Analysis of Four and Five-Way Data and Other Topics in Clustering

ANALYSIS OF FOUR AND FIVE-WAY DATA AND OTHER TOPICS IN CLUSTERING

BY

PETER A. TAIT, M.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2021)	McMaster University
(Mathematics & Statistics)	Hamilton, Ontario, Canada

TITLE:	Analysis of Four and Five-Way Data and Other Topics in Clustering
AUTHOR:	Peter A. Tait M.Sc., (Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	xvii, 139

To my partner in life, Kate Languedoc and our three boys; Xavier, Sam and Lochlan.

Abstract

Clustering is the process of finding underlying group structure in data. As the scale of data collection continues to grow, this “big data” phenomenon results in more complex data structures. These data structures are not always compatible with traditional clustering methods, making their use problematic. This thesis presents methodology for analyzing samples of four-way and higher data, examples of these more complex data types. These data structures consist of samples of continuous data arranged in multidimensional arrays. A large emphasis is placed on clustering this data using mixture models that leverage tensor-variate distributions to model the data. Parameter estimation for all these methods are based on the expectation-maximization algorithm. Both simulated and real data are used for illustration.

Acknowledgments

I would like to start by thanking my supervisor, Dr. Paul McNicholas. His interminable guidance, encouragement, patience and dedication to research have been crucial for the completion of this work and my professional development.

I would also like to show my appreciation to Dr. Sharon McNicholas, and Dr. Antoine Deza who, along with Dr. Paul McNicholas, were on my supervisory committee. We have shared some stimulating conversations. I would also like to thank Dr. Hugh Chipman who served as my external examiner, and Dr. Bartosz Protas who served as the chair for the defence.

I would like to thank my partner Kate, for her encouragement, love and unceasing ability to multi-task, which gave me the time to complete this work. To our boys, ”Get After It”.

Finally, I would like to thank my parents, Bob and Susan, for their love, patience and unending support throughout my education.

Notation

The following is a summary of the mathematical notation used herein:

- $\mathbf{V}$: A matrix.
- $\boldsymbol{\mathfrak{V}}$: An order-D multidimensional array(MDA).
- $\text{vec}(\cdot)$: Vectorization of a matrix or MDA.
- $d \in \{1, 2, \cdots, D\}$.
- $\boldsymbol{\Delta}_d$: scale matrix for dimension/mode d of a MDA.
- $\otimes$: Kronecker product.
- $\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d = \boldsymbol{\Delta}_1 \otimes \boldsymbol{\Delta}_2 \otimes \cdots \otimes \boldsymbol{\Delta}_D$
- $n_1 \times n_2 \times \cdots \times n_D = \mathbf{n}$: Dimensional lengths of a MDA.
- $n^* = \prod_{d=1}^{D} n_d$: The product of all the MDA dimensional lengths.
- $n_g = \sum_{i=1}^{N} z_{ig}$.
- $\|\mathbf{X}\|$: matrix Frobenius norm.
- $\|\boldsymbol{\mathfrak{X}}\|$: array norm.

- $\boldsymbol{\mathfrak{X}} \times_d \boldsymbol{\Delta}_d$: a d-mode matrix product.

- $\boldsymbol{\mathfrak{X}} \times \widehat{\boldsymbol{\Delta}}$: a Tucker product.

- $\mathbf{e}_{i_*}^{n_*}$ is a unit basis vector of length n_*, with the 1 at index i_*.

- $\mathbf{X}_{(1)}$ is a $\prod_{d=2}^{D} n_d \times n_1$ matricization of $\boldsymbol{\mathfrak{X}}$.

- $\breve{\mathbf{X}}_{(1)} = \mathbf{X}_{(1)} - \mathbf{M}_{(1)}$

- $\mathbf{X}_{(1)j} = \left(\mathbf{I}_{n_2} \otimes \mathbf{e}_j^{\top} \bigotimes_{d=3}^{D} \boldsymbol{\Delta}_d^{-\frac{\top}{2}}\right) \breve{\mathbf{X}}_{(1)}$

- $n_{3:D}^{*} = \prod_{d=3}^{D} n_d$

- $\breve{\mathbf{X}}_{(1)gi} = \mathbf{X}_{(1)i} - \mathbf{M}_{(1)g}$

- $\mathbf{X}_{(1)gij} = \left(\mathbf{I}_{n_2} \otimes \mathbf{e}_j^{\top} \bigotimes_{d=3}^{D} \boldsymbol{\Delta}_{gd}^{-\frac{\top}{2}}\right) \breve{\mathbf{X}}_{(1)gi}$

- $\mathbf{A}_{(1)gj} = \left(\mathbf{I}_{n_2} \otimes \mathbf{e}_j^{\top} \bigotimes_{d=3}^{D} \boldsymbol{\Delta}_{gd}^{-\frac{\top}{2}}\right) \mathbf{A}_{(1)g}$

- The superscript $l2$ $\left(\text{e.g. } \mathbf{X}_{(1)gij}^{l2}\right)$ indicates the exchanges of the second and l^{th} elements in a $\otimes$ sequence.

- $n_{2:D/l}^{*} = \prod_{\substack{d=2 \\ d\neq l}}^{D} n_d$: The product of all the MDA dimensional lengths from 2 to D, excluding the l^{th}.

- $\bigotimes_{\substack{d=3 \\ d\neq n}}^{D} \boldsymbol{\Delta}_d$ is the sequence of $\otimes$'s from 3 to D, excluding the n^{th}.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Data and Clustering

New data collection technologies, such as electronic sensors and wearables, are producing rich sources of multivariate data. Such data can be organized in different ways; as vectors, matrices or multidimensional arrays (MDA), a higher order generalization of a matrix. These data structures can be viewed as a hierarchy, where vectors are combined to form matrices and matrices are stacked to form MDAs. A three dimensional MDA can be visualized as a cube formed by stacking matrices, one on top of the other. Historically, data has taken the form of vectors and could be analyzed in a straightforward way, with off-the-shelf statistical techniques. Due to the "big data" phenomenon, these everyday statistical techniques are not well suited for the increased dimensionality and complexity of modern data sources (Bouveyron and Brunet-Saumard, 2014).

Clustering, the practice of elucidating concealed group structure in data, is central to machine learning, computational statistics and exploratory data analysis (Hastie *et al.*, 2009). Modernizing clustering methodology to handle higher order data sources is the prime motivation for the work outlined herein. These modernizations could make important contributions to many scientific problems, e.g.: recommender systems, clustering spatio-temporal processes in neuroimaging studies, and creating subgroups of patients with similar sensor data patterns in clinical trials. Our methodological development, adapt model-based clustering, a popular clustering technique in the literature, to MDA data. This approach relies on the finite mixture model (see Section 2.2) which expects that each observation arises from one of a number, G say, of probability distributions. Using a suitable probability distribution (see Section 2.4 and Chapter 5), we can natively model high-dimensional data, in the form of MDAs.

1.2 Thesis Outline

1.2.1 Chapter 2

Chapter 2 presents the necessary background to understand MDAs, the multilinear normal distribution, model-based clustering and the expectation-maximization (EM) algorithm used to estimate its model parameters. Additionally we lay the ground work for Chapter 5, by detailing how the variance-mean mixture is used to formulate skewed multivariate distributions.

1.2.2 Chapter 3

Chapter 3 outlines a finite mixture of multilinear normal distributions, including parameter estimation, model selection and a simulation study detailing its effectiveness in a variety of sample and MDA sizes. This material is available on arXiv (Tait *et al.*, 2020).

1.2.3 Chapter 4

Chapter 4 describes a framework for adding constraints to the scale matrices in the finite mixture of multilinear normal distributions, making them suitable to model longitudinal data. An applied example is described, were the five-way data is taken from accelerometers used to characterize patterns of physical activity in children. This material is available on arXiv (Tait *et al.*, 2020).

1.2.4 Chapter 5

Chapter 5 presents five new tensor-variate skewed distributions, how to estimate their parameters, a simulation study detailing how they perform across a variety of sample and MDA sizes and an applied analysis of maple tree images.

1.2.5 Chapter 6

Chapter 6 outlines a finite mixture model of the five tensor-variate skewed distributions in Chapter 5. We discuss parameter estimation and show simulation results detailing the models performance on a range of sample and MDA sizes.

1.2.6 Chapter 7

Chapter 7 some concluding remarks and outlines some areas of future research.

Chapter 2

Background

2.1 Clustering

Clustering is a form of unsupervised learning (Hastie *et al.*, 2013), where the goal is to find labels for the observations when known labels are not available — or we behave as if there are no known labels. The labels indicate an observation's membership in a cluster or group. Clustering is also known as unsupervised classification. Many definitions of a cluster have been proposed. We will define a cluster as a unimodal component within a finite mixture model that is appropriate for the data being analyzed (McNicholas, 2016a).

2.2 Finite Mixture Models

One common method of clustering is referred to as model-based clustering, which assumes a random variable $\mathbf{X}$ originates from a population with G separate sub-populations. It is *a pripori* unknown to which of the G sub-populations $\mathbf{X}$ comes, and often the value of G is also unknown. If the number of sub-populations G is finite, the mixture model density of $\mathbf{x}$, a realization of $\mathbf{X}$, is given by,

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g), \tag{2.1}$$

where the π_g are called the mixing proportions and have the following two constraints: $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$. The $f_g(\cdot)$ are the component densities, and $\boldsymbol{\vartheta} = (\pi_1, \pi_2, \ldots, \pi_G, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_G)$ are all the parameters of the mixture model. Overviews of finite mixture models can be found in Fraley and Raftery (2002), McLachlan and Peel (2000a) and McNicholas (2016a,b).

In practice, the normal mixture model has been used most frequently. Early works using the normal mixture models include Wolfe (1965), Baum *et al.* (1970) and Scott and Symons (1971). This early adoption was due to the normal distribution's attractive mathematical properties. The $f_g(\mathbf{x}|\boldsymbol{\theta}_g)$ has a density drawn from the multivariate normal distribution, i.e.,

$$f_g(\mathbf{x}|\boldsymbol{\theta}_g) = f_g(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Delta}_g) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Delta}_g|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)^\top \boldsymbol{\Delta}_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g)\right\}. \tag{2.2}$$

Here $\mathbf{x} \in \mathbb{R}^{p\times 1}$, $\boldsymbol{\mu}_g$ is the mean vector and $\boldsymbol{\Delta}_g$ is the covariance matrix of the distribution.

2.3 Tensors

Tensors are higher order generalizations of matrices. While some might refer to such structures as 'tensors', and so write about clustering tensor-variate data, we prefer the nomenclature multidimensional array (MDA) to avoid confusion with the term ‘tensor’ as used in engineering and physics, e.g., tensor fields.

The number of dimensions an MDA has is referred to as its order. An order-D MDA is equivalent, in our sense, to a D-dimensional array — the $D = 2$ structure is a matrix, the $D = 3$ structure can be regarded as a rectangular cuboid. A rectangular cuboid is defined as a three-dimensional box, where all the angles are right angles, all faces are rectangles, and opposite faces are equal (Harris and Stöcker, 1998). A $D = 4$ structure can be viewed as stacked rectangular cuboids. MDAs can be partitioned into slices or matrices which are two-dimensional sections of a MDA. This is done by fixing all but two dimensions of the MDA. Fibers or vectors are created from an MDA by fixing all but one dimension (Kolda and Bader, 2009). In general, $(D + 1)$-way data can be represented using a sample of D-dimensional MDAs. Herein, we restrict ourselves to MDA data that can be regarded as the realization of continuous random variables.

For example, four-way data can be a useful way to represent repeated measurements of various attributes over time at different years of a study. This is a common scenario in biostatistics, where multiple physical and health traits (physical activity, blood pressure, enzyme levels, etc.) are routinely measured over short periods of time (e.g., one week) and these short periods may be repeated over a longer period of time (e.g., annually).

2.4 Tensor-Variate Normal Distribution

As with the univariate, multivariate and matrix variate cases, the tensor-variate normal distribution (TVND; Hoff *et al.*, 2011), also known as the multilinear normal distribution (MLND; Ohlson *et al.*, 2013), is the most well-known tensor-variate distribution. If $\mathscr{X}$ is a random order-D MDA, following an MLND, with dimensional lengths

$$n_1 \times n_2 \times \cdots \times n_D = \mathbf{n}$$

and realization $\boldsymbol{\mathfrak{X}}$ — which we denote by $\mathscr{X} \sim \mathcal{N}_{\mathbf{n}}\left(\boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d\right)$ — its density can be written

$$\begin{aligned} f(\boldsymbol{\mathfrak{X}}|\boldsymbol{\mathfrak{M}}, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_D) = (2\pi)^{\frac{-n^*}{2}} \prod_{d=1}^{D} |\boldsymbol{\Delta}_d|^{-\frac{n^*}{2n_d}} \\ \times \exp\left\{-\frac{1}{2}\text{vec}(\boldsymbol{\mathfrak{X}} - \boldsymbol{\mathfrak{M}})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}\text{vec}(\boldsymbol{\mathfrak{X}} - \boldsymbol{\mathfrak{M}})\right\}, \end{aligned} \tag{2.3}$$

where vec$(\cdot)$ is the tensor vectorization operator, $\boldsymbol{\mathfrak{M}}$ is the mean MDA,

$$\text{Cov}(\text{vec}(\mathcal{X})) = \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d,$$

$n^* = \prod_{d=1}^{D} n_d$, and

$$\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d = \boldsymbol{\Delta}_1 \otimes \boldsymbol{\Delta}_2 \otimes \ldots \otimes \boldsymbol{\Delta}_D,$$

where $\otimes$ represents the Kronecker product (Ohlson *et al.*, 2013).

One important property is that the exponent in the density function (2.3) can be rewritten as

$$\text{tr}\left[\boldsymbol{\Delta}_1^{-1}\breve{\mathbf{X}}_{(1)}^\top \bigotimes_{d=2}^{D} \boldsymbol{\Delta}_d^{-1}\breve{\mathbf{X}}_{(1)}\right] \tag{2.4}$$

where $\mathbf{V}_{(j)}$ is the matricization along mode j for a MDA $\boldsymbol{\mathfrak{V}}$ and $\breve{\mathbf{X}}_{(1)} = \mathbf{X}_{(1)} - \mathbf{M}_{(1)}$.

This trace can be further decomposed into

$$\sum_{j}^{n^*_{3:D}} \operatorname{tr}\left[\boldsymbol{\Delta}_1^{-1}\mathbf{X}_{(1)j}^{\top}\boldsymbol{\Delta}_2^{-1}\mathbf{X}_{(1)j}\right] \tag{2.5}$$

or

$$\sum_{j}^{n^*_{2:D/n}} \operatorname{tr}\left[\boldsymbol{\Delta}_1^{-1}\left(\mathbf{X}_{(1)j}^{n2}\right)^{\top}\boldsymbol{\Delta}_n^{-1}\mathbf{X}_{(1)j}^{n2}\right], \tag{2.6}$$

where

$$\mathbf{X}_{(1)j} = \left(\mathbf{I}_{n_2} \otimes \mathbf{e}_j^{\top} \bigotimes_{d=3}^{D} \boldsymbol{\Delta}_d^{-\frac{\top}{2}}\right) \breve{\mathbf{X}}_{(1)}$$

and $n^*_{k:D} = \prod_{d=k}^{D} n_d$ for $k \in \{2,3\}$. The details pertaining to these traces are available in Appendices A.2 and A.3.

Note that, if $\mathcal{X}$ is a $m \times n$ random matrix then

$$\mathcal{X} \sim \mathcal{N}_{m\times n}(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) \iff \operatorname{vec}(\mathcal{X}) \sim \phi_{mn}(\operatorname{vec}(\mathbf{M}), \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}), \tag{2.7}$$

where $\mathcal{N}_{m\times n}(\cdot)$ represents the matrix variate normal distribution with mean matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, row covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$, column covariance matrix $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ and $\phi_{mn}(\cdot)$ represents the multivariate normal distribution of dimension mn. Using (2.4) and (2.7), we easily arrive at the following theorem.

Theorem 2.4.1 *If $\mathscr{X}$ is a D-order random MDA of dimension $\mathbf{n}$ then the following statements are equivalent.*

1. $\mathscr{X} \sim \mathcal{N}_{\mathbf{n}}\left(\boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d\right)$

2. $\mathscr{X}_{(j)} \sim \mathcal{N}_{\frac{n^*}{n_j}\times n_j}\left(\boldsymbol{\mathfrak{M}}_{(j)}, \bigotimes_{d\neq j} \boldsymbol{\Delta}_d, \boldsymbol{\Delta}_j\right)$

3. $vec(\mathscr{X}_{(j)}) \sim \phi_{n^*}\left(vec(\boldsymbol{\mathfrak{M}}_{(j)}), \boldsymbol{\Delta}_j \otimes \bigotimes_{d\neq j} \boldsymbol{\Delta}_d\right)$

The TVND, an alternative formulation of the MLND described by Hoff *et al.* (2011), is defined as

$$\begin{aligned} f\left(\boldsymbol{\mathfrak{X}}|\boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d\right) &= (2\pi)^{\frac{-n^*}{2}} \prod_{d=1}^{D} |\boldsymbol{\Delta}_d|^{-\frac{n^*}{2n_d}} \\ &\quad \times \exp\left\{-\frac{1}{2}\left\|(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}}) \times \widehat{\boldsymbol{\Delta}}^{-\frac{1}{2}}\right\|\right\}, \end{aligned} \tag{2.8}$$

where $\times\widehat{\boldsymbol{\Delta}}^{-1}$ is the Tucker product (Kolda and Bader, 2009). The equivalence of the two $\exp(\cdot)$ terms in (2.3) and (2.8) is outlined in Appendix A.4.

2.5 Benefits Over Vectorization

The MDA observations can be vectorized and analyzed as vectorial data; however, this approach has a few drawbacks. The first is that the scale matrices for each mode of the MDA, $\boldsymbol{\Delta}_d$, allow for the modeling of element dependencies within that mode.

Secondly, by modeling each $\boldsymbol{\Delta}_d$ individually, the number of free scale parameters is lessened and the overall Kronecker product structure of $\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d$, leads to sparsity in the covariance matrix. If we consider a D-order MDA of dimension $\mathbf{n}$, the vectorized version of this MDA is an n^* dimensional vector. If no constraints are placed on the scale matrix, then there would be $n^*(n^*+1)/2$ free scale parameters that would need to be estimated. Constraints in the form of eigenvalue decomposition, or implementing a factor analysis could be placed on the scale matrix; however, these methods would not give acceptable results when $n^* \geq 100$, which is easily obtained with MDA data (e.g., $6 \times 6 \times 4$ order-3 MDA). By modeling with a tensor-variate distribution, parameter estimation of the scale parameters is restricted to estimating D lower dimensional scale matrices leading to $\sum_{d=1}^{D} n_d(n_d+1)/2$ free scale parameters. Therefore, in the previous case of a $6 \times 6 \times 4$ order three MDA, there would be only 36 scale parameters when using a tensor-variate distribution in comparison to 10,440 scale parameters in an unconstrained scale matrix when vectorizing.

2.6 Parsimony

In practice, one or more dimensions of the MDA are composed of ordered values, usually some representation of time. Analogous to McNicholas and Murphy (2010) in the multivariate case, we use the modified Cholesky decomposition (MCD; Pourahmadi, 1999) to constrain the number of parameters in the $\mathbf{\Delta}_d$'s modeling temporal dimensions of the MDAs. The Cholesky decomposition(Benoıt, 1924) of a positive definite matrix $\mathbf{\Delta}$ is given as

$$\mathbf{\Delta}_d = \mathbf{A}\mathbf{A}^\top,$$

where $\mathbf{\Delta}_d \in \mathbb{R}^{n_d \times n_d}$ and the Cholesky factor, $\mathbf{A} = (a_{ij})$ is a unique lower triangular matrix. The statistical interpretation of a_{ij}, can be enhanced by considering a modification to the decomposition, where the Cholesky factors are further decomposed as $\mathbf{W} = \mathbf{A}\mathbf{D}^{-1}$, where $\mathbf{D} = \text{diag}(d_{11}, \cdots, d_{n_d n_d})$ and $\mathbf{W}$ is a unit triangular matrix. This alteration to the decomposition is traditionally called the modified Cholesky decomposition (MCD) and takes the form

$$\mathbf{\Delta}_d = \mathbf{A}\mathbf{D}^{-1}\mathbf{D}\mathbf{D}\mathbf{D}^{-1}\mathbf{A}^\top = \mathbf{W}\mathbf{D}^2\mathbf{W}^\top$$

This factorization is commonly rewritten in one of these two forms

$$\boldsymbol{\Gamma}\boldsymbol{\Delta}_d\boldsymbol{\Gamma}^\top = \boldsymbol{\Xi}, \boldsymbol{\Delta}_d^{-1} = \boldsymbol{\Gamma}^\top\boldsymbol{\Xi}^{-1}\boldsymbol{\Gamma},$$

where $\boldsymbol{\Xi} = \mathbf{D}^2$ and $\boldsymbol{\Gamma} = \mathbf{W}^{-1}$. The unique entries in $\boldsymbol{\Gamma}$ are unconstrained and can be modeled statistically. This contrasts with the diagonalization of $\boldsymbol{\Delta}_d$ that results from the eigendecomposition, where the entries of the orthogonal matrix are constrained.

As outlined in Pourahmadi (1999), the unstructured covariance matrix $\boldsymbol{\Delta}_d$ can be modeled in a regression framework, using the MCD. A linear model is used, that assumes there is an ordered random vector, $\mathbf{Z}_d \in \mathbb{R}^{n_d \times 1}$, with mean 0 and covariance $\boldsymbol{\Delta}_d$. The linear model is given by

$$Z_{dt} = \sum_{k=1}^{t-1} \psi_{tk} Z_{dk} + \epsilon_t,$$

where $t = 1, 2, \cdots, n_d$ represents the order of Z_d, ψ_{tk} are the autoregressive (e.g. regression) coefficients, ϵ_t are the uncorrelated prediction errors, $\hat{Z}_{dk} - Z_{dk}$ and the σ_t^2's are the innovation variances (e.g., the variances of ϵ_t). This model can be expressed in matrix form as $\boldsymbol{\Gamma}\mathbf{Z}_d = \boldsymbol{\varepsilon}$, where $\boldsymbol{\Gamma}$ is a unit diagonal lower triangular matrix that contains $-\psi_{tk}$ below the diagonal. The covariance matrix of $\boldsymbol{\varepsilon}$ is given by

$$\text{Cov}(\boldsymbol{\varepsilon}) = \text{Cov}\left[\boldsymbol{\Gamma}\mathbf{Z}_d\right] = \boldsymbol{\Gamma}\boldsymbol{\Delta}_d\boldsymbol{\Gamma}^\top = \text{diag}(\sigma_1^2, \cdots \sigma_{n_d}^2) = \boldsymbol{\Xi}.$$

The regression coefficients, ψ_{tk}, are by their nature, unconstrained. To remove the positive definite constraint on $\boldsymbol{\Delta}_d$, Pourahmadi (1999) take the logarithm of the diagonal entries of $\boldsymbol{\Xi}$. This is makes for a simple relationship between $\log \boldsymbol{\Xi}$ and $\boldsymbol{\Xi}$, and changes $\mathbf{\Delta}_d$ from a positive-definite matrix to a symmetric matrix. This simple relationship is preferable to the relationship between the entries of $\log \mathbf{\Delta}_d$ and

$$\mathbf{\Delta}_d = \mathrm{e}^{\log \mathbf{\Delta}_d},$$

where $\mathrm{e}^{\mathbf{X}}$ and $\log \mathbf{X}$ are the matrix exponential and matrix logarithm, respectively. The parameters of the linear model are modeled with a link function (McCullagh and Nelder, 1989), given by

$$h(\mathbf{\Delta}_d) = (\boldsymbol{\psi}_2^\top, \cdots, \boldsymbol{\psi}_{n_d}^\top, \log \sigma_1^2, \cdots, \log \sigma_{n_d}^2)^\top,$$

where $\boldsymbol{\psi}_t = [\psi_{t1}, \cdots, \psi_{tt-1}]$. Thus we can model the entries of $\mathbf{\Delta}_d$ as unconstrained parameters and we end up with an estimator, $\hat{\mathbf{\Delta}}_g$ that retains its positive definite requirements.

Using the MCD, McNicholas and Murphy (2010) developed an eight member family of mixture models capable of properly modeling the dependence structure present in temporal data. The Cholesky-decomposed Gaussian mixture model family (CDGMM), decomposes each groups precision matrix as

$$\mathbf{\Delta}_{gd}^{-1} = \mathbf{\Gamma}_{gd}^\top \mathbf{\Xi}_{gd}^{-1} \mathbf{\Gamma}_{gd}.$$

The eight members of the CDGMM family come from constraints imposed on either $\mathbf{\Gamma}_{gd}$ and $\mathbf{\Xi}_{gd}$, individually or in combination. These constraints include equality across groups and an isotropic constraint on $\mathbf{\Xi}_{gd}$, $\mathbf{\Xi}_{gd} = \delta_{gd}\mathbf{I}_{n_d}$. These constraints have natural interpretations for temporal data, such as; $\mathbf{\Xi}_{gd} = \mathbf{\Xi}_d$ suggests the variability at each time t is the same for all the groups, $\mathbf{\Gamma}_{gd} = \mathbf{\Gamma}_d$ implies the model coefficients, $\psi_{t,k}$, are the identical for each group g. Of the eight member models, we chose the VVI and EVI models because they have either variable or equal autoregressive coefficients between time points for the groups. They both include an isotropic constraint on the variability at each time point, slightly lowering the number of free parameters being estimated.

For non-temporal dimensions of the MDA, we examine the VVI, EEE and VVV models, three members of the 14 Gaussian parsimonious clustering models (GPCM; Celeux and Govaert, 1995), listed in table 2.1. These models constrain the eigen-decomposition of the associated scale matrix. The decomposition has the following form:

$$\mathbf{\Delta}_{gd} = \lambda_{gd}\mathbf{\Gamma}_{gd}\mathbf{D}_{gd}\mathbf{\Gamma}_{gd}^{\top},$$

where $\lambda_{gd} = |\mathbf{\Delta}_{gd}|^{\frac{1}{n_d}}$, $\mathbf{D}_{gd}$ is a diagonal matrix containing the normalized eigenvalues of $\mathbf{\Delta}_{gd}$ in decreasing order and $\mathbf{\Gamma}_{gd}$ is the corresponding matrix of eigenvectors of $\mathbf{\Delta}_{gd}$.

Table 2.1: Covariance structures for the three GPCM models used in Chapter 4.

Model	**Volume**	**Shape**	**Orientation**	**Free Parameters**
VVI	λ_{gd}	$\mathbf{\Delta}_{gd}$		Gn_d
EEE	λ_d	$\mathbf{\Delta}_d$	$\mathbf{\Gamma}_d$	$n_d(n_d+1)/2$
VVV	λ_{gd}	$\mathbf{\Delta}_{gd}$	$\mathbf{\Gamma}_{gd}$	$Gn_d(n_d+1)/2$

The VVI model assumes a diagonal parameterization of the scale matrix. It has n_d free parameters. It was chosen because it has the most general parameterization with n_d parameters and could give a hint towards the importance of modeling the complete variation in each dimension of the MDA data. The EEE model assumes that each group has the same scale structure for mode-d of the MDA and, of all the GPCM constraints with n_d^2 parameters, is the one with the fewest. The VVV model is unconstrained and each group has its own unique scale matrix.

2.7 Inverse and Generalized Inverse Gaussian Distributions

The MDA distributions outlined in Chapter 5 and the mixtures of these distributions discussed in Chapter 6, rely on the generalized inverse Gaussian (GIG) distribution, and to a lesser extent the inverse Gaussian (IG) distribution. A random variable $\mathcal{Y} \sim \text{IG}(\delta, \gamma)$ has the probability density function

$$f(y|\delta, \gamma) = \frac{\delta}{\sqrt{2\pi}} \exp\{\delta\gamma\} y^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\left(\frac{\delta^2}{y} + \gamma^2 y\right)\right\}$$

for $y > 0$, where $\delta > 0$ and $\gamma > 0$.

We will consider two different parameterizations of the GIG distribution. A random variable $Y \sim \text{GIG}(a, b, \lambda)$, where $a, b > 0$ and $\lambda \in \mathbb{R}$. Its probability density function can be written as

$$f(y|a, b, \lambda) = \frac{(a/b)^{\frac{\lambda}{2}}\, y^{\lambda-1}}{2K_\lambda(\sqrt{ab})} \exp\left\{-\frac{ay + b/y}{2}\right\},$$

where

$$K_\lambda(u) = \frac{1}{2}\int_0^\infty y^{\lambda-1} \exp\left\{-\frac{u}{2}\left(y + \frac{1}{y}\right)\right\} dy$$

is the modified Bessel function of the third kind with index λ.

The expectations of certain functions of a $\text{GIG}(a, b, \lambda)$ random variable are computationally tractable and will be used for parameter estimation of the skewed tensor-variate distributions. Examples of such expectations include

$$\mathbb{E}(Y) = \sqrt{\frac{b}{a}} \frac{K_{\lambda+1}(\sqrt{ab})}{K_\lambda(\sqrt{ab})}, \tag{2.9}$$

$$\mathbb{E}\left(1/Y\right) = \sqrt{\frac{a}{b}} \frac{K_{\lambda+1}(\sqrt{ab})}{K_\lambda(\sqrt{ab})} - \frac{2\lambda}{b}, \tag{2.10}$$

$$\mathbb{E}(\log Y) = \log\left(\sqrt{\frac{b}{a}}\right) + \frac{1}{K_\lambda(\sqrt{ab})} \frac{\partial}{\partial\lambda} K_\lambda(\sqrt{ab}). \tag{2.11}$$

To derive the density of the tensor-variate generalized hyperbolic distribution, this second GIG density function is preferred:

$$g(y|\omega, \eta, \lambda) = \frac{(w/\eta)^{\lambda-1}}{2\eta K_\lambda(\omega)} \exp\left\{-\frac{\omega}{2}\left(\frac{w}{\eta} + \frac{\eta}{w}\right)\right\}, \tag{2.12}$$

where $\omega = \sqrt{ab}$ and $\eta = \sqrt{a/b}$ (Browne and McNicholas, 2015). To reduce confusion, we will denote the GIG parameterization in (2.12) by $\text{I}(\omega, \eta, \lambda)$.

2.8 Variance-Mean Mixtures

A m-variate random vector $\mathbf{X}$ defined in terms of a variance-mean mixture has a probability density function of the form

$$f(\mathbf{x}) = \int_0^\infty \phi_m(\mathbf{x}|\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Delta}) h(w|\boldsymbol{\theta}) dw,$$

where the random variable $W > 0$ has density function $h(w|\boldsymbol{\theta})$ and $\phi_m(\cdot)$ represents the density function of the m-variate Gaussian distribution. Notably, $\mathbf{X}$ can equivalently be expressed as

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{V}, \tag{2.13}$$

where $\boldsymbol{\mu}$ is a location parameter, $\boldsymbol{\alpha}$ is the skewness, $\mathbf{V} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Delta})$ with $\boldsymbol{\Delta}$ as the scale matrix, and $W > 0$ has density function $h(w|\boldsymbol{\theta})$. Note that W and $\mathbf{V}$ are independent. The variance-mean mixture can be used to formulate many multivariate distributions, simply by changing the distribution of W (McNicholas, 2016a).

For example, the m-dimensional normal inverse Gaussian (NIG) distribution, $\text{NIG}_m(\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\Delta}, \delta, \gamma)$ is derived by Karlis and Santourian (2009). They use a variance-mean mixture with $W \sim \text{IG}(\delta, \gamma)$. This derivation has a restriction on the determinant of $\boldsymbol{\Delta}$ to eliminate identifiability problems. To remove this restriction and maintain identifiability, Karlis and Santourian (2009) set $\delta = 1$, and set $\tilde{\gamma} = \gamma$. This formulation was also used by O'Hagan *et al.* (2016). Franczak *et al.* (2014) use a variance-mean mixture representation of a (shifted) asymmetric Laplace distribution for model-based clustering, Murray *et al.* (2014) use a skew-t distribution, Browne and McNicholas (2015) use a generalized hyperbolic distribution, etc.

2.9 EM Algorithm

The parameters of the finite mixture models described herein are estimated by the method of maximum likelihood (ML). The ML estimates are found using the EM algorithm, a two step iterative algorithm used to calculated the parameter estimates in the presence of missing data (Dempster *et al.*, 1977). In the case of the models described in the following chapters, the missing data are represented by the latent variables; w_{ig} used in the variance-mean mixture formulation described in Section 2.8 and z_{ig} which represents the group allocation of each member of the sample, $\{\mathbf{x}_i\}_{i=1}^{N}$, in the finite mixture models described in Section 2.2.

In general, finite mixture models have the following observed log-likelihood

$$\ell_{\text{O}}(\boldsymbol{\vartheta}) = \sum_{i=1}^{N} \log \left[\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_i | \boldsymbol{\theta}_g) \right].$$

The summation over G inside the logarithm makes parameter estimation difficult. To make the parameter estimation more tractable, a latent variable, z_{ig} is included in the likelihood formulation to produce the following complete log-likelihood

$$\ell_{\text{C}}(\boldsymbol{\vartheta}) = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log \pi_g + \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log f_g(\mathbf{x}_i | \boldsymbol{\theta}_g).$$

Note, $\mathbf{z}_i \in \mathbb{R}^{G \times 1}$ is a realization of a random variable, $\mathbf{Z}_i$, which follows a multinomial distribution with one draw on G categories with probabilities $\pi_1, \ldots, \pi_G$. The $\{\mathbf{Z}_i\}_{i=1}^{N}$ random variables are all independent and identically distributed and π_g can be considered the *a priori* probability that group g contains observation $\mathbf{x}_i$.

Using the $\ell_{\text{C}}(\boldsymbol{\vartheta})$, the EM algorithm consists of two steps at each iteration k:

1) E-step: Take $\mathbb{E}_{z_{ig}}\left[\ell_{\text{C}}(\boldsymbol{\vartheta}^{(k)})\right]$. The result of the E-step is often referred to as the Q function, $\mathbb{Q}(\boldsymbol{\vartheta}^{(k)})$.

2) M-step: $\boldsymbol{\vartheta}^{k+1} = \text{argmax}_{\boldsymbol{\vartheta}}\, \mathbb{Q}(\boldsymbol{\vartheta} | \boldsymbol{\vartheta}^{(k)})$.

For a version of the mixture model in Section 2.2 that uses the MLND as its component density, the E-step amounts to replacing the z_{ig} values in the $\ell_C(\boldsymbol{\vartheta})$ equations by their conditional expectations

$$\mathbb{E}[z_{ig}|\boldsymbol{\mathfrak{X}}_i] = \hat{z}_{ig} = \frac{\hat{\pi}_g f_{\text{MLND}}(\boldsymbol{\mathfrak{X}}_i \mid \boldsymbol{\mathfrak{M}}_g, \boldsymbol{\Delta}_{g1}, \ldots, \boldsymbol{\Delta}_{gD})}{\sum_{h=1}^{G} \hat{\pi}_h f_{\text{MLND}}(\boldsymbol{\mathfrak{X}}_i \mid \boldsymbol{\mathfrak{M}}_h, \boldsymbol{\Delta}_{h1}, \ldots, \boldsymbol{\Delta}_{hD})}, \tag{2.14}$$

Note that the $\hat{z}_{ig}$ gives a probability, $\mathbb{P}[z_{ig} = 1|\boldsymbol{\mathfrak{X}}_i]$, of each observation i belonging to a component g, often called the *a posteriori* probability. To harden up these soft classifications, the maximum *a posteriori* (MAP) classification is defined as

$$\text{MAP}_i\left(\hat{z}_{ig}\right) = \begin{cases} 1 & \text{if } g = \text{argmax}_h \left(\{\hat{z}_{ih}\}_{h=1}^{G}\right), \\ 0 & \text{otherwise.} \end{cases}$$

The MAP classifications are often reported as the group labels from a finite mixture model.

A variant of the EM algorithm, called the expectation-conditional maximization algorithm (ECM; Meng and Rubin, 1993) can be an attractive alternative to the EM algorithm when the M-step is computationally complex. The ECM algorithm replaces the M-step by a series of conditional maximization steps that condition the maximization on some of the model parameters. These conditional steps are simpler, which leads to a reduction in the total compute time used to find the final solution. This comes at the expense of an increased number of iterations relative to the EM algorithm. The ECM algorithm preserves the appealing monotone convergence properties of the EM algorithm. The ECM algorithm will be used for parameter estimation in Chapters 5 and 6.

Chapter 3

Finite Mixtures of MLNDs

3.1 Model

Building on the material outlined in Sections 2.2, 2.4 and 2.9, we describe a mixture of MLNDs. In this context, we assume $\mathcal{X}$ comes from a population with G subgroups, all of which come from MLNDs with different parameter values. Given a sample of N iid random D-dimensional arrays $\boldsymbol{\mathfrak{x}}_1, \ldots, \boldsymbol{\mathfrak{x}}_N$, the log complete-data likelihood for the model is given by

$$\begin{aligned}\ell_C(\boldsymbol{\vartheta}) = C + \sum_{g=1}^{G} n_g \log \pi_g - \frac{n^*}{2} \sum_{g=1}^{G} n_g \sum_{d=1}^{D} \frac{1}{n_d} \log(|\boldsymbol{\Delta}_{gd}|) \\ - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} \left[\text{vec}\left(\mathbf{X}_{(1)i} - \mathbf{M}_{(1)g}\right)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}\left(\mathbf{X}_{(1)i} - \mathbf{M}_{(1)g}\right) \right], \quad (3.1)\end{aligned}$$

where C is a constant that does not depend on the parameters.

