications	65
Public lands	68
Legislative rules and procedure	69
Elementary and secondary education	74
Anniversaries	82
Armed forces	83
Defense policy	92
Higher education	103
Foreign policy	104
International affairs	105
Budgets	112
Education	122
House of Representatives	142
Commemorative events and holidays	195
House rules and procedure	329
Commemorations	400
Congressional tributes	541
Congress	693

B.7.5 Corpus preparation

In this section we provide further details of vocabulary selection. We began by considering all phrases with one to five words. From these, we immediately ignored phrases which occurred in more than 10% of bills and fewer than 4 bills, or which occurred as fewer than 0.001% of all phrases. This resulted in a list of 40603 phrases.

We then used a set of features characterizing each word to classify whether it was good or bad to use in the vocabulary. Some of these features were based on corpus statistics, such as the number of bills in which a word appeared. Other features used external data sources, including whether, and how frequently, a word appeared as link text in a Wikipedia article. For training data, we used a manually curated list of 458 "bad" phrases which were semantically awkward or meaningles (such as *the follow bill*, *and sec ammend*, *to a study*, and *pr*) as negative examples in a L_2-penalized logistic regression to select a list of 5,000 "good" words.

Table B.2: Features and coefficients used for predicting "good" phrases. Below, test is a test statistic which measures deviation from a model assuming that words appear independently; large values indicate that they occur more often than expected by chance. We define it as $\text{test} = \frac{\text{Observed count} - \text{Expected count}}{\sqrt{\text{Expected count under a language model assuming independence}}}$.

Coefficient	**Summary**	**Weight**
log(count + 1)	Frequency of phrase in corpus	-0.018
log(number.docs + 1)	Number of bills containing phrase	0.793
anchortext.presentTRUE	Occurs as anchortext in Wikipedia	1.730
anchortext	Frequency of appearing as anchortext in Wikipedia	1.752
frequency.sum.div.number.docs	Frequency divided by number of bills	-0.007
doc.sq	Number of bills containing phrase, squared	-0.294
has.secTRUE	Contains the phrase *sec*	-0.469
has.parTRUE	Contains the phrase *paragra*	-0.375
has.strikTRUE	Contains the phrase *strik*	-0.937
has.amendTRUE	Contains the phrase *amend*	-0.484
has.insTRUE	Contains the phrase *insert*	-0.727
has.clauseTRUE	Contains the phrase *clause*	-0.268
has.provisionTRUE	Contains the phrase *provision*	-0.432
has.titleTRUE	Contains the phrase *title*	-0.841
test.pos	$\ln(max(-\text{test}, 0) + 1)$	0.091
test.zeroTRUE	1 if test = 0	-1.623
test.neg	$\ln(max(\text{test}, 0) + 1)$	0.060
number.terms1	Number of terms in phrase is 1	-1.623
number.terms2	Number of terms in phrase is 2	2.241
number.terms3	Number of terms in phrase is 3	0.315
number.terms4	Number of terms in phrase is 4	-0.478
number.terms5	Number of terms in phrase is 5	-0.454
log(number.docs + 1) * anchortext	ln(Number of bills containing phrase) $\times 1_{\{\text{Appears in Wikipedia anchortext}\}}$	-0.118
log(count + 1) * log(number.docs + 1)	ln(Number of bills containing phrase + 1) $\times$ ln(Frequency of phrase in corpus + 1)	0.246

