Clustering Matrix Variate Data Using Finite Mixture Models with Component-Wise Regularization

CLUSTERING MATRIX VARIATE DATA USING FINITE MIXTURE MODELS WITH COMPONENT-WISE REGULARIZATION

BY

PETER A. TAIT, Hon.B.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

MASTER OF SCIENCE

Master of Science (2017) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Clustering Matrix Variate Data Using Finite Mixture Models with Component-Wise Regularization
AUTHOR:	Peter A. Tait Hon.B.Sc., (Statistics and Genetics) University of Toronto, Toronto, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	viii, 43

To my parents, Susan and Bob. Thank you for your encouragement and support.

To my son Xavier. Thank you for sacrificing some of your "Dad" time.

Abstract

Matrix variate distributions present a innate way to model random matrices. Realizations of random matrices are created by concurrently observing variables in different locations or at different time points. We use a finite mixture model composed of matrix variate normal densities to cluster matrix variate data. The matrix variate data was generated by accelerometers worn by children in a clinical study conducted at McMaster. Their acceleration along the three planes of motion over the course of seven days, forms their matrix variate data. We use the resulting clusters to verify existing group membership labels derived from a test of motor-skills proficiency used to assess the children's locomotion.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Paul McNicholas. His jovial demeanor, continual guidance, support and encouragement have made my academic experience at McMaster enjoyable and a valuable life experience.

I would also like to show my appreciation to Dr. Roman Viveros-Aguilera, and Dr. Ben Bolker who, along with Dr. McNicholas, were on my examination committee. Their feedback on my work has been valuable and improved my dissertation in many ways. I would like to thank all three for making my thesis defense an enjoyable experience.

This research work was partly supported by the Canada Research Chairs program.

The data we used for our analysis would not be available if it was not for the members of the Child Health & Exercise Medicine Program who led the participant recruitment and data collection for the HOPP study, and prepared the accelerometer data for analysis (N. Proudfoot, N. Di Cristofaro, S. King-Dowling, H. Caldwell). The HOPP Study was funded by the Canadian Institutes of Health Research grants awarded to B.W. Timmons (Award no. 102560).

Finally, I would like to thank my family for their love and support during the last 12 months. Coming back to school after a prolonged hiatus involves many sacrifices, not all my own.

Contents

List of Figures

Chapter 1

Introduction

With the development of new data collection technologies, such as electronic sensors, cell phones and web browsers, there are now many rich sources of multivariate data. Much of this data can be represented as matrices, where the rows can describe different time points or spatial locations and the columns can represent different metrics (e.g. heart rate, acceleration, network speed, .).

Statistical methods that can effectively use matrix variate data have gained in popularity with the rise of these new technologies. One common statistical problem statisticians face is finding sub-populations or clusters in multivariate data. They often turn to finite mixture models to accomplish this goal. These statistical methods have a rich history in the statistical literature (McNicholas, 2016).

More recently, finite mixture models have been extended to matrix variate normal (Viroli, 2011), matrix variate t (Doğru *et al.*, 2016) and skew matrix variate distributions (Gallaugher and McNicholas, 2017a). Mixtures of these distributions have been developed to make sense of this plethora of matrix variate data.

We will look at clustering matrix variate data collected by the HOPP study (Timmons *et al.*, 2012) being conducted at McMaster University. The HOPP investigators are using accelerometer sensors to collect data on children's movement patterns in 3-D space over the course of a week. We will use a finite mixture model composed of matrix variate normal densities to cluster the children into different groups based on their movement patterns and compare them to existing group assignments which are based on a test of the children's motor-skills proficiency.

Chapter 2

Background

2.1 Finite Mixture Models

Cluster analysis is an overarching term used to describe statistical methods that look for grouping structures in data. A detailed overview of clustering methods can be found in Hastie *et al.* (2009). One common method of clustering is referred to as model-based, which assumes an observation $\mathbf{X}$ originates from a population with G separate sub-populations. It is unknown which of the G sub-populations $\mathbf{X}$ comes from.

If the number of sub-populations is finite, the mixture model for the density of an observation $\mathbf{X}$ 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's are called the mixing proportions and have the following two constraints, $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$. The $f_g(\cdot)$'s 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) and McLachlan and Peel (2004).

Classically, the normal mixture model has been used most frequently in practice. Some of the first works using the normal mixture models include Wolfe (1965), Baum *et al.* (1970) and Scott and Symons (1971). This early adoption is due to the Normal distributions attractive mathematical properties. In this case, the $f_g(\mathbf{X}|\boldsymbol{\Theta}_g)$ has a density drawn from the multivariate normal distribution where

$$f_g(\mathbf{X}|\boldsymbol{\Theta}_g) = f_g(\mathbf{X}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}_g|}} \exp\left\{ -\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{X} - \boldsymbol{\mu}_g)\right\}. \quad (2.2)$$

Here $\boldsymbol{\mu}_g$ is the mean vector and $\boldsymbol{\Sigma}_g$ is the covariance matrix of the distribution.

In addition to the multivariate normal distribution, many non-normal distributions have been used to formulate finite mixture models. Some examples include the t-distribution (Peel and McLachlan, 2000), the skew-t distribution (Vrbik and McNicholas, 2012, 2014), the shifted asymmetric Laplace distribution (Franczak *et al.*, 2014), the power exponential distribution (Dang *et al.*, 2015) and the generalized hyperbolic distribution (Browne and McNicholas, 2015).

2.2 Matrix Variate Distributions

Matrix variate distributions have an important place in the theory of multivariate analysis. They are used to model matrix valued random variables as random matrices (Gupta and Nagar, 1999). Many multivariate techniques depend on functions of random matrices such as characteristic roots, determinants and traces. The two

most common matrix variate distributions in use by statisticians are the Wishart distribution (Wishart, 1928), used to model the sample covariance distributions and the matrix variate normal distribution, a generalization of the multivariate normal distribution.

2.3 Matrix Variate Normal Distribution

The matrix variate normal distribution (MVN) is an attractive distribution because it retains the mathematical tractability of the multivariate normal distribution and can be used as a reference distribution for many multivariate events because of guarantees made by the central limit theorem. Associations between the MVN and other matrix variate distributions are outlined by Dawid (1981) and Gupta and Nagar (1999).

A random matrix $\mathscr{X}$, of size $n \times p$ follows the MVN, denoted as $N_{n\times p}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi})$. The distribution parameters consist of three matrices, the location parameter matrix $\mathbf{M}$ and two scale parameter matrices, $\mathbf{\Sigma}$, an $n \times n$ matrix and $\mathbf{\Psi}$, an $p \times p$ matrix.

The density of $\mathbf{X}$, a realization of $\mathscr{X}$, can be written as

$$f(\mathbf{X}|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}) = \frac{1}{(2\pi)^{\frac{np}{2}}}|\mathbf{\Sigma}|^{\frac{n}{2}}|\mathbf{\Psi}|^{\frac{p}{2}} \exp\left\{ -\frac{1}{2}\text{tr}(\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{\Psi}^{-1}(\mathbf{X}-\mathbf{M})^T)\right\}. \quad (2.3)$$

A nice property of this MVN density is that it can be decomposed into the multivariate normal density N_{np} with dimensions np (Gupta and Nagar, 1999) as follows:

$$\mathscr{X} \sim N_{n\times p}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}) \Leftrightarrow \text{vec}(\mathscr{X}) \sim N_{np}(\text{vec}(\mathbf{M}), \mathbf{\Sigma} \otimes \mathbf{\Psi}) \quad (2.4)$$

where vec($\mathbf{M}$) is the vectorization of location parameter matrix and $\otimes$ is the Kronecker

product.

Since $N_{n\times p}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi})$ is a special case of $N_{np}(\text{vec}(\mathbf{M}), \mathbf{\Sigma} \otimes \mathbf{\Psi})$, the mean and variance of $N_{n\times p}(\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi})$ can be expressed as

$$\text{E}(\text{vec}(\mathbf{X})|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}) = \text{vec}(\mathbf{M}) \tag{2.5}$$

$$\text{Var}(\text{vec}(\mathbf{X})|\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}) = \mathbf{\Sigma} \otimes \mathbf{\Psi} \tag{2.6}$$

Where vec is a linear transformation, converting a matrix into a column vector. This stacks the columns of a matrix on top of each other. The Kronecker product of the two matrices results in a block matrix.

