Hierarchical multilinear models for multiway data

Peter D. Hoff *

March 8, 2010

Abstract

Reduced-rank decompositions provide descriptions of the variation among the elements of a matrix or array. In such decompositions, the elements of an array are expressed as products of low-dimensional latent factors. This article presents a model-based version of such a decomposition, extending the scope of reduced rank methods to accommodate a variety of data types such as longitudinal social networks and continuous multivariate data that is cross-classified by categorical variables. The proposed model-based approach is hierarchical, in that the latent factors corresponding to a given dimension of the array are not *a priori* independent, but exchangeable. Such a hierarchical approach allows more flexibility in the types of patterns that can be represented.

Some key words: Bayesian, multiplicative model, PARAFAC, regularization, shrinkage.

1 Introduction

Matrix-valued data are prevalent in many scientific disciplines. Studies in social and health sciences often gather social network data that can be represented by square, binary matrices with undefined diagonals. Numerical results from gene expression studies are recorded in matrices with rows representing tissue samples and columns representing genes. Analysis of stock market returns involves data matrices with rows representing stocks and columns representing time. With such data there are often dependencies among both the rows and the columns of the data matrices, and so the standard tools of multivariate analysis, in which patterns along one dimension of the data

*Departments of Statistics and Biostatistics , University of Washington, Seattle, Washington 98195-4322. Web: `http://www.stat.washington.edu/hoff/`. This work was partially supported by NSF grant SES-0631531.

matrix are thought of as i.i.d., may be inadequate for data analysis purposes. As an alternative to the i.i.d. paradigm, patterns of row and column variation in matrix-valued data are often described with reduced-rank matrix decompositions and models. For example, the i,jth entry of an $m_1 \times m_2$ matrix might be expressed as $y_{i,j} = \langle \mathbf{u}_i, \mathbf{v}_j \rangle + \epsilon_{i,j}$, where the heterogeneity among a set of low-dimensional vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_{m_1}\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_{m_2}\}$ is used to represent heterogeneity attributable to the row and column objects respectively. Such models can be described as being bilinear, as the expectation of $y_{i,j}$ is a bilinear function of the parameters. These models are related to biplots [Gabriel, 1971], bilinear regression [Gabriel, 1998] and the singular value decomposition (SVD).

In more complex situations the data take the form of a multidimensional array instead of a matrix. For example, temporal variation in a social network over a discrete set of time points may be represented by a three-way array $\mathbf{Y} = \{y_{i,j,t}\}$, where $y_{i,j,t}$ describes the relationship between nodes i and j at time t. Similarly, gene expression data gathered under a variety of experimental conditions, or multiple variables measured on a set of companies over time are also examples of array-valued or multiway data. Surveys of multiway data analysis include Coppi and Bolasco [1989] and Kroonenberg [2008]. The July-August 2009 issue of the Journal of Chemometrics was dedicated to Richard Harshman, one of the founders of three-way data analysis. Harshman [Harshman, 1970, Harshman and Lundy, 1984] developed a three-way generalization of the SVD known as "parallel factor analysis", or PARAFAC, that has become one of the primary methods of multiway data analysis. The generalization is as follows: The SVD represents the i,jth element of a rank-R matrix $\mathbf{A}$ as $a_{i,j} = \langle \mathbf{u}_i, \mathbf{v}_j \rangle \equiv \sum_{r=1}^{R} u_{i,r} v_{j,r}$. For a three-way array, a PARAFAC decomposition represents the i,j,kth element as $a_{i,j,k} = \langle \mathbf{u}_i, \mathbf{v}_j, \mathbf{w}_k \rangle = \sum_{r=1}^{R} u_{i,r} v_{j,r} w_{k,r}$. Kruskal [1976, 1977] related such decompositions to a precise definition of rank for three-way arrays, in which the rank is the smallest integer R for which the above representation holds. The generalization to arbitrary dimensions is straightforward: A K-dimensional array of rank R is one in which the elements can be expressed as a multilinear function of R-dimensional factors. A compact review of these results and others appears in Kruskal [1989].

While the area of multiway data analysis has been active, most of the focus has been on algorithms for finding least-squares solutions, pre- and post-processing of results, and interpretation of the least-squares parameters. Little has been done in terms of incorporating multilinear representations into statistical models. One exception is the work of Vega-Montoto and Wentzell [2003]

and Vega-Montoto et al. [2005], who develop algorithms for finding maximum likelihood solutions for situations with heteroscedastic or correlated error terms. However, these algorithms assume the error variance is known.

This article develops a hierarchical multilinear model for incorporation into a variety of non-standard multiway data analysis situations, and presents a Bayesian approach for parameter estimation. The motivation is twofold: First, multilinear array representations can involve a large number of parameters. Overfitting of the model can be ameliorated by using shrinkage estimators provided by a Bayesian approach. In particular, a hierarchical Bayesian approach can be used to provide shrinkage patterns that are based primarily on the observed data, rather than relying heavily on a fixed prior distribution. The second motivation is that Bayesian approaches and MCMC estimation methods allow one to incorporate the basic multilinear representation into models for complex data that might involve additional dependence structures or discrete data.

After presenting the hierarchical multilinear model and Bayesian methods for estimation in Sections 2 and 3, a small simulation study is presented in Section 4 to compare mean squared errors of three different parameter estimation methods: least-squares, a simple non-hierarchical Bayesian approach and a Bayesian hierarchical approach. The Bayes estimators are found to outperform the least-squares estimator, with the hierarchical Bayes procedure having the best performance. Also considered is the performance of the estimators when the rank of the model is misspecified. In this situation, the least-squares and non-hierarchical Bayes procedures increasingly overfit the data as the rank is increased, while the hierarchical Bayes procedure is robust to rank misspecification.

Sections 5 and 6 give examples in which it is useful to embed a multilinear model within a larger model for observed data. Section 5 considers estimation of a multivariate mean $\mathrm{E}[\mathbf{y}_{\mathrm{x}}] = \boldsymbol{\mu}_{\mathrm{x}}$ for each possible value of a vector of categorical variables $\mathbf{x}$. Often the number of observations per level of $\mathbf{x}$ is small and varies from level to level. A hierarchical model for the mean, $\boldsymbol{\mu}_{\mathrm{x}} \sim$ multivariate normal$(\boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\Sigma})$, allows for consistent estimation of each $\boldsymbol{\mu}_{\mathrm{x}}$ but shrinkage towards $\boldsymbol{\beta}_{\mathrm{x}}$ when the sample size is small. The values $\mathbf{B} = \{\boldsymbol{\beta}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\}$ can be represented as a multiway array, and a reduced rank multilinear model for $\mathbf{B}$ allows for the modeling of non-additive effects of $\mathbf{x}$ with a relatively small number of parameters.

Section 6 presents an analysis of international cooperation and conflict during the cold war. The data consist of a three-way array with element $y_{i,j,t}$ representing the relationship between

countries i and j in year t. Several features of these data make existing tools from multiway data analysis inappropriate, one being that the data are ordinal. The range of the data includes the integers from -5 to 2, indicating different levels of military cooperation or conflict. Assuming that the $y_{i,j,k}$'s are normally distributed or even continuous would be inappropriate. However, using the tools developed in this article it is reasonably straightforward to embed a multilinear representation within an ordered probit model for these data. A discussion of the results and directions for future research follows in Section 7.

2 Reduced rank models for array data

In this section we review the reduced rank model and an alternating least-squares(ALS) procedure for parameter estimation. For a review of the properties, limitations and alternatives to ALS, see Tomasi and Bro [2006] and Chapter 5 of Kroonenberg [2008].

2.1 Rank and factor representations for arrays

Given an $m_1 \times m_2$ data matrix $\mathbf{Y}$ it is often desirable to separate out the "main features" of $\mathbf{Y}$ from the "patternless noise." This motivates a model of the form $\mathbf{Y} = \boldsymbol{\Theta} + \mathbf{E}$, where $\boldsymbol{\Theta}$ is to be estimated from the data. Interpreting "main features" as those that can be well-approximated by a low-rank matrix, the rank of $\boldsymbol{\Theta}$ is usually taken to be some value $R < m_1 \wedge m_2$. The rank of a matrix $\boldsymbol{\Theta}$ can be defined as the smallest integer R such that there exists matrices $\mathbf{U} \in \mathbb{R}^{m_1 \times R}$ and $\mathbf{V} \in \mathbb{R}^{m_2 \times R}$ such that

$$\boldsymbol{\Theta} \;=\; \sum_{r=1}^{R} \mathbf{u}_r \otimes \mathbf{v}_r = \mathbf{U}\mathbf{V}^T \;, \text{ or equivalently, } \theta_{i,j} = \sum_{r=1}^{R} u_{i,r} v_{j,r} = \langle \mathbf{u}_i, \mathbf{v}_j \rangle,$$

where $\mathbf{u}_r$ is the rth column of $\mathbf{U}$ in the first equation and $\mathbf{u}_i$ is the ith row of $\mathbf{U}$ in the second. Variation among the rows of $\mathbf{U}$ represents the heterogeneity in $\boldsymbol{\Theta}$ attributable to variation in the row objects, and similarly variation among the rows of $\mathbf{V}$ represents heterogeneity attributable to the column objects.

A K-order multiway array $\mathbf{Y}$ with dimension $m_1 \times \cdots \times m_K$ has elements $\{y_{i_1,\ldots,i_K} : i_k \in \{1,\ldots,m_k\}\}$. As with a matrix, we may define a model for a K-order array as $\mathbf{Y} = \boldsymbol{\Theta} + \mathbf{E}$, where $\mathbf{E}$ is an array of uncorrelated, mean-zero noise and $\boldsymbol{\Theta}$ is a reduced rank array to be estimated.

Following Kruskal [1976] and Kruskal [1977], the rank of a K-order array $\boldsymbol{\Theta}$ is simply the smallest integer R such that there exist matrices $\{\mathbf{U}^{(k)} \in \mathbb{R}^{m_k \times R}, k = 1, \ldots, K\}$, such that

$$\begin{aligned} \boldsymbol{\Theta} &= \sum_{r=1}^{R} \mathbf{u}_r^{(1)} \otimes \cdots \otimes \mathbf{u}_r^{(K)} \equiv \langle \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)} \rangle \ , \quad \text{or equivalently} \\ \theta_{i_1,\ldots,i_K} &= \sum_{r=1}^{R} u_{i_1,r}^{(1)} \times \cdots \times u_{i_K,r}^{(K)} \equiv \langle \mathbf{u}_{i_1}^{(1)}, \cdots, \mathbf{u}_{i_K}^{(K)} \rangle, \end{aligned}$$

where $\mathbf{u}_r^{(k)}$ is the rth column of $\mathbf{U}^{(k)}$ in the first equation and $\mathbf{u}_i^{(k)}$ is the ith row of $\mathbf{U}^{(k)}$ in the second. As in the matrix case, variation among the rows of $\mathbf{U}^{(k)}$ represents heterogeneity attributable to the kth set of objects, that is, the kth mode of the array.

