Electronic Journal of Statistics
Vol. 11 (2017) 3703–3737
ISSN: 1935-7524
DOI: 10.1214/17-EJS1330

Adaptive higher-order spectral estimators

David Gerard

Department of Human Genetics
University of Chicago, Chicago, IL, 60637
e-mail: dcgerard@uchicago.edu

and

Peter Hoff

Department of Statistical Science
Duke University, Durham, NC, 27708
e-mail: peter.hoff@duke.edu

Abstract: Many applications involve estimation of a signal matrix from a noisy data matrix. In such cases, it has been observed that estimators that shrink or truncate the singular values of the data matrix perform well when the signal matrix has approximately low rank. In this article, we generalize this approach to the estimation of a tensor of parameters from noisy tensor data. We develop new classes of estimators that shrink or threshold the mode-specific singular values from the higher-order singular value decomposition. These classes of estimators are indexed by tuning parameters, which we adaptively choose from the data by minimizing Stein's unbiased risk estimate. In particular, this procedure provides a way to estimate the multilinear rank of the underlying signal tensor. Using simulation studies under a variety of conditions, we show that our estimators perform well when the mean tensor has approximately low multilinear rank, and perform competitively when the signal tensor does not have approximately low multilinear rank. We illustrate the use of these methods in an application to multivariate relational data.

MSC 2010 subject classifications: Primary 62H12; secondary 15A69, 62C99, 91D30.
Keywords and phrases: Higher-order SVD, network, relational data, shrinkage, SURE, tensor.

Received February 2017.

Contents

1. Introduction

Tensor data arise in fields as diverse as relational data [23], neuroimaging [53, 31], psychometrics [27], chemometrics [42, 4], signal processing [9], and machine learning [45], among others [30]. A tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ with $p_k \in \{1, 2, \ldots\}$ of order K is a K-way array where the elements $\mathcal{X}_{[i_1,\ldots,i_K]}$ are indexed by $i_k \in \{1, 2, \ldots, p_k\}$ for $k = 1, \ldots, K$. For example, a multivariate relational dataset can be expressed as a tensor, where element $\mathcal{X}_{[i,j,t]}$ of the tensor is the tth relation between actors i and j.

Often, a tensor is corrupted by noise. The model we consider for this is:

$$\begin{aligned} &\mathcal{X} = \Theta + \mathcal{E},\ \mathcal{E}_{[i_1,\ldots,i_K]} \sim N(0, \tau^2) \text{ independent} \\ &\quad \text{for } i_k = 1, \ldots, p_k, \text{ and } k = 1, \ldots, K, \end{aligned} \tag{1}$$

where $\Theta \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the signal and $\mathcal{E} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the additive Gaussian measurement error or noise with mean 0 and various τ^2. The performance of an estimator $t(\mathcal{X}) \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ can be evaluated by statistical risk under quadratic loss, i.e. mean squared error (MSE):

$$\text{MSE}(t(\mathcal{X})) = E_\Theta[\|\Theta - t(\mathcal{X})\|^2] = \sum_{\mathbf{i}} E_\Theta[(\Theta_{[\mathbf{i}]} - t(\mathcal{X})_{[\mathbf{i}]})^2], \tag{2}$$

where $\mathbf{i} = (i_1, \ldots, i_K)$ is a K-tuple of tensor indices.

In the matrix variate case, $X \in \mathbb{R}^{p \times n}$, an investigator often believes that the mean is well approximated by a low rank matrix. There has been much work on "denoising" (or mean estimation) in matrix variate data by using this knowledge. A typical estimation scheme begins by computing the singular value decomposition (SVD) of X:

$$X = UDV^T, \tag{3}$$

where, in the case $n \geq p$, $U \in \mathbb{R}^{p \times p}$ is orthogonal, $D = \text{diag}(\sigma_1, \ldots, \sigma_p)$ with $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$, and $V \in \mathbb{R}^{n \times p}$ contains orthonormal columns. The columns of U and V are, respectively, the left and right singular vectors of X and the diagonal elements of D are the singular values. A key property of the SVD is that the number of non-zero singular values of X is precisely the rank of X. One widely studied approach to estimating Θ when it is assumed that Θ has

nearly low rank is to shrink the singular values of X towards 0 while keeping the singular vectors unchanged, thereby inducing an (approximately) low rank estimate. The resulting "spectral" estimator $t(\mathcal{X})$ of Θ then takes the form $t(\mathcal{X}) = Uf(D)V^T$ where $f(D) = \text{diag}(f_1(\sigma_1), \ldots, f_K(\sigma_K))$ and each $f_i(\cdot)$ shrinks the singular values towards 0. These estimators are orthogonally equivariant, meaning that $t(WXZ^T) = Wt(X)Z^T$ for orthogonal matrices W, Z [40].

Early work on singular value shrinkage estimation from a non-statistical perspective began with [14], where they proved that the best rank r approximation to the data matrix $X \in \mathbb{R}^{p\times n}$ (in terms of sum of squared differences from X) is found with the shrinkage function:

$$f_i(\sigma_i) = \sigma_i 1(i \leq r), \tag{4}$$

where $1(\cdot)$ is the indicator function. We call (4) the truncation estimator. However, approximating the data X well is not the same as estimating the underlying signal Θ well. In terms of estimating Θ, the matrix X is unbiased, minimax, and the maximum likelihood estimator under normally distributed errors. However, it is well known that shrinkage estimators, such at that of [44] can uniformly dominate X in terms of risk. This seminal shrinkage estimator, in the context of matrix estimation, is given by

$$f_i(\sigma_i) = \left(1 - \frac{\lambda}{\sum_{i=1}^p \sigma_i^2}\right)\sigma_i, \tag{5}$$

where $\lambda > 0$ is some tuning parameter. For data that exhibit associations between the rows and/or columns of the mean matrix, the estimator of [15], given by

$$f_i(\sigma_i) = \sigma_i - \frac{\lambda}{\sigma_i}, \tag{6}$$

was introduced and results in different amounts of shrinkage for each singular value. [18] improved upon this estimator with a generalization of both (5) and (6), given by

$$f_i(\sigma_i) = \left(1 - \frac{\gamma}{\sum_{i=1}^p \sigma_i^2}\right)\sigma_i - \frac{\lambda}{\sigma_i}, \tag{7}$$

where $\lambda > 0$ and $\gamma > 0$ are tuning parameters.

More recent work has focused on estimators whose functions $f_i(\cdot)$ induce sparsity in the singular values, which may be more appropriate than (5), (6), and (7) in cases where the true signal itself has (approximately) low rank. Motivated by penalized maximum likelihood estimation, the hard-thresholding estimator

$$f_i(\sigma_i) = \sigma_i 1(\sigma_i \geq \lambda) \tag{8}$$

and the soft-thresholding estimator

$$f_i(\sigma_i) = (\sigma_i - \lambda)_+ \tag{9}$$

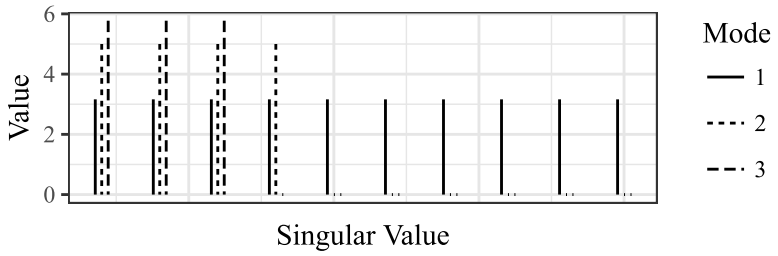

FIG 1. *Mode-specific singular values of simulated tensor with full rank along first mode and low-ranks along second and third modes.*

were introduced [6, for example]. Here, $(y)_+ = \max(y, 0)$ is the "positive part" function. A clever shrinkage function that includes (8), (9), and a truncated version of (6) [50] as special cases is that of [25]:

$$f_i(\sigma_i) = \sigma_i \left(1 - \frac{\lambda^\gamma}{\sigma_i^\gamma}\right)_+ . \tag{10}$$

This estimator was inspired by the adaptive LASSO [56]. A variety of other shrinkage estimators have also been developed [37, 40].

All of these estimators are specific to matrix-variate data. If one were to apply these matrix methods to a tensor, one would first convert the tensor into a matrix. For a K-dimensional tensor, such "matricization" destroys the indexing structure along all but one of the dimensions. This may be detrimental to estimation if, in addition to a data set having approximately low rank, it also has approximately low *multilinear* rank (see Section 2), that is, "matricizing" along each index set, or "mode", results in a low rank matrix.

An extreme simulated example that exhibits this phenomenon is presented in Figure 1. There, we plotted the mode-specific singular values of a tensor that we generated to have full rank along one mode and low ranks along two modes. That is, we plotted the singular values of each matricization of the tensor. If an analyst were presented with a noisy version of this tensor and only matricizing along the first mode, then they would only observe a noisy realization of the solid lines, which would suggest the data are full rank. However, the second and third modes have low-rank structure and shrinking the singular values along these additional modes may improve estimation.

In this article, we introduce a family of estimators that shrink tensor-valued data towards having (approximately) low multilinear rank. We perform this shrinkage on a reparameterization of the higher-order singular value decomposition (HOSVD) of [11], where we shrink the mode-specific singular values of the data tensor towards zero. We consider classes of such "higher-order spectral estimators", where a class is defined by a mode-specific shrinkage function indexed by a tuning parameter. We propose to adaptively select the tuning parameters by minimization of an unbiased estimate of the risk.

Our paper is organized as follows. In Section 2, we review tensors and the HOSVD. We then present how one may define functions that shrink the mode-specific singular values of the HOSVD. In particular, we present two specific estimators that shrink the data tensor towards having (approximately) low multilinear rank and provide some discussion on the intuition behind these estimators. In Section 3, we review Stein's unbiased risk estimates (SURE), then derive the SURE for a broad class of higher-order spectral estimators. In Section 4 we present simulations demonstrating that (1) tensor specific methods perform better when the mean tensor has approximately low multilinear rank; (2) when the mean tensor has low multilinear rank our methods accurately estimate the multilinear rank; and (3) tensor specific methods perform competitively when the signal tensor does not have approximately low multilinear rank. In Section 5 we illustrate the use of these methods in an application to multivariate relational data. We finish with a discussion in Section 6.

2. The higher-order SVD and higher-order spectral estimators

Some tensor data sets have approximately low *multilinear rank*, which we now define. Recall that the rank of a matrix is the dimension of the vector space spanned by its columns and rows. Define the k-mode vectors of a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ as the p_k-dimensional vectors formed from $\mathcal{X}$ by varying i_k and keeping the other indices fixed. The k-mode rank r_k is the dimension of the span of the k-mode vectors, and the multilinear rank of the K-order tensor $\mathcal{X}$ is the K-tuple, $(r_1, \ldots, r_K)$. Define the k-mode matricization [29], or k-mode unfolding, of $\mathcal{X}$ to be $\mathcal{X}_{(k)} \in \mathbb{R}^{p_k \times p/p_k}$ (with $p = \prod_{k=1}^{K} p_k$) where element $(i_1, \ldots, i_K)$ in $\mathcal{X}$ maps to element (i_k, j) in $\mathcal{X}_{(k)}$ where

$$j = 1 + \sum_{\substack{n=1 \\ n \neq k}}^{K} (i_n - 1) J_n \text{ with } J_n = \prod_{\substack{m=1 \\ m \neq k}}^{n-1} p_m.$$

Then, equivalently, r_k is the rank of $\mathcal{X}_{(k)}$.

The SVD, presented in Section 1, has been used to shrink matrix valued data towards low rank. One generalization of the SVD to tensors is the HOSVD of [11], which relates directly to multilinear rank.

Definition 1 (HOSVD of [11])**.** Let $\mathcal{X}_{(k)} = U_k D_k V_k^T$ be the SVD of each k-mode unfolding of $\mathcal{X}$. Let $\mathcal{S} = (U_1^T, \ldots, U_K^T) \cdot \mathcal{X}$, then

$$\mathcal{X} = (U_1, \ldots, U_K) \cdot \mathcal{S} \tag{11}$$

is the higher-order singular value decomposition (HOSVD).