2.4 Finite Mixtures of Multivariate Normal distributions

In the context of finite mixture models, we expect that the random matrix $\boldsymbol{\mathscr{X}}$ follows a MVN distribution and has G sub-populations. The density of $\boldsymbol{\mathscr{X}}$ is

$$f(\mathbf{X}|\pi_1, \pi_2, \ldots, \pi_G, \mathbf{\Theta}_1, \mathbf{\Theta}_2, \ldots, \mathbf{\Theta}_G) = \sum_{g=1}^{G} \pi_g f_{\text{MVN}}(\mathbf{X}|\mathbf{M}_g, \mathbf{\Sigma}_g, \mathbf{\Psi}_g) \tag{2.7}$$

where $\mathbf{\Theta}_g = (\mathbf{M}_g, \mathbf{\Sigma}_g, \mathbf{\Psi}_g)$ represents the parameters of the gth MVN and the weights π_g represent the prior probabilities of belonging to each of the sub-populations $g = 1 \ldots G$ (Glanz and Carvalho, 2013). The a posteriori probability of the observed matrix $\mathbf{X}_g$ belongs to the gth component of the mixture is expressed by Bayes theorem

as

$$\tau_{gi} = \mathbb{P}(\mathbf{X}_i|\pi_1, \pi_2, \ldots, \pi_g, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_g) = \frac{\pi_g f_{\text{MVN}}(\mathbf{X}_i|\boldsymbol{\Theta}_g)}{\sum_{h=1}^{G} \pi_h f_{\text{MVN}}(\mathbf{X}_i|\boldsymbol{\Theta}_h)} \tag{2.8}$$

2.5 Covariance Matrix Estimation

The estimation of a covariance matrix or its inverse is of primary importance for many statistical methods. The covariance matrix derived from an n-dimensional random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ is defined as

$$\boldsymbol{\Sigma}_{n\times n} = \text{E}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})' = (\sigma_{ij}) \tag{2.9}$$

where $\boldsymbol{\mu} = \text{E}(\mathbf{Y})$ is the mean vector and σ_{ij} is the variance of random variable Y_i when $i = j$ and the covariance between Y_i and Y_j when $i \neq j$ (Pourahmadi, 2013).

The values of the matrix $\boldsymbol{\Sigma}$ are by default unconstrained. Each variance and covariance are estimated uniquely from the data. This results in many parameters to estimate, especially when n is large. Imposing some structure on the entries of $\boldsymbol{\Sigma}$ reduces the number of parameters to estimate and makes many problems computationally tractable.

Structured covariance matrices can be modeled in two complementary ways: generalized linear models (GLMs) and regularization (Pourahmadi, 2013). The GLM methods use covariates and different link functions to model the covariance matrices. These methods depend on being able to find unconstrained and statistically important re-parametrization of the covariance matrices. They often employ spectral and Cholesky decompositions to find these re-parametrization (Pourahmadi, 1999; Zhang and Leng, 2011; Lee *et al.*, 2017).

Regularization methods have been a heavily researched topic (Bickel and Levina, 2008b; Pourahmadi, 2013). Two types of regularization have received the most attention. The first involves shrinking either the eigenvalues or the eigenvectors of the covariance matrix. The second involves component-wise regularization of the covariance matrix which shrinks the eigenvalues and the eigenvectors simultaneously. The goal is to replace smaller entries in $\mathbf{\Sigma}$ with zero. We used component-wise regularization, in the form of tapering (Bickel and Gel, 2011; McMurry and Politis, 2010) to regularize the $\mathbf{\Psi}$ covariance matrices used in the mixture models.

Chapter 3

Methodology

3.1 Maximum Likelihood Estimation

Suppose we have N independent observed matrices $\mathbf{X}_i$, where $i = 1, 2, \ldots, N$. We want to cluster these N matrices into one of the G groups. The log-likelihood function can be written as

$$l(\boldsymbol{\pi}, \boldsymbol{\Theta} | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \sum_{i=1}^{N} \log \left\{ \sum_{g=1}^{G} \pi_g f_{\text{MVN}}(\mathbf{X} | \mathbf{M}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\Psi}_g) \right\} \tag{3.1}$$

The parameters in equation 3.1 can be estimated using the EM algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 2008). Different EM algorithms have been developed for the MVN (Glanz and Carvalho, 2013) and mixtures of MVN (Viroli, 2011).

The EM algorithm requires an allocation variable $\mathbf{z}$ for the mixture model defined in equation 2.7. $\mathbf{z}$ is a vector of dimension G, giving the component membership of

each matrix $\mathbf{X}_i$. Note that $\mathbf{z}$ follows a multinomial distribution (Viroli, 2011)

$$f(\mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \prod_{g=1}^{G} \pi_g^{\mathbf{z}_g} \tag{3.2}$$

and when $\mathbf{z}_g = 1$

$$f(\mathbf{z}_g = 1|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \pi_g. \tag{3.3}$$

Using $\mathbf{z}$, we can define the complete density as the product of the conditional densities:

$$f(\mathbf{X}, \mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\Theta}) = f(\mathbf{X}|\mathbf{z}_g = 1; \boldsymbol{\pi}, \boldsymbol{\Theta}) f(\mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\Theta}) \tag{3.4}$$

We can then maximize the conditional expectation of $f(\mathbf{X}, \mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\Theta})$ using a fixed set of parameters, $\boldsymbol{\pi}'$ and $\boldsymbol{\Theta}'$. See Viroli (2011) and Glanz and Carvalho (2013) for details.

The maximization is defined as:

$$\underset{\boldsymbol{\pi}, \boldsymbol{\Theta}}{\operatorname{argmax}} E_{\mathbf{z}|\mathbf{X}, \boldsymbol{\pi}', \boldsymbol{\Theta}'} \big[\log f(\mathbf{X}|\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\Theta}) + \log f(\mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\Theta}) \big], \tag{3.5}$$

which is the same as maximizing the likelihood

$$L(\boldsymbol{\pi}, \boldsymbol{\Theta}|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \tau_{gi}) = \sum_{g=1}^{G} \sum_{i=1}^{N} \tau_{gi} \log \big[\pi_g f_{\text{MVN}}(\mathbf{X}_i|\mathbf{M}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\Psi}_g) \big]. \tag{3.6}$$

3.2 EM Algorithm

To find the maximized values of the mixture parameters, $\mathbf{M}_g$, $\mathbf{\Sigma}_g$ and $\mathbf{\Psi}_g$, the following expectation must be iteratively maximized:

$$E_{\mathbf{z}|\mathbf{X},\boldsymbol{\pi}',\boldsymbol{\Theta}'}\left[\sum_{i=1}^{N}\log f(\mathbf{X}_i|\mathbf{z}_i;\boldsymbol{\pi},\boldsymbol{\Theta})\right], \tag{3.7}$$

where $f(\mathbf{X}|\mathbf{z};\boldsymbol{\pi},\boldsymbol{\Theta})$ has the MVN distribution.

