# Extending Growth Mixture Models and Handling Missing Values via Mixtures of Non-Elliptical Distributions

# EXTENDING GROWTH MIXTURE MODELS AND HANDLING MISSING VALUES VIA MIXTURES OF NON-ELLIPTICAL DISTRIBUTIONS

BY

YUHONG WEI

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

Science Doctor of Philosophy (2017) McMaster University

(Department of Mathematics and Statistics) Hamilton, Ontario, Canada

| | |
|---|---|
| TITLE: | Extending Growth Mixture Models and Handling Missing Values via Mixtures of Non-Elliptical Distributions |
| AUTHOR: | Yuhong Wei<br>Ph.D., (Mathematics and Statistics)<br>McMaster University, Hamilton, Canada |
| SUPERVISOR: | Dr. Paul D. McNicholas |
| NUMBER OF PAGES: | xvi, 149 |

*To my husband Hailiang,*

*and my children,*

*Kaley & Emily*

# Abstract

Growth mixture models (GMMs) are used to model intra-individual change and inter-individual differences in change and to detect underlying group structure in longitudinal studies. Regularly, these models are fitted under the assumption of normality, an assumption that is frequently invalid. To this end, this thesis focuses on the development of novel non-elliptical growth mixture models to better fit real data. Two non-elliptical growth mixture models, via the multivariate skew-t distribution and the generalized hyperbolic distribution, are developed and applied to simulated and real data. Furthermore, these two non-elliptical growth mixture models are extended to accommodate missing values, which are near-ubiquitous in real data.

Recently, finite mixtures of non-elliptical distributions have flourished and facilitated the flexible clustering of the data featuring longer tails and asymmetry. However, in practice, real data often have missing values, and so work in this direction is also pursued. A novel approach, via mixtures of the generalized hyperbolic distribution and mixtures of the multivariate skew-t distributions, is presented to handle missing values in mixture model-based clustering context. To increase parsimony, families of mixture models have been developed by imposing constraints on the component scale matrices whenever missing data occur. Next, a mixture of generalized hyperbolic factor analyzers model is also proposed to cluster high-dimensional data

with different patterns of missing values. Two missingness indicator matrices are also introduced to ease the computational burden. The algorithms used for parameter estimation are presented, and the performance of the methods is illustrated on simulated and real data.

# Acknowledgements

First and foremost, I wish to express my sincerest gratitude to my supervisor, Dr. Paul D. McNicholas, for his support, guidance, understanding and great patience throughout the course of my PhD. I could not have imagined having a better advisor and mentor for my PhD study. I would also like to thank Dr. Douglas L. Steinley and Dr. Emilie Shireman for their collaboration and help in my research.

Thank you to Dr. Benjamin M. Bolker and Dr. Román Viveros-Aguilera for serving as members of my supervisory committee, to Dr. Peter MacDonald for chairing my thesis defence. A special thank you to Dr. Hugh Chipman for agreeing to serve as the external examiner.

Thank you to all of the members of the McNicholas research group for your help, encouragement, and for making this experience all the more enjoyable. My deepest thanks to my family for all of their ongoing support, love, and encouragement over the last few years, especially my husband Hailiang.

Finally, I gratefully acknowledge the financial support I received towards my PhD from a Queen Elizabeth II Graduate Scholarship in Science and Technology, an Ontario Graduate Scholarship (OGS), and a Milos Novotny Fellowship.

# Publications

The following articles are based on the work presented in this thesis and have been published, submitted, or are in preparation for submission for publication:

- Wei, Y., Rausch, E., McNicholas, P. D. and Steinley, D. (2017). Extending growth mixture models using continuous non-elliptical distributions. *Submitted to Psychometrika. Available as arXiv preprint arXiv: 1703.08723.*

- Wei, Y., Rausch, E., McNicholas, P. D. and Steinley, D. (2017). Growth mixture models with non-elliptical random effects with incomplete data. *In preparation.*

- Wei, Y., and McNicholas, P. D. (2017). Mixtures of generalized hyperbolic distributions and mixtures of skew-t distributions for model-based clustering with incomplete data. *Submitted to Computational Statistics and Data Analysis. Available as arXiv preprint arXiv:1703.02177.*

- Wei, Y. and McNicholas, P. D. (2017). Flexible high-dimensional unsupervised learning with missing data. *Submitted to IEEE Transaction on Pattern Analysis and Machine Learning.*

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Finite Mixture Models

Finite mixture models (FMMs) assume that an overall population is made up of a collection of disjoint subpopulations, within which each subpopulation may be modelled by a statistical distribution. Formally, a random vector $\mathbf{X}$ taken from a $G$ component FMM, for all $\mathbf{x} \subset \mathbf{X}$, has density

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \, f_g(\mathbf{x} \mid \boldsymbol{\theta}_g), \tag{1.1}$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$, are the mixing proportions, $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the $g$th component density with component-specific parameters $\boldsymbol{\theta}_g$, and $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ denotes the model parameters with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$. Naturally, the number of mixture components $G$ can be used to model the heterogeneous data, leading to FMMs as extremely powerful and flexible tools for discovering heterogeneity in multivariate datasets. Extensive details on FMMs and their applications can be found in

Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), and Frühwirth-Schnatter (2006).

This thesis focuses on the application of the FMMs in two areas: growth mixture models (GMMs) (Chapters 3 and 4) and model-based clustering (Chapters 5 and 6). Their common aim is to partition the data into meaningful groups of homogeneous observations, where the similarity within groups and the dissimilarity between groups are maximized, i.e., clustering. Note that though both topics evolved from FMMs and had a common goal, they are applicable to different types of data and their statistical models are different. Specifically, GMMs are widely used for the analysis of longitudinal data, where observations are collected over time.