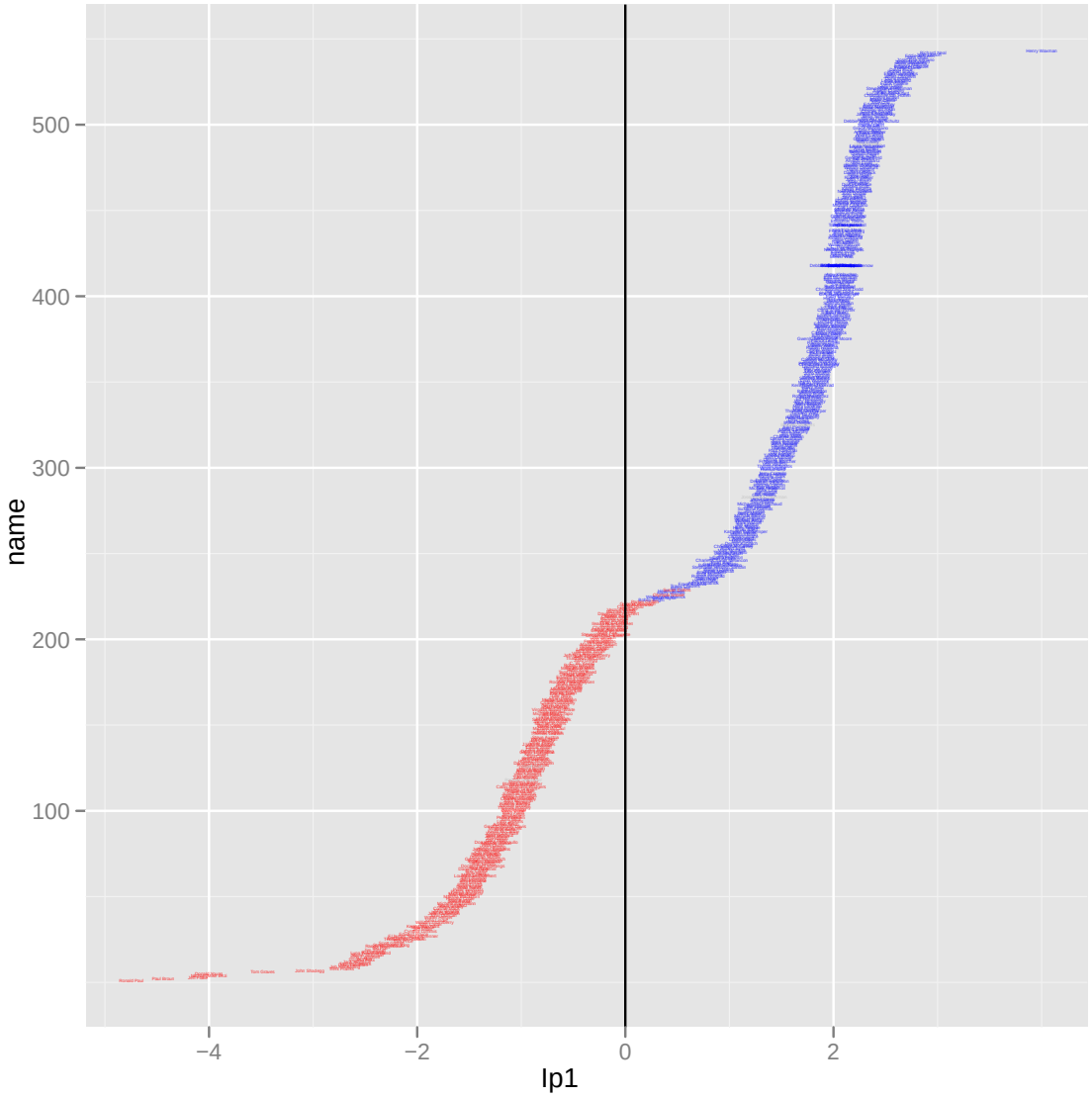


Figure B.4: All legislator ideal points in the 111th Congress. Using votes, ideal points can separate the U.S. political parties Democrats (blue) and Republicans (red). The Y axis contains no information; it is used to stack names for display purposes.

