Theory and Methods for Tensor Data

David Gerard

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Peter Hoff, Chair

Mathias Drton

Michael Perlman

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Theory and Methods for Tensor Data

David Gerard

Chair of the Supervisory Committee:
Professor Peter Hoff
Department of Statistics and Department of Biostatistics - Public Health

We present novel methods and new theory in the statistical analysis of tensor-valued data. A tensor is a multidimensional array. When data come in the form of a tensor, special methods and models are required to capture the dependencies represented by the indexing structure. For such data, it is often reasonable to assume a Kronecker structured covariance model for the random elements within a tensor. A natural type of Kronecker structured covariance model is the array normal model. We develop equivariant and minimax estimators under the array normal model whose risk performances are dramatically better than that of the maximum likelihood estimator. Although we find improved estimators, maximum likelihood estimation is still popular and useful (e.g. for likelihood ratio testing). We study in detail maximum likelihood estimation in separable covariance models, linking it to the relatively modern study of tensor decompositions. This leads us to develop, within this class of Kronecker structured covariance models, likelihood ratio test statistics which are simply represented as the ratio of two scale parameters from two separate tensor decompositions.

We then focus our attention on mean estimation for tensor-valued data. We develop new classes of shrinkage estimators that alter the mode-specific singular values from a tensor generalization of the singular value decomposition. These classes often contain tuning parameters, whose selection is difficult. We choose these tuning parameters by minimizing an unbiased estimate of the mean squared error. From simulations, these new estimators outper-

form matrix-specific estimators when the tensor indexing structure meaningfully represents the heterogeneity of the underlying signal tensor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation to those who guided and funded him as he developed his skills as a researcher. In particular, Laura Kubatko, H. Lisle Gibbs, Dennis Pearl, Joe Verducci, Mathias Drton, Michael Perlman, and his advisor, Peter Hoff.

Chapter 1

INTRODUCTION

This thesis concerns covariance and mean estimation in tensor-variate data. In this introductory chapter, we will begin by defining important operations over tensors that will be used throughout this thesis. We will then briefly describe the Tucker decomposition. We will follow this by reviewing the array normal model, a statistical model useful in describing the tensor-specific patterns in the data. We will finish this chapter with an outline of the thesis.

1.1 Tensors

We define a tensor as a multidimensional array. To be more formal, in the same way that a vector is an element of a vector space, a tensor is an element of a tensor product of vector spaces. In the same way that up to a choice of basis on a real vector space a vector may be represented as a tuple of real numbers, up to a choice of bases on a set of real vector spaces a tensor may be represented as a multidimensional array of real numbers. For this thesis, we will not be concerned with this more formal definition.

We will let $\mathbb{R}^{p_1 \times \cdots \times p_K}$ denote the real vector space of K-order tensors with dimensions $(p_1, \ldots, p_K)$. A tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ contains elements $\mathcal{X}_{[i_1,\ldots,i_K]} \in \mathbb{R}$ for $i_k = 1, \ldots, p_k$ and $k = 1, \ldots, K$. A 1-way tensor ($K = 1$) is a vector and a 2-way tensor ($K = 2$) is a matrix.

Many data sets come in the form of a tensor (beyond $K = 1$ or 2). A multivariate longitudinal network data set is a tensor where an element is the value of relation type k from node i to node j at time t [Hoff, 2011]. A movie can be represented as a tensor where an element of the tensor is the intensity of pixel (i, j) at frame t. The mean estimates of an ANOVA model may be represented as a tensor where an element of the tensor is the mean

value at factor 1 level i, factor 2 level j, and factor 3 level k [Volfovsky and Hoff, 2014] – for example, we might be interested in concentrations of chemical i at location j at time t. There are many other fields where tensors naturally arise [Kroonenberg, 2008, Kolda and Bader, 2009].

In order to analyze tensor data sets, we need tools to manipulate tensors. The first operation we consider is k-mode matricization, or k-mode matrix unfolding. This operation converts a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ into a matrix $\mathcal{X}_{(k)} \in \mathbb{R}^{p_k \times p/p_k}$ where $p = \prod_{k=1}^{K} p_k$. The rows in the resulting matrix $\mathcal{X}_{(k)}$ index the kth mode and the columns index all other modes. The formal definition, due to Kolda and Bader [2009], is below:

Definition 1. *The k-mode matricization of $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$, denoted $\mathcal{X}_{(k)} \in \mathbb{R}^{p_k \times p/p_k}$, maps element $(i_1, \ldots, i_K)$ in $\mathcal{X}$ to element (i_k, j) in $\mathcal{X}_{(k)}$ where*

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

Similarly, we may vectorize a tensor into a vector.

Definition 2. *The vectorization of $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$, denoted $\text{vec}(\mathcal{X}) \in \mathbb{R}^p$, maps element $(i_1, \ldots, i_K)$ in $\mathcal{X}$ to element j in $\text{vec}(\mathcal{X})$ where*

$$j = 1 + \sum_{k=1}^{K} (i_k - 1) J_k \text{ with } J_k = \prod_{m=1}^{k-1} p_m$$

As an example of the matricization and vectorization operators, let

$$\mathcal{X} = \left(\begin{array}{cc|cc} \mathcal{X}_{[1,1,1]} & \mathcal{X}_{[1,2,1]} & \mathcal{X}_{[1,1,2]} & \mathcal{X}_{[1,2,2]} \\ \mathcal{X}_{[2,1,1]} & \mathcal{X}_{[2,2,1]} & \mathcal{X}_{[2,1,2]} & \mathcal{X}_{[2,2,2]} \end{array} \right) \in \mathbb{R}^{2 \times 2 \times 2},$$

where the vertical line | denotes the separation of the third indices. We provide the three possible matricizations:

$$\mathcal{X}_{(1)} = \left(\begin{array}{cccc} \mathcal{X}_{[1,1,1]} & \mathcal{X}_{[1,2,1]} & \mathcal{X}_{[1,1,2]} & \mathcal{X}_{[1,2,2]} \\ \mathcal{X}_{[2,1,1]} & \mathcal{X}_{[2,2,1]} & \mathcal{X}_{[2,1,2]} & \mathcal{X}_{[2,2,2]} \end{array} \right),$$

$$\mathcal{X}_{(2)} = \begin{pmatrix} \mathcal{X}_{[1,1,1]} & \mathcal{X}_{[2,1,1]} & \mathcal{X}_{[1,1,2]} & \mathcal{X}_{[2,1,2]} \\ \mathcal{X}_{[1,2,1]} & \mathcal{X}_{[2,2,1]} & \mathcal{X}_{[1,2,2]} & \mathcal{X}_{[2,2,2]} \end{pmatrix}, \text{ and}$$
$$\mathcal{X}_{(3)} = \begin{pmatrix} \mathcal{X}_{[1,1,1]} & \mathcal{X}_{[2,1,1]} & \mathcal{X}_{[1,2,1]} & \mathcal{X}_{[2,2,1]} \\ \mathcal{X}_{[1,1,2]} & \mathcal{X}_{[2,1,2]} & \mathcal{X}_{[1,2,2]} & \mathcal{X}_{[2,2,2]} \end{pmatrix}.$$

We also have the resulting vectorization:

$$\text{vec}(\mathcal{X}) = (\mathcal{X}_{[1,1,1]}, \mathcal{X}_{[2,1,1]}, \mathcal{X}_{[1,2,1]}, \mathcal{X}_{[2,2,1]}, \mathcal{X}_{[1,1,2]}, \mathcal{X}_{[2,1,2]}, \mathcal{X}_{[1,2,2]}, \mathcal{X}_{[2,2,2]})^T.$$

We will make heavy use the matricization and vectorization operators throughout this thesis.

Recall matrix multiplication: For $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, we have $X = AB \in \mathbb{R}^{m \times p}$ if

$$X_{[i,j]} = \sum_{k=1}^{n} A_{[i,k]} B_{[k,j]}.$$

There are a few types of multiplication between tensors [Bader and Kolda, 2004, Kolda, 2006, Kilmer and Martin, 2011]. For this thesis, we will almost exclusively consider multilinear multiplication, or the Tucker product, between a tensor $\mathcal{A} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ and a list of matrices $B_k \in \mathbb{R}^{q_k \times p_k}$ for $k = 1, \ldots, K$. We have $\mathcal{X} = (B_1, \ldots, B_K) \cdot \mathcal{A} \in \mathbb{R}^{q_1 \times \cdots \times q_K}$ if

$$\mathcal{X}_{[j_1,\ldots,j_K]} = \sum_{i_1,\ldots,i_K=1}^{p_1,\ldots,p_K} \mathcal{A}_{[i_1,\ldots,i_K]} B_{1[j_1,i_1]} \cdots B_{K[j_K,i_K]}. \tag{1.1}$$

The Tucker product has important properties with regard to the matricization and vectorization operators:

$$\mathcal{X} = (B_1, \ldots, B_K) \cdot \mathcal{A} \text{ iff} \tag{1.2}$$
$$\mathcal{X}_{(k)} = B_k \mathcal{A} (B_K^T \otimes \cdots \otimes B_{k+1}^T \otimes B_{k-1}^T \otimes \cdots \otimes B_1^T) = B_k \mathcal{A}_{(k)} B_{-k}^T \text{ iff} \tag{1.3}$$
$$\text{vec}(\mathcal{X}) = (B_K \otimes \cdots \otimes B_1) \, \text{vec}(\mathcal{A}), \tag{1.4}$$

where B^T is the matrix transpose of B and "$\otimes$" denotes the Kronecker product. The Kronecker product between two matrices $A \in \mathbb{R}^{\ell \times m}$ and $B \in \mathbb{R}^{n \times p}$ is the block matrix

$A \otimes B \in \mathbb{R}^{\ell n \times mp}$ where each block is $A_{[i,j]}B$. That is,

$$A \otimes B = \begin{pmatrix} A_{[1,1]}B & A_{[1,2]}B & \cdots & A_{[1,m]}B \\ A_{[2,1]}B & A_{[2,2]}B & \cdots & A_{[2,m]}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{[\ell,1]}B & A_{[\ell,2]} & \cdots & A_{[\ell,m]}B \end{pmatrix}.$$

The Kronecker product has many useful properties:

$$\begin{aligned} (A \otimes B) \otimes C &= A \otimes (B \otimes C) \\ (A \otimes B)(C \otimes D) &= AC \otimes BD \\ (A \otimes B)^T &= \left(A^T \otimes B^T\right) \\ (A \otimes B)^{-1} &= A^{-1} \otimes B^{-1}, \end{aligned} \tag{1.5}$$

where A^{-1} is the inverse of A. Using (1.4) and (1.5), it is trivial to prove that $(C_1, \ldots, C_K) \cdot [(B_1, \ldots, B_K) \cdot \mathcal{A}] = (C_1B_1, \ldots, C_KB_K) \cdot \mathcal{A}$. Hence, we will usually allow "$\cdot$" to also denote component-wise multiplication between two lists of matrices.

The notion of matrix rank extends to tensors in multiple ways. The version that we consider in this thesis is that of *multilinear rank*. 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. Then the multilinear rank of the K-order tensor $\mathcal{X}$ is the the K-tuple, $(r_1, \ldots, r_K)$, where r_k is the dimension of the vector space spanned by the k-mode vectors. Equivalently, r_k is the rank of the k-mode unfolding of $\mathcal{X}$, $\mathcal{X}_{(k)}$. The notion of multilinear rank will be most extensively used in Chapter 4.

Decomposing a matrix extends to tensors in multiple ways. In the same way that matrix decompositions try to represent patterns in matrices in terms of products of lower dimensional matrices, tensor decompositions seek to find patterns by representing tensors in terms of products of lower dimensional tensors. When a tensor is represented as a Tucker product between a list of matrices and a "core" tensor (1.1), this form of decomposition is called a

"Tucker decomposition". We will not be concerned with the myriad of other tensor decompositions [Kolda and Bader, 2009, Kilmer and Martin, 2011, Cichocki et al., 2014].

The matrix singular value decomposition (SVD) can be viewed as a Tucker decomposition. Recall that $X \in \mathbb{R}^{p\times n}$ with $p \leq n$ may be decomposed as the product of an orthogonal matrix $U \in \mathbb{R}^{p\times p}$, a diagonal matrix $D = \text{diag}(\sigma_1, \ldots, \sigma_p)$ for $\sigma_1 \geq \cdots \geq \sigma_p$, and a $n \times p$ matrix with orthonormal columns V. We write the SVD as

$$X = UDV^T = (U, V) \cdot D. \tag{1.6}$$

Hence, D is the core tensor and U and V are the component matrices. Since $X_{(1)} = X = UDV^T$ and $X_{(2)} = X^T = VDU^T$ the SVD may be constructed by calculating the left singular vectors of the two matricizations of X, followed by deriving the core array from $D = U^T XV = (U^T, V^T) \cdot X$. A popular method, then, of generalizing the SVD to tensors is to compute the SVD of $\mathcal{X}_{(k)} = U_k D_k V_k^T$, set $\mathcal{S} = (U_1^T, \ldots, U_K^T) \cdot \mathcal{X}$, and write:

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

This Tucker decomposition is called the higher-order SVD (HOSVD) [De Lathauwer et al., 2000b] and contains many properties which make it seem a natural generalization of the SVD to tensors. It will be considered briefly in Chapter 2 and used extensively in Chapter 4.

1.2 The array normal model

In this section, we review the array normal model. We do so by building up from the multivariate normal model. Let $X \in \mathbb{R}^{p\times n}$ such that

$$X_{[,1]}, \ldots, X_{[,n]} \overset{i.i.d.}{\sim} N_p\left(\theta, \Psi_1 \Psi_1^T\right).$$

This model may be written as

$$X \overset{d}{=} \theta \mathbf{1}_n^T + \Psi_1 Z, \tag{1.7}$$

where $Z \in \mathbb{R}^{p\times n}$ contains standard normal entries. From elementary operations, we have

$$E[(X - \theta \mathbf{1}_n^T)(X - \theta \mathbf{1}_n^T)^T] \propto \Psi_1 \Psi_1^T.$$

That is, $\Psi_1\Psi_1^T$ represents the "row covariance" of X. One natural extension of this model is to allow Z in (1.7) to be multiplied on the right by another matrix Ψ_2.

$$X \overset{d}{=} \Theta + \Psi_1 Z \Psi_2^T, \tag{1.8}$$

where $Z \in \mathbb{R}^{p\times n}$ contains standard normal entries. This is called the matrix normal model [Srivastava and Khatri, 1979, Dawid, 1981]. Under (1.8), it can be shown that

$$\begin{aligned} E[(X-\Theta)(X-\Theta)^T] &\propto \Psi_1\Psi_1^T \text{ and} \\ E[(X-\Theta)^T(X-\Theta)] &\propto \Psi_2\Psi_2^T. \end{aligned} \tag{1.9}$$

Intuitively, we may consider $\Psi_1\Psi_1^T$ as representing the "row covariance" while $\Psi_2\Psi_2^T$ represents the "column covariance". This model contains $p(p+1)/2 + n(n+1)/2 - 1$ covariance parameters. If we were to have allowed for there to be unrestricted covariance between any element in X and any other element in X, then we would have had $np(np+1)/2$ covariance parameters, which is potentially much larger than the number of covariance parameters in the matrix normal model.

Now consider the tensor case $\mathcal{X} \in \mathbb{R}^{p_1\times\cdots\times p_K}$. A natural extension of the matrix normal model is to define the covariance structure through the Tucker product. This was done in Hoff [2011]:

$$\mathcal{X} \overset{d}{=} \Theta + (\Psi_1, \ldots, \Psi_K)\cdot\mathcal{Z},$$

where $\Theta \in \mathbb{R}^{p_1\times\cdots\times p_K}$ and $\mathcal{Z} \in \mathbb{R}^{p_1\times\cdots\times p_K}$ contains standard normal entries. From (1.3) we have

$$\begin{aligned} \mathcal{X} &\overset{d}{=} \Theta + (\Psi_1, \ldots, \Psi_K)\cdot\mathcal{Z} \\ \mathcal{X}_{(k)} &\overset{d}{=} \Theta_{(k)} + \Psi_k\mathcal{Z}_{(k)}\left(\Psi_K^T\otimes\cdots\otimes\Psi_{k+1}^T\otimes\Psi_{k-1}^T\otimes\cdots\otimes\Psi_1^T\right) \\ &= \Theta_{(k)} + \Psi_k\mathcal{Z}_{(k)}\Psi_{-k}^T. \end{aligned}$$

From which, using (1.9), we can show that

$$E\left[(\mathcal{X}_{(k)}-\Theta_{(k)})(\mathcal{X}_{(k)}-\Theta_{(k)})^T\right] \propto \Psi_k\Psi_k^T.$$

And thus, we may interpret $\Psi_k \Psi_k^T$ as being the covariance among the p_k slices of the array $\mathcal{X}$ along the kth mode.

As well as being a generalization of the multivariate normal model, the array normal model may be viewed as a special case of the multivariate normal model. Using (1.4), we have

$$\begin{aligned}
&\mathcal{X} \stackrel{d}{=} \Theta + (\Psi_1, \ldots, \Psi_k) \cdot \mathcal{Z} \\
&\Leftrightarrow \operatorname{vec}(\mathcal{X}) \stackrel{d}{=} \operatorname{vec}(\Theta) + (\Psi_K \otimes \cdots \otimes \Psi_1) \operatorname{vec}(\mathcal{Z}) \\
&\Leftrightarrow \operatorname{vec}(\mathcal{X}) \sim N_p(\operatorname{vec}(\Theta), \Psi_K \Psi_K^T \otimes \cdots \otimes \Psi_1 \Psi_1^T).
\end{aligned}$$

That is, the array normal model is the multivariate normal model with a Kronecker structured covariance matrix.

To summarize, the array normal model is appealing for tensor-variate data sets because of the intuitive interpretation of the mode-specific covariance parameters and because this model is more parsimonious than an unstructured covariance model. That is, the array normal model contains $\frac{1}{2}\sum_{k=1}^{K} p_k(p_k+1) - K + 1$ covariance parameters against the $\frac{1}{2}\prod_{k=1}^{K} p_k \left(\prod_{k=1}^{K} p_k + 1\right)$ covariance parameters of the multivariate normal model. The array normal model will be discussed in more detail in Chapters 2 and 3.

1.3 Contents of chapters

In Chapter 2, we begin by developing a higher-order generalization of the LQ decomposition. We link this decomposition to its role in likelihood-based estimation and testing for Kronecker structured covariance models. This role is analogous to that of the LQ decomposition in likelihood inference for the multivariate normal model. We then extend the literature on tensor decompositions by showing that this higher-order LQ decomposition can be used to construct an alternative version of the popular higher-order singular value decomposition for tensor-valued data. We then develop a novel generalization of the polar decomposition to tensor-valued data.

In Chapter 3, we obtain optimality results for the array normal model that are analogous to some classical results concerning covariance estimation for the multivariate normal model. We show that under a lower triangular product group, a uniformly minimum risk equivariant estimator (UMREE) can be obtained via a generalized Bayes procedure. Although this UMREE is minimax and dominates the MLE, we show that it can be improved upon via an orthogonally equivariant modification. Numerical comparisons of the risks of these estimators show that the equivariant estimators can have substantially lower risks than the MLE.

In Chapter 4, we study mean estimation for tensor-variate data. We generalize existing matrix shrinkage methods to the estimation of a tensor of parameters from noisy tensor data. Specifically, we develop new classes of estimators that shrink or threshold the mode-specific singular values from the higher-order singular value decomposition of De Lathauwer et al. [2000b]. 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 in the absence of low multilinear rank. We illustrate the use of these methods in an application to multivariate relational data.

We conclude this thesis with a discussion and open problems in Chapter 5. In particular, we discuss the existence for the MLE in the array normal model and we discuss minimax estimates of the mean for tensor-variate data.

Chapter 2

A HIGHER-ORDER LQ DECOMPOSITION FOR SEPARABLE COVARIANCE MODELS

2.1 Introduction

There has been a recent surge of interest in methods for tensor-valued data in the machine learning, applied math, and statistical communities. Tensors, or multiway arrays, are higher-order generalizations of vectors and matrices whose elements are indexed by more than two index sets. Analysis methods for tensor-valued data include tensor decompositions and statistical modeling. The former aims to express the tensor in terms of interpretable lower-dimensional components. The latter uncovers patterns through the lens of statistical inference in a parametric statistical model.

The work in the field of tensor decompositions is extensive (see Kolda and Bader [2009] or Cichocki et al. [2014] for a review). A common class of tensor decompositions are Tucker decompositions [Tucker, 1966], which, for an array $X \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ with entries $X_{[i_1,\ldots,i_K]}$, expresses X as a product of a "core" array $S \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ and matrices $U_1, \ldots, U_K$ where $U_k \in \mathbb{R}^{p_k \times p_k}$, expressed as

$$X = (U_1, \ldots, U_K) \cdot S, \tag{2.1}$$

where "$\cdot$" is multilinear multiplication defined in Section 1.1, and again later in Section 2.2. Most Tucker decompositions impose orthogonality constraints on the U_k's. One resulting tensor decomposition with such orthogonality constraints is the higher-order singular value decomposition (HOSVD) of De Lathauwer et al. [2000b,a], a generalization of the singular value decomposition (SVD). There are other generalizations of the SVD to tensors outside the Tucker decomposition framework [de Silva and Lim, 2008, Grasedyck, 2010, Kilmer and

Martin, 2011]. However, our work will focus on Tucker decompositions of the form (2.1), where the U_k's have a variety of forms other than orthogonality.

A different perspective on tensor-valued data analysis uses statistical modeling, which aims to capture the dependencies between the entries of a tensor through a parametric model. One such model is the multilinear normal model [Hoff, 2011, Ohlson et al., 2013, Manceur and Dutilleul, 2013] — also known as the "array normal model" or "tensor normal model" — which is an extension of the matrix normal model [Srivastava and Khatri, 1979, Dawid, 1981]. A $p_1 \times \cdots \times p_K$ tensor X follows a multilinear normal distribution if $\text{vec}(X)$ is normally distributed with covariance $\Sigma_K \otimes \cdots \otimes \Sigma_1$, where "$\otimes$" is the Kronecker product and "$\text{vec}(\cdot)$" is the vectorization operator. For $\Sigma_k = A_k A_k^T$, $k = 1, \ldots, K$, the multilinear normal model may be written

$$X \stackrel{d}{=} (A_1, \ldots, A_K) \cdot Z, \tag{2.2}$$

where $Z \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ contains independent and identically distributed (i.i.d.) standard normal entries. The multilinear normal model "separates" the covariances along the modes, or dimensions of X. That is, the dependencies along the kth mode are represented by a single covariance matrix, Σ_k. Models where the covariance matrix is Kronecker structured are thus often called "separable covariance models". Most results for the multilinear normal model can be easily generalized to array-variate elliptically contoured models with separable covariance [Akdemir and Gupta, 2011].

In Section 2.2, we derive a novel tensor decomposition, a type of Tucker decomposition, whose components provide the maximum likelihood estimators (MLEs) of the parameters in the mean zero multilinear normal model, and array-variate elliptically contoured models with separable covariance in general. This tensor decomposition is a generalization of the LQ matrix decomposition to multiway arrays, and so we call it the incredible Higher-Order LQ decomposition (incredible HOLQ, or just HOLQ). One can view the LQ decomposition as taking the form

$$X = \ell L Q I_n \in \mathbb{R}^{p \times n},$$

where $\ell > 0$, Q has orthonormal rows, L is a lower triangular matrix with positive diagonal elements and unit determinant, and I_n is the identity matrix. The HOLQ takes the form

$$X = \ell(L_1, \ldots, L_K, I_n) \cdot Q \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n},$$

where $\ell > 0$, each L_k is a lower triangular matrix with positive diagonal elements and unit determinant, and $Q \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$ has certain orthogonality properties which generalize the orthonormal rows property of the LQ decomposition. Section 2.3 shows the close relationship between the HOLQ and likelihood inference in the multilinear normal model: In Section 2.3.1, we show that each L_k matrix in the HOLQ is the Cholesky square root of the MLE for the kth component covariance matrix, Σ_k, in the multilinear normal model (2.2). This relationship is analogous to the correspondence between the LQ decomposition and the MLE in the multivariate normal model.

In the same way that likelihood estimation in the multilinear normal model is connected to the HOLQ, likelihood inference in submodels of the unconstrained multilinear normal model is connected to other decompositions where the component matrices have certain structures. In Section 2.3.2, we consider constraining Σ_k to be diagonal. This has the interpretation of statistical independence along the kth mode and corresponds to constraining L_k to be diagonal in the related tensor decomposition. We also consider constraining the diagonal of the lower triangular Cholesky square root of Σ_k to be the vector of ones, which relates to a covariance model used in time series analysis. We label as "HOLQ juniors" the class of decompositions that correspond to submodels of the unrestricted mean zero multilinear normal model. In Section 2.3.3, we use HOLQ juniors to develop a class of likelihood ratio tests for covariance models in elliptically contoured random arrays with separable covariance.

Other tensor decompositions related to the HOLQ are discussed in Section 2.4. In Section 2.4.1 we use the HOLQ to create a new higher-order analogue to the SVD where each mode has singular values and vectors separated from the core array. Since this SVD is derived from the incredible HOLQ, we call it the incredible SVD (ISVD). The ISVD may be viewed as a core rotation of the HOSVD. In Section 2.4.2 we use a novel minimization formulation

of the polar decomposition to generalize it to tensors.

2.2 The incredible HOLQ

Let $X \in \mathbb{R}^{p\times n}$ be of rank p where $p \leq n$. Recall the LQ decomposition,

$$X = LQ,$$

where $L \in G_p^+$, the set of p by p lower triangular matrices with positive diagonal elements, and $Q^T \in \mathcal{V}_{p,n}$, the Stiefel manifold of n by p matrices with orthonormal columns. It is common to formulate the LQ decomposition as a Gram-Schmidt orthogonalization of the rows of X. We instead consider an alternative formulation of the LQ decomposition as a minimization problem:

Theorem 1. *Let $\mathcal{G}_p^+$ denote the set of p by p lower triangular matrices with positive diagonal elements and unit determinant. Let*

$$L = \underset{\tilde{L}\in\mathcal{G}_p^+}{\arg\min} ||\tilde{L}^{-1}X||, \tag{2.3}$$

where $||\cdot||$ is the Frobenius norm. Set $\ell = ||L^{-1}X||$ and $Q = L^{-1}X/\ell$. Then $X = \ell LQ$ is the LQ decomposition of X.

Proof. By the uniqueness of the LQ decomposition [Eaton, 1983, Proposition 5.2], it suffices to show that Q has orthonormal rows. We have $QQ^T = I_p \Leftrightarrow L^{-1}XX^TL^{-T}/\ell^2 = I_p \Leftrightarrow XX^T = \ell^2 LL^T$. Also note that the solution in (2.3) is equivalent to finding the matrix $\tilde{S}$ that satisfies $\tilde{S} = LL^T = \arg\min_{S\in\mathcal{S}_p^1} \mathrm{tr}(S^{-1}XX^T)$, where $\mathcal{S}_p^1$ is the set of p by p positive definite matrices with unit determinant. If we can show that $\tilde{S} = XX^T/|XX^T|^{1/p}$ then we have shown that Q has orthonormal rows. Using Lagrange multipliers, we must minimize $\mathrm{tr}(S^{-1}XX^T) - \lambda \log|S|$ in $S \in \mathcal{S}_p^+$, the set of p by p positive definite matrices, and $\lambda \in \mathbb{R}$. Equivalently, we could also minimize $\mathrm{tr}(VXX^T) - \lambda \log|V|$, where $V = S^{-1}$. Temporarily ignoring the symmetry of V, taking derivatives [Magnus and Neudecker, 1999, chapter 8] and setting equal to zero we have

$$XX^T - \lambda V^{-1} = 0 \text{ and } |V| = 1$$

$$\Leftrightarrow \lambda V^{-1} = XX^T \text{ and } |V| = 1$$
$$\Leftrightarrow V^{-1} = XX^T/|XX^T|^{1/p} \text{ and } \lambda = |XX^T|^{1/p}$$
$$\Leftrightarrow S = XX^T/|XX^T|^{1/p} \text{ and } \lambda = |XX^T|^{1/p}.$$

Since $\log|V|$ is strictly concave (Theorem 25 of Chapter 11 of Magnus and Neudecker [1999] or Theorem 7.6.7 of Horn and Johnson [2013]), $\text{tr}(VXX^T)$ is linear, and $\lambda = |XX^T|^{1/p} > 0$, we have that $\text{tr}(VXX^T) - \lambda \log|V|$ is a convex function in V. Hence, $S = XX^T/|XX^T|^{1/p}$ is a global minimum (c.f. Theorem 13 of Chapter 7 in Magnus and Neudecker [1999]). Since $XX^T/|XX^T|^{1/p}$ is symmetric and positive definite, it is also a global minimum over the space of symmetric positive definite matrices. □

In (2.3), we are "dividing out" L from the rows of X. In this way, we can consider the formulation of the LQ decomposition in Theorem 1 as finding the $L \in \mathcal{G}_p^+$ that accounts for the greatest amount of heterogeneity in the rows of X. The goal of accounting for the heterogeneity in each mode of a multidimensional array will lead to our generalization of the LQ decomposition to tensors, where $X \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$.

Definition 3. *If*

$$(L_1, \ldots, L_K) = \underset{\tilde{L}_k \in \mathcal{G}_{p_k}^+,\ k=1,\ldots,K}{\arg\min} ||(\tilde{L}_1^{-1}, \ldots, \tilde{L}_K^{-1}, I_n) \cdot X|| \tag{2.4}$$

then

$$X = \ell(L_1, \ldots, L_K, I_n) \cdot Q \tag{2.5}$$

is an incredible HOLQ, where $\ell = ||(L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X||$ *and* $Q = (L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X/\ell$.

Here, $(L_1, \ldots, L_K, I_n) \cdot Q$ denotes *multilinear multiplication* of Q by the list of matrices $(L_1, \ldots, L_K, I_n)$ [de Silva and Lim, 2008], also known as the *Tucker product* [Kofidis and Regalia, 2001, Hoff, 2011]. That is, if $X = (L_1, \ldots, L_K, I_n) \cdot Q$ then

$$X_{[j_1,\ldots,j_K,j_{K+1}]} = \sum_{i_1,\ldots,i_K=1}^{p_1,\ldots,p_K} Q_{[i_1,\ldots,i_K,j_{K+1}]} L_{1[j_1,i_1]} \cdots L_{k[j_K,i_K]}.$$

Multilinear multiplication has the following useful properties: If (2.5) holds, then

$$X_{(k)} = L_k Q_{(k)}(I_n \otimes L_K^T \otimes \cdots \otimes L_{k+1}^T \otimes L_{k-1}^T \otimes \cdots \otimes L_1^T) \text{ and} \tag{2.6}$$

$$\text{vec}(X) = (I_n \otimes L_K \otimes \cdots \otimes L_1)\,\text{vec}(Q), \tag{2.7}$$

where $X_{(k)}$ is the unfolding of the array X into a p_k by $n\prod_{i\neq k}^K p_i$ matrix and $\text{vec}(X)$ is the unfolding of the array X into a $n\prod_{k=1}^K p_k$ dimensional vector [Kolda and Bader, 2009]. We will generally denote $I_n \otimes L_K \otimes \cdots \otimes L_{k+1} \otimes L_{k-1} \otimes \cdots \otimes L_1$ by L_{-k} and denote $\prod_{k=1}^K p_k$ by p.

We note that such a minimizing $(L_1, \ldots, L_K)$ in (2.4) may not exist. This is discussed further in Section 2.5. When such a minimizer does exist, we may use (2.6) and Theorem 1 to develop a block coordinate descent algorithm [Tseng, 2001] to solve the minimization problem (2.4): At iteration i, we fix L_k for $k \neq i$. We then find the minimizer in $L_i \in \mathcal{G}_{p_i}^+$ of

$$||L_i^{-1} X_{(i)} L_{-i}^{-T}||,$$

which, by Theorem 1 is the L matrix in the LQ decomposition of $X_{(i)}L_{-i}^{-T} = \ell LQ$. This algorithm is presented in Algorithm 1. A slight improvement on Algorithm 1 is presented in Algorithm 2 where we also update the core array Q of the HOLQ while updating the component lower triangular matrices. Unlike Algorithm 1, Algorithm 2 does not require the calculation of the inverse of L_k or the extra matrix multiplication of $X_{(k)}L_{-k}^{-T}$ at each step. A proof of the equivalence between Algorithms 1 and 2 can be found in Appendix A.1.1.