2.2 Least squares estimation

In the matrix case the least squares estimate of $\boldsymbol{\Theta} = \mathbf{U}\mathbf{V}^T$ (also the MLE assuming normal, i.i.d. errors) can be obtained from the first R components of the singular value decomposition of $\mathbf{Y}$. For arrays of higher order, only iterative methods of estimation are available. Perhaps the simplest method of parameter estimation is the alternating least squares algorithm (ALS), in which factors corresponding to a given mode are updated to minimize the residual sums of squares given the current values for the other modes. In this subsection we review the relevant calculations for ALS, which will also be useful for Bayesian estimation in the next section.

Estimation for a three-way model: We begin with an three-way array so that the main ideas can be understood with a minimal amount of notational complexity. Let $\mathbf{Y}$ be a three-way array modeled as $y_{i,j,k} = \langle \mathbf{u}_i, \mathbf{v}_j, \mathbf{w}_k \rangle + \epsilon_{i,j,k}$, with $\{\epsilon_{i,j,k}\} \sim$ i.i.d. normal$(0, \sigma^2)$. We can write

$$\begin{aligned} \mathbf{y}_{i,j,\cdot} &= \mathbf{W}(\mathbf{u}_i \circ \mathbf{v}_j) + \boldsymbol{\epsilon}_{i,j,\cdot} \\ \mathbf{y}_{i,\cdot,k} &= \mathbf{V}(\mathbf{u}_i \circ \mathbf{w}_k) + \boldsymbol{\epsilon}_{i,\cdot,k} \\ \mathbf{y}_{\cdot,j,k} &= \mathbf{U}(\mathbf{v}_j \circ \mathbf{w}_k) + \boldsymbol{\epsilon}_{\cdot,j,k}, \end{aligned}$$

where $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ are $m_1 \times R$, $m_2 \times R$ and $m_3 \times R$ matrices respectively, $\mathbf{u}_i$, $\mathbf{v}_j$, $\mathbf{w}_k$ are rows of these matrices, $\mathbf{y}_{i,j,\cdot}$, $\mathbf{y}_{i,\cdot,k}$, $\mathbf{y}_{\cdot,j,k}$ are vectors of length m_1, m_2 and m_3, and "$\circ$" denotes the Hadamard product (elementwise multiplication). Some matrix algebra and careful summation shows that, as

a function of $\mathbf{U}$, $p(\mathbf{Y}|\mathbf{U},\mathbf{V},\mathbf{W})$ can be written

$$\begin{aligned} p(\mathbf{Y}|\mathbf{U},\mathbf{V},\mathbf{W}) &\propto \text{etr}(\mathbf{U}^T\mathbf{L}/\sigma^2 - \mathbf{U}^T\mathbf{U}\mathbf{Q}/[2\sigma^2]) \text{ , where} \\ \mathbf{Q} &= (\mathbf{V}^T\mathbf{V}) \circ (\mathbf{W}^T\mathbf{W}) \text{ and} \\ \mathbf{L} &= \sum_{j,k} \mathbf{y}_{\cdot,j,k} \otimes (\mathbf{v}_j \circ \mathbf{w}_k) \text{ .} \end{aligned} \tag{1}$$

With $\mathbf{V}$ and $\mathbf{W}$ fixed, the conditional MLE and least-squares estimate of $\mathbf{U}$ is given by $\hat{\mathbf{U}} = \mathbf{L}\mathbf{Q}^{-1}$. The ALS procedure is to iteratively replace a current value of $\mathbf{U}$ with its conditional least-squares estimate, then replace $\mathbf{V}$ and $\mathbf{W}$ similarly. This procedure is then iterated until a convergence criterion has been met.

Estimation for a K-way model: Now suppose $\mathbf{Y}$ is an $m_1 \times \cdots \times m_K$ array. Let $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}$ be the matrices of factors for the K modes, so that $\mathbf{U}^{(k)}$ is an $m_k \times R$ matrix. The basic results from the three-way model carry over as follows: Let $\mathbf{y}_{\mathbf{i}_1} = (y_{1,i_2,\ldots,i_K}, \ldots, y_{n_1,i_2,\ldots,i_m})$ be a "fiber" along the first dimension of the array. Then we can write $\mathbf{y}_{\mathbf{i}_1} = \mathbf{U}^{(1)}(\mathbf{u}^{(2)}_{i_2} \circ \mathbf{u}^{(3)}_{i_3} \circ \cdots \circ \mathbf{u}^{(K)}_{i_K}) + \epsilon_{\mathbf{i}_1}$. Similar to the three-mode case, as a function of $\mathbf{U}^{(1)}$, $p(\mathbf{Y}|\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(K)})$ can be written

$$\begin{aligned} p(\mathbf{Y}|\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(K)}) &\propto \text{etr}({\mathbf{U}^{(1)}}^T\mathbf{L}/\sigma^2 - {\mathbf{U}^{(1)}}^T\mathbf{U}^{(1)}\mathbf{Q}/[2\sigma^2]) \text{ , where} \\ \mathbf{Q} &= ({\mathbf{U}^{(2)}}^T\mathbf{U}^{(2)}) \circ \cdots \circ ({\mathbf{U}^{(K)}}^T\mathbf{U}^{(K)}) \text{ and} \\ \mathbf{L} &= \sum_{i_2,\ldots,i_m} \mathbf{y}_{\mathbf{i}_1} \otimes (\mathbf{u}^{(2)} \circ \cdots \circ \mathbf{u}^{(k)}) \text{ .} \end{aligned} \tag{2}$$

The conditional MLE and least squares estimator of $\mathbf{U}^{(1)}$ given the factor values for the other modes is thus $\hat{\mathbf{U}}^{(1)} = \mathbf{L}\mathbf{Q}^{-1}$. As with three-way data, the ALS procedure is to iteratively replace the factors matrices with their conditional least-squares estimates until convergence.

3 Bayes and hierarchical Bayes estimation

Compared to least-squares or maximum likelihood methods, Bayesian procedures often provide stable estimation in high-dimensional problems due to regularization via the prior distribution. Using conjugate prior distributions, this section provides a Gibbs sampling scheme that approximates the posterior distribution $p(\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(K)},\sigma^2|\mathbf{Y})$, and by extension, an approximation to the posterior distribution of $\mathbf{\Theta} = \langle \mathbf{U}^{(1)},\ldots,\mathbf{U}^{(K)} \rangle$. The posterior expectation of $\mathbf{\Theta}$ can be used as a Bayesian estimate of the main features of the data array.

3.1 A basic Gibbs sampler

Let the prior distribution for $\mathbf{U}^{(k)}$ be such that that the rows of $\mathbf{U}^{(k)}$ are i.i.d. multivariate normal$(\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k)$ or equivalently, $\mathbf{U}^{(k)} \sim$ matrix normal$(\mathbf{M}_k = \mathbf{1}\boldsymbol{\mu}_k^T, \boldsymbol{\Psi}_k, \mathbf{I})$ with density

$$\begin{aligned} p(\mathbf{U}^{(k)}) &\propto \text{etr}(-(\mathbf{U}^{(k)} - \mathbf{M}_k)^T(\mathbf{U}^{(k)} - \mathbf{M}_k)\boldsymbol{\Psi}_k^{-1}/2) \\ &\propto \text{etr}(\mathbf{U}^{(k)T}\mathbf{M}_k\boldsymbol{\Psi}_k^{-1} - \mathbf{U}^{(k)T}\mathbf{U}^{(k)}\boldsymbol{\Psi}_k^{-1}/2). \end{aligned}$$

Combining this with the likelihood from Equation 2, it follows that if $\mathbf{U}^{(1)} \sim$ matrix normal$(\mathbf{M}_1, \boldsymbol{\Psi}_1, \mathbf{I})$ *a priori*, then the full conditional distribution is also matrix normal with density

$$\begin{aligned} p(\mathbf{U}^{(1)}|\mathbf{Y}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(K)}) &\propto \text{etr}(-(\mathbf{U}^{(1)} - \tilde{\mathbf{M}}_1)^T(\mathbf{U}^{(1)} - \tilde{\mathbf{M}}_1)/2)\tilde{\boldsymbol{\Psi}}_1^{-1}) \\ \tilde{\boldsymbol{\Psi}}_1 &= (\mathbf{Q}/\sigma^2 + \boldsymbol{\Psi}_1^{-1})^{-1} \\ \tilde{\mathbf{M}}_1 &= (\mathbf{L}/\sigma^2 + \mathbf{M}_1\boldsymbol{\Psi}_1^{-1})\tilde{\boldsymbol{\Psi}}_1. \end{aligned}$$

Full conditional distributions for $\mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(m)}$ are derived analogously. Using a conjugate inverse-gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$ prior distribution for σ^2 results in an inverse-gamma(a, b) full conditional distribution where $a = (\nu_0 + \prod_k m_k)/2$ and $b = (\nu_0\sigma_0^2 + ||\mathbf{Y} - \langle \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)} \rangle||^2)/2$.

A Markov chain Monte Carlo approximation to $p(\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}, \sigma^2|\mathbf{Y})$ can be made by iteratively sampling each unknown quantity from its full conditional distribution. This generates a Markov chain, samples from which converge in distribution to $p(\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}, \sigma^2|\mathbf{Y})$. However, it would be inappropriate to estimate $\mathbf{U}^{(k)}$ by its posterior mean $\hat{\mathbf{U}}^{(k)}$, or $\boldsymbol{\Theta}$ with $\langle \hat{\mathbf{U}}^{(1)}, \ldots, \hat{\mathbf{U}}^{(K)} \rangle$, as the values of the latent factors are not separately identifiable. For example, the likelihood is invariant to joint permutations and complementary rescalings of the columns of the $\mathbf{U}^{(k)}$'s (see Kruskal [1989] for a discussion of the uniqueness of reduced-rank array decompositions). Instead, the posterior mean estimate $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$, obtained from the average of $\langle \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)} \rangle$ over iterations of the Markov chain, can be used as a point estimate of $\boldsymbol{\Theta}$. If desired, point estimates of the $\mathbf{U}^{(k)}$'s can then be obtained from a rank-R least-squares approximation of $\hat{\boldsymbol{\Theta}}$.

3.2 Hierarchical modeling of factors

Rarely will we have detailed prior knowledge of an appropriate mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Psi}_k$ for each factor matrix $\mathbf{U}^{(k)}$. Absent these, we may consider a simple "weak" prior distribution such

as $\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_{m_k}^{(k)} \sim$ i.i.d. multivariate normal$(\mathbf{0}, \tau^2\mathbf{I})$, where τ^2 is large. However, doing so would ignore patterns of heterogeneity in the data that could improve parameter estimation.

