Bayesian Analysis (2016) **11**, Number 3, pp. 627–648

Equivariant and Scale-Free Tucker Decomposition Models

Peter D. Hoff*

Abstract. Analyses of array-valued datasets often involve reduced-rank array approximations, typically obtained via least-squares or truncations of array decompositions. However, least-squares approximations tend to be noisy in high-dimensional settings, and may not be appropriate for arrays that include discrete or ordinal measurements. This article develops methodology to obtain low-rank model-based representations of continuous, discrete and ordinal data arrays. The model is based on a parameterization of the mean array as a multilinear product of a reduced-rank core array and a set of index-specific orthogonal eigenvector matrices. It is shown how orthogonally equivariant parameter estimates can be obtained from Bayesian procedures under invariant prior distributions. Additionally, priors on the core array are developed that act as regularizers, leading to improved inference over the standard least-squares estimator, and providing robustness to misspecification of the array rank. This model-based approach is extended to accommodate discrete or ordinal data arrays using a semiparametric transformation model. The resulting low-rank representation is scale-free, in the sense that it is invariant to monotonic transformations of the data array. In an example analysis of a multivariate discrete network dataset, this scale-free approach provides a more complete description of data patterns.

Keywords: factor analysis, rank likelihood, social network, tensor, Tucker product.

1 Introduction

Many datasets are naturally represented as multiway arrays, often referred to as tensors. For example, data gathered under all combinations of levels of three conditions can be expressed as a three-way array $\mathbf{Y} = \{y_{i,j,k} : i \in \{1,\ldots,n_1\}, j \in \{1,\ldots,n_2\}, k \in \{1,\ldots,n_3\}\}$. The index sets are referred to as the modes of the array, and an array with K modes is typically referred to as a K-way array. Such array-valued datasets are common in several disciplines, including chemometrics, signal processing and psychometrics. Another class of array-valued data includes multivariate relational networks, which consist of several types of relational measurements between pairs of nodes. Such a dataset may be represented as a three-way array $\mathbf{Y} \in \mathbb{R}^{n\times n\times p}$, where n is the number of nodes, p is the number of relation types, and the entries of $\mathbf{Y}$ are such that $y_{i,j,k}$ is the value of the kth relation type from node i to j. For example, $y_{i,j,1}$ may give the number of emails sent from person i to person j and $y_{i,j,2}$ may encode an evaluation of i's friendship to j measured on an ordinal scale. In this case, the three modes of the array correspond to the initiator of the relation, the target of the relation and the relation type, respectively.

*Departments of Statistics and Biostatistics, University of Washington, Seattle, WA 98195-4322, pdhoff@uw.edu

 DOI: 10.1214/14-BA934

A common framework for the analysis of array-valued data is a model of the form $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, where $\mathbf{Y}$ is the observed array, $\mathbf{M}$ is a mean array describing a signal of interest, and $\mathbf{E}$ is patternless noise. In many applications, it is assumed that $\mathbf{M}$ is low-dimensional or of low rank, and it is desirable to estimate $\mathbf{M}$ under such an assumption. In other applications, the modeling goal is to decompose the data into interpretable sources of variation. In either case, a useful class of tools for describing heterogeneity in array-valued datasets are array decompositions. One category of decompositions are the "Tucker decompositions" (Tucker, 1964, 1966; Kolda and Bader, 2009), which express a K-way data array $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{S} \times \{\mathbf{U}_1, \ldots, \mathbf{U}_K\}$, where $\mathbf{S}$ is a K-way core array, "$\times$" is a a multilinear operator known as the Tucker product and $\{\mathbf{U}_1, \ldots, \mathbf{U}_K\}$ is a collection of mode-specific factor matrices. De Lathauwer et al. (2000) study a particular type of Tucker decomposition in which the $\mathbf{U}_k$'s are orthogonal, and argue that this "higher-order" singular value decomposition (HOSVD) is a natural extension of the matrix SVD to arrays, with the core array $\mathbf{S}$ playing a role analogous to that of the singular values of a matrix. Data analysis based on this decomposition often proceeds by obtaining a low-rank representation of $\mathbf{Y}$ either via truncation of the core array or with a least-squares approximation, and then using its mode-specific singular vectors to describe the heterogeneity in the entries of $\mathbf{Y}$ along each of its K modes.

While providing a relatively simple approach to exploratory data-analysis, least-squares methods may be limited in terms of their performance and applicability. For example, least-squares methods tend to be noisy in multiparameter estimation problems, leading many researchers to favor regularized procedures instead. Recent work on the analysis of matrix-valued datasets indicates that soft-thresholding the singular values of a data matrix can lead to improved estimation of its mean matrix as compared to a least-squares approach (Mazumder et al., 2010; Cai et al., 2010; Josse and Sardy, 2013). Penalized approaches have also been studied in the context of array-valued data: Recent theoretical work has focused on array completion problems, in which the task is to recover a reduced-rank array based on random linear combinations of its elements (Liu et al., 2009; Mu et al., 2013). The algorithms studied typically involve finding the minimum rank among arrays that match the data at the observed entries. Variants of these procedures include finding arrays that minimize different criteria while still matching the observed data, or by minimizing a residual sum of squares subject to a penalty on the fitted array (Tomioka et al., 2011).

However, such approximations of the raw data may be inappropriate when the data are binary, ordinal or otherwise non-normally distributed. For example, Section 5 of this article considers an analysis of skewed, discrete multivariate relational data. These data, obtained from the GDELT project (Leetaru and Schrodt (2013), gdelt.utdallas.edu), consist of weekly summaries of 20 different types of actions between the 30 most active countries in the GDELT database in 2012. These data can be represented as a $30 \times 30 \times 52 \times 20$ four-way array $\mathbf{Y}$, with entries $\{y_{i,j,k,t} : 1 \leq i, j \leq 30, i \neq j, 1 \leq k \leq 20, 1 \leq t \leq 52\}$, where $y_{i,j,k,t}$ is the number of days in week t in which country i took action k with country j as the target. A least-squares approximation to these data is problematic for several reasons, one of which is that such an approximation predominantly represents the small number of large entries of the array, and is therefore unrepresentative of "most" of the data.

As an alternative to least-squares procedures for estimation of a low-rank mean array $\mathbf{M}$, this article develops a model-based version of a penalized Tucker decomposition, and an extension that can accommodate the analysis of discrete, ordinal or otherwise non-normal data. This approach is distinct from existing least-squares and model-based methods in the following ways: First, in contrast to least-squares or non-model-based approaches, this model-based approach allows for adaptive penalization of mode-specific eigenvalues, a complete inferential framework (allowing, for example, confidence interval construction), and extension to non-normal data structures. In contrast to existing model-based approaches, we derive our procedures using decision-theoretic considerations, which lead to a generalized Bayes framework using invariant prior distributions on the scale parameter and the orthogonal factor matrices $\{\mathbf{U}_1, \ldots, \mathbf{U}_K\}$. In this framework, the factor matrices are orthogonal as in the HOSVD of De Lathauwer et al. (2000). This is unlike existing model-based approaches such as in Chu and Ghahramani (2009), who present a Tucker decomposition model and prior in which the core array $\mathbf{S}$ and factor matrices $\{\mathbf{U}_1, \ldots, \mathbf{U}_K\}$ all have i.i.d. standard normal entries (further resulting in inference that is not scale equivariant). Additional identifiability considerations lead to a particular form for a prior distribution over the core array $\mathbf{S}$. This prior allows for mode-specific penalization of the singular values, and also has an interpretation as a version of normal factor analysis for array-valued data.

The work presented here is related to some recently developed statistical models that make use of the multilinear Tucker product. The core array $\mathbf{S}$ is penalized using a class of array normal distributions, generated by the multilinear Tucker product (Hoff, 2011). Xu et al. (2012) develop a prior over the array normal model in which the mode-specific covariance matrices are functions of a potentially infinite set of latent features. In a similar vein, Fosdick and Hoff (2014) develop a version of factor analysis based on the array normal model. The Tucker product has also been used to construct priors in applications where it is the parameters in the model that are arrays: Bhattacharya and Dunson (2012) use a Tucker product to develop a prior over probability distributions for multivariate categorical data, and Volfovsky and Hoff (2014) use a collection of connected array normal distributions as a prior over parameter arrays in ANOVA decompositions. Regarding penalization, Allen (2012) has proposed a sparsity penalty on the factor matrices of a Tucker decomposition, thereby encouraging zeros in their entries. While appropriate in some applications, procedures based on such a sparsity penalty will not be orthogonally equivariant. In contrast, the uniform priors on the factor matrices used in this article lead to orthogonally equivariant estimates, and penalization is focused on the core array in order to encourage low-rank approximations to the data.

An outline of this paper is as follows: The next section provides a brief review of array rank and Tucker decompositions. In Section 3, a parameterization of the Tucker decomposition model is presented, along with a class of prior distributions that allow for equivariant estimation of the model parameters. Section 4 develops a subclass of priors that allows for mode-specific penalization of the singular values. In a simulation study, estimates obtained using such prior distributions are shown to greatly outperform the popular least-squares approach to estimation. Additionally, the proposed approach performs as well as an "oracle" prior when no mode-specific penalization is warranted,

and greatly outperforms such a prior when the rank of the model is misspecified. This methodology is extended in Section 5 to accommodate discrete, ordinal and non-normal data via a semiparametric transformation model, allowing for scale-free reduced-rank representations of array data of diverse types. This extension is illustrated with an analysis of discrete multivariate international relations data, for which a least-squares approach is shown to be largely uninformative. Some additional model extensions are discussed in Section 6, including an approach to accommodate continuous but heavy-tailed data.

2 Review of array rank and Tucker decompositions

Recall that the rank of a matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ is equal to the dimension of the linear space spanned by the columns (or rows) of $\mathbf{M}$. Now suppose $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is a three-way array, with elements $\{m_{i,j,k} : 1 \leq i \leq n_1, 1 \leq j \leq n_2, 1 \leq k \leq n_3\}$. The notion of array rank considered by Tucker (1964), De Lathauwer et al. (2000) and others is defined by the ranks of various reshapings of $\mathbf{M}$ into matrices, called *matricizations*. For example, the mode-1 matricization $\mathbf{M}_{(1)}$ of $\mathbf{M}$ is the $n_1 \times (n_2 n_3)$ matrix having column vectors of the form $\mathbf{m}_{j,k} = (m_{1,j,k}, \ldots, m_{n_1,j,k})^T$, that is, elements of $\mathbf{M}$ with varying values of the first index and fixed values of the second and third indices. Heterogeneity in the values of $\mathbf{M}$ ascribable to heterogeneity in the first index set can be described in terms of the linear space spanned by the columns of $\mathbf{M}_{(1)}$. The dimension r_1 of this linear space (which is equal to the rank of $\mathbf{M}_{(1)}$) is called the *mode-1 rank* of $\mathbf{M}$. The mode-2 and mode-3 matricizations of $\mathbf{M}$ can be formed similarly, and their ranks provide the mode-2 rank r_2 and mode-3 rank r_3, respectively. The array rank of $\mathbf{M}$ is the vector $\mathbf{r} = (r_1, r_2, r_3)$, and is sometimes referred to as the multilinear rank. Unlike the row and column ranks of a matrix, the ranks corresponding to the different modes of an array are not generally equal.