There are two things to note about these algorithms. First, at each iteration we are reducing the criterion function $||(L_1^{-1}, \ldots, L_K^{-1}, I_n)\cdot X||$. Second, at each iteration of Algorithm 2, we are orthonormalizing the rows of the core array, Q. Hence, the core array Q of any fixed point of this algorithm, including that of the HOLQ, must have a property which we call *scaled all-orthonormality*:

Definition 4. *A $p_1 \times \cdots \times p_K \times n$ tensor Q is* scaled all-orthonormal *if*

$$Q_{(k)}Q_{(k)}^T = I_{p_k}/p_k \text{ for all } k = 1, \ldots, K. \tag{2.8}$$

Algorithm 1 Block coordinate descent for the HOLQ.

Given $X \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$, initialize:

$L_k \leftarrow L_{k0} \in \mathcal{G}_{p_k}^+$ for $k = 1, \ldots, K$.

$\ell \leftarrow ||(L_{10}^{-1}, \ldots, L_{K0}^{-1}, I_n) \cdot X||$

repeat

  for $k \in \{1, \ldots, K\}$ **do**

    LQ decomposition of $X_{(k)} L_{-k}^{-T} = LZ^T$

    $L_k \leftarrow L/|L|^{1/p_k}$

  end for

until Convergence.

Set $\ell \leftarrow ||(L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X||$

Set $Q \leftarrow (L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X/\ell$

return ℓ, Q, and L_k for $k = 1, \ldots, K$.

Theorem 2. *Let $X = \ell(L_1, \ldots, L_K, I_n) \cdot Q$ be an incredible HOLQ. Then the core array Q is scaled all-orthonormal.*

Proof. This is a direct consequence of the LQ step in Algorithm 2. □

Note that we divide by p_k in (2.8) because of the constraint that $||Q|| = 1$. This scaled all-orthonormality property generalizes the orthonormal rows property in the LQ decomposition.

Of course, we could have instead generalized the RQ decomposition, where for $X \in \mathbb{R}^{p \times n}$ we have $X = RZ$ for $R^T \in \mathcal{G}_{p_k}^+$ and $Z^T \in \mathcal{V}_{p,n}$. For $X \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$, if $X = \ell(L_1, \ldots, L_K, I_n) \cdot Q$ is the HOLQ of X, we then take the RQ decomposition of each component $L_k = R_k Z_k$, and set $r = \ell||(Z_1, \ldots, Z_K, I_n) \cdot Q||$ and $Z = \ell(Z_1, \ldots, Z_K, I_n) \cdot Q/r$, then $X = r(R_1, \ldots, R_K, I_n) \cdot Z$ is a higher-order RQ (HORQ) of X, where Z is scaled all-orthonormal. One could instead have started with a similar minimization formulation of the RQ as we did for the LQ (Theorem 1), then generalize to tensors as we did for the HOLQ (2.5), and one would obtain the same HORQ as the one we derive from the HOLQ.

Algorithm 2 Orthogonalized block coordinate descent for the HOLQ.

Given $X \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$, initialize:

$L_k \leftarrow L_{k0} \in \mathcal{G}^+_{p_k}$ for $k = 1, \ldots, K$.

$\ell \leftarrow ||(L_{10}^{-1}, \ldots, L_{K0}^{-1}, I_n) \cdot X||$

$Q \leftarrow (L_{10}^{-1}, \ldots, L_{K0}^{-1}, I_n) \cdot X / \ell$

repeat

  for $k \in \{1, \ldots, K\}$ **do**

    LQ decomposition of $Q_{(k)} = LZ$

    $Q_{(k)} \leftarrow Z$

    $L_k \leftarrow L_k L$

    Re-scale:

    $\ell \leftarrow \ell |L_k|^{1/p_k} ||Q||$

    $L_k \leftarrow L_k / |L_k|^{1/p_k}$

    $Q \leftarrow Q / ||Q||$

  end for

until Convergence.

return ℓ, Q, and L_k for $k = 1, \ldots, K$.

2.3 The incredible HOLQ for separable covariance inference

2.3.1 Maximum likelihood estimation

The LQ decomposition of a data matrix has a close relationship to maximum likelihood inference under the multivariate normal model. Assume a data matrix $X \in \mathbb{R}^{p \times n}$ was generated from a $N_{p \times n}(0, I_n \otimes \Sigma)$ distribution for some Σ symmetric and positive definite. That is, the columns of X are assumed to be independently distributed $N_p(0, \Sigma)$ random vectors. The MLE of Σ is XX^T/n, and so is proportional to $XX^T = LQQ^TL^T = LL^T$, where $X = LQ$ is the LQ decomposition of X.

This result carries over to the multilinear normal model (2.2) using the HOLQ. Assume

the data array $X \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$ follows a multilinear normal model, $X \sim N_{p_1 \times \cdots \times p_K \times n}(0, \sigma^2 I_n \otimes \Sigma_K \otimes \cdots \otimes \Sigma_1)$. That is,

$$X \stackrel{d}{=} \sigma(\Sigma_1^{1/2}, \ldots, \Sigma_K^{1/2}, I_n) \cdot Z, \tag{2.9}$$

where $Z \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$ has i.i.d. standard normal entries and $\Sigma_k^{1/2}$ is the lower triangular Cholesky square root matrix of Σ_k for $k = 1, \ldots, K$. Here, we use the identifiable parameterization of Gerard and Hoff [2015] where $\Sigma_k \in \mathcal{S}_{p_k}^1$ for $k = 1, \ldots, K$ and $\sigma^2 > 0$. The following theorem shows that the MLE of $(\sigma^2, \Sigma_1, \ldots, \Sigma_K)$ can be recovered from the HOLQ of X.

Theorem 3. *Let $X = \ell(L_1, \ldots, L_K, I_n) \cdot Q$ be the incredible HOLQ of X. Then under the model (2.9)*

1. *The MLE of Σ_k is $\hat{\Sigma}_k = L_k L_k^T$ for $k = 1, \ldots, K$,*
2. *The MLE of σ^2 is $\hat{\sigma}^2 = \ell^2/(np)$,*
3. *The maximized likelihood is equal to*

$$\left(2\pi\hat{\sigma}^2\right)^{-np/2} e^{-np/2} = \left(2\pi\ell^2/(np)\right)^{-np/2} e^{-np/2}.$$

Proof. The log-likelihood is proportional to

$$\frac{-np}{2} \log\left(\sigma^2\right) - \frac{1}{2\sigma^2} ||(\Sigma_1^{-1/2}, \ldots, \Sigma_K^{-1/2}, I_n) \cdot X||^2,$$

where $\Sigma_k^{1/2}$ is the lower triangular Cholesky square root matrix of Σ_k. Holding the Σ_k's fixed, taking a derivative of σ^2 and setting equal to zero, we solve for σ^2 and obtain $\hat{\sigma}^2 = ||(\Sigma_1^{-1/2}, \ldots, \Sigma_K^{-1/2}, I_n) \cdot X||^2/(np)$. A second derivative test confirms this is the global maximizer for any fixed $\Sigma_1, \ldots, \Sigma_K$. The profiled likelihood is then

$$\begin{aligned}
&\left(2\pi\hat{\sigma}^2\right)^{-np/2} \exp\left\{-\frac{1}{2\hat{\sigma}^2} ||(\Sigma_1^{-1/2}, \ldots, \Sigma_K^{-1/2}, I_n) \cdot X||^2\right\} \\
&= \left(2\pi\hat{\sigma}^2\right)^{-np/2} \exp\left\{-\frac{1}{2\hat{\sigma}^2} \hat{\sigma}^2 np\right\} \\
&= \left(2\pi\hat{\sigma}^2\right)^{-np/2} e^{-np/2}.
\end{aligned} \tag{2.10}$$

Thus, to maximize the likelihood, we must minimize $\hat{\sigma}^2 = \frac{1}{np}||(\Sigma_1^{-1/2}, \ldots, \Sigma_K^{-1/2}, I_n) \cdot X||^2$ in $\Sigma_k^{1/2} \in \mathcal{G}_{p_k}^+$ for $k = 1, \ldots, K$. This is the same as the minimization problem solved by the HOLQ in (2.4). Hence, the MLE of Σ_k is $\hat{\Sigma}_k = L_k L_k^T$. This in turn implies that $\hat{\sigma}^2 = ||(\hat{\Sigma}_1^{-1/2}, \ldots, \hat{\Sigma}_K^{-1/2}, I_n) \cdot X||^2/(np) = ||(L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X||^2/(np) = \ell^2/(np)$. We may plug $\hat{\sigma}^2 = \ell^2/(np)$ into (2.10) to obtain the final part of the theorem. □

This relationship with the multilinear normal model extends to any array-variate elliptically contoured model with separable covariance. Using our identifiable parameterization, X is a mean zero elliptically contoured random array with separable covariance if its density has the form

$$f(x|\sigma^2, \Sigma_1, \ldots, \Sigma_K) \propto (\sigma^2)^{-p/2} g(||(\Sigma_1^{-1/2}, \ldots, \Sigma_K^{-1/2}) \cdot x||^2/\sigma^2),$$

for some known $g : \mathbb{R}^+ \to \mathbb{R}^+$. Using a general result of Anderson et al. [1986] (see A.1.6), the MLE of $\sigma^2(\Sigma_K \otimes \cdots \otimes \Sigma_1)$ can be shown to be proportional to the MLE under the multilinear normal model. This in turn implies that the MLEs of the component covariance matrices in separable elliptically contoured distributions have the same relationship with the HOLQ as in the multilinear normal model. That is, $\hat{\Sigma}_k = L_k L_k^T$ where $X = \ell(L_1, \ldots, L_K, I_n) \cdot Q$. Only the estimation of the scale σ^2 might be different, depending on the function g.

The MLEs of σ^2 and the Σ_k's depend only on ℓ and the L_k's, not Q. This suggests that the core array Q might be ancillary with respect to the covariance parameters $\Sigma_1, \ldots, \Sigma_K$ and σ^2, that is, the distribution of Q might not depend on the parameter values. In the next paragraph, we will prove that this is indeed the case, but to do so we first introduce a group of transformations that acts transitively on the parameter space. Consider the group

$$\mathcal{G} = \{(a, A_1, \ldots, A_K) : a > 0, A_k \in \mathcal{G}_{p_k}^+ \text{ for } k = 1, \ldots, K\},$$

where the group operation is component-wise multiplication. For example, if $(a, A_1, \ldots, A_K)$, $(b, B_1, \ldots, B_K) \in \mathcal{G}$, then we have

$$(a, A_1, \ldots, A_K)(b, B_1, \ldots, B_K) = (ab, A_1 B_1, \ldots, A_K B_K).$$

The group acts on the sample space by

$$X \mapsto a(A_1, \ldots, A_K, I_n) \cdot X.$$

The following theorem shows that under this group action, the core array of the HOLQ, if unique, is maximally invariant (uniqueness is discussed briefly in Section 2.5). More generally, this theorem states that the set of core arrays of fixed points from Algorithm 2 is a maximally invariant statistic. In other words, two arrays are in the same orbit of $\mathcal{G}$ if and only if the set of core arrays of fixed points of Algorithm 2 are the same.

Theorem 4. *Let X and Y be in $\mathbb{R}^{p_1 \times \cdots p_K \times n}$. Let $\mathcal{Q}_X$ and $\mathcal{Q}_Y$ be the set of core arrays from fixed points of Algorithm 2 for X and Y, respectively. Then $\mathcal{Q}_X = \mathcal{Q}_Y$ if and only if there exist $c > 0$ and $C_k \in \mathcal{G}_{p_k}^+$ for $k = 1, \ldots, K$ such that $c(C_1, \ldots, C_K, I_n) \cdot X = Y$.*

Proof. We first prove the "only if" part. Assume that $\mathcal{Q}_X = \mathcal{Q}_Y$, then we choose one Q in $\mathcal{Q}_X = \mathcal{Q}_Y$. Then there exists $a, b > 0$ and $A_k, B_k \in \mathcal{G}_{p_k}^+$ for $k = 1, \ldots, K$ such that $X = a(A_1, \ldots, A_K, I_n) \cdot Q$ and $Y = b(B_1, \ldots, B_K, I_n) \cdot Q$. One may set $c = b/a$ and $C_k = B_k A_k^{-1}$ to prove that $c(C_1, \ldots, C_K, I_n) \cdot X = Y$.

We now prove the "if" part. Assume there exist $c > 0$ and $C_k \in \mathcal{G}_{p_k}^+$ for $k = 1, \ldots, K$ such that $c(C_1, \ldots, C_K, I_n) \cdot X = Y$. Then for each Q in $\mathcal{Q}_X$ we have that $Y = ca(C_1 A_1, \ldots, C_K A_K, I_n) \cdot Q$ for some $a > 0$ and $A_k \in \mathcal{G}_{p_k}^+$ for $k = 1, \ldots, K$. Since fixed points are entirely determined by the scaled all-orthonormality of the core, Q is also in $\mathcal{Q}_Y$. Likewise any Q in $\mathcal{Q}_Y$ will also be in $\mathcal{Q}_X$. Hence $\mathcal{Q}_X = \mathcal{Q}_Y$. □

By using the above invariance results, we may now prove that $\mathcal{Q}_X$ is ancillary. The group $\mathcal{G}$ acts on the parameter space by [Hoff, 2011]

$$\sigma^2 \mapsto a^2 \sigma^2 \text{ and } \Sigma_k \mapsto A_k \Sigma_k A_k^T.$$

This action is clearly transitive over the parameter space. Hence, the maximally invariant parameter is a constant. Since the distribution of any invariant statistic depends only on

the maximally invariant parameter [Lehmann and Romano, 2005, Theorem 6.3.2], the distribution of $\mathcal{Q}_X$ is ancillary with respect to σ^2 and Σ_k for $k = 1, \ldots, K$. If the MLE is unique, then the core array of the HOLQ is in 1-1 correspondence with $\mathcal{Q}_X$, and so is also maximally invariant. Hence, the core array from a unique HOLQ is ancillary with respect to the covariance parameters, $\Sigma_1, \ldots, \Sigma_K$, and σ^2. This result holds not just for elliptically contoured array-variate models with separable covariance, but also for models of the form

$$X \stackrel{d}{=} \sigma(\Sigma_1^{1/2}, \ldots, \Sigma_K^{1/2}, I_n) \cdot Z, \tag{2.11}$$

where Z has a fixed distribution such that $E[Z] = 0$, $\text{cov}(\text{vec}(Z)) = I_{np}$, and $\Sigma_k^{1/2}$ is the lower triangular Cholesky square root of Σ_k.

2.3.2 HOLQ juniors

If it is believed that the dependencies along a mode follow a particular pattern, then from the perspective of parameter estimation, it would make sense to fit a structured covariance matrix that corresponds to the pattern along that mode. For example, if it is believed that the "slices" of the array along a particular mode k are statistically independent, then one would use a model with Σ_k restricted to be a diagonal matrix. If the p_k slices along the mode k are believed to be i.i.d., then one could restrict Σ_k to be the identity matrix. If one of the modes k corresponded to data gathered over sequential time points, then one could fit Σ_k to correspond to an auto-regressive covariance model, such as that of containing constant prediction error variances and arbitrary autoregressive coefficients. One could then restrict Σ_k to have its lower triangular Cholesky square root to have unit diagonal [Pourahmadi, 1999]. Each of these alternatives corresponds to fitting a submodel of an unrestricted separable covariance model.

We represent such submodels mathematically as follows: Partition the index set $\{1, \ldots, K\}$ into four non-overlapping sets J_1, J_2, J_3, J_4. Let $\mathcal{D}_{p_k}^+$ denote the group of p_k by p_k positive definite diagonal matrices with unit determinant. Also, let $\mathcal{S}_{p_k}^{Ch}$ be the space of p_k by p_k symmetric and positive definite matrices whose lower triangular Cholesky square roots have

unit diagonal. Assume the model $X \sim N_{p_1 \times \cdots \times p_K}(0, \sigma^2 \Sigma_K \otimes \cdots \otimes \Sigma_1)$ where Σ_k is in $\mathcal{S}^1_{p_k}$, $\mathcal{D}^+_{p_k}$, $\mathcal{S}^{Ch}_{p_k}$, or $\{I_{p_k}\}$ when k is in J_1, J_2, J_3, or J_4, respectively. The collection of sets J_1, J_2, J_3, and J_4 corresponds to a submodel where the modes in J_1 have unrestricted covariance, the modes in J_2 have diagonal covariance, the modes in J_3 have constant prediction error variances and arbitrary autoregressive coefficients, and the modes in J_4 have independence and homoscedastic covariance structure. If such a submodel represents a close approximation to the truth, then one would expect to obtain better estimates by fitting this submodel than by fitting an unrestricted multilinear normal model.

In the same way that the HOLQ provides the MLEs in the multilinear normal model, the MLEs in submodels of the unconstrained multilinear normal model are provided by a class of Tucker decompositions we call HOLQ juniors. A HOLQ junior is found by constraining the component matrices in the Tucker decomposition to be in a subspace of $\mathcal{G}^+_{p_k}$. In particular, we consider constraining each L_k in (2.5) to be in $\mathcal{G}^+_{p_k}$, $\mathcal{D}^+_{p_k}$, $\mathcal{G}^{Ch}_{p_k}$, or $\{I_{p_k}\}$, where $\mathcal{G}^{Ch}_{p_k}$ denotes the set of p_k by p_k lower triangular matrices with unit diagonal.

Definition 5 (HOLQ junior)**.** *Let $\mathcal{G}^{(k)} = \mathcal{G}^+_{p_k}$, $\mathcal{D}^+_{p_k}$, $\mathcal{G}^{Ch}_{p_k}$, or $\{I_{p_k}\}$ if k is in J_1, J_2, J_3, or J_4, respectively. If*

$$(L_1, \ldots, L_K) = \underset{\tilde{L}_k \in \mathcal{G}^{(k)},\ k=1,\ldots,K}{\arg\min} ||(\tilde{L}_1^{-1}, \ldots, \tilde{L}_K^{-1}) \cdot X||,$$

then

$$X = \ell(L_1, \ldots, L_K) \cdot Q \tag{2.12}$$

is a HOLQ junior, where $\ell = ||(L_1^{-1}, \ldots, L_K^{-1}) \cdot X||$ and $Q = (L_1^{-1}, \ldots, L_K^{-1}) \cdot X/\ell$.

The core array of a HOLQ junior also has a special structure that we prove in the following theorem.

Theorem 5. *Let $X = \ell(L_1, \ldots, L_K) \cdot Q$ be a HOLQ junior (2.12). Then the core array has the following properties:*

1. *$Q_{(k)} Q_{(k)}^T = I_{p_k}/p_k$ for all $k \in J_1$,*

2. $\operatorname{diag}\left(Q_{(k)}Q_{(k)}^T\right) = \mathbf{1}_{p_k}/p_k$ *for all* $k \in J_2$, *where* $\mathbf{1}_{p_k} \in \mathbb{R}^{p_k}$ *is the vector of* 1*'s, and*
3. $Q_{(k)}Q_{(k)}^T = D_k$ *for some diagonal matrix* D_k *for all* $k \in J_3$.

Proof. We may update the modes for which $k \in J_1$ using Theorem 1 the same way we did in Algorithm 2. The core array of any fixed point must then have the property that $Q_{(k)}Q_{(k)}^T = I_{p_k}/p_k$ for all $k \in J_1$. The proofs for $k \in J_2$ and $k \in J_3$ follow along the same lines as in the proof for $k \in J_1$, and are in Appendices A.1.2 and A.1.3. □

The same arguments as used in Section 2.3.1 show that maximum likelihood inference in multilinear normal submodels has a close connection with HOLQ juniors. The proof of the following is very similar to that of Theorem 3 and is omitted.

Theorem 6. *Let* $X = \ell(L_1, \ldots, L_K) \cdot Q$ *be a HOLQ junior. We assume the model* $X \sim N_{p_1 \times \cdots \times p_K}(0, \sigma^2 \Sigma_K \otimes \cdots \otimes \Sigma_1)$ *where* Σ_k *is in* $\mathcal{S}_{p_k}^1$, $\mathcal{D}_{p_k}$, $\mathcal{S}_{p_k}^{Ch}$, *or* $\{I_{p_k}\}$ *when* k *is in* J_1, J_2, J_3, *or* J_4, *respectively. We have the following:*

1. *The MLE of* Σ_k *is* $L_k L_k^T$ *for* $k = 1, \ldots, K$,
2. *The MLE of* σ^2 *is* $\ell^2/(np)$,
3. *The maximum of the likelihood is equal to*

$$\left(2\pi\hat{\sigma}^2\right)^{-np/2} e^{-np/2} = \left(2\pi\ell^2/(np)\right)^{-np/2} e^{-np/2}.$$

We note here that the same group invariance arguments as used in Section 2.3.1 prove that the core array from a unique HOLQ junior is ancillary with respect to the covariance parameters in separable covariance models. That is, a core array from a unique HOLQ junior (2.12) is ancillary under the model

$$X \stackrel{d}{=} \sigma(\Sigma_1^{1/2}, \ldots, \Sigma_K^{1/2}) \cdot Z, \tag{2.13}$$

where Z has a fixed distribution such that $E[Z] = 0$, $\operatorname{cov}(\operatorname{vec}(Z)) = I_p$, and $\Sigma_k^{1/2}$ is the lower Cholesky square root of Σ_k in $\mathcal{S}_{p_k}^1$, $\mathcal{D}_{p_k}^+$, $\mathcal{S}_{p_k}^{Ch}$, or $\{I_{p_k}\}$ when k is in J_1, J_2, J_3, or J_4, respectively. Equivalently, $\Sigma_k^{1/2}$ is in $\mathcal{G}_{p_k}^+$, $\mathcal{D}_{p_k}^+$, $\mathcal{G}_{p_k}^{Ch}$, or $\{I_{p_k}\}$ when k is in J_1, J_2, J_3, or J_4, respectively

2.3.3 Likelihood ratio testing

One would expect to lose efficiency in covariance estimation when fitting a large model when a submodel is a close approximation to the truth. To aid modeling decisions, we develop a class of likelihood ratio tests (LRTs) for comparing nested separable models. For example, a test of independence across slices of mode k would correspond to $H_0 : \Sigma_k \in \mathcal{D}^+_{p_k}$ versus $H_1 : \Sigma_k \in \mathcal{S}^1_{p_k}$. A test for independence and heteroscedasticity against independence and homoscedasticity along mode k would correspond to $H_0 : \Sigma_k = I_{p_k}$ versus $H_1 : \Sigma_k \in \mathcal{D}^+_{p_k}$. In a longitudinal setting, testing for the presence of non-zero autoregressive coefficients along mode k would correspond to $H_0 : \Sigma_k = I_{p_k}$ versus $H_1 : \Sigma_k \in \mathcal{S}^{Ch}_{p_k}$. As seen in Section 2.3.2, each submodel of the unstructured multilinear normal model corresponds to a HOLQ junior. If we have two models H_0 and H_1, with H_0 nested in H_1, then the likelihood ratio test takes on the simple form of the ratio of the two scale estimates of the HOLQ juniors corresponding to H_0 and H_1.

Theorem 7. *Suppose H_0 is a submodel of H_1. Suppose* $\operatorname{vec}(X) = \ell(L_K \otimes \cdots \otimes L_1)\operatorname{vec}(Q)$ *and* $\operatorname{vec}(X) = a(A_M \otimes \cdots \otimes A_1)\operatorname{vec}(Z)$ *are two HOLQ juniors in vectorized form (2.7) corresponding to H_0 and H_1, respectively. Hence, $\hat{\sigma}_0^2 = \ell^2/p$ and $\hat{\sigma}_1^2 = a^2/p$ are the MLEs of the scale parameters under H_0 and H_1, respectively. Then the LRT of H_0 versus H_1 rejects for large values of $\hat{\sigma}_0^2/\hat{\sigma}_1^2$, or equivalently ℓ/a.*

Proof. Applying Theorem 1 from Anderson et al. [1986] and Theorem 6 (see A.1.6), the LRT rejects for large values of

$$\hat{\sigma}_1^{-p}/\hat{\sigma}_0^{-p} = a^{-p}/\ell^{-p} = \ell^p/a^p,$$

or, equivalently, for large values of ℓ/a. □

The LRT in Theorem 7 includes testing for a Kronecker structured covariance matrix along modes k and j against an unrestricted covariance matrix along the concatenated modes of k and j. That is, it allows for the test $H_0 : \Sigma_{kj} = \Sigma_k \otimes \Sigma_j$ for $\Sigma_k \in \mathcal{S}^1_{p_k}$ and $\Sigma_j \in \mathcal{S}^1_{p_j}$ versus

$H_1 : \Sigma_{ij} \in \mathcal{S}^1_{p_k p_j}$. This is why M may be different from K. For example, if all modes in H_0 and H_1 had the same covariance structure except modes k and j, for which H_0 assumes has separable covariance and for which H_1 assumes has unstructured covariance along the concatenated mode kj, then $M = K-1$. This particular type of test is useful for determining how much separability is reasonable to assume in a covariance matrix.

The likelihood ratio test has a nice intuitive interpretation. Since the MLE of σ^2 under H_0 is $\hat{\sigma}_0^2 = \ell^2/p = ||(L_1^{-1}, \ldots, L_K^{-1}) \cdot X||^2/p$ (Theorem 6), one can consider $\hat{\sigma}_0^2$ as a sort of mean squares left after accounting for covariance/heterogeneity along modes $1, \ldots, K$. Likewise $\hat{\sigma}_1^2$ is a sort of mean squares left after accounting for covariance/heterogeneity along modes $1, \ldots, M$. The likelihood ratio test rejects the null when we can explain significantly more heterogeneity in X by increasing the complexity of the covariance structure.

For many hypothesis tests, the distribution of $p\left(\log\left(\ell^2\right) - \log\left(a^2\right)\right)$, the log-likelihood ratio statistic, can be approximated by a χ^2 distribution. However, this asymptotic approximation would be suspect for small sample sizes. We propose using a Monte Carlo approximation to the null distribution of the LRT statistic. This Monte Carlo approximation can be made arbitrarily precise. The following theorem, whose proof is in Appendix A.1.4, suggests how to sample from the null distribution of the LRT statistic, ℓ/a, or $\hat{\sigma}_0/\hat{\sigma}_1$, in Theorem 7.

Theorem 8. *Under H_0, the distribution of ℓ/a in Theorem 7 does not depend on the parameter values $\Sigma_1, \ldots, \Sigma_K$, and σ^2.*

This property of the LRT statistic was noted by Lu and Zimmerman [2005] for the matrix-normal case. An immediate implication of Theorem 8 is that for tests of these covariance models, a Monte Carlo sample of the LRT statistic under H_0 can be made by simulating values of ℓ/a under H_0. A single value of ℓ/a may be simulated from H_0 as follows:

1. sample $x \sim N_p\left(0, I_p\right)$,
2. construct $X_1 \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ and $X_2 \in \mathbb{R}^{q_1 \times \cdots \times q_M}$ from x,
3. calculate HOLQ juniors $X_1 = \ell(L_1, \ldots, L_K) \cdot Q$ and $X_2 = a(A_1, \ldots, A_M) \cdot Z$,

4. calculate ℓ/a.

2.4 Other tensor decompositions

2.4.1 The incredible SVD

The incredible HOLQ (2.5) may be used to derive a higher-order analogue to the SVD that is related to the HOSVD of De Lathauwer et al. [2000b,a]. From (2.5), we take the SVD of each component lower triangular matrix, $L_k = U_k D_k V_k^T$ for $k = 1, \ldots, K$. Letting $V = (V_1^T, \ldots, V_K^T, I_n) \cdot Q$, we now have an exact decomposition of the data array X which may be viewed as a higher-order generalization of the SVD.

Definition 6. *Suppose*

$$X = \ell(U_1, \ldots, U_K, I_n) \cdot [(D_1, \ldots, D_K, I_n) \cdot V] \tag{2.14}$$

such that

1. $\ell \geq 0$,
2. $U_k \in \mathcal{O}_{p_k}$, *the set of* p_k *by* p_k *orthogonal matrices, for all* $k = 1, \ldots, K$,
3. $D_k \in \mathcal{D}_{p_k}^+$, *for all* $k = 1, \ldots, K$, *and*
4. V *is scaled all-orthonormal.*

Then we say that (2.14) is an incredible SVD (ISVD).

The ISVD can be seen as a type of "core rotation" [Kolda and Bader, 2009] of the HOSVD. The core is rotated to a form where we may separate the "mode specific singular values", $D_1, \ldots, D_K$, from the core. Where the core array in the HOSVD is all-orthogonal (the mode-k unfolding contains orthogonal, but not necessarily orthonormal, rows for all $k = 1, \ldots, K$), the core array in the ISVD is scaled all-orthonormal.

A low rank version of the ISVD can be defined by finding, for $r_k \leq p_k$ for $k = 1, \ldots, K$, the $U_k \in \mathcal{V}_{r_k, p_k}$, $D_k \in \mathcal{D}_{r_k}^+$ for $k = 1, \ldots, K$, $\ell > 0$, and $V \in \mathbb{R}^{r_1 \times \cdots \times r_K \times n}$ that minimize

$$||X - \ell(U_1, \ldots, U_K, I_n) \cdot [(D_1, \ldots, D_K, I_n) \cdot V]||^2. \tag{2.15}$$

We can apply the HOOI [higher-order orthogonal iteration, De Lathauwer et al., 2000a] to obtain the minimizer of (2.15). Let $X = (V_1, \ldots, V_K, I_n) \cdot S$ be the HOOI of X. This minimizes

$$||X - (V_1, \ldots, V_K, I_n) \cdot S||^2,$$

for arbitrary core array $S \in \mathbb{R}^{r_1 \times \cdots \times r_K, n}$ and arbitrary $V_k \in \mathcal{V}_{r_k, p_k}$. We now take the ISVD of $S = \ell(W_1, \ldots, W_K, I_n) \cdot [(D_1, \ldots, D_K, I_n) \cdot V]$. We set $U_k = V_k W_k$ for $k = 1, \ldots, K$. These values now minimize (2.15). The truncated ISVD does not improve the fit of the low rank array to the data array over the HOOI. Rather, the truncated ISVD can be seen as a core rotation of the HOOI, the same as how the ISVD can be seen as a core rotation of the HOSVD. Again, the core is rotated to a form where we may separate the mode specific singular values, $D_1, \ldots, D_K$, from the core.

2.4.2 The IHOP decomposition

In this section, we explore how our minimization approach may lead to another novel Tucker decomposition. Let X be a p by n matrix with $p \leq n$ such that X is of rank p. We may write X as

$$X = PW,$$

where $P \in \mathcal{S}_p^+$ and $W^T \in \mathcal{V}_{p,n}$. This is known as the (left) *polar decomposition* (see, for example, Proposition 5.5 of Eaton [1983]). Following the theme of this chapter, we reformulate the polar decomposition as a minimization problem. Let $\mathcal{S}_p^F$ denote the space of p by p positive definite matrices with unit trace.

Theorem 9. *Let*

$$P = \underset{\tilde{P} \in \mathcal{S}_p^F}{\arg\min} \operatorname{tr}(\tilde{P}^{-1} X X^T). \tag{2.16}$$

Set $\ell = ||P^{-1}X||$ *and* $W = P^{-1}X/\ell$. *Then*

$$X = \ell PW$$

is the polar decomposition of X.

