Latent Factor Regressions for the Social Sciences*

Brandon Stewart†

November 30, 2014

Abstract

In this paper I present a general framework for regression in the presence of complex dependence structures between units such as in time-series cross-sectional data, relational/network data, and spatial data. These types of data are challenging for standard multilevel models because they involve multiple types of structure (e.g. temporal effects and cross-sectional effects) which are interactive. I show that interactive latent factor models provide a powerful modeling alternative that can address a wide range of data types. Although related models have previously been proposed in several different fields, inference is typically cumbersome and slow. I introduce a class of fast variational inference algorithms that allows for models to be fit quickly and accurately.

*For comments and discussions on various portions of this material I thank Michael Gill, Adam Glynn, Justin Grimmer, Gary King, Horacio Larreguy, Chris Lucas, John Marshall, Helen Milner, Brendan O'Connor, Marc Ratkovic, Beth Simmons, and Alex Volfovsky. Molly Roberts provided both enlightening discussions and code from her paper on robust standard errors. Special thanks to Dustin Tingley without whom this paper would not have been possible. An appendix containing additional details is available on my website: scholar.harvard.edu/bstewart

†Graduate Student. Department of Government Harvard University. bstewart@fas.harvard.edu

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). 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 in-feasible 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" (National Research Council, 2013).

When observations are organized along a single partition[2] or "mode"(e.g. the country

[1]Even 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).

[2]By 'partition' I mean a mutually exclusive and exhaustive grouping of the observations.

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 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 2 I describe the general problem including a motivating example from international relations, a subfield of political science; but as will become clear, the implications of this paper span other social sciences. In Section 3, I outline the framework for latent factor regression. In Section 4 I develop an estimation framework for latent factor regression based on variational inference. Section 5 connects my approach to diverse bodies of work, highlighting connections between previous models. Before proceeding to real data, Section 6 provides an overview of simulation evidence demonstrating the effectiveness of the inference framework. Section 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 8 concludes with a short discussion of directions for future research.

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 $d \in \{1 \dots D\}$, we can write the model generically as:

$$y_d \sim f_y(\theta_d, \phi) \tag{1}$$
$$\theta_d = g^{-1}(\eta_d) \tag{2}$$
$$\eta_d = \alpha + X_d\beta \tag{3}$$

Term	Definition	Example
Building Blocks		
Unit	the *level of analysis* interesting to a researcher	country-year, edge in a network
Mode	an observed *partition* over observations	time, cross-section, sender, receiver
Group	the members of a *subset of a partition*	individual years (time), individual countries (cross-section)
Data Structure	*covariance across units* in the dependent variable not captured by the covariates	time-series cross-sectional data, network data, multivariate outcome data
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 1: Definitions and examples for terms used throughout the paper.

where $f_y(\cdot)$ is the exponential family probability density, $g(\cdot)$ is the link function, η_d is the linear predictor for observation d, β are the regression coefficients, α is the intercept and ϕ 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 independence assumption.

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 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.

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 cost in data collection. It is difficult to advocate that an applied researcher bears these costs

[3]In 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.

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.

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:

$$\eta_d = \alpha_d + X_d\beta \tag{4}$$

where α is the now-dyad specific intercept term but β 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 models each cell in A as a separate parameter and thus treats the two modes as *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]

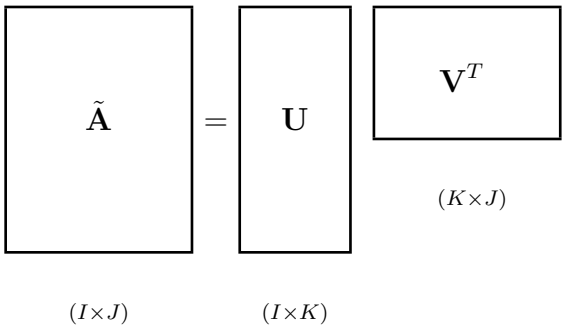
An intermediate between these approaches is a two-way varying intercepts approach in which the two modes are modeled *additively*. Now the dyad's intercept is the sum of the component countries,

$$\eta_{i,j} = \alpha_i + \alpha_j + X_{ij}\beta.$$

[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.

Figure 1: A low-rank matrix $\tilde{A}$ formed from the product of two smaller matrices.



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 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:

$$\eta_{i,j} = \alpha_i + \alpha_j + \left(\sum_k u_{i,k} v_{j,k} \right) + X_{ij}\beta. \tag{5}$$

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.[6] 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, are

[6]Note that K determines the quality of the approximation to the unrestricted matrix A. As $K = N$ we get an exact reconstruction of the matrix (Eckart and Young, 1936).

legislative ideal points); however, they involve fewer parameters than the GKY strategy which involves estimating every element of A separately.[7]

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 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).[8] 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.

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.

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

[7] 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.

[8] 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.

space network models (Hoff and Ward, 2004). The central idea is to replace a constant parameter β in a generalized linear model with a group-specific term formed by

$$\beta_{i,j,\ldots,l} = a + \sum_{m=1}^{M} b_m + \sum_{k=1}^{K} u_{i,k}^{(1)} u_{j,k}^{(2)} \ldots u_{l,k}^{(m)} \tag{6}$$

where a is a globally shared effect, b_m are the additive mode specific effects, $u_{i,k}^{(m)}$ is the latent variable for the ith 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 1 this yields the following form

$$\eta = X \left(U^{(1)} \times \cdots \times U^{(M)} \right) \tag{7}$$

$$U_{,k}^{(m)} \sim \text{Normal}(0, \Sigma) \tag{8}$$

$$\Sigma \sim p(\cdot) \tag{9}$$

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 Σ. By changing the prior distribution for Σ, 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(5.1) and explore the sensitivity of the model to violations of this assumption in the section on simulation (6).

This formulation encompasses an extremely wide range of models (Table 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 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 3.3 I show how different prior distributions for Σ 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.

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

[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 $\beta\text{time}, \text{country} = b_{\text{time}} + b_{\text{country}}$

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

Table 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.

basic model it is severally 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 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 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 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).[10]

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:

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

Many mathematical and computational aspects of tensor analysis are substantially more challenging than the matrix case.[11] 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, 2011*a*), 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 a unified framework.

3.2 Interpretation of Interactive Modes

The basic model in Eq 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.

[10]It 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.

[11]For example there are two natural formulations of the tensor decomposition: the Tucker Decomposition (Tucker, 1964; Hoff, 2011*b*) 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, 2011*b*). 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.

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 away 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, 2010*a*; 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." Rhis 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 5.

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 including Gaussian processes (Rasmussen and Williams, 2006), random effects (Fahrmeir and Lang, 2001), dynamic linear models (West and Harrison, 1997), stochastic volatitlity 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 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 from of the coefficients is:

$$\beta \sim \text{Normal}(0, Q^{-1}) \tag{10}$$

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 β_{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).

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 Estimation

In this section I describe a class of fast approximate inference algorithms for posterior inference in the class of models introduced in Section 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 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 5.

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, 2011*a*; 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 *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 $u_{i,1} \dots u_{i,k}$ for each group $i \in \{1 \dots I\}$ and $v_{j,1} \dots v_{j,k}$ for each group $j \in \{1 \dots J\}$. We then sample the error variance σ^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).

4.2 Variational Approximation

In order to provide a computationally efficient method of posterior inference, I turn to variational approximation (Winn and Bishop, 2005; Bishop, 2006; Grimmer, 2010; 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 (2010) 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.

4.2.1 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 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

$$y|\beta, u \sim \text{Normal}(X\beta + Zu, \sigma_\epsilon^2) \tag{11}$$

$$u_g|\Sigma^R \sim \text{Normal}(0, \Sigma^R) \tag{12}$$

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 Σ^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,

$$\sigma_\epsilon^2 \sim \text{Inverse-Gamma}(a_\epsilon, b_\epsilon) \tag{13}$$

$$\Sigma^R \sim \text{Inverse-Wishart}(A_{\Sigma^R}, B_{\Sigma^R}) \tag{14}$$

$$\beta \sim \text{Normal}(0, \sigma_\beta^2 I_P) \tag{15}$$

[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

where P is the number of columns of $\mathbf{X}$ and σ^2_β is a large value strictly greater than 0.

The approximation to the full joint posterior is

$$p(\beta, u, \Sigma^R, \sigma^2_\epsilon) \approx q(\beta, u)q(\Sigma^R, \sigma^2_\epsilon) \tag{16}$$

$$= q(\beta, u)q(\Sigma^R)q(\sigma^2_\epsilon) \tag{17}$$

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, 2006).