Depending on the permutation of the MDA we use, the final term in (3.1) can be replaced with

$$-\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig} \sum_{j=1}^{n^*_{3:D}} \operatorname{tr}\left[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij}\right] \tag{3.2}$$

or

$$-\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig} \sum_{j=1}^{n^*_{2:D/l}} \operatorname{tr}\left[\boldsymbol{\Delta}_{g1}^{-1}(\mathbf{X}_{(1)gij}^{l2})^{\top}\boldsymbol{\Delta}_{gl}^{-1}\mathbf{X}_{(1)gij}^{l2}\right], \tag{3.3}$$

resulting in three complete-data log-likelihood equations. These equations allow us to isolate the individual model parameters and enable their estimation.

3.2 Parameter Estimation

The model parameters are all estimated by the method of maximum likelihood using the EM algorithm outlined in Section 2.9. The M-step update for $\boldsymbol{\vartheta}$ are available in closed form and follow from taking respective first derivatives of the $\mathcal{Q}$ function and setting the resulting expressions to zero. The update for π_g is given by $\hat{\pi}_g = n_g/N$. The respective M-step updates for $\boldsymbol{\Delta}_{g1}$ and $\boldsymbol{\Delta}_{g2}$ involve taking the first derivative of the $\mathcal{Q}$ function that uses (3.2) as its final term. The updates are

$$\hat{\boldsymbol{\Delta}}_{g1} = \frac{n_1}{n^* n_g}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij} \tag{3.4}$$

and

$$\hat{\boldsymbol{\Delta}}_{g2} = \frac{n_2}{n^* n_g}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top}, \tag{3.5}$$

respectively. The update for $\mathbf{\Delta}_{gl}$ uses the $\mathcal{Q}$ function that adopts (3.3) for its final term. The update is

$$\hat{\mathbf{\Delta}}_{gl} = \frac{n_l}{n^* n_g} \sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n^*_{2:D/l}} \mathbf{X}^{l2}_{(1)gij} \mathbf{\Delta}^{-1}_{g1} (\mathbf{X}^{l2}_{(1)gij})^{\top}. \tag{3.6}$$

The M-step update for $\mathbf{M}_{(1)}$ uses the $\mathcal{Q}$ function equivalent to (3.1) and is given by

$$\hat{\mathbf{M}}_{(1)g} = \frac{1}{n_g} \sum_{i=1}^{N} \hat{z}_{ig} \mathbf{X}_{(1)i}. \tag{3.7}$$

Detailed derivations of the parameter estimates are available in Appendix B.1.

3.3 Model Selection

The number of groups G in a clustering problem is often unknown *a priori*. In such cases, the parameters of a mixture model are typically estimated for different values of G and some criterion is then used to select G. We use the Bayesian information criteria (BIC; Schwarz, 1978) to do the model selection. It can be written

$$\text{BIC} = 2l(\hat{\boldsymbol{\vartheta}}) - \rho \log N, \tag{3.8}$$

where $l(\hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood, ρ is the number of free parameters in the model and N is the number of observations. For our finite mixture of MLNDs,

$$\rho = (G-1) + Gn^* + \frac{G}{2} \sum_{d=1}^{D} n_d(n_d+1). \tag{3.9}$$

3.4 Identifiability

The scale parameters, $\mathbf{\Delta}_{gd}$, in the Kronecker product are unique up to a strictly positive multiplicative constant (Dutilleul, 1999; Anderlucci *et al.*, 2015; Gallaugher and McNicholas, 2018a). Indeed, if we let $d_k > 0$ then

$$\bigotimes_{d=1}^{D} \mathbf{\Delta}_d = \frac{1}{\prod_{k=2}^{D} d_k} \mathbf{\Delta}_1 \otimes \bigotimes_{k=2}^{D} d_k \mathbf{\Delta}_k, \tag{3.10}$$

and the likelihood is unchanged. However, we notice that

$$\bigotimes_{d=1}^{D} \mathbf{\Delta}_d = \bigotimes_{d=1}^{D} \tilde{\mathbf{\Delta}}_d,$$

where $\tilde{\mathbf{\Delta}}_d$ are the terms on the right hand side of (3.10), so the estimate of the Kronecker product would be unique. This constraint is imposed once the EM algorithm has converged. We let $d_k = 1/\delta_{k(1,1)}$, where $\delta_{k(1,1)}$ is the first entry in $\mathbf{\Delta}_k$. Another way of solving this problem of non-identifiability is to set $\operatorname{tr}(\mathbf{\Delta}_d) = n_d$ again for $1 \leq d \leq D-1$ (Anderlucci *et al.*, 2015).

3.5 Stopping rule

To stop our EM algorithm, we use a criterion based on the Aitken acceleration (Aitken, 1926). At iteration t of the EM algorithm, the Aitken acceleration is

$$a^{(t)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}}, \tag{3.11}$$

where $l^{(t)}$ is the (observed) log-likelihood at iteration t. Böhning *et al.* (1994) use $a^{(t)}$ to calculate an asymptotic estimate of the log-likelihood at iteration $t+1$:

$$l_{\infty}^{(t+1)} = l^{(t)} + \frac{1}{1-a^{(t)}}(l^{(t+1)} - l^{(t)}). \tag{3.12}$$

We stop the EM algorithm when

$$l_{\infty}^{(t+1)} - l^{(t)} < \epsilon$$

(Lindsay, 1995; McNicholas *et al.*, 2010).

3.6 Software

We have used version 1.5.3 of the Julia language(`https://julialang.org/`; Bezanson *et al.*, 2017; McNicholas and Tait, 2019), to implement our finite mixture model. Singular $\mathbf{\Delta}_g$ values were numerically regularized by adding a small positive quantity to the diagonal elements of the matrices (Williams and Rasmussen, 2006). The regularization is summarized in the following equation:

$$\tilde{\mathbf{\Delta}} = \hat{\mathbf{\Delta}} + \epsilon\mathbf{I}, \tag{3.13}$$

where $\epsilon \in (0, 0.1]$, $\hat{\mathbf{\Delta}}$ is the estimated singular scale matrix, and $\tilde{\mathbf{\Delta}}$ is the regularized estimate of $\mathbf{\Delta}$. We use $\epsilon = 0.001$ in our implementation. The singularity of $\hat{\mathbf{\Delta}}$ was assessed by checking if its inverse condition number is less than machine epsilon. This regularization is often done implicitly in software implementations such as `scikit-learn`'s GaussianMixture function, written in Python. The value of the regularization parameter ϵ could be tuned. The larger it is, the further the model results are from the true solution. The positive definiteness of the $\mathbf{\Delta}_d$ matrices was checked with the Cholesky decomposition. We used Algorithm 4.1 from Blanchard *et al.* (2019) to implement the log-sum-exp rule.

3.7 Simulation

We conducted a simulation study to investigate the effect of different sample and MDA sizes on the following questions:

- Can we effectively estimate the model parameters?
- Can we effectively capture the original group labels?
- How often do singular scale matrices occur and how do singular scale matrices affect the model results?

Accurately estimating and interpreting the model parameters and labels are common goals of clustering and as such, are important to the assessment of new models.

We used five-way data in our simulations. The simulations were conducted for sample sizes $N \in \{60, 90, 120, 180\}$ subjects with three equal sized groups. The n^* quantity was used to measure the different dimensions of the order four MDAs. Its values included 256, 625, 1296 and 2401. While these values of n^* can equate to any product of dimension lengths, the simplest way to visualize the resulting MDAs is as an order 4 MDA with four equal dimension lengths of 4, 5, 6 or 7. For each combination of N and n^*, 250 simulations were conducted.

Following Definition 2.2 in Ohlson *et al.* (2013), the simulated data was generated using the equation

$$\mathrm{vec}(\boldsymbol{\mathfrak{X}}) = \mathrm{vec}(\boldsymbol{\mathfrak{M}}) + \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{\frac{1}{2}} \mathbf{u}, \tag{3.14}$$

where $\mathbf{u}$ is a vector of iid $\mathcal{N}(0, 1)$ random numbers. This is equivalent to the multivarite normal model for the vectorized version of the MDA data. This formulation has the disadvantage of having to recreate the MDA from the vectorized data and dealing with a potentially large matrix

$$\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{\frac{1}{2}}.$$

If we generate $\mathbf{u}$ and $\boldsymbol{\mathfrak{M}}$ as order-D MDAs, we can use tensor d-mode products to implement the final term on the right hand side of (3.14) (Kolda and Bader, 2009). A tensor d-mode product multiplies an MDA by a matrix in mode d and has the advantage of retaining the mode d structure of the data and not creating one large matrix from the Kronecker product

$$\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{\frac{1}{2}}$$

and then having to permute the data back into an MDA.

A signal-to-noise ratio of a half was applied to the simulated data prior to analysis. The $\boldsymbol{\Delta}_d$ parameters were generated by specifying a diagonal matrix of eigenvalues and a random orthogonal matrix and combining them as you would in an eigen-decomposition of the scale matrix. An $n_d \times n_d$ orthogonal matrix was created by generating n_d^2 i.i.d. $\mathcal{N}(0,1)$ random values, placing them in a matrix and orthogonalizing it with the QR decomposition. We restrict the condition number of these $\mathbf{\Delta}_d$ matrices to be at most 10.

The EM algorithm used to generate the results was initialized with identity matrices for all the scale matrix parameters. The values of $\hat{z}_{ig}$ are initialized with k-means starts, calculated on the vec$(\mathfrak{X}_i)$ version of the data. The BIC was used to select the number of groups in our simulation. We checked $G \in \{2,3,4,5\}$ for each combination of N and n^* and in every instance, $G = 3$ was chosen. We took advantage of the Julia's native distributed computing capabilities to make these simulations computationally feasible.

We use the relative error to determine how close the estimated model parameters were to the true parameters. It is defined as

$$\frac{\left\|\hat{\mathbf{V}} - \mathbf{V}\right\|_F}{\|\mathbf{V}\|_F},$$

where $\|\cdot\|_F$ is the Frobenius matrix norm, $\hat{\mathbf{V}}$ is the estimated parameter matrix, and $\mathbf{V}$ is the true parameter matrix used to generate the simulated data. The smaller this ratio is, the less error is present in the model's parameter estimates.

Figure 3.1 indicates that for the three groups, over 95% of the values are well below one, indicating the mean matrices are being estimated accurately. As noted in Section 3.4, we would expect to be able to accurately estimate

$$\bigotimes_{d=1}^{D} \mathbf{\Delta}_{gd}.$$

Figure 3.2 shows the relative errors are larger than $\mathbf{M}_{(1)g}$ but still consistently between

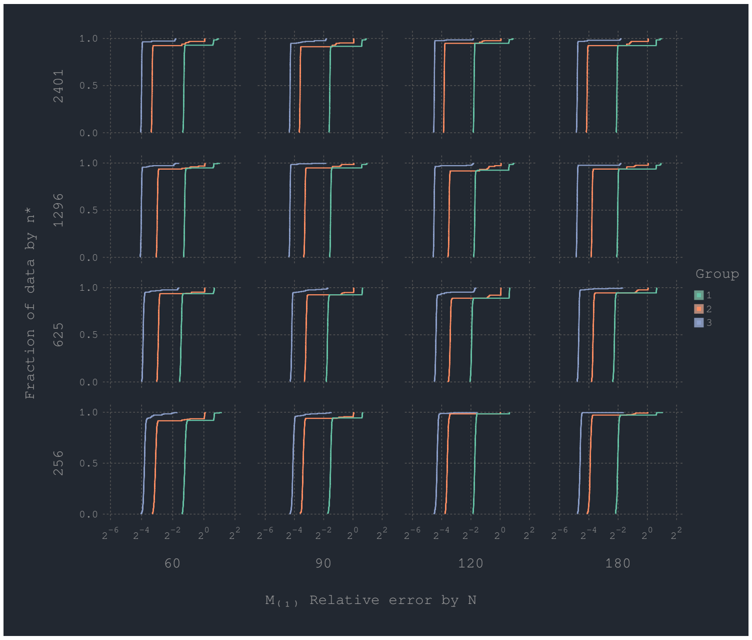


Figure 3.1: Empirical cumulative distribution plots of the relative errors for the $\mathbf{M}_{(1)g}$ parameter matrices. The x-axis is plotted on the log2 scale because the distributions have a very long right tail.

1 and 1.5 for all combinations of N and n^*.

The group labels produced by the finite mixture model were compared to the simulated group labels via the adjusted Rand index (ARI; Hubert and Arabie, 1985). Note that the ARI facilitates a quick assessment of the agreement between two partitions. For our purposes, it suffices to know that an ARI value of 1 indicates perfect class agreement and the expected value of the ARI under random classification is 0. The average ARI for each combination of N and n^* was at least 0.95, with an overall average of 0.969 and standard deviation of 0.118.

Given the high-dimensional data being analyzed and the large number of parameters being estimated by our model, we expected the “curse of dimensionality” to be

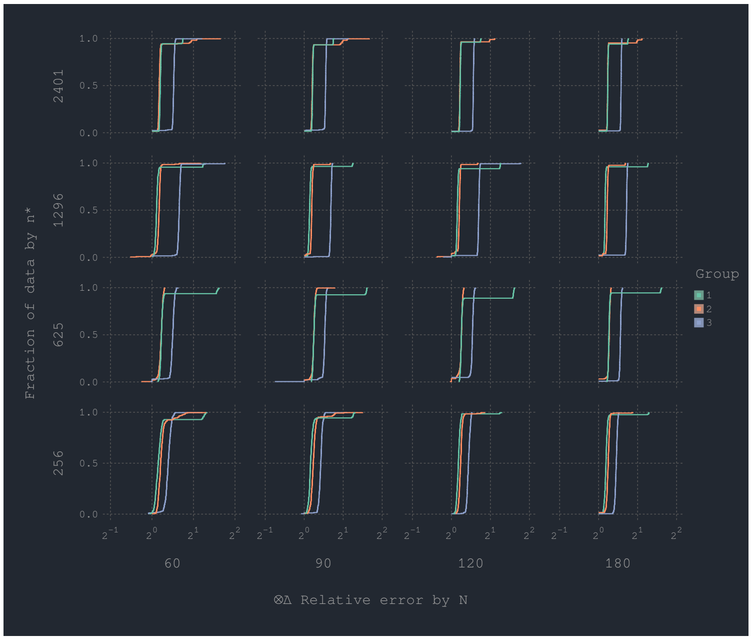


Figure 3.2: Empirical cumulative distribution plots of relative errors for the $\bigotimes_{d=1}^{D} \mathbf{\Delta}_{gd}$ parameter matrices. The x-axis is plotted on the log2 scale because the distributions have a very long right tail.

a problem, manifesting as ill-conditioned scale matrices. This is some-what attenuated in our model because we are estimating individual $\mathbf{\Delta}_d$, which are much lower dimensional than

$$\bigotimes_{d=1}^{D} \mathbf{\Delta}_d$$

and each n_d is less than N. Nevertheless, we count the number of simulations that singular $\mathbf{\Delta}_{gd}$ occur and investigate how their occurrence affects the results. Figure 3.3 indicates no more than 10% of the simulations had singular $\mathbf{\Delta}_d$ matrices and large values of N and n^* resulted in the most occurrences. Singularities happened at the first iteration of the EM algorithm and would occur in a single group. Prior to conducting the simulations, we expected that small N and large n^* values would have had more singular scale matrices.

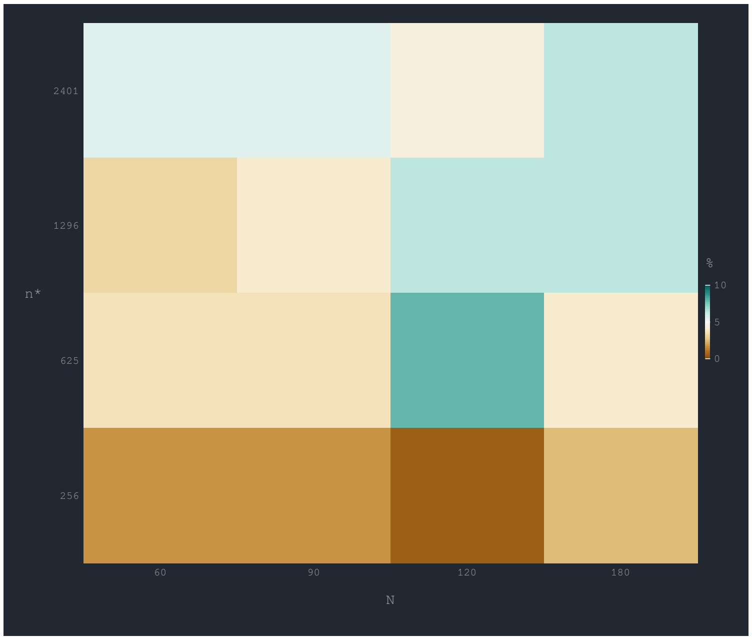
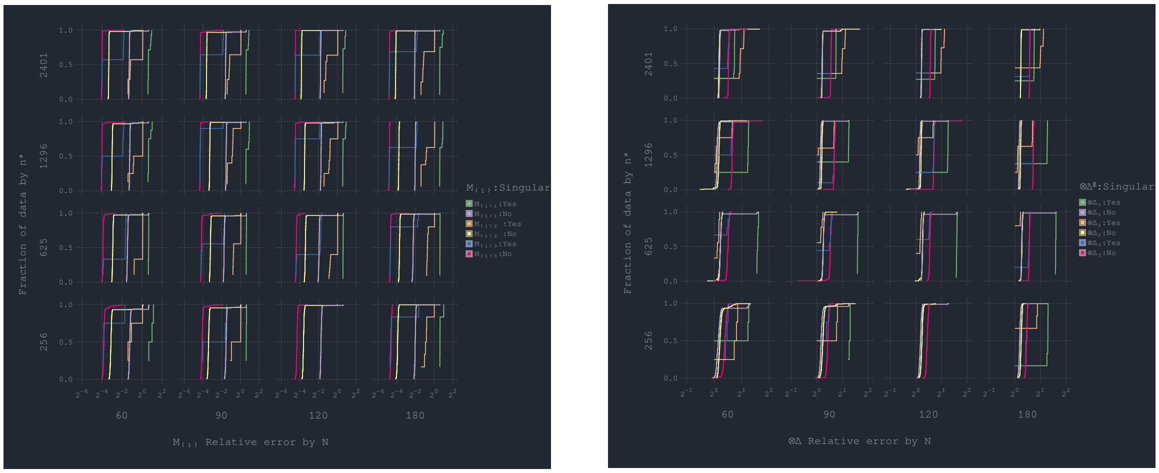


Figure 3.3: Heatmap of the percentage of the 250 simulations with a singular $\mathbf{\Delta}_{gd}$ matrix.

When singular scale matrices are present, the mean matrices $\mathbf{M}_{(1)g}$ can still be reliably estimated, as we can see from Figure 3.4a. The distributions have long right tails but consistently have 95% of the relative errors less than one. As we can see from figure 3.4b, the scale matrices are more adversely affected than $\mathbf{M}_{(1)g}$. When singular $\mathbf{\Delta}_{gd}$'s occur, the overall errors are closer to 2, with group 1 having the worst results in low to moderate n^* values and group 2 at the largest value of n^*.

The mean ARI values are strongly affected by the singular $\mathbf{\Delta}_{gd}$ values. The results are summarized in Table 3.1. The summary statistics are calculated across all the values of N and n^*. When singular $\mathbf{\Delta}_{gd}$ occur, the ARI and estimates of $\mathbf{\Delta}_{gd}$ are adversely affected, making their interpretation suspect.



(a) $\mathbf{M}_{(1)g}$

(b) $\bigotimes_{d=1}^{D} \mathbf{\Delta}_{gd}$

Figure 3.4: Empirical cumulative distribution plots of the relative errors stratified by the occurrence of a singular $\mathbf{\Delta}_{gd}$ matrix.

Table 3.1: ARI summaries for models with and without singular $\mathbf{\Delta}_{gd}$.

Metric	**N**	**Mean**	**SD**
Overall	4000	0.969	0.118
Singular: Yes	168	0.559	0.008
Singular: No	3822	0.987	0.080

3.8 Discussion

The simulations we performed in Section 3.7 can be viewed through the lens of experimental design (ED; Wu and Hamada, 2021) as a computer experiment. In this context, we are doing a 4^2 factorial design, that has two independent variables, N and n^*, each with four levels. The experiment has 250 replicates for each factor level combination. We chose to analyze the results graphically because we are more interested in the practical significance as opposed to the statistical significance of the results. We kept the signal-to-noise ratio, group size n_g and the number of groups G constant for each combination of factor levels.

When viewed as an experimental design, our simulation could be improved in a number of ways. We could use a fractional factorial design to efficiently screen main effects, two-way and three-way interactions of more than two factors; we could use blocking, a set of similar experimental units, to explore variation in nested factors (e.g., n_g changes while N and n^* are held constant); if we had prior knowledge of the accuracy of the parameter we are estimating, we can more accurately calculate the number of simulations we need (Burton *et al.*, 2006). This last point is more difficult in clustering, as few methods report the standard errors of the model parameters. Good general advice concerning designing simulation studies; including random number generation, generating data sets for different types of data (e.g., multivariate, time to event) and performance evaluation are available in Burton *et al.* (2006) and Morris *et al.* (2019).

This work could be extended in several methodological directions. For a given collection of four or five way data, often $\rho > N$, indicating some form of variable selection or dimension reduction could be warranted. Variable selection in model based clustering can be separated into two broad approaches, penalization and model selection (Fop *et al.*, 2018). Model selection approaches have been shown to be superior to penalization in terms of variable selection and classification (Celeux *et al.*, 2014) and would be the preferred avenue of future research. Even without variable selection, our model was able to effectively recover the parameters and labels in our simulations.

Recently, Mai *et al.* (2021), introduced a mixture model that clusters MDA data using a discriminant analysis, where sparsity is applied to the discriminant tensors

$$\left[\boldsymbol{\mathfrak{M}}_{\substack{g\\g\neq 1}} - \boldsymbol{\mathfrak{M}}_1\right] \times \widehat{\boldsymbol{\Delta}}^{-1}.$$

The variation of the EM algorithm they use, called DEEM, has an enhanced E-step, that imposes sparsity on the elements of the discriminant tensor by applying a group lasso penalty (Yuan and Lin, 2006). The reduction in free parameters that resulted from the penalty was never reported in the paper, making it difficult to know how sparse the model parameters were and how much smaller the penalized versus non-penalized ρ was.

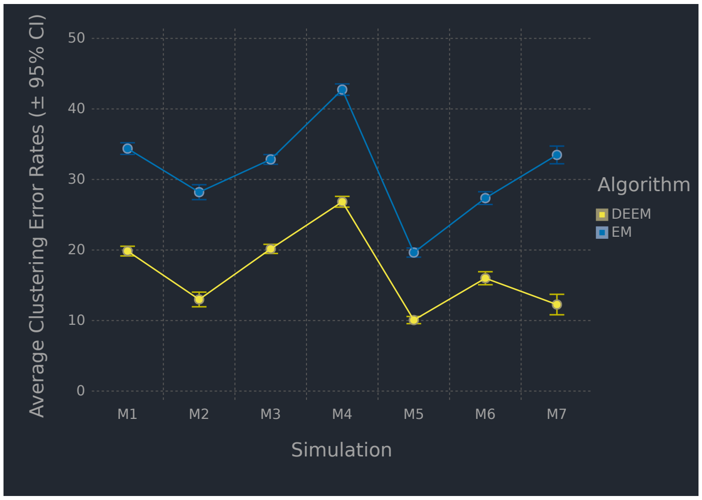


Figure 3.5: Graphical display of table 1 from Mai *et al.* (2021).

Over a range of simulations, their DEEM algorithm outperformed the standard EM algorithm in terms of label (clustering) error. Taking a page from Gelman *et al.* (2002), Figure 3.5 summarizes their simulation results, listed in the Table 1 of the paper. To get a more complete picture of the relative performance, the BIC for each algorithm should be compared and a more clustering specific metric, like the

ARI, should be used to compare the model labels to the known group labels. It would be interesting to see the distribution and median values of the clustering error rates for both algorithms. It seems plausible that the regularization scheme used in the DEEM algorithm reduced the tails of the distribution, which would not affect the median error but would account for the smaller mean error rates. Finally, their model assumed homogeneous covariances across cluster groups and, as we will see in Chapter 4, this is not a realistic assumption for real world data.

Chapter 4

Parsimony Constraints in Finite Mixtures of MLNDs

4.1 Approach

The number of free parameters, given by (3.9), can be substantial as D and n^* increase. We denote the mixture model described in Chapter 3 as VVV. We impose parsimony on the $\mathbf{\Delta}_d$ matrices to improve their interpretability. In practice, one or more modes of the MDA are composed of ordered values, usually some representation of time. To account for this, we extend two members of the CDGMM family, described in Section 2.6, to our mixture model of MLNDs. In our applied problem, described in Section 4.2, modes 1,3 and 4 of the order-4 MDAs represent temporal measurements. Mode 2 represents variables and the scale matrix, $\mathbf{\Delta}_2$, is constrained using the eigen-decomposition associated with the GPCM family of mixture models described in Section 2.6.

To apply the VVI model to $\mathbf{\Delta}_{1g}$, we start with (3.2) and re-express the terms related to $\mathbf{\Delta}_{1g}$ as:

$$\log \mathcal{L}_C(\boldsymbol{\vartheta}) = C + \frac{n^*}{2}\sum_{g=1}^{G} n_g \log(\delta_{g1}^{-1}) - \frac{1}{2}\sum_{g=1}^{G} n_g \delta_{g1}^{-1} \operatorname{tr}\big[\mathbf{T}_{g1}\mathbf{\Lambda}_{g1}\mathbf{T}_{g1}^{\top}\big], \tag{4.1}$$

where

$$\mathbf{\Lambda}_{g1} = \frac{1}{n_g}\sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n_{3:D}^*} \mathbf{X}_{(1)gij}^{\top}\mathbf{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij}.$$

After taking the partial derivatives of the $\mathcal{Q}$ function associated with (4.1), we end up with the following expressions:

$$\begin{aligned}
\hat{\delta}_{g1} &= \frac{1}{n^*}\operatorname{tr}[\mathbf{T}_{g1}\mathbf{\Lambda}_{g1}\mathbf{T}_{g1}^{\top}],\\
\mathbf{\Lambda}_{(r-1)\times(r-1)}^{(g)1\top}\mathbf{\Phi}_{(r-1)\times 1}^{(g)1} &= -\mathbf{\Lambda}_{(r-1)\times 1}^{(g)1},
\end{aligned}$$

where $r = 2, \dots n_1$, $\mathbf{\Phi}_{g1}$ is a portion of the unit lower triangular matrix $\mathbf{T}_{g1}$, which has $n_1(n_1-1)/2$ elements to be estimated. The VVI model for $\mathbf{\Delta}_{1g}$ has $Gn_1(n_1-1)/2+G$ free parameters. For the mathematical details, see appendix C.1.1. A similar derivation for the VVI decomposition of $\mathbf{\Delta}_{gl}$ is available in appendix C.1.1.

To apply the EVI model to $\mathbf{\Delta}_{1g}$, we start with (3.2) and re-express the terms related to $\mathbf{\Delta}_{1g}$ as:

$$\log \mathcal{L}_C(\boldsymbol{\vartheta}) = C + \frac{n^*}{2}\sum_{g=1}^{G} n_g \log(\delta_{g1}^{-1}) - \frac{1}{2}\sum_{g=1}^{G} n_g \delta_{g1}^{-1} \operatorname{tr}\big[\mathbf{T}_{1}\mathbf{\Lambda}_{g1}\mathbf{T}_{1}^{\top}\big],$$

where $\mathbf{\Lambda}_{g1}$ is defined as above. The associated expressions are

$$\hat{\delta}_{g1} = \frac{1}{n^*}\,\mathrm{tr}[\mathbf{T}_1\mathbf{\Lambda}_{g1}\mathbf{T}_1^\top],$$
$$\boldsymbol{\kappa}^{1\top}_{(r-1)\times(r-1)}\mathbf{\Phi}^1_{(r-1)\times 1} = -\boldsymbol{\kappa}^1_{(r-1)\times 1},$$

where the lower triangular elements of $\mathbf{T}_1$ are denoted as the $\boldsymbol{\kappa}^1$ matrix, which has entries

$$\kappa^1_{ij} = \sum_{g=1}^{G}\frac{n_g}{\delta_{g1}}\lambda^{(g)}_{1ij}.$$

The EVI model for $\mathbf{\Delta}_{1g}$ has $n_1(n_1-1)/2+G$ free parameters. The mathematical details are available in Appendix C.1.2. A similar derivation for the EVI decomposition of $\mathbf{\Delta}_{gl}$ is available in Appendix C.1.2.

For non-temporal modes of the MDA, the EEE model assumes that each group has the same scale structure. Starting with (3.2), an EEE model for $\mathbf{\Delta}_{g2}$ can be formulated by reorganizing the equation in terms of $\mathbf{\Delta}_{g2}$ as follows:

$$\log\mathcal{L}_C(\boldsymbol{\vartheta}) = C - \frac{1}{2}\left[N\frac{n^*}{n_2}\log(|\mathbf{\Delta}_2|) + \mathrm{tr}\left\{\left(\sum_{g=1}^{G}\mathbf{\Lambda}_{g2}\right)\mathbf{\Delta}_2^{-1}\right\}\right],$$

where

$$\mathbf{\Lambda}_{g2} = \frac{1}{n_g}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^\top.$$

Using results from Celeux and Govaert (1995), we find that

$$\hat{\mathbf{\Delta}}_2 = \frac{1}{N}\sum_{g=1}^{G}\mathbf{\Lambda}_{g2}\mathbf{\Delta}_2. \tag{4.2}$$

The number of free parameters for the EEE model of $\mathbf{\Delta}_{g2}$ is $n_2(n_2+1)/2$. The mathematical details are available in appendix C.2.

The VVI model assumes a diagonal parameterization of the scale matrix. It has n_d free parameters. It was chosen because it has the most general parameterization with n_d parameters and could give a hint towards the importance of modeling the complete variation in each mode of the MDA data. The VVI model for $\mathbf{\Delta}_{g2}$ represents the scale matrix as $\lambda_{g2}\mathbf{D}_{g2}$. Reorganizing (3.2), we have

$$\log \mathcal{L}_C(\boldsymbol{\vartheta}) = C - \frac{1}{2}\left[\sum_{g=1}^{G} \frac{1}{\lambda_{g2}} \operatorname{tr}\left\{\mathbf{\Lambda}_{g2}\mathbf{D}_{g2}^{-1}\right\} + n^* \sum_{g=1}^{G} n_g \log(\lambda_{g2})\right].$$

Using results from Celeux and Govaert (1995), the estimators are given by:

$$\hat{\mathbf{D}}_{g2} = \frac{\operatorname{diag}(\mathbf{\Lambda}_{g2})}{|\operatorname{diag}(\mathbf{\Lambda}_{g2})|^{\frac{1}{n^*}}}, \tag{4.3}$$

$$\hat{\lambda}_{g2} = \frac{|\operatorname{diag}(\mathbf{\Lambda}_{g2})|^{\frac{1}{n^*}}}{n_g}. \tag{4.4}$$

The number of free parameters for the VVI model of $\mathbf{\Delta}_{g2}$ is Gn_2. The mathematical details are available in Appendix C.2. Parameter estimation for these four parsimonious models is done via the EM algorithm described in Section 3.2.

4.2 CHAMPION Study

The CHAMPION (Cardiovascular Health in children with a chronic inflAMmatory condition: role of Physical activity, fItness, and inflammatiON) study included youth between the ages of 7 and 17 years with a single diagnosis of a chronic inflammatory condition (CIC) including chronic cystic fibrosis (CF), juvenile idiopathic arthritis

(JIA), inflammatory bowel disease (IBD), or type 1 diabetes mellitus (T1D) recruited from the McMaster Children's Hospital. Healthy control participants were recruited from the general community. The study aims to examine the factors affecting heart health in common chronic diseases of childhood.

CHAMPION is a cross-sectional, observational study where each participant was outfitted with an ActiGraph GT3X accelerometer. Accelerations were captured in the vertical (axis1), anterioposterior (axis2), and mediolateral (axis3) planes. The three axis values were combined to form the vector magnitude (VM), which is defined as $\sqrt{\text{axis1}^2 + \text{axis2}^2 + \text{axis3}^2}$. A sample of 83 participants' accelerometer data was analyzed. They had a median of seven wear days. Because youth tend to exhibit short bursts of activity through out the day, understanding intra-day activity patterns could have important implications for health. With this in mind, we aggregated the accelerometer data across each participants wear days into the following nested time periods; four 15 second, six 10 minute and 12 one hour periods (9h–20h). This results in 288 unique epochs per participant. Because the accelerometer data is all ≥ 0 and heavily right skewed, we used the square-root transformation to transform the data before aggregation. The activity counts and VM were aggregated across days by taking their mean values. The participants steps were summed within each day and time period, the transformation was applied and the values were averaged across days. We included the standard deviation (SD) of VM and steps, calculated at the same time as the mean, to capture the variation in these metrics.

The aggregated accelerometer data was transformed into five-way data, with $n_1 = 4$ being the seconds, $n_2 = 7$ being the metrics, $n_3 = 6$ being the minutes and $n_4 = 12$ being the hours. The seven accelerometer metrics include the three axis counts, steps and VM and their respective standard deviations. The n^* value is 2016, making this data comparable to the $n^* = 2401$ and $N = 90$ combination in our simulations. The goal of the analysis was to cluster the youth into groups based on their physical activity profiles and determine if these groups agree with their clinical groupings or are capturing additional information about the participants. A visualization of patients $\mathbf{X}_{(2)i}$ matrices are available in Figure 4.1.

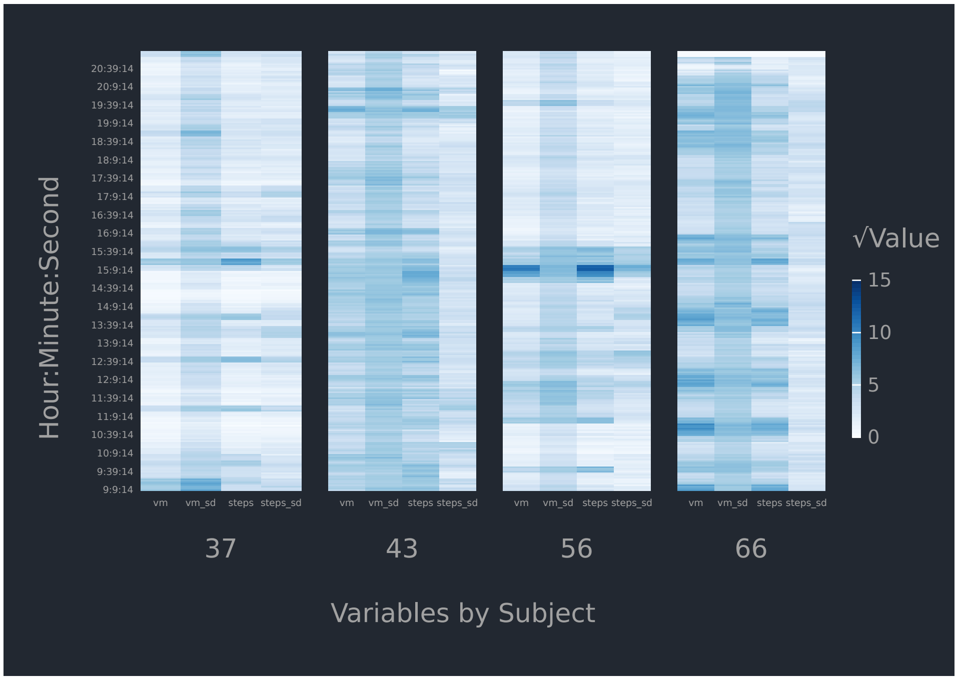


Figure 4.1: Accelerometer data from four random participants in the CHAMPION study. Four columns from the $\mathbf{X}_{(2)i}$ matrices are displayed.

Six different mixture models over 9 different group sizes were assessed. The models BIC values are summarized in Figure 4.2. A two or three group solution with an unstructured $\boldsymbol{\Delta}_{2g}$ resulted in better clustering solutions, as determined by the BIC. One group solutions were investigated but proved inferior to the two and three group models. The model with the largest BIC was the one with two groups and used the

VVI model for the temporal scale modes and VVV for the accelerometer metrics. None of the models we investigated had a singular scale matrix.