Proof. By the uniqueness of the polar decomposition [Eaton, 1983, Proposition 5.5], it suffices to show that W has orthonormal rows. We have that $WW^T = I_p \Leftrightarrow P^{-1}XX^TP^{-1}/\ell^2 = I_p \Leftrightarrow XX^T = \ell^2 PP$. Hence, if we can show that $PP \propto XX^T$ then we have shown that W has orthonormal rows. Using Lagrange multipliers, we must minimize $\text{tr}(P^{-1}XX^T)+\lambda(\text{tr}(P)-1)$. This is equivalent to minimizing $\text{tr}(VXX^T) + \lambda(\text{tr}(V^{-1}) - 1)$ where $V = P^{-1}$. Temporarily ignoring the symmetry, taking derivatives, and setting equal to 0, we have

$$\begin{aligned}
&XX^T - \lambda V^{-1}V^{-1} = 0 \text{ and } \text{tr}(V^{-1}) = 1 \\
&\Leftrightarrow XX^T = \lambda V^{-1}V^{-1} \text{ and } \text{tr}(V^{-1}) = 1 \\
&\Rightarrow V^{-1} = (XX^T)^{1/2}/\,\text{tr}((XX^T)^{1/2}) \text{ and } \lambda = \text{tr}((XX^T)^{1/2})^2 \\
&\Rightarrow P = (XX^T)^{1/2}/\,\text{tr}((XX^T)^{1/2}) \text{ and } \lambda = \text{tr}((XX^T)^{1/2})^2,
\end{aligned}$$

where $(XX^T)^{1/2}$ is any square root matrix of XX^T. Let $(XX^T)^{1/2}$ now be the unique symmetric square root matrix of XX^T, which is a critical point of $\text{tr}(VXX^T)+\lambda(\text{tr}(V^{-1})-1)$ over the space of positive definite matrices. From problem 2 of Section 7.6 in Horn and Johnson [2013], we have that $\text{tr}(V^{-1})$ is strictly convex on the set of positive definite matrices. Since $\lambda = \text{tr}((XX^T)^{1/2})^2 > 0$, we have that $\text{tr}(VXX^T) + \lambda(\text{tr}(V^{-1}) - 1)$ is a convex function for all positive definite V. Therefore $P = (XX^T)^{1/2}/\,\text{tr}((XX^T)^{1/2})$ is a global minimum (c.f. Theorem 13 of Chapter 7 in Magnus and Neudecker [1999]). □

For $X \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$, we now define the incredible higher-order polar decomposition (IHOP).

Definition 7. *If*

$$(P_1, \ldots, P_K) = \underset{P_k \in \mathcal{S}^F_{p_k}, k=1,\ldots,K}{\arg\min} \text{tr}[(P_K^{-1} \otimes \cdots \otimes P_1^{-1})X^T_{(K+1)}X_{(K+1)}], \tag{2.17}$$

then

$$X = \ell(P_1, \ldots, P_K, I_n) \cdot W$$

is an IHOP, where $\ell = ||(P_1^{-1}, \ldots, P_K^{-1}, I_n) \cdot X||$ *and* $W = (P_1^{-1}, \ldots, P_K^{-1}, I_n) \cdot X/\ell$.

Let $\mathcal{G}_p^F$ be the space of lower triangular matrices with positive diagonal elements and unit Frobenius norm. To derive a block coordinate descent algorithm to find the solution to (2.17), we note that (2.16) is equivalent to finding the $L \in \mathcal{G}_p^F$ such that

$$L = \underset{\tilde{L} \in \mathcal{G}_p^F}{\arg\min} ||\tilde{L}^{-1} X||,$$

and then setting $P = LL^T$ for P from (2.16). Hence, (2.17) is equivalent to finding $L_k \in \mathcal{G}_{p_k}^F$ for $k = 1, \ldots, K$ such that

$$(L_1, \ldots, L_K) = \underset{\tilde{L}_k \in \mathcal{G}_{p_k}^F, k=1,\ldots,K}{\arg\min} ||(\tilde{L}_1^{-1}, \ldots, \tilde{L}_K^{-1}, I_n) \cdot X||, \tag{2.18}$$

then setting $P_k = L_k L_k^T$ for $k = 1, \ldots, K$. At iteration i, fix L_k for $k \neq i$. We then find the minimizer in $L_i \in \mathcal{G}_{p_i}^F$ of

$$||L_i^{-1} X_{(i)} L_{-i}^{-T}|| = \text{tr}(P_i^{-1} X_{(i)} P_{-i}^{-1} X_{(i)}^T),$$

which, by Theorem 9 is $L \in \mathcal{G}_{p_k}^F$ such that $LL^T W = X_{(i)} L_{-i}^{-1}$ is the polar decomposition of $X_{(i)} L_{-i}^{-1}$. This algorithm is presented in Algorithm 3. Again following the theme in this chapter, we present a slightly improved algorithm in Algorithm 4. A proof that Algorithm 3 and Algorithm 4 are equivalent can be found in Appendix A.1.5. From the Algorithm 4, we see that any fixed point of R in Algorithm 4 must have the property that $R_{(k)} = L_k Z$ for the current value of L_k and some Z with orthonormal rows.

2.5 Discussion

In this chapter, we have presented a higher-order generalization of the LQ decomposition by reformulating the LQ decomposition as a minimization problem. The orthonormal rows property of the Q matrix in the LQ decomposition generalizes to the scaled all-orthonormal property of the mode-k unfoldings of the core array in the HOLQ. We generalized the HOLQ to HOLQ juniors by constraining the component matrices to subspaces of $\mathcal{G}_{p_k}^+$. One application of the HOLQ (junior) is for estimation and testing in separable covariance models.

Algorithm 3 Block coordinate descent for the IHOP.

Given $X \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$, initialize:

$L_k \leftarrow L_{k0} \in \mathcal{G}_{p_k}^F$ for $k = 1, \ldots, K$.

repeat

for $k \in \{1, \ldots, K\}$ **do**

Polar decomposition of $X_{(k)} L_{-k}^{-1} = PZ^T$

Cholesky decomposition of $P = LL^T$

$L_k \leftarrow L/||L||$

end for

until Convergence.

Set $P_k \leftarrow L_k L_k^T$ for $k = 1, \ldots, K$

Set $\ell \leftarrow ||(P_1^{-1}, \ldots, P_K^{-1}, I_n) \cdot X||$

Set $W \leftarrow (P_1^{-1}, \ldots, P_K^{-1}, I_n) \cdot X/\ell$

return ℓ, W, and P_k for $k = 1, \ldots, K$.

The MLEs of the covariance parameters may be recovered from the HOLQ (junior) and the likelihood ratio test has the simple form of the ratio of two scale estimates from the HOLQ junior. The core array from the HOLQ (junior) is ancillary with respect to the covariance parameters.

We also used the HOLQ to develop a higher-order generalization of the SVD. Our version of the SVD can be viewed as a core rotation for the HOSVD (full rank case) or the HOOI (low rank case), where the core is rotated so that the mode specific singular values may be separated from the core array. We note that one can consider the model of Hoff [2013] as a model based truncated ISVD. He considered the model

$$X \sim N_{p_1 \times \cdots \times p_K}((U_1, \ldots, U_K, I_n) \cdot [(D_1, \ldots, D_K, I_n) \cdot V], \sigma^2 I_p), \text{ where:}$$

U_k is uniformly distributed on $\mathcal{V}_{r_k, p_k}$,

D_k has trace 1 and is uniformly distributed on the r_k simplex,

Algorithm 4 Orthogonalized block coordinate descent for the IHOP.

Given $X \in \mathbb{R}^{p_1 \times \cdots \times p_K \times n}$, initialize:

$L_k \leftarrow L_{k0} \in \mathcal{G}^F_{p_k}$ for $k = 1, \ldots, K$.

$\ell \leftarrow ||(L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X||$

$R \leftarrow (L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X / \ell$

repeat

  for $k \in \{1, \ldots, K\}$ **do**

    Polar decomposition of $L_k R_{(k)} = PZ$

    Cholesky decomposition of $P = LL^T$

    Set $R_{(k)} \leftarrow L^T Z$

    Set $L_k \leftarrow L$

    Re-scale:

    $\ell \leftarrow \ell ||L_k|| ||R||$

    $L_k \leftarrow L_k / ||L_k||$

    $R \leftarrow R / ||R||$

  end for

until Convergence.

Set $P_k \leftarrow L_k L_k^T$ for $k = 1, \ldots, K$

Set $\ell \leftarrow ||(L_1^{-1}, \ldots, L_K^{-1}) \cdot R||$

Set $W = (L_1^{-1}, \ldots, L_K^{-1}) \cdot R / \ell$

return ℓ, W, and P_k for $k = 1, \ldots, K$.

$$V \sim N_{r_1 \times \cdots \times r_K}(0, \tau^2 I_r), \text{ and}$$
$$\tau^2 \sim \text{inverse-gamma}(1/2, \tau_0^2/2),$$

where we changed the notation from his paper to make more clear the connection to the ISVD. In such a model, the core V is scaled all-orthonormal in expectation. That is, $E[V_{(k)} V_{(k)}^T] \propto I_{p_k}$ for all $k = 1, \ldots, K$. One could extend his results by selecting a prior that allows for

non-zero mass for the D_k to be of low rank, as in Hoff [2007] for his model based SVD.

A clear limitation to the utility or the HOLQ or ISVD in practice is that in some dimensions they may not exist, and in other dimensions where they do exist, they may not be unique. The necessary and sufficient conditions for the existence and uniqueness of the HOLQ are not known. Sufficient conditions for existence and uniqueness occur when n is large. When $n \geq p$, the criterion function, $||(L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X||$, is bounded below by the value at the LQ decomposition. For n large enough, the HOLQ is also unique, this follows from the uniqueness of the MLE from Ohlson et al. [2013]. These conditions are equivalently sufficient for the existence and uniqueness of the ISVD. However, in the author's experience, the HOLQ exists and is unique for many dimensions where $n < p$, indeed for many dimensions where $n = 1$. In cases where the HOLQ/ISVD do not exist, the model of Hoff [2013] would be a good alternative. One could also construct a regularized version of the HOLQ.

We note, however, that when a *local* minimum is reached, then the HOLQ exists. This is due to the geodesic convexity results of the log-likelihood in Wiesel [2012b,a]. That is, any local minimum is also a global minimum. These results indicate that, for any particular data set, we can determine if any global minima exist.

Chapter 3

EQUIVARIANT MINIMAX DOMINATORS OF THE MLE IN THE ARRAY NORMAL MODEL

3.1 Introduction

The analysis of array-valued data, or tensor data, is of interest to numerous fields, including psychometrics [Kiers and Mechelen, 2001], chemometrics [Smilde et al., 2005, Bro, 2006], imaging [Vasilescu and Terzopoulos, 2003], signal processing [Cichocki et al., 2014] and machine learning [Tao et al., 2005], among others [Kroonenberg, 2008, Kolda and Bader, 2009]. Such data consist of measurements indexed by multiple categorical factors. For example, multivariate measurements on experimental units over time may be represented by a three-way array $X = \{x_{i,j,t}\} \in \mathbb{R}^{m \times p \times t}$, with i indexing units, j indexing variables and t indexing time. Another example is multivariate relational data, where $x_{i,j,k}$ is the type-k relationship between person i and person j.

Statistical analysis of such data often proceeds by fitting a model such as $X = \Theta + E$, where Θ is low-dimensional and E represents additive residual variation about Θ. Standard models for Θ include regression models, additive effects models (such as those estimated by ANOVA decompositions) and unconstrained mean models if replicate observations are available. Another popular approach is to model Θ as being a low-rank array. For such models, ordinary least-squares estimates of Θ can be obtained via various types of tensor decompositions, depending on the definition of rank being used [De Lathauwer et al., 2000a,b, de Silva and Lim, 2008].

Less attention has been given to the analysis of the residual variation E. However, estimating and accounting for such variation is critical for a variety of inferential tasks, such as prediction, model-checking, construction of confidence intervals, and improved parameter

estimation over ordinary least squares. One model for variation among the entries of an array is the array normal model [Akdemir and Gupta, 2011, Hoff, 2011] which is an extension of the matrix normal model [Srivastava and Khatri, 1979, Dawid, 1981], often used in the analysis of spatial and temporal data [Mardia and Goodall, 1993, Shitan and Brockwell, 1995, Fuentes, 2006]. The array normal model is a class of normal distributions that are generated by a multilinear operator known as the Tucker product: A random K-way array X taking values in $\mathbb{R}^{p_1 \times \cdots \times p_K}$ has an array normal distribution if $X \stackrel{d}{=} \Theta + (A_1, \ldots, A_K) \cdot Z$, where "$\cdot$" denotes the Tucker product (described further in Section 3.2), Z is a random array in $\mathbb{R}^{p_1 \times \cdots \times p_K}$ having i.i.d. standard normal entries, and A_k is a $p_k \times p_k$ nonsingular matrix for each $k \in \{1, \ldots, K\}$. Letting $\Sigma_k = A_k A_k^T$ and "$\otimes$" denote the Kronecker product, we write

$$X \sim N_{p_1 \times \cdots \times p_K}(\Theta, \Sigma_K \otimes \cdots \otimes \Sigma_1). \tag{3.1}$$

A maximum likelihood estimate (MLE) for the parameters in (3.1) can be obtained via an iterative coordinate descent algorithm [Hoff, 2011], which is a generalization of the iterative "flip-flop" algorithm developed in Mardia and Goodall [1993] and Dutilleul [1999], or alternatively the optimization procedures described in Wiesel [2012a]. However, based on results for the multivariate normal model, one might suspect that the MLE lacks desirable optimality properties: In the multivariate normal model, James and Stein [1961] showed that the MLE of the covariance matrix is neither admissible nor minimax. This was accomplished by identifying a minimax and uniformly optimal equivariant estimator that is different from the (equivariant) MLE, and therefore dominates the MLE. As pointed out by James and Stein, this equivariant estimator is itself inadmissible, and improvements to this estimator have been developed and studied by Stein [1975], Takemura [1984], Lin and Perlman [1985], and Haff [1991], among others.

This chapter develops similar results for the array normal model. In particular, we obtain a procedure to obtain the uniformly minimum risk equivariant estimator (UMREE) under a lower-triangular product group of transformations for which the model (3.1) is invariant. Unlike for the multivariate normal model, there is no simple characterization of this class

of equivariant estimators. However, results of Zidek [1969] and Eaton [1989] can be used to show that the UMREE can be obtained from the Bayes decision rule under an improper prior, which we derive in Section 3.2. In Section 3.3 we obtain the posterior distribution under this prior, and show how it can be simulated from using a Markov chain Monte Carlo (MCMC) algorithm. Specifically, the MCMC algorithm is a Gibbs sampler that involves simulation from a class of distributions over covariance matrices, which we call the "mirror-Wishart" distributions.

In Section 3.4.1 we develop a version of Stein's loss function for covariance estimation in the array normal model, and show how the Gibbs sampler of Section 3.3 can be used to obtain the UMREE for this loss. We discuss an orthogonally equivariant improvement to the UMREE in Section 3.4.2, which can be seen as analogous to the estimator studied by Takemura [1984]. Section 3.4.3 compares the risks of the MLE, UMREE and the orthogonally equivariant estimator as a function of the dimension of X in a small simulation study. A discussion follows in Section 3.5. Proofs are contained in an appendix.

3.2 An invariant measure for the array normal model

3.2.1 The array normal model

The array normal model on $\mathbb{R}^{p_1 \times \cdots \times p_K}$ consists of the distributions of random K-arrays $X \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ for which

$$X \stackrel{d}{=} \Theta + (A_1, \ldots, A_K) \cdot Z \tag{3.2}$$

for some $\Theta \in \mathbb{R}^{p_1 \times \cdots \times p_K}$, nonsingular matrices $A_k \in \mathbb{R}^{p_k \times p_k}, k = 1, \ldots, K$ and a random $p_1 \times \cdots \times p_K$ array Z with i.i.d. standard normal entries. Here, "$\cdot$" denotes the *Tucker product*, which is defined by the identity

$$\text{vec}((A_1, \ldots, A_K) \cdot Z) = (A_K \otimes \cdots \otimes A_1)z, \tag{3.3}$$

where "$\otimes$" is the Kronecker product and $z = \text{vec}(Z)$, the vectorization of Z. This identity can be used to find the covariance of the elements of a random array satisfying (3.2): Letting

x, z, θ be the vectorizations of X, Z, Θ, we have

$$\begin{aligned}\mathrm{Cov}[x] &= \mathrm{E}[(x-\theta)(x-\theta)^T] = \mathrm{E}[(A_K \otimes \cdots \otimes A_1) z z^T (A_K^T \otimes \cdots \otimes A_1^T)] \\ &= (A_K \otimes \cdots \otimes A_1)(A_K^T \otimes \cdots \otimes A_1^T) = (A_K A_K^T \otimes \cdots \otimes A_1 A_1^T),\end{aligned}$$

and so the array normal distributions correspond to the multivariate normal distributions with separable (Kronecker structured) covariance matrices.

A useful operation related to the Tucker product is the *matricization* operation, which reshapes an array into a matrix along an index set, or *mode*. For example, the *mode-k matricization* of Z is the $p_k \times (\prod_{l:l\neq k} p_l)$-dimensional matrix $Z_{(k)}$ having rows equal to the vectorizations of the "slices" of Z along the kth index set. An important identity involving the Tucker product is that if $Y = (A_1, \ldots, A_K) \cdot Z$ then

$$Y_{(k)} = A_k Z_{(k)} \left(A_K^T \otimes \cdots \otimes A_{k+1}^T \otimes A_{k-1}^T \otimes \cdots \otimes A_1^T\right). \tag{3.4}$$

As shown in Hoff [2011], a direct application of this identity gives

$$E\left[(X_{(k)} - \Theta_{(k)})(X_{(k)} - \Theta_{(k)})^T\right] = c_k A_k A_k^T,$$

where c_k is a scalar. This shows that $A_k A_k^T$ can be interpreted as the covariance among the p_k slices of the array X along its kth mode.

The array normal model can be parameterized in terms of a mean array $\mathrm{E}[X] = \Theta \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ and covariance $\mathrm{Cov}[\mathrm{vec}(X)] = \sigma^2(\Sigma_K \otimes \cdots \otimes \Sigma_1)$, where $\sigma^2 > 0$ and for each k, $\Sigma_k \in \mathcal{S}_{p_k}^+$, the set of $p_k \times p_k$ positive definite matrices. To make the parameterization identifiable, we restrict the determinant of each Σ_k to be one. Denote by $\mathcal{S}_{\mathbf{p}}^+$ this parameter space, that is, the values of $(\sigma^2, \Sigma_1, \ldots, \Sigma_K)$ for which $|\Sigma_k| = 1$, $k = 1, \ldots, K$. Under this parameterization, we write $X \sim N_{p_1 \times \cdots \times p_K}(\Theta, \sigma^2(\Sigma_K \otimes \cdots \otimes \Sigma_1))$ if and only if $X \overset{d}{=} \Theta + \sigma(\Psi_1, \ldots, \Psi_K) \cdot Z$, where for each k, Ψ_k is a matrix such that $\Psi_k \Psi_k^T = \Sigma_k$.

Given a sample $X_1, \ldots, X_n \sim$ i.i.d. $N_{p_1 \times \cdots \times p_K}(\Theta, \sigma^2(\Sigma_K \otimes \cdots \otimes \Sigma_1))$, the $(K+1)$-array X obtained by "stacking" $X_1, \ldots, X_n$ along a $(K+1)$st mode also has an array normal distribution,

$$X \sim N_{p_1 \times \cdots \times p_K \times n}\left(\Theta \circ \mathbf{1}_n, \sigma^2(I_n \otimes \Sigma_K \otimes \cdots \otimes \Sigma_1)\right),$$

where $\mathbf{1}_n$ is the $n \times 1$ vector of ones and "$\circ$" denotes the outer product. If $n > 1$ then covariance estimation for the array normal model can be reduced to the case that $\Theta = 0$. To see this, let H be a $(n-1) \times n$ matrix such that $HH^T = I_{n-1}$ and $H1_n = 0$. This implies that $H^T H = I_n - \mathbf{1}_n \mathbf{1}_n^T/n$. Letting $Y = (I_{p_1}, \ldots, I_{p_K}, H) \cdot X$, and $Y_{(K+1)}$ be the mode-$(K+1)$ matricization of Y, we have $\mathrm{E}[Y_{(K+1)}] = H\mathrm{E}[X_{(K+1)}] = H\mathbf{1}_n \operatorname{vec}(\Theta)^T = \mathbf{0}$, and so Y is mean-zero. Using identity (3.3), the covariance of $\operatorname{vec}(Y)$ can be shown to be $\sigma^2(HH^T \otimes \Sigma_K \otimes \cdots \otimes \Sigma_1) = \sigma^2(I_{n-1} \otimes \Sigma_K \otimes \cdots \otimes \Sigma_1)$, and so $Y \sim N_{p_1 \times \cdots \times p_K \times (n-1)}(0, \sigma^2(I_{n-1} \otimes \Sigma_K \otimes \cdots \otimes \Sigma_1))$. For the remainder of this chapter, we consider covariance estimation in the case that $\Theta = 0$.

3.2.2 Model invariance and a right invariant measure

Consider the model for an i.i.d. sample of size n from a p-variate mean-zero multivariate normal distribution, $X \sim N_{p \times n}(0, I_n \otimes \Sigma)$, $\Sigma \in \mathcal{S}_p^+$. Recall that $AX \sim N_{p \times n}(0, I_n \otimes A\Sigma A^T)$ for nonsingular matrices A, and so in particular this model is invariant under left multiplication of X by elements of G_p^+, the group of lower triangular matrices with positive diagonals. An estimator $\hat{\Sigma}$ mapping the sample space $\mathbb{R}^{p \times n}$ to $\mathcal{S}_p^+$ is said to be equivariant under this group if $\hat{\Sigma}(AX) = A\hat{\Sigma}(X)A^T$ for all $A \in G_p^+$ and $X \in \mathbb{R}^{p \times n}$. James and Stein [1961] characterized the class of equivariant estimators for this model, identified the UMREE under a particular loss function and showed that the UMREE is minimax. Additionally, as the MLE XX^T/n is equivariant and different from the UMREE, the MLE is dominated by the UMREE.

We pursue analogous results for the array normal model by first reparameterizing in terms of the parameter $\Sigma^{1/2} = (\sigma, \Psi_1, \ldots, \Psi_K)$, so

$$X \sim N_{p_1 \times \cdots \times p_K \times n}\left(0, \sigma^2\left(I_n \otimes \Psi_K \Psi_K^T \otimes \cdots \otimes \Psi_1 \Psi_1^T\right)\right), \tag{3.5}$$

where $\sigma > 0$ and each Ψ_k is in the set $\mathcal{G}_{p_k}^+$ of $p_k \times p_k$ lower triangular matrices with positive diagonals and determinant 1. In this parameterization, Ψ_k is the lower triangular Cholesky square root of the mode-k covariance matrix Σ_k described in Section 3.2.1.

Define the group $\mathcal{G}_{\mathbf{p}}^+$ as

$$\mathcal{G}_{\mathbf{p}}^+ = \left\{A = (a, A_1, \ldots, A_K) \ : \ a > 0, A_k \in \mathcal{G}_{p_k}^+ \text{ for } k = 1, \ldots, K\right\},$$

where the group operation is

$$AT = (a, A_1, \ldots, A_K)(t, T_1, \ldots, T_K) = (at, A_1T_1, \ldots, A_KT_K).$$

Note that $\mathcal{G}^+_{\mathbf{p}}$ consists of the same set as the parameter space for the model, as parameterized in (3.5). If the group $\mathcal{G}^+_{\mathbf{p}}$ acts on the sample space by

$$g : X \mapsto a(A_1, \ldots, A_K, I_n) \cdot X,$$

then as shown in Hoff [2011] it acts on the parameter space by

$$g : (\sigma, \Psi_1, \ldots, \Psi_K) \mapsto (a\sigma, A_1\Psi_1, \ldots, A_K\Psi_K),$$

which we write concisely as $g : \Sigma^{1/2} \mapsto A\Sigma^{1/2}$. An estimator, $\hat{\Sigma}^{1/2} = (\hat{\sigma}, \hat{\Psi}_1, \ldots, \hat{\Psi}_K)$, mapping the sample space $\mathbb{R}^{p_1 \times \cdots \times p_K \times n}$ to the parameter space $\mathcal{G}^+_{\mathbf{p}}$ is equivariant if

$$\hat{\Sigma}^{1/2}\left(a(A_1, \ldots, A_K, I_n) \cdot X\right) = (a, A_1, \ldots, A_K)\hat{\Sigma}^{1/2}(X).$$

For example, if $\hat{\Psi}_k$ is the estimator of Ψ_k when observing X, then $A_k\hat{\Psi}_k$ is the estimator when observing $a(A_1 \ldots, A_K, I_n) \cdot X$.

Unlike the case for the multivariate normal model, the class of $\mathcal{G}^+_{\mathbf{p}}$- equivariant estimators for the array normal model is not easy to characterize beyond the definition given above. However, in cases like the present one where the group space and parameter space are the same, the UMREE under an invariant loss can be obtained as the generalized Bayes decision rule under a (generally improper) prior obtained from a right invariant (Haar) measure over the group [Zidek, 1969, Eaton, 1989]. The first step towards obtaining the UMREE is then to obtain a right invariant measure and corresponding prior. To do this, we first need to define an appropriate measure space for the elements of $\mathcal{G}^+_{\mathbf{p}}$. Recall that matrices A_k in $\mathcal{G}^+_{p_k}$ have determinant 1, and so one of the nonzero elements of A_k can be expressed as a function of the others. For the rest of this section and the next, we parameterize $A_k \in \mathcal{G}^+_{p_k}$ in terms of the elements $\{A_{k[i,j]} : 2 \leq i \leq p_k, 1 \leq j \leq i\}$, and express the upper-left element $A_{k[1,1]}$ as a function of the other diagonal elements, so that

$A_{k[1,1]} = \prod_{i=2}^{p_k} (A_{k[i,i]})^{-1}$. The "free" elements of $A_k \in \mathcal{G}_{p_k}^+$ therefore take values in the space $\mathcal{A}_{p_k} = \{a_{i,i} > 0, a_{i,j} \in \mathbb{R} : 2 \leq i \leq p_k, 1 \leq j < i\}$.

Theorem 10. *A right invariant measure over the group $\mathcal{G}_{\mathbf{p}}^+$ is*

$$d\nu_r\left(a, A_1, \ldots, A_K\right) = \frac{1}{a}\left(\prod_{k=1}^{K}\prod_{i=2}^{p_k} A_{k[i,i]}^{i-2}\right)\; d\mu\left(a, A_1, \ldots, A_K\right),$$

where $d\mu$ is Lebesgue measure over $\mathbb{R}^+ \times \mathcal{A}_{p_1} \times \cdots \times \mathcal{A}_{p_K}$.

We note that although the density given above is specific to the particular parameterization of the $\mathcal{G}_{p_k}^+$'s, the inference results that follow will hold for any parameterization.

Let $L : \mathcal{G}_{\mathbf{p}}^+ \times \mathcal{G}_{\mathbf{p}}^+ \to \mathbb{R}^+$ be an invariant loss function, so that $L(\Sigma^{1/2}, B) = L(A\Sigma^{1/2}, AB)$ for all A, B and $\Sigma^{1/2} \in \mathcal{G}_{\mathbf{p}_k}^+$. Theorem 6.5 of Eaton [1989] implies that the value of the UMREE when the array X is observed is the minimizer in $B = (b, B_1, \ldots, B_K)$ of the integral

$$\int_{\mathcal{G}_{\mathbf{p}}^+} L(A\Sigma_0^{1/2}, B) \times p(X|A\Sigma_0^{1/2})\; d\nu_r(A),$$

where $p(X|A\Sigma_0^{1/2})$ is the array normal density at the parameter value $A\Sigma_0^{1/2}$ and $\Sigma_0^{1/2}$ is an arbitrary element of $\mathcal{G}_{\mathbf{p}}^+$. Since the group action is transitive over the parameter space, and since the integral is right invariant, $\Sigma_0^{1/2}$ can be chosen to be equal to $(1, I_{p_1}, \ldots, I_{p_K})$. Furthermore, since the parameter space and group space are the same, replacing A with $\Sigma^{1/2}$ in the above integral indicates that the UMREE at X is the minimizer in B of

$$\int_{\mathcal{G}_{\mathbf{p}}^+} L(\Sigma^{1/2}, B) \times p(X|\Sigma^{1/2})\; d\nu_r(\Sigma^{1/2}),$$

that is, the UMREE is the Bayes estimator under the (improper) prior ν_r for $\Sigma^{1/2}$. This is summarized in the following corollary:

Corollary 1. *For an invariant loss function $L : \mathcal{G}_{\mathbf{p}}^+ \times \mathcal{G}_{\mathbf{p}}^+ \to \mathbb{R}^+$ the estimator $\hat{\Sigma}^{1/2}$, defined as*

$$\hat{\Sigma}^{1/2}(X) = \underset{B \in \mathcal{G}_{\mathbf{p}}^+}{\arg\min}\, E[L(\Sigma^{1/2}, B)|X], \tag{3.6}$$

uniformly minimizes the risk $E[L(\Sigma^{1/2}, \tilde{\Sigma}^{1/2}(X))|\Sigma^{1/2}]$ *among equivariant estimators* $\tilde{\Sigma}^{1/2}$ *of* $\Sigma^{1/2}$. *The expectation in (3.6) is with respect to the posterior density*

$$p(\sigma, \Psi_1, \ldots, \Psi_K|X) \propto \tag{3.7}$$
$$\sigma^{-np} \exp\left\{-\frac{1}{2\sigma^2}||(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X||^2\right\} \frac{1}{\sigma} \prod_{k=1}^{K} \prod_{i=2}^{p_k} \Psi_{k[i,i]}^{i-2},$$

where $p = \prod_1^K p_k$.

In addition to uniformly minimizing the risk, the UMREE has two additional features. First, since any unique MLE is equivariant [Eaton, 1989, Theorem 3.2], the UMREE dominates any unique MLE, presuming the UMREE is not the MLE. Second, the UMREE under $\mathcal{G}_{\mathbf{p}}^+$ is minimax. This follows because $\mathcal{G}_{\mathbf{p}}^+$ is a subgroup of G_p^+, as $a(A_K \otimes \cdots \otimes A_1) \in G_p^+$ for all $a > 0$ and $A_k \in \mathcal{G}_{p_k}^+$. Since G_p^+ is a solvable group [James and Stein, 1961], this necessarily implies that $\mathcal{G}_{\mathbf{p}}^+$ is solvable [Rotman, 1995, Theorem 5.15]. By the results of Kiefer [1957] and Bondar and Milnes [1981], the equivariant estimator that minimizes (3.6) is minimax.

Note that because the prior ν_r is improper, the posterior (3.7) is not guaranteed to be proper. However, we are able to guarantee propriety if the sample size n is sufficiently large:

Theorem 11. *Let* $n > \prod_{k=1}^K p_k$. *For* $p(\sigma, \Psi_1, \ldots, \Psi_K|X)$ *defined in (3.7),*

$$\int_{\mathbb{R}^+ \times \mathcal{G}_{p_1}^+ \times \cdots \times \mathcal{G}_{p_K}^+} p(\sigma, \Psi_1, \ldots, \Psi_K|X) d\sigma d\Psi_1 \cdots d\Psi_K < \infty$$

The sample size in the Theorem is sufficient for propriety, but empirical evidence suggests that it is not necessary. For example, results from a simulation study in Section 4 suggest that, for some dimensions, a sample size of $n = 1$ is sufficient for posterior propriety and existence of an UMREE.

3.3 Posterior approximation

For the results in Section 3.2 to be of use, we must be able to actually minimize the posterior risk in Equation 3.6 under an invariant loss function of interest. In the next section, we will

show that the posterior risk minimizer under a multiway generalization of Stein's loss is given by posterior expectations of the form $\mathrm{E}[(\sigma^2\Sigma_k)^{-1}|X]$, where $\Sigma_k = \Psi_k\Psi_k^T$. Although these posterior expectations are not generally available in analytic form, they can be approximated using a MCMC algorithm. In this section, we show how a relatively simple Gibbs sampler can be used to simulate a Markov chain of values of $\Sigma^{1/2} = (\sigma, \Psi_1, \ldots, \Psi_K)$, having a stationary distribution equal to the desired posterior distribution given by Equation 3.7. These simulated values can be used to approximate the posterior distribution of $\Sigma^{1/2}$ given X, as well as any posterior expectation, in particular $\mathrm{E}[(\sigma^2\Sigma_k)^{-1}|X]$.

