Three Papers in Political Methodology

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th>Citation</th>
<th>Stewart, Brandon Michael. 2015. Three Papers in Political Methodology. Doctoral dissertation, Harvard University, Graduate School of Arts &amp; Sciences.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:17467206">http://nrs.harvard.edu/urn-3:HUL.InstRepos:17467206</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Three Papers in Political Methodology

A dissertation presented by

Brandon Michael Stewart

to

the Department of Government

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in the subject of Political Science

Harvard University
Cambridge, Massachusetts
May 2015
Three Papers in Political Methodology

Abstract

This collection of three papers develops two statistical techniques for addressing canonical problems in applied computational social science: unsupervised text analysis and regression with dependent data. In both cases I provide a flexible framework that allows the analyst to leverage known structure within the data to improve inference. The first paper introduces the Structural Topic Model (STM) which generalizes and extends a broad class of probabilistic topic models developed in computer science. Crucially for applied social science, STM provides a framework for estimating the factors which drive topical frequency and content within documents. The second paper explores the challenge that non-convex likelihoods pose for applied research with topic models. The paper presents a series of diagnostics and discusses the under-appreciated role of initialization methods. The third paper introduces Latent Factor Regressions (LFR), a new set of tools for regression modeling in the presence of unobserved heterogeneity or dependence between observations. The approach uses interactive latent effects to provide a unified framework for modeling different data structures, including network, time-series cross-sectional and spatial data.

Each of these methods is designed with a focus on applied work. Estimation algorithms are presented which are fast enough for applied work and software is either currently available (STM) or in development (LFR). The use of these techniques is illustrated with a range of applications from across political science.
# Contents

1 Introduction ................................................................. 1

2 A model of text for experimentation in the social sciences ................. 3
   2.1 Introduction .......................................................... 3
   2.2 A model of text that leverages covariate information ................ 7
   2.3 Estimation and Interpretation ..................................... 12
   2.4 Empirical results and data analysis ................................ 18
   2.5 Related Work ....................................................... 32
   2.6 Concluding Remarks ............................................... 36

3 Navigating the Local Modes of Big Data: The Case of Topic Models ....... 39
   3.1 Introduction .......................................................... 39
   3.2 Introduction to Multi-modality .................................... 41
   3.3 The Case of Topic Models ........................................ 46
   3.4 Similarity Between Topics Across Modes ....................... 52
   3.5 Initialization ....................................................... 65
   3.6 Global Solutions .................................................. 68
   3.7 Conclusion .......................................................... 74

4 Latent Factor Regressions for the Social Sciences .......................... 76
   4.1 Introduction .......................................................... 76
   4.2 Regression with Unobserved Heterogeneity ....................... 78
   4.3 Regression with latent factors ................................... 84
   4.4 Estimation ............................................................ 92
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5 Related Work</td>
<td>105</td>
</tr>
<tr>
<td>4.6 Simulation Evidence</td>
<td>113</td>
</tr>
<tr>
<td>4.7 Applications</td>
<td>121</td>
</tr>
<tr>
<td>4.8 Conclusion</td>
<td>131</td>
</tr>
<tr>
<td>4.9 Appendix Road Map</td>
<td>131</td>
</tr>
<tr>
<td>5 Conclusion</td>
<td>133</td>
</tr>
<tr>
<td>5.1 The Future of Structural Topic Model</td>
<td>133</td>
</tr>
<tr>
<td>5.2 The Future of Latent Factor Regressions</td>
<td>134</td>
</tr>
<tr>
<td>6 Appendix</td>
<td>136</td>
</tr>
<tr>
<td>6.1 Variational Inference Algorithms</td>
<td>137</td>
</tr>
<tr>
<td>6.2 Alternative Approaches</td>
<td>167</td>
</tr>
<tr>
<td>6.3 Two-Way Fixed Effects and Latent Factor Regression</td>
<td>169</td>
</tr>
<tr>
<td>6.4 Improving Accuracy of the Variational Framework</td>
<td>175</td>
</tr>
<tr>
<td>6.5 Simulation</td>
<td>178</td>
</tr>
</tbody>
</table>
One of the reasons I love scholarly research is that it is fundamentally a collaborative enterprise. I’ve been fortunate to work with a number of amazing collaborators and have had the pleasure of discussing ideas with many more. This dissertation would simply not be possible without the support of this broader community.

I am grateful for the financial support of a National Science Foundation Graduate Research Fellowship as well as the institutional support at Harvard’s Institute for Quantitative Social Science and Weatherhead Center for International Affairs. I’ve also benefited from conversations with a huge range of scholars including: Ryan Adams, Ken Benoit, Dave Blei, Patrick Brandt, Amy Catalinac, Antonio Coppola, Jimmy Foulds, Sean Gerrish, Michael Gill, Adam Glynn, Christine Kuang, Horacio Larreguy, Jetson Leder-Luis, Chris Lucas, Helen Milner, Brendan O’Connor, Marc Ratkovic, Padhraic Smyth, Arthur Spirling, Alex Storer, Alex Volfovsky, Hanna Wallach, and Daniel Young, as well as the participants of many workshops around Harvard. Of particular note are my office mates John Marshall and Jen Pan, and my frequent collaborators Justin Grimmer and Yuri Zhukov. I’ve also had the support of an incredible dissertation committee: Gary King, Beth Simmons and Dustin Tingley. In addition to providing advice and guidance in these projects, they have also been amazing collaborators on other join projects.

Undoubtedly my largest intellectual debt is to my amazing coauthors on these and many other projects Molly Roberts and Dustin Tingley. They are both phenomenal scholars and wonderful people from whom I have learned a great deal. Completing support of a broader personal community. For their support I’d like to thank Mary Davenport, Luke Davis, Carolina Ferrerosa, Daniel Young and my parents.
Chapter 1

Introduction

This dissertation develops and studies two new statistical methods for analyzing social science data. The innovation in both methods is predicated on the idea that our data are typically structured in the sense that individual observations are grouped together in a known way. The goal of this dissertation is to present a class of methods which can accommodate a variety of different structures. Throughout the work I maintain a focus on applied work and develop estimation procedures that are fast, accurate and have been (or can be) implemented in open-source software.

The Structural Topic Model (STM) is a method for unsupervised text analysis. STM extends the popular Latent Dirichlet Allocation (LDA) model introduced by Blei, Ng and Jordan (2003). Most notably, STM is able to use contextual information about the documents (i.e. metadata) to improve estimation and interpretation. Chapter 2 introduces STM for a statistical audience and describes how to perform estimation. It is based on the manuscript “A model of text for experimentation in the social sciences” coauthored with Molly Roberts and Edo Airoldi and is (at the time of this writing) under review.

Chapter 3 explores the problems posed by non-convex optimization problems that arise in models such as STM. After introducing the necessary technical background, the chapter introduces diagnostics for model stability and provides simulation evidence for the efficacy of different initialization strategies. The final section of the chapter outlines a particularly promising initialization method based on spectral methods. In spectral methods we use a factorization (e.g., singular value decomposition non-negative matrix factorization or tensor decomposition) to recover a method of moments estimate of the model parameters. These estimators have global consistency guarantees and as we show have excellent performance as initializations of likelihood based inference. Because they are deterministic methods...
they also solve many of the stability problems that motivated the chapter. This chapter is based on a book chapter “Navigating the Local Modes of Big Data: The Case of Topic Models” coauthored with Molly Roberts and Dustin Tingley. It will be published in Data Science for Politics, Policy and Government (edited by R. Michael Alvarez) from Cambridge University Press.

Chapter 4 presents the second method of the dissertation, Latent Factor Regressions (LFR), which provides a framework for estimating regressions with dependent or structured data. As with STM, LFR generalizes and extends a large class of models in the networks and spatial statistics literature. A key contribution of this work is computational. I introduce a class of fast variational inference algorithms for computing the approximate posterior which are initialized by a spectral method of the sort described in Chapter 3. This estimation framework allows for a very general model while still being fast enough to estimate for applied use. This chapter is based on the solo-authored manuscript “Latent Factor Regressions for the Social Sciences.”

Chapter 5 provides a short conclusion and sketch of future work. It is followed by a technical appendix supporting Chapter 4.
Chapter 2

A model of text for experimentation in the social sciences

This chapter derived from Roberts, Margaret E., Brandon M. Stewart and Edoardo M. Airoldi. “A model of text for experimentation in the social sciences”

2.1 Introduction

Written documents provide a valuable source of data for the measurement of latent linguistic, political and psychological variables (e.g., Socher et al. 2009; Grimmer 2010c; Quinn et al. 2010; Grimmer and Stewart 2013). Social scientists are primarily interested in how document metadata, i.e., observable covariates such as author or date, influence the content of the text. With the rapid digitization of texts, larger and larger document collections are becoming available for analysis, for which such metadata information is recorded. A fruitful approach for the analysis of text data is the use of mixtures and mixed membership models (Airoldi et al. 2014b), often referred to as topic models in the literature (Blei 2012). While these models can provide insights into the topical structure of a document collection, they cannot easily incorporate the observable metadata information. Here, we develop a framework for modeling text data that can flexibly incorporate a wide range of document-level covariates and metadata, and capture their effect on topical content. We apply our model to learn about how media coverage of China’s rise varies over time and by newswire service.
Quantitative approaches to text data analysis have a long history in the social sciences (Mendenhall 1887; Zipf 1932; Yule 1944; Miller, Newman and Friedman 1958). Today, the most common representation of text data involves representing a document $d$ as a vector of word counts, $\vec{w}_d \in \mathbb{Z}^V_+$, where each of the $V$ entries map to a unique term in a vocabulary of interest (with $V$ in the order of thousands to tens of thousands) specified prior to the analysis. This representation is often referred to as the 

**bag of words** representation, since the order in which words are used across a document is completely disregarded. A milestone in the statistical analysis of text is the analysis of the disputed authorship of “The Federalist” papers (Mosteller and Wallace 1963, 1964, 1984), which featured an in-depth study of the extent to which assumptions used to reduce the complexity of text data representations hold in practice. Because the bag of words representation retains word co-occurrence information, but looses the subtle nuances of grammar and syntax, it is most appropriate for settings where the quantity of interest is a coarse summary such as topical content (Manning, Raghavan and Schütze 2008; Turney and Pantel 2010). In recent years, there has been a surge of interest in methods for text data analysis in the statistics literature, most of which use the bag of words representation (e.g., Blei, Jordan and Ng 2003; Griffiths and Steyvers 2004; Erosheva, Fienberg and Lafferty 2004; Airoldi et al. 2010; Genkin, Lewis and Madigan 2007; Jeske and Liu 2007; Taddy 2013c; Jia et al. 2014). A few studies also test the appropriateness of the assumptions underlying such a representation (e.g., Airoldi and Fienberg 2003; Airoldi et al. 2006).

Perhaps the simplest topic model, to which, arguably, much of the recent interest in statistical text analysis research can be ascribed, is known as the **latent Dirichlet allocation** (LDA henceforth), or also as the **generative aspect model** (Blei, Ng and Jordan 2001, 2003; Minka and Lafferty 2002). Consider a collection of $D$ documents, indexed by $d$, each containing $N_d$ words, a vocabulary of interest of $V$ distinct terms, and $K$ sub-populations, indexed by $k$ and referred to as **topics**. Each topic is associated with a $V$-dimensional probability mass function, $\vec{\beta}_k$, that controls the frequency according to which terms are generated from that topic. Documents are exchangeable. The data generating process for document $d$ assigns terms in the vocabulary to each of the $N_d$ positions; instances of terms that fill these positions are typically referred to as the **words**. The process begins by drawing a $K$-dimensional Dirichlet vector $\theta_d$ that captures the expected proportion of words in document $d$ that can be attributed to each topic. Then for each position (or, equivalently, for each word) in the document, indexed by $n$, it proceeds by sampling an indicator $z_{d,n}$ from a Multinomial$_k(\theta_d, 1)$ whose positive component denotes which topic such position is associated with. The process ends by sampling the actual word indicator $w_{d,n}$ from a Multinomial$_V(B_{z_{d,n}}, 1)$, where the matrix $B = [\vec{\beta}_1 | \ldots | \vec{\beta}_K]$.

\footnote{For the sake of clarity, we wish to emphasize that terms in the vocabulary are unique, while distinct words in a document may instantiate multiple occurrences of the same term.}
form the distribution associated with the active topic.

In practice, social scientists often know more about a document than its word counts. For example, open-ended responses collected as part of a survey experiment include additional information about the respondents (Roberts et al. 2014). From a statistical perspective, it would be desirable to include additional covariates and information about the experimental design into the model to improve estimation of the topics. In addition, the relationships between the observed covariates and latent topics is most frequently the estimand of scientific interest. Here, we allow for such observed covariates to affect two components of the model, the proportion of a document devoted to a topic, which we refer to as topic prevalence and the word rates used in discussing a topic, which we refer to as topical content.

We leverage generalized linear models (GLMs henceforth) to introduce covariate information into the model. Prior distributions with globally shared mean parameters in the latent Dirichlet allocation model are replaced with means parameterized by a linear function of observed covariates. Specifically, for topic prevalence, the Dirichlet distribution that controls the proportion of words in a document attributable to the different topics is replaced with a logistic Normal distribution with a mean vector parametrized as a function of the covariates (Aitchison and Shen 1980). For topical content, we define the distribution over the terms associated with the different topics as an exponential family model, similar to a multinomial logistic regression, parametrized as a function of the marginal frequency of occurrence deviations for each term, and of deviations from it that are specific to topics, covariates and their interactions. We shall often refer to the resulting model as the structural topic model, because the inclusion of covariates is informative about structure in the document collection and its design. From an inferential perspective, including covariate information allows for partial pooling of parameters along the structure defined by the covariates.

As with other topic models, the exact posterior for the proposed model is intractable, and suffers from identifiability issues, in theory and in practice (Airoldi et al. 2014a). Inference is further complicated in our setting by the non-conjugacy of the logistic Normal with the multinomial likelihood. We develop a partially collapsed variational Expectation-Maximization algorithm that uses a Laplace approximation to the non-conjugate portion of the model (Dempster, Laird and Rubin 1977; Liu 1994; Meng and Van Dyk 1997; Blei and Lafferty 2007; Khan and Bouchard 2009; Wang and Blei 2013). This inference strategy provides a computationally efficient approach to model fitting that is sufficiently fast and well behaved to support the analysis of large collections of documents, in practice.

Statistical models of text often have tens of thousands of parameters that can be difficult for users to interpret. We illustrate a series of techniques for examining the model and assessing model fit, including the use of posterior predictive checks (Gelman, Meng and Stern 1996; Mimno and Blei 2011). We also briefly describe additional tools for
model selection and interpretation that we have developed in substantive companion articles that analyze open-ended responses in survey experiments (Roberts et al. 2014), and study the fatwas from Islamic clerics who do and do not support violent Jihad (Lucas et al. 2015).

The central contribution of this article is to introduce a new model of text that can flexibly incorporate various forms of document-level information. We use this model to study differences among newswire services in the frequency with which they cover topics and the vocabulary with which they describes topics. In particular, we are interested in characterizing how Chinese sources represent topics differently than foreign sources, or whether they leave out specific topics completely, thus providing a measure of media slant and censorship in both Chinese and international news media. The rest of the paper is organized as follows. First, we motivate the use of topic models in the social sciences and provide the essential background for our model. Second, we describe the Structural Topic Model and discuss our algorithm for estimation. Third, we use STM to study media coverage of China’s rise by analyzing variations in topic prevalence and content across five different newswire services over time. Fourth, we relate our model to the existing literature. Finally, we conclude with thoughts on future work and collaborations with the social sciences.

To make the model accessible to social scientists, we developed the R package \texttt{stm}, which handles model estimation, summary and visualization \url{(cran.r-project.org/package=stm)}.

### 2.1.1 Background

Our development of the proposed model is motivated by a common structure in the application of topic models within the social sciences. In these settings, the typical application involves estimating latent topics for a corpus of interesting documents and subsequently comparing how topic proportions vary with an external covariate of interest. While informative, these applications raise a practical and theoretical tension. Documents are assumed to be exchangeable under the model and then are immediately shown to be non-exchangeable in order to demonstrate the research finding.

This problem has motivated the development of a series of application-specific models designed to capture particular quantities of interest (Grimmer 2010a; Quinn et al. 2010; Gerrish and Blei 2012; Ahmed and Xing 2010). Many of the models designed to incorporate various forms of meta-data allow the topic mixing proportions ($\theta$) or the observed words ($w$) to be drawn from document-specific prior distributions rather than globally shared priors $\alpha$, $\beta$ in the LDA model. We refer to the distribution over the document-topic proportions as the prior on topical \textit{prevalence} and we refer to the topic-specific distribution over words as the topical \textit{content} prior. For example, the author-topic model allows the prevalence of topics to vary by author (Rosen-Zvi et al. 2004), the geographic topic model allows topical content
to vary by region (Eisenstein et al. 2010) and the dynamic topic model allows topic prevalence and topic content to drift over time (Blei and Lafferty 2006).

However, for the vast majority of social scientists, designing a specific model for each application is prohibitively difficult. These users would need a general model that would balance flexibility to accommodate unique research problems with ease of use.

Our approach to this builds on two prior efforts to incorporate general covariate information into topic models, the Dirichlet-Multinomial Regression topic model of Mimno and McCallum (2008) and the Sparse Additive Generative Model of Eisenstein, Ahmed and Xing (2011). The model of Mimno and McCallum (2008) replaces the Dirichlet prior on the top of the LDA model with a Dirichlet-Multinomial regression over arbitrary covariates. This allows the prior distribution over document-topic proportions to be specific to a set of observed document features through a linear model. Our model extends this approach by allowing covariance among topics and allowing more flexible functional forms by default.

While the Dirichlet-Multinomial Regression model focuses on topical prevalence, the Sparse Additive Generative Model allows topical content to vary by observed categorical covariates. In this framework, topics are modeled as sparse log-space deviations from a baseline distribution over words. Regularization to the corpus mean ensures that rarely occurring words do not produce the most extreme loadings onto topics (Eisenstein, Ahmed and Xing 2011). Because the model is linear in the log-probability it becomes simple to combine several effects (e.g. topic, covariate or topic-covariate interaction) by simply including the deviations additively in the linear predictor. We adopt a similar infrastructure to capture changes in topical content and extend the setting to any covariates.

Our solution to the need for a flexible model combines and extends these existing approaches to create the Structural Topic Model, so-called because we use covariates to structure the corpus beyond a group of exchangeable documents. The premise is simple: we replace the global priors influencing topic prevalence and topical content (α and β in LDA) with document specific priors parameterized by covariates in the spirit of generalized linear models. This allows for problem-specific partial pooling of information in a general framework familiar to applied statisticians and social scientists.

### 2.2 A model of text that leverages covariate information

We introduce the basic structural topic model and notation in Section 2.2.1. We further discuss how covariates can enter the model in the Section 2.2.2. We conclude with details on the prior specifications used for regularization in
Section 2.2.3.

### 2.2.1 Basic model

Recall that we index the documents by \(d \in \{1 \ldots D\}\) and the words (or positions) within the documents by \(n \in \{1 \ldots N_d\}\). Primary observations consist of words \(w_{d,n}\) that are instances of unique terms from a vocabulary of terms, indexed by \(v \in \{1 \ldots V\}\), deemed of interest in the analysis. The model also assumes that the analyst has specified the number of topics \(K\) indexed by \(k \in \{1 \ldots K\}\). Additional observed information is given by two design matrices, one for topic prevalence and one for topical content, where each row defines a vector of covariates for a given document specified by the analyst. The matrix of topic prevalence covariates is denoted by \(X\), and has dimension \(D \times P\). The matrix of topical content covariates is denoted by \(Y\) and has dimension \(D \times A\). Rows of these matrices are denoted by \(\vec{x}_d\) and \(\vec{y}_d\), respectively. Last, we define \(m_v\) to be the marginal log frequency of term \(v\) in the vocabulary, easily estimable from total counts (e.g., see Airoldi, Cohen and Fienberg 2005).

The proposed model can be conceptually divided into three components: (1) a topic prevalence model, which controls how words are allocated to topics as a function of covariates, (2) a topical content model, which controls the frequency of the terms in each topic as a function of covariates, and (3) a core language (or observation) model, which combines these two sources of variation to produce the actual words in each document. Next, we discuss each component of the model in turn. A graphical illustration of the full data generating process for the proposed model is provided in Figure 2.1.

![Figure 2.1: A graphical illustration of the structural topic model.](image-url)
Chapter 2. A model of text for experimentation in the social sciences

In order to illustrate the model clearly, we will specify a particular default set of priors. The model, however, as well as the R package stm, allow for a number of alternative prior specifications, which we discuss in Section 2.2.3.

The data generating process for document $d$, given the number of topics $K$, observed words $\{w_{d,n}\}$, the design matrices for topic prevalence $X$ and topical content $Y$, scalar hyper-parameters $s, r, \rho$, and $K$-dimensional hyper-parameter vector $\vec{\sigma}$, is as follows:

$$\tilde{y}_k \sim \text{Normal}_p(0, \sigma_k^2 I_p),$$

for $k = 1 \ldots K - 1$, (2.1)

$$\tilde{\theta}_d \sim \text{LogisticNormal}_{K-1}(\Gamma' \vec{\gamma}_d, \Sigma),$$

(2.2)

$$\tilde{z}_{d,n} \sim \text{Multinomial}_K(\tilde{\theta}_d),$$

for $n = 1 \ldots N_d$, (2.3)

$$\tilde{w}_{d,n} \sim \text{Multinomial}_{V}(\vec{B} \tilde{z}_{d,n}),$$

for $n = 1 \ldots N_d$, (2.4)

$$\beta_{d,k,v} = \frac{\exp(m_v + k_{k,v}^{(c)} + k_{y,v,k}^{(c)})}{\sum_v \exp(m_v + k_{k,v}^{(c)} + k_{y,v,k}^{(c)})},$$

for $v = 1 \ldots V$ and $k = 1 \ldots K$, (2.5)

where $\Gamma = [\vec{y}_1 \ldots \vec{y}_K]$ is a $P \times (K-1)$ matrix of coefficients for the topic prevalence model specified by Equations 2.1–2.2, and $\{k_{k,v}^{(c)}, k_{y,v,k}^{(c)}\}$ is a collection of coefficients for the topical content model specified by Equation 2.5 and further discussed below. Equations 2.3–2.4 denote the core language model.

The core language model allows for correlations in the topic proportions using the Logistic Normal distribution (Aitchison and Shen 1980; Aitchison 1982). For a model with $K$ topics, we can represent the Logistic Normal by drawing $\vec{\eta} \sim \text{Normal}_K(0, \Sigma)$ and mapping to the simplex, by specifying $\theta_{d,k} = \exp(\eta_k)/\sum_i \exp(\eta_i)$, where $\eta_k$ is fixed to zero in order to render the model identifiable. Given the topic proportion parameter, $\tilde{\theta}_d$, for each word within document $d$ a topic is sampled from a multinomial distribution $\tilde{z}_{d,n} \sim \text{Multinomial}(\tilde{\theta}_d)$, and conditional on such a topic, a word is chosen from the appropriate distribution over terms $\vec{B} \tilde{z}_{d,n}$, also denoted $\vec{B}_{\tilde{z}_{d,n}}$ for simplicity. While in previous research (e.g., Blei and Lafferty 2007) both $\vec{B}$ and $\vec{B}$ are global parameters shared by all documents, in the proposed model they are specified as a function of document-level covariates.

The topic prevalence component of the model allows the expected document-topic proportions to vary as a function of the matrix of observed document-level covariates ($X$), rather than arising from a single prior shared by all documents. We model the mean vector of the Logistic Normal as a simple linear model model such that $\vec{m}_d = \Gamma' \vec{\gamma}_d$, with an additional regularizing prior on the elements of $\Gamma$ to avoid over-fitting. Intuitively, the topic prevalence model takes the form of a multivariate normal linear model with a single shared variance-covariance matrix of parameters. In the absence of covariates, but with a constant intercept, this portion of the mode reduces to the model by Blei and
To model the way covariates affect topical content, we draw on a parameterization that has proved useful in the text analysis literature for modeling differential word usage (e.g., Mosteller and Wallace 1984; Airoldi et al. 2006; Eisenstein, Ahmed and Xing 2011). The idea is to parameterize the (multinomial) distribution of word occurrences in terms of rate deviations, in log-space, from the rates of a corpus-wide background distribution \( \hat{m} \), which can be estimated or fixed to a distribution of interest. The log-rate deviations can then be specified as a function of topics, of observed covariates, and of topic-covariate interactions. In the proposed model, the log-rate deviations are denoted by a collection of parameters \( \{ \kappa \} \), where the super script indicates which set they belong to, i.e., \( t \) topics, \( c \) for covariates, or \( i \) for topic-covariate interactions. In detail, \( \kappa^{(t)} \) is a \( K \)-by-\( V \) matrix containing the log-rate deviations for each topic \( k \) and term \( v \), over the baseline log-rate for term \( v \). These deviations are shared across all \( A \) levels of the content covariate \( Y_d \). The matrix \( \kappa^{(c)} \) has dimension \( A \times V \), and it contains the log-rate deviation for each level of the covariate \( Y_d \) and each term \( v \), over the baseline log-rate for term \( v \). These deviations are shared across all topics. Finally, the array \( \kappa^{(i)} \) has dimension \( A \times K \times V \), and it collects the covariate-topic interaction effects. For example, for the simple case where there is a single covariate \( (Y_d) \) denoting a mutually exclusive and exhaustive group of documents, such as newswire source, the distribution over terms is obtained by adding these effects in log-space such that the rate \( \beta_{d,k,v} \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v}) \), where \( m_v \) is the marginal log-rate of term \( v \). Typically, \( m_v \) is specified as the estimated (marginal) log-rate of occurrence of term \( v \) in the document collection under study (e.g., see Airoldi, Cohen and Fienberg 2005), but can alternatively be specified as any baseline distribution of interest. The content model is completed by positing sparsity inducing priors for the \( \{ \kappa \} \) parameters, so that topic and covariate effects represent sparse deviations from the background distribution over terms. We defer discussion of prior specification to Section 2.2.3. Intuitively, the proposed topical content model replaces the multinomial likelihood for the words with a multinomial logistic regression, where the covariates are the word-level topic latent variables \( \{ z_{d,n} \} \), the user-supplied covariates \( \{ Y_d \} \) and their interactions. In principle, we need not restrict ourselves to models with single categorical covariates; in practice, computational considerations dictate that the number of levels of topical content covariates be relatively small.

### 2.2.2 Covariate model specifications

With the infrastructure developed above, customizing the proposed model to a particular data set requires specifying a model for the covariates informing topic prevalence and content. We discuss the default specifications of such a model
The default specification of the topic prevalence model is inspired by generalized additive models (Hastie and Tibshirani 1990). Each covariate is included with a flexible basis function or spline, which allows non-linearity in the effects on the latent topic prevalence, but the covariates themselves remain additive in the specification. For the examples in this article we use B-splines (De Boor et al. 1978), but in the R package stm any basis function can be used. This agnostic specification is able to capture a diverse array of possible functional forms while remaining intuitive for the end user. The analyst need only focus on the covariates to include within the model. Intuitively, the inclusion of a particular covariate allows the model to draw strength from documents with similar covariate values when estimating the document-topic proportions, analogously to partial pooling in other Bayesian hierarchical modeling contexts (Gelman and Hill 2007).

Analysts can also include covariates that affect the rate at which terms are used within a topic through the topical content model. For example, an indicator for the terms “democrat” and “republican” can be included to estimate how democrats speak differently about a topic than republicans. Although the framework we have developed is quite general, computational considerations do constrain the choices available to an analyst, in practice. Unlike covariates for topical prevalence, for each observed content covariate combination it is necessary to maintain a dense $K \times V$ matrix; namely, the expected number of occurrences of term $v$ attributable to topic $k$, within documents having that observed covariate level. In order to keep the memory requirements practical, we generally recommend specifying topical content models that feature a single mutually exclusive and exhaustive categorical variable.

### 2.2.3 Prior specifications

We complete the model specification with prior specification for the coefficients of the topic prevalence model, parametrized by the $\Gamma$ matrix, and the topical content model, specified by the $\{\kappa\}$ parameters. Here, we present several choices for placing priors on the parameters of each of these models, all of which are implemented in the R package stm.

The default prior specification for the topic prevalence parameters is a zero mean Gaussian distribution with shared variance parameter; that is, $\gamma_{p,k} \sim \text{Normal}(0, \sigma^2_k)$, and $\sigma^2_k \sim \text{Inverse-Gamma}(a, b)$, where $p$ indexes the covariates, $k$ indexes the topics, and $a, b$ are given parameters for a broad inverse-gamma distribution. There is no prior on the intercept, if included as a covariate. This choice of prior will shrink coefficients towards zero, but will not induce sparsity. As a sparsity promoting alternative, the R package stm provides estimation using an $L_1$ penalty, as implemented in the
R package `glmnet` (Friedman, Hastie and Tibshirani 2010).

In the topical content specification, a sparsity inducing prior on the collection of $\{\kappa\}$ parameters is necessary for interpretability. Limiting the number of non-zero coefficients helps to focus the analyst attention on the subset of terms which define the topic. We provide two options based on scale mixtures: the improper Normal-Jeffrey’s prior (Guan and Dy 2009; Figueiredo 2003; Eisenstein, Ahmed and Xing 2011) and the Laplace prior (Friedman, Hastie and Tibshirani 2010). The Normal-Jeffrey’s compound distribution is a parameter-free prior that posits each coefficient is sampled from a Normal distribution, $\kappa_i \sim \text{Normal}(0, \tau_i)$, with its own scale parameter, drawn from an improper Jeffrey’s prior, $\tau_i \sim 1/\tau_i$ (Figueiredo 2003). The Laplace prior also leads to a sparse maximum a-posteriori estimate. Both choices fall into the broader class of priors, often referred to as exponential power distributions with a generalized inverse Gaussian density, of which the Normal-Jeffreys is a particular limiting case (Zhang et al. 2012). This class of priors also includes the Gamma-lasso prior structure which has recently been shown to be a powerful concave penalization option (Taddy 2013b). We use the Laplace prior in the examples that follow.

### 2.3 Estimation and Interpretation

The full posterior of interest is proportional to

$$p(\vec{\eta}, \vec{z}, \vec{\kappa}, \vec{\gamma}, \Sigma | \vec{w}, X, Y) \propto \left( \prod_{d=1}^{D} \mathcal{N}(\vec{\eta}_d | X_d \vec{\gamma}, \Sigma) \left( \prod_{n=1}^{N} \mathcal{M}(z_{nd}, \vec{\theta}_d) \mathcal{M}(w_n | \beta_{d, \kappa, k}) \right) \right) \times \prod p(\kappa) \prod p(\gamma)$$

with $\vec{\theta} = \text{softmax}(\vec{\eta})$ and $\vec{\beta}_d \propto \exp(m_v + \kappa_v^A + \kappa_v^{Y_d} + \kappa_v^{X_d,k})$ and the priors on the prevalence and content coefficients $\gamma, \kappa$ specific to the options chosen by the user. As with most topic models the posterior distribution for the structural topic model is intractable and so we turn to methods of approximate inference. In order to allow for ease of use in iterative model fitting, we use a fast variant of nonconjugate variational Expectation Maximization (EM).

Traditionally topic models have been fit using either collapsed Gibbs sampling or mean field variational Bayes (Griffiths and Steyvers 2004; Blei, Ng and Jordan 2003). Because the Logistic Normal distribution introduces non-conjugacy, these standard methods are not available. The original work on Logistic Normal topic models used an approximate Variational Bayes procedure by maximizing a novel lower bound on the marginal likelihood (Blei and Lafferty 2007) but the bound can be quite loose (Ahmed and Xing 2007; Knowles and Minka 2011). Later work
drew on the statistical literature on Bayesian inference for logistic regression models (Groenewald and Mokgatlhe 2005; Holmes and Held 2006) to develop a Gibbs sampler using auxiliary variable schemes (Mimno, Wallach and McCallum 2008). Recently Chen et al. (2013) developed a scalable Gibbs sampling algorithm by leveraging the Polya-Gamma auxiliary variable scheme of Polson, Scott and Windle (2013).

We opt instead for an approximate variational EM algorithm using a Laplace approximation to the expectations rendered intractable by the nonconjugacy (Wang and Blei 2013). In order to speed convergence we also collapse out the word-level topic indicator $z$ while estimating the variational parameters for the logistic normal latent variable, and then reintroduce it when maximizing the topic-word distributions, $\beta$. Thus inference consists in optimizing the variational posterior for each document’s topic proportions in the E-step, and estimating the topical prevalence and content coefficients in the M-step.

### 2.3.1 Variational expectation-maximization

Recall that we can write the logistic normal document-topic proportions in terms of the $K-1$ dimensional Gaussian random variable such that $\hat{\theta} = \frac{\exp(\eta)}{\sum_{k=1}^{K} \exp(\eta_k)}$ where $\eta \sim \text{Normal}(\tilde{\nu}, \Gamma, \Sigma)$ where $\eta_K$ is set to 0 for identification. Inference under the model involves finding the approximate posterior $q(\eta)q(z)$ which maximizes the approximate Evidence Lower Bound (aELBO),

$$\text{aELBO} \approx \sum_{d=1}^{D} E_q[\log p(\eta_d | \mu_d, \Sigma)] + \sum_{d=1}^{D} \sum_{n=1}^{N} E_q[\log p(z_{n,d} | \eta_d)]$$

$$+ \sum_{d=1}^{D} \sum_{n=1}^{N} E_q[\log p(w_{n,d} | z_{n,d}, \beta_d)] - H(q) \quad (2.7)$$

where $q(\eta)$ is fixed to be Gaussian with mean $\lambda$ and covariance $\nu$ and $q(z)$ is a variational multinomial with parameter $\phi$. $H(q)$ denotes the entropies of the approximating distributions. We affix the word approximate to the ELBO to emphasize that it is not a true bound on the marginal likelihood (due to the Laplace approximation) and strictly speaking it isn’t being directly maximized by the updates (see Wang and Blei (2013) for more discussion).

In the E-step we iterate through each document updating the variational posteriors $q(\eta)q(\phi)$. In the M-step we maximize the aELBO with respect to the model parameters $\gamma, \Sigma$, and $\kappa$. After detailing the E-step and M-step, we discuss convergence, properties and initialization before summarizing the complete algorithm.
Chapter 2. A model of text for experimentation in the social sciences

E-Step. Because the logistic-normal is not conjugate with the multinomial, \( q(\eta) \) does not have a closed form update. We instead adopt the Laplace approximation advocated in Wang and Blei (2013) which involves finding the MAP estimate \( \hat{\eta} \) and approximating the posterior with a quadratic Taylor expansion. This results in a Gaussian form for the variational posterior \( q(\eta) \approx N(\hat{\eta}, -\nabla^2 f(\hat{\eta})^{-1}) \) where \( \nabla^2 f(\hat{\eta}) \) is the hessian of \( f(\eta) \) evaluated at the mode.

In standard variational approximation algorithms for the CTM inference iterates between the word-level latent variables \( q(z) \) and the document-level latent variables \( q(\eta) \) until local convergence. This process can be slow, and so we integrate out the latent variables \( z \) and find the joint optimum using quasi-Newton methods (Khan and Bouchard 2009). Thus solving for \( \hat{\eta} \) for a given document amounts to optimizing the function,

\[
f(\eta) = -\frac{1}{2} \log |\Sigma^{-1}| - \frac{K}{2} \log 2\pi - \frac{1}{2} (\eta - \mu)^T \Sigma^{-1} (\eta - \mu) + \left( \sum_v c_v \log \sum_k \beta_v^k e^{\eta_k} - W \log \sum_k e^{\eta_k} \right)
\]

where \( c_v \) is the count of the \( v \)-th term in the vocabulary and \( W \) is the total count of words in the document. We optimize the objective with quasi-Newton methods using the gradient

\[
\nabla f(\eta)_k = \left( \sum_v c_v \left< \phi_v, k \right> \right) - W \theta_k - \left( \Sigma^{-1} (\eta - \mu_d) \right)_k
\]

where \( \theta \) is the simplex mapped version of \( \eta \) and we define the expected probability of observing a given topic-word as

\[
\left< \phi_v, k \right> = \left( \frac{\exp(\eta_k \beta_v^k)}{\sum \exp(\eta_k \beta_v^w)} \right).
\]

This gives us our variational posterior \( q(\eta) = N(\lambda = \hat{\eta}, \nu = -\nabla^2 f(\hat{\eta})^{-1}) \).

We then solve for \( q(z) \) in closed form,

\[
\phi_{n,k} \propto \exp(\lambda_k) \beta_{k,w,n}
\]

M-Step. In the M-step we update the coefficients in the topic prevalence model, topical content model and the global covariance matrix.

The prior on document-topic proportions maximizes the aELBO with respect to the document specific mean \( \mu_{d,k} = X_d \gamma_k \) and the topic covariance matrix \( \Sigma \). Updates for \( \gamma_k \) correspond to linear regression for each topic under the user specified prior with \( \lambda_k \) as the outcome variable. By default we give the \( \gamma_k \) a Normal(0, \( \sigma_k^2 \)) where \( \sigma_k^2 \) is either manually selected or given a broad inverse-gamma prior. We also provide an option to estimate \( \gamma_k \) using an \( L_1 \) penalty.
Chapter 2. A model of text for experimentation in the social sciences

Σ is then estimated as the convex combination of the MLE and a diagonalized form of the MLE,

$$\hat{\Sigma}_{MLE} = \frac{1}{D} \sum_d \nu_d + (\lambda_d - X_d \hat{\gamma})(\lambda_d - X_d \hat{\gamma})^T$$

$$\hat{\Sigma} = w_\Sigma (\text{diag}(\hat{\Sigma}_{MLE})) + (1 - w_\Sigma)(\hat{\Sigma}_{MLE})$$ (2.11)

where the weight $w_\Sigma \in [0, 1]$ is set by the user and we default to zero.

Updates for the topic-word distributions correspond to estimation of the coefficients ($\kappa$) in a multinomial logistic regression model where the observed words are the output, and the design matrix includes the expectations of the word-level topic assignments $E[q(z)] = \phi$, topical content covariates $Y_d$ and their interactions. The intercept $m$ is fixed to be empirical log probability of the terms in the corpus.

Details of estimating $\kappa$ are dependent on the choice prior. When using the Normal-Jeffrey’s prior we base our estimation strategy off Eisenstein, Ahmed and Xing (2011) which uses a block coordinate ascent strategy and alternates between estimation of $\kappa$ and the penalty $\tau$. When using the Laplace prior we use glmnet leveraging the distributed multinomial regression framework of Taddy (2013a). The idea is to use the Poisson/Multinomial duality and a plugin estimator for document specific effects to perform fast coordinate ascent inference.

Convergence. We monitor convergence as proportional change in the aELBO. This is a sum over the document level contributions and can be dramatically simplified from Equation 2.7 to the following form for each document,

$$L_{aELBO} = \left( \sum_{i=1}^{V} w_i \log(\theta_{\beta_i}) \right) - .5 \log |\Sigma|$$

$$- .5(\lambda - \mu)^T \Sigma^{-1}(\lambda - \mu) + .5 \log(|\nu|)$$ (2.12)

For computational tractability this includes only the document-level contributions to the aELBO implicitly conditioning on the coefficients $\kappa, \gamma$. We monitor relative change in the bound to assess convergence.

Properties. Like most unsupervised learning problems the approximate bound in Equation 2.12 does not have a unique maximum. On a simple level there will be $K!$ equivalent global maxima that arise from the invariance in the posterior to renumbering the topics (i.e. “label switching”). As with other topic models the optimization algorithm can become stuck in local optima (Roberts, Stewart and Tingley Forthcoming; Lancichinetti et al. 2014; Koltcov, Koltsova and Nikolenko 2014). While exactly characterizing the complexity of the optimization problem is beyond the scope of
Chapter 2. A model of text for experimentation in the social sciences

this article, we note that inference for even a two topic LDA model can be shown to be NP-hard (Sontag and Roy 2009; Arora, Ge and Moitra 2012). Theoretical analysis of posterior contraction rates for even the simple LDA analysis is a matter of ongoing research (Tang et al. 2014).

Although it is difficult to establish theoretical properties of the estimator, we observe that it works quite well in practice. Convergence is generally monotonic in approximate bound (although it is not guaranteed to be so) and occurs relatively quickly. In a supplemental appendix we provide a short overview of various simulation studies which demonstrate the efficacy of the approach.

Initialization. Given the possibility of converging to a local maxima of the objective, careful initialization of the algorithm plays an important role in the estimation procedure. We note that in general even the local optima provide useful descriptions of the data, and thus it is not necessary to search for the true global optimum. However to increase the expected quality of solutions and minimize runtime, we initialize the model using a standard LDA model. Our package uses two strategies, one based on collapsed Gibbs Sampling (Griffiths and Steyvers 2004) and a second based on the spectral learning algorithm of Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu and Zhu (2013). A complete description of these approaches and comparison of their performance can be found in Roberts, Stewart and Tingley (Forthcoming). In brief, the Gibbs sampling approach yields substantially better initialization than completely random starting values, and the deterministic spectral algorithm general provides an even better start.