Any matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ can be expressed in terms of its SVD $\mathbf{M} = \mathbf{U}_1 \mathbf{S} \mathbf{U}_2^T$ where $\mathbf{S} = \text{diag}(s_1, \ldots, s_r)$, $\mathbf{U}_1 \in \mathcal{V}_{r,n_1}$, $\mathbf{U}_2 \in \mathcal{V}_{r,n_2}$ and $r \leq n_1 \wedge n_2$ is the rank of $\mathbf{M}$. Here, $\mathcal{V}_{r,n}$ is the space of $n \times r$ matrices with orthonormal columns, known as the Stiefel manifold. As shown by De Lathauwer et al. (2000), an analogous representation holds for any array. The analogy is most easily seen via vectorization: The SVD of a matrix $\mathbf{M}$ yields a representation of $\mathbf{m} = \text{vec}(\mathbf{M})$ as $\mathbf{m} = (\mathbf{U}_2 \otimes \mathbf{U}_1)\,\mathbf{s}$, where $\mathbf{s} = \text{vec}(\mathbf{S})$ and "$\otimes$" is the Kronecker product. Similarly, every K-way array $\mathbf{M}$ of dimension $n_1 \times \cdots \times n_K$ and rank $\mathbf{r} = (r_1, \ldots, r_K)$ can be expressed as

$$\mathbf{m} = (\mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_1)\,\mathbf{s}, \tag{1}$$

where $\mathbf{m}$ is the vectorization of $\mathbf{M}$, $\mathbf{U}_k \in \mathcal{V}_{r_k,n_k}$ for $k \in \{1, \ldots, K\}$ and $\mathbf{s}$ is the vectorization of an $r_1 \times \cdots \times r_K$ array $\mathbf{S}$ known as the "core array." This representation is often referred to as the higher-order SVD (HOSVD). More generally, any representation of $\mathbf{m}$ of the form (1), without $\mathbf{U}_1, \ldots, \mathbf{U}_K$ necessarily being orthogonal, is known as a "Tucker decomposition."

An equivalent representation of $\mathbf{M}$ that retains its array structure is obtained using the so-called "Tucker product" (Tucker, 1964) of the core array $\mathbf{S}$ with the list of factor

matrices $\mathbf{U}_1, \ldots, \mathbf{U}_K$. This representation expresses $\mathbf{M}$ as

$$\mathbf{M} = \mathbf{S} \times \{\mathbf{U}_1, \ldots, \mathbf{U}_K\}, \tag{2}$$

where the Tucker product "$\times$" is defined by the equivalence between (1) and (2). More generally, For $\mathbf{A} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$, $\mathbf{B} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ and $\mathbf{C}_k \in \mathbb{R}^{n_k \times r_k}$, $k = 1, \ldots, K$, $\mathbf{A} = \mathbf{B} \times \{\mathbf{C}_1, \ldots, \mathbf{C}_K\}$ means that $\mathrm{vec}(\mathbf{A}) = (\mathbf{C}_K \otimes \cdots \otimes \mathbf{C}_1)\,\mathrm{vec}(\mathbf{B})$.

For the calculations that follow it will be useful to re-express a Tucker decomposition of $\mathbf{M}$ in terms of its matricizations. If $\mathbf{M}$ can be expressed as in (1) or (2), then it also follows that for each $k \in \{1, \ldots, K\}$,

$$\mathbf{M}_{(k)} = \mathbf{U}_k\, \mathbf{S}_{(k)}\, (\mathbf{U}_K \otimes \cdots \mathbf{U}_{k+1} \otimes \mathbf{U}_{k-1} \otimes \cdots \otimes \mathbf{U}_1)^T \equiv \mathbf{U}_k\, \mathbf{S}_{(k)}\, \mathbf{U}_{-k}^T, \tag{3}$$

where $\mathbf{M}_{(k)}$ and $\mathbf{S}_{(k)}$ are the mode-k matricizations of $\mathbf{M}$ and $\mathbf{S}$, respectively.

3 A model-based Tucker decomposition for arrays

A commonly used model of low-dimensional structure for a matrix-valued dataset $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$ is that $\mathbf{Y}$ is equal to some mean matrix $\mathbf{M}$ of rank $r < n_1 \wedge n_2$, plus an error matrix $\sigma\mathbf{E}$ having i.i.d. mean-zero entries with variance σ^2. Let $\mathbf{M} = \mathbf{U}_1\mathbf{D}\mathbf{U}_2^T$ be the SVD of $\mathbf{M}$ and $\mathbf{S} = \mathbf{D}/\sigma$ be the singular values scaled by the error standard deviation σ. This model can be parameterized as $\mathbf{Y} = \sigma\mathbf{U}_1\mathbf{S}\mathbf{U}_2^T + \sigma\mathbf{E}$, or alternatively in vector form as $\mathbf{y} = \sigma(\mathbf{U}_2 \otimes \mathbf{U}_1)\,\mathbf{s} + \sigma\mathbf{e}$, where $\mathbf{y}$, $\mathbf{s}$ and $\mathbf{e}$ are the vectorizations of $\mathbf{Y}$, $\mathbf{S}$ and $\mathbf{E}$ respectively.

Now consider an analogous model for an array $\mathbf{Y} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$. As in the matrix case, the model is $\mathbf{Y} = \mathbf{M} + \sigma\mathbf{E}$, where $\mathbf{M}$ is an array with array rank $\mathbf{r}$ and $\mathbf{E}$ is a mean-zero error array. Equation (1) says that this model can be expressed as $\mathbf{y} = \sigma(\mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_1)\,\mathbf{s} + \sigma\mathbf{e}$, where $\mathbf{s} \in \mathbb{R}^{r_1 \cdots r_K}$ and $\mathbf{U}_k \in \mathcal{V}_{r_k, n_k}$ for each $k = 1, \ldots, K$. An equivalent representation in terms of the Tucker product is that $\mathbf{Y} = \sigma\mathbf{S} \times \{\mathbf{U}_1, \ldots, \mathbf{U}_K\} + \sigma\mathbf{E}$. This section discusses estimation of the unknown parameters $(\sigma, \mathbf{U}, \mathbf{S})$ in this Tucker decomposition model (TDM) when the error $\mathbf{E}$ is assumed to consist of i.i.d. standard normal random variables. Results on optimal equivariant estimation in the case that $\mathbf{S}$ is known are used to motivate certain priors for equivariant Bayesian inference in the more realistic case that $\mathbf{S}$ is unknown. It is shown that posterior inference under such prior distributions can be made with a relatively straightforward Markov chain Monte Carlo (MCMC) algorithm based on Gibbs sampling.

3.1 Equivariant estimation

First consider the (unrealistic) case that the core array $\mathbf{S}$ is known. Letting $n = n_1 \cdots n_K$, $r = r_1 \cdots r_K$ and $\mathcal{U} = \{\mathbf{U} : \mathbf{U} = \mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_1, \mathbf{U}_k \in \mathcal{V}_{r_k, n_k}\}$, the normal TDM can be expressed as

$$\mathbf{y} = \sigma\mathbf{U}\mathbf{s} + \sigma\mathbf{e},\ \ \mathbf{e} \sim N_n(\mathbf{0}, \mathbf{I}),\ \ (\sigma, \mathbf{U}) \in \mathbb{R}^+ \times \mathcal{U}. \tag{4}$$

Let $\mathcal{W} = \{\mathbf{W} : \mathbf{W} = \mathbf{W}_K \otimes \cdots \otimes \mathbf{W}_1, \mathbf{W}_k \in \mathcal{O}_{n_k}\}$ be the space of Kronecker products of orthogonal matrices, and note that $\mathbf{WU} \in \mathcal{U}$ for all $\mathbf{W} \in \mathcal{W}$ and $\mathbf{U} \in \mathcal{U}$. It follows that the model (4) is invariant under the group of transformations on $\mathcal{Y}$ given by $\mathcal{G} = \{g : \mathbf{y} \to a\mathbf{W}\mathbf{y}, a > 0, \mathbf{W} \in \mathcal{W}\}$, which induces a group $\bar{\mathcal{G}}$ on the parameter space given by $\bar{\mathcal{G}} = \{\bar{g} : (\sigma, \mathbf{U}) \to (a\sigma, \mathbf{WU})\}$. This motivates the use of equivariant estimators of σ and $\mathbf{U}$. For example, it is natural to prefer estimators such that $\hat{\sigma}(a\mathbf{W}\mathbf{y}) = a\hat{\sigma}(\mathbf{y})$, so that a scale change to the data result in the same change to the estimate of the scale parameter σ. Similarly, one may prefer estimators of $\mathbf{U}$ such that $\hat{\mathbf{U}}(a\mathbf{W}\mathbf{y}) = \mathbf{W}\hat{\mathbf{U}}(\mathbf{y})$ and estimators of $\mathbf{m} = \sigma\mathbf{U}\mathbf{s}$ such that $\hat{\mathbf{m}}(a\mathbf{W}\mathbf{y}) = a\mathbf{W}\hat{\mathbf{m}}(\mathbf{y})$.

As with many invariant statistical models, risk-optimal equivariant decision rules can be obtained as Bayes rules under a prior distribution derived from the group:

Proposition 1. *Let $\theta = (\sigma, \mathbf{U})$ and $\Theta = \mathbb{R}^+ \times \mathcal{U}$. Under any invariant loss function $L(d, \theta)$ the minimum risk equivariant decision rule $\delta(\mathbf{y})$ is given for each $\mathbf{y}$ by the minimizer in d of*

$$\int L(d, \theta) p(\mathbf{y}|\theta) \pi_I(d\theta),$$

where for measurable sets $A \subset \mathbb{R}^+$ and $B \subset \mathcal{U}$, $\pi_I(A \times B) = \pi_\sigma(A) \times \pi_U(B)$, with $\pi_\sigma(A) = \int_A \sigma^{-1} d\sigma$ and π_U corresponding to the (proper) probability distribution of $\mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_1$ when each $\mathbf{U}_k$ is uniformly distributed on $\mathcal{V}_{r_k, n_k}$.

This result is an application of more general results from invariant decision theory (a proof is in the Appendix). To put the result more simply, optimal equivariant decision rules can be obtained from the posterior distribution of $(\sigma, \mathbf{U})$ under an improper prior for σ with density $1/\sigma$ and independent uniform priors for $\mathbf{U}_1, \ldots, \mathbf{U}_K$. In what follows, π_σ and π_U will refer to either these measures or their densities, depending on the context.