Recall that the factors represent variance among the elements of the data array $\mathbf{Y}$ that can be attributed to heterogeneity within the various modes. To illustrate, consider three mode data in which the first mode represents a large number of experimental units and the other two modes represent two sets of experimental conditions. In this case, $y_{i,j,k}$ is the measurement for unit i when condition one is at level j and condition two is at level k. Letting the factors corresponding to the three modes be $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$, modeling the rows $\mathbf{u}_1, \ldots, \mathbf{u}_{m_1}$ of the $m_1 \times R$ factor matrix $\mathbf{U}$ as i.i.d. multivariate normal$(\boldsymbol{\mu}, \Psi)$ induces a covariance among the elements of each unit-specific $m_2 \times m_3$ matrix $\mathbf{Y}_i = \{y_{i,j,k}, 1 \leq j \leq m_2, 1 \leq k \leq m_3\}$, given by the following calculation:

$$\begin{aligned}
\mathbf{u}_i &= \boldsymbol{\mu} + \boldsymbol{\gamma}_i,\ \boldsymbol{\gamma}_i \sim \text{multivariate normal}(\mathbf{0}, \boldsymbol{\Psi}) \\
y_{i,j,k} &= \langle \mathbf{u}_i, \mathbf{v}_j, \mathbf{w}_k \rangle + \epsilon_{i,j,k} \\
&= \mathbf{u}_i^T(\mathbf{v}_j \circ \mathbf{w}_k) + \epsilon_{i,j,k} = \boldsymbol{\mu}^T(\mathbf{v}_j \circ \mathbf{w}_k) + \boldsymbol{\gamma}_i^T(\mathbf{v}_j \circ \mathbf{w}_k) + \epsilon_{i,j,k} \\
\text{Cov}[y_{i,j,k}, y_{i,l,m}] &= \text{E}[\boldsymbol{\gamma}_i^T(\mathbf{v}_j \circ \mathbf{w}_k)(\mathbf{v}_l \circ \mathbf{w}_m)^T\boldsymbol{\gamma}_i] \\
&= \text{tr}((\mathbf{v}_j \circ \mathbf{w}_k)(\mathbf{v}_l \circ \mathbf{w}_m)^T\boldsymbol{\Psi}) \\
&= \text{tr}([(\mathbf{v}_j\mathbf{v}_l^T) \circ (\mathbf{w}_k\mathbf{w}_m^T)]\boldsymbol{\Psi})
\end{aligned}$$

Each unit has a measurement under conditions (j, k) and under (l, m), and the covariance of these measurements across experimental units is determined by $\mathbf{v}_j\mathbf{v}_l^T$, $\mathbf{w}_k\mathbf{w}_m^T$ and the covariance matrix $\boldsymbol{\Psi}$. Fixing $\boldsymbol{\Psi}$ in advance places restrictions on the form of this covariance. This suggests the use of a hierarchical model as an alternative, whereby the mean and variance of the factors of each mode are estimated from the observed data. Returning to the general case of K modes, the proposed hierarchical model is as follows:

$$\begin{aligned}
\{\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_{m_k}^{(k)}\} &\overset{\text{iid}}{\sim} \text{multivariate normal}(\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k) \\
\boldsymbol{\Psi}_k &\sim \text{inverse-Wishart}(\mathbf{S}_0, \nu_0) \\
\boldsymbol{\mu}_k|\boldsymbol{\Psi}_k &\sim \text{multivariate normal}(\boldsymbol{\mu}_0, \boldsymbol{\Psi}_k/\kappa_0)
\end{aligned}$$

Readers familiar with factor models for matrices (the case of $K = 2$) may be concerned about the non-orthogonality of the columns of the latent factor matrices in the above model. In the matrix case, the mean matrix for $\mathbf{Y}$ is given by $\boldsymbol{\Theta} = \mathbf{U}^{(1)}\mathbf{U}^{(2)T}$. Letting $\tilde{\mathbf{U}}^{(k)} = \mathbf{U}^{(k)}\mathbf{H}$, $k = 1, 2$ we

see that $\boldsymbol{\Theta} = \tilde{\mathbf{U}}^{(1)}\tilde{\mathbf{U}}^{(2)T}$ for any orthonormal matrix $\mathbf{H}$. This invariance to rotation in the matrix case, however, does not generalize to rotation invariance for multilinear representations of arrays: Kruskal [1977] shows that other than some elementary invariances (such as a common relabeling of the columns of all the factor matrices), multilinear factor representations are generally rotationally unique.

Diffuse priors can be used as a default, such as $\boldsymbol{\mu}_0 = \mathbf{0}$, $\kappa_0 = 1$, $\nu_0 = R + 1$ and $\mathbf{S}_0^{-1} = \mathbf{I}\tau_0^2$, where τ_0^2 is some pre-specified value determined by the scale of the measurements. An alternative default set of priors can be based on unit information prior distributions [Kass and Wasserman, 1995], which weakly center the prior parameters around estimates obtained from the data. For example, τ_0^2 could be obtained as the variance of latent factor estimates obtained from a rank-R least squares approximation to $\mathbf{Y}$, and the prior distribution for σ^2 could be weakly centered around the corresponding residual variance. In either case, the full conditional distributions for all parameters have straightforward derivations, and are summarized in the following Gibbs sampling scheme: Given current values of $\{\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}\}$ and σ^2, new values of these parameters are generated as follows:

1. For each $k \in \{1, \ldots, K\}$ in random order,

   (a) sample $\boldsymbol{\Psi}_k \sim$ inverse-Wishart$([\mathbf{U}^{(k)T}\mathbf{U}^{(k)} + \mathbf{I}\tau_0^2]^{-1}, m_k + R + 1)$

   (b) sample $\boldsymbol{\mu}_k \sim$ multivariate normal$(\mathbf{U}^{(k)T}\mathbf{1}/[m_k + 1], \boldsymbol{\Psi}_k/[m_k + 1])$

   (c) sample $\mathbf{U}^{(k)} \sim$ matrix normal$(\tilde{\mathbf{M}}_k, \tilde{\boldsymbol{\Psi}}_k, \mathbf{I})$, where

   - $\tilde{\boldsymbol{\Psi}}_k = (\mathbf{Q}_k/\sigma^2 + \boldsymbol{\Psi}_k^{-1})^{-1}$
   - $\tilde{\mathbf{M}}_k = (\mathbf{L}_k/\sigma^2 + \mathbf{1}\boldsymbol{\mu}_k^T\boldsymbol{\Psi}_k^{-1})\tilde{\boldsymbol{\Psi}}_k$

2. sample $\sigma^2 \sim$ inverse-gamma$(\tilde{\nu}_0/2, \tilde{\nu}_0\tilde{\sigma}_0^2/2)$, where

   - $\tilde{\nu}_0 = \nu_0 + \prod_{k=1}^K m_k$
   - $\tilde{\nu}_0\tilde{\sigma}_0^2 = \nu_0\sigma_0^2 + ||\mathbf{Y} - \langle \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)} \rangle||^2$

Note that $\{(\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k), k = 1, \ldots, K\}$ will not be separately identifiable since, for example, the scales of $\{\mathbf{U}^{(k)}, k = 1, \ldots, K\}$ are not separately identifiable. However, a non-hierarchical Bayesian approach restricts the overall scale of $\boldsymbol{\Theta}$, as well the shrinkage point for the $\mathbf{U}^{(k)}$'s. In contrast, the hierarchical model allows these things to be determined by the data.

4 Comparison of estimators

This section presents the results of some simulation studies comparing the performance of the hierarchical Bayes procedure to ALS estimation. In the first study, one-hundred random $\boldsymbol{\Theta}$-arrays were generated, each having dimension $m_1 \times m_2 \times m_3 = 10 \times 8 \times 6$ and rank $R = 4$, and each to be estimated from a corresponding "observed" data array $\mathbf{Y}$. Letting $\tilde{R} = m_2 \times m_3$, the $\boldsymbol{\Theta}$ and $\mathbf{Y}$ arrays were generated as follows:

1. For each mode $k \in \{1, 2, 3\}$,
   (a) sample $\boldsymbol{\Psi}_k$ as follows:
      i. sample $\boldsymbol{\Psi}_0 \sim$ Wishart $(\mathbf{I}, \tilde{R} + 1)$,
      ii. set $\nu_0 = \tilde{R} + x$ where $x \sim$ Poisson$(\sqrt{\tilde{R}})$,
      iii. sample $\boldsymbol{\Psi}_k \sim$ inverse-Wishart$(\boldsymbol{\Psi}_0, \nu_0)$;
   (b) sample $\boldsymbol{\mu}_k \sim$ multivariate normal $(\mathbf{0}, \boldsymbol{\Psi}_k)$
   (c) sample $\tilde{\mathbf{U}}^{(k)} \sim$ multivariate normal $(\boldsymbol{\mu}_k, \boldsymbol{\Psi}_k)$.
2. Let $\boldsymbol{\Theta}$ be the rank-R least-squares approximation to $\langle \tilde{\mathbf{U}}^{(1)}, \tilde{\mathbf{U}}^{(2)}, \tilde{\mathbf{U}}^{(3)} \rangle$, but rescaled so that the average squared magnitude of the elements $\sum \theta_{i,j,k}^2/(m_1 m_2 m_3)$ is 1.
3. Set $\mathbf{Y} = \boldsymbol{\Theta} + \mathbf{E}$, where $\{\epsilon_{i,j,k}\} \overset{\text{iid}}{\sim}$ normal$(0, 1/4)$.

We now go through the rationale for this simulation scheme. Working backwards, in steps 2 and 3 the error variance for $\mathbf{E}$ is set to be 1/4 of the average squared magnitude of the elements of $\boldsymbol{\Theta}$. This makes estimation of $\boldsymbol{\Theta}$ feasible but not trivial. In steps 1 and 2, we first generate an array having a maximal rank $\tilde{R}$, and then let $\boldsymbol{\Theta}$ be its rank-4 least-squares approximation. The rationale for this is to make the generated $\boldsymbol{\Theta}$ arrays somewhat different in distribution from the prior distribution that will be used for estimation, thus giving a more fair comparison between the performance of the Bayesian procedure and ALS estimation. Additionally, the "prior" parameters $\boldsymbol{\Psi}_0$ and ν_0 in steps 1.(a) are randomly generated in order to provide a broader range of patterns generated in the $\boldsymbol{\Theta}$ arrays than could be obtained from fixed values of $\boldsymbol{\Psi}_0$ and ν_0.

4.1 Known rank

We first examine the case where the presumed rank of $\boldsymbol{\Theta}$ is equal to the true rank of 4. Two estimates were computed for each of the one-hundred simulated $\boldsymbol{\Theta}$-arrays:

$\hat{\boldsymbol{\Theta}}_{\text{LS}}$ (least squares), an estimate obtained via the alternating least-squares algorithm;

$\hat{\boldsymbol{\Theta}}_{\text{HB}}$ (hierarchical Bayes), a posterior estimate under the hierarchical model and unit information priors described in Section 3.2.