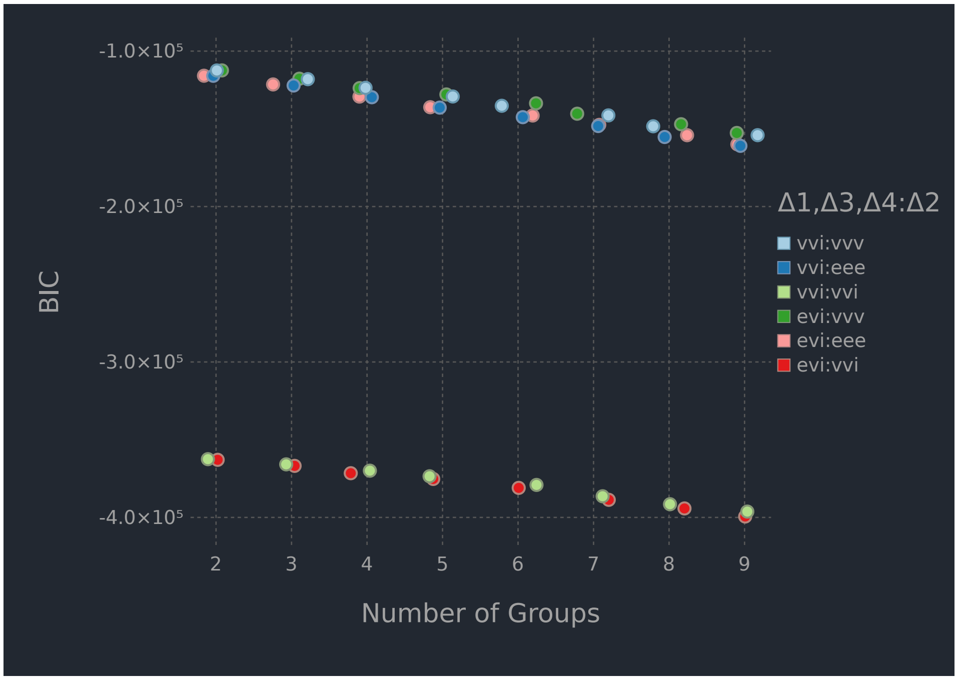


Figure 4.2: BIC by model type and nine different group sizes. The legend is interpreted as follows: the first three letters correspond to the temporal scale model and the last three letters correspond to the scale model used for the accelerometer metrics

To further improve on the clustering solution, we focused on two and three group solutions where each scale model could have any of the three appropriate options. The results are displayed in Figure 4.3. Models with unstructured scale matrices for the second and hour modes are preferred based on the BIC. These modes do not exhibit patterns of variation consistent with an autoregressive model, despite representing measurements taken over time. Models that use an autoregressive model to model the minutes (e.g. VVI or EVI) are consistently preferred. Using a VVV model to model the variation between the accelerometer metrics is preferred, suggesting the groups do not exhibit the same pattern of variation across the metrics. The VVI

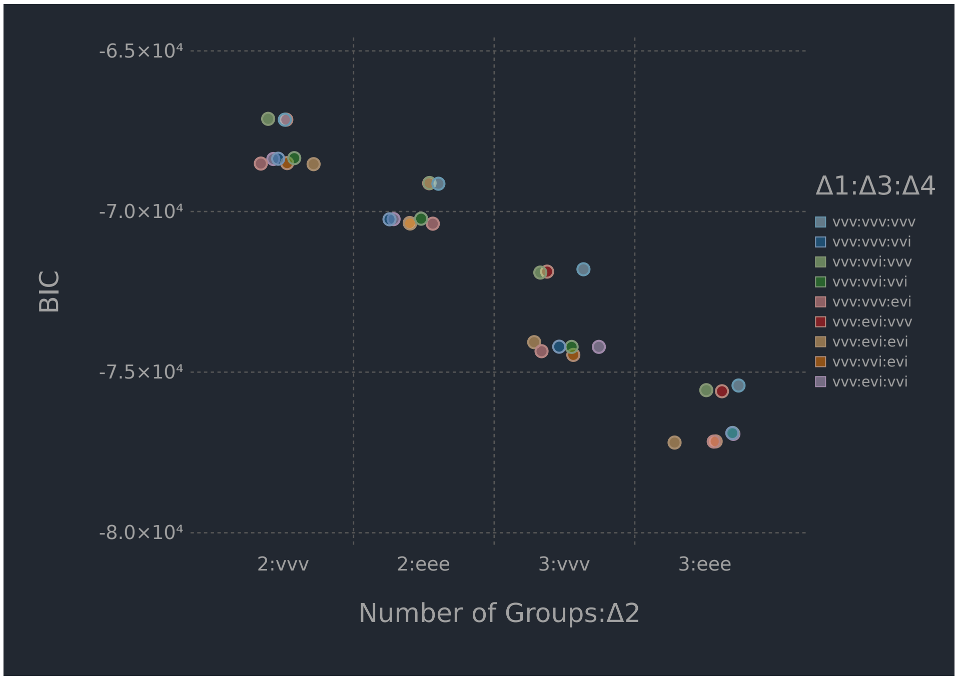


Figure 4.3: BIC by group size and model type. Includes horizontal jittering to clarify the points position. The legend is interpreted as follows: the first three letters correspond to the scale model used for $\mathbf{\Delta}_1$, the second three letters correspond to the scale model used for $\mathbf{\Delta}_3$ and the last three letters correspond to the scale model used for $\mathbf{\Delta}_4$

temporal model for $\mathbf{\Delta}_{3g}$ is preferred over EVI, implying that the groups have different autoregressive patterns in the minute mode of their MDAs. The final model is a two group solution with the VVI temporal model for $\mathbf{\Delta}_{3g}$ and unstructured scale matrices for the other three modes. With such a small number of groups, it is unlikely our mixture model is over-fitting the data.

Figure 4.4 indicates group 1 is the group which on average, has the most consistently active participants, moving at higher intensities through out the day. Group 2 exhibits brief periods of intense activity in the mid morning and late afternoon, illustrated by the deep purple bands in the VM and steps columns. Both groups have similar mean patterns of variation in their steps.

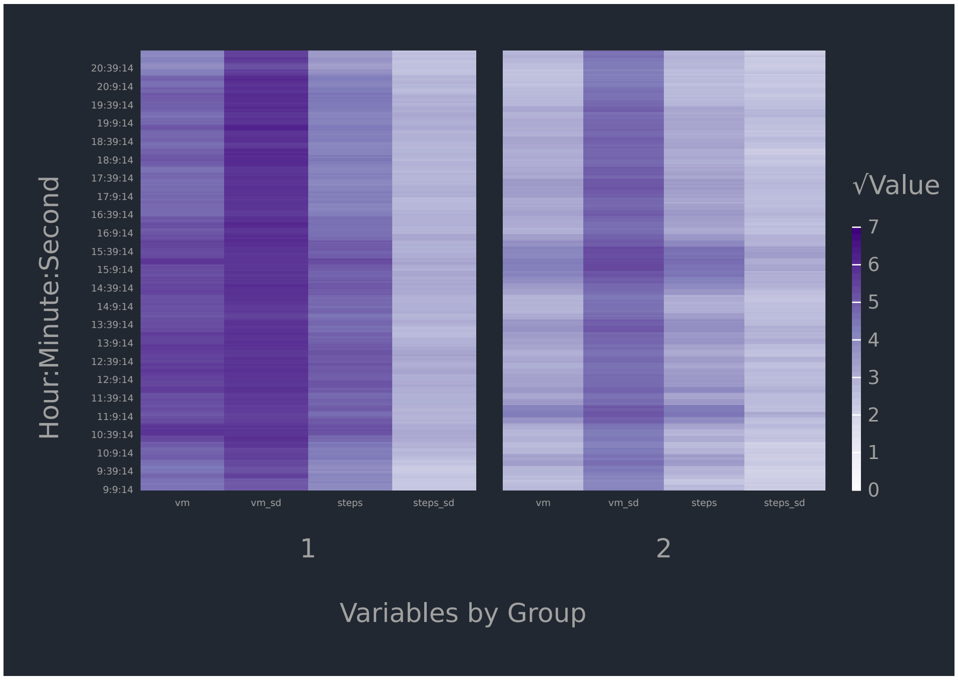


Figure 4.4: The mode 2 matricized mean MDA, $\mathbf{M}_{(2)g}$, which visualizes the mean activity patterns of each group's vector magnitude and steps

Figure 4.5 demonstrates that group 2 has more variation in their accelerometer metrics, as is evident by the deeper shades of blue and green in its heatmap. This makes sense in light of the participants periodic bouts of intense movement interspersed with longer periods of little activity. Steps have the most variation of the five metrics, followed by VM. As expected, VM covaries with the three axis values and steps covary with axis 1. In group 1, the more active participants, variation in VM (vm_sd) does not covary with any metric except VM, unlike group 2. Variation in steps (steps_sd) does not covary with any metric, including steps in group 1.

Figure 4.6 illustrates the pattern exhibited by the AR coefficients for the 10 minute intervals are generally decreasing with increasing lag between time points. This is characteristic of true longitudinal data. In general, the magnitude of the early AR

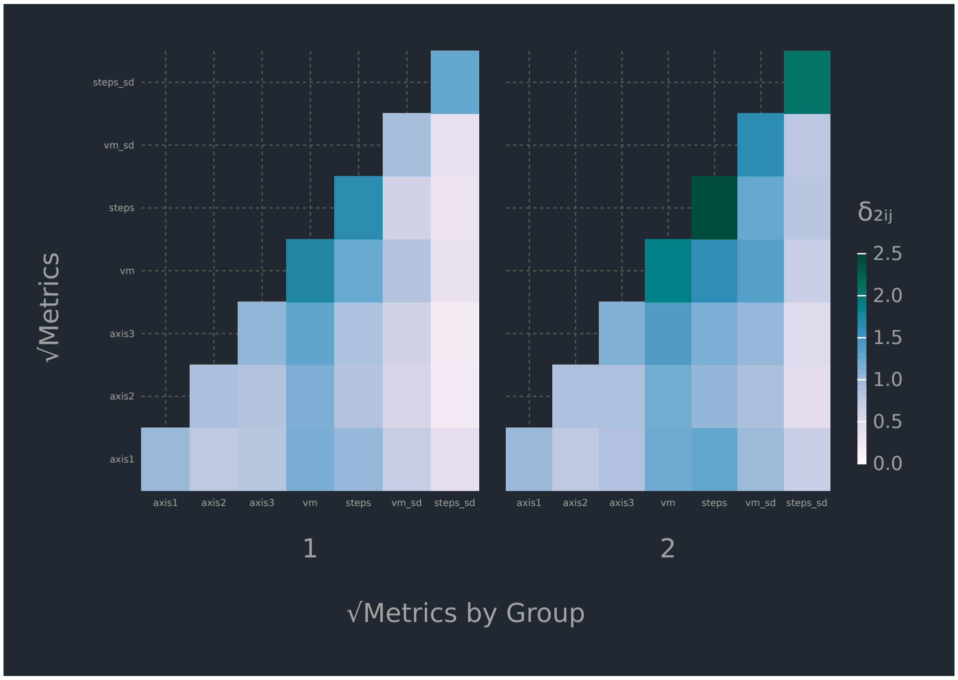


Figure 4.5: Represents the variation of the accelerometer metrics, $\mathbf{\Delta}_2$, plotted by group. The individual entries of the lower triangular portion of the matrix, δ_{2ij} are plotted as a heatmap.

relationships are shifted downward in group 2 relative to group 1. This suggests the AR relationships are slightly stronger in the active participants. Because the two sub-plots are nearly interchangeable, it is not surprising how similar the BIC results were for the models that used the VVI and EVI models for this mode. The two isotropic constraints representing the variation at each time point are nearly identical between the groups ($\delta_{31} = 0.984, \delta_{32} = 0.987$) and relatively small in magnitude.

The variation in the four 15 second intervals is visualized in Figure 4.7a. Group 1 exhibits more variation at this time scale. The level of covariation between the intervals remains nearly constant in both groups. This is contrary to the VVI and EVI temporal models, that would impose decreasing levels of variation as you move away from the main diagonal. Figure 4.7b displays the variation in the hours. There

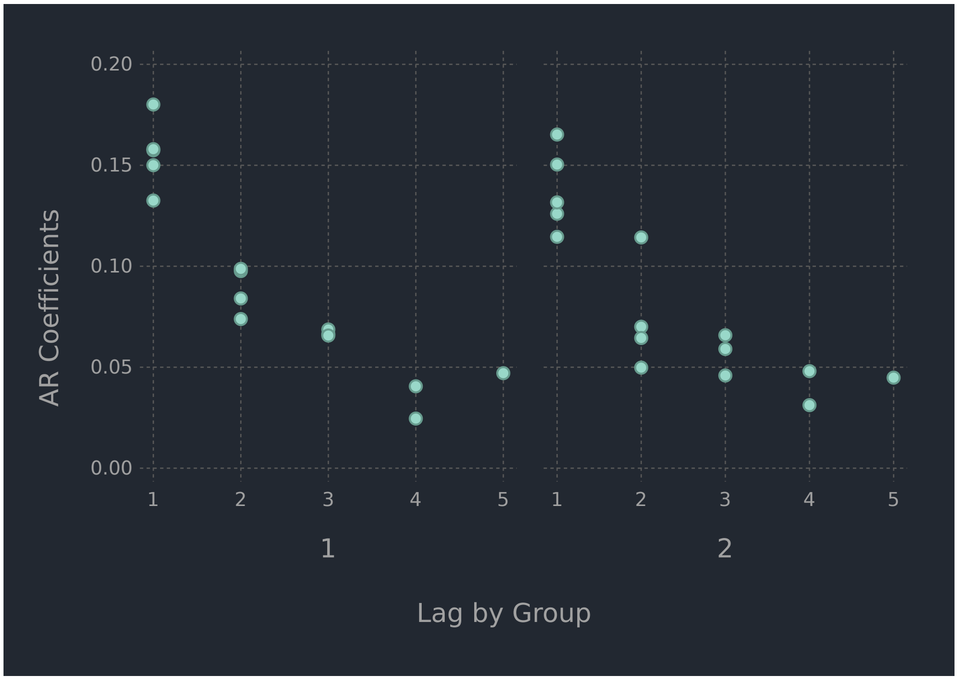
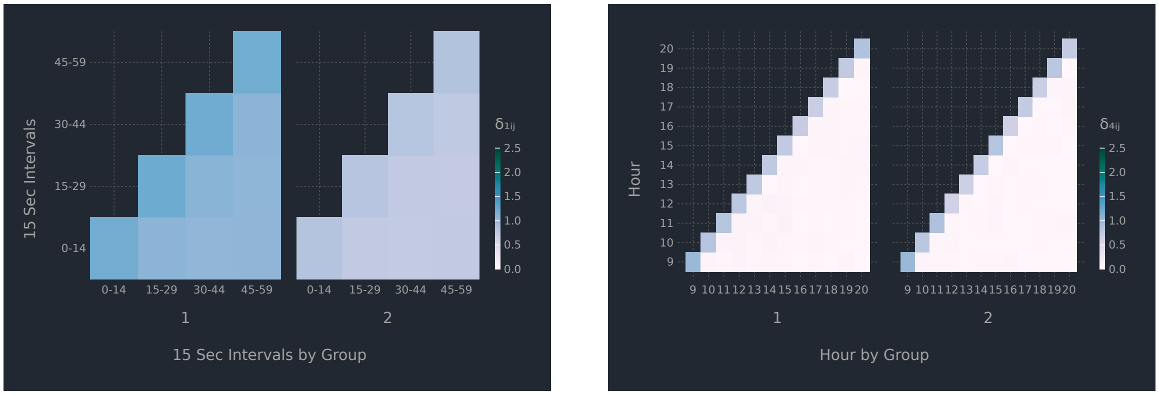


Figure 4.6: The autogressive coefficients from the $\mathbf{T}_{3g}$ matrices derived from the 10 minute intervals.



(a) $\mathbf{\Delta}_{g1}$

(b) $\mathbf{\Delta}_{g4}$

Figure 4.7: Variation in the four 15s intervals, $\mathbf{\Delta}_{g1}$ and the 12 hours, $\mathbf{\Delta}_{g4}$.

is little variation in each hour and it remains consistently low through out the day. Covariation is nearly absent between the hours, which suggests a VVI model from the GPCM family could be appropriate for this mode of the MDA data. These patterns

Table 4.1: Cross-tabulation of the clusters (based on MAP estimates) versus the medical diagnosis groupings.

Cluster	**Control**	**CF**	**IBD**	**JIA**	**T1D**	**Total(%)**
1	7	10	6	12	8	43(52)
2	7	4	11	9	9	40(48)

are evident in both groups and imply that the hour mode does not help distinguish between the participants in each cluster group. This aligns with the existing observations that children move in short bursts of activity through out the day and these patterns are more evident when analyzing higher resolution data at multiple time scales.

A cross-tabulation of the clusters (based on MAP estimates) versus the medical diagnosis groupings is given in Table 4.1. Each cluster represents roughly half the participants. The control and T1D participants are evenly distributed between the two groups. Group 1, the active participants, has the majority of the CF youth (71%) while group 2 has the majority of the IBD (65%) youth. The ARI for the table is 0.0007, indicating that the labels produced by the model are unrelated to the youth's membership in one of the five study groupings. This indicates that within each CIC, there is heterogeneity in the youth's physical activity profiles. Clinicians can use this to inform activity recommendations for low fit youth based on their high fit counterparts that share the same CIC.

The results of the finite mixture of MDAs model could be useful in a number of ways. Beyond characterizing the covariation in each mode of the samples MDAs, the group labels can be used in conjunction with traditional statistical models. We use them as one of four predictors in a quantile regression characterizing the relationship between predictors and aerobic fitness, defined as maximal oxygen uptake

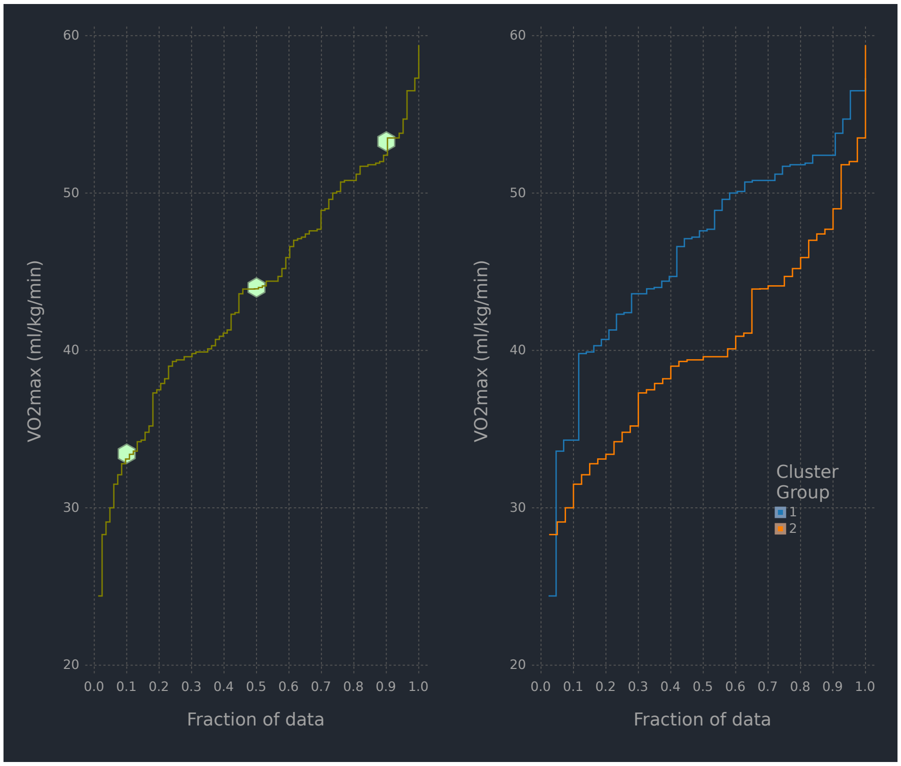


Figure 4.8: Empirical cumulative distribution plots of VO2max for all 83 participants and by cluster group. The green hexagons indicate the following quantiles (τ), used in our regression model: 10th = 33.4, median/50th = 44.0 and the 90th = 53.3.

or VO2max. VO2max is the gold-standard measurement of cardiorespiratory fitness and is the maximum rate of oxygen consumption measured during incremental exercise (VanPutte *et al.*, 2017). Our VO2max values are expressed as a relative rate, millilitres of oxygen per kilogram of body mass per minute (ml/kg/min). Figure 4.8 indicates: group 1, the active youth, have larger VO2max values across the quantiles which aligns with our expectation that the more active youth would have better cardiorespiratory fitness; our finite mixture model effectively captured the salient information about the participants physical activity from the accelerometer data. The 10th = 33.4 and 90th = 53.3 quantiles represent the youth in our sample with low

Table 4.2: Summary statistics by cluster.

Cluster	**Age**(years)		**Females**		**VO2max**(ml/kg/min)	
	Mean	SD	n	%	Mean	SD
1	11.8	2.63	14	33	47.0	6.82
2	14.2	2.05	29	73	40.6	7.10

and high cardiorespiratory fitness.

At the time of writing, we have access to the participants age, sex, study arm and VO2max values. Basic summary statistics by cluster group are listed in table 4.2. Group 1 is slightly younger, predominantly male and has higher cardiorespiratory fitness than group 2. Cluster 2 is older and predominantly female, suggesting daily patterns of physical activity are different for each sex in our sample. This is consistent with existing literature demonstrating that girls engage in less physical activity compared with boys, starting as early as the preschool years, in early childhood, as well as in adolescence (Proudfoot *et al.*, 2019).

The quantile regression results are summarized in Figure 4.9. For the less fit participants, being in the active group vs the inactive group has a relatively large and statistically significant effect on VO2max at the 5% level — $\beta = 8.68$(95%CI $[0.88; 16.36]$) — while being female vs male, trends towards a decreased VO2max — $\beta = -6.25$(95%CI $[-13.56; 1.06]$). When controlled for the other predictors, the effect of the study arm and age do not have an important relationship with VO2max. The effect of age on VO2max is likely muted by the inclusion of our cluster groups, which differ in age. Similar trends are evident at the median VO2max, with both cluster group — $\beta = 4.89$(95%CI $[0.11; 9.67]$) — and sex — $\beta = -6.03$(95%CI $[9.84; -2.21]$) — being statistically significant. At the 90th quantile, the high fit youth do not have any statistically significant relationships between the predictors and VO2max. The same trends in MAP and sex are evident here and being in the control group may confer some increase in cardiorespiratory fitness.

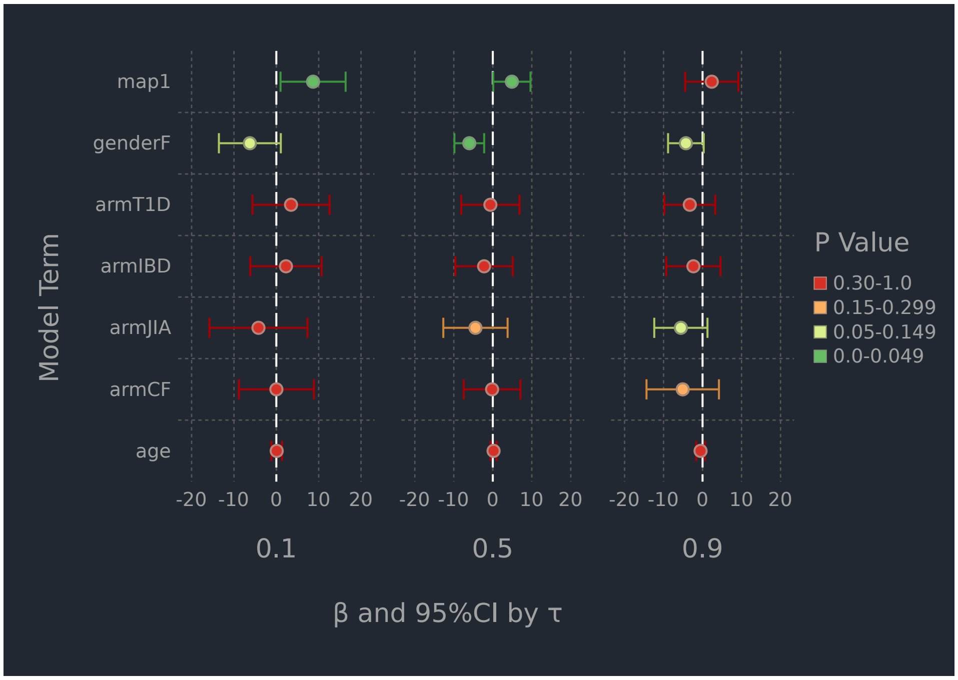


Figure 4.9: Quantile regression coefficients (β), their 95% confidence intervals(CI), and P-values. The β values for the study arms are relative to the control group. Interactions between the models main effects were not included due to sample size restrictions.

In summary, our finite mixture of MDAs model found two groups in the five-way accelerometer data from the CHAMPION study. One group represents participants that are consistently active through out the day and another that exhibits long periods of light activity interspersed with intervals of intense activity. The groups are differentiated by their mean activity profiles, variation at the level of 15 second intervals and between their accelerometer metrics and the strength of the AR relationships between their 10 min intervals. These groupings do not agree with the participants clinical diagnoses, suggesting the accelerometers are capturing information not included in these diagnoses. The group labels produced by our model were good predictors of the participants cardiorespiratory fitness, as measured by VO2max, when the participants had low to median fitness.

4.3 Discussion

Physical activity data, as measured by accelerometers, was conceptualized as a sample of five-way data, where each participant has an order-4 MDA made up of their accelerometer data aggregated over different nested time scales. A finite mixture of MDAs approach is introduced for clustering a sample of MDA data. These models are innovative because they allow MDA data to be analyzed in its natural form, without the need to transform it to meet the limitations of existing model-based clustering methods. Our model provides scientifically relevant parameters for each group, characterizing the mean MDAs, the variation in each mode of the MDA and labels that can be used in conjunction with more traditional statistical methods. These parameters can help non-statisticians make practical decisions and guide their scientific messaging around the MDA sample being analyzed.

Clinical data, like the data from the CHAMPION study, often includes non-continuous covariates. These covariates can be informative (e.g., sex, disease status) and should be included in the clustering solutions whenever possible. One way to do this would be to model the mean of each mixture component with a linear model (McNicholas and Subedi, 2012) that incorporates these covariates. The matrix $\mathbf{M}_{(1)g}$ could be modelled using a matrix-variate regression model that models matrix valued responses (Ding and Cook, 2018). Alternatively, the MDA $\boldsymbol{\mathfrak{M}}_g$ could be modelled using a tensor-variate regression model (Li and Zhang, 2017).

We expect clustering MDA data would benefit from some form of dimension reduction. The VVI model that constrained the eigen-decomposition of $\mathbf{\Delta}_{g2}$, had n_d vs. n_d^2 parameters in the EEE and VVV models. Despite this large reduction in free parameters, the VVI model was not able to improve the performance of the mixture model, as measured by BIC. Flexibility in how the variation of each mode of the MDA is modelled is an important feature of these models and results in better clustering solutions in Section 4.2. We expect this to be true when doing dimension reduction as well. In this vein, a finite mixture of MDA factor analyzers could be developed, and can be viewed as an extension of the work of Tang *et al.* (2013) and Gallaugher and McNicholas (2018b).

Chapter 5

Tensor-Variate Skew Distributions

We use the variance-mean mixture formulation, described in Section 2.8, to describe 5 tensor-variate skewed distributions (Gallaugher *et al.*, 2021). These 5 distributions include the tensor-variate generalized hyperbolic, variance-gamma, shifted asymmetric Laplace, normal inverse Gaussian and skew t distributions. The abbreviations used for these distributions are detailed in Table 5.1. Their matrix variate counterparts are detailed in Gallaugher and McNicholas (2019).

Table 5.1: Distribution abbreviations.

Tensor-Variate Distribution	**Abbreviation**	**Plot Abbreviation**
Normal	MLND	norm
Generalized hyperbolic	TVGH	gh
Normal inverse Gaussian	TVNIG	nig
Shifted asymmetric Laplace	TVSAL	sal
Skew t	TVST	st
Variance-gamma	TVVG	vg

To find the variance-mean mixture formulation, we say an order-D random MDA $\mathscr{X}$, with dimensions $\mathbf{n}$ has a one of the 5 tensor-variate distribution, if $\mathscr{X}$ can be written as

$$\mathscr{X} = \boldsymbol{\mathfrak{M}} + W\boldsymbol{\mathfrak{A}} + \sqrt{W}\mathscr{V}, \tag{5.1}$$

where $\boldsymbol{\mathfrak{M}}$ and $\boldsymbol{\mathfrak{A}}$ are $\mathbf{n}$ dimensional order-D MDAs and $\mathscr{V} \sim \mathcal{N}_{\mathbf{n}}\left(\mathbb{O}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d\right)$. For each of the five tensor-variate distributions, the distribution of W is summarized in Table 5.2.

Table 5.2: The distribution of W for each of the five tensor-variate distributions

TVD	**W**	**Density**
TVGH	$GIG(\omega, 1, \lambda)$	$g(w\mid\omega, 1, \lambda) = \frac{(w)^{\lambda-1}}{2K_\lambda(\omega)} \exp\left\{-\frac{\omega}{2}\left(w + \frac{1}{w}\right)\right\}$
TVVG	Gamma(γ, γ)	$f(w\mid\gamma, \gamma) = \frac{\gamma^\gamma}{\Gamma(\gamma)} w^{\gamma-1} \exp\{-\gamma w\}$
TVSAL	Gamma$(1, 1)$	$f(w\mid 1, 1) = \frac{1}{w} \exp\{-w\}$
TVNIG	Inv-Gaussian$(1, \kappa)$	$f(w\mid 1, \kappa) = 2\pi^{-\frac{1}{2}} \exp\{\kappa\} w^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\left(\frac{1}{w} + \kappa^2 w\right)\right\}$
TVST	Inv-Gamma$\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$	$f(w\mid\frac{\nu}{2}, \frac{\nu}{2}) = \frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \left[\frac{1}{w}\right]^{\frac{\nu}{2}+1} \exp\left\{\frac{\nu}{2w}\right\}$

It then follows that

$$\mathscr{X} | W = w \sim \mathcal{N}_{\mathbf{n}}\left(\boldsymbol{\mathfrak{M}} + w\boldsymbol{\mathfrak{A}}, w\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d\right)$$

and thus the joint density of $\mathscr{X}$ and W is $f(\boldsymbol{\mathfrak{x}}, w|\boldsymbol{\vartheta}) = f(\boldsymbol{\mathfrak{x}}|W = w)f(w)$. This joint density formulation will have a specific form for each of the five tensor-variate distributions.

5.1 Marginal densities

The mathematical details for the derivation of the TVST marginal distribution are given in Appendix D.1.1. The derivations for the other distributions would be similar with the difference being which distribution of W we start with. The marginal densities are listed below, starting with the TVST distribution

$$f_{\text{TVST}}(\boldsymbol{\mathfrak{X}}|\boldsymbol{\vartheta}) = \frac{2\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}} \exp\left\{\text{vec}(\boldsymbol{\mathfrak{X}} - \boldsymbol{\mathfrak{M}})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1} \text{vec}(\boldsymbol{\mathfrak{A}})\right\}}{(2\pi)^{\frac{n^*}{2}} \prod_{d=1}^{D} |\boldsymbol{\Delta}_d|^{\frac{n^*}{2n_d}} \Gamma(\frac{\nu}{2})} \left(\frac{\delta(\boldsymbol{\mathfrak{X}}; \boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}) + \nu}{\rho(\mathbf{A}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1})}\right)^{-\frac{\nu+n^*}{4}}$$
$$\times K_{-\frac{\nu+n^*}{2}}\left(\sqrt{\left[\rho(\boldsymbol{\mathfrak{A}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}))\right]\left[\delta(\boldsymbol{\mathfrak{X}}; \boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}) + \nu\right]}\right) \quad (5.2)$$

for $\nu \in \mathbb{R}^+$. For convenience, we will denote this distribution by $\text{TVST}(\boldsymbol{\mathfrak{M}}, \boldsymbol{\mathfrak{A}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d, \nu)$. The density of the TVGH distribution is given by

$$f_{\text{TVGH}}(\boldsymbol{\mathfrak{X}}|\boldsymbol{\vartheta}) = \frac{\exp\left\{\text{vec}(\boldsymbol{\mathfrak{X}} - \boldsymbol{\mathfrak{M}})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1} \text{vec}(\boldsymbol{\mathfrak{A}})\right\}}{(2\pi)^{\frac{n^*}{2}} \prod_{d=1}^{D} |\boldsymbol{\Delta}_d|^{\frac{n^*}{2n_d}} K_\lambda(\omega)} \left(\frac{\delta(\boldsymbol{\mathfrak{X}}; \boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}) + \omega}{\rho(\mathbf{A}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}) + \omega}\right)^{\frac{\lambda - \frac{n^*}{2}}{2}}$$
$$\times K_{\lambda - n^*/2}\left(\sqrt{\left[\rho(\boldsymbol{\mathfrak{A}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}) + \omega)\right]\left[\delta(\boldsymbol{\mathfrak{X}}; \boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}) + \omega\right]}\right) \quad (5.3)$$

for $\lambda \in \mathbb{R}$, $\omega \in \mathcal{R}^+$. We will denote the tensor-variate generalized hyperbolic distribution by $\text{TVGH}(\boldsymbol{\mathfrak{M}}, \boldsymbol{\mathfrak{A}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d, \lambda, \omega)$.

The density of the TVVG distribution is

$$f_{\text{TVVG}}(\boldsymbol{\mathfrak{X}}|\boldsymbol{\vartheta}) = \frac{2\gamma^\gamma \exp\left\{\text{vec}(\boldsymbol{\mathfrak{X}} - \boldsymbol{\mathfrak{M}})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1} \text{vec}(\boldsymbol{\mathfrak{A}})\right\}}{(2\pi)^{\frac{n^*}{2}} \prod_{d=1}^{D} |\boldsymbol{\Delta}_d|^{\frac{n^*}{2n_d}} \Gamma(\gamma)} \left(\frac{\delta(\boldsymbol{\mathfrak{X}}; \boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1})}{\rho(\mathbf{A}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}) + 2\gamma}\right)^{\frac{\gamma - \frac{n^*}{2}}{2}}$$
$$\times K_{\gamma - n^*/2}\left(\sqrt{\left[\rho(\boldsymbol{\mathfrak{A}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}) + 2\gamma\right]\left[\delta(\boldsymbol{\mathfrak{X}}; \boldsymbol{\mathfrak{M}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1})\right]}\right), \quad (5.4)$$

where $\gamma \in \mathbb{R}^+$. We will denote this distribution by $\text{TVVG}(\boldsymbol{\mathfrak{M}}, \boldsymbol{\mathfrak{A}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d, \gamma)$. The tensor-variate SAL (TVSAL) distribution would naturally arise as a special case of the TVVG with $\gamma = 1$.

Lastly, the TVNIG distribution has the following density function

$$
\begin{aligned}
f_{\text{TVNIG}}(\boldsymbol{\mathfrak{X}}|\boldsymbol{\vartheta}) =& \frac{2\exp\left\{\text{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\text{vec}(\boldsymbol{\mathfrak{A}})+\kappa\right\}}{(2\pi)^{\frac{n^*+1}{2}}\prod_{d=1}^{D}|\boldsymbol{\Delta}_d|^{\frac{n^*}{2n_d}}}\left(\frac{\delta(\boldsymbol{\mathfrak{X}};\boldsymbol{\mathfrak{M}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+1}{\rho(\mathbf{A},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+\kappa^2}\right)^{-\frac{1+n^*}{4}}\\
&\times K_{-\frac{1+n^*}{2}}\left(\sqrt{\left[\rho(\boldsymbol{\mathfrak{A}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+\kappa^2\right]\left[\delta(\boldsymbol{\mathfrak{X}};\boldsymbol{\mathfrak{M}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+1\right]}\right),
\end{aligned}
\tag{5.5}
$$

where $\kappa \in \mathbb{R}^+$. We will use the notation $\text{TVNIG}(\boldsymbol{\mathfrak{M}},\boldsymbol{\mathfrak{A}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d,\kappa)$ to refer to this distribution.

Like with the tensor-variate normal distribution, the skewed distributions are closely related to their lower order counterparts. These relationships are summarized in Corollary 5.1.1 (to Theorem 2.4.1):

Corollary 5.1.1 *Let $TVD_{\mathbf{n}}(\boldsymbol{\mathfrak{M}},\boldsymbol{\mathfrak{A}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d,\boldsymbol{\theta})$ represent one of the four skewed tensor distributions of dimension $\mathbf{n}$, where $\boldsymbol{\theta}$ represents the additional parameters specific to the distribution. Let $MVD_{n\times p}(\mathbf{M},\mathbf{A},\boldsymbol{\Sigma},\boldsymbol{\Psi},\boldsymbol{\theta})$ represent the corresponding matrix variate distribution. Finally, let $D(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\Sigma},\boldsymbol{\theta})$ represent the corresponding multivariate distribution. The following statements are then equivalent.*

1. $\mathscr{X} \sim TVD_{\mathbf{n}}\left(\boldsymbol{\mathfrak{M}},\boldsymbol{\mathfrak{A}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d,\boldsymbol{\theta}\right)$
2. $\mathscr{X}_{(j)} \sim MVD_{\frac{n^*}{n_j}\times n_j}\left(\boldsymbol{\mathfrak{M}}_{(j)},\boldsymbol{\mathfrak{A}}_{(j)},\bigotimes_{d\neq j}\boldsymbol{\Delta}_d,\boldsymbol{\Delta}_j,\boldsymbol{\theta}\right)$
3. $vec(\mathscr{X}_{(j)}) \sim D_{n^*}\left(vec(\boldsymbol{\mathfrak{M}}_{(j)}),vec(\boldsymbol{\mathfrak{A}}_{(j)}),\boldsymbol{\Delta}_j\otimes\bigotimes_{d\neq j}\boldsymbol{\Delta}_d,\boldsymbol{\theta}\right)$

5.2 Expectations

The expectations for these four distributions can be easily calculated using iterative expectation and equation (5.1). The general form is $\mathbb{E}[\mathscr{X}] = \mathfrak{M} + \mathbb{E}[W]\mathfrak{A}$. This leads to the the following expectations:

$$\mathscr{X} \sim \text{TVST}(\mathfrak{M}, \mathfrak{A}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d, \nu) \implies \mathbb{E}[\mathscr{X}] = \mathfrak{M} + \frac{\nu}{\nu - 2}\mathfrak{A} \qquad (\nu > 2), \tag{5.6}$$

$$\mathscr{X} \sim \text{TVGH}(\mathfrak{M}, \mathfrak{A}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d, \lambda, \omega) \implies \mathbb{E}[\mathscr{X}] = \mathfrak{M} + \frac{K_{\lambda+1}(\omega)}{K_{\lambda}(\omega)}\mathfrak{A}, \tag{5.7}$$

$$\mathscr{X} \sim \text{TVVG}(\mathfrak{M}, \mathfrak{A}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d, \gamma) \implies \mathbb{E}[\mathscr{X}] = \mathfrak{M} + \mathfrak{A}, \tag{5.8}$$

$$\mathscr{X} \sim \text{TVNIG}(\mathfrak{M}, \mathfrak{A}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d, \kappa) \implies \mathbb{E}[\mathscr{X}] = \mathfrak{M} + \frac{1}{\kappa}\mathfrak{A}. \tag{5.9}$$

These expectations will be used by our software implementation to find the "mean" MDA for the sample of four or five-way data being analyzed.

