Online Appendix:
Latent Factor Regressions for the Social Sciences

Brandon Stewart[*]

November 30, 2014

Appendix RoadMap

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

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.

[*] Graduate Student. Department of Government Harvard University. bstewart@fas.harvard.edu

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).

A Variational Inference Algorithms

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

Specific novel contributions include:

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

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

3. Initialization Strategies using Spectral Methods
Prior work on variational algorithms has generally not discussed the important role of initialization. I draw on recently developed spectral methods for parameter estimation to develop strong initialization strategies for the model.

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

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

[1] This contribution is about model design and is not further discussed in this appendix. Estimation is treated in the main paper Section 4.3.2 and follows straightforwardly from Algorithms 1-6 below.

	# of Latent Factors		
	1	2 (Matrix)	3+ (Tensor)
Gaussian Reg.	Alg 1	Alg 3*	Alg 5*
Logit	Alg 2	Alg 4*	Alg 6*

Table 1: Algorithms detailed below. Those marked with $*$ are new to this paper.

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

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

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

A.1 Summary of Technical Contributions

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

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

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

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

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

One of the challenges in the use of variational inference algorithms is the presence of many local optima in the objective function. Although initialization of variational algorithms is rarely discussed, it is extremely important in this particular case. I address this by using recent spectral estimators for the parameters. For the matrix case I leverage the results of (Nakajima et al., 2013) which establish a direct connection between a truncated singular value decomposition and the global variational solution for fully-observed Gaussian matrix factorization. Despite the obvious implications for initializing variational inference I've seen no prior work that leverages this connection.[5]

A.1.1 Comparison to Existing Implementations

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

Peter Hoff's group has released software which covers the matrix factorization case with observed covariates where the outcome matrix is symmetric (`eigenmodel` `R` package (Hoff, 2012), `amen` `R` package (Hoff et al., 2014)). These packages are designed for the

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

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

[4] I know of no variational algorithms for binary outcome tensor factorization models. Instead recent work has approached this computational problem by using various tricks to speed up Gibbs sampling (Rai et al., 2014).

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

analysis of (small) undirected networks and use MCMC methods which can be quite slow to converge. They can handle some missingness in the outcome through imputation within the model. These packages are extremely effective for their intended purpose but fail for large data and non-network settings. For example in the FDI example from the main paper the data does not form an undirected network and thus cannot be modeled using either of the aforementioned packages. The democratic peace application can be treated as a collection of 10 networks (one for each time point) but cannot be modeled together.

Bada and Liebel (Bada and Liebl, 2014) have released an `R` package for panel data analysis (`phtt`) which includes an implementation of the interactive fixed effects framework described in Bai (2009). In contrast to the Bayesian estimators considered here this uses maximum likelihood which requires a parameterization of the fixed effects that make the parameter estimates order dependent.[6] It also crucially requires balanced panels (i.e. that all cross-sections contain the same number of observations) and only allows for Gaussian outcomes. While the estimators are substantially faster than the network models, the balanced panel restriction is an extremely demanding requirement in practice. For the FDI analysis the 118 country panel has approximately 30% of the cells missing and would require a substantial drop in either the number of countries or the number of time periods to be estimable under the interactive fixed effects framework. This is a typical problem for social science data particularly in international relations and comparative politics.[7]

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

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

Surprisingly there aren't even any variational implementation of hierarchical linear and generalized linear models. There are several `R` packages for MCMC algorithms such as `MCMCpack` but fast alternatives are limited to quasi-likelihood methods and Laplace approximations such as provided by `lme4`. Thus even implementation of Algorithms 1 and 2 (which are not novel in themselves) constitutes a useful contribution to the research community. Implementations which are sufficiently scalable to accommodate political science data with hundreds of groups and thousands of observations are made possible

[6] That is, the parameter estimate for a country is different if its listed first rather than last. This is less problematic for a naturally ordered set of groups such as time but is more annoying for unordered groups such as countries.

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

by the computational strategies described in Lee and Wand (2014) for which my software will be the first publicly available implementation.[8]

A.2 Algorithm 1: Hierarchical Gaussian Linear Models

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

A.2.1 Preliminaries

In the case of a single mode problem the latent factor regression framework reduces to hierarchical modeling. Algorithm 1 covers the particular case of a Gaussian likelihood with varying intercepts and slopes and weakly informative priors. With groups indexed by g the model is given by

$$y|\beta, u \sim \text{Normal}(X\beta + Zu, \sigma^2_\epsilon) \tag{1}$$

$$u_g|\Sigma^R \sim \text{Normal}(0, \Sigma^R) \tag{2}$$

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

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

$$\sigma^2_\epsilon \sim \text{Inverse-Gamma}(a_\epsilon, b_\epsilon) \tag{3}$$

$$\Sigma^R \sim \text{Inverse-Wishart}(A_{\Sigma^R}, B_{\Sigma^R}) \tag{4}$$

$$\beta \sim \text{Normal}(0, \sigma^2_\beta I_P) \tag{5}$$

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

A.2.2 Non-Conjugate Priors

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

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

$$\rho^2_{i,r} \sim \text{Inverse-Gamma}(.5, 1/a_{i,r}) \tag{6}$$

$$a_{i,r} \sim \text{Inverse-Gamma}(.4, 1/A^2_{i,r}) \tag{7}$$

[8] The algorithms in Lee and Wand (2014) are non-trivial to implement deriving mostly from careful inversion of particular types of sparse matrices. Thus a practitioner is unlikely to find the paper and implement the methods on their own.

where the marginal distribution for $\rho^2_{i,r}$ is now Half-Cauchy($A_{i,r}$). The multivariate extension of the Half Cauchy distribution is given by Huang and Wand (2013)

$$\Sigma^R | a^R_1, \ldots, a^R_{q^R} \sim \text{Inverse-Wishart}(\nu + q^R - 1, 2\nu\text{diag}(1/a^R_1, \ldots, 1/a^R_{q^R}) \tag{8}$$

$$a^R_1 \ldots a^R_{q^R} \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(.5, 1/A^2_{Rr}) \tag{9}$$

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

A.2.3 Variational Approximation

The approximation to the full joint posterior is

$$p(\beta, u, a^R, a_u, a_\epsilon, \Sigma^R, \sigma^2_u \sigma^2_\epsilon) \approx q(\beta, u, a^r, a_u, a_\epsilon) q(\Sigma^R, \sigma^2_u, \sigma^2_\epsilon) \tag{10}$$

$$= q(\beta, u) q(\Sigma^R) q(\sigma^2_\epsilon) q(a_\epsilon) \prod_{r=1}^{q_r} q(a^R_r) \prod_{\ell=1}^{L} q(a_{u\ell}) \prod_{\ell=1}^{L} q(\sigma^2_{u\ell}) \tag{11}$$

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

Thus the approximate evidence lower bound can be given as:

$$\log p(y; q) = E_q\{\log p(\beta.u, a^R, a_u, a_\epsilon, \Sigma^R, \sigma^2_u, \sigma^2_\epsilon) - \log q(\beta.u, a^R, a_u, a_\epsilon, \Sigma^R, \sigma^2_u, \sigma^2_\epsilon)\} \tag{12}$$

Under standard variational inference theory (Bishop, 2006; Grimmer, 2010), the optimal approximating densities to maximize Equation 12 for a generic parameter θ take the form

$$q(\theta) = \exp(E_{q(-\theta)} \text{log}(p(\theta | \text{rest})) \tag{13}$$

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

A.2.4 Optimal Variational Densities

Algebraic manipulations show these forms to be

$$\begin{aligned}
q(\beta, u) &= \text{Normal}(\mu_{q(\beta,u)}, \Sigma_{q(\beta,u)}) \\
q(\sigma^2_\epsilon) &= \text{Inverse-Gamma}(.5(N+1), B_{q(\sigma^2_\epsilon)}) \\
q(a_\epsilon) &= \text{Inverse-Gamma}(1, B_{q(a_\epsilon)}) \\
q(\sigma^2_{u\ell}) &= \text{Inverse-Gamma}(.5(q^G_\ell + 1), B_{q(\sigma^2_{u\ell})} \\
q(a_{u\ell}) &= \text{Inverse-Gamma}(1, B_{q(a_{u\ell})}) \\
q(a^R_r) &= \text{Inverse-Gamma}(.5(\nu + q^R), B_{q(a^R_r)}) \\
q(\Sigma^R) &= \text{Inverse-Wishart}(\nu + m + q^r - 1, B_{q(\Sigma^R)})
\end{aligned}$$

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

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

This allows us to use the following sequence of updates

$$\begin{aligned}
G_i &\leftarrow \mu_{q(1/\sigma^2_\epsilon)}(C^G_i)^T X^R_i && (14)\\
H_i &\leftarrow \{\mu_{q(1/\sigma^2_\epsilon)}(X^R_i)^T(X^R_i) + M_{q((\Sigma^R)^{-1})}\}^{-1} && (15)\\
S &\leftarrow S + G_i H_i G_i^T && (16)\\
s &\leftarrow s + G_i H_i (X^R_i)^T y_i && (17)\\
\Sigma_{q(\beta,u^G)} &\leftarrow \left\{\mu_{q(1/\sigma^2_\epsilon)}(C^G)^T C^G + \begin{bmatrix}\sigma_\beta^{-2} & 0\\ 0 & \text{blockdiag}(\mu_{q(1/\sigma^2_{u\ell})} I_{q^G_\ell})\end{bmatrix} - S\right\}^{-1} && (18)\\
\mu_{q(\beta,u^G)} &\leftarrow \mu_{q(1/\sigma^2_\epsilon)}\Sigma_{q(\beta,u^G)}\left\{(C^G)^T y - s\right\} && (19)\\
\Sigma_{q(u^R_i)} &\leftarrow H_i + H_i G_i^T \Sigma_{q(\beta,u^G)} G_i H_i && (20)\\
\mu_{q(u^R_i)} &\leftarrow H_i\left\{\mu_{q(1/\sigma^2_\epsilon)}(X^R_i)^T y_i - G_i^T \mu_{q(\beta,u^G)}\right\} && (21)
\end{aligned}$$

$$\begin{aligned}
B_{q(\sigma^2_\epsilon)} \leftarrow \mu_{q(1/a_\epsilon)} + .5 &\left[\left\| y - C^G\mu_{q(\beta,u^G)} - \begin{bmatrix} X^R_1\mu_{q(u^R_1)}\\ \vdots \\ X^R_m\mu_{q(u^R_m)}\end{bmatrix}\right\|\right] + \text{tr}\{(C^G)^T C^G \Sigma_{q(\beta,u^G)}\} \qquad (22)\\
&+ \sum_{i=1}^m \text{tr}\{(X^R_i)^T X^R_i \Sigma_{q(u^R_i)}\} - 2\mu^{-1}_{q(1/\sigma^2_\epsilon)} \sum_{i=1}^m \text{tr}\left(G_i H_i G_i^T \Sigma_{q(\beta,u^G)}\right)
\end{aligned}$$

$$\mu_{q(1/\sigma_\epsilon^2)} \leftarrow .5(\sum_{i=1}^{m} n_i + 1)/B_{q(\sigma_\epsilon^2)} \tag{23}$$

$$\mu_{q(1/a_\epsilon)} \leftarrow 1/\{\mu_{q(1/\sigma_\epsilon^2)} + A_\epsilon^{-2}\} \tag{24}$$

$$B_{q(a_r^R)} \leftarrow \nu \left(M_{q((\Sigma^R)^{-1})}\right)_{rr} + A_{Rr}^{-2} \tag{25}$$

$$\mu_{q(1/a_r^R)} \leftarrow .5(\nu + q^R)/B_{q(a_r^R)} \tag{26}$$

$$B_{q(\Sigma^R)} \leftarrow \sum_{i=1}^{m} \left(\mu_{q(u_i^R)}\mu^T_{q(u_i^R)} + \Sigma_{q(u_i^R)}\right) + 2\nu \mathrm{diag}\left(\mu_{q(1/a_1^R)}, \ldots, \mu_{q(1/a_{q^R}^R)}\right) \tag{27}$$

$$M_{q((\Sigma^R)^{-1})} \leftarrow (\nu + m + q^R - 1)B^{-1}_{q(\Sigma^R)} \tag{28}$$

$$\mu_{q(1/a_{u\ell}))} \leftarrow 1/\{\mu_{q(1/\sigma^2_{u\ell})} + A_{u\ell}^{-2}\} \tag{29}$$

$$\mu_{q(1/\sigma^2_{u\ell})} \leftarrow \frac{q_\ell^G + 1}{2\mu_{q(1/a_{u\ell})} + \left|\mu_{q(\mu_\ell^G)}\right|^2 + \mathrm{tr}(\Sigma_{q(u_\ell^G)})} \tag{30}$$

A.2.5 Algorithm 1

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

```
repeat
    $S \leftarrow 0$
    $s \leftarrow 0$
    for $i = 1 \ldots m$ do
        Update $G_i$ (14), Update $H_i$ (15)
        Update $S$ (16), Update $s$ (17)
    end for
    Update $\Sigma_{q(\beta,u^G)}$ (18, Update $\mu_{(q(\beta,u^G)}$ using (19)
    for $i = 1 \ldots m$ do
        Update $\Sigma_{q(u_i^R)}$ (20), Update $\mu_{q(u_i^R)}$ (21)
    end for
    Update $B_{q(\sigma_\epsilon^2)}$ (22)
    Update $\mu_{q(1/\sigma_\epsilon^2)}$ (23)
    Update $\mu_{q(1/a_\epsilon)}$ (24)
    for $r = 1, \ldots, q^R$ do
        Update $B_{q(a_r^R)}$ (25), Update $\mu_{q(1/a_r^R)}$ (26)
    end for
    Update $B_{q(\Sigma^R)}$ (27)
    Update $M_{q(\Sigma^R)^{-1}}$ (28)
    for $\ell = 1 \ldots L$ do
        Update $\mu_{q(1/a_{u\ell})}$ (29), Update $\mu_{q(1/\sigma^2_{u\ell})}$ (30)
    end for
until convergence in $p(y; q)$
```

A.2.6 Computation

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

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

A.3 Algorithm 2: Hierarchical Logistic Regression

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

A.3.1 Variational Approximation

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

$$P(y|\eta) = \sigma(y\eta) \tag{31}$$

where η is the linear predictor and σ is the sigmoid function $\frac{1}{1+\exp(-\eta)}$.[9]

The log-likelihood is then

$$\text{log}p(y) = \sum_n \log(\sigma(y_n \eta_n)) \tag{32}$$

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

$$\sigma(y\eta) \geq \sigma(\xi)\text{exp}\left((y\eta - \xi)/2 - \lambda(\xi)\left((y\eta)^2 - \xi^2\right)\right) \tag{33}$$

$$\lambda(\xi) = \tanh(\xi/2)/(4\xi) \tag{34}$$

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.