The least squares estimates were obtained by running the ALS algorithm using twenty different random starting values and then selecting the one that gave the minimum residual sum of squares. For each starting value, the ALS algorithm was iterated until the magnitude of the change in the estimate, relative to the magnitude of the estimate, was less than 10^{-6}.

The Bayesian estimates were obtained using the Gibbs sampling scheme described in the previous section, with 1000 iterations to allow for convergence to the stationary distribution ("burn-in"), followed by 10000 iterations for estimating the mean matrix. Mixing of the algorithm was assessed by monitoring the value of $||\boldsymbol{\Theta}||^2$ across the 10000 iterations of the Markov chain. Mixing was generally good, with the median effective sample size (the equivalent numbers of independent Monte Carlo samples) for $||\boldsymbol{\Theta}||^2$ being 9422. For each simulated dataset we obtained a posterior mean estimate of $\boldsymbol{\Theta}$. However, this estimate will generally have a rank higher than 4 as rank is not preserved under linear combinations. For this reason, the rank-4 least squares approximation to the posterior mean was also computed as an alternative Bayesian point estimate of $\boldsymbol{\Theta}$.

The results of the simulation study are summarized in Figure 1. For each dataset and estimation method, the ratio of $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||^2 / ||\mathbf{Y} - \boldsymbol{\Theta}||^2$ was computed to assess the performance of $\hat{\boldsymbol{\Theta}}$ relative to the unbiased estimate $\mathbf{Y}$. In this example where the true rank of $\boldsymbol{\Theta}$ is known, using the reduced-rank ALS estimate is superior to using $\mathbf{Y}$, giving reductions of mean squared error of roughly 60 to 80%. However, the first panel of Figure 1 indicates that the Bayesian estimators provide a substantial further reduction in MSE, amounting to an additional reduction of 41% on average and up to 80% for particular datasets. Also, note that the rank-4 Bayesian point estimate performs essentially the same as the posterior mean estimate, even though the latter may be of rank higher than 4.

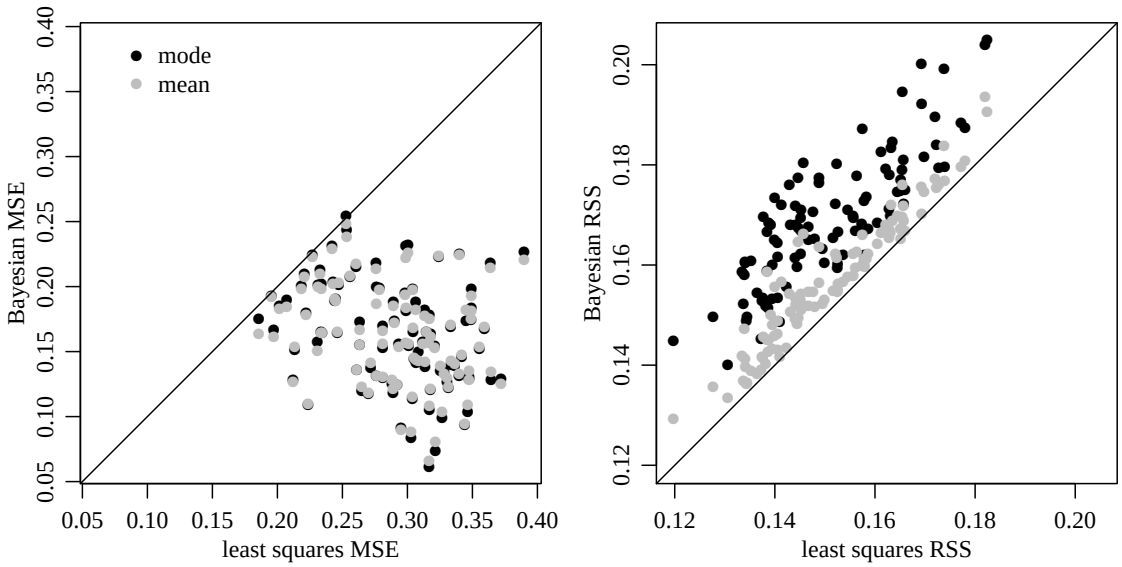

Figure 1: Comparison of MSE and RSS for different estimation methods.

One possible explanation for the superiority of the Bayesian approach over ALS is that the latter does not explore as much of the parameter space as an MCMC algorithm. The second panel of Figure 1, which plots the relative residual sum of squares (RSS) $||\mathbf{Y} - \hat{\mathbf{\Theta}}||^2/||\mathbf{Y}||^2$ for the ALS estimate versus the two Bayes estimates, suggests that this is not the case. This plot indicates that $\hat{\mathbf{\Theta}}_{\text{LS}}$ is in fact closer to $\mathbf{Y}$ than $\hat{\mathbf{\Theta}}_{\text{HB}}$ for every simulated dataset. This observation, together with the superiority of the Bayes estimate in terms of estimating $\mathbf{\Theta}$, suggests that the ALS procedure tends to overfit.

For each of the 100 simulated datasets an alternative Bayesian estimate of $\mathbf{\Theta}$ was also obtained, in which the elements $u_{i,r}^{(k)}$ of the $\mathbf{U}$-matrices were assumed to be *a priori* independent normal$(0, 100)$ random variables. This non-hierarchical approach fixes the amount of regularization, and does not recognize patterns in $\mathbf{\Theta}$ that could be represented by correlations among the latent factors. Not surprisingly, estimates obtained from this approach generally had higher MSEs than the estimates based on the hierarchical model (in 99% of the cases using the posterior mean estimates, and 92% of the cases using rank-4 point estimates).

4.2 Misspecified rank

A more realistic data analysis situation is one in which the true rank of $\boldsymbol{\Theta}$ is not known. In this subsection we investigate the MSEs of $\hat{\boldsymbol{\Theta}}_{\text{LS}}$ $\hat{\boldsymbol{\Theta}}_{\text{HB}}$ for estimating the rank-4 arrays generated as described above, but when the assumed rank is $R \in \{1, \ldots, 8\}$. Using the same simulation

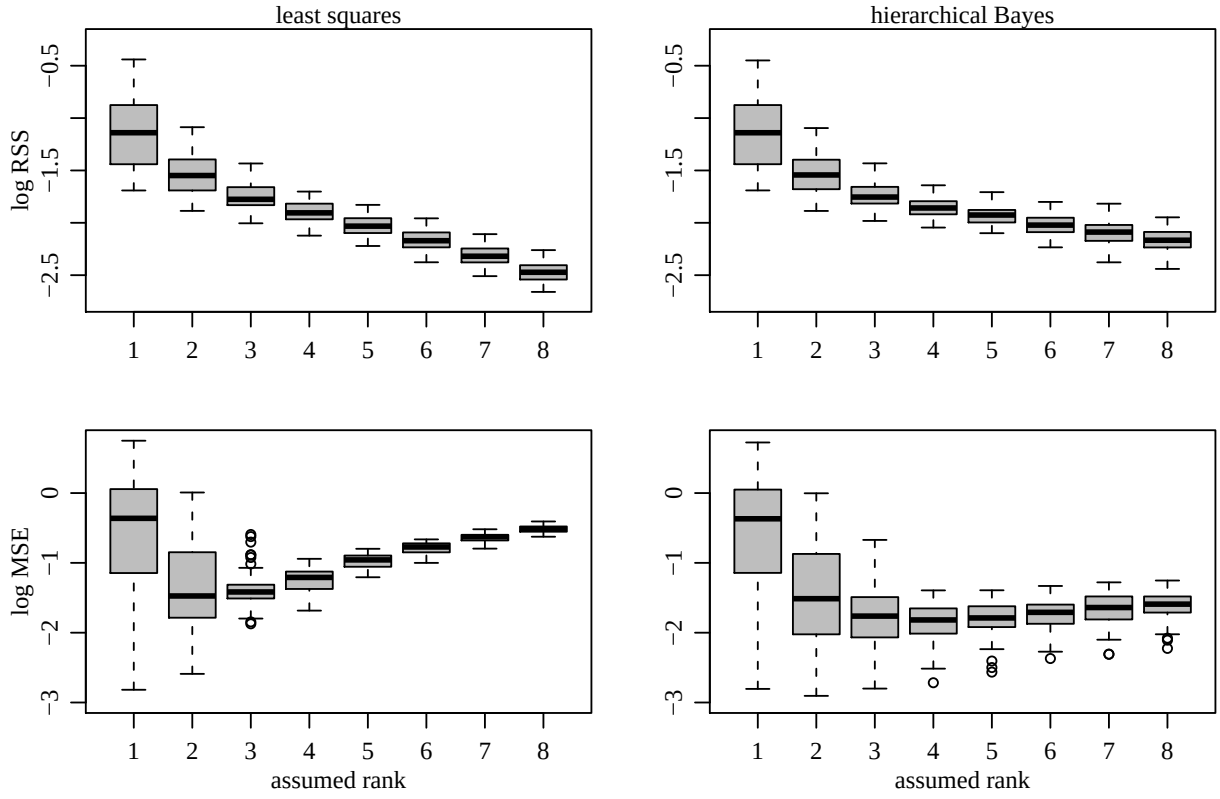

Figure 2: RSSs and MSEs under different presumed ranks and estimation methods.

and estimation procedures as described in the previous subsection, a $\hat{\boldsymbol{\Theta}}$ was obtained for each of the 100 simulated $\boldsymbol{\Theta}$-arrays and for each combination of the two estimation methods and ranks $R \in \{1, \ldots, 8\}$. For each of these $100 \times 2 \times 8$ estimates, a relative MSE $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||^2/||\mathbf{Y} - \boldsymbol{\Theta}||^2$ and RSS $||\mathbf{Y} - \hat{\boldsymbol{\Theta}}||^2/||\mathbf{Y}||^2$ was computed as before. The first of these measures the fidelity of the estimate to the true underlying parameter, and the second to the the data.

A summary of the results are plotted in the four panels of Figure 2. For example, each boxplot in the top row of plots summarizes the 100 RSS values of the ALS estimates assuming a given rank. As expected, as the rank increases the percentage of the variation in $\mathbf{Y}$ explained by the ALS estimate goes up and the RSS goes down. However, the first plot of the bottom row shows

that increasing the rank of the ALS estimate beyond 3 generally increases the MSE. In contrast, the MSE of the hierarchical estimate $\hat{\mathbf{\Theta}}_{\mathrm{HB}}$ generally achieves a minimum at the actual rank of 4, and increases relatively slowly as the assumed rank is increased beyond 4. This suggests that the hierarchical Bayes approach is more robust to overfitting than the least squares method. Since the "true" rank of $\mathbf{\Theta}$ is generally not known, it may be desirable to fit a model with a moderately large rank in the hopes of capturing as much of $\mathbf{\Theta}$ as possible. The above results suggest that a hierarchical Bayes estimate may be preferable in such situations, as it provides a more stable estimate of $\mathbf{\Theta}$ across different choices of the presumed rank.