Under standard variational inference theory (Bishop, 2006; Grimmer, 2010), the optimal approximating densities for a parameter θ take the form

$$q(\theta) = \exp(E_{q(-\theta)}\log(p(\theta|\text{rest})) \tag{18}$$

Algebraic manipulations show these forms to be

$$q(\beta, u) = \text{Normal}(\mu_{q(\beta,u)}, \Sigma_{q(\beta,u)}) \tag{19}$$

$$q(\sigma^2_\epsilon) = \text{Inverse-Gamma}(a_n, b_n) \tag{20}$$

$$q(\Sigma^R) = \text{Inverse-Wishart}(A_N, B_N) \tag{21}$$

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

$$\log(p(y|q)) = E_q\{\log\, p(y, \beta, u, \Sigma^R, \sigma^2_\epsilon) - \log\, q(\beta, u, \Sigma^R, \sigma^2_\epsilon)\} \tag{22}$$

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.

4.2.2 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 (2014*b*) 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:

$$P(y|\eta) = \sigma(y\eta) \tag{23}$$

where η is the linear predictor and σ is the sigmoid function $\frac{1}{1+\exp(-\eta)}$.[13]

The log-likelihood is then

$$\log p(y) = \sum_n \log(\sigma(y_n \eta_n)) \tag{24}$$

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

$$\sigma(y\eta) \geq \sigma(\xi)\exp\left((y\eta - \xi)/2 - \lambda(\xi)\left((y\eta)^2 - \xi^2\right)\right) \tag{25}$$

$$\lambda(\xi) = \tanh(\xi/2)/(4\xi) \tag{26}$$

which introduces a new variational parameter ξ for each data point. The bound is tight at the optimal value of ξ. With the introduction of the parameters ξ 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

$$\xi = \sqrt{E[\eta^2]} \tag{27}$$

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]

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.

4.2.3 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

[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).

to demonstrate the basic inference strategy. Using interactive latent effects only for a varying intercept term and with a Gaussian likelihood yields:

$$y_{i,j} \sim \text{Normal}(x_{ij}\beta + \sum_k u_{i,k}v_{j,k}, \sigma_\epsilon^2) \tag{28}$$

$$u_{i,k} \sim \text{Normal}(0, \rho_k^2) \tag{29}$$

$$v_{i,k} \sim \text{Normal}(0, \tau_k^2) \tag{30}$$

$$\beta \sim \text{Normal}(0, \sigma_\beta^2 I_P) \tag{31}$$

where, for the moment, I treat the variance of the latent factors ρ^2, τ^2 and the noise variance σ_ϵ^2 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:

$$q(U, V, \beta) \approx q(U)q(V)q(\beta) \tag{32}$$

$$= \prod_{i=1}^{I} q(U_i) \prod_{j=1}^{J} q(V_j)q(\beta) \tag{33}$$

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 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:

$$q(U_i) = \text{Normal}(\mu_{q(U_i)}, \Sigma_{q(U_i)}) \tag{34}$$

$$q(V_j) = \text{Normal}(\mu_{q(V_j)}, \Sigma_{q(V_j)}) \tag{35}$$

$$q(\beta) = \text{Normal}(\mu_{q(\beta)}, \Sigma_{(q(\beta)}) \tag{36}$$

The posterior parameters of the approximation are updated as

$$\Sigma_{q(U_i)} = \left(\begin{pmatrix} 1/\tau_1^2 & 0 & \dots & 0 \\ 0 & 1/\tau_2^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1/\tau_k^2 \end{pmatrix} + \sum_{j=1}^{j} \frac{\Sigma_{(q(V_j)} + \mu_{q(V_j)}\mu_{q(V_j)}^T}{\sigma_\epsilon^2} \right)^{-1} \tag{37}$$

$$\mu_{q(U_i)} = \Sigma_{(q(U_i)} \left(\sum_{n \in \Omega} \left(\frac{(y_n - x_n\beta)\mu_{q(V_{j(n)}}}{\sigma_\epsilon^2} \right) \right) \tag{38}$$

where Ω 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 β is even simpler as it corresponds directly to Bayesian linear regression on the residuals

$$\tilde{y_{ij}} = y_{ij} - E[U_i]E[V_j^T] \tag{39}$$

$$= y_{ij} - \mu_{q(U_i)}\mu_{q(V_j)}^T \tag{40}$$

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, N.d.). 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.

4.2.4 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.[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

[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.

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, 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 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,

$$y_{i,j} \sim \text{Normal}(x_{ij}\beta + \sum_k u_{i,k}v_{j,k}, \sigma_\epsilon^2).$$

I start by estimating the model without the latent factors in order to get an initial estimate for β. 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]

4.2.5 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

[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 (Candès 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.

remain tractable and still take the form of bayesian linear regressions. The factorization of the posterior is now:

$$q\left(\beta, U^{(1)} \cdots \times U^{(M)}\right) \approx q(\beta) \prod_i q(U_i^{(1)}) \cdots \prod_j q(U^{(M)}) \tag{41}$$

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, 2011*a*; 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 et al., 2013). 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.

4.3 Estimation for Group Structure Priors

In section 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.

4.3.1 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 β in a linear regression. We place a Gamma(a_0, b_0)

[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 numericall stable (Huang et al., 2013).

prior on the precision parameter of the Normal density of the P dimensional vector β. The optimal q density has the form

$$q(a) = \text{Gamma}(a_N, b_N) \tag{42}$$
$$a_N = a_0 + P/2 \tag{43}$$
$$b_N = b_0 + .5 * E_{\beta,\sigma_\epsilon^2}\left(1/\sigma_\epsilon^2 \beta^T \beta\right) \tag{44}$$

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

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

$$\rho_{i,r}^2 \sim \text{Inverse-Gamma}(.5, 1/a_{i,r}) \tag{45}$$
$$a_{i,r} \sim \text{Inverse-Gamma}(.4, 1/A_{i,r}^2) \tag{46}$$

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.

4.3.2 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 $u \sim$Normal(μ, Q^{-1}). The density is then given by

$$p(u) \propto |Q|^{1/2} \text{exp}(-.5(x-\mu)^T Q (x-\mu)) \tag{47}$$

The cholesky factorization gives $Q = LL^T$ 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 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.[21] 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, 2006).[22] 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:

$$\rho_r^2 = \frac{1}{I-1}\sum_{i=1}^{I}(\Sigma_{(q(U_i))_{r,r}} + \mu^2_{r,q(U_i)}) \tag{48}$$

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).

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 nonparameteric 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.[23] In the applications and simulations below I use the ARD approach for the latent factors.[24]

[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).

[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 multiplicative gamma process prior on this term as in Bhattacharya and Dunson (2011). This prior takes the form $s_k \sim \mathcal{N}(0, \tau_k^{-1})$ with $\tau_k = \prod_{l=1}^{k} \delta_l$ and $\delta_l \sim$Gamma$(a_c, 1)$ for $a_c > 1$. Thus as the rank increases the precision are

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 6 inference is also highly accurate.

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

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.

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 5.1), the second area are the uses of interactive latent factor models of various types (Section 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 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.

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

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.

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.

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.

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.

5.2.1 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 $y_{i,j}$ 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).[25] 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.[26]

Recent work has extended these two-mode models to multiway relational data using tensor decompositions. Hoff (2011*a*) explores the CP decomposition used in this papere, and Hoff (2011*b*) uses a decomposition based on the more general Tucker product.[27] 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, 2011*a*). 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 an 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

[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 `amen` 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 (2011*b*) 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.

case.[28] I address all of these issues in the unified framework here.

5.2.2 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.

5.2.3 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; 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).[29]

[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.

[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

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.[30] Furthermore the regularization in Bayesian approaches often leads to improved parameter estimation for high dimensional cases such these.[31]

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 setting 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.[32]

5.2.4 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), neuro-imaging 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 (Grimmer and Stewart, 2013). 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

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*b*) argues 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).

exactly to the setting here (Hu et al., 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.

5.2.5 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.

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 Katznelson (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 4.3.2.

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.

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.[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.[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.[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.

[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.

[34] I thank Marc Ratkovic for suggesting this strategy.

[35] For example, the variational posterior may be useful for developing proposals in a Hamiltonian Monte Carlo framework (Neal, 2011).

6 Simulation Evidence

In this section I provide simulation evidence that the estimation framework outlined in Section 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 6.1 and then then two mode case (Section 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 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.

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).[36] 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 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.

[36] I 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.

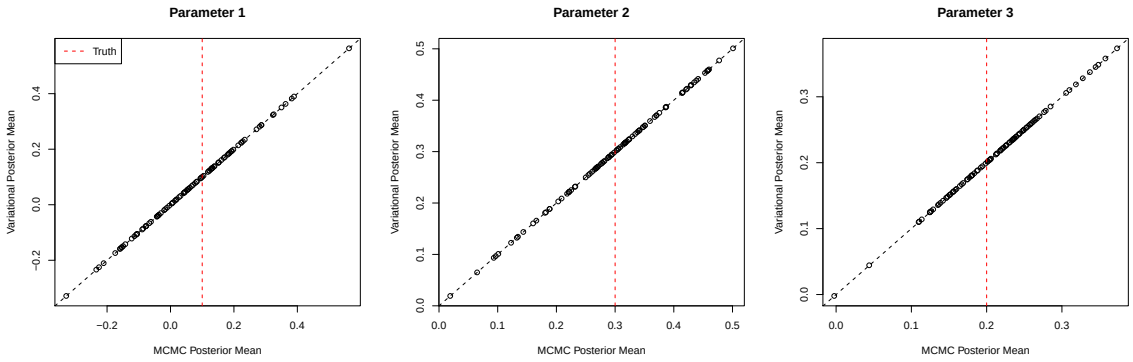


Figure 2: Recovery of the posterior mean compared to MCMC over 100 simulations on each of the three main parameters.