The following expression for the expectation is derived by Viroli (2011):

$$E_{\mathbf{z}|\mathbf{X},\boldsymbol{\pi}',\boldsymbol{\Theta}'}\left[\sum_{i=1}^{N}\log f(\mathbf{X}_i|\mathbf{z}_i;\boldsymbol{\pi},\boldsymbol{\Theta})\right] = \sum_{g=1}^{G} f(\mathbf{z}_{ig}|\mathbf{X}_i;\boldsymbol{\pi}',\boldsymbol{\Theta}')\Bigg[-\frac{npN}{2}\log(2\pi) - \frac{pN}{2}\log|\mathbf{\Psi}_g| - \frac{nN}{2}\log|\mathbf{\Sigma}_g| - \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\mathbf{\Psi}_g^{-1}(\mathbf{X}_i-\mathbf{M}_g)\mathbf{\Sigma}_g^{-1}(\mathbf{X}_i-\mathbf{M}_g)^T\Bigg] \tag{3.8}$$

Closed form estimates of the mixture parameters are obtained by taking the first derivatives of equation 3.8 with respect to the individual parameters. They are denoted as follows:

$$\hat{\mathbf{M}}_g = \frac{\sum_{i=1}^{N}\tau_{gi}\mathbf{X}_i}{\sum_{i=1}^{N}\tau_{gi}} \tag{3.9}$$

$$\hat{\mathbf{\Psi}}_g = \frac{\sum_{i=1}^{N}\tau_{gi}(\mathbf{X}_i-\hat{\mathbf{M}}_g)\mathbf{\Sigma}_g^{-1}(\mathbf{X}_i-\hat{\mathbf{M}}_g)^T}{p\sum_{i=1}^{N}\tau_{gi}} \tag{3.10}$$

$$\hat{\mathbf{\Sigma}}_g = \frac{\sum_{i=1}^{N}\tau_{gi}(\mathbf{X}_i-\hat{\mathbf{M}}_g)^T\mathbf{\Psi}_g^{-1}(\mathbf{X}_i-\hat{\mathbf{M}}_g)}{n\sum_{i=1}^{N}\tau_{gi}} \tag{3.11}$$

The mixture weights can be calculate by evaluating equation 3.8 under the constraints $\hat{\pi}_g > 0$ and $\sum_{g=1}^{G} \hat{\pi}_g = 1$ and are expressed as

$$\hat{\pi}_g = \frac{\sum_{i=1}^{N} \tau_{gi}}{N}. \tag{3.12}$$

In the E-step of the EM algorithm, $\mathbb{P}(z_{gi} = 1|\mathbf{X}_i)$ must be computed at each iteration as a function of the current parameters $\boldsymbol{\pi}'$ and $\boldsymbol{\Theta}'$. In the M-step of the algorithm, the two components of equation 3.5 can be maximized individually because their cross-derivatives are equal to zero.

3.2.1 Convergence Criterion

There are many ways of choosing the convergence criteria for an EM algorithm. We used a criterion based on the Aitken acceleration (Aitken, 1926). At iteration t, the Aitken acceleration is defined as

$$a^{(t)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}} \tag{3.13}$$

where $l^{(t)}$ is the observed log likelihood at iteration t. We use $a^{(t)}$ to calculate an asymptotic estimate of the log-likelihood at iteration $t+1$. This asymptotic estimate is defined as:

$$l_{\infty}^{(t+1)} = l^{(t)} + \frac{1}{1 - a^{(t)}}(l^{(t+1)} - l^{(t)}) \tag{3.14}$$

We stop the EM algorithm when $l_{\infty}^{(t+1)} - l^{(t)} < \epsilon$ as described in McNicholas *et al.* (2010). This criteria is used because the likelihood can flatten out before increasing again. The estimate $l_{\infty}^{(t+1)}$, is used to evaluate if the likelihood will ever increase in

future iterations.

3.2.2 Group Selection

The number of groups, G in a clustering problem is rarely known apriori. Selecting a sufficient number of groups is important to get right. In this work, we used the Bayesian information criteria (BIC) (Schwarz, 1978) to select G. The BIC is defined as follows:

$$\text{BIC} = 2\hat{l} - \rho \log N \tag{3.15}$$

where $\hat{l}$ is the estimated log likelihood, ρ is the number of free parameters in the model and N is the number of observations.

3.3 Covariance Estimation

As noted by Banfield and Raftery (1993), each covariance matrix $\mathbf{\Sigma}_g$ generated from a G component mixture model, modeling a p-dimensional random variable has $\frac{p(p+1)}{2}$ free parameters. In the interest of parsimony, cluster constraints on the Eigen-decomposition of the covariance matrices are introduced by Banfield and Raftery (1993) and Celeux and Govaert (1995).

The Eigen-decomposition takes the form

$$\mathbf{\Sigma}_g = \lambda_g \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g' \tag{3.16}$$

Table 3.1: GPCMs used in the EM algorithm

Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_g$	Free Parameters
EEE	Equal	Equal	Equal	$\lambda\boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}'$	$\frac{p(p+1)}{2}$
VVV	Variable	Variable	Variable	$\lambda_g\boldsymbol{\Gamma}_g\boldsymbol{\Delta}_g\boldsymbol{\Gamma}_g'$	$\frac{Gp(p+1)}{2}$

where $\boldsymbol{\Gamma}_g$ is the matrix of Eigenvectors, $\boldsymbol{\Delta}_g$ is a diagonal matrix containing the normalized Eigenvalues in decreasing order and $\lambda_g = |\boldsymbol{\Sigma}_g|^{\frac{1}{p}}$.

Following the terminology in Celeux and Govaert (1995), we used the Gaussian parsimonious clustering models (GPCM) defined in Table 3.1 to model the covariance matrices, $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\Psi}_g$ in the EM algorithm described above.

It should be noted that, when modeling sequential observations, $\boldsymbol{\Sigma}$ can be represented by a modified Cholesky decomposition (Pourahmadi, 1999, 2000) of the form:

$$\mathbf{D} = \mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' \qquad \text{or equivalently} \qquad \boldsymbol{\Sigma}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T} \tag{3.17}$$

where $\mathbf{T}$ is a unique unit lower triangular matrix and $\mathbf{D}$ is a unique diagonal matrix with positive diagonal elements. As was pointed out in Pourahmadi (1999), $\mathbf{T}$ and $\mathbf{D}$ can be interpreted as generalized auto-regressive parameters and innovation variances.

A family of eight Gaussian mixture models is developed by McNicholas and Murphy (2010) called the Cholesky-decomposed Gaussian mixture model (CDGMM). Each member of the CDGMM family has an interpretation for longitudinal data and belongs to the GLM family of covariance models referenced in Section 2.5. When $\mathbf{T}_g = \mathbf{T}$, the auto-regressive relationship between time points is the same for each group. When $\mathbf{D}_g = \mathbf{D}$, the variability at each time point is the same across groups. It can be shown that models in the CDGMM family have equivalent models in the GPCM family (McNicholas, 2016). These equivalent models are given in Table 3.2.

Table 3.2: GPCM and equivalent CDGMMs

GPCM	CDGMM	$\mathbf{T}_g$	$\mathbf{D}_g$	Free Parameters
EEE	EEA	Equal	Equal / Anisotrophic	$\frac{p(p-1)}{2} + p$
VVV	VVA	Variable	Variable / Anisotrophic	$G[\frac{p(p-1)}{2}] + Gp$

Despite these equivalencies, the GPCM models do not explicitly account for the longitudinal nature of the data.

3.3.1 Tapering

The component-wise regularization, in the form of tapering, starts with the diagonal elements of $\mathbf{\Sigma}_{p\times p}$, and successively adds sub-diagonals if the data determines its warranted. Tapering gradually shrinks the off diagonal elements to zero. Tapering requires the variables that make up the covariance matrix to have a natural ordering, making it appropriate for longitudinal or time-series data.

The tapered estimate of the sample covariance matrix is defined as

$$\mathbf{S}_W = \mathbf{S} * \mathbf{W} = (s_{ij} w_{ij}) \tag{3.18}$$

where $*$ denotes Schur (coordinate-wise) matrix multiplication. A frequently used weight matrix $\mathbf{W}$, called trapezoidal $\mathbf{W}$ is popular in the time series literature (Bickel and Gel, 2011) and is defined as:

$$w_{ij} = \begin{cases} 1, & \text{if } |i-j| \leq l_h, \\ 2 - \frac{|i-j|}{l_h}, & \text{if } l_h < |i-j| < l, \\ 0, & \text{otherwise.} \end{cases} \tag{3.19}$$

where l is the tapering parameter ranging from $0 \leq l \leq p$ and $l_h = \frac{l}{2}$.