4.3 Rank selection

We now consider the possibility of estimating the rank R from the observed data array $\mathbf{Y}$. One popular model selection procedure is to minimize the Bayesian information criterion, or BIC [Schwarz, 1978]. The BIC for a given model and dataset y is $-2\ln p(y|\hat{\theta}) + p\ln n$, where $\hat{\theta}$ is the parameter estimate, p is the dimension of θ and n is the sample size. In practice, the BIC can be computed for a range of different models, and the one giving the smallest BIC is selected. This procedure favors models that fit well (in terms of likelihood) but penalizes model complexity.

As pointed out by Pauler [1998], for hierarchical models the number of parameter can be ambiguous. As a remedy, Spiegelhalter et al. [2002] proposed the deviance information criterion, or DIC which can be computed from output of a Markov chain. The DIC is given by $\bar{D} + \tilde{p}$, where $\bar{D}$ is the average value of $-2\ln p(y|\theta)$ across iterations of the Markov chain, and $\tilde{p}$ is the "effective number of parameters", given by $\tilde{p} = \bar{D} + 2\ln p(y|\hat{\theta})$, where $\hat{\theta}$ is an estimate of θ. For our model the parameters are $\mathbf{\Theta}$ and σ^2, and we take our estimates to be the posterior mean of $\mathbf{\Theta}$ and the mean residual error under the posterior mean, respectively.

For each of the 100 simulated datasets described above we computed the DIC for each value of $R \in \{1, \ldots, 8\}$, and took our "estimate" $\hat{R}$ of R to be the rank for which the DIC was minimized. The fraction of times $\hat{R}$ took on the values $\{1, \ldots, 8\}$ was $\{0.08, 0.15, 0.27, 0.28, 0.06, 0.07, 0.04, 0.05\}$, making the true rank of 4 the most frequently selected, followed closely by 3. The fact that $\hat{R} = 3$ was selected 27 times is somewhat ameliorated by the fact that in 15 of these instances the "best" rank in terms of MSE turned out to be 3 (11 cases) or 2 (4 cases).

To further evaluate the BIC procedure, we also reran the entire simulation study when the true

rank was $R = 2$ and when it was $R = 6$. For the case of $R = 2$, the DIC selection fractions were $\{0.10, 0.74, 0.07, 0.05, 0.02, 0.01, 0.01\}$, indicating that in this case the true rank can be identified with a high degree of accuracy. Rank selection with DIC was more problematic when the true rank was 6, for which the selection proportions were $\{0.07, 0.18, 0.19, 0.17, 0.10, 0.08, 0.09, 0.12\}$. As we would hope, the distribution of ranks selected here is somewhat shifted to the right from the distribution of selected ranks when $R = 4$, but the true rank of 6 can not be identified accurately with DIC. However, the DIC is not as bad in terms of obtaining the rank that gives the best approximation to the true $\boldsymbol{\Theta}$ in terms of MSE. For example, 75% of the 71 simulated datasets for which $\hat{R}$ was less than 6 also attained their minimum MSE at an R-value less than 6. In particular, the seven datasets for which $\hat{R} = 1$ also attained their minimum MSE with a rank 1 model.

5 Example: Multiway means for cross-classified data

Large scale surveys collect data on a variety of numerical and categorical variables. Numerical data are often summarized by computing sample averages for combinations of a set of categorical variables. For example, letting $\mathbf{y}$ be a p-dimensional vector of numerical variables and $\mathbf{x}$ a K-dimensional vector of categorical variables, interest may lie in the population average of $\mathbf{y}$ for a given value of $\mathbf{x}$, which is denoted as $\boldsymbol{\mu}_{\mathrm{x}} \in \mathbb{R}^p$. However, if the number of categorical variables or their number of levels is large compared to the sample size, then we may lack sufficient data to provide stable estimates for each $\boldsymbol{\mu}_{\mathrm{x}}$ separately. For example, the 2008 General Social Survey includes data on the following six variables:

- y_1 (`words`): number of correct answers out of 10 on a vocabulary test;
- y_2 (`tv`): hours of television watched in a typical day;
- x_1 (`deg`) highest degree obtained: none, high school, Bachelor's, graduate;
- x_2 (`age`): 18-34, 35-47, 48-60, 61 and older;
- x_3 (`sex`): male or female;
- x_4 (`child`) number of children: 0, 1, 2, 3 or more.

Complete data for these variables are available for 1116 survey participants. However, there are $4 \times 4 \times 2 \times 4 = 128$ levels of $\mathbf{x}$: More than half of these cells have 5 or fewer observations in them, and about 75% have less than 12 observations. As such, an estimator of $\boldsymbol{\mu}_{\mathrm{x}}$ that uses only data from group $\mathbf{x}$, that is $\{\mathbf{y}_i : \mathbf{x}_i = \mathbf{x}\}$, will be subject to a large sampling variance.

5.1 A multilinear model for group means

Statistical remedies to this problem typically allow the estimate of $\boldsymbol{\mu}_{\mathrm{x}}$ to depend on data from groups other than that corresponding to $\mathbf{x}$. One such approach is to parameterize the set of multivariate means $\{\boldsymbol{\mu}_{\mathrm{x}} : \mathbf{x} \in \mathbb{X}\}$ by a smaller number of parameters. Another approach is via a hierarchical model that allows for the shrinkage of set of parameters towards a common group center. Here we consider the following model which has both of these features:

$$\{\mathbf{y}_i : \mathbf{x}_i = \mathbf{x}\} \overset{\text{iid}}{\sim} \text{multivariate normal}(\boldsymbol{\mu}_{\mathrm{x}}, \boldsymbol{\Sigma}) \tag{3}$$

$$\boldsymbol{\mu}_{\mathrm{x}} = \boldsymbol{\beta}_{\mathrm{x}} + \boldsymbol{\gamma}_{\mathrm{x}} \tag{4}$$

$$\{\boldsymbol{\gamma}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\} \overset{\text{iid}}{\sim} \text{multivariate normal}(\mathbf{0}, \boldsymbol{\Omega}) \tag{5}$$

Equation 3 says that the data within a cell are modeled as multivariate normal, with cell-specific means and a common covariance matrix. Equations 4 and 5 express each $\boldsymbol{\mu}_{\mathrm{x}}$ as equal to a "systematic" component $\boldsymbol{\beta}_{\mathrm{x}}$ plus patternless noise $\boldsymbol{\gamma}_{\mathrm{x}}$. The collection $\{\boldsymbol{\beta}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\}$ can be represented as an $m_1 \times \cdots \times m_K \times p$ array $\mathbf{B}$, where m_k is the number of levels of categorical variable x_k. These values are not separately estimable from the noise $\boldsymbol{\gamma}_{\mathrm{x}}$ unless we assume $\mathbf{B}$ lies in a restricted subset of the set of arrays of this size, such as the set of rank-R arrays. In this setting, where one of the modes of the array represents variables and each other mode represents the different levels of a single categorical variable, it is useful to express the array decomposition as follows:

$$\begin{aligned} \mathbf{B} &= \langle \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}, \mathbf{V} \rangle \ , \quad \text{or equivalently} \\ \boldsymbol{\beta}_x &= \mathbf{V}(\mathbf{u}^{(1)}_{x_1} \circ \cdots \circ \mathbf{u}^{(K)}_{x_K}). \end{aligned}$$

The equations above describe a hierarchical model in which the heterogeneity among $\{\boldsymbol{\mu}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\}$ is centered around a low-dimensional array $\mathbf{B} = \{\boldsymbol{\beta}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\}$. Such a model is similar to representing an interaction term in an ANOVA with a reduced rank matrix [Tukey, 1949, Boik, 1986, 1989]. However, the hierarchical approach used here allows for consistent estimation of each $\boldsymbol{\mu}_{\mathrm{x}}$, but shrinks towards the lower-dimensional representation $\mathbf{B}$ when data are limited.

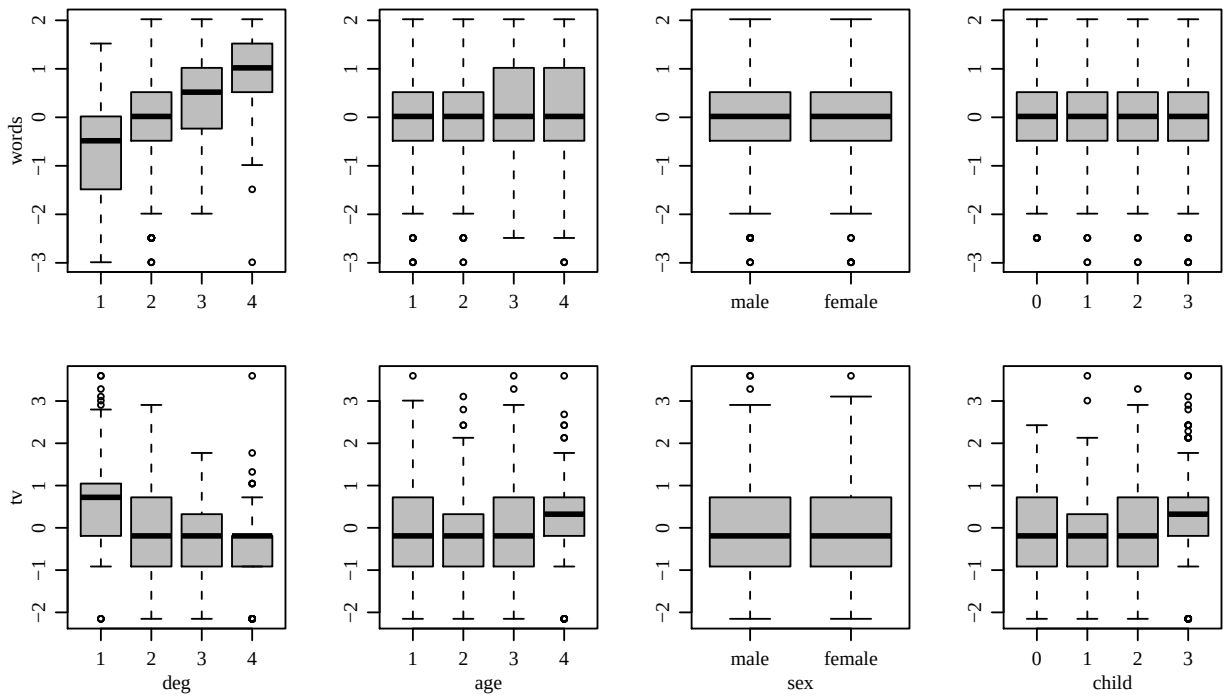

Figure 3: Marginal distributions of vocabulary score and television hours watched for different levels of degree, age sex and number of children.