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 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 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).

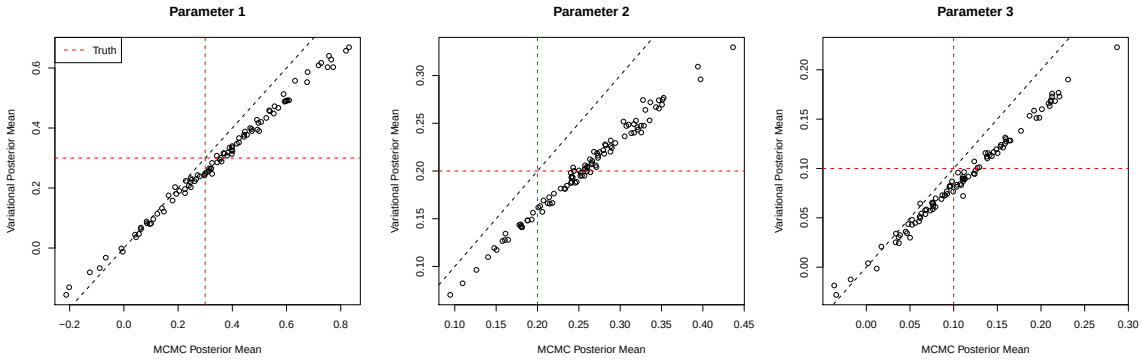


Figure 3: Comparison of posterior means between MCMC and Variational for the logistic regression case.

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 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 118×31 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:

$$y_{ij} = x_{ij}\beta + a_i + b_j + \sum_{k=1}^{K} u_{i,k}v_{j,k} + \epsilon \tag{49}$$

where a and b are country and time varying intercepts, U and V are the interactive factors and ϵ 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.[37] Note that the rank is estimated and not assumed by the variational algorithm.

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.[38]

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 shows the true and estimated parameters for the three blocks of parameters: globally shared regression coefficients (β), the random intercepts (a, b) and the inner product of the latent factors ($u'v$).

[37]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.

[38]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.

For the latter two I use a kernel density smoother (Wand, 2014*a*) 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%.

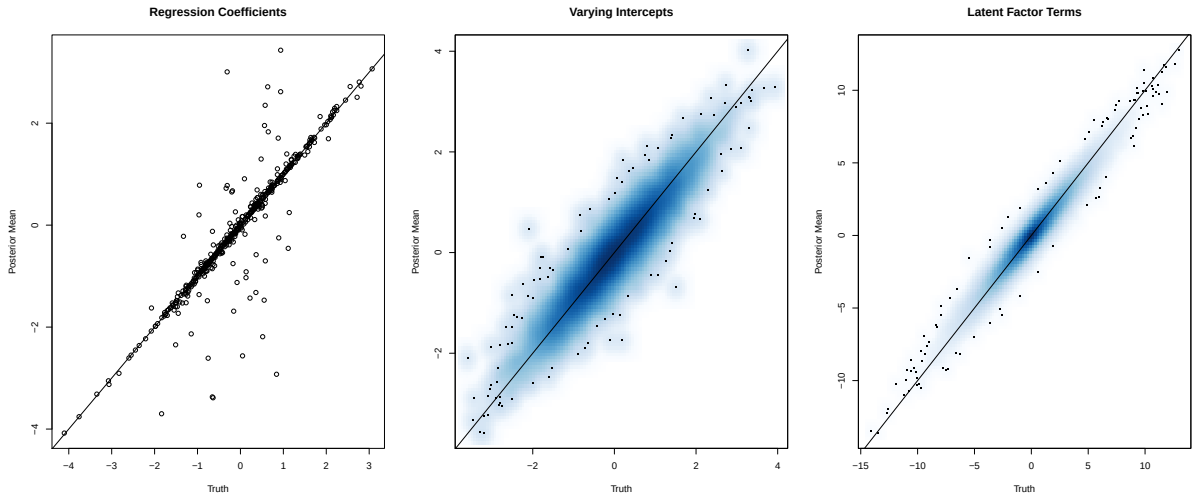


Figure 4: 50 simulations of the two mode model with estimated parameters in the variational algorithm. The left panel shows the globally shared regression coefficients with their estimated and true values across all 50 runs. The middle panel shows the same for the country and time intercepts. The right 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.

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 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 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 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 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.

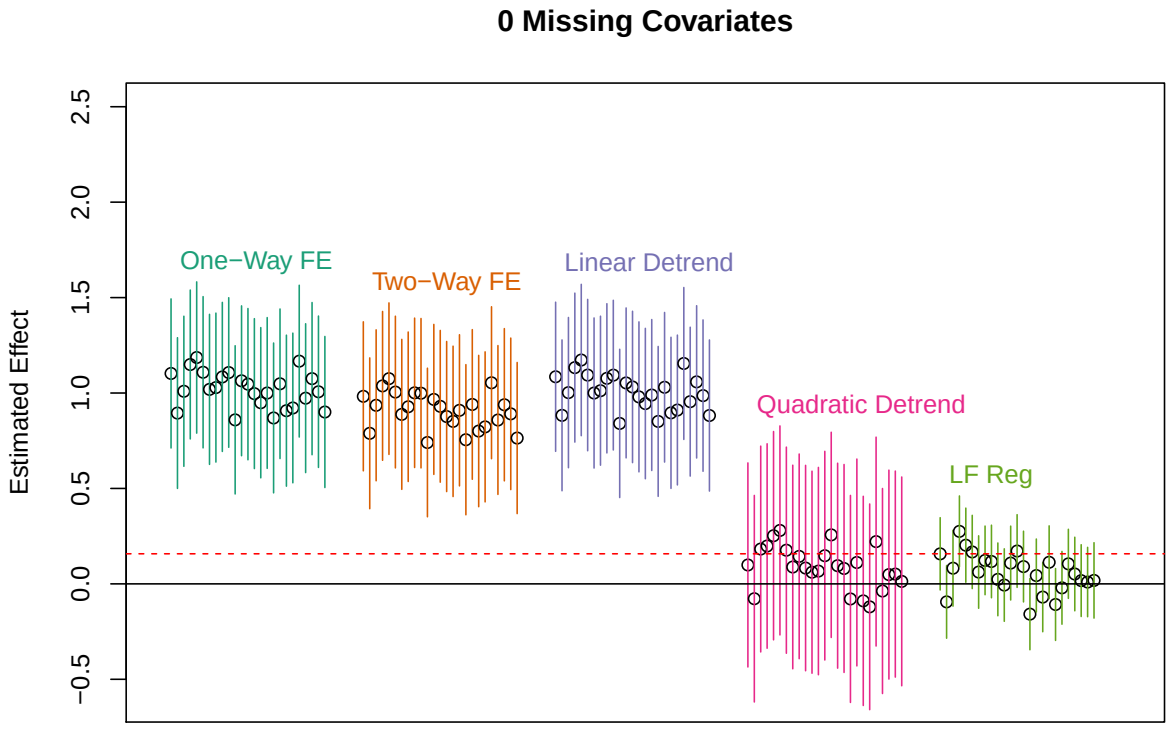


Figure 5: 25 simulations from a two-mode model with full observed covariates. Each of the five estimation strategies is shown with 95% confidence/credible intervals. The red dashed line indicated the true effect to be recovered. All but the quadratic detrending and the latent factor regression strategies massively overestimate the true effect size. The confidence intervals for the quadratic detrending are much too conservative comapred to the latent factor regression.