The Gibbs sampler proceeds by iteratively simulating values of $\{\sigma, \Psi_k\}$ from their full conditional distribution given the current values of $\{\Psi_1, \ldots, \Psi_{k-1}, \Psi_{k+1}, \ldots, \Psi_K\}$. This is done by simulating $\sigma^2\Sigma_k$ from its full conditional distribution, from which σ and Ψ_k can be recovered. One iteration of the Gibbs sampler proceeds as follows:

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

1. simulate $(\sigma^2\Sigma_k)^{-1} \sim \text{mirror-Wishart}_{p_k}(np/p_k, (X_{(k)}\Psi_{-k}^{-T}\Psi_{-k}^{-1}X_{(k)}^T)^{-1})$;

2. set Ψ_k to be the lower triangular Cholesky square root of Σ_k.

In this algorithm, $X_{(k)} \in \mathbb{R}^{p_k \times np/p_k}$ is the mode-k matricization of X and $\Psi_{-k} = \Psi_K \otimes \cdots \otimes \Psi_{k+1} \otimes \Psi_{k-1} \otimes \cdots \otimes \Psi_1$. The mirror-Wishart distribution is a probability distribution on positive definite matrices, related to the Wishart distribution as follows:

Definition 8. *A random $q \times q$ positive definite matrix S has a mirror-Wishart distribution with degrees of freedom $\nu > 0$ and scale matrix $\Phi \in \mathcal{S}_q^+$ if*

$$S \stackrel{d}{=} UV^TVU^T,$$

where VV^T is the lower triangular Cholesky decomposition of a Wishart$_q(\nu, I_q)$-distributed random matrix and UU^T is the upper triangular Cholesky decomposition of Φ.

Some understanding of the mirror-Wishart distribution can be obtained from its expectation:

Lemma 1. *If* $S \sim$ *mirror-Wishart*$_q(\nu, \Phi)$ *then*

$$E[S] = \nu U D U^T$$

where UU^T *is the upper triangular Cholesky decomposition of* Φ *and* D *is a diagonal matrix with entries* $d_j = (\nu + q + 1 - 2j)/\nu,\ j = 1, \ldots, q$.

The calculation follows from Bartlett's decomposition, and is in the appendix. The implications of this for covariance estimation are best understood in the context of the multivariate normal model $X \sim N_{p\times n}(0, I_n \otimes \Sigma)$. In this case, for a given prior the Bayes estimator under Stein's loss is given by $\mathrm{E}[\Sigma^{-1}|X]^{-1}$ (see, for example Yang and Berger [1994]). Under Jeffreys' noninformative prior, $\Sigma^{-1} \sim \mathrm{Wishart}_p(n, (XX^T)^{-1})$ and so the Bayes estimator is XX^T/n. While unbiased, this estimator is generally thought of as not providing appropriate shrinkage of the sample eigenvalues. Note that under Jeffreys' prior, *a posteriori* we have $\Sigma^{-1} \stackrel{d}{=} UVV^TU^T$, where $VV^T \sim \mathrm{Wishart}_p(n, I_p)$ and UU^T is the upper triangular Cholesky decomposition of $(XX^T)^{-1}$. In contrast, under a right invariant measure as our prior we have $\Sigma^{-1} \stackrel{d}{=} UV^TVU^T$. The expectation of VV^T is nI, whereas the expectation of V^TV is nD, which provides a different pattern of shrinkage of the eigenvalues of XX^T. By Lemma 1 , the Bayes estimator under a right invariant measure as our prior in this case is given by $(nUDU^T)^{-1} = U^{-T}D^{-1}U^{-1}/n$, which is the UMREE obtained by James and Stein [1961]. Thus, the UMREE in the multivariate normal model corresponds to a Bayes estimator under a right invariant measure as our prior and mirror-Wishart posterior distribution.

The Gibbs sampler is based on the full conditional distribution of $(\sigma^2\Sigma_k)^{-1}$, which we derive from the full conditional density of $\{\sigma, \Psi_k\}$:

$$p(\sigma, \Psi_k) \propto |\sigma\Psi_k|^{-(np+1)/p_k} \exp\left\{-\operatorname{tr}\left((\sigma^2\Psi_k\Psi_k^T)^{-1}X_{(k)}\Psi_{-k}^{-T}\Psi_{-k}^{-1}X_{(k)}^T\right)/2\right\}\prod_{i=2}^{p_k}\Psi_{k[i,i]}^{i-2},$$

where dependence of the density on $\{\Psi_1, \ldots, \Psi_{k-1}, \Psi_{k+1}, \ldots, \Psi_K, X\}$ has been made implicit. Now set $L_k = \sigma\Psi_k$. The full conditional density of L_k can be obtained from that of $\{\sigma, \Psi_k\}$ and the Jacobian of the transformation.

Lemma 2. *The Jacobian of the transformation* $g(\sigma, \Psi_k) = \sigma\Psi_k$, *mapping* $\mathbb{R}^+ \times \mathcal{G}_{p_k}^+$ *to* $G_{p_k}^+$ *is*

$$J(\sigma, \Psi_k) \propto \sigma^{p_k(p_k+1)/2-1}\Psi_{k[1,1]}.$$

Since $L_k = \sigma\Psi_k$, we have $\sigma = |L_k|^{1/p_k}$ and $\Psi_{k[i,i]} = L_{k[i,i]}/\sigma = L_{k[i,i]}/|L_k|^{1/p_k}$. Lemma 2 implies

$$\begin{aligned} p(L_k) \propto |L_k^T|^{-(np+1)/p_k} \exp\left\{-\operatorname{tr}\left((L_kL_k^T)^{-1}X_{(k)}\Psi_{-k}^{-T}\Psi_{-k}^{-1}X_{(k)}^T\right)/2\right\} \\ \times \prod_{i=2}^{p_k}\left(L_{k[i,i]}/|L_k|^{1/p_k}\right)^{i-2}\left(|L_k|^{1/p_k}\right)^{-p_k(p_k+1)/2+1}\left(L_{k[1,1]}/|L_k|^{1/p_k}\right)^{-1}, \end{aligned}$$

which, through straightforward calculations, can be shown to be proportional to

$$\left(\prod_{i=1}^{p_k} L_{k[i,i]}^{i-np/p_k-p_k-1}\right)\exp\left\{-\operatorname{tr}\left((L_kL_k^T)^{-1}X_{(k)}\Psi_{-k}^{-T}\Psi_{-k}^{-1}X_{(k)}^T\right)/2\right\}.$$

We now "absorb" $X_{(k)}\Psi_{-k}^{-T}\Psi_{-k}^{-1}X_{(k)}^T$ into L_k. First, take the lower triangular Cholesky decomposition of $X_{(k)}\Psi_{-k}^{-T}\Psi_{-k}^{-1}X_{(k)}^T = \Phi_k\Phi_k^T$ so that

$$\left(X_{(k)}\Psi_{-k}^{-T}\Psi_{-k}^{-1}X_{(k)}^T\right)^{-1} = \Phi_k^{-T}\Phi_k^{-1}.$$

We have

$$\begin{aligned} p(L_k) &\propto \left(\prod_{i=1}^{p_k} L_{k[i,i]}^{i-np/p_k-p_k-1}\right)\exp\left\{-\operatorname{tr}\left((L_kL_k^T)^{-1}\Phi_k\Phi_k^T\right)/2\right\} \\ &\propto \left(\prod_{i=1}^{p_k} L_{k[i,i]}^{i-np/p_k-p_k-1}\right)\exp\left\{-\operatorname{tr}\left((\Phi_k^{-1}L_k(\Phi_k^{-1}L_k)^T)^{-1}\right)/2\right\}. \end{aligned}$$

Now let $W_k = \Phi_k^{-1}L_k$, so that $L_k = \Phi_kW_k$. This change of variables has Jacobian $J(W_k) = \prod_{i=1}^{p_k}\Phi_{k[i,i]}^i$ [Eaton, 1983, Proposition 5.13], so that

$$p(W_k) \propto \left(\prod_{i=1}^{p_k} W_{k[i,i]}^{i-np/p_k-p_k-1}\right)\exp\left\{-\operatorname{tr}\left((W_kW_k^T)^{-1}\right)/2\right\}. \tag{3.8}$$

Note that the distribution of W_k does not depend on Ψ_{-k}. Now compare equation (3.8) to the density of the lower triangular Cholesky square root W of an inverse-Wishart distributed random matrix

$$WW^T \sim \text{inverse-Wishart}_{p_k}\left(np/p_k, I_{p_k}\right)$$

, given by

$$p(W) \propto \left(\prod_{i=1}^{p_k} W_{[i,i]}^{-np/p_k - i} \right) \exp\left\{ -\operatorname{tr}((WW^T)^{-1})/2 \right\}. \tag{3.9}$$

The conditional densities of the off-diagonal elements of W_k and W given the diagonal elements clearly have the same form. The diagonal elements of W_k and W in (3.8) and (3.9) are square roots of inverse-gamma distributed random variables, but with different shape parameters. To show this, we first derive the conditional densities of the off-diagonal elements of W:

Lemma 3. *(Bartlett's decomposition for the inverse-Wishart) Let W be the lower triangular Cholesky square root of an inverse-Wishart distributed matrix, so*

$$WW^T \sim \textit{inverse-Wishart}_{p_k}(\nu, I_{p_k}).$$

Then for each $i = 1, \ldots, p_k$,

$$\begin{aligned} W_{[i,i]}^2 \sim\ & \textit{inverse-gamma}([\nu - p_k + i]/2, 1/2), \ \textit{and} \\ & W_{[i,1:(i-1)]} | W_{[i,i]}, W_{[1:(i-1),1:(i-1)]} \\ & \sim N_{i-1}\left(0, W_{[i,i]}^2 W_{[1:(i-1),1:(i-1)]}^T W_{[1:(i-1),1:(i-1)]}\right). \end{aligned}$$

Here, $W_{[1:(i-1),1:(i-1)]}$ denotes the submatrix of W made up of the first $(i-1)$ rows and columns, and $W_{[i,1:(i-1)]}$ is the vector made up of the first $(i-1)$ elements of the ith row.

By Lemma 3, if $WW^T \sim$ inverse-Wishart$(np/p_k, I_{p_k})$ then the squared diagonal elements of W are independent inverse-gamma$((np/p_k - p_k + i)/2, 1/2)$ random variables. This tells us that

$$\int \exp\left\{ -\operatorname{tr}((WW^T)^{-1})/2 \right\} \prod_{i>j} dW_{[i,j]} \ \propto \ \prod_{i=1}^{p_k} W_{[i,i]}^{p_k - 1} \exp\left\{ -1/(2W_{[i,i]}^2) \right\}.$$

This result allows us to integrate (3.8) with respect to the off-diagonal elements of W_k, giving

$$\int \left(\prod_{i=1}^{p_k} W_{k[i,i]}^{i - np/p_k - p_k - 1} \right) \exp\left\{ -\operatorname{tr}\left((W_k W_k^T)^{-1} \right)/2 \right\} \prod_{i>j} dW_{[i,j]}$$

$$\propto W_{k[i,i]}^{i-np/p_k-2} \exp\left\{-1/(2W_{k[i,i]}^2)\right\}.$$

A change of variables implies that the $W_{k[i,i]}^2$'s are independent, and

$$W_{k[i,i]}^2 \sim \text{ inverse-gamma}([np/p_k - i + 1]/2, 1/2). \tag{3.10}$$

This completes the characterization of the distribution of W_k: The distribution of the diagonal elements is given by (3.10) and the conditional distribution of the off-diagonal elements given the diagonal can be obtained from Lemma 3. Finally, this distribution can be related to a Wishart distribution via the following lemma:

Lemma 4. *Let W_k be a random $p_k \times p_k$ lower triangular matrix such that*

$$\begin{aligned}
&W_{k[i,i]}^2 \sim \text{ inverse-gamma}\left([\nu - i + 1]/2, 1/2\right), \text{ and} \\
&W_{k[i,1:(i-1)]} | W_{k[1:(i-1),1:(i-1)]}, W_{k[i,i]} \\
&\sim N_{i-1}\left(0, W_{k[i,i]}^2 W_{k[1:(i-1),1:(i-1)]}^T W_{k[1:(i-1),1:(i-1)]}\right).
\end{aligned}$$

Then the elements of $V_k = W_k^{-1}$ are distributed independently as

$$\begin{aligned}
&V_{k[i,i]}^2 \sim \text{gamma}([\nu - i + 1]/2, 1/2), \ i = 1, \ldots, q \\
&V_{k[i,j]} \sim N(0,1), \ i \neq j.
\end{aligned}$$

Note that the matrix V_k is distributed as the lower triangular Cholesky square root of a Wishart distributed random matrix. Applying the lemma to W_k, for which $\nu = np/p_k$, we have that $V_k = W_k^{-1} = (\Phi_k^{-1} L_k)^{-1} = L_k^{-1}\Phi_k = \frac{1}{\sigma}\Psi_k^{-1}\Phi_k$ is equal in distribution to the lower triangular Cholesky square root of a random matrix which is $\text{Wishart}_{p_k}(np/p_k, I_{p_k})$. That is, the precision matrix $(\sigma^2 \Psi_k \Psi_k^T)^{-1} = \Psi_k^{-T}\Psi_k^{-1}/\sigma^2$ is conditionally distributed as

$$\begin{aligned}
&\frac{1}{\sigma^2}\Psi_k^{-T}\Psi_k^{-1} | \Psi_{-k} \stackrel{d}{=} \Phi_k^{-T} V^T V \Phi_k^{-1}, \text{ where} \\
&VV^T \sim \text{Wishart}_{p_k}(np/p_k, I_{p_k}) \text{ and } \Phi_k \Phi_k^T = X_{(k)} \Psi_{-k}^{-T} \Psi_{-k}^{-1} X_{(k)}^T.
\end{aligned}$$

We say the matrix, $\Phi_k^{-T} V^T V \Phi_k^{-1}$ has a *mirror-Wishart* distribution because $\Phi_k^{-T} V V^T \Phi_k^{-1}$ would have a Wishart distribution. This completes the derivation of the full conditional distribution of $\sigma^2 \Sigma_k = \sigma^2 \Psi_k \Psi_k^T$.

Although not necessary for posterior approximation, the full conditional distribution of σ given $\Psi_1, \ldots, \Psi_K$ and X is easy to derive. The posterior density is

$$p(\sigma) \propto \sigma^{-(np+1)} \exp\left\{ -\left|\left|(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X\right|\right|^2 / (2\sigma^2) \right\}.$$

Letting $\gamma = 1/\sigma^2$, we have

$$p(\gamma) \propto \gamma^{np/2-1} \exp\left\{ -\gamma \left|\left|(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X\right|\right|^2 / 2 \right\},$$

and so the full conditional distribution of $1/\sigma^2$ is

$$\text{gamma}(np/2, ||(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X||^2/2).$$

3.4 Estimation under multiway Stein's loss

3.4.1 The UMREE for multiway Stein's loss

A commonly used loss function for estimation of a covariance matrix Σ is Stein's loss,

$$L_S(S, \Sigma) = \text{tr}\left(S\Sigma^{-1}\right) - \log\left|S\Sigma^{-1}\right| - p, \quad \Sigma, S \in \mathcal{S}_p^+.$$

First introduced by James and Stein [1961], Stein's loss has been proposed as a reasonable and perhaps better alternative to quadratic loss for evaluating performance of covariance estimators. For example, Stein's loss, unlike quadratic loss, does not penalize overestimation of the variances more severely than underestimation.

Recall from Section 3.2 that the array normal model can be parameterized in terms of $\Sigma = (\sigma^2, \Sigma_1, \ldots, \Sigma_K) \in \mathcal{S}_{\mathbf{p}}^+$, where $|\Sigma_k| = 1$ for each $k = 1, \ldots, K$. For estimation of the covariance parameters $\Sigma \in \mathcal{S}_{\mathbf{p}}^+$, we consider the following generalization of Stein's loss, which we call "multiway Stein's loss":

$$L_M(\Sigma, S) = \frac{s^2}{\sigma^2} \sum_{k=1}^{K} \frac{p}{p_k} \text{tr}\left[S_k \Sigma_k^{-1}\right] - Kp \log\left(\frac{s^2}{\sigma^2}\right) - Kp, \quad \Sigma, S \in \mathcal{S}_{\mathbf{p}}^+. \tag{3.11}$$

It is easy to see that for $K = 1$, multiway Stein's loss reduces to Stein's loss. Multiway Stein's loss also has the attractive property of being invariant under multilinear transformations.

To see this, define $SL_{\mathbf{p}}$ to be the set of lists of the form $A = (a, A_1, \ldots, A_K)$ for which $a > 0$ and $A_k \in SL_{p_k}$ for each k, with SL_{p_k} being the special linear group of $p_k \times p_k$ matrices with unit determinant. For two elements A and B of $SL_{\mathbf{p}}$, define $AB = (ab, A_1B_1, \ldots, A_KB_K)$ and $A^T = (a, A_1^T, \ldots, A_K^T)$. Multiway Stein's loss is invariant under transformations of the form $\Sigma \to A\Sigma A^T$, as

$$\begin{aligned}
&L_M\left(A\Sigma A^T, ASA^T\right) \\
&= \frac{a^2s^2}{a^2\sigma^2}\sum_{k=1}^{K}\frac{p}{p_k}\operatorname{tr}\left[A_kS_kA_k^T\left(A_k\Sigma_kA_k^T\right)^{-1}\right] - Kp\log\left(\frac{a^2s^2}{a^2\sigma^2}\right) - Kp \\
&= \frac{s^2}{\sigma^2}\sum_{k=1}^{K}\frac{p}{p_k}\operatorname{tr}\left[S_k\Sigma_k^{-1}\right] - Kp\log\left(\frac{s^2}{\sigma^2}\right) - Kp = L_M\left(\Sigma, S\right).
\end{aligned}$$

Notably, (3.11) is invariant under $\mathcal{G}_{\mathbf{p}}^+$, as $\mathcal{G}_{\mathbf{p}}^+ \subset SL_{\mathbf{p}}$, so the best $\mathcal{G}_{\mathbf{p}}^+$-equivariant estimator under multiway Stein's loss can be found using Corollary 1.

Proposition 1. *(UMREE under multiway Stein's loss) Let*

$$\mathcal{E}_k = \left(E\left[\left(\sigma^2\Sigma_k\right)^{-1}\middle| X\right]\right)^{-1},$$

where the expectation is with respect to the posterior distribution given by Equation 3.7. The minimizer of the posterior expectation

$$E\left[\frac{s^2}{\sigma^2}\sum_{k=1}^{K}\frac{p}{p_k}\operatorname{tr}\left[S_k^T\Sigma_k^{-1}\right] - Kp\log\left(\frac{s^2}{\sigma^2}\right) - Kp\middle| X\right]$$

with respect to s and the S_k's is

$$\hat{\Sigma}_k = \mathcal{E}_k/|\mathcal{E}_k|^{1/(p_k)} \text{ and } \hat{\sigma}^2 = \left(\sum_{k=1}^{K}\frac{1}{K}|\mathcal{E}_k|^{-1/p_k}\right)^{-1}.$$

The posterior expectation $E[(\sigma^2\Sigma_k)^{-1}|X]$ may be approximated by the Gibbs sampler of Section 3.3. That is, if $(\sigma^2\Sigma_k)^{(1)}, \ldots, (\sigma^2\Sigma_k)^{(T)}$ is a long sequence of values of $(\sigma^2\Sigma_k)$ simulated from the Gibbs sampler, then

$$\mathrm{E}[(\sigma^2\Sigma_k)^{-1}|X] \approx \sum_{t=1}^{T}[(\sigma^2\Sigma_k)^{(t)}]^{-1}/T.$$

The form of multiway Stein's loss (3.11) includes a weighted sum of $\text{tr}(S_k\Sigma_k^{-1})$, $k = 1, \ldots, K$. We note that equivariant estimation of Σ is largely unaffected by changes to the weights in this sum:

Proposition 2. *Define weighted multiway Stein's loss as*

$$L_W(\Sigma, S) = \frac{s^2}{\sigma^2} \sum_{k=1}^{K} \frac{w_k}{p_k} \text{tr}\left[S_k\Sigma_k^{-1}\right] - \left(\sum_{k=1}^{K} w_k\right) \log\left(\frac{s^2}{\sigma^2}\right) - \sum_{k=1}^{K} w_k,$$

for known $w_k > 0$, $k = 1, \ldots, K$. Then the UMREE under L_W is given by

$$\hat{\Sigma}_k = \mathcal{E}_k/|\mathcal{E}_k|^{1/(p_k)} \text{ and } \hat{\sigma}^2 = \left(\sum_{k=1}^{K} \frac{w_k}{\sum_{i=1}^{K} w_i} |\mathcal{E}_k|^{-1/p_k}\right)^{-1}.$$

The proof is very similar to that of Proposition 1 and is omitted. This proposition states that only estimation of the scale is affected when we "weight" the loss more heavily for some components of Σ than others.

The posterior distribution may also be used to obtain the UMREE under Stein's original loss L_S, as it too is invariant under transformations of the lower triangular product group. However, risk minimization with respect to L_S requires additional numerical approximations: Let $\mathcal{K}$ be the unique symmetric square root of $E[(\Sigma_K^{-1} \otimes \cdots \otimes \Sigma_1^{-1})/\sigma^2|X]$, which may be approximated by the Gibbs sampler described in Section 3.3. Minimization of the risk with respect to L_S is equivalent to the minimization in $(s^2, S_1, \ldots, S_K)$ of

$$\begin{aligned} E[L_S(S, \Sigma)|X] &= s^2 \,\text{tr}\left(\mathcal{K}\left(S_K \otimes \cdots \otimes S_1\right)\mathcal{K}\right) - p\log\left(s^2\right) + c(\Sigma) \\ &= s^2 \left|\left|(S_1^{1/2}, \ldots, S_K^{1/2}, I_p) \cdot \tilde{\mathcal{K}}\right|\right|^2 - p\log\left(s^2\right) + c(\Sigma) \\ &= \text{tr}\left(s^2 S_k \tilde{\mathcal{K}}_{(k)} S_{-k} \tilde{\mathcal{K}}_{(k)}^T\right) - p\log\left(|s^2 S_k|\right)/p_k + c(\Sigma), \end{aligned}$$

where $\tilde{\mathcal{K}} \in \mathbb{R}^{p_1 \times \cdots \times p_K \times p}$ is the array such that $\tilde{\mathcal{K}}_{(K+1)} = \mathcal{K}$, and $S_k^{1/2}$ is any square root matrix of S_k. Iteratively setting $s^2 S_k = (\tilde{\mathcal{K}}_{(k)} S_{-k} \tilde{\mathcal{K}}_{(k)}^T)^{-1} p/p_k$ will decrease the posterior expected loss at each step. This procedure is analogous to using the iterative flip-flop algorithm to find the MLE based on a sample covariance matrix of $E[(\Sigma_K^{-1} \otimes \cdots \otimes \Sigma_1^{-1})/\sigma^2|X]$. Application

of the results from [Wiesel, 2012b] show that the posterior risk has a property known as geodesic convexity, implying that any local minimizer obtained from this algorithm will also be a global minimizer.

3.4.2 An orthogonally equivariant estimator

The estimator in Proposition 1 depends on the ordering of the indices, and so it is not permutation equivariant. Mirroring the ideas studied in Takemura [1984], in this section we derive a minimax orthogonally equivariant estimator (which is necessarily permutation equivariant) that dominates the UMREE of Proposition 1. First, notice that by transforming the data and then back-transforming the estimator, we can obtain an estimator whose risk is equal to that of the UMREE: For $\Gamma = (1, \Gamma_1, \dots, \Gamma_K) \in \{1\} \times \mathcal{O}_{p_1} \times \cdots \times \mathcal{O}_{p_K}$, where $\mathcal{O}_{p_k}$ is the group of p_k by p_k orthogonal matrices, let $\tilde{X} = (\Gamma_1, \dots, \Gamma_K) \cdot X$. Then $\hat{\Sigma}(\tilde{X})$ is an estimator of $\Gamma\Sigma\Gamma^T$ and $\tilde{\Sigma}(X) = \Gamma^T\hat{\Sigma}(\tilde{X})\Gamma$ is an estimator of Σ. The risk of this estimator is the same as that of the UMREE $\hat{\Sigma}(X)$:

$$\begin{aligned} R\left(\Sigma, \tilde{\Sigma}(X)\right) &= E\left[L_M\left(\Sigma, \Gamma^T\hat{\Sigma}(\tilde{X})\Gamma\right)\middle|\,\Sigma\right] \\ &= E\left[L_M\left(\Gamma\Sigma\Gamma^T, \hat{\Sigma}(\tilde{X})\right)\middle|\,\Sigma\right] \\ &= E\left[L_M\left(\Gamma\Sigma\Gamma^T, \hat{\Sigma}(X)\right)\middle|\,\Gamma\Sigma\Gamma^T\right] \\ &= R\left(\Gamma\Sigma\Gamma^T, \hat{\Sigma}(X)\right) \\ &= R\left(\Sigma, \hat{\Sigma}(X)\right) \end{aligned}$$

where the second equality follows from the invariance of the loss, the third equality follows from a change of variables, and the last equality follows because the risk of $\hat{\Sigma}$ is constant over the parameter space. The UMREE $\hat{\Sigma}$ and the estimator $\tilde{\Sigma}$ have the same risks but are different. Since multiway Stein's loss is convex in each argument, averaging these estimators somehow should produce a new estimator that dominates them both.

In the multivariate normal case in which $K = 1$, averaging the value of $\Gamma^T\hat{\Sigma}(\Gamma X)\Gamma$ with respect to the uniform (invariant) measure for Γ over the orthogonal group results in

the estimator of Takemura [1984]. This estimator is orthogonally equivariant, dominates the UMREE and is therefore also minimax. Constructing an analogous estimator in the multiway case is more complicated, as it is not immediately clear how the back-transformed estimators should be averaged. Direct numerical averaging of estimates of $\sigma^2(\Sigma_1 \otimes \cdots \otimes \Sigma_K)$ will generally produce an estimate that is not separable and therefore outside of the parameter space. Similarly, averaging estimates of each Σ_k separately will not work, as the space of covariance matrices with determinant one is not convex.

Our solution to this problem is to average a transformed version of $\Sigma = (\sigma^2, \Sigma_1, \ldots, \Sigma_K)$ for which each Σ_k lies in the convex set of trace-1 covariance matrices, then transform back to our original parameter space. The resulting estimator, which we call the multiway Takemura estimator (MWTE), is orthogonally equivariant and uniformly dominates the UMREE.

Proposition 3. *Let $\hat{\sigma}^2(\Gamma, X)$ and $\hat{\Sigma}_k(\Gamma, X)$ be the UMREEs of σ^2 and $\Gamma_k \Sigma_k \Gamma_k^T$ based on data $(\Gamma_1, \ldots, \Gamma_K, I_n) \cdot X$. Let*

$$S_k(X) = \int_{\mathcal{O}_{p_K}} \cdots \int_{\mathcal{O}_{p_1}} \frac{\Gamma_k^T \hat{\Sigma}_k(\Gamma, X) \Gamma_k}{\operatorname{tr}\left(\hat{\Sigma}_k(\Gamma, X)\right)} \, d\Gamma_1 \cdots d\Gamma_K$$

and

$$\tilde{\sigma}^2(X) = \int_{\mathcal{O}_{p_K}} \cdots \int_{\mathcal{O}_{p_1}} \hat{\sigma}^2(\Gamma, X) \; d\Gamma_1 \cdots d\Gamma_K.$$

Let $\tilde{\Sigma}_k(X) = S_k(X)/|S_k(X)|^{1/p_k}$ for $k = 1, \ldots, K$. Then $(\tilde{\sigma}^2(X), \tilde{\Sigma}_1(X), \ldots, \tilde{\Sigma}_K(X))$ is orthogonally equivariant and uniformly dominates the UMREE of Proposition 1.

Note that "averaging" over any subset of $\mathcal{O}_{p_1} \times \cdots \times \mathcal{O}_{p_K}$ in the manner of Proposition 3 will uniformly decrease the risk. By averaging with respect to the uniform measure over the orthogonal group, we obtain an estimator that has the attractive property of being orthogonally equivariant.

In practice it is computationally infeasible to integrate over the space of orthogonal matrices. However, we may obtain a stochastic approximation to the MWTE as follows:

Independently for each $t = 1, \ldots, T$ and $k = 1, \ldots, K$, simulate $\Gamma_k^{(t)}$ from the uniform distribution on $\mathcal{O}_{p_k}$. Let

$$S_k(X) = \frac{1}{T}\sum_{t=1}^{T} \frac{{\Gamma_k^{(t)}}^T \hat{\Sigma}_k\left(\Gamma^{(t)}, X\right)\Gamma_k^{(t)}}{\mathrm{tr}\left(\hat{\Sigma}_k\left(\Gamma^{(t)}, X\right)\right)}\ ,\ \tilde{\sigma}_T(X) = \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}(\Gamma^{(t)}, X).$$

Set $\tilde{\Sigma}_{k,T}(X) = S_k(X)/|S_k(X)|^{1/p_k}$ for $k = 1, \ldots, K$. Then an approximation to the MWTE is

$$\tilde{\Sigma}_T = \left(\tilde{\sigma}_T^2(X), \tilde{\Sigma}_{1,T}(X), \ldots, \tilde{\Sigma}_{K,T}(X)\right). \tag{3.12}$$

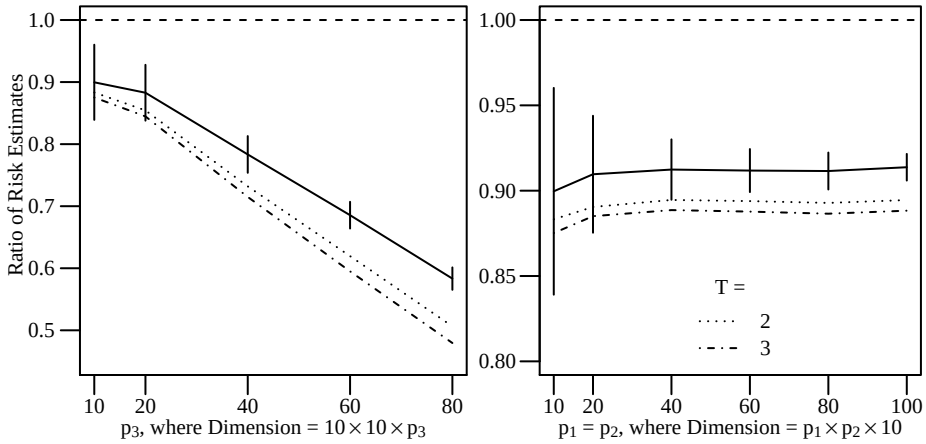
This is a randomized estimator which is orthogonally invariant in the sense of Definition 6.3 of Eaton [1989].

3.4.3 Simulation results

We numerically compared the risks of the MLE, UMREE, and the MWTE under several three-way array normal distributions, using a variety of values of (p_1, p_2, p_3) and with $n = 1$. For each (p_1, p_2, p_3) under consideration, we simulated 100 data arrays from the array normal model. As the risk of both the MLE and the UMREE are constant over the parameter space, it is sufficient to compare their risks at a single point in the parameter space, which we took to be $\Sigma = (1, I_{p_1}, I_{p_2}, I_{p_3})$. Risks were approximated by averaging the losses of each estimator across the 100 simulated data arrays. For each data array, the MLE was obtained from the iterative coordinate descent algorithm outlined in [Hoff, 2011]. Each UMREE was approximated based on 1250 iterations of the Gibbs sampler described in Section 3.3, from which the first 250 iterations were discarded to allow for convergence to the stationary distribution (convergence appeared to be essentially immediate).