Unfortunately, uniformly optimal equivariant decision rules no longer exist under this group when the core array $\mathbf{s}$ is unknown, as the best equivariant rule will depend on $\mathbf{s}$. This article focuses attention on Bayesian inference for $(\sigma, \mathbf{U}, \mathbf{s})$ using prior distributions with densities of the form $\pi(\sigma, \mathbf{U}, \mathbf{s}) = \pi_\sigma(\sigma)\pi_U(\mathbf{U})\pi_s(\mathbf{s})$, where $\pi_s(\mathbf{s})$ is a proper probability density on $\mathbb{R}^{r_1 \cdots r_K}$ for given ranks $r_1, \ldots, r_K$. Although not corresponding to a proper joint prior distribution (because of the improper prior on σ), such densities can be used to construct proper posterior distributions that provide estimates of functions of $(\sigma, \mathbf{U}, \mathbf{s})$ that are equivariant with respect to $\mathcal{G}$ and $\bar{\mathcal{G}} = \{\bar{g} : (\sigma, \mathbf{U}, \mathbf{s}) \to (a\sigma, \mathbf{WU}, \mathbf{s})\}$. Addressing the propriety of such a posterior first, for each $\mathbf{y} \in \mathbb{R}^n$ define a function $f(\sigma, \mathbf{U}, \mathbf{s} : \mathbf{y})$ so that $f(\sigma, \mathbf{U}, \mathbf{s} : \mathbf{y}) \propto p(\mathbf{y}|\sigma, \mathbf{U}, \mathbf{s}) \times \pi(\sigma, \mathbf{U}, \mathbf{s})$, where $p(\mathbf{y}|\sigma, \mathbf{U}, \mathbf{s})$ is the normal sampling density of $\mathbf{y}$, having mean $\sigma\mathbf{U}\mathbf{s}$ and variance $\sigma^2\mathbf{I}$. If f is integrable in $(\sigma, \mathbf{U}, \mathbf{s})$ for the observed value of $\mathbf{y}$, a "posterior" probability distribution can be defined via the density

$$\pi(\sigma, \mathbf{U}, \mathbf{s}|\mathbf{y}) = \frac{f(\sigma, \mathbf{U}, \mathbf{s} : \mathbf{y})}{\int f(\sigma, \mathbf{U}, \mathbf{s} : \mathbf{y})\, d\sigma d\mathbf{U} d\mathbf{s}}\,. \tag{5}$$

That f is generally integrable can be seen by first integrating with respect to σ:

$$\int_0^\infty f(\sigma, \mathbf{U}, \mathbf{s} : \mathbf{y})\, d\sigma = \pi_U(\mathbf{U})\pi_s(\mathbf{s}) \int_0^\infty p(\mathbf{y}|\sigma, \mathbf{U}, \mathbf{s})\pi_\sigma(\sigma)\, d\sigma$$

$$= \pi_U(\mathbf{U})\pi_s(\mathbf{s}) \int_0^\infty (2\pi)^{-n/2}\sigma^{-n-1}\exp(-\sigma^{-2}\|\mathbf{y}-\mathbf{U}\mathbf{s}\|^2/2)\,d\sigma$$
$$= \pi_U(\mathbf{U})\pi_s(\mathbf{s}) \times [\tfrac{1}{2}\pi^{-n/2}\Gamma(n/2)] \times \|\mathbf{y}-\mathbf{U}\mathbf{s}\|^{-n}\,.$$

Now $\|\mathbf{y}-\mathbf{U}\mathbf{s}\| \geq \|\mathbf{y}-\hat{\mathbf{m}}\|$, where $\hat{\mathbf{m}}$ is the least squares estimate of $\mathbf{m}$. Since $\hat{\mathbf{m}}$ is of reduced rank, $\|\mathbf{y}-\hat{\mathbf{m}}\| > 0$ unless the array rank of $\mathbf{y}$ is less than or equal to that of the fitted rank. Presuming this is not the case, it follows that $\|\mathbf{y}-\mathbf{U}\mathbf{s}\|^{-n}$ is bounded above by $\|\mathbf{y}-\hat{\mathbf{m}}\|^{-n}$. Since the priors for $\mathbf{U}$ and $\mathbf{s}$ are proper, the integral of $\|\mathbf{y}-\mathbf{U}\mathbf{s}\|^{-n}$ with respect to $\pi_U(\mathbf{U})$ and $\pi_s(\mathbf{s})$ is finite and so (5) is a proper probability density.

As stated above, the decision rules obtained from such a posterior are not globally risk optimal among equivariant rules, as optimal rules for $(\sigma, \mathbf{U})$ depend on the unknown value of $\mathbf{s}$. However, such posterior distributions still provide equivariant inference in the following sense:

Proposition 2. *Let the prior for $\theta = (\sigma, \mathbf{U}, \mathbf{s})$ be such that the marginal prior for $(\sigma, \mathbf{U})$ is the invariant prior π_I and $\mathbf{s}$ is independent of $(\sigma, \mathbf{U})$. Then for any $a > 0$, $\mathbf{W} \in \mathcal{W}$ and functions $g : \mathbf{y} \to a\mathbf{W}\mathbf{y}$ and $\bar{g} : (\sigma, \mathbf{U}, \mathbf{s}) \to (a\sigma, \mathbf{W}\mathbf{U}, \mathbf{s})$,*

$$\Pr(\theta \in A|\mathbf{y}) = \Pr(\theta \in \bar{g}A|g\mathbf{y})$$

for all measurable subsets A of $\mathbb{R}^+ \times \mathcal{U} \times \mathbb{R}^r$.

A proof is in the Appendix. The result says that, using such a prior, the belief that the correct θ-value is in A having observed $\mathbf{y}$ is the same as the belief that the correct θ-value is in $\bar{g}A$ having observed $g\mathbf{y}$.

3.2 Posterior approximation via the Gibbs sampler

The results in the previous subsection hold as long as $\mathbf{s}$ is *a priori* independent of σ and $\mathbf{U}$ and the prior for $\mathbf{s}$ is proper. The remainder of the article focuses attention on normal priors for $\mathbf{s}$, so that the joint prior distribution of $(\sigma, \mathbf{U}, \mathbf{s})$ has a density of the form $\pi(\sigma, \mathbf{U}, \mathbf{s}) = \pi_I(\sigma, \mathbf{U}) \times \pi_s(\mathbf{s})$, where π_I is density of the invariant prior discussed previously and π_s is a zero-mean multivariate normal prior with covariance matrix $\mathbf{\Psi}$. Not only are such priors for $\mathbf{s}$ computationally convenient, but they lead to an interpretation of the model as a multiway extension to a normal factor analysis model, as will be discussed in the next section.

Posterior inference under such a prior can be made via a reasonably straightforward Gibbs sampling algorithm that approximates the posterior distribution of $(\sigma^2, \mathbf{U}, \mathbf{s})$ given $\mathbf{y}$. The algorithm proceeds by iteratively updating the values of these parameters as follows:

1. Simulate $(\sigma^2, \mathbf{s})$ from $\pi(\sigma^2, \mathbf{s}|\mathbf{y}, \mathbf{U})$ as follows:
   (a) simulate σ^2 from $\pi(\sigma^2|\mathbf{y}, \mathbf{U})$, an inverse-Gamma distribution;
   (b) simulate $\mathbf{s}$ from $\pi(\mathbf{s}|\mathbf{y}, \mathbf{U}, \sigma^2)$, a multivariate normal distribution.

2. For $k \in \{1, \ldots, K\}$, simulate $\mathbf{U}_k$ from $\pi(\mathbf{U}_k|\mathbf{y}, \mathbf{s}, \{\mathbf{U}_j : j \neq k\}, \sigma^2)$, a von Mises–Fisher distribution on $\mathcal{V}_{r_k,n_k}$.

Repeated iteration of the above procedure generates a Markov chain whose stationary distribution is the posterior distribution of $(\sigma^2, \mathbf{U}, \mathbf{s})$ given $\mathbf{y}$.

Full conditional distribution of $(\sigma^2, \mathbf{s})$. Recall that the model for $\mathbf{y}$ is $\mathbf{y} = \sigma\mathbf{U}\mathbf{s} + \sigma\mathbf{e}$, $\mathbf{e} \sim N_n(\mathbf{0}, \mathbf{I})$, where $n = \prod n_k$. The normal prior $\mathbf{s} \sim N_r(\mathbf{0}, \boldsymbol{\Psi})$ implies that, unconditionally on $\mathbf{s}$, $\mathbf{y}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix

$$\mathrm{E}[\mathbf{y}\mathbf{y}^T|\mathbf{U}, \sigma] = \sigma^2\mathrm{E}[\mathbf{U}\mathbf{s}\mathbf{s}^T\mathbf{U}^T + \mathbf{e}\mathbf{e}^T + 2\mathbf{U}\mathbf{s}\mathbf{e}^T] = \sigma^2\left(\mathbf{U}\boldsymbol{\Psi}\mathbf{U}^T + \mathbf{I}\right).$$

Based on this result, standard calculations show that the conditional distribution of σ^2 used in step 1 of the above algorithm is an inverse-gamma$(n/2, \mathbf{y}^T(\mathbf{U}\boldsymbol{\Psi}\mathbf{U}^T + \mathbf{I})^{-1}\mathbf{y}/2)$ distribution. Now given σ and $\mathbf{U}$, the model can be expressed as $\mathbf{y}/\sigma = \mathbf{U}\mathbf{s} + \mathbf{e}$ where the entries of $\mathbf{e}$ are i.i.d. standard normal random variables. This has the same form as a regression model with $\mathbf{s}$ playing the role of the vector of unknown regression coefficients. Combining this "regression likelihood" with the normal prior $\mathbf{s} \sim N_r(\mathbf{0}, \boldsymbol{\Psi})$ gives a normal full conditional distribution for $\mathbf{s}$ with mean and variance given as follows:

$$\mathrm{Var}[\mathbf{s}|\mathbf{y}, \mathbf{U}, \sigma^2, \boldsymbol{\Psi}] = \tilde{\boldsymbol{\Psi}} = (\boldsymbol{\Psi}^{-1} + \mathbf{I})^{-1}, \qquad \mathrm{E}[\mathbf{s}|\mathbf{y}, \mathbf{U}, \sigma^2, \boldsymbol{\Psi}] = \tilde{\boldsymbol{\Psi}}\mathbf{U}^T\mathbf{y}/\sigma.$$

The next section discusses specification and estimation of $\boldsymbol{\Psi}$, and its relationship to the mode-specific singular values of the mean array $\mathbf{M}$.

