HHS Public Access

Author manuscript

J Comput Graph Stat. Author manuscript; available in PMC 2019 June 06.

Published in final edited form as:

J Comput Graph Stat. 2018 ; 27(3): 638–647. doi:10.1080/10618600.2017.1401544.

Tensor-on-tensor regression

Eric F. Lock*

Division of Biostatistics, University of Minnesota

Abstract

We propose a framework for the linear prediction of a multi-way array (i.e., a tensor) from another multi-way array of arbitrary dimension, using the contracted tensor product. This framework generalizes several existing approaches, including methods to predict a scalar outcome from a tensor, a matrix from a matrix, or a tensor from a scalar. We describe an approach that exploits the multiway structure of both the predictors and the outcomes by restricting the coefficients to have reduced CP-rank. We propose a general and efficient algorithm for penalized least-squares estimation, which allows for a ridge (L_2) penalty on the coefficients. The objective is shown to give the mode of a Bayesian posterior, which motivates a Gibbs sampling algorithm for inference. We illustrate the approach with an application to facial image data. An R package is available at https://github.com/lockEF/MultiwayRegression.

Keywords

Multiway data; PARAFAC/CANDECOMP; ridge regression; reduced rank regression

1 Introduction

For many applications data are best represented in the form of a *tensor*, also called a *multi-way* or *multi-dimensional* array, which extends the familiar two-way data matrix (*Samples* × *Variables*) to higher dimensions. Tensors are increasingly encountered in fields that require the automated collection of high-throughput data with complex structure. For example, in molecular “omics” pro ling it is now common to collect high-dimensional data over multiple subjects, tissues, fluids or time points within a single study. For neuroimaging modalities such as fMRI and EEG, data are commonly represented as multi-way arrays with dimensions that can represent subjects, time points, brain regions, or frequencies. In this article we consider an application to a collection of facial images from the Faces in the Wild database (Learned-Miller et al., 2016), which when properly aligned to a 90 90 pixel grid can be represented as a 4-way array with dimension $Faces \times X \times Y \times Colors$, where X and Y give the horizontal and vertical location of each pixel.

*The author gratefully acknowledges the support of NIH grant ULI RR033183/KL2 RR0333182.

SUPPLEMENTARY MATERIAL

MultiwayRegression: R package MultiwayRegression, containing documented code for the methods described in this article. (GNU zipped tar le)

Code: A zipped folder with R scripts to reproduce the simulation and application results in this manuscript. (zipped folder)

This article concerns the prediction of an array of arbitrary dimension $Q_1 \times \cdots Q_M$ from another array of arbitrary dimension $P_1 \times \cdots P_L$. For N training observations, this involves an outcome array $\mathbb{Y}: N \times Q_1 \times \cdots \times Q_M$ and a predictor array $\mathbb{X}: N \times P_1 \times \cdots \times P_L$. For example, we consider the simultaneous prediction of several describable attributes for faces from their images (Kumar et al., 2009), which requires predicting the array $\mathbb{Y}: Faces \times Attributes$ from $\mathbb{X}: Faces \times X \times Y \times Colors$. Another potential application is the prediction of EEG from fMRI data (see De Martino et al. (2011)), or fMRI from EEG data (see Jansen et al. (2012)). The spatial resolution of fMRI is richer than that for EEG, and the temporal resolution of EEG is richer than that for fMRI. Thus, an understanding of the predictive relationship between the two datasets can be used to infer missing temporal information for fMRI (e.g., the order in which certain region activations occur) and missing spatial information for EEG (e.g., the exact spatial location of certain electrical signals) (Huster et al., 2012). Yet another potential application is the prediction of gene expression across multiple tissues from other genomic variables (see Ramasamy et al. (2014)). The association between genetic polymorphisms with gene expression is an important first step to understanding the genetic etiology of a disease, and such associations are known to differ across tissue types (GTEx Consortium, 2015).

The task of tensor-on-tensor regression extends a growing literature on the predictive modeling of tensors under different scenarios. Such methods commonly rely on tensor factorization techniques (Kolda and Bader, 2009), which reconstruct a tensor using a small number of underlying patterns in each dimension. Tensor factorizations extend well known techniques for a matrix, such as the singular value decomposition and principal component analysis, to higher-order arrays. A classical and straightforward technique is the PARAFAC/CANDECOMP (CP) (Harshman, 1970) decomposition, in which the data are approximated as a linear combination of rank-1 tensors. An alternative is the Tucker decomposition (Tucker, 1966), in which a tensor is factorized into basis vectors for each dimension that are combined using a smaller core tensor. The CP factorization is a special case of the Tucker factorization wherein the core tensor is diagonal. Such factorization techniques are useful to account for and exploit multi-way dependence and reduce dimensionality.

Several methods have been developed for the prediction of a scalar outcome from a tensor of arbitrary dimension: $\mathbb{Y}: N \times 1$ and $\mathbb{X}: N \times P_1 \times \cdots \times P_L$. Zhou et al. (2013) and Guo et al. (2012) propose tensor regression models for a single outcome in which the coefficient array is assumed to have a low-rank CP factorization. The proposed framework in Zhou et al. (2013) extends to generalized linear models and allows for the incorporation of sparsity-inducing regularization terms. An analogous approach in which the coefficients are assumed to have a Tucker structure is described by Li et al. (2013). Several methods have also been developed for the classification of multiway data (categorical $Y: N \times 1$) (Tao et al., 2007; Wimalawarne et al., 2016; Lyu et al.), extending well-known linear classification techniques under the assumption that model coefficients have a factorized structure.

There is also a wide literature on the prediction of a matrix from another matrix, $\mathbb{Y}: N \times Q$ and $\mathbb{X}: N \times P$. A classical approach is reduced rank regression, in which the $P \times Q$ coefficient matrix is restricted to have low rank (Izenman, 1975; Mukherjee and Zhu, 2011). Miranda et

al. (2015) describe a Bayesian formulation for regression models with multiple outcome variables and multiway predictors ($\mathbb{Y}: N \times Q$ and $\mathbb{X}: N \times P_1 \times \cdots \times P_L$), which is applied to a neuroimaging study. Conversely, tensor response regression models have been developed to predict a multiway outcome from vector predictors ($\mathbb{Y}: N \times Q_1 \times \cdots \times Q_M$, $\mathbb{X}: N \times P$). Sun and Li (2016) propose a tensor response regression wherein a multiway outcome is assumed to have a CP factorization, and Li and Zhang (2016) propose a tensor response regression wherein a multiway outcome is assumed to have a Tucker factorization with weights determined by vector-valued predictors. For a similar context Lock and Li (2016) describe a supervised CP factorization, wherein the components of a CP factorization are informed by vector-valued covariates. Hoff (2015) extend a bilinear regression model for matrices to the prediction of an outcome tensor from a predictor tensor with the same number of modes (e.g., $\mathbb{Y}: N \times Q_1 \times \cdots \times Q_K$ and $\mathbb{X}: N \times P_1 \times \cdots \times P_K$) via a Tucker product and describe a Gibbs sampling approach to inference.

The above methods address several important tasks, including scalar-on-tensor regression, vector-on-vector regression, vector-on-tensor regression and tensor-on-vector regression. However, there is a lack of methodology to addresses the important and increasingly relevant task of tensor-on-tensor regression, i.e., predicting an array of arbitrary dimension from another array of arbitrary dimension. This scenario is considered within a comprehensive theoretical study of convex tensor regularizers Raskutti and Yuan (2015), including the tensor nuclear norm. However, they do not discuss estimation algorithms for this context, and computing the tensor nuclear norm is NP-hard (Sun and Li, 2016; Friedland and Lim, 2014). In this article we propose a *contracted tensor product* for the linear prediction of a tensor $\mathbb{X}$ from a tensor $\mathbb{Y}$, where both $\mathbb{X}$ and $\mathbb{Y}$ have arbitrary dimension, through a coefficient array $\mathbb{B}$ of dimension $P_1 \times \cdots \times P_L \times Q_1 \times \cdots \times Q_M$. This framework is shown to accommodate all valid linear relations between the variates of $\mathbb{X}$ and the variates of $\mathbb{Y}$. In our implementation $\mathbb{B}$ is assumed to have reduced CP-rank, a simple restriction which simultaneously exploits the multi-way structure of both $\mathbb{X}$ and $\mathbb{Y}$ by borrowing information across the different modes and reducing dimensionality. We propose a general and efficient algorithm for penalized least-squares estimation, which allows for a ridge (L_2) penalty on the coefficients. The objective is shown to give the mode of a Bayesian posterior, which motivates a Gibbs sampling algorithm for inference.

The primary novel contribution of this article is a framework and methodology that allows for tensor-on-tensor regression with arbitrary dimensions. Other novel contributions include optimization under a ridge penalty on the coefficients and Gibbs sampling for inference, and these contributions are also relevant to the more familiar special cases of tensor regression (scalar-on-tensor), reduced rank regression (vector-on-vector), and tensor response regression (tensor-on-vector).

2 Notation and Preliminaries

Throughout this article bold lowercase characters (**a**) denote vectors, bold uppercase characters (**A**) denote matrices, and uppercase blackboard bold characters ($\mathbb{A}$) denote multi-way arrays of arbitrary dimension.

Define a K-way array (i.e., a Kth-order tensor) by $\mathbb{A}: I_1 \times \cdots \times I_K$, where I_k is the dimension of the kth *mode*. The entries of the array are defined by indices enclosed in square brackets, $\mathbb{A}[i_1, \ldots, i_K]$, where $i_k \in \{1, \ldots, I_k\}$ for $k \in 1, \ldots, K$.

For vectors $\mathbf{a}_1, \ldots, \mathbf{a}_K$ of length $I_1, \ldots, I_K$, respectively, define the *outer product*

$$\mathbb{A} = \mathbf{a}_1 \circ \mathbf{a}_2 \cdots \circ \mathbf{a}_K$$

as the K-way array of dimensions $I_1 \times \cdots \times I_K$, with entries

$$\mathbb{A}[i_1, \ldots, i_K] = \prod_{k=1}^{K} \mathbf{a}_k[i_k].$$

The outer product of vectors is defined to have *rank* 1. For matrices $\mathbf{A}_1, \ldots, \mathbf{A}_K$ of the same column dimension R, we introduce the notation

$$[\![\mathbf{A}_1, \ldots, \mathbf{A}_K]\!] = \sum_{r=1}^{R} \mathbf{a}_{1_r} \circ \cdots \circ \mathbf{a}_{K_r}, \quad (1)$$

where $\mathbf{a}_{kr}$ is the rth column of $\mathbf{A}_k$. This gives a CP factorization, and an array that can be expressed in the form (1) is defined to have rank R.

The vectorization operator vec(·) transforms a multiway array to a vector containing the array entries. Specifically, $\mathrm{vec}(\mathbb{A})$ is a vector of length $\prod_{k=1}^{K} I_K$ where

$$\mathrm{vec}(\mathbb{A})\left[i_1 + \sum_{k=2}^{K}\left(\prod_{l=1}^{k-1} I_l\right)\left(i_k - 1\right)\right] = \mathbb{A}[i_1, \ldots, i_K].$$

It is often useful to represent an array in matrix form via *unfolding* it along a given mode. For this purpose we let the rows of $\mathbf{A}^{(k)} : I_k \times (\prod_{j \neq k} I_j)$ give the vectorized versions of each *subarray* in the kth mode.

For two multiway arrays $\mathbb{A}: I_1 \times \cdots \times I_K \times P_1 \cdots P_L$ and $\mathbb{B}: P_1 \times \cdots \times P_L \times Q_1 \times \cdots \times Q_M$ we define the *contracted tensor product*

$$\langle \mathbb{A}, \mathbb{B} \rangle_L : I_1 \times \cdots \times I_K \times Q_1 \times \cdots \times Q_M$$

by

$$\langle \mathbb{A}, \mathbb{B} \rangle_L [i_1, \ldots i_K, q_1, \ldots, q_M] = \sum_{p_1=1}^{P_1} \cdots \sum_{p_L=1}^{P_L} \mathbb{A}[i_1, \ldots i_K, p_1, \ldots, p_L] \mathbb{B}[p_1, \ldots p_L, q_1, \ldots, q_M].$$

An analogous definition of the contracted tensor product, with slight differences in notation, is given in Bader and Kolda (2006). Note that for matrices $\mathbf{A} : I \times P$ and $\mathbf{B} : P \times Q$,

$$\langle \mathbf{A}, \mathbf{B} \rangle_1 = \mathbf{AB},$$

and thus the contracted tensor product extends the usual matrix product to higher-order operands.