Estimation for this model can proceed as described in Section 4 with a few modifications. As before, a Gibbs sampler can be used to approximate the posterior distribution of the unknown parameters. Using a conjugate inverse-Wishart prior distribution for $\boldsymbol{\Sigma}$ and the other prior distributions as in Section 4, one iteration of the Markov chain is as follows:

1. sample $\boldsymbol{\Sigma} \sim p(\boldsymbol{\Sigma}|\{\mathbf{y}_i : i = 1, \ldots, n\}, \{\boldsymbol{\mu}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\})$, an inverse-Wishart distribution;

2. sample $\boldsymbol{\mu}_{\mathrm{x}} \sim p(\boldsymbol{\mu}_{\mathrm{x}}|\{\mathbf{y}_i : \mathbf{x}_i = \mathbf{x}\}, \boldsymbol{\beta}_{\mathrm{x}}, \boldsymbol{\Sigma})$, a multivariate normal distribution for each $\mathbf{x} \in \mathcal{X}$;

3. sample $\boldsymbol{\Omega} \sim p(\boldsymbol{\Omega}|\{\boldsymbol{\mu}_{\mathrm{x}}, \boldsymbol{\beta}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\}, \mathbf{V})$, an inverse-Wishart distribution;

4. iteratively sample $\{\mathbf{U}^{(k)}, k = 1, \ldots, K\}$ as in Section 3;

5. sample $\mathbf{V} \sim p(\mathbf{V}|\mathbf{U}, \{\boldsymbol{\mu}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\}, \boldsymbol{\Omega})$, a matrix normal distribution.

Derivations of the full conditional distributions are straightforward and are available from the author and in the computer code available at the author's website. Provided here are only the

following comments which describe some of the calculations: Let the model for the $p \times R$ matrix $\mathbf{V}$ be such that the R columns are i.i.d. multivariate normal with a zero mean vector and covariance equal to $\boldsymbol{\Omega}$. Doing so links the scale of the factor effects for $\boldsymbol{\mu}_{\mathrm{x}}$ to the scale of the across-group differences $\boldsymbol{\gamma}_{\mathbf{x}}$. Writing $\tilde{\boldsymbol{\mu}}_{\mathrm{x}} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{\mu}_{\mathrm{x}}$ and $\tilde{\mathbf{V}} = \boldsymbol{\Omega}^{-1/2}\mathbf{V}$, we have

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_{\mathbf{x}} &= \tilde{\mathbf{V}}(\mathbf{u}_{x_1}^{(1)} \circ \cdots \circ \mathbf{u}_{x_K}^{(K)}) + \tilde{\boldsymbol{\gamma}}_{\mathbf{x}} \ , \quad \text{with} \\ \{\tilde{\boldsymbol{\gamma}}_{\mathbf{x}}\} &\overset{\text{iid}}{\sim} \text{multivariate normal}(\mathbf{0}, \mathbf{I}). \end{aligned}$$

From this, we see that sampling from the full conditional distribution of $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(K)}$ can be done just as in Section 3.2, with σ^2 replaced by 1 and the observed array data replaced by the values of the array defined by $\{\tilde{\boldsymbol{\mu}}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\}$. Similarly, the full conditional of $\tilde{\mathbf{V}}$ is the matrix normal distribution from Section 3.1, again with σ^2 replaced by 1 and $\{\tilde{\boldsymbol{\mu}}_{\mathrm{x}} : \mathbf{x} \in \mathcal{X}\}$ taking the place of the observed array data. A value of $\mathbf{V}$ can be generated from its full conditional distribution by sampling $\tilde{\mathbf{V}}$ from this matrix normal distribution and then setting $\mathbf{V} = \boldsymbol{\Omega}^{1/2}\tilde{\mathbf{V}}$. Finally, note that the inverse-Wishart full conditional distribution of for $\boldsymbol{\Omega}$ depends on $\mathbf{V}$: If we have $\boldsymbol{\Omega} \sim$ inverse-Wishart$(\boldsymbol{\Omega}_0^{-1}, \eta_0)$ then the full conditional distribution of $\boldsymbol{\Omega}$ is inverse-Wishart$(\boldsymbol{\Omega}_1^{-1}, \eta_1)$ where $\eta_1 = \eta_0 + R + \prod_{k=1}^{K} m_k$ and $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_0 + \mathbf{V}^T\mathbf{V} + \sum_{\mathbf{x}}(\boldsymbol{\mu}_{\mathrm{x}} - \boldsymbol{\beta}_{\mathrm{x}})(\boldsymbol{\mu}_{\mathrm{x}} - \boldsymbol{\beta}_{\mathrm{x}})^T$.

5.2 Posterior analysis of GSS data

We now discuss posterior inference for the GSS data based on the above model and estimation scheme. The numerical variables y_1 (`words`) and y_2 (`tv`) were first centered and scaled to have zero mean and unit variance. Prior distributions for the covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ were taken to be independent inverse-Wishart distributions with $p + 1 = 3$ degrees of freedom each and centered around the sample covariance (correlation) matrix of $\{y_{i,1}, y_{i,2}, i = 1, \ldots, n\}$. Doing so gives these prior distributions an empirical basis while still keeping them relatively weak. Such priors are similar to the "unit information" prior distributions described in Kass and Wasserman [1995]. A rank-2 model for the array of means was used so that the estimated factor effects could represented with a simple two-dimensional plot.

The algorithm described above was used to construct a Markov chain consisting of 22,000 iterations, the first 2000 of which were discarded to allow for convergence to the stationary distribution. Parameter values were saved every 10th iteration, leaving 2000 saved values for Monte

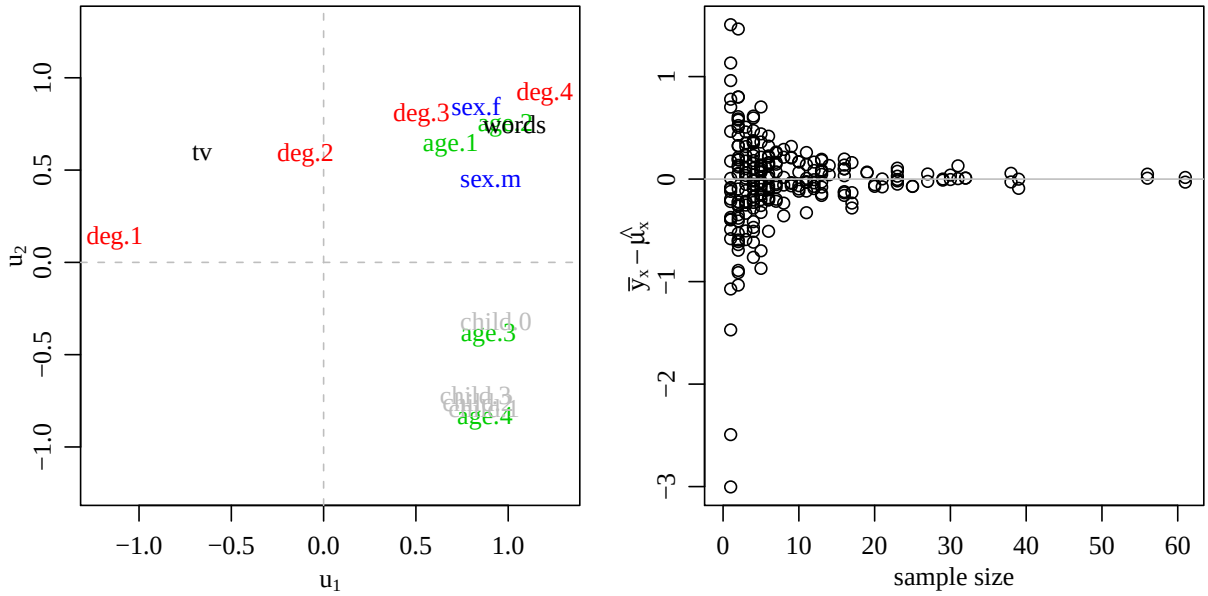

Figure 4: Posterior estimates of factor scores, along with the amount of shrinkage as a function of cell-specific sample size.

Carlo approximation. Mixing of the Markov chain was examined by inspecting the sequences of saved values of $\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$ and the average value of $\{\boldsymbol{\beta}_x\}$ across levels of $\mathbf{x}$. The effective sample sizes for these parameters were all over 1000. Some summary descriptions of the resulting posterior estimates are shown in Figure 4. The first panel plots point estimates of the latent factors $\mathbf{V}$ and $\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(4)}$. These were obtained as follows: A posterior mean array $\bar{\mathbf{B}}$ was obtained from the 2000 saved values of $\mathbf{B}$ from the Markov Chain. This array is not quite a rank-2 array, as rank is not generally preserved under array addition. An alternating least-squares algorithm was performed on $\bar{\mathbf{B}}$ to obtain a rank-2 point estimate $\hat{\mathbf{B}}$ and a multiplicative decomposition in terms of matrices $\hat{\mathbf{V}}$, $\hat{\mathbf{U}}^{(1)}, \ldots, \hat{\mathbf{U}}^{(4)}$. The difference between $\bar{\mathbf{B}}$ and $\hat{\mathbf{B}}$ was small, with $||\bar{\mathbf{B}} - \hat{\mathbf{B}}||^2/||\bar{\mathbf{B}}||^2 = 0.00011$. These point estimates of the latent factors are shown in the first panel in Figure 4: For example, the matrix $\hat{\mathbf{U}}^{(1)}$ represents the multiplicative effects of `deg`, and consists of a two-dimensional vector for each level of this variable. These vectors are plotted in the figure with "deg.1" representing no degree, "deg.2" a high school degree, and so on. Similarly, the matrix $\hat{\mathbf{V}}$ has a two-dimensional vector for each of the two numerical variables. To interpret the figure, note that the estimated mean for either numeric variable in any cell can be obtained by coordinate-wise multiplication and

then addition of the latent factor vectors. For example, the proximity of the "words" vector to the "deg.3" and "deg.4" vectors indicates that these two groups have higher mean vocabulary scores than the other two degree categories. Similarly, the close proximity of the "child.1", "child.2" and "child.3" vectors indicates lack of heterogeneity in the means for three of these four categories across levels of the other $\mathbf{x}$-variables. Finally, note that some care should go into interpreting the figure, as the array $\mathbf{B} = \langle \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(4)}, \mathbf{V} \rangle$ is invariant to certain transformations of the factors: For example, multiplying either the first or second column of each of an even number of factor matrices by -1 does not change the value of $\mathbf{B}$.