Full conditional distribution of $\mathbf{U}$: Let $\mathbf{Y}_{(1)}$, $\mathbf{S}_{(1)}$ and $\mathbf{E}_{(1)}$ be the mode-1 matricizations of the arrays $\mathbf{Y}$, $\mathbf{S}$ and $\mathbf{E}$, respectively. The model can then be written as $\mathbf{Y}_{(1)}/\sigma = \mathbf{U}_1\mathbf{S}_{(1)}\mathbf{U}_{-1}^T + \mathbf{E}_{(1)}$ where $\mathbf{U}_{-1} = (\mathbf{U}_K \otimes \cdots \otimes \mathbf{U}_2)$ and the elements of $\mathbf{E}_{(1)}$ are i.i.d. standard normal random variables. Since the prior for $\mathbf{U}_1$ is the uniform distribution on $\mathcal{V}_{r_1,m_1}$, its full conditional distribution is proportional to the density of $\mathbf{Y}_{(1)}$:

$$\begin{aligned}\pi(\mathbf{U}_1|\ldots) \propto_{\mathbf{U}_1} p(\mathbf{Y}_{(1)}|\mathbf{S}, \mathbf{U}, \sigma_e^2) &\propto_{\mathbf{U}_1} \exp(-\tfrac{1}{2}\|\mathbf{Y}_{(1)}/\sigma - \mathbf{U}_1\mathbf{S}_{(1)}\mathbf{U}_{-1}^T\|^2) \\ &\propto_{\mathbf{U}_1} \mathrm{etr}(\mathbf{U}_1^T\mathbf{Y}_{(1)}\mathbf{U}_{-1}\mathbf{S}_{(1)}^T)/\sigma) \equiv \mathrm{etr}(\mathbf{U}_1^T\mathbf{H})\end{aligned}$$

where $\mathbf{H} = \mathbf{Y}_{(1)}\mathbf{U}_{-1}\mathbf{S}_{(1)}^T/\sigma$, and $\mathrm{etr}(\mathbf{A})$ is $\exp(\mathrm{trace}(\mathbf{A}))$. This is proportional to the matrix-variate von Mises–Fisher distribution vMF$(\mathbf{H})$ on $\mathcal{V}_{r_1,m_1}$. An algorithm for direct simulation from vMF$(\mathbf{H})$ is described in Hoff (2009). The full conditional distributions of $\mathbf{U}_2, \ldots, \mathbf{U}_K$ can be derived analogously.

4 Estimation of $\boldsymbol{\Psi}$

The covariance matrix $\boldsymbol{\Psi}$ of the core array $\mathbf{S}$ can be viewed as a description of the scale of $\mathbf{M}$ relative to the scale σ of the error, or alternatively, as a penalty on the magnitude

of $\mathbf{S}$ that serves to provide a regularized estimator of the mean array $\mathbf{M} = \sigma\mathbf{S} \times \mathbf{U}$. In practice, an appropriate value of $\boldsymbol{\Psi}$ may not be known in advance, and therefore must be estimated from the data. This section discusses estimation of $\boldsymbol{\Psi}$ in the context of two models for $\mathbf{S}$. The first of these is simply that $\text{vec}(\mathbf{S}) = \mathbf{s} \sim N_r(\mathbf{0}, \tau^2\mathbf{I})$, where τ^2 is a scale parameter to be estimated. In a simulation study, it is shown that this model provides better estimates of $\mathbf{M}$ than those obtained by minimizing the residual sum of squares. However, this simple covariance model shrinks all values of $\mathbf{S}$ equally, and does not recognize the array structure of $\mathbf{S}$. As an alternative to this homoscedastic i.i.d. model, a heteroscedastic separable variance model is developed, of the form $\text{Cov}[\mathbf{s}] = \tau^2\boldsymbol{\Lambda}_K \otimes \cdots \otimes \boldsymbol{\Lambda}_1$, where each $\boldsymbol{\Lambda}_k$ is a diagonal matrix with positive entries that sum to 1. Such a model allows for separate penalization of the mode-specific eigenvalues of the array $\mathbf{M}$. Such penalization is useful when it is feared that the fitted rank $\mathbf{r}$ is larger than the actual rank of the mean array for some of the modes. In such cases, it is desirable to have a procedure that can shrink the estimate of $\mathbf{M}$ towards arrays with lower mode-specific ranks. This section first derives this heteroscedastic model and provides some interpretation of the parameters, and then illustrates in a simulation study how estimators based on this model can shrink towards low-rank solutions when the fitted rank is too large.

4.1 Derivation and interpretation of the heteroscedastic model

Even if $\mathbf{s}$ were observed, unrestricted estimation of $\boldsymbol{\Psi}$ based on the model $\mathbf{s} \sim N_r(\mathbf{0}, \boldsymbol{\Psi})$ would be problematic, as $\mathbf{s}$ corresponds to only a single realization from the $N_r(\mathbf{0}, \boldsymbol{\Psi})$ distribution. Instead, consider first estimation of $\boldsymbol{\Psi}$ restricted to the class of separable covariance matrices, so that $\boldsymbol{\Psi} = \boldsymbol{\Psi}_K \otimes \cdots \otimes \boldsymbol{\Psi}_1$, where each $\boldsymbol{\Psi}_k$ is an $r_k \times r_k$ positive definite matrix. Now recall that marginally over $\mathbf{s}$, the distribution for $\mathbf{y} = \text{vec}(\mathbf{Y})$ is a mean-zero n-variate normal distribution with covariance matrix proportional to $\mathbf{U}\boldsymbol{\Psi}\mathbf{U}^T + \mathbf{I}$. As $\mathbf{U}$ and $\boldsymbol{\Psi}$ are both separable, it follows that

$$\text{Cov}[\mathbf{y}|\sigma, \mathbf{U}, \boldsymbol{\Psi}]/\sigma^2 = \mathbf{U}\boldsymbol{\Psi}\mathbf{U}^T + \mathbf{I} = (\mathbf{U}_K\boldsymbol{\Psi}_K\mathbf{U}_K^T \otimes \cdots \otimes \mathbf{U}_1\boldsymbol{\Psi}_1\mathbf{U}_1^T) + \mathbf{I}. \tag{6}$$

This covariance model is similar to that of a factor analysis model, in which the covariance matrix is represented as a reduced-rank positive semidefinite matrix plus a full-rank diagonal matrix of positive entries. As with factor analysis, the covariance model above is not identifiable unless restrictions are placed on the $\boldsymbol{\Psi}_k$'s. First, the eigenvectors of each $\boldsymbol{\Psi}_k$ are not identifiable: If $\boldsymbol{\Psi}_k = \mathbf{V}_k\boldsymbol{\Lambda}_k\mathbf{V}_k^T$ is the eigendecomposition of $\boldsymbol{\Psi}_k$, then $\mathbf{U}_k\boldsymbol{\Psi}_k\mathbf{U}_k^T = \tilde{\mathbf{U}}_k\boldsymbol{\Lambda}_k\tilde{\mathbf{U}}_k{}^T$, where $\tilde{\mathbf{U}}_k = \mathbf{U}_k\mathbf{V}_k^T \in \mathcal{V}_{r_k,n_k}$. Second, the scales of the $\boldsymbol{\Psi}_k$'s are not separately identifiable: For example, replacement of $(\boldsymbol{\Psi}_{k_1}, \boldsymbol{\Psi}_{k_2})$ with $(c\boldsymbol{\Psi}_{k_1}, \boldsymbol{\Psi}_{k_2}/c)$ does not change the covariance matrix. With this in mind, $\boldsymbol{\Psi}$ is parameterized as $\boldsymbol{\Psi} = \tau^2(\boldsymbol{\Lambda}_K \otimes \cdots \otimes \boldsymbol{\Lambda}_1)$ where $\tau^2 > 0$ and for each k, $\boldsymbol{\Lambda}_k$ is an $r_k \times r_k$ diagonal matrix of positive entries that sum to 1.

The parameters $\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K$ can be interpreted in terms of the prior or penalty they induce over the mode-specific eigenvalues of the mean array $\mathbf{M} = \sigma\mathbf{S} \times \mathbf{U}$. These eigenvalues are often of interest in multiway data analysis as they describe the extent to which the variation along a mode can be attributed to a small set of orthogonal

factors. To relate these eigenvalues to the $\mathbf{\Lambda}_k$'s, recall that $\mathbf{M}_{(1)} = \sigma \mathbf{U}_1 \mathbf{S}_{(1)} \mathbf{U}_{(-1)}^T$, and so $\mathbf{M}_{(1)}\mathbf{M}_{(1)}^T = \sigma^2 \mathbf{U}_1 \mathbf{S}_{(1)} \mathbf{S}_{(1)}^T \mathbf{U}_1^T$. Now $\mathbf{S}_{(1)}$ is equal in distribution to $\tau \mathbf{\Lambda}_1^{1/2} \mathbf{Z} \mathbf{\Lambda}_{-1}^{1/2}$, where $\mathbf{\Lambda}_{-1} = \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_2$ and $\mathbf{Z}$ is an $r_1 \times r_{-1}$ matrix of independent standard normal entries. This gives

$$\begin{aligned} \mathrm{E}[\mathbf{M}_{(1)}\mathbf{M}_{(1)}^T] &= \sigma^2\tau^2 \mathbf{U}_1 \mathbf{\Lambda}_1^{1/2} \mathrm{E}[\mathbf{Z}\mathbf{\Lambda}_{-1}\mathbf{Z}^T] \mathbf{\Lambda}_1^{1/2} \mathbf{U}_1^T \\ &= \sigma^2\tau^2 \mathbf{U}_1 \mathbf{\Lambda}_1^{1/2} (\mathrm{tr}(\mathbf{\Lambda}_{-1})\mathbf{I}) \mathbf{\Lambda}_1^{1/2} \mathbf{U}_1^T = \sigma^2\tau^2 \mathbf{U}_1 \mathbf{\Lambda}_1 \mathbf{U}_1^T, \end{aligned} \tag{7}$$

where the last calculation follows because the sum of the entries of each $\mathbf{\Lambda}_k$ is 1, making $\mathrm{tr}(\mathbf{\Lambda}_{-1}) = 1$. Based on this calculation for $\mathbf{M}_{(1)}$ (and analogous calculations for the other $\mathbf{M}_{(k)}$'s), τ^2 is seen to be the expected squared magnitude of the mean array $\mathbf{M}$ relative to the error variance σ^2, and each $\mathbf{\Lambda}_k$ is the (scaled) diagonal eigenvalue matrix of $\mathrm{E}[\mathbf{M}_{(k)}\mathbf{M}_{(k)}^T]$. Additionally, if one or more of the diagonal elements of $\mathbf{\Lambda}_k$ are very close to zero, then $\mathbf{M}_{(k)}$ will be very close to a matrix of rank less than r_k.

The separable model $\mathbf{s} \sim N_r(\mathbf{0}, \tau^2 \mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1)$ also provides a link to the parameterization of the core array used in the HOSVD of De Lathauwer et al. (2000). In this latter approach, the core $\mathbf{S}$ of the data array $\mathbf{Y}$ has the property of "all-orthogonality", in that for each k, $\mathbf{S}_{(k)}\mathbf{S}_{(k)}^T$ is a diagonal matrix whose elements can be thought of as the mode-k eigenvalues of $\mathbf{Y}$. Similarly, the separable covariance model proposed here for the core $\mathbf{S}$ of the mean array $\mathbf{M}$ has the property of all-orthogonality in expectation: For each $k = 1, \ldots, K$, $\mathrm{E}[\mathbf{S}_{(k)}\mathbf{S}_{(k)}^T] = \tau^2 \mathbf{\Lambda}_k$, a diagonal matrix. From (7), $\mathbf{\Lambda}_k$ can be viewed as the eigenvalues of the expected sum of squares matrix $\mathrm{E}[\mathbf{M}_{(k)}\mathbf{M}_{(k)}^T]$, or alternatively as the mode-k eigenvalues in the marginal covariance model for $\mathbf{Y}$ given in (6).

