Multilinear tensor regression for longitudinal relational data

Peter Hoff
Departments of Statistics and Biostatistics
University of Washington

November 12, 2014

Abstract

A fundamental aspect of relational data, such as from a social network, is the possibility of dependence among the relations. In particular, the relations between members of one pair of nodes may have an effect on the relations between members of another pair. This article develops a type of regression model to estimate such effects in the context of longitudinal and multivariate relational data, or other data that can be represented in the form of a tensor. The model is based on a general multilinear tensor regression model, a special case of which is a tensor autoregression model in which the tensor of relations at one time point are parsimoniously regressed on relations from previous time points. This is done via a separable, or Kronecker-structured, regression parameter along with a separable covariance model. In the context of an analysis of longitudinal multivariate relational data, we show how the multilinear tensor regression model can represent patterns that often appear in relational and network data, such as reciprocity and transitivity.

1 Introduction

Longitudinal relational data among a set of m objects or nodes can be represented as a time series of matrices $\{\mathbf{Y}_t : t = 1, \ldots, n\}$, where each $\mathbf{Y}_t$ is an $m \times m$ square matrix. The entries of $\mathbf{Y}_t$ represent directed relationships or actions involving pairs of nodes (dyads) at time t, so $y_{i,j,t}$ is a numerical description of the action taken by node i with node j as the target. For example, in this article we consider weekly data on actions involving country pairs. In this case, $y_{i,j,t}$ represents the intensity of actions taken by country i towards country j in week t. For either descriptive, inferential or

predictive purposes, it is often of interest to analyze $\{\mathbf{Y}_t : t = 1, \ldots, n\}$ via a statistical model. In particular, it is often of interest to quantify and evaluate how one dyad's relations might affect those of another at a later time point.

Foundational development for a class of agent-based models of longitudinal network data appears in Snijders (2001) and is developed further in Snijders et al. (2007). This work focuses on models for binary relational data, that is, social networks. These models represent social links in terms of decisions made by nodes that act to maximize their individual utilities. These models include parameters that can be interpreted as preferences for various types of social structures, such as reciprocated dyads or transitive triads. The parameters in such models are typically homogeneous, in that the preference parameters are common to all individuals in the network (or possibly common to all individuals having common observable attributes). A popular alternative to such homogeneous models utilizes a dynamic latent variable formulation, in which each $\mathbf{Y}_t$ is represented as a function of node-specific latent variables $\mathbf{Z}_t$ that evolve over time. Ward and Hoff (2007), Ward et al. (2013) and Durante and Dunson (2013) model the relationship between nodes i and j at time t as a function of low-dimensional real-valued latent variables $\mathbf{z}_{i,t}$ and $\mathbf{z}_{j,t}$. Similar models considered by Fu et al. (2009) and Xing et al. (2010) assume the latent variables are categorical-valued latent classes. Such models can be viewed as a class of random-effects models, and can represent certain types of dependence often seen in social networks and relational data. Somewhat related to this, Westveld and Hoff (2011) and Hoff (2011) consider different covariance models for longitudinal relational data.

A fundamental feature of network data is the statistical interdependence among the relations between individuals. The two modeling approaches discussed in the previous paragraph both represent certain types of network dependencies, but in different ways. The agent-based approach explicitly models how dyads might affect one another, but generally assumes such influences are homogeneous. Conversely, the latent variable approach allows for across-node heterogeneity in the representation of network behavior, but the interdependence between relations is not explicitly parameterized, and the types of dependence that can be represented are limited by the simple structure of the latent variables. This article presents a modeling approach that is unlike either the agent-based or the random effects models, but like the former has an explicit representation of the dependence between dyads, and like the latter allows for nodal heterogeneity in the model

parameters. The approach is based on a reduced-parameter regression model as follows: Consider modeling the actions $\mathbf{Y}_t$ at time t as a function of their values $\mathbf{X}_t \equiv \mathbf{Y}_{t-1}$ at the previous time point. A conceptually simple model for such data would be a vector autoregressive (VAR) model: Letting $\mathbf{y}_t = \mathrm{vec}(\mathbf{Y}_t)$ and $\mathbf{x}_t = \mathrm{vec}(\mathbf{X}_t)$, a first-order VAR model posits that

$$\mathbf{y}_t = \mathbf{\Theta}\mathbf{x}_t + \mathbf{e}_t \ , \quad \mathrm{E}[\mathbf{e}_t] = \mathbf{0} \ , \quad \mathrm{E}[\mathbf{e}_t \mathbf{e}_s^T] = \begin{cases} \Sigma & \text{if } t = s \\ \mathbf{0} & \text{if } t \neq s, \end{cases}$$

where $\mathbf{\Theta}$ and Σ are parameters to be estimated. For simplicity, here and in what follows we consider models without intercepts, appropriate if the time-series for each pair i, j has been demeaned (so that $\sum_t y_{i,j,t}/n = 0$). Given sufficient data, unrestricted estimates of $\mathbf{\Theta}$ in a VAR model can generally be obtained via ordinary least-squares (OLS) or feasible generalized least squares (GLS). However, such estimates can be unstable or unavailable unless the time series is extremely long: As $\mathbf{y}_t$ and $\mathbf{x}_t$ are each of length m^2 (or $m(m-1)$ if the diagonal of each $\mathbf{Y}_t$ is undefined), the regression matrix $\mathbf{\Theta}$ has m^4 entries (m^2 per pair of nodes).

Estimation stability can be improved by restricting $\mathbf{\Theta}$ to belong to a parameter space of lower dimension. In this article we focus on models where the regression matrix has the form $\mathbf{\Theta} = \mathbf{B} \otimes \mathbf{A}$, where $\mathbf{A}$ and $\mathbf{B}$ are $m \times m$ matrices and "$\otimes$" is the Kronecker product. Such a model is a "bilinear" regression model, as in terms of $\mathbf{Y}_t$ and $\mathbf{X}_t$ the model is

$$\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t\mathbf{B}^T + \mathbf{E}_t, \tag{1}$$

so that the regression model is bilinear in the parameters, that is, linear in $\mathbf{A}$ and linear in $\mathbf{B}$, but not linear in $(\mathbf{A}, \mathbf{B})$. This model appears similar to, but is distinct from the "growth curve" model (Potthoff and Roy, 1964; Gabriel, 1998; Srivastava et al., 2009), in which $\mathrm{E}[\mathbf{Y}|\mathbf{X}, \mathbf{Z}, \mathbf{B}] = \mathbf{X}\mathbf{B}\mathbf{Z}^T$, where $\mathbf{X}$ and $\mathbf{Z}$ are known and $\mathbf{B}$ is a matrix of parameters to be estimated. This latter model is linear in the parameters and bilinear in the two explanatory matrices $\mathbf{X}$ and $\mathbf{Z}$. The model in (1) also bears some resemblance to recently developed reduced-rank regression models (Basu et al., 2012; Li et al., 2013; Shi et al., 2014), in which a scalar response y is regressed on a matrix $\mathbf{X}$ via the mean function $\mathrm{tr}(\mathbf{A}\mathbf{X}\mathbf{B}^T)$, where $\mathbf{A}$ is $r_1 \times p_1$ and $\mathbf{B}$ is $r_2 \times p_2$, with $r_1 < p_1$ and $r_2 < p_2$. In such a model, we can write $\mathrm{tr}(\mathbf{A}\mathbf{X}\mathbf{B}^T) = \boldsymbol{\beta}^T\mathbf{x}$ where $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B}^T\mathbf{A})$. The regression coefficient $\boldsymbol{\beta}$ is therefore the vectorization of rank $r_1 \wedge r_2$ matrix, whereas an unrestricted model would be of rank $p_1 \wedge p_2$. In contrast, in model (1) the response is a matrix, and the matrix of regression coefficients $(\mathbf{B} \otimes \mathbf{A})$ is full-rank (although low-dimensional).

Interpretation of the parameters in (1) is facilitated by noting that for a given ordered pair of nodes (i, j),

$$\mathrm{E}[y_{i,j,t}|\mathbf{X}_t] = \sum_{i'}\sum_{j'} a_{i,i'}b_{j,j'}x_{i',j'}.$$

Generally speaking, $a_{i,i'}$ describes how the actions by i are influenced by previous actions of i', and $b_{j,j'}$ describes how actions towards j are influenced by previous actions towards j'. This model could be referred to as a multiplicative model, as $\theta_{i,i',j,j'} = a_{i,i'}b_{j,j'}$ and so the regression coefficients are multiplicative functions of the parameters. A more familiar alternative to the multiplicative model is an additive model, such as

$$y_{i,j,t} = \sum_{i'}\sum_{j'}(a_{i,i'} + b_{j,j'})x_{i',j'} + \epsilon_{i,j,t},$$

$$\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t\mathbf{1}\mathbf{1}^T + \mathbf{1}\mathbf{1}^T\mathbf{X}_t\mathbf{B}^T + \mathbf{E}_t.$$

While perhaps in an unfamiliar form, this additive model can be expressed as an ordinary linear regression model, although with a complicated design matrix. The additive and multiplicative model have essentially the same number of parameters, but their interpretation is somewhat different. In the multiplicative model, the influence of $x_{i',j'}$ on $y_{i,j}$ is non-negligible if *both* $a_{i,i'}$ and $b_{j,j'}$ are non-negligible. In the additive mode, $x_{i',j'}$ influences $y_{i,j}$ if *either* $a_{i,i'}$ or $b_{j,j'}$ are non-negligible. Which model provides closer approximation to the data-generating process will depend on the application. However, we argue that for many longitudinal relational datasets, and longitudinal international relations data in particular, the effect of $x_{i',j'}$ on $y_{i,j}$ will be small for most values of i, i', j, j', and large only when there is some similarity between both i and i', and j and j'. For example, if i and i' have an alliance, and j and j' have an alliance, then the actions of i' towards j' may influence future actions of i towards j, but perhaps not of i towards k, a country unallied with j'.