Theorem 5.2.1 *If we define the D-order MDA $\mathscr{Z} \sim \mathcal{N}\left(\mathbb{O}, \bigotimes_{d=1}^{D} \mathbf{I}_d\right)$ we can use a Tucker product (Kolda and Bader, 2009) to define*

$$\mathscr{V} = \mathscr{Z} \times \boldsymbol{\Delta}^{\frac{1}{2}} = \mathscr{Z} \times_1 \boldsymbol{\Delta}_1^{\frac{1}{2}} \times_2 \boldsymbol{\Delta}_2^{\frac{1}{2}} \cdots \times_D \boldsymbol{\Delta}_D^{\frac{1}{2}}.$$

Let an equivalent mode 1 matricized version of $\mathscr{V}$ be

$$\mathbf{V}_{(1)} = \boldsymbol{\Delta}_1^{\frac{1}{2}} \mathbf{Z}_{(1)} \left(\bigotimes_{d=D}^{2} \boldsymbol{\Delta}_d^{\frac{1}{2}} \right)^{\top}$$

and of $\mathscr{X}$ *be*

$$\mathbf{X}_{(1)} = \mathbf{M}_{(1)} + W\mathbf{A}_{(1)} + \sqrt{W}\mathbf{V}_{(1)}.$$

Then the following second moments can be found

$$\begin{aligned}
Cov[vec(\mathscr{X})] = \; & vec(\mathbf{M}_{(1)})vec(\mathbf{M}_{(1)})^\top + \mathbb{E}[W]vec(\mathbf{M}_{(1)})vec(\mathbf{A}_{(1)})^\top \\
& + \mathbb{E}[W]vec(\mathbf{A}_{(1)})vec(\mathbf{M}_{(1)})^\top \\
& + \mathbb{E}[W^2]vec(\mathbf{A}_{(1)})vec(\mathbf{A}_{(1)})^\top + \mathbb{E}[W]\left(\bigotimes_{d=D}^{1}\mathbf{\Delta}_d\right) \qquad (5.10) \\
\mathbb{E}\left[\mathbf{X}_{(1)}\mathbf{X}_{(1)}^\top\right] = \; & \mathbf{M}_{(1)}\mathbf{M}_{(1)}^\top + \mathbb{E}\left[W\right]\mathbf{M}_{(1)}\mathbf{A}_{(1)}^\top + \mathbb{E}\left[W\right]\mathbf{A}_{(1)}\mathbf{M}_{(1)}^\top \\
& + \mathbb{E}\left[W^2\right]\mathbf{A}_{(1)}\mathbf{A}_{(1)}^\top + \mathbb{E}[W]\mathbf{\Delta}_1 \times \prod_{d=2}^{D} tr\{\mathbf{\Delta}_d\} \qquad (5.11) \\
\mathbb{E}\left[\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\right] = \; & \mathbf{M}_{(1)}^\top\mathbf{M}_{(1)} + \mathbb{E}\left[W\right]\mathbf{M}_{(1)}^\top\mathbf{A}_{(1)} + \mathbb{E}\left[W\right]\mathbf{A}_{(1)}^\top\mathbf{M}_{(1)} \\
& + \mathbb{E}\left[W^2\right]\mathbf{A}_{(1)}^\top\mathbf{A}_{(1)} + \mathbb{E}[W]\left(\bigotimes_{d=D}^{2}\mathbf{\Delta}_d\right) \times \; tr\{\mathbf{\Delta}_1\} \qquad (5.12)
\end{aligned}$$

The proof of this theorem is given in Appendix D.1.2. Equivalent expressions can be found for different modes of $\mathscr{X}$ by using different matricizations.

5.3 Parameter estimation

We use an expectation conditional maximization (ECM) algorithm (Meng and Rubin, 1993) to estimate the parameters. The ECM algorithm will be described in detail in Chapter 6. This is due to the fact that we can view these distributions as a special case of a mixture, where $G = 1$. The software implementation will be discussed in Chapter 6 as well.

We use the BIC, described in Section 3.3, to do the model selection. For these distributions,

$$\rho = 2n^* + \sum_{d=1}^{D} n_d(n_d + 1) + 1.$$

The identifiability issues and the stopping rule we discussed in Chapter 3 apply here as well.

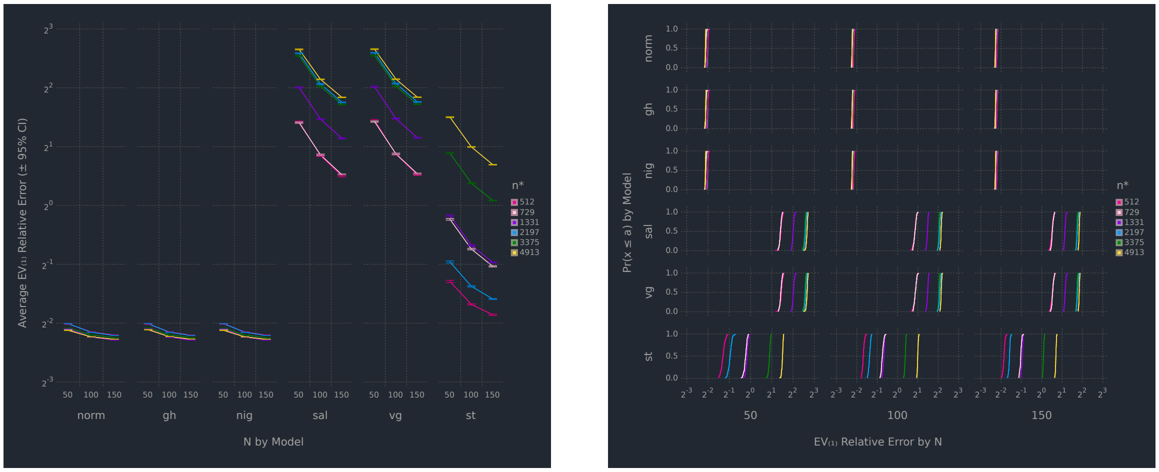
5.4 Simulation study

We conduct a simulation study to investigate the effect of different sample and MDA sizes on our ability to effectively estimate the model parameters. The simulations are conducted using order-3 MDAs. We consider sample sizes $N \in \{50, 100, 150\}$. The n^* quantity is used to measure the different dimensions of the MDAs. Its values include 512, 729, 1331, 2197, 3375 and 4813. The simplest way to visualize the resulting MDAs is an order 3 MDA with three equal dimension lengths of 8, 9, 11, 13, 15 and 17. For each combination of N and n^*, 100 datasets are simulated. We compare the ECM algorithm for the four skewed tensor-variate distributions to the flip-flop algorithm for the tensor-variate normal distribution described in (Manceur and Dutilleul, 2013). Our version of the flip-flop algorithm uses the MLND with the traces described in Appendix A.2. Like Section 3.7, we use the relative error to determine how close the estimated model parameters are to the true parameters.

Normal data

We simulate the tensor-variate Normal data using (3.14). A signal-to-noise ratio of one half was applied to the simulated data prior to analysis. The parameters used to simulate the data were generated in the same way as described in Section 3.7.

The ECM and flip-flop algorithms all converge in three iterations. Figure 5.1a visualizes the mean and 95% confidence intervals for the relative error in $\mathbb{E}[\mathscr{X}]$ across the values of N and n^*. As expected, the flip-flop algorithm (e.g. norm) estimates $\boldsymbol{\mathfrak{M}}$ well. The TVGH and TVNIG have nearly identical performance, which does not degrade as N and n^* increase in size. The other three tensor-variate models do a poor job of estimating $\mathbb{E}[\mathscr{X}]$. Their performance worsens as N decreases and n^* increases. Figure 5.1b indicates that the distributions of relative errors do not have long tails.

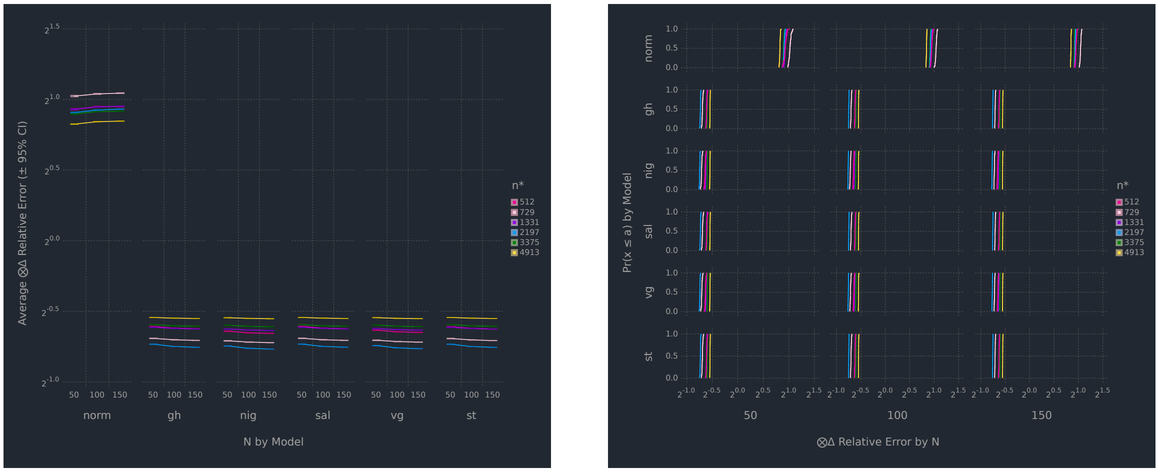


(a) Average and 95% confidence intervals for the relative error.

(b) Empirical distribution plots of the relative error.

Figure 5.1: Simulation results for the mode 1 matricization of $\mathbb{E}[\mathscr{X}]$ (normal).

A different picture emerges when we look at the relative error in $\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d$, visualized in Figure 5.2a. As is clear from (3.10), we can accurately estimate the overall Kronecker product versus the individual scale matrices. The flip-flop algorithm fairs poorly relative the ECM algorithms, all of which perform similarly across the range of sample and MDA sizes. Figure 5.2b indicates all the distributions have small tails and the flip-flop algorithm is shifted to the right of the other models.



(a) Average and 95% confidence intervals for the relative error

(b) Empirical distribution plots of the relative error.

Figure 5.2: Simulation results for $\bigotimes_{d=1}^{D} \mathbf{\Delta}_d$ (normal).

Figure 5.3 summarizes the average BIC values and the models ranks for the 100 simulations, across the values of N and n^* for each of the models. Based on their BIC values, the flip-flop algorithm is consistently the top performer, despite doing a poor job estimating the scale matrices. Of the skewed models, the TVGH and TVNIG models ranked highest.

Skewed data

We used (5.1) to generate the data from a TVST distribution with $\nu = 4$. The different models had a lot of variation in the number of iterations they took to converge to a solution. Typically the flip-flop algorithm converges in a median of 3 iterations and the ECM algorithms converge in a median of 4-6 iterations. Figure 5.4 summarizes the distribution of iterations for all values of N and n^*. The flip-flop algorithm had long tails for all values of N and n^*, with the longest tails occurring for the smallest MDAs ($n^* = 512$). The tails of the ECM algorithm distributions decrease as

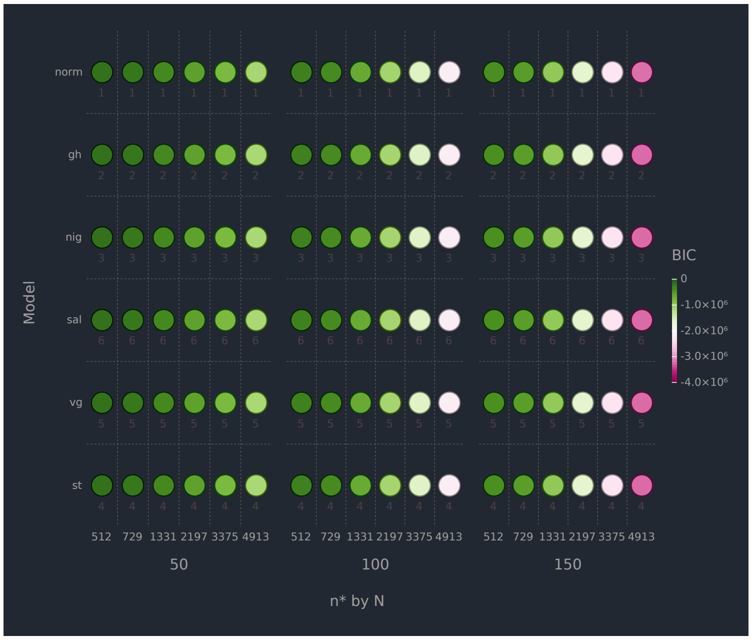


Figure 5.3: Average BIC and rank of the models for each combination of n^* and N (normal).

the sample size increases. Aside from the TVGH, the values of n^* do not affect the distribution of iterations.

Figure 5.5a visualizes the mean and 95% confidence intervals for the relative error in $\mathbb{E}[\mathscr{X}]$ across the values of N and n^*. The flip-flop algorithm and the TVNIG model have reasonable performance. The TVST, TVVG and TVSAL models perform well, accurately estimating $\mathbb{E}[\mathscr{X}]$ in all scenarios. The TVGH results are highly variable. This can be explained by the array of GIG parameter values learned from the data, resulting in GIG distributions that look nothing like the underlying Inv-Gamma $\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ distribution used to generate the data. See Figures 5.8–5.9b for further details. Figure 5.5b indicates the distribution of $\mathbb{E}[\mathscr{X}]$ have long right tails for the flip-flop algorithm, TVNIG and TVGH models.

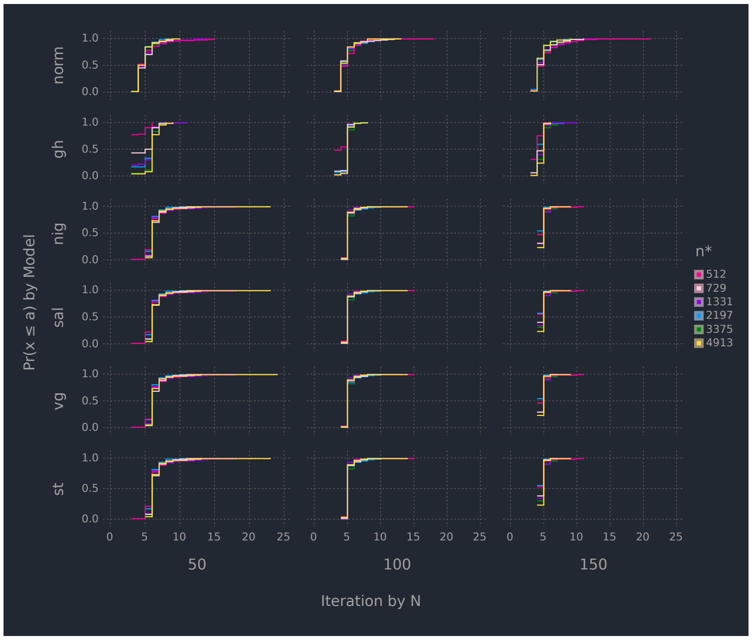
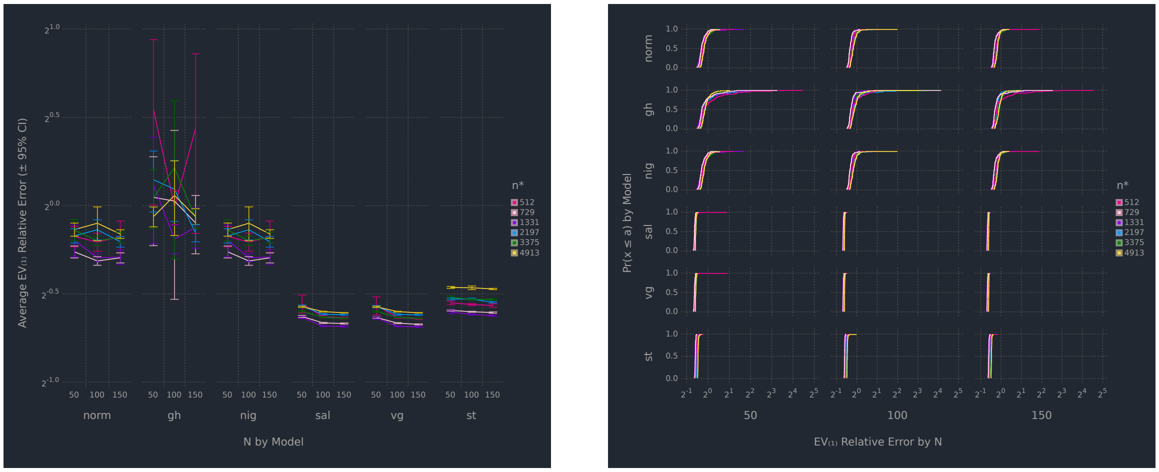


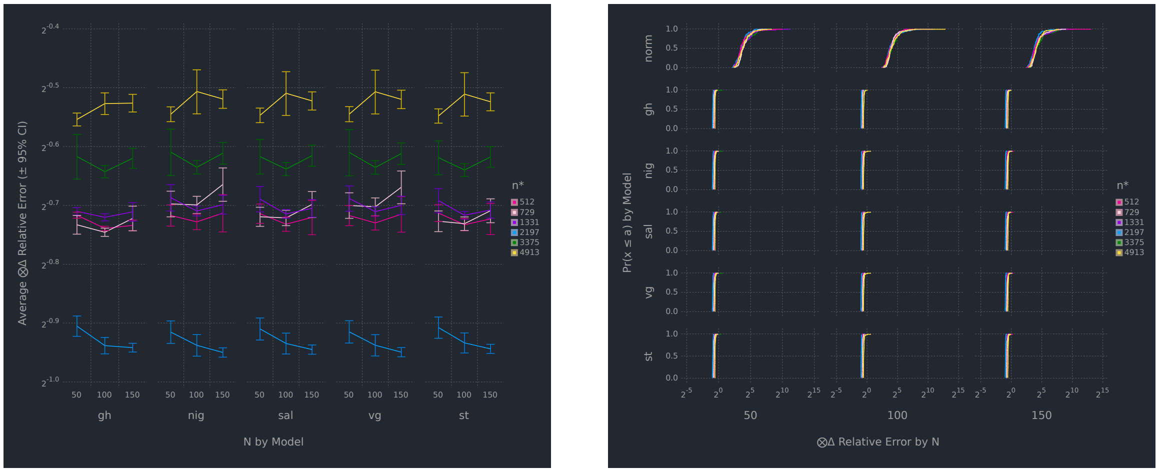
Figure 5.4: Number of iterations for each model by each combination of N and n^* (skewed).



(a) Average and 95% confidence intervals for the relative error.

(b) Empirical distribution plots of the relative error.

Figure 5.5: Simulation results for the mode 1 matricization of $\mathbb{E}[\mathscr{X}]$ (skewed).



(a) Average and 95% confidence intervals for the relative error.

(b) Empirical distribution plots of the relative error.

Figure 5.6: Simulation results for $\bigotimes_{d=1}^{D} \mathbf{\Delta}_d$ (skewed).

Figure 5.6a indicates all the skewed tensor-variate models do an excellent job of estimating $\bigotimes_{d=1}^{D} \mathbf{\Delta}_d$ across the range of N and n^* values. Their performance degrades slightly for the two largest values of n^* . The flip-algorithm has a median relative error of nearly 32. The distributions of the $\bigotimes_{d=1}^{D} \mathbf{\Delta}_d$ relative errors are displayed in Figure 5.6b.

It is also possible to visualize the relative error in $\mathbf{M}_{(1)}$ and $\mathbf{A}_{(1)}$, although we do not do that here. Their empirical cummulative distribution plots look similar to Figure 5.5b.

Figure 5.7 indicates the flip-flop algorithm is consistently the poorest performer among the models, ranking last for each combination of N and n^*. For small to moderate sized MDAs, the TVNIG model consistently ranks the highest. As n^* reaches its maximum size, the TVST model overtakes the TVNIG model in the rankings.

The variability in the TVGH $\mathbb{E}[\mathscr{X}]$ results, visualized in Figures 5.5a and 5.5b, can be explained by the array of GIG parameter values learned from the data. Each

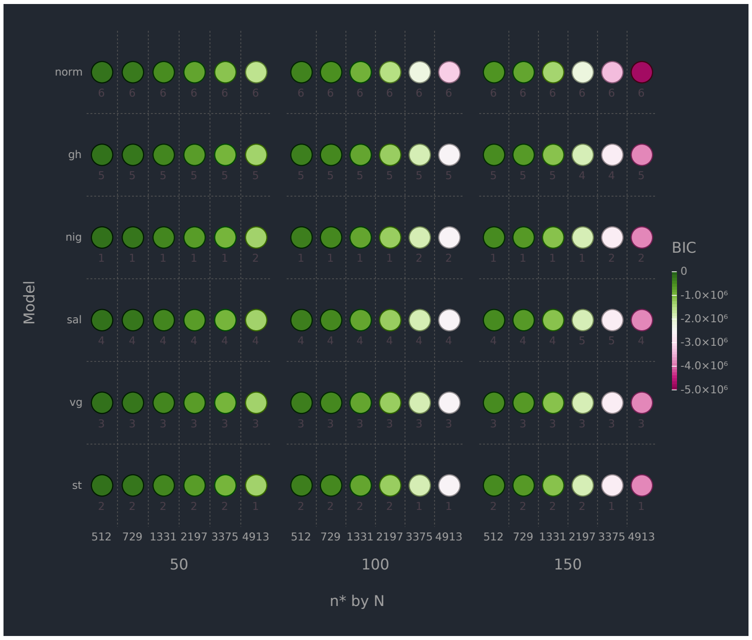


Figure 5.7: Average BIC and rank of the models for each combination of N and n^* (skewed).

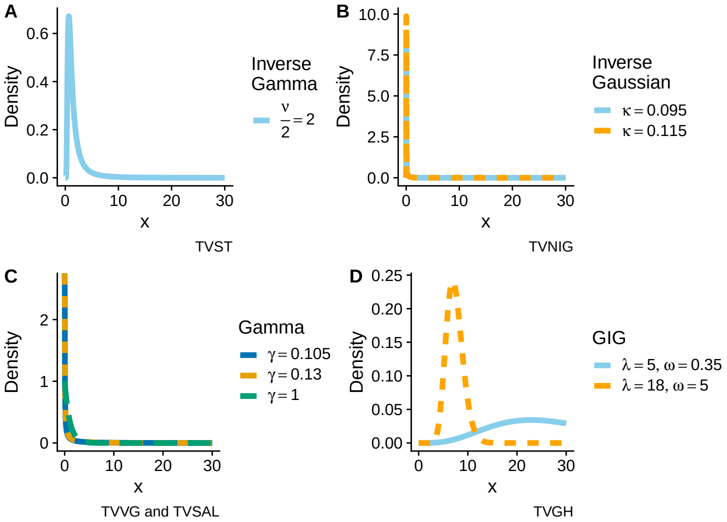
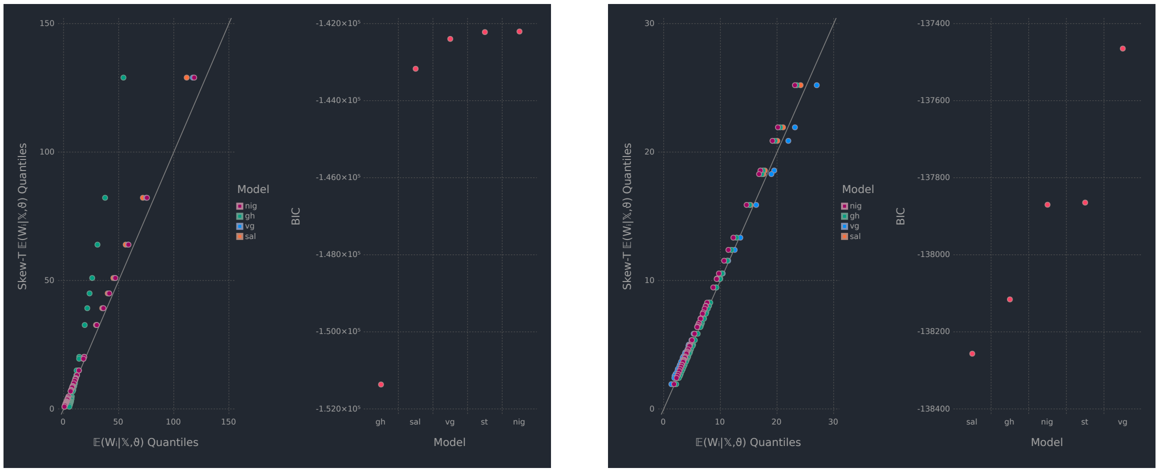


Figure 5.8: The underlying distributions of W_{ig} for each of the five tensor-variate distributions.

of the underlying distributions of W_{ig} are visualized in Figure 5.8. Subplot A represents the distribution that was used to generate the simulated data. Subplots B to D represent the smallest and largest value(s) of the W_{ig} distribution parameters seen in the simulations. The TVNIG, TVVG and TVSAL models are learning parameterizations that create densities resembling the inverse gamma density in subplot A. The shapes of the GIG distributions in subplot D vary considerably, often looking nothing like the distribution in subplot A.



(a) GIG: $\omega = 0.34$, $\lambda = 4.99$ (TVGH).

(b) GIG: h$\omega = 4.92$, $\lambda = 17.95$ (TVGH).

Figure 5.9: Quantile-quantile plot of $\mathbb{E}(W_i \mid \boldsymbol{\mathfrak{X}}_i, \hat{\boldsymbol{\vartheta}})$ from the W_{ig} distributions (skewed).

This dissimilarity of the W_{ig} distributions between the models are reflected in the $\mathbb{E}(W_i \mid \boldsymbol{\mathfrak{X}}_i, \hat{\boldsymbol{\vartheta}})$ values learned from the data. Starting with two data sets from the skewed simulated data, where $n^* = 512$ and $N = 50$, we visualize the distribution of the $\mathbb{E}(W_i \mid \boldsymbol{\mathfrak{X}}_i, \hat{\boldsymbol{\vartheta}})$ values and the resulting model performance in Figures 5.9a and 5.9b. The left hand panel of Figure 5.9a uses a qq-plot to visualize the distribution of the TVST $\mathbb{E}(W_i \mid \boldsymbol{\mathfrak{X}}_i, \hat{\boldsymbol{\vartheta}})$ values verses the distribution of the $\mathbb{E}(W_i \mid \boldsymbol{\mathfrak{X}}_i, \hat{\boldsymbol{\vartheta}})$ values from the other four tensor-variate distributions. Recall that the data was generated from a TVST distribution with $\nu = 4$. The right hand panel includes the model BIC values. The distribution of the W_{ig}'s from the TVGH model resembles the blue curve in Figure 5.8 subplot D. This results in values of $\mathbb{E}(W_i \mid \mathfrak{X}_i, \hat{\boldsymbol{\vartheta}})$ that are divergent from the TVST values and ultimately, in very poor relative model performance, as measured by BIC. Contrast this with the results in figure 5.9b where the distribution of W_{ig}'s from the TVGH model is more akin to the distribution used to generate the simulated data. In this instance, the qq-plot indicates the distribution of the $\mathbb{E}(W_i \mid \mathfrak{X}_i, \hat{\boldsymbol{\vartheta}})$ values between the TVST model and the other models is similar and the model performance between the 5 models is very comparable.

5.5 Image analysis

As our data analysis example, we chose to analyse RGB images, defined as order-3 MDAs. The images come from the CIFAR-100 data set(Krizhevsky and Hinton, 2009). We chose images of maple trees that had green or yellow leaves and came from the following CIFAR-100 class hierarchy: superclass trees → class maple. These MDAs had an $n^* = 3072$, making them comparable to the $n^* = 3375$ results in our simulation. Figure 5.10 is an example of one of the images in our sample of 207 MDAs.

We used the BIC to select the best model for this data. We can see from Figure 5.11, the skewed models all outperformed the flip-flop algorithm. The TVNIG model had the best performance, with a BIC $= 2.978 \times 10^6$.

Figure 5.12a is the image that results from the estimated $\mathbb{E}[\mathscr{X}]$ MDA. The sky, tree trunk and branches are clearly visible. Contrast this to the heatmap in Figure 5.12b that visualizes each slice of the estimated $\mathbb{E}[\mathscr{X}]$ MDA. The combinations of R, G

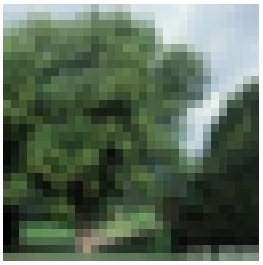


Figure 5.10: An image of a maple tree from the CIFAR-100 data set.

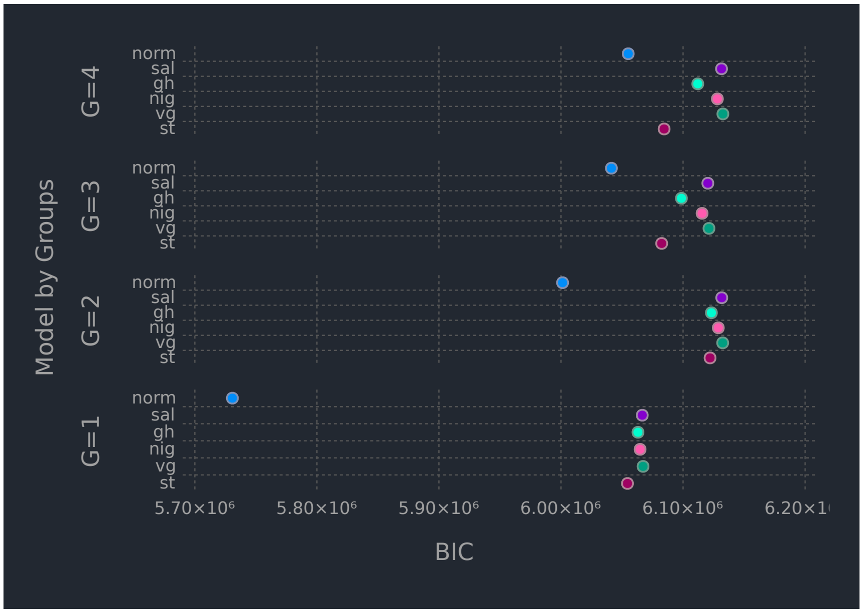
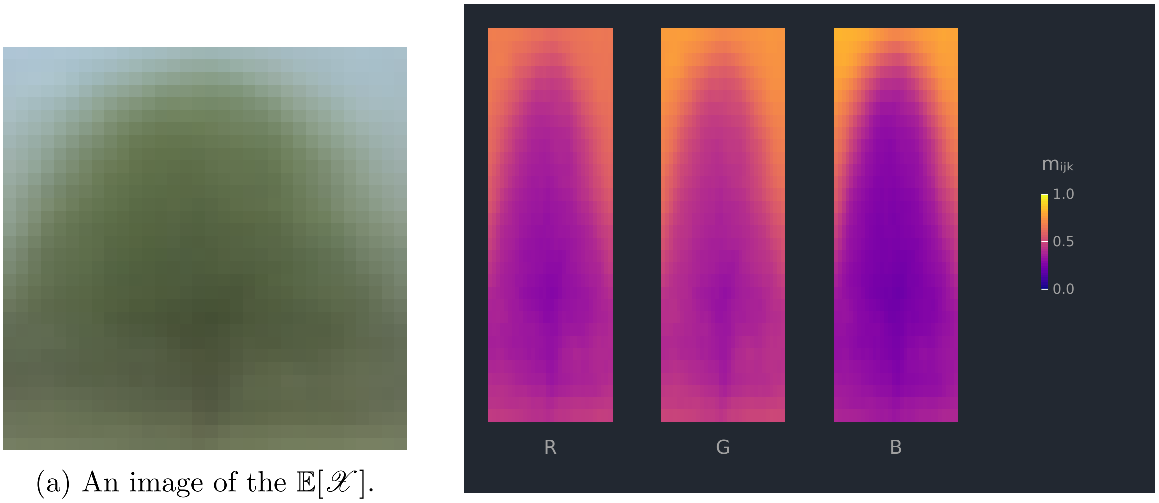


Figure 5.11: BIC results from the image analysis.

and B values result in the colors displayed in Figure 5.12a. The sky and tree trunk remain clearly distinguishable.

The location MDA, $\mathfrak{M}$, is visualized in Figure 5.13a. Even without the addition of the scaled skewness MDA $\mathfrak{A}$ from (5.9), the visualization is clearly a green tree with a brown trunk and blue sky. The estimated skewness MDA, $\mathfrak{A}$, is visualized in Figure 5.13b. It is clear that each colour slice has different skewness patterns. The

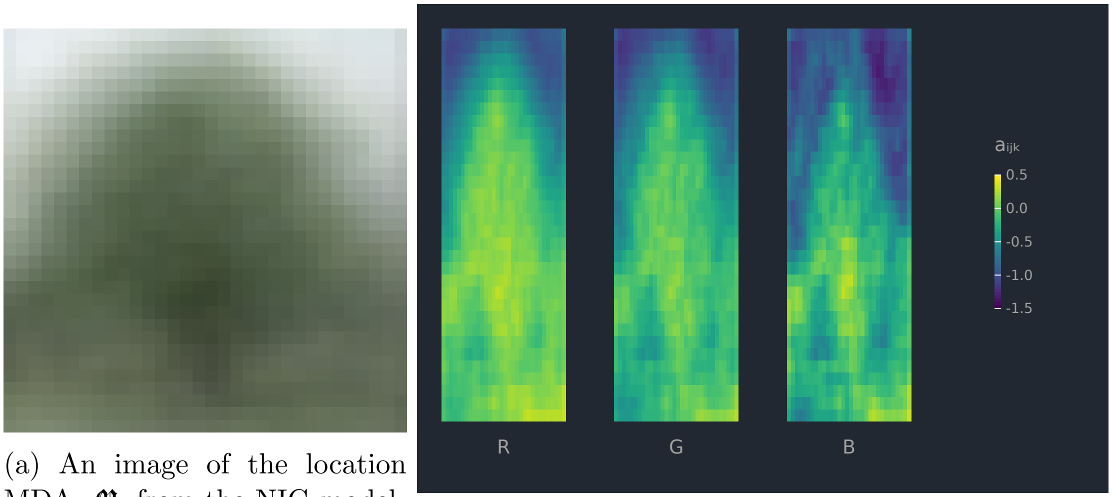


(a) An image of the $\mathbb{E}[\mathscr{X}]$.

(b) Slices of the $\mathbb{E}[\mathscr{X}]$ MDA.

Figure 5.12: Visualizations of the $\mathbb{E}[\mathscr{X}]$ MDA from the TVNIG model.

sky tends to have the lowest skewness, a pattern accentuated in the "B" slice. The "R" slice has the most positive skewness, concentrated in the trunk and body of the trees.



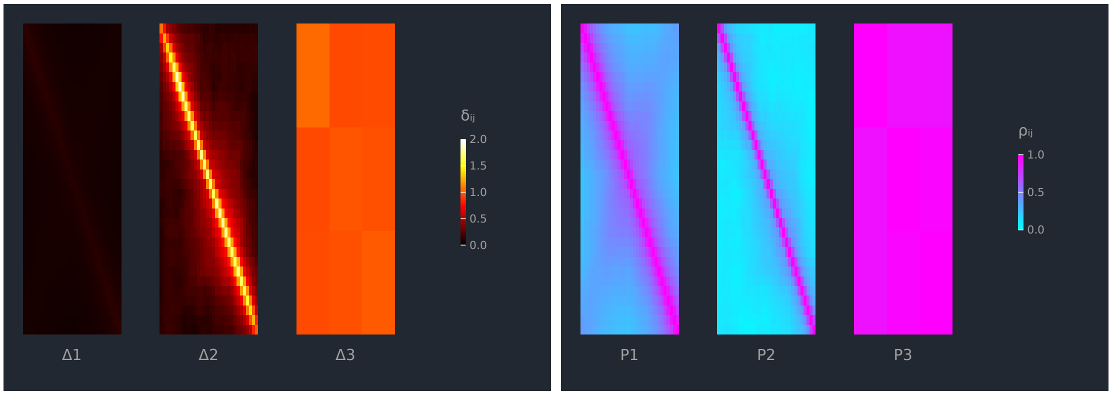
(a) An image of the location MDA, $\mathfrak{M}$, from the NIG model.

(b) Slices of the $\mathfrak{A}$ MDA from the NIG model

Figure 5.13: Visualizations of the $\mathfrak{M}$ and $\mathfrak{A}$ MDAs.

The estimated variability in each of the three modes is visualized in Figure 5.14a.

Each scale matrix, $\mathbf{\Delta}_d$, is visualized as a heatmap. The rows ($\mathbf{\Delta}_1$), have little variation. The columns ($\mathbf{\Delta}_2$) exhibit a pattern of covariation consistent with images, one that decreases as the distance between pixels increases. All three slices ($\mathbf{\Delta}_3$) have a moderate level of variation. Figure 5.14b displays the three scale matrices $\{\mathbf{\Delta}_d\}_{d=1}^{3}$ as correlation matrices $\{\mathbf{P}_d\}_{d=1}^{3}$. Despite having little variation, the row dimension of the image MDAs have a clear correlation pattern. The pattern indicates that entries close together are positively correlated, as one would expect with image data.



(a) Scale matrices, Δ_d

(b) Correlation matrices, $\mathbf{P}_d$

Figure 5.14: Visualizations of the variability in each dimension of the MDAs.