The ratio of risk estimates across several values of (p_1, p_2, p_3) are are plotted in solid lines in Figure 3.1. We considered array dimensions in which the first two dimensions were identical. This scenario could correspond to, for example, data arrays representing longitudinal relational or network measurements between $p_1 = p_2$ nodes at p_3 time points. The

Figure 3.1: Risk comparisons for the MLE, UMREE and MWTE. Both panels plot Monte Carlo estimates of the risk ratios of the UMREE to the MLE in solid lines, and the approximate MWTE to the MLE in dashed lines. The width of the vertical bars is one standard deviation of the ratio of the UMREE loss to the MLE loss, across the 100 data sets.



first panel of the figure considers the relative performance of the estimators as the "number of time points" (p_3) increases. The results indicate that the UMREE provides substantial and increasing risk improvements compared to the MLE as p_3 increases. However, the right panel indicates that the gains are not as dramatic and not increasing when the "number of nodes" ($p_1 = p_2$) increases while p_3 remains fixed. Even so, the variability in the ratio of losses (shown with vertical bars) decreases as the number of nodes increases, indicating an increasing probability that the UMREE will beat the MLE in terms of loss.

We also compared these risks to the risk of the approximate MWTE given in (3.12), with $T \in \{2, 3\}$. The risks for the approximate MWTE relative to those of the MLE are shown in dashed lines in the two panels of the Figure, and indicate non-trivial improvements in risk as compared to the UMREE. We examined values of T greater than 3 but found no appreciable further reduction in the risk. Note, however, that the MWTE does not have constant risk over the parameter space (though MWTE will have constant risk over the orbits of the orthogonal product group).

3.5 Discussion

This chapter has extended the results of James and Stein [1961] and Takemura [1984] by developing equivariant and minimax estimators of the covariance parameters in the array normal model. Considering the class of estimators equivariant with respect to a special lower triangular group, we showed that the uniform minimum risk equivariant estimator (UMREE) can be viewed as a generalized Bayes estimator that can be obtained from a simple Gibbs sampler. We obtained an orthogonally equivariant estimator based on this UMREE by combining values of the UMREE under orthogonal transformations of the data. Both the UMREE and the orthogonally equivariant estimator are minimax, and both dominate any unique MLE in terms of risk.

Empirical results in Section 4 indicate that the risk improvements of the UMREE over the MLE can be substantial, while the improvements of the orthogonally equivariant estimator over the UMREE are more modest. However, the risk improvements depend on the array dimensions in a way that is not currently understood. Furthermore, we do not yet know the minimal conditions necessary for the propriety of the posterior or the existence of the UMREE. Empirical results from the simulations in Section 4 suggest that the UMREE exists for sample sizes as low as $n = 1$, at least for the array dimensions in the study. This is similar to the current state of knowledge for the existence of the MLE: The array normal likelihood is trivially bounded for $n \geq p$ (as it is bounded by the maximized likelihood under the unconstrained p-variate normal model), and some sufficient conditions for uniqueness of the MLE are given in Ohlson et al. [2013]. However, empirical results (not shown) suggest that a unique MLE may exist for $n = 1$ for some array dimensions (although not for others). Obtaining necessary and sufficient conditions for the existence of the UMREE and the MLE is an ongoing area of research of the author.

Chapter 4

ADAPTIVE HIGHER-ORDER SPECTRAL ESTIMATORS

4.1 Introduction

Tensor data arise in fields as diverse as relational data [Hoff, 2014], neuroimaging [Zhang et al., 2014, Li and Zhang, 2015], psychometrics [Kiers and Mechelen, 2001], chemometrics [Smilde et al., 2005, Bro, 2006], signal processing [Cichocki et al., 2014], and machine learning [Tao et al., 2005], among others [Kroonenberg, 2008]. 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 movie can be expressed as a tensor, where element $\mathcal{X}_{[i,j,t]}$ of the tensor is the intensity of pixel (i, j) at frame t (Figure 4.1).

Often, a tensor is corrupted by noise (Figure 4.2). The model we consider for this is:

$$\mathcal{X} = \Theta + \mathcal{E},\ \mathcal{E}_{[i_1,\ldots,i_K]} \sim N(0, \tau^2) \text{ independent for } i_k = 1, \ldots, p_k, \text{ and } k = 1, \ldots, K, \quad (4.1)$$

where $\Theta \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the signal (Figure 4.1) and $\mathcal{E} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is the additive measurement error or noise. 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):

$$\text{MSE}(t(\mathcal{X})) = E_\Theta[||\Theta - t(\mathcal{X})||^2] = \sum_{\mathbf{i}} E_\Theta[(\Theta_{[\mathbf{i}]} - t(\mathcal{X})_{[\mathbf{i}]})^2], \quad (4.2)$$

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. Consider the "Casorati" matrix of a movie, where column t contains the vectorized image of frame t. Since adjacent frames usually contain only minor changes, the columns of this Casorati matrix exhibit strong co-linearity. It would be reasonable, then, to assume that the Casorati matrix is well approximated by a low rank matrix.

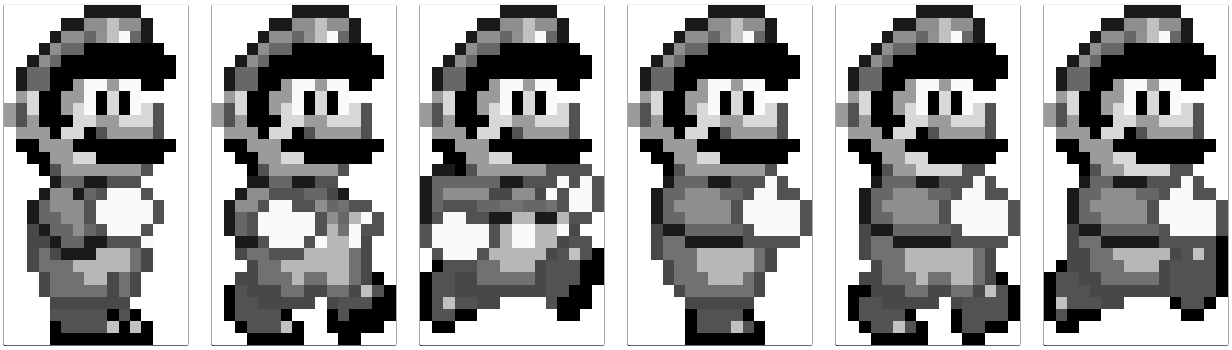


Figure 4.1: Frames of a movie [Nintendo, 1990].

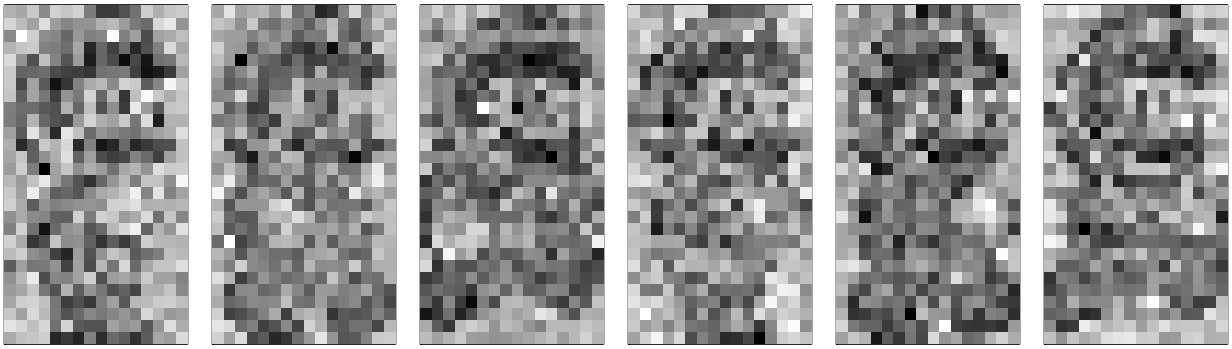


Figure 4.2: Frames of a noisy movie [Nintendo, 1990].

There has been much work on "denoising" (or mean estimation) in matrix variate data by using our knowledge that the mean has approximately low rank. A typical estimation scheme begins by computing the singular value decomposition (SVD) of X:

$$X = UDV^T, \tag{4.3}$$

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 [Shabalin and Nobel, 2013].

Early work on singular value shrinkage estimation from a non-statistical perspective began with Eckart and Young [1936], where they proved that the best rank r approximation to the data matrix $X \in \mathbb{R}^{p \times n}$ is found with the shrinkage function:

$$f_i(\sigma_i) = \sigma_i 1(i \leq r), \tag{4.4}$$

where $1(\cdot)$ is the indicator function. We call (4.4) 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 Stein [1981] can uniformly dominate X in terms of risk. This seminal shrinkage estimator, in the context of matrix estimation, is given by

$$f_i(\sigma_i) = \left(1 - \frac{\lambda}{\sum_{i=1}^p \sigma_i^2}\right) \sigma_i. \tag{4.5}$$

For data that exhibit associations between the rows and/or columns of the mean matrix, the estimator of Efron and Morris [1972a], given by

$$f_i(\sigma_i) = \sigma_i - \frac{\lambda}{\sigma_i}, \tag{4.6}$$

was introduced and results in different amounts of shrinkage for each singular value. Efron and Morris [1976] improved upon this estimator with a generalization of both (4.5) and (4.6), given by

$$f_i(\sigma_i) = \left(1 - \frac{\gamma}{\sum_{i=1}^{p} \sigma_i^2}\right) \sigma_i - \frac{\lambda}{\sigma_i}. \tag{4.7}$$

More recent work has focused on estimators whose functions $f_i(\cdot)$ induce sparsity in the singular values, which may be more appropriate than (4.5), (4.6), and (4.7) in cases where the true signal itself has (approximately) low rank. Motivated by penalized maximum likelihood estimation, the hard-thresholding estimator

$$f_i(\sigma_i) = \sigma_i 1(\sigma_i \geq \lambda) \tag{4.8}$$

and the soft-thresholding estimator

$$f_i(\sigma_i) = (\sigma_i - \lambda)_+ \tag{4.9}$$

were introduced [Candès et al., 2013, for example]. Here, $(y)_+ = \max(y, 0)$ is the "positive part" function. A clever shrinkage function that includes (4.8), (4.9), and a truncated version of (4.6) [Verbanck et al., 2015] as special cases is that of Josse and Sardy [2015]:

$$f_i(\sigma_i) = \sigma_i \left(1 - \frac{\lambda^\gamma}{\sigma_i^\gamma}\right)_+. \tag{4.10}$$

This estimator was inspired by the adaptive LASSO [Zou, 2006]. A variety of other shrinkage estimators have also been developed [Nadakuditi, 2014, Shabalin and Nobel, 2013].

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 example, by constructing the Casorati matrix. For a K-dimensional tensor, such "matricization" destroys the indexing structure along all but one of the dimensions. This may be detrimental

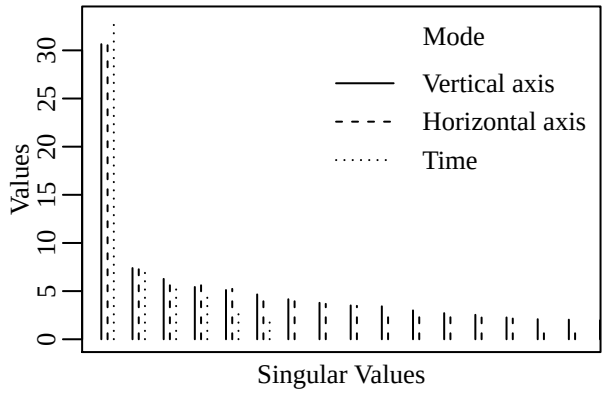


Figure 4.3: Mode specific singular values of the Mario movie of Figure 4.1.

to estimation if, in addition to a data set having approximately low rank, it also has approximately low *multilinear* rank (see Section 4.2), that is, "matricizing" along each index set, or "mode", results in a low rank matrix. For a movie data set, this might occur if horizontally or vertically adjacent pixels tend to contain similar intensities.

Figure 4.3 illustrates an example of this phenomenon. This figure plots the first sixteen singular values for each of the three different matricizations of the Mario tensor of Figure 4.1 (see Section 4.2). If we were only matricizing along the time dimension, as we do when forming the Casorati matrix, we would only observe the dotted lines, which seem to suggest the data have approximately rank 1. However, the vertical and horizontal dimensions of the images also seem to have approximately low rank, as evidenced by the solid and dashed lines. Shrinking the singular values along these additional modes may also 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 De Lathauwer et al. [2000b], 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 4.2, we introduce 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 4.3, we review Stein's unbiased risk estimates (SURE), then derive the SURE for a broad class of higher-order spectral estimators. In Section 4.4 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. We also compare the performance of our estimators with estimators that do not take into account the tensor indexing on the Mario movie in Figure 4.2. In Section 4.5 we illustrate the use of these methods in an application to multivariate relational data. We finish with a discussion in Section 4.6.

4.2 The higher-order SVD and higher-order spectral estimators

Many 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 the K-tuple, $(r_1, \ldots, r_K)$. Define the k-mode matricization [Kolda and Bader, 2009], 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 4.1, has been used to shrink matrix valued data towards low rank. One generalization of the SVD to tensors is the HOSVD of De Lathauwer et al. [2000b], which relates directly to multilinear rank.

Definition 9 (HOSVD of De Lathauwer et al. [2000b]). *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*

$$\mathcal{X} = (U_1, \ldots, U_K) \cdot \mathcal{S} \tag{4.11}$$

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

The product "$\cdot$" in (4.11) 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:

$$\begin{aligned} \mathcal{X} &= (U_1, \ldots, U_K) \cdot (D_1, \ldots, D_K) \cdot (D_1^{-1}, \ldots, D_K^{-1}) \cdot \mathcal{S} \\ &= (U_1, \ldots, U_K) \cdot (D_1, \ldots, D_K) \cdot \mathcal{V}, \end{aligned} \tag{4.12}$$

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 (4.12), we now define higher-order spectral estimators of Θ under the model (4.1).

Definition 10. *Let* $\mathcal{X} = (U_1, \ldots, U_K)\cdot(D_1, \ldots, D_K)\cdot\mathcal{V}$ *as in (4.12) with* $D_k = \mathrm{diag}(\sigma_1^k, \ldots, \sigma_{p_k}^k)$. *An estimator* $t(\mathcal{X})$ *of the form*

$$t(\mathcal{X}) = (U_1, \ldots, U_K) \cdot (f^1(D_1), \ldots, f^K(D_K)) \cdot \mathcal{V}, \tag{4.13}$$

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

Each of the matrix shrinkage functions listed in Section 4.1 (4.4)-(4.10) may, in principle, be applied to each mode in our higher-order spectral estimator (4.13). We focus on two examples of higher-order spectral estimators. One of these is a generalization of the matrix truncation estimator (4.4) and the other is a generalization of the matrix soft-thresholding estimator (4.9). 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 [Timmerman and Kiers, 2000, Kiers and Kinderen, 2003, Ceulemans and Kiers, 2006]. 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 [De Lathauwer et al., 2000b]. Given multilinear rank $(r_1, \ldots, r_K)$, it is found by taking the HOSVD (4.11) 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 (4.13), where

$$f_i^k(\sigma_i^k) = \sigma_i^k 1(i \leq r_k). \tag{4.14}$$

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 (4.9). The second estimator we explore is obtained by applying soft-thresholding to the mode-specific singular values:

$$f_i^k(\sigma_i^k) = (\sigma_i^k - \lambda_k)_+. \tag{4.15}$$

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 (4.15). 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 (4.15) 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 the value of λ_k to the singular values the less dispersed the resulting mode-specific singular values tend to be. To gain intuition

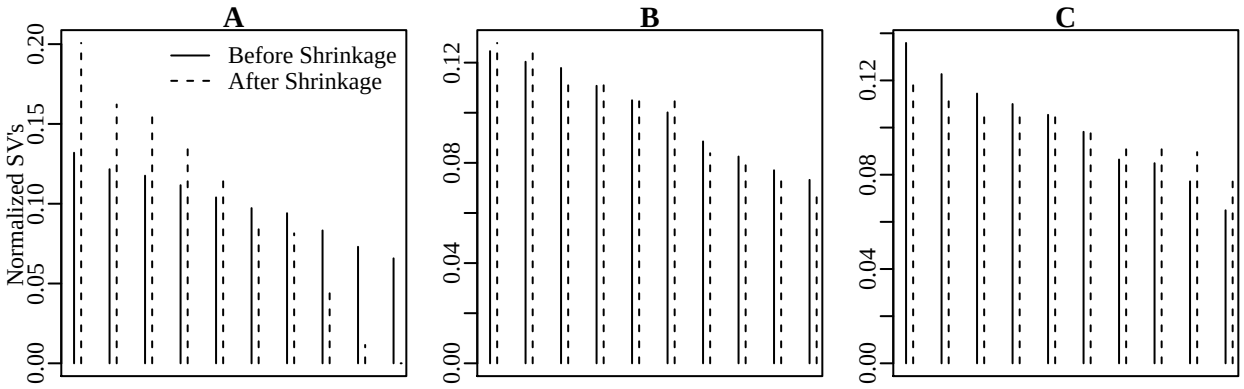


Figure 4.4: Singular values for the three modes, before and after shrinkage, normalized to sum to one.

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 $(947, 873, 844, 801, 746, 698, 675, 597, 524, 472)$. We applied the mode specific soft-thresholding function (4.15) to each mode with $\lambda_1 = 500$, $\lambda_2 = 0$, $\lambda_3 = -10000$. 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 4.4. 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.

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:

$$t(\mathcal{X}) = c\ (U_1, \ldots, U_K) \cdot (f^1(D_1)D_1^{-1}, \ldots, f^K(D_K)D_K^{-1}) \cdot \mathcal{S}, \tag{4.16}$$

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 (4.15).

4.3 Stein's unbiased risk estimate

Both shrinkage function (4.14) and (4.16) define classes of estimators, indexed by tuning parameters. Ideally, we would like to choose these tuning parameters by minimizing the risk (4.2). However, because the mean Θ is unknown, minimization of (4.2) 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 12 (Stein [1981]). *Under the model (4.1), 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*

$$E_\Theta \left[\sum_{\mathbf{i}} \left| \frac{d}{d\mathcal{X}_{[\mathbf{i}]}} t_{\mathbf{i}}(\mathcal{X}_{[\mathbf{i}]}) \right| \right] < \infty. \tag{4.17}$$

Then

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

where $\operatorname{div}(\cdot)$ *is the divergence of* $t(\cdot)$*. We denote Stein's unbiased risk estimate (SURE) as*

$$\mathrm{SURE}(t) = ||t(\mathcal{X}) - \mathcal{X}||^2 + 2\tau^2 \operatorname{div}(t(\mathcal{X})) - p\tau^2. \tag{4.18}$$

"Almost differentiable" basically means differentiable everywhere except on a set of Lebesgue measure zero [Stein, 1981, Definition 1]. Because the SURE (4.18) 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 Candès et al. [2013] in the matrix case.

The difficult part of (4.18) 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 (4.12) in Subsection 4.3.1. Then we will use these differentials to derive the divergence of estimators of the form (4.13) in Subsection 4.3.2. This divergence can then be inserted into (4.18) to obtain the SURE.

4.3.1 Differentials of the HOSVD

In this subsection, we calculate the differentials for the elements in the altered HOSVD (4.12). 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 $\mathbb{R}^{p_1 \times \cdots \times p_K}$ [de Silva and Lim, 2008, Proposition 7.2].

Theorem 13. *The differentials of D_k, U_k, and $\mathcal{V}$ from (4.12) are given in equations (4.19), (4.21), and (4.25), 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 13. 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 Candès et al. [2013] and have a closed form solution, given by

$$d\sigma_i^k[\Delta] = (U_k^T \Delta_{(k)} U_{-k} \mathcal{S}_{(k)} D_k^{-1})_{[i,i]} \text{ for } i = 1, \ldots, p_k \text{ and } k = 1, \ldots, K, \tag{4.19}$$

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 Candès et al. [2013] to get (4.19).

Let $\Omega_{U_k}[\Delta] = U_k^T dU_k[\Delta]$. Then from (4.8) of Candès et al. [2013] we have

$$\begin{aligned}&\Omega_{U_k}[\Delta]_{[i,j]}\\&= -1(i \neq j)\left[\sigma_j^k (U_k^T \Delta_{(k)} U_{-k} S_{(k)}^T D_k^{-1})_{[i,j]} + \sigma_i^k (U_k^T \Delta_{(k)} U_{-k} S_{(k)}^T D_k^{-1})_{[j,i]}\right] / ((\sigma_i^k)^2 - (\sigma_j^k)^2),\end{aligned} \tag{4.20}$$

and so

$$dU_k[\Delta] = U \Omega_{U_k}[\Delta]. \tag{4.21}$$

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 Magnus and Neudecker [1999] for the differential of matrix multiplication and the Kronecker product, we have

$$\begin{aligned}\Delta = d\mathcal{X}[\Delta] &= d(U \cdot D \cdot \mathcal{V})[\Delta] \\ &= \sum_{k=1}^{K} d\underline{U}_k[\Delta] \cdot D \cdot \mathcal{V} + \sum_{k=1}^{K} U \cdot d\underline{D}_k[\Delta] \cdot \mathcal{V} + U \cdot D \cdot d\mathcal{V}[\Delta],\end{aligned} \tag{4.22}$$

where

$$d\underline{U}_k[\Delta] = (U_1, \ldots, U_{k-1}, dU_k[\Delta], U_{k+1}, \ldots, U_K) \text{ and} \tag{4.23}$$

$$d\underline{D}_k[\Delta] = (D_1, \ldots, D_{k-1}, dD_k[\Delta], D_{k+1}, \ldots, D_K). \tag{4.24}$$

From (4.22), we solve for $d\mathcal{V}[\Delta]$ and have

$$d\mathcal{V}[\Delta] = D^{-1} \cdot U^T \cdot \Delta - \sum_{k=1}^{K} dF_k[\Delta] \cdot \mathcal{V} - \sum_{k=1}^{K} dG_k[\Delta] \cdot \mathcal{V}, \tag{4.25}$$

where

$$dF_k[\Delta] = (I_{p_1}, \ldots, I_{p_{k-1}}, D_k^{-1} \Omega_{U_k}[\Delta] D_k, I_{p_{k+1}}, \ldots, I_{p_K}) \text{ and} \tag{4.26}$$

$$dG_k[\Delta] = (I_{p_1}, \ldots, I_{p_{k-1}}, D_k^{-1} dD_k[\Delta], I_{p_{k+1}}, \ldots, I_{p_K}). \tag{4.27}$$

□

4.3.2 Divergence of higher-order spectral estimators

In this section, we show that the divergence of higher-order spectral estimators of the form (4.13) can be found in the following theorem.

Theorem 14. *The divergence of estimators of the form (4.13) is*

$$\text{Sum}\left(f(D)\cdot D^{-1}\cdot\mathcal{C}+\sum_{k=1}^{K}H_k\cdot\mathcal{S}^2\right), \tag{4.28}$$

where $\text{Sum}(\mathcal{A})$ *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* $(\mathcal{S}^2)_{[\mathbf{i}]}=(\mathcal{S}_{[\mathbf{i}]})^2$,

$$H_k=(f^1(D_1)D_1^{-1},\ldots,f^{k-1}(D_{k-1})D_{k-1}^{-1},D_k^{-1}df^k(D_k)D_k^{-1},f^{k+1}(D_{k+1}),\ldots,f^K(D_K)), \tag{4.29}$$

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

$$\mathcal{C}_{[\mathbf{i}]}=1+\sum_{k=1}^{K}\sum_{j=1,j\neq i_k}^{p_k}\frac{\mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma^k_{i_k})^2-(\sigma^k_j)^2}-\mathcal{S}^2_{[\mathbf{i}]}\sum_{k=1}^{K}\left(\frac{1}{(\sigma^k_{i_k})^2}+\sum_{m=1,m\neq i_k}^{p_k}\frac{1}{(\sigma^k_m)^2-(\sigma^k_{i_k})^2}\right). \tag{4.30}$$

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 Candès et al. [2013], also note that $\Delta^{\mathbf{i}}$ forms an orthonormal basis for $\mathbb{R}^{p_1\times\cdots\times p_K}$, and so

$$\text{div}(t(\mathcal{X}))=\sum_{\mathbf{i}}\langle\Delta^{\mathbf{i}},df[\Delta^{\mathbf{i}}]\rangle$$

$$
\begin{aligned}
&= \sum_{\mathbf{i}} \langle (U_1^T, \ldots, U_K^T) \cdot \Delta^{\mathbf{i}}, (U_1^T, \ldots, U_K^T) \cdot df[\Delta^{\mathbf{i}}] \rangle \\
&= \sum_{\mathbf{i}} \langle E^{\mathbf{i}}, (U_1^T, \ldots, U_K^T) \cdot df[\Delta^{\mathbf{i}}] \rangle, \\
&= \sum_{\mathbf{i}} ((U_1^T, \ldots, U_K^T) \cdot df[\Delta^{\mathbf{i}}])_{[\mathbf{i}]}, \qquad (4.31)
\end{aligned}
$$

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

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

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

$$
U^T \cdot df[\Delta^{\mathbf{i}}] = \sum_{k=1}^{K} d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(D) \cdot \mathcal{V} + \sum_{k=1}^{K} df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V} + f(D) \cdot d\mathcal{V}[\Delta^{\mathbf{i}}], \qquad (4.32)
$$

where

$$
d\tilde{U}_k[\Delta^{\mathbf{i}}] = (I_{p_1}, \ldots, I_{p_{k-1}}, \Omega_{U_k}[\Delta^{\mathbf{i}}], I_{p_{k+1}}, \ldots, I_{p_K}). \qquad (4.33)
$$

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 (4.32) and plug that into (4.31). We will first calculate all of the differentials that are in (4.32), then we will determine the $(i_1, \ldots, i_K)$th element of $U^T \cdot df[\Delta^{\mathbf{i}}]$. Then we will simplify (4.31). These latter two steps may be found in Appendix A.3.

We begin with the differentials. From (4.19), we have

$$
\begin{aligned}
d\sigma_j^k[\Delta^{\mathbf{i}}] &= (U_k^T \Delta^{\mathbf{i}}_{(k)} U_{-k} S^T_{(k)} D_k^{-1})_{[j,j]} \\
&= (E^{\mathbf{i}}_{(k)} S^T_{(k)} D_k^{-1})_{[j,j]}
\end{aligned}
$$

$$= 1(j = i_k) S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}/\sigma_j^k. \tag{4.34}$$

This is since $E^{\mathbf{i}}_{(k)} S^T_{(k)} \in \mathbb{R}^{p_k \times p_k}$ such that

$$\left(E^{\mathbf{i}}_{(k)} S^T_{(k)}\right)_{[\ell,j]} = \begin{cases} 0 & \text{if } \ell \neq i_k \\ S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} & \text{if } \ell = i_k. \end{cases} \tag{4.35}$$

Similarly, from (4.20), we have

$$\begin{aligned} &\Omega_{U_k}[\Delta^{\mathbf{i}}]_{[\ell,j]} \\ &= -1(\ell \neq j)\left[\sigma_j^k (U_k^T \Delta_{(k)} U_{-k} S^T_{(k)} D_k^{-1})_{[\ell,j]} + \sigma_\ell^k (U_k^T \Delta_{(k)} U_{-k} S^T_{(k)} D_k^{-1})_{[j,\ell]}\right] / ((\sigma_\ell^k)^2 - (\sigma_j^k)^2) \\ &= -1(\ell \neq j)\left[\sigma_j^k (E^{\mathbf{i}}_{(k)} S^T_{(k)} D_k^{-1})_{[\ell,j]} + \sigma_\ell^k (E^{\mathbf{i}}_{(k)} S^T_{(k)} D_k^{-1})_{[j,\ell]}\right] / ((\sigma_\ell^k)^2 - (\sigma_j^k)^2) \\ &= -1(\ell \neq j)\left[S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} 1(\ell = i_k) + S_{[i_1,\ldots,i_{k-1},\ell,i_{k+1},\ldots,i_K]} 1(j = i_k)\right] / ((\sigma_\ell^k)^2 - (\sigma_j^k)^2). \end{aligned} \tag{4.36}$$

Also, from the chain rule, we have that

$$\begin{aligned} d(f_j^k \circ \sigma_j^k)[\Delta^{\mathbf{i}}] &= \left(\frac{d}{d\sigma_j^k} f_j^k(\sigma_j^k)\right) d\sigma_j^k[\Delta^{\mathbf{i}}] \\ &= \delta_{j,i_k} \left(\frac{d}{d\sigma_j^k} f_j^k(\sigma_j^k)\right) S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}/\sigma_j^k. \end{aligned} \tag{4.37}$$

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 (4.25), (4.31), (4.32), (4.34), (4.36), and (4.37) to calculate a closed-form expression for the divergence. This simplification is relegated to Appendix A.3. □

We now present the formula for the SURE for all higher-order spectral estimators of the form (4.13):

Theorem 15 (SURE for (4.13))**.** *Under the model (4.1), suppose $t(\cdot)$ in (4.13) is almost differentiable and for which (4.17) holds. Then*

$$\text{SURE}(t) = ||t(\mathcal{X}) - \mathcal{X}||^2 + 2\tau^2 \,\text{Sum}\left(f(D) \cdot D^{-1} \cdot \mathcal{C} + \sum_{k=1}^{K} H_k \cdot \mathcal{S}^2\right) - p\tau^2. \tag{4.38}$$

This SURE formula is applicable for all shrinkage functions of the form (4.13) where $f^k(D_k) = \text{diag}(f_1^k(\sigma_1^k), \ldots, f_{p_k}^k(\sigma_{p_k}^k))$. 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 (4.5) or (4.7). For such estimators, we prove in Appendix A.5 that the form of the divergence is very similar as in (4.28). 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 $\text{diag}(D_k) \mapsto \text{diag}(f(D_k))$.

4.4 Simulation studies

In this section, we consider four competitors to the mode-specific soft-thresholding estimator (4.16) and the truncated HOSVD (4.14). 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 (4.5) [Stein, 1981]. This estimator may be derived from an empirical Bayes argument where $\Theta_{[\mathbf{i}]} \sim N(0, \gamma^2)$ [Efron and Morris, 1972b]. As such, it should perform well when the entries of Θ are centered about 0. For a matrix parameter Θ, Efron and Morris [1972a] 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 (4.6) to the first mode matricization of the data tensor. However, the Efron-Morris estimator does not induce low rank estimates, and so our fourth and final competitor is the matrix soft-thresholding estimator (4.9) applied to the first mode matricization of $\mathcal{X}$, and whose tuning parameter is chosen with the SURE formula from Candès et al. [2013]. This estimator should improve on the Efron-Morris estimator when $\Theta_{(1)}$ has approximately low rank.

We now describe the design of the simulation study. We evaluated the risk of the mode-

specific soft-thresholding, truncated HOSVD, 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 six 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 4.5.

The James-Stein estimator (4.5) 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 4.5 (**A**), the James-Stein estimator does perform best, but the 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 (4.9) 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 4.5 (**B**), the matrix soft-thresholding estimator does perform best, but the mode-specific soft-thresholding estimator does surprisingly well,

and performs just below that of the matrix soft-thresholding estimator.

For Scenario **C**, we expect the matrix soft-thresholding estimator (4.9) and the Efron-Morris estimator (4.6) to perform well. There is temporal correlation along one of the modes of the mean tensor. This scenario corresponds to the ideas behind forming the Casorati matrix — we take into account the temporal correlation of the mean by performing soft-thresholding along this mode. However, from Figure 4.5 (**C**), we see that the mode-specific soft-thresholding estimator performed best.

The matrix soft-thresholding estimator (4.9) 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 4.5 (**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 (4.16) to do well, as the mean tensor has approximately low multilinear rank, but it is not exactly low multilinear rank. Figure 4.5 (**E**) reveals the the mode-specific soft-thresholding estimator does indeed perform better than the other estimators.