We evaluate this claim empirically with a brief comparison of the two models. Our data of interest consist of weekly relational measures between pairs of 25 countries over the roughly ten and a half year period from 2004 to mid 2014, giving $n = 543$ weeks of data. The value of $y_{i,j,t}$ is a transformed count of the number of positive verbal statements of country i towards country j during week t (a fuller description of the data appears in a more complete analysis in Section 4). We fit both of these models to the data using a least-squares criterion. Both models explain only a small fraction of the data variation, but the multiplicative model explains over twice as much: The R^2 coefficients (one minus the ratio of the residual sum of squares to the total sum of squares) are

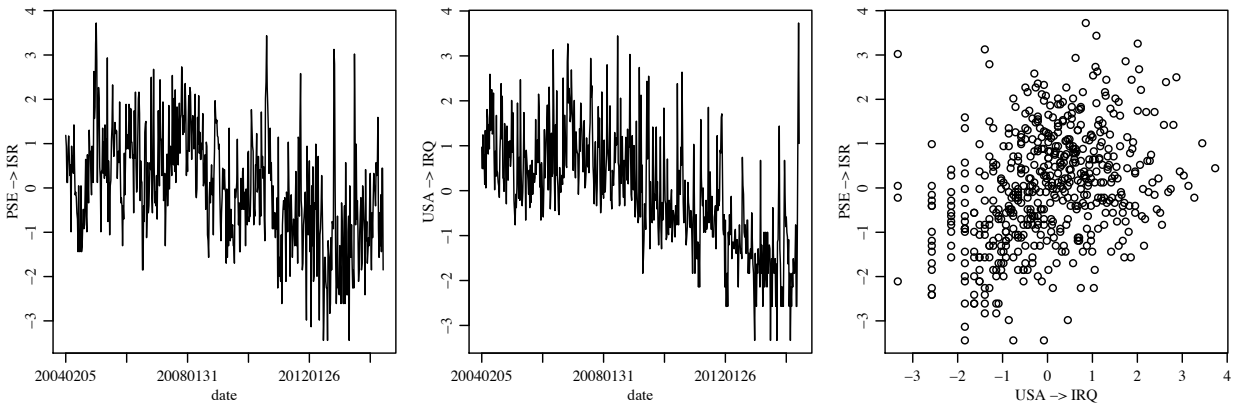

Figure 1: Example data. From left to right, positive verbal relations versus time from PSE to ISR and USA to IRQ, and a scatterplot.

5.8% for the additive model and 13.2% for the multiplicative model. As the models have the same number of parameters, this suggests we should favor the multiplicative model.

As there are a large number of parameters in the multiplicative model (essentially $2 \times m^2$), it is natural to wonder if they are simply representing noise in the data, or a consistent signal. To examine this, we identified the i, j pairs for which the values of $\hat{a}_{i,j}$ and $\hat{b}_{i,j}$ are largest. This information is depicted graphically in Figure 2, in which a link is drawn between countries i and j if $\hat{a}_{i,j}$ is among the largest 10% of values of $\hat{\mathbf{A}}$ (in the left panel), or $\hat{b}_{i,j}$ is among the largest 10% of values of $\hat{\mathbf{B}}$ (the left panel). The figure indicates a strong geographic component to the off-diagonal elements of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. This is empirical evidence that relations between a pair (i', j') are in some cases predictive of future relations between other pairs (i, j). Otherwise, these off-diagonal components would be representing noise, and there would be no discernible geographic pattern.

We examined this claim further with a small cross-validation study. We randomly generated 10 cross validation datasets, each consisting of a training set and test set of 488 and 55 values of $\{\mathbf{Y}_t, \mathbf{X}_t\}$, respectively. For each dataset, least-squares parameter estimates for the additive and multiplicative models were obtained from each training set, and then used to make predictions of each $\mathbf{Y}_t$ in the test set. The multiplicative model outperformed the additive model for all datasets: The average predictive R^2 for the multiplicative model was 12.3% (with a range of 10.9% to 13.7 %), compared to 4.5% (with a range of 3.7% to 5.4%) for the additive model.

These results motivate further study and development of multiplicative models of this form. In

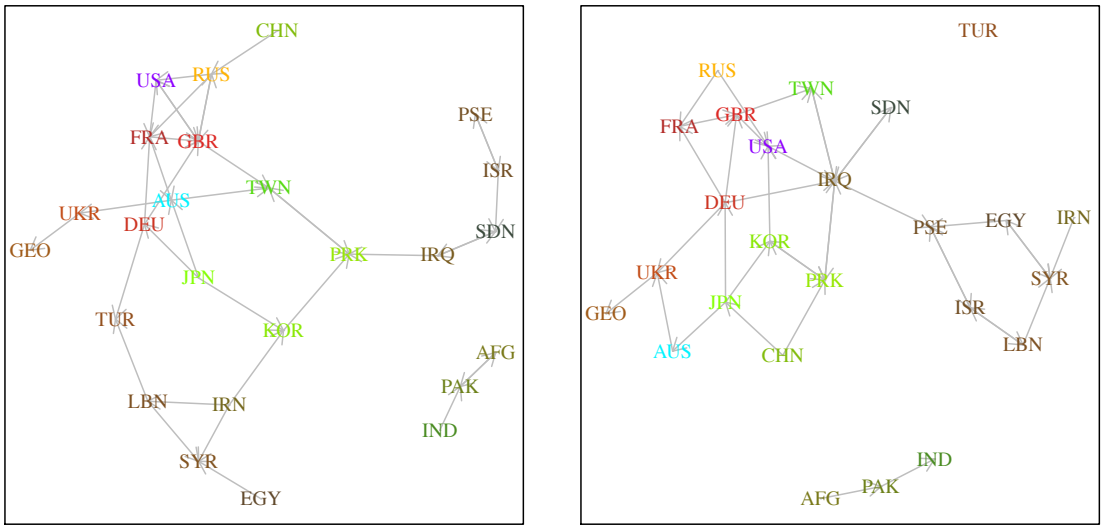

Figure 2: Relatively large entries of $\hat{\mathbf{A}}$ (left) and $\hat{\mathbf{B}}$ (right).

the next section, we present some basic theory for this model including results on identifiability, convergence of OLS estimates and parameter interpretation under model misspecification. We then extend this model to a general multilinear regression model that can accommodate longitudinal measurements of multiway arrays, or tensors. Such models are motivated by the fact that a more complete version of the dataset includes information on four different relation types, and so the data $\mathbf{Y}_t$ at week t consist of a 25×4 three-way tensor. The regression problem than becomes one of regressing the relational tensor $\mathbf{Y}_t$ from time t on the tensor $\mathbf{X}_t = \mathbf{Y}_{t-1}$ from time $t-1$ in a parsimonious way. To accomplish this, in Section 3 we propose and develop the following a multilinear generalization of the bilinear regression model: To relate an $m_1 \times \cdots \times m_K$ tensor $\mathbf{Y}_t$ to a $p_1 \times \cdots \times p_K$ tensor $\mathbf{X}_t$, we use the model

$$\mathbf{Y}_t = \mathbf{X}_t \times \{\mathbf{B}_1, \ldots, \mathbf{B}_K\} + \mathbf{E}_t \text{ or equivalently,}$$

$$\mathbf{y}_t = (\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x}_t + \mathbf{e}_t,$$

where "$\times$" is a multilinear operator known as the "Tucker product," "$\otimes$" is the Kronecker product and $\mathbf{y}_t, \mathbf{x}_t, \mathbf{e}_t$ are the vectorizations of $\mathbf{Y}_t, \mathbf{X}_t, \mathbf{E}_t$, respectively. We present least-squares and Bayesian approaches to parameter estimation, including methods for joint inference on the regression coefficients and the error variance, $\text{Cov}[\mathbf{e}_t] = \Sigma$. Sample size limitations will generally preclude unconstrained estimation of Σ, a $\prod_k m_k$ by $\prod_k m_k$-dimensional error covariance matrix. As a par-

simonious alternative, we use an array normal model for $\mathbf{e}_t$, which is a multivariate normal model with a Kronecker structured covariance matrix, $\mathrm{Cov}[\mathbf{e}_t] = \Sigma_K \otimes \cdots \otimes \Sigma_1$ (Akdemir and Gupta, 2011; Hoff, 2011). Bayesian estimation for the resulting general multilinear tensor regression model with Kronecker structured error covariance can be made using semi-conjugate priors and a Gibbs sampler.

A detailed analysis of the longitudinal relational data presented above is given in Section 4. This includes a cross-validation study to evaluate different models, development of a parsimonious model that allows for network reciprocity and transitivity, and a summary of a Bayesian analysis of the data using this latter model. A discussion of model limitations and possible extensions follows in Section 5.

2 The bilinear regression model

In this section and the next we consider the general problem of regressing one tensor $\mathbf{Y}$ on another tensor $\mathbf{X}$, where $\mathbf{Y}$ and $\mathbf{X}$ are of potentially different sizes. We start with the matrix case: A bilinear regression model of a matrix $\mathbf{Y} \in \mathbb{R}^{m_1 \times m_2}$ on a matrix $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ takes the form

$$\mathbf{Y} = \mathbf{A}\mathbf{X}\mathbf{B}^T + \mathbf{E}, \tag{2}$$

where $\mathbf{E}$ is an $m_1 \times m_2$ matrix of mean-zero disturbance terms, and $\mathbf{A} \in \mathbb{R}^{m_1 \times p_1}$ and $\mathbf{B} \in \mathbb{R}^{m_2 \times p_2}$ are unknown matrices to be estimated. As discussed in the Introduction, this model can be equivalently represented as

$$\mathbf{y} = (\mathbf{B} \otimes \mathbf{A})\mathbf{x} + \mathbf{e}, \tag{3}$$

where "$\otimes$" is the Kronecker product and $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{e}$ are the vectorizations of $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{E}$. Both representations (2) and (3) will be useful in what follows. Note that the parameters $\mathbf{A}$ and $\mathbf{B}$ are not separately identifiable, in that $\mathrm{E}[\mathbf{y}|\mathbf{x}, c\mathbf{A}, \mathbf{B}/c] = \mathrm{E}[\mathbf{y}|\mathbf{x}, \mathbf{A}, \mathbf{B}]$ for any nonzero scalar c. However, these parameters are identifiable up to scale, in the sense that if $(\mathbf{B} \otimes \mathbf{A})\mathbf{x} = (\tilde{\mathbf{B}} \otimes \tilde{\mathbf{A}})\mathbf{x}$ for all $\mathbf{x}$, then $\tilde{\mathbf{A}} = c\mathbf{A}$ and $\tilde{\mathbf{B}} = \mathbf{B}/c$ for some $c \neq 0$ unless all entries of either $\mathbf{A}$ or $\mathbf{B}$ are zero.