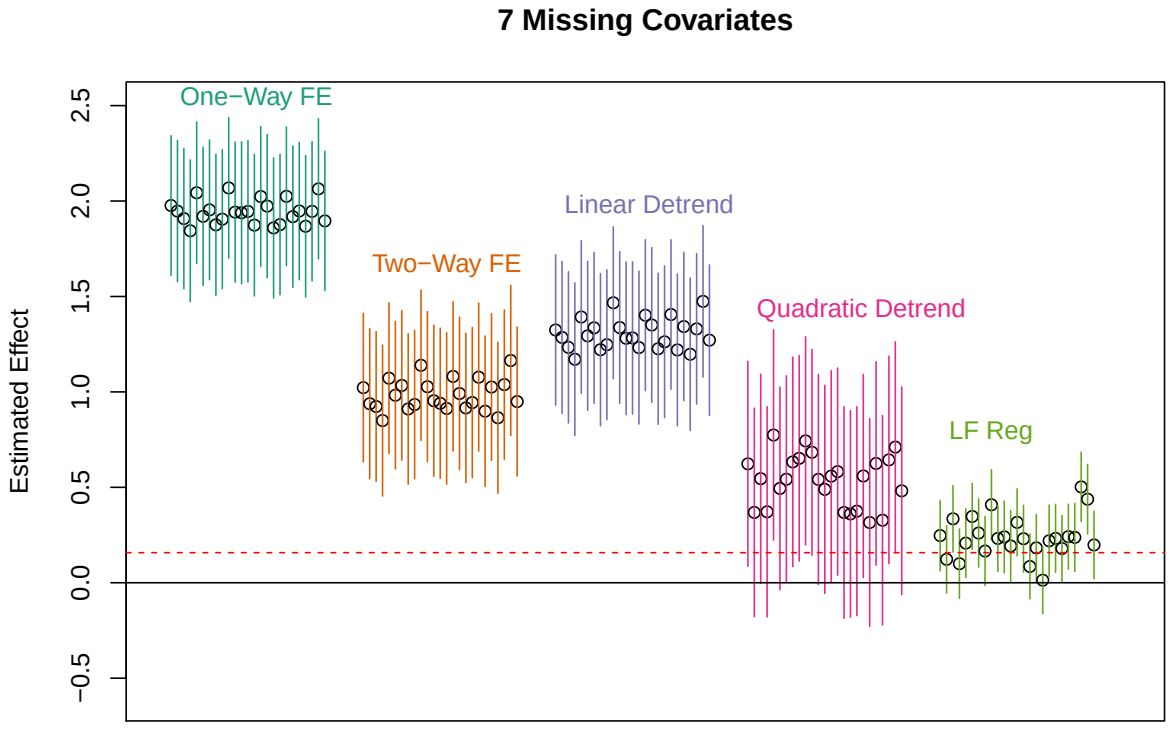


Figure 6: 25 simulations from a two-mode model with all but the covariate of interest missing. Each of the five estimation strategies is shown with 95% confidence/credible intervals. The red dashed line indicated the true effect to be recovered.

6.4 Overview and Limitations

The sequence of simulations above demonstrate the inference framework proposed in Section 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 estimate of 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.

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 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*.

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).

7.1.1 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.[39] The resulting model does not exhibit the strong temporal trends in the original model and 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.

7.1.2 Applying the Latent Factor Model

Using the framework presented in Section 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:

$$\text{fdi}_{c,t} = X_{c,t}\beta + a_c + b_t + \sum_k u_{c,k} v_{t,k} + \epsilon \tag{50}$$

[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 removed.

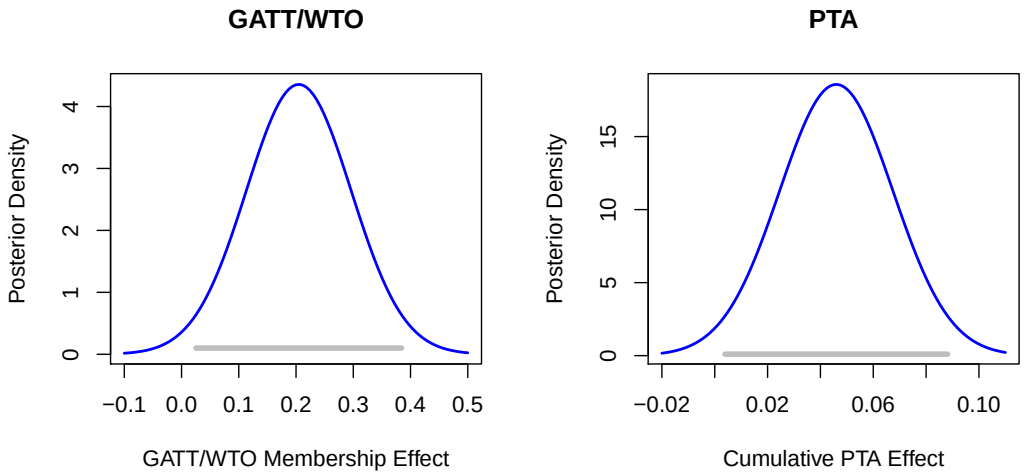


Figure 7: Posterior density of the two coefficients of theoretical interst: GATT/WTO and Cumulative PTAs for the latent factor model on data of Büthe and Milner (2008).

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 ϵ 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 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 substantively 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 8 plots a projection of the countries into a

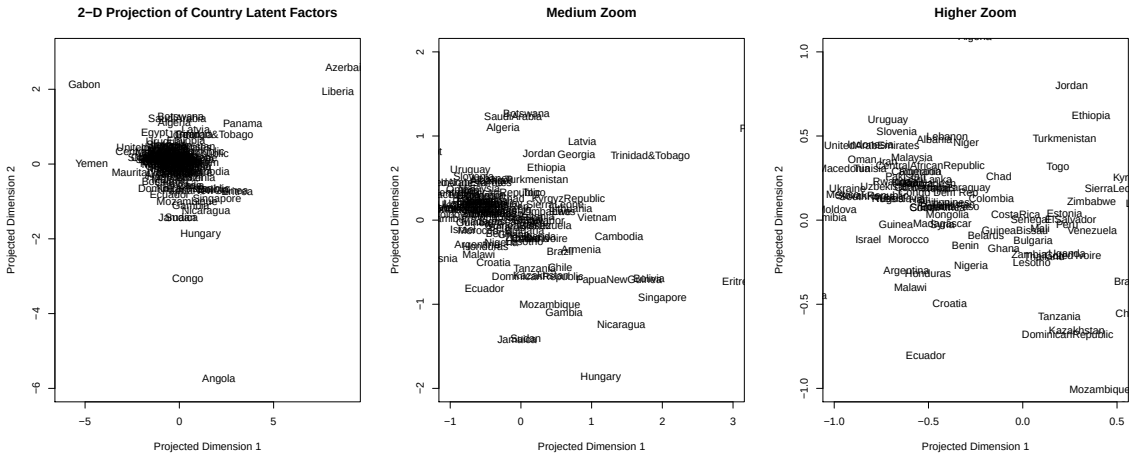


Figure 8: Posterior means of the country-specific interactive latent effects projected into 2 dimensions using Sammon multidimensional scaling. Each successive panel from left to right is further zoomed in. Countries which are close to each other respond to global temporal shocks in a similar way.

two dimensional space.[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 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.

7.1.3 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

[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.

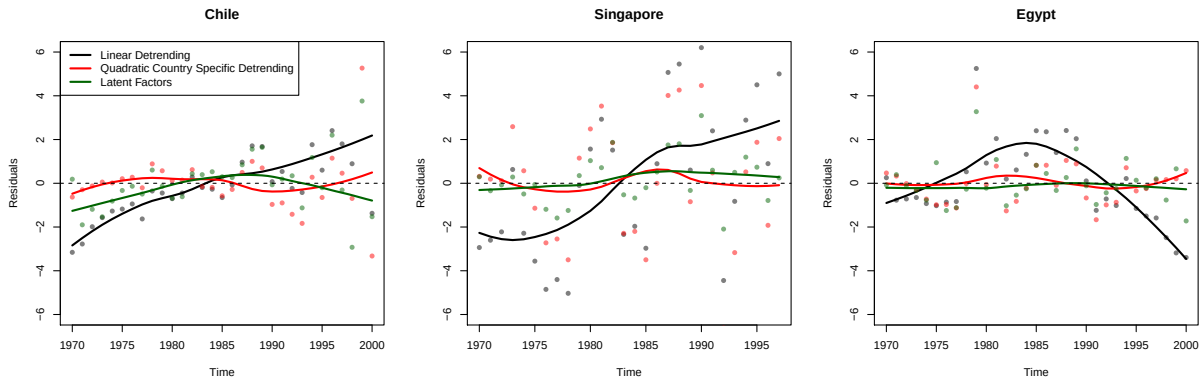


Figure 9: Time-series residual plots for three countries comparing the linear detrending with country fixed effects (black), country-specific quadratic detrending (red) and latent factor model (green). Lines give loess smoothed trends with a span of 3/4. This is a reproduction of Figure 9 in King and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

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).

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.[41] The Oneal and Russett (1999) model has been the subject of intense interest and scrutiny, including a large number of challenges and replies.[42] A particularly prominent case of such a challenge is the "Dirty Pool" debate discussed above.

[41] Earlier versions of this finding include Maoz and Russett (1993) and Oneal and Russet (1997)

[42] The 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).

7.2.1 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 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 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]

7.2.2 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

[43] It 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 (2011*a*).