Algorithm. We can summarize the entire inference procedure as:

1: repeat
2: for $d \leftarrow 1, D$ do
3: Maximize $f(\eta^*)$ (2.8)
4: Set $\lambda$ equal to the maximum.
5: Set $\nu$ equal to $-\nabla^2(f(\hat{\eta}^*)^{-1})$
6: Solve in closed form for $\phi$ using (2.10)
7: end for
8: Update $\gamma$ (specific to prior, see appendix)
9: Update $\Sigma$ using (2.11)
10: Update $\kappa$ (specific to prior, see appendix)
11: until convergence
2.3.2 Interpretation

After fitting the model we are left with the task of summarizing the topics in an interpretable way (Chang et al. 2009). The majority of topic models are summarized by the most frequent terms within a topic, although there are several methods for choosing higher order phrases (Blei and Lafferty 2009; Mei, Shen and Zhai 2007). We recommend the use of two different methods for generating lists of representative features. The first is the simple metric of the most frequent terms in the topic which provides a general overview. The second is to look at terms with the largest deviations from the marginal distribution (|κ|) which provides a sense of what makes the topic distinct from the baseline of the corpus. In our software, we provide implementations of several alternative metrics for word labeling that trade off term frequency within the topic and exclusivity to that topic (Bischof and Airoldi 2012).

2.3.3 Goodness-of-fit diagnostics

Model checking is an important part of the analysis process, particularly in the high-dimensional settings which are common with text data. In social science settings, model validation techniques also provide an important mechanism for building trust in new findings (Grimmer and Stewart 2013; Chuang et al. 2012). Here we briefly overview some of the tools for diagnosing problems and assessing model fit, deferring details to applied companion pieces (Roberts, Stewart and Tingley Forthcoming; Roberts et al. 2014).

The most common method for evaluating model quality is the calculation of the distribution of the log-likelihood over a held-out set of documents. Due to the presence of latent variables, a direct calculation of the held-out log-likelihood involves an intractable integration problem and numerous strategies have been proposed to deal with this complication (Wallach et al. 2009). The conceptually simplest of these strategies is a “document-completion” approximation. For this metric, a portion of the documents are selected to have a fraction of their words withheld from the training sample. With the observed portion we estimate the document latent variable \( \hat{\theta}_d \), and then at test time we evaluate the log likelihood of the unobserved portion using our estimate of \( \hat{\theta}_d \) and the global parameters \( B \). More accurate, but computationally intensive, strategies have been proposed (Wallach et al. 2009; Foulds and Smyth 2014; Scott and Baldridge 2013).

In addition to held-out likelihood there are a variety of additional automated metrics that have been proposed for evaluating the goodness-of-fit of models of text, which would also work for the proposed model. Among these, Taddy (2011) uses a test for overdispersion in the model residuals to choose the number of topics. Mimno and Blei (2011) leverage posterior predictive checks (Gelman, Meng and Stern 1996) to assess model fit. In Section 2.4.3, we will also
use posterior predictive checks to assess model fit in our case study on China.

A popular metric in the computational linguistic literature is “semantic coherence”, which measures the empirical co-occurrence of terms with high probability under a given topic (Newman et al. 2010; Mimno et al. 2011). The semantic coherence metric and its variants have been shown to correlate reasonably well with expert judgments of topic quality (Mimno et al. 2011). Roberts et al. (2014) propose to calculate semantic coherence together with a measure the degree of differentiation between topics, and choose a model that scores high on both dimensions.

Automatic metrics such as semantic coherence are attractive because they are typically easy to compute and make the interpretation of results somewhat objective. However, experiments on human interpretation of models of text show that the best scoring models on automated metrics are often not rated as highly interpretable by humans (Chang et al. 2009). Semi-automated approaches using a series of crowd-sourced experiments on Amazon’s Mechanical Turk are a promising strategy to fully automate human evaluation of models of text (Lau, Newman and Baldwin 2014). Ultimately, the most effective validation measure is careful reading of the underlying texts and of the model output by domain experts.

2.4 Empirical results and data analysis

In this section we demonstrate that the our proposed model is useful with a combination of simulation evidence and an example application in political science. From a social science perspective, we are interested in studying how media coverage of China’s rise varies between mainstream Western news sources and the Chinese state-owned news agency, Xinhua. We use the STM on a corpus of newswire reports to analyze the differences in both topic prevalence and topical content across five major news agencies.

Before proceeding to our application, we present series of simulation studies. In Section 2.4.1, we start with a very simple simulation that captures the intuition of why we expect the model to be useful in practice. This section also lays the foundation for our simulation procedures. In Section 2.4.2 we demonstrate that the model is able to recover parameters of interest in a more complicated simulation setting which closely parallels our real data. In Section 2.4.3 we further motivate our applied question and present our data. Using the China data we perform a held-out likelihood comparison to three competing models (Section 2.4.3) and check model fit using posterior predictive checks (Section 2.4.3). Finally having validated the model through simulation, held-out experiments and model checking, we present our results in Section 2.4.3.
2.4.1 Estimating non-linear covariate effects

In this simulation we build intuition for why including covariate information into the topic model is useful for recovering trends in topical prevalence. We compare STM with Latent Dirichlet Allocation (LDA) using a very simple data generating process which generates 100 documents using 3 topics and a single continuous covariate. We start by drawing the topic word distributions for each topic \( \tilde{\beta}_k \sim \text{Dirichlet}_{49}(0.05) \). Collecting the topic word distributions into the 3 by 50 matrix \( B \), each document is simulated from

\[
\begin{align*}
N_d & \sim \text{Pois}(50) \\
x_d & \sim \text{Uniform}(0, 1) \\
\tilde{\theta}_d & \sim \text{LogisticNormal}_2(\mu = (.5, \cos(10x_d)), \Sigma = I.5) \\
w_{d,n} & \sim \text{Multinomial}(B\tilde{\theta}_d)
\end{align*}
\]

where we have omitted the token level latent variable \( z \) in order to reduce sampling variance.

We simulate from this data generating 50 times. For each simulated dataset we fit an LDA model using collapsed Gibbs sampling and an STM model. For both cases we use the correctly specified number of topics. For STM we specify the model with the covariate \( x_d \) for each document using a B-spline with 10 degrees of freedom. Crucially we do not provide it any information about the true functional form. LDA cannot use the covariate information by construction.\(^2\)

Interpreting the simulation results is slightly complicated due to posterior invariance to label switching. For both LDA and STM we match the estimated topics to the simulated parameters using the Hungarian algorithm to maximize the dot product of the true \( \theta \) and the MAP estimate (Papadimitriou and Steiglitz 1998; Hornik 2005).

In Figure 2.2 we plot the Loess-smoothed (span= 1/3) relationship between the covariate and the MAP estimate for \( \tilde{\theta}_d \) of the second topic. Each line corresponds to one run of the model and the true relationship is depicted with a thick black line. For comparison the third panel shows the case using the true values of \( \theta \). While the fits based on the LDA model vary quite widely, the proposed model fits essentially all 50 samples with a recognizable representation of the true functional form. This is in some sense not at all surprising, the proposed model has access to valuable information about the covariate that LDA does not incorporate. The result is a very favorable bias-variance tradeoff in

\(^2\)Collapsed Gibbs sampling for LDA is implemented using the \texttt{lda} package in \texttt{R} (Chang 2012). The hyperparameters are fixed to \( \alpha = \eta = .05 \). We run the chain for 500 iterations and average over the last 100 draws.
which our prior produces a very mild bias in the estimate of the covariate effects in return for a substantial variance reduction across simulations.

This simulation demonstrates that STM is able to capture a non-linear covariate effect on topical prevalence. The focus here on the document-topic proportions ($\theta$) differs from prior work in computer science which typically focuses on the recovery of the topic-word distributions ($\beta$). Recovery of $\beta$ is an easier task in the sense that the parameters are global and our estimates can be expected to improve as the number of documents increases (Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu and Zhu 2013). By contrast $\theta$ is a document level parameter where it makes less sense to speak of the number of words increasing towards infinity. Nevertheless, estimates of covariate relationships based on the document level parameters $\theta$ are often the primary focus for applied social scientists and thus we emphasize them here.

### 2.4.2 Frequentist coverage evaluation in a realistic setting

In this section we expand the quantitative evaluation of the proposed model to a more complex and realistic setting. Using the fitted model from the application in Section 2.4.3 as a reference, we simulate synthetic data from the estimated model parameters. The simulated data set includes 11,980 documents, a vocabulary of $V = 2518$ terms, $K = 100$ topics, and covariates for both topic prevalence and topical content. We set the true values of $\theta$ and $\beta$ to the MAP estimates of the reference model and simulate new observed words as above. We then fit the model to the synthetic documents using the same settings (and observed covariates) as we did in estimating the reference model. We
repeat this process 100 times, and, as above, align the topics to the reference model using the Hungarian algorithm. This is a substantially more rigorous test of the inference procedure. With 100 topics, a content covariate with 5 levels and 2518 vocabulary terms, there are over 1.2 million topic-word probabilities that need to be estimated. The documents themselves are on average 167 words long, and for each one of them over 100 topic proportions need to be estimated.

We evaluate the simulations by examining the frequentist coverage of the credible interval for $\bar{\theta}$ and the expected error between the MAP estimate and the truth.

Calculating Credible Intervals  The most straightforward method for defining credible intervals for $\bar{\theta}$ is using the Laplace approximation to the unnormalized topic proportions $\eta$. By simulating draws from the variational posterior over $\eta$ and applying the softmax transformation, we can recover the credible intervals for $\theta$. However, this procedure poses a computational challenge as the covariance matrix $\nu_d$, which is of dimension $K - 1 \times K - 1$ cannot easily be stored for each document, and recalculating $\nu_d$ can be computationally unfeasible.

Instead, we introduce a simpler global approximation of the covariance matrix $\nu_d$, which leverages the MLE of the global covariance matrix $\Sigma$

$$\bar{\nu} = \hat{\Sigma} - (\lambda_d - X_d \hat{\gamma})(\lambda_d - X_d \hat{\gamma})^T \quad (2.13)$$

$$= \frac{1}{D} \sum_d \nu_d. \quad (2.14)$$

The approximation $\bar{\nu}$ equals the sample average of the estimated document-specific covariance matrices $\{\nu_d\}$. Under this approximation it is still necessary to simulate from the multivariate Normal variational posterior, but there are substantial computational gains from avoiding the need to recalculate the covariance matrix for each document. As we show next, this approximation yields credible intervals with good coverage properties.

To summarize, for each document we simulate 2500 draws from the variational posterior $N(\lambda_d, \bar{\nu})$ using the document-specific variational mode $\lambda_d$ and the global approximation to the covariance matrix $\bar{\nu}$. We then apply the softmax transformation to these draws and recover the 95% credible interval of $\theta$. We calculate coverage along each topic separately.

Results  The left panel of Figure 2.3 shows boxplots of the coverage rates grouped by size of the true $\theta$ with the dashed line indicating the nominal 95% coverage. We can see that for very small values of $\theta (< .05)$ and moderate to large
Figure 2.3: Coverage rates for a 95% credible interval on the document-topic proportions $\hat{\theta}$ in a simulated $K = 100$ topic model. The left panel shows the distribution of coverage rates on a nominal 95% credible interval grouped by the size of the true $\theta$. The right panel shows the distribution of the $L_1$ errors, $E[|\hat{\theta} - \tilde{\theta}|]$, where the $\tilde{\theta}$ is the MAP estimate.

values (> .15) coverage is extremely close to the nominal 95% level. The observed discrepancies between empirical and nominal coverage are reasonable. There are several sources of variability that contribute to these deviations. First the variational posterior is conditional on the point estimates of the topic-word distributions $\hat{\beta}$, which are estimated with error. Many of the documents are quite short relative to the total number of topics, thus the accuracy of the Laplace approximation may suffer. Finally, the optimization procedure only finds a local optimum.

Next we consider how well the MAP estimates of $\theta$ compare to the true values. The right panel of Figure 2.3 provides a series of boxplots of the expected $L_1$ error grouped by the true $\theta$. For very small values of $\theta$ the estimates are extremely accurate, and the size of the errors grows little as the true parameter value increases. For very large values of $\theta$ there is a small, but persistent, negative bias that results in underestimation of the large elements of $\theta$.

This simulation represents a challenging case but the model performs well. Additional simulation results can be found in (Roberts et al. 2014). The stm package also automates a permutation style test in which a covariate is randomly permuted and the model is repeatedly re-estimated. This can help determine if there is a risk of overfitting in reported covariate effects. In the next section we validate the model using real data.
2.4.3 Media coverage of China’s rise

Over the past decade, “rising” China has been a topic of conversations, news sources, speeches, and lengthy books. However, what rising China and Chinese governance means for China, the West and the rest of the world is subject to much intense debate (Ikenberry 2008; Ferguson 2010). Tellingly, both Western countries and China accuse each other of slanting their respective medias to obfuscate the true quality of Chinese governance or meaning of China’s newfound power (Fang 2001). Western “slant” and Chinese censorship and propaganda have been blamed for polarizing views among the American and Chinese public (Roy 1996; Johnston and Stockmann 2007), possibly increasing the probability of future conflict between the two countries.

In Section 2.4.3, we study both Western and Chinese media slant about China’s rise through a collection of newspapers containing the word China over a decade of its development. We give a brief analysis of how different media agencies have characterized China’s rise, focusing particularly on key differences in the way the Chinese news agency, Xinhua, represents and covers news topics differently than mainstream Western sources. In doing so, we seek to measure “slant” on a large scale. Proceeding this substantive analysis, in Section 2.4.3 we first show the extent to which our model leads to better prediction out-of-sample than existing models on the data, and the extent to which the proposed model fits the data (using posterior predictive checks).

To explore how different news agencies have treated China’s rise differently, we analyze a stratified random sample (Rosenbaum 1987) of 11,980 news reports containing the term “China” dated from 1997-2006 and originating from 5 different international news sources. For each document in our sample we observe the day it was written and the news wire service publishing the report. Our data include five news sources: Agence France Presse (AFP), the Associated Press (AP), British Broadcasting Corporation (BBC), Japan Economic Newswire (JEN), and Xinhua (XIN), the state-owned Chinese news agency. We include the month a document was written and the news agency as covariates on topical prevalence. We also include news agency as a covariate affecting topical content in order to estimate how topics are discussed in different ways by different news agencies. In our case study we set the number of topics to 100.

Quantitative evaluation

To provide a fully automated comparison of our model to existing alternatives, we estimate the heldout likelihood using the document completion approach (Asuncion et al. 2009; Wallach et al. 2009). To demonstrate that the covariates provide useful predictive information we compare the proposed structural topic model (STM) to latent Dirichlet

---

Chapter 2. A model of text for experimentation in the social sciences

Allocaiton (LDA), the Dirichlet Multinomial Regression topic model (DMR), and the Sparse Additive Generative text model (SAGE). We use a measure of predictive power to evaluate comparative performance among these models: for a subset of the documents we hold back half of the document and evaluate the likelihood of the held out words (Asuncion et al. 2009; Paisley, Wang and Blei 2012). Higher numbers indicate a more predictive model.

Figure 2.4 shows the heldout likelihood for a variety of topic values. We show two plots. On the left is the average heldout likelihood for each model on 100 datasets, and their 95% quantiles. At first glance, in this plot, it seems that STM is doing much better or about the same as the other three models. However, looking at the second plot, the paired differences between the models on each individual dataset, we see that STM consistently outperforms all other models when the models are run on the same dataset. With the exception of the 40 topic run, STM does better than all models in every dataset for every topic number. By focusing on the paired comparison, we see that STM is indeed the preferred choice for prediction.

The main takeaway from this table is that STM performs significantly better than competing models, except for the case of 40 topics, when it has comparable predictive ability to Dirichlet Multinomial Regression model. This suggests that including information on topical prevalence and topical content aids in prediction. Further, STM has more interpretable quantities of interest than its closest competitor because it allows correlations between topics and covariates on topic content. We cover these qualitative advantages in the next section.

![Figure 2.4: STM vs. SAGE, LDA and DMR Heldout Likelihood Comparison. On the left is the mean heldout likelihood and 95% quantiles. On the right is the mean paired difference between the three comparison models and STM.](image-url)
Chapter 2. A model of text for experimentation in the social sciences

Assessing model fit

The most effectively method for assessing model fit is to carefully read documents which are closely associated with particular documents in order to verify that the semantic concept covered by the topic is reflected in the text. The parameter $\theta_d$ provides an estimate of each document’s association with every topic making it straightforward to effective direct analyst engagement with the texts. Our software package makes it easy to identify these documents, read them and create images of components of the text (for examples see: Roberts, Stewart and Tingley (2014)). An overview of manual validation procedures can be found in Grimmer and Stewart (2013).

When automated tools are required we can use the framework of posterior predictive checks to assess components of model fit (Gelman, Meng and Stern 1996; Mimno, Blei and Engelhardt 2014). Mimno and Blei (2011) outlines a framework for posterior predictive checks for the latent Dirichlet Allocation model using mutual information between document indices and observed words as the realized discrepancy function. Under the data generating process, knowing the document index would provide us no additional information about the terms it contains after conditioning on the topic. In practice, we know that natural language is “bursty” and thus we may not expect independence to hold (Doyle and Elkan 2009).

As in Mimno and Blei (2011) we operationalize the check using the instantaneous mutual information between words and document indices conditional on the topic: $\text{IMI}(w, D|k) = H(D|k) - H(D|W = w, k)$ where $D$ is the document index, $w$ is the observed word, $k$ is the topic and $H()$ is the entropy. When the model assumptions hold we expect this quantity to be close to zero because the entropy of each word should be the same as the entropy of the topic distribution under the data generating process. In order to provide a reference for the observed value we plot the value for the top 10 words for three different topics along with 20 draws from the simulated posterior predictive distribution (Gelman, Meng and Stern 1996; Mimno and Blei 2011).

Figure 2.5 gives an example of these checks for three topics. The posterior predictive checks give us an indication of where the model assumptions do and do not hold. Cases where there is a large gap between the observed value (dark circle) and the reference distribution (open circles) indicate cases where the model assumptions do not hold. Generally these discrepancies occur for terminology which is specific to a sub-component of the topic. For example, in the left plot on SARS/Avian flu the two terms with the greatest discrepancies are the word stems for “SARS” and “bird.” These terms would naturally be “bursty” in the sense that once we have observed one use of the term it is likely that we would see it again. A model which split SARS and Avian Flu into separate topics would be unlikely to have this problem. However for our purposes here combining them into one topic is not a problem.
Figure 2.5: *Posterior Predictive Checks using the methodology outlined in Mimno and Blei (2011). The plot shows the top ten most probable words for each of three topics marginalizing over the covariate-specific word distributions. The x-axis gives the instantaneous mutual information which would be 0 in the true data generating process. The black closed circle gives the observed value.*

Another method for performing similar checks involves using the model residuals to estimate the multinomial dispersion parameter (Taddy 2011). Under the model the dispersion parameter of the multinomial will be equal to 1. Taddy (2011) argues that any overdispersion is plausible evidence that the number of topics is set too low. These methods are also implemented within our software.

**Substantive analysis of differential newswire reporting**

While it is useful to demonstrate that STM shows predictive gains, our primary motivation in developing the STM is to create a tool which can help us answer social science questions. Specifically we want to study how the various news sources cover topics related to the last ten years of China’s development and the vocabulary with which these newswires describes the same events. We are mainly interested in how Chinese and Western sources represent prominent international events during this time period differently, or in other words, to describe discrepancies between Chinese and Western portrayals of the same events. Accusations of “slant” have been largely anecdotal, and the STM provides us with a unique opportunity to measure characterizations of news about China on a large scale.

We start with a general topic related to Chinese governance, a topic about Chinese government strategy, leadership transitions, and future policy. We might call this a “China trajectory” topic. Figure 2.6 shows the highest probability of words in this topic for each of the news sources. The news sources have vastly different accounts of China’s trajectory. AFP and AP talk about China’s rule with words like “Tiananmen”, referring to the 1989 Tiananmen student movement, and “Zhao”, referring to the reformer Zhao Ziyang who fell out of power during that incident due to his support of the
students. Even though Tiananmen occurred 10 years before our sample starts, these Western news sources discuss it as central to China’s current trajectory.

Xinhua, on the other hand has more positive view of China’s direction, with words like “build” and “forward”, omitting words like “corrupt” or mentions of the Tiananmen crackdown. Interestingly, the BBC and JEN also have a forward-looking view on China’s trajectory, discussing “reform”, “advancing”, and references to the formation of laws in China. The analysis provides clear evidence of varying perspectives in both Western and Chinese sources on China’s political future, and surprisingly shows significant variation within Western sources.

Second, we turn to what is likely the most controversial event within China during our time period, the crackdown on Falungong. Falungong is a spiritual group that became very popular in China during the 1990s. Due to the scale and organization of the group, the Chinese government outlawed Falungong beginning in 1999, arresting followers, and dismantling the organization.

This topic appears within all of our news sources, since the crackdown occurred within the time period we are studying. Figure 2.7 shows the different ways in which the news sources portray the Falungong incident. Again, we see that the AP and AFP have the most “Western” view of the incident, using words like “dissident”, “crackdown”, and “activist”. The BBC, on the other hand takes a much milder language to talk about the incident, with words such as “illegal”, or “according”. JEN talks a lot about asylum for those fleeing China, with words such as “asylum”, “refugee”, and “immigration”. Xinhua, on the other hand, talks about the topic using exclusively language about crime, for example “crime”, “smuggle”, “suspect”, and “terrorist”. Again, we see not only the difference between Western and Chinese sources, but interestingly large variation in language within Western sources.

Since we included news source as a covariate in estimating topical prevalence part within the model, we can estimate the differences in frequency, or how much each of the news sources discussed the Falungong topic. As shown in Figure 2.8, we see unsurprisingly that Xinhua talks significantly less about the topic than Western news sources. Interestingly, the Western news sources we would identify to have the most charged language, AFP and AP, also talk about the topic more. Slant has a fundamental relationship with topical prevalence, where those with a positive slant on China talk about negative topics less, and those with negative slant on China talk about negative topics more.

Last, we turn to the differing news coverage of SARS during the outbreak of the disease during 2003 in China. First, in Figure 2.9 we show that by smoothing over time, our model is able to capture the SARS and subsequent Avian flu events, described above. The topic model shows how the news both quickly picked up outbreaks of SARS and
Avian flu and quickly stopped talking about them when the epidemics were resolved.4

The Chinese government received a lot of international criticism for its news coverage of SARS, mostly because it reported on the disease much later than it knew that the epidemic was occurring. As shown in Figure 2.10, our model picks up small differences in news coverage between Chinese and Western sources once news coverage began.

4In general, our model picks up both short-lived events like the Olympics and invasion of Iraq, and long-term topical trends, such as discussion about North Korea and nuclear weapons over time and discussion of the environment, both increasing over time.
Figure 2.7: Falungong Topic. Each group of words are the highest probability words for the news source.

happening, although not substantial. In particular, while Western news sources seemed to talk a lot about death, Chinese news sources mainly focused on policy-related words, such as “control”, “fight”, and “aid”, and avoided mentions of death by the disease.

Finally, because the model allows for the inclusion of correlated topics, we can also visualize the relationship between China-related topics in the 1997-2006 period. In particular, we can see how topics are correlated differently
for different news wires. Since political events are all very related, some topics within our dataset are highly correlated. Figures 2.11 and 2.12 are visual representations of the relationship between topics in the BBC newswire and the Xinhua newswire. We depict an edge between topics when they exhibit a positive correlation above 0.1. In separate work (Lucas et al. 2015), we consider more nuanced procedures based on the literature on graph recovery (Blei and Lafferty 2007; Meinshausen and Bühlmann 2006; Liu et al. 2012).

Since 100 topics was too many to include in one plot, we included topics that were substantively most cohesive. The two plots shows three main groups of discussion: economic discussion including manufacturing, taxes, banks and loans, politics including the Communist party, and law in China, and international relations including the military, ASEAN, and Japanese relations. Interestingly, by comparing the two plots we also can get a sense of “slant” by which topics are highly correlated in the different newswire services. For example, in the BBC newswire, the return
Chapter 2. A model of text for experimentation in the social sciences

Figure 2.9: SARS and Avian Flu. Each dot represents the average topic proportion in a document in that month and the line is a smoothed average across time.

of Hong Kong is related to foreign diplomacy. In China, the return of Hong Kong is mainly associated with economic development topics. Similarly, in the Xinhua news corpus, the environmental pollution topic is not highly correlated with other topics. In the BBC corpus, the environmental pollution topic is highly central, associated with topics such as the economy, development, and disasters.

In conclusion, the STM allows us to measure slant of the various newswire services over a ten year period of China’s rise. We see much variation in how the different newswires discuss the rise of China. Unsurprisingly, Xinhua news services omits negative news, and focuses on the positive aspects of China’s rise. Interestingly, however, we see high variation within Western news sources, with AFP and AP taking a much more negative slant on China’s rise than
Chapter 2. A model of text for experimentation in the social sciences

Figure 2.10: SARS and Avian Flu, comparisons between news sources.

the BBC. We believe we are the first to quantify media slant in news sources all over the world on China’s rise, adding to the discussion of how slant will influence perceptions of China in many different countries.

2.5 Related Work

Here we briefly relate our work to the existing literature on social science applications of topic models as well as the broader literature on LDA and its extensions.
2.5.1 Topic models for measurement in social science

Unlike the common applications of information retrieval in computer science, social scientists are often interested in the use of topic models for measurement and applications which fall under the general category of ‘text as data’ (Grimmer and Stewart 2013). The analyst is generally interested in how topic proportions vary with a covariate of
interest, for example: how party manifestos change over time (Catalinac 2011), how the healthcare debate differs by party (Hopkins 2012), how international event types differ by country (Bagozzi and Schrodt 2011), how supreme court justices vary by ideology (Lauderdale and Clark 2014), and how historical news coverage varies over time (Newman
and Block 2006; Yang, Torget and Mihalcea 2011). In each case, we have a tension between the model’s assumption of document exchangeability and the trends demonstrated in the results.

This tension has not gone unacknowledged. In political science, this has led to a series of single membership models where topic proportions are able to vary by author (Grimmer 2010a) or over time (Quinn et al. 2010). These models provide better statistical efficiency through partial pooling as well as a more theoretically based data generation process, but they are limited to the single membership case. In computer science, a series of models have been developed to address specific substantive problems in the study of property law (Wang et al. 2012), historiography (Mimno 2012), voting behavior (Gerrish and Blei 2012), partisan ideology (Ahmed and Xing 2010), language diffusion (Eisenstein et al. 2010) and countless other application areas.

The Structural Topic Model provides a generic framework for estimating these types of models by including the application-specific structure of the problem as covariates within either the topic prevalence or topical content GLMs.

### 2.5.2 Extensions to the latent Dirichlet allocation

The literature on topic models has exploded in recent years with dozens of implementations, hundreds of model extensions and thousands of citations to the original paper. The majority of this work has focused on weakening assumptions of the standard model, with a particular emphasis on model selection by nonparametric approaches and the incorporation of side information to weaken exchangeability assumptions. The STM focuses on the incorporation of side information due to its salience for applied work.

Our work conditions on covariate values, using them to structure the corpus of documents. A separate class of models assume a generative model of the covariates in order to facilitate prediction as in the supervised LDA model of Blei and McAuliffe (2007). In the social science context, we are typically more interested in explanation than prediction and can generally assume that the covariates are always observed (for example see Grimmer (2013a)). Further the generative models create a theoretical tension, the generative story that (for example) the author comes first and generates the text is more believable than that the text comes first and then generates the author (Hopkins and King 2010). As such we choose to condition on the observed values of the covariates.

Numerous additional frameworks have been proposed, many of which appeared during development of our model. These include Factorial LDA (Paul and Dredze 2012), Conditional Topic Random Fields (Zhu and Xing 2010), Labeled LDA (Ramage et al. 2009), sparse mixed effects SAGE (Wang et al. 2012), Kernel Topic Models (Hennig et al. 2011), discriminative LDA (Lacoste-Julien, Sha and Jordan 2008), the doubly correlated nonparametric topic model (Kim...
Chapter 2. A model of text for experimentation in the social sciences

and Sudderth 2011), multinomial regression LDA (Law, Settles and Mitchell 2010), and the inverse regression topic model (Rabinovich and Blei 2014).

2.6 Concluding Remarks

In this paper we have outlined a new mixed membership model for topical content with structural information and demonstrated its use to address questions about the variation in news coverage of China’s rise. We also have outlined some of the features of topic models which are important for the social sciences. We conclude by highlighting some areas of work that would be fruitful for expanding the role of topic models for social science inference.

A productive line of inquiry has focused on the interpretation of topic models (Chang et al. 2009; Mimno et al. 2011). These methods are aided by techniques for dealing with the practical threats to interpretation such as excessive stop-words and categories with overlapping keywords (Wallach, Mimno and McCallum 2009; Zou and Adams 2012). In addition to fully automated approaches, work on interactive topic modeling and user-specified constraints is particularly appropriate to social scientists who may have a deep knowledge of their particular document sets (Andrzejewski, Zhu and Craven 2009; Andrzejewski et al. 2011; Ramage et al. 2009; Hu, Boyd-Graber and Satinoff 2011). We believe that the key to adoption of these methods is a greater emphasis on freely available code and clear non-technical explanations of user-specified values. One significant advantage of our approach is that the prior structure is specified in the language of generalized linear models which is already familiar to many applied social scientists.

A second area we want to emphasize is the recent work on general methods for evaluation and model checking (Wallach et al. 2009; Mimno and Blei 2011). As noted in both the computer-science literature (Blei 2012) and the political-science literature (Grimmer and Stewart 2013), validation of the model becomes even more important when using unsupervised methods for inference or measurement than it is when used for prediction or exploration. While model-based fit statistics are an important part of the process, we also believe that recent work in the automated visualization of topic models (Chaney and Blei 2012; Chuang et al. 2012; Chuang, Manning and Heer 2012) are of equal or greater importance for helping users to substantively engage with the underlying texts. The engagement between the user and the texts is important both for theoretical reasons but also for harnessing the relative strengths of the applied user (Grimmer and King 2011).

In additional work we have applied the proposed model to the study of open-ended survey response (Roberts et al. 2014), applications in comparative politics (Lucas et al. 2015) and student-generated text in massive open online courses (Reich et al. Forthcoming). In Roberts et al. (2014) we also show how to appropriate include measurement
uncertainty from the variational posterior in regressions where the latent topic is used as the outcome variable. The R package, \texttt{stm}, implements all the methods described here in addition to a suite of visualization and post-estimation tools.

Appendix: Estimation of topic prevalence and content

Optimization of coefficients governing the topical content and prevalence models are dependent on the choice of priors. We briefly outline estimation using the default priors.

The default specification for topic prevalence coefficients is:

$$\gamma_{p,k} \sim \mathcal{N}(0, \sigma_k^2)$$  \hspace{1cm} (2.15)

where \( p \) indexes the covariates in the design matrix \( X \). The variance parameter for the prior \( \sigma_k^2 \) can either be set by the user or given a prior distribution \( \sigma_k^2 \sim \text{InvGamma} (s, r) \) where \( s, r \) are shape and rate hyperparameters set to 1 by default. We leave the intercept unpenalized and compute the MAP estimate of \( \gamma \). The \( \hat{\gamma}_k \) update for a given penalty parameter takes the form of a penalized linear regression

$$\hat{\gamma}_k = \left( X^T X + \text{diag}(1/\sigma_k^2) \right)^{-1} X^T \lambda_k$$  \hspace{1cm} (2.16)

When the penalty parameter is unspecified we update the point estimate as

$$\hat{\sigma}^2 = \left( .5 r + \sum_p \hat{\gamma}^2_{p,k} \right) / (.5 s + p)$$

We iterate between the penalty parameter and the coefficients until convergence. We could also adopt a full variational treatment of \( \gamma \) and \( \sigma^2 \) although this would induce coupling between the equations which make it more difficult to parallelize computations.

The default specification for the topical content coefficients is the Normal-Jeffreys prior. Estimation involves alternating between maximization of \( \kappa \) given \( \tau \) which is equivalent to MAP estimation of a multinomial logistic regression with a normal prior. Following Eisenstein, Ahmed and Xing (2011) we use a block relaxation approach where each \( V \)-length vector in \( \kappa \) is updated using BFGS in turn. Taking as an example the vector for topic \( k \) we obtain the objective

37
and gradient

\[
L_{\kappa_k} = \langle \hat{c}_k \rangle \kappa_k - \langle C_k \rangle \log \sum_v \exp(\kappa_{k,v} + m_v)) - \frac{1}{2} \kappa_k^2 / \tau_k
\]  

(2.17)

\[
\nabla L_{\kappa_k} = \langle \hat{c}_k \rangle - \sum_j \langle C_{jk} \rangle \beta_{jk} - \kappa_k / \tau_k
\]  

(2.18)

where \( \langle c_k \rangle \) is \( V \)-length vector of expected counts for each term in the vocabulary for topic \( k \). \( C_k \) is the summation over that vector producing a scalar equal to the expected number of words assigned to topic \( k \). Updates for each covariate and interaction proceed analogously.

After each update of \( \hat{\kappa} \) we update the corresponding penalty vector \( \hat{\tau} \) with its variational expectation

\[
\hat{\tau}_{v,k} = \kappa_{v,k}^2
\]  

(2.19)

When using the Laplace prior for topical content we use the distributed multinomial regression approach of Taddy (2013a). The basic idea is to use a plugin estimator for the document fixed effects to decouple the parameters into independent poisson regression. Thus all the parameters for a particular word in the vocabulary are updated jointly. The regularization parameter controlling the sparsity is chosen using an information criterion approach. See Taddy (2013b,a) for details.
Chapter 3

Navigating the Local Modes of Big Data:
The Case of Topic Models

This chapter derived from Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. “Navigating the Local Modes of Big Data: The Case of Topic Models”. In Data Science for Politics, Policy and Government, R. Michael Alvarez, editor. Cambridge University Press. Forthcoming.

3.1 Introduction

Each day humans generate massive volumes of data in a variety of different forms (Lazer et al. 2009). For example, digitized texts provide a rich source of political content through standard media sources such as newspapers, as well as newer forms of political discourse such as tweets and blog posts. In this chapter we analyze a corpus of 13,246 posts that were written for 6 political blogs during the course of the 2008 U.S. presidential election. But this is just one small example. An aggregator of nearly every document produced by the US federal government, voxgov.com, has collected over 8 million documents from 2010-2014 including over a million tweets from members of Congress. These data open new possibilities for studies of all aspect of political life from public opinion (Hopkins and King 2010) to political control (King, Pan and Roberts 2013) to political representation (Grimmer 2013b).

The explosion of new sources of political data has been met by the rapid development of new statistical tools for meeting the challenges of analyzing “big data.” (Council 2013; Grimmer and Stewart 2013; Fan, Han and Liu 2014).
A prominent example in the field of text analysis is Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003; Blei 2012), a topic model which uses patterns of word co-occurrences to discover latent themes across documents. Topic models can help us to deal with the reality that large datasets of text are also typically unstructured. In this chapter we focus on a particular variant of LDA, the Structural Topic Model (STM) (Roberts et al. 2014), which provides a framework to relate the corpus structure we do have (in the form of document-level metadata) with the inferred topical structure of the model.

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

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

The outline of this chapter is as follows. In Section 3.2 we introduce the problem of multi-modality and provide several examples of models with multiple modes. In Section 3.3 we focus on the particular case of topic models and highlight some of the practical problems that can arise in applied research. In Section 3.4 we introduce a set of tools that allow users to explore the consequences of multi-modality in topic models by assessing stability of findings across multiple runs of the model. In Section 3.5 and Section 3.6 we discuss procedures for carefully initializing models which may produce better solutions. Finally Section 3.7 concludes by returning to the constraints and opportunities afforded by big data in light of the statistical tools we have to analyze this data.
3.2 Introduction to Multi-modality

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

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

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

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

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

We begin with a simple Gaussian (Normal) model with mean \(\mu\) and variance \(\sigma^2\).\(^3\) In the next section we will show how we can generalize this basic setup to a more flexible Gaussian mixture model.

\[ Y \sim N(\mu, \sigma^2) \]

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

\[
L(\mu|y) \propto N(y|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i - \mu)^2}{2\sigma^2}\right)
\]

\[
\ln L(\mu|y) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2} + \frac{\sum_{i=1}^{n} y_i^2}{\sigma^2} + \left(\frac{n}{2}\right)\mu^2
\]

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

---

\(^2\)There exist a large number of optimization procedures for finding optima of a particular function, see Boyd and Vandenberghe (2009) for a review.

\(^3\)This model is equivalent to a Normal linear regression only modeling the intercept; without regressors.

\(^4\)See King (1998) for a more in-depth discussion of this example.

\(^5\)For multi-dimensional likelihoods, if the Hessian is positive definite the model will be strictly convex (only has one optimum); if it is positive semi-definite, it will be convex (two points may share a optimum, on the same plane.)

\(^6\)Other normal linear regression models that are sometimes used in big data applications include lasso (Tibshirani 1996).
3.2.2 Mixture Models

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

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

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

1. Randomly select a distribution $d_i$ with probability $P(d_i) = w_i$, where $\sum w_i = 1$.
2. From the selected distribution, draw $y \sim N(\mu_i, \sigma_i^2)$.

The log likelihood for this model becomes:

$$\ln L(y | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} w_k N(y_n | \mu_k, \sigma_k^2) \right)$$

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

\textsuperscript{7} Although see additional strategies for the lower dimensional case in Kalai, Moitra and Valiant (2012).
Chapter 3. Navigating the Local Modes of Big Data: The Case of Topic Models

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

3.2.3 Latent Dirichlet Allocation

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

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

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

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

1. First, the length of the document is chosen from a Poisson, with prior \( \eta: N \sim \text{Poisson}(\eta) \).
2. Next, the proportion of the document in each topic is drawn, with prior \( \alpha: \theta \sim \text{Dir}(\alpha) \)
3. Last, for each of the \( N \) words:
Chapter 3. Navigating the Local Modes of Big Data: The Case of Topic Models

- A topic for the word is chosen: \( z_n \sim \text{Multinomial}(\theta) \).
- The word is chosen from the topic matrix \( \beta \), selecting the topic that was chosen \( z_n \): \( w_n \sim \text{Multinomial}(\beta^{z_n}) \)

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

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

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

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

To find the probability of the words marginalized over the latent parameters, we integrate over \( z_n \) and \( \theta \).

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

The marginal likelihood itself is intractable in the case of LDA because of the coupling of \( \beta \) and \( \theta \) which leads to an intractable integration problem. The variational EM approach uses Jensen’s Inequality to create a lower bound on the marginal likelihood which we can maximize via coordinate ascent. That is, the algorithm is alternating between

\(^8\) For more information on variational EM, see Jordan et al. (1998); Grimmer (2010b); Bishop et al. (2006).

\(^9\) The posterior distribution of LDA can also be estimated using Gibbs Sampling, see Griffiths and Steyvers (2004) for more information.
updating the content of the topics ($\beta$) and the topical makeup of a document ($\theta$). It is this alternating maximization strategy that leads to multiple local optima. If we could jointly optimize $\beta$ and $\theta$ we would likely have fewer issues of local modes, but the coupling in the marginal likelihood makes this unfeasible.

### 3.3 The Case of Topic Models

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

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

The posterior of the LDA model cannot be computed in closed form. Two popular approximate inference algorithms are collapsed Gibbs sampling (Griffiths and Steyvers 2004) and variational inference (Blei, Ng and Jordan 2003). In this context, both methods can be seen as a form of alternating maximization; in Gibbs sampling we ran-
domly draw from a single parameter conditional on the others and in variational inference we update a single parameter averaging over the other parameters with respect to the approximating distribution (Grimmer 2010b). This process of alternating conditional updates, necessitated by the inability to directly integrate over the posterior, leads to a sensitivity to the starting values of the parameters. The myriad solutions which can result from different starting points is well known amongst computer scientists but infrequently discussed.\footnote{For example, Blei (2012) provides an excellent overview of LDA and related models but does not mention the issue of local optima at all. The original paper introducing LDA, mentions local optima only in passing to warn against degenerate initializations (Blei, Ng and Jordan 2003). Notable exceptions to this trend are Koltcov, Koltsova and Nikolenko (2014); Lancichinetti et al. (2014) which investigate the stability more directly, as our the efforts in this chapter.}

In fact, we can be more precise about the difficulty of the LDA inference problem by introducing some terminology from theoretical computer science. Non-deterministic Polynomial-time hard (NP-hard) problems are a class of problems which it is strongly suspected cannot be solved in polynomial time.\footnote{That is, if $P \neq NP$ then this is the case. However, there is no formal proof that $P \neq NP$.} A more complete definition is beyond the scope of this article, but the classification conveys a sense of the difficulty of a problem. Maximum likelihood estimation can be shown to be NP-hard even for LDA models with only two topics (Sontag and Roy 2011; Arora, Ge and Moitra 2012). These hardness results suggest not only why local optima are a characteristic of the LDA problem but also why they cannot easily be addressed by changes in the inference algorithm. That is, we can reasonably conjecture from these results that without additional assumptions to make the problem tractable, it would be impossible to develop a computationally practical, globally optimal inference algorithm for LDA.\footnote{The exact connection between NP-hard complexity and local modes is difficult to concisely state. Not all convex problems can be provably solved in polynomial time (de Klerk and Pasechnik 2002). However it is sufficient for the argument here to establish that the hardness results imply that there is something inherently difficult about the nature of the problem which makes it unlikely that a computationally practical algorithm with global convergence properties exists without adding additional assumptions.}

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

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

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

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

The implication of this argument is to say that if we found the global optimum we might not choose to use it. This seems counter-intuitive at first, but various forms of the argument have a long tradition in statistics. Consider the argument that we should choose a model on the basis of cross-validation or other forms of held-out prediction. This is the most commonly used evaluation metric for topic models (Wallach et al. 2009; Foulds and Smyth 2014) and also has a strong tradition in political science (Beck, King and Zeng 2000; De Marchi, Gelpi and Grynaviski 2004; Ward, Greenhill and Bakke 2010). Selecting a model which maximizes a held-out predictive measure implies that we may not choose the model which maximizes the \textit{in-sample} objective function. In settings where forecasting is the primary goal the ability to predict a held-out sample is the clear gold standard; however, in the case of topic models, prediction
Chapter 3. Navigating the Local Modes of Big Data: The Case of Topic Models

is not the only relevant standard.