4.2 Simulation study

A natural estimator of the reduced-rank mean array $\mathbf{M}$ based on the data array $\mathbf{Y}$ is the minimizer of the residual sum of squares $\|\mathbf{Y} - \mathbf{M}\|^2$. If $K > 2$ the least-squares estimator of $\mathbf{M}$ is not available in closed form, and so standard practice is to obtain a local minimizer $\hat{\mathbf{M}}_{\mathrm{ALS}}$ via an alternating least-squares (ALS) algorithm. The algorithm minimizes the sum of squares iteratively in the mode-specific eigenvectors of $\mathbf{M}$, a process that has been called "higher order orthogonal iteration" (HOOI) (De Lathauwer et al., 2000).

One might anticipate that estimates of the mean array $\mathbf{M}$ based on the homoscedastic model for $\mathbf{S}$, in which $\mathbf{s} \sim N_r(\mathbf{0}, \tau^2\mathbf{I})$, will outperform $\hat{\mathbf{M}}_{\mathrm{ALS}}$ due to the ability of the former to shrink the values of $\mathbf{S}$ and the tendency of least-squares estimators to overfit, particularly for large values of $\mathbf{r}$. It might be further anticipated that the heteroscedastic covariance model for $\mathbf{S}$, in which $\mathbf{s} \sim N_r(\mathbf{0}, \tau^2(\mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1))$, will outperform the homoscedastic model when $\mathbf{r}$ is chosen to be too large, as the heteroscedastic model allows for mode-specific shrinkage of the mean array towards estimates of lower rank. However, such desirable performance in the case of a misspecified rank may come at the expense of poorer performance when the rank is correctly specified.

These possibilities were investigated with a simulation study comparing three different estimators of the mean array $\mathbf{M}$:

1. $\hat{\mathbf{M}}_{\text{ALS}}$, obtained with the ALS algorithm;
2. $\hat{\mathbf{M}}_{\text{HOM}}$, the posterior mean under the homoscedastic model $\mathbf{s} \sim N_r(\mathbf{0}, \tau^2\mathbf{I})$;
3. $\hat{\mathbf{M}}_{\text{HET}}$, the posterior mean under the heteroscedastic model $\mathbf{s} \sim N_r(\mathbf{0}, \tau^2\mathbf{\Lambda}_K \otimes \cdots \otimes \mathbf{\Lambda}_1)$.

The Bayes estimator $\hat{\mathbf{M}}_{\text{HOM}}$ was obtained using a conjugate inverse-gamma$(\nu_0/2, \tau_0^2/2)$ prior for τ^2, where $\nu_0 = 1$ and $\tau_0^2 = \prod_{k=1}^K n_k/r_k$. This value of τ_0^2 makes the expected prior magnitude of the mean array equal to that of the error, so that $\text{E}[\|\mathbf{M}\|^2] = \text{E}[\|\mathbf{E}\|^2]$ *a priori*. The Bayes estimator $\hat{\mathbf{M}}_{\text{HET}}$ was obtained under a prior on $(\tau^2, \mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_K)$ in which τ^2 has an inverse-gamma$(1/2, \tau_0^2/2)$ distribution and the diagonal elements of each $\mathbf{\Lambda}_k$ are uniform on the r_k-dimensional simplex. The value of $\tau_0^2 = \prod_{k=1}^K n_k$ was chosen so that $\text{E}[\|\mathbf{M}\|^2] = \text{E}[\|\mathbf{E}\|^2]$ *a priori*, as with the prior used to obtain $\hat{\mathbf{M}}_{\text{HOM}}$. The uniform priors on the $\mathbf{\Lambda}_k$'s are not conjugate, and so the Markov chain for posterior estimation in this model relies on a Metropolis–Hastings update for these parameters.

Three-dimensional data arrays $\mathbf{Y} \in \mathbb{R}^{60\times 50\times 40}$ were simulated according to the following procedure: For a given rank vector $\mathbf{r}_0 = (r_{01}, r_{02}, r_{03})$,

1. Simulate $\mathbf{U}_k \sim \text{uniform}(\mathcal{V}_{r_{0k},n_k})$ for each $k \in \{1, 2, 3\}$;
2. Simulate $\mathbf{s} \sim N_r(\mathbf{0}, \psi \times \left(\prod_{k=1}^K r_{0k}^2\right)^{-1/3} \times \mathbf{I})$;
3. Let $\mathbf{M} = \mathbf{S} \times \{\mathbf{U}_1, \ldots, \mathbf{U}_K\}$, where $\text{vec}(\mathbf{S}) = \mathbf{s}$;
4. Let $\mathbf{Y} = \mathbf{M} + \mathbf{E}$, where $\mathbf{E}$ has i.i.d. standard normal entries.

Data were generated under two values of $\mathbf{r}_0$ and two values of ψ for a total of four different conditions. The values of $\mathbf{r}_0$ included a "low-rank" condition $\mathbf{r}_0 = (6, 5, 4)$ and a "high-rank" condition $\mathbf{r}_0 = (30, 25, 20)$, and the values of ψ included a "low-signal" condition $\psi = 1000$ and a "high-signal" condition $\psi = 2000$. Ten datasets were generated under each of these four conditions, for a total of forty simulated datasets. For each dataset, $\hat{\mathbf{M}}_{\text{ALS}}$, $\hat{\mathbf{M}}_{\text{HOM}}$ and $\hat{\mathbf{M}}_{\text{HET}}$ were obtained with the assumed rank $\mathbf{r}$ equal to the true rank $\mathbf{r}_0$. Each Bayesian estimate was obtained via 11,000 iterations of the MCMC algorithm described in the previous section. The first 1000 iterations of each Markov chain were dropped to allow for convergence to the stationary distribution, and parameter values were saved every 10th iteration thereafter, resulting in 1000 simulated values of $\mathbf{M}$ with which to approximate its posterior mean. Convergence and mixing of the Markov chains were monitored via traceplots of the simulated values of σ^2 and τ^2, as well as their effective sample sizes, which roughly measure the approximation variability of the posterior mean estimates relative to those that would be obtained from independent Monte Carlo simulations. Effective sample sizes for σ^2 and τ^2 were above 300 for all scenarios and datasets, and close to half the Markov chains attained the maximum possible value of 1000.

rank	$\mathbf{r}_0 = (6,5,4)$		$\mathbf{r}_0 = (30,25,20)$	
signal	low	high	low	high
RSE($\hat{\mathbf{M}}_{\text{ALS}}$)	0.195	0.088	0.848	0.379
RSE($\hat{\mathbf{M}}_{\text{HOM}}$)	0.165	0.082	0.485	0.280
RSE($\hat{\mathbf{M}}_{\text{HET}}$)	0.165	0.082	0.489	0.281

Table 1: Relative squared estimation errors.

rank	$\mathbf{r}_0 = (6,5,4)$		$\mathbf{r}_0 = (30,25,20)$	
signal	low	high	low	high
RSE($\hat{\mathbf{M}}_{\text{ALS}}$)	0.855	0.404	4.840	2.420
RSE($\hat{\mathbf{M}}_{\text{HOM}}$)	0.260	0.141	1.364	0.840
RSE($\hat{\mathbf{M}}_{\text{HET}}$)	0.166	0.082	0.495	0.284

Table 2: Relative squared estimation errors when the fitted rank is twice that of $\mathbf{r}_0$.

For each estimator and each simulation condition, a relative squared estimation error (RSE) was computed by averaging the value of $\|\mathbf{M}-\hat{\mathbf{M}}\|^2/\|\mathbf{M}\|^2$ across the 10 datasets. These values are given in Table 1. Note that $\hat{\mathbf{M}}_{\text{HOM}}$ is to some extent an "oracle" estimator, in that it is based on a prior distribution that was used to simulate the data (although $\hat{\mathbf{M}}_{\text{HOM}}$ requires estimation of τ^2). Nevertheless, in the low-rank case ($\mathbf{r}_0 = (6,5,4)$), the two Bayes estimators performed nearly identically in terms of RSE, and the ALS estimator performed slightly worse. In terms of variability across datasets, $\hat{\mathbf{M}}_{\text{HOM}}$ outperformed $\hat{\mathbf{M}}_{\text{ALS}}$ for all datasets, and outperformed $\hat{\mathbf{M}}_{\text{HET}}$ in 10 of the 20 datasets. The story is similar for the 20 high-rank datasets ($\mathbf{r}_0 = (30,25,20)$), except that ALS performs more poorly in this case than in the low-rank case, presumably because of the much larger number of parameters and the general tendency of least-squares estimators to overfit the data. Regarding this, the residual squared error $\|\mathbf{Y}-\hat{\mathbf{M}}\|^2$ was lower for the ALS estimator than the Bayes estimators across all datasets and scenarios.

For the same 40 simulated datasets, estimates $\hat{\mathbf{M}}_{\text{ALS}}$, $\hat{\mathbf{M}}_{\text{HOM}}$ and $\hat{\mathbf{M}}_{\text{HET}}$ were also obtained using a fitted rank of $\mathbf{r} = 2 \times \mathbf{r}_0$, that is, twice the actual rank of $\mathbf{M}$. Note that in the high-rank scenario the fitted rank is $\mathbf{r} = (60,50,40)$, which is the dimension of the data array. In this case, the estimates are of full rank and so in particular the ALS estimate is simply $\mathbf{Y}$. Also, the Bayes estimates in this full rank case were obtained using a proper gamma$(1/2,1/2)$ prior distribution for σ^2 to guarantee the propriety of the posterior (recall the discussion in Section 2). Relative squared errors (RSEs) for these misspecified-rank estimators are given in Table 2. Not surprisingly, $\hat{\mathbf{M}}_{\text{ALS}}$ performs poorly across all scenarios, and roughly 4 to 6 times worse than it does when the rank is correctly specified. The Bayes estimator $\hat{\mathbf{M}}_{\text{HOM}}$ performs reasonably well in the low-rank scenario, but roughly 3 times worse than it does in the high-rank scenario with correctly specified rank. In contrast, the performance of $\hat{\mathbf{M}}_{\text{HET}}$ with a misspecified rank is nearly identical to its performance with a correctly specified rank. This suggests that the heteroscedastic model for $\mathbf{S}$ is able to shrink the estimate of $\mathbf{M}$ towards arrays of the correct rank.