[9] Although this representation is less standard in the social sciences, the symmetric form of the likelihood simplifies the notation below.

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

$$\xi = \sqrt{E[\eta^2]} \tag{35}$$

$$= \sqrt{\text{diagonal}\left(C\{\Sigma_{q(\beta,u,\xi)} + \mu_{q(\beta,u;\xi)}\mu^T_{q(\beta,u,\xi)}\}C^T\right)} \tag{36}$$

Thus the entire procedure contains only closed form updates and does not need to resort to numerical optimization. Because the approximation to the sigmoid function is a lower bound, the Evidence Lower Bound is still a true lower bound on $\log(p(y))$.

The justification of Jaakkola and Jordan (2000) is based on constructing a lower bound for the marginal likelihood using convex duality. Additionally, recent work by Scott and Sun (2013) has given a probabilistic interpretation showing the connection to data augmentation using the Polya-Gamma latent variable family (Polson, Scott and Windle, 2013).

A.3.2 Optimal Densities

Conditional latent variables ξ we get a normal density for $q(\beta, u)$ which means that optimization proceeds much as in Algorithm 1. I define some new terms (in the same style as above) before defining Algorithm 2.

$$G_i \leftarrow 2(C^G_i)^T \text{diag}\{\lambda(\xi_i)\} X^R_i \tag{37}$$

$$H_i \leftarrow \{2(X^R_i)^T \text{diag}\{\lambda(\xi_i)\}(X^R_i) + M_{q((\Sigma^R)^{-1})}\}^{-1} \tag{38}$$

$$S \leftarrow S + G_i H_i G_i^T \tag{39}$$

$$s \leftarrow s + G_i H_i (X^R_i)^T (y_i - .5) \tag{40}$$

$$\Sigma_{q(\beta,u^G;\xi)} \leftarrow \left\{ 2(C^G)^T \text{diag}\{\lambda(\xi_i)\} C^G + \begin{bmatrix} \sigma_\beta^{-2} & 0 \\ 0 & \text{blockdiag}(\mu_{q(1/\sigma^2_{u\ell})} I_{q^G_\ell}) \end{bmatrix} - S \right\}^{-1} \tag{41}$$

$$\mu_{q(\beta,u^G;\xi)} \leftarrow \mu_{q(1/\sigma^2_\epsilon)} \Sigma_{q(\beta,u^G)} \left\{ (C^G)^T (y - .5) - s \right\} \tag{42}$$

$$\Sigma_{q(u^R_i)} \leftarrow H_i + H_i G_i^T \Sigma_{q(\beta,u^G)} G_i H_i \tag{43}$$

$$\mu_{q(u^R_i)} \leftarrow H_i \left\{ \mu_{q(1/\sigma^2_\epsilon)} (X^R_i)^T (y_i - .5) - G_i^T \mu_{q(\beta,u^G)} \right\} \tag{44}$$

$$\xi^2 \leftarrow \text{diagonal} \left\{ C^G (\Sigma_{q(\beta,u;\xi)} + \mu_{q(\beta,u;\xi)} \mu^T_{q(\beta,u,\xi)})(C^G)^T \right\} \tag{45}$$

$$\begin{aligned} \xi_i^2 \leftarrow {} & 2\text{diagonal} \left(C^G \left(-\Sigma_{q(\beta,u;\xi)} G_i H_i + \mu_{q(\beta,u;\xi)} \mu^T_{q(\beta,u,\xi)} \right) (X^R_i)^T \right) \\ & + \text{diagonal} \left(X^R_i \left(\Sigma_{q(u^R_i;\xi)} + \mu_{q(u_i;\xi)} \mu^T_{q(u_i,\xi)} \right) (X^R_i \right) \end{aligned} \tag{46}$$

$$B_{q(a_r^R)} \leftarrow \nu \left(M_{q((\Sigma^R)^{-1})}\right)_{rr} + A_{Rr}^{-2} \tag{47}$$

$$\mu_{q(1/a_r^R)} \leftarrow .5(\nu + q^R)/B_{q(a_r^R)} \tag{48}$$

$$B_{q(\Sigma^R)} \leftarrow \sum_{i=1}^{m} \left(\mu_{q(u_i^R)}\mu_{q(u_i^R)}^T + \Sigma_{q(u_i^R)}\right) + 2\nu \mathrm{diag}\left(\mu_{q(1/a_1^R)}, \ldots, \mu_{q(1/a_{q^R}^R)}\right) \tag{49}$$

$$M_{q((\Sigma^R)^{-1})} \leftarrow (\nu + m + q^R - 1)B_{q(\Sigma^R)}^{-1} \tag{50}$$

$$\mu_{q(1/a_{u\ell})} \leftarrow 1/\{\mu_{q(1/\sigma_{u\ell}^2)} + A_{u\ell}^{-2}\} \tag{51}$$

$$\mu_{q(1/\sigma_{u\ell}^2)} \leftarrow \frac{q_\ell^G + 1}{2\mu_{q(1/a_{u\ell'\xi})} + \left|\mu_{q(\mu_\ell^G)}\right|^2 + \mathrm{tr}(\Sigma_{q(u_\ell^G;\xi)})} \tag{52}$$

A.3.3 Algorithm 2

With the updates above we can now state Algorithm 2.

```
repeat
    $S \leftarrow 0$
    $s \leftarrow 0$
    for $i = 1 \ldots m$ do
        Update $G_i$ (37), Update $H_i$ (38)
        Update $S$ (39), Update $s$ (40)
    end for
    Update $\Sigma_{q(\beta,u^G;\xi)}$ (41, Update $\mu_{(q(\beta,u^G;\xi)}$ using (42)
    for $i = 1 \ldots m$ do
        Update $\Sigma_{q(u_i^R;\xi)}$ (43), Update $\mu_{q(u_i^R;\xi)}$ (44)
    end for
    Update $\xi^2$ (45)
    for $i = 1 \ldots m$ do
        Update $\xi_i^2$ (46)
    end for
    for $r = 1, \ldots, q^R$ do
        Update $B_{q(a_r^R)}$ (47), Update $\mu_{q(1/a_r^R)}$ (48)
    end for
    Update $B_{q(\Sigma^R)}$ (49)
    Update $M_{q(\Sigma^R)^{-1}}$ (50)
    for $\ell = 1 \ldots L$ do
        Update $\mu_{q(1/a_{u\ell})}$ (51), Update $\mu_{q(1/\sigma_{u\ell}^2)}$ (52)
    end for
until convergence in $p(y;q)$
```

A.4 Algorithm 3: Gaussian Outcome Matrix Factorization

In Algorithm 3 I show how to connect the Gaussian linear regression in Algorithm 1 with a matrix factorization model. The model can be stated as:

$$\begin{aligned}
y_{i,j} &\sim \text{Normal}(x_{ij}\beta + Z_{ij}\gamma + \sum_k u_{i,k}v_{j,k}, \sigma^2_\epsilon) && (53)\\
u_{i,k} &\sim \text{Normal}(0, \rho^2_k) && (54)\\
v_{i,k} &\sim \text{Normal}(0, \tau^2_k) && (55)\\
\beta &\sim \text{Normal}(0, \sigma^2_\beta I_P) && (56)\\
\gamma &\sim \text{Normal}(0, \Sigma^R) && (57)\\
\Sigma^R | a^R_1, \dots, a^R_{q^R} &\sim \text{Inverse-Wishart}(\nu + q^R - 1, 2\nu \text{diag}(1/a^R_1, \dots, 1/a^R_{q^R}) && (58)\\
a^R_1 \dots a^R_{q^R} &\overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(.5, 1/A^2_{Rr}) && (59)\\
\sigma^2_\epsilon &\sim \text{Inverse-Gamma}(.5, 1/a_\epsilon) && (60)\\
a_\epsilon &\sim \text{Inverse-Gamma}(.4, 1/A^2_\epsilon) && (61)
\end{aligned}$$

where I have switched the notation of the random effect to γ to reserve u as one of the latent factors.

In order to get the dimensionality selection effects of Automatic Relevance Determination we will point estimate the variances ρ^2, τ^2 as explained below.

When notationally convenient I collect the latent factors u into a matrix U where each row i contains the k factors for group i. We denote the row of matrix U contain the latent factors of group i as U_i. V follows similarly.

A.4.1 Matrix Factorization Component

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

$$\begin{aligned}
q(U, V, \beta) &\approx q(U)q(V)q(\beta) && (62)\\
&= \prod_{i=1}^{I} q(U_i) \prod_{j=1}^{J} q(V_j) q(\beta, \gamma) && (63)
\end{aligned}$$

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

The consequence of the factorization assumption is that the approximation is unable to capture the posterior covariance between the latent factor matrices $q(U)$ and $q(V)$. In the true posterior these effects are going to be negatively correlated, and it indeed it is exactly this feature which makes Gibbs sampling challenging. This hurts the accuracy of the approximation and will in general cause the approximation to understate the variance. That said, this does not appear to substantially detract from the quality of the approximation for the other parameters $q(\beta)$.[10]

[10] Note that in practice I always include standard additive effects for the rows and columns of the

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

$$q(U_i) = \text{Normal}(\mu_{q(U_i)}, \Sigma_{q(U_i)}) \tag{64}$$
$$q(V_j) = \text{Normal}(\mu_{q(V_j)}, \Sigma_{q(V_j)}) \tag{65}$$
$$q(\beta) = \text{Normal}(\mu_{q(\beta)}, \Sigma_{(q(\beta)}) \tag{66}$$

The posterior parameters of the approximation are updated as

$$\Sigma_{q(U_i)} = \left(\begin{pmatrix} 1/\tau_1^2 & 0 & \dots & 0 \\ 0 & 1/\tau_2^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1/\tau_k^2 \end{pmatrix} + \sum_{j=1}^{j} \frac{\Sigma_{(q(V_j)} + \mu_{q(V_j)}\mu_{q(V_j)}^T}{\sigma_\epsilon^2} \right)^{-1} \tag{67}$$

$$\mu_{q(U_i)} = \Sigma_{(q(U_i)} \left(\sum_{n \in \Omega} \left(\frac{(y_n - x_n\beta)\mu_{q(V_{j(n)})}}{\sigma_\epsilon^2} \right) \right) \tag{68}$$

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

$$\tilde{y_{ij}} = y_{ij} - E[U_i]E[V_j^T] \tag{69}$$
$$= y_{ij} - \mu_{q(U_i)}\mu_{q(V_j)}^T \tag{70}$$

Note that this has a tractable form due to the factorization assumption that defines $E[U_i, V_j^T] = E[U_i]E[V_j^T]$.

A.4.2 Optimal Densities

Most of the optimal densities follow by replacing the outcome y with a working response. In addition the following three updates are required for algorithm 3.

$$\tau_k^2 = \frac{1}{I-1} \sum_{i=1}^{I} \left((\Sigma_{q(U_i)})_{kk} + (\mu_{q(U_i)})^T \mu_{q(U_i)} \right) \tag{71}$$

$$\rho_k^2 = \frac{1}{J-1} \sum_{j=1}^{J} \left((\Sigma_{q(V_j)})_{kk} + (\mu_{q(V_j)})^T \mu_{q(V_j)} \right) \tag{72}$$

The update for the error rate parameter is substantially more complicated due to the inclusion of the latent factors. It is useful to define it in several pieces, using $\eta_{C\beta}$

matrix into the components β, γ. This has the benefit of weakening the dependence on the factorization assumption.

to indicate the fitted portion of the linear predictor due to the observed parameters and additive effects and η_{UV^T} to indicate the fitted portion due to the latent factors.

$$\begin{aligned}\eta^{(C\beta)} &= y - C^G \mu_{q(\beta,\gamma^G)} - \begin{bmatrix} X_1^R \mu_{q(\gamma_1^R)} \\ \vdots \\ X_m^R \mu_{q(\gamma_m^R)} \end{bmatrix} \\ \eta_{i,j}^{(UV)} &= \mu_{q(U_i)} \mu_{q(V_j)}^T \end{aligned}$$

Also define a term ζ to capture a portion of the regression context.

$$\zeta_{C\beta} = \mathrm{tr}\{(C^G)^T C^G \Sigma_{q(\beta,\gamma^G)}\} + \sum_{i=1}^{m} \mathrm{tr}\{(X_i^R)^T X_i^R \Sigma_{q(\gamma_i^R)}\} - 2\mu_{q(1/\sigma_\epsilon^2)}^{-1} \sum_{i=1}^{m} \mathrm{tr}\left(G_i H_i G_i^T \Sigma_{q(\beta,\gamma^G)}\right)$$

We also define a term to capture a piece from the matrix factorization component. Here we double index y as though it is arranged into a matrix.

$$\zeta_{UV^T} = \sum_{i,j\in\Omega} y_{i,j} - 2y_{i,j}\mu_{q(U_i)}\mu_{q(V_j)}^T + \mathrm{tr}\left[\left(\Sigma_{q(U_i)} + \mu_{q(U_i)}\mu_{q(U_i)}^T\right)\left(\Sigma_{q(V_j)} + \mu_{q(V_j)}\mu_{q(V_j)}^T\right)\right]$$

The trace term can be efficiently computed because both matrices are symmetric and thus the trace is the sum over their elementwise product:

$$\mathrm{tr}(AB) = \sum_{ij} A_{ij} B_{ij}$$

With these components together we can right the update as

$$B_{q(\sigma_\epsilon^2)} \leftarrow \zeta_{UV^T} - 2\left(\sum_{i,j\in\Omega} (y_{ij} - \eta_{ij}^{(UV^T)})\eta_{ij}^{(CB)}\right) + \left|\eta_{ij}^{(C\beta)}\right|^2 + \zeta_C\beta \qquad (73)$$

A.4.3 Algorithm 3

Updates after y_{working} should use the working version of the response in place of y.

```
repeat
    $S \leftarrow 0$
    $s \leftarrow 0$
    $y_{\text{working}} \leftarrow y - \eta^{(UV)}$
    for $i = 1 \ldots m$ do
        Update $G_i$ (14), Update $H_i$ (15)
        Update $S$ (16), Update $s$ (17)
    end for
    Update $\Sigma_{q(\beta,\gamma^G)}$ (18, Update $\mu_{(q(\beta,\gamma^G)}$ using (19)
    for $i = 1 \ldots m$ do
```

```
        Update $\Sigma_{q(\gamma_i^R)}$ (20), Update $\mu_{q(\gamma_i^R)}$ (21)
    end for
    Update $y_{\text{working}} \leftarrow y - \eta^{(C\beta)}$
    for $i = 1 \dots I$ do
        Update $\Sigma_{q(U_i)}$ (67)
        Update $\mu_{q(U_i)}$ (68)
    end for
    for $j = 1 \dots J$ do
        Update $\Sigma_{q(V_j)}$ (67)
        Update $\mu_{q(V_j)}$ (68)
    end for
    Update $\tau^2$ (71)
    Update $\rho^2$ (72)
    $y_{\text{working}} \leftarrow y$
    Update $B_{q(\sigma^2_\epsilon)}$ (73)
    Update $\mu_{q(1/\sigma^2_\epsilon)}$ (23)
    Update $\mu_{q(1/a_\epsilon)}$ (24)
    for $r = 1, \dots, q^R$ do
        Update $B_{q(a^R_r)}$ (25), Update $\mu_{q(1/a^R_r)}$ (26)
    end for
    Update $B_{q(\Sigma^R)}$ (27)
    Update $M_{q(\Sigma^R)^{-1}}$ (28)
    for $\ell = 1 \dots L$ do
        Update $\mu_{q(1/a_{\gamma\ell})}$ (29), Update $\mu_{q(1/\sigma^2_{\gamma\ell})}$ (30)
    end for
until convergence in $p(y; q)$
```

A.4.4 Initialization Methods

Due to the multimodality in the posterior distribution it is helpful to initialize the variational algorithm carefully. Here and in Algorithm 4-6, I initialize by updating the coefficients β, γ and then using the spectral algorithm of Nakajima et al. (2013) to estimate $q(U)q(V)$ from $y - E[X\beta + Z\gamma]$.