Implicit in this argument is the claim that the objective function need not directly correspond with human judgment. In human evaluations of topic coherence, selecting model parameters to maximize predictive log-likelihood can actually lead to a mild decrease in assessment of human interpretability (Chang et al. 2009; Lau, Newman and Baldwin 2014). Domain expert assessment (Mimno et al. 2011) and alignment to reference concepts (Chuang et al. 2013) have consistently shown that selecting on the objective function alone does not necessarily yield the same model as human selection.

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

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

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

3.3.2 Finding a Reference Model

Choosing a single reference model for analysis is challenging. The ideal selection criterion is the utility of the model for the analyst, which is an inherently subjective and application specific assessment (Grimmer and King 2011; Grimmer and Stewart 2013). There is an inherent tradeoff in selection criteria between how time intensive the criterion is for the analyst and how closely it approximates the theoretical ideal. In this section we outline methods which span the range of high quality to highly automated.
Manual Review The most thorough and time intensive process is a manual review and validation of the model. This entails reading several example documents for each topic and carefully the topic-word distributions to verify that the topics are capturing a single well-defined concept. Depending on the number of topics and the length of the documents this may be a daunting task in itself.

We may also want to consider information beyond the content of the documents themselves. In the social sciences we often have a rich source of additional information in document metadata. Mapping the relations between topics and a document’s author (Grimmer 2010a) or date (Quinn et al. 2010) is an important part of understanding if the model is functioning. When an existing typology of the documents is available, we can evaluate how well it corresponds to the inferred topics (Chuang et al. 2013). Ideally we hope that the model will convey some things we already know, allowing us to validate it, while also providing us with some novel insights. The different types of validation criteria have been well developed in the literature for measurement models and content analysis (Quinn et al. 2010; Grimmer and Stewart 2013; Krippendorff 2012, e.g). 14

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

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

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

---

14 Quinn et al. (2010) present five types of validity for topic models: external, semantic, discriminant, predictive and hypothesis.
Chapter 3. Navigating the Local Modes of Big Data: The Case of Topic Models

and Smith 2014). Visualization of topic models is an active area of research which promises to vastly improve the analyst’s interaction with the model.

**Complete Automated Approaches** The fastest evaluation metrics are those which are completely automated. The most natural metric is the objective function which is generally either a bound or an approximation to the marginal likelihood (Grimmer 2010b). The default standard within the computer science literature is held-out likelihood which provides a measure of how predictive the model is on unseen documents (Wallach et al. 2009; Foulds and Smyth 2014). Evaluating how well the model predicts new data is appealing in its simplicity, but a predictive model need not be the most semantically interpretable.

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

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

### 3.3.3 Assessing Stability

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

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

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

3.4 Similarity Between Topics Across Modes

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

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

3.4.1 Political Blogs

In order to make our discussion concrete we turn to a specific data set. We use a collection of 13,246 blog posts from American political blogs written during the 2008 presidential election (Eisenstein and Xing 2010). Six different blogs, American Thinker, Digby, Hot Air, Michelle Malkin, Think Progress, and Talking Points Memo, were used to construct the corpus. Each blog is given a rating: liberal or conservative. For each blog post the day of the post is recorded. We stemmed, removed a standard list of stopwords and words which appeared in fewer than 1% of the

\footnote{The CMU Poliblog corpus is available at \url{http://sailing.cs.cmu.edu/socialmedia/blog2008.html} and documentation on the blogs is available at \url{http://www.sailing.cs.cmu.edu/socialmedia/blog2008.pdf}. A sample of 5000 posts is also available in the \texttt{stm} package.}
documents. This results in a vocabulary of 2653 words.

To analyze these texts we use STM (Roberts et al. 2014). STM is a mixed-membership topic model in the style of LDA which allows for the inclusion of document-level covariates, in this case rating (liberal/conservative) and time (day of the post). We use the stm package in R which uses a fast variational EM algorithm. We specify topic prevalence as a function of the partisan rating and a smooth function of time. We estimated the model 685 times initializing with a short run of LDA (we return to this in Section 3.5).\footnote{Each model is run to convergence (a relative change of less than $10^{-5}$ in the objective).}

We briefly define a minimal amount of notation for use in later sections. Let $K = 100$ be the user-selected number of topics, $V = 2653$ be the size of the vocabulary and $D = 13246$ be the number of documents. Mixed membership topic models including LDA and STM can be summarized by two matrices of parameters. $\beta$ is a row-normalized $K$-by-$V$ matrix of topic-word distributions. The entry $\beta_{k,v}$ can be interpreted as the probability of observing the $v$-th word in topic $k$. $\theta$ is a row-normalized $D$-by-$K$ matrix of the document-topic distributions. The entry $\theta_{d,k}$ can be interpreted as the proportion of words in document $d$ which arise from topic $k$. Both LDA and STM can be framed as a factorization of the row-normalized $D$-by-$V$ empirical word count matrix $W$, such that $W \approx \theta \beta$. We will use the $\theta$ and $\beta$ matrices to compare the models.

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

The hundred topics estimated in the model cover a huge range of issues spanning the political dimensions of the 2008 presidential election. We select five topics which illustrate different properties of stability to use as running examples. Figure 3.1 shows the top 20 most probable words for each of the example topics. The topics cover Supreme Court rulings, Vice President Cheney, Global Warming Research, Nuclear Weapons issues in Iran and North Korea and the controversy surrounding Barack Obama’s former pastor, Jeremiah Wright.

### 3.4.2 Comparing Topics

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

We can compare the similarity of two models by comparing the topic word distribution $\beta$ or the document-topic
Chapter 3. Navigating the Local Modes of Big Data: The Case of Topic Models

Figure 3.1: Five example topics from the reference model. These are given the labels Supreme Court, Cheney, Global Warming, Iran/N.K. Nukes, and Wright respectively.

distribution $\theta$. Using $\beta$ implies that two topics are considered similar if they generate similar observed words. Using $\theta$ assesses two topics as similar if they load in the same patterns across the corpus. While both approaches are useful, $\beta$ will tend to contract on the true posterior faster than $\theta$ resulting in a less noisy measure. This is because the number of documents will tend to grow faster than the number of unique words in the vocabulary. Before proceeding to pairwise similarity metrics we need to align topics across runs.
**Alignment**  Consider a simple case where we have two runs of the model. We first need to establish which two topics from each run to compare. The topic numbers are arbitrary across each run, which on its own is unproblematic but means that we need to do something additional in order to compare topics to each other across runs. We call the process of deciding which topics to compare, the process of alignment. The alignment itself is determined by some metric of similarity typically on the topic-word distribution. Here we use the inner product between the rows of $\beta$.

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

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

$$L_1 = \sum_v |\beta_{k,v}^{\text{ref}} - \beta_{k,v}^{\text{cand}}|$$

and has a range: $[0,2]$. We discuss alternate metrics, but we use $L_1$ as the result is easy to conceptualize. We discuss the implications of alternative distance metrics in Section 3.4.5.

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

We can establish the comparable metric for documents. Ranking documents by their use of a particular topic, we can count the overlap in the number of the ten documents most strongly associated with a topic. This metric ranges

---

**Footnote:** The Hungarian algorithm is a polynomial time algorithm for solving the linear sum assignment problem. Given a $K$ by $K$ matrix, where entry $i,j$ gives the cost of matching row $i$ to columns $j$, the Hungarian algorithm finds the optimal assignment of rows to columns such that the cost is minimized. The Hungarian algorithm guarantees that this can be solved in $O(K^3)$ time (Papadimitriou and Steiglitz 1998). We use the implementation in the clue package in R (Hornik 2005).

---

55
from \{0, \ldots, 10\} with 10 indicating complete agreement in the two sets.

Figure 3.2: Relation between three measures of topic similarity across all topics and modes. Plotted surface is a kernel smoothed density estimate.

Figure 3.2 plots the relations between each of these three metrics across the aligned topics. Each pair of metrics is strongly correlated in the theoretically anticipated direction. Also as expected, the measure based on the documents is somewhat noisier than the corresponding measure based on the words.

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

### 3.4.3 Aggregations

The pairwise similarities shown in Figure 3.2 are useful for contextualizing the full range of topic pairs; however, to make these metrics more interpretable it is helpful to aggregate up to either the model level or the topic level. Aggregation at the model level gives us a sense of how well the local modes approximate the reference model summing over each topic. Aggregation to the topic level gives us information about how stable a given topic in the reference model is across runs.

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

In Figure 3.3 we plot each of the three similarity metrics on the Y-axis against the approximate bound on the X-axis. The outlier (upper right corner of the first two plots, and lower right of the third) is the reference model which is, by definition, an exact match for itself. The dashed line marks a natural reference point (5 of 10 words or documents
in the left two plots, and an $L_1$ distance in the middle of the range for the third). The blue line gives a simple linear trend line.

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

**Topic Level Aggregations** Aggregation to the topic level provides us with a measure of how stable a topic within the reference model is across different runs. This helps address the applied situation where a researcher has identified a topic of interest, but wants some understanding of how frequent it is across multiple runs of the model. The distribution over topics is plotted in Figure 3.4. The five example topics are each denoted by the dashed lines and a label.

In each plot the distribution varies over essentially the full range of the metric, indicating that some topics are extremely stable across all of the runs while others are essentially unique to the reference model.

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

One of the least stable topics across runs is the Cheney topic. In the reference model the topic is primarily about

---

Figure 3.3: Comparison between the approximation to the bound on the marginal likelihood (the objective function) with similarity metrics aggregated to the model level.
Chapter 3. Navigating the Local Modes of Big Data: The Case of Topic Models

Vice President Cheney whereas other models include broader coverage of the Bush presidency. As an example we chose the local model which is furthest away from the reference model in $L_1$ distance. In Table 3.1 we compare the two version of the topic by comparing the topic-specific probabilities of observing eighteen terms. These terms define the set of words which have probability of at least 0.01 in one of the two models. We can see that while both topics discuss Cheney, the local model discusses President Bush using words such as Bush, Bush’s, George which have negligible probability under the reference model version of the topic.

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

### 3.4.4 Covariate Effect Stability

In applied use of STM, we are often interested in the role played by covariates in driving topical prevalence. Indeed this is a principle advantage of the STM framework: it allows for the inclusion of covariate information in the estimation process and facilitates the estimation of covariate effects on the resulting model. In the Poliblog corpus, we can examine the role of partisanship in topical coverage.
Table 3.1: The topic-specific probabilities of observing 18 words in the Cheney topic in both the reference model and a local solution far away from it. Included words have a probability of at least 0.01 under one of the two versions of the topics. The reference model topic is focused primarily on Vice President Cheney whereas the local mode includes broader coverage of the Bush presidency.

<table>
<thead>
<tr>
<th>Term</th>
<th>Ref. Model</th>
<th>Local Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>administr</td>
<td>&lt; .0005</td>
<td>0.104</td>
</tr>
<tr>
<td>bush</td>
<td>&lt; .0005</td>
<td>0.275</td>
</tr>
<tr>
<td>bush'</td>
<td>&lt; .0005</td>
<td>0.0191</td>
</tr>
<tr>
<td>cheney</td>
<td>0.0464</td>
<td>0.0279</td>
</tr>
<tr>
<td>decis</td>
<td>0.0178</td>
<td>0.0060</td>
</tr>
<tr>
<td>dick</td>
<td>0.0195</td>
<td>0.0109</td>
</tr>
<tr>
<td>execut</td>
<td>0.0226</td>
<td>0.0022</td>
</tr>
<tr>
<td>first</td>
<td>0.0253</td>
<td>0.0001</td>
</tr>
<tr>
<td>georg</td>
<td>&lt; .0005</td>
<td>0.0480</td>
</tr>
<tr>
<td>histori</td>
<td>0.0104</td>
<td>0.0099</td>
</tr>
<tr>
<td>leader</td>
<td>0.0134</td>
<td>&lt; .0005</td>
</tr>
<tr>
<td>nation</td>
<td>0.0102</td>
<td>&lt; .0005</td>
</tr>
<tr>
<td>offic</td>
<td>0.0414</td>
<td>0.0209</td>
</tr>
<tr>
<td>presid</td>
<td>0.5302</td>
<td>0.2868</td>
</tr>
<tr>
<td>presidenti</td>
<td>0.0254</td>
<td>0.003</td>
</tr>
<tr>
<td>role</td>
<td>0.0129</td>
<td>0.0001</td>
</tr>
<tr>
<td>term</td>
<td>0.0025</td>
<td>0.0130</td>
</tr>
<tr>
<td>vice</td>
<td>0.0512</td>
<td>0.0251</td>
</tr>
</tbody>
</table>

We start by unpacking the partisanship effects for our example topics in the reference model. We then show how to assess the stability of these findings across other local modes.

**Unpacking Covariate Effects**  Figure 3.5 plots the expected proportion of topic use in Conservative blogs minus the expected proportion of topic use in Liberal blogs under the reference model. Thus topics more associated with the Conservative blogs appear to the right of zero.

We briefly contextualize the partisan effects in this set of topics. Conservative attention to the Supreme Court topic is primarily driven by the June 2008 *Heller v. District of Columbia* case which struck down parts of the *Firearms Control Regulations Act of 1975* on 2nd Amendment grounds. As discussed in the previous section the Cheney topic is primarily about Dick Cheney's legacy on the Vice Presidency. The coverage is mainly from liberal blogs and is predominantly critical in tone.

The greater conservative attention to global warming is initially surprising given that it is typically a more liberal issue, but it should be remembered that these blogs came from 2008 which was prior to the more recent trend (at time of writing) in liberal assertiveness. We explore this further by examining the posts most associated with this topic.
Figure 3.5: Differences in topical coverage by rating (controlling for time). Effects to the right of 0 indicate a topic more heavily used by Conservatives. Lines indicate 95% confidence intervals using the “global” approximation to measurement uncertainty (Roberts et al. 2014).

Figure 3.6 shows the first 300 characters of the three posts most associated with the topic. The first and third posts are critical of global warming, while the second post describes a report warning against climate change. The first and third are as expected from a Conservative blog and the second is from a Liberal blog.

The Iran and North Korea Nuclear Weapons topic shows a Conservative effect consistent with increased attention to security topics, consistent with conventional views that issue ownership of security is much greater for Republicans. Finally the scandal involving Reverend Jeremeiah Wright, which is critical of then Democratic Primary candidate Barack Obama, is more prevalent on Conservative blogs.

Stability Across Models How stable are these effects across other plausible local modes. A simple way to evaluate this is to align the topics to the reference model and then calculate the effect for each topic.  

18 This is similar to the permutation test methodology developed in Roberts et al. (2014). In Roberts et al. (2014) we are interested in testing whether our finding on the effect of binary treatment indicator is driven by including it as a topic prevalence covariate (that is, are we at risk of baking in our conclusion). We randomly permute the treatment indicator across documents and rerun the model. In each case we calculate the largest treatment effect observed within the data across all topics and compare this distribution to the observed level. If we were baking in the conclusion, the model would discover large treatment effects despite that the the treatment indicator had been randomly assigned. In practice the observed effect is substantially larger than the randomly permuted datasets suggesting that the model is working as expected. Here we are aligning
Deathly news for the religion of Global Warming. Looks like at least one prominent scientific group has changed its mind about the irrefutability of evidence regarding man made climate change. The American Physical Society representing nearly 50,000 physicists has reversed its stance on climate change. NASA has confirmed that a developing natural climate pattern will likely result in much colder temperatures. Of course, the climate alarmists’ favorite dubious data source was also quick to point out that such natural phenomena should not confuse the issue of manmade greenhouse gas induced global warming. Climate change report forecasts global sea levels to rise up to 4 feet by 2100. According to a new report led by the U.S. Geological Survey, the U.S. faces the possibility of much more rapid climate change by the end of the century than previous studies have suggested. The report, Figure 3.6: The first 300 characters of the three posts most associated with the global warming topic. Posts 1 and 3 come from American Thinker and post 2 comes from Think Progress. process produces a distribution over effect sizes, it is important to emphasize the conceptual challenges in interpreting the results. Each model is estimating the effect of the partisan rating but on a slightly different version of the topic. Thus differences arise for two reasons: the document-topic assignments may be different, but also because the topics themselves capture different concepts. The alignment ensures that this concept is the most similar to our reference model (given the alignment method and the similarity metric) but they are not necessarily conceptually identical. Figure 3.7 plots the distribution of effect sizes. Beginning with the first plot on the top-left, we see that the partisan effect for the Supreme Court topic in the reference model has one of the largest observed values across all of the local modes. Not only is the reference model effect out in the tail, but the distribution over effect sizes includes negative as well as positive values. What accounts for this difference? Comparing the most probable words in the reference model with those in an aligned topic for one of the models with a strong liberal effect provides an indication of the differences:

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

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

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

The second plot in Figure 3.7 shows the Cheney topic. Here we see a distribution with three modes where the reference model sits directly on top of the most typical point. Following the discussion in the previous section this reflects the difference between having the topic focus exclusively on Vice President Cheney as opposed to including the broader Bush Presidency.

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

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

Finally, the topic on Reverend Wright shows another case where the reference model is largely consistent with the local modes. There is some distinction between topics which contain coverage of the scandal and those which also contain elements of the positive liberal coverage that followed Barack Obama’s speech on the matter (“A more perfect union”).

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

Differences from Alignment  While most of the analyses above are insensitive to the method of aligning topics, we do observe significant differences in the covariates effects. Global alignments tend to result in more cases where there are several clusters of effect sizes. Consider for example, the Cheney Topic (top-center of Figure 3.7). In the example discussed in Section 3.4.3 we saw that the matching topic in another model included both discussion of the Bush Presidency and Cheney. If the global alignment had assigned that topic to the Bush reference model topic, that would leave it unavailable for the Cheney reference model topic. This tends to manifest in the covariate effect distributions as clusters of certain covariate effect sizes. We still find the global alignment the most useful though as it ensures that we are not omitting any topics from the comparison models.

3.4.5 Additional Comparisons and Related Work

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

Different number of topics  We ran the above set of experiments under the same dataset with $K = 50$ topics and observed essentially the same patterns and trends reported. Smaller experiments at $K = 20$ reveal higher levels of instability across runs with increased instances of topics that are very poorly aligned. We conjecture that this is primarily a matter of how well the number of topics fit the specific dataset rather than a statement about small numbers of topics in general.\footnote{In Roberts et al. (2014) we examined a small open-ended survey response dataset with $K = 3$ and found results to be extremely stable even under a more demanding permutation test.} If instability was solely a function of the number of topics we would expect substantially poorer
performance in this extreme case. That the instability would be connected to selecting too few topics for a given dataset certainly makes intuitive sense, but additional investigation would be necessary to make conclusive statements.

**Alternative distance measures** In the results above we use two basic measures of distance between the topic-word distributions. We align the topics using a dot product measure and we presented calculations based on $L_1$ distance. We also performed experiments using a cosine similarity metric (essentially the dot product rescaled by the $L_2$ norm of the vectors). The results, depicted in Figure 3.8 show slightly less clear correlations between the similarity metric and the top words and top documents measure. Specifically there are many cases with high cosine topic appears with a comparatively low number of top words or documents in common. Manual examination of topics in these settings demonstrated that this was primarily connected with topics where the majority of the probability mass loaded onto fewer than 10 words.\footnote{Koltcov, Koltsova and Nikolenko (2014) in a similar investigation of stability in LDA, guard against the possibility of $L_1$ style calculations being dominated by the long-tail of infrequently occurring words. To guard against this we tested a version where we only calculated the distance over the minimal set of words accounting for 75% of a topic’s probability mass within the reference model. The results are substantially the same but with slightly less noise. We opted to maintain the versions we presented above to allow for simpler interpretation.}

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

**Alternative Approaches** The similarity metrics described here are automated approximations to semantic similarity. All of the metrics equally penalize deviations from the reference model regardless of whether it is in the direction of a

\footnote{Chuang et al. (2013) presented a number of different distance metrics (e.g., testing KL divergence, cosine metric and Spearman rank coefficient) against human judgments of similarity. They find that the cosine metric most directly matches human judgment and that it could even be further improved using a rescaled dot product measure which they introduced. The strong findings for the cosine metric provide an interesting contrast to Figure 3.8 and suggest that it may be perform better in other circumstances.}

\hfill

Figure 3.8: Comparison of metric based on cosine similarity.
One solution would be to embed words within a vector space such that semantically related words are close together and then calculate differences relative to this space (Mikolov et al. 2013). This has the advantage of more sharply penalizing differences between topics that involve words which are semantically unrelated. However, in order to perform the word embeddings we need an extremely large text corpus which limits the applicability to smaller document settings.\footnote{An alternate strategy is to cast the notion of distance between topics entirely in the realm of human judgments. This is essentially the approach of Grimmer and King (2011) which offers experimental protocols for evaluating similarity between topics.}

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

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

### 3.5 Initialization

When the function we are optimizing is well-behaved and globally concave, any starting point will result in the same global solution. Thus initialization of the parameters becomes a trivial detail, possibly chosen to save on computational costs.\footnote{We clarify well-behaved because in practice even globally convex problems can be sensitive to starting values due to practical issues in numerical optimization.} In the multimodal setting, our initialization influences our final solution. When the computational cost of inference in the model is extremely low, we can simply randomly initialize the parameters and repeat until we have identified the same maximum several times. However, in latent variable models not only may we never encounter a repeat solution, but each solution to the model may be very computationally expensive, a problem which is exacerbated in big data settings. If fitting a topic model on a million documents takes a week of computational time, rerunning it a thousand different times is not a reasonable strategy. A well-known but little-discussed aspect of statistical optimization is that careful initialization can be an incredibly powerful tool (McLachlan and Peel 2004; Murphy 2012,
Before returning to the case of topic models, we consider the simpler case of $k$-means, a central algorithm in the clustering literature closely related to the Normal mixture model discussed in Section 3.2.2. The $k$-means example helps to provide some intuition about the role of “smart” initialization. In Section 3.5.2, we return to the case of topic models and discuss how simpler models such as LDA can be used to initialize more complex models such as STM. In Section 3.5.3, we provide a simulation study which shows that the LDA based initialization yields higher values of the approximate evidence lower bound than random initialization.

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

An alternative approach is to explore deterministic approaches to initialization. In Section 3.6 we will outline very recent research which yields deterministic initializations with excellent performance.

### 3.5.1 $k$-means

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

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

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

---

23By easily parallelized, we mean that it can be easily fit into the Map-Reduce paradigm (Dean and Ghemawat 2008). The algorithm is still serial in the iterations, but the expensive calculations within each iteration can be performed in parallel.
Chapter 3. Navigating the Local Modes of Big Data: The Case of Topic Models

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

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

3.5.2 What makes a good initialization?

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

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

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

In the case of STM, there is a natural simpler model, LDA. Due to the Dirichlet-Multinomial conjugacy in LDA we can perform inference using a fast collapsed Gibbs sampler (Griffiths and Steyvers 2004). They key here is that the conjugacy of the model allows for all parameters except the token-level topic latent variables to be integrated out. The result is a very fast sampler which has been heavily optimized (Yao, Mimno and McCallum 2009). The cost of
Chapter 3. Navigating the Local Modes of Big Data: The Case of Topic Models

Inference is linear in the number of individual words (tokens) in the text. Because LDA is itself multimodal the result is an initialization which is different each time. Thus like k-means++ this approach places STM in a good region of the parameter space but still allows for variation across runs. The initialization for the LDA algorithm itself is just a random assignment of the tokens, so we don’t have a problem of infinite regress.

### 3.5.3 The effects of initialization

Unlike the case of k-means++ we cannot make theoretical guarantees on the quality of LDA as a method for initializing STM. This naturally leads us to ask about how it performs as an initialization in practice. To investigate this issue we compared the objective function values in the 685 model runs initialized with LDA to a set of 50 runs initialized from random starting values. Figure 3.9 plots the resulting distributions over the final level of the objective function.

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

### 3.6 Global Solutions

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

---

24 Also crucially the collapsed sampler mixes dramatically faster than an uncollapsed version (Carpenter 2010; Asuncion Jr 2011). By integrating out the topic-word distribution $\beta$ we are implicitly updating the global parameters every time we take a new sample at the document level. As a result we only need a few passes through the data to reach a good region of the parameter space.

25 Such a theoretical analysis is likely possible under a certain set of assumptions but would lead to a lengthy and technical digression here.

26 Specifically we initialize topic-word distributions with random draws from a Dirichlet distribution and set the document-topic proportion prior mean to zero. This is the commonly used initialization procedure in many variational algorithms for LDA.
optimization of the model may be so costly that we want to strictly limit the number of times we run the model.

In this section we introduce recent innovations in theoretical computer science which allow for global optimization of non-convex models using spectral learning. As we will show these algorithms introduce additional assumptions into the model in order to achieve tractable inference with provable guarantees of recovering the globally optimal parameters. Following the logic of Section 3.5, we use an algorithm for LDA as an initialization to the STM. Our results suggest that this hybrid strategy can be a useful technique for tackling big data problems.

We remind the reader that these techniques are very much “on the frontier” and so the substantive implications for applied projects have not been charted out, something that is beyond the scope of this chapter. Furthermore, we emphasize that these initialization strategies do not “solve” the multimodality problem. These techniques do not yield a correct answer, and even though they do very well at maximizing the approximate evidence lower bound, this does not mean the solution is optimal with respect to other criteria (as discussed above). The types of robustness exercises

Figure 3.9: A comparison of initialization strategies for the $K = 100$ STM models.
discussed above should continue to be an important part of the research process. Nevertheless, we find that these deterministic initialization procedures are a promising contribution to the topic modeling toolkit.

3.6.1 Introduction to Spectral learning

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

Here we describe a class of spectral learning algorithms for estimating the parameters of latent variable models while retaining guarantees of globally optimal convergence. The key insight is that by using matrix (or array) decomposition techniques we can recover the parameters from low order moments of the data. This approach relies on a method of moments inferential framework, as opposed to the likelihood based framework we have adopted thus far (Pearson 1894; King 1998; Anandkumar, Ge, Hsu, Kakade and Telgarsky 2012). In models with certain structures this can lead to procedures with provable theoretical guarantees of recovering the true parameters as well as algorithms which are naturally scalable.

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

The discussion of these methods is unavoidably more technical than the previous material. However, the common theme is straightforward: we are making stronger assumptions about the model in order to obtain an algorithm that does not suffer from problems of local modes. Importantly for our case we use the spectral algorithm as an initial-

---

27Spectral methods derive their name from the use of tools from linear algebra which are connected to the spectral theorem. Here we use an inclusive definition of spectral learning which includes methods using a variety of matrix and array decomposition techniques beyond the canonical Singular Value Decomposition.
ization rather than as a procedure to fit the model. In doing so we weaken our reliance on the assumptions in the spectral algorithm while still achieving its desirable properties. In this sense the spectral learning algorithms are complementary to the likelihood based approach we have considered here (Anandkumar, Ge, Hsu, Kakade and Telgarsky 2012).

### 3.6.2 An Algorithm for LDA

Here we briefly describe the intuition behind the inference algorithm of Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu and Zhu (2013) which uses a non-negative matrix factorization (NMF)\(^\text{28}\) to recover the model parameters from the word co-occurrence matrix, as we show below, to separate the \(\beta\) parameter (the topic distributions) from the data. The main input to the algorithm is a matrix of word-word co-occurrences which is of size \(V\)-by-\(V\) where \(V\) is the number of the words in the vocabulary. Normalizing this matrix so all entries sum to 1, we get the matrix \(Q\). If we assume that \(Q\) is constructed from an infinite number of documents then it is the second order moment matrix and the element \(Q_{i,j}\) has the interpretation as the probability of observing word \(i\) and word \(j\) in the same document. We can write the \(Q\) matrix in terms of the model parameters as,

\[
Q = \mathbb{E} [\theta^\top \theta^\top \theta^\top \theta^\top \beta^\top]
\]

\[
= \beta^\top \mathbb{E} [\theta^\top \theta] \beta
\]

where the second line follows by treating the parameters as fixed but unknown. Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu and Zhu (2013) show that we can recover \(\beta^\top\) from the rest of the parameters using a non-negative Matrix Factorization.

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

\(^{28}\)NMF is similar to a Singular Value Decomposition except that all elements of the decomposition are constrained to be non-negative.
Halpern, Mimno, Moitra, Sontag, Wu and Zhu 2013). Thus if we can identify the anchors, we can solve for $\beta$ using convex optimization methods.

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

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

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

The second major concern is that spectral methods lean more heavily on the model assumptions which can lead to somewhat less interpretable models in real data (Nguyen, Hu and Boyd-Graber 2014). Finally, as a practical matter the spectral method only recovers the topic word distributions $\beta$ so additional methods are still required to infer

---

29 Anchor selection methods use either a sparse regression framework (Recht et al. 2012) or appeal to geometric properties of the anchors (Kumar, Sindhwani and Kambadur 2012). See Gillis (2014) for a summary of these approaches. For our experiments here we focus the approach defined in Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu and Zhu (2013) which falls into the geometric properties camp. They use a combinatorial search based on a modified Gram Schmidt orthogonalization process for the anchor selection. Parameter recovery then uses an exponentiated gradient descent algorithm (Kivinen and Warmuth 1997) with an $L_2$ norm loss.

30 A good example is the mixed membership stochastic blockmodel, which is loosely speaking LDA for community detection on a network (Airoldi et al. 2009). Huang et al. (2013) give a spectral algorithm which learns hundreds of communities in a network of millions of nodes in under 10 minutes.
the document-topic proportions. These can be obtained by a single pass of Gibbs sampling or variational inference (Roberts et al. 2014).

### 3.6.3 Spectral Learning as Initialization

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

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

![Comparing Initialization Strategies](image)

**Figure 3.10:** A comparison of the spectral initialization strategy to random and LDA for the $K = 100$ STM models. The green dashed denotes the result of the spectral initialized solution.
Figure 3.10 shows the results of the model with the spectral initialization. Not only is the result dramatically better with respect to the lower bound than the random and LDA initializations but the model converged considerably faster as well.\footnote{It took 25 iterations to converge after the spectral initialization compared to 60 iterations for LDA initialization and close to 200 iterations for random initialization.} Because our focus here is on introducing this class of algorithms we do not go through the process of reinterpreting the 100 topics model.

### 3.6.4 Future Directions

Spectral algorithms are a very active area of current research. Here we have focused on a particular algorithm which leverages nonnegative matrix factorization under a separability assumption. There have been several algorithmic improvements since Arora, Ge, Kannan and Moitra (2012) introduced the anchor based method (Recht et al. 2012; Kumar, Sindhwani and Kambadur 2012; Ding, Rohban, Ishwar and Saligrama 2013; Gillis and Luce 2013; Gillis 2014; Zhou, Bilmes and Guestrin 2014). There has also been substantial work applying the approach to other problem domains (Arora, Ge, Moitra and Sachdeva 2012; Arora, Ge and Moitra 2013; Arora, Bhaskara, Ge and Ma 2013; Zhou, Bilmes and Guestrin 2014).

A separate line of work uses higher order moments of the data along with tools for array (tensor) decomposition (Anandkumar, Ge, Hsu, Kakade and Telgarsky 2012). These methods have also resulted in algorithms for an incredibly rich set of applications and models. Importantly we can also use this framework to develop algorithms for LDA with provable global convergence guarantees (Anandkumar, Liu, Hsu, Foster and Kakade 2012; Anandkumar, Hsu, Javanmard and Kakade 2013).\footnote{Technically the work in Anandkumar, Liu, Hsu, Foster and Kakade (2012) uses an approach called Excess Correlation Analysis which involves two singular value decompositions on the second and third moments of the data. The approach based on the tensor method of moments strategy is described in Anandkumar, Ge, Hsu, Kakade and Telgarsky (2012) and applies to a wider class of models. We collect them together here because they emerged from the same research group and use similar techniques.} This work differs in both the assumptions and methods used. Crucially the tensor method of moments approach uses the third moments of the data which may require an even higher sample size to accurately estimate.\footnote{An excellent discussion of differing assumptions of spectral methods is given in Ding, Ishwar, Rohban and Saligrama (2013).}

### 3.7 Conclusion

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

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

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

We see great potential for the analysis of “big” data in the social sciences, but rather than focus on the data we have taken a more methodological focus. We think this has important implications for both methodological development but also could structure the types of questions we ask, and the types of data sets we seek to build. Methodologically, we think that there will be important advances in areas such as optimal initialization strategies, which is especially important as our data sets grow in size. From an applied perspective, users will be unlikely to want to wait for extended periods of time in order to get even a single set of results. Advances in computational power needs to be matched with smart ways to leverage that power. From a research design perspective, we think more focus should be put on bringing greater structure to so called “unstructured” data. In the STM we focus on the inclusion of metadata for modeling and hypothesis testing, but this is only one possible use. Can more direct supervision help us with issues of multimodality? Of course, in the end, big data will be at its best when there is active dialogue between those who pose the big question and those who might provide the big answers.
Chapter 4

Latent Factor Regressions for the Social Sciences

4.1 Introduction

Most datasets analyzed in the social sciences have some form of structure. Whether such data includes information about countries over time, students in a classroom, or friends in a network, the reality of social research is that our units of observation are often deeply interconnected. From a theoretical perspective these connections are frequently the primary quantity of interest (Fowler and Christakis 2009; Maoz 2011). From a methodological perspective, the structure of the data complicates the standard statistical toolkit and threatens our ability to empirically test hypotheses.

Regression, in particular the generalized linear model (GLM), plays a central role in the social sciences as the default statistical method for the analysis of relationships in quantitative data. GLMs leverage the assumption that observations are conditionally independent given the covariates in order to allow for tractable inference. Methodologists have periodically warned of the inaccuracy of these standard regression tools in the presence of unmodeled dependence between units.¹ Common concerns within political science are temporal dependence (Beck and Katz 1995), spatial dependence (Franzese Jr and Hays 2007) and network dependence (Hoff and Ward 2004). Analogous contributions exist in other social sciences such as economics (Wooldridge 2010) and sociology (Snijders and Bosker 1999).

¹Even in just the last five years there have been a number of such articles in political science alone (Pang 2014; Erikson, Pinto and Rader 2014; Gaibulloev, Sandler and Sul 2014; King and Roberts 2014; Beck 2012; Bell and Jones 2012; Wawro and Katznelson 2013; Arceneaux and Nickerson 2009; Aronow and Samii 2013; Park 2012; Pang 2010; Beck and Katz 2011; Dorff and Ward 2013).
Previous approaches have addressed a single form of dependence at a time, often with solutions which are mutually incompatible. Here I provide a unifying characterization of these problems which leads naturally to a single solution.

One way to think about dependence is as arising due to unobserved heterogeneity between repeated units within the data. Thus if we had the right set of control variables, we could treat the remaining stochastic error as independent across observations. Subject matter experts often have an implicit understanding of unmodeled dependence and are able to specify the important groups within the data. I refer to these natural groups as the “modes” of dependence. For example, in the analysis of country-year data, it is difficult to justify the claim that there is a set of variables which eliminate country-level correlation in the residuals. However, even when the analyst believes that there is unmodeled cross-sectional heterogeneity, it can be difficult to translate that belief into a practical modeling choice. The technical literature on the subject is vast, spanning numerous related fields, and yields often contradictory instructions. Furthermore many of the existing methods are customized to a particular type of dependence and can be computationally infeasible in the applied setting. Unfortunately most statistical problems arising from unobserved heterogeneity will not vanish asymptotically as the size of our data increases. Indeed addressing the problems of dependence between units and data set structure has been identified as a particularly pressing issue in the era of “big data” (Council 2013).

When observations are organized along a single partition\(^2\) or “mode”(e.g. the country or year), the analyst can model heterogeneity by replacing a constant parameter with a set of parameters that vary by group. When the constant intercept is replaced with a group specific intercept, this results in the familiar “fixed effects” model (Angrist and Pischke 2008). The broader set of methods for analyzing grouped data are often referred to as multilevel or hierarchical modeling (Gelman and Hill 2007). When the data are structured by multiple cross-cutting modes (such as dependence between observations in the same year, and observations within the same country) the problem becomes substantially more challenging. Existing solutions can model the two modes additively, but this often fails to capture important facets of the data generating process.

In this article, I present a unified framework that allows for multiple interactive forms of structure using interactive latent factors. The modeling framework has three principal advantages: (1) it models a wider variety of dependence types than previous approaches (which are subsumed in this framework), (2) it is less demanding on the data than previous approaches, and (3) inference is sufficiently fast to be practical for applied use. The core idea of modeling complex structured data using latent factor models has been repeatedly and independently reinvented across the social and natural sciences. However, the approaches invented across different fields have primarily been developed in

\(^2\)By “partition” I mean a mutually exclusive and exhaustive grouping of the observations.
isolation. To the best of my knowledge there has been no unifying treatment that connects the similar approaches across fields. The framework here combines the best features of these disparate approaches and is coupled with new inference algorithms which will be made available in a forthcoming R package.

In Section 4.2 I describe the general problem including a motivating example from international relations, a sub-field of political science; but as will become clear, the implications of this paper span other social sciences. In Section 4.3, I outline the framework for latent factor regression. In Section 4.4 I develop an estimation framework for latent factor regression based on variational inference. Section 4.5 connects my approach to diverse bodies of work, highlighting connections between previous models. Before proceeding to real data, Section 4.6 provides an overview of simulation evidence demonstrating the effectiveness of the inference framework. Section 4.7 illustrates the proposed method with applications that have been the focus of methodological debates within international relations, but whose features nevertheless extend to a broad set of social science fields. Finally, Section 4.8 concludes with a short discussion of directions for future research.

4.2 Regression with Unobserved Heterogeneity

The generalized linear model (GLM) is the standard regression model in quantitative political science (King 1998; McCullagh and Nelder 1989). The starting point for the GLM is the assumption of independence of the observations ($y$) conditional on the covariates ($X$). Indexing the observations as $d \in \{1 \ldots D\}$, we can write the model generically as:

$$y_d \sim f_y(\theta_d, \phi)$$  \hspace{1cm} (4.1)

$$\theta_d = g^{-1}(\eta_d)$$  \hspace{1cm} (4.2)

$$\eta_d = \alpha + X_d \beta$$  \hspace{1cm} (4.3)

where $f_y(\cdot)$ is the exponential family probability density, $g(\cdot)$ is the link function, $\eta_d$ is the linear predictor for observation $d$, $\beta$ are the regression coefficients, $\alpha$ is the intercept and $\phi$ collects incidental parameters. This encompasses nearly all the regression models used in political science include OLS, logit, probit, poisson and negative binomial regression. Due to the conditional independence assumption, the likelihood can be expressed as a product over the density $f_y(\cdot)$. While mathematically convenient, this assumption may not be plausible when the data are grouped. That is, there is some heterogeneity within the data not captured within the covariates $X$ which threaten the conditional
Chapter 4. Latent Factor Regressions for the Social Sciences

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Building Blocks</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit</td>
<td>the level of analysis interesting to a researcher</td>
<td>country-year, edge in a network</td>
</tr>
<tr>
<td>Mode</td>
<td>an observed partition over observations</td>
<td>time, cross-section, sender, receiver</td>
</tr>
<tr>
<td>Group</td>
<td>the members of a subset of a partition</td>
<td>individual years (time), individual countries (cross-section)</td>
</tr>
<tr>
<td>Data Structure</td>
<td>covariance across units in the dependent variable not captured by the covariates</td>
<td>time-series cross-sectional data, network data, multivariate outcome data</td>
</tr>
</tbody>
</table>

| Model            |                                                                            |                                              |
| Single Mode      | a model where only one mode of dependence is modeled                      | fixed effects, random effects                |
| Additive Modes   | a model of mode effects which enter the linear predictor additively        | two-way effects such as time and country intercepts |
| Jointly Unique   | a model of mode effects where each combination of groups across modes is estimated separately | a country-year specific intercept            |

Table 4.1: Definitions and examples for terms used throughout the paper.

Most methodological approaches for modeling heterogeneity are variants on one simple idea: taking a set of constant parameters, and allow them to vary with some observed partition of the data. I call this observed partition a “mode” of the data and the collection of units sharing a parameter a “group”. For example, for a dataset where observations are repeated within years, time is a “mode” of the data and the observations within the same year form a “group.” Table 4.1 provides a reference for these terms as well as other terms used throughout the paper.

Approaches to modeling heterogeneity can be differentiated along two dimensions. The first is how the groups within a given mode are related to one another. For example, time is naturally ordered and the analyst may want to impose some smoothness such that neighboring years are estimated to have similar parameters. Groups of other modes may not be naturally ordered, such as countries in the international system. The second dimension captures how each mode is related to other modes. This can be intuitively thought of as a unit being a member of multiple groups. Thus in a time-series cross-sectional dataset, every observation is both a member of time point and a member of a cross-section. The way these mode effects interact is the second dimension. The simplest type of interaction is an additive model, such as in two-way fixed effects where, for example, the effect for a particularly country-year is the
country effect plus the year effect. The other extreme of mode interaction is a model where effects are jointly unique. This corresponds to estimating a country-specific effect for each year.

The preceding literature has mostly been concerned with different approaches to capturing how groups are related to each other rather than modeling multiple modes. When multiple modes are modeled it is often in the context of extreme views where the mode effects are additive or completely jointly unique. To see how this extreme view poses a problem for applied work, I turn to a debate in the applied international relations literature.