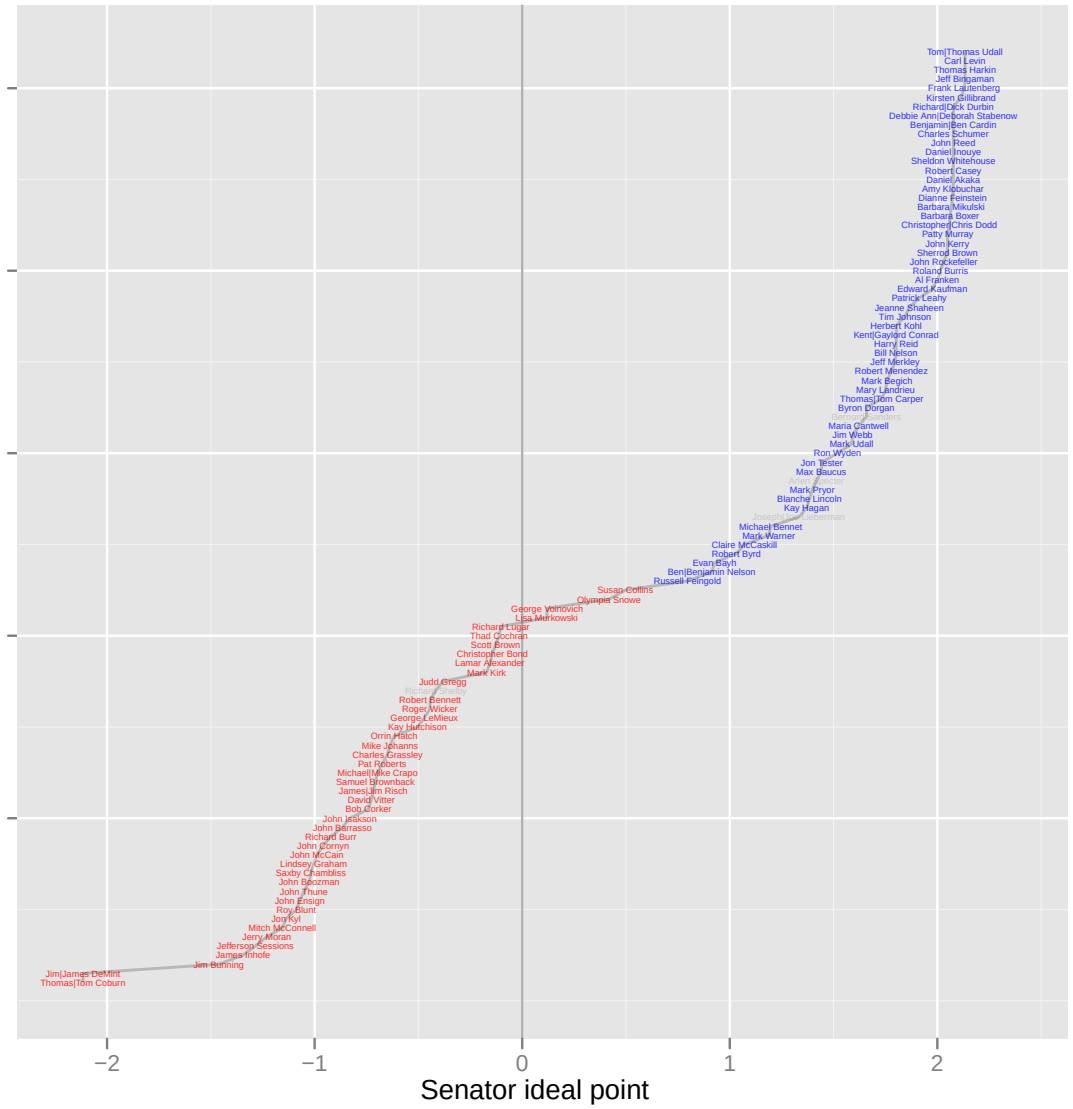


Figure B.5: All Senator ideal points in the 111th Congress. Using votes, ideal points can separate the U.S. political parties Democrats (blue) and Republicans (red).

Bibliography

Adams, R. P. and MacKay, D. J. C. (2007). Bayesian online changepoint detection. Technical report, University of Cambridge, Cambridge, UK.

Agarwal, D. and Chen, B.-C. (2010). fLDA: matrix factorization through latent Dirichlet allocation. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 91–100.