3 General framework

Consider predicting a multiway array $\mathbb{Y}: N \times Q_1 \times \cdots \times Q_M$ from a multiway array $\mathbb{X}: N \times P_1 \times \cdots \times P_L$ with the model

$$\mathbb{Y} = \langle \mathbb{X}, \mathbb{B} \rangle_L + \mathbb{E} \quad (2)$$

where $\mathbb{B}: P_1 \times \cdots \times P_L \times Q_1 \times \cdots \times Q_M$ is a coefficient array and $\mathbb{E}: N \times Q_1 \times \cdots \times Q_M$ is an error array. The first L modes of $\mathbb{B}$ contract the dimensions of $\mathbb{X}$ that are not in $\mathbb{Y}$, and the last M modes of $\mathbb{B}$ expand along the modes in $\mathbb{Y}$ that are not in $\mathbb{X}$. The predicted outcome indexed by $(q_1, \ldots, q_M)$ is

$$\mathbb{Y}[n, q_1, \ldots, q_M] \approx \sum_{p_1}^{P_1} \cdots \sum_{p_L}^{P_L} \mathbb{X}[N, p_1, \ldots p_L] \mathbb{B}[p_1, \ldots, p_L, q_1, \ldots q_M] \quad (3)$$

for observations $n = 1, \ldots, N$. In (2) we forgo the use of an intercept term for simplicity, and assume that $\mathbb{X}$ and $\mathbb{Y}$ are each centered to have mean 0 over all their values.

Let P be the total number of predictors for each observation, $P = \prod_{l=1}^{L} P_L$, and Q be the total number of outcomes for each observation, $Q = \prod_{m=1}^{M} Q_M$. Equation (2) can be reformulated by rearranging the entries of $\mathbb{X}$, $\mathbb{Y}$, $\mathbb{B}$ and $\mathbb{E}$ into matrix form

$$\mathbf{Y}^{(1)} = \mathbf{X}^{(1)}\mathbf{B} + \mathbf{E}^{(1)} \quad (4)$$

where $\mathbf{Y}^{(1)} : N \times Q$, $\mathbf{X}^{(1)} : N \times P$, and $\mathbf{E}^{(1)} : N \times Q$ are the arrays $\mathbb{Y}$, $\mathbb{X}$ and $\mathbb{E}$ unfolded along the first mode. The columns of $\mathbf{B} : P \times Q$ vectorize the first L modes of $\mathbb{B}$ (collapsing $\mathbb{X}$), and the rows of $\mathbf{B}$ vectorize the last M modes of $\mathbb{B}$ (expanding to $\mathbb{Y}$):

$$\mathrm{B}\left[p_1 + \sum_{l=2}^{l-1}\left(\prod_{i=1}^{l-1} P_l\right)(p_l - 1), q_1 + \sum_{m=2}^{M}\left(\prod_{i=1}^{m-1} Q_m\right)(q_m - 1)\right] = \mathbb{B}[p_1, \ldots, p_L, q_1, \ldots, q_M].$$

From its matrix form (4) it is clear that the general framework (2) supports all valid linear relations between the P variates of $\mathbb{X}$ and the Q variates of $\mathbb{Y}$.

4 Estimation criteria

Consider choosing $\mathbb{B}$ to minimize the sum of squared residuals

$$\left\| \mathbb{Y} - \langle \mathbb{X}, \mathbb{B} \rangle_L \right\|_F^2.$$

The unrestricted solution for $\mathbb{B}$ is given by separate OLS regressions for each of the Q outcomes in $\mathbb{Y}$, each with design matrix $\mathbf{X}^{(1)}$; this is clear from (4), where the columns of $\mathbf{B}$ are given by separate OLS regressions of $\mathbf{X}^{(1)}$ on each column of $\mathbf{Y}^{(1)}$. Therefore, the unrestricted solution is not well-defined if $Q > N$ or more generally if $\mathbf{X}^{(1)}$ is not of full column rank. The unrestricted least squares solution may be undesirable even if it is well-defined, as it does not exploit the multi-way structure of $\mathbb{X}$ or $\mathbb{Y}$, and requires fitting

$$\prod_{l=1}^{L} P_l \prod_{m=1}^{M} Q_m \quad (5)$$

unknown parameters. Alternatively, the multi-way nature of $\mathbb{X}$ and $\mathbb{Y}$ suggests a low-rank solution for $\mathbb{B}$. The rank R solution can be represented as

$$\mathbb{B} = \quad [\![\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M]\!] \ , \quad (6)$$

where $\mathbf{U}_l : P_l \times R$ for $l = 1, \ldots, L$ and $\mathbf{V}_m : Q_m \times R$ for $m = 1, \ldots, M$. The dimension of this model is

$$R(P_1 + \cdots + P_L + Q_1 + \cdots + Q_M), \quad (7)$$

which can be a several order reduction from the unconstrained dimensionality (5). Moreover, the reduced rank solution allows for borrowing of information across the different dimensions of both $\mathbb{X}$ and $\mathbb{Y}$. However, the resulting least-squares solution

$$\underset{\text{rank}(\mathbb{B}) \le R}{\arg\min} \left\| \mathbb{Y} - \langle \mathbb{X}, \mathbb{B} \rangle_L \right\|_F^2. \quad (8)$$

is still prone to over-fitting and instability if the model dimension (7) is high relative to the number of observed outcomes, or if the predictors $\mathbb{X}$ have multicollinearity that is not addressed by the reduced rank assumption (e.g., multicollinearity within a mode). High-dimensionality and multicollinearity are both commonly encountered in application areas

that involve multi-way data, such as imaging and genomics. To address these issues we incorporate an L_2 penalty on the coefficient array $\mathbb{B}$,

$$\underset{\text{rank}(\mathbb{B}) \leq R}{\arg\min} \left\| \mathbb{Y} - \langle \mathbb{X}, \mathbb{B} \rangle_L \right\|_F^2 . + \lambda \left\| \mathbb{B} \right\|_F^2, \quad (9)$$

where λ controls the degree of penalization. This objective is equivalent to that of ridge regression when predicting a vector outcome $\mathbb{Y}: N \times 1$ from a matrix $\mathbb{X}: N \times P$, where necessarily $R = 1$.

5 Identifiability

The general predictive model (2) is identifiable for $\mathbb{B}$, in that $\mathbb{B} \neq \mathbb{B}^*$ implies

$$\left\langle \widetilde{\mathbb{X}}, \mathbb{B} \right\rangle_L \neq \left\langle \widetilde{\mathbb{X}}, \mathbb{B}^* \right\rangle_L$$

for some $\widetilde{\mathbb{X}} \in \mathbb{R}^{P_1 \times \cdots \times P_L}$. To show this, note that if $\widetilde{\mathbb{X}}$ is an array with 1 in position [p_1, …, p_L] and zeros elsewhere, then

$$\left\langle \widetilde{\mathbb{X}}, \mathbb{B} \right\rangle_L [q_1, \ldots q_M] = \mathbb{B}[p_1, \ldots, p_M, q_1, \ldots, q_M].$$

However, the resulting components $\mathbf{U}_1$, …, $\mathbf{U}_L$, $\mathbf{V}_1$, …, $\mathbf{V}_M$ in the factorized representation of $\mathbb{B}$ (6) are not readily identified. Conditions for their identifiability are equivalent to conditions for the identifiability of the CP factorization, for which there is an extensive literature. To account for arbitrary scaling and ordering of the components, we impose the restrictions

1. $\|\mathbf{u}_{1r}\| = \cdots = \|\mathbf{u}_{Lr}\| = \|\mathbf{v}_{1r}\| = \cdots = \|\mathbf{v}_{Mr}\|$ for $r = 1 \ldots, R$, and
2. $\|\mathbf{u}_{11}\| \geq \|\mathbf{u}_{12}\| \geq \cdots \geq \|\mathbf{u}_{1R}\|$.

The above restrictions are generally enough to ensure identifiability when $L+M \geq 3$ under verifiable conditions (Sidiropoulos and Bro, 2000). If $L + M = 2$ (i.e., when predicting a matrix from a matrix, a 3-way array from a vector, or a vector from a 3-way array), then $\mathbb{B}$ is a matrix and we require additional orthogonality restrictions:

3. $\mathbf{u}_{1r}^T \mathbf{u}_{1r^*} = 0$ for all $r \neq r^*$, or $\mathbf{v}_{1r}^T \mathbf{v}_{1r^*} = 0$ for all $r \neq r^*$.

In practice these restrictions can be imposed post-hoc, after the estimation procedure detailed in Section 7. For $L + M \geq 3$, restrictions (a) and (b) can be imposed via a re-ordering and re-scaling of the components. For $L + M = 2$, components that satisfy restrictions (a), (b) and (c) can be identified via a singular value decomposition of $\mathbb{B}$.

6 Special cases

Here we describe other methods that fall within the family given by the reduced rank ridge objective (9). When predicting a vector from a matrix ($Q = 0$, $P = 1$), this framework is equivalent to standard ridge regression (Hoerl and Kennard, 1970), which is equivalent to OLS when $\lambda = 0$. Moreover, a connection between standard ridge regression and continuum regression (Sundberg, 1993) implies that the coefficients obtained through ridge regression are proportional to partial least squares regression for some $\lambda = \lambda^*$, and the coefficients are proportional to the principal components of $\mathbb{X}$ when $\lambda \to \infty$.

When predicting a matrix from another matrix ($Q = 1$, $P = 1$), the objective given by (9) is equivalent to reduced rank regression (Izenman, 1975) when $\lambda = 0$. For arbitrary λ the objective is equivalent to a recently proposed reduced rank ridge regression (Mukherjee and Zhu, 2011).

When predicting a scalar from a tensor of arbitrary dimension ($Q = 0$, arbitrary P), (9) is equivalent to tensor ridge regression (Guo et al., 2012). Guo et al. (2012) use an alternating approach to estimation but claim that the subproblem for estimation of each $\mathbf{U}_l$ cannot be computed in closed form and resort to gradient style methods instead. On the contrary, our optimization approach detailed in Section 7 does give a closed form solution to this subproblem (11). Alternatively, Guo et al. (2012) suggest the separable form of the objective,

$$\underset{\text{rank}(\mathbb{B}) \leq R}{\arg\min} \left\| \mathbf{y} - \langle \mathbb{X}, \mathbb{B} \rangle_L \right\|_F^2 . + \lambda \sum_{l=1}^{L} \left\| \mathbf{U}_l \right\|_F^2 . \quad (10)$$

This separable objective is also used by Zhou et al. (2013), who consider a power family of penalty functions for predicting a vector from a tensor using a generalized linear model; their objective for a Gaussian response under L_2 penalization is equivalent to (10). The solution of the separable L_2 penalty depends on arbitrary scaling and orthogonality restrictions for identifiability of the $\mathbf{U}_l$'s. For example, the separable penalty (10) is equivalent to the non-separable L_2 penalty (9) if the columns of $\mathbf{U}_2, \ldots, \mathbf{U}_L$ are restricted to be orthonormal.

Without scale restrictions on the columns of $\mathbf{U}_l$, the solution to the separable L_2 penalty is equal to the solution for the non-separable penalty $\left\| \mathbb{B} \right\|_*$ for $L = 2$, where $\left\| \mathbb{B} \right\|_*$ defines the nuclear norm (i.e., the sum of the singular values of $\mathbb{B}$). This interesting result is given explicitly in Proposition 1, and its proof is given in Appendix C.

Proposition 1

For $\mathbb{B} = [\![\mathbf{U}_1, \mathbf{U}_2]\!] = \mathbf{U}_1 \mathbf{U}_2^T$, where the columns of $\mathbf{U}_1$ and $\mathbf{U}_2$ are orthogonal,

$$\underset{\text{rank}(\mathbb{B}) \le R}{\arg\min} \left\| \mathbf{y} - \langle \mathbb{X}, \mathbb{B} \rangle_2 \right\|_F^2 + \lambda \sum_{l=1}^{2} \left\| \mathbf{U}_l \right\|_F^2 = \underset{\text{rank}(\mathbb{B}) \le R}{\arg\min} \left\| \mathbf{y} - \langle \mathbb{X}, \mathbb{B} \rangle_2 \right\|_F^2 + 2\lambda \left\| \mathbb{B} \right\|_*.$$

7 Optimization

We describe an iterative procedure to estimate $\mathbb{B}$ that alternatingly solves the objective (9) for the component vectors in each mode, $\{\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M\}$, with the others fixed.