[44] The 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.

receiver by r and the time by t, the model is

$$y_{s,r,t} \sim \text{Bernoulli}(\text{InvLogit}(\eta_{s,r,t})) \tag{51}$$

$$\eta_{s,r,t} = x_{s,r,t}\beta + x_{s,r,t}\gamma_t + a_s + b_r + c_t + \sum_k u^{(S)}_{s,k} u^{(R)}_{r,k} u^{(T)}_{t,k} \tag{52}$$

where γ_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 $u^{(S)}_{s,k}$ is the kth 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 γ_t I use a weakly informative hierarchical multivariate prior for the P covariates in the model:

$$\gamma_t|\Sigma \sim \text{Normal}(0, \Sigma) \tag{53}$$

$$\Sigma|\alpha_1 \dots \alpha_P \sim \text{Inverse-Wishart}\left(\nu + P - 1, 2\nu\text{diag}(1/\alpha_1, \dots, 1/\alpha_P)\right) \tag{54}$$

$$\alpha_p \sim \text{Inverse-Gamma}(.5, 1/A_p^2) \tag{55}$$

where p indexes the covariates and $\nu, A_1^2 \dots A_p^2$ 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.[45]

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.[46] 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 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

[45] With only 10 groups the covariance in the random effects Σ 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.

[46] 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.

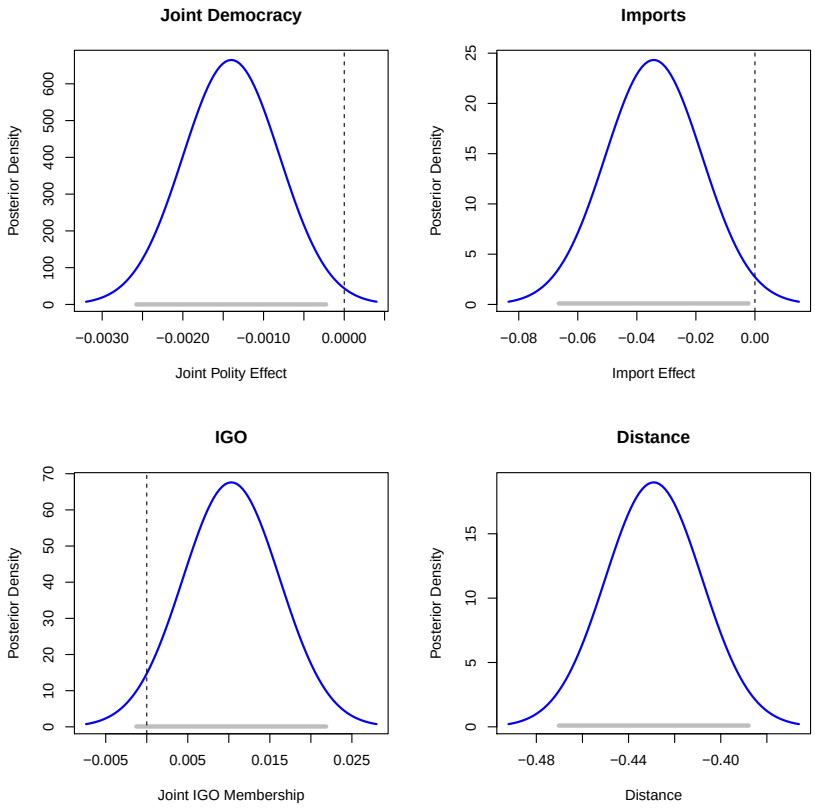


Figure 10: Posterior distribution of the main dyadic effects in the model of militarized interstate disputes. The grey line gives the 95% credible intervals and the dashed line marks 0.

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.

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.

A Online Appendix Road Map

In the appendix (accessible at scholar.harvard.edu/bstewart) 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).

References

Agarwal, Deepak and Bee-Chung Chen. 2009. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM pp. 19–28.

Agarwal, Deepak, Bee-Chung Chen and Bo Pang. 2011. Personalized recommendation of user comments via factor models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics pp. 571–582.

Aguilar, Omar and Mike West. 2000. "Bayesian dynamic factor models and portfolio allocation." *Journal of Business & Economic Statistics* 18(3):338–357.

Airoldi, Edoardo M, David M Blei, Stephen E Fienberg and Eric P Xing. 2008. "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research* 9(1981-2014):3.

Airoldi, Edoardo M, Thiago B Costa and Stanley H Chan. 2013. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*. pp. 692–700.

Allen, Genevera I and Robert Tibshirani. 2012. "Inference with transposable data: modelling the effects of row and column correlations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74(4):721–743.

Anandkumar, Anima, Rong Ge, Daniel Hsu and Sham M Kakade. 2013. "A tensor spectral approach to learning mixed membership community models." *arXiv:1302.2684* .

Anandkumar, Anima, Rong Ge, Daniel Hsu, Sham M Kakade and Matus Telgarsky. 2012. "Tensor decompositions for learning latent variable models." *arXiv:1210.7559* .

Anandkumar, Anima, Yi-kai Liu, Daniel J Hsu, Dean P Foster and Sham M Kakade. 2012. A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*. pp. 917–925.

Anandkumar, Animashree, Rong Ge and Majid Janzamin. 2014*a*. "Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates." *arXiv preprint arXiv:1402.5180* .

Anandkumar, Animashree, Rong Ge and Majid Janzamin. 2014*b*. "Provable Learning of Overcomplete Latent Variable Models: Semi-supervised and Unsupervised Settings." *arXiv preprint arXiv:1408.0553* .

Ando, Tomohiro and Jushan Bai. 2013. "Panel data models with grouped factor structure under unknown group membership." *Available at SSRN 2373629* .

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arceneaux, Kevin and David W Nickerson. 2009. "Modeling certainty with clustered data: A comparison of methods." *Political Analysis* 17(2):177–190.

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-groups Estimators*." *Oxford bulletin of Economics and Statistics* 49(4):431–434.

Aronow, Peter and Cyrus Samii. 2013. "Cluster-Robust Variance Estimation for Dyadic Data." *arXiv preprint arXiv:1312.3398* .

Bafumi, Joseph and Andrew E Gelman. 2006. "Fitting multilevel models when predictors and group effects correlate.".

Baglama, James and Lothar Reichel. 2005. "Augmented implicitly restarted Lanczos bidiagonalization methods." *SIAM Journal on Scientific Computing* 27(1):19–42.

Bai, Jushan. 2009. "Panel data models with interactive fixed effects." *Econometrica* 77(4):1229–1279.

Bai, Jushan and Kunpeng Li. 2014. "Theory and methods of panel data models with interactive effects." *The Annals of Statistics* 42(1):142–170.

Bai, Jushan and Serena Ng. 2002. "Determining the number of factors in approximate factor models." *Econometrica* 70(1):191–221.

Bai, Jushan and Serena Ng. 2008. *Large dimensional factor analysis*. Now Publishers Inc.

Bates, Douglas M. 2010. "lme4: Mixed-effects modeling with R.".

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* .

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving quantitative studies of international conflict: A conjecture." *American Political Science Review* pp. 21–35.

Beck, Nathaniel and Jonathan N Katz. 1995. "What to do (and not to do) with time-series cross-section data." *American political science review* pp. 634–647.

Beck, Nathaniel and Jonathan N Katz. 2001. "Throwing out the baby with the bath water: A comment on Green, Kim, and Yoon." *International Organization* 55(2):487–495.

Beck, Nathaniel and Jonathan N Katz. 2007. "Random coefficient models for time-series–cross-section data: Monte Carlo experiments." *Political Analysis* 15(2):182–195.

Beck, Nathaniel and Jonathan N Katz. 2011. "Modeling dynamics in time-series-cross-section political economy data." *Annual Review of Political Science* 14:331–352.

Beck, Nathaniel, Jonathan N Katz and Richard Tucker. 1998. "Taking time seriously: Time-series-cross-section analysis with a binary dependent variable." *American Journal of Political Science* 42(4):1260–1288.

Bell, Andrew and Kelvyn Jones. 2012. "Explaining fixed effects: random effects modeling of time-series cross-sectional and panel data." *Political Science Research and Methods* pp. 1–21.

Bennett, D Scott and Allan C Stam III. 2004. *The behavioral origins of war*. University of Michigan Press.

Besag, Julian, Jeremy York and Annie Mollié. 1991. "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics* 43(1):1–20.

Bhattacharya, Anirban and David B Dunson. 2011. "Sparse Bayesian infinite factor models." *Biometrika* 98(2):291–306.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. springer New York.

Blackwell, Matthew. 2013. "A framework for dynamic causal inference in political science." *American Journal of Political Science* 57(2):504–520.

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

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

Box-Steffensmeier, Janet M and Bradford S Jones. 2004. *Event history modeling: A guide for social scientists*. Cambridge University Press.

Box-Steffensmeier, Janet M, Dan Reiter and Christopher Zorn. 2003. "Nonproportional hazards and event history analysis in international relations." *Journal of Conflict Resolution* 47(1):33–53.

Brandt, Patrick T and John Williams. 2007. *Multiple time series models*. Sage.

Breslow, Norman E and David G Clayton. 1993. "Approximate inference in generalized linear mixed models." *Journal of the American Statistical Association* 88(421):9–25.

Brouhns, Natacha, Michel Denuit and Jeroen K Vermunt. 2002. "A Poisson log-bilinear regression approach to the construction of projected lifetables." *Insurance: Mathematics and Economics* 31(3):373–393.