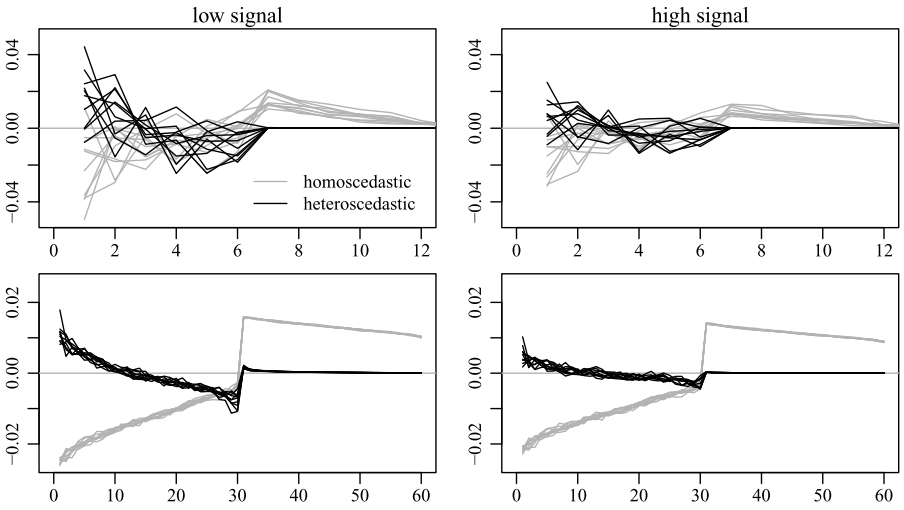

Figure 1: Difference of eigenvalues between $\hat{\mathbf{M}}_{(1)}\hat{\mathbf{M}}_{(1)}^T$ and $\mathbf{M}_{(1)}\mathbf{M}_{(1)}^T$. Estimates in the first row are based on a true rank of $\mathbf{r}_0 = (6, 5, 4)$ and a fitted rank of $\mathbf{r} = (12, 10, 8)$. Estimates in the second row are based on a true rank of $\mathbf{r}_0 = (30, 25, 20)$ and a fitted rank of $\mathbf{r} = (60, 50, 40)$.

This is explored further in Figure 1. For each Bayesian estimate $\hat{\mathbf{M}}$ obtained with a misspecified rank, its mode-1 matricization $\hat{\mathbf{M}}_{(1)}$ was constructed and the normalized eigenvalues of $\hat{\mathbf{M}}_{(1)}\hat{\mathbf{M}}_{(1)}^T$ were computed, from which the normalized eigenvalues of $\mathbf{M}_{(1)}\mathbf{M}_{(1)}^T$ were subtracted off, where $\mathbf{M}_{(1)}$ is the mode-1 matricization of the true mean array $\mathbf{M}$. These eigenvalue differences are plotted across datasets and conditions in Figure 1. For example, the plot in the upper-left corner of the figure shows results under the low-signal low-rank condition, for which the true rank is $\mathbf{r} = (6, 5, 4)$ but the fitted rank is $\mathbf{r} = (12, 10, 8)$. Each black line corresponds to the eigenvalues of $\hat{\mathbf{M}}_{(1)}\hat{\mathbf{M}}_{(1)}^T$ obtained under the heteroscedastic model minus the eigenvalues of $\mathbf{M}_{(1)}\mathbf{M}_{(1)}^T$, for one of the 10 simulated datasets. The gray lines correspond to the analogous differences under the homoscedastic model. The results indicate that the homoscedastic model generally underestimates non-zero eigenvalues and substantially overestimates zero eigenvalues. In contrast, the heteroscedastic model generally does a very good job of estimating the zero eigenvalues as being very nearly zero. However, for the non-zero eigenvalues, the estimated eigenvalues for the the heteroscedastic model are somewhat too "steep", overestimating the true large non-zero eigenvalues and underestimating the small non-zero eigenvalues. A larger signal appears to ameliorate these biases, as the differences between estimated and true eigenvalues is diminished in going from the low-signal to the high-signal scenario. However, the presence of such biases suggests exploration of more complex adaptive penalties or hierarchical priors, i.e., ones that could more flexibly adapt to the shape of the eigenspectra in the observed data. For example, a beta(a, b) prior over the diagonal elements of $\mathbf{\Lambda}_k$ could be used instead of the uniform prior. However, in the absence of prior information about the eigenspectra, the values of a and b

would need to be obtained from the data. Such an empirical Bayes approach would be similar in spirit to the two-parameter matrix regularizer of Josse and Sardy (2013).

Regarding computation time, iteration of the ALS algorithm is generally faster than iteration of the Gibbs sampler. For example, for the low-rank scenario ($\mathbf{r}_0 = \mathbf{r} = (6, 5, 4)$), each scan of the Gibbs sampler is about 2.5 times slower than each iteration of the ALS algorithm (about 0.26 seconds versus 0.11 seconds in R on a desktop computer). For the high-rank scenario ($\mathbf{r}_0 = \mathbf{r} = (30, 25, 20)$), the Gibbs sampler is a little over four times slower. The increase is due to the fact that simulation of the components of $\mathbf{U}$ from their full conditional distribution involves a rejection sampler for each column of each $\mathbf{U}_k$ matrix. While the per-iteration computational burdens of these two approaches are comparable, obtaining a precise approximation to the posterior distribution of $\mathbf{M}$ will generally take substantially longer than obtaining an approximate least-squares estimate: The former may require thousands of iterations of the Gibbs sampler, while the latter generally requires an order of magnitude fewer iterations of the ALS algorithm. However, if the goal is only to obtain a posterior mean estimate of $\mathbf{M}$, a substantially shortened Gibbs sampler may be sufficient: In the low-rank case, a 10,000-iteration Gibbs sampler gave an RSE of $\hat{\mathbf{M}}_{\text{HET}}$ that was about 15% lower than that of $\hat{\mathbf{M}}_{\text{ALS}}$ (see Table 1). However, a Gibbs sampler with only 1,000 iterations provided an RSE for $\hat{\mathbf{M}}_{\text{HET}}$ that was 14% lower than that of $\hat{\mathbf{M}}_{\text{ALS}}$. In other words, most of the improvement of $\hat{\mathbf{M}}_{\text{HET}}$ over $\hat{\mathbf{M}}_{\text{ALS}}$ can be obtained using a relatively short Gibbs sampler.

5 A scale-free Tucker decomposition model

In this section the TDM is extended in order to analyze data arrays for which the assumption of normally distributed errors is inappropriate. The approach presented is based upon a transformation model in which the observed data array is modeled as an unknown increasing function of a latent array that follows a normal TDM. The model fitting procedure provides parameter estimates that are invariant to monotonic transformations of the data array, thereby giving a "scale-free" TDM. This approach is motivated and illustrated with an analysis of discrete multivariate data on relations between countries in the year 2012.

5.1 Data description

The motivating application of this section is to obtain a low-rank representation of a relational dataset on actions between countries, available from the GDELT project (gdeltproject.org). The data analyzed consist of a weekly summary of 20 different types of actions between the 30 most active countries in the GDELT database in 2012. These data can be represented as a $30 \times 30 \times 20 \times 52$ four-way array $\mathbf{Y}$, with entries $\{y_{i,j,k,t} : 1 \leq i, j \leq 30, i \neq j, 1 \leq k \leq 20, 1 \leq t \leq 52\}$ where $y_{i,j,k,t}$ is the number of days in week t in which country i took action k with country j as the target. The types of actions include "positive" actions such as diplomatic cooperation and the provision of aid, as well as "negative" actions such as the expression of disapproval, military threats and military conflict (a list of the action types is given in Table 3). Figure 2 provides a

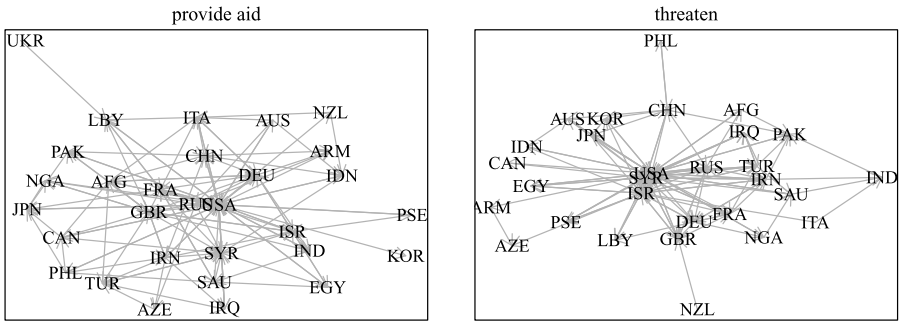

Figure 2: Networks corresponding to two of the twenty action types.

graphical summary of the array $\mathbf{Y}$ for two of the twenty action types. To construct this figure, counts for each of the action types between each ordered pair of countries were summed across the 52 weeks of the year and then dichotomized, so that a link between two countries indicates the presence of the action type for at least one day of the year.

The data array $\mathbf{Y}$ has nearly one million entries but is very sparse, with just over 2% of the entries being non-zero. This sparsity varies by action type from a high of about 12% for the action "consult" to a low of less than 0.01% for the action "use unconventional mass violence." Sparsity also varies considerably by country: The first panel of Figure 3 plots outdegrees and indegrees of each country, computed (for country i) as $\sum_{jkt} y_{i,j,k,t}$ and $\sum_{jkt} y_{j,i,k,t}$, respectively. These two measures of activity are highly correlated across countries, with Syria being somewhat of an outlier, being the target of more actions than it initiates. Additionally, the counts for each action are highly skewed: There are more counts of zero than counts of one, more counts of one than counts of two, and so on. This is illustrated in the second panel of Figure 3, which gives the empirical distribution of the non-zero entries of $\mathbf{Y}$.

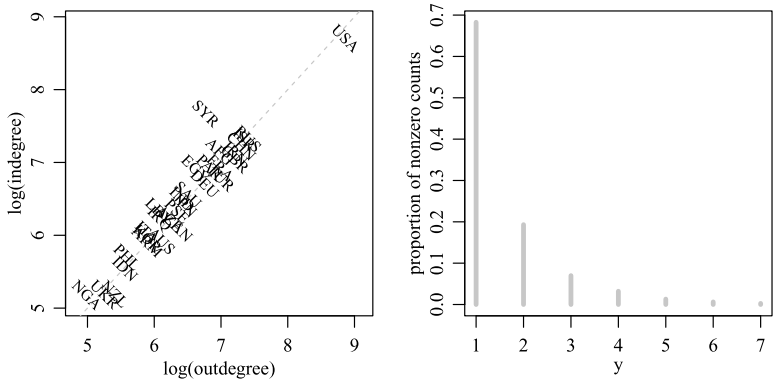

Figure 3: Descriptive data plots. The left panel shows country-specific outdegrees and indegrees on the log-scale. The right panel gives a histogram of the non-zero action counts.

5.2 Scale free TDM

Existing array decomposition methods applied directly to these data would be problematic for several reasons. One particular issue in applying matrix or array decomposition methods to relational datasets is that self-relations are typically undefined, that is, $y_{i,i,k,t}$ is not defined for any i, k or t. This issue can be addressed via an alternating least-squares algorithm that iterates between fitting a reduced-rank model and replacing any missing values with fitted values (see, for example, Ward and Hoff (2007) for details on such an algorithm applied to matrix-valued relational data). A more serious problem is that the discrete or ordinal nature of many relational datasets makes least-squares methods of limited use. For example, as will be illustrated at the end of this section, a reduced rank representation of the GDELT data array $\mathbf{Y}$ obtained via alternating least squares generally represents the largest data values at the expense of other interesting features of the data.

While the normal TDM model presented in the previous section may not be appropriate for ordinal or discrete data, the normal model can be extended to accommodate such data via a latent variable formulation, in which the entries of $\mathbf{Y}$ are modeled as a non-decreasing function of the elements of a latent array $\mathbf{Z}$ that follows the Tucker decomposition model. If the elements of $\mathbf{Y}$ take on a known finite number of possible values, then such an approach can be viewed as similar to an ordered probit model.