Chapter 6

Finite Mixtures of Tensor-Variate Skewed Distributions

6.1 Overview

Recall from Section 2.2, in the mixture model framework, the random variable $\mathbf{X}$ originates from a population with G separate sub-populations. Each subgroup has the same density function, with different parameter values. Given a sample of N i.i.d. random D-dimensional arrays, $\boldsymbol{\mathfrak{X}}_1, \ldots, \boldsymbol{\mathfrak{X}}_N$, the observed-data likelihood is

$$\mathcal{L}_{\text{O}} = \prod_{i=1}^{N} \sum_{g=1}^{G} \pi_g f_{\text{TVD}} \left(\boldsymbol{\mathfrak{X}}_i | \boldsymbol{\mathfrak{M}}_g, \boldsymbol{\mathfrak{A}}_g, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}, \boldsymbol{\theta}_g \right),$$

where $\boldsymbol{\theta}_g$ are the parameters associated with the distribution of $W_{ig} \in \mathbb{R}^{+}$.

Similar to Chapter 3, we proceed as if the data are incomplete by treating the group allocation z_{ig} and W_{ig} as latent variables. This results in a general form for the complete likelihood that can be generalized to all 5 tensor-variate distributions

$$\ell_C(\boldsymbol{\vartheta}) = \ell_{C1} + C_1 + \ell_{C2} + C_2 + \ell_{C3},$$

where each term is defined as follows:

$$\begin{aligned}
\ell_{C1} &= \sum_{g=1}^{G} n_g \log \pi_g \\
C_1 &= -\frac{Nn^*}{2} \log(2\pi) \\
\ell_{C2} &= \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} \log h(w_{ig}|\boldsymbol{\theta}_g) \\
C_2 &= \text{Constants related to } \log h(w_{ig}|\boldsymbol{\theta}_g) \\
\ell_{C3} &= -\frac{n^*}{2} \sum_{g=1}^{G} n_g \sum_{d=1}^{D} \frac{1}{n_d} \log |\boldsymbol{\Delta}_{gd}| + \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} \text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}(\boldsymbol{\mathfrak{A}}_g) \\
&\quad + \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} \text{vec}(\boldsymbol{\mathfrak{A}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
&\quad - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \frac{z_{ig}}{w_{ig}} \text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
&\quad - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} w_{ig} \text{vec}(\boldsymbol{\mathfrak{A}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}(\boldsymbol{\mathfrak{A}}_g)
\end{aligned}$$

The mathematical details leading up these expressions are listed in Appendix E.2.

6.2 Parameter Estimation

Parameter estimation is based on an ECM algorithm, as described below.

1) Initialization: Initialize the parameters $\mathbf{M}_{(1)g}, \mathbf{A}_{(1)g}, \boldsymbol{\Delta}_{gd}, \boldsymbol{\theta}_g$ and $\hat{z}_{ig}$. Set $t = 0$.

2) E Step: Update $\hat{z}_{ig}, a_{ig}, b_{ig}, c_{ig}$, where

$$a_{ig} = \mathbb{E}(W_{ig} \mid \mathbf{X}_{(1)i}, \hat{\boldsymbol{\theta}}_g), \quad b_{ig} = \mathbb{E}\left(\frac{1}{W_{ig}} \;\middle|\; \mathbf{X}_{(1)i}, \hat{\boldsymbol{\theta}}_g\right), \quad c_{ig} = \mathbb{E}(\log W_{ig} \mid \mathbf{X}_{(1)i}, \hat{\boldsymbol{\theta}}_g),$$

$$\hat{z}_{ig} = \frac{\hat{\pi}_g f_{\text{TVD}}\left(\mathbf{X}_{(1)i}|\hat{\boldsymbol{\theta}}_g\right)}{\sum_{h=1}^{G} \hat{\pi}_h f_{\text{TVD}}\left(\mathbf{X}_{(1)i}|\hat{\boldsymbol{\theta}}_g\right)}$$

All the expectations are conditional on current parameter estimates; however, we do not use iteration-specific notation. Although a_{ig}, b_{ig}, c_{ig} are dependent on the distribution of W_{ig}, it can be shown that

$$W_{ig}^{\text{ST}} \mid \mathbf{X}_{(1)i} \sim \text{GIG}\left(\rho\left(\mathbf{A}_{(1)g}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\right), \delta\left(\mathbf{X}_{(1)i}; \mathbf{M}_{(1)g}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\right) + \nu, -(\nu + n^*)/2\right),$$
$$W_{ig}^{\text{GH}} \mid \mathbf{X}_{(1)i} \sim \text{GIG}\left(\rho\left(\mathbf{A}_{(1)g}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\right) + \omega, \delta\left(\mathbf{X}_{(1)i}; \mathbf{M}_{(1)g}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\right) + \omega, \lambda - n^*/2\right),$$
$$W_{ig}^{\text{VG}} \mid \mathbf{X}_{(1)i} \sim \text{GIG}\left(\rho\left(\mathbf{A}_{(1)g}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\right) + 2\gamma, \delta\left(\mathbf{X}_{(1)i}; \mathbf{M}_{(1)g}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\right), \gamma - n^*/2\right),$$
$$W_{ig}^{\text{NIG}} \mid \mathbf{X}_{(1)i}, \sim \text{GIG}\left(\rho\left(\mathbf{A}_{(1)g}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\right) + \kappa^2, \delta\left(\mathbf{X}_{(1)i}; \mathbf{M}_{(1)g}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\right) + 1, -(1 + n^*)/2\right).$$

These expectations can be calculated using the results given in (2.9)–(2.11).

3) First CM Step: Update the parameters $\pi_g, \mathbf{M}_{(1)g}, \mathbf{A}_{(1)g}$ and $\boldsymbol{\theta}_g$.

$$\hat{\pi}_g = \frac{n_g}{N}$$

$$\hat{\mathbf{M}}_{(1)g} = \frac{\sum_{i=1}^{N} \hat{z}_{ig} \mathbf{X}_{(1)i} \left\{ \bar{a} b_{ig} - 1 \right\}}{\sum_{i=1}^{N} \hat{z}_{ig} \bar{a} b_{ig} - \hat{n}_g} \tag{6.1}$$

$$\hat{\mathbf{A}}_{(1)g} = \frac{\sum_{i=1}^{N} \hat{z}_{ig} \mathbf{X}_{(1)i} \left\{ \bar{b} - b_{ig} \right\}}{\sum_{i=1}^{N} \hat{z}_{ig} a_{ig} \bar{b} - \hat{n}_g} \tag{6.2}$$

The mathematical details related to these updates are given in Appendix E.4. The updates for the additional parameters, $\boldsymbol{\theta}_g$, are detailed in Section 6.3.

4) Additional CM Steps: Update $\boldsymbol{\Delta}_{gd}$

$$\begin{aligned} \hat{\boldsymbol{\Delta}}_{g1} = \frac{n_1}{n^* \hat{n}_g} \Bigg[\sum_{i=1}^{N} \hat{z}_{ig} \Bigg\{ & b_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{X}^{\top}_{(1)gij} \boldsymbol{\Delta}^{-1}_{g2} \mathbf{X}_{(1)gij} + a_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{A}^{\top}_{(1)gj} \boldsymbol{\Delta}^{-1}_{g2} \mathbf{A}_{(1)gj} \\ & - \sum_{j=1}^{n^*_{3:D}} \mathbf{A}^{\top}_{(1)gj} \boldsymbol{\Delta}^{-1}_{g2} \mathbf{X}_{(1)gij} - \sum_{j=1}^{n^*_{3:D}} \mathbf{X}^{\top}_{(1)gij} \boldsymbol{\Delta}^{-1}_{g2} \mathbf{A}_{(1)gj} \Bigg\} \Bigg] \end{aligned} \tag{6.3}$$

$$\begin{aligned} \hat{\boldsymbol{\Delta}}_{g2} = \frac{n_2}{n^* \hat{n}_g} \Bigg[\sum_{i=1}^{N} \hat{z}_{ig} \Bigg\{ & b_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{X}_{(1)gij} \boldsymbol{\Delta}^{-1}_{g1} \mathbf{X}^{\top}_{(1)gij} + a_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{A}_{(1)gj} \boldsymbol{\Delta}^{-1}_{g1} \mathbf{A}^{\top}_{(1)gj} \\ & - \sum_{j=1}^{n^*_{3:D}} \mathbf{X}_{(1)gij} \boldsymbol{\Delta}^{-1}_{g1} \mathbf{A}^{\top}_{(1)gj} - \sum_{j=1}^{n^*_{3:D}} \mathbf{A}_{(1)gj} \boldsymbol{\Delta}^{-1}_{g1} \mathbf{X}^{\top}_{(1)gij} \Bigg\} \Bigg] \end{aligned} \tag{6.4}$$

$$\hat{\mathbf{\Delta}}_{gl} = \frac{n_l}{n^* \hat{n}_g} \left[\sum_{i=1}^{N} \hat{z}_{ig} \left\{ b_{ig} \sum_{j=1}^{n^*_{2:D/l}} \mathbf{X}^{l2}_{(1)gij} \mathbf{\Delta}^{-1}_{g1} \left(\mathbf{X}^{l2}_{(1)gij} \right)^\top + a_{ig} \sum_{j=1}^{n^*_{2:D/l}} \mathbf{A}^{l2}_{(1)gj} \mathbf{\Delta}^{-1}_{g1} \left(\mathbf{A}^{l2}_{(1)gj} \right)^\top \right.\right.$$
$$\left.\left. - \sum_{j=1}^{n^*_{2:D/l}} \mathbf{X}^{l2}_{(1)gij} \mathbf{\Delta}^{-1}_{g1} \left(\mathbf{A}^{l2}_{(1)gj} \right)^\top - \sum_{j=1}^{n^*_{2:D/l}} \mathbf{A}^{l2}_{(1)gj} \mathbf{\Delta}^{-1}_{g1} \left(\mathbf{X}^{l2}_{(1)gij} \right)^\top \right\} \right] \tag{6.5}$$

where $n^*_{3:D} = \prod_{d=3}^{D} n_d$,

$$\mathbf{X}_{(1)gij} = \left(\mathbf{I}_{n_2} \otimes \mathbf{e}_j^\top \bigotimes_{d=3}^{D} \mathbf{\Delta}_{gd}^{-\frac{\top}{2}} \right) \breve{\mathbf{X}}_{(1)gi},$$

$\breve{\mathbf{X}}_{(1)gi} = \mathbf{X}_{(1)i} - \mathbf{M}_{(1)g}$,

$$\mathbf{A}_{(1)gj} = \left(\mathbf{I}_{n_2} \otimes \mathbf{e}_j^\top \bigotimes_{d=3}^{D} \mathbf{\Delta}_{gd}^{-\frac{\top}{2}} \right) \mathbf{A}_{(1)g},$$

$\mathbf{\Delta}_d^{-\frac{1}{2}}$ is the Cholesky decomposition of $\mathbf{\Delta}_d$ and $\mathbf{e}_j$ is a Kronecker product of unit basis vectors. The superscript $l2$ indicates the exchanges of the second and l^{th} elements in the sequence, where $3 \leq l \leq D$.

The derivations for these updates are given in appendix E.5. The Q functions, $\mathbb{Q}(\boldsymbol{\vartheta})$, used for these derivations are given in equations E.31 and E.34. They take advantage of the vector to trace conversion we outlined in appendices A.2 and A.3 .

5) Check Convergence: If not converged repeat steps 2–4 until convergence.

6.3 Updates for the Additional Parameters

TVST

In the case of the TVST distribution, the degrees of freedom ν_g needs to be updated. The update for the degrees of freedom cannot be obtained in closed form. Instead we solve (6.6) for ν_g to obtain the update.

$$\log\left(\frac{\nu_g}{2}\right) + 1 - \varphi\left(\frac{\nu_g}{2}\right) - \frac{1}{N}\sum_{i=1}^{N}(b_{ig} + c_{ig}) = 0, \tag{6.6}$$

where $\varphi(\cdot)$ is the digamma function.

TVGH

In the case of the TVGH distribution, we update λ_g and ω_g. In this case,

$$\ell_{C2g} = N\log(K_{\lambda_g}(\omega_g)) - \lambda_g\sum_{i=1}^{N} c_{ig} - \frac{1}{2}\omega_g\sum_{i=1}^{N}(a_{ig} + b_{ig}) \tag{6.7}$$

The updates for λ_g and ω_g cannot be obtained in closed form. Numerical methods for these updates are discussed in Browne and McNicholas (2015) and because the portion of the likelihood function that includes these parameters is the same as in the multivariate case, the updates these authors describe can be used directly here.

The updates for λ_g and ω_g rely on the log convexity of $K_{\lambda_g}(\omega_g)$, Baricz (2010), in both λ_g and ω_g and maximizing (6.7) via conditional maximization. The resulting updates are

$$\hat{\lambda}_g^{(t+1)} = \bar{c}_g \hat{\lambda}_g^{(t)} \left[\frac{\partial}{\partial s} \log(K_s(\hat{\omega}_g^{(t)})) \Big|_{s=\hat{\lambda}_g^{(t)}} \right]^{-1} \tag{6.8}$$

$$\hat{\omega}_g^{(t+1)} = \hat{\omega}_g^{(t)} - \left[\frac{\partial}{\partial s} q(\hat{\lambda}_g^{(t+1)}, s) \Big|_{s=\hat{\omega}_g^{(t)}} \right] \left[\frac{\partial^2}{\partial s^2} q(\hat{\lambda}_g^{(t+1)}, s) \Big|_{s=\hat{\omega}_g^{(t)}} \right]^{-1} \tag{6.9}$$

where the derivative in (6.8) is calculated numerically and $\bar{c}_g = \sum_{i=1}^{N} c_{ig}/N$. The partials in (6.9) are described in Browne and McNicholas (2015), and can be written as

$$\frac{\partial}{\partial \omega_g} q(\lambda_g, \omega_g) = \frac{1}{2}[R_{\lambda_g}(\omega_g) + R_{-\lambda_g}(\omega_g) - (\bar{a}_g + \bar{b}_g)],$$

and

$$\frac{\partial^2}{\partial \omega_g^2} q(\lambda_g, \omega_g) = \frac{1}{2} \left[R_{\lambda_g}(\omega_g)^2 - \frac{1+2\lambda_g}{\omega_g} R_{\lambda_g}(\omega_g) - 1 + R_{-\lambda_g}(\omega_g)^2 - \frac{1-2\lambda_g}{\omega_g} R_{-\lambda_g}(\omega_g) - 1 \right],$$

where $R_{\lambda_g}(\omega_g) = K_{\lambda_g+1}(\omega_g)/K_{\lambda_g}(\omega_g)$.

TVVG

In the case of the TVVG, the update for γ_g is needed. This update, like the TVST and TVGH, cannot be obtained in closed form. Instead, the update, $\gamma_g^{(t+1)}$, is obtained by solving (6.10) for γ_g:

$$\log(\gamma_g) + 1 - \varphi(\gamma_g) + \bar{c}_g - \bar{a}_g = 0. \tag{6.10}$$

TVNIG

Finally, in the TVNIG case, the update for κ_g can be written in closed form as

$$\hat{\kappa}_g = \frac{N}{\sum_{i=1}^{N} a_{ig}}.$$

6.4 Software

The ECM algorithm is implemented in version 1.5.3 of the Julia programming language (`https://julialang.org/`; Bezanson *et al.*, 2017). Bessel function values are calculated using 100 digit numbers, made possible by version 1.2.4 of the `ArbNumerics.jl` library. We use numerical differentiation to find

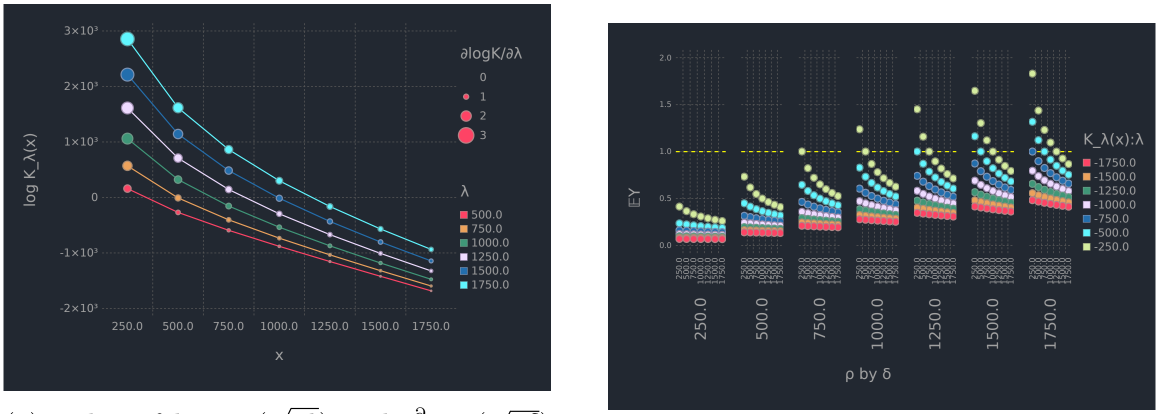
$$\frac{\partial}{\partial \lambda} K_\lambda(\sqrt{\rho\delta})$$

in (2.11). To do the numerical differentiation, the complex step method, (Squire and Trapp, 1998) is implemented which uses the following approximation for the derivative of $f(x)$

$$f'(x) \approx \operatorname{Im}\left\{\frac{f(x+ih)}{h}\right\},$$

where $i = \sqrt{-1}$. This approximation has $O(h^2)$ error and h is not restricted by rounding errors (e.g., $h = 10^{-50}$). The only restriction is that the algorithm for $f(x)$ cannot use complex arithmetic. Similar to Chapter 3, singular $\mathbf{\Delta}_{gd}$ values were numerically regularized by adding a small positive quantity to the diagonal elements of the matrices (Williams and Rasmussen, 2006).

Given the complexity of implementing all five of the mixtures, a lot of effort was



(a) Value of $\log K_\lambda(\sqrt{ab})$ and $\frac{\partial}{\partial\lambda}K_\lambda(\sqrt{\rho\delta})$ as λ and $x = \sqrt{\rho\delta}$ change

(b) Values of $\mathbb{E}(W_{ig})$ as λ, ρ and δ change

Figure 6.1: An overview of the variation in $\log K_\lambda(\sqrt{ab})$, $\frac{\partial}{\partial\lambda}K_\lambda(\sqrt{\rho\delta})$ and $\mathbb{E}(W_{ig})$ values we see in our models.

spent on software engineering, to ensuring the code quality was high. Tables such as 6.4 proved useful for implementing the various likelihood functions in a generic and reusable way.

Table 6.1: Terms for the Log-likelihood `Julia` Function(s)

Term	ST	GH	VG	NIG
1	$N\log\{2\} + \frac{N\nu}{2}\log\left\{\frac{\nu}{2}\right\}$	0	$N\log\{2\} + N\gamma\log\{\gamma\}$	$N\log\{2\} + N\kappa$
2	$\sum_{i=1}^{N}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}})^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\text{vec}(\boldsymbol{\mathfrak{A}})$			
3	$-\frac{Nn^*}{2}\sum_{d=1}^{D}\frac{1}{n_d}\log\{\lvert\boldsymbol{\Delta}_d\rvert\}$			
4	$-\frac{Nn^*}{2}\log\{2\pi\} - N\log\left\{\Gamma(\frac{\nu}{2})\right\}$	$-\frac{Nn^*}{2}\log\{2\pi\} - N\log\{K_\lambda(\omega)\}$	$-\frac{Nn^*}{2}\log\{2\pi\} - N\log\{\Gamma(\gamma)\}$	$-N\left(\frac{n^*+1}{2}\right)\log\{2\pi\}$
5	$-\left(\frac{\nu+n^*}{4}\right)\sum_{i=1}^{N}\log\{\delta_i(\cdot)+\nu\}$	$\left(\frac{\lambda-\frac{n^*}{2}}{2}\right)\sum_{i=1}^{N}\log\{\delta_i(\cdot)+\omega\}$	$\left(\frac{\gamma-\frac{n^*}{2}}{2}\right)\sum_{i=1}^{N}\log\{\delta_i(\cdot)\}$	$-\left(\frac{n^*+1}{4}\right)\sum_{i=1}^{N}\log\{\delta_i\ (\cdot)+1\}$
6	$N\left(\frac{\nu+n^*}{4}\right)\log\{\rho(\cdot)\}$	$-N\left(\frac{\lambda-\frac{n^*}{2}}{2}\right)\log\{\rho(\cdot)+\omega\}$	$-N\left(\frac{\gamma-\frac{n^*}{2}}{2}\right)\log\{\rho(\cdot)+2\gamma\}$	$N\left(\frac{n^*+1}{4}\right)\log\left\{\rho(\cdot)+\kappa^2\right\}$
7	$\sum_{i=1}^{N}\log K_{-\frac{\nu+n^*}{2}}\left(\sqrt{\rho(\cdot)\left(\delta_i(\cdot)+\nu\right)}\right)$	$\sum_{i=1}^{N}\log K_{\lambda-\frac{n^*}{2}}\left(\sqrt{\left(\rho(\cdot)+\omega\right)\left(\delta_i(\cdot)+\omega\right)}\right)$	$\sum_{i=1}^{N}\log K_{\gamma-\frac{n^*}{2}}\left(\sqrt{\left(\rho(\cdot)+2\gamma\right)\delta_i(\cdot)}\right)$	$\sum_{i=1}^{N}\log K_{-\frac{n^*+1}{2}}\left(\sqrt{\left(\rho(\cdot)+\kappa^2\right)\left(\delta_i(\cdot)+1\right)}\right)$

To avoid numerical instabilities, where possible, the computations were done on the log scale. For example, to calculate $\mathbb{E}\left(1/Y\right)$, we use the log identity

$$\log(a-c) = \log(a) + \log\left(1 - \frac{c}{a}\right)$$

to convert (2.10) to the log scale as follows

$$\log \mathbb{E}\left(1/Y\right) = \left[\frac{1}{2}\log(a) - \frac{1}{2}\log(b) + \log\left(K_{\lambda+1}(\sqrt{ab})\right) - \log\left(K_{\lambda}(\sqrt{ab})\right)\right] \\ - \left[\log(2) + \log(\lambda) - \log(b)\right]$$

When

$$\sqrt{\frac{a}{b}}\frac{K_{\lambda+1}(\sqrt{ab})}{K_{\lambda}(\sqrt{ab})} > \frac{2\lambda}{b},$$

$$\log \mathbb{E}\left(1/Y\right) = \frac{1}{2}\log(a) - \frac{1}{2}\log(b) + \log\left(K_{\lambda+1}(\sqrt{ab})\right) - \log\left(K_{\lambda}(\sqrt{ab})\right) + \\ \log\left[1 - \frac{2\lambda}{\sqrt{ab}} \times \exp\left\{\log K_{\lambda}(\sqrt{ab}) - \log K_{\lambda+1}(\sqrt{ab})\right\}\right]$$

and, when

$$\frac{2\lambda}{b} > \sqrt{\frac{a}{b}}\frac{K_{\lambda+1}(\sqrt{ab})}{K_{\lambda}(\sqrt{ab})},$$

$$\log \mathbb{E}\left(1/Y\right) = \log(2) + \log(\lambda) - \log(b) + \log\left[1 - \frac{\sqrt{ab}}{2\lambda} \times \exp\left\{\log K_{\lambda+1}(\sqrt{ab}) - \log K_{\lambda}(\sqrt{ab})\right\}\right].$$

It should be noted that

$$\log\left(1 - \frac{c}{a}\right)$$

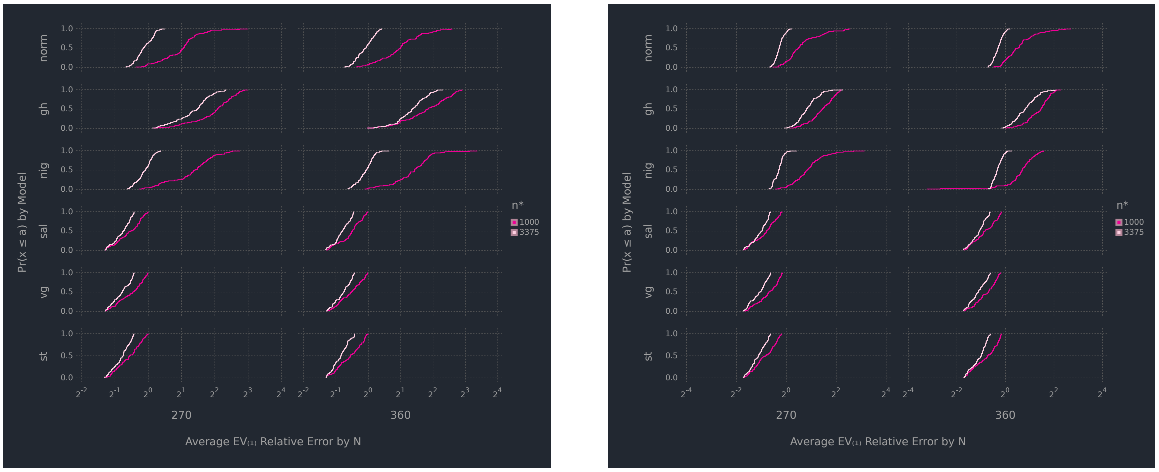
should be implemented with the `log1p(x)` function, where $x = -c/a$.

6.5 Simulation

Given the large increase in free parameters that results from taking $G = 1$ in Chapter 5 to $G \geq 2$ in the mixture model, we evaluated the effect of the signal-to-noise ratio (s2n $\in \{0.5, 2.0\}$) on the models performance across the following values of $N \in \{270, 360\}$ and $n^* \in \{1000, 3375\}$. Similar to the simulation in Chapter 3, we set $G = 3$ and had equal groups sizes, $n_g \in \{90, 120\}$. We created 100 replicates per combination of signal-to-noise ratio, N and n^*. Similar to the simulation in Section 5.4, we generate the data from a TVST distribution with $\nu = 4$. Unlike the results presented in Chapter 3, we use order-3 MDAs and averaged the results across the groups to simplify the graphical displays. The normal model is the finite mixture of MLNDs described in Chapter 3.

Figures 6.2a and 6.2b plot the distribution of the average relative errors for the mode 1 matricization of $\mathbb{E}\left[\mathfrak{X}\right]$, for each signal-to-noise level. When the data is noisy, smaller MDA sizes have larger errors. These are particularly acute for the normal model, the TVGH and TVNIG models. This can be explained by the discrepancy in the underlying distributions of W_{ig} vs the Inv-Gamma$\left(\nu/2, \nu/2\right)$ distribution used to generate the data, as visualized in Figure 5.8. When there is more signal than noise, all the three aforementioned models improve their performance. The TVST, TVVG and TVSAL models are not impacted by any of the changes in the three factors we varied.

When estimating the average relative error in $\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}$, the normal fairs relatively poorly compared to its skewed counter parts. This is not surprising given it has



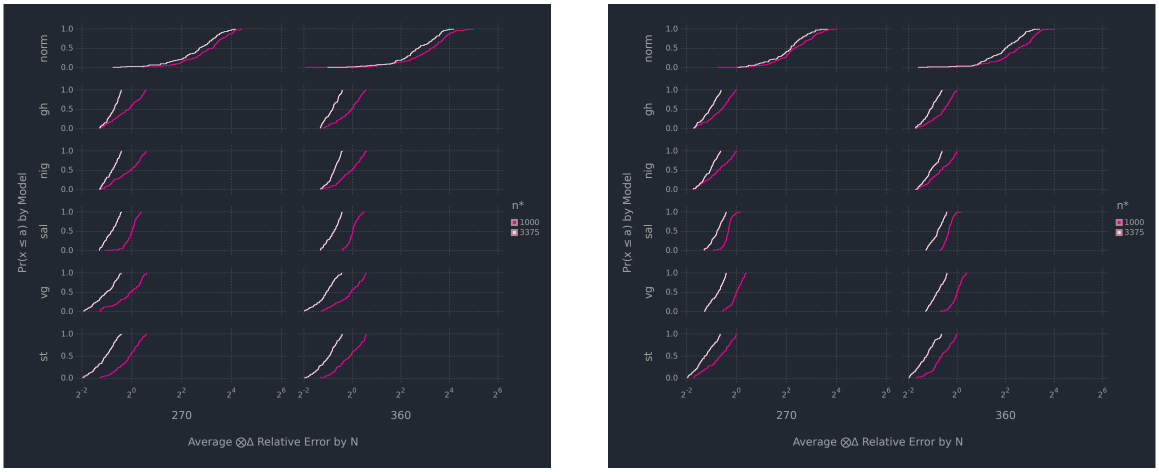
(a) Empirical distribution plots of the average relative error, signal-to-noise = 0.5.

(b) Empirical distribution plots of the average relative error, signal-to-noise = 2.0.

Figure 6.2: Simulation results for the mode 1 matricization of $\mathbb{E}\left[\boldsymbol{\mathfrak{X}}\right]$.

no way of modeling the skewness as captured by $\boldsymbol{\mathfrak{A}}_g$ and the underlying distribution of W_{ig}. MDA size n^* is more impact-full on the error estimates than the sample size N. The dynamics between these two sizes should be further explored, to give users of these models usage guidelines.

The group labels produced by the finite mixture model were compared to the simulated group labels via the ARI. The average ARI for each combination of N and n^* was at least 0.93 and 0.84 for the signal-to-noise ratios of 2 and 0.5. The average ARI values are both close to 1, indicating that in both high and low noise scenarios, the models can generate group labels that are in agreement with the true groupings in the data.



(a) Empirical distribution plots of the average relative error, signal-to-noise = 0.5.

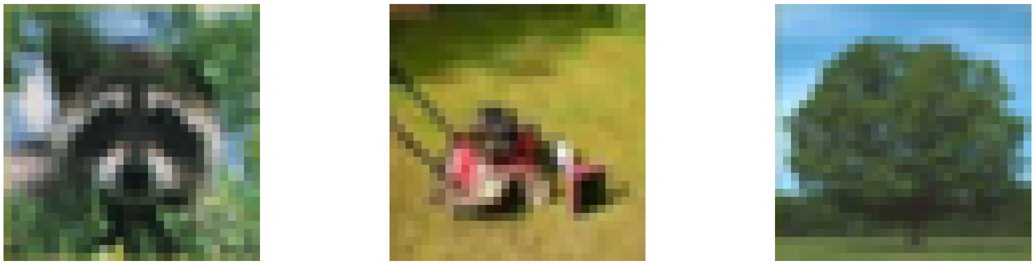
(b) Empirical distribution plots of the average relative error, signal-to-noise = 2.0.

Figure 6.3: Simulation results for $\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}$.

6.6 Image analysis

As our data analysis example, we chose to analyse RGB images, defined as order-3 MDAs. Analogous to Chapter 5, the images come from the CIFAR-100 data set and consist of raccoons, maple trees and lawnmowers. This sample of 420 MDAs has three equal sized groups of 140 images (e.g., $n_g = 140$) were each image has an n^* equal to 3072. The raw data is visualized in Figure 6.4. This is a difficult clustering problem. The images share many of the same colors and the colors occur in similar locations within each image.

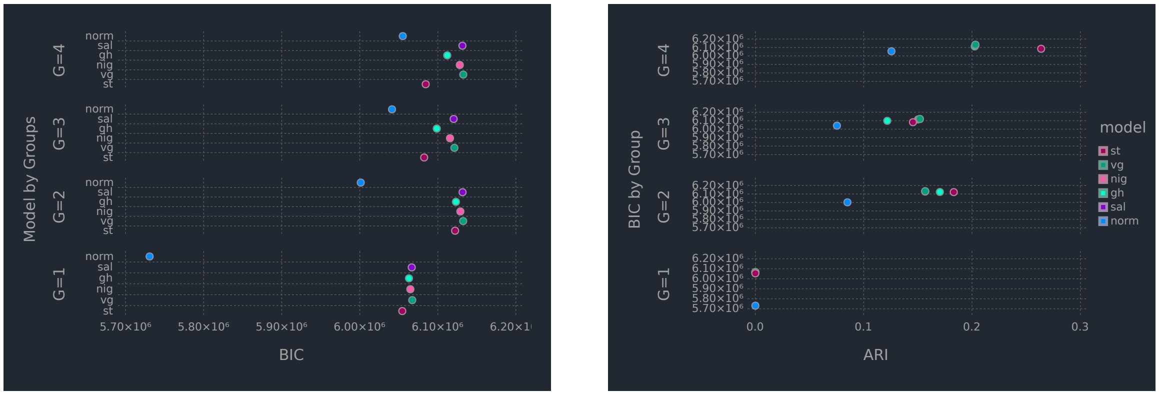
Models with $G \in \{1, 2, 3, 4\}$ were compared using the BIC in Figure 6.5a. Two and four group solutions are preferred. The model with the largest BIC was the TVVG with $G = 4$. In Figure 6.5b, the BIC values are plotted against the ARI values capturing the pairwise agreement between the cluster and image labels. Here the $G = 4$ models are clearly preferred when judged by their ARI values. The TVST



(a) Raccoon (b) Lawn Mower (c) Maple Tree

Figure 6.4: Samples of the raw images being clustered.

model with $G = 4$ has the best combination of BIC and ARI values and will be used in the following analysis.



(a) BIC (b) BIC vs ARI

Figure 6.5: Model selection results from the image analysis.

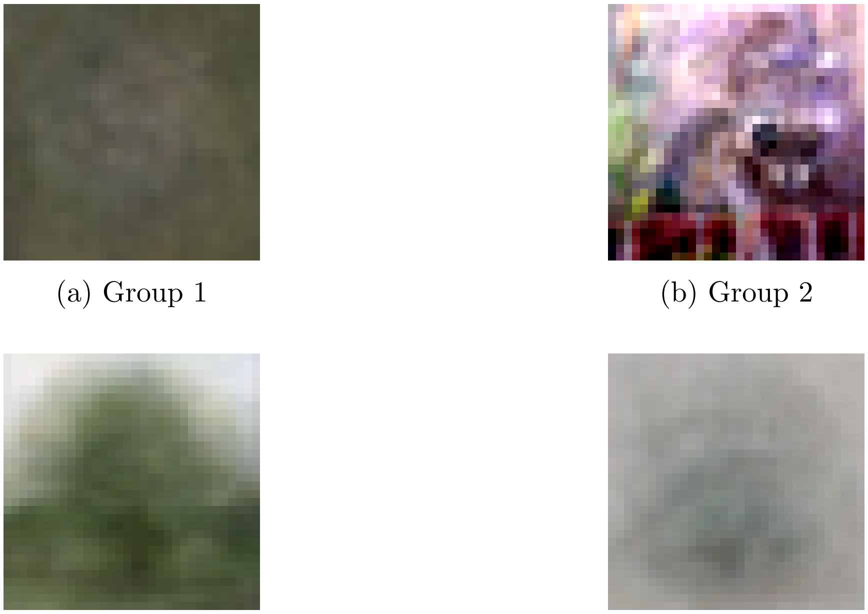
The cross-tabulations of the clusters versus the image labels for the TVST model with $G = 4$ are listed in Table 6.2. Clearly clusters 2 and 3 represent raccoons and maple trees, respectively. Clusters 1 and 4 represent mixed groupings, with cluster 4 accounting for the majority of the lawnmowers. The ARI value for this table is 0.264.

Each cluster group's location tensor $\boldsymbol{\mathfrak{M}}_g$ is visualized in Figure 6.6. Cluster groups 1 and 4 do not resemble any one of the three image types. This is expected after examining Table 6.2. Cluster group 2 clearly captures images of raccoons that contain

Table 6.2: Cross-tabulation of the clusters (based on MAP estimates) versus the image labels.

Cluster	**Lawn Mower**	**Maple Tree**	**Raccoon**
1	24	12	45
2	0	0	7
3	1	101	6
4	115	27	82

red hues.



(a) Group 1

(b) Group 2

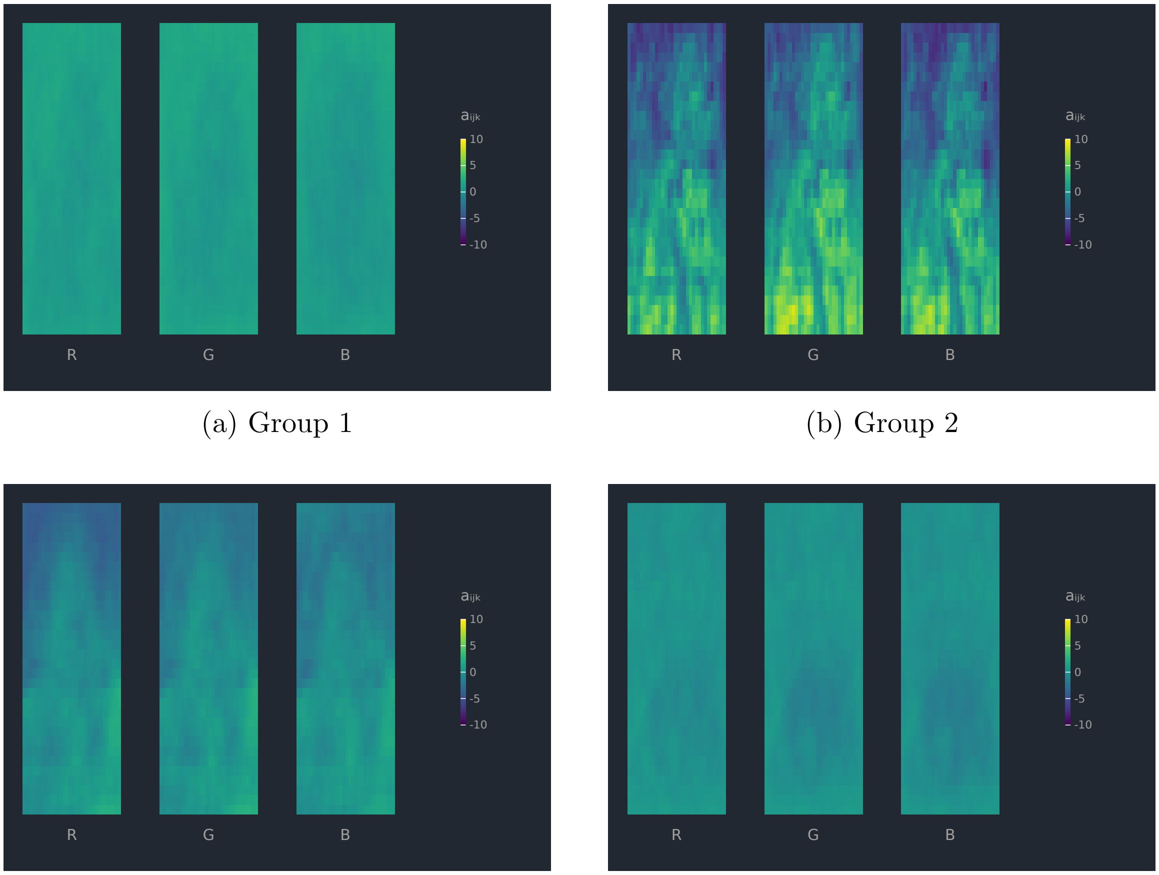
(c) Group 3

(d) Group 4

Figure 6.6: Visualizations of the $\boldsymbol{\mathfrak{M}}_g$ MDAs from the TVST model.