4.2.1 ‘The Dirty Pool’ Debate

The Spring 2001 issue of International Organization, a prominent international relations journal, contained a symposium on pooled estimators within international relations. The symposium contained an introduction by Gourevitch and Lake (2001), the main article entitled ‘Dirty Pool’ by Green, Kim and Yoon (2001) (hereafter GKY), as well as replies by Beck and Katz (2001), Oneal and Russett (2001) and a summary by King (2001). The central argument of GKY is that by ignoring unobserved heterogeneity in cross-sectional data, findings in quantitative international relations are biased.3 They argue for the inclusion of dyad-level varying intercepts (fixed effects). They demonstrate that including these terms results in democracy being negatively related to trade and unconnected to militarized interstate disputes. These findings if true would undermine enormous portions of the international relations literature.

The replies took a staunchly different approach. Oneal and Russett (2001) demonstrated that the findings are robust under a series of alternate specification and emphasizes the role of the shorter time period in the GKY data in generating the original result. Beck and Katz (2001) argue that for the typical setting of international relations data analysis the proposed solution is “profoundly misleading in assessing the impacts of important independent variables” (Beck and Katz 2001, p.488).

The King (2001) summary of the debate emphasized the central contribution of GKY as identifying the “complex dependence structures” in international relations and how those structures contribute to unmeasured heterogeneity. In total the methodological evaluation of the symposium is somewhat grim. All participants agree that unobserved heterogeneity is a consequential issue, but there is little in the way of clear solutions. King concludes his methodological assessment by simply stating “Getting better data is usually the best advice, and it clearly is here” (King 2001, pg. 504). While better data can obviate the need for more complex methods, there will always be an opportunity

3In a two-page long table, GKY extensively detail the quantitative analyses of international dyads undertaken within the preceding 3-year period in 10 of the top journals. This totaled 51 articles of which nearly all used pooled estimators. These patterns have not changed substantially in recent years.
cost in data collection. It is difficult to advocate that an applied researcher bears these costs to address unobserved heterogeneity which by definition is not the quantity of primary theoretical interest.

This paper provides a methodological framework for applied scholars that goes well beyond the current debate. However, the key ideas from this debate have repeatedly surfaced in the applied literature across a wide variety of fields. Thus the debate serves as a useful concrete example of a relevant application.

4.2.2 Modeling Heterogeneity

GKY is concerned with a particular type of cross-sectional heterogeneity. Specifically they argue that there is a single relevant mode in the data, the dyad pair, which affects only the intercept. In the analysis of conflict data this has the theoretical intuitive description: each dyad has a different ex-ante probability of conflict but the contribution of each covariate to the linear predictor is constant across all cross-sections.\(^4\) Letting \(d\) index the dyad, this results in a linear predictor which can be written as:

\[
\eta_d = \alpha_d + X_d \beta
\]  

(4.4)

where \(\alpha\) is the now-dyad specific intercept term but \(\beta\) remains constant.

The central problem with the GKY approach is that it is too demanding on the data. Because every pair of countries is given a separate parameter, we require repeated observations of that dyad which can only be acquired through time. We can instead describe the data as containing two modes, one which indexes the source of the action and one which indexes the receiver. To visualize this, imagine that we collected all the intercept terms into \(A\), a square, symmetric matrix where the dimension is the number of countries (\(N\)). Entry \(A_{i,j}\) is the intercept for the dyad containing country \(i\) and country \(j\). The completely pooled estimator approximates this matrix with a single number, that is the intercept for every dyad is the same. The GKY model in Equation 4.4 models each cell in \(A\) as a separate parameter and thus treats the two modes as \textit{jointly unique}. This means that the estimate of the probability of war between country \(i\) and country \(j\) offers us no information for our estimate of the probability of war between country \(i\) and country \(l\). The implication is that the GKY model implicitly takes the epistemological position that nothing about the causes of peace can be learned from a dyad which has never gone to war.\(^5\)

\(^4\)The simplicity of this interpretation is slightly marred by the non-linearity of the link function. The real quantity of interest here is the probability of the outcome, not the change in the linear predictor. With a non-linear link function a shift in the intercept changes the effect of the covariates on the scale of the predicted probabilities.

\(^5\)We can of course use a model without believing that all the assumptions are true. However it is worth emphasizing that dyads which have never gone to war are dropped from the dataset under the GKY model and thus cannot contribute to our estimates of the effects.
Chapter 4. Latent Factor Regressions for the Social Sciences

An intermediate between these approaches is a two-way varying intercepts approach in which the two modes are modeled additively. Now the dyad’s intercept is the sum of the component countries,

\[ j_{i,j} = \alpha_i + \alpha_j + X_{ij}\beta. \]

This model approximates the matrix \( A \) by its margins (i.e. a column effect plus a row effect). In the conflict setting this would intuitively capture how bellicose country \( i \) is, but will not distinguish whether it is more belligerent towards any particular country. The advantage of the two-way additive model is that it places fewer demands on the data by allowing more observations to inform each parameter. In estimating the intercept for dyad \( i, j \) the model draws information from all the dyads of which countries \( i \) and \( j \) are members. This means that the model can be identified even if we only observe a single year of data for each dyad.

The additive model is extremely limited in the type of relationships it can capture. We can see this by noting that only a small class of parameter matrices \( A \) could be represented by column and row effects. The GKY model by contrast estimates every cell of \( A \) completely separately which may neglect important structure in the parameters. An intermediate between these two extremes is accessible through the idea of a low-rank approximation. The central idea is to replace the complete parameter matrix \( A \) with a low-rank approximation \( \tilde{A} \) which we can estimate using fewer parameters.

The key to low-rank approximation is that a low-rank matrix can be formed through the matrix multiplication of two smaller matrices (see Figure 4.1). This form encompasses a much wider range of matrices (and hence parameters) than a simple additive model. Political scientists may be most familiar with the idea of a low-rank approximation in legislative ideal points (Clinton, Jackman and Rivers 2004) which seeks to describe the legislator by bills matrix of votes with the product of (usually) \( K = 2 \) dimensional matrices representing the legislators’ ideal points and the bills’ ideal points. We can write the model in vector notation as:

\[ j_{i,j} = \alpha_i + \alpha_j + \left( \sum_k u_{i,k}v_{j,k} \right) + X_{ij}\beta. \] (4.5)

where \( k = 1 \ldots K \) is the rank of the approximating matrix and \( U \) is the \( I \)-by-\( K \) factor matrix and \( V \) is the \( J \)-by-\( K \) factor matrix. Here the intercepts capturing the row and column margins are included as separate parameters but could easily be absorbed into the latent factor matrix. Crucially the matrices \( U \) and \( V \) are unobserved (as, for example,

\[ \text{Note that } K \text{ determines the quality of the approximation to the unrestricted matrix } A. \text{ As } K = N \text{ we get an exact reconstruction of the matrix (Eckart and Young 1936).} \]
are legislative ideal points); however, they involve fewer parameters than the GKY strategy which involves estimating every element of $A$ separately.\footnote{It is worth emphasizing that $U$ and $V$ are not identified in the formulation here without further constraints or a prior distribution due to a rotational invariance in the posterior (West 2003). Below I will use prior distributions to essentially make an arbitrary choice of a rotation of $U$ and $V$. This is not problematic as our interest is in the inner product $UV^T$ which is identified.}

In the low-rank model the latent effects for dyad $i$ and dyad $j$ enter the model through an inner product (i.e. a multiplication of the parameters) and consequently these models are often described as interactive effects models (in contrast to the additive fixed or random effects models). Interactive latent effects will serve as the basis for the framework I develop in this paper. Models of this sort have previously been used in political science for special cases such as the analysis of fixed rank network data most notably by Hoff and Ward (2004). I defer a more comprehensive survey of the related work and the differences with my framework here until Section 4.5.

The approach advocated by GKY and the subsequent discussion focuses on varying intercept models estimated as “fixed effects.” This places their work in line with a well-developed literature on panel data methods in econometrics (Angrist and Pischke 2008; Wooldridge 2010; Greene 2012). These approaches also straightforwardly allow for heterogeneity in the covariate effects. When the analyst specifies a probabilistic model for the varying parameters the result is often called a multilevel model (Western 1998; Gelman and Hill 2007; Gelman et al. 2013; Snijders and Bosker 1999). The extension of varying slope and varying intercept models to the GLM setting go by the moniker Generalized Linear Mixed effects Models (GLMM) and are heavily used in a wide variety of fields such as epidemiology, bio-statistics, sociology and economics (Breslow and Clayton 1993).\footnote{The “mixed” reference here alludes to the idea that some coefficients are “fixed” in the sense of being shared across the entirety of the data while others are “random” in that they can vary by subgroup.} When the data is structured along a single mode this class of models can be quite effective at recovering the effect of interest. Recent advances in Bayesian statistics, particularly the development of the Stan software library for Hamiltonian Monte Carlo (Stan Development Team 2014; Hoffman and Gelman 2013; Neal 2011), have made these models straightforward to design and fit.

---

**Figure 4.1:** A low-rank matrix $\tilde{A}$ formed from the product of two smaller matrices.
Chapter 4. Latent Factor Regressions for the Social Sciences

However as the ‘Dirty Pool’ case illustrates, the dependence structures that characterize data in the social sciences are often significantly more complex than single mode. The latent factor framework I present offers a general solution to these problems and provides a unifying approach to modeling data types as diverse as time-series cross sectional, network, and spatial data.

4.3 Regression with latent factors

In this section I show how interactive latent factors can be incorporated within a broad class of generalized linear models. This allows us to extend beyond the non-interactive single mode models such as varying intercept fixed/random effect models which are the predominant applied approach to modeling heterogeneity. By simply allowing for interaction in the latent factors, the framework can recover an enormous range of models from ideal point models of roll call voting (Clinton, Jackman and Rivers 2004) to latent space network models (Hoff and Ward 2004). The central idea is to replace a constant parameter $\beta$ in a generalized linear model with a group-specific term formed by

$$
\beta_{i,j,...,l} = a + \sum_{m=1}^{M} b_m + \sum_{k=1}^{K} u^{(1)}_{i,k} u^{(2)}_{j,k} \cdots u^{(m)}_{l,k}
$$

(4.6)

where $a$ is a globally shared effect, $b_m$ are the additive mode specific effects, $u^{(m)}_{i,k}$ is the latent variable for the $i$th group of mode $m$, and $K$ is the dimensionality of our latent variable. The dimensionality of the latent variable controls the models ability to represent more complex interactions between the modes of the data. Note that the limiting case of $k = 0$ results in the two-way effects model.$^9$ With no modes, the model collapses to the completely pooled estimator having a single globally shared parameter.

I take a Bayesian approach to modeling the latent factors, giving each latent effect a normal prior. In the form of generalized linear model from Equation 4.1 this yields the following form

$$
\eta = X \left( U^{(1)} \times \cdots \times U^{(M)} \right)
$$

(4.7)

$$
U^{(m)}_{i,k} \sim \text{Normal}(0, \Sigma)
$$

(4.8)

$$
\Sigma \sim p(\cdot)
$$

(4.9)

$^9$By two-way effects I mean additive varying effects. So in the context of a varying intercept model for two modes time and cross section, we would get $\beta_{\text{time}, \text{country}} = b_{\text{time}} + b_{\text{country}}$.
Chapter 4. Latent Factor Regressions for the Social Sciences

<table>
<thead>
<tr>
<th>Modes</th>
<th>Group Structure</th>
<th>Model/Citation</th>
<th>This Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Mode</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>panel</td>
<td>none</td>
<td>fixed effects (Angrist and Pischke 2008)</td>
<td>✓</td>
</tr>
<tr>
<td>panel</td>
<td>common distribution</td>
<td>random effects / multilevel models (Gelman and Hill 2007)</td>
<td>✓</td>
</tr>
<tr>
<td>time</td>
<td>random walk</td>
<td>dynamic linear models (West and Harrison 1997; Martin and Quinn 2002)</td>
<td>✓</td>
</tr>
<tr>
<td>geography</td>
<td>spatial auto-regressive</td>
<td>spatial regressions (Gleditsch and Ward 2008)</td>
<td>✓</td>
</tr>
<tr>
<td>Two Mode</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>time and panel</td>
<td>none</td>
<td>interactive fixed effects (Bai 2009)</td>
<td>?/✓†</td>
</tr>
<tr>
<td>source, receiver</td>
<td>common distribution</td>
<td>generalized bilinear mixed effects model (Hoff and Ward 2004; Hoff 2005)</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 4.2: Example models and their relationship to the framework presented here. “Common distribution” indicates that the parameters are drawn from a shared prior and thus exhibits partial pooling. Models within each number of modes are ordered in increasing complexity of the group structure. † Interactive fixed effects use no prior distributions on the latent factors. These can be estimated in the current framework but require stronger assumptions for identification.

where the intercept has been absorbed into the covariate matrix $X$. The latent factors are given zero-mean Gaussian priors with variance controlled through the covariance matrix $\Sigma$. By changing the prior distribution for $\Sigma$, the model can capture different types of group structure and perform dimensionality selection for the rank of the latent factors. When using the Bayesian approach, it is necessary to make the standard random effects assumption of independence between the effects on observed covariates and the latent factors. I discuss this issue in more detail in the section related work(4.5.1) and explore the sensitivity of the model to violations of this assumption in the section on simulation (4.6).

This formulation encompasses an extremely wide range of models (Table 4.2 gives some examples). As I will show the number of latent interactions $M$ corresponds to the modes of the data that the analyst wishes to model. In the next section, I start with the familiar case of modeling a single mode structure and show how the construction naturally generalizes to two dimensions (matrices), three dimensions (arrays) and arbitrary $M$-dimensional data. Then in Section 4.3.2 I discuss how the latent factors $U^{(m)}$ have been substantively interpreted across a few of the diverse fields where they have been applied. In Section 4.3.3 I show how different prior distributions for $\Sigma$ lead to different models for how groups within a mode are related. By modeling the relations between groups in particular ways, the framework can mimic a broad class of spatial and time series methods.
4.3.1 Interactive Modes with Multilinear Latent Factors

In the simplest data analysis setting, the observations are treated as completely independent, resulting in the standard pooled regression estimator. While this is an important basic model it is severely limited by the assumption that effects are constant across modes of the data. Even in the case where the analyst’s interest is in the average population effect of a particular variable, accurately accounting for heterogeneity in other portions of the model can be vital for accurately estimating the effect. (Angrist and Pischke 2008). In the rest of this section I show how the model changes with the addition of each mode, moving from the single mode case to an arbitrary \( M \) mode setting.

Single Mode Setting The most familiar single mode model is the varying intercept “fixed/random effects” model where each group within the single mode of the data is given a separate intercept term. These models are attractive because they are easy to estimate and interpret. With only one mode there is no equivalent of rank and thus no need to infer the dimensionality of the latent effect. The available methods for the single mode settings are well described by existing textbooks on multilevel and longitudinal modeling and are heavily used throughout the social sciences (Snijders and Bosker 1999; Gelman and Hill 2007). Although the single mode setting does not make use of the interactive effects structure we describe here, the fast inference algorithms developed in Section 4.4 apply to this setting and provide a computationally attractive estimation alternative in large data settings.

Two Mode Setting Introducing a second mode into the model requires the analyst to make a choice about how mode effects interact with each other. Imagine for example that the data are time-series cross-sectional with each observation indexed by a country and year. It is reasonable to believe that unobserved heterogeneity causes dependence in the outcome for each of the years within a country, and for each of the countries within a year. Furthermore it may be that the effects are interactive. Analogues in other social science disciplines are immediate.

To gain some intuition for this mathematically, imagine as in Section 4.2.2 arranging the outcome variable into a matrix where the rows index the countries and columns index the time. A model with additive country and time effects would estimate a parameter for each country (row) and each year (column). Then to get a particular country-year parameter we simply add the components together. This is equivalent to approximating the matrix of parameters by its margins. Substantively this means that the temporal effects of a particular year are experienced in the same way across all cross-sections, and the cross-section effects are experienced the same way across all time periods. While this will sometimes be plausible we often will want to mode cross-section specific temporal effects.

The model can capture interactive effects, if the analyst is willing to estimate the entire matrix of parameters.
However, since we generally only get one observation per cell of the matrix (i.e. each country-year combination is observed only once), it will be necessary to find an approximation. By assuming the matrix of parameters has a low-rank structure the complete matrix can be approximated as the product of two smaller matrices. Note that even if in the true model of the world the parameters do not follow this low-rank structure, the procedure will still return the best low-rank approximation of the truth.

The covariance of the outcome implied by the low-rank solution is limited compared to estimating every cell separately, but it is substantially less restrictive than estimating the effects additively. From a probabilistic perspective this low-rank model yields the matrix-variate normal distribution which is the extension of the multivariate normal to matrix data (Dawid 1981; Allen and Tibshirani 2012).

### Three Mode Setting

What happens when a third mode is introduced into the model? This frequently arises (for example) in longitudinal network data where each observation is an action and we observe source-receiver-time triples. The framework extends easily to this case as well. Now instead of arranging the parameters into a matrix, they can be arranged into a stack of matrices. This object is called a tensor (also sometimes called an array), and like matrices, a low-rank tensor can be represented as the product of smaller tensors (Kolda and Bader 2009). Although the tensors make the computation and notation substantially more complicated the construction of the parameter is simply the product of three latent variables as in:

$$\beta_{i,j,l} = \sum_k u_{i,k}^{(1)} u_{j,k}^{(2)} u_{l,k}^{(3)}$$

Many mathematical and computational aspects of tensor analysis are substantially more challenging than the matrix case. However, the tensor formulation allows us to extend the model to an arbitrary number of modes. This approach has proven useful in a wide variety of applications, such as constructing deep interaction priors (Volfovsky and Hoff 2012), modeling multivariate event counts (Hoff 2011a), and analyzing neural images (Zhou, Li and Zhu 2013). Fortunately the estimation strategy in this paper extends to these more complicated models, and hence provides

---

10It is useful to contrast the two mode setting with the difference-in-differences estimator common in econometrics (Angrist and Pischke 2008). For binary treatments Heckman, Ichimura and Todd (1997); Heckman et al. (1998) show that difference-in-differences can be interpreted as a matching estimator. Imai and Kim (2012) prove that that it is equivalent to weighted two-way fixed effects where the weighting helps to avoid treatment-control mismatches in comparison sets. The additive two mode model without priors would be equivalent to the unweighted two way fixed effects estimator. It is unclear whether the interactive model presented in this paper has a direct interpretation under this framework.

11For example there are two natural formulations of the tensor decomposition: the Tucker Decomposition (Tucker 1964; Hoff 2011b) and the CP/Parafac model (Harshman and Lundy 1994). The Tucker decomposition is more general has the natural natural interpretation of an array normal model for separable data (Hoff 2011b). The CP/Parafac is comparatively simpler and is a special case of the array normal model with a superdiagonal core array. In what follows I make use of the CP/Parafac form although all the described methods could be applied to the more general case.
a unified framework.

4.3.2 Interpretation of Interactive Modes

The basic model in Eq 4.6 has been proposed in a variety of different fields. The different substantive interpretations of the modes provide a helpful guide for understanding their potential role in the data analysis. Here I give a brief overview of different interpretations of the two-mode context as applied in three different fields: computer science, economics and network analysis. I note that none of these versions is more correct, only more natural in different contexts. Each case helps to explain how the latent factor model is able to model heterogeneity.

**Low Rank Approximations (Computer Science)** A relatively atheoretical framework is to see the latent factors as purely a statistical approximation. The true model may involve a matrix (or tensor) of different parameters and the goal is to select the best rank $k$ approximations to that object (Lim and Teh 2007; Zhao, Zhang and Cichocki 2014). This view is prevalent in, for example, the computer science literature on recommendation systems. If Netflix wants to make a recommendation to you about a particular movie that you haven’t seen, they use a low rank approximation of the user/rating matrix to make a best guess. Viewed in this way the interactive latent variables play a role similar to linear regression in finding a linear dimensionality reduction of the data. Instead of projecting the data onto the column space of the covariates, the goal is to find the best rank $k$ approximation (Hastie, Tibshirani and Friedman 2009; Cunningham and Ghahramani 2014).

**Common Shocks and Varying Response (Econometrics)** In the literature on panel data econometrics, latent factors are given a more explicit substantive interpretation in terms of time-series cross-sectional data (Bai 2009). Here the idea is to view the latent factor for time as capturing a common global shock and the latent factor for country as capturing the varying responses to those shocks. Countries with similar loadings have similar unobserved characteristics that cause them to respond similarly to a certain type of shock, but crucially each country may respond differently. In this sense the models are used to introduce a covariance structure amongst the outcomes and are often framed as an alternative to spatial models (Bai 2009; Pesaran, Shin and Smith 1999; Moon and Weidner 2010a; Pang 2014). Like the spatial models they are compared to the econometrics models often assume that the panels are balanced in the sense that a complete time series is observed for each cross-sectional unit.
A Latent Space (Networks)  In networks, a common interpretation originating from Hoff, Raftery and Handcock (2002), is to view the latent variables as defining a “social space” where nodes who are nearby in the space are more likely to have ties. In the model of Hoff, Raftery and Handcock (2002) the latent variables are explicitly parameterized as distance in this space, but we can conceptualize the interaction of the latent variables as defining an inner product space. Hoff (2005) shows that this inner product space captures attractive properties of third order dependence such as clustering or transitivity, allowing the model to encapsulate logic like a “friend of a friend is a friend.” This is the same interpretation given to ideal point scaling in political science when we speak of legislators and bills be projected into a low-dimensional common space (Clinton, Jackman and Rivers 2004; Martin and Quinn 2002; Poole and Rosenthal 1997).

One contribution of this paper is to demonstrate that these interpretations describe the same class of models even though they differ in their goals. In computer science, analysts are often primarily interested in prediction of unobserved elements of the matrix. By contrast many of the econometric and network models explicitly require that the data matrix be fully observed. In the econometrics case this requires that every country must be observed for the entire length of the time-series. In the network case this means that the edges between every pair of nodes must be observed.

In practice this means that the analyst can generally only use a tiny subset of the available data.

The goals lead to different practices of interpretation. The emphasis on prediction in the computer science context means that there is little interest in interpreting the latent dimensions. In the econometrics setting the goal is typically to adequately control for some unobserved heterogeneity to accurately estimate a different set of parameters. On the other extreme, for the networks literature there are typically not other parameters and interpretation of the latent factors is the sole quantity of interest. I discuss these approaches and their relations to the extant literature in more depth in Section 4.5.

4.3.3 Modeling Group Structure

In the preceding section I discussed interpretations for interactive modes of the data. In this case each observation has a membership in each mode and the joint effect of that membership is allowed to be more than the sum of its parts. Thus in the case of time-series cross-sectional data we are able to capture temporal shocks that do not affect all cross-sectional units in the same way.

In this section I show how the prior distributions for the latent factors can be used to model how the groups within a mode are related. This unifies the single mode model with a wide variety of time-series and spatial regression models.
Chapter 4. Latent Factor Regressions for the Social Sciences

including Gaussian processes (Rasmussen and Williams 2006), random effects (Fahrmeir and Lang 2001), dynamic linear models (West and Harrison 1997), stochastic volatility models (Chib, Nardari and Shephard 2002) and spatial autoregressive models (Besag, York and Mollié 1991; Held et al. 2005).

**Unordered Groups** When groups with in a mode have no natural ordering or the analyst does not wish to model the ordering, the central choice is the degree to which to pool the parameter estimates. Classical fixed effects use no pooling at all, each group uses only the observations within that group to estimate the group’s parameters (Angrist and Pischke 2008; Wooldridge 2010). By contrast the multilevel modeling literature uses partial pooling in which estimates are drawn to a common mean with the strength of pooling determined by the variance parameter (Gelman and Hill 2007). In the limit as the variance of the latent variables goes to zero, we force parameters across all groups to have the same value and recover the pooled regression estimator.

The general framework described above is able to support the broad range of options available in the literature on multilevel and longitudinal modeling (Gelman and Hill 2007; Snijders and Bosker 1999). As a default choice I use a class of weakly informative folded half-$t$ distributions as recommended in Gelman (2006). In cases where we allow multiple sets of parameters to vary by group (such as an intercept and multiple covariates), I also use the multivariate extension of the half-$t$ prior, the scaled inverse Wishart distribution (Huang and Wand 2013). These priors make it feasible to effectively model a large number of groups each containing relatively few members. Under the estimation framework I propose in Section 4.4 computation remains tractable in of these settings.

**Ordered Groups** In certain cases the groups of a mode will be naturally ordered. For example, the analyst may believe that the values of a parameter should be smooth through time or across space. In geography this is neatly captured in Tobler’s law “everything is related to everything else, but near things are more related than distant things” (Tobler 1970) which essentially suggests that parameters of geographically proximate areas should be related. It is also the basic premise of autoregressive time series models where we believe the past influences the present (Hamilton 1994; Brandt and Williams 2007). These notions of inter-related groups allow us to model a mode with a large number of groups even if each group has relatively few observations. If we are willing to assume, for example, that parameter values of neighboring time points are related, we can infer parameter which vary over time even with a relatively small number of units (Martin and Quinn 2002; Park 2012).

The can incorporate ordered group structure information using a broad class of prior distributions called Gaussian Markov Random Fields (GMRFs) (Rue and Held 2004; Rue, Martino and Chopin 2009). GMRFs are simply high
dimensional multivariate normal prior distributions where the precision matrix is a sparse matrix, $Q$. Thus the form of the coefficients is:

$$\beta \sim \text{Normal}(0, Q^{-1})$$ (4.10)

The precision matrix encodes conditional independence properties on the parameters. For example, in time series models we often make the assumption that parameters have a conditional independence structure such that:

$$\beta_{t+1} \sim \text{Normal}(\beta_t, \sigma^2)$$

where implicitly the value of the parameter at $p(\beta_{t+1}|\beta_t)$ is conditionally independent of $\beta_{t-1}$. In a GMRF we specify this by making the precision matrix $Q$ tri-diagonal. The sparsity in the precision matrix arises due to the conditional independence assumptions. Crucially as long as the matrix remains sparse computation is tractable even for extremely high dimensional parameters (Rue and Held 2004).

Many of the previous approaches to modeling complex data structures have focused on specifying a single mode model with a carefully constructed group structure. These frameworks can still be in the interactive latent factor setting described in this paper through the use of the GMRF priors. I direct interested readers to Rue and Held (2004) for the theoretical framework as well as relations to existing models. A shorter discussion directed towards political scientists can be found in Wawro and Katznelson (2013).

### 4.3.4 Summary

Modeling complex structures in the regression framework can be divided into two related components: the way different modes of the data interact and the way groups within a mode are related to one another. I have argued that we move beyond simple additive forms for models with multiple modes and instead have advocated interactive modes based on multilinear latent factors. These models gracefully extend from the two-mode case to an arbitrary number of modes. I have also shown that we can still incorporate rich information about group structure within a mode which often arises in time-series or spatial models.
4.4 Estimation

In this section I describe a class of fast approximate inference algorithms for posterior inference in the class of models introduced in Section 4.3. These algorithms use the framework of variational approximation which is a deterministic form of Bayesian inference (Jordan et al. 1999; Wainwright and Jordan 2008). This results in dramatic speed improvements over existing approaches in the social sciences. These computational improvements are not a simple novelty; they open a broader class of models as viable alternatives for use in exploratory data analysis (Gelman 2004) and iterative model fitting (Blei 2014).

Variational approximations are made tractable by strong assumptions on the nature of the dependence in the posterior distribution. This may at first seem at odds with the more flexible modeling strategy advocated in this paper. However, as I show using simulation evidence in Section 4.6, when the estimation algorithms are carefully designed we can still achieve excellent estimates of the true posterior. Furthermore, we can also use variational algorithms as a complement to more traditional Markov Chain Monte Carlo (MCMC) methods, both as a way to quickly explore possible models and as a highly accurate method of initializing the simulation state.

In the next section I briefly describe the inference problem, current state of the art, and why it is so challenging for the common tools of Bayesian inference. In the sections that follow I outline variational approximation treating in turn estimation of interactive modes, group structure prior and methods for automatically determining the rank of the approximation. This section is unavoidably more technical than the preceding portions of the paper and a reader uninterested in the details can safely skip to Section 4.5.

4.4.1 State of the Art

Our estimation goal is to calculate the posterior distribution of the latent variables given the data.

\[ p(\theta|y) \propto p(y|\theta)p(\theta) \]

Each literature has approached this problem in a different way including MCMC approaches based on Gibbs sampling (Aguilar and West 2000; Hoff and Ward 2004; Hoff 2011a; Pang 2014), variational inference (Lim and Teh 2007), Monte Carlo Expectation Maximization (Agarwal and Chen 2009), maximum likelihood (Bai and Li 2014), and a variety of algorithms based on the singular value decomposition (Bai 2009; Fithian and Mazumder 2013; Nakajima et al. 2013). Among these choices, Gibbs sampling is the \textit{de facto} standard for performing Bayesian inference in the
social sciences (Jackman 2000). Thus I give a brief explanation of how Gibbs sampling works in this context and motivate the move to alternative inference framework.

Gibbs sampling consists of sequentially sampling from the complete conditionals of each block of latent variables. As an example, consider the basic two-mode interactive latent factor models with a Gaussian outcome and no group structure. The model can be stated as:

\[ y_{i,j} = \sum_k u_{i,k}v_{j,k} + \epsilon \]
\[ \epsilon \sim \text{Normal}(0, \sigma^2) \]

MCMC proceeds by iterating between sampling \( u_{i,1} \ldots u_{i,k} \) for each group \( i \in \{1 \ldots I\} \) and \( v_{j,1} \ldots v_{j,k} \) for each group \( j \in \{1 \ldots J\} \). We then sample the error variance \( \sigma^2 \) and repeat. This sampler is easy to implement because the updates for each latent factor takes the form of a regression. So when updating \( \vec{u}_i \) we obtain a complete conditional which has the same form as a normal regression on the observations in unit \( i \) where the corresponding values of \( V \) play the role of covariates. For non-Gaussian likelihoods, we can introduce a Metropolis step to deal with the conjugacy (Hoff 2005). For cases with \( M \) modes the same basic structure holds where the “covariate” matrix is simply a product over the latent variables that we are conditioning on.

Gibbs sampling is attractive because it retains asymptotic guarantees of recovering the true posterior. However these guarantees only hold if the chain converges on all parameters which can be extremely difficult to assess in the high dimensional cases given here (Gill 2008). Furthermore convergence is typically extremely slow, with convergence times on the scale of hours to weeks being common. Slow mixing of the samplers arise because parameter updates for the interactive latent factors are strongly coupled. The result is an estimation framework that is not amenable to applied work.

In this section I develop an alternate estimation framework based on Variational Bayes. I emphasize that this is a complement to more traditional Gibbs sampling strategy. In the ideal case variational methods can be used to quickly fit and explore new models. Once a model has been selected we can run the time consuming, but asymptotically more accurate Gibbs sampler. This allows applied users to try out new specifications, inspect model fit and re-specify the model (Gelman 2004; Blei 2014).
4.4.2 Variational Approximation

In order to provide a computationally efficient method of posterior inference, I turn to variational approximation (Winn and Bishop 2005; Bishop et al. 2006; Grimmer 2010b; Ormerod and Wand 2010). In variational inference we estimate the parameters of an approximating set of distributions to make our approximation as close as possible to the true posterior in terms of the KL-divergence. Variational inference turns posterior inference into an optimization problem. The procedure is deterministic given the initialization and convergence is typically quite fast and easily assessed.

Before moving to derive the inference algorithms, it is worth emphasizing how the variational algorithms are able to provide computational efficient estimation where Gibbs sampling does not. The core posterior inference problem is one of integration, where we seek to integrate over the latent variables. Due to the interaction of the latent factors this integration is intractable. Gibbs sampling solves the problem with Monte Carlo integration. Variational inference solves the problem by constructing an approximate posterior which factorizes in a way that makes the integration tractable. Thus whereas in Gibbs sampling we condition on the value of \( v \) while sampling \( u \), in variational inference we take the expectation over \( v \) with respect to the approximate posterior.

The downside of the factorization assumption is that the resulting posterior is an approximation and is theoretically guaranteed to understate the true variance (Wainwright and Jordan 2008). Gibbs sampling also provides an approximation, however (theoretically) we can always increase the accuracy by running the sampler for longer. In standard variational methods there is no simple method for trading off computational time for accuracy. \(^{12}\) Nevertheless, as I will show the variational approach has excellent accuracy and yields dramatic computational gains.

For the sake of space, I assume a basic familiarity with variational inference methods. See Grimmer (2010b) for an excellent short introduction directed at social scientists. In the next sections I describe variational inference for interactive latent factors in the one, two and \( M \) mode settings. I then discuss computation for group structure priors.

**Single Mode Settings**

In single mode settings, the modeling framework described above reduces to a Generalized Linear Mixed Model (GLMMs) also called multilevel or longitudinal models (Gelman and Hill 2007). GLMMs are widely applied and inference procedures for them have been comprehensively studied. I note that even though MCMC methods are relatively straightforward for these models (Hadfield 2010; Martin, Quinn and Park 2011; Pham and Wand 2014), less

\(^{12}\) Although in the appendix I describe a few methods that allow for more accurate approximations at the expense of computational time. These methods are particularly geared towards difficult cases such as the non-conjugacy induced in logistic regression models.
accurate maximum likelihood methods are still extremely popular (Pinheiro and Bates 2000; Bates 2010) due to their computational convenience (Shor et al. 2007).

The framework for the single mode setting encompasses a fairly wide range of models. I refer to Zhao et al. (2006) for an explanation using similar notation. For the Gaussian outcome with groups indexed by $g$ the model is given by

$$y|\beta, u \sim \text{Normal}(X\beta + Zu, \sigma^2_{\epsilon}) \quad (4.11)$$

$$u_g|\Sigma^R \sim \text{Normal}(0, \Sigma^R) \quad (4.12)$$

where $X$ collects the covariates with globally shared effects and $Z$ is a block diagonal matrix over groups containing effects which are group specific. The positive definite covariance matrix $\Sigma^R$ captures the covariance across the group-specific effects. Note that the $R$ superscript is only a notational convenience to remind us that these are the covariances of the “random” effects.

With conjugate priors for $\beta, \sigma^2, \Sigma^R$ the entire model is conditionally conjugate which significantly simplifies inference. To keep the setup we use a simple set of conjugate priors,

$$\sigma^2_{\epsilon} \sim \text{Inverse-Gamma}(a_{\epsilon}, b_{\epsilon}) \quad (4.13)$$

$$\Sigma^R \sim \text{Inverse-Wishart}(A_{\Sigma^R}, B_{\Sigma^R}) \quad (4.14)$$

$$\beta \sim \text{Normal}(0, \sigma^2_{\beta}I_P) \quad (4.15)$$

where $P$ is the number of columns of $X$ and $\sigma^2_{\beta}$ is a large value strictly greater than 0.

The approximation to the full joint posterior is

$$p(\beta, u, \Sigma^R, \sigma^2_{\epsilon}) \approx q(\beta, u)q(\Sigma^R, \sigma^2_{\epsilon}) \quad (4.16)$$

$$= q(\beta, u)q(\Sigma^R)q(\sigma^2_{\epsilon}) \quad (4.17)$$

where in the first line we give the approximate posterior under a minimal product restriction (Menictas and Wand 2013) and the second line follows due to induced factorizations (Bishop et al. 2006).

Under standard variational inference theory (Bishop et al. 2006; Grimmer 2010b), the optimal approximating
densities for a parameter $\theta$ take the form

$$q(\theta) = \exp(E_{q(-\theta)} \log(p(\theta|\text{rest}))) \quad (4.18)$$

Algebraic manipulations show these forms to be

$$q(\beta, u) = \text{Normal}(\mu_{q(\beta, u)}, \Sigma_{q(\beta, u)}) \quad (4.19)$$

$$q(\sigma^2_\epsilon) = \text{Inverse-Gamma}(a_n, b_n) \quad (4.20)$$

$$q(\Sigma^R) = \text{Inverse-Wishart}(A_N, B_N) \quad (4.21)$$

where the exact forms of the posterior parameters $(a_n, b_n, A_N, B_N)$ are defined in the appendix. The algorithm proceeds by updating each of the quantities in turn until convergence. Convergence can be assessed by monitoring the Evidence Lower Bound given by

$$\log(p(y|q)) = E_q[\log p(y, \beta, u, \Sigma^R, \sigma^2_\epsilon) - \log q(\beta, u, \Sigma^R, \sigma^2_\epsilon)] \quad (4.22)$$

In practice the algorithm outlined above can be computationally intensive for data containing a large number of groups. Following Lee and Wand (2014) I leverage the block diagonal structure of $Z$ to calculate the necessary inverses. This is cumbersome in terms of notation but results in substantial computational benefits. I use this basic approach in the algorithms described below but continue to use the simpler notation as above.

**Extensions to Non-Gaussian Settings**

For the Gaussian outcome model all the priors conjugate or can be made written in a conjugate form using data augmentation. This is not true for the broader class of generalized linear models. These models require a slightly more complicated inference scheme as a result of the nonconjugate prior. Here I describe inference for the logistic regression setting. Algorithms for the Poisson and negative binomial model are also available in Luts and Wand (2013); Wand (2014b) and are comparatively straightforward.

In logistic regression, a Bernoulli likelihood over $y \in \{-1, 1\}$ is parameterized by the sigmoid (inverse-logit)
function of the parameters:

\[ P(y|\eta) = \sigma(y\eta) \] (4.23)

where \( \eta \) is the linear predictor and \( \sigma \) is the sigmoid function \( \frac{1}{1+\exp(-\eta)} \).

The log-likelihood is then

\[ \log p(y) = \sum_n \log(\sigma(y_n\eta_n)) \] (4.24)

However this leads to an intractable expectation. Instead I introduce an additional local variational bound on the marginal likelihood. Following Jaakkola and Jordan (2000) I approximate the sigmoid term using a quadratic lower bound such that

\[ \sigma(y\eta) \geq \sigma(\xi)\exp\left(\frac{(y\eta - \xi)/2 - \lambda(\xi)\left((y\eta)^2 - \xi^2\right)}{2}\right) \] (4.25)

\[ \lambda(\xi) = \tanh(\xi/2)/(4\xi) \] (4.26)

which introduces a new variational parameter \( \xi \) for each data point. The bound is tight at the optimal value of \( \xi \). With the introduction of the parameters \( \xi \) the data likelihood is now a quadratic function of the parameters to be optimized and thus we get a normal variational distribution for our regression coefficients with closed form mean and variances. \( \lambda(\xi) \) ends up playing the role of inverse error variances in a regression style update.

Jaakkola and Jordan (2000) show that the optimal values of the variational parameters can also be solved in closed form by

\[ \xi = \sqrt{E[\eta^2]} \] (4.27)

Thus the entire procedure contains only closed form updates and thus does not need to resort to numerical optimization. Because the approximation to the sigmoid function is a lower bound, the Evidence Lower Bound is still a true lower bound on \( \log(p(y)) \). Further details are given in Appendix A.\(^{14}\)

\(^{13}\)Although this representation is less standard in the social sciences, the symmetric form of the likelihood simplifies the notation below.

\(^{14}\)The justification of Jaakkola and Jordan (2000) is based on constructing a lower bound for the marginal likelihood using convex duality. However, recent work by Scott and Sun (2013) has given a probabilistic interpretation showing the connection to data augmentation using the Polya-Gamma latent variable family (Polson, Scott and Windle 2013).
Chapter 4. Latent Factor Regressions for the Social Sciences

There are numerous other approaches to nonconjugate variational inference (Salimans and Knowles 2013; Wang and Blei 2013; Knowles and Minka 2011; Ranganath, Gerrish and Blei 2013; Tan and Nott 2013; Marlin, Khan and Murphy 2011). However, I choose the lower bound approach for its relative simplicity and computational efficiency. In Appendix D I describe alternative approaches for handling the nonconjugate terms in the variational bound including approaches using quadrature (Tan and Nott 2013) and piecewise bounds (Marlin, Khan and Murphy 2011) both of which allow the analyst to tradeoff computational time for accuracy.

Two Mode Settings

In the two mode case estimation becomes complicated by the interaction between the latent variables. Consequently a stronger factorization assumption is needed to make the expectations tractable. Again I start with the simplest version of the model in order to demonstrate the basic inference strategy. Using interactive latent effects only for a varying intercept term and with a Gaussian likelihood yields:

\[ y_{i,j} \sim \text{Normal}(x_{i,j}\beta + \sum_k u_{i,k}v_{j,k}, \sigma^2_\epsilon) \] (4.28)

\[ u_{i,k} \sim \text{Normal}(0, \rho^2_k) \] (4.29)

\[ v_{i,k} \sim \text{Normal}(0, \tau^2_k) \] (4.30)

\[ \beta \sim \text{Normal}(0, \sigma^2_\beta I_P) \] (4.31)

where, for the moment, I treat the variance of the latent factors \( \rho^2, \tau^2 \) and the noise variance \( \sigma^2_\epsilon \) as fixed. When notationally convenient I collect the latent factors \( u \) into a matrix \( U \) where each row \( i \) contains the \( k \) factors for group \( i \). We denote the row of matrix \( U \) contain the latent factors of group \( i \) as \( U_i \). \( V \) follows similarly.

Following the computer science literature (Lim and Teh 2007), I assume a factorization over the latent factors:

\[ q(U,V,\beta) \approx q(U)q(V)q(\beta) \] (4.32)

\[ = \prod_{i=1}^I q(U_i) \prod_{j=1}^J q(V_j)q(\beta) \] (4.33)

Note that this is not a minimal product restriction on the variational parameters as either \( q(U) \) or \( q(V) \) could be combined with \( q(\beta) \) but I separate them in order to keep the treatment of the two modes symmetric.

The consequence of the stronger factorization assumption is that the approximation is unable to capture the pos-
terior covariance between the latent factor matrices $q(U)$ and $q(V)$. In the true posterior these effects are going to be negatively correlated, and it indeed it is exactly this feature which makes Gibbs sampling challenging. This hurts the accuracy of the approximation and will in general cause the approximation to understate the variance. That said, this does not appear to substantially detract from the quality of the approximation for the other parameters $q(\beta)$.