In many datasets one of the indices of the array $\mathbf{Y}$ represents variables having different scales. For example, the large heterogeneity in sparsity between the 20 different action types in the GDELT dataset suggests modeling the different types on different scales. As another example, consider an $n \times n \times 2$ relational array where $y_{i,j,1}$ is the number of emails sent from person i to person j, and $y_{i,j,2}$ encodes an evaluation of i's friendship to j on an ordinal scale. In such a case, it may not make sense to model $y_{i,j,1}$ and $y_{i,j,2}$ as the same transformation of the latent variables $z_{i,j,1}$ and $z_{i,j,2}$. In particular, the number of levels of the two variables may be different. For cases such as these, a more appropriate transformation model may be one with with variable-specific transformations, so that

$$\mathbf{Z} = \mathbf{S} \times \{\mathbf{U}_1, \ldots, \mathbf{U}_K\} + \mathbf{E}, \quad \text{vec}(\mathbf{E}) \sim N_n(\mathbf{0}, \mathbf{I}), \quad y_{\boldsymbol{i},j} = g_j(z_{\boldsymbol{i},j}), \tag{8}$$

where $\boldsymbol{i} \in \{1, \ldots, n_1\} \times \cdots \times \{1, \ldots, n_{K-1}\}$, $j \in \{1, \ldots, n_K\}$, and for notational convenience the variables to be modeled on different scales are indexed by the Kth mode of the array. Note that the scale parameter σ from the TDM in the previous sections would be confounded with the transformations $g_1, \ldots, g_{n_K}$, and so can be set to 1.

In the case that the transformations $g_1, \ldots, g_{n_K}$ are nuisance parameters, scale-free estimation of $(\mathbf{S}, \mathbf{U})$ can be obtained using a rank likelihood L_R, defined as

$$L_R(\mathbf{S}, \mathbf{U} : \mathbf{Y}) = \Pr(\mathbf{Z} \in R(\mathbf{Y}) | \mathbf{S}, \mathbf{U}),$$

where $R(\mathbf{Y})$ is the set of $\mathbf{Z}$-values consistent with the observed data $\mathbf{Y}$ and the fact that the functions $g_1, \ldots, g_{n_K}$ are non-decreasing. This set can be expressed as

$$R(\mathbf{Y}) = \{\mathbf{Z} : \max\{z_{\boldsymbol{i}',j} : y_{\boldsymbol{i}',j} < y_{\boldsymbol{i},j}\} < z_{\boldsymbol{i},j} < \min\{z_{\boldsymbol{i}',j} : y_{\boldsymbol{i},j} < y_{\boldsymbol{i}',j}\}\}.$$

A feature of estimates obtained from the rank likelihood is that they are scale-free: The set $R(\mathbf{Y})$ is invariant to strictly increasing transformations of the data, and therefore so is L_R.

While maximum likelihood estimation using the rank likelihood is generally computationally intractable, Bayesian inference using this likelihood is feasible via the Gibbs sampler (see Hoff (2007) and Hoff (2008) for applications of the rank likelihood to semiparametric copula and regression models, respectively). Under a prior distribution for $(\mathbf{S}, \mathbf{U})$ from the previous section, posterior estimates for this scale-free TDM can be obtained via a simple extension of the previous algorithm. The extended algorithm can be roughly understood as follows: If $\mathbf{Z}$ were observed, parameter estimates could be obtained from the MCMC algorithm for the normal TDM. As $\mathbf{Z}$ is not observed, the algorithm requires additional steps in order to integrate over the possible values of $\mathbf{Z}$. This can be done by simulating values of the elements of $\mathbf{Z}$ from their full conditional distributions at each step of the Markov chain. Specifically, posterior approximation for this scale-free TDM can proceed by iterating the following steps: Given current values $(\mathbf{Z}, \mathbf{S}, \mathbf{U})$,

1. Update $(\mathbf{S}, \mathbf{U})$ as in the case of the normal TDM, with $\mathbf{Z}$ taking on the role of $\mathbf{Y}$;
2. Update the elements of $\mathbf{Z}$ given $\mathbf{Y}$, $\mathbf{S}$ and $\mathbf{U}$ as follows:
   (a) Compute $\mathbf{M} = \mathbf{S} \times \{\mathbf{U}_1, \ldots, \mathbf{U}_K\}$;
   (b) Simulate each $z_{\boldsymbol{i},j}$ from the constrained normal$(m_{\boldsymbol{i},j}, 1)$ distribution, constrained so that $\max\{z_{\boldsymbol{i}',j} : y_{\boldsymbol{i}',j} < y_{\boldsymbol{i},j}\} < z_{\boldsymbol{i},j} < \min\{z_{\boldsymbol{i}',j} : y_{\boldsymbol{i},j} < y_{\boldsymbol{i}',j}\}$.

Iteration of steps 1 and 2 generates a Markov chain, samples from which approximate the posterior distribution proportional to $L_R(\mathbf{S}, \mathbf{U} : \mathbf{Y}) \times \pi(\mathbf{S}, \mathbf{U})$. As mentioned above, parameter estimates obtained from this posterior distribution are invariant to monotonic transformations of each variable along the Kth mode of the array. For this reason, this estimation procedure and the resulting estimates can be referred to as a scale-free Tucker decomposition (SFTD).

5.3 Analysis of GDELT data

A rank $\mathbf{r} = (4, 4, 4, 4)$ representation of the GDELT data was obtained from the SFTD procedure described above using the heteroscedastic prior described in Section 4 and modeling the 20 action types on different scales. A rank of 4 for each mode was chosen because of the substantial heterogeneity in the degrees as displayed in the first panel of Figure 3. A standard approach to representing such heterogeneity would be with an additive model in which the entries of $\mathbf{M}$ are expressed as the sum of mode-specific effects, for example, $m_{i,j,k,t} = a_i + b_j + c_k + d_t$. Such an additive effects model has a rank of $(2, 2, 2, 2)$. A rank $(4, 4, 4, 4)$ approximation was fit to $\mathbf{Y}$ in order to capture the rank $(2, 2, 2, 2)$ additive effects along with two additional dimensions of non-additive data patterns, which are shown below.

The MCMC algorithm described above was run for 55,000 iterations. The first 5,000 iterations were dropped to allow for convergence, and parameter values were saved every 10th iteration thereafter. This resulted in 5,000 simulated values of the parameters with which to approximate posterior quantities of interest. Mixing of the Markov chain was evaluated with traceplots and effective sample sizes of τ^2 and the eigenvalue parameters $\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_4$. The effective sample size for τ^2 was 1197. Effective sample sizes for the eigenvalues ranged between 371 and 1266, with a mean of 678. The first eigenvalues of the third and fourth modes (corresponding to action type and week) are close to one with high posterior probability, meaning that $\mathbf{M}_{(3)}$ and $\mathbf{M}_{(4)}$ are both close to being rank-1 matrices. Eigenspectra of the first and second modes (corresponding to initiators and targets of the actions) were more evenly distributed. For both of these two modes, the first two eigenvectors predominantly represented the heterogeneity in outdegrees and indegrees. To examine non-additive patterns in the data, the posterior mean array $\hat{\mathbf{M}}$ was centered along each index of each mode, creating an array $\tilde{\mathbf{M}}$ representing the non-additive patterns in the data.

The first two left singular vectors of $\tilde{\mathbf{M}}_{(1)}$, $\tilde{\mathbf{M}}_{(2)}$ and $\tilde{\mathbf{M}}_{(3)}$ are displayed in Figure 4. The first two plots indicate strong geographic patterns in the first two modes of $\tilde{\mathbf{M}}$. These patterns indicate that, after accounting for additive effects, countries that have similar patterns of activity in the dataset are typically close to one another geographically. The converse is not generally true: PSE and ISR are far apart from SYR, IRQ and IRN on the plot, indicating heterogeneity in the dataset that is non-geographic. The third plot in Figure 4 displays the singular vectors of $\tilde{\mathbf{M}}_{(3)}$ corresponding to the different action types. Plotting symbols "+" and "−" are used to indicate actions that are categorized as "positive" or "negative", respectively. The singular vectors of $\tilde{\mathbf{M}}_{(3)}$ distinguish somewhat the two types of actions, but there is considerable overlap. This is not too surprising, since countries that interact frequently with each other generally relate both positively and negatively during the course of the year.

The utility of the SFTD in comparison to a least-squares approach can be seen by contrasting this scale-free representation of $\mathbf{Y}$ given in Figure 4 to an analogous least-squares representation shown in in Figure 5. This plot gives the first two singular vectors

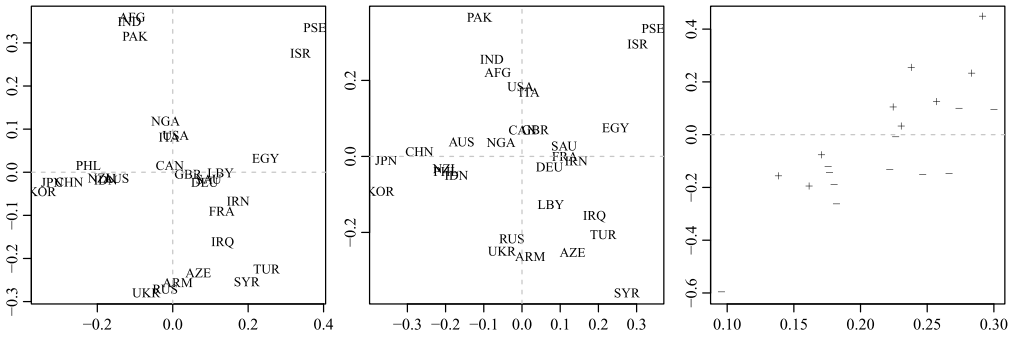

Figure 4: The first two left singular vectors of $\tilde{\mathbf{M}}_{(1)}$, $\tilde{\mathbf{M}}_{(2)}$ and $\tilde{\mathbf{M}}_{(3)}$, from the SFTD of $\mathbf{Y}$.