In short, the Nakajima et al. (2013) approach involves using a truncated singular value decomposition to estimate the parameters of the variational posterior. In the case of imbalanced panels this requires filling in the missing elements of the matrix. Using Ω to denote the observed indices, I use a simple mean imputation:

$$Y_{!\Omega} \leftarrow E[Y_\Omega] \tag{74}$$

This is the same procedure as used in Chatterjee (2012) which provides a theoretical justification for the choice. An alternative strategy would be to impute the values using an algorithm such as Soft-Impute (Mazumder, Hastie and Tibshirani, 2010) which is itself based on a singular value decomposition. Extensive study of the properties of these estimators is left to future work.

The exact version of the algorithm used is in Nakajima et al. (2012) with supporting details in Nakajima and Sugiyama (2011); Nakajima et al. (2013). Further details will be filled in here as well in future drafts.

A.4.5 Dimensionality Reduction

In algorithms 3-6 I use Automatic Relevance Determination to choose the dimensionality of the latent factors. This involves point estimating the factor variances ρ and τ which produces a model-induced regularization effect (see for example Nakajima et al. (2013) on the origins of this effect). In practice this means estimating the model with the maximal number of latent factors and dropping them out as their variance parameters go to zero.

An added benefit of using the spectral initialization is that we get an initial estimate of the dimensionality which can help defray computational costs. Then as the variances fall below a certain threshold they are dropped from the model.

A.5 Algorithm 4: Logistic Regression with Matrix Factorization

A.5.1 Preliminaries

The model can be given as:

$$y_{i,j} \sim \text{Bernoulli}(\sigma(\eta)) \tag{75}$$

$$\eta = x_{ij}\beta + Z_{ij}\gamma + \sum_k u_{i,k}v_{j,k} \tag{76}$$

$$u_{i,k} \sim \text{Normal}(0, \rho_k^2) \tag{77}$$

$$v_{i,k} \sim \text{Normal}(0, \tau_k^2) \tag{78}$$

$$\beta \sim \text{Normal}(0, \sigma_\beta^2 I_P) \tag{79}$$

$$\gamma \sim \text{Normal}(0, \Sigma^R) \tag{80}$$

$$\Sigma^R | a_1^R, \ldots, a_{q^R}^R \sim \text{Inverse-Wishart}(\nu + q^R - 1, 2\nu \text{diag}(1/a_1^R, \ldots, 1/a_{q^R}^R) \tag{81}$$

$$a_1^R \ldots a_{q^R}^R \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(.5, 1/A_{Rr}^2) \tag{82}$$

The logistic regression with matrix factorization problem is slightly more complicated than the Gaussian regression because it is slightly more difficult to define the working response. Using the formulation in Equation 33 we can show the working response for the estimation of $q(\beta, \gamma)$ is $y/2 - 2y^2\eta^*\lambda(\xi)$ where η^* is the portion of the linear predictor to be removed.

A.5.2 Optimal Densities

All the optimal densities remain in the same families as given in Algorithms 2 and 3.

$$G_i \leftarrow 2(C_i^G)^T \mathrm{diag}\{\lambda(\xi_i)\} X_i^R \tag{83}$$

$$H_i \leftarrow \{2(X_i^R)^T \mathrm{diag}\{\lambda(\xi_i)\}(X_i^R) + M_{q((\Sigma^R)^{-1})}\}^{-1} \tag{84}$$

$$S \leftarrow S + G_i H_i G_i^T \tag{85}$$

$$s \leftarrow s + G_i H_i (X_i^R)^T (y_{\text{working}}) \tag{86}$$

$$\Sigma_{q(\beta,\gamma^G;\xi)} \leftarrow \left\{ 2(C^G)^T \mathrm{diag}\{\lambda(\xi_i)\} C^G + \begin{bmatrix} \sigma_\beta^{-2} & 0 \\ 0 & \mathrm{blockdiag}(\mu_{q(1/\sigma^2_{\gamma\ell})} I_{q_\ell^G}) \end{bmatrix} - S \right\}^{-1} \tag{87}$$

$$\mu_{q(\beta,\gamma^G;\xi)} \leftarrow \mu_{q(1/\sigma_\epsilon^2)} \Sigma_{q(\beta,\gamma^G)} \left\{ (C^G)^T (y_{\text{working}}) - s \right\} \tag{88}$$

$$\Sigma_{q(\gamma_i^R)} \leftarrow H_i + H_i G_i^T \Sigma_{q(\beta,\gamma^G)} G_i H_i \tag{89}$$

$$\mu_{q(\gamma_i^R)} \leftarrow H_i \left\{ (X_i^R)^T (y_{\text{working}}) - G_i^T \mu_{q(\beta,\gamma^G)} \right\} \tag{90}$$

$$\Sigma_{q(U_i)} \leftarrow \left(\begin{pmatrix} 1/\tau_1^2 & 0 & \dots & 0 \\ 0 & 1/\tau_2^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1/\tau_k^2 \end{pmatrix} + \sum_{j=1}^{j} \frac{\Sigma_{(q(V_j)} + \mu_{q(V_j)} \mu_{q(V_j)}^T}{\lambda(\xi)} \right)^{-1} \tag{91}$$

$$\mu_{q(U_i)} \leftarrow \Sigma_{(q(U_i)} \left(\sum_{n \in \Omega} \left(\frac{(y_{\text{working}} - x_n \beta) \mu_{q(V_{j(n)})}}{\lambda(\xi)} \right) \right) \tag{92}$$

$$\xi^2 \leftarrow \text{diagonal} \left\{ E[(C\beta + UV^T)(C\beta + UV^T)^T] \right\} \tag{93}$$

$$B_{q(a_r^R)} \leftarrow \nu \left(M_{q((\Sigma^R)^{-1})} \right)_{rr} + A_{Rr}^{-2} \tag{94}$$

$$\mu_{q(1/a_r^R)} \leftarrow .5(\nu + q^R)/B_{q(a_r^R)} \tag{95}$$

$$B_{q(\Sigma^R)} \leftarrow \sum_{i=1}^{m} \left(\mu_{q(u_i^R)} \mu_{q(u_i^R)}^T + \Sigma_{q(u_i^R)} \right) + 2\nu \mathrm{diag} \left(\mu_{q(1/a_1^R)}, \dots, \mu_{q(1/a_{q^R}^R)} \right) \tag{96}$$

$$M_{q((\Sigma^R)^{-1})} \leftarrow (\nu + m + q^R - 1) B_{q(\Sigma^R)}^{-1} \tag{97}$$

$$\mu_{q(1/a_{u\ell})} \leftarrow 1/\{\mu_{q(1/\sigma_{u\ell}^2)} + A_{u\ell}^{-2}\} \tag{98}$$

$$\mu_{q(1/\sigma_{u\ell}^2)} \leftarrow \frac{q_\ell^G + 1}{2\mu_{q(1/a_{u\ell'\xi})} + \left| \mu_{q(\mu_\ell^G)} \right|^2 + \mathrm{tr}(\Sigma_{q(u_\ell^G;\xi)})} \tag{99}$$

$$\tau_k^2 = \frac{1}{I-1} \sum_{i=1}^{I} \left((\Sigma_{q(U_i)})_{kk} + (\mu_{q(U_i)})^T \mu_{q(U_i)} \right) \tag{100}$$

$$\rho_k^2 = \frac{1}{J-1} \sum_{j=1}^{J} \left((\Sigma_{q(V_j)})_{kk} + (\mu_{q(V_j)})^T \mu_{q(V_j)} \right) \tag{101}$$

A.5.3 Algorithm 4

With the statements above we can now give Algorithm 4

```
repeat
    $S \leftarrow 0$
    $s \leftarrow 0$
    Update $y_{\text{working}} \leftarrow y/2 - 2y^2(\mu_{q(U)}\mu_{q(V)}^T)\lambda(\xi)$
    for $i = 1 \ldots m$ do
        Update $G_i$ (83), Update $H_i$ (84)
        Update $S$ (85), Update $s$ (86)
    end for
    Update $\Sigma_{q(\beta,\gamma^G;\xi)}$ (87, Update $\mu_{(q(\beta,\gamma^G;\xi)}$ using (88)
    for $i = 1 \ldots m$ do
        Update $\Sigma_{q(\gamma_i^R;\xi)}$ (89), Update $\mu_{q(\gamma_i^R;\xi)}$ (90)
    end for
    Update $y_{\text{working}} \leftarrow y/2 - 2y^2(C\mu_{q(\beta,\gamma)})\lambda(\xi)$
    for $i = 1 \ldots I$ do
        Update $\Sigma_{q(U_i)}$ (67)
        Update $\mu_{q(U_i)}$ (68)
    end for
    for $j = 1 \ldots J$ do
        Update $\Sigma_{q(V_j)}$ (67)
        Update $\mu_{q(V_j)}$ (68)
    end for
    Update $\tau^2$ (71)
    Update $\rho^2$ (72)
    $y_{\text{working}} \leftarrow y$
    Update $\xi^2$ (93)
    for $r = 1, \ldots, q^R$ do
        Update $B_{q(a_r^R)}$ (94), Update $\mu_{q(1/a_r^R)}$ (95)
    end for
    Update $B_{q(\Sigma^R)}$ (96)
    Update $M_{q(\Sigma^R)^{-1}}$ (97)
    for $\ell = 1 \ldots L$ do
        Update $\mu_{q(1/a_{u\ell})}$ (98), Update $\mu_{q(1/\sigma^2_{u\ell})}$ (99)
    end for
until convergence in $p(y; q)$
```

A.6 Algorithm 5: Gaussian Tensor Factorization

Algorithms 3 covers the case of two types of interactive latent factors. This addresses the case when the data can be arranged into a matrix and adds a matrix factorization to the linear predictor. When more than two types of interactive latent factors are present the data can be arranged into a tensor and a tensor decomposition can be added to the linear predictor(Kolda and Bader, 2009). The multilinear form presented in this article corresponds to a type of tensor decomposition called the CANDECOMP/PARAFAC (CP) tensor factorization (Kolda and Bader, 2009; Hoff, 2011*a*; Zhao, Zhang and Cichocki, 2014).

A.6.1 Notation

Because this model allows for tensors with an arbitrary number of modes it will be useful to change notation. Denote the tensor containing the outcome variable as $\mathcal{Y}$. The interactive latent factors will form a tensor of equivalent dimensions which collects their contribution to the linear predictor. Denote this latent tensor $\mathcal{U}$. For each mode of the tensor indexed by $m \in 1 \ldots M$ there exists a factor matrix $U^{(m)} \in \mathbb{R}^{(n_m \times K)}$ where n_m is the dimension of the m-th mode and K is the dimensionality of the latent factor. A column in this matrix is denoted $u_k^{(m)}$. The latent tensor $\mathcal{U}$ can be formed by taking the sum over the kroeneker product of the modes such that

$$\mathcal{U} = \sum_{k=1}^{K} u_k^{(1)} \otimes \ldots \otimes u_k^{(M)} \tag{102}$$

In general we will work with the collection of matrix representations but it will often be simpler to index terms form the latent tensor.

Observed covariates in the form of X, Z and C are left as capital letters with the understanding that when multiply indexed they still return a vector as in the standard regression case.

A.6.2 Model

Using the new notation we can state the model as

$$y_{i,j,\ldots,r} \sim \text{Normal}(X_{i,j,\ldots,r}\beta + Z_{i,j,\ldots,r}\gamma + \mathcal{U}_{i,j,\ldots,r}, \sigma_\epsilon^2) \tag{103}$$

$$u_k^{(m)} \overset{\text{ind.}}{\sim} \text{Normal}(0, \tau_{m,k}^2) \tag{104}$$

$$\beta \sim \text{Normal}(0, \sigma_\beta^2 I_P) \tag{105}$$

$$\gamma \sim \text{Normal}(0, \Sigma^R) \tag{106}$$

$$\Sigma^R | a_1^R, \ldots, a_{q^R}^R \sim \text{Inverse-Wishart}(\nu + q^R - 1, 2\nu \text{diag}(1/a_1^R, \ldots, 1/a_{q^R}^R) \tag{107}$$

$$a_1^R \ldots a_{q^R}^R \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(.5, 1/A_{Rr}^2) \tag{108}$$

$$\sigma_\epsilon^2 \sim \text{Inverse-Gamma}(.5, 1/a_\epsilon) \tag{109}$$

$$a_\epsilon \sim \text{Inverse-Gamma}(.4, 1/A_\epsilon^2) \tag{110}$$

where I emphasize that the model in Equation 104 uses a shared variance for a given mode and factor dimension thus making it an analogous to the matrix case. τ^2 is now a matrix which collects these variances for each mode (along the rows) and each dimension of the latent factor (along the columns).

A.6.3 Variational Approximation

Again we factorize the density over the modes and the

$$q\left(\beta, U^{(1)} \ldots U^{(M)}\right) \approx q(\beta) \prod_i q(U^{(1)}) \ldots q(U^{(M)}) \tag{111}$$

with induced factorization further factorizing the posterior over the rows of the factor matrices (Zhao, Zhang and Cichocki, 2014). Due to the conjugacy in the model these rows are again multivariate Gaussian distributions.