Standard calculations lead to the following Gaussian forms of the approximate densities:

$$q(U_i) = \text{Normal}(\mu_{q(U_i)}, \Sigma_{q(U_i)})$$ (4.34)
$$q(V_j) = \text{Normal}(\mu_{q(V_j)}, \Sigma_{q(V_j)})$$ (4.35)
$$q(\beta) = \text{Normal}(\mu_{q(\beta)}, \Sigma_{q(\beta)})$$ (4.36)

The posterior parameters of the approximation are updated as

$$\Sigma_{q(U_i)} = \left( \frac{1}{\tau_i^2} 0 \ldots 0 \right) + \sum_{j=1}^J \Sigma_{q(V_j)} + \mu_{q(V_j)} \mu_{q(V_j)}^T + \frac{\sigma_e^2}{\sigma_e^2} \left( \sum_{j=1}^J \Sigma_{q(V_j)} \right)$$ (4.37)

$$\mu_{q(U_i)} = \Sigma_{q(U_i)} \sum_{n \in \Omega} \left( \frac{(y_n - x_n \beta) \mu_{q(V_n)}}{\sigma_e^2} \right)$$ (4.38)

where $\Omega$ indicates the set of observations for which $y$ is observed. The form of $q(V)$ following analogously. Although the form seems complicated at first, it is simply Bayesian linear regression with two distinctions. First, we are now fitting the model to the residuals $(y - x\beta)$ and second we have to include the covariance of the variational distribution when calculating the cross products.

The variational distribution for $\beta$ is even simpler as it corresponds directly to Bayesian linear regression on the residuals

$$\tilde{y}_{ij} = y_{ij} - E[U_i]E[V_j^T]$$ (4.39)
$$= y_{ij} - \mu_{q(U_i)} \mu_{q(V_j)}^T$$ (4.40)

The logistic regression case is essentially analogous to the derivation here so I defer details to the appendix. A
particular feature of the logistic regression case is that further computational speedup is possible through the use of a case control approximate likelihood (Raftery et al. 2012). I plan to explore this in future work.

The introduction of a second interactive mode leads to an optimization problem that contains many local optima. Consequently the choice of how to initialize the algorithm is particularly important in determining the estimated solutions (Roberts, Stewart and Tingley Forthcoming). From an applied perspective this is problematic because we might eliminate the benefits of our speed improvements by repeatedly fitting the model in order to find the global optimum.

For the two-mode case we can use recent results in theoretical computer science to find the global optimum of the Evidence Lower Bound in particular special cases (Nakajima and Sugiyama 2011; Nakajima et al. 2012, 2013). Even in cases which are not covered by this analysis, similar techniques to generate a strong deterministic initialization. I discuss this approach next.

Initialization for the Two Mode Setting

The initialization procedure defined here is based on the theoretical analysis of Nakajima et al. (2013) which shows that for fully observed matrices a simple algorithm can find the global optimum of the variational objective for the Gaussian probabilistic matrix factorization model with no additional observed covariates. The analysis not only yields a useful algorithm for initializing the model but it also clarifies some of the properties of the variational estimation strategy.

The core result of Nakajima et al. (2013) is to show that the globally optimal variational parameters can be recovered by soft-thresholding the singular value decomposition (SVD) of the matrix.\textsuperscript{15} The idea of soft-thresholding an SVD has arisen across various applications in statistics (Donoho and Johnstone 1994; Chatterjee 2012; Fithian and Mazumder 2013). The key to the analysis of Nakajima et al. (2013) is to show the exact correspondence with the factorized variational Bayes solution. They also show that we can recover the variances of the latent factors which correspond with the optimal MAP estimation of those parameters under an empirical Bayes strategy. This allows for an automatic rank selection of the decomposition, often called Automatic Relevance Determination in the machine learning literature (Bishop et al. 2006).

The algorithm involves calculating the SVD of the outcome data matrix. The singular values are then soft-thresholded using a threshold which is estimated from the data based on the dimensionality of the data and the error

\textsuperscript{15}Soft-thresholding is an operation that appears frequently in the literature on sparse estimation. It means that we shrink the parameter towards zero unless it is sufficiently small at which point we set it to exactly zero.

100
variance. The procedure involves only a single SVD calculation and two uni-dimensional parameter optimization. The computational cost is dominated by the SVD calculation which for moderate sized matrices is usually quite small.

The theoretical results in Nakajima et al. (2013) only guarantee that the procedure finds the global solution in a very restrictive setting: a Gaussian likelihood and the ability to arrange the data into a fully observed matrix. For the many cases not covered by this setup the results still provide a useful initialization. Consider for example the model described in the previous section,

\[ y_{i,j} \sim \text{Normal}(x_{i,j}\beta + \sum_k u_{i,k} v_{j,k}, \sigma^2). \]

I start by estimating the model without the latent factors in order to get an initial estimate for \( \beta \). Then calculate residuals \((y - x\beta)\) and arrange them into a matrix. If a particular combination of groups \( i, j \) does not appear in the data I replace this cell with the mean of the remaining residuals.\(^{16}\) I then calculate our estimates of \( q(U)q(V) \) and use these to initialize the model.

The SVD procedure can also be embedded into the update process. When the likelihood is Gaussian and the matrix is fully observed these conditional updates on the residuals are exact. With minor levels of missingness in the matrix or a non-Gaussian likelihood the updates can be used as proposals which are accepted only if they increase the value of the Evidence Lower Bound.\(^{17}\) Crucially these moves are joint in \( q(U)q(V) \) which can be helpful when there are a large number of groups. When using the updates iteratively I compute the SVD using a relatively recent algorithm, the implicitly-restarted Lanczos bidiagonalization algorithm (Baglama and Reichel 2005). This approach allows us to warm start the algorithm with the previously calculated values and only compute the number of singular values required by the model. This makes the process significantly faster.\(^{18}\)

**M Mode Settings**

The move to the general \( M \)-way mode setting is straightforward for the basic variational algorithms. Due to the assumed factorization of the posterior, the required expectations remain tractable and still take the form of Bayesian

\(^{16}\)This seemingly *ad hoc* procedure is given a rigorous theoretical defense in Chatterjee (2012) in terms of the Frobenius norm of a partially observed matrix. See also the extensive work on matrix completion using convex optimization methods (Candes and Recht 2009; Cai, Candès and Shen 2010; Mazumder, Hastie and Tibshirani 2010).

\(^{17}\)For related approaches to using SVD to update parameters see Seeger and Bouchard (2012); Fithian and Mazumder (2013).

\(^{18}\)In practice the SVD procedure is valuable as an initialization but typically unnecessary within iterative updates of the algorithm itself. Rigorous analysis of the quality of this procedure as an initialization and update procedure are left to future work.
linear regressions. The factorization of the posterior is now:

$$q(\beta, U^{(1)} \times U^{(M)}) \approx q(\beta) \prod_i q(U^{(1)}_i) \ldots \prod_j q(U^{(M)}_j)$$ (4.41)

with the latent factors still taking multivariate Gaussian forms.

In the two mode case I exploited matrix decompositions for initializing the model. In the general $M$ mode setting the SVD and related theoretical results are no longer available. The data can instead be arranged into a tensor and tensor decomposition can be used as an initialization (Kolda and Bader 2009). The particular multilinear form presented in this article corresponds to a particular type of tensor decomposition called the CANDECOMP/PARAFAC (CP) tensor factorization (Kolda and Bader 2009; Hoff 2011a; Zhao, Zhang and Cichocki 2014).\(^{19}\)

Recent work in theoretical computer science has explored the use of tensor decompositions for estimating parameters for latent variable models using a method of moments framework (Anandkumar, Ge, Hsu, Kakade and Telgarsky 2012; Anandkumar, Liu, Hsu, Foster and Kakade 2012; Anandkumar, Ge, Hsu and Kakade 2013\(^{b}\)). This work has in turn driven work on tensor decomposition methods which have provable guarantees (Anandkumar, Ge and Janzamin 2014\(^{a}\); Suzuki 2014). Here I adopt the procedure of (Anandkumar, Ge and Janzamin 2014\(^{a}\)) for the CP decomposition of non-orthogonal tensors.\(^{20}\) The Anandkumar, Ge and Janzamin (2014\(^{a}\)) procedure provides global convergence guarantees under the presence of incoherent tensor components which are essentially a soft orthogonality constraint.

For now I refer the interested reader to the original papers (Anandkumar, Ge and Janzamin 2014\(^{a,b}\)), simply noting that the procedure works well for initializing the higher order models in practice. Further development of this procedure as well as the circumstances where guarantees can be made is left to future work.

### 4.4.3 Estimation for Group Structure Priors

In section 4.3 I outlined several options for prior distributions on the latent factors that allow us to control the way groups within a mode are interconnected. Here I briefly describe computation for these priors in the variational setting.

\(^{19}\)Unlike matrix decompositions, the tensor decomposition can be challenging to compute in general, and the workhorse method Alternating Least Squares is not even guaranteed to converge in general (Kolda and Bader 2009).

\(^{20}\)When the tensors are orthogonal and symmetric the decomposition can be computed using tensor eigen decomposition (Anandkumar, Ge, Hsu, Kakade and Telgarsky 2012). Thus the state of the art is to "whiten" the tensor so that it is orthogonal symmetric and then estimate the decomposition (Anandkumar, Ge and Janzamin 2014\(^{a}\)). However whitening is often the most computationally expensive part of the process and, most importantly for applied use, the least numerically stable (Huang et al. 2013).
Unordered Groups

The conjugate prior for the latent factor variances is the Inverse-Gamma distribution. Due to the conjugacy the q-density is also an Inverse-Gamma. Consider for concreteness an Inverse-Gamma prior on coefficients $\beta$ in a linear regression. We place a $\text{Gamma}(a_0, b_0)$ prior on the precision parameter of the Normal density of the $P$ dimensional vector $\beta$. The optimal q density has the form

$$ q(a) = \text{Gamma}(a_N, b_N) \quad (4.42) $$

$$ a_N = a_0 + P/2 \quad (4.43) $$

$$ b_N = b_0 + .5 \cdot E_{\beta, \sigma^2} \left( 1/(\sigma^2 \beta^T \beta) \right) \quad (4.44) $$

The expectation will in turn be a function of the posterior variance as well the mean and covariance of the coefficients $\beta$.

In the single mode case it is popular to see more weakly informative priors such as the Half Cauchy (Gelman 2006) or the Scaled Inverse Wishart (Huang and Wand 2013). These priors are also available within this framework by using data augmentation. Wand et al. (2011) shows that the Half Cauchy can be represented by

$$ \rho_{i,r}^2 \sim \text{Inverse-Gamma}(5, 1/a_{i,r}) \quad (4.45) $$

$$ a_{i,r} \sim \text{Inverse-Gamma}(4, 1/A_{i,r}^2) \quad (4.46) $$

where the marginal distribution for $\rho_{i,r}^2$ is now Half-Cauchy($A_{i,r}$). Similar results are available for sparsity promoting priors such as the Laplace distribution, Horseshoe distribution and Generalized Double Pareto (Wand et al. 2011; Neville, Ormerod and Wand 2012). In summary, the variational framework provided here is able to encapsulate the full range of priors for unordered groups which are typically used in longitudinal data analysis.

Ordered Groups

When groups are ordered the analyst can use the class of Gaussian Markov Random Fields (GMRFs) to perform inference. As shown in Rue and Held (2004) the key to tractable computation is the sparsity of the precision matrix $Q$ which encodes conditional independence assumptions in the model. They key to computation is that the sparsity properties of the precision matrix are inherited into the cholesky decomposition of $Q$. I briefly sketch the strategy
differing readers to Rue, Martino and Chopin (2009) for more details.

Consider a GMRF prior on a random variable $u$ such that $u \sim \text{Normal}(\mu, Q^{-1})$. The density is then given by

$$p(u) \propto |Q|^{1/2} \exp(-0.5 (x - \mu)^T Q (x - \mu)) \quad (4.47)$$

The cholesky factorization gives $Q = L L^T$ where again $L$ remains sparse. We can solve equations of the form $Qu = b$ by solving $Lv = b$ and then $L^T u = v$. These fast system solutions can form the basis of a Fisher scoring method for finding the posterior mode (Rue and Held 2004).

### 4.4.4 Rank Determination

In all but the single mode case the rank of the interactive factors needs to be selected. The issue of setting the model dimensionality is a common concern in latent variable models (McLachlan and Peel 2004). Intuitively rank selection places the model on a continuum between the case where the effects are purely additive and the case where they are jointly unique. However rarely is there specific knowledge about this continuum and thus forcing the analyst to fix the rank a priori is unappealing. This is particularly problematic for MCMC methods where model estimation can take days to weeks which precludes the possibility of testing alternative specifications.

A simple approach the rank determination problem, which I use throughout the applications, is to allow each dimension of the latent factor matrix to have a variance parameter for each rank $k$ which is point estimated in an empirical Bayes framework. A particular property of this estimation strategy is that it will set unnecessary factors to exactly zero resulting in a type of rank selection called Automated Relevance Determination (ARD) (Bishop et al. 2006). In the two mode case, this is called model-based regularization and is shown to arise due to the nature of variational approximation (Nakajima et al. 2012). Point estimates of the variances are:

$$\rho^2_r = \frac{1}{I-1} \sum_{i=1}^I (\Sigma_{ij(q(U_i))})_{r,r} + \mu^2_{r,q(U_i)} \quad (4.48)$$

and can also be computed through the SVD based method of (Nakajima et al. 2013). The ARD dimensionality selection also works in the tensor case (Zhao, Zhang and Cichocki 2014).

---

21 I say forcing because it may be that optionally fixing the rank is desirable in circumstances where the analyst wishes to interpret the latent factors. In that setting having a low dimensional rank can make visual inspection easier. In this paper, I am primarily concerned with settings where the structure in the dataset is a nuisance that we use the latent effects to marginalize over rather than something to be interpreted. Nevertheless, the framework is completely compatible with fixing the rank.

22 The ARD approach under a Normal prior as I have used here is closely related to a convex relaxation of the rank selection problem using the nuclear norm of $uv^T$ (Fithian and Mazumder 2013).
The ARD approach has two major limitations which motivate the possibility of alternative approaches. First, because the ARD approach uses point estimates for the variances it will necessarily understate the variance of the model (because we cannot be certain about the true rank). Second, the ARD approach is not compatible with structured priors on the latent factors as would arise in models of group structure. In both cases it is necessary to adopt a more explicit model of rank selection.

A Bayesian nonparametric approach to this problem which has been shown to be successful in related work is based on the multiplicative gamma process (Bhattacharya and Dunson 2011). I plan to develop variational algorithms for this approach in future work. In the applications and simulations below I use the ARD approach for the latent factors.

4.4.5 Summary

In this section I’ve outlined computation for the broad class of latent factor models using variational methods. These algorithms are fast to estimate and unlike simulation based methods convergence is easily assessed by monitoring the lower bound on the marginal likelihood. As I will show through simulation evidence in Section 4.6 inference is also highly accurate.

In the near future, I will release software implementing these methods for the R language.

4.5 Related Work

As I suggested in the introduction the core ideas in the methodological framework I describe here have been repeatedly reinvented throughout a variety of disciplines. For the most part these methods seems to have primarily been developed in isolation with few connections between different sections of the literature. Although a complete review of the related literature would be impossible, I highlight the related work, drawing connections which to the best of my knowledge have not been made.

---

23 Only very recently have variational inference methods for the gamma process begun to emerge in the literature (Roychowdhury and Kulis 2014).

24 I briefly summarize the multiplicative gamma process prior for the interested reader. The idea is to write the model in a form that explicitly introduces a scaling parameter which is comparable to the singular value. So for the two mode case we have: \( \eta_{i,j} = \sum_k s_k u_{i,k} v_{j,k} \) where \( s \) plays the role of the singular values. We place a multiplicative gamma process prior on this term as in Bhattacharya and Dunson (2011). This prior takes the form \( s_k \sim N(0, \tau_k^{-1}) \) with \( \tau_k = \prod_{l=1}^k \delta_l \) and \( \delta_l \sim \text{Gamma}(a, 1) \) for \( a > 1 \). Thus as the rank increases the precision are shrunk towards zero forcing rank selection. This approach is developed in Rai et al. (2014) using MCMC inference with either a truncation or adaptation strategy for handling the dimensionality. These methods should be straightforward to extend to the variational setting. In this way we could place group structure priors on the latent factors \( u \) and \( v \) without interfering with the rank selection prior on \( s \).
In reviewing the related work I give a practitioners view, dividing the literature into three broad areas that reflect approaches to modeling structured data. The first area covers standard regression methods such fixed/random effects and standard error corrections (Section 4.5.1), the second area are the uses of interactive latent factor models of various types (Section 4.5.2), the third area reviews the relevant work on structured group priors with a particular focus on models which can be framed as Gaussian Markov Random Fields (Section 4.5.3). Finally, I discuss some limitations of this approach.

The key feature across all areas is that although hundreds of different models have been developed for different types of data there are a relatively small number of common themes. The models I have presented in this paper can apply to nearly all the settings described below.

### 4.5.1 Fixed/Random Effects and Standard Error corrections

The most common approach to dealing with unobserved heterogeneity is the use of standard error corrections. These approaches typically use some form of sandwich estimators for the variance and have been developed for a huge number of data types such as time-series cross-sectional (Beck and Katz 1995), clusters (Arellano 1987), spatial correlation (Driscoll and Kraay 1998), dyadic (Aronow and Samii 2013) and a host of others. In recent years it has been argued that these corrections do not adequately address the problem (Beck 2012; King and Roberts 2014). I return to one of these critiques in more depth below.

The second most common approach to modeling heterogeneity is the use of fixed effects (Angrist and Pischke 2008), random effects (Wooldridge 2010), random coefficient models (Hsiao and Pesaran 2008) and multilevel modeling (Gelman and Hill 2007). An extensive literature has developed around the relative merits of these approaches, much of which focuses on the choice between fixed or random effects (Browne and Draper 2006; Beck and Katz 2007; Shor et al. 2007; Arceneaux and Nickerson 2009; Bell and Jones 2012; Stegmueller 2013). These approaches are all special cases of the single mode case of the framework presented here. In the presence of more than one mode, random/fixed effects models are limited to simple additive forms over the modes. As such these models are easily replicated in the estimation procedures described in Section 4.4.

What both the standard error and fixed/random effects literature have in common is the willingness of scholars to implicitly specify the modes of the data. For example, scholars are willing to add ‘country fixed effects’ or use standard error corrections which purportedly address temporal auto-correlation. This suggests that analysts are generally open to modeling dependence in their data but simply lack easily available approaches to more sophisticated modeling. The
next three sections discuss some of the more sophisticated approaches, none of which have enjoyed the widespread use of these simpler approaches.

### 4.5.2 Latent Factors

The use of interactive latent factors has surfaced in a number of distinct areas of the literature. The idea that differentiates many latent factor models is the parametric form given to the latent factor. In this work I have presented a latent Gaussian model but other distributional forms are common in particular applications. Here I give a brief summary of the most relevant work across disciplines by application area. I focus primarily on latent Gaussian models as they are the most relevant to this work.

**Networks**

Latent factor models have been particularly popular in the burgeoning literature on the analysis of networks. The network literature itself covers a large range of applications from social networks, protein networks, dyadic analysis and applications in genomics. In each case the analyst is concerned with modeling a binary outcome $y_{ij}$ which indicates a link between node $i$ and node $j$. An extensive survey form a computer science and statistical perspective is given in Goldenberg et al. (2010).

One of the fundamental models in the network literature is the stochastic block model (Wang and Wong 1987). Here the latent factor is assumed to be a discrete variable which is interpreted to represent membership in a latent community. Thus if two nodes share a community they have a higher probability of having an edge. Extending the model to latent variables which lie on the simplex results in the Mixed Membership Stochastic Blockmodel (Airoldi et al. 2008) where each node has proportional membership across all $K$ communities.

A different parametric form for the latent factors was created by Hoff, Raftery and Handcock (2002) based on the notion of a social space. Each node is projected into a low-dimensional latent space where nodes that are closer together are more likely to have an edge. This approach was later extended in Hoff (2005) to a latent factor model of the type considered in the two mode case here, named the Generalized Bilinear Mixed Effects model. Although these three models are different in their assumptions of the functional form they are actually quite similar in their implied mathematical form.

Crucially Hoff’s work gives a rigorous statistical motivation for the latent factor network models from the perspective exchangeable latent variables (Hoff 2008). Hoff (2005) shows that the latent factor model is able to characterize a
number of properties of the network that are missed by the additive latent effects framework including transitivity, balance and clusterability. These models have been imported into political science through collaborations with Michael Ward and his students (Hoff and Ward 2004; Ward, Siverson and Cao 2007; Ward, Stovel and Sacks 2011; Dorff and Ward 2013). Because the interest has primarily been in networks, the majority of this work considers the two-mode case for symmetric square matrices with undefined diagonals (i.e. source-receiver structures in an undirected network). The likelihoods are typically binary (tie or no tie) although alternative network likelihoods have been considered (Hoff et al. 2013). A notable extension of these two-mode models to the case of networks over time is considered by Ward, Ahlquist and Rozenas (2013) where the latent factors within each time period are pooled together using a dynamic linear model.

Recent work has extended these two-mode models to multiway relational data using tensor decompositions. Hoff (2011a) explores the CP decomposition used in this paper, and Hoff (2011b) uses a decomposition based on the more general Tucker product. These $M$-mode models have been applied to a variety of relational data settings including ANOVA priors with deep interactions (Volfovsky and Hoff 2012), factor analysis for multivariate outcome data (Fosdick and Hoff 2014) and event count models (Hoff 2011a). Recent work has moved beyond latent Gaussian priors to consider prior distributions on the Stiefel manifold which allow for equivariant and scale-free estimation (Hoff 2013). Corresponding statistical theory based on the exchangeable random structures is given in Lloyd et al. (2013); Orbanz and Roy (2013).

The work by Peter Hoff and his coauthors is the single biggest influence on the current work and thus it is worth explicitly contrasting it to the models considered here. The network models described in this section share four limitations which limit their applicability for the cases considered here: MCMC algorithms which are slow to estimate, the use of interactive latent effects only on intercept terms, no ability to place group structure priors on the latent factors and implementations which are limited to the square symmetric case. I address all of these issues in the

---

25 Hoff et al. (2013) considers likelihoods for fixed rank nomination schemes. Hoff (2005) gives extensions to the ordered probit case for ordinal relations. The amem package in R implements these two approaches in addition to normal relational data and binary data all for the square symmetric setting.

26 Estimation in the Ward, Ahlquist and Rozenas (2013) is performed sequentially over each time step using the previous time step as the prior for the next. Crucially, the current paper shows how we can do joint estimation for this model in the framework given here by using the GMRF representation of the dynamic linear model in the two-mode case. For a related formulation see Durante and Dunson (2013).

27 Specifically the Tucker product represents the tensor decomposition as the product of a core-array (analogous to singular values of a matrix) with factor matrices for each mode. Hoff (2011b) shows that this corresponds with an array normal distribution having separable covariance structure. The CP decomposition used here is a special case of the Tucker Product where the core array is super-diagonal. I opt for the simpler CP decomposition form to maintain a simpler inference structure. What is lost in this process is the ability to have different rank approximations along each mode, which should not be a substantial sacrifice except in the very high dimensional case.

28 A notable exception is the variational algorithm of Salter-Townshend and Murphy (2013) for the latent space model of Hoff, Raftery and Handcock (2002). This approach uses a structured mean-field variational algorithm that requires numerical optimization of several of the parameters whereas the algorithms I employ here use entirely closed form updates.
unified framework here.

**Recommendation Systems in Computer Science**

Two related applications in computer science have been strong proponents of latent factor models: collaborative filtering (recommending items to people) and link prediction (filling in missing edges in a network). An excellent review of the link between the two is given in Menon and Elkan (2011). The applications in computer science are the most foreign to the settings commonly found in social science paper. However, the nature of the problems addressed in recommendation systems creates a distinctive focus on scalable methods applicable to large data settings, and the ability to handle incomplete or missing data. These two features are essential components of the framework I develop in this paper and play an important role in making these methods broadly applicable. Thus I provide a brief overview of the relevant literature.

The basic probabilistic matrix factorization model is given in Salakhutdinov and Mnih (2007) with the corresponding probabilistic tensor decomposition described by Chu and Ghahramani (2009). Variational algorithms are considered in Lim and Teh (2007) and Zhao, Zhang and Cichocki (2014) respectively.

The collaborative filtering literature has also considered the inclusion of covariates where it is used to address the “cold start” problem (i.e. how do you recommend a movie to a user who has not rated any movies yet). These models consider mode specific covariates which are used to inform the priors over the latent factors (Agarwal and Chen 2009; Zhang, Agarwal and Chen 2011; Agarwal, Chen and Pang 2011; Chen et al. 2011). Additional covariate models under different probabilistic assumptions are given in (Miller, Jordan and Griffiths 2009; Porteous, Asuncion and Welling 2010).

Although arising from a distinct literature these models are essentially of the same form as the network models considered above. However they provide useful insights on approaches to scalable computation that provide a helpful complement to the network literature.

**Interactive Fixed Effects in Econometrics**

In Econometrics, a class of models related to the two-mode case have been considered under the moniker of interactive fixed effects as a way to model time-series cross-sectional data (Pesaran 2006; Bai 2009). The interpretation given to the models is country-specific responses to global economic shocks and is often presented as an alternative to a spatial weights model which side steps the need to choose the weights matrix (Bai 2009; Sarafidis and Wansbeek 2012;
Chapter 4. Latent Factor Regressions for the Social Sciences

Zhukov and Stewart (2013). Notably the interactive fixed effects models have recently been introduced to political science by Gaibulloev, Sandler and Sul (2014) and Pang (2014). These models are distinctive from the ones considered here in that they assume no prior distributions on the latent factors. This precludes the use of partial pooling, group structure priors and model-based methods of selecting dimensionality. Furthermore the regularization in Bayesian approaches often leads to improved parameter estimation for high dimensional cases such these.

From an applied perspective a significant weakness in all of the above models is that they require balanced panels (each cross-sectional unit has a fully observed time series of the same length). In most practical settings this limits the analyst to either a very small set of cases or a very short time series. As far as I am aware none of the above literature makes the connection to the network or computer science literature although they are essentially the same model.

Additional applications

The latent factor structure surfaces in a variety of other fields including demography (Lee and Carter 1992; Brouhns, Denuit and Vermunt 2002), forecasting (Mammen, Nielsen and Fitzenberger 2011; Aguilar and West 2000), neuroimaging analysis (Zhou, Li and Zhu 2013; Zhou and Li 2014) and gene expression analysis (Carvalho et al. 2008). A few general frameworks have been proposed for particular cases: notably in the two mode case for generalized bilinear regression (Gabriel 1998) and matrix-variate data (Allen and Tibshirani 2012).

A small subset of work considers the combination of latent factor models with the kind of group structure priors we describe here. Lopes, Salazar and Gamerman (2008); Lopes, Gamerman and Salazar (2011) consider an interactive two mode model for time-series spatial data where one of the factors is given a spatial Gaussian Random Field prior. Durante and Dunson (2013) and Ward, Ahlquist and Rozenas (2013) study a two mode relational case with dynamic linear model priors.

Finally I note that the same latent factor structure underpins a variety of models in automated text analysis (Grim-

---

29 A variety of estimation approaches have been proposed primarily using the framework of maximum likelihood or estimators based on the singular value decomposition. Rigorous theory for the maximum likelihood estimation of these models is given in Bai and Li (2014). Pesaran (2006) presents an estimation framework based on common correlated effects (CCE) which can be estimated using OLS. This framework is further extended by (Castagnetti, Rossi and Trapani 2012).

30 By model-based, I mean selection of the dimensionality within the context of the model. A variety of different post-hoc selection metrics have been proposed for choosing the dimensionality of the latent factors. Most of these involve the use of various information criteria applied to the principal components of the error structure (Bai and Ng 2002, 2008). Moon and Weidner (2010) argue that in the interactive fixed effects model we don’t need to worry about setting the number too high, only too low, prompting an investigation of lower bounds.

31 Gerard and Hoff (2014) give results that show that the Bayes procedure for the array decomposition dominates the MLE. This parallels well-known results for the multivariate Gaussian where the MLE of the covariance matrix is neither admissible nor minimax (James and Stein 1961).

32 Numerous extensions have been proposed that parallel models in the network literature. Such extensions include measurement error (Lee, Moon and Weidner 2012), temporal lags (Fang, Chen and Zhang 2013), group shrinkage (Lu and Su 2013), bayesian versions (Liu, Sickles and Tsionas 2013), unknown group membership (Ando and Bai 2013), and diagnostic tests (Su, Jin and Zhang 2012).
mixed membership topic models, for example, can be framed as a two mode case (documents and words) where the latent factors are assumed to lie on the simplex and the likelihood is Poisson (Blei 2012; Gopalan, Hofman and Blei 2013). When given latent gaussian priors these models correspond exactly to the setting here (Hu, Ryu, Carlson, Wang and Carin 2014).

The framework also encompasses ideal point models of vote counts common in political science (Clinton, Jackman and Rivers 2004). Here the two modes are bill and legislator with a logistic or probit link bernoulli likelihood.

Summary

The core insights of using latent factor structures for modeling complex data has arisen independently across a truly astonishing number of fields. This review of the literature is of course not exhaustive but provides a sense of the breadth of applications, inference procedures and interpretations given to these models. A key goal of this paper is to highlight the commonality in these myriad approaches and leverage the best features of different traditions.

4.5.3 Structured Gaussian Priors

In latent factor models groups are typically un-ordered, with the interactive modes doing the work of modeling the dependence structures in the data. By contrast, the time-series, spatial statistics and multilevel modeling literature use the structure between groups to model dependence in the data. In this paper I’ve advocated the use of Gaussian Markov Random Fields (GMRFs) as a way of representing group structure within the model. The technical details including the broad range of models GMRFs encapsulate is given by Rue, Martino and Chopin (2009) and Rue and Held (2004). The advantage is that this infrastructure allows us to use the extensive work on modeling group structure in spatial statistics (Besag, York and Mollié 1991; Franzese Jr and Hays 2007; Gleditsch and Ward 2008), time series analysis (Brandt and Williams 2007; West and Harrison 1997; Hamilton 1994) and multilevel modeling (Gelman and Hill 2007; Snijders and Bosker 1999) all within the context of the interactive latent factor models described here.

Explicit use of GMRFs has been relatively rare in political science. Girosi and King (2008) argue for a GMRF prior structure for applications in demographic forecasting. Wawro and Katzenelson (2013) advocate the use of GMRFs as a tool for modeling parameter heterogeneity in service of bridging methodological divides between quantitative and qualitative approaches. Fortunately, this is easily incorporated into this paper’s setup, as discussed in Section 4.4.3.
4.5.4 Alternative Approaches

In the supplemental appendix I provide a short description of how the latent factor regression framework relates to a number of other related approaches. These include exponential random graph models (Cranmer and Desmarais 2011), survival analysis (Box-Steffensmeier and Jones 2004), binary treatment causal inference (Imai and Kim 2012; Blackwell 2013), mixture models Park (2012); Imai and Tingley (2012), flexible regressions (Wahba 1990; Gu 2013; Beck, King and Zeng 2000; Hainmueller and Hazlett 2014) and graphon estimation (Chatterjee 2012; Airoldi, Costa and Chan 2013). Many of these methods address a specific type of data or take a fundamentally different approach. The contrast between these alternative methods and the framework developed in this paper help to highlight the distinctive features of my approach.

4.5.5 Limitations

The latent factor regression framework provides a very flexible modeling strategy for capturing dependence but it nevertheless has some limitations. The most important limitation is that the framework relies on the ability of the analyst to specify the number of modes as well as the group membership. In many applications this is a reasonable limitation; indeed, this is the same information which scholars are implicitly providing when they specify fixed effects, clustered standard errors or other types of corrections. Crucially these decisions are natural to make on theoretical grounds as they correspond to the analyst’s identification of the salient units in the data. An alternative view of this requirement is an assumption of exchangeability, which guarantees that the data are conditionally independent given the group-specific latent variables.

The bayesian modeling framework adopted in this paper also requires that the latent factors are uncorrelated with the effects on the observed covariates. This is the usual “random effects” assumption and initially seems quite unrealistic. In practice, it does not appear that the model is particularly sensitive to this assumption (as I will show via Simulation in the next section). Furthermore we can include covariates containing group level averages of predictors of interest in order to break this correlation as suggested in the multilevel modeling literature (Mundlak 1978; Bafumi and Gelman 2006; Bell and Jones 2012). An exact theoretical characterization of the size of this problem is beyond the scope of the present work but is an area of interest for future investigation.

Finally, the estimation framework proposed introduces strong independence assumptions in the posterior. However, as I will show in the next section, we can still obtain extremely accurate approximations to the posterior that have
favorable frequentist coverage properties on the main effects of interest.\textsuperscript{33}

In many cases the variational approximation will be sufficiently accurate to provide posterior inference on the quantities of interest to applied researchers. When a higher accuracy approximation is needed, the variational approximation can always be used to initialize a sampling based approach which will asymptotically recover the true posterior.\textsuperscript{34} Thus we can have the best of both worlds: the fast variational methods can be used to quickly explore and re-specify models and can then be used to help speed convergence of the asymptotically exact sampling algorithms. In future work I will pursue MCMC algorithms which are able to leverage the variational posterior directly.\textsuperscript{35} While the latent factor regressions framework does require strong assumptions, we can often weaken our reliance on these assumptions in various ways. In the next section I provide simulation evidence which addresses many of the concerns above.

### 4.6 Simulation Evidence

In this section I provide simulation evidence that the estimation framework outlined in Section 4.4 provides accurate estimation of model parameters and is sufficiently fast to enable applied use in an interactive setting. I start with a set of simple simulations for the single mode (Section 4.6.1 and then then two mode case (Section 4.6.2). In each case I demonstrate that the variational estimation algorithm runs hundreds of times faster than MCMC while also correctly recovering posterior means and factor rank. I also show that the 95\% credible intervals have excellent frequentist coverage in the single mode case and are only slightly too narrow in the two mode case.

In both cases, I give some general timing results to provide a sense of the relative speed of the variational algorithms compared to MCMC. The code I use for variational inference is unoptimized native R code. I expect the public release of the software to have substantial speedups over the timing results presented here.

In Section 4.6.3 I test model sensitivity to the assumption that the latent effects are uncorrelated with the covariate effects. I also provide a comparison to standard fixed effects strategies. Additional details for the simulations are included in Appendix E.

\textsuperscript{33}Of course some quantities of interest will be completely unavailable; notable, the posterior covariance between distributions assumed to factorize. However these terms are rarely of interest in applied work. When they are, alternative MCMC estimation strategies will be necessary.

\textsuperscript{34}I thank Marc Ratkovic for suggesting this strategy.

\textsuperscript{35}For example, the variational posterior may be useful for developing proposals in a Hamiltonian Monte Carlo framework (Neal 2011).
### 4.6.1 Single Mode Case

**Simulation** I start by considering the single mode case with unordered groups and a normal likelihood. I consider a case with three covariates which have both population level and group specific effects. The example is taken from the help file of `MCMCpack` and is reproduced in full in the appendix (Martin, Quinn and Park 2011). In the first simulation I use 20 groups and 1000 observations. In all cases I use the uninformative half-Cauchy priors the variances and a scaled inverse-Wishart prior for the random effects covariance matrix.

**Speed** It is difficult to compare timings between MCMC and deterministic methods because it is unclear how long one should run the MCMC chain. Here I simply use the default parameters in the help file which uses 1,000 passes of burnin followed by 10,000 draws from the posterior thinned at intervals of 10. It is also worth noting that `MCMCpack` uses highly optimized C++ code compared to the unoptimized native R code for the variational approach. Even with these caveats the timings are incredibly clear. The variational solution takes on average 0.205 seconds and `MCMCpack` takes 27.86 seconds. Thus the variational solution is 136 times faster. In order to match this speed `MCMCpack` would have to use 80 samples to characterize the posterior with no burnin, which is clearly an unrealistic option.

**Accuracy** Next I demonstrate recovery of the posterior mean. Figure 4.2 shows that the variational algorithm is extremely accurate at recovering the posterior mean (which is expected given the theoretical properties of variational inference). The posterior credible intervals also have excellent frequentist coverage with the 95% credible interval covering the truth in 97, 95 and 96 simulations for the three parameters respectively.

---

Figure 4.2: *Recovery of the posterior mean compared to MCMC over 100 simulations on each of the three main parameters.*

---

36I use this data generating process not for any theoretical reason but to signal that the inference method is applicable to a simulation which I did not design for this purpose.
**Scalability**  The above results were for 20 groups under 1,000 observations to mirror the existing data generating process. I ran a second simulation using 122 groups and 2,673 observations to match the application in Section 4.7.1. Again the variational approach was substantially faster taking just under 2 seconds compared to 101 seconds from MCMC.

**Logistic Regression**  The logistic case does not enjoy the theoretical guarantees of the normal likelihood due to the introduction of the additional lower bound on the marginal likelihood. However, results remain quite strong with coverage of 93%, 90% and 96% on the main three effects and computational time ranging between 0.25-3 seconds. Figure 4.3 shows the comparison of the posterior means. The posterior means are slightly, but systematically, biased towards zero for the variational method which is consistent with previous findings in the literature (Ormerod and Wand 2012; Tan and Nott 2013).

![Figure 4.3: Comparison of posterior means between MCMC and Variational for the logistic regression case.](image)

### 4.6.2 Two Mode Case

Once we move to a model with interactive latent factors it becomes more difficult to produce a gold standard reference. Many existing inference methods are inapplicable to the case outlined here because the matrix is both non-square (thus ruling out network inference methods) and partially missing (thus ruling out the econometric approaches). Assessing convergence in custom MCMC algorithms is challenging even for very small cases and thus I focus here on assessing recovery of simulated parameters.
Simulation Details  I simulate a synthetic dataset based on the actual data from the application in Section 4.7.1. This is a time-series cross-sectional analysis where the two modes contain 31 groups (years) and 118 groups (countries). The outcome data can be organized into a matrix dimension $118 \times 31$ with approximately 30% of the entries being missing. This missing data makes the estimation more challenging because the initialization procedure no longer possesses any formal guarantees of global optimization. I chose this setting because it actively reflects the common state of data in the social sciences.

We also observe a set of 8 covariates collected with an intercept into the matrix $X$. Thus the model is:

$$y_{ij} = x_{ij}\beta + a_i + b_j + \sum_{k=1}^{K} u_{ik}v_{jk} + \epsilon$$  \hspace{1cm} (4.49)

where $a$ and $b$ are country and time varying intercepts, $U$ and $V$ are the interactive factors and $\epsilon$ is the normal error term. I simulate all parameters from standard normal distributions and fix the group indexes and covariates $X$ to their observed values. To address rank selection, I also randomly draw $K \sim \text{Pois}(\lambda = 3) + 1$ which gives a distribution over integer values ranging primarily from 2-5.\footnote{The observed distribution of ranks in my random sample was: Rank 1: 2, Rank 2: 16, Rank 3: 9, Rank 4: 8, Rank 5: 8, Rank 6: 4, Rank 7: 1, Rank 8: 0, Rank 9: 1.}

Speed  The unoptimized variational code takes about 5 seconds to estimate the model. Clearly any MCMC timing is going to depend entirely on the number of simulations, but replicating procedures in the literature I can estimate that sampling each model would take approximately 2 days.\footnote{Here I base the MCMC time on the Gibbs sampling code for the Generalized Bilinear Mixed Effects model (Hoff 2005). Ward, Siverson and Cao (2007) in a similar application with only $K = 3$ latent factors, ran the sampler for 500,000 iterations which is necessary due to the high auto-correlation in the chain. In a similar setting Fosdick and Hoff (2013) report running the chain for 500,000 iterations in a $K = 3$ latent factor models and calculating effective sample sizes between 734 and 2607. By running a smaller sample I was able to estimate the average cost of each iteration as approximately 0.373 seconds. This places the cost of running the simulation for 500,000 iterations at 2.16 days. Clearly one could run the chain for shorter periods of time but at best one could get about 15 samples in the time necessary for the variational algorithm to complete.}

Accuracy  I consider two accuracy properties. The ability to recover the true parameters and the ability to learn the true rank of the latent factors. In every one of the 50 simulations the algorithm correctly inferred the true rate. Figure 4.4 shows the true and estimated parameters for the three blocks of parameters: globally shared regression coefficients ($\beta$), the random intercepts ($a, b$) and the inner product of the latent factors ($u'v$). For the latter two I use a kernel density smoother (Wand 2014a) so that the distribution of points will be more easily visible. We can see clearly by the way the points in all three panels hug the diagonal that the estimates are extremely accurate. Average frequentist coverage
for a 95% credible interval across all regression coefficients was 92%.

![Figure 4.4: 50 simulations of the two mode model with estimated parameters in the variational algorithm. The left panel shows the globally shared regression coefficients with their estimated and true values across all 50 runs. The middle panel shows the same for the country and time intercepts. The right panel shows the estimated and true product of the interactive latent factors.

**Scalability** We note that convergence is extremely rapid even for larger numbers of factors. By contrast MCMC algorithms mix substantially slower as the latent dimensionality rises, requiring dramatically more computational time.

### 4.6.3 Simulated Model Misspecification

In this section I examine the performance of the latent factor regression framework in a case of model misspecification and compare it to the performance of several alternative strategies. It is difficult to effectively simulate data with the complexities of real-world covariates; thus, as in the previous section, I use the observed covariates from my first application in Section 4.7.1. In contrast to the previous simulation I also use the fitted values from the model to populate the parameters and latent variables. This allows the parameters to be arbitrarily correlated and “realistic” in the sense that they represent an actual model fit. I generate a new error term in order to simulate the outcome (using a larger error variance than originally estimated).