Given data $\{(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_n, \mathbf{X}_n)\}$, least-squares parameter estimates $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ of $(\mathbf{A}, \mathbf{B})$ are

minimizers of the residual mean squared error:

$$\begin{aligned}(\hat{\mathbf{A}}, \hat{\mathbf{B}}) &= \arg\min_{\mathbf{A},\mathbf{B}} \sum_i ||\mathbf{Y}_i - \mathbf{A}\mathbf{X}_i\mathbf{B}^T||^2/n \\ &= \arg\min_{\mathbf{A},\mathbf{B}} \sum_i ||\mathbf{Y}_i||^2/n - 2\sum_i \mathrm{tr}(\mathbf{Y}_i^T\mathbf{A}\mathbf{X}_i\mathbf{B}^T/n) + \sum_i \mathrm{tr}(\mathbf{A}\mathbf{X}_i\mathbf{B}^T\mathbf{B}\mathbf{X}_i\mathbf{A}^T/n) \\ &= \arg\min_{\mathbf{A},\mathbf{B}} \ \mathrm{tr}\left(\mathbf{A}^T\mathbf{A}\sum \mathbf{X}_i\mathbf{B}^T\mathbf{B}\mathbf{X}_i/n\right) - 2\,\mathrm{tr}\left(\mathbf{A}^T\sum \mathbf{Y}_i\mathbf{B}\mathbf{X}_i^T/n\right), \qquad (4)\end{aligned}$$

where $\mathrm{tr}(\mathbf{H})$ denotes the trace of a square matrix $\mathbf{H}$, and the term $\sum_i ||\mathbf{Y}_i||^2/n$ has been dropped as it doesn't affect the minimization. Equivalently, using representation (3) we have

$$\begin{aligned}(\hat{\mathbf{A}}, \hat{\mathbf{B}}) &= \arg\min_{\mathbf{A},\mathbf{B}} \sum_i ||\mathbf{y}_i - (\mathbf{B}\otimes\mathbf{A})\mathbf{x}_i||^2/n \\ &= \arg\min_{\mathbf{A},\mathbf{B}} \sum_i ||\mathbf{y}_i||^2/n - 2\,\mathrm{tr}\left((\mathbf{B}\otimes\mathbf{A})\sum \mathbf{x}_i\mathbf{y}_i^T/n\right) + \mathrm{tr}\left((\mathbf{B}^T\mathbf{B}\otimes\mathbf{A}^T\mathbf{A})\sum \mathbf{x}_i\mathbf{x}_i^T/n\right) \\ &= \arg\min_{\mathbf{A},\mathbf{B}} f(\mathbf{A},\mathbf{B},\mathbf{S}_{xx},\mathbf{S}_{xy}),\end{aligned}$$

where

$$f(\mathbf{A},\mathbf{B},\mathbf{S}_{xx},\mathbf{S}_{xy}) = \mathrm{tr}((\mathbf{B}^T\mathbf{B}\otimes\mathbf{A}^T\mathbf{A})\mathbf{S}_{xx}) - 2\,\mathrm{tr}\left((\mathbf{B}\otimes\mathbf{A})\mathbf{S}_{xy}\right), \qquad (5)$$

with $\mathbf{S}_{xx} = \sum \mathbf{x}_i\mathbf{x}_i^T/n$ and $\mathbf{S}_{xy} = \sum \mathbf{x}_i\mathbf{y}_i^T/n$.

Taking derivatives of the objective function in (4) or (5) with respect to $\mathbf{A}$ indicates that for a non-zero value of $\mathbf{B}$, the minimizer of the residual mean squared error in $\mathbf{A}$ is given by

$$\tilde{\mathbf{A}}(\mathbf{B}) = \left(\sum \mathbf{Y}_i\mathbf{B}\mathbf{X}_i^T\right)\left(\sum \mathbf{X}_i\mathbf{B}^T\mathbf{B}\mathbf{X}_i^T\right)^{-1}.$$

A similar calculation shows that for a non-zero value of $\mathbf{A}$, the minimizer in $\mathbf{B}$ is given by

$$\tilde{\mathbf{B}}(\mathbf{A}) = \left(\sum \mathbf{Y}_i^T\mathbf{A}\mathbf{X}_i\right)\left(\sum \mathbf{X}_i^T\mathbf{A}^T\mathbf{A}\mathbf{X}_i\right)^{-1}.$$

This suggests the following alternating least-squares (ALS) algorithm to locate local minima of (5): Given values $\{\hat{\mathbf{A}}^{(s)}, \hat{\mathbf{B}}^{(s)}\}$ at iteration s, new values are generated as $\hat{\mathbf{A}}^{(s+1)} = \tilde{\mathbf{A}}(\hat{\mathbf{B}}^{(s)})$ and $\hat{\mathbf{B}}^{(s+1)} = \tilde{\mathbf{B}}(\hat{\mathbf{A}}^{(s+1)})$. Such a procedure is a block coordinate descent algorithm, and will converge to a local minimum of (5) if certain conditions on the data are met (such as $\sum \mathbf{X}_i\mathbf{B}^T\mathbf{B}\mathbf{X}_i^T$ and $\sum \mathbf{X}_i^T\mathbf{A}^T\mathbf{A}\mathbf{X}_i$ being invertible for all nonzero $\mathbf{A}$ and $\mathbf{B}$ - see Luenberger and Ye (2008, Section 8.9)).

One would hope that, given sufficient data, the parameter estimates would bear some resemblance to the true data-generating mechanism. We investigate this by examining the critical points of a large-sample version of the objective function (5). Consider a scenario in which $\mathbf{S}_{xx} = \sum \mathbf{x}_i \mathbf{x}_i^T / n$ converges almost surely to a positive definite matrix $\Sigma_{xx} = \mathrm{E}[\mathbf{x}\mathbf{x}^T]$ and $\mathbf{S}_{xy} = \sum \mathbf{x}_i \mathbf{y}_i^T / n$ converges almost surely to a matrix $\Sigma_{xy} = \mathrm{E}[\mathbf{x}\mathbf{y}^T]$. This implies almost sure convergence of $f(\mathbf{A}, \mathbf{B}, \mathbf{S}_{xx}, \mathbf{S}_{xy})$ to $f(\mathbf{A}, \mathbf{B}, \Sigma_{xx}, \Sigma_{xy})$, and so we would expect that a minimizer of $f(\mathbf{A}, \mathbf{B}, \mathbf{S}_{xx}, \mathbf{S}_{xy})$ would resemble a minimizer of $f(\mathbf{A}, \mathbf{B}, \Sigma_{xx}, \Sigma_{xy})$, given sufficient data. In particular, results of White (1981) imply that if estimation of $\{\mathbf{A}, \mathbf{B}\}$ is restricted to a compact subset of $\mathbb{R}^{m_1 \times p_1} \times \mathbb{R}^{m_2 \times p_2}$, then a sequence of local minimizers $\{\hat{\mathbf{A}}_n, \hat{\mathbf{B}}_n\}$ of $f(\mathbf{A}, \mathbf{B}, \mathbf{S}_{xx}, \mathbf{S}_{xy})$ will converge almost surely to the global minimizer of $f(\mathbf{A}, \mathbf{B}, \Sigma_{xx}, \Sigma_{xy})$, if one exists. This motivates an investigation of minimizers of $f(\mathbf{A}, \mathbf{B}, \Sigma_{xx}, \Sigma_{xy})$ under various conditions on Σ_{xx} and Σ_{xy}. Such minimizers are referred to as "pseudotrue" parameters in the literature on nonlinear least squares estimates and misspecified models (see, for example, White (1981, 1982)).

The ideal condition is, of course, when the model is correct. In this case, $\mathrm{E}[\mathbf{y}|\mathbf{x}] = (\mathbf{B}_0 \otimes \mathbf{A}_0)\mathbf{x}$ and so $\Sigma_{xy} = \mathrm{E}[\mathbf{x}\mathbf{x}^T(\mathbf{B}_0 \otimes \mathbf{A}_0)^T] = \Sigma_{xx}(\mathbf{B}_0 \otimes \mathbf{A}_0)^T$. The large sample objective function is then

$$\mathrm{tr}((\mathbf{B}^T\mathbf{B} \otimes \mathbf{A}^T\mathbf{A})\Sigma_{xx}) - 2\,\mathrm{tr}((\mathbf{B} \otimes \mathbf{A})\Sigma_{xx}(\mathbf{B}_0 \otimes \mathbf{A}_0)^T).$$

If Σ_{xx} is positive definite, then the objective function is uniquely minimized in $(\mathbf{B} \otimes \mathbf{A})$ by the truth $(\mathbf{B}_0 \otimes \mathbf{A}_0)$. The pseudotrue parameters are equal to the true parameters, and the least-squares estimator is asymptotically consistent.

If the model is incorrect, we may still hope that $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ conveys meaningful information about the data generating mechanism. For example, recall that $a_{i,j}$, the i,jth element of $\mathbf{A}$, represents a measure of the conditional dependence of $\mathbf{y}_{[i,]} = (y_{i,1}, \ldots, y_{i,m_2})^T$, the ith row of $\mathbf{Y}$, on $\mathbf{x}_{[j,]} = (x_{j,1}, \ldots, x_{j,p_2})^T$, the jth row of $\mathbf{X}$, given the other rows of $\mathbf{X}$. If there is no such dependence, then we would hope that the pseudotrue parameter for $a_{i,j}$ would be zero as well. It can be shown that this is true, under some additional conditions:

Proposition 1. *If* $E[\mathbf{x}_{[j,]}\mathbf{y}_{[i,]}^T] = \mathbf{0}$ *and* $E[\mathbf{x}_{[j,]}\mathbf{x}_{[k,]}^T] = \mathbf{0}$ *for all* $k \neq j$*, then then the pseudotrue parameter for* $a_{i,j}$ *is zero.*

A similar result holds if the conditional expectation of $\mathbf{Y}$ given $\mathbf{X}$ is truly linear, although not necessarily Kronecker structured. In this case we can write $\mathrm{E}[\mathbf{y}|\mathbf{x}] = \Theta\mathbf{x}$, where here $\mathbf{y}$ and $\mathbf{x}$ are

the vectorizations of $\mathbf{Y}$ and $\mathbf{X}$.

Proposition 2. *Let* $E[\mathbf{y}|\mathbf{x}] = \Theta\mathbf{x}$ *and* $E[\mathbf{x}\mathbf{x}^T] = \Omega \otimes \Psi$ *for some positive definite matrices* Ω *and* Ψ*. Then if the entries of* Θ *corresponding to the elements of* $\mathbf{y}_{[i,]}$ *and* $\mathbf{x}_{[j,]}$ *are zero, then the pseudotrue parameter for* $a_{i,j}$ *is zero.*

Proofs of both propositions are in the appendix. The conditions of both results correspond to $\mathbf{y}_{[i,]}$ being "conditionally uncorrelated" with $\mathbf{x}_{[j,]}$ in some way: Under the conditions of the first proposition, the inverse of a covariance matrix of the elements of $\mathbf{Y}$ and $\mathbf{X}$ would have zeros for all entries corresponding to elements of $\mathbf{y}_{[i,]}$ and $\mathbf{x}_{[j,]}$, that is, the partial correlations are zero. In the second proposition, Θ represents the conditional relationship directly.