7.1 Least-squares

Here we consider the case without ridge regularization, $\lambda = 0$, wherein the component vectors in each mode are updated via separate OLS regressions.

To simplify notation we describe the procedure to update $\mathbf{U}_1$ with $\{\mathbf{U}_2, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M\}$ fixed. The procedure to update each of $\mathbf{U}_2, \ldots, \mathbf{U}_L$ is analogous, because the loss function is invariant under permutation of the L modes of $\mathbb{X}$.

Define $\mathbb{C}_r : N \times P_1 \times Q_1 \times \cdots \times Q_M$ to be the contracted tensor product of $\mathbb{X}$ and the r'th component of the CP factorization without $\mathbf{U}_1$:

$$\mathbb{C}_r = \left\langle \mathbb{X}, \mathbf{u}_{2r} \circ \cdots \circ \mathbf{u}_{Lr} \circ \mathbf{v}_{1r} \circ \cdots \circ \mathbf{v}_{Mr} \right\rangle_{L-1}.$$

Unfolding $\mathbb{C}_r$ along the dimension corresponding to P_1 gives the design matrix to predict $\text{vec}(Y)$ for the r'th column of $\mathbf{U}_1$, $\mathbf{C}_r : NQ \times P_1$. Thus, concatenating these matrices to define $\mathbf{C} : NQ \times RP_1$ by $\mathbf{C} = [\mathbf{C}_1 \ldots \mathbf{C}_R]$ gives the design matrix for all of the entries of $\mathbf{U}_1$, which are updated via OLS:

$$\text{vec}(\mathbf{U}_1) = \left(\mathbf{C}^T\mathbf{C}\right)^{-1}\mathbf{C}^T\text{vec}(\mathbb{Y}). \quad (11)$$

For the outcome modes we describe the procedure to update $\mathbf{V}_M$ with $\{\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_{M-1}\}$ fixed. The procedure to update each of $\mathbf{V}_1, \ldots, \mathbf{V}_{L-1}$ is analogous, because the loss function is invariant under permutation of the M modes of $\mathbb{Y}$.

Let $\mathbf{Y}_M : Q_M \times N \prod_{m=1}^{M-1} Q_m$ be $\mathbb{Y}$ unfolded along the mode corresponding to Q_M. Define $\mathbf{D} : N \prod_{m=1}^{M-1} Q_m \times R$ so that the r'th column of D, $\mathbf{d}_r$, gives the entries of the contracted tensor product of $\mathbb{X}$ and the r'th component of the CP factorization without $\mathbf{V}_M$:

$$\mathbf{d}_r = \text{vec}\left(\left\langle \mathbb{X}, \mathbf{u}_{1r} \circ \cdots \circ \mathbf{u}_{Lr} \circ \mathbf{v}_{1r} \circ \cdots \circ \mathbf{v}_{(M-1)r} \right\rangle_L\right).$$

The entries of $\mathbf{V}_M$ are then updated via Q_M separate OLS regressions:

$$\mathbf{V}_M = \left(\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{Y}_M^T. \quad (12)$$

7.2 Ridge-regularization

For $\lambda > 0$, note that the objective (9) can be equivalently represented as an unregularized least squares problem with modified predictor and outcome arrays $\widetilde{\mathbb{X}}$ and $\widetilde{\mathbb{Y}}$:

$$\underset{\text{rank}(\mathbb{B}) \le R}{\arg\min} \left\| \widetilde{\mathbb{Y}} - \left\langle \widetilde{\mathbb{X}}, \mathbb{B} \right\rangle_L \right\|_F^2.$$

Here $\widetilde{\mathbb{X}}: (N+P) \times P_1 \times \cdots \times P_L$ is the concatenation of $\mathbb{X}$ and a tensor wherein each $P_1 \times \cdots P_L$ dimensional slice has $\sqrt{\lambda}$ for a single entry and zeros elsewhere; $\widetilde{\mathbb{Y}}: (N+P) \times Q_1 \times \cdots \times Q_M$ is the concatenation of $\mathbb{Y}$ and a $P \times Q_1 \times \cdots Q_M$ tensor of zeros. Unfolding $\widetilde{\mathbb{X}}$ and $\widetilde{\mathbb{Y}}$ along the first dimension yields the matrices

$$\widetilde{\mathbf{X}}^{(1)} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \sqrt{\lambda}\mathbf{I}_{P \times P} \end{bmatrix} \text{ and } \widetilde{\mathbf{Y}}^{(1)} = \begin{bmatrix} \mathbf{Y}^{(1)} \\ 0_{P \times \prod_{m=1}^{L} Q_m} \end{bmatrix},$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{0}$ is a matrix of zeros.

Thus, one can optimize the objective (9) via alternating least squares by replacing $\widetilde{\mathbb{X}}$ for $\mathbb{X}$ and $\widetilde{\mathbb{Y}}$ for $\mathbb{Y}$ in the least-squares algorithm of Section 7.1. However, $\widetilde{\mathbb{X}}$ and $\widetilde{\mathbb{Y}}$ can be very large for high-dimensional $\mathbb{X}$. Thankfully, straightforward tensor algebra shows that this is equivalent to a direct application of algorithm in Section 7.1 to the original data $\mathbb{X}$ and $\mathbb{Y}$, with computationally efficient modifications to the OLS updating steps (11) and (12). The updating step for $\mathbf{U}_1$ (11) is

$$\text{vec}(\mathbf{U}_1) = \left(\mathbf{C}^T\mathbf{C} + \lambda\left(\mathbf{U}_2^T\mathbf{U}_2 \cdots \cdot \mathbf{U}_L^T\mathbf{U}_L \cdot \mathbf{V}_1^T\mathbf{V}_1 \cdots \cdot \mathbf{V}_M^T\mathbf{V}_M\right) \otimes \mathbf{I}_{P_1 \times P_1}\right)^{-1}\mathbf{C}^T\text{vec}(\mathbb{Y}) \quad (13)$$

where $\cdot$ defines the dot product and $\otimes$ defines the Kronecker product. The updating step for $\mathbf{V}_M$ (12) is

$$\mathbf{V}_M = \left(\mathbf{D}^T\mathbf{D} + \lambda\left(\mathbf{U}_1^T\mathbf{U}_1 \cdots \cdot \mathbf{U}_L^T\mathbf{U}_L \cdot \mathbf{V}_1^T\mathbf{V}_1 \cdots \cdot \mathbf{V}_{M-1}^T\mathbf{V}_{M-1}\right)\right)^{-1}\mathbf{D}^T\mathbf{Y}_M^T. \quad (14)$$

This iterative procedure is guaranteed to improve the regularized least squares objective (9) at each sub-step. The algorithm is “multi-convex”, i.e., the objective is convex and the parameter domains that are iteratively optimized are each convex. However, this is not enough to guarantee convergence to a global optimum, and the full space of low-rank tensors is not a convex space. The algorithm also may not converge to a local optimum, in that for a

given distance metric it is possible that there exists alternative solutions within an infinitesimally small ϵ-ball of the converged solution with a better objective (see (Chen et al., 2012) for a discussion on iterative tensor optimization and stationary points). However, it is straightforward that the algorithm is guaranteed to converge to a *coordinate-wise* optimum, wherein the solution cannot be improved by changing the parameters in any single dimension. Higher levels of regularization ($\lambda \to \infty$) tend to convexify the objective and facilitate convergence to a global optimum, a similar phenomenon is observed in Zhou et al. (2013). In practice we find that robustness to initial values and local optima is improved by a tempered regularization, starting with larger values of λ that gradually decrease to the desired level of regularization.

7.3 Tuning parameter selection

Selection of λ and R (if unknown) can be accomplished by assessing predictive accuracy with a training and test set, as illustrated in Section 10. More generally, these parameters can be selected via K-fold cross-validation. This approach has the advantage of being free of model assumptions, and is assessed via a simulation study in Appendix F. Alternatively, it is straightforward to compute the deviance information criterion (DIC) (Spiegelhalter et al., 2014) for posterior draws under the Bayesian inference framework of Section 8 and use this as a model-based heuristic to select both λ and R.

8 Inference

In the previous sections we have considered optimizing a given criteria for point estimation, without specifying a distributional form for the data or even a philosophical framework for inference. Indeed, the estimator given by the objective (9) is consistent under a wide variety of distributional assumptions, including those that allow for correlated responses or predictors. See Appendix A for more details on its consistency.

For inference and uncertainty quantification for this point estimate, we propose a Markov chain Monte Carlo (MCMC) simulation approach. This approach is theoretically motivated by the observation that (9) gives the mode of a Bayesian probability distribution. There are several other reasons to use MCMC simulation for inference in this context, rather than (e.g.,) asymptotic normality of the global optimizer under an assumed likelihood model (see Zhou et al. (2013) and Zhang et al. (2014) for related results). The algorithm in Section 7 may converge to a local minimum, which can still be used as a starting value for MCMC. Moreover, our approach gives a framework for full posterior inference on $\hat{\beta}$ over its rank R support, the conditional mean for observed responses, and the predictive distribution for the response array given new realizations of the predictor array without requiring the identifiability of $\theta = \{\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M\}$. Inference for θ is also possible under the conditions of Section 5.

If the errors $\mathbb{E}$ have independent N(0, σ^2) entries, the log-likelihood of $\mathbb{Y}$ implied by the general model (2) is

$$\log\text{pr}\left(\mathbb{Y}\middle|\sigma^2,\mathbb{B},\mathbb{X}\right)=\text{constant}-\frac{1}{2\sigma^2}\left\|\mathbb{Y}-\langle\mathbb{X},\mathbb{B}\rangle_L\right\|_F^2,$$

and thus the unregularized objective (8) gives the maximum likelihood estimate under the restriction $\text{rank}(\mathbb{B})=R$. For $\lambda>0$, consider a prior distribution for $\mathbb{B}$ that is proportional to the spherical Gaussian distribution with variance σ^2/λ over the support of rank R tensors:

$$\text{pr}(\mathbb{B})\propto\begin{cases}\exp\left(-\frac{\lambda}{2\sigma^2}\left\|\mathbb{B}\right\|_F^2\right) & \text{if rank}(\mathbb{B})\le R\,.\\ 0 & \text{otherwise,}\end{cases}\qquad(15)$$

The log posterior distribution for $\mathbb{B}$ is

$$\log\text{pr}\left(\mathbb{B}\middle|\mathbb{Y},\mathbb{X},\sigma^2\right)=\text{constant}-\frac{1}{2\sigma^2}\left(\left\|\mathbb{Y}-\langle\mathbb{X},\mathbb{B}\rangle_L\right\|_F^2+\lambda\left\|\mathbb{B}\right\|_F^2\right)\qquad(16)$$

where $\text{rank}(\mathbb{B})=R$, which is maximized by (9).