Browne, William J and David Draper. 2006. "A comparison of Bayesian and likelihood-based methods for fitting multilevel models." *Bayesian Analysis* 1(3):473–514.

Büthe, Tim and Helen V Milner. 2008. "The politics of foreign direct investment into developing countries: increasing FDI through international trade agreements?" *American Journal of Political Science* 52(4):741–762.

Cai, Jian-Feng, Emmanuel J Candès and Zuowei Shen. 2010. "A singular value thresholding algorithm for matrix completion." *SIAM Journal on Optimization* 20(4):1956–1982.

Candès, Emmanuel J and Benjamin Recht. 2009. "Exact matrix completion via convex optimization." *Foundations of Computational mathematics* 9(6):717–772.

Carvalho, Carlos M, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang and Mike West. 2008. "High-dimensional sparse factor modeling: applications in gene expression genomics." *Journal of the American Statistical Association* 103(484).

Castagnetti, Carolina, Eduardo Rossi and Lorenzo Trapani. 2012. Inference on Factor Structures in Heterogeneous Panels. Technical report University of Pavia, Department of Economics and Management.

Chatterjee, Sourav. 2012. "Matrix estimation by universal singular value thresholding." *arXiv preprint arXiv:1212.1247* .

Chen, Bee-Chung, Jian Guo, Belle Tseng and Jie Yang. 2011. User reputation in a comment rating environment. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM pp. 159–167.

Chib, Siddhartha, Federico Nardari and Neil Shephard. 2002. "Markov chain Monte Carlo methods for stochastic volatility models." *Journal of Econometrics* 108(2):281–316.

Chu, Wei and Zoubin Ghahramani. 2009. "Probabilistic models for incomplete multi-dimensional arrays.".

Clarke, Kevin A., H. E. Goemans and Michael Peress. 2010. "Time and the Study of Conflict Problems of Temporal Aggregation in Quantitative International Relations.".

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98(02):355–370.

Cranmer, Skyler J and Bruce A Desmarais. 2011. "Inferential network analysis with exponential random graph models." *Political Analysis* 19(1):66–86.

Cunningham, John P and Zoubin Ghahramani. 2014. "Unifying linear dimensionality reduction." *arXiv preprint arXiv:1406.0873* .

Dafoe, Allan. 2011. "Statistical critiques of the democratic peace: Caveat emptor." *American Journal of Political Science* 55(2):247–262.

Dafoe, Allan, John R Oneal and Bruce Russett. 2013. "The democratic peace: Weighing the evidence and cautious inference." *International Studies Quarterly* 57(1):201–214.

Dawid, A Philip. 1981. "Some matrix-variate distribution theory: notational considerations and a Bayesian application." *Biometrika* 68(1):265–274.

De Marchi, Scott, Christopher Gelpi and Jeffrey D Grynaviski. 2004. "Untangling neural nets." *American Political Science Review* 98(02):371–378.

Donoho, David L and Jain M Johnstone. 1994. "Ideal spatial adaptation by wavelet shrinkage." *Biometrika* 81(3):425–455.

Dorff, Cassy and Michael D Ward. 2013. "Networks, Dyads, and the Social Relations Model." *Political Science Research and Methods* 1(02):159–178.

Driscoll, John C and Aart C Kraay. 1998. "Consistent covariance matrix estimation with spatially dependent panel data." *Review of economics and statistics* 80(4):549–560.

Durante, Daniele and David B Dunson. 2013. "Nonparametric Bayes dynamic modeling of relational data." *arXiv preprint arXiv:1311.4669* .

Eckart, Carl and Gale Young. 1936. "The approximation of one matrix by another of lower rank." *Psychometrika* 1(3):211–218.

Erikson, Robert S., Pablo M. Pinto and Kelly T. Rader. 2014. "Dyadic Analysis in International Relations: A Cautionary Tale." *Political Analysis* .

Fahrmeir, Ludwig and Stefan Lang. 2001. "Bayesian inference for generalized additive mixed models based on Markov random field priors." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50(2):201–220.

Fang, Guobin, Kani Chen and Bo Zhang. 2013. "Estimation of Dynamic Mixed Double Factors Model in High Dimensional Panel Data." *arXiv preprint arXiv:1301.2079* .

Farber, Henry S and Joanne Gowa. 1997. "Common interests or common polities? Reinterpreting the democratic peace." *The Journal of Politics* 59(02):393–417.

Fithian, William and Rahul Mazumder. 2013. "Scalable Convex Methods for Flexible Low-Rank Matrix Modeling." *arXiv preprint arXiv:1308.4211* .

Fosdick, Bailey K and Peter D Hoff. 2013. "Testing and modeling dependencies between a network and nodal attributes." *arXiv preprint arXiv:1306.4708* .

Fosdick, Bailey K and Peter D Hoff. 2014. "Separable factor analysis with applications to mortality data." *The Annals of Applied Statistics* 8(1):120–147.

Fowler, James and Nicholas A. Christakis. 2009. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Little, Brown and Company.

Franzese Jr, Robert J and Jude C Hays. 2007. "Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data." *Political Analysis* 15(2):140–164.

Freedman, David A. 2006. "On the so-called Huber sandwich estimator and robust standard errors." *The American Statistician* 60(4).

Gabriel, K Ruben. 1998. "Generalised bilinear regression." *Biometrika* 85(3):689–700.

Gaibulloev, Khusrav, Todd Sandler and Donggyu Sul. 2014. "Dynamic Panel Analysis under Cross-Sectional Dependence." *Political Analysis* 22(2):258–273.

Gartzke, Erik. 2000. "Preferences and the democratic peace." *International Studies Quarterly* 44(2):191–212.

Gartzke, Erik. 2007. "The capitalist peace." *American Journal of Political Science* 51(1):166–191.

Gelman, Andrew. 2004. "Exploratory data analysis for complex models." *Journal of Computational and Graphical Statistics* 13(4).

Gelman, Andrew. 2006. "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian analysis* 1(3):515–534.

Gelman, Andrew and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

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

Gerard, David and Peter Hoff. 2014. "Equivariant minimax dominators of the MLE in the array normal model." *arXiv preprint arXiv:1408.0424* .

Gill, Jeff. 2008. "Is partial-dimension convergence a problem for inferences from MCMC algorithms?" *Political Analysis* 16(2):153–178.

Girosi, Federico and Gary King. 2008. *Demographic forecasting*. Princeton University Press.

Gleditsch, Kristian S. and Michael D. Ward. 2008. *An Introduction to Spatial Regression Models in the Social Sciences*. Sage Publications.

Goldenberg, Anna, Alice X Zheng, Stephen E Fienberg and Edoardo M Airoldi. 2010. "A survey of statistical network models." *Foundations and Trends® in Machine Learning* 2(2):129–233.

Gopalan, Prem, Jake M Hofman and David M Blei. 2013. "Scalable Recommendation with Poisson Factorization." *arXiv preprint arXiv:1311.1704* .

Gourevitch, Peter and David A Lake. 2001. "Research Design and Method in International Relations: Editors' Introduction." *International Organization* 55(02):439–440.

Green, Donald P, Soo Yeon Kim and David H Yoon. 2001. "Dirty pool." *International Organization* 55(2):441–468.

Greene, William H. 2012. *Econometric Analysis*. Pearson.

Grimmer, Justin. 2010. "An introduction to Bayesian inference via variational approximations." *Political Analysis* .

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* .

Gu, Chong. 2013. *Smoothing spline ANOVA models*. Vol. 297 Springer.

Hadfield, Jarrod D. 2010. "MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package." *Journal of Statistical Software* 33(2):1–22.

Hainmueller, Jens and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22(2):143–168.

Hamilton, James Douglas. 1994. *Time series analysis*. Vol. 2 Princeton university press Princeton.

Harshman, Richard A and Margaret E Lundy. 1994. "PARAFAC: Parallel factor analysis." *Computational Statistics & Data Analysis* 18(1):39–72.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The elements of statistical learning*. Springer.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd. 1998. Characterizing selection bias using experimental data. Technical report National bureau of economic research.

Heckman, James J, Hidehiko Ichimura and Petra E Todd. 1997. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme." *The review of economic studies* 64(4):605–654.

Hegre, Håvard, John R Oneal and Bruce Russett. 2010. "Trade does promote peace: New simultaneous estimates of the reciprocal effects of trade and conflict." *Journal of Peace Research* 47(6):763–774.

Held, Leonhard, Isabel Natário, Sarah Elaine Fenton, Håvard Rue and Nikolaus Becker. 2005. "Towards joint disease mapping." *Statistical methods in medical research* 14(1):61–82.

Hoff, P.D., A.E. Raftery and M.S. Handcock. 2002. "Latent space approaches to social network analysis." *Journal of the american Statistical association* 97(460):1090–1098.

Hoff, P.D. and M.D. Ward. 2004. "Modeling dependencies in international relations networks." *Political Analysis* 12(2):160–175.

Hoff, Peter. 2008. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*. pp. 657–664.