The product "$\cdot$" in (11) between a list of matrices, $\{U_1, \ldots, U_K\}$ for $U_k \in \mathbb{R}^{p_k \times p_k}$, and a tensor, $\mathcal{S} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is called the *Tucker product*. The Tucker product is defined through the k-mode matricizations of $(U_1, \ldots, U_K) \cdot \mathcal{S}$:

$$\begin{aligned} &\mathcal{X} = (U_1, \ldots, U_K) \cdot \mathcal{S} \\ &\Leftrightarrow \mathcal{X}_{(k)} = U_k \mathcal{S}_{(k)} (U_K^T \otimes \cdots \otimes U_{k+1}^T \otimes U_{k-1}^T \otimes \cdots \otimes U_1^T) = U_k \mathcal{S}_{(k)} U_{-k}^T, \end{aligned}$$

where "$\otimes$" is the Kronecker product. The "core array", $\mathcal{S}$ has the property of *all-orthogonality* where

$$\mathcal{S}_{(k)}\mathcal{S}_{(k)}^T = D_k^2 \text{ for all } k = 1, \ldots, K.$$

The HOSVD is multilinear rank-revealing in the same way the SVD is rank-revealing. That is, let $D_k = (\mathcal{S}_{(k)}\mathcal{S}_{(k)}^T)^{1/2} = \text{diag}(\sigma_1^k, \ldots, \sigma_{p_k}^k)$ be the mode specific singular values of $\mathcal{X}$. Then the multilinear rank of $\mathcal{X}$ is $(r_1, \ldots, r_K)$ if D_k contains r_k non-zero mode-specific singular values. In the core array, this is equivalent to $\mathcal{S}$ containing zeros everywhere except in one of the "corners": $\mathcal{S}_{[1:r_1,\ldots,1:r_K]}$, where $1 : r_k = 1, \ldots, r_k$. It is possible, then, to shrink $\mathcal{S}$ towards having (approximately) low multilinear rank by shrinking the elements in $\mathcal{S}$ towards 0. We propose doing this via a re-parameterization of $\mathcal{S}$, given as follows:

$$\begin{aligned} \mathcal{X} &= (U_1, \ldots, U_K) \cdot (D_1, \ldots, D_K) \cdot (D_1^{-1}, \ldots, D_K^{-1}) \cdot \mathcal{S} \\ &= (U_1, \ldots, U_K) \cdot (D_1, \ldots, D_K) \cdot \mathcal{V}, \end{aligned} \tag{12}$$

where $\mathcal{S} = (D_1, \ldots, D_K) \cdot \mathcal{V}$. Our higher-order spectral estimators shrink $\mathcal{S}$ by shrinking each mode-specific D_k. We abuse notation a little by allowing "$\cdot$" to also represent a binary operator between two lists of matrices whose operation is component-wise multiplication. This should not cause confusion because $(A_1B_1, \ldots, A_KB_K) \cdot \mathcal{C} = (A_1, \ldots, A_K) \cdot [(B_1, \ldots, B_K) \cdot \mathcal{C}]$.

Using reparameterization (12), we now define higher-order spectral estimators of Θ under the model (1).

Definition 2. Let $\mathcal{X} = (U_1, \ldots, U_K) \cdot (D_1, \ldots, D_K) \cdot \mathcal{V}$ as in (12) with $D_k = \text{diag}(\sigma_1^k, \ldots, \sigma_{p_k}^k)$. An estimator $t(\mathcal{X})$ of the form

$$t(\mathcal{X}) = (U_1, \ldots, U_K) \cdot (f^1(D_1), \ldots, f^K(D_K)) \cdot \mathcal{V}, \tag{13}$$

where $f^k(D_k) = \text{diag}(f_1^k(\sigma_1^k), \ldots, f_{p_k}^k(\sigma_{p_k}^k))$, is called a *higher-order spectral estimator*.

Each of the matrix shrinkage functions listed in Section 1 (4)-(10) may, in principle, be applied to each mode in our higher-order spectral estimator (13). We focus on two examples of higher-order spectral estimators. One of these is a generalization of the matrix truncation estimator (4) and the other is a generalization of the matrix soft-thresholding estimator (9). The former can be used to choose the multilinear rank of Θ, the latter is for estimation of Θ when we suspect that the mean tensor has approximately low multilinear rank.

Example: Truncated HOSVD to find the multilinear rank. The first step in many tensor applications is to choose the multilinear rank of the underlying signal, a difficult task [46, 26, 7]. The methods in this paper present a way to choose the multilinear rank. The truncated HOSVD is one popular way to induce low multilinear rank [11]. Given multilinear rank $(r_1, \ldots, r_K)$, it is

found by taking the HOSVD (11) and setting all elements in $\mathcal{S}$ except the "corner" $\mathcal{S}_{[1:r_1,\ldots,1:r_K]}$ to 0. The truncated HOSVD may be viewed as a higher-order spectral estimator (13), where

$$f_i^k(\sigma_i^k) = \sigma_i^k 1(i \leq r_k). \tag{14}$$

This sets to 0 all but r_k of the mode-specific singular values, resulting in an estimate of Θ that has multilinear rank $(r_1, \ldots, r_K)$. The set of all possible multilinear ranks defines a class of reduced rank estimators of Θ. In this paper, we suggest adaptively selecting an estimator from this class by minimizing an unbiased estimate of the risk.

Example: Mode-specific soft-thresholding. Shrinking all of the singular values can generally improve estimation over just truncating the smallest few singular values. A popular form of shrinkage that accomplishes this, a result of nuclear-norm regularization, is the soft-thresholding estimator (9). The second estimator we explore is obtained by applying soft-thresholding to the mode-specific singular values:

$$f_i^k(\sigma_i^k) = (\sigma_i^k - \lambda_k)_+. \tag{15}$$

As with the previous example, the set of $(\lambda_1, \ldots, \lambda_K)$ defines a class of estimators. We propose adaptively selecting a member of this class by minimizing an unbiased estimate of the risk.

A few words are in order about the mode-specific soft-thresholding estimator in (15). First, we note that the resulting core array,

$$(f^1(D_1)D_1^{-1}, \ldots, f^K(D_K)D_K^{-1}) \cdot \mathcal{S},$$

is not generally all-orthogonal. Hence, the $f^k(D_k)$ are not actually the new mode-specific singular values of the estimator $t(\mathcal{X})$. That is, it would be incorrect to think that subtracting off λ_1 from the first-mode singular values means that the new first-mode singular values are $\sigma_{i_1}^1 - \lambda_1$. We are altering the mode-specific singular values, but the relationship is complex. Rather, the proper intuition for shrinkage functions of the form (15) is that the larger the value of λ_k, the more dispersed the resulting mode-specific singular values tend to be on a normalized scale. Likewise, the more negative the value of λ_k to the singular values the less dispersed the resulting mode-specific singular values tend to be. To gain intuition regarding this phenomenon, we provide an extreme case. We generated a $10 \times 10 \times 10$ tensor where each mode had approximately the same singular values. The first-mode specific singular values were $(9.5, 8.7, 8.4, 8.0, 7.5, 7.0, 6.8, 6.0, 5.2, 4.7)$. We applied the mode specific soft-thresholding function (15) to each mode with $\lambda_1 = 5$, $\lambda_2 = 0$, $\lambda_3 = -100$. We then calculated the mode-specific singular values of the resulting tensor and compared these to the original mode-specific singular values, scaled to sum to one. The comparisons can be found in Figure 2. The changed (and normalized) singular values are more dispersed for the first mode, remain relatively unchanged for the second, and are less dispersed for the third.

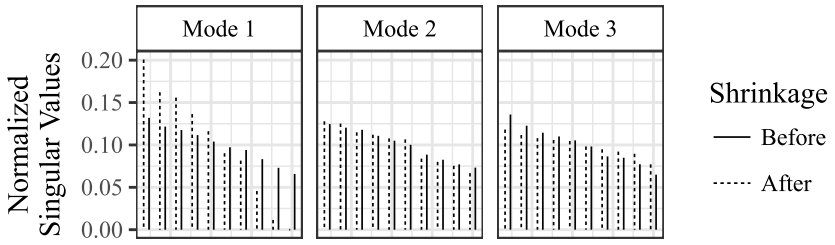

FIG 2. *Singular values for the three modes, before and after shrinkage, normalized to sum to one.*

We have found that we can improve performance (with respect to MSE) by adding an overall scale tuning parameter. That is, we consider a shrinkage estimator of the form:

$$t(\mathcal{X}) = c\ (U_1, \ldots, U_K) \cdot (f^1(D_1)D_1^{-1}, \ldots, f^K(D_K)D_K^{-1}) \cdot \mathcal{S}, \tag{16}$$

where $c > 0$ is the overall scale parameter, $f^k(D_k) = \text{diag}(f_1^k(\sigma_1^k), \ldots, f_{p_k}^k(\sigma_{p_k}^k))$, and $f_i^k(\cdot)$ is from (15).

3. Stein's unbiased risk estimate

Both shrinkage function (14) and (16) define classes of estimators, indexed by tuning parameters. Ideally, we would like to choose these tuning parameters by minimizing the risk (2). However, because the mean Θ is unknown, minimization of (2) with respect to the tuning parameters is not possible. One approach for selecting an estimator from one of these classes is to minimize a risk estimate that does not depend on the unknown parameter. One such estimate is Stein's unbiased risk estimate:

Theorem 1 ([44]). *Under the model (1), suppose* $t : \mathbb{R}^{p_1 \times \cdots \times p_K} \to \mathbb{R}^{p_1 \times \cdots \times p_K}$ *is an almost differentiable function for which*

$$E_\Theta \left[\sum_{\mathbf{i}} \left| \frac{d}{d\mathcal{X}_{[\mathbf{i}]}} t_{\mathbf{i}}(\mathcal{X}_{[\mathbf{i}]}) \right| \right] < \infty. \tag{17}$$

Then

$$\text{MSE}(t(\mathcal{X})) = E_\Theta[\|\Theta - t(\mathcal{X})\|^2] = E_\Theta\left[\|t(\mathcal{X}) - \mathcal{X}\|^2 + 2\tau^2 \,\text{div}(t(\mathcal{X})) - p\tau^2\right],$$

where $\text{div}(\cdot)$ *is the divergence of* $t(\cdot)$, τ^2 *is the variance of each* $\mathcal{X}_{[\mathbf{i}]}$, *and* $p = \prod_{k=1}^K p_k$. *We denote Stein's unbiased risk estimate (SURE) as*

$$\text{SURE}(t) = \|t(\mathcal{X}) - \mathcal{X}\|^2 + 2\tau^2 \,\text{div}(t(\mathcal{X})) - p\tau^2. \tag{18}$$

"Almost differentiable" basically means differentiable everywhere except on a set of Lebesgue measure zero [44, Definition 1]. Because the SURE (18) does not depend on the parameter values Θ, we can minimize the SURE and use this minimization as a proxy for minimizing the risk. In many cases, adaptive estimators obtained by minimizing SURE over a class of estimators yields improved risk performance, as was observed by [6] in the matrix case.

The difficult part of (18) is calculating the divergence. We will spend the next two subsections performing this task. First, we will calculate the differentials for the elements of the altered HOSVD (12) in Subsection 3.1. Then we will use these differentials to derive the divergence of estimators of the form (13) in Subsection 3.2. This divergence can then be inserted into (18) to obtain the SURE.

3.1. Differentials of the HOSVD

In this subsection, we calculate the differentials for the elements in the altered HOSVD (12). In what follows, we will assume that $\mathcal{X}$ has full multilinear rank. Given that $p_k \leq p/p_k$ for all $k = 1, \ldots, K$, where $p = \prod_{k=1}^{K} p_k$, this rank condition is fulfilled almost surely for data $\mathcal{X}$ that have a p.d.f. that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^{p_1 \times \cdots \times p_K}$ [13, Proposition 7.2].

Theorem 2. *The differentials of D_k, U_k, and $\mathcal{V}$ from (12) are given in equations (19), (21), and (25), respectively.*

An outline of the derivation is as follows: Because each U_k and D_k from the HOSVD is from the SVD of $\mathcal{X}_{(k)} = U_k D_k V_k^T$, the calculation begins by recognizing that the differentials of the U_k's and the D_k's are the same as in the matrix case. The differentials can then be re-written as functions of the terms in the HOSVD. To obtain the differential of $\mathcal{V}$, we write $\mathcal{X} = (U_1, \ldots, U_K) \cdot (D_1, \ldots, D_K) \cdot \mathcal{V}$ and apply the chain rule to each U_k, each D_k, then to $\mathcal{V}$. We then solve for the differential of $\mathcal{V}$, which may be written in terms of the differentials of the U_k's and the D_k's.

Proof of Theorem 2. Denote the differential of a function g at $\mathcal{X}$ with increment Δ as $dg[\Delta]$. Since U_k and D_k are the left singular vectors and the singular values, respectively, of $\mathcal{X}_{(k)}$ for each $k = 1, \ldots, K$, the differentials, $dU_k[\Delta]$ and $dD_k[\Delta]$, are the same as in [6] and have a closed form solution, given by

$$d\sigma_i^k[\Delta] = (U_k^T \Delta_{(k)} U_{-k} \mathcal{S}_{(k)} D_k^{-1})_{[i,i]} \text{ for } i = 1, \ldots, p_k \text{ and } k = 1, \ldots, K, \quad (19)$$

where

$$U_{-k} = U_K \otimes \cdots \otimes U_{k+1} \otimes U_{k-1} \otimes \cdots \otimes U_1.$$

This follows because the SVD of $\mathcal{X}_{(k)}$ is $U_k D_k V_k^T = U_k \mathcal{S}_{(k)} U_{-k}^T$ which implies that $V_k = U_{-k} \mathcal{S}_{(k)}^T D_k^{-1}$. We plug in V_k into equation (4.7) of [6] to get (19).