Under the factorized form (6) the full conditional distributions implied by (16) for each of $\mathbf{U}_1,\ldots,\mathbf{U}_L,\mathbf{V}_1,\ldots,\mathbf{V}_M$ are multivariate normal. For example, the full conditional for $\mathbf{U}_1$ is

$$\text{pr}\left(\text{vec}\left(\mathbf{U}_1\right)\middle|\mathbf{U}_2,\ldots,\mathbf{U}_L,\mathbf{V}_1,\ldots,\mathbf{V}_M,\mathbb{Y},\mathbb{X},\sigma^2\right)=N\left(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1\right),$$

where $\boldsymbol{\mu}_1$ is the right hand side of (13) and

$$\boldsymbol{\Sigma}_1=\sigma^2\left(\mathbf{C}^T\mathbf{C}+\lambda\left(\mathbf{U}_2^T\mathbf{U}_2\cdots\cdot\mathbf{U}_L^T\mathbf{U}_L\cdot\mathbf{V}_1^T\mathbf{V}_1\cdots\cdot\mathbf{V}_M^T\mathbf{V}_M\right)\otimes\mathbf{I}_{P_1\times P_1}\right)^{-1}$$

where $\mathbf{C}$ is defined as in Section 7.1. The full conditional for $\mathbf{V}_M$ is

$$\text{pr}\left(\text{vec}\left(\mathbf{V}_M\right)\middle|\mathbf{U}_2,\ldots,\mathbf{U}_L,\mathbf{V}_1,\ldots,\mathbf{V}_M,\mathbb{Y},\mathbb{X},\sigma^2\right)=N\left(\boldsymbol{\mu}_{L+M},\boldsymbol{\Sigma}_{L+M}\right),$$

where $\boldsymbol{\mu}_{L+M}$ is given by the right hand side of (14) and

$$\boldsymbol{\Sigma}_{L+M}=\sigma^2\left(\mathbf{D}^T\mathbf{D}+\lambda\left(\mathbf{U}_1^T\mathbf{U}_1\cdots\cdot\mathbf{U}_L^T\mathbf{U}_L\cdot\mathbf{V}_1^T\mathbf{V}_1\cdots\cdot\mathbf{V}_{M-1}^T\mathbf{V}_{M-1}\right)\right)^{-1}\otimes\mathbf{I}_{Q_M\times Q_M}$$

The derivations of the conditional means and variances $\left\{\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right\}_{i=1}^{L+M}$ are given in Appendix B. When $\lambda=0$ the full conditionals correspond to a at prior on $\mathbb{B}$, $\text{pr}(\mathbb{B})\propto 1$ for $\text{rank}(\mathbb{B})=R$, and the posterior mode is given by the unregularized objective (8).

In practice we use a flexible Jeffrey's prior for σ^2, $\mathrm{pr}(\sigma^2) \propto 1/\sigma^2$, which leads to an inverse-gamma (IG) full conditional distribution,

$$\mathrm{pr}\left(\sigma^2 \middle| \mathbb{B}, \mathbb{Y}, \mathbb{X}\right) = \mathrm{IG}\left(\frac{NQ}{2}, \frac{1}{2}\left\|\mathbb{Y} - \langle \mathbb{X}, \mathbb{B} \rangle_L\right\|_F^2\right). \quad (17)$$

We simulate dependent samples from the marginal posterior distribution of $\mathbb{B}$ by Gibbs sampling from the full conditionals of $\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M$, and σ^2:

1. Initialize $\mathbb{B}^{(0)}$ by the posterior mode (9) using the procedure in Section 7.

   For samples $t = 1, \ldots, T$, repeat (b) and (c):

2. Draw $\sigma^{2(t)}$ from $P\left(\sigma^2 \middle| \mathbb{B}^{(t-1)}, \mathbb{Y}, \mathbb{X}\right)$ as in (17).

3. Draw $\mathbb{B}^{(t)} = [\![\mathbf{U}_1^{(t)}, \ldots, \mathbf{U}_L^{(t)}, \mathbf{V}_1^{(t)}, \ldots, \mathbf{V}_M^{(t)}]\!]$, as follows:

$$\begin{aligned}
&\mathbf{U}_1^{(t)} \sim P\left(\mathbf{U}_1 \middle| \mathbf{U}_2^{(t-1)}, \ldots, \mathbf{U}_L^{(t-1)}, \mathbf{V}_1^{(t-1)}, \ldots, \mathbf{V}_M^{(t-1)}, \mathbb{Y}, \mathbb{X}, \sigma^{2(t)}\right) \\
&\vdots \\
&\mathbf{U}_L^{(t)} \sim P\left(\mathbf{U}_L \middle| \mathbf{U}_1^{(t)}, \ldots, \mathbf{U}_{L-1}^{(t-1)}, \mathbf{V}_1^{(t-1)}, \ldots, \mathbf{V}_M^{(t-1)}, \mathbb{Y}, \mathbb{X}, \sigma^{2(t)}\right) \\
&\mathbf{V}_1^{(t)} \sim P\left(\mathbf{V}_1 \middle| \mathbf{U}_2^{(t)}, \ldots, \mathbf{U}_L^{(t)}, \mathbf{V}_2^{(t-1)}, \ldots, \mathbf{V}_M^{(t-1)}, \mathbb{Y}, \mathbb{X}, \sigma^{2(t)}\right) \\
&\vdots \\
&\mathbf{V}_L^{(t)} \sim P\left(\mathbf{V}_M \middle| \mathbf{U}_1^{(t)}, \ldots, \mathbf{U}_L^{(t)}, \mathbf{V}_1^{(t-1)}, \ldots, \mathbf{V}_M^{(t-1)}, \mathbb{Y}, \mathbb{X}, \sigma^{2(t)}\right).
\end{aligned}$$

For the above algorithm $\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M$ serve as a parameter augmentation to facilitate sampling for $\mathbb{B}$. Interpreting the marginal distribution of each of the $\mathbf{U}'_l s$ or $\mathbf{V}'_m s$ separately requires careful consideration of their identifiability (see Section 5). One approach is to perform a post-hoc transformation of the components at each sampling iteration

$$\mathbb{B}^{(t)} = [\![\mathbf{U}_1^{*(t)}, \ldots, \mathbf{U}_L^{*(t)}, \mathbf{V}_1^{*(t)}, \ldots, \mathbf{V}_M^{*(t)}]\!],$$

where $\left\{\mathbf{U}_1^{*(t)}, \ldots, \mathbf{U}_L^{*(t)}, \mathbf{V}_1^{*(t)}, \ldots, \mathbf{V}_M^{*(t)}\right\}$ satisfy given restrictions for identifiability.

For $\tilde{N}$ out-of-sample observations with predictor array $\mathbb{X}_{\text{new}}: \tilde{N} \times P_1 \times \cdots \times P_L$, the point prediction for the responses is

$$\widehat{\mathbb{Y}}_{\text{new}} = \left\langle \mathbb{X}_{\text{new}}, \widehat{\mathbb{B}} \right\rangle_L \quad (18)$$

where $\mathbb{B}$ is given by (9). Uncertainty in this prediction can be assessed using samples from the posterior predictive distribution of $\mathbb{Y}_{\text{new}}$:

$$\mathbb{Y}_{\text{new}}^{(t)} = \left\langle \mathbb{X}_{\text{new}}, \mathbb{B}^{(t)} \right\rangle_L + \mathbb{E}_{\text{new}}^{(t)}, \quad (19)$$

where $\mathbb{E}_{\text{new}}^{(t)}$ is generated with independent N(0, $\sigma^{2(t)}$) entries.

9 Simulation study

9.1 Approach

We conduct a simulation study to predict a three-way array $\mathbb{Y}$ from another three-way array $\mathbb{X}$ under various conditions. We implement a fully crossed factorial simulation design with the following manipulated conditions:

- Rank $R = 0, 1, 2, 3, 4$ or 5 (6 levels)
- Sample size $N = 30$ or 120 (2 levels)
- Signal-to-noise ratio SNR = 1 or 5 (2 levels).

For each of the 24 scenarios, we simulate data as follows:

1. Generate $\mathbb{X} : N \times P_1 \times P_2$ with independent $\mathcal{N}(0, 1)$ entries.
2. Generate $\mathbf{U}_l : P_l \times R$ for $l = 1, \ldots, L$ and $\mathbf{V}_m : Q_m \times R$ for $m = 1, \ldots, M$, each with independent $\mathcal{N}(0, 1)$ entries.
3. Generate error $\mathbb{E} : N \times Q_1 \times Q_2$ with independent $\mathcal{N}(0, 1)$ entries.
4. Set $\mathbb{Y} = \langle \mathbb{X}, \mathbb{B} \rangle_L + \mathbb{E}$, where

$$\mathbb{B} = c \left[\!\left[\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M \right]\!\right]$$

and c is the scalar giving

$$\frac{\left\| \langle \mathbb{X}, \mathbb{B} \rangle_L \right\|_F^2}{\left\| \mathbb{E} \right\|_F^2} = \text{SNR}.$$

We x the dimensions $p_1 = 15$, $p_2 = 20$, $q_1 = 5$, $q_2 = 10$, and generate 10 replicated datasets as above for each of the 24 scenarios, yielding 240 simulated datasets. For each simulated dataset, we estimate $\mathbb{B}$ as in Section 7 under each combination of the following parameters:

- Assumed rank $\hat{R} = 1, 2, 3, 4$ or 5 (5 levels)
- Regularization term $\lambda = 0, 0{:}5, 1, 5$ or 50 (5 levels).

For each of the 240 simulated datasets and 5 5 = 25 estimation procedures, we compute the relative out-of-sample prediction error of the resulting coefficient estimate $\widehat{\mathbb{B}}$. This is done empirically by generating a new dataset with $\tilde{N} = 500$ observations:

$$\mathbb{Y}_{\text{new}} = \left\langle \mathbb{X}_{\text{new}}, \mathbb{B} \right\rangle_L + \mathbb{E}_{\text{new}}$$

where $\mathbb{X}_{\text{new}}$ and $\mathbb{E}_{\text{new}}$ have independent $\mathcal{N}(0, 1)$ entries. The relative prediction error (RPE) for these test observations is

$$\text{RPE} = \frac{\left\| \mathbb{Y}_{\text{new}} - \left\langle \mathbb{X}_{\text{new}}, \widehat{\mathbb{B}} \right\rangle_L \right\|_F^2}{\left\| \mathbb{Y}_{\text{new}} \right\|_F^2}. \quad (20)$$

Symmetric 95% credible intervals are created for each value of $\mathbb{Y}_{\text{new}}$ using $T = 1000$ outcome arrays simulated from the posterior (19).

9.2 Results

First we consider the results for those cases with no signal, $R = 0$, where the oracle RPE is 1. The marginal mean RPE across the levels of N, λ, and $\hat{R}$ are shown in Table 1. Overall, simulations with a higher training sample size N resulted in lower RPE, estimation with higher regularization parameter λ resulted in lower RPE, and estimation with higher assumed rank resulted in higher RPE. These results are not surprising, as a lower sample size, higher assumed rank and less regularization all encourage over-fitting.

Table 2 shows the effect of the regularization parameter λ on the accuracy of the estimated model, in terms of RPE and coverage rates, for different scenarios. As expected, prediction error is generally improved in scenarios with a higher training sample size and higher signal-to-noise ratio. Higher values of λ generally improve predictive performance when the sample size and signal-to-noise ratio are small, as these scenarios are prone to over-fitting without regularization. However, large values of λ can lead to over-shrinkage of the estimated coefficients and introduce unnecessary bias, especially in scenarios that are less prone to over-fitting. Coverage rates of the 95% credible intervals are generally appropriate, especially with a higher training sample size. However, for the scenario with low sample size and high signal ($N = 30$, $\lambda = 120$) coverage rates for moderate values of λ are poor, as inference is biased toward smaller values of $\mathbb{B}$.

Table 3 illustrates the effects of rank misspecification on performance, under the scenario with $N = 120$, SNR= 1 and no regularization ($\lambda = 0$). For each possible value of the true rank $R = 1, \ldots, 5$, the RPE is minimized when the assumed rank is equal to the true rank. Predictive performance is generally more robust to assuming a rank higher than the true rank than it is to assuming a rank lower than the true rank.

See Appendix D for additional simulation results when the predictors $\mathbb{X}$ or response $\mathbb{Y}$ are correlated, and see Appendix E for a comparison with ad-hoc approaches that do not account for low rank dependence in $\mathbb{X}$ or $\mathbb{Y}$.

10 Application

We use the tensor-on-tensor regression model to predict attributes from facial images, using the Labeled Faces in the Wild database (Learned-Miller et al., 2016). The database includes over 13000 publicly available images taken from the internet, where each image includes the face of an individual. Each image is labeled only with the name of the individual depicted, often a celebrity, and there are multiple images for each individual. The images are unposed and exhibit wide variation in lighting, image quality, angle, etc. (hence “in the wild”).

Low-rank matrix factorization approaches are commonly used to analyze facial image data, particularly in the context of facial recognition (Sirovich and Kirby, 1987; Turk and Pentland, 1991; Vasilescu and Terzopoulos, 2002; Kim and Choi, 2007). Although facial images are not obviously multi-linear, the use of multi-way factorization techniques has been shown to convey advantages over simply vectorizing images (e.g., from a P_1 P_2 array of pixels to a vector of length $P_1 \times P_2$) (Vasilescu and Terzopoulos, 2002). Kim and Choi (2007) show that treating color as another mode within a tensor factorization framework can improve facial recognition tasks with different lighting. Moreover, the CP factorization has been shown to be much more efficient as a dimension reduction tool for facial images than PCA, and marginally more efficient than the Tucker and other multiway factorization techniques (Lock et al., 2011).