A.6.4 Optimal Densities

In algorithms 3 and 4 the latent factors for mode 1 were updated by a Bayesian linear regression using the mode 2 factors as covariates. Because of the expectations of the quadratic terms these forms also include the covariance of the variational posterior over the mode 2 factors. Thus to rewrite Equations 67 and 68 in the current notation. To do so we denote $u_i^{(m)}$ to be the i'th row of the factor matrix $U^{(m)}$

$$\Sigma_{q(U_i^{(1)})} \leftarrow \left(\begin{pmatrix} 1/\tau_{1,1}^2 & 0 & \dots & 0 \\ 0 & 1/\tau_{1,2}^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1/\tau_{1,k}^2 \end{pmatrix} + \sum_{i=1}^{n_2} \frac{\Sigma_{(q(U^{(2)}))_i} + \mu_{q(U_i^{(2)})} \mu_{q(U_i^{(2)})}^T}{\sigma_\epsilon^2} \right)^{-1} \tag{112}$$

$$\mu_{q(U_i^{(1)})} \leftarrow \Sigma_{(q(U_i^{(1)})} \left(\sum_{n \in \Omega} \left(\frac{(y_n - x_n \beta) \mu_{q(U_{i(n)}^{(2)})}}{\sigma_\epsilon^2} \right) \right) \tag{113}$$

Notice the numerator in Equation 112 contains the terms related to the second mode of the tensor. In the general M-mode tensor case this simply becomes an elementwise product. Denote the elementwise product of a collection of matrices as $\underset{(m)}{\odot}$ where m denotes the index we are taking the elementwise product over.

$$\Sigma_{q(U_i^{(1)})} \leftarrow \left(\begin{pmatrix} 1/\tau_{1,1}^2 & 0 & \dots & 0 \\ 0 & 1/\tau_{1,2}^2 & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1/\tau_{1,k}^2 \end{pmatrix} + \sum_{i=1}^{n_2} \frac{\underset{(m)}{\odot} \left(\Sigma_{(q(U^{(m)}))_i} + \mu_{q(U_i^{(m)})} \mu_{q(U_i^{(m)})}^T \right)}{\sigma_\epsilon^2} \right)^{-1} \tag{114}$$

$$\mu_{q(U_i^{(1)})} \leftarrow \Sigma_{q(U_i^{(1)})} \left(\sum_{n \in \Omega} \left(\frac{(y_n - x_n \beta) \mu_{q(U_{i(n)}^{(2)})}}{mu_{q(\sigma_\epsilon^2)}} \right) \right) \tag{115}$$

As with the matrix factorization model the noise parameter requires some care. Define $\zeta_{C\beta}$ as before and introduce the more general form $\zeta_{\mathcal{U}}$ as

$$\zeta_{\mathcal{U}} = \sum_{i \in \Omega} |\mathcal{Y}_i|^2 - 2\mathcal{Y}_i \mathcal{U}_i + \underset{(m)}{\odot} \left(\Sigma_{(q(U^{(m)}))_i} + \mu_{q(U_i^{(m)})} \mu_{q(U_i^{(m)})}^T \right) \tag{116}$$

This leads to the update

$$B_{q(\sigma_\epsilon^2)} \leftarrow \zeta_{\mathcal{U}} - 2 \left(\sum_{i \in \Omega} (\mathcal{Y}_i - \mathcal{U}_i) C_i \beta \right) + |C_i \beta|^2 + \zeta_C \beta \tag{117}$$

A.6.5 Algorithm 5

With these updates in hand we can present algorithm 5.

```
repeat
    $S \leftarrow 0$
    $s \leftarrow 0$
    $y_{\text{working}} \leftarrow \mathcal{Y}_\Omega - \mathcal{U}_\Omega$
    for $i = 1 \ldots m$ do
        Update $G_i$ (14), Update $H_i$ (15)
        Update $S$ (16), Update $s$ (17)
    end for
    Update $\Sigma_{q(\beta,\gamma^G)}$ (18, Update $\mu_{(q(\beta,\gamma^G)}$ using (19)
    for $i = 1 \ldots m$ do
        Update $\Sigma_{q(\gamma_i^R)}$ (20), Update $\mu_{q(\gamma_i^R)}$ (21)
    end for
    Update $y_{\text{working}} \leftarrow \mathcal{Y}_\Omega - \mathcal{C}_\Omega \beta$
    for $m = 1 \ldots M$ do
        for $i = 1 \ldots N_m$ do
            Update $\Sigma_{q(U_i^{(m)})}$ (114)
            Update $\mu_{q(U_i^{(m)})}$ (115)
            Update $\tau^2_{i,k}$ (71)
        end for
    end for
    $y_{\text{working}} \leftarrow y$
    Update $B_{q(\sigma^2_\epsilon)}$ (117)
    Update $\mu_{q(1/\sigma^2_\epsilon)}$ (23)
    Update $\mu_{q(1/a_\epsilon)}$ (24)
    for $r = 1, \ldots, q^R$ do
        Update $B_{q(a_r^R)}$ (25), Update $\mu_{q(1/a_r^R)}$ (26)
    end for
    Update $B_{q(\Sigma^R)}$ (27)
    Update $M_{q(\Sigma^R)^{-1}}$ (28)
    for $\ell = 1 \ldots L$ do
        Update $\mu_{q(1/a_{u\ell})}$ (29), Update $\mu_{q(1/\sigma^2_{u\ell})}$ (30)
    end for
until convergence in $p(y;q)$
```

A.6.6 Initialization

While Nakajima and Sugiyama (2011) provides a direct connection between matrix factorizations and variational bayes there is no such clean theoretical result for the tensor case. Indeed while matrix factorizations are often easy to compute by the singular value decomposition, low-rank tensor decompositions need not even exist and commonly used algorithms for finding them often do not have convergence guarantees (Kolda and Bader, 2009). This is an active area of research that I don't delve into extensively here (but see Hoff (2011*a*); Anandkumar, Ge and Janzamin (2014); Suzuki (2014)).

In order to initialize the model I perform the Nakajima et al. (2012) estimator on the matricization of each mode of the tensor. Matricization involves unfolding the tensor to create a matrix in which the rows represent one mode of the tensor and the columns represent all other modes (Kolda and Bader, 2009). The spectral estimator can be applied to each matricization and the latent factors for the preserved dimension are then used as initializations for the tensor factorization algorithm. This works well in practice although further study is needed.

A.6.7 Computation

Care must be taken to avoid memory issues with tensor latent variables. Often the tensors are very sparsely observed and thus the complete tensor should not be explicitly formed when at all possible.

A.7 Algorithm 6: Logistic Regression Tensor Factorization

Algorithm 6 provides the tensor variant of Algorithm 4. It provides no unique challenges beyond those in moving from Algorithm 3 to Algorithm 5.

A.7.1 Model

We work with the model

$$\mathcal{Y} \sim \text{Bernoulli}(\sigma(\eta)) \tag{118}$$

$$\eta = X\beta + Z\gamma + \mathcal{U} \tag{119}$$

$$u_k^{(m)} \overset{\text{ind.}}{\sim} \text{Normal}(0, \tau_{m,k}^2) \tag{120}$$

$$\beta \sim \text{Normal}(0, \sigma_\beta^2 I_P) \tag{121}$$

$$\gamma \sim \text{Normal}(0, \Sigma^R) \tag{122}$$

$$\Sigma^R | a_1^R, \dots, a_{q^R}^R \sim \text{Inverse-Wishart}(\nu + q^R - 1, 2\nu\text{diag}(1/a_1^R, \dots, 1/a_{q^R}^R) \tag{123}$$

$$a_1^R \dots a_{q^R}^R \overset{\text{ind.}}{\sim} \text{Inverse-Gamma}(.5, 1/A_{Rr}^2) \tag{124}$$

A.7.2 Optimal Densities

Again we denote the elementwise product of a collection of matrices as $\underset{(m)}{\odot}$ where m denotes the index we are taking the elementwise product over. Then the updates for the latent factor matrices are

$$\Sigma_{q(U_i^{(1)})} \leftarrow \left(\left(\begin{pmatrix} 1/\tau^2_{1,1} & 0 & \dots & 0 \\ 0 & 1/\tau^2_{1,2} & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 1/\tau^2_{1,k} \end{pmatrix} \right) + \sum_{i=1}^{n_2} \frac{\bigodot_{(m)} \left(\Sigma_{(q(U^{(m)}))_i} + \mu_{q(U_i^{(m)})} \mu^T_{q(U_i^{(m)})} \right)}{\lambda(\xi)} \right)^{-1} \tag{125}$$

$$\mu_{q(U_i^{(1)})} \leftarrow \Sigma_{q(U_i^{(1)})} \left(\sum_{n \in \Omega} \left(\frac{(y_n - x_n \beta) \mu_{q(U^{(2)}_{i(n)})}}{\lambda(\xi)} \right) \right) \tag{126}$$

where y is replaced with y_{working}. I've omitted the "working" subscript here to avoid confusion with the observation index.

The update for the the variational parameter ξ becomes

$$\xi^2 \leftarrow \text{diagonal} E[(\mathcal{C}\beta + \mathcal{U})(\mathcal{C}\beta + \mathcal{U})^T] \tag{127}$$

A.7.3 Algorithm 6

With the above updates we can specify the algorithm as

```
repeat
    $S \leftarrow 0$
    $s \leftarrow 0$
    Update $y_{\text{working}} \leftarrow y/2 - 2y^2\mathcal{U}\lambda(\xi)$
    for $i = 1 \dots m$ do
        Update $G_i$ (83), Update $H_i$ (84)
        Update $S$ (85), Update $s$ (86)
    end for
    Update $\Sigma_{q(\beta,\gamma^G;\xi)}$ (87, Update $\mu_{(q(\beta,\gamma^G;\xi)}$ using (88)
    for $i = 1 \dots m$ do
        Update $\Sigma_{q(\gamma_i^R;\xi)}$ (89), Update $\mu_{q(\gamma_i^R;\xi)}$ (90)
    end for
    Update $y_{\text{working}} \leftarrow y/2 - 2y^2(\mathcal{C}\beta)\lambda(\xi)$
    for $m = 1 \dots M$ do
        for $i = 1 \dots N_m$ do
            Update $\Sigma_{q(U_i^{(m)})}$ (125)
            Update $\mu_{q(U_i^{(m)})}$ (126)
            Update $\tau^2_{i,k}$ (71)
        end for
    end for
    $y_{\text{working}} \leftarrow y$
    Update $\xi^2$ (127)
    for $r = 1, \dots, q^R$ do
```

```
        Update $B_{q(a_r^R)}$ (94), Update $\mu_{q(1/a_r^R)}$ (95)
    end for
    Update $B_{q(\Sigma^R)}$ (96)
    Update $M_{q(\Sigma^R)^{-1}}$ (97)
    for $\ell = 1 \ldots L$ do
        Update $\mu_{q(1/a_{u\ell})}$ (98), Update $\mu_{q(1/\sigma^2_{u\ell})}$ (99)
    end for
until convergence in $p(y; q)$
```

A.7.4 Computational Notes

As with the Gaussian tensor factorization case the complete tensor may be very sparsely observed. However in the binary case it may also be that instances of a $y = 1$ are very rare. This will often occur with various types of longitudinal network structure. In this setting substantial computational gains can be realized by using a case control likelihood as suggested in (Raftery et al., 2012).

B Alternative Approaches

In this appendix I briefly overview several approaches to modeling heterogeneity that are not encompassed by the framework presented here. In each case I have highlighted the contrast with the framework that I have provided in the main text.

Exponential Random Graph Models (ERGMS) ERGMs provide an alternative approach to modeling networks. Here we forego the conditional independence assumption and instead model the entire graph as a single draw from a joint (Gibbs) distribution (Cranmer and Desmarais, 2011). ERGMs are notoriously difficult to estimate and require careful specification of the sufficient statistics of the graph. They also only apply to graphs with unweighted edges. However, in settings where the conditional independence is untenable they are effectively the only option.

Survival Analysis Semiparametric approaches to survival analysis, such as the Cox model are essentially single mode models with varying intercepts (where time is the mode) (Box-Steffensmeier and Jones, 2004). The spline based method of Beck, Katz and Tucker (1998) for analyzing binary duration data can be placed into the GMRF framework. see also Jackman (1998) for relevant connections.

Binary Treatment Causal Inference Recent work on causal inference in time-series and time-series cross sectional data addresses related problems in the potential outcomes framework (Blackwell, 2013; Blackwell and Glynn, 2013; Xu, 2014). Additional work provides the infrastructure for causal inference under interference between units or with contagion (Bowers, Fredrickson and Panagopoulos, 2013; Ogburn and VanderWeele, 2014). Imai and Kim (2012) illuminate the connection between difference-in-differences design and two-way weighted fixed effects estimators.

Mixture Models In the models described in this paper, the groups within a mode are assumed to be observed. When even this information is unavailable mixture models can be used to model the heterogeneity. Kyung, Gill and Casella (2010, 2011) use a Dirichlet process random-effects model to allow varying intercepts over unknown groupings in the data. Park (2012) provides a model for time-series cross-sectional data where which uses a non-recurrent Hidden Markov Model for parameters. This treats time as a mode where all members of a group are temporally contiguous but the breakpoints between groups are estimated by the model. Imai and Tingley (2012) use a finite mixture of GLMs model which allow each data point to be drawn from an *a priori* fixed number of possible regression models. While presented in the context of theory testing, we can motivate the same infrastructure as a way to model a single model with unknown group membership with each group represented by one of the regressions. And extended to the infinite mixture model is provided by the Dirichlet Process-GLM framework (Hannah, Blei and Powell, 2011). Finally recent econometric work has focused on learning unknown group membership in the interactive fixed effects framework (Ando and Bai, 2013).

Flexible Regressions A separate approach to modeling heterogeneity avoids explicit models for the heterogeneity and simply uses an extremely flexible regression (Wahba, 1990; Gu, 2013). By allowing effects to be highly non-linear and context dependent we can side step the issue of explicitly producing models of the models. In the international relations context this was Beck, King and Zeng (2000) argue for the use of neural networks for estimating these types of functions. Hainmueller and Hazlett (2014) present a more interpretable approach based on Kernel Regularized Least Squares. Both methods face the challenge of interpreting the resulting models which is complicated by the non-linear forms.

Nonparametric Estimation For the particular case of network data there has been increasing interest in nonparametric approaches to estimation.[11] Chatterjee (2012) provides a consistent estimator based on a truncated singular value decomposition that applies to a wide range of models including many of the two mode network models. Airoldi, Costa and Chan (2013); Chan and Airoldi (2014); Yang, Han and Airoldi (2014) design a consistent estimator that uses a histogram of stochastic block models. These approaches are extremely new but they point to approaches to nonparametric estimation of particular versions of the model. The statistical characterization of graphons also provides insights to the extent to which the latent effects models are identifiable (Bickel and Chen, 2009; Bickel, Chen and Levina, 2011).