Albert, J. (1992). Bayesian estimation of normal Ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, **17**, 251–269.

Alvarez, M., Luengo, D., and Lawrence, N. D. (2009). Latent force models. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Apple, R. W. J. (1995). Clinton's moscow visit: A climate of mutual doubt. *The New York Times*.

Armagan, A. and Zaretzki, R. L. (2011). A note on mean-field variational approximations in Bayesian probit models. *Computational Statistics and Data Analysis*, **55**(1), 641 – 643.

Beim correspondence (2011). Deborah Beim. Personal correspondence. Personal correspondence.

Bickel, P. J. and Doksum, K. A. (2007). *Mathematical statistics: Basic ideas and selected topics*, volume 1. Pearson Prentice Hall, 2nd ed edition.

Bird, S., Dale, R., Dorr, B. J., Gibson, B., Joseph, M. T., Kan, M.-Y., Lee, D., Powley, B., Radev, D. R., and Tan, Y. F. (2008). The ACL Anthology reference corpus: A reference dataset for bibliographic research. In *Language Resources and Evaluation*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Blei, D. and Lafferty, J. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*.

Blei, D. M. and McAuliffe, J. D. (2008). Supervised topic models. *Advances in Neural Information Processing Systems*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *JMLR*.

Borner, K., Chen, C., and Boyack, K. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, **37**.

Bottou, L. and LeCun, Y. (2004). Large scale online learning. In *Advances in Neural Information Processing Systems*.

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the Americal Statistical Association*, **105**.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117.

Brown, P. F., Pietra, V. J., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, **19**, 263–311.

Carbonetto, P., King, M., and Hamze, F. (2009). A stochastic approximation method for inference in probabilistic graphical models.

Chang, J. and Blei, D. M. (2009). Relational topic models for document networks. *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AIStats) 2009*, **5**.

Chang, J., Boyd-Graber, J., and Blei, D. M. (2009). Connections between the lines: Augmenting social networks with text. In *Refereed Conference on Knowledge Discovery and Data Mining*.

Chemudugunta, C., Smyth, P., and Steyvers, M. (2006). Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *Advances in Neural Information Processing Systems*.

CIA Factbook (2012). CIA World Factbook page for Iran. Available 21 October 2012 at `<http://www.state.gov/e/eb/esc/iransanctions/index.htm>`.

Clinton, J., Jackman, S., and Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review*, **98**(2), 355–370.

Cohn, D. and Hofmann, T. (2001). The missing link - a probabilistic model of document content and hypertext connectivity.

Congressional Research Service (2011). Available September 2012 at <http://www.loc.gov/crsinfo/>.

CoW Homepage (2012). Correlates of War Homepage. Available September 2012 at <http://www.correlatesofwar.org/>.

Cox, G. W. and McCubbins, M. D. (1993). *Legislative Leviathon*. University of California Press.

Cox, G. W. and McCubbins, M. D. (2005). *Setting the Agenda: Responsible Party Government in the U.S. House of Representatives*. Cambridge University Press.

Cox, G. W. and Poole, K. T. (2002). On measuring partisanship in roll-call voting: The U.S. House of Representatives, 1877-1999. *American Journal of Political Science*, **46**(3), pp. 477–489.

Cushman, J. H. J. (1991). War in the gulf: Sea mines; allied ships hunt gulf for iraqi mines. *The New York Times*.

Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. *Proc. of the 24th ICML*.

Enelow, J. and Hinich, M. (1984). *The Spatial Theory of Voting: an Introduction*. Cambridge University Press, New York.

Garfield, E. (1992). Of nobel class: a citation perspective on high impact research authors. *Theoretical Medicine*, **13**, 117–135.

Garfield, E. (2002). Algorithmic citation-linked historiography - mapping the literature of science. *ASIST 2002: Information, Connections, and Community*.

Gartzke, E. (1998). Kant we all just get along? opportunity, willingness, and the origins of the democratic peace. *American Journal of Political Science*, **42**(1), 1–27.