Let $\Omega_{U_k}[\Delta] = U_k^T dU_k[\Delta]$. Then from (4.8) of [6] we have

$$\Omega_{U_k}[\Delta]_{[i,j]} = \frac{-1(i \neq j)\left[\sigma_j^k (U_k^T \Delta_{(k)} U_{-k} S_{(k)}^T D_k^{-1})_{[i,j]} + \sigma_i^k (U_k^T \Delta_{(k)} U_{-k} S_{(k)}^T D_k^{-1})_{[j,i]}\right]}{((\sigma_i^k)^2 - (\sigma_j^k)^2)}, \tag{20}$$

and so

$$dU_k[\Delta] = U \Omega_{U_k}[\Delta]. \tag{21}$$

We now derive $d\mathcal{V}[\Delta]$. Let $U = (U_1, \ldots, U_K)$ and $D = (D_1, \ldots, D_K)$. Also note that $d\mathcal{X}[\Delta] = \Delta$. Using the chain rule, and following Chapter 8, Section 1, Equations (15) and (16) of [34] for the differential of matrix multiplication and the Kronecker product, we have

$$\begin{aligned} \Delta = d\mathcal{X}[\Delta] &= d(U \cdot D \cdot \mathcal{V})[\Delta] \\ &= \sum_{k=1}^K d\underline{U}_k[\Delta] \cdot D \cdot \mathcal{V} + \sum_{k=1}^K U \cdot d\underline{D}_k[\Delta] \cdot \mathcal{V} + U \cdot D \cdot d\mathcal{V}[\Delta], \end{aligned} \tag{22}$$

where

$$d\underline{U}_k[\Delta] = (U_1, \ldots, U_{k-1}, dU_k[\Delta], U_{k+1}, \ldots, U_K) \text{ and} \tag{23}$$

$$d\underline{D}_k[\Delta] = (D_1, \ldots, D_{k-1}, dD_k[\Delta], D_{k+1}, \ldots, D_K). \tag{24}$$

From (22), we solve for $d\mathcal{V}[\Delta]$ and have

$$d\mathcal{V}[\Delta] = D^{-1} \cdot U^T \cdot \Delta - \sum_{k=1}^K dF_k[\Delta] \cdot \mathcal{V} - \sum_{k=1}^K dG_k[\Delta] \cdot \mathcal{V}, \tag{25}$$

where

$$dF_k[\Delta] = (I_{p_1}, \ldots, I_{p_{k-1}}, D_k^{-1} \Omega_{U_k}[\Delta] D_k, I_{p_{k+1}}, \ldots, I_{p_K}) \text{ and} \tag{26}$$

$$dG_k[\Delta] = (I_{p_1}, \ldots, I_{p_{k-1}}, D_k^{-1} dD_k[\Delta], I_{p_{k+1}}, \ldots, I_{p_K}). \tag{27}$$

□

3.2. Divergence of higher-order spectral estimators

In this section, we show that the divergence of higher-order spectral estimators of the form (13) can be found in the following theorem.

Theorem 3. *The divergence of estimators of the form (13) is*

$$\text{Sum}\left(f(D) \cdot D^{-1} \cdot \mathcal{C} + \sum_{k=1}^K H_k \cdot \mathcal{S}^2\right), \tag{28}$$

where $\mathrm{Sum}(\mathcal{A})$ *is the sum of all elements in the tensor* $\mathcal{A}$, $\mathcal{S}^2 \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ *such that* $(\mathcal{S}^2)_{[\mathbf{i}]} = (\mathcal{S}_{[\mathbf{i}]})^2$,

$$\begin{aligned} H_k = (&f^1(D_1)D_1^{-1}, \ldots, f^{k-1}(D_{k-1})D_{k-1}^{-1}, D_k^{-1}df^k(D_k)D_k^{-1}, \\ &f^{k+1}(D_{k+1}), \ldots, f^K(D_K)), \end{aligned} \tag{29}$$

and $\mathcal{C} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ *such that*

$$\begin{aligned} \mathcal{C}_{[\mathbf{i}]} = 1 + \sum_{k=1}^{K} \sum_{j=1, j \neq i_k}^{p_k} \frac{\mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma^k_{i_k})^2 - (\sigma^k_j)^2} - \\ \mathcal{S}^2_{[\mathbf{i}]} \sum_{k=1}^{K} \left(\frac{1}{(\sigma^k_{i_k})^2} + \sum_{m=1, m \neq i_k}^{p_k} \frac{1}{(\sigma^k_m)^2 - (\sigma^k_{i_k})^2} \right). \end{aligned} \tag{30}$$

Proof. Let

$$\Delta^{i_1,\ldots,i_K} = \Delta^{\mathbf{i}} = U_{1[:,i_1]} \circ \cdots \circ U_{K[:,i_K]},$$

where $\circ$ is the outer product and $U_{k[:,i_k]}$ is the i_kth column of U_k. Note that

$$(U_1^T, \ldots, U_K^T) \cdot \Delta^{\mathbf{i}} = E^{\mathbf{i}},$$

where $E^{\mathbf{i}}$ is the $p_1 \times \cdots \times p_K$ array with a one in position $(i_1, \ldots, i_K)$ and zeros everywhere else. Similar to the arguments of [6], also note that $\Delta^{\mathbf{i}}$ forms an orthonormal basis for $\mathbb{R}^{p_1 \times \cdots \times p_K}$, and so

$$\begin{aligned} \mathrm{div}(t(\mathcal{X})) &= \sum_{\mathbf{i}} \langle \Delta^{\mathbf{i}}, df[\Delta^{\mathbf{i}}] \rangle \\ &= \sum_{\mathbf{i}} \langle (U_1^T, \ldots, U_K^T) \cdot \Delta^{\mathbf{i}}, (U_1^T, \ldots, U_K^T) \cdot df[\Delta^{\mathbf{i}}] \rangle \\ &= \sum_{\mathbf{i}} \langle E^{\mathbf{i}}, (U_1^T, \ldots, U_K^T) \cdot df[\Delta^{\mathbf{i}}] \rangle, \\ &= \sum_{\mathbf{i}} ((U_1^T, \ldots, U_K^T) \cdot df[\Delta^{\mathbf{i}}])_{[\mathbf{i}]}, \end{aligned} \tag{31}$$

where $\langle , \rangle$ is the usual Euclidean inner product. From the chain rule, we have:

$$df[\Delta^{\mathbf{i}}] = \sum_{k=1}^{K} d\underline{U}_k[\Delta^{\mathbf{i}}] \cdot f(D) \cdot \mathcal{V} + \sum_{k=1}^{K} U \cdot df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V} + U \cdot f(D) \cdot d\mathcal{V}[\Delta^{\mathbf{i}}],$$

where

$$\begin{aligned} f(D) &= (f^1(D_1), \ldots, f^K(D_K)) \text{ and} \\ df(\tilde{D})_k[\Delta^{\mathbf{i}}] &= (f^1(D_1), \ldots, f^{k-1}(D_{k-1}), d(f^k \circ D_k)[\Delta^{\mathbf{i}}], \\ &\qquad f^{k+1}(D_{k+1}), \ldots, f^K(D_K)), \end{aligned}$$

where "$\circ$" now means composition. Hence,

$$\begin{aligned} &U^T \cdot df[\Delta^{\mathbf{i}}] \\ &= \sum_{k=1}^{K} d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(D) \cdot \mathcal{V} + \sum_{k=1}^{K} df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V} + f(D) \cdot d\mathcal{V}[\Delta^{\mathbf{i}}], \end{aligned} \tag{32}$$

where

$$d\tilde{U}_k[\Delta^{\mathbf{i}}] = (I_{p_1}, \ldots, I_{p_{k-1}}, \Omega_{U_k}[\Delta^{\mathbf{i}}], I_{p_{k+1}}, \ldots, I_{p_K}). \tag{33}$$

The outline of the derivation of the divergence is as follows. The ultimate goal is to obtain the $(i_1, \ldots, i_K)$th element of $U^T \cdot df[\Delta^{\mathbf{i}}]$ in (32) and plug that into (31). We will first calculate all of the differentials that are in (32), then we will determine the $(i_1, \ldots, i_K)$th element of $U^T \cdot df[\Delta^{\mathbf{i}}]$. Then we will simplify (31). These latter two steps may be found in Appendix A.

We begin with the differentials. From (19), we have

$$\begin{aligned} d\sigma_j^k[\Delta^{\mathbf{i}}] &= (U_k^T \Delta_{(k)}^{\mathbf{i}} U_{-k} S_{(k)}^T D_k^{-1})_{[j,j]} \\ &= (E_{(k)}^{\mathbf{i}} S_{(k)}^T D_k^{-1})_{[j,j]} \\ &= 1(j = i_k) S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} / \sigma_j^k. \end{aligned} \tag{34}$$

This is since $E_{(k)}^{\mathbf{i}} S_{(k)}^T \in \mathbb{R}^{p_k \times p_k}$ such that

$$\left(E_{(k)}^{\mathbf{i}} S_{(k)}^T\right)_{[\ell,j]} = \begin{cases} 0 & \text{if } \ell \neq i_k \\ S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} & \text{if } \ell = i_k. \end{cases} \tag{35}$$

Similarly, from (20), we have

$$\begin{aligned} &\Omega_{U_k}[\Delta^{\mathbf{i}}]_{[\ell,j]} \\ &= \frac{-1(\ell \neq j)\left[\sigma_j^k (U_k^T \Delta_{(k)} U_{-k} S_{(k)}^T D_k^{-1})_{[\ell,j]} + \sigma_\ell^k (U_k^T \Delta_{(k)} U_{-k} S_{(k)}^T D_k^{-1})_{[j,\ell]}\right]}{(\sigma_\ell^k)^2 - (\sigma_j^k)^2} \\ &= \frac{-1(\ell \neq j)\left[\sigma_j^k (E_{(k)}^{\mathbf{i}} S_{(k)}^T D_k^{-1})_{[\ell,j]} + \sigma_\ell^k (E_{(k)}^{\mathbf{i}} S_{(k)}^T D_k^{-1})_{[j,\ell]}\right]}{(\sigma_\ell^k)^2 - (\sigma_j^k)^2} \\ &= \left[S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} 1(\ell = i_k) + S_{[i_1,\ldots,i_{k-1},\ell,i_{k+1},\ldots,i_K]} 1(j = i_k)\right] \\ &\times \frac{-1(\ell \neq j)}{(\sigma_\ell^k)^2 - (\sigma_j^k)^2}. \end{aligned} \tag{36}$$

Also, from the chain rule, we have that

$$\begin{aligned} d(f_j^k \circ \sigma_j^k)[\Delta^{\mathbf{i}}] &= \left(\frac{d}{d\sigma_j^k} f_j^k(\sigma_j^k)\right) d\sigma_j^k[\Delta^{\mathbf{i}}] \\ &= \delta_{j,i_k} \left(\frac{d}{d\sigma_j^k} f_j^k(\sigma_j^k)\right) S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} / \sigma_j^k. \end{aligned} \tag{37}$$

We have just completed all of the calculus necessary to obtain the divergence, and the remainder of the calculation is simplification. That is, we can use equations (25), (31), (32), (34), (36), and (37) to calculate a closed-form expression for the divergence. This simplification is relegated to Appendix A. □

We now present the formula for the SURE for all higher-order spectral estimators of the form (13):

Theorem 4 (SURE for (13))**.** *Under the model (1), suppose $t(\cdot)$ in (13) is almost differentiable and for which (17) holds. Then*

$$\begin{aligned}&\mathrm{SURE}(t)\\&= \|t(\mathcal{X}) - \mathcal{X}\|^2 + 2\tau^2 \,\mathrm{Sum}\left(f(D)\cdot D^{-1}\cdot \mathcal{C} + \sum_{k=1}^{K} H_k \cdot \mathcal{S}^2\right) - p\tau^2.\end{aligned} \tag{38}$$

For higher-order spectral estimators, the "almost differentiability" condition is satisfied if each function $f^k(D_k)$ is almost differentiable. To see this, note that the singular vectors (the U_k's) and the singular values (the D_k's) are almost differentiable [34, Chapter 8, Section 8]. Since $\mathcal{V} = (D_1^{-1}, \ldots, D_K^{-1}) \cdot \mathcal{S}$, and $\mathcal{S}$ is itself just derived from the right singular vectors of each matricization $\mathcal{X}_{(k)}$ multiplied with $\mathcal{X}$ [11], this implies that $\mathcal{V}$ is also almost differentiable. So we consider the higher-order spectral estimators

$$t(\mathcal{X}) = (U_1(\mathcal{X}), \ldots, U_K(\mathcal{X})) \cdot (f^1(D_1(\mathcal{X})), \ldots, f^K(D_K(\mathcal{X}))) \cdot \mathcal{V}(\mathcal{X}), \tag{39}$$

where we emphasize in (39) that the U_k's, D_k's, and $\mathcal{V}$ are all functions of $\mathcal{X}$. It is now clear that the higher-order spectral estimators (39) are almost differentiable if the f_k's are almost differentiable (since the compositions and products of differentiable functions are differentiable). The soft-thresholding (15) and truncation (14) functions are both almost differentiable.