The effectiveness of tapering depends on the choice of l. l is usually chosen by a a resampling scheme like k-fold cross-validation or subsampling (Bickel and Levina, 2008b). We used the subsampling procedure described by Bickel and Levina (2008a) to choose l. The sample is split randomly into two chunks of size $n_1 = n(1 - \frac{1}{\log n})$ and $n_2 = \frac{n}{logn}$, where $\mathbf{\Sigma}$ is of dimension $n \times n$. This is repeated k times. We let $\hat{\mathbf{\Sigma}}_{1,v}$ and $\hat{\mathbf{\Sigma}}_{2,v}$ be the empirical covariance matrices based on samples of size n_1 and n_2 from the vth split. Using these covariance matrices, we minimize the risk, $\hat{R}(l)$ in the following equation:

$$\hat{R}(l) = \frac{1}{k} \sum_{v=1}^{k} \parallel T_l(\hat{\mathbf{\Sigma}}_{1,v}) - \hat{\mathbf{\Sigma}}_{2,v} \parallel_F^2 \tag{3.20}$$

where $T_l()$ is the tapered covariance matrix with tapering parameter l and $\|\|_F^2$ is the squared Frobenius matrix norm of the difference between the tapered and un-tapered covariance matrices. This is repeated for different values of l and the value of l with the smallest $\hat{R}(l)$ is used in the mixture model.

3.4 The Julia Programming Language

We have used the Julia language (Bezanson *et al.*, 2012) to implement all the methods described in Section 3. Typically statisticians choose the R language (R Core Team, 2017) to implement their algorithms. We have deviated from the norm here because we felt Julia offered a number of advantages over R.

Julia is a language specifically designed for numerical computing. It is a dynamic language which checks data types, modifies objects and functions at run-time and not

compile time, making the users programming experience similar to R. Despite this, it has performance approaching statically typed languages like C and FORTRAN. Its speed at doing numerical computations is within a factor of 2 relative to optimized C code and an order of magnitude faster than R.

Julia has extensive mathematical function libraries and does not require wrappers to call existing C or FORTRAN code. Programmers do not have to vectorize code for performance like they do in R. Julia was designed for parallel and distributed computing from the ground up, making it ideal for implementing algorithms that will be used on large data sets.

The R language has many attractive features including a plethora of packages geared towards statisticians, built in support for NA's, a large user community and a very mature development environment (RStudio).

Julia is a more desirable choice than R when the developer requires fast run time speed, is implementing an algorithm from scratch or the statistical model would benefit from a parallel or distributed implementation.

Chapter 4

Analyses

4.1 HOPP Study

The motivating clinical study for this work is called HOPP (Timmons *et al.*, 2012). HOPP stands for Health Outcomes and Physical activity in Preschoolers. It is the only Canadian longitudinal study of preschool children to examine the relationships between physical activity, fitness, nutrition and health outcomes. The study is following 414 children for three years. The children are tested once per year over the course of the study.

After each study visit, the children wear accelerometer belts for seven consecutive days to measure their physical activity. The accelerometer is only removed for sleeping or when the child goes in the water. Every 3 seconds, the accelerometer measures how the child accelerates in the three planes of motion:

- Vertical (Axis 1)
- Anteroposterior or Forward-Backward (Axis 2)

- Mediolateral or Side to side (Axis 3)

A forth measurement, the vector magnitude (vm), is included in the analysis along with the three axis measurements. It is defined as follows:

$$VM = \sqrt{\text{axis1}^2 + \text{axis2}^2 + \text{axis3}^2}. \tag{4.1}$$

4.1.1 HOPP Data

The HOPP study data we analyzed was a subset of the main study. It included the accelerometer data from 49 children taken from the third year of the study. These children were divided into two groups, based a motor-skills assessment called the Bruinitisky-Oserestky test of motor proficiency (BOT-2) (Cools *et al.*, 2009; Timmons *et al.*, 2012). This is a 14-item test that gives is a composite score from 4 areas: fine manual control, manual coordination, body coordination, and strength and agility. The composite score is converted into a sex- age-specific percentile, where the included participants were either < 15th percentile (which is indicative of a motor deficiency) or > 80th percentile for motor skills. The two groups consisted of 25 children determined to have Normal motor-skills and 24 determined to have Abnormal motor-skills. The researchers were interested in seeing if their groupings were supported by the accelerometer data.

The HOPP data has measurements defined at each 3 second interval over the course of the 7 day measurement period. This allows us some latitude in how we aggregate the data for analysis. After consulting with the HOPP researchers, we decided on aggregating each childs measurements in two distinct ways. The first aggregation scheme was by monitoring day. This results in between 6 and 8 time points per child.

The vm and axis values were summed for each day. The second aggregation scheme involved aggregating the measurements by each hour the accelerometer was worn.

The three axis measurements and the vector magnitude captured over the course of the monitoring period, form each childs matrix variate data. The rows of the matrix $\mathbf{X}$ are the time points being used, in this case days or hours. The columns of $\mathbf{X}$ are the four measurements. There are 49 $\mathbf{X}$'s in total, one per child.

Despite the fact that the children were supposed to wear their accelerometers for seven consecutive days, there was a lot of variability in their wear time. The variability in the days worn is illustrated in Figure 4.1. The children are ordered by their median wear time, which is represented by the circles in the plot. The lines range from the minimum to the maximum wear day. The majority of the children wear their accelerometers for seven consecutive days. There are some outliers, like child A11, who wore their accelerometer on days 1,2,8,9,10,13,14 and 18. While the sequence has eight total days, there are wide gaps between successive days. This could potentially pose a problem when estimating $\mathbf{\Psi}$, the between days covariance matrix.

When aggregating the four metrics by wear hour, the number of measurement times between children and the size of the gaps between wear times were even more pronounced. The children's hours of wear time ranged from 107 to 492 hours. In order to reduce this variability, hours was changed to percent of the maximum hour worn. This reduced the median number of unique time points from 89 to 59 per child. The resulting variability is displayed in Figure 4.2. The gaps between successive wear times are still evident in the children's data (e.g.: A11 vs A10), but the overall number of time points is much more uniform.