Hoff, Peter, Bailey Fosdick, Alex Volfovsky and Katherine Stovel. 2013. "Likelihoods for fixed rank nomination networks." *Network Science* 1(03):253–277.

Hoff, Peter D. 2005. "Bilinear mixed-effects models for dyadic data." *Journal of the American Statistical Association* 100(469):286–295.

Hoff, Peter D. 2011*a*. "Hierarchical multilinear models for multiway data." *Computational Statistics & Data Analysis* 55(1):530–543.

Hoff, Peter D. 2011*b*. "Separable covariance arrays via the Tucker product, with applications to multivariate relational data." *Bayesian Analysis* 6(2):179–196.

Hoff, Peter David. 2013. "Equivariant and scale-free Tucker decomposition models." *arXiv preprint arXiv:1312.6397* .

Hoffman, Matthew D. and Andrew Gelman. 2013. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* in press.

Hsiao, Cheng and M Hashem Pesaran. 2008. *Random coefficient models*. Springer.

Hu, Changwei, Eunsu Ryu, David Carlson, Yingjian Wang and Lawrence Carin. 2014. Latent Gaussian Models for Topic Modeling. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. pp. 393–401.

Huang, Alan and Matthew P. Wand. 2013. "Simple marginally noninformative prior distributions for covariance matrices." *Bayesian Analysis* 8(2):439–452.

Huang, Furong, UN Niranjan, M Hakeem and Animashree Anandkumar. 2013. "Fast Detection of Overlapping Communities via Online Tensor Methods.".

Imai, Kosuke and Dustin Tingley. 2012. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56(1):218–236.

Imai, Kosuke and In Song Kim. 2012. On the use of linear fixed effects regression models for causal inference. Technical report Technical Report, Department of Politics, Princeton University. available at http://imai. princeton. edu/research/FEmatch. html.

Jaakkola, Tommi S and Michael I Jordan. 2000. "Bayesian parameter estimation via variational methods." *Statistics and Computing* 10(1):25–37.

Jackman, Simon. 2000. "Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo." *American Journal of Political Science* pp. 375–404.

James, William and Charles Stein. 1961. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Number 1961 *in* "1" pp. 361–379.

Jenke, Elizabeth and Christopher Gelpi. 2012. "Theme and Variations: Historical Contingencies in the Causal Model of Inter–State Conflict.".

Jordan, Michael I, Zoubin Ghahramani, Tommi S Jaakkola and Lawrence K Saul. 1999. "An introduction to variational methods for graphical models." *Machine learning* 37(2):183–233.

Kacowicz, Arie M. 1995. "Explaining zones of peace: democracies as satisfied powers?" *Journal of Peace Research* 32(3):265–276.

Keshk, Omar MG, Brian M Pollins and Rafael Reuveny. 2004. "Trade still follows the flag: The primacy of politics in a simultaneous model of interdependence and armed conflict." *Journal of Politics* 66(4):1155–1179.

Kim, Hyung Min and David L Rousseau. 2005. "The Classical Liberals Were Half Right (or Half Wrong): New Tests of the Liberal Peace, 1960-88." *Journal of Peace Research* 42(5):523–543.

King, Gary. 1998. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: University of Michigan Press.

King, Gary. 2001. "Proper nouns and methodological propriety: Pooling dyads in international relations data." *International Organization* 55(2):497–507.

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* .

Knowles, David A and Tom Minka. 2011. Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*. pp. 1701–1709.

Kolda, Tamara G and Brett W Bader. 2009. "Tensor decompositions and applications." *SIAM review* 51(3):455–500.

Lee, Nayoung, Hyungsik Roger Moon and Martin Weidner. 2012. "Analysis of interactive fixed effects dynamic linear panel regression with measurement error." *Economics Letters* 117(1):239–242.

Lee, Ronald D and Lawrence R Carter. 1992. "Modeling and forecasting US mortality." *Journal of the American statistical association* 87(419):659–671.

Lee, Yuen Yi Cathy and Matt Wand. 2014. "Streamlined Mean Field Variational Bayes for Longitudinal and Multilevel Data Analysis.".

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

Liu, Junrong, Robin Sickles and EG Tsionas. 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.".

Lopes, Hedibert Freitas, Dani Gamerman and Esther Salazar. 2011. "Generalized spatial dynamic factor models." *Computational Statistics & Data Analysis* 55(3):1319–1330.

Lopes, Hedibert Freitas, Esther Salazar and Dani Gamerman. 2008. "Spatial dynamic factor analysis." *Bayesian Analysis* 3(4):759–792.

Lu, Xun and Liangjun Su. 2013. "Shrinkage estimation of dynamic panel data models with interactive fixed effects.".

Luts, Jan and Matt P Wand. 2013. "Variational inference for count response semiparametric regression." *arXiv preprint arXiv:1309.4199* .

Mammen, Enno, Jens Perch Nielsen and Bernd Fitzenberger. 2011. "Generalized linear time series regression." *Biometrika* p. asr044.

Maoz, Zeev. 2011. *Networks of nations*. Cambridge University Press, Cambridge, UK.

Maoz, Zeev and Bruce Russett. 1993. "Normative and structural causes of democratic peace, 1946-1986." *American Political Science Review* pp. 624–638.

Marlin, Benjamin M, Mohammad Emtiyaz Khan and Kevin P Murphy. 2011. Piecewise Bounds for Estimating Bernoulli-Logistic Latent Gaussian Models. In *ICML*. pp. 633–640.

Martin, Andrew D and Kevin M Quinn. 2002. "Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis* 10(2):134–153.

Martin, Andrew D, Kevin M Quinn and Jong Hee Park. 2011. "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software* 42(9):1–21.

Mazumder, Rahul, Trevor Hastie and Robert Tibshirani. 2010. "Spectral regularization algorithms for learning large incomplete matrices." *The Journal of Machine Learning Research* 11:2287–2322.

McCullagh, P. and John A. Nelder. 1989. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability Taylor & Francis.

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

Menictas, Marianne and Matt P Wand. 2013. "Variational inference for marginal longitudinal semiparametric regression." *Stat* 2(1):61–71.

Menon, Aditya Krishna and Charles Elkan. 2011. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases*. Springer pp. 437–452.

Miller, Kurt, Michael I Jordan and Thomas L Griffiths. 2009. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*. pp. 1276–1284.

Moon, H and Martin Weidner. 2010*a*. "Dynamic linear panel regression models with interactive fixed effects." *manuscript, UCL* .

Moon, Hyungsik and Martin Weidner. 2010*b*. "Linear regression for panel with unknown number of factors as interactive fixed effects." *Unpublished manuscript, University of Southern California* .

Mousseau, Michael. 2013. "The Democratic Peace Unraveled: Its the Economy1." *International Studies Quarterly* 57(1):186–197.

Mousseau, Michael, Omer F Orsun, Jameson Lee Ungerer and Demet Yalcin Mousseau. 2013. "Capitalism and peace: Its Keynes, not Hayek." *Assessing the Capitalist Peace* pp. 80–109.

Mundlak, Yair. 1978. "On the pooling of time series and cross section data." *Econometrica: journal of the Econometric Society* pp. 69–85.

Nakajima, Shinichi and Masashi Sugiyama. 2011. "Theoretical analysis of Bayesian matrix factorization." *The Journal of Machine Learning Research* 12:2583–2648.

Nakajima, Shinichi, Masashi Sugiyama, S Derin Babacan and Ryota Tomioka. 2013. "Global analytic solution of fully-observed variational Bayesian matrix factorization." *The Journal of Machine Learning Research* 14(1):1–37.

Nakajima, Shinichi, Ryota Tomioka, Masashi Sugiyama and S. D. Babacan. 2012. Perfect Dimensionality Recovery by Variational Bayesian PCA. In *Advances in Neural Information Processing Systems 25*. pp. 971–979.

National Research Council. 2013. *Frontiers in Massive Data Analysis*. The National Academies Press.

Neal, Radford. 2011. "MCMC Using Hamiltonian Dynamics." *Handbook of Markov Chain Monte Carlo* pp. 113–162.

Neville, Sarah, John Ormerod and Matt Wand. 2012. "Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies.".

Oneal, John R and Bruce M Russet. 1997. "The classical liberals were right: Democracy, interdependence, and conflict, 1950–1985." *International Studies Quarterly* 41(2):267–294.

Oneal, John R and Bruce Russett. 1999. "Assessing the liberal peace with alternative specifications: Trade still reduces conflict." *Journal of Peace Research* 36(4):423–442.

Oneal, John R and Bruce Russett. 2001. "Clear and clean: The fixed effects of the liberal peace." *International Organization* 55(2):469–485.

Oneal, John R and Bruce Russett. 2005. "Rule of three, let it be? When more really is better." *Conflict Management and Peace Science* 22(4):293–310.

Orbanz, Peter and Daniel M Roy. 2013. "Bayesian models of graphs, arrays and other exchangeable random structures." *arXiv preprint arXiv:1312.7857* .