This SURE formula is applicable for all shrinkage functions of the form (13) where $f^k(D_k) = \mathrm{diag}(f_1^k(\sigma_1^k), \ldots, f_{p_k}^k(\sigma_{p_k}^k))$. For such shrinkage functions, the shrinkage being applied to each singular value is a function only of that singular value. However, it is possible to construct estimators which use all of the mode k singular values to shrink each mode k singular value, e.g. if we were to use a shrinkage function analogous to those of (5) or (7). For such estimators, we prove in Appendix C that the form of the divergence is very similar as in (28). The only difference is that one replaces $\frac{d}{d\sigma_{i_k}^k} f_{i_k}^k(\sigma_{i_k}^k)$ with $\frac{d}{d\sigma_{i_k}^k} f_{i_k}^k(\sigma_1^k, \ldots, \sigma_{p_k}^k)$. That is, for such shrinkage functions, $df^k(D_k)$ is a diagonal matrix containing only the diagonal of the Jacobian matrix of the transformation $\mathrm{diag}(D_k) \mapsto \mathrm{diag}(f(D_k))$.

Recall the overall scale tuning parameter we included in (16). Given the SURE of an estimator $t(\mathcal{X})$, it is trivial to derive the SURE of a new estimator $u(\mathcal{X}) := ct(\mathcal{X})$, where c is some constant. That is, since $\mathrm{div}(u(\mathcal{X})) = c\,\mathrm{div}(t(\mathcal{X}))$, we can merely replace $t(\mathcal{X})$ with $ct(\mathcal{X})$ in (18). We can then optimize the SURE over c as well as any parameters in $t(\cdot)$.

Details of the optimization procedure for the soft-thresholding estimator (16) may be found in Appendix B. In Appendix B, we also briefly discuss the computational complexity of the optimization procedure.

4. Simulation studies

In this section, we consider five competitors to the mode-specific soft-thresholding estimator (with the overall scale tuning parameter) (16) and the truncated HOSVD (14). We will compare these estimators assuming the error variance τ^2 is one. The first competitor is $\mathcal{X}$, which is the maximum likelihood estimator and the uniformly minimum variance unbiased estimator. However, the risk-performance of this estimator is known to be dominated by our second competitor, the James-Stein estimator (5) [44]. This estimator may be derived from an empirical Bayes argument where $\Theta_{[\mathbf{i}]} \sim N(0, \gamma^2)$ [16]. As such, it should perform well when the entries of Θ are centered about 0. For a matrix parameter Θ, [15] developed an empirical Bayes estimator that performs better than the James-Stein estimator when Θ exhibits empirical correlation along the rows. With this in mind, our third estimator is obtained by applying the Efron-Morris estimator (6) to the first mode matricization of the data tensor. However, the Efron-Morris estimator does not induce low rank estimates, and so our fourth competitor is the matrix soft-thresholding estimator (9) applied to the first mode matricization of $\mathcal{X}$, and whose tuning parameter is chosen with the SURE formula from [6]. This estimator should improve on the Efron-Morris estimator when $\Theta_{(1)}$ has approximately low rank. Our final estimator is the least-squares low-multilinear rank approximation to the data tensor [10], where the multilinear rank is chosen by SURE using our truncated HOSVD estimator. We call this estimator the HOOI for the optimization procedure used to compute it (the higher-order orthogonal iteration).

We now describe the design of the simulation study. We evaluated the risk of the mode-specific soft-thresholding, truncated HOSVD, HOOI, maximum likelihood, James-Stein, Efron-Morris, and matrix soft-thresholding estimators under six different values of $\Theta \in \mathbb{R}^{10\times10\times10}$, constructed as follows:

A. $\text{vec}(\Theta) \sim N_p(0, I_{1000})$.
B. $\text{vec}(\Theta) \sim N_p(0, I_{10} \otimes I_{10} \otimes F)$, where $F = \text{diag}(1^2, 2^2, \ldots, 10^2)$.
C. $\text{vec}(\Theta) \sim N_{1000}(0, I_{10} \otimes I_{10} \otimes \Sigma)$ where $\Sigma \in \mathbb{R}^{10\times10}$ has an AR-1 (0.7) covariance structure. That is, $\Sigma_{[i,j]} = 0.7^{|i-j|}$.
D. $\Theta_{(1)} = U_{[:,1:5]} D_{[1:5,1:5]} V^T_{[:,1:5]}$ where UDV^T is the SVD of a 10×100 matrix that has standard normal entries.
E. $\text{vec}(\Theta) \sim N_p(0, F \otimes F \otimes F)$, where $F = \text{diag}(1^2, 2^2, \ldots, 10^2)$.
F. Θ is a rank $(5, 5, 5)$ tensor where all of the non-zero mode-specific singular values are the same along all modes.

For each scenario, we re-scaled Θ to have Frobenius norm $\sqrt{1000}$, so that $E[\|\mathcal{E}\|^2] = 1000 = \|\Theta\|^2$. For each Θ, we simulated $\mathcal{X}_{[\mathbf{i}]} \sim N(\Theta_{[\mathbf{i}]}, 1)$, calculated the seven estimators given this data tensor, and calculated the squared error loss for each estimator. We repeated this process 500 times. Box plots of the losses for each of the six Θ values are given in Figure 3.

The James-Stein estimator (5) is expected to perform well in Scenario **A** as it can be viewed as an empirical Bayes procedure for the prior with which Θ was actually generated. Indeed, from Figure 3 (**A**), the James-Stein estimator

does perform best, but the mode-specific soft-thresholding estimator performs almost as well, even though there is no correlation along any of the modes of the mean tensor.

For scenario **B**, we expect the matrix soft-thresholding estimator (9) to do well. Since the mean tensor in this scenario has approximately low rank only along the first mode, estimators that shrink towards the space of low multilinear rank tensors should be over-fitting and should not perform well. From Figure 3 (**B**), the matrix soft-thresholding estimator does perform best, but surprisingly, the mode-specific soft-thresholding estimator does equally well.

For Scenario **C**, we expect the matrix soft-thresholding estimator (9) and the Efron-Morris estimator (6) to perform well. There is temporal correlation along one of the modes of the mean tensor. We take into account the temporal correlation of the mean by performing soft-thresholding along this mode. However, from Figure 3 (**C**), we see that the mode-specific soft-thresholding estimator performed best.

The matrix soft-thresholding estimator (9) was designed to do well when the mean matrix is of low rank. This is exactly the situation in Scenario **D**, as a tensor with low rank along one mode may be matricized to form a low rank matrix. However, from Figure 3 (**D**), for our one Θ value, the mode-specific soft-thresholding estimator performs best.

As for Scenario **E**, we expect the mode-specific soft-thresholding estimator (16) to do well, as the mean tensor has approximately low multilinear rank, but it is not exactly low multilinear rank. Figure 3 (**E**) reveals the mode-specific soft-thresholding estimator does indeed perform better than the other estimators.

We expect the truncated HOSVD (14) and the HOOI to do well in Scenario **F** because the mean tensor has low multilinear rank, and the truncated HOSVD and HOOI are correctly shrinking toward this structure. From Figure 3 (**F**), we see that the truncated HOSVD and HOOI do indeed perform best in terms of loss. The HOOI performs slightly better (note, though, that for every other scenario the HOOI and truncated HOSVD have comparable performances). The mode-specific soft-thresholding estimator does not perform much worse. The estimators that do not take into account the tensor indexing perform about twice as bad as these tensor-specific estimators.

For scenarios **C** and **D**, we emphasize here that we are looking at the risk only at a few points in the parameter space. There are likely points where the matrix-soft thresholding estimator performs better than the tensor estimators. However our mode-specific soft-thresholding estimator did not perform poorly under any of our simulated mean tensors.

Our procedure for the truncated HOSVD produces a multilinear rank with the smallest SURE. It is of interest to know if this multilinear rank provides a good estimate of the true rank of Θ. We evaluated this possibility in simulation Scenarios **D** and **F**. We also included two matrix-specific rank estimators for comparison: using either a minimum-description length (MDL) criterion [51] or parallel analysis [5]. We also tried to implement the methods from Yokota et al. [52] but found that our implementation of it resulted in poor behavior (at least for these data) and so we omit the results. In Scenario **F**, where the

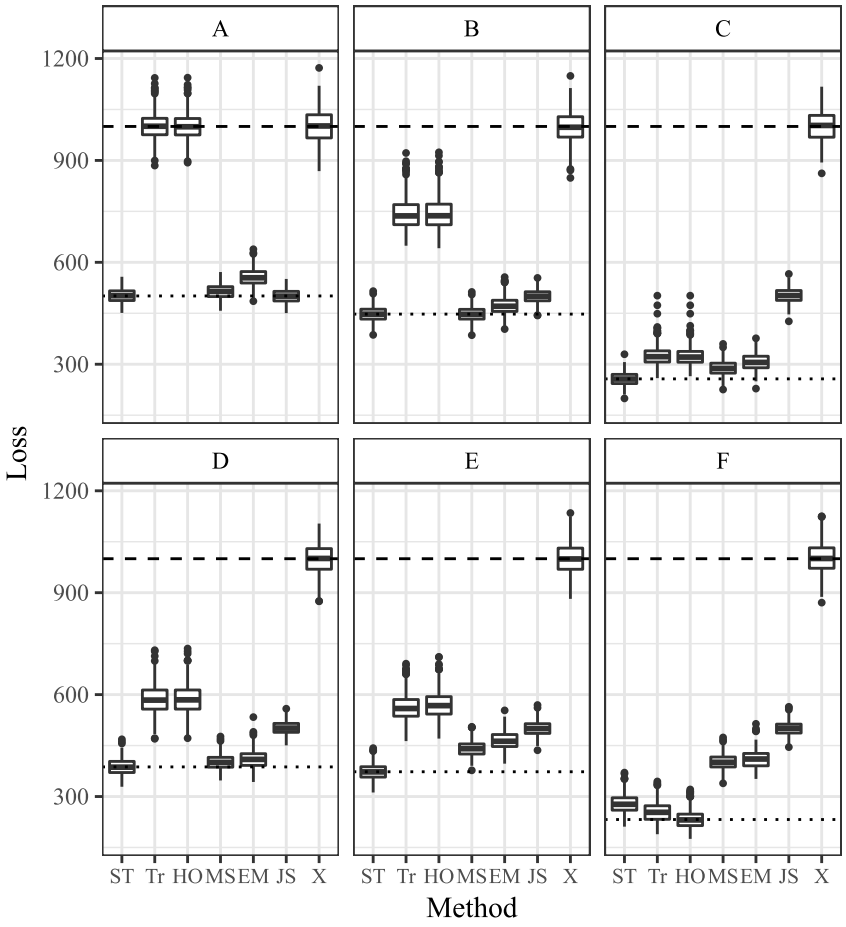

FIG 3. *Box plots of losses for the seven estimators under different scenarios. The estimators include the mode-specific soft-thresholding (ST), truncated HOSVD (Tr), least-squares low-multlinear rank approximation (HO), matrix soft-thresholding (MS), Efron-Morris (EM), James-Stein (JS), and maximum likelihood (X) estimators. In the scenarios, the mean tensor was simulated to have* (***A***) *uncorrelated elements,* (***B***) *full rank but dispersed singular values only along mode 1,* (***C***) *AR-1 covariance along mode 1,* (***D***) *low rank only along mode 1,* (***E***) *full rank but dispersed singular values along all modes, and* (***F***) *rank* $(5, 5, 5)$ *with all the same non-zero singular values.*

tensor had dimension $(10, 10, 10)$ and the true multilinear rank was $(5, 5, 5)$, our SURE method correctly estimated the multilinear rank in 95% of trials, parallel analysis correctly estimated the multilinear rank in 71% of trials, and the MDL method correctly estimated the multilinear rank in 31% of the trials (Table 1). In Scenario **D**, where the true multilinear rank was $(5, 10, 10)$, the results of the simulation study can be found in Figure 4. There, we see that the rank of the first mode is correctly estimated in 96% of trials using our SURE method.