[11] This literature makes use of the a statistical function called a graphon. Graphons are the limiting objects that describe random network objects. Define a graph with N nodes which are each given a latent variable $u_i \sim$Uniform$(0, 1)$ for all $i \in \{1 \dots N\}$. Two vertices are connected with probability $w(u_i, u_j)$ where the function w is the graphon. That is, it is a function which maps $[0, 1]^2 \to [0, 1]$. This describes a wide class of models over exchangeable graphs (Lloyd et al., 2013; Orbanz and Roy, 2013). Interest is in the circumstances in which we can non-parametrically estimate this function and the circumstances in which we can prove consistency of the resulting estimator (Airoldi, Costa and Chan, 2013).

C Two-Way Fixed Effects and Latent Factor Regression

The goal of this appendix is to illuminate the connection between the Latent Factor Regression framework and two-way fixed effects. In doing so I also address the three related questions: 'where is the variation coming from?', 'what is the counterfactual?' and 'what's the analogous experiment?' In order to connect to the literature on causal inference I focus on settings where the effect of interest is a binary treatment with other covariates serving as continuous pre-treatment confounders. However, I emphasize that the framework naturally extends to non-binary effects. I also only consider varying intercepts although the framework naturally extends to varying coefficients.

C.1 Two-way Fixed Effects

For concreteness and without loss of generality, I consider the case of data organized in a time-series cross-sectional format with $t \in \{1 \dots T\}$ years and $c \in \{1 \dots C\}$ countries. The two-way fixed effects estimator considers a model of the form:

$$y_{c,t} = x_{c,t}\beta + \alpha_c + \gamma_t + \epsilon \tag{128}$$

where α and γ are vectors of country and time specific intercepts respectively, $x_{c,t}$ are the covariates for unit c at time t and ϵ is random Gaussian noise.

Arranging the outcome y into a C by T matrix it is clear that the fixed effects α and γ are estimated using the rows and columns of the matrix respectively. The unmeasured confounding in the data is then controlled through the additive combination of the row and column effect. What the two-way fixed effects setup implicitly assumes is that the effect of country c is constant across all time periods and analogously the effect of time period t is the same across all countries. We can give this an economic interpretation by saying that there are global shocks to the system represented by γ_t which affect all countries in the same way, and different base levels of the outcome for each country measured by α_c.

C.2 Latent Factor Regressions

The analogous latent factor regression model is:

$$y_{c,t} = x_{c,t}\beta + \alpha_c + \gamma_t + \sum_{k}^{K} u_{c,k}v_{t,k} + \epsilon \tag{129}$$

where $u_{c,k}$ is an element of factor matrix U having dimension $C \times K$ (with V analogously defined). The rank of the approximation, K, can be estimated or fixed by the analyst.

The two-fixed effects estimator is a special case of the latent factor regression where $k = 0$ and thus is directly nested within the framework proposed in the main paper. The latent factor matrices U and V represent deviations from the additive form of the model and thus under the interpretation above can be seen as capturing the degree to which common shocks to the system elicit varying responses from countries.

It is helpful to define a third model which serves as an illustrative limiting case which I call the *joint* fixed effects model. This formulation can be given as,

$$y_{c,t} = x_{c,t}\beta + \phi_{c,t} + \epsilon \tag{130}$$

Φ is a matrix where each cell $\phi_{c,t}$ represents a fixed effect for that country-year combination. Clearly the model in Equation 130 is only identified if the data contain more than one observation for each pair of indices. Whereas the two-way fixed effects estimator of Equation 128 uses the entire row and entire column to estimate the intercepts for an observation, the joint fixed effects estimator uses only the information within the cell.

The latent factor regression model provides an intermediate point between the two fixed effects models. The degree of the tradeoff is controlled by the rank of the latent factor model, K. When $K = 0$ the model recovers exactly the two-fixed effects model. When $K = \min(C, T)$ the model recovers exactly the joint fixed effects model.[12] The solution (α, γ, u, v) for a fixed value K is the best rank K approximation to the joint parameter matrix Φ.[13]

Thus, even in cases where we have replications at the cell (e.g. country/year) level the latent factor approach is an attractive alternative to the joint effects model because if we believe there is *any* structure in the matrix of parameters Φ (e.g. the effect for country c at time t has any information to offer us on the effect for country c at time $t-1$) then the latent factor regression framework provides us with a favorable bias-variance tradeoff. This is because structure of the matrix allows the model to bring additional observations beyond the replications at the cell level.[14]

Having addressed the basic infrastructure, I now consider direct answers to a few common questions.

C.3 Connections to Causal Inference

A natural question for any new method is 'what is this doing?' That is, we want to be able to articulate which treated units are being compared to which control units. This is closely related to the question of what variation is being removed from the data by the procedure. Before moving into the latent factor regression framework I briefly explain what information is being removed in the two-way fixed effects framework as well as the joint fixed effects.

Again it is helpful to arrange the data Y into a C by T matrix. In two-way fixed effects we are removing from each cell variation which is common to the row and variation which is common to the column. The variation removed is a function of three averages:

[12] The proof of this follows immediately from the singular value decomposition which guarantees the existence of the decomposition as well as the exact reconstruction of the full matrix (Eckart and Young, 1936).

[13] This is the best approximation in terms of the Frobenius norm of the matrix as again guaranteed by appeal to the singular value decomposition. When estimating the rank using the ARD priors as used in the paper this corresponds exactly to the best approximation of the joint parameter matrix under the nuclear norm, a common convex relaxation of the rank selection problem (Fithian and Mazumder, 2013).

[14] The bias is determined by the degree to which Φ cannot be captured by the low-rank structure and the variance improvement comes as a result of using additional observations in estimating $\phi_{c,t}$ as $\alpha_c + \gamma_t + \sum_k u_{c,k}v_{c,k}$. Clearly for a non-random matrix Φ the bias decreases in the rank of the approximation.

the average of all observations in the same row (country), the average of all observations in the same column (time) and the average of all remaining observations (Imai and Kim, 2012).[15]

In the joint fixed effects estimator we consider only a single cell of the matrix when removing variation. Again if we observe only a single observation per country-time combination this is unidentified, having a single free parameter for each observation. In cases where we have replication at this level, such as in the 'Dirty Pool' example (Green, Kim and Yoon, 2001), we are simply subtracting off the mean of all observations within that cell. They key difference here is that the two-way fixed effects estimator uses data from every country and every time period in constructing each counterfactual, whereas the joint estimator uses only the country and year for that cell.

In the latent factor regression, we consider the entire row and column of the matrix (as in the two-fixed effects) but each observation is not weighted equally. Instead the model implicitly assigns higher weight to countries (or years) which have trends in the dependent variable which are similar to the country (year) of the cell. I make this comparison more precise below by first considering which units are stochastically equivalent under the model and then by giving an augmented data interpretation of the estimator.

C.3.1 Stochastic equivalence

An intuitive way to think of how the latent factor framework models dependence is by considering the components of the inner product term U and V as forming a K dimensional vector space in which both countries and years are projected. Countries which have similar projections u_c in the space are approximately stochastically equivalent and respond to shocks in a similar fashion. This provides us with insight into where the variation comes from. We are implicitly comparing a country-time unit with a country-time unit having a similar projection into the K dimensional space. This produces a continuous weighting over units in computing the counterfactuals in a potential outcomes framework. In the matching framework, this continuous weighting can be seen as analogous to synthetic control methods which matches treated units to a reweighted collection of controls (Abadie, Diamond and Hainmueller, 2010).[16]

Note that the appeal here to the approximate stochastic equivalence of two units sharing similar values of the continuous factors is in principle no different from approximate stochastic equivalence of two units with similar pre-treatment covariates. That is, when we control for a series of (non-categorical) observed covariates we are invoking an assumption of approximate stochastic equivalence of two units with similar covariate profiles.

A further advantage of the low-rank framework is that two countries can be similar along one dimension but different along another. Put another way, there can be a type of global shock for which two countries can have a similar response, but a different type of

[15] The third term is necessary to adjust for the fact that we are adjusting the data based on two margins rather than one. The full form of the two-way fixed effects as an adjusted matching estimator is given with proof as Proposition 4 of Imai and Kim (2012).

[16] The nature of this correspondence is developed in Xu (2014) with reference to the interactive fixed effects formulation of (Bai, 2009).

global shock for which their responses will be different. Mathematically this is represented by $u_{i,k} \approx u_{j,k}$ for $k = i$ but not for $k \neq i$.[17] This provides a substantially more flexible framework for modeling heterogeneous units than a model assuming units must be alike on all dimensions.

C.3.2 Interpretations as OLS on augmented data

Part of the reason simple additive fixed effects are so easy to understand is that we can give a representation of the model as a simpler procedure on an augmented dataset. The "least squares dummy variable" method can be written by using OLS on a transformed datasets formed by subtracting off the mean by group.

Using the results in Bai (2009), we can give a a similar interpretation for the latent factor regression as OLS on augmented data. Here we consider the special case of latent factor regression where the latent factors are given no priors which is simply the limiting special case of uninformative priors.

Start by defining the projection matrices:

$$M_V = I_T - V'V/T \tag{131}$$

$$M_U = I_C - U'U'/C \tag{132}$$

then writing the model as:

$$Y = \beta_1 X^1 + \cdots + \beta_p X^p + UV' + \epsilon \tag{133}$$

where we have simply absorbed the additive fixed effects into the factor matrices.

Then the left multiplying by M_V and right multiplying by M_U we get

$$M_V Y M_U = \beta_1 (M_V X^1 M_U) + \cdots + \beta_p (M_V X^p M_U) + M_V \epsilon M_U \tag{134}$$

This leads to the following least squares estimator under a given factor structure,

$$\hat{\beta} = \begin{bmatrix} \mathrm{tr}[M_U X^{1'} M_V X^1] & \dots & \mathrm{tr}[M_U X^{1'} M_V X^p] \\ \vdots & \vdots & \vdots \\ \mathrm{tr}[M_U X^{p'} M_V X^1] & \dots & \mathrm{tr}[M_U X^{p'} M_V X^p] \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{tr}[M_U X^{1'} M_V Y] \\ \vdots \\ \mathrm{tr}[M_U X^{p'} M_V Y] \end{bmatrix} \tag{135}$$

Thus the projection matrices M_U and M_V play a role analogous to projection matrices in least square dummy variables estimation.

Bai (2009) also gives an instrumental variables interpretation that holds in this setting as well. Define

$$\sum_c^C Z_c' Z_c = \begin{bmatrix} \mathrm{tr}[M_U X^{1'} M_V X^1] & \dots & \mathrm{tr}[M_U X^{1'} M_V X^p] \\ \vdots & \vdots & \vdots \\ \mathrm{tr}[M_U X^{p'} M_V X^1] & \dots & \mathrm{tr}[M_U X^{p'} M_V X^p] \end{bmatrix} \tag{136}$$

The form of the estimator for beta is the IV estimator with Z_c' as the instrument.

[17] For example, imagine a model rank K=2. Let us suppose that the two factors capture distinct economic shocks such as a sharp increase in energy prices (k=1) and a new technological development in a high tech sector (k=2). Two countries may respond similarly to changes in energy prices and thus $u_{i,1} \approx u_{j,1}$ but respond very differently to technological developments $u_{i,2} \neq u_{j,2}$. More generally when the number of factors is relatively high this flexibility of the model allows every country to be distinct while sharing qualities that substantially overlap.

C.4 Connections to Existing Work

Here I briefly connect this work to several existing articles in order to further illuminate the features of the model.

C.4.1 Grouped Fixed Effects

One way to deal with the problem of not observing multiple observations for each cell in our data matrix Y is to aggregate over the rows (countries) to create groups. We can then consider group-time specific fixed effects. These groups could be determined*a priori* or estimated as in Bonhomme and Manresa (2012). If the group membership is estimated under fixed K then the problem becomes a latent class (or finite mixture) model.

The latent factor regression framework and the mixture model framework differ in the parametric assumption on the latent variable. For the mixture model case the latent variable is categorical whereas in the latent factor regression it is continuous. The grouped fixed effects model has the distinct advantage that the latent variable is naturally interpretable as a partition over units. However, assuming that the latent variable is categorical is substantially less general than the latent factor framework as members of the same group are assumed to be exactly stochastically equivalent. This is a stronger assumption which is unnecessary under the latent factor regression model.

C.4.2 Correlated Error Models

One way to view the latent factor regression framework is as inducing a low rank decomposition of the error structure. Writing the regression model in matrix notation:

$$y = X\beta + \epsilon \tag{137}$$

$$\epsilon \sim \mathcal{N}(0, \Sigma) \tag{138}$$

the standard regression model assumes that $\Sigma = \sigma^2 I$. If instead we treat Σ as unstructured we essentially observe a single draw from a multivariate Gaussian. Unfortunately MLE is known to perform poorly for covariance estimation in this setting (James and Stein, 1961).

Numerous proposals have been made for estimating Σ under some particularly assumptions, e.g. time-series models often assume that Σ^{-1} is tri-diagonal (West and Harrison, 1997). A particularly general case is given by spatial error models which assume that the variance is rescaled by a known weights matrix, W. However the appropriate form of W is often not known and reasonable choices can produce different results (Zhukov and Stewart, 2013). The latent factor regression provides a quite general parametric form for Σ.[18]

This interpretation of the model also makes clearer where the two way random effects model will be insufficient. The two-way fixed effects model in Equation 128 is unable correlation in the second moments of the data but not the third order moments. This notion has the clearest expression in network data where properties of third order dependence have been well characterized (Hoff, 2005; Wasserman and Faust, 1994). These intuitions extend reasonably well to the spatial and cross-sectional case, where intuitively we can

[18] For more comparisons to various spatial models with Monte Carlo experiments see Pang (2014).

think of third order dependence as capturing the consistency of the pairings A-B, B-C, A-C where A, B, C are nodes in a networks or locations in a space.

This covariance structure can also be given a probabilistic interpretation akin to the linear regressions product of univariate normals. Under a two-mode model the latent factor regression provides a parameterization of a matrix normal distribution (Dawid, 1981; Hoff, 2005; Allen and Tibshirani, 2012). Under a general m-mode we obtain an array normal distribution under separable covariance structure(Hoff, 2011*b*). The key assumption in these settings is a weak row-column exchangeability which is substantially more general than the exchangeability assumption typically invoked (Hoff, 2005; Orbanz and Roy, 2013).

C.5 Concluding Thoughts

In this appendix I've tried to illuminate what the latent factor regression framework is doing for particular cases. I've primarily discussed the two mode framework but a key advantage of the model is the ability to extend easily to an arbitrary number of modes.