Kumar et al. (2009) developed an attribute classifier, which gives describable attributes for a given facial image. These attributes include characteristics that describe the individual (e.g., gender, race, age), that describe their expression (e.g., smiling, frowning, eyes open), and that describe their accessories (e.g., glasses, make-up, jewelry). These attribute were determined on the Faces in the Wild dataset, as well as other facial image databases. In total 72 attributes are measured for each image. The attributes are measured on a continuous scale; for example, for the smiling attribute, higher values correspond to a more obvious smile and lower values correspond to no smile.

Our goal is to create an algorithm to predict the 72 describable and correlated attributes from a given image that contains a face. First, the images are *frontalized* as described in Hassner et al. (2015). In this process the unconstrained images are rotated, scaled, and cropped so that all faces appear forward-facing and the image shows only the face. After this step images are aligned over the coordinates, in that we expect the nose, mouth and other facial features to be in approximately the same location. Each frontalized image is 90 × 90 pixels, and each pixel gives the intensity for colors red, green and blue, resulting in a multiway array of dimensions 90 × 90 × 3. We center the array by subtracting the “mean face” from each image, i.e., we center each pixel triplet ($x \times y \times$ color) to have mean 0 over the collection of frontalized images. We standardize the facial attribute data by converting the measurements to z-scores, wherein each attribute has mean zero and standard deviation 1 over the collection of faces.

To train the predictive model we use a use a random sample of 1000 images from unique individuals. Thus the predictor array of images $\mathbb{X}$ is of dimension 1000 × 90 × 90 × 3, and the outcome array of attributes $\mathbb{Y}$ is of dimension 1000 × 72. Another set of 1000 images

from unique individuals are used as a validation set, $\mathbb{X}_{new}: 1000 \times 90 \times 90 \times 3$ and $\mathbb{Y}_{new}: 1000 \times 72$.

We run the optimization algorithm in Section 7 to estimate the coefficient array $\mathbb{B}: 90 \times 90 \times 3 \times 72$ under various values for the rank R and regularization parameter λ. We consider all combinations of the values $\lambda = \{0, 0{:}1, 1, 10, 100, 1000, 10^4, 10^5\}$ and $R = \{1, 2, \ldots, 16\}$. We also consider the full rank model that ignores multi-way structure, where the coefficients are given by separate ridge regressions for each of the 72 outcomes on the $90 \cdot 90 \cdot 3 = 24300$ predictors. For each estimate we compute the relative prediction error (RPE) for the test set (see (20)). The resulting RPE values over the different estimation schemes are shown in Figure 1. The minimum RPE achieved was 0:568, for $R = 15$ and $\lambda = 10^5$. The performance of models with no regularization ($\lambda = 0$), or without rank restriction (rank=FULL), were much worse in comparison. This illustrates the benefits of simultaneous rank restriction and ridge regularization for high-dimensional multi-way prediction problems.

In what follows we use $R = 15$ and $\lambda = 10^5$. Figure 2 shows the predicted values vs. the given values, for the test data, over all 72 characteristics. The plot shows substantial residual variation but a clear trend, with correlation $r = 0{:}662$.

To assess predictive uncertainty we generate 5000 posterior samples as in Section 8, yielding samples from the posterior predictive distribution of the 72 characteristics for each of the 1000 test images. Symmetric credible intervals were computed for each characteristic of each image. The empirical coverage rates for the given values were 0:934 for 95% credible intervals and 0:887 for 90% credible intervals. The full posterior distributions for a small number of select characteristics, for a single test image, are shown in Figure 3 as an illustration of the results.

11 Discussion

In this article we have proposed a general framework for predicting one multiway array from another using the contracted tensor product. The simulation studies and facial image application illustrate the advantages of CP-rank regularization and ridge regularization in this framework. These two parameters define a broad class of models that are appropriate for a wide variety of scenarios. The CP assumption accounts for multi-way dependence in both the predictor and outcome array, and the ridge penalty accounts for auxiliary high-dimensionality and multi-collinearity of the predictors. However, several alternative regularization strategies are possible. The coefficient array can be restricted to have a more general Tucker structure (as in Li et al. (2013)), rather than a CP structure. A broad family of separable penalty functions, such as the separable L_2 penalty in (10), are straightforward to impose within the general framework using an alternating estimation scheme similar to that described in Zhou et al. (2013). In particular, a separable L_1 penalty has advantages when a solution that includes sparse subregions of each mode is desired. The alternating estimation scheme described herein for the non-separable L_2 penalty is not easily extended to alternative non-separable penalty functions.

We have described a simple Gaussian likelihood and a prior distribution for $\mathbb{B}$ that are motivated by the least-squares objective with non-separable L_2 penalty. The resulting probability model involves many simplifying assumptions, which may be over-simplified for some situations. In particular, the assumption of independent and homoscadastic error in the outcome array can be inappropriate for applications with auxiliary structure in the outcome. The array normal distribution (Akdemir and Gupta, 2011; Hoff et al., 2011) allows for multiway dependence and can be used as a more flexible model for the error covariance. Alternatively, envelope methods (Cook and Zhang, 2015) rely on a general technique to account for and ignore immaterial structure in the response and/or predictors of a predictive model. A tensor envelope is defined in Li and Zhang (2016), and its use in the tensor-on-tensor regression framework is an interesting direction for future work.

The approach to inference used herein is “semi-Bayesian”, in that the prior is limited to facilitate inference under the given penalized least-squares objective and is not intended to be subjective. Fully Bayesian approaches, such as using a prior for both the rank of the coefficient array and the shrinkage parameter, are another interesting direction of future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

A Consistency

Here we establish the consistency of the minimizer of the objective (9), for fixed dimension as $N \to \infty$, under general conditions.

Theorem 1

Assume model (2) holds for, where

1. For each response index $(q_1, \ldots, q_M)$, the errors $\mathbb{E}[n, q_1, \ldots, q_M]$ are independent and identically distributed (iid) for $n = 1, \ldots, N$, with mean 0 and finite second moment.
2. For each predictor index $(p_1, \ldots, p_L)$, $\mathbb{X}[n; p_1, \ldots, p_L]$ are iid for $n = 1, \ldots, N$ from a bounded distribution.
3. $\mathbb{B}_0$ has a rank R_0 factorization (6), where $\theta_0 = \{\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M\}$ is in the interior of a compact space Θ and is identifiable under the restrictions of Section 5.

For $R = R_0$ and fixed ridge penalty $\lambda \geq 0$, the minimizer of the objective (9), $\hat{\beta}_N$, converges to in probability as $N \to \infty$. Moreover, under the restrictions of Section 5 the factorization parameters $\hat{\theta}_N$ converge to θ_0 in probability as $N \to \infty$.

For Theorem 1 we require that the observations are iid, but within an observation the elements of the error array $\mathbb{E}$ or predictor array $\mathbb{X}$ may be correlated or from different distributions. Also, note that the correct rank is assumed for the estimator, but the result holds for any fixed penalty $\lambda \geq 0$. Theorem 1 applies to the global minimizer of objective (9), which may not be attained by the iterative algorithm of Section 7. The requirements that the predictors $\mathbb{X}$ are bounded and that Θ is compact are similar to those used to show the constancy of tensor regression under a normal likelihood model in Zhou et al. (2013). These requirements facilitate the use of Glivenko-Cantelli theory with a classical result on the asymptotic consistency of M-estimators (Van der Vaart, 2000); the proof is given below.

Proof

Let $\mathbb{B}(\theta)$ be the coefficient array resulting from $\theta = \{\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M\}$:

$$\mathbb{B}(\theta) = [\![\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M]\!] .$$

Let $M(\theta)$ be the expected squared error for a single observation:

$$M(\theta) = E\left(\left\| \mathbb{Y}_n - \left\langle \mathbb{X}_n, \mathbb{B}(\theta) \right\rangle \right\|_F^2\right),$$

which exists for all $\theta \in \Theta$ because the entries of $\mathbb{E}$ are assumed to have finite second moment. Let $M_N^\lambda(\theta)$ be the penalized empirical squared error loss (9) divided by N:

$$M_N^\lambda(\theta) = \frac{1}{N}\left(\sum_{n=1}^{N} \left\| \mathbb{Y}_n - \left\langle \mathbb{X}_n, \mathbb{B}(\theta) \right\rangle \right\|_F^2 \right) + \frac{\lambda}{N} \|\mathbb{B}(\theta)\|_F^2 .$$

From Theorem 5.7 of Van der Vaart (2000), the following three properties imply $\hat{\theta}$ is a consistent estimator of θ_0:

1. 1. $\inf_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) > M(\theta_0)$ $\varepsilon > 0$,
2. 2. $M_N^\lambda(\hat{\theta}_n) \leq M_N^\lambda(\theta_0) - O_P(1)$, where $O_P(1)$ defines a stochastically bounded sequence of random variables, and
3. 3. $\sup_{\theta \in \Theta} \left| M_N^\lambda(\theta) - M(\theta) \right| \xrightarrow{P} 0$.

Because $E(\mathbb{Y}) = \left\langle \mathbb{X}, \mathbb{B}(\theta_0) \right\rangle_L$, the coefficient array $\mathbb{B}(\theta_0)$ minimizes the expected squared error. Property 1 then follows from the identifiability of θ_0. For any $\theta \in \Theta$, $M_N^\lambda(\theta) - M(\theta)$ almost surely by the strong law of large numbers and the fact

$$\lim_{N \to \infty} \sup_{\theta \in \Theta} \frac{\lambda}{N} \|\mathbb{B}(\theta)\|_F^2 = 0. \quad (21)$$

Also, $M(\theta)$ is necessarily bounded over the compact space Θ. Thus, both $M_N(\hat{\theta}_n)$ and $M_n(\theta_0)$ are stochastically bounded, and property 2 follows. For property 3 it suffices to show uniform convergence of the unpenalized squared error $M_N^0(\theta)$, by

$$\sup_{\theta \in \Theta} \left| M_N^{\lambda}(\theta) - M(\theta) \right| \leq \sup_{\theta \in \Theta} \left| M_N^0(\theta) - M(\theta) \right| + \sup_{\theta \in \Theta} \frac{\lambda}{N} \|\mathbb{B}(\theta)\|_F^2$$

and (21). The uniform convergence of $M_N^0(\theta)$ can be verified by Glivenko-Cantelli theory.

Define

$$m_0(\mathbb{X}_n, \mathbb{Y}_n) = \left\| \mathbb{Y}_n - \langle \mathbb{X}_n, \mathbb{B}(\theta) \rangle \right\|_F^2.$$

The class $\{m : \theta \in \Theta\}$ is Glivenko-Cantelli, because Θ is compact and $m_0(\mathbb{X}_n, \mathbb{Y}_n)$ is continuous as a function of θ and bounded on Θ for any $(\mathbb{X}_n, \mathbb{Y}_n)$. Thus, property 3 holds and $\hat{\theta}$ is a consistent estimator of θ.

By the continuous mapping theorem, $\mathbb{B}(\hat{\theta})$ is also consistent estimator of the true coefficient array $\mathbb{B}(\theta)$ □.

B Posterior derivations

Here we derive the full conditional distributions of the factorization components for $\mathbb{B} = [\![\mathbf{U}_1, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M]\!]$, used in Section 8.