Gerrish, S. and Blei, D. (2011). Predicting legislative roll calls from text. *Proceedings of the International Conference on Machine Learning*.

Ghahramani, Z. and Hinton, G. E. (1996). Switching state-space models. Technical report, Kings College Road, Toronto M5S 3H5.

Google Blog (2008). The Official Google Blog. Available 2012 at <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>.

Govtrack website (2010). Govtrack.us website. Available March 2010 at <http://www.govtrack.us>.

Graves, A. (2011). Practical variational inference for neural networks. In *Neural Information Processing Systems*, pages 2348–2356.

Grimmer, J. and Stewart, B. M. (2012). Text as data: The promise and pitfalls of automatic content analysis. *Political Analysis*. Submitted.

Haberman, C. (2005). Failed truck-bomb plot chills israel-p.l.o. autonomy talks. *The New York Times*.

Heckman, J. and Snyder, J. (1996). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *RAND Journal of Economics*, **27**(0), 142–189.

Hensel, P. R. (2001). Contentious issues and world politics: The management of territorial claims in the americas, 1816-1992. *International Studies Quarterly*, **45**(1), 81–109.

Herszenhorn, D. M. (2010). Congress wraps up session early as midterm races loom. In *The New York Times*. Available 17 Jan 2011 at <http://www.nytimes.com/2010/09/30/us/politics/-30cong.html>.

Hoff, P., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, **97**, 1090–1098.

Hoffman, M. D., Blei, D. M., and Bach, F. (2010). Online learning for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proceedings of the Twenty-Second Annual International SIGIR Converence on Research and Development in Information Retrieval*.

Ibáñez, A., Larrañaga, P., and Bielza, C. (2009). Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics*, **25**, 3303–3309.

ICoW Homepage (2012). Available September 2012 at <http://www.paulhensel.org/icow.html>.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian logistic regression: a variational approach. *Statistics and Computing*, **10**, 25–37.

Jackman, S. (2001). Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, **14**(9).

Jehl, D. (1996). Middle east talks are effort to aid peres and arafat. *The New York Times*.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. Springer-Verlag, New York.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Learning in Graphical Models*.

Jr., R. F. F. (1965). The congress and america's future.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**(1), 3545.

Kogan, S., Levin, D., Routledge, B., Sagi, J., and Smith, N. (2009). Predicting risk from financial reports with regression. In *ACL Human Language Technologies*, pages 272–280. Association for Computational Linguistics.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Alstyne, M. V. (2009). Computational social science. *Science*, **323**.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proc. 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA. ACM.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, **2**, 164–168.

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., and Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *The British Medical Journal*.

Mann, G., Mimno, D., and McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. In *Joint Conference on Digital Libraries (JCDL)*.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1993). Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, **19**.

Markoff, J. (2011). Armies of expensive lawyers, replaced by cheaper software. In *The New York Times*. Available October 27 2012 at <http://www.nytimes.com/2011/03/05/science/05legal.html?pagewanted=all>.

Martin, A. D. and Quinn, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis*, **10**, 134–153.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Konstan, J. A., and Riedl, J. (2002). On the recommending of citations for research papers. *Proc. 2002 ACM conference on Computer supported cooperative work*, pages 116–125.

Murphy, K. P. (1998). Switching Kalman filters. Technical report.

Nallapati, R. and Cohen, W. (2008). Link-plsa-lda: A new unsupervised model for topics and influence of blogs. *International Conference for Weblogs and Social Media*.

National Archives Press Release (2012). National Archives press release. <http://www.archives.gov/press/press-releases/2012/nr12-83.html>.

National Archives Workshop (2012). National Archives workshop announcement. <http://www.archives.gov/pacific-alaska/seattle/public/workshops.html>.

Nature (1969). How to be overtaken by success. *Nature*, **222**.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial and Applied Mathematics.

NSF Website (2010). A timeline of NSF history. Available January 31 2010 at <http://www.nsf.gov/about/history/>.