Method	Proportion Correct
MDL	.31
PA	.74
SURE	.95

TABLE 1
Proportion of times the multilinear rank is estimated correctly in Scenario $\boldsymbol{F}$ *using either the minimum description length criterion (MDL), parallel analysis (PA), or Stein's unbiased risk estimate (SURE).*

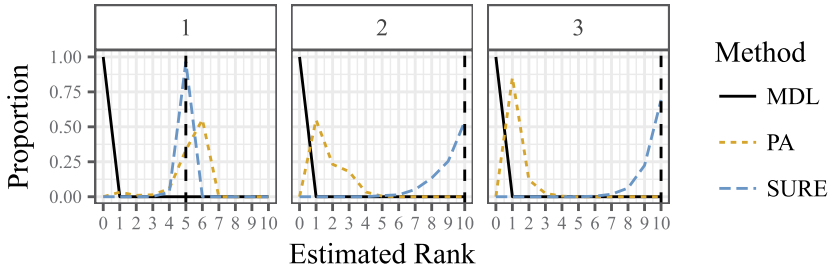

FIG 4. *Proportion of trials in Scenario* $\boldsymbol{D}$ *a rank is estimated when using either the minimum description length criterion (MDL), parallel analysis (PA), or Stein's unbiased risk estimate (SURE). The column facets distinguish between the different modes and the dashed vertical lines indicate the true mode-specific rank.*

The rank of the second and third modes are correctly estimated using SURE a majority of the time. Parallel analysis and MDL have much worse performance, particularly in modes 2 and 3 where the matricization of the mean tensor has full rank.

5. Multivariate relational data example

In this section, we demonstrate the applicability of our estimators to multivariate relational data. Such data may be viewed as a three-way tensor $\mathcal{X}$ where entry $\mathcal{X}_{[i,j,k]}$ is the value of relation type k from node i to node j. One example of such a data set is a social network in which multiple types of relations are measured between individuals. As another example, in sports statistics, round robin interaction data consist of outcomes of competitions between teams. In this section we illustrate our methods with round robin data from the 2014-2015 regular season of the National Basketball Association (NBA). The NBA consists of a Western conference and an Eastern conference of fifteen teams each, where intra-conference play has three to four games per year per pair of teams and inter-conference play is limited to two games a season per pair of teams. For each conference, we created a four dimensional tensor where element $\mathcal{Y}_{[i,j,k,\ell]}$ is statistic k obtained by team i while playing team j either during team i's first home ($\ell = 1$) or first away ($\ell = 2$) game against team j during the season. The statistics we considered were free-throw percentage, two-point field goal percentage, and three-point field goal percentage. We thus have two tensors each

of dimension $15 \times 15 \times 3 \times 2$, one for each of the two conferences. In this section, we illustrate the utility of tensor shrinkage by predicting late season relational basketball statistics from early season data. Our approach is analogous to that of [17], who illustrated the utility of vector shrinkage estimation by predicting late season baseball batting averages from data on early season batting averages.

The statistics in our data set are all empirical proportions. We model the elements of $\mathcal{Y}$ with a binomial model,

$$n_{i,j,k,\ell}\mathcal{Y}_{[i,j,k,\ell]} \sim \text{Bin}(n_{i,j,k,\ell}, p_{i,j,k,\ell}),$$

where all elements are independent, given the $p_{i,j,k,\ell}$'s. We apply an arc-sin transformation to the data tensor to stabilize the variance:

$$\mathcal{X}_{[i,j,k,\ell]} = (n_{i,j,k,\ell})^{1/2} \arcsin(2\mathcal{Y}_{[i,j,k,\ell]} - 1).$$

From the central limit theorem, we have approximately

$$\mathcal{X}_{[i,j,k,\ell]} \sim N(\Theta_{[i,j,k,\ell]}, 1),$$

where $\Theta_{[i,j,k,\ell]} = (n_{i,j,k,\ell})^{1/2} \arcsin(2p_{i,j,k,\ell} - 1)$, resulting in the model in (1).

A commonly used representation of a mean tensor Θ is an ANOVA decomposition, such as

$$\Theta_{[i,j,k,\ell]} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell + \tilde{\Theta}_{[i,j,k,\ell]},$$

where $\tilde{\Theta}_{[i,j,k,\ell]}$ contains all of the interaction effects. Note that $\mathbf{1}_{p_1}^T \alpha = 0$, $\mathbf{1}_{p_2}^T \beta = 0$, $\mathbf{1}_{p_3} \gamma = 0$, and $\mathbf{1}_{p_4}^T \delta = 0$, where $\mathbf{1}_{p_k}$ is the vector of ones of length p_k. The tensor $\tilde{\Theta}$ also satisfies $\tilde{\Theta}_{(k)} \mathbf{1}_{p/p_k} = 0$ for all $k = 1, 2, 3, 4$. Suppose we obtain the maximum likelihood estimates of μ, α, β, γ, and δ by fitting a main-effects ANOVA model. We then calculate the residual tensor,

$$\begin{aligned}\mathcal{R}_{[i,j,k,\ell]} =& \mathcal{X}_{[i,j,k,\ell]} - \frac{p_1}{p} \sum_{j',k',\ell'} \mathcal{X}_{[i,j',k',\ell']} - \frac{p_2}{p} \sum_{i',k',\ell'} \mathcal{X}_{[i',j,k',\ell']} - \\ & \frac{p_3}{p} \sum_{i',j',\ell'} \mathcal{X}_{[i',j',k,\ell']} - \frac{p_4}{p} \sum_{i',j',k'} \mathcal{X}_{[i',j',k',\ell]} + \frac{3}{p} \sum_{i',j',k',\ell'} \mathcal{X}_{[i',j',k',\ell']}.\end{aligned}$$

This residual tensor has an expected value of $\tilde{\Theta}$. It was proposed in [43] and [15] that we estimate the interaction effects $\tilde{\Theta}$ with a vector shrinkage-type estimator on the residuals. If the interactions $\tilde{\Theta}$ are close to zero — when the interaction effects are small — then such estimators will adaptively shrink the residuals towards zero. However, these estimators were developed to adapt to patterns in vectors or matrices of residuals, and not tensors of residuals. In contrast, our approach should be able to adapt to these patterns along any of the four modes of the residual tensor.

We applied mode-specific soft-thresholding and the truncated HOSVD to the array of residuals $\mathcal{R}$ from the main effects ANOVA model. These methods

suggest that the residual tensor should be heavily shrunk both towards zero and towards low multilinear rank structure. For the West, the Frobenius norm of the residual tensor was 38.38, while the Frobenius norm of the resulting shrunken residual tensor using the mode-specific soft-thresholding estimator was 7.81. In the East, the values were 38.95 and 6.97, respectively. We also used SURE to estimate the multilinear rank of each residual tensor using the truncated HOSVD. The estimated multilinear rank of the residual tensor of the Western conference was $2 \times 3 \times 1 \times 2$, and for the Eastern conference the estimated multilinear rank was $4 \times 2 \times 1 \times 1$. These are very small ranks compared to the dimensions of the tensors $15 \times 15 \times 3 \times 2$.

An ad hoc evaluation of the performance of our estimators can be obtained by predicting game statistics after the first home and first away games. Since some teams only play each other three times, we do not have late season data on all possible combinations of team pairs by home versus away games. For the late season data we do have, we present the squared error losses for predicting the statistics of the remaining part of the season for each conference in Table 2. The different estimators are (1) the raw data array $\mathcal{X}$, (2) the mean estimates of the main-effects ANOVA model, (3) the mode-specific soft-thresholding shrunken residual tensor added to the mean estimates of the main-effects ANOVA model, (4) the truncated HOSVD shrunken residual tensor added to the mean estimates of the main-effects ANOVA model, and (5) an estimator derived from logistic regression using the main-effects of each mode. The losses are with respect to the arc-sin transformed data. The poor performance of $\mathcal{X}$ is unsurprising. The amount of shrinkage that our estimators produce indicates that the fully saturated model is over-fitting and that most of the information is contained in the main-effects. However, our mode-specific soft-thresholding estimator is also fitting the fully saturated model and it performs comparable to the main-effects ANOVA model, even improving the predictions for the Eastern conference.

Estimator	East	West
$\mathcal{X}$	2410	2476
ANOVA	1344	1364
Mode-specific Soft-thresholding	1327	1385
Truncated HOSVD	1391	1451
Logistic Regression	1481	1552

TABLE 2

Squared error losses when predicting the statistics of the remaining games of the season.

6. Discussion

This paper introduced new classes of shrinkage estimators for tensor-valued data that are higher-order generalizations of existing matrix spectral estimators. Each class is indexed by tuning parameters whose values we chose by minimizing an unbiased estimate of the risk. In terms of MSE, these estimators outperform their matrix counterparts when the mean has approximately low

multilinear rank and they perform competitively when the mean does not have low multilinear rank.

There has been some recent work on penalized optimization methods for estimating signal tensors in the presence of Gaussian noise [41, 48, 49, 33, 47]. Usually, these estimators are defined as the minimizers of a penalized squared error empirical loss, where the penalty is usually some generalization of the nuclear norm to tensors (for example, the sum of the nuclear norms of the K matricizations of a tensor). These estimators, though similar in spirit, are very different from our approach. The main advantage of our estimators is their simplicity — they are simply functions of the HOSVD (13) for which there are efficient and accurate numerical procedures to compute.

We have presented a way to adaptively choose the tuning parameters of our higher-order spectral estimators by minimizing the SURE. This approach is applicable, not just for the truncated HOSVD (14) and the mode-specific soft-thresholding (16) estimators, but also for *all* estimators of the form (13) that satisfy the conditions of Theorem 1. Although we found that adaptively choosing the tuning parameters by minimizing the SURE worked well under the scenarios we studied, there are other ways to select tuning parameters. In the case of matrix spectral estimators, others have chosen the amount of shrinkage by minimax considerations [15, 44], cross-validation [3, 38, 24], and asymptotic considerations [20, 19]. Exploring these methods for our higher-order spectral estimators (13) is a current research area of the authors.

In this paper, we focused on estimators of the form (13). If the mean tensor is believed to have approximately low multilinear rank, we should shrink the core array through the Tucker product along the modes to obtain this low multilinear rank. The form of our higher-order spectral estimators (13) allows us to use the mode-specific singular values to determine the form and amount of shrinkage that should be performed to each mode of the core array. However, different classes of higher-order spectral estimators can be studied. In Appendix D, we explore functions that shrink each element of the core array individually:

$$t(\mathcal{X}) = (U_1, \ldots, U_K) \cdot g(\mathcal{S}), \text{ where } g(\mathcal{S})_{[\mathbf{i}]} = g_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]}).$$

This class of estimators can be used, for example, to induce zeros in the core array, which has applications in increasing the interpretability of a higher-order generalization of principal components analysis [22, 28, 36, 1, 12, 35].

Although the error variance τ^2 in (1) might be known in some settings, such as fMRI data sets [6], in most applied situations the variance would be unknown. There are matrix-specific estimates of the variance that can be applied to tensor-variate datasets by first matricizing along each mode. In our software, we have implemented the methods described in Choi et al. [8] and Gavish and Donoho [20]. Though, instead of plugging in an estimate of the variance into the SURE formula (18), there has been a recent suggestion to use a generalized SURE formula [39, 25]:

$$\text{GSURE}(t) = \frac{||t(\mathcal{X}) - \mathcal{X}||^2}{(1 - \text{div}(t(\mathcal{X}))/p)^2}.$$

This formula is motivated by generalized cross-validation [21] and is an approximation to SURE [25]. Importantly, GSURE does not require the variance to be known, and so its minimization may be accomplished without an estimate of τ^2. For our higher-order spectral estimators, we have already accomplished the hard work of calculating the divergence in this paper, and implementing GSURE is an easy application of this result. Our software allows for GSURE implementation for the estimators discussed in this article.

There has been recent work in exploring matrix and tensor decomposition methods in regression [55, 32, 54]. There, the authors assume the coefficient matrix/tensor has some low rank structure. That is, they model

$$y_i = \langle \mathcal{B}, \mathcal{X}_i \rangle + \mathcal{E}_i,$$

where y_i is a scalar response for individual i, $\mathcal{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is a tensor of covariates, $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the coefficient tensor, $\mathcal{E}_i \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is some (possibly Gaussian) noise, and $\langle \cdot, \cdot \rangle$ is the usual Euclidean inner product. They then assume that $\mathcal{B}$ has low rank structure. These approaches require either the selection of the rank of the coefficient matrix/tensor or, in the case of penalized regression, the selection of a tuning parameter — non-trivial tasks that might benefit from the use of SURE-like methods. If one had enough samples ($n > \prod p_k$), then one could apply our higher-order spectral estimators to the ordinary least squares (OLS) estimates $\hat{\mathcal{B}}$ to obtain new estimated coefficients. If the design is orthogonal ($\langle \mathcal{X}_i, \mathcal{X}_j \rangle = 0$ for all $i \neq j$), then our methods would be directly applicable. If not, then one could decorrelate the OLS estimates using the design matrix (though this might destroy the tensor structure). However, in most tensor settings we would expect $n \ll \prod p_k$, and so exploring penalization methods might be of worth. In penalized regression settings with vector coefficients, others have detailed how to use SURE for model selection [57] and something similar might be applicable here in the tensor-regression setting. However, as our estimators are based on the HOSVD and were not formulated in a penalization setting, extending the ideas from this paper to penalized coefficient matrices/tensors is not trivial, and is thus a subject of future research.