We expect the truncated HOSVD (4.14) to do well in Scenario **F** because the mean tensor has low multilinear rank, and the truncated HOSVD is correctly shrinking toward this structure. From Figure 4.5 (**F**), we see that the truncated HOSVD does indeed perform best in terms of loss. However, 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

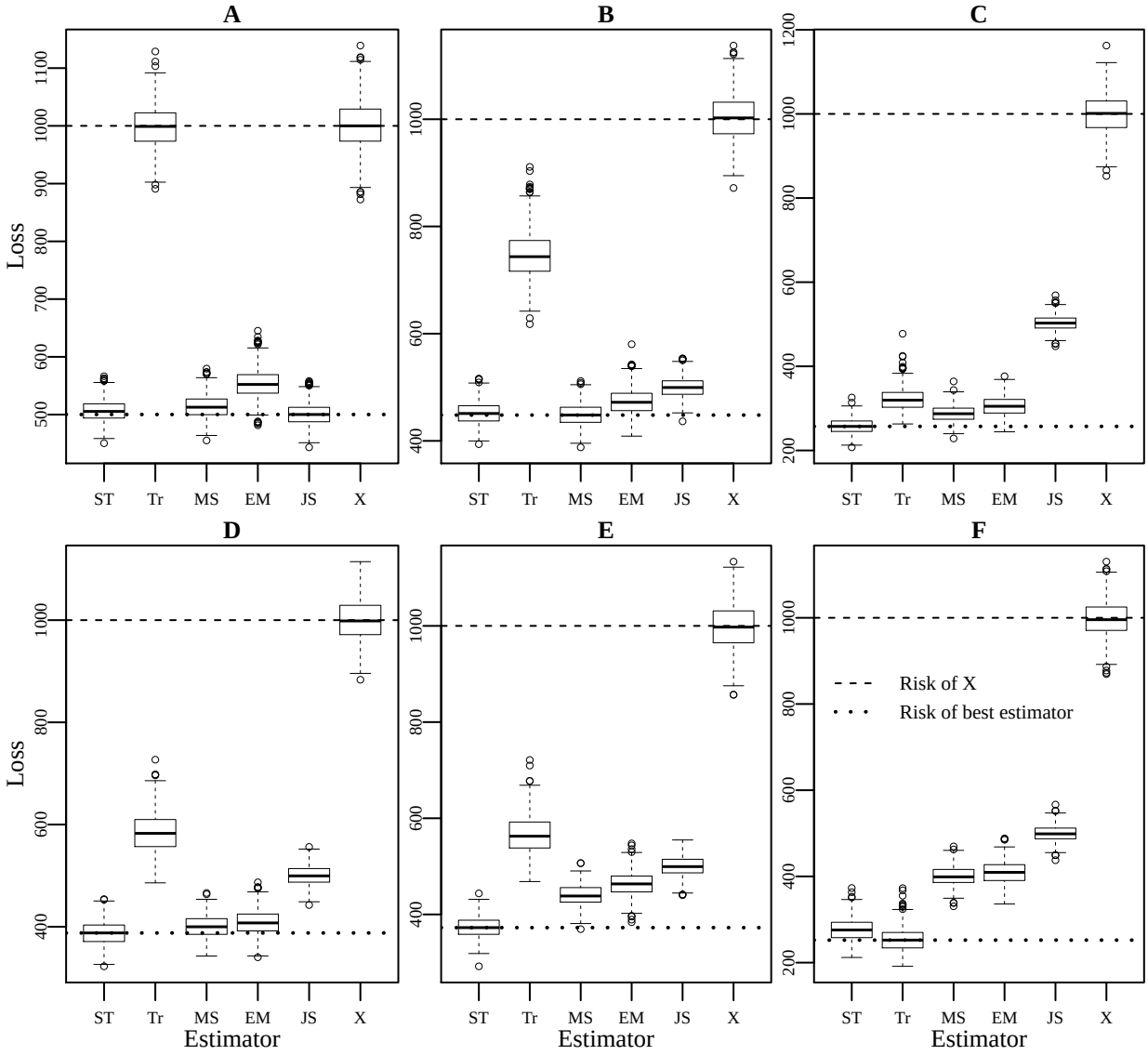


Figure 4.5: Box plots of losses for the six estimators under different scenarios. The estimators include the mode-specific soft-thresholding (ST), truncated HOSVD (Tr), matrix soft-thresholding (MS), Efron-Morris (EM), James-Stein (JS), and maximum likelihood (X) estimators. In the scenarios, the mean 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.

rank of Θ. We evaluated this possibility in simulation Scenarios **D** and **F**. In Scenario **F**, where the tensor had dimension $(10, 10, 10)$ and the true multilinear rank was $(5, 5, 5)$, this SURE method correctly estimated the multilinear rank in 94.4% of trials. In Scenario **D**, where the true multilinear rank was $(5, 10, 10)$, the results of the simulation study can be found in Table 4.1. There, we see that the rank of the first mode is correctly estimated in 96% of trials. The rank of the second and third modes are correctly estimated a majority of the time.

Estimated Rank	4	5	6	7	8	9	10
Mode 1	.04	.96	0	0	0	0	0
Mode 2	0	.01	.02	.05	.12	.25	.56
Mode 3	0	0	0	.02	.06	.20	.71

Table 4.1: Proportion of times each rank is estimated based on SURE for each mode over 500 repetitions when the true multilinear rank is $(5, 10, 10)$.

For the Mario movie in Figure 4.2, we calculated each estimator described in Section 4.4. We compared each estimator with the true movie in Figure 4.1 under squared error loss over 1000 repetitions. A box plot of these losses is in Figure 4.6. The mode-specific soft-thresholding estimator (4.16) performs best among this set of estimators, having a MSE 81% that of the matrix soft-thresholding estimator chosen with SURE from Candès et al. [2013].

4.5 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

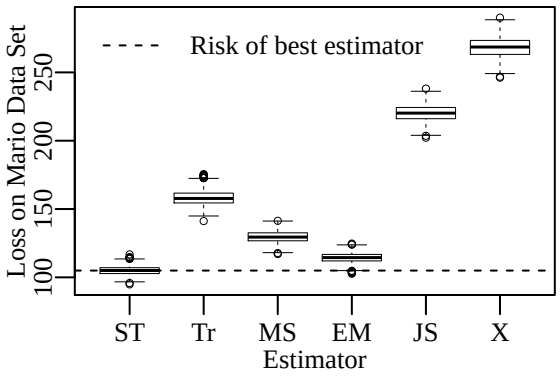


Figure 4.6: Box plot of losses on the Mario data set for each estimator. Abbreviations are the same as in Figure 4.5.

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 Efron and Morris [1975], 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 (4.1).

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 Stein [1966] and Efron and Morris [1972a] 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 $2 \times 2 \times 1 \times 1$. 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 4.2. 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 4.2: Squared error losses when predicting the statistics of the remaining games of the season.

4.6 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 [Signoretto et al., 2010, Tomioka et al., 2011a,b, Liu et al., 2013, Tomioka and Suzuki, 2013]. 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. Our estimators can be written in closed form (4.13) and are not the solution of a minimization problem. While both our class of estimators and the class of penalized optimization estimators contain tuning parameters, the penalized optimization

approaches are not totally adaptive as they do not provide methods to select the tuning parameters. 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 (4.14) and the mode-specific soft-thresholding (4.16) estimators, but also for *all* estimators of the form (4.13) that satisfy the conditions of Theorem 12.

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 [Efron and Morris, 1972a, Stein, 1981], cross-validation [Bro et al., 2008, Owen and Perry, 2009, Josse and Husson, 2012], and asymptotic considerations [Gavish and Donoho, 2014a,b]. Exploring these methods for our higher-order spectral estimators (4.13) is a current research area of the author.

In this paper, we focused on estimators of the form (4.13). 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 (4.13) 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 the Appendix A.6, 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 [Henrion, 1993, Kiers et al., 1997, Murakami et al., 1998, Andersson and Henrion, 1999, De Lathauwer et al., 2001, Martin and Van Loan, 2008].

Although the error variance τ^2 in (4.1) might be known in some settings, such as fMRI data sets [Candès et al., 2013], in most applied situations the variance would not be unknown.

Instead of plugging in an estimate of the variance into the SURE formula (4.18), there has been a recent suggestion to use a generalized SURE formula [Sardy, 2012, Josse and Sardy, 2015]:

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

This formula is motivated by generalized cross-validation [Golub et al., 1979] and is an approximation to SURE [Josse and Sardy, 2015]. 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. The code available on the author's web page allows for GSURE implementation for the estimators discussed in this article.

Chapter 5

DISCUSSION

In this thesis, we have developed novel covariance and mean estimators for tensor-variate data. In Chapter 2, we linked likelihood inference in the array normal model with novel tensor decompositions that are analogues to common matrix decompositions. This also resulted in a simple form for the likelihood ratio test statistic for testing hypotheses within a Kronecker structured covariance model framework. Using the history of covariance estimation in the multivariate normal model as a guide, in Chapter 3 we then developed improved covariance estimators over the MLE's in the array normal model. These estimators were derived using the classical theory of equivariance. In order to derive these estimators, we used a generalized Bayes approach. This led us to develop a novel class of distributions over the space of symmetric and positive definite matrices to form a simple Gibbs sampler to make draws from the posterior. We also developed a generalization of Stein's loss that allowed us to find a simple form of the best equivariant estimator.

In Chapter 4, we then approached mean estimation for tensor-variate data. Common mean estimators in matrix-variate data sets shrink or threshold the singular values of the data matrix. We extended this approach to tensors where we shrink or threshold the mode-specific singular values of a data tensor. These estimators have tuning parameters which we chose by minimizing an unbiased estimate of the risk. These estimators performed extremely well when the underlying mean tensor had approximately low multilinear rank.

Existence of MLE's in the array normal model

As noted in Section 2.5, the current state of knowledge for the conditions under which the MLE exists under the array normal model are not entirely known. In practice, this is not

too important as the results of Wiesel [2012b] suggest that for any single data set, we can see if the MLE exists. That is, Wiesel [2012b] proves that any local minima for Kronecker structured covariance matrices are also global minima. They do this by proving that the negative log-likelihood for the array normal model is "g-convex". Hence, if we find a local minimum, then the likelihood is bounded. However, it would still be interesting to find conditions (preferably on the dimensions) for the existence of the MLE for the array normal model (and, thus, the incredible HOLQ). This would allow us to see under what dimensions the HOLQ is applicable without resorting to an optimization method.

Extending the Efron-Morris estimator to tensor-variate data

The estimators that we introduced in Chapter 4 have good risk performances under the scenarios that we have studied. However, we have not proven that any these estimators are minimax. A current research interest of the author is in developing minimax shrinkage estimators for tensor-variate data. One approach would be an empirical Bayes approach similar in spirit to that of Efron and Morris [1972a]. The approach taken in that paper was, for $X \in \mathbb{R}^{p\times n}$ with $p < n$,

$$X_{[,i]}|\theta_i \sim N_p(\theta_i, I_p) \text{ independent for } i = 1, \ldots, n \tag{5.1}$$

$$\theta_i \overset{i.i.d.}{\sim} N_p(0, A), \tag{5.2}$$

for some $A \in \mathbb{R}^{p\times p}$ symmetric and positive definite. Let $\Theta = (\theta_1, \ldots, \theta_P)$. The Bayes rule under (5.1) and (5.2) is

$$\hat{\Theta} = (I_p - (A + I_p)^{-1})X.$$

Since marginally we have

$$X_{[,i]} \sim N_p(0, A + I_p),$$

the empirical Bayes approach of Efron and Morris [1972a] was to estimate $A + I_p$ with the sample covariance matrix $XX^T/(n - p - 1)$. A natural extension to tensor-variate

data is to replace A in (5.2) with a Kronecker structured covariance matrix. That is, for $\mathcal{X}, \Theta \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ and $p = \prod_{k=1}^{K} p_k$, let

$$\begin{aligned} \operatorname{vec}(\mathcal{X})|\Theta &\sim N_p(\operatorname{vec}(\Theta), I_p) \\ \operatorname{vec}(\Theta) &\sim N_p(0, \Sigma_K \otimes \cdots \otimes \Sigma_1). \end{aligned}$$

The Bayes rule is

$$\operatorname{vec}(\hat{\Theta}) = (I_p - [\Sigma_K \otimes \cdots \otimes \Sigma_1 + I_p]^{-1}) \operatorname{vec}(\mathcal{X}),$$

and marginally we have

$$\operatorname{vec}(\mathcal{X}) \sim N_p(0, I_p + \Sigma_K \otimes \cdots \otimes \Sigma_1).$$

The goal is now to choose an estimator of $I_p + \Sigma_K \otimes \cdots \otimes \Sigma_1$ such that the empirical Bayes estimator is minimax. Finding such estimators and proving under what conditions (presumably on the dimension) that such an empirical Bayes estimator is minimax is not clear. As the Efron-Morris estimator may be viewed as a spectral estimator (4.6), perhaps a natural generalization will exist that is a higher-order spectral estimator (4.13). Of particular interest would be under what conditions (presumably on the Σ_k's) that such tensor-variate empirical Bayes estimators dominate their vector (James-Stein estimator) and matrix (Efron-Morris estimator) counterparts.

BIBLIOGRAPHY

Deniz Akdemir and Arjun K. Gupta. Array variate random variables with multiway Kronecker delta covariance matrix structure. *J. Algebr. Stat.*, 2(1):98–113, 2011. ISSN 1309-3452.

T. W. Anderson, Kai Tai Fang, and Huang Hsu. Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions. *Canad. J. Statist.*, 14(1):55–59, 1986. ISSN 0319-5724. doi: 10.2307/3315036. URL http://dx.doi.org/10.2307/3315036.

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. doi: doi:10.1016/S0167-9473(99)00017-1. URL http://dx.doi.org/doi:10.1016/S0167-9473(99)00017-1.

Brett W. Bader and Tamara G. Kolda. MATLAB tensor classes for fast algorithm prototyping. Technical Report SAND2004-5187, Sandia National Laboratories, October 2004. URL http://www.osti.gov/scitech/biblio/974890.

MS Bartlett. On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, 53:260–283, 1933.

James V. Bondar and Paul Milnes. Amenability: a survey for statistical applications of Hunt-Stein and related conditions on groups. *Z. Wahrsch. Verw. Gebiete*, 57(1):103–128, 1981. ISSN 0044-3719. doi: 10.1007/BF00533716. URL http://dx.doi.org/10.1007/BF00533716.

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. doi: 10.1007/s00216-007-1790-1. URL http://dx.doi.org/10.1007/s00216-007-1790-1.

Rasmus Bro. Review on multiway analysis in chemistry - 2000–2005. *Critical reviews in analytical chemistry*, 36(3-4):279–293, 2006.

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. doi: 10.1109/TSP.2013.2270464. URL http://dx.doi.org/10.1109/TSP.2013.2270464.

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. doi: 10.1348/000711005X64817. URL http://dx.doi.org/10.1348/000711005X64817.

A Cichocki, D Mandic, C Caiafa, AH Phan, G Zhou, Q Zhao, and L De Lathauwer. Tensor decompositions for signal processing applications. *From Two-way to Multiway Component Analysis, ESAT-STADIUS Internal Report*, pages 13–235, 2014.

Thomas M. Cover and Joy A. Thomas. Determinant inequalities via information theory. *SIAM J. Matrix Anal. Appl.*, 9(3):384–392, 1988. ISSN 0895-4798. doi: 10.1137/0609033. URL http://dx.doi.org/10.1137/0609033.

A. P. Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 1981. ISSN 0006-3444. doi: 10.1093/biomet/68.1.265. URL http://dx.doi.org/10.1093/biomet/68.1.265.

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), 2000a. ISSN 0895-4798. doi: 10.1137/S0895479898346995. URL http://dx.doi.org/10.1137/S0895479898346995.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278 (electronic), 2000b. ISSN 0895-4798. doi: 10.1137/S0895479896305696. URL http://dx.doi.org/10.1137/S0895479896305696.

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. doi: 10.1109/78.950782. URL http://dx.doi.org/10.1109/78.950782.

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. doi: 10.1137/06066518X. URL http://dx.doi.org/10.1137/06066518X.

Pierre Dutilleul. The MLE algorithm for the matrix normal distribution. *Journal of statistical computation and simulation*, 64(2):105–123, 1999.

Morris L. Eaton. *Multivariate statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1983. ISBN 0-471-02776-6. A vector space approach.

Morris L. Eaton. *Group invariance applications in statistics*. NSF-CBMS Regional Conference Series in Probability and Statistics, 1. Institute of Mathematical Statistics, Hayward, CA; American Statistical Association, Alexandria, VA, 1989. ISBN 0-940600-15-3.

Morris L. Eaton and Ingram Olkin. Best equivariant estimators of a Cholesky decomposition. *Ann. Statist.*, 15(4):1639–1650, 1987. ISSN 0090-5364. doi: 10.1214/aos/1176350615. URL http://dx.doi.org/10.1214/aos/1176350615.

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

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

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, 1972b. ISSN 0162-1459. doi: 10.1080/01621459.1972.10481215. URL http://dx.doi.org/10.1080/01621459.1972.10481215.

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. doi: 10.1080/01621459.1975.10479864. URL http://dx.doi.org/10.1080/01621459.1975.10479864.

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

Montserrat Fuentes. Testing for separability of spatial-temporal covariance functions. *J. Statist. Plann. Inference*, 136(2):447–466, 2006. ISSN 0378-3758. doi: 10.1016/j.jspi.2004.07.004. URL http://dx.doi.org/10.1016/j.jspi.2004.07.004.

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, 2014a. ISSN 0018-9448. doi: 10.1109/TIT.2014.2323359. URL http://dx.doi.org/10.1109/TIT.2014.2323359.

Matan Gavish and David L Donoho. Optimal shrinkage of singular values. *arXiv preprint arXiv:1405.7511*, 2014b.

David Gerard and Peter Hoff. Equivariant minimax dominators of the MLE in the array normal model. *J. Multivariate Anal.*, 137:32–49, 2015. doi: 10.1016/j.jmva.2015.01.020. URL http://dx.doi.org/10.1016/j.jmva.2015.01.020.

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. doi: 10.2307/1268518. URL http://dx.doi.org/10.2307/1268518.

Lars Grasedyck. Hierarchical singular value decomposition of tensors. *SIAM J. Matrix Anal. Appl.*, 31(4):2029–2054, 2010. ISSN 0895-4798. doi: 10.1137/090764189. URL http://dx.doi.org/10.1137/090764189.

L. R. Haff. The variational form of certain Bayes estimators. *Ann. Statist.*, 19(3):1163–1190, 1991. ISSN 0090-5364. doi: 10.1214/aos/1176348244. URL http://dx.doi.org/10.1214/aos/1176348244.

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. doi: 10.1002/cem.1180070604. URL http://dx.doi.org/10.1002/cem.1180070604.

Peter D. Hoff. Model averaging and dimension selection for the singular value decomposition. *J. Amer. Statist. Assoc.*, 102(478):674–685, 2007. ISSN 0162-1459. doi: 10.1198/016214506000001310. URL http://dx.doi.org/10.1198/016214506000001310.

Peter D. Hoff. Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.*, 6(2):179–196, 2011. ISSN 1936-0975. doi: 10.1214/11-BA606. URL http://dx.doi.org/10.1214/11-BA606.

Peter D Hoff. Multilinear tensor regression for longitudinal relational data. *arXiv preprint arXiv:1412.0048*, 2014. URL http://arxiv.org/abs/1412.0048.

Peter David Hoff. Equivariant and scale-free Tucker decomposition models. *arXiv preprint arXiv:1312.6397*, 2013.

Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013. ISBN 978-0-521-54823-6.

W. James and Charles Stein. Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 361–379. Univ. California Press, Berkeley, Calif., 1961.

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. doi: 10.1016/j.csda.2011.11.012. URL http://dx.doi.org/10.1016/j.csda.2011.11.012.

Julie Josse and Sylvain Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, 2015. doi: 10.1007/s11222-015-9554-9. URL http://dx.doi.org/10.1007/s11222-015-9554-9.

J. Kiefer. Invariance, minimax sequential estimation, and continuous time processes. *Ann. Math. Statist.*, 28:573–601, 1957. ISSN 0003-4851.

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. doi: 10.1348/000711003321645386. URL http://dx.doi.org/10.1348/000711003321645386.

Henk AL Kiers and Iven Van Mechelen. Three-way component analysis: Principles and illustrative application. *Psychological methods*, 6(1):84, 2001.

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. doi: 10.1007/BF02294556. URL http://dx.doi.org/10.1007/BF02294556.

Misha E. Kilmer and Carla D. Martin. Factorization strategies for third-order tensors. *Linear Algebra Appl.*, 435(3):641–658, 2011. ISSN 0024-3795. doi: 10.1016/j.laa.2010.09.020. URL http://dx.doi.org/10.1016/j.laa.2010.09.020.

Eleftherios Kofidis and Phillip A. Regalia. Tensor approximation and signal processing applications. In *Structured matrices in mathematics, computer science, and engineering, I (Boulder, CO, 1999)*, volume 280 of *Contemp. Math.*, pages 103–133. Amer. Math. Soc., Providence, RI, 2001. doi: 10.1090/conm/280/04625. URL http://dx.doi.org/10.1090/conm/280/04625.

Tamara G. Kolda. Multilinear operators for higher-order decompositions. Technical Report SAND2006-2081, Sandia National Laboratories, April 2006. URL http://www.osti.gov/scitech/biblio/923081.

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

Samuel Kotz and Saralees Nadarajah. *Multivariate t distributions and their applications*. Cambridge University Press, Cambridge, 2004. ISBN 0-521-82654-3. doi: 10.1017/CBO9780511550683. URL http://dx.doi.org/10.1017/CBO9780511550683.

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. doi: 10.1002/9780470238004. URL http://dx.doi.org/10.1002/9780470238004. With a foreword by Willem J. Heiser and Jarqueline Meulman.

E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

Lexin Li and Xin Zhang. Parsimonious tensor response regression. *arXiv preprint arXiv:1501.07815*, 2015.

Shang P. Lin and Michael D. Perlman. A Monte Carlo comparison of four estimators of a covariance matrix. In *Multivariate analysis VI (Pittsburgh, Pa., 1983)*, pages 411–429. North-Holland, Amsterdam, 1985.

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.

Nelson Lu and Dale L. Zimmerman. The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.*, 73(4):449–457, 2005. ISSN 0167-7152. doi: 10.1016/j.spl.2005.04.020. URL http://dx.doi.org/10.1016/j.spl.2005.04.020.

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.

Ameur M. Manceur and Pierre Dutilleul. Maximum likelihood estimation for the tensor normal distribution: algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.*, 239:37–49, 2013. ISSN 0377-0427. doi: 10.1016/j.cam.2012.09.017. URL http://dx.doi.org/10.1016/j.cam.2012.09.017.

Kanti V. Mardia and Colin R. Goodall. Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate environmental statistics*, volume 6 of *North-Holland Ser. Statist. Probab.*, pages 347–386. North-Holland, Amsterdam, 1993.

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. doi: 10.1137/060655924. URL http://dx.doi.org/10.1137/060655924.

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. doi: 10.1007/BF02294854. URL http://dx.doi.org/10.1007/BF02294854.

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. doi: 10.1109/TIT.2014.2311661. URL http://dx.doi.org/10.1109/TIT.2014.2311661.

Nintendo. Super Mario World, 1990.

Martin Ohlson, M. Rauf Ahmad, and Dietrich von Rosen. The multilinear normal distribution: introduction and some basic properties. *J. Multivariate Anal.*, 113:37–47, 2013. ISSN 0047-259X. doi: 10.1016/j.jmva.2011.05.015. URL http://dx.doi.org/10.1016/j.jmva.2011.05.015.

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. doi: 10.1214/08-AOAS227. URL http://dx.doi.org/10.1214/08-AOAS227.

Mohsen Pourahmadi. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999. ISSN 0006-3444. doi: 10.1093/biomet/86.3.677. URL http://dx.doi.org/10.1093/biomet/86.3.677.

Joseph J. Rotman. *An introduction to the theory of groups*, volume 148 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, fourth edition, 1995. ISBN 0-387-94285-8. doi: 10.1007/978-1-4612-4176-8. URL http://dx.doi.org/10.1007/978-1-4612-4176-8.

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. doi: 10.1080/01621459.2012.664527. URL http://dx.doi.org/10.1080/01621459.2012.664527.

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. doi: 10.1016/j.jmva.2013.03.005. URL http://dx.doi.org/10.1016/j.jmva.2013.03.005.

Mahendran Shitan and Peter J. Brockwell. An asymptotic test for separability of a spatial autoregressive model. *Comm. Statist. Theory Methods*, 24(8):2027–2040, 1995. ISSN 0361-0926. doi: 10.1080/03610929508831600. URL http://dx.doi.org/10.1080/03610929508831600.

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.

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

Muni Shanker Srivastava and C. G. Khatri. An introduction to multivariate statistics, 1979.

Charles Stein. Estimation of a covariance matrix. *Rietz Lecture*, 1975.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981. ISSN 0090-5364. URL http://links.jstor.org/sici?sici=0090-5364(198111)9:6<1135:EOTMOA>2.0.CO;2-5&origin=MSN.

CM 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.

Akimichi Takemura. An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population. *Tsukuba J. Math.*, 8(2):367–376, 1984. ISSN 0387-4982.

Dacheng Tao, Xuelong Li, Weiming Hu, Stephen Maybank, and Xindong Wu. Supervised tensor learning. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

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. doi: 10.1348/000711000159132. URL http://dx.doi.org/10.1348/000711000159132.

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. URL http://papers.nips.cc/paper/4985-convex-tensor-decomposition-via-structured-schatten-norm-regularization.pdf.

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

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., 2011b. URL http://papers.nips.cc/paper/4453-statistical-performance-of-convex-tensor-decomposition.pdf.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001. ISSN 0022-3239. doi: 10.1023/A:1017501703105. URL http://dx.doi.org/10.1023/A:1017501703105.

Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966. ISSN 0033-3123.

M Alex O Vasilescu and Demetri Terzopoulos. Multilinear subspace analysis of image ensembles. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–93. IEEE, 2003.

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. doi: 10.1007/s11222-013-9444-y. URL http://dx.doi.org/10.1007/s11222-013-9444-y.

Alexander Volfovsky and Peter D. Hoff. Hierarchical array priors for ANOVA decompositions of cross-classified data. *Ann. Appl. Stat.*, 8(1):19–47, 2014. ISSN 1932-6157. doi: 10.1214/13-AOAS685. URL http://dx.doi.org/10.1214/13-AOAS685.

Ami Wiesel. On the convexity in Kronecker structured covariance estimation. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 880–883. IEEE, 2012a.

Ami Wiesel. Geodesic convexity and covariance estimation. *IEEE Trans. Signal Process.*, 60(12):6182–6189, 2012b. ISSN 1053-587X. doi: 10.1109/TSP.2012.2218241. URL http://dx.doi.org/10.1109/TSP.2012.2218241.

Ruo-yong Yang and James O. Berger. Estimation of a covariance matrix using the reference prior. *Ann. Statist.*, 22(3):1195–1211, 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325625. URL http://dx.doi.org/10.1214/aos/1176325625.

Xiang Zhang, Lexin Li, Hua Zhou, Dinggang Shen, et al. Tensor generalized estimating equations for longitudinal imaging analysis. *arXiv preprint arXiv:1412.6592*, 2014.

James V. Zidek. A representation of Bayes invariant procedures in terms of Haar measure. *Ann. Inst. Statist. Math.*, 21:291–308, 1969. ISSN 0020-3157.

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

Appendix A

A.1 Proofs of Chapter 2

A.1.1 Equivalence of Algorithms 1 and 2

In Algorithm 1, the core array is $Y = (L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X$. Let MZ be the LQ decomposition of $L_k Y_{(k)} = X_{(k)} L_{-k}^{-T}$. Then Algorithm 1 updates the core array by

$$\begin{aligned} Y_{(k)} &\leftarrow (M/|M|^{1/p_k})^{-1} X_{(k)} L_{-k}^{-T} \\ &= (M/|M|^{1/p_k})^{-1} L_k L_k^{-1} X_{(k)} L_{-k}^{-T} \\ &= (M/|M|^{1/p_k})^{-1} L_k Y_{(k)} \\ &= c_1 M^{-1} L_k Y_{(k)}, \end{aligned}$$

for $c_1 = |M|^{1/p_k}$. Note that $Y_{(k)}/||Y_{(k)}|| = c_2 L_k^{-1} X_{(k)} L_{-k}^{-T} = c_2 L_k^{-1} MZ$, for $c_2 = ||Y_{(k)}||^{-1}$. That is, set $L = c_2 L_k^{-1} M$ so that LZ is the LQ decomposition of $Y_{(k)}/||Y_{(k)}||$. Then Algorithm 2 updates the core array by

$$\begin{aligned} Q_{(k)} = Y_{(k)}/||Y_{(k)}|| &\leftarrow Z/||Z|| \\ &\propto L^{-1} LZ \\ &\propto (L_k^{-1} M)^{-1} Y_{(k)} \\ &\propto M^{-1} L_k Y_{(k)}, \end{aligned}$$

Hence, each update of the core array is the same for both algorithms at each iteration up to a scale difference. For each iteration of Algorithm 1, $L_k \leftarrow M/|M|^{1/p_k}$. But for each iteration of Algorithm 2, $L_k \leftarrow L_k L/|L|^{1/p_k} = L_k c_2 L_k^{-1} M/|c_2 L_k^{-1} M|^{1/p_k} = M/|M|^{1/p_k}$. Hence, L_k is being updated the same in both algorithms at each iteration.

In this paragraph, we prove that we are updating the scale in Algorithm 2 correctly. Noting that $M = ||Y||L_k L = \ell L_k L$, for ℓ the current value of the scale, the update for the scale is

$$\begin{aligned}\tilde{\ell} &= ||(M/|M|^{1/p_k})^{-1} X_{(k)} L_{-k}^{-T}|| \\ &= ||(M/|M|^{1/p_k})^{-1} MZ|| \\ &= |(||Y||L_k L)|^{1/p_k}\, ||Z|| \\ &= \ell |L|^{1/p_K} ||Z||.\end{aligned}$$

A.1.2 Update of $k \in J_2$ in HOLQ Junior

For $X \in \mathbb{R}^{p\times n}$, consider finding the minimizer in $D \in \mathcal{D}_p^+$ of $||D^{-1}X||$. Using Lagrange multipliers, and letting $S = XX^T$, this is equivalent to minimizing in D

$$\operatorname{tr}(D^{-2}S) + \lambda(|D| - 1) = \sum_{i=1}^{p} D_{[i,i]}^{-2} S_{[i,i]} + \lambda(\prod_{i=1}^{p} D_{[i,i]} - 1),$$

for D a diagonal p by p matrix with positive diagonal elements. The solution to this optimization problem is

$$\tilde{D}_{[i,i]} = \left(S_{[i,i]} / \prod_{i=1}^{p} S_{[i,i]}^{1/p} \right)^{-1/2} \quad \text{for } i = 1, \ldots, p$$

or

$$\tilde{D} = \operatorname{diag}(S_{[1,1]}, \ldots, S_{[p,p]})^{-1/2} / |\operatorname{diag}(S_{[1,1]}, \ldots, S_{[p,p]})^{-1/2}|^{1/p}.$$

So for the block coordinate descent algorithm, for step $k \in J_2$,

$$\begin{aligned}&\text{Set } S_k = X_{(k)} L_{-k}^{-T} L_{-k}^{-1} X_{(k)}^T \\ &\text{Set } E = \operatorname{diag}(S_{k[1,1]}, \ldots, S_{k[p,p]})^{-1/2} \\ &\text{Set } L_k \leftarrow E/|E|^{1/p}.\end{aligned} \tag{A.1}$$

This block coordinate descent algorithm is equivalent to the following steps of simultaneously updating the core array along with the component matrix:

$$\begin{aligned}
&\text{Set } R_k = Q_{(k)}Q_{(k)}^T \\
&\text{Set } F = \text{diag}(R_{k[1,1]}, \ldots, R_{k[p_k,p_k]})^{1/2} \\
&\text{Set } \ell \leftarrow \ell |F|^{1/p_k} ||F^{-1}Q_{(k)}|| \\
&\text{Set } L_k \leftarrow L_k F/|F|^{1/p_k} \\
&\text{Set } Q_{(k)} \leftarrow F^{-1}Q_{(k)}/||F^{-1}Q_{(k)}||.
\end{aligned} \tag{A.2}$$

We'll now prove the equivalence of using step (A.1) or step (A.2) to find the HOLQ junior. At each step of (A.1), the core array is $Y = (L_1^{-1}, \ldots, L_K^{-1}) \cdot X$. Hence, the core array is updated at each iteration of $k \in J_2$ by

$$Y_{(k)} \leftarrow (E/|E|^{1/p_k})^{-1} X_{(k)} L_{-k}^{-T} = |E|^{1/p_k} E^{-1} X_{(k)} L_{-k}^{-T}.$$

Note that, since $Q = Y/||Y||$, we have

$$FF \propto \text{diag}(Y_{(k)} Y_{(K)}^T) \propto \text{diag}(L_k^{-1} S_k L_k^{-1}) \propto L_k^{-1} E E L_k^{-1}.$$

This implies that $F = c_2 L_k^{-1} E$ for some constant c_2. Hence, the core in (A.2) is being updated by:

$$\begin{aligned}
Q_{(k)} = Y_{(k)}/||Y|| \leftarrow F^{-1}Q_{(k)} &\propto F^{-1}Y_{(k)} = F^{-1}L_k^{-1}X_{(k)}L_{-k}^{-T} \\
&\propto E^{-1}L_k L_k^{-1} X_{(k)} L_{-k}^{-T} = E^{-1} X_{(k)} L_{-k}^{-T}.
\end{aligned}$$

So the core array is being updated the same at each step, up to a scale difference. Likewise, in (A.1), we have $L_k \leftarrow E/|E|^{1/p_k}$ whereas in (A.2) we have $L_k \leftarrow L_k F/|F|^{1/p_k} = L_k c_2 L_k^{-1} E/|c_2 L_k^{-1} E|^{1/p_k} = E/|E|^{1/p_k}$, so each component matrix is being updated the same at each iteration.