D Improving Accuracy of the Variational Framework

In this appendix I provide a brief summary of relevant results on accuracy of posterior inference in the variational framework including approaches to improve accuracy.

The variational algorithms presented in the main text produce quite faithful approximations to the true posterior. The results are particularly strong in the case of the normal likelihood (which is conjugate) and in the single mode case (where we do not have to make strong factorization assumptions). Theoretical results in Ormerod and Wand (2012) and previous empirical results reported in the literature support this observation (Wand, 2014; Lee and Wand, 2014; Tan and Nott, 2013).

In the next two sections I step through three settings where accuracy can be improved and sketch some methods for trading off computational time for improved quality of approximation. None of the methods described below have been implemented in the results reported above. I have included the discussion of these approaches here in order to demonstrate that improvements to the quality of the approximation are possible within the variational framework. However, in the main text I chose to maintain the simplest version of the available methods that also produced accurate results.

D.1 Non-Gaussian Likelihoods

The posterior approximation used here for the logistic regression case uses the Jaakkola and Jordan (2000) bound on the sigmoid function. This is a quadratic bound which is tight but only at the value of the variational parameters ξ. Previous empirical studies suggest that the bound produces a small bias towards zero in the random effects which disappears as the cluster size grows (Ormerod and Wand, 2012; Tan and Nott, 2013; Scott and Sun, 2013). This accords with the findings in the simulations i ran for this paper.

Numerous alternative strategies have been proposed but here I highlight two in particular: the Non-conjugate Variational Message Passing scheme of Knowles and Minka (2011), and piecewise bounds (Marlin, Khan and Murphy, 2011).

Knowles and Minka (2011) generalize the variational message passing scheme (Bishop, 2006) to handle non-conjugate factors by approximating them using an approximating distribution in the same family as the prior. For the case of a multivariate Gaussian approximation Wand (2014) provides a simplified update structure which enables efficient computation. These updates were used in Tan and Nott (2013) to derive variational algorithms for GLMMs which show excellent results.

For the logistic regression case, this scheme involves the calculation of an analytically intractable expectation. However, Ormerod and Wand (2012) show that it can be reduced to a uni-dimensional problem and evaluated accurately using Adaptive Gauss-Hermite quadrature. This results in a slightly slower algorithm but yields more faithful representations of the posterior.

For the case of count models the required expectation can be evaluated in closed form resulting in little tradeoff in speed (Luts and Wand, 2013; Wand, 2014).

An alternative strategy is similar to the bounding approach of Jaakkola and Jordan (2000). Rather than using a quadratic bound that is tight only at a single point, Marlin, Khan and Murphy (2011) advocate the use of piecewise bounds. By increasing the number of pieces, the bound on the nonconjugate term can be made arbitrarily tight resulting in more accurate inference. This has the advantage of allowing variational inference to take on a bit of the quality of MCMC where a continuous increase in computational cost results in a continuous increase in accuracy. Experimental evaluations are given in (Khan et al., 2010).

A particularly compelling application of the bounding approach is the extension to the multinomial outcome case. The Jaakkola and Jordan (2000) does not extend to this setting nor does the use of quadrature. Khan et al. (2012) develop a stick breaking likelihood suitable for a categorical outcome for which efficient bounds could be constructed. This would allow for the implementation of an approximate analogue of a multinomial logistic regression that admits a tight bound.

D.2 Factorization

The key assumption in variational inference is the product density factorization of the joint posterior. Stronger factorizations make the model more tractable but also less accurate. In the single mode case the key assumed factorization is between the regression coefficients and their variance parameters. In the case of unordered groups this assumption is relatively mild but for Gaussian Markov Random Fields it is somewhat stronger.

Luckily we can always make our approximations arbitrarily more accurate by using conditional approximations. Whereas standard variational bayes might approximate $q(x, \theta|y) = q(x|y)q(\theta|y)$ for a latent field x and hyperparameters θ we instead do $q(x, \theta|y) = q(x|y, \theta)q(\theta|y)$ this is tractable for hyperparameters of low dimension like the variances in a hierarchical model. This is the essential insight of Rue, Martino and Chopin (2009) and is formalized in the VB context by Han, Liao and Carin (2013). It is also raised Salimans and Knowles (2013) who describe it as a hierarchical approximation.

A milder form of this strategy is considered by Tan and Nott (2013) in their use of partial non-centering of GLMMs. Here the partial non-centering parameters give the model extra flexibility to handle the assumed factorization.

An open question is whether there is a straightforward equivalent of this to the multilinear case. If there were it would be tremendously useful across a wide variety of models but it seems unlikely as there is no single low-dimensional parameter to be conditioned on. A reasonable strategy might be to instead us a mixture of distributions to capture the joint approximation. A setting using a mixture of multivariate normals is considered by Gershman, Hoffman and Blei (2012). By combining this nonparametric approach with the gridding strategies in Rue, Martino and Chopin (2009) it may be possible to construct a tighter approximation to the factorized terms.

E Simulation

In this section I provide the necessary details to replicate the simulation results.

E.1 Single Mode Setting

In the first simulation of the single mode setting, I demonstrate the ability of the variational algorithm to recover parameters in hierarchical linear regression model with a Gaussian outcome. After reporting details for the Gaussian outcome I describe estimation for the hierarchical logistic regression model.

Gaussian Hierarchical Regression I use the data generating process from the help file of `MCMChregress` in `MCMCpack` version 1.3-3 (Martin, Quinn and Park, 2011). Simulations were run in R version 3.1.1 on a 3.2Ghz quadcore processor with 7GB RAM. The code including the model estimation is given below.

```
nobs <- 1000
nspecies <- 20
species <- c(1:nspecies,sample(c(1:nspecies),(nobs-nspecies),replace=TRUE))

 # Covariates
 X1 <- runif(n=nobs,min=0,max=10)
 X2 <- runif(n=nobs,min=0,max=10)
 X <- cbind(rep(1,nobs),X1,X2)
 W <- X

 # Target parameters
 # beta
 beta.target <- matrix(c(0.1,0.3,0.2),ncol=1)
 # Vb
 Vb.target <- c(0.5,0.2,0.1)
 # b
 b.target <- cbind(rnorm(nspecies,mean=0,sd=sqrt(Vb.target[1])),
                   rnorm(nspecies,mean=0,sd=sqrt(Vb.target[2])),
                   rnorm(nspecies,mean=0,sd=sqrt(Vb.target[3])))
 # sigma2
 sigma2.target <- 0.02

 # Response
 Y <- vector()
 for (n in 1:nobs) {
   Y[n] <- rnorm(n=1,
                 mean=X[n,]%*%beta.target+W[n,]%*%b.target[species[n],],
                 sd=sqrt(sigma2.target))
 }
```

```
  # Data-set
  Data <- as.data.frame(cbind(Y,X1,X2,species))

  model <- MCMChregress(fixed=Y~X1+X2, random=~X1+X2, group="species",
                      data=Data, burnin=1000, mcmc=10000, thin=10,verbose=1,
                      seed=NA, beta.start=0, sigma2.start=1,
                      Vb.start=1, mubeta=0, Vbeta=1.0E6,
                      r=3, R=diag(c(1,0.1,0.1)), nu=0.001, delta=0.001)
```

The results for the Gaussian case are given in the main text.

Logistic Regression Here I present results for a hierarchical logistic regression which parallels the Gaussian outcome model. Again I use the data generating process given in MCMCpack with the exception that I extend the burnin and number of posterior samples to match the normal regression case. The full code is:

```
# Constants
nobs <- 1000
nspecies <- 20

simresults <- vector(mode="list", length=100)
for(s in 1:100) {
  # Covariates
  species <- c(1:nspecies,sample(c(1:nspecies),(nobs-nspecies),replace=TRUE))
  X1 <- runif(n=nobs,min=-10,max=10)
  X2 <- runif(n=nobs,min=-10,max=10)
  X <- cbind(rep(1,nobs),X1,X2)
  W <- X

  # Target parameters
  # beta
  beta.target <- matrix(c(0.3,0.2,0.1),ncol=1)
  # Vb
  Vb.target <- c(0.5,0.05,0.05)
  # b
  b.target <- cbind(rnorm(nspecies,mean=0,sd=sqrt(Vb.target[1])),
                    rnorm(nspecies,mean=0,sd=sqrt(Vb.target[2])),
                    rnorm(nspecies,mean=0,sd=sqrt(Vb.target[3])))

  # Response
  theta <- vector()
  Y <- vector()
  for (n in 1:nobs) {
    theta[n] <- inv.logit(X[n,]%*%beta.target+W[n,]%*%b.target[species[n],])
    Y[n] <- rbinom(n=1,size=1,prob=theta[n])
  }

  # Data-set
  Data <- as.data.frame(cbind(Y,theta,X1,X2,species))
  plot(Data$X1,Data$theta)

  #== Call to MCMChlogit
  model <- MCMChlogit(fixed=Y~X1+X2, random=~X1+X2, group="species",
                    data=Data, burnin=1000, mcmc=10000, thin=10,verbose=1,
                    seed=NA, beta.start=0, sigma2.start=1,
                    Vb.start=1, mubeta=0, Vbeta=1.0E6,
                    r=3, R=diag(c(1,0.1,0.1)), nu=0.001, delta=0.001, FixOD=1)
```

Run times are slightly more variable for the logistic regression case and are plotted in Figure 1. MCMC runs for 41 seconds on average with variational running for 2 seconds.

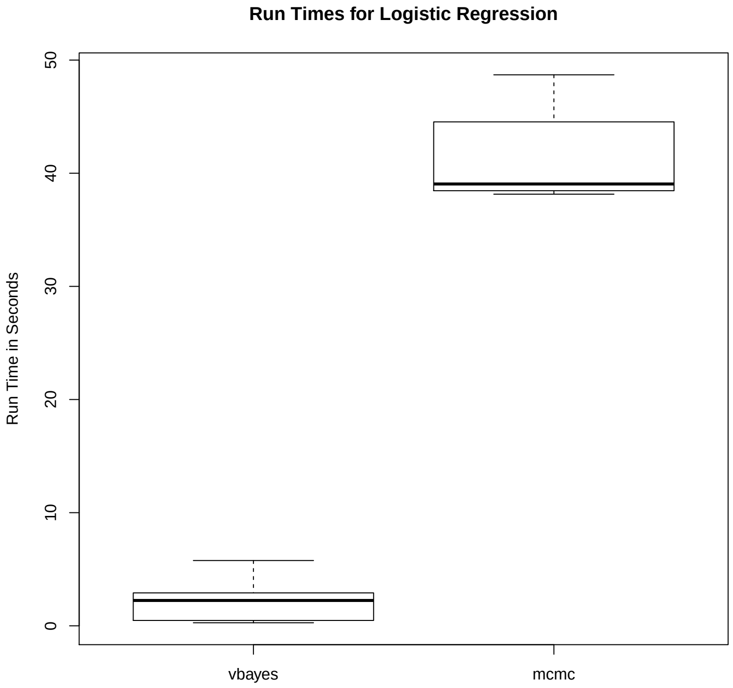


Figure 1: Distribution of run times for 100 simulations of the single mode logistic regression case.

E.2 Two Mode Setting

In this simulation I demonstrate the ability of the variational algorithm to recover simulated parameters in the Gaussian outcome case with two modes and interactive latent factors. In order to match the simulation to the FDI application, I use the observed covariates from Büthe and Milner (2008). Each simulation then follows the following procedure where `nc` is the number of countries in `cindex` and `nt` is the number of time periods in `tindex`

```
k <- rpois(1,lambda=3) + 1
factor1 <- matrix(rnorm(k*nc), nrow=nc,ncol=k)
factor2 <- matrix(rnorm(k*nt), nrow=nt,ncol=k)
uv <-rowSums(factor1[cindex,,drop=FALSE]*
     factor2[tindex,,drop=FALSE])
cint <- rnorm(nc)
tint <- rnorm(nt)
beta <- matrix(rnorm(9),ncol=1)
y <- c(X%*%beta) + cint[cindex] + tint[tindex] +
      uv + rnorm(length(y))
y <- c(y)
```

All parameters are simulated from Normal$(0, 1)$.

E.3 Model Misspecification

Here I provide some additional results for the models presented where covariates in the true data generating process are omitted from the estimated models. In the paper I gave two extreme examples: no covariates omitted and all but one covariate omitted. Here I provide the cases between. In each case a random selection of covariates was dropped whereas in the two extreme examples the observed covariates are the same across all simulations.

Recall that I compare four alternative estimation strategies in addition to the latent factor regressions:

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

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

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

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

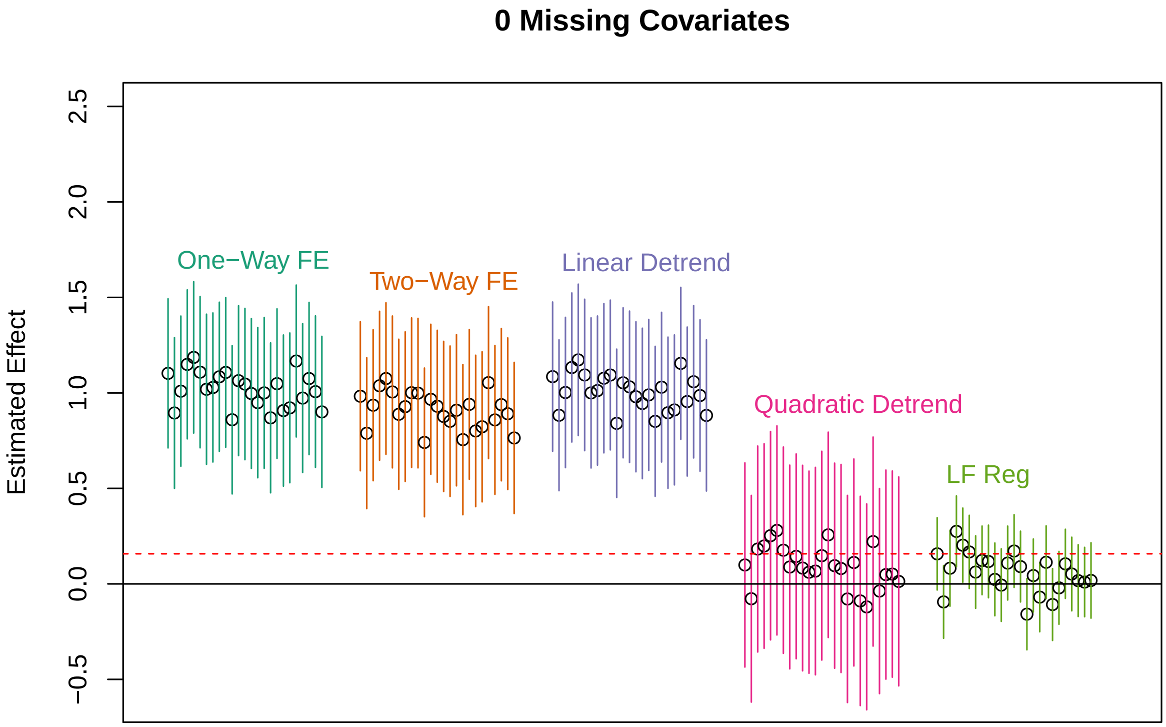


Figure 2: Reprinted from the main paper, here the estimating models use all of the covariates in the true DGP. 25 simulations from a two-mode model with full observed covariates. Each of the five estimation strategies is shown with 95% confidence/credible intervals. The red dashed line indicated the true effect to be recovered.