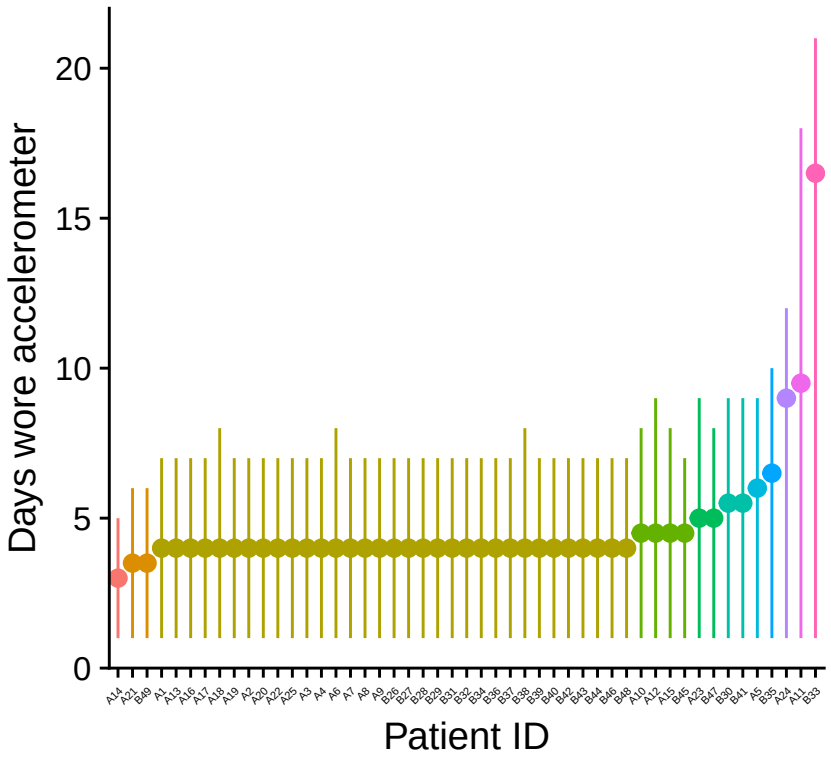

Figure 4.1: Days worn accelerometer. Variability across children

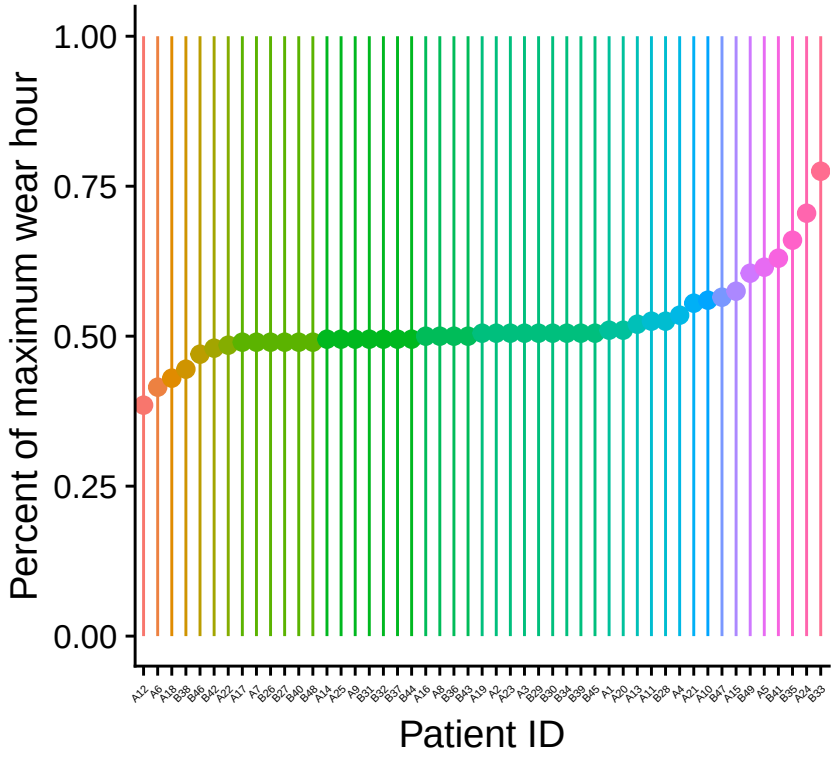

Figure 4.2: Percent of maximum wear hour. Variability across children

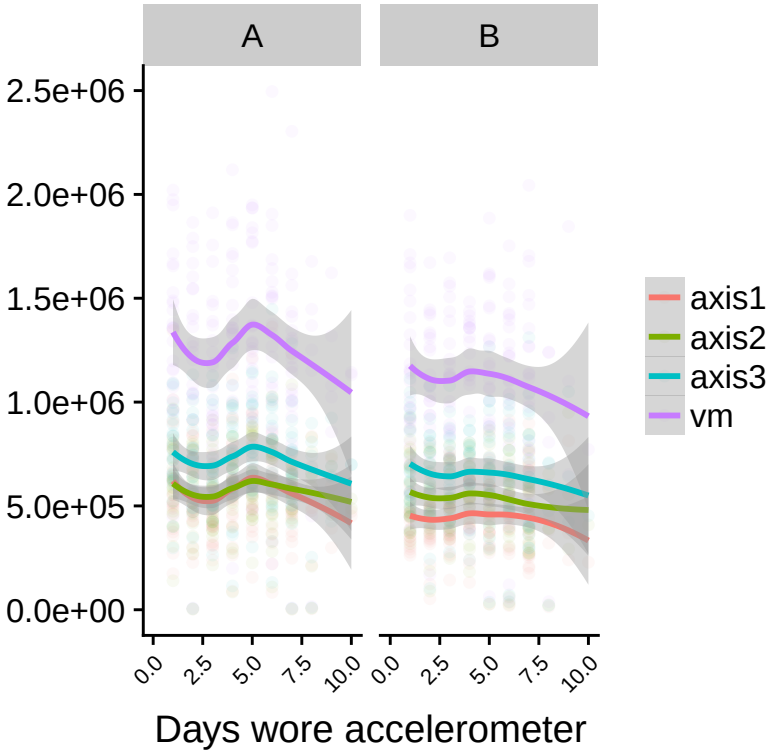

Figure 4.3: Smoothed mean trends across days

Next we look at the mean trends of the four metrics over time. Figure 4.3 shows the trends for the two investigator identified groups by wear day. Each line corresponds to a smoothed trend and the grey area surrounding the line corresponds to a 95% confidence band. Group A has a peak in the four trends at day 5 which does not occur in the group B. In group B, the axis 2 trend is clearly distinguishable from the axis 1 trend, although the two confidence bands mostly overlap. The overlap indicates there is likely not a statistically significant difference (at the 5% level) in the curves at a given time point. This difference in axis 1 and 2 trends is not evident in group A. Qualitatively, the groups have different patterns of wear across the days.

The mean trends across wear hours is displayed in Figure 4.4. The trends appear

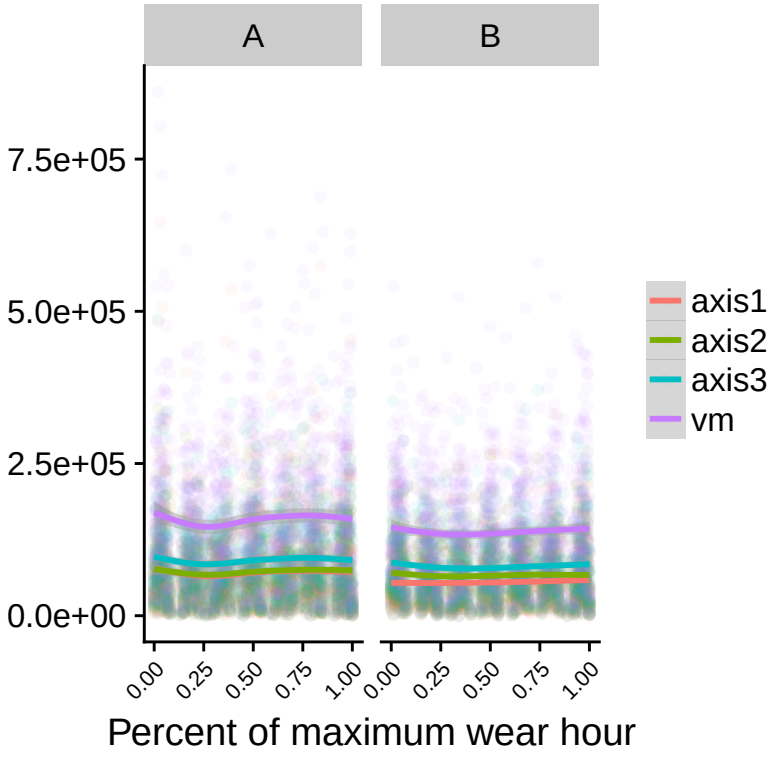

Figure 4.4: Smoothed mean trends across percent maximum wear hour

to be shifted up in magnitude in group A as compared to group B. The peak we observed in Figure 4.3 is no longer evident in group A's trend line. On this time scale, the two groups mean trends look very similar.

4.2 Clustering of the HOPP Data

We used a finite mixture model of MVN's to investigate the number of clusters/sub-populations in the HOPP data. The clustering algorithm was initialized using k-means clustering values (McNicholas, 2016). We looked for between 2 and 5 groups.

We chose to use the two GPCMs, VVV and EEE listed in Table 3.1 because of their

ease of implementation and their connections to the CDGMMs listed in Table 3.2. Tapering the $\mathbf{\Psi}$ covariance matrices would impose additional structural constraints on them, freeing us from using the other members of the CDGMM family to model the covariances between time points. When tapering was used, it was trained via subsampling using 5 distinct subsamples for each value of l and applied to all $\mathbf{\Psi}_g$'s used in the model. Tapering was not used on the $\mathbf{\Sigma}_g$ matrices, as the variables do not have a natural ordering.

The groups produced by the clustering algorithm were compared to the existing groups via the adjusted Rand index. The unadjusted Rand index (RI) is the ratio of the pairs agreement to the total number of pairs (Rand, 1971). A RI of one indicates perfect agreement. Chance agreement can enlarge the RI, making it problematic in some cases. The adjusted RI (ARI) was developed to overcome this problem (Hubert and Arabie, 1985). The ARI has an expected value of zero when the classification is purely random and retains the property of being equal to 1 when there is complete class agreement.

4.2.1 Imputation Strategy

Given that many children had a different number of time measurements / rows in their $\mathbf{X}$ matrix, an imputation strategy was required to ensure all the $\mathbf{X}$'s had the same number of rows, as this is required by the MVN clustering algorithm detailed in Chapter 3.

The general imputation strategy consisted of the following steps:

1. Decide on the number of time measurements to include in the analysis
2. Calculate the overall median vm and axis values per child

3. Restrict the number of rows in each X to the value in 1.

4. Replace any missing rows with the median values from 2.

In step 1, we used 8 days and 45 hours respectively. We chose these numbers because they kept the number of imputed values low (4 total or 8.2%) and still allowed us to look at a sizable number of time points.