Note that the diagonal elements of $Q_{(k)}Q_{(k)}^T$ in (A.2) are being scaled to be $1/p_k$ at each iteration. Hence, any fixed point of this algorithm must have the property that $\text{diag}\left(Q_{(k)}Q_{(k)}^T\right) = \mathbf{1}_{p_k}/p_k$ for all $k \in J_2$. In other words,the rows of $Q_{(k)}$ have Frobenius norm $1/p_k$.

A.1.3 Update of $k \in J_3$ in HOLQ Junior

For $K = 1$ and $n \geq p$, we require finding the $L_k \in \mathcal{G}_{p_k}^{Ch}$ that minimizes $||L^{-1}X||^2$. Using Lagrange multipliers, this is equivalent to finding the V in the general linear group of p by p non-singular matrices that minimizes

$$\text{tr}(VV^TXX^T) - \text{tr}(\Lambda_1(V - I_p)) - \text{tr}(\mathbf{1}_p\mathbf{1}_p^T(\Lambda_2 * V)), \tag{A.3}$$

where $\Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_p)$, $\mathbf{1}_p$ is the p-dimensional vector of 1's, "$*$" is the Hadamard (element-wise) product, and Λ_2 is upper triangular with 0's in the diagonal. That is,

$$\Lambda_2 = \begin{pmatrix} 0 & \lambda_{1,2} & \lambda_{1,3} & \cdots & \cdots & \lambda_{1,p} \\ 0 & 0 & \lambda_{2,3} & \cdots & \cdots & \lambda_{2,p} \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & 0 & \lambda_{p-1,p} \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

The idea is that the Lagrange multipliers in Λ_1 are constraining the diagonal elements of V to be 1, and the Lagrange multipliers in Λ_2 are constraining the upper triangular elements of V to be 0. Once we find the minimizer, we can set $L = V^{-1}$. Taking derivatives of (A.3) and setting equal to zero, we have

$$\begin{aligned} &2XX^TV - \Lambda_1 - \Lambda_2 = 0 \text{ and } V \in \mathcal{G}_p^{Ch} \\ &\Leftrightarrow V = (\Lambda_1 + \Lambda_2)(XX^T)^{-1}/2 \text{ and } V \in \mathcal{G}_p^{Ch}. \end{aligned}$$

Note that $\Lambda_1 + \Lambda_2$ is upper triangular. By the uniqueness of the LDU decomposition of $XX^T = U^TDU$ where $U^T \in \mathcal{G}_{p_k}^{Ch}$ [Horn and Johnson, 2013, Corollary 3.5.5], the only critical point occurs at $V = U^{-T}$ and $\Lambda_1 + \Lambda_2 = 2DU$.

Since the constraints are all linear, to prove that this minimizer is a global minimizer, it suffices to prove that $\text{tr}(VV^TXX^T)$ is convex in V. But by Exercise 1 of Section 10.6 in Magnus and Neudecker [1999], the Hessian matrix is $2XX^T \otimes I_p$, which is clearly positive definite.

To summarize, the minimizer of $||\tilde{L}^{-1}X||^2$ in $\tilde{L} \in \mathcal{G}_p^{Ch}$ is U^T from the LDU decomposition of $XX^T = U^TDU$. This is equivalent to taking the LQ decomposition of $X = LQ$, then setting $F = \text{diag}(L_{[1,1]}, \ldots, L_{[p,p]})$. The minimizer is then LF^{-1}. What's "left over" after multiplying out LF^{-1} is then FQ, which has orthogonal (though not necessarily orthonormal) rows.

For modes where $L_k \in \mathcal{G}_{p_k}^{Ch}$, we thus update L_k and the core Q by:

$$\begin{aligned}
&\text{Take LQ decomposition of core } Q_{(k)} = LZ \\
&\text{Set } F = \text{diag}(L_{[1,1]}, \ldots, L_{[p,p]}) \\
&L_k \leftarrow L_k L F^{-1} \\
&Q_{(k)} \leftarrow FZ/||FZ|| \\
&\ell \leftarrow \ell ||FZ||.
\end{aligned}$$

Hence, any fixed point, including the HOLQ junior, must have the property that $Q_{(k)}$ has orthogonal, though not necessarily orthonormal, rows. This proves the last part of Theorem 5.

A.1.4 Proof of Theorem 8

We can represent models H_0 and H_1 in Theorem 7 as being generated under two different groups. Let $\{1, \ldots, K\} = J_1 \cup J_2 \cup J_3 \cup J_4$. Let $\Psi_k \in \mathcal{G}_0^{(k)}$ where $\mathcal{G}_0^{(k)} = \mathcal{G}_{p_k}^+$, $\mathcal{D}_{p_k}^+$, $\mathcal{G}_{p_k}^{Ch}$, or $\{I_{p_k}\}$ if k is in J_1, J_2, J_3, or J_4, respectively. Let $\{1, \ldots, M\} = \tilde{J}_1 \cup \tilde{J}_2 \cup \tilde{J}_3 \cup \tilde{J}_4$. Let $\Phi_m \in \mathcal{G}_1^{(m)}$ and $\mathcal{G}_1^{(m)} = \mathcal{G}_{q_m}^+$, $\mathcal{D}_{q_m}^+$, $\mathcal{G}_{p_k}^{Ch}$, or $\{I_{q_m}\}$ if m is in $\tilde{J}_1$, $\tilde{J}_2$, $\tilde{J}_3$, or $\tilde{J}_4$, respectively. Then the models of H_0 and H_1 can be represented by

$$\begin{aligned}
H_0 &: \text{vec}(X) \stackrel{d}{=} \sigma_0(\Psi_1, \ldots, \Psi_K) \cdot \text{vec}(Z) \\
H_1 &: \text{vec}(X) \stackrel{d}{=} \sigma_1(\Phi_1, \ldots, \Phi_M) \cdot \text{vec}(Z),
\end{aligned}$$

where $\text{vec}(Z)$ is a p vector with standard normal entries. Saying that H_0 is a submodel of H_1 is equivalent to saying that $\mathcal{G}_0$ is a subgroup of $\mathcal{G}_1$. Hence, we are in the situation of

having a hypothesis testing problem that is invariant under $\mathcal{G}_0$ [Eaton, 1989, Definition 3.2]. The LRT statistic is an invariant function [Eaton, 1983, Proposition 7.13]. The distribution of any invariant function depends only on the maximally invariant parameter [Lehmann and Romano, 2005, Theorem 6.3.2]. Under the null, the maximally invariant parameter is a constant because the group action is transitive over the parameter space (since the model is generated by $\mathcal{G}_0$).

A.1.5 Proof of Equivalence of Algorithms 3 and 4.

The core array from Algorithms 3 and 4 is $R \propto (L_1^{-1}, \ldots, L_K^{-1}, I_n) \cdot X$. Let LL^TZ be the polar decomposition of $X_{(k)}L_{-k}^{-1}$ for L lower triangular with positive diagonal elements and $Z^T \in \mathcal{V}_{p_k, np/p_k}$. This is, equivalently, the polar decomposition of $L_kR_{(k)}$ as in Algorithm 4. Thus we have

$$\begin{aligned} R &\leftarrow cL^{-1}X_{(k)}L_{-k}^{-1} \\ &\propto L^{-1}LL^TZ \\ &\propto L^TZ. \end{aligned}$$

Hence, we are updating the core array correctly. L_k is trivially being updated correctly in the Algorithm 4. To see that we are updating the scale correctly, note that

$$\begin{aligned} \ell &= ||(L/||L||)^{-1}X_{(k)}L_{-k}^{-1}|| \\ &= ||L||||L^{-1}LL^TZ|| \\ &= ||L||||L^TZ|| \\ &= ||L||||R||. \end{aligned}$$

A.1.6 Theorem 1 from Anderson et al. [1986]

The following is a simplified version of Theorem 1 from Anderson et al. [1986].

Theorem 16. *Let Ω be a set in the space of $\mathcal{S}_p^+$ such that if $S \in \Omega$ then $cS \in \Omega$ for all $c > 0$. For $X \in \mathbb{R}^{n\times p}$, suppose that g is such that $g(tr(XX^T))$ is a density in $\mathbb{R}^{n\times p}$ and*

$y^{np/2}g(y)$ has a finite positive maximum at y_g. Suppose on the basis of an observation of X from $|\Sigma|^{-n/2}g(tr(X\Sigma^{-1}X^T))$, the MLE under normality $\hat{\Sigma} \in \Omega$ exists and is unique and that $\hat{\Sigma}$ is positive definite with probability 1. Then the MLE for g is proportional to $\hat{\Sigma}$ and the maximum of the likelihood is $|\hat{\Sigma}|^{-n/2}g(y_g)$.

In this chapter, Ω is the cone of Kronecker structured covariance matrices. This result says that the MLE under elliptically contoured distributions is proportional to the MLE under normality. In this chapter, we use the parameterization where $|\sigma^2\Sigma_K \otimes \cdots \otimes \Sigma_1| = (\sigma^2)^p$. Hence, $|\hat{\Sigma}|^{-n/2}g(y_g) = \hat{\sigma}^{-np}g(y_g)$. For the HOLQ junior, we are implying that $n = 1$, with the understanding that some modes might be the identity.

A.2 Proofs of Chapter 3

A.2.1 Proof of Theorem 10

Proof. Let $a > 0$, $A_k \in \mathcal{G}_{p_k}^+$ for all $k = 1, \ldots, K$. Let t be a fixed element in $\mathbb{R}^+$ and T_k be fixed elements in $\mathcal{G}_{p_k}^+$ for $k = 1, \ldots, K$. In the terminology of Definition 1.7 of Eaton [1989], the integral with respect to Lebesgue measure is relatively right invariant with multiplier $\chi(t, T_1, \ldots, T_K) = t\prod_{k=1}^K \prod_{i=2}^{p_k} T_{k[i,i]}^{2-i}$ if the following holds:

$$\begin{aligned} &\int_{\mathcal{G}_{\mathbf{p}}^+} f(a/t, A_1T_1^{-1}, \ldots, A_KT_K^{-1}) d\mu\,(a, A_1, \ldots, A_K) \\ &= \left(t\prod_{k=1}^K \prod_{i=2}^{p_k} T_{k[i,i]}^{2-i}\right) \int_{\mathcal{G}_{\mathbf{p}}^+} f(a, A_1, \ldots, A_K) d\mu\,(a, A_1, \ldots, A_K)\,, \end{aligned} \tag{A.4}$$

for arbitrary $f()$. If (A.4) holds, then by Theorem 1.6 of Eaton [1989], a right invariant measure over the group $\mathcal{G}_{\mathbf{p}}^+$ is

$$\chi(a, A_1, \ldots, A_K)^{-1} = \frac{1}{a}\prod_{k=1}^K \prod_{i=2}^{p_k} A_{k[i,i]}^{i-2} d\mu.$$

It remains to make a change of variables to show that (A.4) holds. For $E_k, T_k \in \mathcal{G}_{p_k}^+$ with T_k fixed for $k = 1, \ldots, K$, let $g_k(E_k) = E_kT_k$ for $k = 1, \ldots, K$. For $e, t > 0$ with t fixed let

$g(e) = et$. The Jacobian for transforming the scale, $g(e) = et$, is t. The Jacobian for the transformation $g_k(E_k) = E_k T_k$ is

$$J(E_k) = \prod_{i=2}^{p_k} T_{k[i,i]}^{2-i}. \tag{A.5}$$

To see this, note that this transformation is equivalent to $p_k(p_k+1)/2 - 1$ linear transformations of the form:

$$g_{i,j} : E_{k[i,j]} \mapsto \sum_{j \le m \le i} E_{k[i,m]} T_{k[m,j]} \text{ for all } 1 \le j \le i \le p_k \text{ s.t. } (i,j) \ne (1,1).$$

Stack the elements of E_k into the following vector:

$$\begin{aligned} s =& (E_{k[p_k,p_k]}, E_{k[p_k,p_k-1]}, E_{k[p_k-1,p_k-1]}, E_{k[p_k,p_k-2]}, \\ & E_{k[p_k-1,p_k-2]}, E_{k[p_k-2,p_k-2]}, E_{k[p_k,p_k-3]}, \ldots, E_{k[2,1]}), \end{aligned}$$

and notice that the matrix of the linear transformation is lower triangular where in the diagonal, each $T_{k[i,i]}$ is repeated $p_k - i + 1$ times for $i = 2, 3, \ldots, p_k$, and $T_{k[1,1]}$ is repeated $p_k - 1$ times. Call this matrix of linear transformation u. Then the linear transformation can be written as: $g_k(s) = us$. Hence the determinant of the Jacobian is

$$|u| = T_{k[1,1]}^{p_k-1} \prod_{i=2}^{p_k} T_{k[i,i]}^{p_k-i+1} = \prod_{i=2}^{p_k} T_{k[i,i]}^{2-i},$$

where the second equality results from our unit determinant parameterization of $\mathcal{G}_{p_k}^+$, $\prod_{i=2}^{p_k} T_{k[i,i]}^{-1} = T_{k[1,1]}$. □

A.2.2 Proof of Theorem 11

Consider the reformulation of the problem to a parameterization of $\Sigma = \sigma^2(\Psi_K \Psi_K^T \otimes \cdots \otimes \Psi_1 \Psi_1^T)$ where $\Psi_{k[1,1]} = 1$ for $k = 1, \ldots, K$. That is, we now work with the group $\mathcal{G}_{\mathbf{p}}^1 = \{(a, A_1, \ldots, A_K) | a > 0, A_k \in \mathcal{G}_{p_k}^1 \text{ for } k = 1, \ldots, K\}$ where $\mathcal{G}_{p_k}^1$ is the group of p_k by p_k lower triangular matrices with positive diagonal elements and 1 in the $(1,1)$ position. The group operation in $\mathcal{G}_{p_k}^1$ is matrix multiplication, and that of $\mathcal{G}_{\mathbf{p}}^1$ is component-wise multiplication. The left and right Haar measures over $\mathcal{G}_{p_k}^1$ are easy to derive:

Lemma 5. *For $E_k, T_k \in \mathcal{G}^1_{p_k}$ with T_k fixed, the Jacobian for the transformation $g(E_k) = E_k T_k$ is*

$$J(E_k) = \prod_{i=2}^{p_k} T_{k[i,i]}^{p_k - i + 1}$$

the Jacobian for the transformation $g(E_k) = T_k E_k$ is

$$J(E_k) = \prod_{i=2}^{p_k} T_{k[i,i]}^{i}$$

So the right Haar measure is $d\nu_r(E_k) = \prod_{i=2}^{p_k} E_{k[i,i]}^{-p_k+i-1}$.

Proof. The proof is very similar to those in Propositions 5.13 and 5.14 of Eaton [1983], noting that $T_{k[1,1]} = 1$. □

We'll eventually need the inverse transformation, which follows directly from Theorem 3 of chapter 8 section 4 of Magnus and Neudecker [1999].

Lemma 6. *For $E_k \in \mathcal{G}^1_{p_k}$, the Jacobian for the transformation $g(E_k) = E_k^{-1}$ is*

$$\prod_{i=2}^{p_k} E_{k[i,i]}^{-p_k-1} \tag{A.6}$$

Proof. From Magnus and Neudecker [1999], $d(E_k^{-1}) = -E_k^{-1}(dE_k)E_k^{-1}$. Using Lemma 5, the Jacobian of the first transformation, $g_1(dE_k) = E_k^{-1}(dE_k)$ is $\prod_{i=2}^{p_k} E_{k[i,i]}^{-i}$. Jacobian of the second transformation $g_2(dE_k) = (dE_k)E_k^{-1}$ is $\prod_{i=2}^{p_k} E_{k[i,i]}^{-p_k+i-1}$. Hence, overall Jacobian is (A.6). □

Under this new parameterization, the likelihood is

$$\begin{aligned}
&p(X|\sigma, \Psi_1, \ldots, \Psi_K) \\
&= (2\pi)^{np/2}|\sigma^2(\Psi_K\Psi_K^T \otimes \cdots \otimes \Psi_1\Psi_1^T)|^{-n/2} \\
&\times \exp\{-||(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X||^T/(2\sigma^2)\}\} \\
&\propto \sigma^{-np} \prod_{k=1}^{K}\prod_{i=2}^{p_k} \Psi_{k[i,i]}^{-np/p_k} \exp\{-||(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X||^T/(2\sigma^2)\}\},
\end{aligned}$$

where $p = \prod_{k=1}^{K} p_k$. The (improper) prior is

$$\pi(\sigma, \Psi_1, \ldots, \Psi_K) \propto \frac{1}{\sigma} \prod_{k=1}^{K} \prod_{i=2}^{p_k} \Psi_{k[i,i]}^{i-p_k-1}.$$

Hence, the posterior is

$$\sigma^{-np-1} \prod_{k=1}^{K} \prod_{i=2}^{p_k} \Psi_{k[i,i]}^{i-np/p_k-p_k-1} \exp\{-||(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X||^T/(2\sigma^2)\}.$$

Since $\sigma^2|\Psi \sim$ inverse-gamma$(np/2, ||(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X||^2/2)$, we can integrate out σ^2, obtaining

$$\pi(\Psi_1, \ldots, \Psi_K|X) \propto ||(\Psi_1^{-1}, \ldots, \Psi_K^{-1}, I_n) \cdot X||^{-np} \prod_{k=1}^{K} \prod_{i=2}^{p_k} \Psi_{k[i,i]}^{i-np/p_k-p_k-1}.$$

Let $S = X_{(K+1)}^T X_{(K+1)}$, the sample covariance matrix, then

$$\begin{aligned} &\pi(\Psi_1, \ldots, \Psi_K|X) \\ &\propto \text{tr}[S(\Psi_K^{-T}\Psi_K^{-1} \otimes \cdots \otimes \Psi_1^{-T}\Psi_1^{-1})]^{-np/2} \prod_{k=1}^{K} \prod_{i=2}^{p_k} \Psi_{k[i,i]}^{i-np/p_k-p_k-1}. \end{aligned}$$

Let $L_k = \Psi_k^{-1}$ for $k = 1, \ldots, K$. Then, using Lemma 6, we have

$$\begin{aligned} &\pi(L_1, \ldots, L_K|X) \\ &\propto \text{tr}[S(L_K^T L_K \otimes \cdots \otimes L_1^T L_1)]^{-np/2} \prod_{k=1}^{K} \prod_{i=2}^{p_k} L_{k[i,i]}^{np/p_k-i} \end{aligned} \tag{A.7}$$

The posterior density is integrable if and only if (A.7) is integrable. We will now prove that when $n > \prod_{k=1}^{K} p_k$ then (A.7) is integrable. First, consider, consider the integral over $\mathcal{G}_p^1$, where $p = \prod_{k=1}^{K} p_k$,

$$\int_{\mathcal{G}_p^1} \text{tr}\left(VSV^T\right)^{-np/2} \prod_{i=2}^{p} V_{[i,i]}^{np-p-1} dV \tag{A.8}$$

Let $e = (1, 0, \ldots, 0)^T$, the vector of length p with a 1 in the first position and 0's everywhere else. Then $V = (e^T, V_2^T)^T$ and

$$\text{tr}(VSV^T) = tr(e_1^T S e_1) + tr(V_2 S V_2^T) = S_{[1,1]} + tr(V_2 S V_2^T)$$

$$= (1 + tr(V_2 S V_2^T)/S_{[1,1]})S_{[1,1]} = (1 + tr(V_2 S_T (V_2 S_T)^T)/S_{[1,1]})S_{[1,1]},$$

where $S = S_T S_T^T$ is the lower triangular Cholesky decomposition of S. Let $W = V_2 S_T$, so $V_2 = W S_T^{-1}$. The Jacobian of this transformation is $S_{T[1,1]}^{1-p} \prod_{i=2}^{p} S_{T[i,i]}^{i-p-1}$ (same as the Jacobian in Proposition 5.14 of Eaton [1983] except with one less $S_{T[1,1]}$ term). Then Equation (A.8) is proportional to

$$\begin{aligned}
&\int_{\mathcal{G}_p^1} \left((1 + \mathrm{tr}(WW^T)/S_{[1,1]}\right)^{-np/2} \prod_{i=2}^{p} W_{[i,i]}^{np-p-1} dW \\
&= \int_{\mathcal{G}_p^1} \left((1 + \mathbf{w} D \mathbf{w}/(np-p)\right)^{-(np-p+p)/2} \prod_{i=2}^{p} W_{[i,i]}^{np-p-1} dW,
\end{aligned}$$

where $\mathbf{w}$ is a vector containing all the non-zero elements of W and $D = (n-p)I_p/S_{[1,1]}$. Notice that $\left((1 + \mathbf{w} D \mathbf{w}/(np-p)\right)^{-(np-p+p)/2}$ is the kernel of a multivariate T distribution with degrees of freedom $np-p$ and scale matrix $D^{-1} = S_{[1,1]} I_p/(np-p)$ [Kotz and Nadarajah, 2004, equation (1.1)]. Note that $E[W_{[i,j]}^{\nu}] < \infty$ if $\nu < n-p$ [Kotz and Nadarajah, 2004, section 1.7]. In particular, $n-p-1 < n-p$. Hence

$$\int_{\mathcal{G}_p^1} \mathrm{tr}\left(VSV^T\right)^{-np/2} \prod_{i=2}^{p} V_{[i,i]}^{np-p-1} dV < \infty$$

Using this, we have the following inequalities:

$$\begin{aligned}
\infty &> \int_{\mathcal{G}_p^1} \mathrm{tr}\left[VSV^T\right]^{-np/2} \prod_{i=1}^{p} V_{k[i,i]}^{np-p-1} dV \\
&= \int_{\mathcal{G}_p^1} \mathrm{tr}\left[VSV^T\right]^{-np/2} |V|^{np-p-1} dV \\
&\geq \int_{\mathcal{G}_{p_1}^1 \times \cdots \times \mathcal{G}_{p_K}^1} \mathrm{tr}\left[(L_K \otimes \cdots \otimes L_1) S (L_K \otimes \cdots \otimes L_1)^T\right]^{-np/2} \\
&\times |L_K \otimes \cdots \otimes L_1|^{np-p-1} dL_1 \cdots dL_K \\
&= \int_{\mathcal{G}_{p_1}^1 \times \cdots \times \mathcal{G}_{p_K}^1} \mathrm{tr}\left[S(L_K^T L_K \otimes \cdots \otimes L_1^T L_1)\right]^{-np/2} \\
&\times \prod_{k=1}^{K} \prod_{i=2}^{p_k} L_{k[i,i]}^{(np-p-1)p/p_k} dL_1 \cdots dL_K,
\end{aligned}$$

where the second inequality results from integrating over a smaller space. Note the following results: (1) $(np-p-1)p/p_k \geq np/p_k - i_k$ for all $k = 1, \ldots, K$ and $i_k = 2, \ldots, p_k$ if $n \geq p$, (2) $L_{k[i,i]} > 0$, and (3) $E[|X|^{r_1}] < \infty$ and $r_1 > r_2 \Rightarrow E[|X|^{r_2}] < \infty$. Hence,

$$\begin{aligned}\infty > & \int_{\mathcal{G}^1_{p_1} \times \cdots \times \mathcal{G}^1_{p_K}} \operatorname{tr}[S(L_K^T L_K \otimes \cdots \otimes L_1^T L_1)]^{-np/2} \\ & \times \prod_{k=1}^{K} \prod_{i=2}^{p_k} L_{k[i,i]}^{np/p_k - i} dL_1 \cdots dL_K\end{aligned}$$

and the result is proved.

A.2.3 Proof of Lemma 1

Proof. Let VV^T be the lower triangular Cholesky decomposition of a Wishart$_p(\nu, I_p)$-distributed random matrix. Recall from Bartlett's decomposition [Bartlett, 1933] that the elements of V are independent with

$$V^2_{[i,i]} \sim \chi^2_{\nu-i+1} \text{ and } V_{[i,j]} \sim N(0,1).$$

Let $S = V^T V$. For $i \neq j$, we have

$$E\left[S_{[i,j]}\right] = E\left[\sum_{k=1}^{p} V_{[k,i]} V_{[k,j]}\right] = \sum_{k=1}^{p} E\left[V_{[k,i]}\right] E\left[V_{[k,j]}\right].$$

For $i \neq j$, we have either $E[V_{[k,i]}] = 0$ or $E[V_{[k,j]}] = 0$ for all $k = 1, \ldots, p$. Hence, $E[S_{[i,j]}] = 0$ for all $i \neq j$.

For $i = j$, we have

$$\begin{aligned}E\left[S_{[i,j]}\right] = E\left[\sum_{k=1}^{p} V_{[k,i]} V_{[k,j]}\right] &= \sum_{k=1}^{p} E\left[V^2_{[k,i]}\right] = E\left[V^2_{[i,i]}\right] + \sum_{k=i+1}^{p} E\left[V^2_{[k,i]}\right] \\ &= \nu - i + 1 + \sum_{k=i+1}^{p} 1 = \nu - i + 1 + p - i = \nu + p + 1 - 2i.\end{aligned}$$

This expectation has been calculated in other papers [James and Stein, 1961, Eaton and Olkin, 1987, for example]. □

A.2.4 Proof of Lemma 2

Proof. We proceed by invariance arguments. The Jacobian, $J(\sigma, \Psi)$, is the unique continuous function that satisfies

$$\begin{aligned}\int_{G_{p_k}^+} f(L)\frac{dL}{\prod_{i=1}^{p_k} L_{[i,i]}^{p_k-i+1}} &= \int_{\mathbb{R}\times\mathcal{G}_{p_k}^+} f(\sigma\Psi)\frac{J(\sigma,\Psi)d\sigma d\Psi}{\prod_{i=1}^{p_k}(\sigma\Psi_{[i,i]})^{p_k-i+1}} \\ &= \int_{\mathbb{R}\times\mathcal{G}_{p_k}^+} f(\sigma\Psi)\frac{J(\sigma,\Psi)d\sigma d\Psi}{\sigma^{p_k(p_k+1)/2}\prod_{i=1}^{p_k}\Psi_{[i,i]}^{p_k-i+1}},\end{aligned}$$

where $dL/(\prod_{i=1}^{p_k} L_{[i,i]}^{p_k-i+1})$ is a right invariant measure with respect to the action $L \mapsto LA$ on $G_{p_k}^+$ for $A \in G_{p_k}^+$ [Eaton, 1983, Proposition 5.14]. Hence, this invariance property must also hold for the right integral. So for $b > 0$ and $B \in \mathcal{G}_{p_k}^+$, we have that $bB \in G_{p_k}^+$ and

$$\int_{\mathbb{R}\times\mathcal{G}_{p_k}^+} f(\sigma\Psi)\frac{J(\sigma,\Psi)d\sigma d\Psi}{\sigma^{p_k(p_k+1)/2}\prod_{i=1}^{p_k}\Psi_{[i,i]}^{p_k-i+1}} = \int_{\mathbb{R}\times\mathcal{G}_{p_k}^+} f(b\sigma\Psi B)\frac{J(\sigma,\Psi)d\sigma d\Psi}{\sigma^{p_k(p_k+1)/2}\prod_{i=1}^{p_k}\Psi_{[i,i]}^{p_k-i+1}}.$$

So making the change of variables $\sigma = e/b$ and $\Psi = EB^{-1}$, we have

$$\begin{aligned}&\int_{\mathbb{R}\times\mathcal{G}_{p_k}^+} f(b\sigma\Psi B)\frac{J(\sigma,\Psi)d\sigma d\Psi}{\sigma^{p_k(p_k+1)/2}\prod_{i=1}^{p_k}\Psi_{[i,i]}^{p_k-i+1}} \\ &= \int_{\mathbb{R}\times\mathcal{G}_{p_k}^+} f(eE)\frac{\frac{1}{b}\prod_{i=2}^{p_k} B_{[i,i]}^{i-2} J(e/b, EB^{-1})dedE}{(e/b)^{p_k(p_k+1)/2}\prod_{i=1}^{p_k} E_{[i,i]}^{p_k-i+1} B_{[i,i]}^{i-p_k-1}} \\ &= \int_{\mathbb{R}\times\mathcal{G}_{p_k}^+} f(eE)\frac{b^{p_k(p_k+1)/2-1} B_{[1,1]}^{p_k}\prod_{i=2}^{p_k} B_{[i,i]}^{p_k-1} J(e/b, EB^{-1})dedE}{e^{p_k(p_k+1)/2}\prod_{i=1}^{p_k} E_{[i,i]}^{p_k-i+1}} \\ &= \int_{\mathbb{R}\times\mathcal{G}_{p_k}^+} f(eE)\frac{b^{p_k(p_k+1)/2-1} B_{[1,1]} J(e/b, EB^{-1})dedE}{e^{p_k(p_k+1)/2}\prod_{i=1}^{p_k} E_{[i,i]}^{p_k-i+1}},\end{aligned}$$

where we used (A.5) for the first equality and our parameterization of $\mathcal{G}_{p_k}^+$, $\prod_{i=2}^{p_k} B_{[i,i]}^{-1} = B_{[1,1]}$, for the last equality. So we must have that

$$J(\sigma,\Psi) = b^{p_k(p_k+1)/2-1} B_{[1,1]} J(\sigma/b, \Psi B^{-1}).$$

Set $B = \Psi$ and $b = \sigma$ to obtain: $J(\sigma,\Psi) = \sigma^{p_k(p_k+1)/2-1}\Psi_{[1,1]} J(1, I)$, where $J(1, I)$ is a constant. □

A.2.5 Proof of Lemma 3

Let $S^{-1} \sim \text{Wishart}_p(\nu, I_p)$ and partition S^{-1} and $S \sim \text{inverse-Wishart}_p(\nu, I_p)$ conformably such that $p_1 + p_2 = p$:

$$S^{-1} = \begin{pmatrix} S^{11} & S^{12} \\ S^{21} & S^{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

Denote $S^{11\bullet 2} = S^{11} - S^{12}(S^{22})^{-1}S^{21}$, the Schur complement. The following are well known properties of the Wishart distribution (see, for example, Proposition 8.7 of Eaton [1983])

$$S^{22} \sim \text{Wishart}_{p_2}(I_{p_2}, \nu),\ S^{21}|S^{22} \sim N_{p_2 \times p_1}(0, S^{22} \otimes I_{p_1}),$$
$$S^{11\bullet 2} \sim \text{Wishart}_{p_1}(I_{p_1}, \nu - p_2),\ \text{and } S^{11\bullet 2} \text{ is independent of } \{S^{22}, S^{21}\}$$

The relationship of the inverse of a partitioned matrix (see, for example, Section 0.7.3 of Horn and Johnson [2013]) implies that

$$S_{11} = (S^{11\bullet 2})^{-1} \sim \text{inverse-Wishart}_{p_1}(I_{p_1}, \nu - p_2) \tag{A.9}$$
$$S_{22\bullet 1} = (S^{22})^{-1} \sim \text{inverse-Wishart}_{p_2}(I_{p_2}, \nu) \tag{A.10}$$
$$\begin{aligned} S_{21}|S_{11}, S_{22\bullet 1} &\stackrel{d}{=} -(S^{22})^{-1}S^{21}(S^{11\bullet 2})^{-1} \\ &\sim N_{p_2 \times p_1}(0, (S^{22})^{-1} \otimes (S^{11\bullet 2})^{-1}(S^{11\bullet 2})^{-1}) \\ &= N_{p_2 \times p_1}(0, S_{22\bullet 1} \otimes S_{11}S_{11}). \end{aligned} \tag{A.11}$$

It is also well known that

$$\text{if } p = 1 \text{ then } S \sim \text{ inverse-gamma}(\nu/2, 1/2). \tag{A.12}$$

We should be able to use these results to come up with the distribution of the elements of the lower triangular Cholesky decomposition from an inverse-Wishart distributed random matrix, which seems surprisingly difficult to find in the literature.