The second plot in Figure 4 highlights how the estimated cell means $\{\hat{\boldsymbol{\mu}}_{\mathrm{x}}\}$ differ from the empirical cell means $\{\bar{\mathbf{y}}_{\mathrm{x}}\}$ as a function of sample size. This plot indicates what we would expect from a hierarchical model: The difference between estimated cell mean and empirical cell mean decreases with increasing sample size. A cell with a large sample size will have $\bar{\mathbf{y}}_{\mathrm{x}} \approx \hat{\boldsymbol{\mu}}_{\mathrm{x}}$, whereas a cell with a small sample size will have an estimated mean $\hat{\boldsymbol{\mu}}_{\mathrm{x}}$ shrunk towards the reduced-rank value $\hat{\boldsymbol{\beta}}_{\mathrm{x}}$. Note that without the multiplicative effects in Equation 4 of the hierarchical model, the cell means would all be shrunk towards a common vector, regardless of the value of $\mathbf{x}$. In contrast, the hierarchical multiplicative effects model allows cell-specific shrinkage, as estimated by the reduced rank array $\hat{\mathbf{B}}$.

An alternative approach to the analysis of these data might involve MANOVA or a hierarchical model similar to the one above but in which $\boldsymbol{\beta}_{\mathrm{x}}$ is parameterized in terms of additive effects, so that $\boldsymbol{\beta}_{\mathrm{x}} = \mathbf{u}^{(1)}_{x_1} + \cdots + \mathbf{u}^{(K)}_{x_K}$ with each $\mathbf{u}^{(k)}_{x_k} \in \mathbb{R}^p$. Such additive models have representations as multilinear models, although of course they are restricted to be additive. For comparison, an additive MANOVA model was fit and the average value of $(\bar{\mathbf{y}}_{\mathrm{x}} - \hat{\boldsymbol{\beta}}_{\mathrm{x}})^2$ was computed, measuring the lack-of-fit of the additive model. This value was about the same as the corresponding value for the multilinear model for the `tvhours` variable, but 15% larger for the `words` variable. This indicates that some patterns among the cell means for `words` cannot be represented with an additive model. In general, we may expect that some aspects of the heterogeneity among the $\boldsymbol{\mu}_{\mathbf{x}}$'s will not be additive. In such situations, it may be preferable to use a multiplicative model whose complexity can be controlled with the choice of the rank R rather than to have to consider the inclusion and estimation of a variety of higher-order interaction terms.

6 Example: Analysis of longitudinal conflict data

The theory of the Kantian peace holds that militarized interstate disputes are less likely to occur between democratic countries. Ward et al. [2007] evaluate this theory using international cooperation and conflict data from the cold war period. The data include records of militarized conflict and cooperation every five years from 1950 to 1985, along with economic and political characteristics of the countries. In this section we analyze a subset of the data from Ward et al. [2007]. These data include cooperation, conflict and gross domestic product data (gdp) for each of $m = 66$ countries every fifth year, $t \in \{1950, 1955, \ldots, 1980, 1985\}$. Additionally, each country in each of these years has a polity score, measuring the level of openness in government. A positive polity score is given to democratic states, while a negative score is given to authoritarian states.

The cooperation and conflict data form a three-way array with two modes representing country pairs and one mode representing time. In this section we will fit an ordered probit model of cooperation and conflict data as a function of gdp and polity. Specifically, for each unordered pair $\{i, j\}$ of countries and each time t, our data are as follows:

$y_{i,j,t} \in \{-5, -4, \ldots, +1, +2\}$, indicating the level of military cooperation (positive) or conflict (negative) between countries i and j in year t;

$x_{i,j,t,1} = \log \text{gdp}_i + \log \text{gdp}_j$, the sum of the log gdps of the two countries;

$x_{i,j,t,2} = (\log \text{gdp}_i) \times (\log \text{gdp}_j)$, the product of the log gdps;

$x_{i,j,t,3} = \text{polity}_i \times \text{polity}_j$, where $\text{polity}_i \in \{-1, 0, +1\}$;

$x_{i,j,t,4} = (\text{polity}_i > 0) \times (\text{polity}_j > 0)$.

The sample space for $y_{i,j,t}$ is ordered but the scale is not meaningful: The difference between $y = 0$ and $y = 1$ is not comparable to the difference between $y = -5$ and $y = -4$. For this reason we use the following ordered probit model to relate $y_{i,j,t}$ to $\mathbf{x}_{i,j,t}$:

$$\begin{aligned} z_{i,j,t} &= \boldsymbol{\beta}^T \mathbf{x}_{i,j,t} + \gamma_{i,j,t} \\ y_{i,j,t} &= \max\{k : z_{i,j,t} > c_k, k \in \{-5, -4, \ldots, +1, +2\}\} \end{aligned}$$

In this model the parameters to estimate include the regression coefficients $\boldsymbol{\beta}$ and the cutoffs $\mathbf{c} = (c_{-4}, \ldots, c_{+2})$, with $c_{-5} = -\infty$. The usual probit regression model would assume the $\gamma_{i,j,t}$'s are

independent standard normal variables (standard, as the scale of these error terms is not separately identifiable from $\boldsymbol{\beta}$ and $\mathbf{c}$). However, results of Ward et al. [2007] suggest that the residuals from regression models of international relations data are generally not patternless. For example, we might expect $\gamma_{i,1,t}, \ldots, \gamma_{i,66,t}$ to exhibit statistical correlation, as these residuals are all associated with country i. More subtle might be higher order patterns common in relational data: If i and j have a positive relationship and j and k have a positive relationship, then a positive relationship between i and k is more likely.

Hoff [2008] describes how two-way factor models can be used to represent patterns in ordinal matrix-valued relational and social network data. Here we extend this idea, using a three-way factor model to represent the longitudinal relational patterns represented by the array $\boldsymbol{\Gamma} = \{\gamma_{i,j,t}\}$. Specifically, the following factor model is proposed:

$$\begin{aligned} \gamma_{i,j,t} &= \langle \mathbf{u}_i, \mathbf{u}_j, \mathbf{v}_t \rangle + \epsilon_{i,j,t} \ , \text{ with} \\ \{\epsilon_{i,j,t} = \epsilon_{j,i,t}\} &\overset{\text{iid}}{\sim} \text{normal}(0,1). \end{aligned}$$

The $\mathbf{u}_j$'s are vectors representing heterogeneity among the countries and the $\mathbf{v}_t$'s represent heterogeneity over time. This is a modification of the usual three-way PARAFAC representation to accommodate the fact that the data are symmetric ($y_{i,j,t} = y_{j,i,t}$). This model has a simple interpretation: Letting $\boldsymbol{\Gamma}_t = \{\gamma_{i,j,t} : (i,j) \in \{1,\ldots,m\}^2\}$, we have

$$\boldsymbol{\Gamma}_t = \mathbf{U}\boldsymbol{\Lambda}_t\mathbf{U}^T + \mathbf{E}_t \ , \text{ where } \mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)^T \text{ and } \boldsymbol{\Lambda}_t = \text{diag}(\mathbf{v}_t).$$

This symmetric version of the PARAFAC model is analogous to a type of eigenvalue decomposition of a collection of square matrices $\{\boldsymbol{\Gamma}_{1950}, \ldots, \boldsymbol{\Gamma}_{1985}\}$ in which the eigenvectors are held constant across matrices, but the eigenvalues are allowed to vary.

The unobserved quantities in this model include the latent variable array $\mathbf{Z}$ as well as the parameters $\mathbf{U}, \mathbf{V}$ and $\boldsymbol{\beta}$. Using the same hierarchical prior distributions for $\mathbf{U}$ and $\mathbf{V}$ described in Section 3.2 and a diffuse multivariate normal$(\mathbf{0}, 100 \times \mathbf{I})$ prior distribution for $\boldsymbol{\beta}$, we can implement a Gibbs sampler to approximate the joint posterior distribution $p(\mathbf{Z}, \mathbf{U}, \mathbf{V}, \boldsymbol{\beta} | \mathbf{Y}, \mathbf{X})$. All full conditionals are standard, and are available from the supplementary material at the author's website. Using a rank-2 model, the Gibbs sampler was run for 505,000 iterations, dropping the first 5,000 to allow for burn-in and then saving the parameter values every 10th iteration. Convergence of the Markov chain was monitored via the sampled values of $\boldsymbol{\beta}$. The effective sample sizes for the

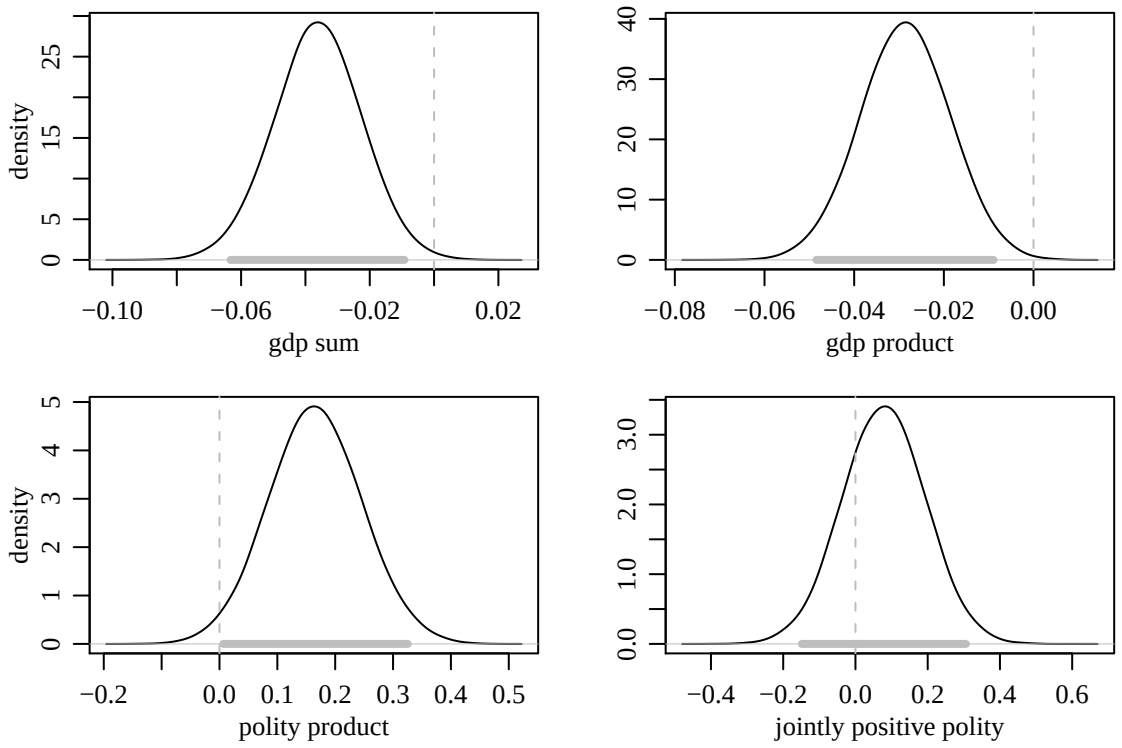

Figure 5: Posterior densities for the elements of $\boldsymbol{\beta}$. Gray lines are 95% confidence intervals.

four regression coefficients based on the 50,000 saved scans were 12,548, 16,622, 1,386 and 8,878 respectively.