Admittedly this is not an optimal imputation strategy. It is likely very conservative, does not account for the uncertainty related to the imputation and could attenuate any correlations between the variables or time points. A superior strategy involves modeling the covariance matrix using the GLM framework mentioned in Chapter 2. A generalized EM algorithm is developed by (Huang *et al.*, 2012) in the context of the modified Cholesky decomposition, to evaluate the maximum likelihood estimates of the GLMs parameters.

4.2.2 Results for the Days Data

Clustering the matrix variate data over days resulted in the discovery of two groups in the data. Figure 4.5 shows the two group solution as the clear favorite based on the size of its BIC value. Dasgupta and Raftery (1998) suggest that a difference of 10 between BIC values constitutes very strong evidence in favor of the model with the larger BIC value. These authors defined the BIC to be -1 times our definition in equation 3.15. This plot is constructed for the results of the VVV:VVV model with different group sizes. Here the first VVV corresponds to the covariance structure of the $\mathbf{\Sigma}_g$ matrix while the second one corresponds to the structure of the $\mathbf{\Psi}_g$ matrix.

Figure 4.6 compares different covariance structures for the two group solution. The smallest BIC values belong to the model with same $\mathbf{\Sigma}_g$ structure across the groups

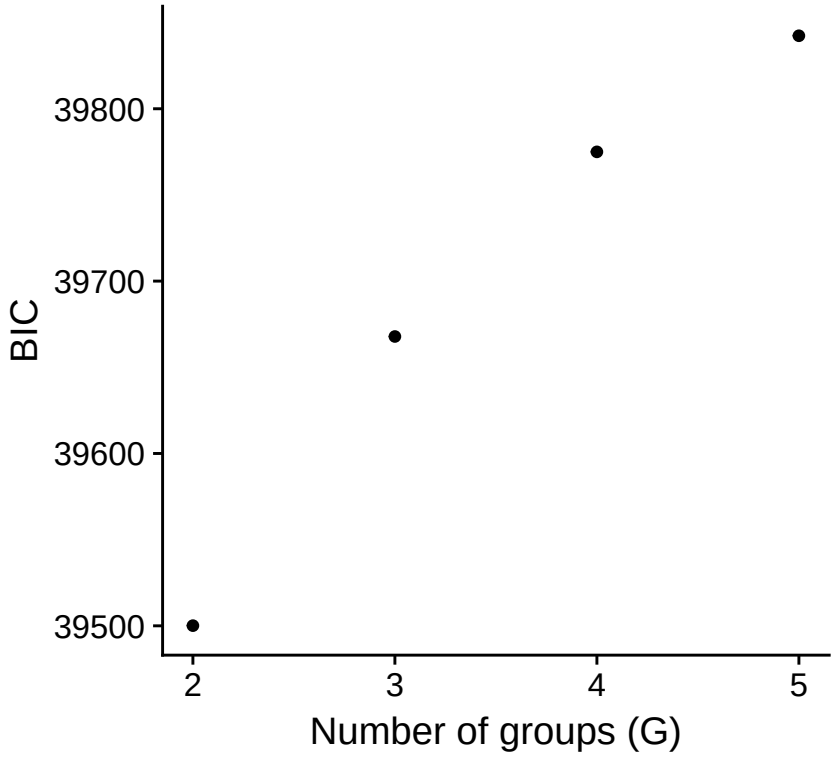

Figure 4.5: BICs for different group sizes using the days data

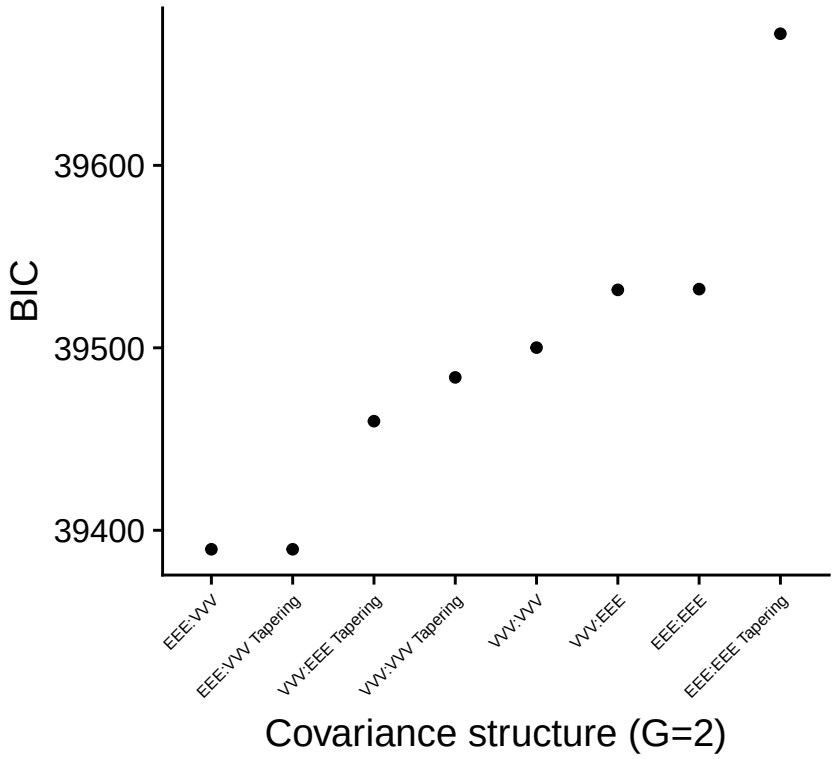

Figure 4.6: BICs for different covariance structures using the days data

and different $\mathbf{\Psi}_g$ structures between the groups. In this configuration, tapering did not improve the model fit.

The cluster labels for the EEE:VVV with tapering model are compared to the investigators groups in Table 4.1. There are 21 of the 49 entries in the off diagonal cells, indicating poor agreement. The RI is 0.5 but when compared to the ARI, which is equal to 0.002, it is clear that it is inflated by random agreements. The ARI indicates that we have poor pairwise agreement between the groups and the cluster labels.

This lack of agreement is interesting. The results suggest that there are two groups in the data but they do not correspond to the groups identified by the investigators.

Table 4.1: Classification performance of the EEE:VVV tapering model for the days data

	Actual Groups	
Cluster Labels	1	2
1	18	14
2	7	10

Clearly the accelerometer data is capturing information that is not being gathered by the BOT-2 assessment.

The mean trends of the two groups identified by the model are illustrated in Figure 4.7. The EEE:VVV with tapering model is displayed because the tapering smoothed the mean curves, making their differences more apparent. The results suggest that group 1 is larger because of the narrower confidence bands, it indicates an increase in the variables over the closing days of monitoring and the curves are shifted down compared to group 2.

In-order to visualize the effect of tapering on the $\mathbf{\Psi}$ matrices, we converted them into correlation matrices ($\boldsymbol{\rho}$) using the well known matrix identity $\boldsymbol{\rho} = D'\mathbf{\Psi}D'$, where $D = \sqrt{\text{diag}(\mathbf{\Psi})}$. Figure 4.8 displays the $\mathbf{\Psi}$ correlation matrices for the VVV:VVV and EEE:VVV Tapering model. It is clear that tapering reduced the size of the off diagonal elements in the later model. When we examine the eighth day, we see that the magnitude of the correlation does not decrease with time as we would expect. This is due to our single imputation strategy. The eighth day contains the majority of the imputed values, tainting the correlations between it and the other time points.