Proof of Lemma 3. We proceed by induction on the dimension. It is clearly true for $n = 1$. Assume it is true for $n - 1$. Then partition $S_{[1:n,1:n]} \sim \text{ inverse-Wishart}_n(I_n, \nu - p + n)$ such

that the top left submatrix, S_{11}, is $n-1$ by $n-1$.

$$S_{[1:n,1:n]} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & s_{22} \end{pmatrix} = \begin{pmatrix} W_1 & 0 \\ S_{21}W_1^{-T} & s_{22\bullet 1}^{1/2} \end{pmatrix} \begin{pmatrix} W_1^T & W_1^{-1}S_{12} \\ 0 & s_{22\bullet 1}^{1/2} \end{pmatrix}.$$

Note that $S_{11} = W_1 W_1^T$. Using (A.9)-(A.12), we have that:

$$\begin{aligned} W_{[n,n]}^2 = s_{22\bullet 1} &\sim \text{ inverse-gamma}\left((\nu - p + n)/2, 1/2\right) \\ S_{21}W_1^{-T}|W_1, s_{22\bullet 1} &= S_{21}S_{11}^{-1/2^T}|S_{11}, s_{22\bullet 1} \sim N_{1\times n-1}\left(0, \left(s_{22\bullet 1} \otimes W_1^T W_1\right)\right) \\ &= N_{n-1}(0, s_{22\bullet 1}W_1^T W_1) = N_{n-1}(0, W_{[n,n]}^2 W_1^T W_1). \end{aligned}$$

□

A.2.6 Proof of Lemma 4

Proof. We proceed by induction on the dimension. It is clearly true for $n = 1$. Assume it is true for $n-1$. Note that for lower triangular matrices, the $[1:n, 1:n]$ submatrix of the inverse is the inverse of the $[1:n, 1:n]$ submatrix. Hence, partition $W_{k[1:n,1:n]} = V_{k[1:n,1:n]}^{-1}$ by:

$$V_{k[1:n,1:n]} = \begin{pmatrix} V_{11} & 0 \\ V_{21} & v_{22} \end{pmatrix}, \quad W_{k[1:n,1:n]} = \begin{pmatrix} W_{11} & 0 \\ W_{21} & w_{22} \end{pmatrix},$$

where the top left submatrix is $n-1$ by $n-1$. Then $v_{22}^2 = 1/w_{22}^2$ is clearly $\chi_{\nu-n+1}^2$. We have that $V_{21} = -w_{22}^{-1}W_{21}W_{11\bullet 2}^{-1}$. Also, $W_{11\bullet 2} = W_{11} - W_{21} * 0/w_{22} = W_{11}$. Since

$$W_{21}|W_{11}, w_{22} \sim N_{n-1}\left(0, w_{22}^2 W_{11}^T W_{11}\right),$$

we have that

$$-w_{22}^{-1}W_{21}W_{11}^{-1}|W_{11}, w_{22} \sim N_{n-1}\left(0, I\right).$$

Hence, the result is proved. □

A.2.7 Proof of Proposition 1

Proof. This minimization problem is equivalent to minimizing

$$\begin{aligned} & s^2 \sum_{k=1}^{K} \frac{p}{p_k} \operatorname{tr}\left(S_k E\left[\left(\sigma^2 \Sigma_k\right)^{-1}\right]\right) - Kp \log\left(s^2\right) \\ & = s^2 \sum_{k=1}^{K} \frac{p}{p_k} \operatorname{tr}\left(S_k \mathcal{E}_k^{-1}\right) - Kp \log\left(s^2\right). \end{aligned}$$

Let us absorb the scale parameter into S_k. That is, let $\tilde{S}_k = s^2 S_k$, then $s^2 = |\tilde{S}_k|^{1/p_k}$, and we wish to minimize with respect to $\tilde{S}_k$:

$$\frac{p}{p_k} \operatorname{tr}\left(\tilde{S}_k \mathcal{E}_k^{-1}\right) - \frac{Kp}{p_k} \log\left(|\tilde{S}_k|\right) + |\tilde{S}_k|^{1/p_k} \sum_{j \neq k} \frac{p}{p_j} \operatorname{tr}\left(S_j^T \mathcal{E}_j^{-1}\right).$$

Letting $\lambda = \frac{p_k}{p} \sum_{j \neq k} \frac{p}{p_j} \operatorname{tr}\left(S_j^T \mathcal{E}_j^{-1}\right)$, this is equivalent to minimizing:

$$\operatorname{tr}\left(\tilde{S}_k \mathcal{E}_k^{-1}\right) - K \log\left(|\tilde{S}_k|\right) + |\tilde{S}_k|^{1/p_k} \lambda$$

with respect to $\tilde{S}_k$.

Since the mapping $\tilde{S}_k \mapsto \mathcal{E}_k^{-1/2} \tilde{S}_k \mathcal{E}_k^{-1/2} = \Omega$ is a bijection of the set of $p_k \times p_k$ symmetric positive definite matrices, we can write:

$$\begin{aligned} & \min_{\tilde{S}_k > 0} \left\{ \operatorname{tr}\left(\tilde{S}_k \mathcal{E}_k^{-1}\right) - K \log\left(|\tilde{S}_k|\right) + |\tilde{S}_k|^{1/p_k} \lambda \right\} \\ & = \min_{\Omega > 0} \left\{ \operatorname{tr}(\Omega) - K \log(|\Omega|) + |\Omega|^{1/p_k} \lambda^* + K \log(|\mathcal{E}_k|) \right\} \\ & = \min_{\omega_1 \geq \cdots \geq \omega_{p_k} > 0} \left\{ \sum_{i=1}^{p_k} \omega_i - K \sum_{i=1}^{p_k} \log(\omega_i) + \lambda^* \prod_{i=1}^{p_k} \omega_i^{1/p_k} \right\}, \end{aligned}$$

where $\lambda^* = \lambda |\mathcal{E}_k|^{1/p_k}$ and $\omega_1, \omega_2, \ldots, \omega_{p_k}$ are the ordered eigenvalues of Ω. Taking derivatives with respect to ω_j and setting equal to 0, we have:

$$1 - \frac{K}{\omega_j} + \frac{1}{p_k} \omega_j^{1/p_k - 1} \lambda^* \prod_{i \neq j} \omega_i^{1/p_k} = 0$$

$$\Leftrightarrow \omega_j = K - \frac{1}{p_k}\lambda^* \prod_{i=1}^{p_k} \omega_i^{1/p_k} \text{ for all } j = 1, \ldots, p_k.$$

So all of the eigenvalues have the same critical value.

Taking second derivatives, we have:

$$\begin{aligned}
&\frac{K}{\omega_j^2} - \frac{p_k - 1}{p_k^2}\lambda^* \omega^{1/p_k - 2} \prod_{i \neq j}^{p_k} \omega_j^{1/p_k} > 0 \Leftrightarrow K - \frac{p_k - 1}{p_k^2}\lambda^* \prod_{j=1}^{p_k} \omega_j^{1/p_k} > 0 \\
\Leftrightarrow &K + \frac{p_k - 1}{p_k}\left(K - \frac{1}{p_k}\lambda^* \prod_{j=1}^{p_k} \omega_j^{1/p_k} - K\right) > 0 \\
\Leftrightarrow &K + \frac{p_k - 1}{p_k}(\omega_j - K) > 0 \Leftrightarrow \frac{p_k - 1}{p_k}\omega_j + K\frac{1}{p_k} > 0.
\end{aligned}$$

Hence, by a second derivative test, this critical value is a minimizer for all ω_j. This is a global minimum since

$$\text{as } \omega_1 \to \infty \text{ we have that } \left\{\sum_{i=1}^{p_k} \omega_i - K\sum_{i=1}^{p_k} \log(\omega_i) + \lambda^* \prod_{i=1}^{p_k} \omega_i^{1/p_k}\right\} \to \infty$$

and

$$\text{as } \omega_{p_k} \to 0 \text{ we have that } \left\{\sum_{i=1}^{p_k} \omega_i - K\sum_{i=1}^{p_k} \log(\omega_i) + \lambda^* \prod_{i=1}^{p_k} \omega_i^{1/p_k}\right\} \to \infty.$$

This implies that all of the ω_j are equal. In particular, that $\omega_j = (Kp_k)/(p_k + \lambda^*)$ for all $j = 1, \ldots, p_k$. This in turn implies that Ω is a constant multiple of the identity. Thus, the $\tilde{S}_k$ that minimizes the risk given all S_j such that $j \neq k$ is:

$$\tilde{S}_k = \frac{Kp_k}{p_k + \lambda^*}\mathcal{E}_k.$$

But this means that the S_k that minimizes this risk, no matter what the other S_j's are, is $\hat{\Sigma}_k = \mathcal{E}_k/|\mathcal{E}_k|^{1/(p_k)}$.

It remains to minimize with respect to s. The minimizer is the s such that

$$2s\sum_{k=1}^{K} \frac{p}{p_k} \operatorname{tr}\left(\hat{\Sigma}_k \mathcal{E}_k^{-1}\right) - \frac{2Kp}{s} = 0.$$

And solving for s we get

$$\hat{\sigma}^2 = \frac{K}{\sum_{k=1}^{K} \frac{1}{p_k} \operatorname{tr}\left(\hat{\Sigma}_k \mathcal{E}_k^{-1}\right)}.$$

But since $\hat{\Sigma}_k = \mathcal{E}_k / |\mathcal{E}_k|^{1/(p_k)}$, we have that

$$\hat{\sigma}^2 = \frac{K}{\sum_{k=1}^{K} |\mathcal{E}_k|^{-1/p_k}}.$$

□

A.2.8 Proof of Proposition 3

Proof. Let $\Phi_k = \Sigma_k / \operatorname{tr}(\Sigma_k)$, $D_k = S_k / \operatorname{tr}(S_k)$ for $k = 1, \ldots, K$. So $\Sigma_k = \Phi_k / |\Phi_k|^{1/p_k}$ and $S_k = D_k / |D_k|^{1/p_k}$ for $k = 1, \ldots, K$. Φ_k and D_k both have trace 1. The space of trace 1 symmetric positive definite matrices is convex. Let $\Phi = (\sigma^2, \Phi_1, \ldots, \Phi_K)$ and $D = (s^2, D_1, \ldots, D_K)$. Define

$$L_2(\Phi, D) = \frac{s^2}{\sigma^2} \sum_{k=1}^{K} \frac{p}{p_k} |D_k \Phi_k^{-1}|^{-1/p_k} \operatorname{tr}\left(D_k \Phi_k^{-1}\right) - Kp \log\left(\frac{s^2}{\sigma^2}\right) - Kp.$$

So, $L_M(\Sigma, S) = L_2(\Phi, D)$.

Hence, $E[L_M(\Sigma, S) | X] = E[L_2(\Phi, D) | X]$.

So if L_2 is convex in each D_k, we can uniformly decrease the risk. That is, given $B_k, E_k \in \mathcal{G}_{p_k}^+$ are two estimators from two different special linear group transformations, an estimator that uniformly decreases the risk is found by setting $F_k = (B_k / \operatorname{tr}(B_k) + E_k / \operatorname{tr}(E_k))/2$ and using $F_k / |F_k|^{1/p_k}$ as our estimator. Averaging over the whole space of orthogonal matrices will result in an orthogonally equivariant estimator.

It remains to prove that L_2 is convex in each D_k. It suffices to show that $|D_k|^{-1/p_k} \operatorname{tr}(D_k \Phi_k^{-1})$ is convex in D_k. Since, for $\alpha \in [0, 1]$, $\operatorname{tr}((\alpha D_k + (1-\alpha) E_k)\Phi_k^{-1}) = \alpha \operatorname{tr}(D_k \Phi_k^{-1}) + (1-\alpha) \operatorname{tr}(E_k \Phi_k^{-1})$ is convex in D_k, if $|D_k|^{-1/p_k}$ is also convex, then we are done. We have $\log(|D_k|)$ is a concave function [Cover and Thomas, 1988, Theorem 1], so $-\log(|D_k|)/p_k$ is convex, so $\exp(-\log(|D_k|)/p_k) = |D_k|^{-1/p_k}$ is convex.

We also have that $cb^2 - h\log(b^2)$ is convex in b^2 for $c, h > 0$, so we can average the scale estimates to decrease risk as well.

To summarize, we have:

$$\begin{aligned}
&L_M\left(\Sigma, \left(f^2, F_1/|F_1|^{1/p_1}, \ldots, F_K/|F_K|^{1/p_K}\right)\right)\\
=&L_2\left(\Phi, \left(f^2, F_1, \ldots, F_K\right)\right)\\
=&L_2\left(\Phi, \left((b^2+e^2)/2, B_1/\operatorname{tr}(B_1) + E_1/\operatorname{tr}(E_1)\right)/2\right.,\\
&\left.\ldots, (B_K/\operatorname{tr}(B_K) + E_K/\operatorname{tr}(E_K))/2)\right)\\
\leq&\frac{1}{2}L_2\left(\Phi, \left(b^2, B_1/\operatorname{tr}(B_1), \ldots, B_K/\operatorname{tr}(B_K)\right)\right)\\
&+\frac{1}{2}L_2\left(\Phi, \left(e^2, E_1/\operatorname{tr}(E_1), \ldots, E_K/\operatorname{tr}(E_K)\right)\right)\\
=&\frac{1}{2}L_M(\Sigma, B) + \frac{1}{2}L_M(\Sigma, E).
\end{aligned}$$

If B and E have the same (constant) risk as the UMREE, $\hat{\Sigma}(X)$, then

$$\begin{aligned}
&E\left[L_M\left(\Sigma, \left(f^2, F_1/|F_1|^{1/p_1}, \ldots, F_K/|F_K|^{1/p_K}\right)\right)\right]\\
&\leq \frac{1}{2}E\left[L_M(\Sigma, B)\right] + \frac{1}{2}\left[L_M(\Sigma, E)\right]\\
&= E\left[L_M\left(\Sigma, \hat{\Sigma}(X)\right)\right]
\end{aligned}$$

□

A.3 Simplification of the divergence

We will need the $(i_1, \ldots, i_K)$th element of $U^T \cdot df[\Delta^{\mathbf{i}}]$ in (4.32). There are three terms in (4.32). We will deal with them one by one. First, we will work with the first term of (4.32), $\sum_{k=1}^{K} d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(D) \cdot \mathcal{V}$. Note that, for $\mathcal{A} = f(D) \cdot \mathcal{V}$, we have

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

$$= -\sum_{j=1,j\neq i_k}^{p_k}\left(\prod_{\ell=1,\ell\neq k}^{K} f_{i_\ell}^{\ell}(\sigma_{i_\ell}^{\ell})\right) f_j^k(\sigma_j^k)\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}\mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}/[(\sigma_{i_k}^k)^2-(\sigma_j^k)^2]$$

$$= -\left(\prod_{\ell=1,\ell\neq k}^{K} f_{i_\ell}^{\ell}(\sigma_{i_\ell}^{\ell})\right)\sum_{j=1,j\neq i_k}^{p_k} f_j^k(\sigma_j^k)\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}\mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}/[(\sigma_{i_k}^k)^2-(\sigma_j^k)^2].$$

Now we work with the second term of (4.32), $\sum_{k=1}^{K} df(\tilde{D})_k[\Delta^{\mathbf{i}}]\cdot\mathcal{V}$. We have that:

$$\left(df(\tilde{D})_k[\Delta^{\mathbf{i}}]\cdot\mathcal{V}\right)_{[\mathbf{i}]} = \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j)\right) d(f_{i_k}^k\circ\sigma_{i_k}^k)[\Delta^{\mathbf{i}}]\mathcal{V}_{[\mathbf{i}]} \tag{A.13}$$

$$= \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j)\right)\left(\frac{d}{d\sigma_{i_k}^k} f_{i_k}^k(\sigma_{i_k}^k)\right)\mathcal{V}_{[\mathbf{i}]}\mathcal{S}_{[\mathbf{i}]}/\sigma_{i_k}^k$$

$$= \left(\prod_{j\neq k} f_{i_j}^j(\sigma_{i_j}^j)/\sigma_{i_j}^j\right)\left(\frac{d}{d\sigma_{i_k}^k} f_{i_k}^k(\sigma_{i_k}^k)\right)\mathcal{S}_{[\mathbf{i}]}^2/(\sigma_{i_k}^k)^2, \tag{A.14}$$

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 (4.32), $f(D)\cdot d\mathcal{V}[\Delta^{\mathbf{i}}]$. We have:

$$\left(f(D)\cdot d\mathcal{V}[\Delta^{\mathbf{i}}]\right)_{[\mathbf{i}]} = \left(\prod_{k=1}^{K} f_{i_k}^k(\sigma_{i_k}^k)\right) d\mathcal{V}[\Delta^{\mathbf{i}}]_{[\mathbf{i}]}. \tag{A.15}$$

We now need to obtain $d\mathcal{V}[\Delta^{\mathbf{i}}]_{[\mathbf{i}]}$. From (4.25), we have

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

$$= D^{-1}\cdot E^{\mathbf{i}} - \sum_{k=1}^{K} dF_k[\Delta^{\mathbf{i}}]\cdot\mathcal{V} - \sum_{k=1}^{K} dG_k[\Delta^{\mathbf{i}}]\cdot\mathcal{V}. \tag{A.16}$$

There are three terms in (A.16). Let us deal with them one by one. The first term in (A.16) is

$$\left(D^{-1}\cdot E^{\mathbf{i}}\right)_{[\mathbf{i}]} = \left(\prod_{k=1}^{K}\sigma_{i_k}^k\right)^{-1}. \tag{A.17}$$

The second term in (A.16) is

$$\left(dF_k[\Delta^{\mathbf{i}}]\cdot\mathcal{V}\right)_{[\mathbf{i}]}$$

$$
\begin{aligned}
&= \left((I_{p_1}, \ldots, I_{p_{k-1}}, D_k^{-1}\Omega_{U_k}[\Delta^{\mathbf{i}}]D_k, I_{p_{k+1}}, \ldots, I_{p_K}) \cdot \mathcal{V}\right)_{[\mathbf{i}]} \\
&= \sum_{j=1}^{p_k} \left(D_k^{-1}\Omega_{U_k}[\Delta^{\mathbf{i}}]D_k\right)_{[i_k,j]} \mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \\
&= -\sum_{j=1, j\neq i_k}^{p_k} \frac{\sigma_j^k}{\sigma_{i_k}^k} S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} / [(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2] \\
&= -\sum_{j=1, j\neq i_k}^{p_k} \frac{\sigma_j^k}{\sigma_{i_k}^k} S_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} \mathcal{V}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]} / [(\sigma_{i_k}^k)^2 - (\sigma_j^k)^2]. \qquad \text{(A.18)}
\end{aligned}
$$

The third term in (A.16) is

$$
\begin{aligned}
\left(dG_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}\right)_{[\mathbf{i}]} &= \left(\mathcal{V} \times_k D_k^{-1} dD_k[\Delta^{\mathbf{i}}]\right)_{[\mathbf{i}]} \\
&= d\sigma_{i_k}^k[\Delta] \mathcal{V}_{[\mathbf{i}]} / \sigma_{i_k}^k \\
&= \mathcal{S}_{[\mathbf{i}]} \mathcal{V}_{[\mathbf{i}]} / (\sigma_{i_k}^k)^2. \qquad \text{(A.19)}
\end{aligned}
$$

To obtain the third term in (4.32), we need only plug in (A.17), (A.18), and (A.19) into (A.16). And then we need to plug in (A.16) into (A.15).

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

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

for H_k in (4.29) and $\mathcal{C} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ in (4.30). The term $f(D) \cdot D^{-1} \cdot \mathcal{C}$ is from the first and second parts of (4.32), whereas the terms $\sum_{k=1}^{K} H_k \cdot \mathcal{S}^2$ are from the second part of (4.32) and were already derived in (A.14). 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 (4.32), we have that the sum of the first and third terms in (4.32) is equal

to:

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

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,

$$\mathcal{C}_{[\mathbf{i}]}=1+\sum_{k=1}^{K}\sum_{j=1,j\neq i_k}^{p_k}\frac{\mathcal{S}^2_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}}{(\sigma^k_{i_k})^2-(\sigma^k_j)^2}-\mathcal{S}^2_{[\mathbf{i}]}\sum_{k=1}^{K}\left(\frac{1}{(\sigma^k_{i_k})^2}+\sum_{m=1,m\neq i_k}^{p_k}\frac{1}{(\sigma^k_m)^2-(\sigma^k_{i_k})^2}\right). \tag{A.20}$$

A.4 Newton step for optimization

Let $f_{\mathbf{i}}=\prod_{k=1}^{K}f^k_{i_k}(\sigma^k_{i_k})$ and $\tilde{\sigma}_{\mathbf{i}}=\prod_{k=1}^{K}\sigma^k_{i_k}$. The SURE is equal to:

$$||f(D)\cdot D^{-1}\cdot\mathcal{S}-\mathcal{S}||^2+2\tau^2\sum_{\mathbf{i}}\left[\left(f(D)\cdot D^{-1}\cdot\mathcal{C}\right)_{[\mathbf{i}]}+\sum_{k=1}^{K}\left(H_k\cdot\mathcal{S}^2\right)_{[\mathbf{i}]}\right]-p\tau^2 \tag{A.21}$$

$$= \sum_{\mathbf{i}} \left[\left(f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]} - \mathcal{S}_{[\mathbf{i}]} \right)^2 + 2\tau^2 f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{C}_{[\mathbf{i}]} + 2\tau^2 f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]}^2 \sum_{k=1}^{K} \frac{\frac{d}{d\sigma_{i_k}^k} f_{i_k}^k(\sigma_{i_k}^k)}{\sigma_{i_k}^k f_{i_k}^k(\sigma_{i_k}^k)} \right] - p\tau^2. \quad \text{(A.22)}$$

Considering only the mode-specific soft-thresholding estimator, we need to calculate the gradient given $\mathbf{i}$. One can show that if any $\lambda_k \geq \sigma_{i_k}^k$, then its contribution to the gradient is zero. Hence, given $\mathbf{i}$, assume that all $\lambda_k < \sigma_{i_k}^k$ for $k = 1, \ldots, K$. Then $(\sigma_{i_k}^k - \lambda_k)_+ = \sigma_{i_k}^k - \lambda_k$ and $\frac{d}{d\sigma_{i_k}^k} f_{i_k}^k(\sigma_{i_k}^k) = 1$. We have

$$\begin{aligned}
&\frac{d}{d\lambda_k} \left[\left(c f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]} - \mathcal{S}_{[\mathbf{i}]} \right)^2 + 2\tau^2 c f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{C}_{[\mathbf{i}]} + 2\tau^2 c f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]}^2 \sum_{k=1}^{K} \frac{1}{\sigma_{i_k}^k f_{i_k}^k(\sigma_{i_k}^k)} \right] \\
&= \frac{d}{d\lambda_k} \left[c^2 f_{\mathbf{i}}^2 \tilde{\sigma}_{\mathbf{i}}^{-2} \mathcal{S}_{[\mathbf{i}]}^2 - 2c f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]}^2 + 2\tau^2 c f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{C}_{[\mathbf{i}]} + 2\tau^2 c f_{\mathbf{i}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]}^2 \sum_{k=1}^{K} \frac{1}{\sigma_{i_k}^k f_{i_k}^k(\sigma_{i_k}^k)} \right] \\
&= -2c^2 (\sigma_{i_k}^k - \lambda_k) f_{\mathbf{i}_{-k}}^2 \tilde{\sigma}_{\mathbf{i}}^{-2} \mathcal{S}_{[\mathbf{i}]}^2 + 2c f_{\mathbf{i}_{-k}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]}^2 - 2\tau^2 c f_{\mathbf{i}_{-k}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{C}_{[\mathbf{i}]} \\
&- 2\tau^2 c f_{\mathbf{i}_{-k}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]}^2 \sum_{j=1, j \neq k}^{K} \frac{1}{\sigma_{i_j}^j f_{i_j}^j(\sigma_{i_j}^j)} \\
&= -2c^2 \sigma_{i_k}^k f_{\mathbf{i}_{-k}}^2 \tilde{\sigma}_{\mathbf{i}}^{-2} \mathcal{S}_{[\mathbf{i}]}^2 + \lambda_k 2c^2 f_{\mathbf{i}_{-k}}^2 \tilde{\sigma}_{\mathbf{i}}^{-2} \mathcal{S}_{[\mathbf{i}]}^2 + 2c f_{\mathbf{i}_{-k}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]}^2 - 2\tau^2 c f_{\mathbf{i}_{-k}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{C}_{[\mathbf{i}]} \\
&- 2\tau^2 c f_{\mathbf{i}_{-k}} \tilde{\sigma}_{\mathbf{i}}^{-1} \mathcal{S}_{[\mathbf{i}]}^2 \sum_{j=1, j \neq k}^{K} \frac{1}{\sigma_{i_j}^j f_{i_j}^j(\sigma_{i_j}^j)},
\end{aligned} \quad \text{(A.23)}$$

where $f_{\mathbf{i}_{-k}} = \prod_{j=1, j \neq k}^{K} f_{i_j}^j(\sigma_{i_j}^j)$. Let

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

where we are summing over the set of i_k's such that $\sigma^k_{i_k} > \lambda_k$ for $k = 1, \ldots, K$. Then the gradient (A.23) is equal to

$$-a + \lambda_k b + d - e - h.$$

We now have a Newton step for a gradient descent algorithm:

$$\begin{aligned}\lambda_k^{NEW} &= \lambda_k - (-a + \lambda_k b + d - e - h)/b \\ &= \lambda_k - \lambda_k + (a - d + e + h)/b \\ &= (a - d + e + h)/b\end{aligned}$$

To update c, we have

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

Let

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

where we are summing over the set of i_k's such that $\sigma^k_{i_k} > \lambda_k$ 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 $a > 0$.

A.5 General spectral functions

In Section 4.3.1, we assumed that the spectral functions were of the form:

$$f^k(D_k) = \text{diag}(f^k_1(\sigma^k_1), \ldots, f^k_{p_k}(\sigma^k_{p_k})).$$

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 (4.7). 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

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

The derivation of the SURE is the same as in Section 4.3.1 except for the second term in (4.32):

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

We have:

$$\begin{aligned}\left(df(\tilde{D})_k[\Delta^{\mathbf{i}}] \cdot \mathcal{V}\right)_{[\mathbf{i}]} &= \left(\prod_{j \neq k} f_{i_j}^j(\sigma_{i_j}^j)\right) d(f^k \circ D_k)[\Delta^{\mathbf{i}}]_{[i_k, i_k]} \mathcal{V}_{[\mathbf{i}]} \\ &= \left(\prod_{j \neq k} f_{i_j}^j(\sigma_{i_j}^j)\right) d(g^k \circ \mathbf{s}_k)[\Delta^{\mathbf{i}}]_{[i_k]} \mathcal{V}_{[\mathbf{i}]} \end{aligned} \tag{A.24}$$

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 (4.37) 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

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

And so

$$\begin{aligned} d(g^k \circ \mathbf{s}_k)[\Delta^{\mathbf{i}}]_{[i_k]} &= (J_{g^k}(\mathbf{s}_k)d\mathbf{s}_k[\Delta])_{[i_k]} \\ &= J_{g^k}(\mathbf{s}_k)_{[i_k,i_k]}S_{[\mathbf{i}]}/\sigma^k_{i_k}. \end{aligned} \tag{A.25}$$

Inserting (A.25) into (A.24), we get:

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

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

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

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

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

The divergence is now of the form:

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

A.6 SURE for estimators that shrink elements in $\mathcal{S}$

Consider the HOSVD (4.11). In this section, we will find the SURE for estimators of the form:

$$t(\mathcal{X}) = U\cdot g(\mathcal{S}), \tag{A.26}$$

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 [Henrion, 1993, Kiers et al., 1997, Murakami et al., 1998, Andersson and Henrion, 1999, De Lathauwer et al., 2001, Martin and Van Loan, 2008]. 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 17. *The differentials of U_k and $\mathcal{S}$ are given in equations (4.21) and (A.27), respectively.*

Proof. We have already calculated $dU_k[\Delta]$ in Theorem 13. To obtain $d\mathcal{S}[\Delta]$, we apply the chain rule to the HOSVD (4.11) 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 (4.23). Hence,

$$d\mathcal{S}[\Delta] = U^T \cdot \Delta - \sum_{k=1}^{K} d\tilde{U}_k[\Delta] \cdot \mathcal{S} \tag{A.27}$$

where $d\tilde{U}_k[\Delta]$ is defined in (4.33). □

The derivation of the divergence for functions of the form (A.26) is very similar to that in Section 4.3.2. The divergence may still be found from (4.31). 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 (4.23). Hence,

$$U^T \cdot dt[\Delta^{\mathbf{i}}] = \sum_{k=1}^{K} d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot g(\mathcal{S}) + d(g \circ \mathcal{S})[\Delta^{\mathbf{i}}], \tag{A.28}$$

where $d\tilde{U}_k[\Delta^{\mathbf{i}}]$ is from (4.33), noting that the relationship in (4.36) 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

$$\begin{aligned}
&\left(U^T \cdot df[\Delta^{\mathbf{i}}]\right)_{[\mathbf{i}]} \\
&= \left(\sum_{k=1}^{K} d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S}) + d(f \circ \mathcal{S})[\Delta^{\mathbf{i}}]\right)_{[\mathbf{i}]} \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) d\mathcal{S}_{[\mathbf{i}]}[\Delta^{\mathbf{i}}] \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) d\mathcal{S}[\Delta^{\mathbf{i}}]_{[\mathbf{i}]} \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(\left(U^T \cdot \Delta^{\mathbf{i}}\right)_{[\mathbf{i}]} - \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot \mathcal{S}\right)_{[\mathbf{i}]}\right) \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(E^{\mathbf{i}}_{[\mathbf{i}]} - \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot \mathcal{S}\right)_{[\mathbf{i}]}\right) \\
&= \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot f(\mathcal{S})\right)_{[\mathbf{i}]} + \left(\frac{d}{d\mathcal{S}_{[\mathbf{i}]}} f_{\mathbf{i}}(\mathcal{S}_{[\mathbf{i}]})\right) \left(1 - \sum_{k=1}^{K} \left(d\tilde{U}_k[\Delta^{\mathbf{i}}] \cdot \mathcal{S}\right)_{[\mathbf{i}]}\right).
\end{aligned} \tag{A.29}$$

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

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

Hence, from (A.29) we have,

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

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

$$=\sum_{\mathbf{i}}\Bigg[-\sum_{k=1}^{K}\sum_{j=1,j\neq i_k}^{p_k}\frac{\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}f_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]}(\mathcal{S}_{[i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_K]})}{(\sigma^k_{i_k})^2-(\sigma^k_j)^2} \tag{A.30}$$

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

We can rearrange the summations in the left part of (A.30) by switching the order of the j and the i_k and then altering the notation of the dummy variables to obtain:

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

Hence, the SURE for these higher-order spectral functions (A.26) is:

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

VITA

David Gerard grew up in Columbus, Ohio. He liked it enough there to go to the Ohio State University, both for a Bachelors in Mathematics and Molecular Genetics and for a Masters in Statistics. In June 2015, he obtained a PhD in Statistics from the University of Washington. He enjoys football, coffee, and board games.