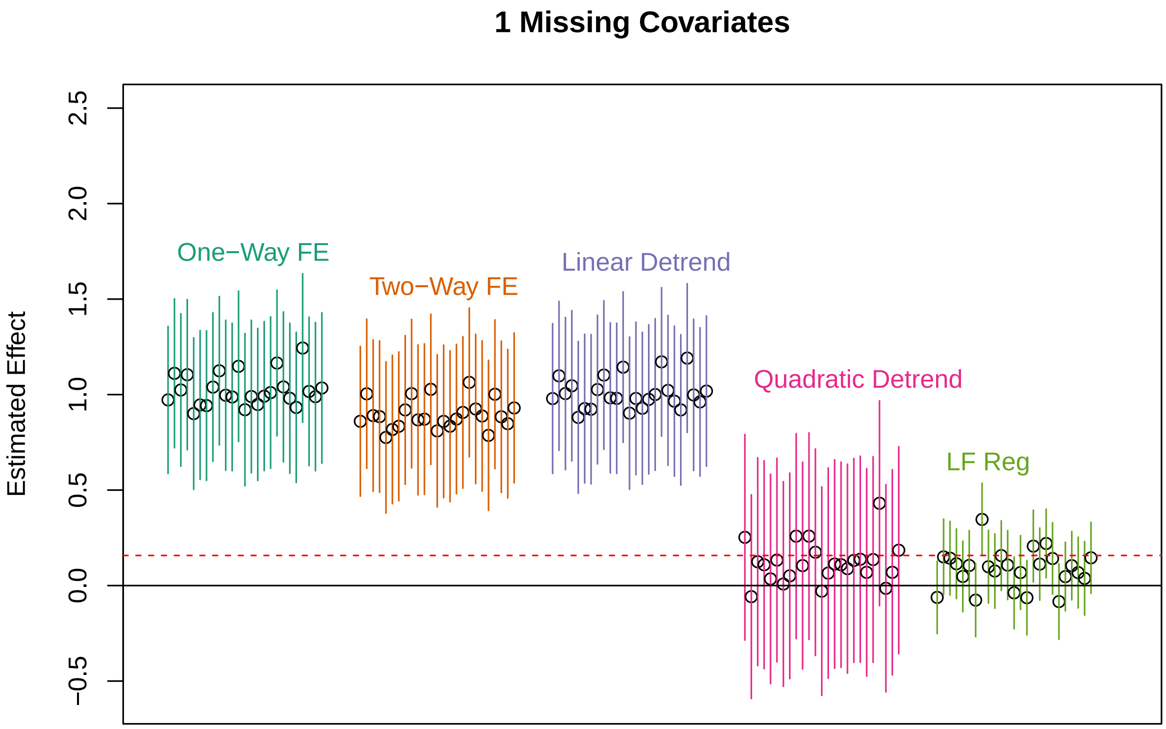


Figure 3: One random covariate is dropped in each simulation.

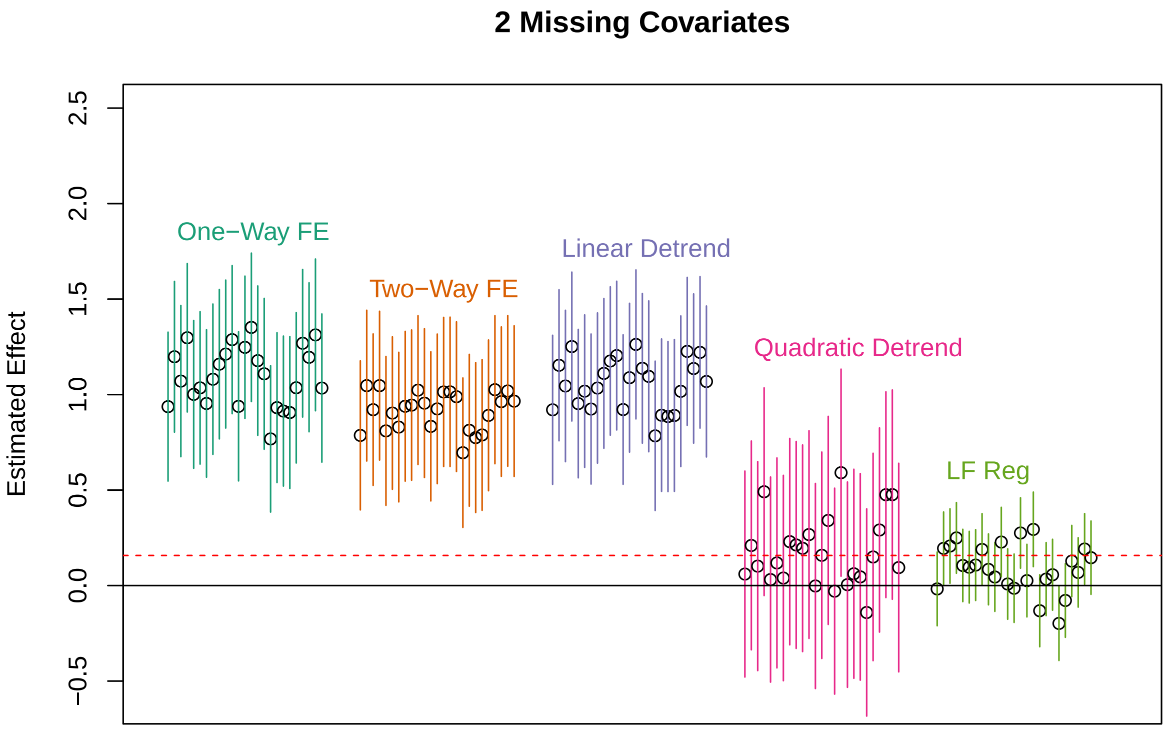


Figure 4: Two random covariates are dropped in each simulation.

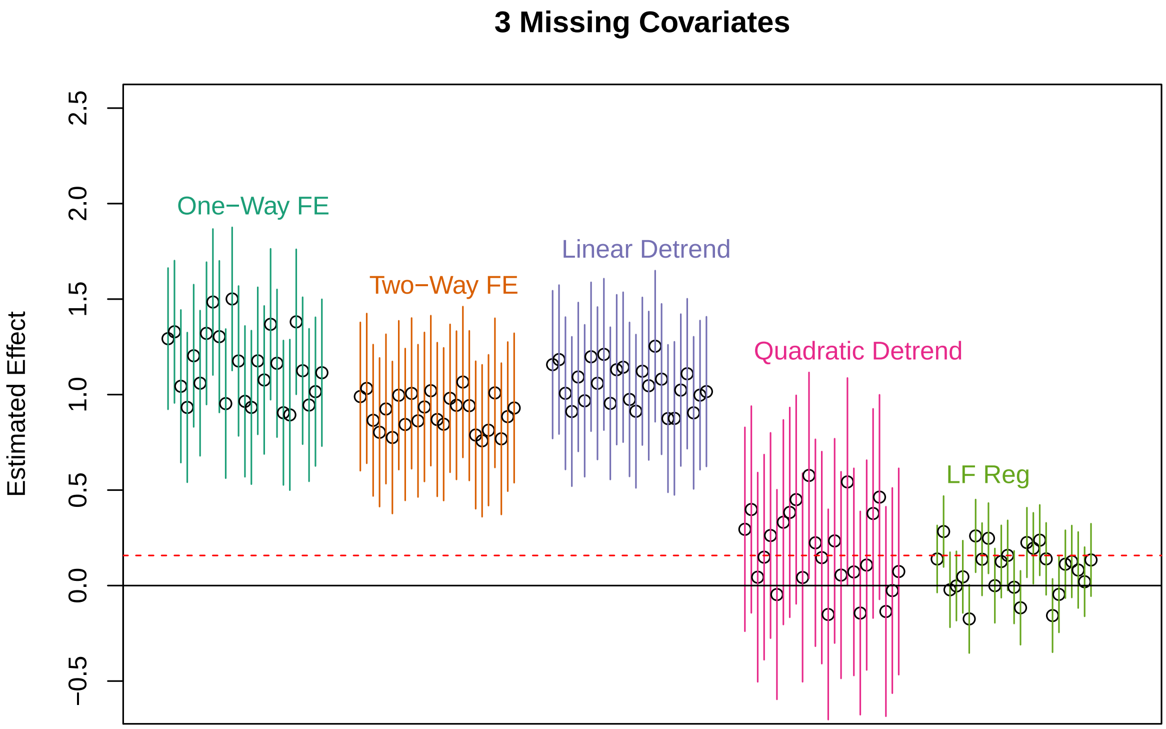


Figure 5: Three random covariates are dropped in each simulation.

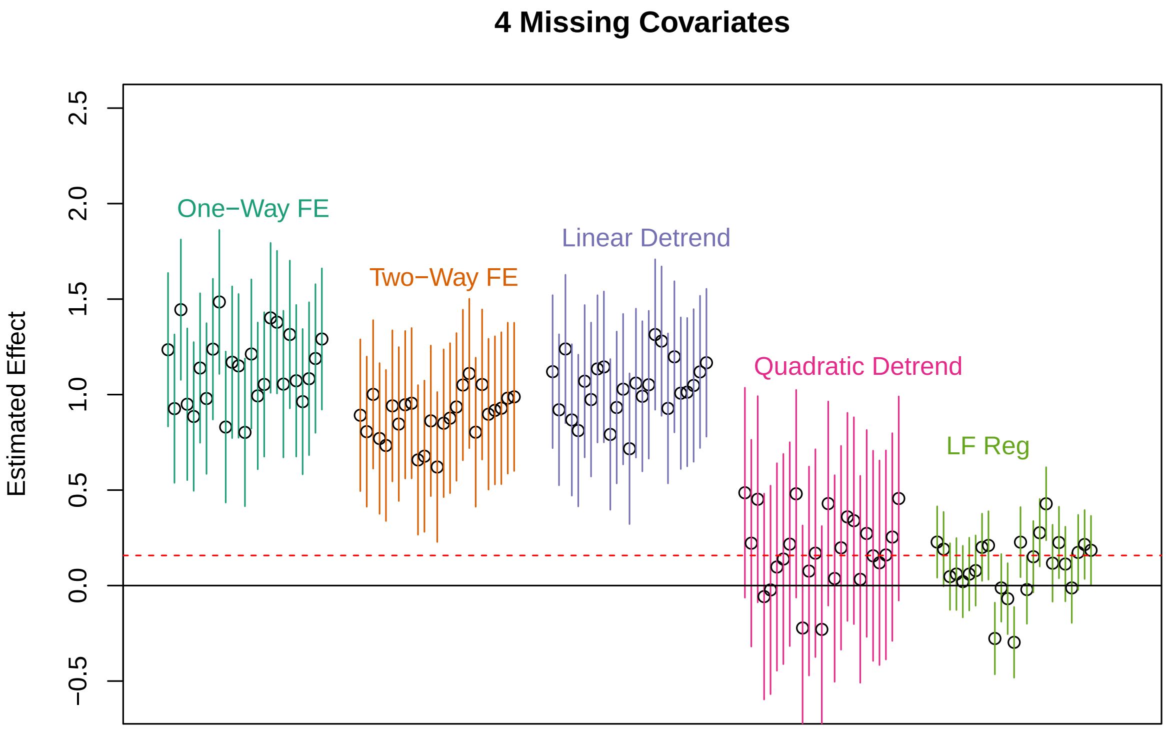


Figure 6: Four random covariates are dropped in each simulation.

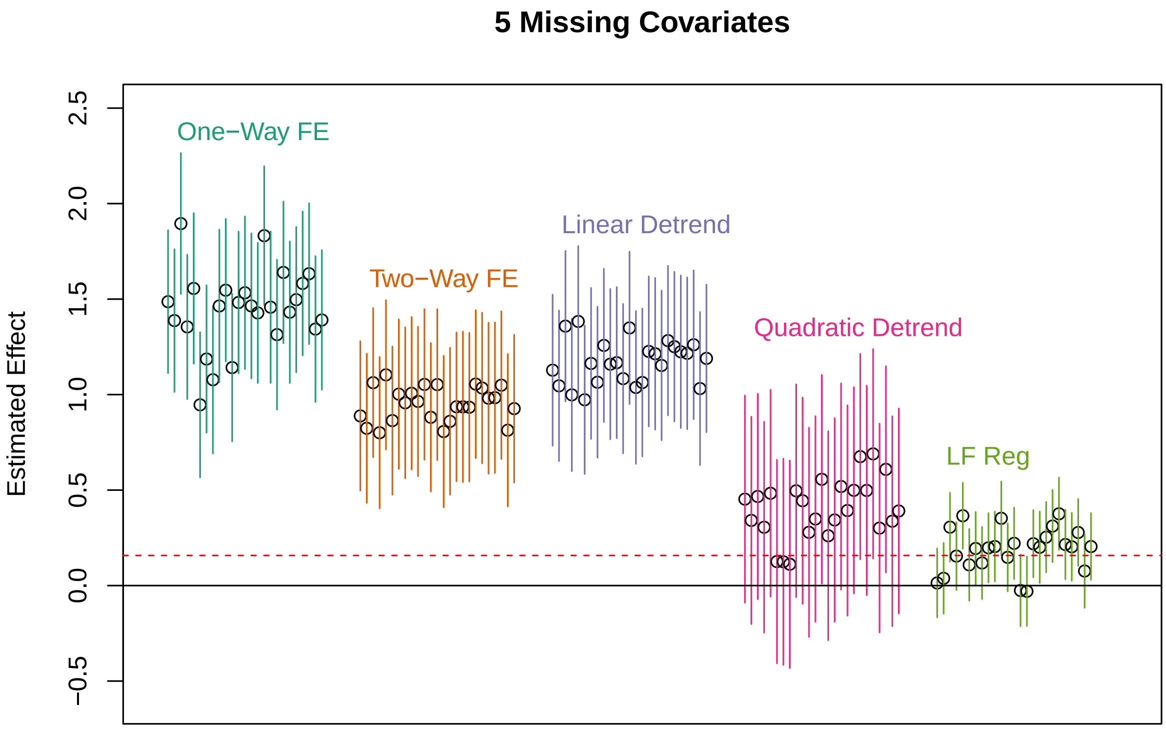


Figure 7: Five random covariates are dropped in each simulation.