3 Extension to correlated multiway data

In this section we extend the bilinear regression model in two directions: First, we show that the bilinear model is a special case of a more general type of multilinear tensor regression model that can be applied to tensor valued data. Such a model can accommodate, for example, multivariate longitudinal relational data of the type described in Section 1, where we have multiple relation types measured between pairs of countries over time. Such data can be represented as a time series of three-way tensors. A second extension of the model allows for covariance in the error term. As sample size limitations will generally preclude unrestricted estimation of the covariance, we consider instead a reduced-dimension multilinear covariance model that allows for correlation along each mode of the tensor. The covariance model, like the mean model, is obtained from a multilinear transformation, so we refer to the combined mean and covariance model as a general multilinear regression model. The joint multilinear structure of the mean and covariance facilitates parameter estimation. In particular, a Bayesian approach to generalized least-squares (GLS) is available via a straightforward Gibbs sampling algorithm.

3.1 Multilinear tensor regression

The bilinear regression model maps a covariate matrix $\mathbf{X} \in \mathbb{R}^{p_1 \times p_2}$ to a mean matrix $\mathbf{M} = \mathbf{A}\mathbf{X}\mathbf{B}^T \in \mathbb{R}^{m_1 \times m_2}$. Equivalently, the model maps $\mathbf{x}$, the vectorization of $\mathbf{X}$, to $\mathbf{m} = (\mathbf{B} \otimes \mathbf{A})\mathbf{x}$, the vectorization of $\mathbf{M}$. Such a map between spaces of matrices is a special case of a more general class

of maps between spaces of multiway arrays, or tensors. Specifically, given matrices $\mathbf{B}_1, \ldots, \mathbf{B}_K$, with $\mathbf{B}_k \in \mathbb{R}^{p_k \times m_k}$, we can define a mapping from $\mathbb{R}^{p_1 \times \cdots \times p_K}$ to $\mathbb{R}^{m_1 \times \cdots \times m_K}$ by first obtaining the vectorization $\mathbf{x}$, computing $\mathbf{m} = (\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x}$, and then forming an $m_1 \times \cdots \times m_K$-dimensional array $\mathbf{M}$ from $\mathbf{m}$. This transformation is known as the "Tucker product" (Tucker, 1964) of the array $\mathbf{X}$ and the list of matrices $\mathbf{B}_1, \ldots, \mathbf{B}_K$, which we write as

$$\mathbf{M} = \mathbf{X} \times \{\mathbf{B}_1, \ldots, \mathbf{B}_K\}.$$

An important class of operations related to the Tucker product are *matricizations*, which reshape an array $\mathbf{M}$ into matrices of various dimensions. For example, the mode-1 matricization of a $m_1 \times m_2 \times m_3$-dimensional array $\mathbf{M}$ is an $m_1 \times (m_2 m_3)$-dimensional matrix denoted $\mathbf{M}_{(1)}$. More generally, the mode-k matricization of an $m_1 \times \cdots \times m_K$-dimensional array $\mathbf{M}$ is an $m_k \times (\prod_{j:j \neq k} m_j)$-dimensional matrix denoted $\mathbf{M}_{(k)}$. The matricization operation facilitates both understanding and computation of the Tucker product via the following set of equivalencies:

$$\begin{aligned} \mathbf{M} &= \mathbf{X} \times \{\mathbf{B}_1, \ldots, \mathbf{B}_K\}, \\ \mathbf{m} &= (\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x}, \qquad (6) \\ \mathbf{M}_{(k)} &= \mathbf{B}_k \mathbf{X}_{(k)} (\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_{k+1} \otimes \mathbf{B}_{k-1} \otimes \cdots \otimes \mathbf{B}_1). \qquad (7) \end{aligned}$$

In particular, (7) can be used to compute the Tucker product via a series of reshapings and matrix multiplications. Additionally, this result indicates that the Tucker product consists of a series of linear transformations along the different modes of the array. More on the Tucker product and related operations can be found in, for example, De Lathauwer et al. (2000), Kolda and Bader (2009) and Hoff (2011).

Given an explanatory tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_k}$ and outcome tensor $\mathbf{Y} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$, the Tucker product can be used to define the following multilinear tensor regression model,

$$\mathbf{Y} = \mathbf{X} \times \{\mathbf{B}_1, \ldots, \mathbf{B}_K\} + \mathbf{E}, \qquad (8)$$

where $\mathbf{B}_k \in \mathbb{R}^{m_k \times p_k}$, $k = 1, \ldots, K$. If $K = 3$ for example, the model for the i, j, kth element of $\mathbf{Y}$ is

$$y_{i,j,k} = \sum_{i'} \sum_{j'} \sum_{k'} b_{1,i,i'} b_{2,j,j'} b_{3,k,k'} x_{i',j',k'} + \epsilon_{i,j,k},$$

and so $b_{1,i,i'}$ can be viewed as the multiplicative effect of the i''th "slice" of $\mathbf{X}$ on the ith slice of $\mathbf{Y}$. The similarity of this model to the bilinear regression model is most easily seen via the vectorized version of (8), which takes the following form

$$\mathbf{y} = (\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x} + \mathbf{e}. \tag{9}$$

With this notation, replicate observations $\{(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_n, \mathbf{X}_n)\}$ are easily handled by "stacking" the arrays to form two $(K+1)$-way arrays $\mathbf{Y} \in \mathbb{R}^{m_1 \times \cdots \times m_K \times n}$ and $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$, where the $(K+1)$st mode indexes the replications. If each slice follows model (8), then the model for the stacked data is

$$\begin{aligned} \mathbf{Y} &= \mathbf{X} \times \{\mathbf{B}_1, \ldots, \mathbf{B}_K, \mathbf{I}_n\} + \mathbf{E}, \text{ or equivalently} \\ \mathbf{y} &= (\mathbf{I}_n \otimes \mathbf{B}_k \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x} + \mathbf{e}, \end{aligned} \tag{10}$$

where $\mathbf{I}_n$ is an $n \times n$ diagonal matrix, $\mathbf{E}$ is a mean-zero array of the same dimension as $\mathbf{Y}$, and $\mathbf{e}$ is the vectorization of $\mathbf{E}$. However, in what follows we work with model (8), while recognizing that estimation with replications can be handled as a special case by stacking the replications and fixing the parameter matrix for the last mode to be the identity matrix.

Estimation is facilitated by application of identity (7). For example, matricizing each term in (8) along the first mode gives

$$\mathbf{Y}_{(1)} = \mathbf{B}_1 \tilde{\mathbf{X}}_{(1)} + \mathbf{E}_{(1)}, \tag{11}$$

where $\tilde{\mathbf{X}}_{(1)} = \mathbf{X}_{(1)}(\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_2)$. In terms of $\mathbf{B}_1$, this is simply a multivariate linear regression model (Mardia et al., 1979, Chapter 6). The least-squares criterion in $\mathbf{B}_1$ is $||\mathbf{Y} - \mathbf{B}_1\tilde{\mathbf{X}}||^2$, which is uniquely minimized in $\mathbf{B}_1$ by $\mathbf{Y}\tilde{\mathbf{X}}^T(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)^{-1}$ (if $\tilde{\mathbf{X}}$ has full row rank). Similar forms result from matricizing along any of the other K modes. It follows that estimates of $\{\mathbf{B}_1, \ldots, \mathbf{B}_K\}$ can be obtained by generalizing the block coordinate descent algorithm described in Section 2. Given starting values of $\mathbf{B}_1, \ldots, \mathbf{B}_K$, the algorithm is to iterate the following steps until convergence:

For for each $k \in \{1, \ldots, K\}$

1. compute $\tilde{\mathbf{X}} = \mathbf{X} \times \{\mathbf{B}_1, \ldots, \mathbf{B}_{k-1}, \mathbf{I}_{p_k}, \mathbf{B}_{k+1}, \ldots, \mathbf{B}_K\}$;
2. form $\mathbf{Y}_{(k)}$ and $\tilde{\mathbf{X}}_{(k)}$, the mode-k matricizations of $\mathbf{Y}$ and $\tilde{\mathbf{X}}$;

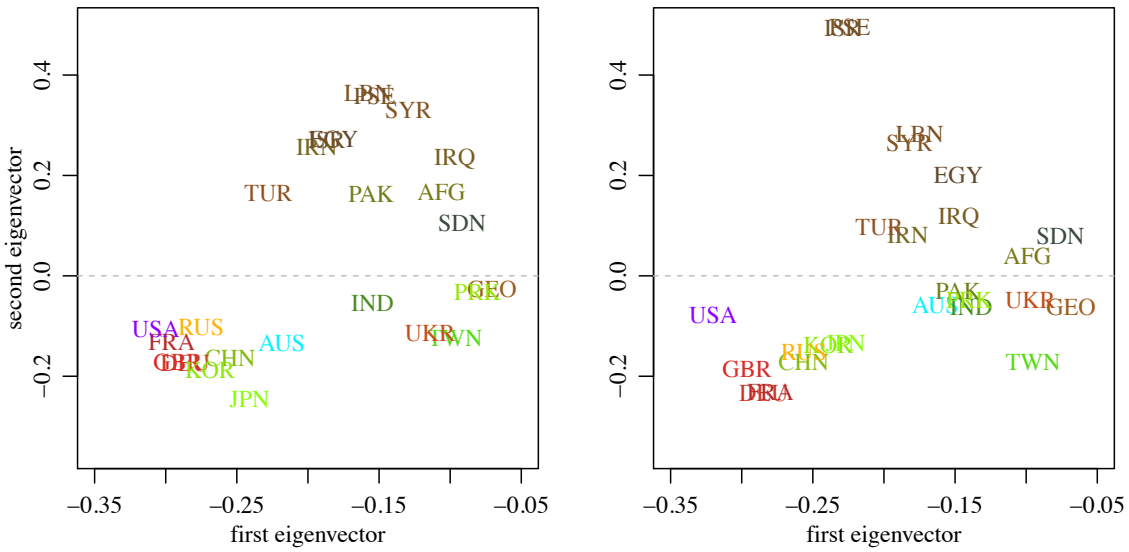

Figure 3: Eigenvectors of mode-specific residual correlation matrices.

3. set $\mathbf{B}_k = \mathbf{Y}_{(k)}\tilde{\mathbf{X}}_{(k)}\left(\tilde{\mathbf{X}}_{(k)}\tilde{\mathbf{X}}_{(k)}^T\right)^{-1}$.

Note that in the algorithm we are computing $\tilde{\mathbf{X}}_{(1)}$, for example, by first computing $\mathbf{X} \times \{\mathbf{I}_{p_1}, \mathbf{B}_2, \ldots, \mathbf{B}_K\}$ and then matricizing, rather than matricizing $\mathbf{X}$ and then multiplying on the right by $(\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_2)$. The two approaches give the same result, but the former can be accomplished with $K-1$ "small" matrix multiplications, whereas the latter requires construction of and multiplication by $(\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_2)$, which can be unmanageably large in some applications.