The slices of each cluster groups skewness tensor $\boldsymbol{\mathfrak{A}}_g$ is visualized in Figure 6.7. Cluster groups 1 and 4 do not contain any specific skewness patterns. Cluster group 3 has negative skewness around the transition from the trees foliage to the sky. The images in cluster group 2 have a pronounced skewness pattern, with both large positive and negative values, which help differentiate them from the other images.

Each group's three scale matrices are displayed as correlation matrices $\{\mathbf{P}_{gd}\}_{d=1}^{3}$,



(a) Group 1

(b) Group 2

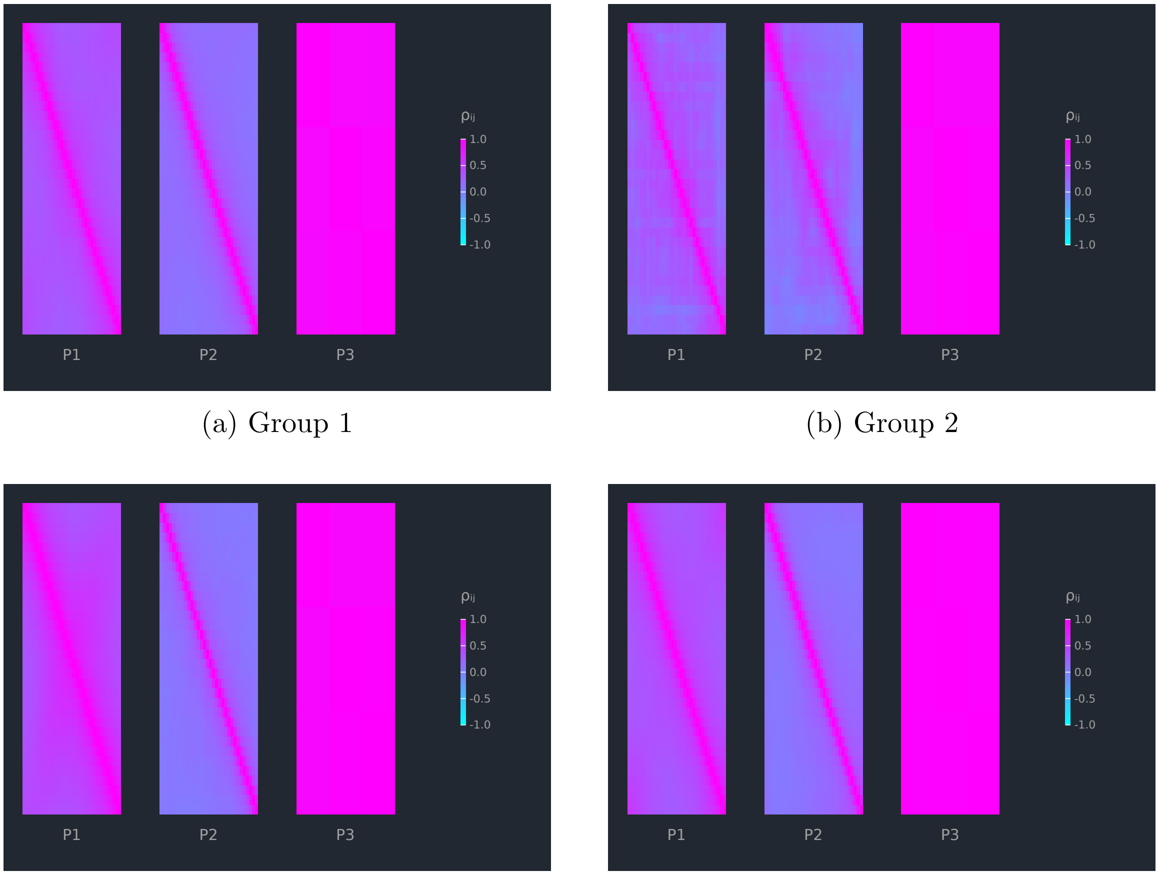
(c) Group 3

(d) Group 4

Figure 6.7: The slices of the $\mathfrak{A}_g$ MDAs from the TVST model.

in Figure 6.8. The correlation between the slices, $\mathbf{P}_{g3}$, are all highly positive. The column correlations, $\mathbf{P}_{g2}$, exhibit the expected pattern for image data in cluster groups 1, 3 and 4. On the other hand, the row correlations, visualized by $\mathbf{P}_{g1}$, have an atypical correlation pattern. Entries spaced farther apart exhibit roughly the same correlation as those located close together. Contrasting skewness patterns and variation in the rows and columns of the MDA are primarily responsible for producing the cluster groups we see in Table 6.2.

Given the modest ARI value of the chosen model, we expect that this clustering



(a) Group 1

(b) Group 2

(c) Group 3

(d) Group 4

Figure 6.8: The correlation matrices, derived from each $\boldsymbol{\Delta}_{gd}$ from the TVST model.

solution could be improved. As discussed in Chapter 3, a regularization scheme that imposes some kind of sparsity on the model parameters could be beneficial. Given these images have labels, a semi-supervised classification approach (McNicholas, 2010; McLachlan and Peel, 2000b) could be another good alternative.

In a semi-supervised classification scenario, suppose we know the labels for K of the N MDAs and the sample is ordered as $\boldsymbol{\mathfrak{X}}_1, \ldots, \boldsymbol{\mathfrak{X}}_K, \boldsymbol{\mathfrak{X}}_{K+1}, \ldots, \boldsymbol{\mathfrak{X}}_N$. We know the first K z_{ig} values and can write the observed-data likelihood as

$$\begin{aligned}\mathcal{L}_{\text{O}} = &\prod_{i=1}^{K}\prod_{g=1}^{G}\left[\pi_g f_{\text{TVD}}\left(\mathfrak{X}_i|\boldsymbol{\mathfrak{M}}_g, \mathfrak{A}_g, \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}, \boldsymbol{\theta}_g\right)\right]^{z_{ig}}\\ &\times \prod_{j=K+1}^{N}\sum_{h=1}^{G}\pi_h f_{\text{TVD}}\left(\mathfrak{X}_j|\boldsymbol{\mathfrak{M}}_h, \mathfrak{A}_h, \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{hd}, \boldsymbol{\theta}_h\right),\end{aligned}$$

where $\boldsymbol{\theta}_g$ are the parameters associated with the distribution of $W_{ig} \in \mathbb{R}^+$. Parameter estimation, identifiability, etc., follow in a comparable fashion to the clustering case described above. As the ratio K/N increases, we would expect the ARI of the solution to approach one.

Chapter 7

Conclusions

In this thesis, we present novel methods to cluster MDA data in its native form without the need for vectorization to satisfy the dated requirements of off-the-shelf clustering techniques. The first topic was formulating a mixture model for four-way and higher data that uses the MDA structure to estimate each groups mean MDA and the variation in each mode of the MDA data. Additionally we produce labels that can be used to define homogenous subgroups in the sample. We further extend this model to properly model temporal patterns of variation in specific modes of the MDA data, a common scenario with real world data.

We then characterize five novel tensor-variate skewed distributions and construct a mixture model for each of them. These distributions have the advantage of being able to model skewness in the MDA data and have heavier tails than the MLND. This provides the users of our models the advantage of being able to properly cluster their data in the presence of non-normal MDA data. Given the paucity of techniques for assessing tensor-variate normality and whether it is a valid assumption for a given sample of data, this could have important implications for applied data analysis.

As we mentioned in Chapter 3, our models, especially the skewed ones, could benefit from some form of sparsity constraints. Existing approaches to this problem come in two flavors, keeping the data in MDA form and sparsifying the estimated parameters (Mai *et al.*, 2021) or using some form of tensor decomposition (Kolda and Bader, 2009) on the MDA data and developing a probabilistic model for the components of the decomposition (Hoff *et al.*, 2016; Hinrich and Mørup, 2019). Given that model selection approaches have been shown to be superior to penalization in terms of variable selection (Celeux *et al.*, 2014), methodology akin to the latter is our preferred direction of future research.

Another promising avenue of research would be to use envelope methods (Cook, 2018) to formulate a finite mixture model for MDA data. Envelope methods are a relatively new research area in multivariate analysis, with the goal of jointly modeling parameters of interest and the covariance structure in the data using a low-dimensional subspace. This subspace captures all the relevant information about the parameters and could be an effective and concise means of doing regularization and parameter estimation in one model.

Appendix A

Tensor Manipulations

A.1 Vectorization Equivalences

If we have an order-3 MDA, $\mathfrak{X}$, its vectorization can be represented as:

$$\text{vec}(\mathfrak{X}) = \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3} x_{i_1 i_2 i_3}(\mathbf{e}_{i_1}^{n_1} \otimes \mathbf{e}_{i_2}^{n_2} \otimes \mathbf{e}_{i_3}^{n_3}),$$

where $x_{i_1 i_2 i_3}$ is $(i_1, i_2, i_3)^{\text{th}}$ element of $\mathfrak{X}$ and $\mathbf{e}_{i_1}^{n_1}$, $\mathbf{e}_{i_2}^{n_2}$ and $\mathbf{e}_{i_3}^{n_3}$ are unit basis vectors of size n_1, n_2 and n_3 respectively.

Noting $\text{vec}(\mathbf{a}\mathbf{b}^\top) = \mathbf{b} \otimes \mathbf{a}$, we can see that

$$\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3} x_{i_1 i_2 i_3}(\mathbf{e}_{i_1}^{n_1} \otimes \mathbf{e}_{i_2}^{n_2} \otimes \mathbf{e}_{i_3}^{n_3}) = \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\sum_{i_3=1}^{n_3} x_{i_1 i_2 i_3}(\mathbf{e}_{i_2}^{n_2} \otimes \mathbf{e}_{i_3}^{n_3})\mathbf{e}_{i_1}^{n_1 \top} = \text{vec}(\mathbf{X}_{(1)}),$$

where $\mathbf{X}_{(1)}$ is the mode one matricization of $\mathfrak{X}$. Note that this is the transpose of $\mathbf{X}_{(1)}$ in Kolda and Bader (2009), due to the unit basis vectors being in the opposite order. This hints at the fact that Kronecker products are permutation invariant. This result generalizes to order-D MDAs by expressing $\text{vec}(\boldsymbol{\mathfrak{x}})$ as follows:

$$\text{vec}(\boldsymbol{\mathfrak{x}}) = \sum_{I_D} x_{I_D} \left(\mathbf{e}_{i_1}^{n_1} \otimes \bigotimes_{d=2}^{D} \mathbf{e}_{i_d}^{n_d} \right) \tag{A.1}$$

where $I_D = \{i_1, \ldots, i_D : 1 \leq i_j \leq n_j, 1 \leq j \leq D\}$.

A.2 From Vectorization to Trace

Noting that $\text{tr}\mathbf{A}^\top\mathbf{S} = \text{vec}(\mathbf{A})^\top\text{vec}(\mathbf{S})$,

$$\begin{aligned}
&\text{vec}(\breve{\boldsymbol{\mathfrak{x}}})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}\text{vec}(\breve{\boldsymbol{\mathfrak{x}}}) = \text{vec}(\breve{\mathbf{X}}_{(1)})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1}\text{vec}(\breve{\mathbf{X}}_{(1)}) = \\
&\text{vec}(\breve{\mathbf{X}}_{(1)})^\top\text{vec}\left(\bigotimes_{d=2}^{D} \boldsymbol{\Delta}_d^{-1}\breve{\mathbf{X}}_{(1)}\boldsymbol{\Delta}_1^{-1} \right) = \text{tr}\left[\breve{\mathbf{X}}_{(1)}^\top \bigotimes_{d=2}^{D} \boldsymbol{\Delta}_d^{-1}\breve{\mathbf{X}}_{(1)}\boldsymbol{\Delta}_1^{-1} \right] = \\
&\text{tr}\left[\boldsymbol{\Delta}_1^{-1}\breve{\mathbf{X}}_{(1)}^\top \bigotimes_{d=2}^{D} \boldsymbol{\Delta}_d^{-1}\breve{\mathbf{X}}_{(1)} \right].
\end{aligned} \tag{A.2}$$

In (A.2), we have isolated $\boldsymbol{\Delta}_1$. To isolate $\boldsymbol{\Delta}_2$, the following manipulations are done to (A.2):

$$\begin{aligned}
&\bigotimes_{d=2}^{D} \boldsymbol{\Delta}_d^{-1} = \boldsymbol{\Delta}_2^{-1} \otimes \bigotimes_{d=3}^{D} \boldsymbol{\Delta}_d^{-1} = (\mathbf{I}_{n_2}\boldsymbol{\Delta}_2^{-1}\mathbf{I}_{n_2}) \otimes \left[\bigotimes_{d=3}^{D}\boldsymbol{\Delta}_d^{-\frac{1}{2}}\mathbf{I}_{n^*_{3:D}}\bigotimes_{d=3}^{D}\boldsymbol{\Delta}_d^{-\frac{\top}{2}}\right] = \\
&(\mathbf{I}_{n_2}\boldsymbol{\Delta}_2^{-1}\mathbf{I}_{n_2}) \otimes \left[\bigotimes_{d=3}^{D}\boldsymbol{\Delta}_d^{-\frac{1}{2}}\sum_{j}^{n^*_{3:D}}\mathbf{e}_j 1\mathbf{e}_j^\top \bigotimes_{d=3}^{D} \boldsymbol{\Delta}_d^{-\frac{\top}{2}}\right] \implies \\
&\sum_{j}^{n^*_{3:D}} \operatorname{tr}\left[\boldsymbol{\Delta}_1^{-1}\breve{\mathbf{X}}_{(1)}^\top \left(\mathbf{I}_{n_2}\otimes\bigotimes_{d=3}^{D}\boldsymbol{\Delta}_d^{-\frac{1}{2}}\mathbf{e}_j\right)\boldsymbol{\Delta}_2^{-1}\left(\mathbf{I}_{n_2}\otimes\mathbf{e}_j^\top\bigotimes_{d=3}^{D}\boldsymbol{\Delta}_d^{-\frac{\top}{2}}\right)\breve{\mathbf{X}}_{(1)}\right] = \\
&\sum_{j}^{n^*_{3:D}} \operatorname{tr}\left[\boldsymbol{\Delta}_1^{-1}\mathbf{X}_{(1)j}^\top\boldsymbol{\Delta}_2^{-1}\mathbf{X}_{(1)j}\right] \qquad \text{(A.3)}
\end{aligned}$$

where $\mathbf{X}_{(1)j} = \left(\mathbf{I}_{n_2}\otimes\mathbf{e}_j^\top\bigotimes_{d=3}^{D}\boldsymbol{\Delta}_d^{-\frac{\top}{2}}\right)\breve{\mathbf{X}}_{(1)}$.

A.3 Tensor Commutation Operator

The tensor commutation operator (TCO), $\boldsymbol{\mathfrak{K}}$, is introduced in Ohlson *et al.* (2013). It is an orthogonal matrix that interchanges basis vectors and scale matrices in a sequence of Kronecker products. The TCO has two useful properties we will leverage below; $\boldsymbol{\mathfrak{K}}_{gn} = \boldsymbol{\mathfrak{K}}_{ng}^\top$ and $\boldsymbol{\mathfrak{K}}_{gn} \times \boldsymbol{\mathfrak{K}}_{ng} = \mathbf{I}_{n^*}$. For example,

$$\begin{aligned}
\boldsymbol{\mathfrak{K}}_{i_g i_n}&\left(\bigotimes_{j=i_1}^{i_{g-1}}\mathbf{e}_{i_j}^{n_j}\otimes\mathbf{e}_{i_g}^{n_g}\otimes\bigotimes_{k=i_{g+1}}^{i_{n-1}}\mathbf{e}_{i_k}^{n_k}\otimes\mathbf{e}_{i_n}^{n_n}\otimes\bigotimes_{l=i_{n+1}}^{i_D}\mathbf{e}_{i_l}^{n_l}\right) = \\
&\left(\bigotimes_{j=i_1}^{i_{g-1}}\mathbf{e}_{i_j}^{n_j}\otimes\mathbf{e}_{i_n}^{n_n}\otimes\bigotimes_{k=i_{g+1}}^{i_{n-1}}\mathbf{e}_{i_k}^{n_k}\otimes\mathbf{e}_{i_g}^{n_g}\otimes\bigotimes_{l=i_{n+1}}^{i_D}\mathbf{e}_{i_l}^{n_l}\right)
\end{aligned}$$

where $i_1 < i_g < i_n < i_D$, and

$$\boldsymbol{\mathfrak{K}}_{gn} \left(\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d \right) \boldsymbol{\mathfrak{K}}_{ng} = \\ \left(\bigotimes_{d_1=1}^{g-1} \boldsymbol{\Delta}_{d_1} \otimes \boldsymbol{\Delta}_n \otimes \bigotimes_{d_2=g+1}^{n-1} \boldsymbol{\Delta}_{d_2} \otimes \boldsymbol{\Delta}_g \otimes \bigotimes_{d_3=n+1}^{D} \boldsymbol{\Delta}_{d_3} \right)$$

where $1 < g < n < D$.

The TCO is similar to the commutation matrix, outlined in Abadir and Magnus (2005). We use the TCO to permute the elements of equations A.2 and A.3

$$\begin{aligned} \text{vec}(\boldsymbol{\mathfrak{X}})^{n2} &= \boldsymbol{\mathfrak{K}}_{2n} \left[\sum_{I_D} x_{I_D} \left(\mathbf{e}_{i_1}^{n_1} \otimes \bigotimes_{d=2}^{D} \mathbf{e}_{i_d}^{n_d} \right) \right] = \sum_{I_D} x_{I_D} \boldsymbol{\mathfrak{K}}_{2n} \left(\mathbf{e}_{i_1}^{n_1} \otimes \bigotimes_{d=2}^{D} \mathbf{e}_{i_d}^{n_d} \right) \\ &= \sum_{I_D} x_{I_D} \left(\mathbf{e}_{i_1}^{n_1} \otimes \mathbf{e}_{i_n}^{n_n} \otimes \bigotimes_{j=3}^{n-1} \mathbf{e}_{i_j}^{n_j} \otimes \mathbf{e}_{i_2}^{n_2} \otimes \bigotimes_{k=n+1}^{D} \mathbf{e}_{i_k}^{n_k} \right) \\ &= \sum_{I_D} x_{I_D} \left(\mathbf{e}_{i_1}^{n_1} \otimes \mathbf{e}_{i_n}^{n_n} \otimes \bigotimes_{\substack{d=3 \\ d \neq n}}^{D} \mathbf{e}_{i_d}^{n_d} \right) \end{aligned}$$

This formulation results in a permutation of the mode 1 matrix, $\mathbf{X}_{(1)}$ that we designate $\mathbf{X}_{(1)}^{n2}$. We can use the TCO to modify the initial quadratic form in (A.2) as follows:

$$\text{vec}(\breve{\boldsymbol{\mathfrak{X}}})^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1} \text{vec}(\breve{\boldsymbol{\mathfrak{X}}}) = \text{vec}(\breve{\boldsymbol{\mathfrak{X}}})^{\top} \mathbf{I}_{n^*} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1} \mathbf{I}_{n^*} \text{vec}(\breve{\boldsymbol{\mathfrak{X}}}) = \\ \text{vec}(\breve{\boldsymbol{\mathfrak{X}}})^{\top} \boldsymbol{\mathfrak{K}}_{2n}^{\top} \boldsymbol{\mathfrak{K}}_{2n} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d^{-1} \boldsymbol{\mathfrak{K}}_{n2} \boldsymbol{\mathfrak{K}}_{2n} \text{vec}(\breve{\boldsymbol{\mathfrak{X}}}) = \\ \left(\text{vec}(\boldsymbol{\mathfrak{X}})^{n2} \right)^{\top} \left[\boldsymbol{\Delta}_1 \otimes \boldsymbol{\Delta}_n \otimes \bigotimes_{\substack{d=3 \\ d \neq n}}^{D} \boldsymbol{\Delta}_d^{-1} \right] \text{vec}(\boldsymbol{\mathfrak{X}})^{n2} \qquad \text{(A.4)}$$

Following the same steps between (A.2) and (A.3), we get the following trace:

$$\sum_{j}^{n^*_{2:D/n}} \operatorname{tr}\left[\boldsymbol{\Delta}_1^{-1}\left(\mathbf{X}^{n2}_{(1)j}\right)^\top \boldsymbol{\Delta}_n^{-1}\mathbf{X}^{n2}_{(1)j}\right],$$

which can be used to isolate and produce a closed form estimate of $\boldsymbol{\Delta}_n$.

A.4 From Vectorization to Array Norm

When writing the density of the tensor-variate normal, many papers (e.g. Hoff *et al.*, 2011) use an array norm as an alternative to the initial quadratic form in equation A.3. The array norm is comparable to the matrix Frobenius norm, which is commonly denoted as $\|\mathbf{X}\| = \sqrt{\langle \mathbf{X}, \mathbf{X}\rangle}$, where $\langle\cdot\rangle$ is the array inner product. The details of the equivalence is detailed below:

$$\begin{aligned}
&\operatorname{vec}(\breve{\boldsymbol{\mathfrak{X}}})^\top \bigotimes_{d=D}^{1} \boldsymbol{\Delta}_d^{-1}\operatorname{vec}(\breve{\boldsymbol{\mathfrak{X}}}) = \operatorname{vec}(\breve{\boldsymbol{\mathfrak{X}}})^\top \operatorname{vec}\left(\boldsymbol{\Delta}_1^{-1}\breve{\mathbf{X}}_{(1)} \bigotimes_{d=D}^{2} \boldsymbol{\Delta}_d^{-1}\right) = \\
&\operatorname{vec}(\breve{\boldsymbol{\mathfrak{X}}})^\top \operatorname{vec}\left(\breve{\boldsymbol{\mathfrak{X}}} \times_1 \boldsymbol{\Delta}_1^{-1} \times_2 \boldsymbol{\Delta}_2^{-1} \cdots \times_d \boldsymbol{\Delta}_D^{-1}\right) = \operatorname{vec}(\breve{\boldsymbol{\mathfrak{X}}})^\top \operatorname{vec}\left(\breve{\boldsymbol{\mathfrak{X}}} \times \widehat{\boldsymbol{\Delta}}^{-1}\right) = \\
&\langle \breve{\boldsymbol{\mathfrak{X}}}, \breve{\boldsymbol{\mathfrak{X}}} \times \widehat{\boldsymbol{\Delta}}^{-1}\rangle = \langle \breve{\boldsymbol{\mathfrak{X}}}, \breve{\boldsymbol{\mathfrak{X}}} \times \widehat{\boldsymbol{\Delta}}^{-\frac{1}{2}} \times \widehat{\boldsymbol{\Delta}}^{-\frac{\top}{2}}\rangle = \langle \breve{\boldsymbol{\mathfrak{X}}} \times \widehat{\boldsymbol{\Delta}}^{-\frac{1}{2}}, \breve{\boldsymbol{\mathfrak{X}}} \times \widehat{\boldsymbol{\Delta}}^{-\frac{1}{2}}\rangle = \\
&\left\|\breve{\boldsymbol{\mathfrak{X}}} \times \widehat{\boldsymbol{\Delta}}^{-\frac{1}{2}}\right\|^2, \qquad \text{(A.5)}
\end{aligned}$$

where $\times_d \boldsymbol{\Delta}_d^{-1}$ indicates a d-mode matrix product (e.g., $\boldsymbol{\Delta}_d^{-1}\breve{\mathbf{X}}_{(d)}$) and $\times\widehat{\boldsymbol{\Delta}}^{-1}$ is the Tucker product. The mathematical details associated with these products are outlined in Kolda (2006).

Appendix B

Mixtures of Tensor-Variate Normal Distributions

B.1 Mixture of Tensor-Variate Normal Distributions

B.1.1 $\ell_C(\boldsymbol{\vartheta})$

Noting that $n_g = \sum_{i=1}^{N} z_{ig}$ and $N = \sum_{g=1}^{G} n_g$,

$$
\begin{aligned}
\ell_C(\boldsymbol{\vartheta}) &= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\log\pi_g + \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\log f(\boldsymbol{\mathfrak{X}}_i|\boldsymbol{\Theta}_g) \\
&= \sum_{g=1}^{G} n_g\log\pi_g + \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\left[-\frac{n^*}{2}\log(2\pi) - \frac{n^*}{2}\sum_{d=1}^{D}\frac{1}{n_d}\log|\boldsymbol{\Delta}_{gd}|\right. \\
&\qquad \left. -\frac{1}{2}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)\right],
\end{aligned}
$$

which we can write as

$$\begin{aligned}\ell_C(\boldsymbol{\vartheta}) = \sum_{g=1}^{G} n_g \log \pi_g - \frac{Nn^*}{2}\log(2\pi) - \frac{n^*}{2}\sum_{g=1}^{G} n_g \sum_{d=1}^{D} \frac{1}{n_d}\log|\boldsymbol{\Delta}_{gd}| \\ - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g).\end{aligned}$$

B.2 $\mathbb{Q}$ function

We have

$$\begin{aligned}\mathbb{Q}(\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \hat{n}_g \log \pi_g - \frac{Nn^*}{2}\log(2\pi) - \frac{n^*}{2}\sum_{g=1}^{G} \hat{n}_g \sum_{d=1}^{D} \frac{1}{n_d}\log|\boldsymbol{\Delta}_{gd}| \\ - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g),\end{aligned} \tag{B.6}$$

where $\hat{z}_{ig}$ and $\hat{n}_g$ are estimates of their respective quantities.

B.2.1 $\mathbf{M}_{(1)g}$ update

We can write

$$\begin{aligned}\mathbb{Q}(\boldsymbol{\vartheta}) &= C - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig}\left[\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)\right] \\ &= C - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig}\left[-\text{vec}(\mathbf{X}_{(1)i})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\mathbf{M}_{(1)g}) - \text{vec}(\mathbf{M}_{(1)g})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\mathbf{X}_{(1)i})\right. \\ &\qquad \left.+\text{vec}(\mathbf{M}_{(1)g})^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\mathbf{M}_{(1)g})\right].\end{aligned}$$

After a little work, we have

$$\mathbb{Q}(\boldsymbol{\vartheta}) = C - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\left[-2\,\mathrm{tr}\left\{\mathbf{X}_{(1)i}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g}\right\} + \ \mathrm{tr}\left\{\mathbf{M}_{(1)g}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g}\right\}\right].$$

Now,

$$\begin{aligned}\frac{\partial}{\partial\mathbf{M}_{(1)g}}\,\mathbb{Q}(\boldsymbol{\vartheta}) &= -\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\left[-2\,\mathrm{tr}\left\{\mathbf{X}_{(1)i}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{dM}_{(1)g}\right\} + \ \mathrm{tr}\left\{\mathbf{dM}_{(1)g}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g}\right\}\right.\\ &\qquad\left. + \,\mathrm{tr}\left\{\mathbf{M}_{(1)g}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{dM}_{(1)g}\right\}\right]\\ &= -\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\left[-2\,\mathrm{tr}\left\{\mathbf{X}_{(1)i}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{dM}_{(1)g}\right\} + 2\,\mathrm{tr}\left\{\mathbf{M}_{(1)g}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{dM}_{(1)g}\right\}\right]\\ &= \sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\left[\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{X}_{(1)i} - \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g}\right].\end{aligned}$$

Solving for $\mathbf{M}_{(1)g}$,

$$\begin{aligned}\hat{n}_g\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g} &= \sum_{i=1}^{N}\hat{z}_{ig}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{X}_{(1)i}\\ \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}\times\hat{n}_g\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g} &= \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}\times\sum_{i=1}^{N}\hat{z}_{ig}\left[\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{X}_{(1)i}\right]\\ \mathbf{M}_{(1)g} &= \frac{1}{\hat{n}_g}\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}.\end{aligned}$$

The estimate of $\mathbf{M}_{(1)g}$ is

$$\hat{\mathbf{M}}_{(1)g} = \frac{1}{\hat{n}_g}\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}.$$

B.3 $\boldsymbol{\Delta}_{gd}$ updates

Recall that

$$\mathbb{Q}(\boldsymbol{\vartheta}) = -\frac{n^*}{2}\sum_{g=1}^{G}\hat{n}_g\sum_{d=1}^{D}\frac{1}{n_d}\log|\boldsymbol{\Delta}_{gd}| - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\text{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}_{(1)ij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}],$$

and note the following differentials:

$$\mathbf{dX}^{-1} = -\mathbf{X}^{-1}\mathbf{dXX}^{-1}, \qquad \mathbf{d}\log|\mathbf{X}| = \text{tr}\mathbf{X}^{-1}\mathbf{dX}.$$

B.3.1 $\boldsymbol{\Delta}_{g1}$ update

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\Delta}_{g1}}\mathbb{Q}(\boldsymbol{\vartheta}) &= -\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_1}\log|\boldsymbol{\Delta}_{g1}| - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\text{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}_{(1)ij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}] \\
&= -\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_1}\text{tr}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1} + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\text{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}_{(1)ij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}] \\
&= -\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_1}\text{tr}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1} + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\text{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}_{(1)ij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1}] \\
&= -\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_1}\boldsymbol{\Delta}_{g1}^{-\top} + \frac{1}{2}\sum_{g=1}^{G}\boldsymbol{\Delta}_{g1}^{-\top}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)ij}^{\top}\boldsymbol{\Delta}_{g2}^{-\top}\mathbf{X}_{(1)ij}\boldsymbol{\Delta}_{g1}^{-\top}
\end{aligned}$$

Note $\boldsymbol{\Delta}_{g*}^{-\top} = \boldsymbol{\Delta}_{g*}^{-1}$.

$$\frac{n^*}{2}\frac{\hat{n}_g}{n_1}\boldsymbol{\Delta}_{g1}^{-1} = \frac{1}{2}\boldsymbol{\Delta}_{g1}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)ij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}\boldsymbol{\Delta}_{g1}^{-1}$$

To solve for $\boldsymbol{\Delta}_{g1}$, we start by multiplying both sides by $\boldsymbol{\Delta}_{g1}$ from the left and the right:

$$\begin{aligned}
\frac{n^*}{2}\frac{\hat{n}_g}{n_1}\boldsymbol{\Delta}_{g1} &= \frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}^{\top}_{(1)ij}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}\\
\hat{\boldsymbol{\Delta}}_{g1} &= \frac{n_1}{n^*\hat{n}_g}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}^{\top}_{(1)ij}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}
\end{aligned}$$

B.3.2 $\boldsymbol{\Delta}_{g2}$ update

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\Delta}_{g2}}\mathbb{Q}(\boldsymbol{\vartheta}) &= -\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_2}\log|\boldsymbol{\Delta}_{g2}| - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\operatorname{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}^{\top}_{(1)ij}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}]\\
&= -\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_2}\operatorname{tr}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g2} + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\operatorname{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}^{\top}_{(1)ij}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g2}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}]\\
&= -\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_2}\operatorname{tr}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g2} + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\operatorname{tr}[\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}^{\top}_{(1)ij}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g2}]\\
&= -\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_2}\boldsymbol{\Delta}_{g2}^{-\top} + \frac{1}{2}\sum_{g=1}^{G}\boldsymbol{\Delta}_{g2}^{-\top}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)ij}\boldsymbol{\Delta}_{g1}^{-\top}\mathbf{X}^{\top}_{(1)ij}\boldsymbol{\Delta}_{g2}^{-\top}.
\end{aligned}$$

Solving gives

$$\frac{n^*}{2}\frac{\hat{n}_g}{n_2}\boldsymbol{\Delta}_{g2}^{-1} = \frac{1}{2}\boldsymbol{\Delta}_{g2}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)ij}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}^{\top}_{(1)ij}\boldsymbol{\Delta}_{g2}^{-1}.$$

To solve for $\mathbf{\Delta}_{g2}$, we start by multiplying both sides by $\mathbf{\Delta}_{g2}$ from the left and the right:

$$\begin{aligned}\frac{n^*}{2}\frac{\hat{n}_g}{n_2}\mathbf{\Delta}_{g2} &= \frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)ij}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)ij}^{\top} \\ \hat{\mathbf{\Delta}}_{g2} &= \frac{n_2}{n^*\hat{n}_g}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)ij}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)ij}^{\top}.\end{aligned}$$

B.3.3 $\mathbf{\Delta}_{gl}$ update

This derivation starts with a different version of $\mathbb{Q}(\boldsymbol{\vartheta})$, given by:

$$\mathbb{Q}(\boldsymbol{\vartheta}) = -\frac{n^*}{2}\sum_{g=1}^{G}\hat{n}_g\sum_{d=1}^{D}\frac{1}{n_d}\log|\mathbf{\Delta}_{gd}| - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{2:D/l}}\operatorname{tr}\left[\mathbf{\Delta}_{g1}^{-1}\left(\mathbf{X}_{(1)ij}^{l2}\right)^{\top}\mathbf{\Delta}_{gl}^{-1}\mathbf{X}_{(1)ij}^{l2}\right] \tag{B.7}$$

Starting with (B.7) and following the steps laid out above to find $\hat{\mathbf{\Delta}}_{g2}$, we get the following estimator of $\mathbf{\Delta}_{gl}$:

$$\hat{\mathbf{\Delta}}_{gl} = \frac{n_l}{n^*\hat{n}_g}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)ij}^{l2}\mathbf{\Delta}_{g1}^{-1}\left(\mathbf{X}_{(1)ij}^{l2}\right)^{\top}.$$

Appendix C

Scale Matrix Modifications

C.1 Cholesky decomposition

These aside from the $\mathbb{Q}(\boldsymbol{\vartheta})$ expressions, these derivations are similar to those in McNicholas and Murphy (2010). In our analysis, the $\boldsymbol{\Delta}_{g2}$ matrix does not represent a temporal dimension of the multi-dimensional array and will not be represented by a Cholesky decomposition in our models.

C.1.1 VVI

$\boldsymbol{\Delta}_{g1}$

To derive the VVI model for $\boldsymbol{\Delta}_{g1}$, we start with the following two terms from $\mathbb{Q}(\boldsymbol{\vartheta})$:

$$\overset{\text{I}}{-\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_1}\log(|\boldsymbol{\Delta}_{g1}|)} \qquad \overset{\text{II}}{-\frac{1}{2}\sum_{g=1}^{G}\operatorname{tr}\left[\boldsymbol{\Delta}_{g1}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)ij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)ij}\right]} \tag{C.8}$$

We define:

- $\mathbf{\Lambda}_{g1} = \frac{1}{\hat{n}_g} \sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{X}_{(1)ij}^\top \mathbf{\Delta}_{g2}^{-1} \mathbf{X}_{(1)ij}$
- $\mathbf{A}_{g1} = \mathbf{T}_{g1}^\top \delta_{g1}^{-1} \mathbf{I}_{n_1 \times n_1} \mathbf{T}_{g1}$

The $\mathbb{Q}(\boldsymbol{\vartheta})$ function is now:

$$\mathbb{Q}(\boldsymbol{\vartheta}) = C + \frac{n^*}{2} \sum_{g=1}^{G} \hat{n}_g \log(\delta_{g1}^{-1}) - \frac{1}{2} \sum_{g=1}^{G} \hat{n}_g \delta_{g1}^{-1} \operatorname{tr}\left[\mathbf{T}_{g1} \mathbf{\Lambda}_{g1} \mathbf{T}_{g1}^\top\right]. \tag{C.9}$$

To find the estimator of δ_{g1}, $\hat{\delta}_{g1}$ we take the partial derivative of (C.9) w.r.t to δ_{g1}^{-1}, equate the result to zero and solve for δ_{g1}^{-1}. We find

$$\hat{\delta}_{g1} = \frac{1}{n^*} \operatorname{tr}[\mathbf{T}_{g1} \mathbf{\Lambda}_{g1} \mathbf{T}_{g1}^\top].$$

To find the estimator of $\mathbf{T}_{g1}$, $\hat{\mathbf{T}}_{g1}$, we take the partial derivative of (C.9) w.r.t to $\mathbf{T}_{g1}$:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{T}_{g1}} \mathbb{Q}(\boldsymbol{\vartheta}) &= -\frac{1}{2} \sum_{g=1}^{G} \hat{n}_g \delta_{g1}^{-1} \left[\operatorname{tr}\left\{\mathbf{dT}_{g1} \mathbf{\Lambda}_{g1} \mathbf{T}_{g1}^\top\right\} + \operatorname{tr}\left\{\mathbf{T}_{g1} \mathbf{\Lambda}_{g1} \left(\mathbf{dT}_{g1}\right)^\top\right\}\right] \\ &= -\sum_{g=1}^{G} \hat{n}_g \delta_{g1}^{-1} \left[\operatorname{tr}\left\{\mathbf{\Lambda}_{g1} \mathbf{T}_{g1}^\top \left(\mathbf{dT}_{g1}\right)^\top\right\}\right] \\ &= -\sum_{g=1}^{G} \frac{\hat{n}_g}{\delta_{g1}} \left[\mathbf{T}_{g1} \mathbf{\Lambda}_{g1}\right].\end{aligned}$$

The matrix, $\mathbf{T}_{g1}$, is a unit-lower diagonal matrix. So for $r = 2, 3, \ldots, n_1$ we have:

$$\begin{bmatrix} \phi^{(g)1}_{r1} \\ \phi^{(g)1}_{r2} \\ \vdots \\ \phi^{(g)1}_{r,r-1} \end{bmatrix} = - \begin{bmatrix} \lambda^{(g)1}_{11} & \lambda^{(g)1}_{21} & \cdots & \lambda^{(g)1}_{r-1,1} \\ \lambda^{(g)1}_{12} & \lambda^{(g)1}_{22} & \cdots & \lambda^{(g)1}_{r-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda^{(g)1}_{1,r-1} & \lambda^{(g)1}_{2,r-1} & \cdots & \lambda^{(g)1}_{r-1,r-1} \end{bmatrix}^{-1} \begin{bmatrix} \lambda^{(g)1}_{r1} \\ \lambda^{(g)1}_{r2} \\ \vdots \\ \lambda^{(g)1}_{r,r-1} \end{bmatrix}.$$

This is equivalent to solving the following system of equations for $\mathbf{\Phi}^{(g),1}_{(r-1)\times 1}$:

$$\mathbf{\Lambda}^{(g)1\top}_{(r-1)\times(r-1)}\mathbf{\Phi}^{(g)1}_{(r-1)\times 1} = -\mathbf{\Lambda}^{(g)1}_{(r-1)\times 1}.$$

Note: $\mathbf{\Lambda}^{(g)1}_{(r-1)\times(r-1)}$ is symmetric because $\lambda^{(g)1}_{ij} = \lambda^{(g)1}_{ji}$.