NY CA Website (2012). Available March 2012 at <http://www.nycourts.gov/ctapps/>.

Osareh, F. (1996). Bibliometrics, citation analysis and co-citation analysis: A review of literature I. *Libri*, **46**, 149–158.

Owen, A. B. (1998). Latin supercube sampling for very high-dimensional simulations. *ACM Trans. Model. Comput. Simul.*, **8**(1), 71–102.

Paisley, J., Gerrish, S., and Blei, D. (2010). Dynamic modeling with the collaborative kalman filter. *Fifth Annual NYAS Machine Learning Symposium*.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine*, pages 329–334.

Poole, K. T. and Rosenthal, H. (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.

Poole, K. T. and Rosenthal, H. (1991). Patterns of congressional voting. *American Journal of Political Science*, **35**(1), 228–278.

Porter, A. L., Chubin, D. E., and Jin, X.-Y. (1988). Citations and scientific progress: Comparing bibliometric measures with scientist judgments. *Scientometrics*, **13**(3-4), 103–124.

Qazvinian, V. and Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 689–696.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2006). An automated method of topic-coding legislative speech over time with application to the 105th-108th U.S. Senate. In *In Midwest Political Science Association Meeting*, pages 1–61.

Radev, D. R., Joseph, M. T., Gibson, B., and Muthukrishnan, P. (2009). A Bibliometric and Network Analysis of the field of Computational Linguistics. *Journal of the American Society for Information Science and Technology*.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, **22**(3).

Ruanaidh, J. and Fitzgerald, W. J. (1996). Numerical Bayesian methods applied to signal processing (Statistics and Computing).

Salakhutdinov, R. R. and Mnih, A. (2008). Probabilistic matrix factorization. *Advances in Neural Information Processing Systems*, **20**, 1257–1264.

Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, **7**(2), 31–40.

Sarkees, M. R. and Warman, F. (2012). Resort to war: 1816 - 2007. *CQ Press*.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Shaparenko, B. and Joachims, T. (2007). Information genealogy: Uncovering the flow of ideas in non-hyperlinked document databases. In *Proc. ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

State Department (2012a). Iran sanctions (state department website). Available 21 October 2012 at <http://www.state.gov/e/eb/esc/iransanctions/index.htm>.

State Department (2012b). Ukraine. Available 21 October 2012 at <http://www.state.gov/r/pa/ei/bgn/3211.htm>.

Taddy, M. (2012). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*. To appear.

Tang, B. (1993). Orthogonal array-based latin hypercubes. *Journal of the American Statistical Association*.

Tang, J. and Zhang, J. (2009). A discriminative approach to topic-based citation recommendation. *Advances in Knowledge Discovery and Data Mining*, **5476**, 572–579.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *Proceedings of EMNLP*, pages 327–335.

Thompson Reuters (2009). Journal Citation Reports - Science Edition.

Toole, J. J., Knopf, J. L., Wozney, J. M., Sultzman, L. A., Buecker, J. L., Pittman, D. D., Kaufman, R. J., Brown, E., Shoemaker, C., Orr, E. C., Amphlett, G. W., Foster, W. B., Coe, M. L., Knutson, G. J., Fass, D. N., and Hewick, R. M. (1984). Molecular cloning of a cDNA encoding human antihaemophilic factor. *Nature*, **312**, 342–347.

Wainwright, M. J. and Jordan, M. I. (2003). Graphical models, exponential families, and variational inference. Technical report, Dept. of Statistics.

Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th international conference on Knowledge Discovery and Data mining*, pages 448–456, New York, NY, USA. ACM.

Wang, E., Liu, D., Silva, J., Dunson, D., and Carin, L. (2010). Joint analysis of time-evolving binary matrices and associated documents. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 2370–2378.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, **85**(411), 699–704.

Winn, J. and Bishop, C. M. (2001). Variational message passing. **5**.

Zimmer, J. and Stewart, B. (2012). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis (conditionally accepted)*.