3.2 Inference under a separable covariance model

The international relations data that will be analyzed in Section 4 consists of time series of four different relational measurements between pairs of 25 countries. These data can be represented as a four-way array $\mathbf{Y} \in \mathbb{R}^{25\times25\times4\times543}$. Using the algorithm described in Section 3.1, we obtained least-squares estimates of $\{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3\}$ in the model $\mathbf{Y} = \mathbf{X} \times \{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \mathbf{I}\} + \mathbf{E}$, where $\mathbf{X}$ is a lagged version of $\mathbf{Y}$. These estimates are equivalent to maximum likelihood estimates under the assumption of i.i.d. residual variation. The plausibility of this assumption is examined graphically in Figure 3. This plot shows eigenvectors of the sample correlation matrices of $\mathbf{R}_{(1)}$ and $\mathbf{R}_{(2)}$, which are the mode-1 and mode-2 matricizations of the residual array $\mathbf{R} = \mathbf{Y} - \mathbf{X} \times \{\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2, \hat{\mathbf{B}}_3, \mathbf{I}\}$. These plots should appear patternless under the assumption of i.i.d. residuals. Instead, clear patterns are

exhibited, most of which are geographic, indicating correlations among the residuals among certain groups of countries. In cases like this where residual variation is not well represented by an i.i.d. assumption, it may be preferable to use an estimation method that accounts for residual correlation or heteroscedasticity.

A separable covariance model: Given multiple observations, we might model the residuals in the vectorized version of the model (9) as $\mathbf{e}_1, \ldots, \mathbf{e}_n \sim$ i.i.d. $N_m(\mathbf{0}, \Sigma)$, where $m = \prod m_k$ and Σ is an unknown covariance matrix to be estimated. The difficulty with this, as with an unrestricted regression model, is that the sample size will generally be too small to reliably estimate Σ without making some restrictions on its form. A flexible, reduced-parameter covariance model that retains the tensor structure of the data is the array normal model (Akdemir and Gupta, 2011; Hoff, 2011), which assumes a separable (Kronecker structured) covariance matrix. For example, we say that $\mathbf{E}$ has a mean-zero array normal distribution, and write $\mathbf{E} \sim N_{m_1 \times \cdots \times m_K}(\mathbf{0}, \Sigma_1, \ldots, \Sigma_K)$, if the distribution of the vectorization $\mathbf{e}$ of $\mathbf{E}$ is given by

$$\mathbf{e} \sim N_m(\mathbf{0}, \Sigma_K \otimes \cdots \otimes \Sigma_1),$$

where Σ_k is a positive definite $m_k \times m_k$ matrix for each $k = 1, \ldots, K$. Each Σ_k can be interpreted as the covariance along the kth mode of $\mathbf{E}$. For example, if $\mathbf{E} \sim N_{m_1 \times \cdots \times m_K}(\mathbf{0}, \Sigma_1, \ldots, \Sigma_K)$, then it is straightforward to show that $\mathrm{E}[\mathbf{E}_{(k)}\mathbf{E}_{(k)}^T] \propto \Sigma_k$, where $\mathbf{E}_{(k)}$ is the mode-k matricization of $\mathbf{E}$.

Combining this error model with the mean model in (8), and applying identities (6) and (7) gives three equivalent forms for this general multilinear tensor regression model:

$$\begin{array}{lll} \text{Tensor form:} & \mathbf{Y} = \mathbf{X} \times \{\mathbf{B}_1, \ldots, \mathbf{B}_K\} + \mathbf{E} & \mathbf{E} \sim N_{m_1 \times \cdots \times m_K}(\mathbf{0}, \Sigma_1, \ldots, \Sigma_K) \\ \text{Vector form:} & \mathbf{y} = (\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x} + \mathbf{e} & \mathbf{e} \sim N_m(\mathbf{0}, \Sigma_K \otimes \cdots \otimes \Sigma_1) \\ \text{Matrix form:} & \mathbf{Y}_{(k)} = \mathbf{B}_k \mathbf{X}_{(k)} \mathbf{B}_{-k}^T + \mathbf{E}_{(k)} & \mathbf{E}_{(k)} \sim N_{m_k \times m_{-k}}(\mathbf{0}, \Sigma_k, \Sigma_{-k}), \end{array} \tag{12}$$

where in the matrix form, $\mathbf{B}_{-k} = \mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_{k+1} \otimes \mathbf{B}_{k-1} \otimes \cdots \otimes \mathbf{B}_1$, Σ_{-k} is defined similarly and $m_{-k} = \prod_{l:l \neq k} m_k$. As before, we note that n replications from a K-mode model can be represented by stacking the data arrays and using a $(K+1)$-mode model with the restriction that $\mathbf{B}_{K+1} = \Sigma_{K+1} = \mathbf{I}_n$.

As in the uncorrelated case, the matrix form of the model can be used to obtain iterative algorithms for parameter estimation. For example, multiplying the terms in the matrix form on

the right by $\Sigma_{-k}^{-1/2}$ allows us to express the model as

$$\tilde{\mathbf{Y}}_{(k)} = \mathbf{B}_k \tilde{\mathbf{X}}_{(k)} + \tilde{\mathbf{E}}_{(k)}\,, \quad \tilde{\mathbf{E}}_{(k)} \sim N_{m_k \times m_{-k}}(\mathbf{0}, \Sigma_k, \mathbf{I}_{m_{-k}}) \tag{13}$$

where now $\tilde{\mathbf{Y}}_{(k)} = \mathbf{Y}_{(k)}\Sigma_{-k}^{-1/2}$ and $\tilde{\mathbf{X}}_{(k)} = \mathbf{X}_{(k)}\mathbf{B}_{-k}^T\Sigma_{-k}^{-1/2}$. Given the parameters other than $\mathbf{B}_k$, this is a multivariate linear regression model with dependent errors. The (conditional) MLE and generalized least squares estimator is $\hat{\mathbf{B}}_k = \tilde{\mathbf{Y}}_{(k)}\tilde{\mathbf{X}}_{(k)}^T(\tilde{\mathbf{X}}_{(k)}\tilde{\mathbf{X}}_{(k)}^T)^{-1}$, which has the same form as the OLS estimator (see, for example, Mardia et al. (1979, Section 6.6.3)), except here the covariance along the modes other than k have been incorporated into the construction of $\tilde{\mathbf{Y}}_{(k)}$ and $\tilde{\mathbf{X}}_{(k)}$. Generalized least-squares estimates of the $\mathbf{B}_k$'s, conditional on values of the Σ_k's, can thus be found via the coordinate descent algorithm in the previous subsection, modulo the modification to $\tilde{\mathbf{Y}}_{(k)}$ and $\tilde{\mathbf{X}}_{(k)}$. Analogously, given current values of the $\mathbf{B}_k$'s, the likelihood can be minimized in the Σ_k's by applying a similar iterative algorithm, described in Hoff (2011).

Bayesian estimation: Generally speaking, maximum likelihood estimates in high dimensional problems can be unstable and overfit to the data. Philosophical issues aside, Bayes methods provide parameter estimates that ameliorate these problems to some extent. Additionally, Markov chain Monte Carlo (MCMC) methods that approximate posterior distributions are useful for exploring the parameter space in a way that is often more informative than computing a matrix of second derivatives at a local mode, especially if the dimension of the parameter space is large. With this in mind, we present a class of semiconjugate prior distributions for the model (12), and obtain a Gibbs sampler that can be used to simulate parameter values from the corresponding posterior distribution.

Recall from the previous subsection that, given $\{\mathbf{B}_l : l \neq k\}$ and $\{\Sigma_l : l \neq k\}$, the model in terms of $(\mathbf{B}_k, \Sigma_k)$ can be expressed as an ordinary multivariate regression model,

$$\mathbf{Y} \sim N_{m\times n}(\mathbf{B}\mathbf{X}, \Sigma, \mathbf{I}_n) \tag{14}$$

where $\mathbf{B} \in \mathbb{R}^{m\times p}$ and $\Sigma \in \mathcal{S}_p^+$ are to be estimated from $\mathbf{Y} \in \mathbb{R}^{m\times n}$ and $\mathbf{X} \in \mathbb{R}^{p\times n}$. As such, Bayesian inference for $\{(\mathbf{B}_k, \Sigma_k), k = 1, \ldots, K\}$ can be made via a Gibbs sampler that iteratively re-expresses the model in terms of (14) for each k, and simulates $(\mathbf{B}_k, \Sigma_k)$ from the corresponding posterior distribution.

Posterior inference for (14) is facilitated by choosing a conjugate prior, which for this model is $\Sigma \sim$ inverse-Wishart$(\mathbf{S}_0^{-1}, \nu_0)$ and $\mathbf{B}|\Sigma \sim N_{m\times p}(\mathbf{M}_0, \Sigma, \mathbf{I}_p)$, where the inverse-Wishart distribution

is parameterized so that $\mathrm{E}[\Sigma^{-1}] = \nu_0 \mathbf{S}_0^{-1}$. Under this prior and model (14), the joint posterior density of $(\mathbf{B}, \Sigma)$ given $\mathbf{Y}$ can be expressed as $p(\mathbf{B}, \Sigma|\mathbf{Y}) = p(\mathbf{B}|\Sigma, \mathbf{Y}) \times p(\mathbf{\Sigma}|\mathbf{Y})$, where the first density on the right-hand side is a matrix normal density, and the second is an inverse-Wishart density. Specifically,

$$\Sigma|\mathbf{Y} \sim \text{inverse-Wishart}(\mathbf{S}_n^{-1}, \nu_0 + n), \quad \text{where } \mathbf{S}_n = \mathbf{S}_0 + \mathbf{Y}(\mathbf{I}_n + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y}^T; \tag{15}$$

$$\mathbf{B}|\Sigma, \mathbf{Y} \sim N_{m\times p}\left(\mathbf{M}_n, \Sigma, (\mathbf{I}_p + \mathbf{X}\mathbf{X}^T)^{-1}\right), \quad \text{where } \mathbf{M}_n = (\mathbf{M}_0 + \mathbf{Y}\mathbf{X}^T)(\mathbf{I}_p + \mathbf{X}\mathbf{X}^T)^{-1}. \tag{16}$$

Typically, n will be much larger than p, in which case $\mathbf{S}_n$ is more efficiently calculated as $\mathbf{S}_n = \mathbf{S}_0 + \mathbf{Y}(\mathbf{I}_n - \mathbf{X}^T(\mathbf{I} + \mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X})\mathbf{Y}^T$, which requires inversion of a $p \times p$ matrix rather than an $n \times n$ matrix.