All methods discussed in this paper are implemented in the R package `hose` available at

https://github.com/dcgerard/hose.

Code and instructions to reproduce all of the results of this paper are available at

https://github.com/dcgerard/reproduce_sure.

Appendix A: Simplification of the divergence

We will need the $(i_1, \ldots, i_K)$th element of $U^T \cdot df[\Delta^{\mathbf{i}}]$ in (32). There are three terms in (32). We will deal with them one by one. First, we will work with the

first term of (32), $\sum_{k=1}^{K} d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(D) \cdot \mathcal{V}$. Note that, for $\mathcal{A} = f(D) \cdot \mathcal{V}$, we have

$$\begin{aligned}
\left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot \mathcal{A}\right)_{[\mathbf{i}]} &= \left((I_{p_1}, \ldots, I_{p_{k-1}}, \Omega_{U_k}[\Delta^{\mathbf{i}}], I_{p_{k+1}}, \ldots, I_{p_K}) \cdot \mathcal{A}\right)_{[\mathbf{i}]} \\
&= -\sum_{j=1, j\neq i_k}^{p_k} \mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \mathcal{A}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} / [(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2] \\
&= -\sum_{j=1, j\neq i_k}^{p_k} \left[\frac{f_j^k(\sigma_j^k) \mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2} \right. \\
&\times \left. \left(\prod_{\ell=1, \ell\neq k}^{K} f_{i_\ell}^{\ell}(\sigma_{i_\ell}^{\ell}) \right) \right] \\
&= -\left(\prod_{\ell=1, \ell\neq k}^{K} f_{i_\ell}^{\ell}(\sigma_{i_\ell}^{\ell}) \right) \\
&\times \sum_{j=1, j\neq i_k}^{p_k} \frac{f_j^k(\sigma_j^k) \mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2}.
\end{aligned}$$

Now we work with the second term of (32), $\sum_{k=1}^{K} df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}$. We have that:

$$\begin{aligned}
\left(df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}\right)_{[\mathbf{i}]} &= \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j) \right) d(f_{i_k}^k \circ \sigma_{i_k}^k)[\Delta^{\mathbf{i}}] \mathcal{V}_{[\mathbf{i}]} \qquad (40) \\
&= \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j) \right) \left(\frac{d}{d\sigma_{i_k}^k} f_{i_k}^k(\sigma_{i_k}^k) \right) \mathcal{V}_{[\mathbf{i}]} \mathcal{S}_{[\mathbf{i}]} / \sigma_{i_k}^k \\
&= \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j) / \sigma_{i_j}^j \right) \left(\frac{d}{d\sigma_{i_k}^k} f_{i_k}^k(\sigma_{i_k}^k) \right) \mathcal{S}_{[\mathbf{i}]}^2 / (\sigma_{i_k}^k)^2, \qquad (41)
\end{aligned}$$

since $\mathcal{V}_{[\mathbf{i}]} = \left(\prod_{k=1}^{K} \sigma_{i_k}^k \right)^{-1} \mathcal{S}_{[\mathbf{i}]}$.

It remains to work with the third term in (32), $f(D) \cdot d\mathcal{V}[\Delta^{\mathbf{i}}]$. We have:

$$\left(f(D) \cdot d\mathcal{V}[\Delta^{\mathbf{i}}]\right)_{[\mathbf{i}]} = \left(\prod_{k=1}^{K} f_{i_k}^k(\sigma_{i_k}^k) \right) d\mathcal{V}[\Delta^{\mathbf{i}}]_{[\mathbf{i}]}. \qquad (42)$$

We now need to obtain $d\mathcal{V}[\Delta^{\mathbf{i}}]_{[\mathbf{i}]}$. From (25), we have

$$\begin{aligned}
d\mathcal{V}[\Delta^{\mathbf{i}}] &= D^{-1} \cdot U^T \cdot \Delta^{\mathbf{i}} - \sum_{k=1}^{K} dF_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V} - \sum_{k=1}^{K} dG_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}, \\
&= D^{-1} \cdot E^{\mathbf{i}} - \sum_{k=1}^{K} dF_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V} - \sum_{k=1}^{K} dG_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}. \qquad (43)
\end{aligned}$$

There are three terms in (43). Let us deal with them one by one. The first term in (43) is

$$\left(D^{-1}\cdot E^{\mathbf{i}}\right)_{[\mathbf{i}]}=\left(\prod_{k=1}^{K}\sigma_{i_k}^k\right)^{-1}. \tag{44}$$

The second term in (43) is

$$\begin{aligned}
&\left(dF_k[\Delta^{\mathbf{i}}]\cdot\mathcal{V}\right)_{[\mathbf{i}]}\\
&=\left((I_{p_1},\ldots,I_{p_{k-1}},D_k^{-1}\Omega_{U_k}[\Delta^{\mathbf{i}}]D_k,I_{p_{k+1}},\ldots,I_{p_K})\cdot\mathcal{V}\right)_{[\mathbf{i}]}\\
&=\sum_{j=1}^{p_k}\left(D_k^{-1}\Omega_{U_k}[\Delta^{\mathbf{i}}]D_k\right)_{[i_k,j]}\mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}\\
&=-\sum_{j=1,j\neq i_k}^{p_k}\frac{\sigma_j^k}{\sigma_{i_k}^k}\frac{S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}\mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2-(\sigma_j^k)^2}\\
&=-\sum_{j=1,j\neq i_k}^{p_k}\frac{\sigma_j^k}{\sigma_{i_k}^k}\frac{S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}\mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2-(\sigma_j^k)^2}.
\end{aligned} \tag{45}$$

The third term in (43) is

$$\begin{aligned}
\left(dG_k[\Delta^{\mathbf{i}}]\cdot\mathcal{V}\right)_{[\mathbf{i}]}&=\left(\mathcal{V}\times_k D_k^{-1}dD_k[\Delta^{\mathbf{i}}]\right)_{[\mathbf{i}]}\\
&=d\sigma_{i_k}^k[\Delta]\mathcal{V}_{[\mathbf{i}]}/\sigma_{i_k}^k\\
&=\mathcal{S}_{[\mathbf{i}]}\mathcal{V}_{[\mathbf{i}]}/(\sigma_{i_k}^k)^2.
\end{aligned} \tag{46}$$

To obtain the third term in (32), we need only plug in (44), (45), and (46) into (43). And then we need to plug in (43) into (42).

We will now show that the divergence is of the form:

$$\begin{aligned}
&\sum_{i_1,\ldots,i_K}\left[\mathcal{C}_{[\mathbf{i}]}\prod_{k=1}^{K}f_{i_k}^k(\sigma_{i_k}^k)/\sigma_{i_k}^k\right.\\
&\left.+\sum_{k=1}^{K}\left(\prod_{j\neq k}f_{i_j}^j(\sigma_{i_j}^j)/\sigma_{i_j}^j\right)\left(\frac{d}{d\sigma_{i_k}^k}f_{i_k}^k(\sigma_{i_k}^k)\right)\mathcal{S}^2_{[i_1,\ldots,i_k]}/(\sigma_{i_k}^k)^2\right]\\
&=\mathrm{Sum}\left(f(D)\cdot D^{-1}\cdot\mathcal{C}+\sum_{k=1}^{K}H_k\cdot\mathcal{S}^2\right),
\end{aligned}$$

for H_k in (29) and $\mathcal{C}\in\mathbb{R}^{p_1\times\cdots\times p_K}$ in (30). The term $f(D)\cdot D^{-1}\cdot\mathcal{C}$ is from the first and second parts of (32), whereas the terms $\sum_{k=1}^{K}H_k\cdot\mathcal{S}^2$ are from the second part of (32) and were already derived in (41). Let us find $\mathcal{C}$. Let $\mathbf{f}_{i_1,\ldots,i_k}=\mathbf{f}_{\mathbf{i}}=\prod_{k=1}^{K}f_{i_k}^k(\sigma_{i_k}^k)$. Ignoring the second term in (32), we have that

the sum of the first and third terms in (32) is equal to:

$$\sum_{\mathbf{i}} \Bigg\{ -\sum_{k=1}^{K} \sum_{m=1, m\neq i_k}^{p_k} \Bigg[\mathbf{f}_{i_1,\ldots,i_{k-1},m,i_{k+1},\ldots,i_K} \times \frac{\mathcal{S}_{[i_1,\ldots,i_{k-1},m,i_{k+1},\ldots,i_K]} \mathcal{V}_{[i_1,\ldots,i_{k-1},m,i_{k+1},\ldots,i_K]}}{(\sigma^k_{i_k})^2 - (\sigma^k_m)^2} \Bigg] + \mathbf{f}_{\mathbf{i}} \Bigg[\left(\prod_{k=1}^{K} \sigma^k_{i_k} \right)^{-1} + \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\sigma^k_j}{\sigma^k_{i_k}} \frac{\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma^k_{i_k})^2 - (\sigma^k_j)^2} - \mathcal{S}_{[\mathbf{i}]} \mathcal{V}_{[\mathbf{i}]} \sum_{k=1}^{K} \frac{1}{(\sigma^k_{i_k})^2} \Bigg] \Bigg\}.$$

After rearranging summands, we obtain:

$$\sum_{\mathbf{i}} \mathbf{f}_{\mathbf{i}} \Bigg[\left(\prod_{k=1}^{K} \sigma^k_{i_k} \right)^{-1} + \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\sigma^k_j}{\sigma^k_{i_k}} \frac{\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma^k_{i_k})^2 - (\sigma^k_j)^2} - \mathcal{S}_{[\mathbf{i}]} \mathcal{V}_{[\mathbf{i}]} \sum_{k=1}^{K} \left(\frac{1}{(\sigma^k_{i_k})^2} + \sum_{m=1, m\neq i_k}^{p_k} \frac{1}{(\sigma^k_m)^2 - (\sigma^k_{i_k})^2} \right) \Bigg].$$

And after factoring out $\prod_{k=1}^{K} (\sigma^k_{i_k})^{-1}$, we get:

$$\sum_{\mathbf{i}} \mathbf{f}_{\mathbf{i}} \left(\prod_{k=1}^{K} \sigma^k_{i_k} \right)^{-1} \Bigg[1 + \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma^k_{i_k})^2 - (\sigma^k_j)^2} - \mathcal{S}^2_{[\mathbf{i}]} \sum_{k=1}^{K} \left(\frac{1}{(\sigma^k_{i_k})^2} + \sum_{m=1, m\neq i_k}^{p_k} \frac{1}{(\sigma^k_m)^2 - (\sigma^k_{i_k})^2} \right) \Bigg].$$

That is,

$$\mathcal{C}_{[\mathbf{i}]} = 1 + \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma^k_{i_k})^2 - (\sigma^k_j)^2} - \mathcal{S}^2_{[\mathbf{i}]} \sum_{k=1}^{K} \left(\frac{1}{(\sigma^k_{i_k})^2} + \sum_{m=1, m\neq i_k}^{p_k} \frac{1}{(\sigma^k_m)^2 - (\sigma^k_{i_k})^2} \right). \tag{47}$$

Appendix B: Details of optimization

We now provide some brief details on our optimization strategy when considering only the mode-specific soft-thresholding estimator. Let $f_{\mathbf{i}} = \prod_{k=1}^{K} f_{i_k}^k(\sigma_{i_k}^k)$ and $\tilde{\sigma}_{\mathbf{i}} = \prod_{k=1}^{K} \sigma_{i_k}^k$. The SURE is equal to:

$$\begin{aligned} &\|f(D)\cdot D^{-1}\cdot\mathcal{S}-\mathcal{S}\|^2 \\ &+\ 2\tau^2\sum_{\mathbf{i}}\left[\left(f(D)\cdot D^{-1}\cdot\mathcal{C}\right)_{[\mathbf{i}]}+\sum_{k=1}^{K}\left(H_k\cdot\mathcal{S}^2\right)_{[\mathbf{i}]}\right]-p\tau^2 \qquad (48)\\ &=\sum_{\mathbf{i}}\Bigg[\left(f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{S}_{[\mathbf{i}]}-\mathcal{S}_{[\mathbf{i}]}\right)^2+2\tau^2 f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{C}_{[\mathbf{i}]} \\ &+\ 2\tau^2 f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{S}_{[\mathbf{i}]}^2\sum_{k=1}^{K}\frac{\frac{d}{d\sigma_{i_k}^k}f_{i_k}^k(\sigma_{i_k}^k)}{\sigma_{i_k}^k f_{i_k}^k(\sigma_{i_k}^k)}\Bigg]-p\tau^2. \qquad (49)\end{aligned}$$