First, we consider the *a priori* conditional distributions that are implied by the spherical Gaussian prior for $\mathbb{B}$ (15). Here we derive the prior conditional for $\mathbf{U}_1$, pr($\mathbf{U}_1 / \mathbf{U}_2, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M$); the prior conditionals for $\mathbf{U}_2, \ldots, \mathbf{U}_L, \mathbf{V}_1, \ldots, \mathbf{V}_M$ are analogous, because the prior for $\mathbb{B}$ is permutation invariant over its $L + M$ modes. Let $\mathbf{b}_r^{(1)}$ give the vectorized form of the CP factorization without $\mathbf{U}_1$,

$$\mathbf{b}_r^{(1)} = \text{vec}(\mathbf{u}_{2r} \circ \cdots \circ \mathbf{u}_{Lr} \circ \mathbf{v}_{1r} \circ \cdots \circ \mathbf{v}_{Mr}),$$

and define the matrix $\mathbf{B}^{(1)} : Q \prod_{l=2}^{L} P_l \times R$ by $\mathbf{B}^{(1)} = \left[\mathbf{b}_1^{(1)} \cdots \mathbf{b}_R^{(1)} \right]$. Then

$$\text{vec}\left(\mathbf{U}_1 \mathbf{B}^{(1)^T} \right) = \text{vec}(\mathbb{B}) \sim N\left(0, \frac{\sigma^2}{\lambda} \mathbf{I}_{PQ \times PQ} \right),$$

and it follows that

$$\mathrm{pr}(\mathrm{vec}(\mathbf{U}_1)\,|\,\mathbf{U}_2,\ldots,\mathbf{U}_L,\mathbf{V}_1,\ldots,\mathbf{V}_M) = N\left(0,\left(\mathbf{B}^{(1)^T}\mathbf{B}^{(1)}\right)^{-1}\otimes\frac{\sigma^2}{\lambda}\mathbf{I}_{P_1\times P_1}\right).$$

The general model (2) implies

$$\mathbf{C}\mathrm{vec}(\mathbf{U}_1)+\mathrm{vec}(\mathbb{E})=\mathrm{vec}(\mathbb{Y}),$$

where $\mathbf{C}$ is defined as in (11). If $\mathbb{E}$ has independent $N(0,\sigma^2)$ entries, a direct application of the Bayesian linear model (Lindley and Smith, 1972) gives

$$\mathrm{pr}(\mathrm{vec}(\mathbf{U}_1)\,|\,\mathbf{U}_2,\ldots,\mathbf{U}_L,\mathbf{V}_1,\ldots,\mathbf{V}_M,\mathbb{Y},\mathbb{X},\sigma^2) = N(\boldsymbol{\mu}_1,\Sigma_1)$$

where

$$\boldsymbol{\mu}_1=\left(\mathbf{C}^T\mathbf{C}+\lambda\mathbf{B}^{(1)^T}\mathbf{B}^{(1)}\otimes I_{P_1\times P_1}\right)^{-1}\mathbf{C}^T\ \mathrm{vec}(\mathbb{Y})$$

and

$$\Sigma_1=\sigma^2\left(\mathbf{C}^T\mathbf{C}+\lambda\mathbf{B}^{(1)^T}\mathbf{B}^{(1)}\otimes I_{P_1\times P_1}\right)^{-1}.$$

Basic tensor algebra shows

$$\mathbf{B}^{(1)^T}\mathbf{B}^{(1)}=\mathbf{U}_2^T\mathbf{U}_2\cdot\cdots\cdot\mathbf{U}_L^T\mathbf{U}_L\cdot\mathbf{V}_1^T\mathbf{V}_1\cdot\cdots\cdot\mathbf{V}_M^T\mathbf{V}_M.$$

The posterior mean and variance for $\mathbf{U}_2,\ldots,\mathbf{U}_L$ are derived in an analogous way.

For the $\mathbf{V}_m's$ it suffices to consider $\mathbf{V}_M$, as the posterior derivations for $\mathbf{V}_1,\ldots,\mathbf{V}_{M-1}$ are analogous. The prior conditional for $\mathbf{V}_M$ is

$$\mathrm{pr}(\mathrm{vec}(\mathbf{V}_M)\,|\,\mathbf{U}_1,\ldots,\mathbf{U}_L,\mathbf{V}_1,\ldots,\mathbf{V}_{M-1}) = N\left(0,\left(\mathbf{B}^{(L+M)^T}\mathbf{B}^{(L+M)}\right)^{-1}\otimes\frac{\sigma^2}{\lambda}\mathbf{I}_{Q_M\times Q_M}\right)$$

and the general model (2) implies

$$\mathbf{D}\mathbf{V}_M+\mathbf{E}=\mathbf{Y}_m,$$

where $\mathbf{D}$ and $\mathbf{Y}_M$ are defined as in (12), and $\mathbf{E}$ has independent $N(0,\sigma^2)$ entries. Separate applications of the Bayesian linear model for each row of $\mathbf{V}_M$ gives

$$\mathrm{pr}(\mathrm{vec}(\mathbf{V}_M)\,|\,\mathbf{U}_1,\ldots,\mathbf{U}_L,\mathbf{V}_1,\ldots,\mathbf{V}_{M-1},\mathbb{Y},\mathbb{X},\sigma^2) = N(\boldsymbol{\mu}_{L+M},\Sigma_{L+M})$$

where

$$\boldsymbol{\mu}_{L+M} = \mathrm{vec}((\mathbf{D}^T\mathbf{D} + \lambda\mathbf{B}^{(L+M)^T}\mathbf{B}^{(L+M)})^{-1}\mathbf{D}^T\mathbf{Y}_M^T)$$

and

$$\Sigma_{L+M} = \sigma^2\left(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{B}^{(L+M)^T}\mathbf{B}^{(L+M)}\right)^{-1} \otimes \mathbf{I}_{Q_M \times Q_M}.$$

Basic tensor algebra shows

$$\mathbf{B}^{(L+M)^T}\mathbf{B}^{(L+M)} = \mathbf{U}_1^T\mathbf{U}_1 \cdot \cdots \cdot \mathbf{U}_L^T\mathbf{U}_L \cdot \mathbf{V}_1^T\mathbf{V}_1 \cdot \cdots \cdot \mathbf{V}_{M-1}^T\mathbf{V}_{M-1}.$$

C Proof of Proposition 1

Here we prove the equivalence of separable L_2 penalization and nuclear norm penalization stated in Proposition 1. The result is shown for predicting a vector from a three-way array, in which $\mathbb{B} = \mathbf{U}_1\mathbf{U}_2^T$. Analogous results exist for predicting a matrix from a matrix ($\mathbb{B} = \mathbf{U}_1\mathbf{V}_1^T$) and predicting a three-way array from a vector ($\mathbb{B} = \mathbf{V}_1\mathbf{V}_2^T$).

In the solution to

$$\underset{\mathrm{rank}(\mathbb{B}) \le R}{\arg\min} \left\|\mathbf{y} - \langle\mathbb{X},\mathbb{B}\rangle_2\right\|_F^2 + \lambda\sum_{l=1}^{2}\left\|\mathbf{U}_l\right\|_F^2 \quad (22)$$

the columns of $\mathbf{U}_1$ and $\mathbf{U}_2$, $\{\mathbf{u}_{11}, \ldots, \mathbf{u}_{1R}\}$ and $\{\mathbf{u}_{21}, \ldots, \mathbf{u}_{2R}\}$, must satisfy

$$\left\|\mathbf{u}_{1r}\right\|^2 = \left\|\mathbf{u}_{2r}\right\|^2 = \left\|\mathbf{u}_{1r}\mathbf{u}_{2r}^T\right\|_F \text{ for } r = 1, \ldots, R. \quad (23)$$

Here (23) follows from the general result that for $c > 0$,

$$\arg\min\{(a,b): ab = c\}a^2 + b^2 = (\sqrt{c},\sqrt{c}),$$

where $a = \|\mathbf{u}_{1r}\|^2$, $b = \|\mathbf{u}_{2r}\|^2$, and $c = \left\|\mathbf{u}_{1r}\mathbf{u}_{2r}^T\right\|_F^2$. Thus,

$$\sum_{l=1}^{2} \|\mathbf{U}_l\|_F^2 = \sum_{l=1}^{2} \sum_{r=1}^{R} \|\mathbf{u}_{lr}\|^2 \quad (24)$$
$$= 2 \sum_{r=1}^{R} \|\mathbf{u}_{1r}\mathbf{u}_{2r}^T\|_F,$$

Under orthogonality of the columns of $\mathbf{U}_1$ and $\mathbf{U}_2$, the non-zero singular values of $\mathbb{B}$ are $\{\|\mathbf{u}_{1r}\mathbf{u}_{2r}^T\|_F\}_{r=1}^R$, and therefore (24) is equal to $2\|\mathbb{B}\|_*$. It follows that (22) is equivalent to

$$\underset{\text{rank}(\mathbb{B}) \le R}{\arg\min} \|\mathbf{y} - \langle \mathbb{X}, \mathbb{B} \rangle_2\|_F^2 + 2\lambda \|\mathbb{B}\|_*.$$

D Correlated data simulation

Here we describe the results of a simulation study analogous to that in Section 9, but with correlation in the predictors $\mathbb{X}$ or in the response $\mathbb{Y}$. We simulate Gaussian data with an exponential spatial correlation structure using the R package `fields` (Douglas Nychka et al., 2015). The entries of $\mathbb{E}$ are assumed to be on a $Q_1 \times Q_2$ grid ($Q_1 = 5$, $Q_2 = 10$) with adjacent entries having distance 1. The entries of $\mathbb{X}$ are assumed to be on a $P_1 \times P_2$ grid ($P_1 = 15$, $P_2 = 20$) with adjacent entries having distance 1. The correlation between adjacent locations is $\rho = 0.6$ for each scenario, and the marginal variance of the entries is 1. Thus, data are simulated exactly as in Section 9.1, except for the correlation structure of $\mathbb{X}$ (step 1.) or $\mathbb{E}$ (step 3.).

The resulting RPE and credible interval coverage rates are shown in Table 4, which is analogous to Table 2 for the uncorrelated case. Interestingly, for penalized estimation and $n = 30$ the scenario with correlated $\mathbb{X}$ gives significantly better performance in terms of RPE than the scenario without correlation. This was unexpected, but may be because correlation in $\mathbb{X}$ discourages the algorithm from converging to a local minimum. For correlated $\mathbb{E}$ the results are often similar to the uncorrelated scenario but tend toward lower accuracy. In particular, the credible intervals tend to undercover more than for the uncorrelated scenario, or for the scenario with correlated $\mathbb{X}$. This is probably because correlation in $\mathbb{E}$ violates the assumed likelihood model for inference, while correlation in $\mathbb{X}$ does not.

E Full-rank comparison

Here we describe the results of a simulation study in which data are generated as in Section 9, but the model is estimated without low-rank constraints. For each simulation scenario, we consider two additional estimation approaches:

1. Unconstrained $\mathbb{X}$ and unconstrained $\mathbb{Y}$, in which the solution is given by independent ridge regression for each location of $\mathbb{Y}$ on the vectorized entries of $\mathbb{X}$.

2. Rank constraint for $\mathbb{X}$, but not for $\mathbb{Y}$, in which the solution is given by separate tensor regressions with ridge regularization for each location of $\mathbb{Y}$ on $\mathbb{X}$. That is, each location in $\mathbb{Y}$ is considered independently as a univariate outcome, and criteria (9) is optimized separately for each of the $Q_1 Q_2 = 50$ outcomes ($Q_1 = 5$, $Q_2 = 10$).

The resulting RPE under the different simulation scenarios and estimation approaches are shown in Table 5. Recall that each outcome has $N = 30$ or 120 observations; for approach 1 the total number of parameters for each outcome is $P_1 P_2 = 300$ ($P_1 = 15$ and $P_2 = 20$), and for approach 2 the total number of parameters for each outcome is $R(P_1 + P_2)$. Thus, without ridge regularization ($\lambda = 0$), the solution is always undefined for approach 1, and is also undefined under approach 2 for most values of N and R considered. With ridge regularization ($\lambda > 0$), the results under approach 1 are generally inferior to those under approach 2, and both approaches are inferior to the tensor-on-tensor approach presented in Section 9 with a low-rank constraint for both $\mathbb{X}$ and $\mathbb{Y}$. Moreover, while approach 1 is very fast because it is non-iterative, approach 2 was generally much slower than the tensor-on-tensor approach because it requires fitting a separate model for each location; an average, model estimation for approach 2 took 12 minutes per simulated dataset, and the tensor-on-tensor model took 3 minutes per dataset. This demonstrates the potential advantages of simultaneously allowing for low-rank dependence in both the predictors $\mathbb{X}$ and the outcomes $\mathbb{Y}$ in the tensor-on-tensor context.