The number of free parameters is

$$G\left[\frac{n_1(n_1-1)}{2}\right] + G.$$

$\mathbf{\Delta}_{gl}$

To find the VVI model for $\mathbf{\Delta}_{gl}$, we start with the following two terms from the permuted version of (C.8):

$$\overset{\text{I}}{-\frac{n^*}{2n_l}\sum_{g=1}^{G}\hat{n}_g \log(|\mathbf{\Delta}_{gl}|)} \qquad \overset{\text{II}}{-\frac{1}{2}\sum_{g=1}^{G} \operatorname{tr}\left[\mathbf{\Delta}^{-1}_{gl}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{2:D/l}}\mathbf{X}^{l2}_{(1)ij}\mathbf{\Delta}^{-1}_{g1}(\mathbf{X}^{l2}_{(1)ij})^\top\right]}$$

We define:

- $\mathbf{\Lambda}_{gl} = \frac{1}{\hat{n}_g}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{2:D/l}}\mathbf{X}^{l2}_{(1)ij}\mathbf{\Delta}^{-1}_{g1}(\mathbf{X}^{l2}_{(1)ij})^\top$
- $\mathbf{A}_{gl} = \mathbf{T}^\top_{gl}\delta^{-1}_{gl}\mathbf{I}_{n_l\times n_l}\mathbf{T}_{gl}$

and following similar steps outlined for $\mathbf{\Delta}_{g1}$ above, we find the $\mathbb{Q}(\boldsymbol{\vartheta})$ function is

$$\mathbb{Q}(\boldsymbol{\vartheta}) = C + \frac{n^*}{2}\sum_{g=1}^{G} \hat{n}_g \log(\delta_{gl}^{-1}) - \frac{1}{2}\sum_{g=1}^{G} \hat{n}_g \delta_{gl}^{-1} \operatorname{tr}\left[\mathbf{T}_{gl}\mathbf{\Lambda}_{gl}\mathbf{T}_{gl}^{\top}\right], \tag{C.10}$$

the estimator for δ_{gl} is

$$\hat{\delta}_{gl} = \frac{1}{n^*}\operatorname{tr}[\mathbf{T}_{gl}\mathbf{\Lambda}_{gl}\mathbf{T}_{gl}^{\top}], \tag{C.11}$$

and the second score function is:

$$\mathrm{S}_2(\delta_{gl}, \mathbf{T}_{gl}) = -\frac{\hat{n}_g}{\delta_{gl}}\mathbf{T}_{gl}\mathbf{\Lambda}_{gl}. \tag{C.12}$$

We can find the sub-diagonal components of $\mathbf{T}_{gl}$ by solving the following system of equations for $\mathbf{\Phi}^{(g)l}_{(r-1)\times 1}$:

$$\mathbf{\Lambda}^{(g)l\top}_{(r-1)\times(r-1)}\mathbf{\Phi}^{(g)l}_{(r-1)\times 1} = -\mathbf{\Lambda}^{(g)l}_{(r-1)\times 1}$$

The number of free parameters is:

$$G\left[\frac{n_l(n_l-1)}{2}\right] + G.$$

If $\mathbf{\Delta}_{g2}$ was VVI, the derivation would be similar to $\mathbf{\Delta}_{gl}$ except we would start with the first two terms in (C.8).

C.1.2 EVI

$\mathbf{\Delta}_{g1}$

Starting with (C.8) and Λ_{g1} from Section C.1.1, we define

$$\mathbf{A}_{g1} = \mathbf{T}_1^{\top}\delta_{g1}^{-1}\mathbf{I}_{n_1\times n_1}\mathbf{T}_1$$

and use these quantities to find the following $\mathbb{Q}(\boldsymbol{\vartheta})$ function

$$\mathbb{Q}(\boldsymbol{\vartheta}) = C + \frac{n^*}{2}\sum_{g=1}^{G}\hat{n}_g\log(\delta_{g1}^{-1}) - \frac{1}{2}\sum_{g=1}^{G}\hat{n}_g\delta_{g1}^{-1}\,\mathrm{tr}\left[\mathbf{T}_1\mathbf{\Lambda}_{g1}\mathbf{T}_1^{\top}\right]. \tag{C.13}$$

The estimator of δ_{g1}, $\hat{\delta}_{g1}$, is found by taking the partial derivative of (C.13) w.r.t to δ_{g1}^{-1}, equating it to zero and solving for δ_{g1}:

$$\hat{\delta}_{g1} = \frac{1}{n^*}\,\mathrm{tr}[\mathbf{T}_1\mathbf{\Lambda}_{g1}\mathbf{T}_1^{\top}]. \tag{C.14}$$

Taking the partial derivative of (C.13) w.r.t to $\mathbf{T}_1$, we find the second score function is:

$$\mathrm{S}_2(\delta_{g1},\mathbf{T}_1) = -\mathbf{T}_1\sum_{g=1}^{G}\frac{\hat{n}_g}{\delta_{g1}}\mathbf{\Lambda}_{g1} \tag{C.15}$$

Note the following:

$\mathbf{\Phi}_1 = \{\phi_{1ij}\}$, elements of $\mathbf{T}_1$ to be estimated.

$$\kappa_{ij}^{1} = \sum_{g=1}^{G}\frac{n_g\lambda_{1ij}^{(g)}}{\delta_{g1}}.$$

Then, for $r = 1, 2, \ldots, n_1$, we have:

$$\begin{bmatrix} \phi^1_{r1} \\ \phi^1_{r2} \\ \vdots \\ \phi^1_{r,r-1} \end{bmatrix} = - \begin{bmatrix} \kappa^1_{11} & \kappa^1_{21} & \cdots & \kappa^1_{r-1,1} \\ \kappa^1_{12} & \kappa^1_{22} & \cdots & \kappa^1_{r-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa^1_{1,r-1} & \kappa^1_{2,r-1} & \cdots & \kappa^1_{r-1,r-1} \end{bmatrix}^{-1} \begin{bmatrix} \kappa^1_{r1} \\ \kappa^1_{r2} \\ \vdots \\ \kappa^1_{r,r-1} \end{bmatrix},$$

where $r = 2, 3, \ldots, n_1$. This is equivalent to solving the following system of equations for $\mathbf{\Phi}^1_{(r-1)\times 1}$:

$$\boldsymbol{\kappa}^{1,\top}_{(r-1)\times(r-1)} \mathbf{\Phi}^1_{(r-1)\times 1} = -\boldsymbol{\kappa}^1_{(r-1)\times 1}$$

The number of free parameters is:

$$\frac{n_1(n_1-1)}{2} + G$$

$\mathbf{\Delta}_{gl}$

Starting with the permuted version of (C.8) and Λ_{gl} defined in Section C.1.1, we define

$$\mathbf{A}_{gl} = \mathbf{T}_l^\top \delta_{gl}^{-1} \mathbf{I}_{n_l \times n_l} \mathbf{T}_l$$

and use these quantities to find the following $\mathbb{Q}(\boldsymbol{\vartheta})$ function:

$$\mathbb{Q}(\boldsymbol{\vartheta}) = C + \frac{n^*}{2} \sum_{g=1}^{G} \hat{n}_g \log(\delta_{gl}^{-1}) - \frac{1}{2} \sum_{g=1}^{G} \hat{n}_g \delta_{gl}^{-1} \operatorname{tr}\left[\mathbf{T}_l \mathbf{\Lambda}_{gl} \mathbf{T}_l^\top\right], \tag{C.16}$$

the estimator of δ_{gl}:

$$\hat{\delta}_{gl} = \frac{1}{n^*} \operatorname{tr}[\mathbf{T}_l \mathbf{\Lambda}_{gl} \mathbf{T}_l^\top], \tag{C.17}$$

and the second score function $S_2(\delta_{g1}, \mathbf{T}_1)$:

$$\mathrm{S}_2(\delta_{gl}, \mathbf{T}_l) = -\mathbf{T}_l \sum_{g=1}^{G} \frac{\hat{n}_g}{\delta_{gl}} \mathbf{\Lambda}_{gl}. \tag{C.18}$$

If we note,

$$\kappa_{ij}^{l} = \sum_{g=1}^{G} \frac{n_g \lambda_{lij}^{(g)}}{\delta_{gl}},$$

then we can solve the following system of equations for $\mathbf{\Phi}^{l}_{(r-1)\times 1}$, for $r = 1, 2, \ldots, n_l$.

$$\boldsymbol{\kappa}^{l\top}_{(r-1)\times(r-1)} \mathbf{\Phi}^{l}_{(r-1)\times 1} = -\boldsymbol{\kappa}^{l}_{(r-1)\times 1}$$

This gives the sub-diagonal elements of $\mathbf{T}_l$.

Note: $\boldsymbol{\kappa}^{l}_{(r-1)\times(r-1)}$ is symmetric because $\kappa_{ij}^{l} = \kappa_{ji}^{l}$.

The number of free parameters is:

$$\frac{n_l(n_l - 1)}{2} + G.$$

Note: If $\mathbf{\Delta}_{g2}$ was EVI, the derivation would be similar to $\mathbf{\Delta}_{gl}$ except we would start with a modified version of (C.13).

C.2 Eigen-Decomposition

$$\mathbf{\Delta}_{n_* \times n_*} = \mathbf{\Gamma}\mathbf{A}\mathbf{\Gamma}^{\top} = \lambda \mathbf{\Gamma}\mathbf{D}\mathbf{\Gamma}^{\top}$$

where:

- $\lambda = |\mathbf{\Delta}|^{\frac{1}{n_*}}$ and represents the volume.

- $\mathbf{D} = |\boldsymbol{\Delta}|^{-\frac{1}{n_*}} \mathbf{I}_{n_*} \mathbf{A}$ and represents the shape.
  - $\mathbf{D}$ is a diagonal matrix that contains the normalized eigen-values of $\boldsymbol{\Delta}$ in decreasing order.
  - $|\mathbf{D}| = 1$
- $\boldsymbol{\Gamma}$ are the ordered eigen-vectors of $\boldsymbol{\Delta}$ and represent the orientation

In our applied example, the eigen decompositions describe below are applied to $\boldsymbol{\Delta}_{g2}$ because they do not represent a temporal dimension of the multidimensional arrays.

VVI

Starting with (C.8), we define:

- $\boldsymbol{\Lambda}_{g2} = \sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n_{3:D}^{*}} \mathbf{X}_{(1)ij} \boldsymbol{\Delta}_{g1}^{-1} \mathbf{X}_{(1)ij}^{\top}$
- $\mathbf{A}_{g2} = \boldsymbol{\Delta}_{g2}^{-1} = \lambda_{g2}^{-1} \mathbf{D}_{g2}^{-1}$

The two terms in (C.8) can be decomposed as follows:

$$
\begin{array}{rr}
\text{I} & \text{II} \\
-\frac{n^*}{2n_2}\sum_{g=1}^{G}\hat{n}_g\log(|\mathbf{A}_{g2}^{-1}|) & -\frac{1}{2}\sum_{g=1}^{G}\operatorname{tr}\left[\boldsymbol{\Delta}_{g2}^{-1}\boldsymbol{\Lambda}_{g2}\right] \\
\frac{n^*}{2n_2}\sum_{g=1}^{G}\hat{n}_g\log(|\mathbf{A}_{g2}|) & -\frac{1}{2}\sum_{g=1}^{G}(\lambda_{g2}^{-1})\operatorname{tr}\left[\boldsymbol{\Lambda}_{g2}\mathbf{D}_{g2}^{-1}\right] \\
\frac{n^*}{2n_2}\sum_{g=1}^{G}\hat{n}_g\log(|(\lambda_{g2}^{-1})\mathbf{D}_{g2}^{-1}|) & \cdots \\
-\frac{n^*}{2n_2}\sum_{g=1}^{G}\hat{n}_g\log((\lambda_{g2}^{-1})^{n_2}) & \cdots \\
-\frac{n^*}{2}\sum_{g=1}^{G}\hat{n}_g\log(\lambda_{g2}) & \cdots
\end{array}
$$

The $\mathbb{Q}(\boldsymbol{\vartheta})$ function is now

$$
\mathbb{Q}(\boldsymbol{\vartheta}) = C + -\frac{1}{2}\left[\sum_{g=1}^{G}\frac{1}{\lambda_{g2}}\operatorname{tr}\left\{\boldsymbol{\Lambda}_{g2}\mathbf{D}_{g2}^{-1}\right\} + n^*\sum_{g=1}^{G}\hat{n}_g\log(\lambda_{g2})\right]. \tag{C.19}
$$

Using the expressions for $\lambda_k\mathbf{B}_k$ on page 12 of Celeux and Govaert (1995), the estimators of $\mathbf{D}_{g2}$ and λ_{g2} are:

$$
\begin{aligned}
\hat{\mathbf{D}}_{g2} &= \frac{\operatorname{diag}(\boldsymbol{\Lambda}_{g2})}{|\operatorname{diag}(\boldsymbol{\Lambda}_{g2})|^{\frac{1}{n^*}}}, \\
\hat{\lambda}_{g2} &= \frac{|\operatorname{diag}(\boldsymbol{\Lambda}_{g2})|^{\frac{1}{n^*}}}{n_g},
\end{aligned}
$$

where $d = n^*$, $\mathbf{W}_k = \boldsymbol{\Lambda}_{g2}$ and $n_k = \hat{n}_g$.

EEE

Starting with equation C.8 for the $\mathbb{Q}(\boldsymbol{\vartheta})$, we can define

- $\boldsymbol{\Lambda}_{g2} = \sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{X}_{(1)ij} \boldsymbol{\Delta}_{g1}^{-1} \mathbf{X}_{(1)ij}^{\top}$
- $\boldsymbol{\Delta}_{g2} = \lambda_2 \boldsymbol{\Gamma}_2 \mathbf{D}_2 \boldsymbol{\Gamma}_2^{\top}$

The $\mathbb{Q}(\boldsymbol{\vartheta})$ can be rewritten as:

$$\mathbb{Q}(\boldsymbol{\vartheta}) = C - \frac{1}{2}\left[N\frac{n^*}{n_2}\log(|\boldsymbol{\Delta}_2|) + \text{tr}\left\{\left(\sum_{g=1}^{G}\boldsymbol{\Lambda}_{g2}\right)\boldsymbol{\Delta}_2^{-1}\right\}\right]. \tag{C.20}$$

The common variance matrix is defined in Celeux and Govaert (1995) as:

$$\hat{\boldsymbol{\Delta}}_2 = \frac{1}{N}\sum_{g=1}^{G}\boldsymbol{\Lambda}_{g2}$$

where $\mathbf{W} = \sum_{g=1}^{G} \boldsymbol{\Lambda}_{g2}$ and $d = \frac{n^*}{n_2}$.

The number of free parameters is

$$\frac{n_2(n_2+1)}{2}.$$

Appendix D

Skew Distributions

D.1 Skewed Distributions

D.1.1 Derivation details for the tensor variate skew t distribution

$\mathscr{X}$ is a random MDA with a $\text{TVST}_{\mathbf{n}}(\boldsymbol{\mathfrak{M}}, \boldsymbol{\mathfrak{A}}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d, \nu)$. $\mathscr{X}$ can be written as

$$\mathscr{X} = \boldsymbol{\mathfrak{M}} + W\boldsymbol{\mathfrak{A}} + \sqrt{W}\mathscr{V},$$

where $\boldsymbol{\mathfrak{M}}$ and $\boldsymbol{\mathfrak{A}}$ are $\mathbf{n}$ dimensional MDAs, $\mathscr{V} \sim \mathcal{N}_{\mathbf{n}}\left(\mathbb{O}, \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_d\right)$ and $W \sim \text{Inv-Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$. The inverse Gamma density has the form:

$$f(w|a,b) = \frac{b^a}{\Gamma(a)} w^{-(a+1)} \exp\left[-\frac{b}{w}\right].$$

It then follows that

$$\mathscr{X}|W = w \sim \mathcal{N}_{\mathbf{n}}\left(\boldsymbol{\mathfrak{M}} + w\boldsymbol{\mathfrak{A}}, w\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d\right).$$

Joint density

The joint density of $\mathscr{X}$ and W is

$$\begin{aligned} f(\boldsymbol{\mathfrak{x}}, w|\boldsymbol{\vartheta}) &= f(\boldsymbol{\mathfrak{x}}|W = w)f(w) \\ &= (2\pi)^{\frac{-n^*}{2}}\prod_{d=1}^{D}[w^{n_d}|\boldsymbol{\Delta}_d|]^{-\frac{n^*}{2n_d}} \\ &\quad\times \exp\left\{-\frac{1}{2}\text{vec}(\boldsymbol{\mathfrak{x}} - \boldsymbol{\mathfrak{M}} - w\boldsymbol{\mathfrak{A}})^{\top}\frac{1}{w}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\text{vec}(\boldsymbol{\mathfrak{x}} - \boldsymbol{\mathfrak{M}} - w\boldsymbol{\mathfrak{A}})\right\} \\ &\quad\times \frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}\left[\frac{1}{w}\right]^{\frac{\nu}{2}+1}\exp\left\{\frac{\frac{\nu}{2}}{w}\right\} \\ &= \frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{(2\pi)^{\frac{n^*}{2}}\prod_{d=1}^{D}|\boldsymbol{\Delta}_d|^{\frac{n^*}{2n_d}}\Gamma(\frac{\nu}{2})}\cdot w^{-\left(\frac{\nu+n^*}{2}+1\right)} \\ &\quad\times \exp\left\{-\frac{1}{2w}\left(\text{vec}(\boldsymbol{\mathfrak{x}} - \boldsymbol{\mathfrak{M}} - w\boldsymbol{\mathfrak{A}})^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\text{vec}(\boldsymbol{\mathfrak{x}} - \boldsymbol{\mathfrak{M}} - w\boldsymbol{\mathfrak{A}}) + \nu\right)\right\} \end{aligned} \tag{D.21}$$

The intermediate steps pertaining to the determinant of the scale matrices are :

$$\prod_{d=1}^{D}|w\boldsymbol{\Delta}_d|^{-\frac{n^*}{2n_d}} = \prod_{d=1}^{D}[w^{n_d}|\boldsymbol{\Delta}_d|]^{-\frac{n^*}{2n_d}} = w^{-\frac{n^*}{2}}\prod_{d=1}^{D}|\boldsymbol{\Delta}_d|^{-\frac{n^*}{2n_d}}$$

Using the following identities:

- $\text{vec}(\mathbf{A} + \mathbf{E}) = \text{vec}(\mathbf{A}) + \text{vec}(\mathbf{E})$

- $\text{vec}(\alpha\mathbf{A}) = \alpha\text{vec}(\mathbf{A})$

the exponential term in (D.21) can be written as:

$$\exp\left\{-\frac{1}{2w}\left(\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}}-w\boldsymbol{\mathfrak{A}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}}-w\boldsymbol{\mathfrak{A}})+\nu\right)\right\}=$$

$$\exp\left\{-\frac{1}{2w}\left(\left[\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top-w\mathrm{vec}(\boldsymbol{\mathfrak{A}})^\top\right]\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\left[\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})-w\mathrm{vec}(\boldsymbol{\mathfrak{A}})\right]+\nu\right)\right\}=$$

$$\exp\left\{-\frac{1}{2w}\left(\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})-w\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})\right.\right.$$
$$\left.\left.-w\mathrm{vec}(\boldsymbol{\mathfrak{A}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})+w^2\mathrm{vec}(\boldsymbol{\mathfrak{A}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})+\nu\right)\right\}=$$

$$\exp\left\{-\frac{1}{2w}\left(\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})-2w\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})\right.\right.$$
$$\left.\left.+w^2\mathrm{vec}(\boldsymbol{\mathfrak{A}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})+\nu\right)\right\}$$

Grouping the terms that incorporate w facilitates the integration in Section D.1.1.

$$\exp\left\{\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})\right\}\times$$
$$\exp\left\{-\frac{1}{2}\left[\frac{\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})+\nu}{w}+w\mathrm{vec}(\boldsymbol{\mathfrak{A}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})\right]\right\}$$

If we define

$$\delta(\cdot)=\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\textstyle\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})\qquad\rho(\cdot)=\mathrm{vec}(\boldsymbol{\mathfrak{A}})^\top\textstyle\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}}),$$

then the following expression is obtained:

$$\exp\left\{\left(\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})\right)\right\}$$
$$\times\exp\left\{-\frac{1}{2}\left[\frac{\delta(\boldsymbol{\mathfrak{X}};\boldsymbol{\mathfrak{M}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+\nu}{w}+w\rho(\boldsymbol{\mathfrak{A}};\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})\right]\right\} \tag{D.22}$$

Marginal Density

Building on (D.21) and (D.22), the marginal density of $\boldsymbol{\mathfrak{X}}$ is

$$\begin{aligned}f(\boldsymbol{\mathfrak{X}})&=\int_0^\infty f(\boldsymbol{\mathfrak{X}},w)dw\\&=\frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{(2\pi)^{\frac{n^*}{2}}\prod_{d=1}^{D}|\boldsymbol{\Delta}_d|^{\frac{n^*}{2n_d}}\Gamma(\frac{\nu}{2})}\exp\left\{\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})\right\}\\&\quad\times\int_0^\infty w^{-\left(\frac{\nu+n^*}{2}+1\right)}\exp\left\{-\frac{1}{2}\left[\frac{\delta(\boldsymbol{\mathfrak{X}};\boldsymbol{\mathfrak{M}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+\nu}{w}+w\rho(\boldsymbol{\mathfrak{A}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})\right]\right\}dw.\end{aligned} \tag{D.23}$$

Using the following change of variables, we can rearrange the integral in (D.23):

$$\begin{array}{cc}\mathrm{u} & \mathrm{du}\\ \frac{\sqrt{\rho(\boldsymbol{\mathfrak{A}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})}}{\sqrt{\delta(\boldsymbol{\mathfrak{X}};\boldsymbol{\mathfrak{M}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+\nu}}w & \frac{\sqrt{\rho(\boldsymbol{\mathfrak{A}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})}}{\sqrt{\delta(\boldsymbol{\mathfrak{X}};\boldsymbol{\mathfrak{M}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+\nu}}dw\end{array}$$

$$\begin{array}{lr}
\text{Integral term 1} & \text{Integral term 2} \\
w^{-\left(\frac{\nu+n^*}{2}+1\right)} & \exp\left\{-\frac{1}{2}\left[\frac{\delta(\cdot)+\nu}{w}+w\rho(\cdot)\right]\right\}dw \\
\left[\frac{\sqrt{\delta(\cdot)+\nu}}{\sqrt{\rho(\cdot)}}u\right]^{-\left(\frac{\nu+n^*}{2}+1\right)} & \exp\left\{-\frac{1}{2}\left[\frac{\sqrt{\delta(\cdot)+\nu}\sqrt{\delta(\cdot)+\nu}\sqrt{\rho(\cdot)}}{w\sqrt{\rho(\cdot)}}+\frac{w\sqrt{\rho(\cdot)}\sqrt{\rho(\cdot)}\sqrt{\delta(\cdot)+\nu}}{\sqrt{\delta(\cdot)+\nu}}\right]\right\}dw \\
\left[\left[\frac{\delta(\cdot)+\nu}{\rho(\cdot)}\right]^{\frac{1}{2}}\right]^{-\left(\frac{\nu+n^*}{2}+1\right)}u^{-\left(\frac{\nu+n^*}{2}+1\right)} & \exp\left\{-\frac{1}{2}\left[\frac{\sqrt{\delta(\cdot)+\nu}\sqrt{\rho(\cdot)}}{u}+\sqrt{\rho(\cdot)}\sqrt{\delta(\cdot)+\nu u}\right]\right\}dw \\
\left[\frac{\delta(\cdot)+\nu}{\rho(\cdot)}\right]^{-\left(\frac{\nu+n^*}{4}+\frac{1}{2}\right)}u^{-\left(\frac{\nu+n^*}{2}+1\right)} & \exp\left\{-\frac{\sqrt{\delta(\cdot)+\nu}\sqrt{\rho(\cdot)}}{2}\left[u+\frac{1}{u}\right]\right\}\left[\frac{\delta(\cdot)+\nu}{\rho(\cdot)}\right]^{\frac{1}{2}}du
\end{array}$$

Putting the two terms together, we have:

$$\int_0^\infty\left[\frac{\delta(\cdot)+\nu}{\rho(\cdot)}\right]^{-\left(\frac{\nu+n^*}{4}+\frac{1}{2}\right)}u^{-\left(\frac{\nu+n^*}{2}+1\right)}\left[\frac{\delta(\cdot)+\nu}{\rho(\cdot)}\right]^{\frac{1}{2}}\exp\left\{-\frac{\sqrt{\delta(\cdot)+\nu}\sqrt{\rho(\cdot)}}{2}\left[u+\frac{1}{u}\right]\right\}du =$$
$$\left[\frac{\delta(\cdot)+\nu}{\rho(\cdot)}\right]^{-\frac{\nu+n^*}{4}}\int_0^\infty u^{-\left(\frac{\nu+n^*}{2}+1\right)}\exp\left\{-\frac{\sqrt{\delta(\cdot)+\nu}\sqrt{\rho(\cdot)}}{2}\left[u+\frac{1}{u}\right]\right\}du$$

The integral is now a Bessel function of the second kind. We can write the marginal density, $f(\boldsymbol{\mathfrak{X}})$ as:

$$f_{\mathrm{TVST}}(\boldsymbol{\mathfrak{X}}|\boldsymbol{\vartheta}) =\frac{2\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\exp\left\{\mathrm{vec}(\boldsymbol{\mathfrak{X}}-\boldsymbol{\mathfrak{M}})^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}})\right\}}{(2\pi)^{\frac{n^*}{2}}\prod_{d=1}^{D}|\boldsymbol{\Delta}_d|^{\frac{n^*}{2n_d}}\Gamma(\frac{\nu}{2})}\left(\frac{\delta(\boldsymbol{\mathfrak{X}};\boldsymbol{\mathfrak{M}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})+\nu}{\rho(\mathbf{A},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1})}\right)^{-\frac{\nu+n^*}{4}}$$
$$\times K_{-\frac{\nu+n^*}{2}}\left(\sqrt{\left[\rho(\boldsymbol{\mathfrak{A}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}))\right]\left[\delta(X;\boldsymbol{\mathfrak{M}},\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_d^{-1}))+\nu\right]}\right). \quad \text{(D.24)}$$

D.1.2 Expectations

We define the following terms; $\mathbf{n}_{2:D} = \prod_{d=2}^{D} n_d$, $\breve{\boldsymbol{\Delta}} = \bigotimes_{d=D}^{2} \boldsymbol{\Delta}_d$ and $\breve{\boldsymbol{\Delta}}^{\frac{1}{2}} = \bigotimes_{d=D}^{2} \boldsymbol{\Delta}_d^{\frac{1}{2}}$.

$$\begin{aligned}
\mathbb{E}\left[\text{vec}(\mathscr{X})\text{vec}(\mathscr{X})^\top\right] &= \mathbb{E}\left[\text{vec}(\mathbf{X}_{(1)})\text{vec}(\mathbf{X}_{(1)})^\top\right] \\
&= \mathbb{E}\Big[\Big\{\text{vec}(\mathbf{M}_{(1)}) + W\text{vec}(\mathbf{A}_{(1)}) + \sqrt{W}(\breve{\boldsymbol{\Delta}}^{\frac{1}{2}} \otimes \boldsymbol{\Delta}_1^{\frac{1}{2}})\text{vec}(\mathbf{Z}_{(1)})\Big\} \times \\
&\qquad \Big\{\text{vec}(\mathbf{M}_{(1)})^\top + W\text{vec}(\mathbf{A}_{(1)})^\top + \sqrt{W}\text{vec}(\mathbf{Z}_{(1)})^\top(\breve{\boldsymbol{\Delta}}^{\frac{1}{2}} \otimes \boldsymbol{\Delta}_1^{\frac{1}{2}})^\top\Big\}\Big] \\
&= \text{vec}(\mathbf{M}_{(1)})\text{vec}(\mathbf{M}_{(1)})^\top + \mathbb{E}[W]\text{vec}(\mathbf{M}_{(1)})\text{vec}(\mathbf{A}_{(1)})^\top \\
&\qquad + \mathbb{E}[W]\text{vec}(\mathbf{A}_{(1)})\text{vec}(\mathbf{M}_{(1)})^\top + \mathbb{E}[W^2]\text{vec}(\mathbf{A}_{(1)})\text{vec}(\mathbf{A}_{(1)})^\top \\
&\qquad + \mathbb{E}[W](\breve{\boldsymbol{\Delta}}^{\frac{1}{2}} \otimes \boldsymbol{\Delta}_1^{\frac{1}{2}})\,\mathbb{E}\left[\text{vec}(\mathbf{Z}_{(1)})\text{vec}(\mathbf{Z}_{(1)})^\top\right](\breve{\boldsymbol{\Delta}}^{\frac{1}{2}} \otimes \boldsymbol{\Delta}_1^{\frac{1}{2}})^\top \\
&= \text{vec}(\mathbf{M}_{(1)})\text{vec}(\mathbf{M}_{(1)})^\top + \mathbb{E}[W]\text{vec}(\mathbf{M}_{(1)})\text{vec}(\mathbf{A}_{(1)})^\top \\
&\qquad + \mathbb{E}[W]\text{vec}(\mathbf{A}_{(1)})\text{vec}(\mathbf{M}_{(1)})^\top \\
&\qquad + \mathbb{E}[W^2]\text{vec}(\mathbf{A}_{(1)})\text{vec}(\mathbf{A}_{(1)})^\top + \mathbb{E}[W]\left(\bigotimes_{d=D}^{1}\boldsymbol{\Delta}_d\right).
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\mathbf{X}_{(1)}\mathbf{X}_{(1)}^\top\right] &= \mathbb{E}\left[\left\{\mathbf{M}_{(1)} + W\mathbf{A}_{(1)} + \sqrt{W}\boldsymbol{\Delta}_1^{\frac{1}{2}}\mathbf{Z}_{(1)}\breve{\boldsymbol{\Delta}}^{\frac{1}{2}}\right\} \times \left\{\mathbf{M}_{(1)} + W\mathbf{A}_{(1)} + \sqrt{W}\boldsymbol{\Delta}_1^{\frac{1}{2}}\mathbf{Z}_{(1)}\breve{\boldsymbol{\Delta}}^{\frac{1}{2}}\right\}^\top\right] \\
&= \mathbf{M}_{(1)}\mathbf{M}_{(1)}^\top + \mathbb{E}\left[W\right]\mathbf{M}_{(1)}\mathbf{A}_{(1)}^\top + \mathbb{E}\left[W\right]\mathbf{A}_{(1)}\mathbf{M}_{(1)}^\top + \mathbb{E}\left[W^2\right]\mathbf{A}_{(1)}\mathbf{A}_{(1)}^\top \\
&\qquad + \mathbb{E}[W]\boldsymbol{\Delta}_1^{\frac{1}{2}}\,\mathbb{E}\left[\mathbf{Z}_{(1)}\breve{\boldsymbol{\Delta}}\mathbf{Z}_{(1)}^\top\right]\boldsymbol{\Delta}_1^{\frac{\top}{2}}
\end{aligned}$$

The $\mathbf{Z}_{(1)}$ term can be broken into row vectors, $\mathbf{z}_{(1)} \in \mathbb{R}^{1\times \mathbf{n}_{2:D}}$.

$$\begin{aligned}
\mathbb{E}\left[\mathbf{z}_{(1)}\breve{\boldsymbol{\Delta}}\mathbf{z}_{(1)}^{\top}\right] &= \mathbb{E}\left[\operatorname{tr}\left\{\mathbf{z}_{(1)}^{\top}\mathbf{z}_{(1)}\breve{\boldsymbol{\Delta}}\right\}\right] = \operatorname{tr}\left\{\mathbb{E}\left[\mathbf{z}_{(1)}^{\top}\mathbf{z}_{(1)}\right]\breve{\boldsymbol{\Delta}}\right\} = \operatorname{tr}\left\{\mathbf{I}_{\mathbf{n}_{2:D}}\breve{\boldsymbol{\Delta}}\right\} = \operatorname{tr}\left\{\breve{\boldsymbol{\Delta}}\right\},\\
\mathbb{E}\left[\mathbf{Z}_{(1)}\breve{\boldsymbol{\Delta}}\mathbf{Z}_{(1)}^{\top}\right] &= \mathbf{I}_{n_1} \times \operatorname{tr}\left\{\breve{\boldsymbol{\Delta}}\right\} = \mathbf{I}_{n_1} \times \operatorname{tr}\left\{\bigotimes_{d=D}^{2}\boldsymbol{\Delta}_d\right\} = \mathbf{I}_{n_1} \times \prod_{d=2}^{D}\operatorname{tr}\left\{\boldsymbol{\Delta}_d\right\},\\
\mathbb{E}\left[\mathbf{X}_{(1)}\mathbf{X}_{(1)}^{\top}\right] &= \mathbf{M}_{(1)}\mathbf{M}_{(1)}^{\top} + \mathbb{E}\left[W\right]\mathbf{M}_{(1)}\mathbf{A}_{(1)}^{\top} + \mathbb{E}\left[W\right]\mathbf{A}_{(1)}\mathbf{M}_{(1)}^{\top} + \mathbb{E}\left[W^2\right]\mathbf{A}_{(1)}\mathbf{A}_{(1)}^{\top}\\
&\quad + \mathbb{E}[W]\boldsymbol{\Delta}_1 \times \prod_{d=2}^{D}\operatorname{tr}\left\{\boldsymbol{\Delta}_d\right\}\\
\mathbb{E}\left[\mathbf{X}_{(1)}^{\top}\mathbf{X}_{(1)}\right] &= \mathbb{E}\left[\left\{\mathbf{M}_{(1)} + W\mathbf{A}_{(1)} + \sqrt{W}\boldsymbol{\Delta}_1^{\frac{1}{2}}\mathbf{Z}_{(1)}\breve{\boldsymbol{\Delta}}^{\frac{1}{2}}\right\}^{\top} \times \left\{\mathbf{M}_{(1)} + W\mathbf{A}_{(1)} + \sqrt{W}\boldsymbol{\Delta}_1^{\frac{1}{2}}\mathbf{Z}_{(1)}\breve{\boldsymbol{\Delta}}^{\frac{1}{2}}\right\}\right]\\
&= \mathbf{M}_{(1)}^{\top}\mathbf{M}_{(1)} + \mathbb{E}\left[W\right]\mathbf{M}_{(1)}^{\top}\mathbf{A}_{(1)} + \mathbb{E}\left[W\right]\mathbf{A}_{(1)}^{\top}\mathbf{M}_{(1)} + \mathbb{E}\left[W^2\right]\mathbf{A}_{(1)}^{\top}\mathbf{A}_{(1)}\\
&\quad + \mathbb{E}[W]\breve{\boldsymbol{\Delta}}^{\frac{\top}{2}}\,\mathbb{E}\left[\mathbf{Z}_{(1)}^{\top}\boldsymbol{\Delta}_1\mathbf{Z}_{(1)}\right]\breve{\boldsymbol{\Delta}}^{\frac{1}{2}}
\end{aligned}$$

The $\mathbf{Z}_{(1)}^{\top}$ term can be broken into row vectors, $\mathbf{z}_{(1)}^{\top} \in \mathbb{R}^{1\times n_1}$.