Returning to the tensor regression model, for Bayesian analysis we parameterize the model as

$$\mathbf{Y} = \mathbf{X} \times \{\mathbf{B}_1, \ldots, \mathbf{B}_K\} + \tau\mathbf{E}$$

$$\mathbf{E} \sim N_{m_1\times\cdots\times m_K}(\mathbf{0}, \Sigma_1, \ldots, \Sigma_K),$$

where τ is an additional scale parameter that decouples the magnitude of the error variance from the prior variance of the $\mathbf{B}_k$'s (both of which would otherwise be determined by the Σ_k's). An inverse-gamma$(\eta_0/2, \eta_0\tau_0^2/2)$ prior distribution for τ^2 results in an inverse-gamma$([\eta_0 + m]/2, [\eta_0\tau_0^2 + ||\mathbf{Y} - \mathbf{X} \times \{\Sigma_1^{-1/2}\mathbf{B}_1, \ldots, \Sigma_K^{-1/2}\mathbf{B}_K\}||^2]/2)$ full conditional distribution. Based on these results, a Gibbs sampler with stationary distribution equal to the posterior distribution of $\{\mathbf{B}_1, \ldots, \mathbf{B}_K, \Sigma_1, \ldots, \Sigma_K, \tau^2\}$ can be constructed by iterating the following steps:

1. Iteratively for each $k = 1, \ldots, K$:
   (a) compute $\tilde{\mathbf{Y}} = \mathbf{Y}_{(k)}\mathbf{\Sigma}_{-k}^{-1/2}/\tau$ and $\tilde{\mathbf{X}} = \mathbf{X}_{(k)}\mathbf{B}_{-k}^T\mathbf{\Sigma}_{-k}^{-1/2}/\tau$;
   (b) simulate $(\Sigma_k, \mathbf{B}_k)$ from (15) and (16), replacing $\mathbf{Y}$ and $\mathbf{X}$ with $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$.
2. Simulate $\tau^2 \sim$ inverse-gamma$([\eta_0 + m]/2, [\eta_0\tau_0^2 + ||\mathbf{Y} - \mathbf{X} \times \{\Sigma_1^{-1/2}\mathbf{B}_1, \ldots, \Sigma_K^{-1/2}\mathbf{B}_K\}||^2]/2)$.

Parameter values simulated from this Markov chain can be used to make Monte Carlo approximations to posterior quantities of interest.

4 Analysis of longitudinal multirelational IR data

In this section we analyze weekly counts of four different action types between 25 countries over the ten and a half-year period from 2004 through the middle of 2014. These data were obtained from the ICEWS project (`http://www.lockheedmartin.com/us/products/W-ICEWS/iData.html`), which records time-stamped actions taken by one country with another country as the target. The 25 countries included in this analysis consist of the most active countries during the time period. The action types correspond to the four "quad classes" often used in international relations event analysis, which are negative material actions,which are positive material actions, negative verbal actions and positive verbal actions, and are denoted m-, m+, v-, v+, respectively. Examples of events that would fall into each of these four categories are: imposing a blockade (m–), providing humanitarian aid (m+), demanding a change in leadership (v–), and granting diplomatic recognition (v+). These data can be expressed as a $25 \times 25 \times 4 \times 543$-dimensional array $\mathbf{Y}$, where entry $y_{i_1,i_2,j,t}$ corresponds to the number of actions of type j, taken by country i_1 with country i_2 as the target, during week t. A normal quantile-quantile transformation was applied to each time-series corresponding to an actor-target-type triple, so that for each i_1, i_2, j, the empirical distribution of the time-series $\{y_{i_1,i_2,j,t} : t = 1, \ldots, 543\}$ is approximately standard normal.

In this section we consider several candidate models for these data, and present in detail the estimation results for the one providing the best fit in terms of predictive R^2. Perhaps the simplest modeling approach is to fit four separate bilinear regression models to each of the four action types, that is, to fit $\mathbf{Y}^{(j)} = \mathbf{X}^{(j)} \times \{\mathbf{B}_1^{(j)}, \mathbf{B}_2^{(j)}, \mathbf{I}\} + \mathbf{E}^{(j)}$, where $\mathbf{Y}^{(j)}$ is the $25 \times 25 \times 543$ array of between-country relations of type j, and $\mathbf{X}^{(j)}$ is a lagged version of $\mathbf{Y}^{(j)}$, for each $j \in \{1, \ldots, 4\}$. A competing model is the joint multilinear model $\mathbf{Y} = \mathbf{X} \times \{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \mathbf{I}\} + \mathbf{E}$, where $\mathbf{Y}$ is the complete $25 \times 25 \times 4 \times 543$ data array, and $\mathbf{B}_3$ is a 4×4 matrix of coefficients representing the effects of the different event types on one another. One possible advantage of using separate bilinear fits is that separate coefficient matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ can be estimated for each event type. Two disadvantages of this approach, as compared to the joint multilinear procedure, are that (1) the bilinear approach does not make use of one relation type to help predict another, and (2) if the coefficient matrices are sufficiently similar across event types, fitting them to be equal (as in the multilinear model) could improve estimation.

Inspection of the OLS estimates of $\{(\mathbf{B}_1^{(j)}, \mathbf{B}_2^{(j)}), j = 1, \ldots, 4\}$, indicated a high degree of sim-

model	material -	material +	verbal -	verbal +
separate additive	2.7 (2.3,3.6)	1.0 (0.4,1.5)	2.7 (1.9,3.4)	4.5 (3.7,5.5)
separate bilinear	7.9 (7.0,9.5)	2.9 (1.8,3.5)	7.8 (6.9,9.0)	12.3 (10.9,13.7)
joint multilinear	8.9 (7.9,10.3)	3.6 (2.9,4.4)	9.5 (8.5,11.1)	12.5 (11.5,13.7)
relational multilinear	11.0 (9.6,12.6)	4.5 (3.5,5.0)	11.5 (10.7,12.9)	13.6 (12.6,14.7)

Table 1: Results from the cross-validation study: Averages (and ranges in parentheses) of predictive R^2-values across the ten cross-validation datasets, for each model.

ilarity across the four action types, suggesting that the joint bilinear model may be appropriate. More formally, we compared the separate and joint models using a 10-fold cross validation study as described in the Introduction: For each of the 10 training and test set pairs, OLS estimates for each model were obtained using the algorithm described in Section 3.1. Average predictive R^2-values, as well as their ranges across the 10 test sets, are presented in Table 4 (for completeness, the table also includes results from separate fits of the additive model, described in the Introduction). The results indicate that, in terms of out-of-sample predictive performance for each action type, the benefits of the joint multilinear model outweigh the flexibility of having separate bilinear fits for each action type.

4.1 Reciprocity and transitivity

We now extend the explanatory tensor $\mathbf{X}$ to account for certain types of patterns often seen in relational data and social networks. One such pattern is the tendency for actions from one node i_1 to another node i_2 to be reciprocated over time, so that if $y_{i_1,i_2,j,t}$ is large, we may expect $y_{i_2,i_1,j,t+1}$ to be large as well. To estimate such an effect from the data, we add four "slices" to the tensor $\mathbf{X}$ along its third mode as follows: Redefine $\mathbf{X}$ so that $\mathbf{X} \in \mathbb{R}^{25\times25\times8\times543}$, with elements $x_{i_1,i_2,j,t} = y_{i_1,i_2,j,t-1}$ for $k \in \{1,\ldots,4\}$ as before, and $x_{i_1,i_2,j,t} = y_{i_2,i_1,j-4,t-1}$ for $j \in \{5,\ldots,8\}$. A multilinear regression model of the form $\mathbf{Y} = \mathbf{X} \times \{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \mathbf{I}\} + \mathbf{E}$ then has $\mathbf{B}_3 \in \mathbb{R}^{4\times8}$, the first four columns of which describe the effects of $y_{i_1,i_2,j_1,t-1}$ on $y_{i_1,i_2,j_2,t}$, and the last four columns of which describe the effects of $y_{i_2,i_1,j_1,t-1}$ on $y_{i_1,i_2,j_2,t}$, that is, the tendencies of actions of various types to be reciprocated by other actions at the next time point.

Other network effects can be accommodated similarly. One common pattern in network and

relational data is a type of third-order dependence known as transitivity, which describes how the simultaneous presence of relations between nodes i_1 and i_3, and between i_2 and i_3, might lead to a relation from i_1 to i_2. Based on this idea, we construct a transitivity predictors for each action type, and add them to the third mode of $\mathbf{X}$. Specifically, we let $x_{i_1,i_2,j-8,t} = \sum_{i_3}(y_{i_1,i_3,j,t} + y_{i_3,i_1,j,t})(y_{i_2,i_3,j,t} + y_{i_3,i_2,j,t})$ for each $j \in \{1,2,3,4\}$, so that now $\mathbf{X} \in \mathbb{R}^{25\times25\times12\times543}$, and the last four columns of the coefficient matrix $\mathbf{B}_3 \in \mathbb{R}^{4\times12}$ represent how the relations of nodes i_1 and i_2 with common targets leads to actions between i_1 and i_2 at the next time point. Note that this is a simplified measure of transitivity, in that the directions of the actions are not accounted for.

In what follows, we refer to this regression model as a relational multilinear regression, as it includes terms that allow estimation of patterns of reciprocity and transitivity that are often observed in relational data.

4.2 Longer-term dependence

Finally, we illustrate how to extend the relational multilinear model to account for longer-term longitudinal dependence. The appropriateness of doing so for these data is suggested by Figure 1: While the week-t observation is predictive of that at week $t+1$, some trends in the time series appear to persist beyond one week. In a separate exploratory analysis (not presented here), we considered using lagged monthly averages as predictors, along with the one-week lag observation. We found that after including a one week lag and a one month lag (the latter being an average of four weeks of previous data), the effects of lagged data from earlier months were minimal. For this reason, in what follows we model the data at time $t+1$ as a function of the data from the previous week t, as well as the average of the data from the previous month (weeks $t-1, t-2, t-3, t-4$).

One possibility for incorporating the one-month lagged data would be to add 12 more variables along the third mode of $\mathbf{X}$ as in the previous subsection. Each of these 12 variables would represent a monthly lagged version of the existing 12 variables along this mode. Such an approach would double the dimension of $\mathbf{B}_3$, and also make the interpretation of parameter values more cumbersome. A more parsimonious alternative is to assume separability of the effects of the two lag scales (weekly and monthly). Specifically, we reconstruct $\mathbf{X}$ to be a $25 \times 25 \times 12 \times 2 \times 543$-dimensional tensor, where $x_{i_1,i_2,j,1,t}$ corresponds to the previously existing entries of $\mathbf{X}$, and $x_{i_1,i_2,j,2,t}$ corresponds to the average of $x_{i_1,i_2,j,1,t-1}, \ldots, x_{i_1,i_2,j,1,t-4}$, that is, the average of the previous month's predictors.