The plots in Figure 5 show the marginal posterior distributions of the four regression coefficients, along with 95% highest posterior density confidence intervals. The results indicate a negative association between gdp and the latent variable z, reflecting the fact that a majority of the conflicts over the cold war period involved economically large countries. The plots in the second row indicate that $z_{i,j}$ tends to be larger if both i and j have polity scores of the same sign, but that there is not strong evidence for a further increase if the polities of i and j are both positive.

Figure 6 displays a summary of the posterior distribution of $\mathbf{U}$ and $\mathbf{V}$. This summary was obtained as follows: First, a Monte Carlo approximation $\hat{\boldsymbol{\Theta}}$ of the posterior mean of the three-way array $\boldsymbol{\Theta} = \langle \mathbf{U}, \mathbf{U}, \mathbf{V} \rangle$ was obtained using the values generated from the Markov chain. The alternating least-squares algorithm was then applied to $\hat{\boldsymbol{\Theta}}$ to obtain values $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$. The columns of $\hat{\mathbf{U}}$ were normalized to be unit vectors, and the columns of $\hat{\mathbf{V}}$ were then rescaled accordingly. The columns of the latent factor matrices were then permuted so that the magnitude of the columns

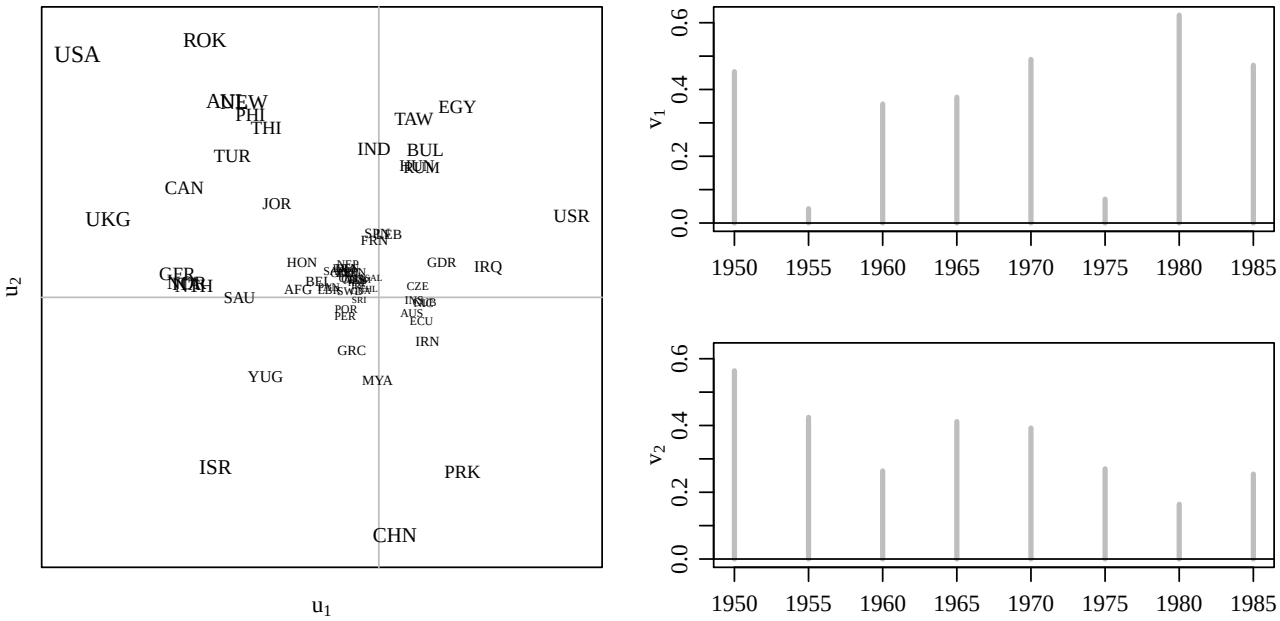

Figure 6: Posterior estimates of the country- and time-specific factors.

of $\hat{\mathbf{V}}$ were in decreasing order. The resulting values are plotted in Figure 6. The large square plot shows the estimates of the two-dimensional latent factor vectors $\{\hat{\mathbf{u}}_i\}$ for each country, with a larger font used for those countries with larger vectors. The second column gives the values of $\hat{v}_{t,k}$, sorted chronologically. Since all of these values are positive, two latent vectors $\{\hat{\mathbf{u}}_{i_1}, \hat{\mathbf{u}}_{i_2}\}$ being in similar directions indicates a tendency for countries i_1 and i_2 to cooperate militarily, whereas vectors in opposite directions indicate a tendency for conflict. For example, the vectors corresponding to USA and South Korea are similar to each other and in the opposite direction of China and North Korea. The heterogeneity of the $\hat{\mathbf{v}}_t$'s over time allows for different patterns of conflict across the years. For example, cooperation and conflict in 1980 and 1985 are described primarily by the first dimension of the factors (u_1), whereas events in 1955 and 1975 primary by the second (u_2).

7 Discussion

This article has presented a hierarchical version of a reduced-rank multilinear model for array data and a Bayesian method for parameter estimation. Unlike least-squares estimation, a Bayesian approach allows for regularized estimates of the potentially large number of parameters in a mul-

tilinear model. Unlike a non-hierarchical Bayesian approach, the hierarchical approach provides a data-driven method of regularization, and a more flexible representation of the patterns in the data array. Additionally, in a simulation study the estimates provided by the hierarchical approach showed robustness to rank misspecification, as compared those obtained from a least-squares or non-hierarchical approach.

Another advantage of the Bayesian approach is that it allows for the incorporation of multilinear structure into a broad class of statistical models. For example, a least-squares approach would be inappropriate for the ordinal cooperation and conflict data in Section 6, but Bayesian estimation for these data, using a probit model with multilinear effects, is relatively straightforward. As another example, the survey data presented in Section 5 was not in the form of an array, but the cell means corresponding to the 128 levels of the 4 categorical variables can be represented as such. A reduced-rank multilinear model provides a parsimonious representation of the cell means, but also is more flexible than a simple additive effects model.

An important line of future research is the study of the theoretical properties of hierarchical Bayesian approaches to parameter estimation for multiway data arrays. For a matrix model in which $\mathbf{Y} = \mathbf{\Theta} + \mathbf{E}$ and $\mathbf{E}$ is a matrix of normally-distributed noise, Tsukuma [2008, 2009] studies Bayesian and hierarchical Bayesian approaches to providing admissible and minimax estimates of $\mathbf{\Theta}$. One aspect of this work shows that under certain prior distributions on the singular vectors of $\mathbf{\Theta}$, the Bayes estimates are equivariant and can be obtained by shrinking the singular values of $\mathbf{Y}$. Such estimates are somewhat analogous to those presented in this article for multiway data, as shrinking the singular values of a matrix is similar to regularizing the variance of a set of multiplicative factors. The author is currently investigating the extent to which such similarities between the matrix and array models lead to similar theoretical properties of Bayesian estimates in the two cases.

Replication code and data for the numerical results in this paper are available at the author's website: `http://www.stat.washington.edu/~hoff`

References

Robert J. Boik. Testing the rank of a matrix with applications to the analysis of interaction in ANOVA. *J. Amer. Statist. Assoc.*, 81(393):243–248, 1986. ISSN 0162-1459.

Robert J. Boik. Reduced-rank models for interaction in unequally replicated two-way classifications. *J. Multivariate Anal.*, 28(1):69–87, 1989. ISSN 0047-259X.

R. Coppi and S. Bolasco, editors. *Multiway data analysis*. North-Holland Publishing Co., Amsterdam, 1989. ISBN 0-444-87410-0. Papers from the International Meeting on the Analysis of Multiway Data Matrices held in Rome, March 28–30, 1988.

K. R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58:453–467, 1971. ISSN 0006-3444.

K. Ruben Gabriel. Generalised bilinear regression. *Biometrika*, 85(3):689–700, 1998. ISSN 0006-3444.

R.A. Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA working papers in phonetics*, 16(1):84, 1970.

R.A. Harshman and M.E. Lundy. The PARAFAC model for three-way factor analysis and multidimensional scaling. *Research methods for multimode data analysis*, pages 122–215, 1984.

Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 657–664. MIT Press, Cambridge, MA, 2008.

Robert E. Kass and Larry Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.*, 90(431):928–934, 1995. ISSN 0162-1459.

Pieter M. Kroonenberg. *Applied multiway data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2008. ISBN 978-0-470-16497-6. With a foreword by Willem J. Heiser and Jarqueline Meulman.

J. B. Kruskal. Rank, decomposition, and uniqueness for 3-way and N-way arrays. In *Multiway data analysis (Rome, 1988)*, pages 7–18. North-Holland, Amsterdam, 1989.

Joseph B. Kruskal. More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281–293, 1976. ISSN 0033-3123.

Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Appl.*, 18(2):95–138, 1977.

Donna K. Pauler. The Schwarz criterion and related methods for normal linear models. *Biometrika*, 85(1):13–27, 1998. ISSN 0006-3444. doi: 10.1093/biomet/85.1.13. URL `http://dx.doi.org/10.1093/biomet/85.1.13`.

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978. ISSN 0090-5364.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(4):583–639, 2002. ISSN 1369-7412. doi: 10.1111/1467-9868.00353. URL `http://dx.doi.org/10.1111/1467-9868.00353`.

Giorgio Tomasi and Rasmus Bro. A comparison of algorithms for fitting the PARAFAC model. *Comput. Statist. Data Anal.*, 50(7):1700–1734, 2006. ISSN 0167-9473.

Hisayuki Tsukuma. Admissibility and minimaxity of Bayes estimators for a normal mean matrix. *J. Multivariate Anal.*, 99(10):2251–2264, 2008. ISSN 0047-259X.

Hisayuki Tsukuma. Generalized Bayes minimax estimation of the normal mean matrix with unknown covariance matrix. *J. Multivariate Anal.*, 100(10):2296–2304, 2009.

J.W. Tukey. One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242, 1949.

L. Vega-Montoto and P.D. Wentzell. Maximum likelihood parallel factor analysis (MLPARAFAC). *Journal of Chemometrics*, 17(4):237–253, 2003.

L. Vega-Montoto, H. Gu, and P.D. Wentzell. Mathematical improvements to maximum likelihood parallel factor analysis: theory and simulations. *Journal of chemometrics*, 19(4), 2005.

M.D. Ward, R.M. Siverson, and X. Cao. Disputes, democracies, and dependencies: A reexamination of the Kantian peace. *American Journal of Political Science*, pages 583–601, 2007.