$$\begin{aligned}
\mathbb{E}\left[\mathbf{z}_{(1)}^{\top}\boldsymbol{\Delta}_1\mathbf{z}_{(1)}\right] &= \mathbb{E}\left[\operatorname{tr}\left\{\mathbf{z}_{(1)}\mathbf{z}_{(1)}^{\top}\boldsymbol{\Delta}_1\right\}\right] = \operatorname{tr}\left\{\mathbb{E}\left[\mathbf{z}_{(1)}\mathbf{z}_{(1)}^{\top}\right]\boldsymbol{\Delta}_1\right\} = \operatorname{tr}\left\{\mathbf{I}_{n_1}\boldsymbol{\Delta}_1\right\} = \operatorname{tr}\left\{\boldsymbol{\Delta}_1\right\}\\
\mathbb{E}\left[\mathbf{Z}_{(1)}^{\top}\boldsymbol{\Delta}_1\mathbf{Z}_{(1)}\right] &= \mathbf{I}_{\mathbf{n}_{2:D}} \times \operatorname{tr}\left\{\boldsymbol{\Delta}_1\right\},\\
\mathbb{E}\left[\mathbf{X}_{(1)}^{\top}\mathbf{X}_{(1)}\right] &= \mathbf{M}_{(1)}^{\top}\mathbf{M}_{(1)} + \mathbb{E}\left[W\right]\mathbf{M}_{(1)}^{\top}\mathbf{A}_{(1)} + \mathbb{E}\left[W\right]\mathbf{A}_{(1)}^{\top}\mathbf{M}_{(1)} + \mathbb{E}\left[W^2\right]\mathbf{A}_{(1)}^{\top}\mathbf{A}_{(1)}\\
&\quad + \mathbb{E}[W]\breve{\boldsymbol{\Delta}} \times \operatorname{tr}\left\{\boldsymbol{\Delta}_1\right\}
\end{aligned}$$

Appendix E

Mixtures of Tensor-Variate Skew Distributions

E.1 $\log f(\boldsymbol{\mathfrak{X}}_i, w_{ig}|\boldsymbol{\Theta}_g)$ for the TVST

$$
\begin{aligned}
\log f(\boldsymbol{\mathfrak{X}}_i,w_{ig}|\boldsymbol{\Theta}_g) = &-\frac{n^*}{2}\log(2\pi) - \frac{n^*}{2}\sum_{d=1}^{D}\frac{1}{n_d}\log|\boldsymbol{\Delta}_{gd}| \\
&+\frac{\nu_g}{2}\log(\frac{\nu_g}{2}) - \log\left(\Gamma\left(\frac{\nu_g}{2}\right)\right) - \left(\frac{n^*}{2}+\frac{\nu_g}{2}+1\right)\log(w_{ig}) - \frac{\nu_g}{2w_{ig}} \\
&+\frac{1}{2}\text{vec}(\boldsymbol{\mathfrak{X}}_i-\boldsymbol{\mathfrak{M}}_g)^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{A}}_g)+\frac{1}{2}\text{vec}(\boldsymbol{\mathfrak{A}}_g)^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{X}}_i-\boldsymbol{\mathfrak{M}}_g) \\
&-\frac{1}{2w_{ig}}\text{vec}(\boldsymbol{\mathfrak{X}}_i-\boldsymbol{\mathfrak{M}}_g)^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{X}}_i-\boldsymbol{\mathfrak{M}}_g)-\frac{w_{ig}}{2}\text{vec}(\boldsymbol{\mathfrak{A}}_g)^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{A}}_g).
\end{aligned}
$$

E.2 $\ell_C(\boldsymbol{\vartheta})$

Noting that $n_g = \sum_{i=1}^{N} z_{ig}$ and $N = \sum_{g=1}^{G} n_g$, the complete log-likelihood for the TVST distribution is

$$\begin{aligned}
\ell_C(\boldsymbol{\vartheta}) &= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\log\pi_g + \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\log f(\boldsymbol{\mathfrak{X}}_i, w_{ig}|\boldsymbol{\Theta}_g) \\
&= \sum_{g=1}^{G} n_g\log\pi_g + \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\left[-\frac{n^*}{2}\log(2\pi) - \frac{n^*}{2}\sum_{d=1}^{D}\frac{1}{n_d}\log|\boldsymbol{\Delta}_{gd}|\right. \\
&\quad +\frac{\nu_g}{2}\log(\frac{\nu_g}{2}) - \log\left(\Gamma\left(\frac{\nu_g}{2}\right)\right) - \left(\frac{n^*}{2}+\frac{\nu_g}{2}+1\right)\log(w_{ig}) - \frac{\nu_g}{2w_{ig}} \\
&\quad +\frac{1}{2}\mathrm{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}}_g) + \frac{1}{2}\mathrm{vec}(\boldsymbol{\mathfrak{A}}_g)^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
&\quad \left. -\frac{1}{2w_{ig}}\mathrm{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) - \frac{w_{ig}}{2}\mathrm{vec}(\boldsymbol{\mathfrak{A}}_g)^\top\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}}_g)\right]
\end{aligned} \tag{E.25}$$

The complete-data log-likelihood, $\ell_C(\boldsymbol{\vartheta})$, can be broken into a general form for all five tensor-variate mixture models by making the following groupings of the terms; constants unrelated to the parameters, $h(w_{ig}|\vartheta)$ terms and terms with tensor/matrix parameters:

$$
\begin{aligned}
\ell_{C1} &= \sum_{g=1}^{G} n_g \log \pi_g, \\
C_1 &= -\frac{Nn^*}{2}\log(2\pi) \\
\ell_{C2} &= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\log h(w_{ig}|\boldsymbol{\theta}_g), \\
C_2 &= \text{Constants related to } \log h(w_{ig}|\boldsymbol{\theta}_g), \\
\ell_{C3} &= -\frac{n^*}{2}\sum_{g=1}^{G} n_g \sum_{d=1}^{D}\frac{1}{n_d}\log|\boldsymbol{\Delta}_{gd}| + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{A}}_g) \\
&\quad + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\text{vec}(\boldsymbol{\mathfrak{A}}_g)^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
&\quad - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} \frac{z_{ig}}{w_{ig}}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
&\quad - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}w_{ig}\text{vec}(\boldsymbol{\mathfrak{A}}_g)^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\text{vec}(\boldsymbol{\mathfrak{A}}_g).
\end{aligned}
$$

In the case of the TVST, ℓ_{C2} and C_2 are:

$$
\begin{aligned}
\ell_{C2} &= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\left[\frac{\nu_g}{2}\log\left(\frac{\nu_g}{2}\right) - \log\left(\Gamma\left(\frac{\nu_g}{2}\right)\right) - \left(\frac{\nu_g}{2}\right)\log(w_{ig}) - \frac{\nu_g}{2w_{ig}}\right] \\
&= \sum_{g=1}^{G} n_g\left[\frac{\nu_g}{2}\log\left(\frac{\nu_g}{2}\right)\log\left(\Gamma\left(\frac{\nu_g}{2}\right)\right)\right] - \sum_{g=1}^{G}\frac{\nu_g}{2}\sum_{i=1}^{N}\left[z_{ig}\log(w_{ig}) - \frac{z_{ig}}{w_{ig}}\right] \\
C_2 &= \sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\left[-\left(\frac{n^*}{2}\right)\log(w_{ig}) - \log(w_{ig})\right] \\
&= -\left(\frac{n^*}{2}+1\right)\sum_{g=1}^{G}\sum_{i=1}^{N} z_{ig}\log(w_{ig}),
\end{aligned}
$$

where w_{ig} is a constant and ν_g is the parameter. Now we have

$$\ell_C(\boldsymbol{\vartheta}) = \ell_{C1} + C_1 + \ell_{C2} + C_2 + \ell_{C3}.$$

E.3 $\mathbb{Q}$ function

Note that

$$a_{ig} = \mathbb{E}(W_{ig} \mid \boldsymbol{\mathfrak{X}}_i, \hat{\boldsymbol{\vartheta}}) \qquad b_{ig} = \mathbb{E}\left(\frac{1}{W_{ig}} \,\middle|\, \boldsymbol{\mathfrak{X}}_i, \hat{\boldsymbol{\vartheta}}\right) \qquad c_{ig} = \mathbb{E}(\log W_{ig} \mid \boldsymbol{\mathfrak{X}}_i, \hat{\boldsymbol{\vartheta}}).$$

Now, the $\mathbb{Q}(\boldsymbol{\vartheta})$ function for the TVST is given by

$$\begin{aligned}
\mathbb{Q}(\boldsymbol{\vartheta}) = {} & \sum_{g=1}^{G} \hat{n}_g \log \pi_g - \frac{Nn^*}{2} \log(2\pi) \\
& + \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} \left[\frac{\nu_g}{2} \log\left(\frac{\nu_g}{2}\right) - \log\left(\Gamma\left(\frac{\nu_g}{2}\right)\right) - \left(\frac{n^*}{2} + \frac{\nu_g}{2} + 1\right) c_{ig} - \frac{\nu_g}{2} b_{ig} \right] \\
& - \frac{n^*}{2} \sum_{g=1}^{G} \hat{n}_g \sum_{d=1}^{D} \frac{1}{n_d} \log|\boldsymbol{\Delta}_{gd}| + \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} \text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}(\boldsymbol{\mathfrak{A}}_g) \\
& + \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} \text{vec}(\boldsymbol{\mathfrak{A}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
& - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} b_{ig} \text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
& - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} a_{ig} \text{vec}(\boldsymbol{\mathfrak{A}}_g)^\top \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \text{vec}(\boldsymbol{\mathfrak{A}}_g), \qquad \text{(E.26)}
\end{aligned}$$

where $\hat{z}_{ig}$ and $\hat{n}_g$ are estimates of their respective quantities.

E.4 $\mathbf{M}_{(1)g}$ and $\mathbf{A}_{(1)g}$ updates

Note:

$$\mathrm{vec}(\boldsymbol{\mathfrak{X}} - \boldsymbol{\mathfrak{M}})^\top = \mathrm{vec}(\boldsymbol{\mathfrak{X}})^\top - \mathrm{vec}(\boldsymbol{\mathfrak{M}})^\top \quad \mathrm{tr}\mathbf{A}^\top\mathbf{S} = \mathrm{vec}(\mathbf{A})^\top\mathrm{vec}(\mathbf{S}) \quad \mathrm{tr}\mathbf{A}^\top\mathbf{S} = \mathrm{tr}\mathbf{S}^\top\mathbf{A}$$

E.4.1 $\mathbf{M}_{(1)g}$ update

We have

$$\begin{aligned}
\mathbb{Q}(\boldsymbol{\vartheta}) = C &+ \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\mathrm{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{A}}_g) \\
&+ \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\mathrm{vec}(\boldsymbol{\mathfrak{A}}_g)^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
&- \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathrm{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathrm{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \\
= C &- \sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\left[\mathrm{tr}\left\{\mathbf{A}_{(1)g}^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g}\right\}\right] \\
&- \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\left[-2\,\mathrm{tr}\left\{\mathbf{X}_{(1)i}^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g}\right\} + \mathrm{tr}\left\{\mathbf{M}_{(1)g}^\top \bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g}\right\}\right]
\end{aligned}$$

and so

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{M}_{(1)g}} \mathbb{Q}(\boldsymbol{\vartheta}) = &- \sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig} \left[\operatorname{tr}\left\{ \mathbf{A}_{(1)g}^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \mathbf{dM}_{(1)g} \right\}\right] \\
&- \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig} b_{ig} \left[-2\operatorname{tr}\left\{ \left(\mathbf{X}_{(1)i}^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} - \mathbf{M}_{(1)g}^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \right) \mathbf{dM}_{(1)g} \right\}\right] \\
= &- \sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig} \left[\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \mathbf{A}_{(1)g} \right] \\
&+ \sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig} b_{ig} \left[\bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \mathbf{X}_{(1)i} - \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \mathbf{M}_{(1)g} \right].
\end{aligned}
$$

We now have

$$
\hat{\mathbf{M}}_{(1)g} = \frac{\sum_{i=1}^{N} \hat{z}_{ig} b_{ig} \mathbf{X}_{(1)i} - \hat{n}_g \hat{\mathbf{A}}_{(1)g}}{\sum_{i=1}^{N} \hat{z}_{ig} b_{ig}}. \tag{E.27}
$$

E.4.2 $\mathbf{A}_{(1)g}$ update

Now, we can write

$$
\begin{aligned}
\mathbb{Q}(\boldsymbol{\vartheta}) = &\, C + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig} \left[\operatorname{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g)^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \operatorname{vec}(\boldsymbol{\mathfrak{A}}_g) + \operatorname{vec}(\boldsymbol{\mathfrak{A}}_g)^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \operatorname{vec}(\boldsymbol{\mathfrak{X}}_i - \boldsymbol{\mathfrak{M}}_g) \right] \\
&- \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig} a_{ig} \operatorname{vec}(\boldsymbol{\mathfrak{A}}_g)^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \operatorname{vec}(\boldsymbol{\mathfrak{A}}_g) \\
= &\, C + \sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig} \left[\operatorname{tr}\left\{ \mathbf{X}_{(1)i}^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \mathbf{A}_{(1)g} \right\} - \operatorname{tr}\left\{ \mathbf{M}_{(1)g}^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \mathbf{A}_{(1)g} \right\}\right] \\
&- \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N} \hat{z}_{ig} a_{ig} \left[\mathbf{A}_{(1)g}^{\top} \bigotimes_{d=1}^{D} \boldsymbol{\Delta}_{gd}^{-1} \mathbf{A}_{(1)g} \right],
\end{aligned}
$$

and so

$$\begin{aligned}\frac{\partial}{\partial \mathbf{A}_{(1)g}}\,\mathbb{Q}(\boldsymbol{\vartheta}) &= \sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\left[\operatorname{tr}\left\{\left(\mathbf{X}_{(1)i}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}-\mathbf{M}_{(1)g}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\right)\mathbf{dA}_{(1)g}\right\}\right]\\ &\quad-\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\left[\mathbf{A}_{(1)g}^{\top}\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{dA}_{(1)g}\right]\\ &=\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\left[\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{X}_{(1)i}-\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{M}_{(1)g}\right]-\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\left[\bigotimes_{d=1}^{D}\boldsymbol{\Delta}_{gd}^{-1}\mathbf{A}_{(1)g}\right].\end{aligned}$$

We have

$$\hat{\mathbf{A}}_{(1)g}=\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}-\hat{n}_g\hat{\mathbf{M}}_{(1)g}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}}. \tag{E.28}$$

E.4.3 Solve Equations for Final Updates

We have two equations with two unknown parameters, $\mathbf{A}_{(1)g}$ and $\mathbf{M}_{(1)g}$. We define the following quantities:

$$\bar{a}=\frac{1}{\hat{n}_g}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig} \qquad\qquad \bar{b}=\frac{1}{\hat{n}_g}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}.$$

Substituting (E.28) into (E.27),

$$\begin{aligned}\hat{\mathbf{M}}_{(1)g}&=\frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}-\hat{n}_g\left[\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}-\hat{n}_g\hat{\mathbf{M}}_{(1)g}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}}\right]}{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}\\ &=\frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}-\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}}{\bar{a}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}+\frac{\hat{n}_g\hat{\mathbf{M}}_{(1)g}}{\bar{a}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}.\end{aligned}$$

We can write

$$\hat{\mathbf{M}}_{(1)g}\left[1-\frac{\hat{n}_g}{\bar{a}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}\right]=\frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}-\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}}{\bar{a}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}$$

and so

$$\begin{aligned}\hat{\mathbf{M}}_{(1)g}&=\frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}-\frac{\bar{a}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}}{\hat{n}_g}-\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}}{\bar{a}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}\\&\quad+\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}}{\hat{n}_g}\\&=\left[\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}\left\{\bar{a}b_{ig}-1\right\}}{\sum_{i=1}^{N}\hat{z}_{ig}\bar{a}b_{ig}}\right]-\left[\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}\left\{\bar{a}b_{ig}-1\right\}}{\hat{n}_g}\right].\end{aligned}$$

That is,

$$\hat{\mathbf{M}}_{(1)g}=\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}\left\{\bar{a}b_{ig}-1\right\}}{\sum_{i=1}^{N}\hat{z}_{ig}\bar{a}b_{ig}-\hat{n}_g}. \tag{E.29}$$

Now, substituting (E.27) into (E.28),

$$\begin{aligned}\hat{\mathbf{A}}_{(1)g}&=\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}-\hat{n}_g\left[\frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}-\hat{n}_g\hat{\mathbf{A}}_{(1)g}}{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}}\right]}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}}\\&=\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}}-\frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\bar{b}}+\frac{\hat{n}_g\hat{\mathbf{A}}_{(1)g}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\bar{b}}.\end{aligned}$$

We can write

$$\hat{\mathbf{A}}_{(1)g}\left[1-\frac{\hat{n}_g}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\bar{b}}\right]=\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}}-\frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\bar{b}}$$

and so

$$\begin{aligned}\hat{\mathbf{A}}_{(1)g} &= \frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}} - \frac{\bar{b}\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}}{\hat{n}_g} - \frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\bar{b}} \\ &\quad + \frac{\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\mathbf{X}_{(1)i}}{\hat{n}_g} \\ &= \left[\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}\left\{\bar{b}-b_{ig}\right\}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\bar{b}}\right] - \left[\frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}\left\{\bar{b}-b_{ig}\right\}}{\hat{n}_g}\right].\end{aligned}$$

That is,

$$\hat{\mathbf{A}}_{(1)g} = \frac{\sum_{i=1}^{N}\hat{z}_{ig}\mathbf{X}_{(1)i}\left\{\bar{b}-b_{ig}\right\}}{\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\bar{b}-\hat{n}_g}. \tag{E.30}$$

E.5 $\mathbf{\Delta}_{gd}$ updates

$$\begin{aligned}\mathbb{Q}(\boldsymbol{\vartheta}) = \mathbb{Q}_1 + C_1 + \mathbb{Q}_2 + C_2 - \frac{n^*}{2}\sum_{g=1}^{G}\hat{n}_g\sum_{d=1}^{D}\frac{1}{n_d}\log|\mathbf{\Delta}_{gd}| \\ + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\text{tr}[\mathbf{\Delta}_1^{-1}\mathbf{X}_{(1)gij}^{\top}\mathbf{\Delta}_2^{-1}\mathbf{A}_{(1)gj}] \\ + \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\text{tr}[\mathbf{\Delta}_1^{-1}\mathbf{A}_{(1)gj}^{\top}\mathbf{\Delta}_2^{-1}\mathbf{X}_{(1)gij}] \\ - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\sum_{j=1}^{n^*_{3:D}}\text{tr}[\mathbf{\Delta}_1^{-1}\mathbf{X}_{(1)gij}^{\top}\mathbf{\Delta}_2^{-1}\mathbf{X}_{(1)gij}] \\ - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\sum_{j=1}^{n^*_{3:D}}\text{tr}[\mathbf{\Delta}_1^{-1}\mathbf{A}_{(1)gj}^{\top}\mathbf{\Delta}_2^{-1}\mathbf{A}_{(1)gj}]\end{aligned} \tag{E.31}$$

where $\mathbb{Q}_*$ is the $\mathbb{Q}$ function version of ℓ_{C*}.

E.5.1 $\boldsymbol{\Delta}_{g1}$ update

Note the following differentials:

$$\mathbf{d}\mathbf{X}^{-1} = -\mathbf{X}^{-1}\mathbf{d}\mathbf{X}\mathbf{X}^{-1} \qquad\qquad \mathbf{d}\log|\mathbf{X}| = \operatorname{tr}\mathbf{X}^{-1}\mathbf{d}\mathbf{X}$$

We can write

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\Delta}_{g1}}\mathbb{Q}(\boldsymbol{\vartheta}) = &-\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_1}\operatorname{tr}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1} - \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\operatorname{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{A}_{(1)gj}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1}]\\
&-\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\operatorname{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1}]\\
&+\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\sum_{j=1}^{n^*_{3:D}}\operatorname{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1}]\\
&+\frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\sum_{j=1}^{n^*_{3:D}}\operatorname{tr}[\boldsymbol{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{A}_{(1)gj}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{d}\boldsymbol{\Delta}_{g1}]\\
= &-\frac{n^*}{2}\sum_{g=1}^{G}\frac{\hat{n}_g}{n_1}\boldsymbol{\Delta}_{g1}^{-\top} - \frac{1}{2}\sum_{g=1}^{G}\boldsymbol{\Delta}_{g1}^{-\top}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-\top}\mathbf{X}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-\top}\\
&-\frac{1}{2}\sum_{g=1}^{G}\boldsymbol{\Delta}_{g1}^{-\top}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-\top}\mathbf{A}_{(1)gj}\boldsymbol{\Delta}_{g1}^{-\top}\\
&+\frac{1}{2}\sum_{g=1}^{G}\boldsymbol{\Delta}_{g1}^{-\top}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-\top}\mathbf{X}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-\top}\\
&+\frac{1}{2}\sum_{g=1}^{G}\boldsymbol{\Delta}_{g1}^{-\top}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\sum_{j=1}^{n^*_{3:D}}\boldsymbol{\Delta}_{g1}^{-\top}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-\top}\mathbf{A}_{(1)gj}\boldsymbol{\Delta}_{g1}^{-\top}
\end{aligned}$$

Note $\boldsymbol{\Delta}_{g1}^{-\top} = \boldsymbol{\Delta}_{g1}^{-1}$ and solve for $\boldsymbol{\Delta}_{g1}$:

$$\begin{aligned}\frac{n^*}{2}\frac{\hat{n}_g}{n_1}\boldsymbol{\Delta}_{g1}^{-1} = &-\frac{1}{2}\boldsymbol{\Delta}_{g1}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-1}\\ &-\frac{1}{2}\boldsymbol{\Delta}_{g1}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{A}_{(1)gj}\boldsymbol{\Delta}_{g1}^{-1}\\ &+\frac{1}{2}\boldsymbol{\Delta}_{g1}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-1}\\ &+\frac{1}{2}\boldsymbol{\Delta}_{g1}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\sum_{j=1}^{n^*_{3:D}}\boldsymbol{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{A}_{(1)gj}\boldsymbol{\Delta}_{g1}^{-1}\end{aligned}$$

Multiply both sides by $\boldsymbol{\Delta}_{g1}$ from the left and the right:

$$\begin{aligned}\frac{n^*}{2}\frac{\hat{n}_g}{n_1}\boldsymbol{\Delta}_{g1} = &-\frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij} - \frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{A}_{(1)gj}\\ &+\frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij} + \frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{A}_{(1)gj}\end{aligned}$$

The update for $\boldsymbol{\Delta}_{g1}$, i.e., $\hat{\boldsymbol{\Delta}}_{g1}$, is:

$$\begin{aligned}\hat{\boldsymbol{\Delta}}_{g1} = \frac{n_1}{n^*\hat{n}_g}\Bigg[\sum_{i=1}^{N}\hat{z}_{ig}\Bigg\{&b_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij} + a_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{A}_{(1)gj}\\ &-\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{X}_{(1)gij} - \sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}^{\top}\boldsymbol{\Delta}_{g2}^{-1}\mathbf{A}_{(1)gj}\Bigg\}\Bigg]. \qquad \text{(E.32)}\end{aligned}$$

E.5.2 $\boldsymbol{\Delta}_{g2}$ update

Now,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Delta}_{g2}} \mathbb{Q}(\boldsymbol{\vartheta}) = &-\frac{n^*}{2} \sum_{g=1}^{G} \frac{\hat{n}_g}{n_2} \operatorname{tr} \boldsymbol{\Delta}_{g2}^{-1} \mathbf{d} \boldsymbol{\Delta}_{g2} - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n^*_{3:D}} \operatorname{tr}[\boldsymbol{\Delta}_{g2}^{-1} \mathbf{A}_{(1)gj} \boldsymbol{\Delta}_{g1}^{-1} \mathbf{X}_{(1)gij}^{\top} \boldsymbol{\Delta}_{g2}^{-1} \mathbf{d} \boldsymbol{\Delta}_{g2}] \\
&- \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n^*_{3:D}} \operatorname{tr}[\boldsymbol{\Delta}_{g2}^{-1} \mathbf{X}_{(1)gij} \boldsymbol{\Delta}_{g1}^{-1} \mathbf{A}_{(1)gj}^{\top} \boldsymbol{\Delta}_{g2}^{-1} \mathbf{d} \boldsymbol{\Delta}_{g2}] \\
&+ \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} b_{ig} \sum_{j=1}^{n^*_{3:D}} \operatorname{tr}[\boldsymbol{\Delta}_{g2}^{-1} \mathbf{X}_{(1)gij} \boldsymbol{\Delta}_{g1}^{-1} \mathbf{X}_{(1)gij}^{\top} \boldsymbol{\Delta}_{g2}^{-1} \mathbf{d} \boldsymbol{\Delta}_{g2}] \\
&+ \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} \hat{z}_{ig} a_{ig} \sum_{j=1}^{n^*_{3:D}} \operatorname{tr}[\boldsymbol{\Delta}_{g2}^{-1} \mathbf{A}_{(1)gj} \boldsymbol{\Delta}_{g1}^{-1} \mathbf{A}_{(1)gj}^{\top} \boldsymbol{\Delta}_{g2}^{-1} \mathbf{d} \boldsymbol{\Delta}_{g2}] \\
= &-\frac{n^*}{2} \sum_{g=1}^{G} \frac{\hat{n}_g}{n_2} \boldsymbol{\Delta}_{g2}^{-\top} - \frac{1}{2} \sum_{g=1}^{G} \boldsymbol{\Delta}_{g2}^{-\top} \sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{X}_{(1)gij} \boldsymbol{\Delta}_{g1}^{-\top} \mathbf{A}_{(1)gj}^{\top} \boldsymbol{\Delta}_{g2}^{-\top} \\
&- \frac{1}{2} \sum_{g=1}^{G} \boldsymbol{\Delta}_{g2}^{-\top} \sum_{i=1}^{N} \hat{z}_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{A}_{(1)gj} \boldsymbol{\Delta}_{g1}^{-\top} \mathbf{X}_{(1)gij}^{\top} \boldsymbol{\Delta}_{g2}^{-\top} \\
&+ \frac{1}{2} \sum_{g=1}^{G} \boldsymbol{\Delta}_{g2}^{-\top} \sum_{i=1}^{N} \hat{z}_{ig} b_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{X}_{(1)gij} \boldsymbol{\Delta}_{g1}^{-\top} \mathbf{X}_{(1)gij}^{\top} \boldsymbol{\Delta}_{g2}^{-\top} \\
&+ \frac{1}{2} \sum_{g=1}^{G} \boldsymbol{\Delta}_{g2}^{-\top} \sum_{i=1}^{N} \hat{z}_{ig} a_{ig} \sum_{j=1}^{n^*_{3:D}} \mathbf{A}_{(1)gj} \boldsymbol{\Delta}_{g1}^{-\top} \mathbf{A}_{(1)gj}^{\top} \boldsymbol{\Delta}_{g2}^{-\top}
\end{aligned}$$

Note $\mathbf{\Delta}_{g2}^{-\top} = \mathbf{\Delta}_{g2}^{-1}$. Solve for $\mathbf{\Delta}_{g2}$:

$$\begin{aligned}\frac{n^*}{2}\frac{\hat{n}_g}{n_2}\mathbf{\Delta}_{g2}^{-1} = &-\frac{1}{2}\mathbf{\Delta}_{g2}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}\mathbf{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top}\mathbf{\Delta}_{g2}^{-1}\\ &-\frac{1}{2}\mathbf{\Delta}_{g2}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top}\mathbf{\Delta}_{g2}^{-1}\\ &+\frac{1}{2}\mathbf{\Delta}_{g2}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top}\mathbf{\Delta}_{g2}^{-1}\\ &+\frac{1}{2}\mathbf{\Delta}_{g2}^{-1}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}\mathbf{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top}\mathbf{\Delta}_{g2}^{-1}\end{aligned}$$

Multiply both sides by $\mathbf{\Delta}_{g2}$ from the left and the right:

$$\begin{aligned}\frac{n^*}{2}\frac{\hat{n}_g}{n_2}\mathbf{\Delta}_{g2} = &-\frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}\mathbf{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top} - \frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top}\\ &+\frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top} + \frac{1}{2}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}\mathbf{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top}.\end{aligned}$$

The update for $\mathbf{\Delta}_{g2}$, i.e., $\hat{\mathbf{\Delta}}_{g2}$ is:

$$\begin{aligned}\hat{\mathbf{\Delta}}_{g2} = \frac{n_2}{n^*\hat{n}_g}\Bigg[\sum_{i=1}^{N}\hat{z}_{ig}\Bigg\{&b_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top} + a_{ig}\sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}\mathbf{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top}\\ &-\sum_{j=1}^{n^*_{3:D}}\mathbf{X}_{(1)gij}\mathbf{\Delta}_{g1}^{-1}\mathbf{A}_{(1)gj}^{\top} - \sum_{j=1}^{n^*_{3:D}}\mathbf{A}_{(1)gj}\mathbf{\Delta}_{g1}^{-1}\mathbf{X}_{(1)gij}^{\top}\Bigg\}\Bigg]. \qquad \text{(E.33)}\end{aligned}$$

E.5.3 $\boldsymbol{\Delta}_{gl}$ update

$$
\begin{aligned}
\mathbb{Q}(\boldsymbol{\vartheta}) = \ell_{C1} + C_1 + \mathbb{Q}_{C2} + C_2 - \frac{n^*}{2}\sum_{g=1}^{G}\hat{n}_g\sum_{d=1}^{D}\frac{1}{n_d}\log|\boldsymbol{\Delta}_{gd}| \\
+ \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{2:D/l}} \operatorname{tr}\left[\boldsymbol{\Delta}_{g1}^{-1}\left(\mathbf{X}^{l2}_{(1)gij}\right)^\top \boldsymbol{\Delta}_{gl}^{-1}\mathbf{A}^{l2}_{(1)gj}\right] \\
+ \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}\sum_{j=1}^{n^*_{2:D/l}} \operatorname{tr}\left[\boldsymbol{\Delta}_{g1}^{-1}\left(\mathbf{A}^{l2}_{(1)gj}\right)^\top \boldsymbol{\Delta}_{gl}^{-1}\mathbf{X}^{l2}_{(1)gij}\right] \\
- \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}b_{ig}\sum_{j=1}^{n^*_{2:D/l}} \operatorname{tr}\left[\boldsymbol{\Delta}_{g1}^{-1}\left(\mathbf{X}^{l2}_{(1)gij}\right)^\top \boldsymbol{\Delta}_{gl}^{-1}\mathbf{X}^{l2}_{(1)gij}\right] \\
- \frac{1}{2}\sum_{g=1}^{G}\sum_{i=1}^{N}\hat{z}_{ig}a_{ig}\sum_{j=1}^{n^*_{2:D/l}} \operatorname{tr}\left[\boldsymbol{\Delta}_{g1}^{-1}\left(\mathbf{A}^{l2}_{(1)gj}\right)^\top \boldsymbol{\Delta}_{gl}^{-1}\mathbf{A}^{l2}_{(1)gj}\right]. \qquad \text{(E.34)}
\end{aligned}
$$

Using equation E.34, we follow the same steps as we outlined for $\hat{\boldsymbol{\Delta}}_{g2}$ to get the update for $\boldsymbol{\Delta}_{gl}$, i.e., $\hat{\boldsymbol{\Delta}}_{gl}$:

$$
\begin{aligned}
\hat{\boldsymbol{\Delta}}_{gl} = \frac{n_l}{n^*\hat{n}_g}\left[\sum_{i=1}^{N}\hat{z}_{ig}\left\{b_{ig}\sum_{j=1}^{n^*_{2:D/l}}\mathbf{X}^{l2}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-1}\left(\mathbf{X}^{l2}_{(1)gij}\right)^\top + a_{ig}\sum_{j=1}^{n^*_{2:D/l}}\mathbf{A}^{l2}_{(1)gj}\boldsymbol{\Delta}_{g1}^{-1}\left(\mathbf{A}^{l2}_{(1)gj}\right)^\top\right.\right. \\
\left.\left. - \sum_{j=1}^{n^*_{2:D/l}}\mathbf{X}^{l2}_{(1)gij}\boldsymbol{\Delta}_{g1}^{-1}\left(\mathbf{A}^{l2}_{(1)gj}\right)^\top - \sum_{j=1}^{n^*_{2:D/l}}\mathbf{A}^{l2}_{(1)gj}\boldsymbol{\Delta}_{g1}^{-1}\left(\mathbf{X}^{l2}_{(1)gij}\right)^\top\right\}\right]. \\
\text{(E.35)}
\end{aligned}
$$

Bibliography

Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra*, volume 1. Cambridge University Press.

Aitken, A. C. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh*, **45**, 14–22.

Anderlucci, L., Viroli, C., *et al.* (2015). Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics*, **9**(2), 777–800.

Baricz, A. (2010). Turán type inequalities for some probability density functions. *Studia Scientiarum Mathematicarum Hungarica*, **47**, 175–189.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

Benoıt, C. (1924). Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrésa un systeme d'équations linéaires en nombre inférieura celui des inconnues. application de la méthodea la résolution d'un systeme défini d'équations linéaires (procédé du commandant cholesky). *Bulletin Géodésique*, **2**(1), 67–77.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, **59**(1), 65–98.

Blanchard, P., Higham, D. J., and Higham, N. J. (2019). Accurate computation of the log-sum-exp and softmax functions. *arXiv preprint arXiv:1909.03469*.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**, 373–388.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, **71**, 52–78.

Browne, R. P. and McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, **43**(2), 176–198.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, **25**(24), 4279–4292.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.

Celeux, G., Martin-Magniette, M.-L., Maugis-Rabusseau, C., and Raftery, A. E. (2014). Comparing model selection and regularization approaches to variable selection in model-based clustering. *Journal de la Société Française de Statistique*, **155**(2), 57–71.

Cook, R. D. (2018). *An introduction to envelopes: dimension reduction for efficient estimation in multivariate statistics*. John Wiley & Sons, Hoboken, NJ.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**(1), 1–38.

Ding, S. and Cook, R. D. (2018). Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**(2), 387–408.

Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, **64**(2), 105–123.

Fop, M., Murphy, T. B., *et al.* (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, **12**, 18–65.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**(6), 1149–1157.

Gallaugher, M. P. B. and McNicholas, P. D. (2018a). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, **80**, 83–93.

Gallaugher, M. P. B. and McNicholas, P. D. (2018b). Mixtures of skewed matrix variate bilinear factor analyzers. In *Proceedings of the Joint Statistical Meetings*. American Statistical Association, Alexandria, VA. Also available as arXiv:1809.02385.

Gallaugher, M. P. B. and McNicholas, P. D. (2019). Three skewed matrix variate distributions. *Statistics and Probability Letters*, **145**, 103–109.

Gallaugher, M. P. B., Tait, P. A., and McNicholas, P. D. (2021). Four skewed tensor distributions. arXiv:2106.08984.

Gelman, A., Pasarica, C., and Dodhia, R. (2002). Let's practice what we preach: turning tables into graphs. *The American Statistician*, **56**(2), 121–130.

Harris, J. W. and Stöcker, H. (1998). *Handbook of Mathematics and Computational Science*. Springer Science & Business Media.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, 2nd edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, chapter 14, pages 501–527. Springer Series in Statistics. Springer New York.

Hinrich, J. L. and Mørup, M. (2019). Probabilistic tensor train decomposition. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE.

Hoff, P. D. *et al.* (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, **6**(2), 179–196.

Hoff, P. D. *et al.* (2016). Equivariant and scale-free tucker decomposition models. *Bayesian Analysis*, **11**(3), 627–648.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Karlis, D. and Santourian, A. (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, **19**(1), 73–83.

Kolda, T. (2006). Multilinear operators for higher-order decompositions technical report. *Sandia National Laboratories, Albuquerque, NM and Livermore, CA.*

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, **51**(3), 455–500.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical Report TR-2009, Department of Computer Science, University of Toronto, Toronto, Ontario, Canada.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, **112**(519), 1131–1146.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–163. JSTOR.

Mai, Q., Zhang, X., Pan, Y., and Deng, K. (2021). A doubly enhanced em algorithm for model-based tensor clustering. *Journal of the American Statistical Association*, pages 1–15.

Manceur, A. M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics*, **239**, 37–49.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

McLachlan, G. and Peel, D. (2000a). *Finite Mixture Models*. Wiley, New York.

McLachlan, G. J. and Peel, D. (2000b). *Finite Mixture Models*. John Wiley & Sons, New York.

McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference*, **140**(5), 1175–1181.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Chapman & Hall/CRC Press, Boca Raton.

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification*, **33**(3), 331–373.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, **38**(1), 153–168.

McNicholas, P. D. and Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, **142**(5), 1114–1127.

McNicholas, P. D. and Tait, P. A. (2019). *Data Science with Julia*. Chapman and Hall/CRC Press, Boca Raton.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**(3), 711–723.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**(11), 2074–2102.

Murray, P. M., Browne, R. B., and McNicholas, P. D. (2014). Mixtures of skew-t factor analyzers. *Computational Statistics and Data Analysis*, **77**, 326–335.

O'Hagan, A., Murphy, T. B., Gormley, I. C., McNicholas, P. D., and Karlis, D. (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Computational Statistics and Data Analysis*, **93**, 18–30.

Ohlson, M., Ahmad, M. R., and Von Rosen, D. (2013). The multilinear normal distribution: Introduction and some basic properties. *Journal of Multivariate Analysis*, **113**, 37–47.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**(3), 677–690.

Proudfoot, N. A., King-Dowling, S., Cairney, J., Bray, S. R., MacDonald, M. J., and Timmons, B. W. (2019). Physical activity and trajectories of cardiovascular health indicators during early childhood. *Pediatrics*, **144**(1), e20182242.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387–397.

Squire, W. and Trapp, G. (1998). Using complex variables to estimate derivatives of real functions. *SIAM Review*, **40**(1), 110–112.

Tait, P. A., McNicholas, P. D., and Obeid, J. (2020). Clustering higher order data: An application to pediatric multi-variable longitudinal data. arXiv:1907.08566.

Tang, Y., Salakhutdinov, R., and Hinton, G. (2013). Tensor analyzers. In *International Conference on Machine Learning*, pages 163–171.

VanPutte, C. L., Regan, J. L., and Russo, A. F. (2017). *Seeley's Anatomy & Physiology*. McGraw-Hill Education.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA.

Wolfe, J. H. (1965). A computer program for the maximum likelihood analysis of types. Technical Bulletin 65-15, U.S. Naval Personnel Research Activity.

Wu, C. and Hamada, M. (2021). *Experiments: Planning, Analysis, and Optimization*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.