In order to introduce correlation between the covariate effects and the latent variables, I randomly drop from zero to seven of the covariates before estimating the model. This forces the model to capture the covariate effects within
the latent factors. I always leave in one main theoretical variable and evaluate the ability of the model to recover this parameter.

I also compare the method to four alternative specifications including two used in prior work. These specifications are:

1. One-Way Fixed Effects
   “country” level intercepts which are the largest source of variation in the model.

2. Two-Way Fixed Effects
   “time” and “country” intercepts. This is the additive two-mode model.

3. Global Linear Detrending with One-Way Fixed Effects
   “country” intercepts and a linear time trend shared by countries

4. Country-Specific Quadratic Detrending
   “country” specific quadratic time trends

The last two specifications are chosen to mirror the empirical strategies used by previous work. I discuss these in more detail in the applications section.

Figure 4.5 shows the baseline case with full observed covariates. The fixed effects and linear detrending strategies are unable to model the dependence and consequently dramatically overestimate the effect size of the covariate of interest. Quadratic country-specific detrending does better but has confidence intervals that are entirely too large whereas the latent factor regression does extremely well covering the interval in 23 of the 25 simulations which is just shy of the 95% coverage rate.

Figure 4.6 shows the process repeated with seven missing covariates, leaving only the theoretical variable of interest. Here we can see that the latent factor regression does extremely well, again having 23 of the 25 intervals covering the truth and only a small upward bias. The other four estimators perform analogously to the fully observed case with the notable exception that the quadratic detrending now exhibits a strong positive bias. The remaining cases of missing one to six covariates are presented in the supplemental appendix.

I emphasize that this simulation does not demonstrate that latent factor regression is always a superior method. We should expect it to perform the best in this situation as it is the closest to the true data generating process. The simulation does however illustrate two important points. First, the latent factor regression performs well in cases where the latent effects are correlated the observed covariate effects. This corroborates analogous findings for multilevel models.
Figure 4.5: 25 simulations from a two-mode model with full observed covariates. Each of the five estimation strategies is shown with 95% confidence/credible intervals. The red dashed line indicated the true effect to be recovered. All but the quadratic detrending and the latent factor regression strategies massively overestimate the true effect size. The confidence intervals for the quadratic detrending are much too conservative compared to the latent factor regression.

under other simulation strategies (Bafumi and Gelman 2006; Bell and Jones 2012). Second, inadequate modeling of dependence can cause us to dramatically overestimate our effect of interest.
Figure 4.6: 25 simulations from a two-mode model with all but the covariate of interest missing. Each of the five estimation strategies is shown with 95% confidence/credible intervals. The red dashed line indicated the true effect to be recovered.
4.6.4 Overview and Limitations

The sequence of simulations above demonstrate the inference framework proposed in Section 4.4 is able to estimate simulated parameters with both high accuracy and remarkable speed. These results hold with interactive latent factors, partially missing data and nonconjugate likelihoods. Uncertainty estimation is also handled well by the estimation framework with the credible intervals shown to have excellent frequentist coverage properties. Finally the automatic rank selection for the interactive latent factors performed perfectly.

Although accuracy is quite high throughout the most noticeable weakness is in the logistic regression setting. However, in this setting coefficients are biased towards zero making the procedure more conservative in the expected effect size. In future work I plan to improve the logistic regression approximation, and approaches to doing so are discussed in Appendix D.

The simulations considered here are limited in that the randomness in simulating the parameters masks some of the complexity of real applications. I address this partially by conditioning on existing covariate values and estimated parameters. Nevertheless, an exploration of more complex simulation studies is ongoing.

Having established the excellent performance of the estimators on simulated data we now turn to two real applications.

4.7 Applications

The modeling framework in this paper has broad applicability across the social sciences. In this framework I focus on two particular applications in the fields of international relations which motivated the development of this framework. Both cases follow a similar pattern in the literature of an initial finding of theoretical interest and a methodological response. They also both use a type of dataset structure that is common in the literature: time-series cross-sectional and longitudinal network data. I show how these can both be addressed in the common framework of this paper.

The first application is based on Büthe and Milner (2008)’s study of the role of international trade agreements in increasing foreign direct investment (FDI). They argue that membership in the General Agreement on Tariffs and Trade (GATT) and successor the World Trade Organization (WTO) increase FDI. They use a linear detrending strategy with country fixed effects and to control for temporal and cross-sectional heterogeneity. In a methodological critique of robust standard errors, King and Roberts (2014) replicate one of the models in Büthe and Milner (2008). They introduce a country-specific quadratic time trend which eliminates the positive and statistically significant effect of
joining the GATT/WTO on FDI. In the framework I have developed here, King and Roberts (2014) is arguing that the two modes time and cross-section are jointly unique rather than additive as assumed in Büthe and Milner (2008). I show that by using my modeling framework we can recover a positive effect of the GATT/WTO on FDI while satisfying the criteria of King and Roberts (2014).

The second application is an examination of the democratic peace hypothesis and the subsequent critique in Green, Kim and Yoon (2001) described in Section 4.2.1. Recall that the data is organized in a source-receiver-time structure. The essence of the critique is that additive mode effects for each country are insufficient to address unobserved heterogeneity and that each dyad must be considered jointly unique. I build off the extension of this work by Ward, Siverson and Cao (2007) who consider a two-mode interactive latent space model estimated separately at five year intervals across the data. I show that not only can the latent space model be estimated dramatically faster than in Ward, Siverson and Cao (2007), but also that it is possible to jointly model the time dimension as well through a three-mode interactive latent factor model. This means that we do not have to estimate the model by five year intervals, and thereby more naturally partially pool the data through time-specific random effects. In both cases and contrary to the critique of Green, Kim and Yoon (2001), a pacific effect of democracy is identified in keeping with the democratic peace hypothesis.

Both applications share a common theme. The original work identifies two modes of dependence which are controlled for additively. The critiques in King and Roberts (2014) and Green, Kim and Yoon (2001) correctly identify that problematic levels of dependence still remain along these modes and threatens the evidence for the main findings. However, these problems are not easily corrected. Following previous best practice, both critiques consider the modes of dependence as jointly unique which in both cases causes the original effect to disappear. The latent factor regression framework occupies a middle ground between the additive and jointly unique modeling strategies. Crucially it is able to sufficiently address the dependence in the data while also identifying a significant effect. This is an important improvement on previous best practice because we should ideally use statistical procedures that demand the least from the data while still satisfying the key requirements of conditional independence.

### 4.7.1 Political Determinants of FDI

Büthe and Milner (2008) study the political factors which affect foreign direct investment. Specifically they argue that joining international trade agreements institutionalizes commitments to liberal economic policies which are attractive to the potential investors. Using a variety of empirical strategies they analyze a dataset of 122 developing countries
from 1970-2000 concluding that participation in the GATT/WTO has a positive impact on FDI inflows.

The observations in the Büthe and Milner (2008) study are at the country-year level. The authors use two strategies for addressing unobserved heterogeneity within an OLS framework. First they linearly detrend the outcome and independent variables. To capture cross-sectional variation they use country fixed effects estimated by demeaning the dependent and independent variables within groups and adjusting the degrees of freedom appropriately. Remaining heteroskedasticity in the errors is addressed via the use of robust standard errors (Arellano 1987). Büthe and Milner (2008) implements a variety of robustness checks to validate these findings including estimating via generalized least squares, allowing for an AR(1) process, using panel corrected standard errors (Beck and Katz 1995), bootstrapped standard errors, and instrumental variable analysis (Wooldridge 2010). Here I focus on the political and economic factors model (Model 4, Table 1 in the original paper).

The Critique

The Büthe and Milner (2008) example is part of a larger critique of King and Roberts (2014) on the overuse of robust standard errors. They argue

when misspecification is bad enough to make classical and robust standard errors diverge, assuming that it is nevertheless not so bad as to bias everything else requires considerable optimism. And even if the optimism is warranted, settling for a misspecified model, with or without robust standard errors, will still bias estimators of all but a few quantities of interest (King and Roberts 2014, pg. 1).

The argument that we are generally better modeling the data rather than relying on standard error corrections has been echoed throughout the methodological literature (Freedman 2006; Beck 2012; Dorff and Ward 2013). King and Roberts (2014) show that rather than jettisoning robust standard errors entirely we can use them as a diagnostic test of model misspecification. When robust standard errors differ from their classical counterparts it is indicative of some feature of the data that needs better modeling. To formalize the notion of ‘difference’ between classical and robust standard errors, they develop a generalized information matrix (GIM) test.

King and Roberts (2014) replicate three articles which use robust standard errors including Büthe and Milner (2008). After using the GIM test to demonstrate misspecification, they identify the source of the problem as the detrending strategy. Given the highly heterogeneous set of countries they use a detrending strategy that is both country specific and quadratic. The resulting model does not exhibit the strong temporal trends in the original model and

39 In Appendix F I provide a comparison of the two detrending strategies in the original feature. The crux of the matter is that because the detrending in King and Roberts (2014) is country-specific, persistent covariates such as WTO/GATT membership have most of their variance
passes a GIM test for heteroskedasticity and autocorrelation. The new model gives an estimate of a slightly negative effect of GATT/WTO membership with a confidence interval that covers zero, changing the conclusions of the original paper. King and Roberts (2014) conclude by noting that they chose a detrending strategy in order to stay close to the original text, but that “an alternative and probably more substantive approach would be to drop the detrending strategy altogether and to model the time series process in the data more directly” (King and Roberts 2014, pg. 27). This paper proposes a methodology that does exactly this.

**Applying the Latent Factor Model**

Using the framework presented in Section 4.3, I show that we can avoid the detrending procedure entirely and directly model the interactive effects of time and cross section. I use an interactive two mode model with country effects (indexed by \( c \)) and time effects (indexed by \( t \)). Thus the model can be given as:

\[
\text{fdi}_{c,t} = X_{c,t} \beta + a_c + b_t + \sum_k u_{c,k} v_{t,k} + \epsilon
\]

(4.50)

where \( a_c \) and \( b_t \) are country and time specific intercept terms, \( U \) and \( V \) are latent factor matrices for country and time respectively, and \( \epsilon \) is the normal error term. The country and time specific intercept terms are given Gaussian priors with Half-Cauchy priors on the associate variance terms. The factor matrices are given Gaussian priors with point estimated variances which allows us to infer dimensionality by Automatic Relevance Determination. A complete statement of the model is given in Appendix F.

This approach differs from the previous models in a few key ways. First I do not need to assume a strong parametric form for the temporal effects. The yearly effects are treated as unstructured parameters allowing for the possibility of abrupt economic shocks. Second, the temporal and cross-sectional effects are estimated within the model and thus are available for analysis and interpretation along with their associated credible intervals. Finally, the model occupies a conceptual middle ground between the two prior solutions. Time effects can be country-specific but information sharing through the inner product term \( u'_c v_t \) allows for more efficient use of information.

Figure 4.7 plots the variational posterior of the effects of GATT/WTO and Cumulative Preferential Trade Agreements (PTAs). The expected benefit of GATT/WTO membership corresponds to an expected increase of 0.205 in log FDI. This corresponds to an expected 23% increase in FDI for members vs non-members (with a 95% credible interval of 2% to 46%). The effect is smaller than the original finding Büthe and Milner (2008) but is still both sub-removed.
stantively and statistically significant. A similar pattern holds for cumulative Preferential Trade Agreements where each additional PTA is associated with a 5% increase in expected FDI with a 95% credibly interval of 0.4% to 9%.

An added benefit of the latent factor model is that we can use the estimates of the interactive latent factors to explore the nature of unobserved heterogeneity and hopefully in the future further refine our theory. Figure 4.8 plots a projection of the countries into a two dimensional space.\textsuperscript{40} Countries which are near each other in the latent space respond to common shocks in time in a similar way. This provides a sense of what the interactive latent effects model is capturing beyond the covariates and two-way country and year intercepts.

Importantly for the methodological debate over the findings, the residuals in the latent factor model exhibit temporal correlation similar to the correction proposed in King and Roberts (2014). In Figure 4.9 I reproduce the time series residual plot (Figure 9) of King and Roberts (2014) for the same selection of cases and parameter settings. This shows that the residuals in the latent factor model are comparable to the country specific quadratically detrended model while still finding the relevant effect. Further comparisons across all countries support the finding in the these three sample cases.

\textsuperscript{40}The model estimates the latent factors to be of rank 9 which would mean that a completely faithful reproduction would require 9 dimensions. Instead I project the effects down to two dimensions using Sammon scaling of the euclidean distances between the factor loadings (Sammon 1969). Sammon scaling has the property of accurately preserving small distances at the expense of larger distances. This increases the likelihood that countries which are close together in the low dimensional space are actually close together in the high dimensional space.
**Chapter 4. Latent Factor Regressions for the Social Sciences**

![Figure 4.8](image1)

**Figure 4.8:** Posterior means of the country-specific interactive latent effects projected into 2 dimensions using Sammon multidimensional scaling. Each successive panel from left to right is further zoomed in. Countries which are close to each other respond to global temporal shocks in a similar way.

![Figure 4.9](image2)

**Figure 4.9:** Time-series residual plots for three countries comparing the linear detrending with country fixed effects (black), country-specific quadratic detrending (red) and latent factor model (green). Lines give loess smoothed trends with a span of 3/4. This is a reproduction of Figure 9 in King and Roberts (2014) with the addition of the latent factor model.

**Summary**

In this example I demonstrated that the latent factor modeling framework can be used to effectively model a dataset with complex time series and cross-sectional dependence. I emphasize that the original article by Büthe and Milner (2008) is a carefully performed study which uses a large number of robustness checks to establish the validity of their finding. King and Roberts (2014) are also correct in pointing out the remaining correlations in the error structures suggest we should reconsider the evidence for the finding. Crucially it does not appear that there is any way of
modeling dependence that is strictly additive in the time and cross-section effects that would adequately control the dependence. Furthermore, existing latent factor models from other disciplines are unavailable due to the unbalanced panels and asymmetry of the time and cross-sectional structure. This leaves treating the time series and cross-sectional effects as jointly unique as the only available option. My approach provides a new intermediate point which allows us to provide new evidence for the argument of Büthe and Milner (2008).

4.7.2 The democratic peace

The democratic peace is arguably the most robust empirical finding in the study of conflict. Maoz and Russet describe it as “one of the most significant nontrivial products of the scientific study of world politics” (Maoz and Russett 1993, pg. 624). The canonical reference is Oneal and Russett (1999) which uses dyad-year data to establish the pacific effect of trade, joint involvement in international organizations and democracy. The Oneal and Russett (1999) model has been the subject of intense interest and scrutiny, including a large number of challenges and replies. A particularly prominent case of such a challenge is the “Dirty Pool” debate discussed above.

The Critique

The challenge of Green, Kim and Yoon (2001) in the “Dirty Pool” debate was to appropriately model the joint effects of a particular pair of countries in explaining militarized interstate disputes. Ward, Siverson and Cao (2007) take up this challenge using the social relations model to model dyadic dependence (Hoff and Ward 2004; Ward, Stovel and Sacks 2011; Dorff and Ward 2013). As we showed in Section 4.5, the social relations model is closely related to the two mode interactive latent factor model. Because the model is defined for static networks Ward, Siverson and Cao (2007) estimate the model separately on 11 specific years rather than model all of the time periods at once. This strategy of modeling distinct “snapshots” of the data has the advantage of allowing for parameters to be different within each time period, an issue that has been raised before in the context of the democratic peace (Clarke, Goemans and Peress 2010; Jenke and Gelpi 2012). However, it analyzes only a small portion of the data (a total of 11 years in

---

41Earlier versions of this finding include Maoz and Russett (1993) and Oneal and Russett (1997)
42The literature typically has the followed the pattern of a challenge and reply reaffirming the original findings. These critiques tend to fall into two camps. The first are methodological concerns such as simultaneity bias (Keshk, Pollins and Reuveny 2004; Kim and Rousseau 2005), non-linearity of effects (Beck, King and Zeng 2000) and temporal dependence (Box-Steffensmeier, Reiter and Zorn 2003; Beck, Katz and Tucker 1998). The second are questions of interpretation such as the claim that the finding is driven by the economy (Gartzke 2007; Mousseau et al. 2013; Mousseau 2013) or other confounding effects (Kacowicz 1995; Farber and Gowa 1997; Gartzke 2000). Generally these concerns are met with direct responses that reaffirm the original finding (De Marchi, Gelpi and Grynaviski 2004; Oneal and Russett 2005; Hegre, Oneal and Russett 2010; Dafoe, Oneal and Russett 2013; Dafoe 2011). The findings of the democratic peace have also held when directly tested against a wide variety of alternate theories (Bennett and Stam III 2004).
their 50 year period) and does not allow for efficient to be shared across years. Nevertheless, in most, but not all years, they find support for the hypothesis that higher levels of democracy reduce the probability of conflict.

In the conclusion to their article, Ward, Siverson and Cao (2007) write

“One weakness of work on this topic to date is the absence of any substantial consideration of time dependencies despite our demonstration that other dependencies are important... In the long run it will be important to include temporal as well as higher-order dependencies in our models of interstate interaction. However no one has yet solved this problem (Ward, Siverson and Cao 2007, pg. 598).”

Next I show how we can do exactly this by considering the three mode latent factor model.43

Applying the Three Mode Model

In order to be able to make direct comparisons with Ward, Siverson and Cao (2007), I consider only the 10 years considered in their snapshots using the same explanatory covariates.44 The outcome variable is a binary variable indicating the presence of a militarized interstate dispute for a source, receiver, year triple. I estimate a three-mode interactive latent factor structure (source, receiver and time). In order to capture temporal heterogeneity in the parameters I allow each time slice to be governed by a separate set of covariate effects and pool them together using a hierarchical prior.

Collecting all the covariates and an intercept together in a matrix $X$ and indexing the sender by $s$, the receiver by $r$ and the time by $t$, the model is

$$y_{s,r,t} \sim \text{Bernoulli}(\text{InvLogit}(\eta_{s,r,t}))$$ (4.51)

$$\eta_{s,r,t} = x_{s,r,t}\beta + x_{s,r,t}\gamma_t + a_s + b_r + c_t + \sum_k u_{s,k}^{(s)} u_{r,k}^{(r)} u_{t,k}^{(T)}$$ (4.52)

where $\gamma_t$ are the time specific covariate effects, $a, b, c$ are source, receiver and time specific intercepts, $k$ indexes the dimensionality of the latent factors, and $u_{s,k}^{(s)}$ is the $k$th element of the source mode latent factor for country $s$ with the other terms following analogously.

The prior structures are similar to the previous example with a Normal prior and point estimated variances on all latent factors. For the time random effects vector $\gamma_t$ I use a weakly informative hierarchical multivariate prior for the

---

43It bears emphasizing here that since the publication of Ward, Siverson and Cao (2007) several pieces of work have proposed solutions to this general problem, many of which were highly influential on my current enterprise here. In particular my framework here can be seen as generalization of the approaches in Ward, Ahlquist and Rozenas (2013); Hoff (2011a).

44The model specification includes the product of polity scores, trade imports, common IGO membership, distance, as well as population GDP and Polity for each of the sender and receivers. Note I use 10 years rather than the full 11 because the replication file is missing the data file for 1970.
Chapter 4. Latent Factor Regressions for the Social Sciences

$P$ covariates in the model:

$$
\gamma_i | \Sigma \sim \text{Normal}(0, \Sigma) \quad (4.53)
$$

$$
\Sigma | \alpha_1 \ldots \alpha_P \sim \text{Inverse-Wishart}(\nu + P - 1, 2\nu \text{diag}(1/\alpha_1, \ldots, 1/\alpha_P)) \quad (4.54)
$$

$$
\alpha_p \sim \text{Inverse-Gamma}(5, 1/A_p^2) \quad (4.55)
$$

where $p$ indexes the covariates and $\nu, A_1^2 \ldots A_p^2$ are hyper-parameters which are fixed. Huang and Wand (2013) show that this prior structure is the multivariate equivalent of the Half-t prior proposed by Gelman (2006) and when $\nu = 2$ as I use here this corresponds to a uniform prior over the correlation parameters and each of the standard deviations having Half-t distributions with 2 degrees of freedom.\footnote{With only 10 groups the covariance in the random effects $\Sigma$ is unlikely to be particularly informative. Nevertheless for many settings this will be an attractive feature of the model and consequently I include the full multivariate prior form here.}

The data consists of 160,052 observations across 165 source countries, 165 receiver countries and 10 time periods. The size of the data creates a challenging inference problem and the current implementation of the variational algorithm was quite a bit slower than in previous cases.\footnote{Ultimately model fitting takes between 10 and 15 minutes for this which is still dramatically faster than the comparable sampling algorithm. In future work I hope to explore approaches to speeding up the necessary calculations even further.} The model selects a 7 dimensional latent factor.

The posterior distribution of the average effects across the four main dyadic variables is given in Figure 4.10. The findings support the basic tenets of the democratic peace hypothesis with joint democracy showing a pacific effect and trade also decreases the probability of war. However as in Ward, Siverson and Cao (2007), I find that Joint IGO membership increases the probability of war which runs counter to the Kantian peace argument. Finally, I emphasize as did Ward, Siverson and Cao (2007) that the most dominant effect is a simple measure of distance, reflecting that in the latter half of the 20th century geographic proximity plays the largest role in the probability of conflict.

In sum these results broadly support the Kantian peace hypothesis. Pooling together the available data allows for a clearer support of the joint democracy finding than the mostly mixed results from the separate analyses reported by Ward, Siverson and Cao (2007). Analysis of the residuals by time, dyad and individual source and receiver country reveals no remaining systematic correlations. Taken with the positive findings for the democratic peace, this suggests that the dyad-specific fixed effects solution of Green, Kim and Yoon (2001) imposed too stringent a demand on the data, eliminating heterogeneity at the expense of the ability to identify an interesting finding. The method presented here does not have this problem.
Figure 4.10: Posterior distribution of the main dyadic effects in the model of militarized interstate disputes. The grey line gives the 95% credible intervals and the dashed line marks 0.
4.8 Conclusion

In this paper I have introduced a framework for regression with structured data using interactive latent factors, demonstrating its utility through simulation and two applications. The framework generalizes and extends previous efforts across a variety of different fields. As such this paper provides an important unifying framework to statistical methodology (King 1998) for many data sets applied practitioners now face. I have also developed fast variational inference algorithms which make the estimation of these models feasible for applied use and which will soon be available for the open source community.

There are several useful ways in which the current work can be extended. For practical use it would be helpful to have a suite of diagnostic measures for assessing when heterogeneity has been insufficiently modeled in the data. Several measures have been proposed in the literature and a systematic effort to collect and implement these would be useful for practitioners. On the algorithmic side, I intend to explore methods for further characterizing the accuracy of the variational posterior and providing methods to improve accuracy where computationally feasible. Perhaps most importantly publicly available software will allow a much broader range of applications of the method which will in turn drive new innovations.

4.9 Appendix Road Map

In the appendix, I provide additional details of materials omitted from the main paper. Appendix A includes the technical details of the estimation algorithms. Appendices B-D provide additional insights into particular areas of the literature. Appendices E-F provide additional details on simulations and applications.

A Variational Inference Algorithms

This appendix details the six algorithms employed in the main text along with a short discussion of the technical contributions of the paper and a comparison to existing software implementations.

B Alternative Approaches

This section extends the literature review to include alternative approaches to modeling heterogeneity. Many of these models take a fundamentally different approach than I have taken here and the contrast clarifies the benefits and tradeoffs of the latent factor framework.

C Two-Way Fixed Effects and Latent Factor Regression
This appendix outlines the connection between special cases of the latent factor regression framework and two-way and joint fixed effects estimator. The connections help to illuminate how the model works with a particular focus on causal estimation in a potential outcomes framework.

D Improving Accuracy of the Variational Framework

This appendix discusses possible approaches for improving accuracy in the variational inference framework. It covers two possible improvements: those geared towards improved modeling on non-Gaussian (and thus non-conjugate) models, and those geared towards weakening the factorization assumptions in the approximate posterior.

E Simulation

Here I provide the details to replicate the simulation results in the main paper.

F Additional Application Details

This section collects additional details and results from the applications. Currently it includes a comparison of the two different temporal detrending strategies in Büthe and Milner (2008) and King and Roberts (2014).
Chapter 5

Conclusion

This dissertation has presented two methods for analyzing the kinds of complex datasets which frequently arise in computational social science. Although the methods target quite different data analysis situations they both strive to incorporate all available information about the structure in which documents or observations are nested. Here I briefly expand on future directions for each project.

5.1 The Future of Structural Topic Model

Although it is the more mature of the projects, STM is still under constant development and expansion. Below I describe current/future directions in three areas: software, estimation, and applications.

Software  We are continuing to update the open-source R package stm with additional functionality and faster code. We have also begun to develop an eco-system of “helper” packages which augment the functionality of the main software with visualizations for topic correlation (stmCorrViz) and interactive web-browsers for covariate-topic effects (stmBrowser). We also have two additional projects in development to build a graphical user interface for the package functionality (to support our users without programming background) and add parallel processing capability (to allow use on a cluster computing environment).

Estimation  In order to support ongoing work with Molly Roberts which studies propaganda in China, we have developed a stochastic variational inference algorithm for STM. This allows the model to be fit on corpora containing
Chapter 5. Conclusion

millions of documents and hundreds of topics. In future work we hope to move to the recently developed framework of stochastic structured variational inference (Hoffman and Blei 2015) which restores some of the dependencies in the posterior assumed away by mean-field assumptions.

Applications An exciting part of the STM project is that other scholars have begun to pickup the model for use in their own projects. Work in progress with Molly Roberts and Rich Nielsen on matching where the pre-treatment confounder is represented by text suggests that STM may have additional types of applications which we have not yet considered.

5.2 The Future of Latent Factor Regressions

The latent factor regression project is also undergoing rapid development. Future work can be roughly divided into advances in software, data preprocessing and model extensions.

Software The most important next goal is to release a software package in R which implements the algorithms described. The methods have been designed with eventual implementation in mind but the generality of the approach requires an intricate software design. Guided by the experience of developing the stm package, I plan to keep the workflow of the end-user in mind throughout all stages of development.

Data Pre-processing One key insight from developing the applications is that the model as written can be quite sensitive to how the data is pre-processed. For example, in modeling time-series cross-sectional data if a particular cross-section is on a dramatically different scale it can cause the model to overfit. Identifying and implementing standard methods of pre-processing the data, whether by rescaling a priori or including scaling parameters in the model would help address these concerns.

Model Extensions In LFR the decomposition of the latent space captures the part of the data generating process which is not well explained by the covariates. That is, we often want to estimate a particular covariate effect marginalizing out the heterogeneity in the units. A different goal would be to explore the covariate effects on the latent factor projection itself. This would be closer to the STM use case but with unnormalized latent variables and applicable to more complex data generating processes. This suggests a version of the model where the latent factors generate the
observed data directly and the covariates are in the prior for the latent factors. Such a model would have applications to exploratory analysis of all types of relational data.
Chapter 6

Appendix

In this appendix I provide additional details of materials omitted from Chapter 4 on Latent Factor Regressions. Appendix A includes a summary of technical contributions as well as details of the estimation algorithms. Appendices B-D provide additional insights into particular areas of the literature. Appendices E-F provide additional details on simulations and applications.

1. Variational Inference Algorithms
   This appendix details the six algorithms employed in the main text along with a short discussion of the technical contributions of the paper and a comparison to existing software implementations.

2. Alternative Approaches
   This section extends the literature review to include alternative approaches to modeling heterogeneity. Many of these models take a fundamentally different approach than I have taken here and the contrast clarifies the benefits and tradeoffs of the latent factor framework.

3. Two-Way Fixed Effects and Latent Factor Regression
   This appendix outlines the connection between special cases of the latent factor regression framework and two-way and joint fixed effects estimator. The connections help to illuminate how the model works with a particular focus on causal estimation in a potential outcomes framework.

4. Improving Accuracy of the Variational Framework
   This appendix discusses possible approaches for improving accuracy in the variational inference framework.
Chapter 6. Appendix

It covers two possible improvements: those geared towards improved modeling on non-Gaussian (and thus non-conjugate) models, and those geared towards weakening the factorization assumptions in the approximate posterior.

5. Simulation
Here I provide the details to replicate the simulation results in the main paper.

6. Additional Application Details
This section collects additional details and results from the applications. Currently it includes a comparison of the two different temporal detrending strategies in Büthe and Milner (2008) and King and Roberts (2014).

6.1 Variational Inference Algorithms
The goal of this appendix is to clarify the technical contributions of the latent factor regressions project and offer details on the estimation algorithms. Both the model design and estimation strategy draw from existing components in the literature, but combine them in novel ways. This work is distinguished from prior work by the development of a more general framework and attention to the demands of applied data analysis.

Specific novel contributions include:

1. Gaussian Markov Random Field (GMRF) priors with factorization models
   This allows the user to optionally include information about how units are connected (such as temporal or spatial smoothness). Thus they form a complement to the factorization models: GMRF’s provide a flexible framework for encapsulating known information and factorization models infer unknown information.\(^1\)

2. Variational Inference for Factorization Models with Observed Covariates
   Prior work used Gibbs sampling methods which are often too slow for applied use. I build on and extend variational inference algorithms for the class of latent factor models. These contributions are discussed in more detail below.

3. Initialization Strategies using Spectral Methods
   Prior work on variational algorithms has generally not discussed the important role of initialization. I draw on

\(^1\)This contribution is about model design and is not further discussed in this appendix. Estimation is treated in the main paper Section 4.3.2 and follows straightforwardly from Algorithms 1-6 below.
recently developed spectral methods for parameter estimation to develop strong initialization strategies for the model.

Taken together the above contributions make the latent factor model a practical approach to modeling heterogeneity in social science data.

In the next section I provide a summary of technical contributions as well as a comparison to existing software implementations. In the sections that follow I summarize the six estimation algorithms that are used throughout the paper. They are briefly summarized in Table 6.1. Algorithms 1 and 2 are direct translations of Lee and Wand (2014) but are detailed here because they serve as building blocks for the more complicated models in Algorithms 3-6.

<table>
<thead>
<tr>
<th># of Latent Factors</th>
<th>1</th>
<th>2 (Matrix)</th>
<th>3+ (Tensor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Reg.</td>
<td>Alg 1</td>
<td>Alg 3*</td>
<td>Alg 5*</td>
</tr>
<tr>
<td>Logit</td>
<td>Alg 2</td>
<td>Alg 4*</td>
<td>Alg 6*</td>
</tr>
</tbody>
</table>

*Table 6.1: Algorithms detailed below. Those marked with * are new to this paper.*

**Brief Reminder:**

In the sections that follow I assume a general familiarity with the main text of the latent factor regressions paper. The paper uses primarily two examples which have the following salient details:

1. Foreign Direct Investment (FDI) (Büthe and Milner 2008)
   A time-series cross sectional dataset with 118 countries and 31 years. The panels are unbalanced with approximately 30% of the entries missing.

2. Democratic Peace (Ward, Siverson and Cao 2007)
   A directed-dyadic time dataset with 165 source countries, 165 receiver countries and 10 time periods.

**6.1.1 Summary of Technical Contributions**

The goal of this section is clarify the technical contributions of the latent factor regressions project before proceeding to the specific algorithms. Both the model design and estimation strategy draw from existing components in the literature, but combine them in novel ways. This work is distinguished from prior work by the development of a more general framework and attention to the demands of applied data analysis.

Estimation in the latent factor regression framework is the combination of estimation strategies for three components of the model: the matrix/tensor factorization component, the GLM/regression component and the prior structure.
Each of these three has been considered in the variational inference literature although never in combination. Variational approaches to matrix factorization (Lim and Teh 2007) and tensor factorization (Zhao, Zhang and Cichocki 2014) have been developed for a squared loss function. Separate work has considered generalized linear models, such as logistic regression (Jaakkola and Jordan 2000), and varying intercept/coefficient hierarchical extensions (Lee and Wand 2014). Finally, variational inference algorithms have been proposed for various prior structures such as the Half-Cauchy and Scaled Inverse-Wishart priors (Wand et al. 2011; Huang and Wand 2013) both of which I use in the main text.

Algorithms 1-6 of the latent factor regression cover the estimation strategy. These are summarized in Table 6.1 and are briefly summarized below before being more extensively covered in later sections.

Algorithms 1 and 2 (used in simulations) are direct translations of previous work done in Matt Wand’s research group (Menictas and Wand 2013; Wand 2014; Lee and Wand 2014). They are included because they serve as useful building blocks for the models which include latent factors. Algorithms 3/4 involve the combination of Algorithms 1/2 with prior work on variational algorithms for Gaussian matrix factorization (Lim and Teh 2007). The combination of the matrix factorization with observed covariates is novel and thus required new derivations. Furthermore to the best of my knowledge there had not been a simple variational treatment for the binary outcome matrix factorization model (even without observed covariates). Finally Algorithms 5 and 6 (hierarchical linear regression and logistic regression models with tensor factorization components) involve the combination of prior work on Gaussian tensor factorization (Zhao, Zhang and Cichocki 2014) with Algorithms 1 and 2. Here again the combination of the factorization models with regressions required new derivations both due to the inclusion of covariates as well as the extension to binary outcomes.

Use of alternative prior structures involve a fairly direct translation of work in Wand et al. (2011) and Huang and Wand (2013). These had not yet been combined with factorization models however doing so poses no major technical challenges.

One of the challenges in the use of variational inference algorithms is the presence of many local optima in the objective function. Although initialization of variational algorithms is rarely discussed, it is extremely important in this particular case. I address this by using recent spectral estimators for the parameters. For the matrix case I leverage the

---

2The work of Wand’s group extends prior results particularly Jaakkola and Jordan (2000) and Jordan et al. (1999) to address various practical concerns in implementation.

3Some caveats are in order here. Notably Salter-Townshend and Murphy (2013) and Bailey Fosdick’s dissertation Fosdick (2013) contain variational algorithms for binary outcomes with Gaussian latent factors. However neither has closed form updates due to nonconjugacy resorting to either gradient descent (Salter-Townshend and Murphy 2013) or Gibbs sampling (Fosdick 2013).

4I know of no variational algorithms for binary outcome tensor factorization models. Instead recent work has approached this computational problem by using various tricks to speed up Gibbs sampling (Rai et al. 2014).
results of (Nakajima et al. 2013) which establish a direct connection between a truncated singular value decomposition and the global variational solution for fully-observed Gaussian matrix factorization. Despite the obvious implications for initializing variational inference I’ve seen no prior work that leverages this connection.\(^5\)

**Comparison to Existing Implementations**

Despite a plethora of previous articles there are very few available implementations of existing methods. This dramatically limits their use in applied work. Here I highlight the only publicly available tools for estimating related models. Each is designed to a particular task that is too narrow to accommodate many applications in the social sciences (including the two applications in the paper).

Peter Hoff’s group has released software which covers the matrix factorization case with observed covariates where the outcome matrix is symmetric (*eigenmodel* R package (Hoff 2012), *amen* R package (Hoff et al. 2014)). These packages are designed for the analysis of (small) undirected networks and use MCMC methods which can be quite slow to converge. They can handle some missingness in the outcome through imputation within the model. These packages are extremely effective for their intended purpose but fail for large data and non-network settings. For example in the FDI example from the main paper the data does not form an undirected network and thus cannot be modeled using either of the aforementioned packages. The democratic peace application can be treated as a collection of 10 networks (one for each time point) but cannot be modeled together.

Bada and Liebel (Bada and Liebl 2014) have released an R package for panel data analysis (*phtt*) which includes an implementation of the interactive fixed effects framework described in Bai (2009). In contrast to the Bayesian estimators considered here this uses maximum likelihood which requires a parameterization of the fixed effects that make the parameter estimates order dependent.\(^6\) It also crucially requires balanced panels (i.e. that all cross-sections contain the same number of observations) and only allows for Gaussian outcomes. While the estimators are substantially faster than the network models, the balanced panel restriction is an extremely demanding requirement in practice. For the FDI analysis the 118 country panel has approximately 30% of the cells missing and would require a substantial drop in either the number of countries or the number of time periods to be estimable under the interactive fixed effects framework. This is a typical problem for social science data particularly in international relations and comparative

\(^5\)Although a similar strategy has been used in related areas (Zhang et al. 2014) including my own work on topic models (Roberts, Stewart and Tingley Forthcoming). See also Seeger and Bouchard (2012) which uses the Nakajima and Sugiyama (2011) estimator within an EM algorithm as a means of updating the parameters.

\(^6\)That is, the parameter estimate for a country is different if its listed first rather than last. This is less problematic for a naturally ordered set of groups such as time but is more annoying for unordered groups such as countries.
politics.\textsuperscript{7}

In both the network and panel data models the number of latent factors must be set by hand. This is a huge practical obstruction as it requires information from the analyst that they are ill-prepared to provide (having little information about what the factors are). This is slightly less problematic in the panel data setting where the speed of the estimator makes it possible to simply run the model many times and use model fit statistics to adjudicate amongst the solutions. By contrast, my approach integrates selection of the number of latent factors into the model itself.

It is worth emphasizing that all three packages do an excellent job for the types of data for which they were designed. However, my observation in the main text is that the same model structure is applicable to a broader range of problems. It is to these newer applications that the existing software implementations are not well suited.

Surprisingly there aren’t even any variational implementation of hierarchical linear and generalized linear models. There are several \texttt{R} packages for MCMC algorithms such as \texttt{MCMCpack} but fast alternatives are limited to quasi-likelihood methods and Laplace approximations such as provided by \texttt{lme4}. Thus even implementation of Algorithms 1 and 2 (which are not novel in themselves) constitutes a useful contribution to the research community. Implementations which are sufficiently scalable to accommodate political science data with hundreds of groups and thousands of observations are made possible by the computational strategies described in Lee and Wand (2014) for which my software will be the first publicly available implementation.\textsuperscript{8}

6.1.2 Algorithm 1: Hierarchical Gaussian Linear Models

Algorithm 1 for hierarchical Gaussian linear models is a direct translation of Algorithms 1 and 2 in Lee and Wand (2014). It serves as a core building block for the later algorithms as well.

Preliminaries

In the case of a single mode problem the latent factor regression framework reduces to hierarchical modeling. Algorithm 1 covers the particular case of a Gaussian likelihood with varying intercepts and slopes and weakly informative

\textsuperscript{7}It is worth noting that the problem may arise even when data is not, strictly speaking, “missing.” For example when a new country is created it enters the dataset at a particular time. The previous years aren’t “missing” because it is an ill-defined quantity. However, the interactive fixed effects framework will still fail to work in this instance.

\textsuperscript{8}The algorithms in Lee and Wand (2014) are non-trivial to implement deriving mostly from careful inversion of particular types of sparse matrices. Thus a practitioner is unlikely to find the paper and implement the methods on their own.
priors. With groups indexed by $g$ the model is given by

$$y|\beta, u \sim \text{Normal}(X\beta + Zu, \sigma_e^2)$$ (6.1)

$$u_g|\Sigma^R \sim \text{Normal}(0, \Sigma^R)$$ (6.2)

where $X$ collects the covariates with globally shared effects and $Z$ is a block diagonal matrix over groups containing effects which are group specific. The positive definite covariance matrix $\Sigma^R$ captures the covariance across the group-specific effects. Note that the $R$ superscript is only a notational reminder that these are the covariances of the random effects.

With conjugate priors for $\beta, \sigma^2, \Sigma^R$ the entire model is conditionally conjugate which significantly simplifies inference,

$$\sigma_e^2 \sim \text{Inverse-Gamma}(a_\epsilon, b_\epsilon)$$ (6.3)

$$\Sigma^R \sim \text{Inverse-Wishart}(A_{\Sigma^R}, B_{\Sigma^R})$$ (6.4)

$$\beta \sim \text{Normal}(0, \sigma_\beta^2 I_P)$$ (6.5)

where $P$ is the number of columns of $X$ and $\sigma_\beta^2$ is a large value strictly greater than 0.

Non-Conjugate Priors

However, in practice it may be better to use a more weakly informative prior for variance components (Gelman 2006). By using the data augmentation results in Wand et al. (2011) we can adopt non-conjugate priors such as the Half-Cauchy distribution (Gelman 2006) and the scaled inverse Wishart (Huang and Wand 2013).

Wand et al. (2011) shows that the Half Cauchy can be represented by

$$\rho^2_{i,r} \sim \text{Inverse-Gamma}(5, 1/a_{i,r})$$ (6.6)

$$a_{i,r} \sim \text{Inverse-Gamma}(4, 1/A_{i,r}^2)$$ (6.7)
where the marginal distribution for $\rho_{i,r}^2$ is now Half-Cauchy($A_{i,r})$. The multivariate extension of the Half Cauchy distribution is given by Huang and Wand (2013)

$$
\Sigma | a_r^1, \ldots, a_r^q \sim \text{Inverse-Wishart}(\nu + q - 1, 2\nu \text{diag}(1/a_r^1, \ldots, 1/a_r^q)) \quad (6.8)
$$

$$
a_r^1 \ldots a_r^q \sim \text{Inverse-Gamma}(0.5, 1/A_{r,r}^2) \quad (6.9)
$$

where $\nu$ is a parameter set by the user. When $\nu = 2$ the correlation parameters have uniform distributions over (-1,1) and the standard deviations have Half-$t$ distributions with 2 degrees of freedom (Huang and Wand 2013).