F Cross-validation simulation

Here we describe the results of a simulation study to assess the use of cross-validation to select the parameters R and λ. Data were generated as in Section 9. For each dataset generated, we perform five-fold cross-validation for each pair (R, λ) over the grid $R = \{1, 2, 3, 4, 5\}$ and $\lambda = \{0, 0{:}5, 1, 5, 50\}$. The estimates $\left(\hat{R}, \hat{\lambda}\right)$ are selected as the pair that gives the lowest mean squared prediction error across test folds. The resulting estimates are summarized in Table 6. The accuracy of the selected ranks varied widely depending on the simulation scenario. For the scenario with larger sample size and larger signal, $N = 120$, SNR = 5, the true underlying rank was selected for every replication. With lower sample size ($N = 30$) the true rank was selected for only approximately 1/3 of the replications, and for remaining replications the results were evenly split between overestimation and underestimation of the rank. The selected value for $\hat{\lambda}$ generally increased with lower signal and lower sample size, which is consistent with the values of λ that gave optimal performance for the larger test set (see Table 2).

References

Akdemir D, Gupta AK. Array variate random variables with multiway Kronecker delta covariance matrix structure. J Algebr Stat. 2011; 2(1):98–113.

Bader BW, Kolda TG. Algorithm 862: Matlab tensor classes for fast algorithm prototyping. ACM Transactions on Mathematical Software (TOMS). 2006; 32(4):635–653.

Chen B, He S, Li Z, Zhang S. Maximum block improvement and polynomial optimization. SIAM Journal on Optimization. 2012; 22(1):87–107.

Cook RD, Zhang X. Foundations for envelope models and methods. Journal of the American Statistical Association. 2015; 110(510):599–611.

De Martino F, De Borst AW, Valente G, Goebel R, Formisano E. Predicting EEG single trial responses with simultaneous fMRI and relevance vector machine regression. Neuroimage. 2011; 56(2):826–836. [PubMed: 20691274]

Douglas NychkaReinhard FurrerJohn PaigeStephan Sain. fields: Tools for spatial data. 2015 R package version 9.0.

Friedland S, Lim LH. Nuclear norm of higher-order tensors. arXiv preprint arXiv: 1410.6072. 2014

GTEx Consortium. The genotype-tissue expression (GTEx) pilot analysis: Multi-tissue gene regulation in humans. Science. 2015; 348(6235):648–660. [PubMed: 25954001]

Guo W, Kotsia I, Patras I. Tensor learning for regression. IEEE Transactions on Image Processing. 2012; 21(2):816–827. [PubMed: 21859620]

Harshman RA. Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multi-modal factor analysis. UCLA Working Papers in Phonetics. 1970; 16:1–84.

Hassner T, Harel S, Paz E, Enbar R. Effective face frontalization in unconstrained images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4295–4304.

Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1):55–67.

Hoff PD. Multilinear tensor regression for longitudinal relational data. The Annals of Applied Statistics. 2015; 9(3):1169–1193. [PubMed: 27458495]

Hoff PD, et al. Separable covariance arrays via the Tucker product, with applications to multivariate relational data. Bayesian Analysis. 2011; 6(2):179–196.

Huster RJ, Debener S, Eichele T, Herrmann CS. Methods for simultaneous eeg-fmri: an introductory review. Journal of Neuroscience. 2012; 32(18):6053–6060. [PubMed: 22553012]

Izenman AJ. Reduced-rank regression for the multivariate linear model. Journal of multivariate analysis. 1975; 5(2):248–264.

Jansen M, White TP, Mullinger KJ, Liddle EB, Gowland PA, Francis ST, Bowtell R, Liddle PF. Motion-related artefacts in eeg predict neuronally plausible patterns of activation in fmri data. Neuroimage. 2012; 59(1):261–270. [PubMed: 21763774]

Kim YD, Choi S. International Conference on Biometrics. Springer; 2007. Color face tensor factorization and slicing for illumination-robust recognition; 19–28.

Kolda TG, Bader BW. Tensor decompositions and applications. SIAM review. 2009; 51(3):455–500.

Kumar N, Berg AC, Belhumeur PN, Nayar SK. 2009 IEEE 12th International Conference on Computer Vision. IEEE; 2009. Attribute and simile classifiers for face verification; 365–372.

Learned-Miller E, Huang GB, RoyChowdhury A, Li H, Hua G. Advances in Face Detection and Facial Image Analysis. Springer; 2016. Labeled faces in the wild: A survey; 189–248.

Li L, Zhang X. Parsimonious tensor response regression. Journal of the American Statistical Association. 2016 (just-accepted).

Li X, Zhou H, Li L. Tucker tensor regression and neuroimaging analysis. arXiv preprint arXiv: 1304.5637. 2013

Lindley DV, Smith AF. Bayes estimates for the linear model. Journal of the Royal Statistical Society. Series B (Methodological). 1972; 34(1):1–41.

Lock EF, Li G. Supervised multiway factorization. arXiv preprint arXiv: 1609.03228. 2016

Lock EF, Nobel AB, Marron JS. Comment on Population Value Decomposition, a framework for the analysis of image populations. Journal of the American Statistical Association. 2011; 106(495): 798–802.

Lyu T, Lock EF, Eberly LE. Discriminating sample groups with multi-way data. Biostatistics. 18:434–450.

Miranda M, Zhu H, Ibrahim JG. TPRM: Tensor partition regression models with applications in imaging biomarker detection. arXiv preprint arXiv: 1505.05482. 2015

Mukherjee A, Zhu J. Reduced rank ridge regression and its kernel extensions. Statistical analysis and data mining. 2011; 4(6):612–622. [PubMed: 22993641]

Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, De T, Coin L, de Silva R, Cookson MR, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nature neuroscience. 2014; 17(10):1418–1428. [PubMed: 25174004]

Raskutti G, Yuan M. Convex regularization for high-dimensional tensor regression. arXiv preprint arXiv: 1512.01215. 2015

Sidiropoulos ND, Bro R. On the uniqueness of multilinear decomposition of N-way arrays. Journal of Chemometrics. 2000; 14(3):229–239.

Sirovich L, Kirby M. Low-dimensional procedure for the characterization of human faces. Journal of the Optical Society of America A. 1987; 4(3):519–524.

Spiegelhalter DJ, Best NG, Carlin BP, Linde A. The deviance information criterion: 12 years on. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2014; 76(3):485–493.

Sun WW, Li L. Sparse low-rank tensor response regression. arXiv preprint arXiv: 1609.04523. 2016

Sundberg R. Continuum regression and ridge regression. Journal of the Royal Statistical Society. Series B (Methodological). 1993:653–659.

Tao D, Li X, Wu X, Hu W, Maybank SJ. Supervised tensor learning. Knowledge and information systems. 2007; 13(1):1–42.

Tucker LR. Some mathematical notes on three-mode factor analysis. Psychometrika. 1966; 31(3):279–311. [PubMed: 5221127]

Turk M, Pentland A. Eigenfaces for recognition. Journal of cognitive neuroscience. 1991; 3(1):71–86. [PubMed: 23964806]

Van der Vaart AW. Asymptotic statistics. Vol. 3. Cambridge university press; 2000.

Vasilescu MAO, Terzopoulos D. European Conference on Computer Vision. Springer; 2002. Multilinear analysis of image ensembles: Tensorfaces; 447–460.

Wimalawarne K, Tomioka R, Sugiyama M. Theoretical and experimental analyses of tensor-based regression and classification. Neural computation. 2016; 28(4):686–715. [PubMed: 26890354]

Zhang X, Li L, Zhou H, Shen D, et al. Tensor generalized estimating equations for longitudinal imaging analysis. arXiv preprint arXiv: 1412.6592. 2014

Zhou H, Li L, Zhu H. Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association. 2013; 108(502):540–552. [PubMed: 24791032]

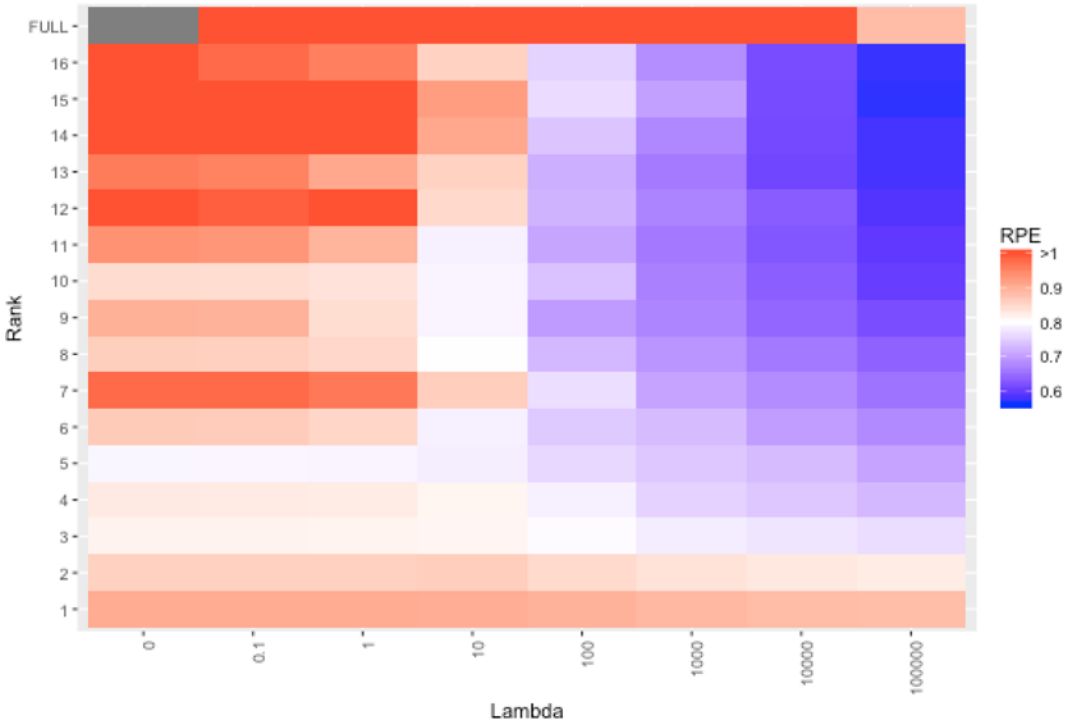

Figure 1.

Relative prediction error for characteristics of out-of-sample images for different parameter choices. The top row (full rank) gives the results under separate ridge regression models for each outcome without rank restriction.

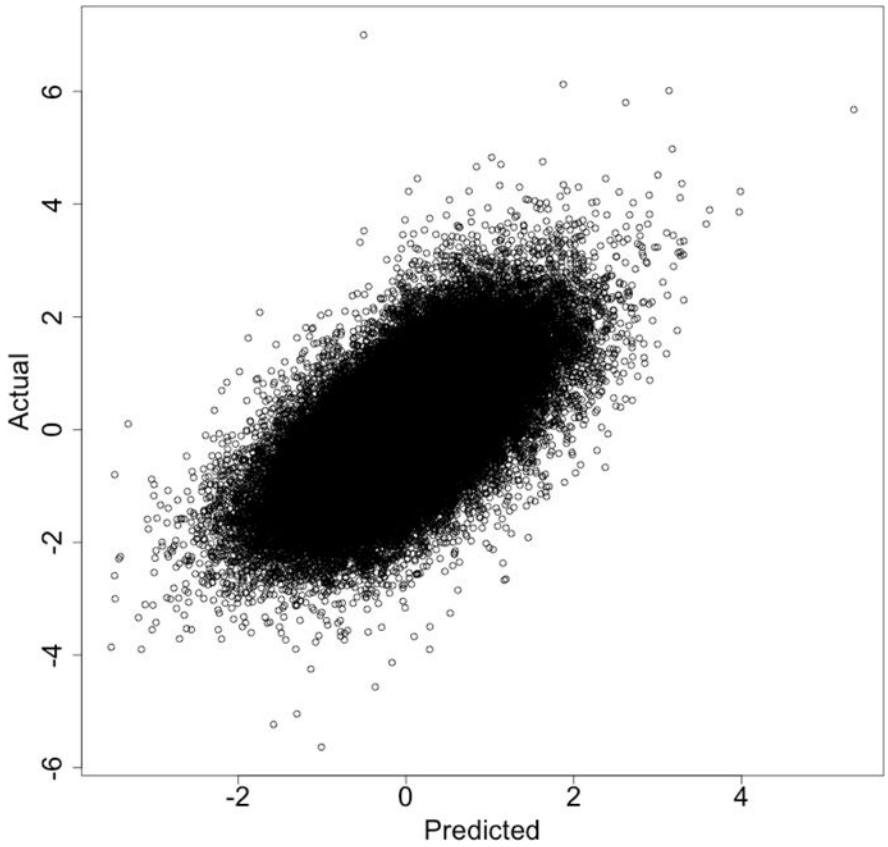

Figure 2.
Actual vs. predicted values for 1000 test images across 72 characteristics.