Treating $\mathbf{Y}$ as a $25 \times 25 \times 4 \times 1 \times 543$ dimensional array, the multilinear regression model of $\mathbf{Y}$ on $\mathbf{X}$ is expressed as

$$\mathbf{Y} = \mathbf{X} \times \{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \mathbf{B}_4, \mathbf{I}\} + \mathbf{E}, \tag{17}$$

where $\mathbf{B}_4$ is a 1×2 matrix (or vector) that describes the relative magnitude of the effect of 1-week lagged data to that of the 1-month lagged data.

4.3 Parameter estimation and interpretation

We first compare the predictive performance of the least-squares estimates from the relational multilinear model (17) to the performance of the previously discussed models, using the 10-fold cross validation procedure described above. As shown in Table 4, model (17) outperforms all of the others in terms of predictive performance, and in fact outperformed its closest competitor, the joint multilinear model, on each of the 10 test datasets. These results suggest that this model is not overfitting relative to these simpler models.

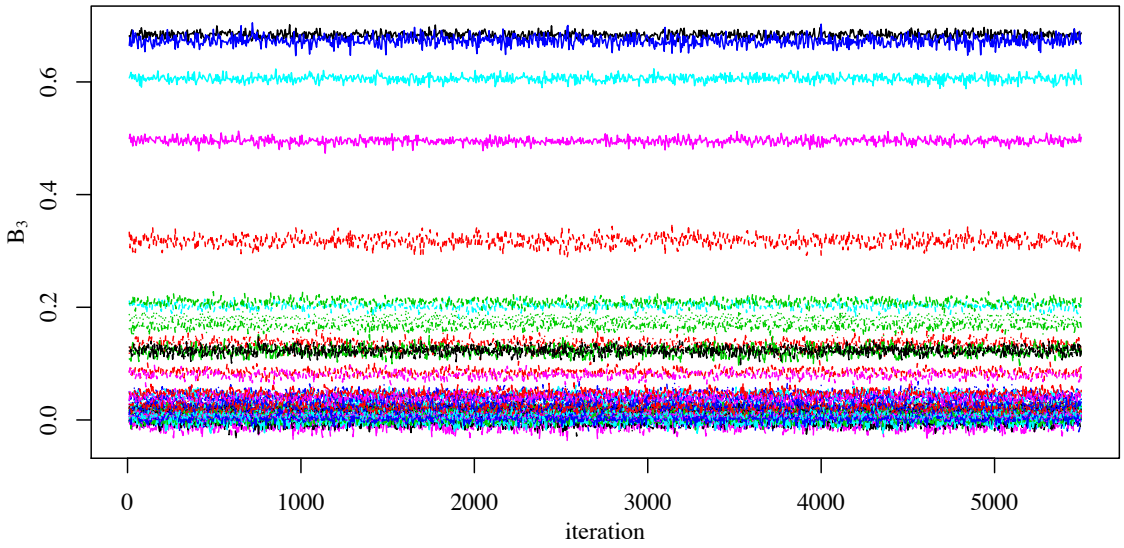

Figure 4: Values of the 48 entries of $\mathbf{B}_3$ simulated from the Gibbs sampler, using the least-squares estimates as starting values.

A more complete description of these data can be obtained via a Bayesian analysis of (17) using a separable model for residual covariance, as described in Section 3.2. Such an analysis accommodates residual dependence and provides an assessment of parameter uncertainty using, for

example, Bayesian confidence intervals. For this analysis, we used diffuse but proper priors, with $(\nu_0, \tau_0^2) = (1, 1)$, and for each mode k, $\mathbf{M}_{0k} = \mathbf{0}$, $\mathbf{S}_{0k} = \mathbf{I}_{m_k}$ and $\nu_{0k} = m_k + 1$. We ran four separate Gibbs samplers with which to approximate the posterior distribution: three with random starting values and one starting at the least squares estimates. Each sampler was run for 5500 iterations, allowing for 500 iterations for convergence to the stationary distribution. The sampler that started at the least squares estimates appeared to converge essentially immediately, whereas the samplers with random starting values appeared to take between about 50 and 250 iterations to arrive at the same part of the parameter space. Recalling that the separate magnitudes of the $\mathbf{B}_k$'s (and the Σ_k's) are not separately identifiable (as $\mathbf{F} \otimes \mathbf{G} = (c\mathbf{F}) \otimes (\mathbf{G}/c)$), we saved normalized versions of these parameters from the MCMC output. The normalization maintained a constant relative magnitude among $||\mathbf{B}_1||^2, ||\mathbf{B}_2||^2, ||\mathbf{B}_3||^2||\mathbf{B}_4||^2$, but the leaves the magnitude of $\mathbf{B}_4 \otimes \mathbf{B}_3 \otimes \mathbf{B}_2 \otimes \mathbf{B}_1$ unchanged as compared to doing no normalization. The Σ_k's were rescaled similarly. Further details on this post-processing of the MCMC output is available from the replication code available at the authors website. Mixing of the Gibbs sampler was very good: Figure 4 shows traceplots of the elements of $\mathbf{B}_3$, the coefficients describing the effects of the different action types, from the Gibbs sampler starting at the least squares estimates. After convergence, traceplots from the other Gibbs samplers looked nearly identical. For example, posterior mean estimates of the $\mathbf{B}_k$'s across the four different samplers differed from the across-sampler average by no more than 0.0003.

The posterior distributions of $\mathbf{B}_1$ and $\mathbf{B}_2$ are summarized in Figure 5. In each panel, nominally significant positive effects are shown by drawing a directed link from country i_1 to country i_2 if the lower 99% posterior quantile for entry i_1, i_2 of $\mathbf{B}_1$ or $\mathbf{B}_2$ is greater than zero. Not shown in the graph is that the lower 99% posterior quantile of each diagonal entry of $\mathbf{B}_1$ and $\mathbf{B}_2$ were positive. The 99th quantile was used instead of the 95th to ensure readability of the graphs. Also, there were very few negative coefficients of $\mathbf{B}_1$ and $\mathbf{B}_2$: only approximately 1% had their upper 99% posterior quantile below zero. Interpretation of these coefficients is illustrated with the following example: Letting i denote the index of Iran for example, the largest value of $\{b_{1,i,j} : j \in \{1, \ldots, 25\} \setminus \{i\}\}$ corresponds to that of Syria. These parameter estimates thus predict that actions of Syria towards a third party will increase the probability of actions of Iran towards that third party, at a future time point.

The posterior distribution of the $\mathbf{B}_3$ coefficients, which describe the main, reciprocal and tran-

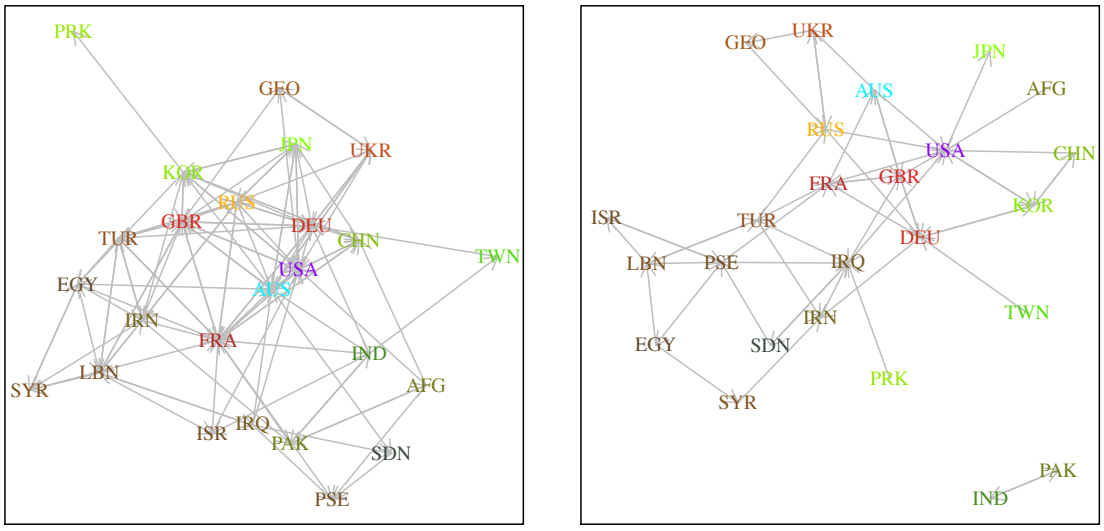

Figure 5: Summary of the posterior distributions of $\mathbf{B}_1$ and $\mathbf{B}_2$.

sitive effects of the four action types on future actions, is summarized in Table 4.3. This table gives posterior mean estimates of those coefficients of $\mathbf{B}_3$ for which zero is not included in their 95% posterior confidence interval. The first four columns of this matrix represent the direct effects of action variable j_1 from i_1 to i_2 on the future value of action variable j_2 from i_1 to i_2, for $j_1, j_2 \in \{1, 2, 3, 4\}$. Not surprisingly, the largest estimated coefficients are along the diagonal, indicating that the strongest predictor of action variable j_1 is the previous value of this variable. Other "significant" coefficients include effects of actions on actions of a common valence: The second most important predictors of "m–", "m+" and "v–" are "v–", "v+" and "m–", respectively. The variable "v+" (verbal positive) represents an exception to this pattern. However, many of the actions that fall into this category are bilateral negotiations and diplomatic resolutions that often occur as a result of diplomatic disputes that are in the "verbal negative" category. The second four columns of $\mathbf{B}_3$ represent the reciprocal effects of action from i_2 to i_1 on future actions from i_1 to i_2. Similar to the direct effects, the largest coefficients for three of the four action types are along the main diagonal. The exception is the "m+" category (material positive), for which the 95% posterior confidence interval contained zero. This reflects the fact that this category is largely comprised of actions that involve the provision of economic, military and humanitarian aid. Such actions are typically initiated by wealthy countries with less-developed countries as the target, and

outcome	predictor											
	direct				reciprocal				transitive			
	m–	m+	v–	v+	m–	m+	v–	v+	m–	m+	v–	v+
m–	0.68	0.04	0.17		0.20	0.02	0.12	0.02	0.02			
m+	0.09	0.50	0.04	0.13	0.04		0.02				0.04	
v–	0.18		0.61	0.12	0.13	0.03	0.21			0.01	0.02	
v+	0.05	0.03	0.08	0.67	0.05	0.02	0.03	0.32		0.02	0.02	

Table 2: Summary of the posterior distribution of $\mathbf{B}_3$.

so are often unreciprocated. The final four columns of $\mathbf{B}_3$ represent the transitivity effects. While the results indicate some evidence of transitivity, the magnitude of such effects is small compared to the direct and reciprocal effects.

The matrix $\mathbf{B}_4$ consists of two coefficients representing the multiplicative effects of one-week lagged data as compared to one-month lagged data. Both coefficients of $\mathbf{B}_4$ were positive in every iteration of the Gibbs sampler, and posterior distribution of the ratio of the former coefficient to the latter had a mean of 1.98 and a 95% posterior confidence interval of (1.94, 2.03), indicating the effect of the one-week lagged data was roughly twice that of the one-month lagged data.