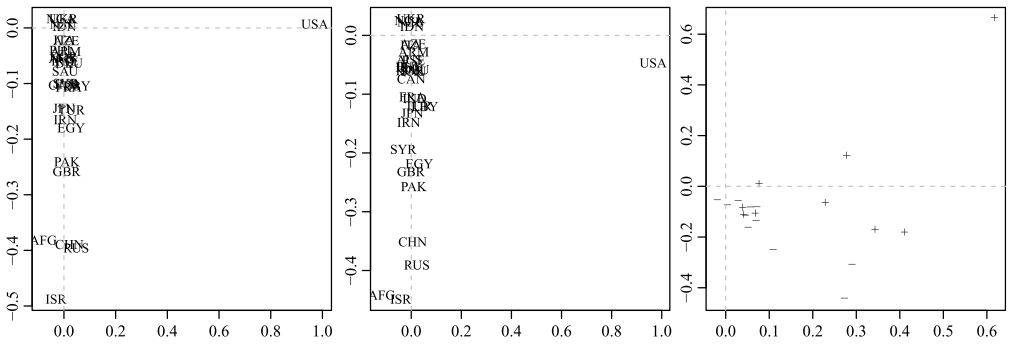

Figure 5: The first two left singular vectors of $\tilde{\mathbf{M}}_{\mathrm{ALS}(1)}$, $\tilde{\mathbf{M}}_{\mathrm{ALS}(2)}$, and $\tilde{\mathbf{M}}_{\mathrm{ALS}(3)}$.

of the first three modes of $\tilde{\mathbf{M}}_{\mathrm{ALS}}$, where $\tilde{\mathbf{M}}_{\mathrm{ALS}}$ was constructed as with the SFTD except using a rank (4,4,4,4) alternating least-squares approximation $\hat{\mathbf{M}}_{\mathrm{ALS}}$ to $\mathbf{Y}$ instead of the posterior mean array $\hat{\mathbf{M}}$. The least squares approach is primarily identifying the countries that have the most number of data values of 7 (the highest value possible), at the expense of representing the other patterns in the data. For example, the first singular vectors of the both the first- and second-mode matricizations of $\tilde{\mathbf{M}}_{\mathrm{ALS}}$ are essentially devoted to distinguishing the USA from the other countries.

The posterior mean array $\hat{\mathbf{M}}$ and the least squares representation $\hat{\mathbf{M}}_{\mathrm{ALS}}$ can also be evaluated in terms of how well they represent the rank ordering of the values of $\mathbf{Y}$. This is done by computing Kendall's τ, a scale-free measure of association, between the entries of $\mathbf{Y}$ and each of the two low-rank representations $\hat{\mathbf{M}}$ and $\hat{\mathbf{M}}_{\mathrm{ALS}}$. This is done separately for each of the 20 action types in order to evaluate any heterogeneity in performance. As shown in Table 3, the SFTD representation has a higher degree of association with the ranks of $\mathbf{Y}$ than the least-squares representation for all action types. This is perhaps not too surprising – the SFTD is inherently scale-free, and so $\hat{\mathbf{M}}$ is only representing information about the rank ordering of the entries of $\mathbf{Y}$. In contrast, $\hat{\mathbf{M}}_{\mathrm{ALS}}$ must also represent differences in magnitude. For these highly skewed data, a good representation of large differences in magnitude comes at the cost of a poorer representation of small differences, which constitute most of the differences in the entries of $\mathbf{Y}$.

action	$\hat{\mathbf{M}}_{\mathrm{ALS}}$	$\hat{\mathbf{M}}$	action	$\hat{\mathbf{M}}_{\mathrm{ALS}}$	$\hat{\mathbf{M}}$	action	$\hat{\mathbf{M}}_{\mathrm{ALS}}$	$\hat{\mathbf{M}}$
statement	0.66	0.74	yield	0.65	0.77	exhibit force	0.76	0.88
appeal	0.65	0.75	investigate	0.63	0.75	reduce relations	0.61	0.77
cooperative intent	0.63	0.7	demand	0.68	0.84	coerce	0.57	0.69
consult	0.58	0.65	disapprove	0.69	0.8	assault	0.58	0.76
diplomatic coop	0.57	0.67	reject	0.76	0.84	fight	0.65	0.77
material coop	0.58	0.7	threaten	0.65	0.82	mass violence	0.83	0.91
aid	0.66	0.79	protest	0.66	0.81			

Table 3: Kendall's τ measure of association between $\mathbf{Y}$ and $\hat{\mathbf{M}}_{\mathrm{ALS}}$ and $\hat{\mathbf{M}}$.

6 Discussion

While the objectives of an array-valued data analysis may be primarily descriptive, model-based approaches may be appealing for a variety of reasons. For example, regularized data descriptions may be obtained using model-based Bayesian procedures, with the prior acting as a penalty term. This article has developed a parameterization of the normal Tucker decomposition model that allows for scale-equivariant and orthogonally-equivariant estimates and data descriptions, while still allowing for penalization of mode-specific singular values. Such regularized estimates can greatly improve upon least-squares estimates in situations where the data array is equal to a reduced-rank mean array plus noise. Another benefit of the model-based approach is its extensibility to a variety of different data types and data analysis scenarios. For example, the semiparametric transformation model developed in Section 5 provides a scale-free reduced-rank representation for data arrays that consist of discrete, ordinal or other types of measurements for which a least squares criterion is not appropriate.

The Gaussian model described in Sections 3 and 4 can also be extended to accommodate non-normal data that is continuous but heavy-tailed, using scale mixtures of Gaussian error distributions (Fernández and Steel, 2000). Operationally, the error structure is represented as $\mathbf{E} = \mathbf{G} \circ \mathbf{W}$, where $\mathbf{G}$ is a Gaussian array, $\mathbf{W}$ is an array of latent variables and "$\circ$" is the Hadamard (elementwise) product. For example, gamma-distributed entries for $\mathbf{W}$ lead to t-distributed errors. Posterior inference for such a model can be obtained via an additional Gibbs sampling step in the MCMC algorithm presented in Section 4.

An additional extension of the model would be to data analysis situations in which it is desired to account for known explanatory factors or patterns in the data. For example, one extension of the model used to analyze the GDELT data in Section 5 takes the form

$$\begin{aligned} \mathbf{Z} &= \langle \mathbf{X}, \mathbf{B} \rangle + \mathbf{S} \times \{\mathbf{U}_1, \ldots, \mathbf{U}_K\} + \mathbf{E}, \\ \text{vec}(\mathbf{E}) &\sim N(\mathbf{0}, \boldsymbol{\Sigma}(\rho) \otimes \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{I}) \end{aligned}$$

where $\mathbf{X}$ and $\mathbf{B}$ represent arrays of known explanatory variables and unknown regression coefficients, respectively, and $\boldsymbol{\Sigma}(\rho)$ is some simple one-parameter model that accounts for some of the temporal dependence in the data. In such a model, the reduced rank term $\mathbf{S} \times \{\mathbf{U}_1, \ldots, \mathbf{U}_K\}$ would express data patterns not accounted for by $\langle \mathbf{X}, \mathbf{B} \rangle$ or $\boldsymbol{\Sigma}(\rho)$. Bayesian inference for parameters in such a model could be obtained by adding steps to the MCMC algorithm outlined in this article.

Replication code for the results in Sections 4 and 5 is available at the author's website: `www.stat.washington.edu/~hoff`.

References

Allen, G. (2012). "Regularized Tensor Factorizations and Higher-Order Principal Components Analysis." arXiv:1202.2476. 629

Bhattacharya, A. and Dunson, D. B. (2012). "Simplex factor models for multivariate unordered categorical data." *Journal of the American Statistical Association*, 107(497):

362–377. MR2949366. doi: http://dx.doi.org/10.1080/01621459.2011.646934. 629

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). "A singular value thresholding algorithm for matrix completion." *SIAM Journal on Optimization*, 20(4): 1956–1982. MR2600248. doi: http://dx.doi.org/10.1137/080738970. 628

Chu, W. and Ghahramani, Z. (2009). "Probabilistic models for incomplete multi-dimensional arrays." In: *12th International Conference on Artificial Intelligence and Statistics*, volume 5, 89–96. 629

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). "A multilinear singular value decomposition." *SIAM Journal on Matrix Analysis and Applications*, 21(4): 1253–1278. MR1780272. doi: http://dx.doi.org/10.1137/S0895479896305696. 628, 629, 630, 636

Fernández, C. and Steel, M. F. J. (2000). "Bayesian regression analysis with scale mixtures of normals." *Econometric Theory*, 16(1): 80–101. MR1749020. doi: http://dx.doi.org/10.1017/S0266466600161043. 646

Fosdick, B. K. and Hoff, P. D. (2014). "Separable factor analysis with applications to mortality data." *The Annals of Applied Statistics*, 8(1): 120–147. MR3191985. doi: http://dx.doi.org/10.1214/13-AOAS694. 629

Hoff, P. D. (2007). "Extending the rank likelihood for semiparametric copula estimation." *The Annals of Applied Statistics*, 1(1): 265–283. MR2393851. doi: http://dx.doi.org/10.1214/07-AOAS107. 643

— (2008). "Rank Likelihood Estimation for Continuous and Discrete Data." *ISBA Bulletin*, 15(1): 8–10. URL: http://bayesian.org/sites/default/files/fm/bulletins/0803.pdf 643

— (2009). "Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data." *Journal of Computational and Graphical Statistics*, 18(2): 438–456. MR2749840. doi: http://dx.doi.org/10.1198/jcgs.2009.07177. 634

— (2011). "Separable Covariance Arrays Via the Tucker Product, with Applications to Multivariate Relational Data." *Bayesian Analysis*, 6(2): 179–196. MR2806238. doi: http://dx.doi.org/10.1214/11-BA606. 629

Josse, J. and Sardy, S. (2013). "Reduced rank matrix estimation by adaptive trace norm regularization." arXiv:1310.6602. 628, 640

Kolda, T. G. and Bader, B. W. (2009). "Tensor decompositions and applications." *SIAM Review*, 51(3): 455–500. MR2535056. doi: http://dx.doi.org/10.1137/07070111X. 628

Leetaru, K. and Schrodt, P. (2013). "GDELT: Global Data on Events, Language, and Tone, 1979–2012." In: *International Studies Association Annual Conference*. San Diego, CA. 628

Liu, J., Musialski, P., Wonka, P., and Ye, J. (2009). "Tensor completion for estimating missing values in visual data." In: *2009 IEEE 12th International Conference on Computer Vision*, 2114–2121. IEEE. 628

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). "Spectral regularization algorithms for learning large incomplete matrices." *Journal of Machine Learning Research*, 11: 2287–2322. MR2719857. 628

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2013). "Square deal: Lower bounds and improved relaxations for tensor recovery." arXiv:1311.5870. 628

Tomioka, R., Suzuki, T., Hayashi, K., and Kashima, H. (2011). "Statistical performance of convex tensor decomposition." In: *Advances in Neural Information Processing Systems*, 972–980. 628

Tucker, L. (1966). "Some mathematical notes on three-mode factor analysis." *Psychometrika*, 31(3): 279–311. 628

Tucker, L. R. (1964). "The extension of factor analysis to three-dimensional matrices." In: Gulliksen, H. and Frederiksen, N. (eds.), *Contributions to mathematical psychology*, 110–127. New York: Holt, Rinehart and Winston. 628, 630

Volfovsky, A. and Hoff, P. D. (2014). "Hierarchical array priors for ANOVA decompositions of cross-classified data." *The Annalls of Applied Statistics*, 8(1): 19–47. MR3191981. doi: http://dx.doi.org/10.1214/13-AOAS685. 629

Ward, M. D. and Hoff, P. D. (2007). "Persistent patterns of international commerce." *Journal of Peace Research*, 44(2): 157–175. 642

Xu, Z., Yan, F., and Qi, A. (2012). "Infinite Tucker Decomposition: Nonparametric Bayesian Models for Multiway Data Analysis." In: Langford, J. and Pineau, J. (eds.), *Proceedings of the 29th International Conference on Machine Learning*, ICML '12, 1023–1030. 629

Acknowledgments

This research was supported by NI-CHD grant R01HD067509.