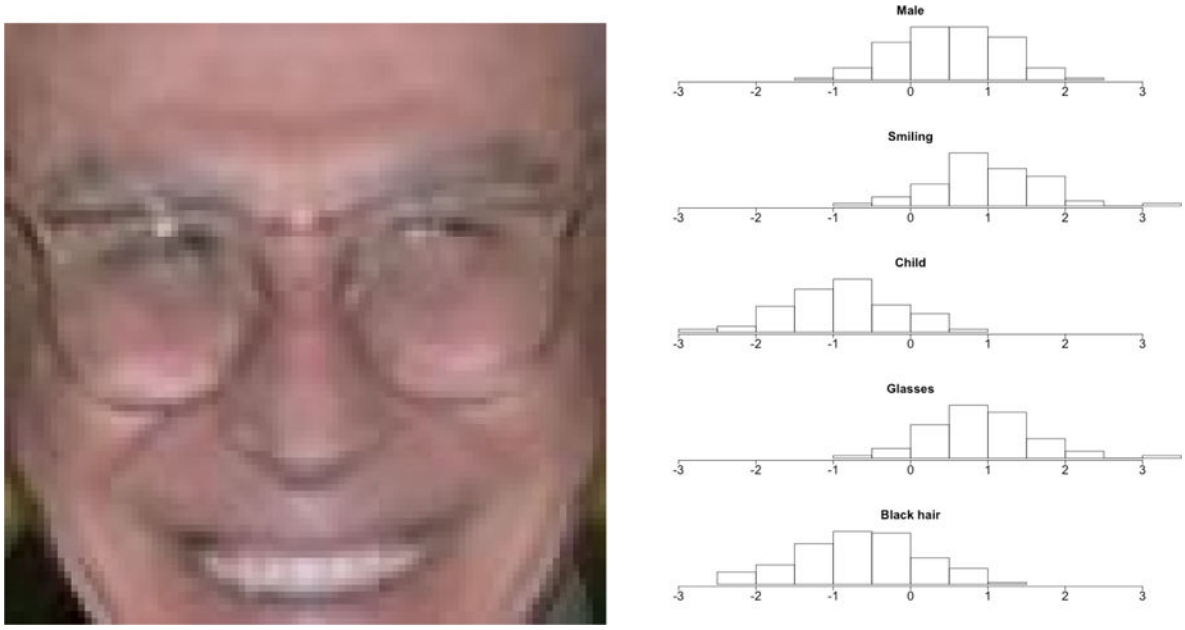

Figure 3.
Example test image (left), and its posterior samples for 5 select characteristics (right).

Table 1

Marginal mean RPE for no signal, $R = 0$.

Training samples	***N* = 30** **1.48**		***N* = 120** **1.08**		
Regularization	$\lambda = 0$ 2.00	$\lambda = 0.5$ 1.14	$\lambda = 1$ 1.13	$\lambda = 5$ 1.08	$\lambda = 50$ 1.02
Assumed rank	$\hat{R} = 1$ 1.04	$\hat{R} = 2$ 1.12	$\hat{R} = 3$ 1.20	$\hat{R} = 4$ 1.32	$\hat{R} = 5$ 1.70

Table 2

The top panel shows mean RPE by regularization for different scenarios using correct assumed ranks. The bottom panel shows the coverage rate for 95% credible intervals, and their mean length relative to the standard deviation of $\mathbb{Y}$.

RPE (std error)	**λ = 0**	**λ = 0.5**	**λ = 1**	**λ = 5**	**λ = 50**
N = 120, SNR = 5	**0.04** (0.01)	**0.04** (0.01)	**0.04** (0.01)	**0.05** (0.01)	**0.20** (0.01)
N = 120, SNR = 1	**0.52** (0.01)	**0.52** (0.01)	**0.52** (0.01)	**0.52** (0.01)	**0.59** (0.01)
N = 30, SNR = 5	**1.64** (0.12)	**0.74** (0.05)	**0.70** (0.04)	**0.63** (0.02)	**0.77** (0.01)
N = 30, SNR = 1	**1.90** (0.15)	**1.07** (0.04)	**1.03** (0.04)	**0.92** (0.02)	**0.91** (0.01)
Coverage (length)	λ = **0**	λ = **0.5**	λ = **1**	λ = **5**	λ = **50**
N = 120, SNR = 5	**0.95** (0.77)	**0.95** (0.77)	**0.95** (0.77)	**0.94** (0.80)	**0.91** (1.43)
N = 120, SNR = 1	**0.95** (2.79)	**0.95** (2.79)	**0.95** (2.79)	**0.95** (2.79)	**0.94** (2.90)
N = 30, SNR = 5	**0.91** (3.10)	**0.68** (1.11)	**0.65** (1.04)	**0.68** (1.10)	**0.84** (2.10)
N = 30, SNR = 1	**0.98** (4.75)	**0.95** (3.56)	**0.94** (3.36)	**0.91** (3.05)	**0.91** (3.18)

Table 3

Mean RPE by assumed rank for different true ranks for $N = 120$, S2N= 1, and $\lambda = 0$

	$\hat{R} = 1$	$\hat{R} = 2$	$\hat{R} = 3$	$\hat{R} = 4$	$\hat{R} = 5$
$R = 1$	**0.50**	0.54	0.55	0.56	0.57
$R = 2$	0.64	**0.50**	0.53	0.53	0.54
$R = 3$	0.71	0.58	**0.53**	0.55	0.57
$R = 4$	0.78	0.67	0.57	**0.51**	0.53
$R = 5$	0.81	0.68	0.61	0.55	**0.53**

Table 4

Mean RPE by regularization and coverage rate for correlated $\mathbb{X}$ or correlated $\mathbb{E}$ using correct assumed ranks. The coverage rates are for 95% credible intervals, and their mean length relative to the standard deviation of $\mathbb{Y}$ is shown.

Correlated $\mathbb{X}$					
RPE (std error)	**λ = 0**	**λ = 0.5**	**λ = 1**	**λ = 5**	**λ = 50**
N = 120, SNR = 5	**0.04** (0.01)	**0.04** (0.01)	**0.04** (0.01)	**0.05** (0.01)	**0.19** (0.01)
N = 120, SNR = 1	**0.52** (0.01)	**0.52** (0.01)	**0.52** (0.01)	**0.52** (0.01)	**0.58** (0.01)
N = 30, SNR = 5	**1.70** (0.17)	**0.43** (0.03)	**0.43** (0.03)	**0.45** (0.02)	**0.58** (0.01)
N = 30, SNR = 1	**1.62** (0.11)	**0.87** (0.02)	**0.83** (0.02)	**0.79** (0.02)	**0.80** (0.01)
Coverage (length)	**λ = 0**	**λ = 0.5**	**λ = 1**	**λ = 5**	**λ = 50**
N = 120, SNR = 5	**0.95** (0.78)	**0.95** (0.78)	**0.95** (0.78)	**0.94** (0.83)	**0.92** (1.42)
N = 120, SNR = 1	**0.95** (2.80)	**0.95** (2.80)	**0.95** (2.80)	**0.95** (2.80)	**0.94** (2.92)
N = 30, SNR = 5	**0.93** (3.10)	**0.76** (1.11)	**0.74** (1.04)	**0.74** (1.10)	**0.84** (2.10)
N = 30, SNR = 1	**0.98** (4.75)	**0.95** (3.56)	**0.94** (3.36)	**0.92** (3.05)	**0.92** (3.18)
Correlated $\mathbb{E}$					
RPE (std error)	**λ = 0**	**λ = 0.5**	**λ = 1**	**λ = 5**	**λ = 50**
N = 120, SNR = 5	**0.04** (0.01)	**0.04** (0.01)	**0.04** (0.01)	**0.04** (0.01)	**0.19** (0.01)
N = 120, SNR = 1	**0.59** (0.01)	**0.59** (0.01)	**0.59** (0.01)	**0.58** (0.01)	**0.60** (0.01)
N = 30, SNR = 5	**2.03** (0.22)	**0.72** (0.03)	**0.68** (0.03)	**0.63** (0.02)	**0.77** (0.01)
N = 30, SNR = 1	**1.86** (0.13)	**1.10** (0.02)	**1.04** (0.02)	**0.93** (0.01)	**0.91** (0.01)
Coverage (length)	**λ = 0**	**λ = 0.5**	**λ = 1**	**λ = 5**	**λ = 50**
N = 120, SNR = 5	**0.95** (0.77)	**0.95** (0.77)	**0.95** (0.77)	**0.94** (0.79)	**0.91** (1.41)
N = 120, SNR = 1	**0.93** (2.75)	**0.93** (2.75)	**0.93** (2.75)	**0.93** (2.75)	**0.93** (2.88)
N = 30, SNR = 5	**0.91** (2.98)	**0.68** (1.10)	**0.66** (1.01)	**0.67** (1.05)	**0.84** (2.09)
N = 30, SNR = 1	**0.98** (4.89)	**0.94** (3.51)	**0.93** (3.28)	**0.89** (2.92)	**0.91** (3.12)

Table 5

Mean RPE by regularization parameter (λ), under three estimation approaches: 1. with no rank constraints, 2. with rank constraint for $\mathbb{X}$ but independent models for each entry of $\mathbb{Y}$, and 3. with low-rank dependence for both $\mathbb{X}$ and $\mathbb{Y}$ (reproduced from Table 2).

1. Unconstrained $\mathbb{X}$, unconstrained $\mathbb{Y}$

RPE (std error)	**λ = 0**	**λ = 0.5**	**λ = 1**	**λ = 5**	**λ = 50**
N = 120, SNR = 5	NA	**0.64** (0.01)	**0.64** (0.01)	**0.63** (0.01)	**0.64** (0.01)
N = 120, SNR = 1	NA	**1.14** (0.01)	**1.14** (0.01)	**1.11** (0.01)	**0.99** (0.01)
N = 30, SNR = 5	NA	**0.91** (0.01)	**0.91** (0.01)	**0.91** (0.01)	**0.91** (0.01)
N = 30, SNR = 1	NA	**1.00** (0.01)	**1.00** (0.01)	**1.00** (0.01)	**0.99** (0.01)

2. Constrained $\mathbb{X}$, unconstrained $\mathbb{Y}$

RPE (std error)	λ = **0**	λ = **0.5**	λ = **1**	λ = **5**	λ = **50**
N = 120, SNR = 5	NA	**0.30** (0.01)	**0.29** (0.01)	**0.27** (0.01)	**0.38** (0.01)
N = 120, SNR = 1	NA	**1.72** (0.01)	**1.62** (0.01)	**1.30** (0.01)	**0.88** (0.01)
N = 30, SNR = 5	NA	**0.91** (0.01)	**0.90** (0.01)	**0.87** (0.01)	**0.86** (0.01)
N = 30, SNR = 1	NA	**1.20** (0.01)	**1.18** (0.01)	**1.12** (0.01)	**1.00** (0.01)

3. Constrained $\mathbb{X}$, Constrained $\mathbb{Y}$

RPE (std error)	λ = **0**	λ = **0.5**	λ = **1**	λ = **5**	λ = **50**
N = 120, SNR = 5	**0.04** (0.01)	**0.04** (0.01)	**0.04** (0.01)	**0.05** (0.01)	**0.20** (0.01)
N = 120, SNR = 1	**0.52** (0.01)	**0.52** (0.01)	**0.52** (0.01)	**0.52** (0.01)	**0.59** (0.01)
N = 30, SNR = 5	**1.64** (0.12)	**0.74** (0.05)	**0.70** (0.04)	**0.63** (0.02)	**0.77** (0.01)
N = 30, SNR = 1	**1.90** (0.15)	**1.07** (0.04)	**1.03** (0.04)	**0.92** (0.02)	**0.91** (0.01)

Table 6

Rank selection accuracy and selected regularization parameter (λ) under five-fold cross-validation, averaged over replications for four different scenarios.

	$\hat{R} = R$	$\hat{R} < R$	$\hat{R} > R$	**Mean** $\hat{\lambda}$
N = 120, SNR = 5	100%	0%	0%	0.03
N = 120, SNR = 1	88%	12%	0%	2.14
N = 30, SNR = 5	34%	32%	34%	16.1
N = 30, SNR = 1	30%	38%	32%	27.8