To update each λ_k, we simply apply a general purpose univariate optimizer (e.g. Brent's method [2]). To update c, we have

$$\begin{aligned} &\frac{d}{dc}\Bigg[c^2 f_{\mathbf{i}}^2\tilde{\sigma}_{\mathbf{i}}^{-2}\mathcal{S}_{[\mathbf{i}]}^2-2cf_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{S}_{[\mathbf{i}]}^2+2\tau^2 cf_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{C}_{[\mathbf{i}]} \\ &+\ 2\tau^2 cf_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{S}_{[\mathbf{i}]}^2\sum_{k=1}^{K}\frac{1}{\sigma_{i_k}^k f_{i_k}^k(\sigma_{i_k}^k)}\Bigg] \\ &=2cf_{\mathbf{i}}^2\tilde{\sigma}_{\mathbf{i}}^{-2}\mathcal{S}_{[\mathbf{i}]}^2-2f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{S}_{[\mathbf{i}]}^2+2\tau^2 f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{C}_{[\mathbf{i}]}+2\tau^2 f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{S}_{[\mathbf{i}]}^2\sum_{k=1}^{K}\frac{1}{\sigma_{i_k}^k f_{i_k}^k(\sigma_{i_k}^k)}.\end{aligned}$$

Let

$$\begin{aligned} a&=\sum_{\mathbf{i}}f_{\mathbf{i}}^2\tilde{\sigma}_{\mathbf{i}}^{-2}\mathcal{S}_{[\mathbf{i}]}^2,\\ b&=\sum_{\mathbf{i}}f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{S}_{[\mathbf{i}]}^2,\\ d&=\sum_{\mathbf{i}}\tau^2 f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{C}_{[\mathbf{i}]},\ \text{and}\\ e&=\sum_{\mathbf{i}}\tau^2 f_{\mathbf{i}}\tilde{\sigma}_{\mathbf{i}}^{-1}\mathcal{S}_{[\mathbf{i}]}^2\sum_{k=1}^{K}\frac{1}{\sigma_{i_k}^k f_{i_k}^k(\sigma_{i_k}^k)},\end{aligned}$$

where we are summing over the set of i_k's such that $\sigma_{i_k}^k > \lambda_k$ for $k = 1, \ldots, K$. Then the minimum c occurs at $(b-d-e)/a$. This is a global minimizer, conditional on the λ_k's, since $a > 0$.

Our optimization procedure then merely iterates between updating c and each λ_k. Though a non-convex optimization problem, we have found that by

starting each λ_k at 0 and starting c at 1 (the case of no shrinkage) works well enough in practice that we generally only run the optimization procedure once.

We now briefly discuss the computational complexity of this optimization procedure. Obtaining the HOSVD has compuational complexity

$$\mathcal{O}\left(\max(p_1,\ldots,p_K)\prod_{k=1}^{K} p_k\right),$$

since this is the computational complexity when iterating through the K SVD's we need to calculate (assuming that $p_i \leq \prod_{k:k\neq i} p_k$ for all i). Thus, this is also the computational complexity of calculating higher-order spectral estimators. Obtaining the $\mathcal{C}$ array (30) given the HOSVD also has computational complexity $\mathcal{O}\left(\max(p_1,\ldots,p_K)\sum_{k=1}^{K} p_k\right)$ since each $\mathcal{C}_{[\mathbf{i}]}$ has on the order of $\sum_k p_k$ computations and there are $\prod_k p_k$ elements in $\mathcal{C}$. However, the computational complexity of calculating the SURE conditioned on having the HOSVD and the $\mathcal{C}$ array (30) is merely linear in the number of elements in the data tensor, $\mathcal{O}(\prod_{k=1}^{K} p_k)$ (e.g. see (49)). To see this, note that in (38) $f(D)$, D, and H_k are all diagonal matrices. Also note that

$$\|t(\mathcal{X}) - \mathcal{X}\| = \|(f^1(D_1)D_1^{-1},\ldots,f^K(D_K)D_K^{-1})\cdot\mathcal{S} - \mathcal{S}\|. \tag{50}$$

During the optimization of the SURE, we do not need to recalculate the HOSVD and $\mathcal{C}$.

Appendix C: General spectral functions

In Section 3.1, we assumed that the spectral functions were of the form:

$$f^k(D_k) = \operatorname{diag}(f_1^k(\sigma_1^k),\ldots,f_{p_k}^k(\sigma_{p_k}^k)).$$

That is, we only used σ_i^k when determining the amount of shrinkage to perform on σ_i^k. In this section, we will extend these results to weakly differentiable functions of the form:

$$f^k : \mathcal{D}_{p_k}^+ \to \mathcal{D}_{p_k}^+,$$

where $\mathcal{D}_{p_k}^+$ is the space of p_k by p_k diagonal matrices with non-negative diagonal elements. This will allow us to use $\sigma_1^k,\ldots,\sigma_{p_k}^k$ to determine the amount of shrinkage to perform on σ_i^k. These types of spectral functions might be desirable if, for example, we wished to develop a generalization of estimator (7). Let $\mathbf{s}_k = (\sigma_1^k,\ldots,\sigma_{p_k}^k)^T$ be the vector of the kth mode specific singular values. We look at functions

$$g^k : \mathbb{R}^{p_k+} \to \mathbb{R}^{p_k+},$$

where $\mathbb{R}^{p_k+}$ is the space of p_k vectors with non-negative elements. Then

$$f^k(D_k) = \operatorname{diag}(g^k(\mathbf{s}_k))$$

The derivation of the SURE is the same as in Section 3.1 except for the second term in (32):

$$\sum_{k=1}^{K} df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}.$$

We have:

$$\begin{aligned}\left(df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}\right)_{[\mathbf{i}]} &= \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j)\right) d(f^k \circ D_k)[\Delta^{\mathbf{i}}]_{[i_k,i_k]}\mathcal{V}_{[\mathbf{i}]} \\ &= \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j)\right) d(g^k \circ \mathbf{s}_k)[\Delta^{\mathbf{i}}]_{[i_k]}\mathcal{V}_{[\mathbf{i}]} \qquad (51)\end{aligned}$$

By the chain rule:

$$d(g^k \circ \mathbf{s}_k)[\Delta^{\mathbf{i}}] = J_{g^k}(\mathbf{s}_k) d\mathbf{s}_k[\Delta],$$

where $J_{g^k}(\mathbf{s}_k)$ is the Jacobian matrix of g_k evaluated at $\mathbf{s}_k$. We know from (37) that

$$d\mathbf{s}_k[\Delta^{\mathbf{i}}]_{[j]} = 1(j = i_k) S_{[\mathbf{i}]}/\sigma_j^k \text{ for } j = 1, \ldots, p_k.$$

So $d\mathbf{s}_k[\Delta^{\mathbf{i}}]$ contains zeros except in the i_kth position. Hence

$$(J_{g^k}(\mathbf{s}_k) d\mathbf{s}_k[\Delta])_{[j]} = J_{g^k}(\mathbf{s}_k)_{[j,i_k]} S_{[\mathbf{i}]}/\sigma_{i_k}^k \text{ for } j = 1, \ldots, p_k$$

And so

$$\begin{aligned} d(g^k \circ \mathbf{s}_k)[\Delta^{\mathbf{i}}]_{[i_k]} &= (J_{g^k}(\mathbf{s}_k) d\mathbf{s}_k[\Delta])_{[i_k]} \\ &= J_{g^k}(\mathbf{s}_k)_{[i_k,i_k]} S_{[\mathbf{i}]}/\sigma_{i_k}^k. \qquad (52)\end{aligned}$$

Inserting (52) into (51), we get:

$$\left(df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}\right)_{[\mathbf{i}]} = \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j)\right) J_{g^k}(\mathbf{s}_k)_{[i_k,i_k]} S_{[\mathbf{i}]}/\sigma_{i_k}^k \mathcal{V}_{[\mathbf{i}]}.$$

That is, we only need the (i_k, i_k)th element of the Jacobian matrix of the spectral function. Let

$$J^k(D_k) = \text{diag}(J_{g^k}(\mathbf{s}_k)_{[1,1]}, \ldots, J_{g^k}(\mathbf{s}_k)_{[p_k,p_k]}) \text{ for } k = 1, \ldots, K.$$

Then

$$\sum_{k=1}^{K} df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V} = \sum_{k=1}^{K} Q_k \cdot \mathcal{S}^2$$

where

$$Q_k = (f^1(D_1)D_1^{-1}, \ldots, f^{k-1}(D_{k-1})D_{k-1}^{-1}, J_k(D_k)D_k^{-2}, \\ f^{k+1}(D_{k+1})D_{k+1}^{-1}, \ldots, f^K(D_K)D_K^{-1}).$$

The divergence is now of the form:

$$\text{Sum}\left(f(D)\cdot D^{-1}\cdot\mathcal{C} + \sum_{k=1}^{K} Q_k\cdot\mathcal{S}^2\right).$$

Appendix D: SURE for estimators that shrink elements in $\mathcal{S}$

Consider the HOSVD (11). In this section, we will find the SURE for estimators of the form:

$$t(\mathcal{X}) = U\cdot g(\mathcal{S}), \tag{53}$$

where

$$(g(\mathcal{S}))_{[\mathbf{i}]} = g_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]}).$$

That is, we shrink each element of $\mathcal{S}$ separately. An example of such a function is to soft-threshold each element of $\mathcal{S}$:

$$g_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]}) = \text{sign}(\mathcal{S}_{[\mathbf{i}]})(|\mathcal{S}_{[\mathbf{i}]}| - \lambda)_+,$$

where $\text{sign}(x)$ is -1 of $x<0$, 1 if $x>0$, and 0 if $x=0$. Such a function induces 0's in the core array, which has applications to increasing interpretability of higher-order PCA [22, 28, 36, 1, 12, 35]. Inducing 0's in the core array is usually performed by applying orthogonal rotations along each mode. Our approach provides an alternative mechanism to induce 0's in the core array.

Theorem 5. *The differentials of U_k and $\mathcal{S}$ are given in equations (21) and (54), respectively.*

Proof. We have already calculated $dU_k[\Delta]$ in Theorem 2. To obtain $d\mathcal{S}[\Delta]$, we apply the chain rule to the HOSVD (11) and solve for $d\mathcal{S}[\Delta]$.

$$\Delta = d\mathcal{X}[\Delta] = d(U\cdot\mathcal{S})[\Delta] = \sum_{k=1}^{K} d\underline{U}_k[\Delta]\cdot\mathcal{S} + U\cdot d\mathcal{S}[\Delta],$$

where $d\underline{U}_k[\Delta]$ is defined in (23). Hence,

$$d\mathcal{S}[\Delta] = U^T\cdot\Delta - \sum_{k=1}^{K} d\tilde{U}_k[\Delta]\cdot\mathcal{S} \tag{54}$$

where $d\tilde{U}_k[\Delta]$ is defined in (33). □

The derivation of the divergence for functions of the form (53) is very similar to that in Section 3.2. The divergence may still be found from (31). From the chain rule, we have:

$$dt[\Delta^{\mathbf{i}}] = \sum_{k=1}^{K} d\underline{U}_k[\Delta^{\mathbf{i}}] \cdot g(\mathcal{S}) + U \cdot d(g \circ \mathcal{S})[\Delta^{\mathbf{i}}],$$

where this "$\circ$" means composition and $d\underline{U}_k[\Delta^{\mathbf{i}}]$ is from (23). Hence,

$$U^T \cdot dt[\Delta^{\mathbf{i}}] = \sum_{k=1}^{K} d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot g(\mathcal{S}) + d(g \circ \mathcal{S})[\Delta^{\mathbf{i}}], \tag{55}$$

where $d\tilde{U}_k[\Delta^{\mathbf{i}}]$ is from (33), noting that the relationship in (36) still holds.