Ormerod, JT and MP Wand. 2010. "Explaining variational approximations." *The American Statistician* 64(2):140–153.

Ormerod, JT and MP Wand. 2012. "Gaussian variational approximate inference for generalized linear mixed models." *Journal of Computational and Graphical Statistics* 21(1):2–17.

Pang, Xun. 2010. "Modeling heterogeneity and serial correlation in binary time-series cross-sectional data: A Bayesian multilevel model with AR (p) Errors." *Political Analysis* 18(4):470–498.

Pang, Xun. 2014. "Varying Responses to Common Shocks and Complex Cross-Sectional Dependence: Dynamic Multilevel Modeling with Multifactor Error Structures for Time-Series Cross-Sectional Data." *Political Analysis* .

Park, Jong Hee. 2012. "A Unified Method for Dynamic and Cross-Sectional Heterogeneity: Introducing Hidden Markov Panel Models." *American Journal of Political Science* 56(4):1040–1054.

Pesaran, M Hashem. 2006. "Estimation and inference in large heterogeneous panels with a multifactor error structure." *Econometrica* 74(4):967–1012.

Pesaran, M Hashem, Yongcheol Shin and Ron P Smith. 1999. "Pooled mean group estimation of dynamic heterogeneous panels." *Journal of the American Statistical Association* 94(446):621–634.

Pham, Tung H and Matt P Wand. 2014. "Generalized Additive Mixed Model Analysis via gammSlice.".

Pinheiro, José C and Douglas M Bates. 2000. *Mixed-effects models in S and S-PLUS*. Springer.

Polson, Nicholas G, James G Scott and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108(504):1339–1349.

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

Porteous, Ian, Arthur U Asuncion and Max Welling. 2010. Bayesian Matrix Factorization with Side Information and Dirichlet Process Mixtures. In *AAAI*.

Raftery, Adrian E, Xiaoyue Niu, Peter D Hoff and Ka Yee Yeung. 2012. "Fast inference for the latent space network model using a case-control approximate likelihood." *Journal of Computational and Graphical Statistics* 21(4):901–919.

Rai, Piyush, Yingjian Wang, Shengbo Guo, Gary Chen, David Dunson and Lawrence Carin. 2014. "Scalable Bayesian Low-Rank Decomposition of Incomplete Multiway Tensors.".

Ranganath, Rajesh, Sean Gerrish and David M Blei. 2013. "Black box variational inference." *arXiv preprint arXiv:1401.0118* .

Rasmussen, Carl Edward and CKI Williams. 2006. "Gaussian processes for machine learning." *The MIT Press, Cambridge, MA, USA* 38:715–719.

Roberts, Margaret E, Brandon M Stewart and Dustin Tingley. N.d. Navigating the Local Modes of Big Data: The Case of Topic Models. Technical report Harvard University.

Roychowdhury, Anirban and Brian Kulis. 2014. "Gamma Processes, Stick-Breaking, and Variational Inference." *arXiv preprint arXiv:1410.1068* .

Rue, Havard and Leonhard Held. 2004. *Gaussian Markov random fields: theory and applications*. CRC Press.

Rue, Håvard, Sara Martino and Nicolas Chopin. 2009. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the royal statistical society: Series b (statistical methodology)* 71(2):319–392.

Salakhutdinov, Ruslan and Andriy Mnih. 2007. Probabilistic Matrix Factorization. In *NIPS*.

Salimans, Tim and David A Knowles. 2013. "Fixed-form variational posterior approximation through stochastic linear regression." *Bayesian Analysis* 8(4):837–882.

Salter-Townshend, Michael and Thomas Brendan Murphy. 2013. "Variational Bayesian inference for the latent position cluster model for network data." *Computational Statistics & Data Analysis* 57(1):661–671.

Sammon, John W. 1969. "A nonlinear mapping for data structure analysis." *IEEE Transactions on computers* 18(5):401–409.

Sarafidis, Vasilis and Tom Wansbeek. 2012. "Cross-sectional dependence in panel data analysis." *Econometric Reviews* 31(5):483–531.

Scott, James G and Liang Sun. 2013. "Expectation-maximization for logistic regression." *arXiv preprint arXiv:1306.0040* .

Seeger, Matthias and Guillaume Bouchard. 2012. Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Proceedings of the 15th international conference on artificial intelligence and statistics*.

Shor, Boris, Joseph Bafumi, Luke Keele and David Park. 2007. "A Bayesian multilevel modeling approach to time-series cross-sectional data." *Political Analysis* 15(2):165–181.

Snijders, Tom AB and Roel J Bosker. 1999. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications Limited.

Stan Development Team. 2014. "Stan: A C++ Library for Probability and Sampling, Version 2.2.".

Stegmueller, Daniel. 2013. "How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches." *American Journal of Political Science* 57(3):748–761.

Su, Liangjun, Sainan Jin and Yonghui Zhang. 2012. "Specification test for panel data models with interactive fixed effects.".

Suzuki, Taiji. 2014. "Convergence rate of Bayesian tensor estimator: Optimal rate without restricted strong convexity." *arXiv preprint arXiv:1408.3092* .

Tan, Linda SL and David J Nott. 2013. "Variational inference for generalized linear mixed models using partially noncentered parametrizations." *Statistical Science* 28(2):168–188.

Tobler, Waldo R. 1970. "A computer movie simulating urban growth in the Detroit region." *Economic geography* pp. 234–240.

Tucker, Ledyard R. 1964. "The extension of factor analysis to three-dimensional matrices." *Contributions to mathematical psychology* pp. 109–127.

Volfovsky, Alexander and Peter D Hoff. 2012. "Hierarchical array priors for ANOVA decompositions." *arXiv preprint arXiv:1208.1726* .

Wahba, Grace. 1990. *Spline models for observational data*. Vol. 59 Siam.

Wainwright, Martin J and Michael I Jordan. 2008. "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning* 1(1-2):1–305.

Wand, Matt. 2014*a*. *KernSmooth: Functions for kernel smoothing for Wand and Jones (1995)*. R package version 2.23-12.

Wand, Matt P. 2014*b*. "Fully simplified multivariate normal updates in non-conjugate variational message passing." *Journal of Machine Learning Research* 15:1351–1369.

Wand, Matthew P, John T Ormerod, Simone A Padoan and Rudolf Fuhrwirth. 2011. "Mean field variational Bayes for elaborate distributions." *Bayesian Analysis* 6(4):847–900.

Wang, Chong and David M Blei. 2013. "Variational inference in nonconjugate models." *The Journal of Machine Learning Research* 14(1):1005–1031.

Wang, Yuchung J and George Y Wong. 1987. "Stochastic blockmodels for directed graphs." *Journal of the American Statistical Association* 82(397):8–19.

Ward, Michael D, John S Ahlquist and Arturas Rozenas. 2013. "Gravity's Rainbow: A dynamic latent space model for the world trade network." *Network Science* 1(01):95–118.

Ward, Michael D, Katherine Stovel and Audrey Sacks. 2011. "Network analysis and political science." *Annual Review of Political Science* 14:245–264.

Ward, Michael D, Randolph M Siverson and Xun Cao. 2007. "Disputes, democracies, and dependencies: A reexamination of the Kantian peace." *American Journal of Political Science* 51(3):583–601.

Wawro, Gregory J and Ira Katznelson. 2013. "Designing Historical Social Scientific Inquiry: How Parameter Heterogeneity Can Bridge the Methodological Divide between Quantitative and Qualitative Approaches." *American Journal of Political Science* .

West, M. and J. Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics Springer.

West, Mike. 2003. "Bayesian factor regression models in the large p, small n paradigm." *Bayesian statistics* 7:733–742.

Western, Bruce. 1998. "Causal heterogeneity in comparative research: A Bayesian hierarchical modelling approach." *American Journal of Political Science* 42:1233–1259.

Winn, John M and Christopher M Bishop. 2005. Variational message passing. In *Journal of Machine Learning Research*. pp. 661–694.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. MIT press.

Zhang, Liang, Deepak Agarwal and Bee-Chung Chen. 2011. Generalizing matrix factorization through flexible regression priors. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM pp. 13–20.

Zhao, Qibin, Liqing Zhang and Andrzej Cichocki. 2014. "Bayesian CP Factorization of Incomplete Tensors with Automatic Rank Determination." *arXiv preprint arXiv:1401.6497* .

Zhao, Yihua, John Staudenmayer, Brent A Coull and Matthew P Wand. 2006. "General design Bayesian generalized linear mixed models." *Statistical Science* pp. 35–51.

Zhou, Hua and Lexin Li. 2014. "Regularized matrix regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(2):463–483.

Zhou, Hua, Lexin Li and Hongtu Zhu. 2013. "Tensor regression with applications in neuroimaging data analysis." *Journal of the American Statistical Association* 108(502):540–552.

Zhukov, Yuri M and Brandon M Stewart. 2013. "Choosing Your Neighbors: Networks of Diffusion in International Relations1." *International Studies Quarterly* 57(2):271–287.