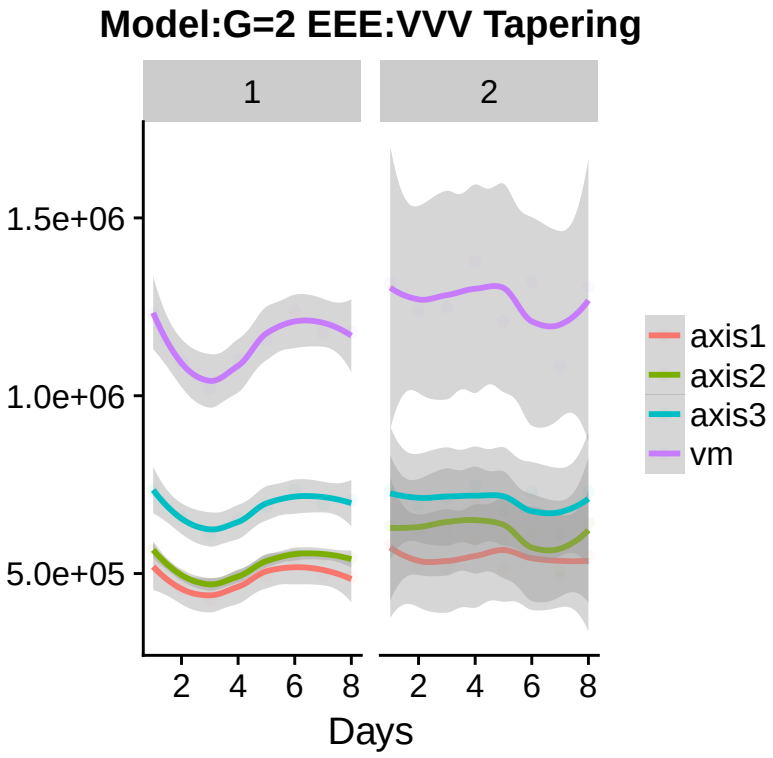

Figure 4.7: Clustering mean trends for the days data

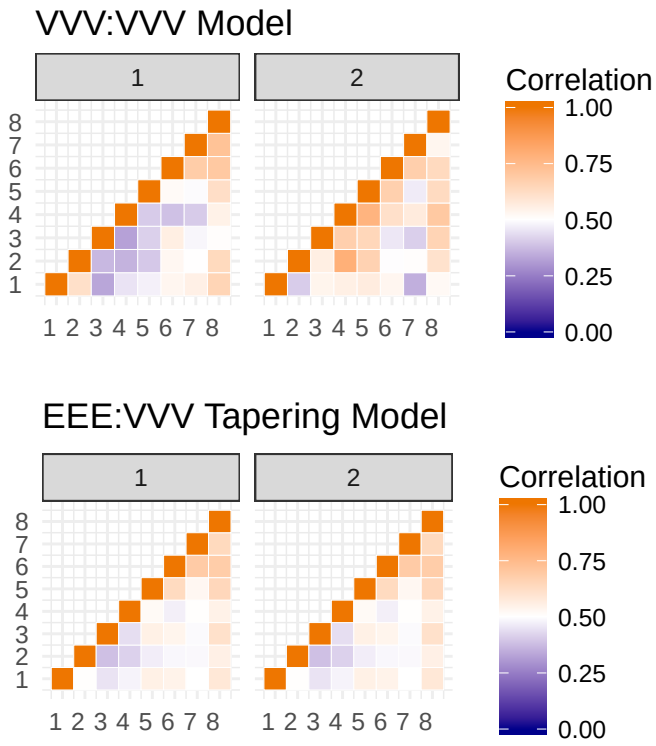

Figure 4.8: $\mathbf{\Psi}$ correlation matrix comparison for the days data

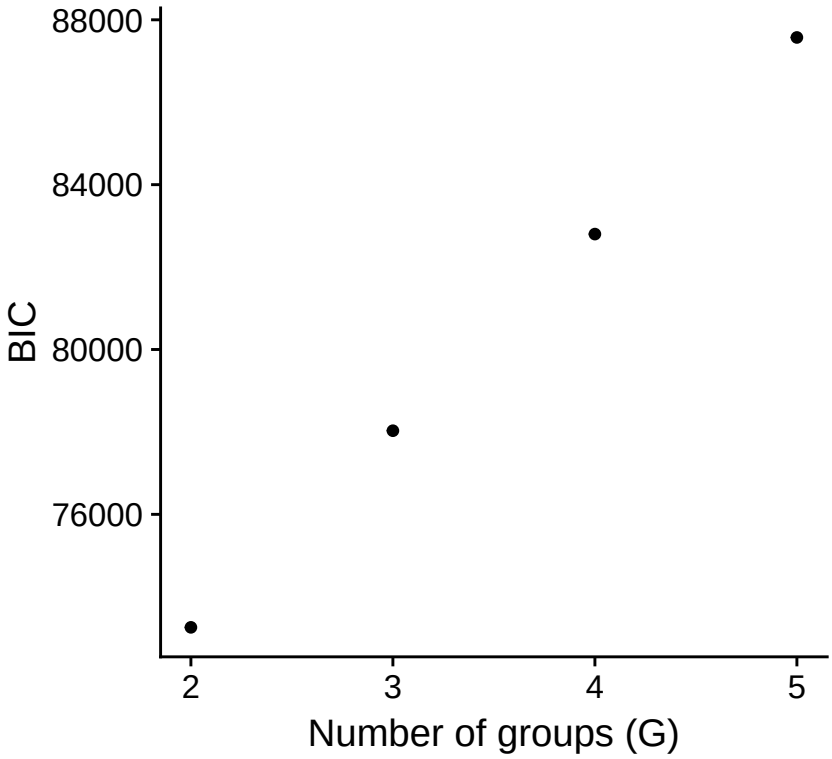

Figure 4.9: BICs for different group sizes using the percent maximum wear hours data

4.2.3 Results for the Hours Data

Clustering the matrix variate data over the percent maximum wear hours resulted in the discovery of two groups in the data. Figure 4.9 indicates that the two group solution is the overwhelming favorite based on the size of its BIC value. In contrast to the days data, Figure 4.10 indicates an EEE model of $\mathbf{\Psi}_g$ was preferred.

The results in Table 4.2 indicate that the mixture model actually found only one group in the data. The RI was 0.49 and the ARI was equal to 0.0. This contradicts the results found in the days data. We feel that this contradiction can be explained by the model fit being poor and is not due to any actual homogeneity in the study

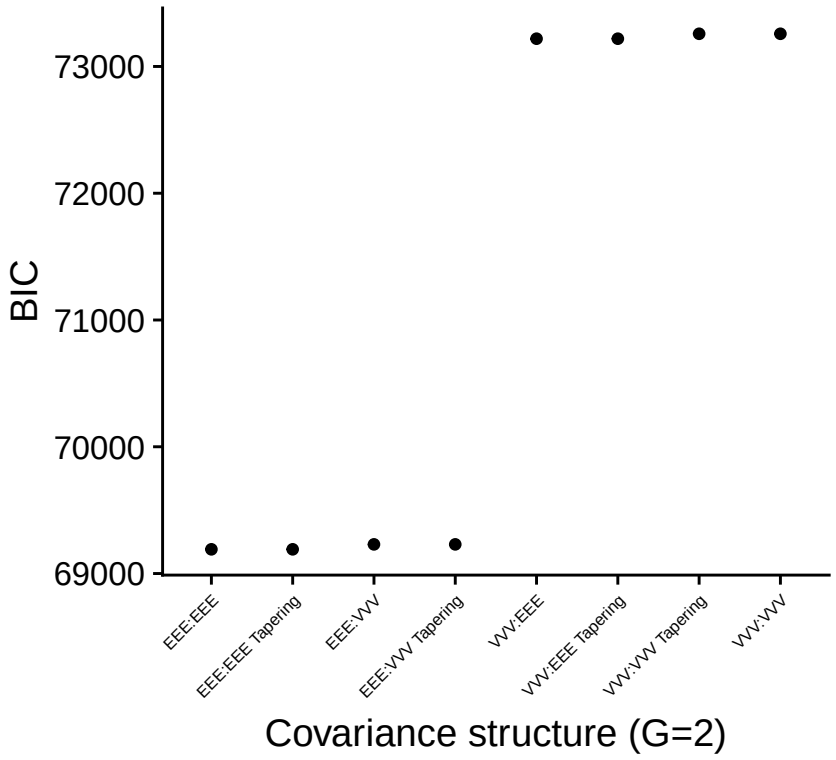

Figure 4.10: BICs for different covariance structures using the percent maximum wear hours data

Table 4.2: Classification performance of the EEE:EEE tapering model for the percent maximum wear hours data

	Actual Groups	
Cluster Labels	1	2
1	25	24

subjects movement patterns. The two group model did not capture any variation between the clusters within the hours or the four variables. This can be seen by the EEE:EEE model having the lowest BIC value. Better models of the $\mathbf{\Psi}$ matrix, that could capture the non consecutive nature of the measurement times would improve the model fit and likely change the results as well.