From the chain rule we have:

$$d(f_{[\mathbf{i}]} \circ \mathcal{S}_{[\mathbf{i}]})[\Delta^{\mathbf{i}}]_{[\mathbf{i}]} = \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) d\mathcal{S}_{[\mathbf{i}]}[\Delta^{\mathbf{i}}].$$

We need the $(i_1, \ldots, i_K)$th element of

$$\begin{aligned}
&\left(U^T \cdot df[\Delta^{\mathbf{i}}]\right)_{[\mathbf{i}]} \\
&= \left(\sum_{k=1}^{K} d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S}) + d(f \circ \mathcal{S})[\Delta^{\mathbf{i}}]\right)_{[\mathbf{i}]} \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) d\mathcal{S}_{[\mathbf{i}]}[\Delta^{\mathbf{i}}] \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) d\mathcal{S}[\Delta^{\mathbf{i}}]_{[\mathbf{i}]} \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} \\
&+ \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(\left(U^T \cdot \Delta^{\mathbf{i}}\right)_{[\mathbf{i}]} - \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot \mathcal{S}\right)_{[\mathbf{i}]}\right) \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} \\
&+ \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(E^{\mathbf{i}}_{[\mathbf{i}]} - \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot \mathcal{S}\right)_{[\mathbf{i}]}\right) \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(1 - \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot \mathcal{S}\right)_{[\mathbf{i}]}\right).
\end{aligned} \tag{56}$$

Note that for any $\mathcal{A} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$

$$\begin{aligned}\left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot \mathcal{A}\right)_{[\mathbf{i}]} &= \left((I_{p_1}, \ldots, I_{p_{k-1}}, d\Omega_{U_k}[\Delta^{\mathbf{i}}], I_{p_{k+1}}, \ldots, I_{p_K}) \cdot \mathcal{A}\right)_{[\mathbf{i}]} \\ &= -\sum_{j=1, j\neq i_k}^{p_k} \mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \mathcal{A}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} / [(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2].\end{aligned}$$

Hence, from (56) we have,

$$\begin{aligned}&\operatorname{div}(g) \\ &= \sum_{\mathbf{i}} \left[-\sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} f(\mathcal{S})_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2} \right. \\ &\quad \left. + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(1 + \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2}\right) \right] \\ &= \sum_{\mathbf{i}} \left[-\sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2} \right. \\ &\quad \times f_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}(\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}) \\ &\quad + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \\ &\quad \left. \times \left(1 + \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} / [(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2]\right) \right].\end{aligned} \tag{57}$$

We can rearrange the summations in the left part of (57) by switching the order of the j and the i_k and then altering the notation of the dummy variables to obtain:

$$\begin{aligned}\operatorname{div}(g) = \sum_{\mathbf{i}} \left[\mathcal{S}_{[\mathbf{i}]} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]}) \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} 1/[(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2] \right. \\ \left. + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(1 + \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2}\right) \right].\end{aligned}$$

Hence, the SURE for these higher-order spectral functions (53) is:

$$\begin{aligned}\operatorname{SURE}(g(\mathcal{X})) = -p\tau^2 + \|f(\mathcal{S}) - \mathcal{S}\|^2 \\ + 2\tau^2 \sum_{\mathbf{i}} \left[\mathcal{S}_{[\mathbf{i}]} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]}) \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} 1/[(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2] \right. \\ \left. + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(1 + \sum_{k=1}^{K} \sum_{j=1, j\neq i_k}^{p_k} \frac{\mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2}\right) \right].\end{aligned}$$

Acknowledgments

Peter Hoff's research was partially supported by NSF grant DMS-1505136.

References

[1] Claus A Andersson and Rene Henrion. A general algorithm for obtaining simple structure of core arrays in n-way PCA with application to fluorometric data. *Computational statistics & data analysis*, 31(3):255–278, 1999. https://doi.org/10.1016/S0167-9473(99)00017-1.

[2] RP Brent. An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, 14(4):422–425, 1971. https://doi.org/10.1093/comjnl/14.4.422. MR0339475

[3] R Bro, Karin Kjeldahl, AK Smilde, and HAL Kiers. Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry*, 390 (5):1241–1251, 2008. ISSN 1618-2650. https://doi.org/10.1007/s00216-007-1790-1.

[4] Rasmus Bro. Review on multiway analysis in chemistry - 2000–2005. *Critical reviews in analytical chemistry*, 36 (3-4):279–293, 2006. https://doi.org/10.1080/10408340600969965.

[5] Andreas Buja and Nermin Eyuboglu. Remarks on parallel analysis. *Multivariate behavioral research*, 27(4): 509–540, 1992. https://doi.org/10.1207/s15327906mbr2704_2.

[6] Emmanuel J. Candès, Carlos A. Sing-Long, and Joshua D. Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans. Signal Process.*, 61(19): 4643–4657, 2013. ISSN 1053-587X. https://doi.org/10.1109/TSP.2013.2270464. MR3105401

[7] Eva Ceulemans and Henk AL Kiers. Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology*, 59(1):133–150, 2006. https://doi.org/10.1348/000711005X64817. MR2246998

[8] Yunjin Choi, Jonathan Taylor, and Robert Tibshirani. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *arXiv preprint arXiv:1410.8260*, 2014.

[9] Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, and Huy Anh Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2): 145–163, 2015. https://doi.org/10.1109/MSP.2013.2297439.

[10] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-$(R_1, R_2, \cdots, R_N)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4): 1324–1342 (electronic), 2000. ISSN 0895-4798. https://doi.org/10.1137/S0895479898346995. MR1780276

[11] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4): 1253–

1278 (electronic), 2000. ISSN 0895-4798. https://doi.org/10.1137/S0895479896305696. MR1780272

[12] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. Independent component analysis and (simultaneous) third-order tensor diagonalization. *Signal Processing, IEEE Transactions on*, 49 (10):2262–2271, 2001. https://doi.org/10.1109/78.950782.

[13] Vin de Silva and Lek-Heng Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.*, 30(3): 1084–1127, 2008. ISSN 0895-4798. https://doi.org/10.1137/06066518X. MR2447444

[14] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. ISSN 0033-3123. https://doi.org/10.1007/BF02288367.

[15] Bradley Efron and Carl Morris. Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika*, 59(2):335–347, 1972. ISSN 0006-3444. https://doi.org/10.1093/biomet/59.2.335. MR0334386

[16] Bradley Efron and Carl Morris. Limiting the risk of Bayes and empirical Bayes estimators — Part II: The empirical Bayes case. *J. Amer. Statist. Assoc.*, 67:130–139, 1972. ISSN 0162-1459. https://doi.org/10.1080/01621459.1972.10481215. MR0323015

[17] Bradley Efron and Carl Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70 (350):311–319, 1975. https://doi.org/10.1080/01621459.1975.10479864. MR0391403

[18] Bradley Efron and Carl Morris. Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.*, 4(1):22–32, 1976. ISSN 0090-5364. https://doi.org/doi:10.1214/aos/1176343345. MR0394960

[19] M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63 (4):2137–2152, April 2017. ISSN 0018-9448. https://doi.org/10.1109/TIT.2017.2653801. MR3626861

[20] Matan Gavish and David Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60 (8):5040–5053, 2014. ISSN 0018-9448. https://doi.org/10.1109/TIT.2014.2323359. MR3245370

[21] Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979. ISSN 0040-1706. https://doi.org/10.2307/1268518. MR0533250

[22] René Henrion. Body diagonalization of core matrices in three-way principal components analysis: Theoretical bounds and simulation. *Journal of Chemometrics*, 7(6):477–494, 1993. https://doi.org/10.1002/cem.1180070604.

[23] Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics*, 9(3): 1169–1193, 2015. https://doi.org/doi:10.1214/15-AOAS839. MR3418719

[24] Julie Josse and François Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Comput. Statist. Data Anal.*, 56(6): 1869–1879, 2012. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2011.11.012. MR2892383

[25] Julie Josse and Sylvain Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, 2015. https://doi.org/10.1007/s11222-015-9554-9. MR3489867

[26] Henk AL Kiers and Albert Kinderen. A fast method for choosing the numbers of components in Tucker3 analysis. *British Journal of Mathematical and Statistical Psychology*, 56(1):119–125, 2003. https://doi.org/10.1348/000711003321645386. MR2101798

[27] Henk AL Kiers and Iven Van Mechelen. Three-way component analysis: Principles and illustrative application. *Psychological methods*, 6(1):84–110, 2001. https://doi.org/10.1037/1082-989X.6.1.84.

[28] Henk AL Kiers, Jos MF Ten Berge, and Roberto Rocci. Uniqueness of three-mode factor models with sparse cores: The 3× 3× 3 case. *Psychometrika*, 62(3):349–374, 1997. ISSN 0033-3123. https://doi.org/10.1007/BF02294556. MR1475181

[29] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009. ISSN 0036-1445. https://doi.org/10.1137/07070111X. MR2535056

[30] Pieter M. Kroonenberg. *Applied multiway data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2008. ISBN 978-0-470-16497-6. https://doi.org/10.1002/9780470238004. With a foreword by Willem J. Heiser and Jarqueline Meulman. MR2378349

[31] Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 0 (0):1–16, 2017. https://doi.org/10.1080/01621459.2016.1193022.

[32] Xiaoshan Li, Hua Zhou, and Lexin Li. Tucker tensor regression and neuroimaging analysis. *arXiv preprint arXiv:1304.5637*, 2013.

[33] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):208–220, 2013. https://doi.org/10.1109/TPAMI.2012.39.

[34] Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1999. ISBN 0-471-98633-X. Revised reprint of the 1988 original. MR1698873

[35] Carla D Moravitz Martin and Charles F Van Loan. A Jacobi-type method for computing orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30 (3):1219–1232, 2008. https://doi.org/10.1137/060655924. MR2447449

[36] Takashi Murakami, Jos MF Ten Berge, and Henk AL Kiers. A case of extreme simplicity of the core matrix in three-mode principal components

analysis. *Psychometrika*, 63(3):255–261, 1998. ISSN 0033-3123. https://doi.org/10.1007/BF02294854.

[37] Raj Rao Nadakuditi. Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60 (5):3002–3018, 2014. ISSN 0018-9448. https://doi.org/10.1109/TIT.2014.2311661. MR3200641

[38] Art B. Owen and Patrick O. Perry. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.*, 3(2):564–594, 2009. ISSN 1932-6157. https://doi.org/10.1214/08-AOAS227. MR2750673

[39] Sylvain Sardy. Smooth blockwise iterative thresholding: a smooth fixed point estimator based on the likelihood's block gradient. *J. Amer. Statist. Assoc.*, 107(498): 800–813, 2012. ISSN 0162-1459. https://doi.org/10.1080/01621459.2012.664527. MR2980086

[40] Andrey A. Shabalin and Andrew B. Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivariate Anal.*, 118:67–76, 2013. ISSN 0047-259X. https://doi.org/10.1016/j.jmva.2013.03.005. MR3054091

[41] Marco Signoretto, Lieven De Lathauwer, and Johan AK Suykens. Convex multilinear estimation and operatorial representations. In *NIPS2010 Workshop: Tensors, Kernels and Machine Learning (TKML)*, 2010.

[42] Age Smilde, Rasmus Bro, and Paul Geladi. *Multi-way analysis: applications in the chemical sciences*. John Wiley & Sons, 2005.

[43] Charles Stein. An approach to the recovery of interblock information in balanced incomplete block designs. *Research paper in statistics: Festschrift for J. Neyman*, pages 351–366, 1966. MR0210232

[44] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981. ISSN 0090-5364. https://doi.org/doi:10.1214/aos/1176345632. MR0630098

[45] Dacheng Tao, Xuelong Li, Weiming Hu, Stephen Maybank, and Xindong Wu. Supervised tensor learning. In *Fifth IEEE International Conference on Data Mining*, pages 450–457. IEEE, 2005. https://doi.org/10.1109/ICDM.2005.139.

[46] Marieke E Timmerman and Henk AL Kiers. Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima. *British Journal of Mathematical and Statistical Psychology*, 53(1):1–16, 2000. https://doi.org/10.1348/000711000159132.

[47] Ryota Tomioka and Taiji Suzuki. Convex tensor decomposition via structured Schatten norm regularization. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1331–1339. Curran Associates, Inc., 2013.

[48] Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. Estimation of low-rank tensors via convex optimization. *arXiv:1010.0789*, 2011. URL http://arxiv.org/abs/1010.0789.

[49] Ryota Tomioka, Taiji Suzuki, Kohei Hayashi, and Hisashi Kashima. Statistical performance of convex tensor decomposition. In J. Shawe-Taylor, R.S.

Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 972–980. Curran Associates, Inc., 2011.

[50] Marie Verbanck, Julie Josse, and François Husson. Regularised PCA to denoise and visualise data. *Statistics and Computing*, 25(2):471–468, 2015. ISSN 0960-3174. https://doi.org/10.1007/s11222-013-9444-y. MR3306719

[51] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):387–392, Apr 1985. ISSN 0096-3518. https://doi.org/10.1109/TASSP.1985.1164557. MR0788604

[52] T. Yokota, N. Lee, and A. Cichocki. Robust multilinear tensor rank estimation using higher order singular value decomposition and information criteria. *IEEE Transactions on Signal Processing*, 65 (5):1196–1206, March 2017. ISSN 1053-587X. https://doi.org/10.1109/TSP.2016.2620965. MR3584316

[53] Xiang Zhang, Lexin Li, Hua Zhou, Dinggang Shen, et al. Tensor generalized estimating equations for longitudinal imaging analysis. *arXiv preprint arXiv:1412.6592*, 2014. URL http://arxiv.org/abs/1412.6592.

[54] Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483, 2014. ISSN 1467-9868. https://doi.org/10.1111/rssb.12031. MR3164874

[55] Hua Zhou, Lexin Li, and Hongtu Zhu. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108 (502):540–552, 2013. https://doi.org/10.1080/01621459.2013.776499. MR3174640

[56] Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476): 1418–1429, 2006. ISSN 0162-1459. https://doi.org/10.1198/016214506000000735. MR2279469

[57] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192, 10 2007. https://doi.org/10.1214/009053607000000127. MR2363967