**Variational Approximation**

The approximation to the full joint posterior is

$$
p(\beta, u, a^R, a_u, a_\epsilon, \Sigma^R, \sigma^2_u, \sigma^2_\epsilon) \approx q(\beta, u, a^R, a_u, a_\epsilon)q(\Sigma^R, \sigma^2_u, \sigma^2_\epsilon) \quad (6.10)
$$

$$
= q(\beta, u)q(\Sigma^R)q(\sigma^2_u)q(\sigma^2_\epsilon) \prod_{r=1}^{q} q(a_r^R) \prod_{\ell=1}^{L} q(a_u^\ell) \prod_{\ell=1}^{L} q(\sigma^2_u) \quad (6.11)
$$

where in the first line we give the approximate posterior under a minimal product restriction (Menictas and Wand 2013) and the second line follows due to induced factorizations (Bishop et al. 2006; Lee and Wand 2014).

Thus the approximate evidence lower bound can be given as:

$$
\log p(y; q) = E_q \{ \log p(\beta, u, a^R, a_u, a_\epsilon, \Sigma^R, \sigma^2_u, \sigma^2_\epsilon) \} - \log q(\beta, u, a^R, a_u, a_\epsilon, \Sigma^R, \sigma^2_u, \sigma^2_\epsilon) \quad (6.12)
$$

Under standard variational inference theory (Bishop et al. 2006; Grimmer 2010b), the optimal approximating densities to maximize Equation 6.12 for a generic parameter $\theta$ take the form

$$
q(\theta) = \exp(E_{q(\theta)} \log(p(\theta|\text{rest}))) \quad (6.13)
$$

Because the model is in the conjugate exponential family (after data augmentation for the priors) the approximate posteriors are in the same family as their prior distributions. Each approximating family is described below.
Optimal Variational Densities

Algebraic manipulations show these forms to be

\[
\begin{align*}
q(\beta, u) &= \text{Normal}(\mu_{q(\beta, u)}, \Sigma_{q(\beta, u)}) \\
q(\sigma^2_e) &= \text{Inverse-Gamma}(0.5(N + 1), B_{q(\sigma^2_e)}) \\
q(a_e) &= \text{Inverse-Gamma}(1, B_{q(a_e)}) \\
q(\sigma^2_{ui}) &= \text{Inverse-Gamma}(0.5(q^G + 1), B_{q(\sigma^2_{ui})}) \\
q(a_{ui}) &= \text{Inverse-Gamma}(1, B_{q(a_{ui})}) \\
q(a^R) &= \text{Inverse-Gamma}(0.5(\nu + q^R), B_{q(a^R)}) \\
q(\Sigma^R) &= \text{Inverse-Wishart}(\nu + m + q^R - 1, B_{q(\Sigma^R)})
\end{align*}
\]

where \(\mu_{q(\beta, u)}, \Sigma_{q(\beta, u)}\) are the mean and covariance of \(q(\beta, u)\) and the parameters \(B\) are the rate parameters of the various approximating distributions. Estimation proceeds through cyclical coordinate ascent on the variational distributions.

The above algorithm only works well when the number of groups is not too large. To make the algorithm scalable I follow Lee and Wand (2014) in defining a streamlined algorithm which uses some matrix algebra tricks to allow for effective inversion of the large matrices. For this we partition parameters into groups \(G\) and \(R\) where the \(R\) group are the random intercepts and coefficients for a large number of groups and the \(G\) parameters collect any remaining parameters. Further define the matrix \(C^G = [XZ^G]\).

This allows us to use the following sequence of updates
\( G_i \leftarrow \mu_{q(1/\sigma^2)}(C_i^G)^T X_i^R \) \hspace{1cm} (6.14)

\( H_i \leftarrow [\mu_{q(1/\sigma^2)}(X_i^R)^T (X_i^R) + M_{q(2\gamma_i+1)}]^{-1} \) \hspace{1cm} (6.15)

\( S \leftarrow S + G_i H_i G_i^T \) \hspace{1cm} (6.16)

\( s \leftarrow s + G_i H_i (X_i^R)^T y_i \) \hspace{1cm} (6.17)

\[ \Sigma_{(\beta, u^2)} \leftarrow \mu_{q(1/\sigma^2)}(C_i^G)^T C_i^G + \begin{bmatrix} \sigma_\beta^2 & 0 \\ 0 & \text{blockdiag}(\mu_{q(1/\sigma^2)} I_{q_i^G}) \end{bmatrix} - S \] \hspace{1cm} (6.18)

\[ \mu_{q(\beta, u^2)} \leftarrow \mu_{q(1/\sigma^2)} \Sigma_{q(\beta, u^2)} \left\{ (C_i^G)^T y - s \right\} \] \hspace{1cm} (6.19)

\[ \Sigma_{(u^2)} \leftarrow H_i + H_i G_i^T \Sigma_{q(\beta, u^2)} G_i H_i \] \hspace{1cm} (6.20)

\[ \mu_{q(\beta)} \leftarrow H_i \left\{ \mu_{q(1/\sigma^2)}(X_i^R)^T y_i - G_i^T \mu_{q(\beta, u^2)} \right\} \] \hspace{1cm} (6.21)

\[ B_{q(\sigma^2)} \leftarrow \mu_{q(1/\sigma^2)} + 5 \begin{bmatrix} \left\| y - C_i^G \mu_{q(\beta, u^2)} \right\| \\ \vdots \\ \left\| X_i^R \mu_{q(\beta, u^2)} \right\| \\ \left\| X_m^R \mu_{q(\beta, u^2)} \right\| \end{bmatrix} + \text{tr}\left( (C_i^G)^T C_i^G \Sigma_{q(\beta, u^2)} \right) + \sum_{i=1}^m \text{tr}\left( (X_i^R)^T X_i^R \Sigma_{q(\beta, u^2)} \right) \]

\[ - 2\mu_{q(1/\sigma^2)} \sum_{i=1}^m \text{tr}\left( G_i H_i G_i^T \Sigma_{q(\beta, u^2)} \right) \] \hspace{1cm} (6.22)
Chapter 6. Appendix

\[
\mu_{q(1/\sigma^2)} \leftarrow 5(\sum_{i=1}^{m} n_i + 1)/B_{q(\sigma^2)} \quad (6.23)
\]

\[
\mu_{q(1/a)} \leftarrow 1/\{\mu_{q(1/\sigma^2)} + A_{e}^{-2}\} \quad (6.24)
\]

\[
B_{q(d^2)} \leftarrow \nu \left(M_{q(\Sigma^{-1})} + A_{Rr}^{-2}\right) \quad (6.25)
\]

\[
\mu_{q(1/a^2)} \leftarrow 5(\nu + q^R)/B_{q(d^2)} \quad (6.26)
\]

\[
B_{q(\Sigma^2)} \leftarrow \sum_{i=1}^{m} \left(\mu_{q(d^2)} \Sigma_{q(d^2)} + \Sigma_{q(d^2)}\right) + 2\nu \text{diag} \left(\mu_{q(1/a^2)}, \ldots, \mu_{q(1/a''_{q^R})}\right) \quad (6.27)
\]

\[
M_{q(\Sigma^{-1})} \leftarrow (\nu + m + q^R - 1)B_{q(\Sigma^{-1})}^{-1} \quad (6.28)
\]

\[
\mu_{q(1/a_{\ell}^2)} \leftarrow 1/\{\mu_{q(1/\sigma_{\ell}^2)} + A_{u_{\ell}}^{-2}\} \quad (6.29)
\]

\[
\mu_{q(1/(\sigma_{\ell}^2 + 1))} \leftarrow \frac{q_{\ell}^G + 1}{2\mu_{q(1/a_{\ell}^2)} + \left|\mu_{q(1/a_{\ell}^2)}\right|^2 + \text{tr}(\Sigma_{q(d^2)})} \quad (6.30)
\]

Algorithm 1

With the updates given we can now state algorithm 1. Numbers in parentheses indicate the equation number for the update.

1: repeat
2: \(S \leftarrow 0\)
3: \(s \leftarrow 0\)
4: for \(i = 1 \ldots m\) do
5: \(\text{Update } G_i (6.14), \text{ Update } H_i (6.15)\)
6: \(\text{Update } S (6.16), \text{ Update } s (6.17)\)
7: end for
8: Update \(\Sigma_{q(\beta, u^2)} (6.18), \text{ Update } \mu_{q(\beta, u^2)}\) using (6.19)
9: for \(i = 1 \ldots m\) do
10: \(\text{Update } \Sigma_{q(d^2)} (6.20), \text{ Update } \mu_{q(d^2)} (6.21)\)
11: end for
12: Update \(B_{q(\sigma^2)} (6.22)\)
13: Update \(\mu_{q(1/\sigma^2)} (6.23)\)
14: Update \(\mu_{q(1/a)} (6.24)\)
15: for \( r = 1, \ldots, q^R \) do
16: Update \( B_{q(x^r)} \) (6.25), Update \( \mu_{q(1/x^r)} \) (6.26)
17: end for
18: Update \( B_{q(2^q)} \) (6.27)
19: Update \( M_{q(2^q)}^{-1} \) (6.28)
20: for \( \ell = 1 \ldots L \) do
21: Update \( \mu_{q(1/\sigma^L)} \) (6.29), Update \( \mu_{q(1/\sigma^L)} \) (6.30)
22: end for
23: until convergence in \( p(y; q) \)

Computation

A few quick notes that help to speed implementation in practice:

- Use R’s native recycling to avoid matrix multiplication with a diagonal matrix.
- The matrix inverses all involve matrices which are guaranteed to be positive definite and thus can be inverted quickly through the Cholesky decomposition.
- Many of the operations particularly \((C^G)^T(C^G)\) can be cached.
- Many of the symmetric matrices can be computed more rapidly using R’s crossprod function to avoid computing both off-diagonals.

6.1.3 Algorithm 2: Hierarchical Logistic Regression

Algorithm 2 is a direct translation of Algorithm 3 in Lee and Wand (2014). Again it serves as a useful building block for later algorithms. Minor changes are made to the notation and presentation to fit the context.

Variational Approximation

In logistic regression, a Bernoulli likelihood over \( y \in \{-1, 1\} \) is parameterized by the sigmoid (inverse-logit) function of the parameters:

\[
P(y|\eta) = \sigma(\eta) \quad (6.31)
\]
where $\eta$ is the linear predictor and $\sigma$ is the sigmoid function $\frac{1}{1+e^{-\eta}}$.

The log-likelihood is then

$$\log p(y) = \sum_n \log(\sigma(y_n\eta_n)) \quad (6.32)$$

However this leads to an intractable expectation in the variational approximation. Instead I introduce an additional local variational bound on the marginal likelihood. Following Jaakkola and Jordan (2000) I approximate the sigmoid term using a quadratic lower bound such that

$$\sigma(y\eta) \geq \sigma(\xi)\exp\left(\frac{(y\eta - \xi)}{2} - \lambda(\xi)\left((y\eta)^2 - \xi^2\right)\right) \quad (6.33)$$

$$\lambda(\xi) = \tanh(\xi/2)/(4\xi) \quad (6.34)$$

which introduces a new variational parameter $\xi$ for each data point. The bound is tight at the optimal value of $\xi$. With the introduction of the parameters $\xi$ the data likelihood is now a quadratic function of the parameters to be optimized and thus we get a normal variational distribution for our regression coefficients with closed form mean and variances. $\lambda(\xi)$ ends up playing the role of inverse error variances in a regression style update.

Jaakkola and Jordan (2000) show that the optimal values of the variational parameters can also be solved in closed form by

$$\xi = \sqrt{E[\eta^2]} \quad (6.35)$$

$$= \sqrt{\text{diagonal}\left( C^{\frac{1}{2}}(\Sigma_q(\beta,\eta,\xi) + \mu_q(\beta,\eta,\xi)\mu_q^T(\beta,\eta,\xi))C^{\frac{1}{2}} \right)} \quad (6.36)$$

Thus the entire procedure contains only closed form updates and does not need to resort to numerical optimization. Because the approximation to the sigmoid function is a lower bound, the Evidence Lower Bound is still a true lower bound on $\log(p(y))$.

The justification of Jaakkola and Jordan (2000) is based on constructing a lower bound for the marginal likelihood using convex duality. Additionally, recent work by Scott and Sun (2013) has given a probabilistic interpretation showing the connection to data augmentation using the Polya-Gamma latent variable family (Polson, Scott and Windle 2013).

---

9 Although this representation is less standard in the social sciences, the symmetric form of the likelihood simplifies the notation below.
Optimal Densities

Conditional latent variables $\xi$ we get a normal density for $q(\beta, u)$ which means that optimization proceeds much as in Algorithm 1. I define some new terms (in the same style as above) before defining Algorithm 2.

$$G_i \leftarrow 2(C^G_i)^\top \text{diag}(\Lambda(\xi_i)) X_i^R$$ (6.37)

$$H_i \leftarrow [2(X_i^R)^\top \text{diag}(\Lambda(\xi_i))(X_i^R) + M_{q(\beta,u)}]^{-1}$$ (6.38)

$$S \leftarrow S + G_i H_i G_i^\top$$ (6.39)

$$s \leftarrow s + G_i H_i X_i^R (y_i - .5)$$ (6.40)

$$\Sigma_{q(\beta,u)} \leftarrow 2(C^G)^\top \text{diag}(\Lambda(\xi_i))C^G + \begin{pmatrix} \sigma^2_{\beta} & 0 \\ 0 & \text{blockdiag}(\mu_{q(1/\sigma^2_u)}I_{d_u}) \end{pmatrix} - S $$ (6.41)

$$\mu_{q(\beta,u)} \leftarrow \mu_{q(1/\sigma^2_u)}\Sigma_{q(\beta,u)} \left\{ (C^G)^\top (y_i - .5) - s \right\}$$ (6.42)

$$\Sigma_{q(u)} \leftarrow H_i + H_i G_i^\top \Sigma_{q(\beta,u)} G_i H_i$$ (6.43)

$$\mu_{q(u)} \leftarrow H_i \left\{ \mu_{q(1/\sigma^2_u)}(X_i^R)^\top (y_i - .5) - G_i^\top \mu_{q(\beta,u)} \right\}$$ (6.44)

$$\xi_i^2 \leftarrow \text{diagonal} \left\{ C^G(\Sigma_{q(\beta,u)} + \mu_{q(\beta,u)}\mu_{q(\beta,u)}^\top)(C^G)^\top \right\}$$ (6.45)

$$\xi_i^2 \leftarrow 2\text{diagonal} \left\{ C^G \left( -\Sigma_{q(\beta,u)} G_i + \mu_{q(\beta,u)}\mu_{q(\beta,u)}^\top \right) (X_i^R)^\top \right\} + \text{diagonal} \left\{ X_i^R \left( \Sigma_{q(u)} + \mu_{q(u)}\mu_{q(u)}^\top \right) (X_i^R) \right\}$$ (6.46)

$$B_{q(u)} \leftarrow \nu \left( M_{q(\beta,u)} \right)_{rr} + A_{Rr}^{-2}$$ (6.47)

$$\mu_{q(1/u)} \leftarrow .5(\nu + q_R^R) / B_{q(u)}$$ (6.48)

$$B_{q(\xi)} \leftarrow \sum_{i=1}^m \left( \mu_{q(u)}\mu_{q(u)}^\top + \Sigma_{q(u)} \right) + 2\text{diag} \left( \mu_{q(1/u)}^2, \ldots, \mu_{q(1/u)}^2 \right)$$ (6.49)

$$M_{q(\xi)^{-1}} \leftarrow (\nu + m + q_R^R - 1)B_{q(\xi)}^{-1}$$ (6.50)

$$\mu_{q(1/a_u)} \leftarrow 1 / \left( \mu_{q(1/\sigma^2_u)} + A_{u}^{-2} \right)$$ (6.51)

$$\mu_{q(1/\sigma^2_u)} \leftarrow \frac{q_R^2 + 1}{2\mu_{q(1/a_u)} + \left| \mu_{q(u)} \right|^2 + \text{tr}(\Sigma_{q(\xi)})}$$ (6.52)
Algorithm 2

With the updates above we can now state Algorithm 2.

1: \textbf{repeat}
2: \hspace{1em} \textbf{S} \leftarrow 0
3: \hspace{1em} s \leftarrow 0
4: \hspace{1em} \textbf{for} i = 1 \ldots m \textbf{do}
5: \hspace{2em} \text{Update } G_i \text{ (6.37)}, \text{ Update } H_i \text{ (6.38)}
6: \hspace{2em} \text{Update } S \text{ (6.39)}, \text{ Update } s \text{ (6.40)}
7: \hspace{1em} \textbf{end for}
8: \hspace{1em} \text{Update } \Sigma_{q(\beta, u; \xi)} \text{ (6.41)}, \text{ Update } \mu_{q(\beta, u; \xi)} \text{ using (6.42)}
9: \hspace{1em} \textbf{for} i = 1 \ldots m \textbf{do}
10: \hspace{2em} \text{Update } \Sigma_{q(u_R^i; \xi)} \text{ (6.43)}, \text{ Update } \mu_{q(u_R^i; \xi)} \text{ (6.44)}
11: \hspace{1em} \textbf{end for}
12: \hspace{1em} \text{Update } \xi^2 \text{ (6.45)}
13: \hspace{1em} \textbf{for} i = 1 \ldots m \textbf{do}
14: \hspace{2em} \text{Update } \xi_{i}^2 \text{ (6.46)}
15: \hspace{1em} \textbf{end for}
16: \hspace{1em} \textbf{for} r = 1, \ldots, q^R \textbf{do}
17: \hspace{2em} \text{Update } B_{q(u^r)} \text{ (6.47)}, \text{ Update } \mu_{q(1/a^r)} \text{ (6.48)}
18: \hspace{1em} \textbf{end for}
19: \hspace{1em} \text{Update } B_{q(2^r)} \text{ (6.49)}
20: \hspace{1em} \text{Update } M_{q(\Sigma^{-1})} \text{ (6.50)}
21: \hspace{1em} \textbf{for} \ell = 1 \ldots L \textbf{do}
22: \hspace{2em} \text{Update } \mu_{q(1/a_{u \ell})} \text{ (6.51)}, \text{ Update } \mu_{q(1/\sigma^2_{u \ell})} \text{ (6.52)}
23: \hspace{1em} \textbf{end for}
24: \hspace{1em} \textbf{until} convergence in } p(y; q)
6.1.4 Algorithm 3: Gaussian Outcome Matrix Factorization

In Algorithm 3 I show how to connect the Gaussian linear regression in Algorithm 1 with a matrix factorization model. The model can be stated as:

\[
y_{i,j} \sim \text{Normal}(x_{i,j}\beta + Z_{i,j}\gamma + \sum_k u_{i,k}v_{j,k}, \sigma^2_e) \tag{6.53}
\]

\[
u_{i,k} \sim \text{Normal}(0, \rho^2_k) \tag{6.54}
\]

\[
v_{i,k} \sim \text{Normal}(0, \tau^2_k) \tag{6.55}
\]

\[
\beta \sim \text{Normal}(0, \sigma^2_{\beta}I_p) \tag{6.56}
\]

\[
\gamma \sim \text{Normal}(0, \Sigma^R_{\gamma}) \tag{6.57}
\]

\[
\Sigma^R_{\gamma} | a^R_{\gamma}, \ldots, a^R_{\gamma} \sim \text{Inverse-Wishart}(v + q^R - 1, 2v\text{diag}(1/a^R_1, \ldots, 1/a^R_q)) \tag{6.58}
\]

\[
a^R_1 \ldots a^R_q \text{ ind.} \sim \text{Inverse-Gamma}(5, 1/A^2_{ae}) \tag{6.59}
\]

\[
\sigma^2_e \sim \text{Inverse-Gamma}(5, 1/a_e) \tag{6.60}
\]

\[
a_e \sim \text{Inverse-Gamma}(4, 1/A^2_e) \tag{6.61}
\]

where I have switched the notation of the random effect to \(\gamma\) to reserve \(u\) as one of the latent factors.

In order to get the dimensionality selection effects of Automatic Relevance Determination we will point estimate the variances \(\rho^2, \tau^2\) as explained below.

When notationally convenient I collect the latent factors \(u\) into a matrix \(U\) where each row \(i\) contains the \(k\) factors for group \(i\). We denote the row of matrix \(U\) contain the latent factors of group \(i\) as \(U_i\). \(V\) follows similarly.

Matrix Factorization Component

Following the computer science literature (Lim and Teh 2007), I assume a factorization over the latent factors:

\[
q(U, V, \beta) = q(U)q(V)q(\beta) \tag{6.62}
\]

\[
= \prod_{i=1}^I q(U_i) \prod_{j=1}^J q(V_j)q(\beta, \gamma) \tag{6.63}
\]

Note that this is not a minimal product restriction on the variational parameters as either \(q(U)\) or \(q(V)\) could be combined with \(q(\beta, \gamma)\) but I separate them in order to keep the treatment of the two modes symmetric.
The consequence of the factorization assumption is that the approximation is unable to capture the posterior covariance between the latent factor matrices $q(U)$ and $q(V)$. In the true posterior these effects are going to be negatively correlated, and it indeed it is exactly this feature which makes Gibbs sampling challenging. This hurts the accuracy of the approximation and will in general cause the approximation to understate the variance. That said, this does not appear to substantially detract from the quality of the approximation for the other parameters $q(\beta)$.

Standard calculations lead to the following Gaussian forms of the approximate densities:

$$
q(U) = \text{Normal}(\mu_{q(U)}, \Sigma_{q(U)}) \quad (6.64)
$$

$$
q(V) = \text{Normal}(\mu_{q(V)}, \Sigma_{q(V)}) \quad (6.65)
$$

$$
q(\beta) = \text{Normal}(\mu_{q(\beta)}, \Sigma_{q(\beta)}) \quad (6.66)
$$

The posterior parameters of the approximation are updated as

$$
\Sigma_{q(U)} = \begin{pmatrix}
\frac{1}{\tau_1^2} & 0 & \cdots & 0 \\
0 & \frac{1}{\tau_2^2} & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & \cdots & \cdots & \frac{1}{\tau_k^2}
\end{pmatrix} + \sum_{j=1}^J \frac{\Sigma_{q(V_j)} + \mu_{q(V_j)} \mu_{q(V_j)^T}}{\sigma_e^2} \quad (6.67)
$$

$$
\mu_{q(U)} = \Sigma_{q(U)} \left( \sum_{n \in \Omega} \frac{(y_n - x_n \beta)}{\sigma_e^2} \right) \quad (6.68)
$$

where $\Omega$ indicates the set of observations for which $y$ is observed. The form of $q(V)$ following analogously. Although the form seems complicated at first, it is simply Bayesian linear regression with two distinctions. First, we are now fitting the model to the residuals $(y - x \beta)$ and second we have to include the covariance of the variational distribution when calculating the cross products.

The variational distribution for $\beta$ is even simpler as it corresponds directly to Bayesian linear regression on the residuals

$$
\tilde{y}_{ij} = y_{ij} - E[U_i] E[V_j^T] \quad (6.69)
$$

$$
= y_{ij} - \mu_{q(U)} \mu_{q(V_j)^T} \quad (6.70)
$$

10Note that in practice I always include standard additive effects for the rows and columns of the matrix into the components $\beta, \gamma$. This has the benefit of weakening the dependence on the factorization assumption.
Note that this has a tractable form due to the factorization assumption that defines $E[U_i, V_j^T] = E[U_i]E[V_j^T]$.

**Optimal Densities**

Most of the optimal densities follow by replacing the outcome $y$ with a working response. In addition the following three updates are required for algorithm 3.

$$
\tau^2_k = \frac{1}{T-1} \sum_{i=1}^T \left( \Sigma_{q(U_i)k} + (\mu_{q(U_i)})^T \mu_{q(U_i)} \right) \tag{6.71}
$$

$$
\rho^2_k = \frac{1}{J-1} \sum_{j=1}^J \left( \Sigma_{q(V_j)k} + (\mu_{q(V_j)})^T \mu_{q(V_j)} \right) \tag{6.72}
$$

The update for the error rate parameter is substantially more complicated due to the inclusion of the latent factors. It is useful to define it in several pieces, using $\eta_{C\beta}$ to indicate the fitted portion of the linear predictor due to the observed parameters and additive effects and $\eta_{UV^T}$ to indicate the fitted portion due to the latent factors.

$$
\eta^{(C\beta)} = y - C^G \mu_{q(\beta, \gamma^c)} - \begin{bmatrix} X_1^R \mu_{q(\gamma^c_1)} \\ \vdots \\ X_m^R \mu_{q(\gamma^c_m)} \end{bmatrix}
$$

$$
\eta_{i,j}^{(UV)} = \mu_{q(U_i)} \mu_{q(V_j)}^T
$$

Also define a term $\zeta$ to capture a portion of the regression context.

$$
\zeta_{C\beta} = \text{tr}\left( (C^G)^T C^G \Sigma_{q(\beta, \gamma^c)} \right) + \sum_{i=1}^m \text{tr}\left( (X_i^R)^T X_i^R \Sigma_{q(\gamma^c)} \right) - 2 \mu_{q(1/\sigma^2)}^{-1} \sum_{i=1}^m \text{tr}\left( G_i H_i G_i^T \Sigma_{q(\beta, \gamma^c)} \right)
$$

We also define a term to capture a piece from the matrix factorization component. Here we double index $y$ as though it is arranged into a matrix.

$$
\zeta_{UV^T} = \sum_{i,j \in \Omega} y_{i,j} - 2y_{i,j} \mu_{q(U_i)} \mu_{q(V_j)}^T + \text{tr}\left[ \left( \Sigma_{q(U_i)} + \mu_{q(U_i)} \mu_{q(V_j)}^T \right) \left( \Sigma_{q(V_j)} + \mu_{q(V_j)} \mu_{q(U_i)}^T \right) \right]
$$

The trace term can be efficiently computed because both matrices are symmetric and thus the trace is the sum over
their elementwise product:

\[ \text{tr}(AB) = \sum_{ij} A_{ij}B_{ij} \]

With these components together we can write the update as

\[
B_{q(i)}^{(2)} \leftarrow \zeta UV^T - 2 \left( \sum_{i,j \in \Omega} \left( y_{ij} - \eta_{ij}(UV^T) \right) \eta_{ij}(CM) \right) + \left\| \eta_{ij}(CB) \right\|^2 + \zeta C\beta \quad (6.73)
\]

**Algorithm 3**

Updates after \( y_{\text{working}} \) should use the working version of the response in place of \( y \).

1: repeat
2: \( S \leftarrow 0 \)
3: \( s \leftarrow 0 \)
4: \( y_{\text{working}} \leftarrow y - \eta^{(UV)} \)
5: for \( i = 1 \ldots m \) do
6: \hspace{1em} Update \( G_i \) (6.14), Update \( H_i \) (6.15)
7: \hspace{1em} Update \( S \) (6.16), Update \( s \) (6.17)
8: end for
9: Update \( \Sigma_{q(i)^{y\varphi}} \) (6.18), Update \( \mu_{q(i)^{y\varphi}} \) using (6.19)
10: for \( i = 1 \ldots m \) do
11: \hspace{1em} Update \( \Sigma_{q(i)^{y\varphi}} \) (6.20), Update \( \mu_{q(i)^{y\varphi}} \) (6.21)
12: end for
13: Update \( y_{\text{working}} \leftarrow y - \eta^{(C\beta)} \)
14: for \( i = 1 \ldots I \) do
15: \hspace{1em} Update \( \Sigma_{q(t_i)} \) (6.67)
16: \hspace{1em} Update \( \mu_{q(t_i)} \) (6.68)
17: end for
18: for \( j = 1 \ldots J \) do

154
 Initialization Methods

Due to the multimodality in the posterior distribution it is helpful to initialize the variational algorithm carefully. Here and in Algorithm 4-6, I initialize by updating the coefficients $\beta, \gamma$ and then using the spectral algorithm of Nakajima et al. (2013) to estimate $q(U)q(V)$ from $y - E[X\beta + Z\gamma]$.

In short, the Nakajima et al. (2013) approach involves using a truncated singular value decomposition to estimate the parameters of the variational posterior. In the case of imbalanced panels this requires filling in the missing elements of the matrix. Using $\Omega$ to denote the observed indices, I use a simple mean imputation:

$$Y_{\Omega} \leftarrow E[Y_{\Omega}] \quad (6.74)$$

This is the same procedure as used in Chatterjee (2012) which provides a theoretical justification for the choice. An
alternative strategy would be to impute the values using an algorithm such as Soft-Impute (Mazumder, Hastie and Tibshirani 2010) which is itself based on a singular value decomposition. Extensive study of the properties of these estimators is left to future work.

The exact version of the algorithm used is in Nakajima et al. (2012) with supporting details in Nakajima and Sugiyama (2011); Nakajima et al. (2013). Further details will be filled in here as well in future drafts.

Dimensionality Reduction

In algorithms 3-6 I use Automatic Relevance Determination to choose the dimensionality of the latent factors. This involves point estimating the factor variances \( \rho \) and \( \tau \) which produces a model-induced regularization effect (see for example Nakajima et al. (2013) on the origins of this effect). In practice this means estimating the model with the maximal number of latent factors and dropping them out as their variance parameters go to zero.

An added benefit of using the spectral initialization is that we get an initial estimate of the dimensionality which can help defray computational costs. Then as the variances fall below a certain threshold they are dropped from the model.

6.1.5 Algorithm 4: Logistic Regression with Matrix Factorization

Preliminaries

The model can be given as:

\[
y_{i,j} \sim \text{Bernoulli}(\sigma(\eta))
\]

\[
\eta = x_{i} \beta + Z_{i} \gamma + \sum_{k} u_{i,k} v_{j,k}
\]

\[
u_{i,k} \sim \text{Normal}(0, \rho_{k}^2)
\]

\[v_{i,k} \sim \text{Normal}(0, \tau_{k}^2)
\]

\[
\beta \sim \text{Normal}(0, \sigma_{\beta}^2 I_P)
\]

\[
\gamma \sim \text{Normal}(0, \Sigma^{R})
\]

\[
\Sigma^{R} \sim \text{Inverse-Wishart}(\nu + q^{R} - 1, \nu \text{diag}(1/\alpha_{1}^{R}, \ldots, 1/\alpha_{q^{R}}^{R}))
\]

\[
a_{1}^{R} \ldots a_{q^{R}}^{R} \sim \text{Inverse-Gamma}(0.5, 1/A_{\alpha_{1}^{R}}^{R})
\]

156
The logistic regression with matrix factorization problem is slightly more complicated than the Gaussian regression because it is slightly more difficult to define the working response. Using the formulation in Equation 6.33 we can show the working response for the estimation of \( q(\beta, \gamma) \) is \( y/2 - 2y^2 \eta' \lambda(\xi) \) where \( \eta' \) is the portion of the linear predictor to be removed.

**Optimal Densities**

All the optimal densities remain in the same families as given in Algorithms 2 and 3.
\[ G_t \leftarrow 2(C^G_t)^T \text{diag}(\Lambda(\xi_t))X_t^R \]  
(6.83)

\[ H_t \leftarrow [2(X_t^R)^T \text{diag}(\Lambda(\xi_t))(X_t^R) + M_{q(2\beta^y-1)})^{-1} \]  
(6.84)

\[ S \leftarrow S + G_t H_t G_t^T \]  
(6.85)

\[ s \leftarrow s + G_t H_t (X_t^R)^T (y_{\text{working}}) \]  
(6.86)

\[ \Sigma_{q(\beta^y-\xi)} \leftarrow 2(C^G)^T \text{diag}(\Lambda(\xi))C^G + \begin{bmatrix} \sigma^2 & 0 \\ 0 & \text{blockdiag}(\mu_{q(1/\tau^2_{\beta^y})}) \end{bmatrix} - S \]  
(6.87)

\[ \mu_{q(\beta^y-\xi)} \leftarrow \mu_{q(1/\tau^2_{\beta^y})} \Sigma_{q(\beta^y-\xi)} \left( (C^G)^T (y_{\text{working}}) - s \right) \]  
(6.88)

\[ \Sigma_{q(\gamma^y)} \leftarrow H_t + H_t G_t^T \Sigma_{q(\beta^y-\xi)} G_t H_t \]  
(6.89)

\[ \mu_{q(\gamma^y)} \leftarrow H_t \left( (X_t^R)^T (y_{\text{working}}) - G_t^T \mu_{q(\gamma^y)} \right) \]  
(6.90)

\[ \Sigma_{q(U_j)} \leftarrow \Sigma_{q(U_j)} \left( \sum_{j=1}^{J} \frac{\sum_{j=1}^{J} (y_{\text{working}} - x_{q(\beta^y)}) \mu_{q(V_{j,n})}}{\lambda(\xi)} \right) \]  
(6.92)

\[ \xi^2 \leftarrow \text{diagonal} \left[ E[(C^2 + U^2)(C^2 + U^2)^T] \right] \]  
(6.93)

\[ B_{q(\alpha^y)} \leftarrow \nu \left( M_{q(2\beta^y-1)} \right)_{rr} + A_{Rr}^2 \]  
(6.94)

\[ \mu_{q(1/\alpha^y)} \leftarrow .5(\nu + q^R)/B_{q(\alpha^y)} \]  
(6.95)

\[ B_{q(2\alpha^y)} \leftarrow \sum_{i=1}^{m} \left( \mu_{q(\alpha^y)} \mu_{q(\alpha^y)}^T + \Sigma_{q(\alpha^y)} \right) + 2 \text{diag} \left( \mu_{q(1/\alpha^y)}, \ldots, \mu_{q(1/\alpha^y)} \right) \]  
(6.96)

\[ M_{q(2\gamma^y-1)} \leftarrow (\nu + m + q^R - 1)B_{q(2\alpha^y)}^{-1} \]  
(6.97)

\[ \mu_{q(1/m^y)} \leftarrow 1/\left[ \mu_{q(1/\tau^2_{\gamma^y})} + A_{M}^2 \right] \]  
(6.98)

\[ \mu_{q(1/\tau^2_{\gamma^y})} \leftarrow \frac{q^R + 1}{2q_{1/(\alpha^y)} + \left| \mu_{q(1/\alpha^y)} \right|^2 + \text{tr} \Sigma_{q(1/\tau^2_{\gamma^y})}} \]  
(6.99)

\[ \tau^2_k = \frac{1}{T-1} \sum_{i=1}^{J} (\Sigma_{q(U_j)})_{kk} + (\mu_{q(U_j)})^T \mu_{q(U_j)} \]  
(6.100)

\[ \rho^2_k = \frac{1}{T-1} \sum_{j=1}^{J} (\Sigma_{q(V_j)})_{kk} + (\mu_{q(V_j)})^T \mu_{q(V_j)} \]  
(6.101)
Algorithm 4

With the statements above we can now give Algorithm 4

1: repeat
2: \( S \leftarrow 0 \)
3: \( s \leftarrow 0 \)
4: Update \( y_{\text{working}} \leftarrow y/2 - 2y^2(\mu_{q(U)}\mu^T_{q(U)})\lambda(\xi) \)
5: \( \text{for } i = 1 \ldots m \text{ do} \)
6: \( \quad \text{Update } G_i \text{ (6.83), Update } H_i \text{ (6.84)} \)
7: \( \quad \text{Update } S \text{ (6.85), Update } s \text{ (6.86)} \)
8: \( \text{end for} \)
9: Update \( \Sigma_{q(\beta, \gamma; \xi)} \) (6.87, Update \( \mu_{q(\beta, \gamma; \xi)} \) using (6.88)
10: \( \text{for } i = 1 \ldots m \text{ do} \)
11: \( \quad \text{Update } \Sigma_{q(\gamma; \xi)} \) (6.89), Update \( \mu_{q(\gamma; \xi)} \) (6.90)
12: \( \text{end for} \)
13: Update \( y_{\text{working}} \leftarrow y/2 - 2y^2(C\mu_{q(\beta, \gamma)})\lambda(\xi) \)
14: \( \text{for } i = 1 \ldots I \text{ do} \)
15: \( \quad \text{Update } \Sigma_{q(U_i)} \) (6.67)
16: \( \quad \text{Update } \mu_{q(U_i)} \) (6.68)
17: \( \text{end for} \)
18: \( \text{for } j = 1 \ldots J \text{ do} \)
19: \( \quad \text{Update } \Sigma_{q(V_j)} \) (6.67)
20: \( \quad \text{Update } \mu_{q(V_j)} \) (6.68)
21: \( \text{end for} \)
22: Update \( r^2 \) (6.71)
23: Update \( \rho^2 \) (6.72)
24: \( y_{\text{working}} \leftarrow y \)
25: Update \( \xi^2 \) (6.93)
26: \( \text{for } r = 1, \ldots, q^R \text{ do} \)
27: \( \quad \text{Update } B_{q(a^R)} \) (6.94), Update \( \mu_{q(1/a^R)} \) (6.95)

159
6.1.6 Algorithm 5: Gaussian Tensor Factorization

Algorithms 3 covers the case of two types of interactive latent factors. This addresses the case when the data can be arranged into a matrix and adds a matrix factorization to the linear predictor. When more than two types of interactive latent factors are present the data can be arranged into a tensor and a tensor decomposition can be added to the linear predictor (Kolda and Bader 2009). The multilinear form presented in this article corresponds to a type of tensor decomposition called the CANDECOMP/PARAFAC (CP) tensor factorization (Kolda and Bader 2009; Hoff 2011a; Zhao, Zhang and Cichocki 2014).

Notation

Because this model allows for tensors with an arbitrary number of modes it will be useful to change notation. Denote the tensor containing the outcome variable as $Y$. The interactive latent factors will form a tensor of equivalent dimensions which collects their contribution to the linear predictor. Denote this latent tensor $U$. For each mode of the tensor indexed by $m \in 1\ldots M$ there exists a factor matrix $U^{(m)} \in \mathbb{R}^{(n_{m} \times K)}$ where $n_{m}$ is the dimension of the $m$-th mode and $K$ is the dimensionality of the latent factor. A column in this matrix is denoted $u_{k}^{(m)}$. The latent tensor $U$ can be formed by taking the sum over the kroeneker product of the modes such that

$$ U = \sum_{k=1}^{K} u_{k}^{(1)} \otimes \ldots \otimes u_{k}^{(M)} $$

(6.102)

In general we will work with the collection of matrix representations but it will often be simpler to index terms form the latent tensor.

Observed covariates in the form of $X, Z$ and $C$ are left as capital letters with the understanding that when multiply
indexed they still return a vector as in the standard regression case.

Model

Using the new notation we can state the model as

\[ y_{i,j,...,r} \sim \text{Normal}(X_{i,j,...,r}\beta + Z_{i,j,...,r}\gamma + U_{i,j,...,r}, \sigma^2_{\epsilon}) \] (6.103)

\[ u_k^{(m)} \text{ ind.} \sim \text{Normal}(0, \tau_{m,k}^2) \] (6.104)

\[ \beta \sim \text{Normal}(0, \sigma^2_{\beta} I_P) \] (6.105)

\[ \gamma \sim \text{Normal}(0, \Sigma_R) \] (6.106)

\[ \Sigma_R | a_R^1, \ldots, a_R^q \sim \text{Inverse-Wishart}(\nu + q - 1, 2v\text{diag}(1/a_R^1, \ldots, 1/a_R^q)) \] (6.107)

\[ a_R^k \sim \text{Inverse-Gamma}(0.5, 1/a_R^2) \] (6.108)

\[ \sigma^2_{\epsilon} \sim \text{Inverse-Gamma}(0.5, 1/a_{\epsilon}) \] (6.109)

\[ a_{\epsilon} \sim \text{Inverse-Gamma}(4, 1/A_{\epsilon}^2) \] (6.110)

where I emphasize that the model in Equation 6.104 uses a shared variance for a given mode and factor dimension thus making it an analogous to the matrix case. \( \tau^2 \) is now a matrix which collects these variances for each mode (along the rows) and each dimension of the latent factor (along the columns).

Variational Approximation

Again we factorize the density over the modes and the

\[ q(\beta, U^{(1)} \ldots U^{(M)}) \approx q(\beta) \prod_i q(U^{(i)}) \ldots q(U^{(M)}) \] (6.111)

with induced factorization further factorizing the posterior over the rows of the factor matrices (Zhao, Zhang and Cichocki 2014). Due to the conjugacy in the model these rows are again multivariate Gaussian distributions.