5 Discussion

This article has developed a multilinear tensor regression (MLTR) model for regressing a tensor of outcome data on a tensor of explanatory variables. The regression coefficients in such a model are multiplicative in the parameters, rather than additive as in the more standard class of linear regression models. As was shown in an example analysis of longitudinal relational data, in some cases a multiplicative effects model provides a better representation of the data than a comparable but more standard additive effects model. Additionally, it was shown how the MLTR model can be extended to estimate a variety of network effects, such as reciprocity and transitivity, as well as temporal effects of lagged data beyond that in a first-order autoregressive model.

Like any regression model, the multilinear tensor regression model could be extended or modified in many different ways. Of particular use would be an extension to accommodate data that is

binary, ordinal or generally of a form for which a least-squares criteria/normal error model would be inappropriate. One possible approach for doing this would be via various link functions, as is done with generalized linear models. An alternative approach would be to use a semiparametric transformation model, in which the observed data are modeled as being a nondecreasing function of a latent tensor that follows a normal multilinear tensor regression model. However, for some data types, such as if $\mathbf{Y}$ were a sparse binary tensor, there might not be enough information in the data to provide stable parameter estimates: Although a MLTR model where $\mathrm{E}[\mathbf{y}] = (\mathbf{B}_K \otimes \cdots \otimes \mathbf{B}_1)\mathbf{x}$ constitutes a great simplification as compared to a full model where $\mathrm{E}[\mathbf{y}] = \Theta\mathbf{x}$ with Θ unstructured, it still has a large number of parameters. One possible remedy in cases with limited data information is to use sparsity-inducing penalties, such as L_1 penalties on the $\mathbf{B}_k$'s. This would have to be done with some care, as the overall scale of each $\mathbf{B}_k$ matrix is not identifiable.

Replication code for the results in Section 4 is available at the authors website: `www.stat.washington.edu/~hoff`. This research was supported by NIH grant R01HD067509. The author thanks Michael Ward for guidance with the data.

A Proofs

Proof of Proposition 1. Let $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ be a pseudotrue parameter for $(\mathbf{A}, \mathbf{B})$. If $\tilde{\mathbf{B}} = \mathbf{0}$, then setting $\tilde{a}_{i,j}$, the i,jth element of $\tilde{\mathbf{A}}$, to zero does not change the asymptotic criterion function, and so $\tilde{a}_{i,j} = 0$ is a pseudotrue value. If $\tilde{\mathbf{B}} \neq \mathbf{0}$, then $\mathrm{E}[\mathbf{X}\tilde{\mathbf{B}}^T\tilde{\mathbf{B}}\mathbf{X}^T]$ is invertible (assuming, for example, the distribution of $\mathbf{X}$ has full support on $\mathbb{R}^{p_1 \times p_2}$), and the pseudotrue parameter $\tilde{\mathbf{A}}$ will satisfy

$$\tilde{\mathbf{A}} = \mathrm{E}[\mathbf{Y}\tilde{\mathbf{B}}\mathbf{X}^T]\mathrm{E}[\mathbf{X}\tilde{\mathbf{B}}^T\tilde{\mathbf{B}}\mathbf{X}^T]^{-1}. \tag{18}$$

Let $\mathbf{y}_i$ and $\mathbf{x}_j$ be rows i and j of $\mathbf{Y}$ and $\mathbf{X}$, respectively. If $\mathbf{x}_j$ is mean zero and independent of the other rows of $\mathbf{X}$, so that $\mathrm{E}[\mathbf{x}_j\mathbf{x}_k^T] = \mathbf{0}$, then the i,jth element of $\tilde{\mathbf{A}}$ is given by

$$\tilde{a}_{i,j} = \mathrm{E}[\mathbf{y}_i^T\tilde{\mathbf{B}}\mathbf{x}_j]/\mathrm{E}[\mathbf{x}_j^T\tilde{\mathbf{B}}^T\tilde{\mathbf{B}}\mathbf{x}_j].$$

If $\mathbf{y}_i$ is uncorrelated with $\mathbf{x}_j$, then the numerator, and the coefficient is zero. □

Proof of Proposition 2. As in the proof of Proposition 1, if $\mathrm{E}[\mathbf{X}\tilde{\mathbf{B}}^T\tilde{\mathbf{B}}\mathbf{X}^T]$ is invertible, then the pseudotrue parameter is given by $\tilde{\mathbf{A}}$ in (18). Under the assumption that $\mathrm{E}[\mathbf{x}\mathbf{x}^T] = \Omega \otimes \Psi$, we have

$\mathrm{E}[\mathbf{X}\tilde{\mathbf{B}}^T\tilde{\mathbf{B}}\mathbf{X}^T] = c\Psi$ with $c = \mathrm{tr}(\Omega\mathbf{B}^T\mathbf{B})$, and

$$\mathrm{E}[\mathbf{Y}\mathbf{B}\mathbf{X}^T] = (\mathbf{1}_{m_2}^T \otimes \mathbf{I}_{m_1})[\Sigma_{yx} \circ (\mathbf{B} \otimes \mathbf{1}\mathbf{1}^T)](\mathbf{1}_{p_2} \otimes \mathbf{I}_{p_1}),$$

where "$\circ$" is the Hadamard (elementwise) product. Under the assumption of the proposition, $\Sigma_{yx} = \mathrm{E}[\mathbf{y}\mathbf{x}^T] = \mathrm{E}[\mathrm{E}[\mathbf{y}|\mathbf{x}]\mathbf{x}^T] = \Theta\mathrm{E}[\mathbf{x}\mathbf{x}^T] = \Theta(\Omega \otimes \mathbf{\Psi})$, which can be expressed as

$$\begin{pmatrix} \Theta_{1,1} & \cdots & \Theta_{1,p_2} \\ \vdots & & \vdots \\ \Theta_{m_2,1} & \cdots & \Theta_{m_2,p_2} \end{pmatrix} \begin{pmatrix} \omega_{1,1}\Psi & \cdots & \omega_{1,p_2}\Psi \\ \vdots & & \vdots \\ \omega_{p_2,1}\Psi & \cdots & \omega_{p_2,p_2}\Psi \end{pmatrix} = \begin{pmatrix} \sum_{j_2=1}^{p_2} \omega_{j_2,1}\Theta_{1,j_2}\Psi & \cdots & \sum_{j_2=1}^{p_2} \omega_{j_2,p_2}\Theta_{1,j_2}\Psi \\ \vdots & & \vdots \\ \sum_{j_2=1}^{p_2} \omega_{j_2,1}\Theta_{m_2,j_2}\Psi & \cdots & \sum_{j_2=1}^{p_2} \omega_{j_2,p_2}\Theta_{m_2,j_2}\Psi \end{pmatrix},$$

where Θ_{i_2,j_2} is the $m_1 \times p_1$ matrix describing the effects of the column j_2 of $\mathbf{X}$ on column i_2 of $\mathbf{Y}$. The expectation $\mathrm{E}[\mathbf{Y}\mathbf{B}\mathbf{X}^T]$ is obtained by multiplying each block of the form $\sum_{j_2=1}^{p_2} \omega_{j_2,j_2'}\Theta_{i_2,j_2}\Psi$ by element i_2, j_2' of $\mathbf{B}$, and summing the blocks. This results in an $m_1 \times p_1$ matrix given by

$$\mathrm{E}[\mathbf{Y}\mathbf{B}\mathbf{X}^T] = \left(\sum_{i_2=1}^{m_2} \sum_{j_2=1}^{p_2} \left(\sum_{j_2'=1}^{p_2} \omega_{j_2,j_2'} b_{i_2,j_2'} \right) \Theta_{i_2,j_2} \right) \Psi.$$

Multiplying by the inverse of $\mathrm{E}[\mathbf{X}\tilde{\mathbf{B}}^T\tilde{\mathbf{B}}\mathbf{X}^T]$ on the left gives the pseudotrue parameter $\mathbf{A}$ as

$$\tilde{\mathbf{A}} = c^{-1} \sum_{i_2=1}^{m_2} \sum_{j_2=1}^{p_2} \left(\sum_{j_2'=1}^{p_2} \omega_{j_2,j_2'} b_{i_2,j_2'} \right) \Theta_{i_2,j_2}.$$

The effects of the jth row of $\mathbf{X}$ on the ith row of $\mathbf{Y}$ consist of the i,jth elements of the Θ_{i_2,j_2}'s. These are all zero under the assumption of the proposition, and thus so is $\tilde{a}_{i,j}$. □

References

Akdemir, D. and A. K. Gupta (2011). Array variate random variables with multiway Kronecker delta covariance matrix structure. *J. Algebr. Stat. 2*(1), 98–113.

Basu, S., J. Dunagan, K. Duh, and K.-K. Muniswamy-Reddy (2012). Blr-d: applying bilinear logistic regression to factored diagnosis problems. *ACM SIGOPS Operating Systems Review 45*(3), 31–38.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications 21*(4), 1253–1278.

Durante, D. and D. B. Dunson (2013). Nonparametric bayes dynamic modeling of relational data.

Fu, W., L. Song, and E. P. Xing (2009). Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th annual international conference on machine learning*, pp. 329–336. ACM.

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

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

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Rev. 51*(3), 455–500.

Li, X., H. Zhou, and L. Li (2013). Tucker tensor regression and neuroimaging analysis.

Luenberger, D. G. and Y. Ye (2008). *Linear and nonlinear programming* (Third ed.). International Series in Operations Research & Management Science, 116. Springer, New York.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate analysis*. London: Academic Press [Harcourt Brace Jovanovich Publishers]. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

Potthoff, R. F. and S. N. Roy (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika 51*, 313–326.

Shi, J. V., Y. Xu, and R. G. Baraniuk (2014). Sparse bilinear logistic regression. *arXiv preprint arXiv:1404.4104*.

Snijders, T., C. Steglich, and M. Schweinberger (2007). *Modeling the coevolution of networks and behavior*. na.

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological methodology 31*(1), 361–395.

Srivastava, M. S., T. von Rosen, and D. von Rosen (2009). Estimation and testing in general multivariate linear models with Kronecker product covariance structure. *Sankhyā 71*(2, Ser. A), 137–163.

Tucker, L. R. (1964). The extension of factor analysis to three-dimensional matrices. *Contributions to mathematical psychology*, 109–127.

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

Ward, M. D. and P. D. Hoff (2007). Persistent patterns of international commerce. *Journal of Peace Research 44*(2), 157–175.

Westveld, A. H. and P. D. Hoff (2011). A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *Ann. Appl. Stat. 5*(2A), 843–872.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc. 76*(374), 419–433.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Xing, E. P., W. Fu, and L. Song (2010). A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics 4*(2), 535–566.