The mean trends for the single group is displayed in Figure 4.11. The axis trends are very linear, with the axis 2 and axis 3 trends being shifted upwards relative to the axis 1 trend. This indicates the children are moving side to side the most, followed by forwards and backwards and vertically.

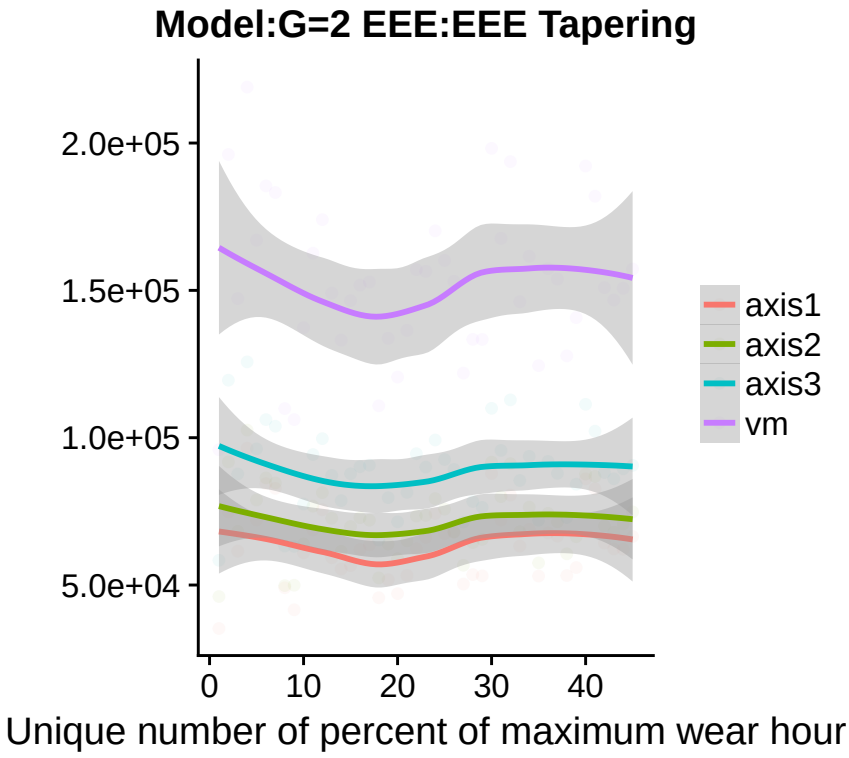

Figure 4.11: Mean trends from model using the percent maximum wear hours data

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Using a finite mixture model consisting of MVNs, we were able to confirm that the HOPP study data we analyzed did indeed consist of two groups of children. These two groups did not coincide with the group labels provided by the study investigators. This suggests that the accelerometer data does provide additional information, not captured by the BOT-2 assessment that should be used to identify children with abnormal movement patterns.

5.2 Future Work

This work suggests many avenues for future investigation. The first would be covariance estimation. The covariance structures we used could be improved by taking advantage of the GLM framework described in (Pourahmadi, 2013). This would allow us to estimate missing entries in the covariance matrices using an EM algorithm

(Huang *et al.*, 2012) as opposed to using single imputation of the raw data as we did here. In addition, using the modified Cholesky decomposition would allow us to model autoregressive (Pourahmadi, 1999), moving average (Zhang and Leng, 2011) and ARMA (Lee *et al.*, 2017) covariance structures. Finally the GLM framework could help us model the nonconsecutive time points we observed in the accelerometer data (Pan and MacKenzie, 2006; Zhang *et al.*, 2015).

A second topic to investigate is the over penalization of the mixture models BIC values when tapering is used. Tapering is reducing the number of free parameters in the model by setting some entries of $\boldsymbol{\Psi}$ to zero. This is not currently being captured by ρ in equation 3.15. Some work has been done on modifying the BIC value in the context of banding the covariance matrix (Leng and Li, 2011) but to our knowledge, this has not been extended to tapering.

A third avenue of investigation could be looking at mixtures of non-Normal matrix variate distributions. Mixtures of matrix variate t distributions (Doğru *et al.*, 2016) offer a model of the data that is more robust to outliers due to its heavy tails. Mixtures of skewed matrix variate distributions (Gallaugher and McNicholas, 2017a) could offer similar robustness to outliers (Gallaugher and McNicholas, 2017b), with the additional benefit of modeling asymmetric (e.g. fatter) clusters.

Finally, we could investigate how to incorporate additional covariates (e.g.: age, gender, etc.) into the models (Anderlucci and Viroli, 2015) and modeling the component means (McNicholas and Subedi, 2012) using linear, cubic and non-parametric trends.

Bibliography

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

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**(3), 803–821.

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.

Bezanson, J., Karpinski, S., Shah, V. B., and Edelman, A. (2012). Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*.

Bickel, P. J. and Gel, Y. R. (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(5), 711–728.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, **36**(6), 2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, **36**(1), 199–227.

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

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

Cools, W., De Martelaer, K., Samaey, C., and Andries, C. (2009). Movement skill assessment of typically developing preschool children: A review of seven movement skill assessment tools. *Journal of sports science & medicine*, **8**(2), 154.

Dang, U. J., Browne, R. P., and McNicholas, P. D. (2015). Mixtures of multivariate power exponential distributions. *Biometrics*, **71**(4), 1081–1089.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, **93**, 294–302.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a bayesian application. *Biometrika*, **68**(1), 265–274.

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.

Doğru, F. Z., Bulut, Y. M., and Arslan, O. (2016). Finite mixtures of matrix variate t distributions. *Gazi University Journal of Science*, **29**(2), 335–341.

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. and McNicholas, P. D. (2017a). Finite mixtures of skewed matrix variate distributions. *arXiv preprint arXiv:1703.08882v2*.

Gallaugher, M. P. and McNicholas, P. D. (2017b). A matrix variate skew-t distribution. *Stat*, **6**(1), 160–170.

Glanz, H. and Carvalho, L. (2013). An Expectation-Maximization Algorithm for the Matrix Normal Distribution. *ArXiv e-prints*.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix variate distributions*, volume 104. CRC Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York.

Huang, J. Z., Chen, M., Maadooliat, M., and Pourahmadi, M. (2012). A cautionary note on generalized linear models for covariance of unbalanced longitudinal data. *Journal of Statistical Planning and Inference*, **142**(3), 743–751.

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

Lee, K., Baek, C., and Daniels, M. J. (2017). Arma cholesky factor models for the covariance matrix of linear models. *Computational Statistics & Data Analysis*, **115**, 267–280.

Leng, C. and Li, B. (2011). Forward adaptive banding for estimating large covariance matrices. *Biometrika*, **98**(4), 821–830.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley, New York, 2 edition.

McMurry, T. L. and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, **31**(6), 471–482.

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

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. *The 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., 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.

Pan, J. and MacKenzie, G. (2006). Regression models for covariance structures in longitudinal studies. *Statistical Modelling*, **6**, 43–57.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

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

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**(2), 425–435.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation: with High-Dimensional Data*. John Wiley and Sons, Inc., New York.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

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.

Timmons, B. W., Proudfoot, N. A., MacDonald, M. J., Bray, S. R., and Cairney, J. (2012). The health outcomes and physical activity in preschoolers (hopp) study: rationale and design. *BMC Public Health*, **12**(1), 284–291.

Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**(4), 511–522.

Vrbik, I. and McNicholas, P. D. (2012). Analytic calculations for the EM algorithm for multivariate skew-t mixture models. *Statistics and Probability Letters*, **82**(6), 1169–1174.

Vrbik, I. and McNicholas, P. D. (2014). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics and Data Analysis*, **71**, 196–210.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, **20A**(1-2), 32–52.

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

Zhang, W. and Leng, C. (2011). A moving average cholesky factor model in covariance modelling for longitudinal data. *Biometrika*, **99**(1), 141–150.

Zhang, W., Leng, C., and Tang, C. Y. (2015). A joint modelling approach for longitudinal studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**(1), 219–238.