Optimal Densities

In algorithms 3 and 4 the latent factors for mode 1 were updated by a Bayesian linear regression using the mode 2 factors as covariates. Because of the expectations of the quadratic terms these forms also include the covariance of the
variational posterior over the mode 2 factors. Thus to rewrite Equations 6.67 and 6.68 in the current notation. To do so we denote $u_i^{(m)}$ to be the $i$’th row of the factor matrix $U^{(m)}$.

$$
\Sigma_{q(U^{(m)})} \leftarrow \begin{pmatrix}
\begin{array}{cccc}
1/\tau_{1,1} & 0 & \ldots & 0 \\
0 & 1/\tau_{1,2} & \ldots & \ldots \\
\vdots & \vdots & \ddots & \vdots \\
0 & \ldots & \ldots & 1/\tau_{1,k}
\end{array}
\end{pmatrix}^{-1} + \sum_{i=1}^{n_i} \frac{\Sigma_{q(U^{(m)})} + \mu_{q(U^{(m)})}^T \mu_{q(U^{(m)})}}{\sigma_e^2}
$$

(6.112)

$$
\mu_{q(U^{(m)})} \leftarrow \Sigma_{q(U^{(m)})} \left( \sum_{i \in \Omega} \left( y_n - x_n \beta \right) \mu_{q(U^{(m)})} \right)
$$

(6.113)

Notice the numerator in Equation 6.112 contains the terms related to the second mode of the tensor. In the general $M$-mode tensor case this simply becomes an elementwise product. Denote the elementwise product of a collection of matrices as $\odot$ where $m$ denotes the index we are taking the elementwise product over.

$$
\Sigma_{q(U^{(m)})} \leftarrow \begin{pmatrix}
\begin{array}{cccc}
1/\tau_{1,1} & 0 & \ldots & 0 \\
0 & 1/\tau_{1,2} & \ldots & \ldots \\
\vdots & \vdots & \ddots & \vdots \\
0 & \ldots & \ldots & 1/\tau_{1,k}
\end{array}
\end{pmatrix}^{-1} + \sum_{i=1}^{n_i} \odot_{(m)} \left( \Sigma_{q(U^{(m)})} + \mu_{q(U^{(m)})}^T \mu_{q(U^{(m)})} \right)
$$

(6.114)

$$
\mu_{q(U^{(m)})} \leftarrow \Sigma_{q(U^{(m)})} \left( \sum_{i \in \Omega} \left( y_n - x_n \beta \right) \mu_{q(U^{(m)})} \right)
$$

(6.115)

As with the matrix factorization model the noise parameter requires some care. Define $\zeta_{CB}$ as before and introduce the more general form $\zeta_{U}$ as

$$
\zeta_{U} = \sum_{i \in \Omega} |Y_i|^2 - 2Y_i^T U_i + \odot_{(m)} \left( \Sigma_{q(U^{(m)})} + \mu_{q(U^{(m)})}^T \mu_{q(U^{(m)})} \right)
$$

(6.116)

This leads to the update

$$
B_{q(U)} \leftarrow \zeta_{U} - 2 \left( \sum_{i \in \Omega} (Y_i - U_i) C_i \beta \right) + |C_i \beta|^2 + \zeta_{U}$$

(6.117)
Algorithm 5

With these updates in hand we can present algorithm 5.

1: repeat
2: \( S \leftarrow 0 \)
3: \( s \leftarrow 0 \)
4: \( y_{\text{working}} \leftarrow \mathcal{Y}_\Omega - \mathcal{U}_\Omega \)
5: for \( i = 1 \ldots m \) do
6: \( \text{Update } G_i \) (6.14), \( \text{Update } H_i \) (6.15)
7: \( \text{Update } S \) (6.16), \( \text{Update } s \) (6.17)
8: end for
9: \( \text{Update } \Sigma_{q(\beta,\gamma^c)} \) (6.18), \( \text{Update } \mu_{q(\beta,\gamma^c)} \) using (6.19)
10: for \( i = 1 \ldots m \) do
11: \( \text{Update } \Sigma_{q(\gamma^c)} \) (6.20), \( \text{Update } \mu_{q(\gamma^c)} \) (6.21)
12: end for
13: \( \text{Update } y_{\text{working}} \leftarrow \mathcal{Y}_\Omega - C_\beta \)
14: for \( m = 1 \ldots M \) do
15: \( \text{for } i = 1 \ldots N_m \) do
16: \( \text{Update } \Sigma_{q(U_m^i)} \) (6.114)
17: \( \text{Update } \mu_{q(U_m^i)} \) (6.115)
18: \( \text{Update } r_{i,k}^2 \) (6.71)
19: end for
20: end for
21: \( y_{\text{working}} \leftarrow y \)
22: \( \text{Update } B_{q(\sigma^2)} \) (6.117)
23: \( \text{Update } \mu_{q(1/\sigma^2)} \) (6.23)
24: \( \text{Update } \mu_{q(1/\tau_k^2)} \) (6.24)
25: for \( r = 1, \ldots, q^R \) do
26: \( \text{Update } B_{q(\alpha^R)} \) (6.25), \( \text{Update } \mu_{q(1/\alpha^R)} \) (6.26)
27: end for
Initialization

While Nakajima and Sugiyama (2011) provides a direct connection between matrix factorizations and variational bayes there is no such clean theoretical result for the tensor case. Indeed while matrix factorizations are often easy to compute by the singular value decomposition, low-rank tensor decompositions need not even exist and commonly used algorithms for finding them often do not have convergence guarantees (Kolda and Bader 2009). This is an active area of research that I don’t delve into extensively here (but see Hoff (2011a); Anandkumar, Ge and Janzamin (2014a); Suzuki (2014)).

In order to initialize the model I perform the Nakajima et al. (2012) estimator on the matricization of each mode of the tensor. Matricization involves unfolding the tensor to create a matrix in which the rows represent one mode of the tensor and the columns represent all other modes (Kolda and Bader 2009). The spectral estimator can be applied to each matricization and the latent factors for the preserved dimension are then used as initializations for the tensor factorization algorithm. This works well in practice although further study is needed.

Computation

Care must be taken to avoid memory issues with tensor latent variables. Often the tensors are very sparsely observed and thus the complete tensor should not be explicitly formed when at all possible.

6.1.7 Algorithm 6: Logistic Regression Tensor Factorization

Algorithm 6 provides the tensor variant of Algorithm 4. It provides no unique challenges beyond those in moving from Algorithm 3 to Algorithm 5.
Chapter 6. Appendix

Model

We work with the model

\[ Y \sim \text{Bernoulli}(\sigma(\eta)) \] (6.118)

\[ \eta = X\beta + Z\gamma + \mathcal{U} \] (6.119)

\[ u_{(m)}^{(m)} \overset{\text{ind.}}{\sim} \text{Normal}(0, \tau_{m,2}^2) \] (6.120)

\[ \beta \overset{\text{Normal}}{\sim} (0, \sigma_\beta^2 I_P) \] (6.121)

\[ \gamma \overset{\text{Normal}}{\sim} (0, \Sigma_R) \] (6.122)

\[ \Sigma^R | \alpha^R \sim \text{Inverse-Wishart}(\nu + q^R - 1, 2\nu \text{diag}(1/\alpha^R_1, \ldots, 1/\alpha^R_q)) \] (6.123)

\[ \alpha^R_1, \ldots, \alpha^R_q \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(.5, 1/A^2_R) \] (6.124)

Optimal Densities

Again we denote the elementwise product of a collection of matrices as \( \odot \), where \( m \) denotes the index we are taking the elementwise product over. Then the updates for the latent factor matrices are

\[
\begin{align*}
\Sigma^R | d_1^R, \ldots, d_q^R & \sim \text{Inverse-Wishart}(\nu + q^R - 1, 2\nu \text{diag}(1/d_1^R, \ldots, 1/d_q^R)) \\
d_1^R, \ldots, d_q^R & \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(.5, 1/A^2_R)
\end{align*}
\] (6.125)

\[
\begin{align*}
\mu^{(1)} & \sim \text{Normal}(0, \sigma^2 \beta^T) \\
\mu^{(2)} & \sim \text{Normal}(0, \text{diag}(\mu^T \Sigma^R^{-1} \mu))
\end{align*}
\] (6.126)

where \( y \) is replaced with \( y_{\text{working}} \). I’ve omitted the “working” subscript here to avoid confusion with the observation index.

The update for the the variational parameter \( \xi \) becomes

\[
\xi^2 \leftarrow \text{diagonal} E[(C\beta + \mathcal{U})(C\beta + \mathcal{U})^T]
\] (6.127)
Algorithm 6

With the above updates we can specify the algorithm as

1: repeat
2: \( S \leftarrow 0 \)
3: \( s \leftarrow 0 \)
4: Update \( y_{\text{working}} \leftarrow y/2 - 2y^2U_\lambda(\xi) \)
5: for \( i = 1 \ldots m \) do
6: Update \( G_i \) (6.83), Update \( H_i \) (6.84)
7: Update \( S \) (6.85), Update \( s \) (6.86)
8: end for
9: Update \( \Sigma_{q(y; \beta; \xi)} \) (6.87), Update \( \mu_{q(y; \beta; \xi)} \) using (6.88)
10: for \( i = 1 \ldots m \) do
11: Update \( \Sigma_{q(y; \beta; \xi)} \) (6.89), Update \( \mu_{q(y; \beta; \xi)} \) (6.90)
12: end for
13: Update \( y_{\text{working}} \leftarrow y/2 - 2y^2(C\beta)\lambda(\xi) \)
14: for \( m = 1 \ldots M \) do
15: for \( i = 1 \ldots N_m \) do
16: Update \( \Sigma_{q(U_j^m)} \) (6.125)
17: Update \( \mu_{q(U_j^m)} \) (6.126)
18: Update \( r_{i,j}^2 \) (6.71)
19: end for
20: end for
21: \( y_{\text{working}} \leftarrow y \)
22: Update \( \xi^2 \) (6.127)
23: for \( r = 1, \ldots, q^R \) do
24: Update \( B_{q(r; \xi)} \) (6.94), Update \( \mu_{q(1/r; \xi)} \) (6.95)
25: end for
26: Update \( B_{q(2^R; \xi)} \) (6.96)
27: Update \( M_{q(2^R; \xi)} \) (6.97)
28: \textbf{for } \ell = 1 \ldots L \textbf{ do}
29: \hspace{1em} \text{Update } \mu_{q(1/a_{ul})} (6.98), \text{ Update } \mu_{q(1/\sigma_{ul}^2)} (6.99)
30: \textbf{end for}
31: \textbf{until} convergence in \( p(y; q) \)

\textbf{Computational Notes}

As with the Gaussian tensor factorization case the complete tensor may be very sparsely observed. However in the binary case it may also be that instances of a \( y = 1 \) are very rare. This will often occur with various types of longitudinal network structure. In this setting substantial computational gains can be realized by using a case control likelihood as suggested in (Raftery et al. 2012).

\section{6.2 Alternative Approaches}

In this appendix I briefly overview several approaches to modeling heterogeneity that are not encompassed by the framework presented here. In each case I have highlighted the contrast with the framework that I have provided in the main text.

\textbf{Exponential Random Graph Models} (ERGMS) ERGMs provide an alternative approach to modeling networks. Here we forego the conditional independence assumption and instead model the entire graph as a single draw from a joint (Gibbs) distribution (Cranmer and Desmarais 2011). ERGMs are notoriously difficult to estimate and require careful specification of the sufficient statistics of the graph. They also only apply to graphs with unweighted edges. However, in settings where the conditional independence is untenable they are effectively the only option.

\textbf{Survival Analysis} Semiparametric approaches to survival analysis, such as the Cox model are essentially single mode models with varying intercepts (where time is the mode) (Box-Steensmeier and Jones 2004). The spline based method of Beck, Katz and Tucker (1998) for analyzing binary duration data can be placed into the GMRF framework. see also Jackman (1998) for relevant connections.

\textbf{Binary Treatment Causal Inference} Recent work on causal inference in time-series and time-series cross sectional data addresses related problems in the potential outcomes framework (Blackwell 2013; Blackwell and Glynn 2013;
Chapter 6. Appendix


**Mixture Models** In the models described in this paper, the groups within a mode are assumed to be observed. When even this information is unavailable mixture models can be used to model the heterogeneity. Kyung, Gill and Casella (2010, 2011) use a Dirichlet process random-effects model to allow varying intercepts over unknown groupings in the data. Park (2012) provides a model for time-series cross-sectional data where which uses a non-recurrent Hidden Markov Model for parameters. This treats time as a mode where all members of a group are temporally contiguous but the breakpoints between groups are estimated by the model. Imai and Tingley (2012) use a finite mixture of GLMs model which allow each data point to be drawn from an *a priori* fixed number of possible regression models. While presented in the context of theory testing, we can motivate the same infrastructure as a way to model a single model with unknown group membership with each group represented by one of the regressions. And extended to the infinite mixture model is provided by the Dirichlet Process-GLM framework (Hannah, Blei and Powell 2011). Finally recent econometric work has focused on learning unknown group membership in the interactive fixed effects framework (Ando and Bai 2013).

**Flexible Regressions** A separate approach to modeling heterogeneity avoids explicit models for the heterogeneity and simply uses an extremely flexible regression (Wahba 1990; Gu 2013). By allowing effects to be highly non-linear and context dependent we can side step the issue of explicitly producing models of the models. In the international relations context this was Beck, King and Zeng (2000) argue for the use of neural networks for estimating these types of functions. Hainmueller and Hazlett (2014) present a more interpretable approach based on Kernel Regularized Least Squares. Both methods face the challenge of interpreting the resulting models which is complicated by the non-linear forms.

**Nonparametric Estimation** For the particular case of network data there has been increasing interest in nonparametric approaches to estimation. Chatterjee (2012) provides a consistent estimator based on a truncated singular

---

11 This literature makes use of the statistical function called a graphon. Graphons are the limiting objects that describe random network objects. Define a graph with *N* nodes which are each given a latent variable *u* ~ Uniform(0, 1) for all *i* ∈ {1...*N*}. Two vertices are connected with probability *w*(*u* *i*, *u* *j*) where the function *w* is the graphon. That is, it is a function which maps [0, 1]^2 → [0, 1]. This describes a wide class of models over exchangeable graphs (Lloyd et al. 2013; Orbanz and Roy 2013). Interest is in the circumstances in which we can non-parametrically estimate this function and the circumstances in which we can prove consistency of the resulting estimator (Airoldi, Costa and Chan 2013).
value decomposition that applies to a wide range of models including many of the two mode network models. Airoldi, Costa and Chan (2013); Chan and Airoldi (2014); Yang, Han and Airoldi (2014) design a consistent estimator that uses a histogram of stochastic block models. These approaches are extremely new but they point to approaches to nonparametric estimation of particular versions of the model. The statistical characterization of graphons also provides insights to the extent to which the latent effects models are identifiable (Bickel and Chen 2009; Bickel, Chen and Levina 2011).

6.3 Two-Way Fixed Effects and Latent Factor Regression

The goal of this appendix is to illuminate the connection between the Latent Factor Regression framework and two-way fixed effects. In doing so I also address the three related questions: ‘where is the variation coming from?’, ‘what is the counterfactual?’ and ‘what’s the analogous experiment?’ In order to connect to the literature on causal inference I focus on settings where the effect of interest is a binary treatment with other covariates serving as continuous pre-treatment confounders. However, I emphasize that the framework naturally extends to non-binary effects. I also only consider varying intercepts although the framework naturally extends to varying coefficients.

6.3.1 Two-way Fixed Effects

For concreteness and without loss of generality, I consider the case of data organized in a time-series cross-sectional format with \( t \in \{1 \ldots T\} \) years and \( c \in \{1 \ldots C\} \) countries. The two-way fixed effects estimator considers a model of the form:

\[
y_{c,t} = x_{c,t}\beta + \alpha_c + \gamma_t + \epsilon
\]

where \( \alpha \) and \( \gamma \) are vectors of country and time specific intercepts respectively, \( x_{c,t} \) are the covariates for unit \( c \) at time \( t \) and \( \epsilon \) is random Gaussian noise.

Arranging the outcome \( y \) into a \( C \times T \) matrix it is clear that the fixed effects \( \alpha \) and \( \gamma \) are estimated using the rows and columns of the matrix respectively. The unmeasured confounding in the data is then controlled through the additive combination of the row and column effect. What the two-way fixed effects setup implicitly assumes is that the effect of country \( c \) is constant across all time periods and analogously the effect of time period \( t \) is the same across all countries. We can give this an economic interpretation by saying that there are global shocks to the system represented.
by $\gamma_t$ which affect all countries in the same way, and different base levels of the outcome for each country measured by $\alpha_c$.

### 6.3.2 Latent Factor Regressions

The analogous latent factor regression model is:

$$y_{ct} = x_{ct}\beta + \alpha_c + \gamma_t + \sum_k u_{c,k} v_{t,k} + \epsilon$$  \hspace{1cm} (6.129)

where $u_{c,k}$ is an element of factor matrix $U$ having dimension $C \times K$ (with $V$ analogously defined). The rank of the approximation, $K$, can be estimated or fixed by the analyst.

The two-fixed effects estimator is a special case of the latent factor regression where $k = 0$ and thus is directly nested within the framework proposed in the main paper. The latent factor matrices $U$ and $V$ represent deviations from the additive form of the model and thus under the interpretation above can be seen as capturing the degree to which common shocks to the system elicit varying responses from countries.

It is helpful to define a third model which serves as an illustrative limiting case which I call the joint fixed effects model. This formulation can be given as,

$$y_{ct} = x_{ct}\beta + \phi_{ct} + \epsilon$$  \hspace{1cm} (6.130)

$\Phi$ is a matrix where each cell $\phi_{ct}$ represents a fixed effect for that country-year combination. Clearly the model in Equation 6.130 is only identified if the data contain more than one observation for each pair of indices. Whereas the two-way fixed effects estimator of Equation 6.128 uses the entire row and entire column to estimate the intercepts for an observation, the joint fixed effects estimator uses only the information within the cell.

The latent factor regression model provides an intermediate point between the two fixed effects models. The degree of the tradeoff is controlled by the rank of the latent factor model, $K$. When $K = 0$ the model recovers exactly the two-fixed effects model. When $K = \min(C, T)$ the model recovers exactly the joint fixed effects model.\(^{12}\) The solution $(\alpha, \gamma, u, v)$ for a fixed value $K$ is the best rank $K$ approximation to the joint parameter matrix $\Phi$.\(^{13}\)

\(^{12}\)The proof of this follows immediately from the singular value decomposition which guarantees the existence of the decomposition as well as the exact reconstruction of the full matrix (Eckart and Young 1936).

\(^{13}\)This is the best approximation in terms of the Frobenius norm of the matrix as again guaranteed by appeal to the singular value decomposition. When estimating the rank using the ARD priors as used in the paper this corresponds exactly to the best approximation of the joint parameter matrix under the nuclear norm, a common convex relaxation of the rank selection problem (Fithian and Mazumder 2013).
Thus, even in cases where we have replications at the cell (e.g. country/year) level the latent factor approach is an attractive alternative to the joint effects model because if we believe there is any structure in the matrix of parameters $\Phi$ (e.g. the effect for country $c$ at time $t$ has any information to offer us on the effect for country $c$ at time $t - 1$) then the latent factor regression framework provides us with a favorable bias-variance tradeoff. This is because structure of the matrix allows the model to bring additional observations beyond the replications at the cell level.\(^{14}\)

Having addressed the basic infrastructure, I now consider direct answers to a few common questions.

### 6.3.3 Connections to Causal Inference

A natural question for any new method is ‘what is this doing?’ That is, we want to be able to articulate which treated units are being compared to which control units. This is closely related to the question of what variation is being removed from the data by the procedure. Before moving into the latent factor regression framework I briefly explain what information is being removed in the two-way fixed effects framework as well as the joint fixed effects.

Again it is helpful to arrange the data $Y$ into a $C$ by $T$ matrix. In two-way fixed effects we are removing from each cell variation which is common to the row and variation which is common to the column. The variation removed is a function of three averages: the average of all observations in the same row (country), the average of all observations in the same column (time) and the average of all remaining observations (Imai and Kim 2012).\(^{15}\)

In the joint fixed effects estimator we consider only a single cell of the matrix when removing variation. Again if we observe only a single observation per country-time combination this is unidentified, having a single free parameter for each observation. In cases where we have replication at this level, such as in the ‘Dirty Pool’ example (Green, Kim and Yoon 2001), we are simply subtracting off the mean of all observations within that cell. They key difference here is that the two-way fixed effects estimator uses data from every country and every time period in constructing each counterfactual, whereas the joint estimator uses only the country and year for that cell.

In the latent factor regression, we consider the entire row and column of the matrix (as in the two-fixed effects) but each observation is not weighted equally. Instead the model implicitly assigns higher weight to countries (or years) which have trends in the dependent variable which are similar to the country (year) of the cell. I make this comparison more precise below by first considering which units are stochastically equivalent under the model and then by giving

\(^{14}\)The bias is determined by the degree to which $\Phi$ cannot be captured by the low-rank structure and the variance improvement comes as a result of using additional observations in estimating $\phi_{i,j}$ as $\alpha_i + \gamma_j + \sum_k u_{ik} v_{jk}$. Clearly for a non-random matrix $\Phi$ the bias decreases in the rank of the approximation.

\(^{15}\)The third term is necessary to adjust for the fact that we are adjusting the data based on two margins rather than one. The full form of the two-way fixed effects as an adjusted matching estimator is given with proof as Proposition 4 of Imai and Kim (2012).
an augmented data interpretation of the estimator.

Stochastic equivalence

An intuitive way to think of how the latent factor framework models dependence is by considering the components of the inner product term $U$ and $V$ as forming a $K$ dimensional vector space in which both countries and years are projected. Countries which have similar projections $u_c$ in the space are approximately stochastically equivalent and respond to shocks in a similar fashion. This provides us with insight into where the variation comes from. We are implicitly comparing a country-time unit with a country-time unit having a similar projection into the $K$ dimensional space. This produces a continuous weighting over units in computing the counterfactuals in a potential outcomes framework. In the matching framework, this continuous weighting can be seen as analogous to synthetic control methods which matches treated units to a reweighted collection of controls (Abadie, Diamond and Hainmueller 2010).

Note that the appeal here to the approximate stochastic equivalence of two units sharing similar values of the continuous factors is in principle no different from approximate stochastic equivalence of two units with similar pre-treatment covariates. That is, when we control for a series of (non-categorical) observed covariates we are invoking an assumption of approximate stochastic equivalence of two units with similar covariate profiles.

A further advantage of the low-rank framework is that two countries can be similar along one dimension but different along another. Put another way, there can be a type of global shock for which two countries can have a similar response, but a different type of global shock for which their responses will be different. Mathematically this is represented by $u_{i,k} \approx u_{j,k}$ for $k = i$ but not for $k \neq i$. This provides a substantially more flexible framework for modeling heterogeneous units than a model assuming units must be alike on all dimensions.

Interpretations as OLS on augmented data

Part of the reason simple additive fixed effects are so easy to understand is that we can give a representation of the model as a simpler procedure on an augmented dataset. The “least squares dummy variable” method can be written by using OLS on a transformed datasets formed by subtracting off the mean by group.

Using the results in Bai (2009), we can give a a similar interpretation for the latent factor regression as OLS on
augmented data. Here we consider the special case of latent factor regression where the latent factors are given no priors which is simply the limiting special case of uninformative priors.

Start by defining the projection matrices:

\[ M_V = I_T - V'V/T \]  \hspace{1cm} (6.131)

\[ M_U = I_C - U'U'/C \]  \hspace{1cm} (6.132)

then writing the model as:

\[ Y = \beta_1X^1 + \cdots + \beta_pX^p + UV' + \epsilon \]  \hspace{1cm} (6.133)

where we have simply absorbed the additive fixed effects into the factor matrices.

Then the left multiplying by \( M_V \) and right multiplying by \( M_U \) we get

\[ M_VYM_U = \beta_1(M_VX^1M_U) + \cdots + \beta_p(M_VX^pM_U) + M_V\epsilon M_U \]  \hspace{1cm} (6.134)

This leads to the following least squares estimator under a given factor structure,

\[
\hat{\beta} = \begin{bmatrix}
\text{tr}[M_UX^1M_VX^1] & \cdots & \text{tr}[M_UX^1M_VX^p] \\
\vdots & \ddots & \vdots \\
\text{tr}[M_UX^pM_VX^1] & \cdots & \text{tr}[M_UX^pM_VX^p]
\end{bmatrix}^{-1}
\begin{bmatrix}
\text{tr}[M_UX^YM_VY] \\
\vdots \\
\text{tr}[M_UX^YM_VY]
\end{bmatrix}
\]  \hspace{1cm} (6.135)

Thus the projection matrices \( M_U \) and \( M_V \) play a role analogous to projection matrices in least square dummy variables estimation.

Bai (2009) also gives an instrumental variables interpretation that holds in this setting as well. Define

\[
\sum_c Z'_c Z_c = \begin{bmatrix}
\text{tr}[M_UX^1M_VX^1] & \cdots & \text{tr}[M_UX^1M_VX^p] \\
\vdots & \ddots & \vdots \\
\text{tr}[M_UX^pM_VX^1] & \cdots & \text{tr}[M_UX^pM_VX^p]
\end{bmatrix}
\]  \hspace{1cm} (6.136)

The form of the estimator for beta is the IV estimator with \( Z'_c \) as the instrument.
6.3.4 Connections to Existing Work

Here I briefly connect this work to several existing articles in order to further illuminate the features of the model.

**Grouped Fixed Effects**

One way to deal with the problem of not observing multiple observations for each cell in our data matrix $Y$ is to aggregate over the rows (countries) to create groups. We can then consider group-time specific fixed effects. These groups could be determined *a priori* or estimated as in Bonhomme and Manresa (2012). If the group membership is estimated under fixed $K$ then the problem becomes a latent class (or finite mixture) model.

The latent factor regression framework and the mixture model framework differ in the parametric assumption on the latent variable. For the mixture model case the latent variable is categorical whereas in the latent factor regression it is continuous. The grouped fixed effects model has the distinct advantage that the latent variable is naturally interpretable as a partition over units. However, assuming that the latent variable is categorical is substantially less general than the latent factor framework as members of the same group are assumed to be exactly stochastically equivalent. This is a stronger assumption which is unnecessary under the latent factor regression model.

**Correlated Error Models**

One way to view the latent factor regression framework is as inducing a low rank decomposition of the error structure. Writing the regression model in matrix notation:

$$y = X\beta + \epsilon$$  \hfill (6.137)

$$\epsilon \sim \mathcal{N}(0, \Sigma)$$  \hfill (6.138)

the standard regression model assumes that $\Sigma = \sigma^2 I$. If instead we treat $\Sigma$ as unstructured we essentially observe a single draw from a multivariate Gaussian. Unfortunately MLE is known to perform poorly for covariance estimation in this setting (James and Stein 1961).

Numerous proposals have been made for estimating $\Sigma$ under some particularly assumptions, e.g. time-series models often assume that $\Sigma^{-1}$ is tri-diagonal (West and Harrison 1997). A particularly general case is given by spatial error models which assume that the variance is rescaled by a known weights matrix, $W$. However the appropriate form of $W$ is often not known and reasonable choices can produce different results (Zhukov and Stewart 2013). The latent factor
This interpretation of the model also makes clearer where the two way random effects model will be insufficient. The two-way fixed effects model in Equation 6.128 is unable correlation in the second moments of the data but not the third order moments. This notion has the clearest expression in network data where properties of third order dependence have been well characterized (Hoff 2005; Wasserman and Faust 1994). These intuitions extend reasonably well to the spatial and cross-sectional case, where intuitively we can think of third order dependence as capturing the consistency of the pairings A-B, B-C, A-C where A, B, C are nodes in a networks or locations in a space.

This covariance structure can also be given a probabilistic interpretation akin to the linear regressions product of univariate normals. Under a two-mode model the latent factor regression provides a parameterization of a matrix normal distribution (Dawid 1981; Hoff 2005; Allen and Tibshirani 2012). Under a general m-mode we obtain an array normal distribution under separable covariance structure (Hoff 2011b). The key assumption in these settings is a weak row-column exchangeability which is substantially more general than the exchangeability assumption typically invoked (Hoff 2005; Orbanz and Roy 2013).

6.3.5 Concluding Thoughts

In this appendix I’ve tried to illuminate what the latent factor regression framework is doing for particular cases. I’ve primarily discussed the two mode framework but a key advantage of the model is the ability to extend easily to an arbitrary number of modes.

6.4 Improving Accuracy of the Variational Framework

In this appendix I provide a brief summary of relevant results on accuracy of posterior inference in the variational framework including approaches to improve accuracy.

The variational algorithms presented in the main text produce quite faithful approximations to the true posterior. The results are particularly strong in the case of the normal likelihood (which is conjugate) and in the single mode case (where we do not have to make strong factorization assumptions). Theoretical results in Ormerod and Wand (2012) and previous empirical results reported in the literature support this observation (Wand 2014b; Lee and Wand 2014; Tan and Nott 2013).

18 For more comparisons to various spatial models with Monte Carlo experiments see Pang (2014).
In the next two sections I step through three settings where accuracy can be improved and sketch some methods for trading off computational time for improved quality of approximation. None of the methods described below have been implemented in the results reported above. I have included the discussion of these approaches here in order to demonstrate that improvements to the quality of the approximation are possible within the variational framework. However, in the main text I chose to maintain the simplest version of the available methods that also produced accurate results.

### 6.4.1 Non-Gaussian Likelihoods

The posterior approximation used here for the logistic regression case uses the Jaakkola and Jordan (2000) bound on the sigmoid function. This is a quadratic bound which is tight but only at the value of the variational parameters $\xi$. Previous empirical studies suggest that the bound produces a small bias towards zero in the random effects which disappears as the cluster size grows (Ormerod and Wand 2012; Tan and Nott 2013; Scott and Sun 2013). This accords with the findings in the simulations I ran for this paper.

Numerous alternative strategies have been proposed but here I highlight two in particular: the Non-conjugate Variational Message Passing scheme of Knowles and Minka (2011), and piecewise bounds (Marlin, Khan and Murphy 2011).

Knowles and Minka (2011) generalize the variational message passing scheme (Bishop et al. 2006) to handle non-conjugate factors by approximating them using an approximating distribution in the same family as the prior. For the case of a multivariate Gaussian approximation Wand (2014b) provides a simplified update structure which enables efficient computation. These updates were used in Tan and Nott (2013) to derive variational algorithms for GLMMs which show excellent results.

For the logistic regression case, this scheme involves the calculation of an analytically intractable expectation. However, Ormerod and Wand (2012) show that it can be reduced to a uni-dimensional problem and evaluated accurately using Adaptive Gauss-Hermite quadrature. This results in a slightly slower algorithm but yields more faithful representations of the posterior.

For the case of count models the required expectation can be evaluated in closed form resulting in little tradeoff in speed (Luts and Wand 2013; Wand 2014b).

An alternative strategy is similar to the bounding approach of Jaakkola and Jordan (2000). Rather than using a quadratic bound that is tight only at a single point, Marlin, Khan and Murphy (2011) advocate the use of piecewise
bounds. By increasing the number of pieces, the bound on the nonconjugate term can be made arbitrarily tight resulting in more accurate inference. This has the advantage of allowing variational inference to take on a bit of the quality of MCMC where a continuous increase in computational cost results in a continuous increase in accuracy. Experimental evaluations are given in (Khan et al. 2010).

A particularly compelling application of the bounding approach is the extension to the multinomial outcome case. The Jaakkola and Jordan (2000) does not extend to this setting nor does the use of quadrature. Khan et al. (2012) develop a stick breaking likelihood suitable for a categorical outcome for which efficient bounds could be constructed. This would allow for the implementation of an approximate analogue of a multinomial logistic regression that admits a tight bound.

6.4.2 Factorization

The key assumption in variational inference is the product density factorization of the joint posterior. Stronger factorizations make the model more tractable but also less accurate. In the single mode case the key assumed factorization is between the regression coefficients and their variance parameters. In the case of unordered groups this assumption is relatively mild but for Gaussian Markov Random Fields it is somewhat stronger.

Luckily we can always make our approximations arbitrarily more accurate by using conditional approximations. Whereas standard variational bayes might approximate \( q(x, \theta | y) = q(x|y)q(\theta | y) \) for a latent field \( x \) and hyperparameters \( \theta \) we instead do \( q(x, \theta | y) = q(x|y, \theta)q(\theta | y) \) this is tractable for hyperparameters of low dimension like the variances in a hierarchical model. This is the essential insight of Rue, Martino and Chopin (2009) and is formalized in the VB context by Han, Liao and Carin (2013). It is also raised Salimans and Knowles (2013) who describe it as a hierarchical approximation.

A milder form of this strategy is considered by Tan and Nott (2013) in their use of partial non-centering of GLMMs. Here the partial non-centering parameters give the model extra flexibility to handle the assumed factorization.

An open question is whether there is a straightforward equivalent of this to the multilinear case. If there were it would be tremendously useful across a wide variety of models but it seems unlikely as there is no single low-dimensional parameter to be conditioned on. A reasonable strategy might be to instead use a mixture of distributions to capture the joint approximation. A setting using a mixture of multivariate normals is considered by Gershman, Hoffman and Blei (2012). By combining this nonparametric approach with the gridding strategies in Rue, Martino and Chopin (2009) it may be possible to construct a tighter approximation to the factorized terms.
6.5 Simulation

In this section I provide the necessary details to replicate the simulation results.

6.5.1 Single Mode Setting

In the first simulation of the single mode setting, I demonstrate the ability of the variational algorithm to recover parameters in hierarchical linear regression model with a Gaussian outcome. After reporting details for the Gaussian outcome I describe estimation for the hierarchical logistic regression model.

**Gaussian Hierarchical Regression**  I use the data generating process from the help file of `MCMChregress` in `MCMCpack` version 1.3-3 (Martin, Quinn and Park 2011). Simulations were run in R version 3.1.1 on a 3.2Ghz quadcore processor with 7GB RAM. The code including the model estimation is given below.

```r
nobs <- 1000
nspecies <- 20
species <- c(1:nspecies,sample(c(1:nspecies),(nobs-nspecies),replace=TRUE))

# Covariates
X1 <- runif(n=nobs,min=0,max=10)
X2 <- runif(n=nobs,min=0,max=10)
X <- cbind(rep(1,nobs),X1,X2)
W <- X

# Target parameters
# beta
beta.target <- matrix(c(0.1,0.3,0.2),ncol=1)
# Vb
Vb.target <- c(0.5,0.2,0.1)
# b
b.target <- cbind(rnorm(nspecies,mean=0,sd=sqrt(Vb.target[1])),
                  rnorm(nspecies,mean=0,sd=sqrt(Vb.target[2])),
```

178
\begin{verbatim}

rnorm(nspecies,mean=0, sd=sqrt(Vb.target[3])))

# sigma2
sigma2.target <- 0.02

# Response
Y <- vector()
for (n in 1:nobs) {
  Y[n] <- rnorm(n=1,
                mean=X[n,]%*%beta.target+W[n,]%*%b.target[species[n],],
                sd=sqrt(sigma2.target))
}

# Data-set
Data <- as.data.frame(cbind(Y,X1,X2,species))

model <- MCMChregress(fixed=Y˜X1+X2, random=˜X1+X2, group="species",
                      data=Data, burnin=1000, mcmc=10000, thin=10, verbose=1,
                      seed=NA, beta.start=0, sigma2.start=1,
                      Vb.start=1, mubeta=0, Vbeta=1.0E6,
                      r=3, R=diag(c(1,0.1,0.1)), nu=0.001, delta=0.001)

The results for the Gaussian case are given in the main text.

Logistic Regression  Here I present results for a hierarchical logistic regression which parallels the Gaussian outcome model. Again I use the data generating process given in \texttt{MCMCpack} with the exception that I extend the burnin and number of posterior samples to match the normal regression case. The full code is:

# Constants
nobs <- 1000
nspecies <- 20
\end{verbatim}
simresults <- vector(mode="list", length=100)
for(s in 1:100) {
  # Covariates
  species <- c(1:nspecies, sample(c(1:nspecies), (nobs-nspecies), replace=TRUE))
  X1 <- runif(n=nobs, min=-10, max=10)
  X2 <- runif(n=nobs, min=-10, max=10)
  X <- cbind(rep(1, nobs), X1, X2)
  W <- X

  # Target parameters
  # beta
  beta.target <- matrix(c(0.3, 0.2, 0.1), ncol=1)
  # Vb
  Vb.target <- c(0.5, 0.05, 0.05)
  # b
  b.target <- cbind(rnorm(nspecies, mean=0, sd=sqrt(Vb.target[1])),
                    rnorm(nspecies, mean=0, sd=sqrt(Vb.target[2])),
                    rnorm(nspecies, mean=0, sd=sqrt(Vb.target[3])))

  # Response
  theta <- vector()
  Y <- vector()
  for (n in 1:nobs) {
    theta[n] <- inv.logit(X[n,]%*%beta.target+W[n,]%*%b.target[species[n],])
    Y[n] <- rbinom(n=1, size=1, prob=theta[n])
  }

  # Data-set
  Data <- as.data.frame(cbind(Y, theta, X1, X2, species))
plot(Data$X1, Data$theta)

#== Call to MCMChlogit
model <- MCMChlogit(fixed=Y ~ X1 + X2, random=~X1+X2, group="species",
data=Data, burnin=1000, mcmc=10000, thin=10, verbose=1,
seed=NA, beta.start=0, sigma2.start=1,
Vb.start=1, mubeta=0, Vbeta=1.0E6,
r=3, R=diag(c(1,0.1,0.1)), nu=0.001, delta=0.001, FixOD=1)

Run times are slightly more variable for the logistic regression case and are plotted in Figure 6.1. MCMC runs for 41 seconds on average with variational running for 2 seconds.

![Run Times for Logistic Regression](image)

**Figure 6.1:** Distribution of run times for 100 simulations of the single mode logistic regression case.
6.5.2 Two Mode Setting

In this simulation I demonstrate the ability of the variational algorithm to recover simulated parameters in the Gaussian outcome case with two modes and interactive latent factors. In order to match the simulation to the FDI application, I use the observed covariates from Büthe and Milner (2008). Each simulation then follows the following procedure where nc is the number of countries in cindex and nt is the number of time periods in tindex.

\[
\begin{align*}
k &\leftarrow \text{rpois}(1, \text{lambda}=3) + 1 \\
\text{factor1} &\leftarrow \text{matrix}(\text{rnorm}(k \times nc), \text{nrow}=nc, \text{ncol}=k) \\
\text{factor2} &\leftarrow \text{matrix}(\text{rnorm}(k \times nt), \text{nrow}=nt, \text{ncol}=k) \\
\text{uv} &\leftarrow \text{rowSums}(\text{factor1}[\text{cindex}, \text{,drop=FALSE}] \times \\
&\quad \text{factor2}[\text{tindex}, \text{,drop=FALSE}]) \\
\text{cint} &\leftarrow \text{rnorm}(nc) \\
\text{tint} &\leftarrow \text{rnorm}(nt) \\
\beta &\leftarrow \text{matrix}(\text{rnorm}(9), \text{ncol}=1) \\
y &\leftarrow \text{c}(X \times \% \% \beta) + \text{cint}[\text{cindex}] + \text{tint}[\text{tindex}] + \\
&\quad \text{uv} + \text{rnorm}(\text{length}(y)) \\
y &\leftarrow \text{c}(y)
\end{align*}
\]

All parameters are simulated from Normal(0, 1).

6.5.3 Model Misspecification

Here I provide some additional results for the models presented where covariates in the true data generating process are omitted from the estimated models. In the paper I gave two extreme examples: no covariates omitted and all but one covariate omitted. Here I provide a sample of the cases between. In each case a random selection of covariates was dropped whereas in the two extreme examples the observed covariates are the same across all simulations.

Recall that I compare four alternative estimation strategies in addition to the latent factor regressions:

1. One-Way Fixed Effects
   “country” level intercepts which are the largest source of variation in the model.

2. Two-Way Fixed Effects
   “time” and “country” intercepts. This is the additive two-mode model.
3. Global Linear Detrending with One-Way Fixed Effects
   “country” intercepts and a linear time trend shared by countries

4. Country-Specific Quadratic Detrending
   “country” specific quadratic time trends

Figure 6.2: Two, Four and Six random covariates are dropped in each simulation.
Bibliography


Ando, Tomohiro and Jushan Bai. 2013. “Panel data models with grouped factor structure under unknown group membership.” Available at SSRN 2373629.


Bafumi, Joseph and Andrew E Gelman. 2006. “Fitting multilevel models when predictors and group effects correlate.”.


Beck, Nathaniel. 2012. “Sweeping fewer things under the rug: tis often (usually?) better to model than be robust.” *The Society for Political Methodology POLMETH XXIX*.


BIBLIOGRAPHY


BIBLIOGRAPHY


192

Genkin, Alexander, David D. Lewis and David Madigan. 2007. “Large-Scale Bayesian Logistic Regression for Text Categorization.” *Technometrics* 49.


BIBLIOGRAPHY


Hoff, Peter. 2012. eigenmodel: Semiparametric factor and regression models for symmetric relational data. R package version 1.01. URL: http://CRAN.R-project.org/package=eigenmodel

Hoff, Peter, Bailey Fosdick, Alex Volfovsky and Kate Stovel. 2014. amen: Additive and multiplicative effects modeling of networks and relational data. R package version 0.999. URL: http://CRAN.R-project.org/package=amen


Huang, Furong, UN Niranjan, M Hakeem and Animashree Anandkumar. 2013. “Fast Detection of Overlapping Communities via Online Tensor Methods.”.


King, Gary and Margaret Roberts. 2014. “How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It.” *Political Analysis* .


Lee, Yuen Yi Cathy and Matt Wand. 2014. “Streamlined Mean Field Variational Bayes for Longitudinal and Multilevel Data Analysis.”


Liu, Junrong, Robin Sickles and EG Tsonias. 2013. “Bayesian Treatments to Panel Data Models.”
Lloyd, James, Peter Orbanz, Zoubin Ghahramani and Daniel Roy. 2013. “Random function priors for exchangeable arrays with applications to graphs and relational data.”


Lu, Xun and Liangjun Su. 2013. “Shrinkage estimation of dynamic panel data models with interactive fixed effects.”


Mimno, David and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In UAI.


Moon, H and Martin Weidner. 2010a. “Dynamic linear panel regression models with interactive fixed effects.” manuscript, UCL.

Moon, Hyungsik and Martin Weidner. 2010b. “Linear regression for panel with unknown number of factors as interactive fixed effects.” Unpublished manuscript, University of Southern California.


URL: http://books.google.com/books?id=0In7ugAACAJ


201


Pham, Tung H and Matt P Wand. 2014. “Generalized Additive Mixed Model Analysis via gammSlice.”.


Rai, Piyush, Yingjian Wang, Shengbo Guo, Gary Chen, David Dunson and Lawrence Carin. 2014. “Scalable Bayesian Low-Rank Decomposition of Incomplete Multiway Tensors.”


Reich, Justin, Dustin Tingley, Jetson Leder-Luis, Margaret E. Roberts and Brandon M. Stewart. Forthcoming. “Computer Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses.” *Journal of Learning Analytics*.