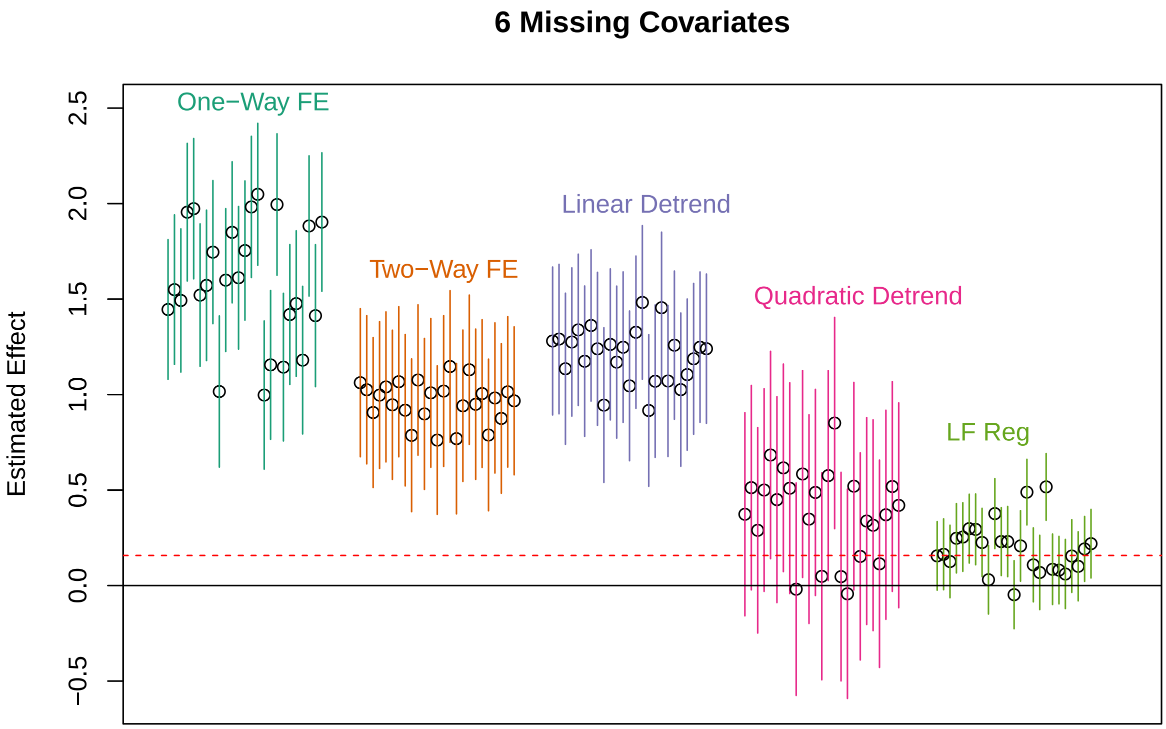


Figure 8: Six random covariates are dropped in each simulation.

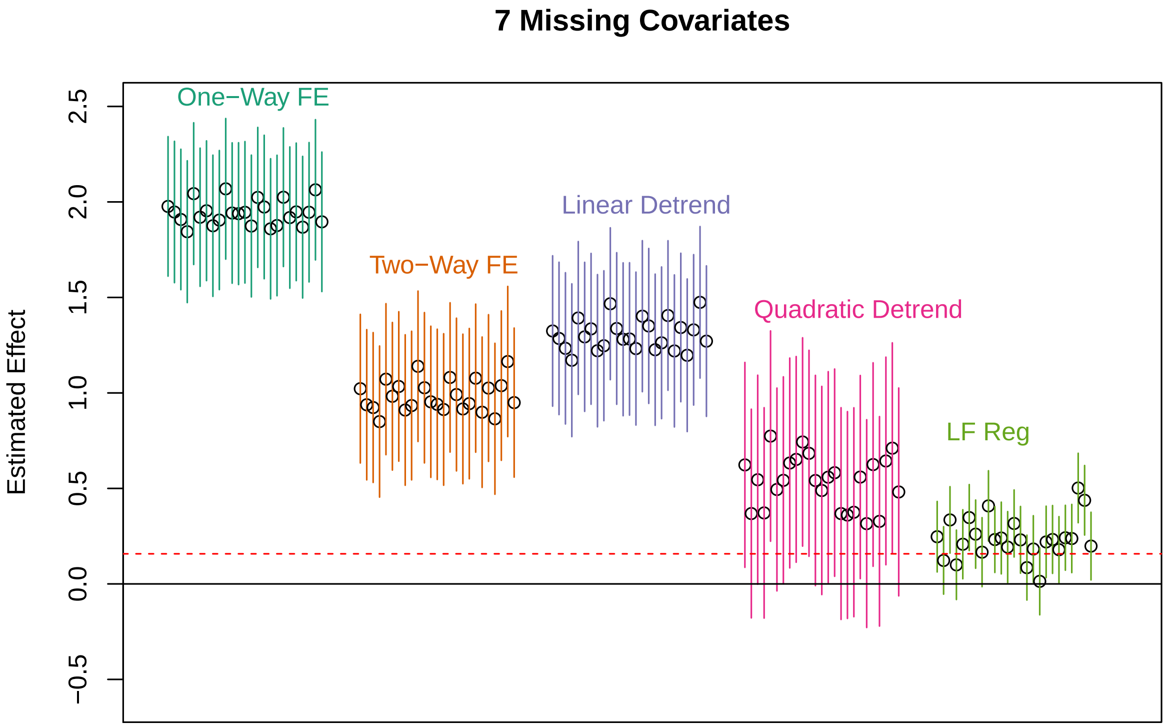


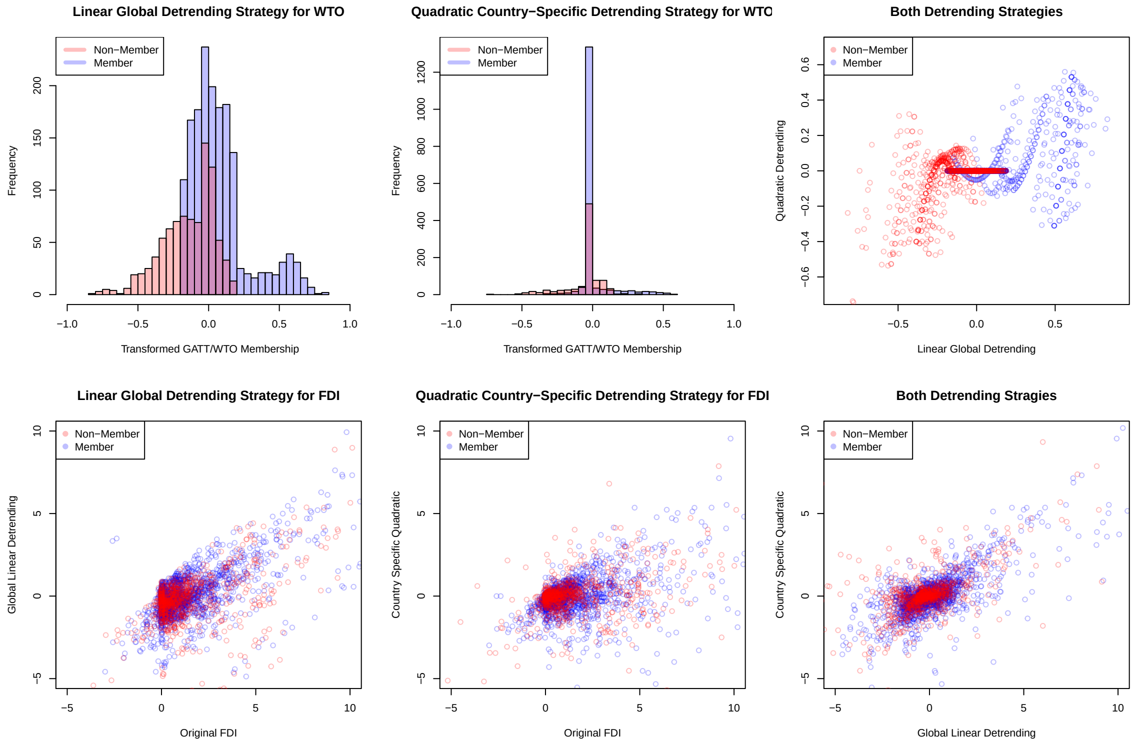
Figure 9: Seven random covariates are dropped in each simulation. This is all but the theoretical variable of interest.

F Additional Application Details

This appendix collects additional details on the applications.

F.1 Understanding the Detrending Strategies for FDI

Before proposing a third model for the FDI data it is helpful to understand the differences in the detrending strategies proposed by both sets of authors. Figure 10 shows the detrending patterns for GATT/WTO membership (the main independent variable) and FDI inflows (the outcome). The first two columns on the top left show the linear and quadratic detrending strategies on the GATT/WTO membership. What this plot makes clear is that both strategies have the effect of creating overlap in the member and non-member distributions such that a non-member country can have a higher value than a member country of the membership variable. It also shows that the quadratic detrending draws many more of the observations directly to zero including many observations where there is membership. The effects of detrending on GATT/WTO membership are particularly severe because the covariate of interest is temporally persistent (once attaining membership countries are not leaving). The differences in the FDI trends, which were the main justification for the quadratic detrending, show more minor differences. However it is clear that the distribution is flattened out a bit more under the quadratic detrending.


Linear Global Detrending Strategy for WTO

Quadratic Country−Specific Detrending Strategy for WTO

Both Detrending Strategies

Linear Global Detrending Strategy for FDI

Quadratic Country−Specific Detrending Strategy for FDI

Both Detrending Stragies

Figure 10: A comparison of detrending strategies. From left to right we show the global linear detrending (Büthe and Milner, 2008) with original, the country-specific quadratic detrending (King and Roberts, 2014) with original and a comparison of the two detrending strategies. The first row shows the GATT/WTO membership variable and the second row shows the outcome. All observations are color coded by their membership status. Note that in the second row we have zoomed in on the main portion of the data but there are large outliers outside the range of the plot.

References

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of Californias tobacco control program." *Journal of the American Statistical Association* 105(490).

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

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

Bada, Oualid and Dominik Liebl. 2014. "The R Package phtt: Panel Data Analysis with Heterogeneous Time Trends." *Journal of Statistical Software* 59(6):1–34.
URL: *http://www.jstatsoft.org/v59/i06/*

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

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, 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.

Bickel, Peter J and Aiyou Chen. 2009. "A nonparametric view of network models and Newman–Girvan and other modularities." *Proceedings of the National Academy of Sciences* 106(50):21068–21073.

Bickel, Peter J, Aiyou Chen and Elizaveta Levina. 2011. "The method of moments and degree distributions for network models." *The Annals of Statistics* 39(5):2280–2301.

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.

Blackwell, Matthew and Adam Glynn. 2013. How to Make Causal Inferences with Time-Series Cross-Sectional Data. Technical report Harvard University.

Bonhomme, Stéphane and Elena Manresa. 2012. Grouped patterns of heterogeneity in panel data. Technical report Citeseer.

Bowers, Jake, Mark M Fredrickson and Costas Panagopoulos. 2013. "Reasoning about interference between units: a general framework." *Political Analysis* 21(1):97–124.

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

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.

Chan, Stanley H and Edoardo M Airoldi. 2014. "A Consistent Histogram Estimator for Exchangeable Graph Models." *arXiv preprint arXiv:1402.1888* .

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

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

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

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

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

Fosdick, Bailey Kathryn. 2013. Modeling Heterogeneity within and between Matrices and Arrays PhD thesis University of Washington.

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

Gershman, Samuel, Matt Hoffman and David Blei. 2012. "Nonparametric variational inference." *arXiv preprint arXiv:1206.4665* .

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

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

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

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.

Han, Shaobo, Xuejun Liao and Lawrence Carin. 2013. Integrated Non-Factorized Variational Inference. In *Advances in Neural Information Processing Systems.* pp. 2481–2489.

Hannah, Lauren A, David M Blei and Warren B Powell. 2011. "Dirichlet process mixtures of generalized linear models." *The Journal of Machine Learning Research* 12:1923–1953.

Hoff, Peter. 2012. *eigenmodel: Semiparametric factor and regression models for symmetric relational data.* R package version 1.01.
URL: *http://CRAN.R-project.org/package=eigenmodel*

Hoff, Peter, Bailey Fosdick, Alex Volfovsky and Kate Stovel. 2014. *amen: Additive and multiplicative effects modeling of networks and relational data.* R package version 0.999.
URL: *http://CRAN.R-project.org/package=amen*

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.

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

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. 1998. "Time Series Models for Discrete Data: solutions to a problem with quantitative studies of international conflict." *Manuscript, Department of Political Science, Stanford University* .

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.

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.

Khan, Mohammad E, Guillaume Bouchard, Kevin P Murphy and Benjamin M Marlin. 2010. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems.* pp. 1108–1116.

Khan, Mohammad E, Shakir Mohamed, Benjamin M Marlin and Kevin P Murphy. 2012. A stick-breaking likelihood for categorical data analysis with latent Gaussian models. In *International conference on Artificial Intelligence and Statistics.* pp. 610–618.

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.

Kyung, Minjung, Jeff Gill and George Casella. 2010. "Estimation in Dirichlet random effects models." *The Annals of Statistics* 38(2):979–1009.

Kyung, Minjung, Jeff Gill and George Casella. 2011. "New findings from terrorism data: Dirichlet process random-effects models for latent groups." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60(5):701–721.

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.

Lloyd, James, Peter Orbanz, Zoubin Ghahramani and Daniel Roy. 2013. "Random function priors for exchangeable arrays with applications to graphs and relational data.".

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

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, 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.

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

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.

Ogburn, Elizabeth L and Tyler J VanderWeele. 2014. "Vaccines, Contagion, and Social Networks." *arXiv preprint arXiv:1403.1241* .

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

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.

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.

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.".

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.

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.

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.

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.*

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.

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

Wand, Matt P. 2014. "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.

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.

Wasserman, Stanley and Katherine Faust. 1994. *Social network analysis: Methods and applications.* Vol. 8 Cambridge university press.

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

Xu, Yiqing. 2014. "Generalized Synthetic Control Method for Causal Inference with Time-Series Cross-Sectional Data.".

Yang, Justin J, Qiuyi Han and Edoardo M Airoldi. 2014. Nonparametric estimation and testing of exchangeable graph models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics.* pp. 1060–1067.

Zhang, Yuchen, Xi Chen, Dengyong Zhou and Michael I Jordan. 2014. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems.* pp. 1260–1268.

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

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