GMMs incorporate both conventional random effects growth modeling (Laird and Ware, 1982) and latent trajectory classes as in finite mixture modeling (1.1); therefore, they offer a way to handle the unobserved heterogeneity between subjects in their development. One common fundamental assumption for GMMs is that model errors are normally distributed (e.g., Verbeke and Lesaffre, 1996; Muthén and Shedden, 1999; Nagin, 1999; Muthén and Muthén, 2000; Muthén, 2001a,b; Muthén and Asparouhov, 2008). When the data are asymmetric and/or have heavier tails, more than one latent class is required to capture the observed variable distribution. This thesis focuses on the development of a GMM with continuous non-elliptical distributions that allow for parameterization of skewness and heavier tails, in addition to location and scale as in Gaussian GMM.

Assuming no prior knowledge of class labels, the application of FMMs (1.1) to perform clustering in this way is known as model-based clustering. As McNicholas (2016a) points out, the association between mixture models and clustering goes back

at least as far as Tiedeman (1955), who uses the former as a means of defining the latter. Gaussian mixture models are historically the most popular tool for model-based clustering and dominated the literature for quite some time (e.g., Celeux and Govaert, 1995; Fraley and Raftery, 1998; McLachlan et al., 2003; Bouveyron et al., 2007; McNicholas and Murphy, 2008, 2010). Recently, finite mixtures of non-elliptical distributions have flourished and facilitated the flexible clustering of the data featuring longer tails and asymmetry (e.g., Lin, 2010; Vrbik and McNicholas, 2012; Lee and McLachlan, 2014; Murray et al., 2014; Franczak et al., 2014; Dang et al., 2015; Karlis and Santourian, 2009; O'Hagan et al., 2016; Tortora et al., 2016). A comprehensive review of model-based clustering work, up to and including some recent work on non-Gaussian mixtures, is given by McNicholas (2016b). However, unobserved or missing observations are frequently a hindrance in multivariate datasets and so developing mixture models that can accommodate incomplete data is an important issue in model-based clustering. Therefore, work in this direction is also pursued in this thesis.

## 1.2 Thesis Structure

### 1.2.1 Chapter 2

Background information is given including details on growth mixture models, missing data mechanism, the EM algorithm and variants thereof, and some well-known non-elliptical distributions. Methods for model-selection and performance assessment are also discussed.

### 1.2.2 Chapter 3

A GMM with continuous non-elliptical distribution is introduced to capture skewness and heavier tails in the dataset, via the multivariate skew-t distribution and the generalized hyperbolic distribution. When extending GMMs, four statistical models are considered with different distributions of measurement errors and random effects. Algorithms for model parameter estimation are presented. The performance of our proposed GMMs with non-elliptical distributions is illustrated on simulated and real data.

### 1.2.3 Chapter 4

The growth mixture models with non-elliptical random effects are generalized to accommodate missing values under missing at random mechanism. Two indicator variables are introduced to facilitate the computation procedure for model parameter estimation. The methods are compared to the competing algorithms though simulation studies and real data analysis.

### 1.2.4 Chapter 5

Flexible methods and algorithms for model-based clustering with incomplete data are presented via mixture of the generalized hyperbolic and skew-t distributions. The statistical properties of the generalized hyperbolic and skew-t distributions are presented. An analytically tractable and computational feasible algorithm is formulated for parameter estimation and imputation of missing values for mixture models employing missing at random mechanisms.

### 1.2.5 Chapter 6

A generalization of the mixture of generalized hyperbolic factor analyzers (MGHFA) is presented for handling high-dimensional data in the presence of missing values. Under a missing at random mechanism, we develop a computationally efficient EM algorithm for parameter estimation of the MGHFA model with different patterns of missing values. As a by-product, the proposed procedure provides a conditional predictor to impute the missing values and a classifier to cluster partially observed vectors. The performance of our proposed methodology is illustrated through the analysis of simulated and real data.

### 1.2.6 Chapter 7

A summary of the work demonstrated in this thesis is presented and possible research prospects for future direction are also discussed.

## 1.3 The Contribution of this Work

The impact of the work proposed in this thesis on the body of current growth mixture models and model-based clustering literature is summarized here. Firstly, growth mixture models with non-elliptical distributions are introduced. This model extends the current literature on Gaussian growth mixture models, via the generalized hyperbolic distribution and the multivariate skew-t distribution, to allow for parameterization of skewness and heavy tails in a dataset. The parameter estimation procedure for the proposed models is shown to be both mathematically elegant and computationally appealing. These models show greater flexibility and ability to recover the true data

structure when compared to the Gaussian growth mixture models, due to the fact that the generalized hyperbolic distribution is a flexible distribution including many well-known distributions as its limiting and special cases.

Next, following the positive results using growth mixture models with non-elliptical distributions, we generalized our proposed models to accommodate missing values which are very common for longitudinal data. Our proposed models are demonstrated using simulated and real data and perform favourably compared to their counterpart Gaussian growth mixture models.

The second part of the thesis deals with model-based clustering in the presence of unobserved or missing values. Mixtures of the generalized hyperbolic distributions and mixtures of multivariate skew-t distributions for model-based clustering are presented to tackle missing values under missing at random mechanism. In addition to considering missing data, we develop families of MGHD and MST mixture models, each with 14 parsimonious eigen-decomposed scale matrices corresponding to the famous Gaussian parsimonious clustering models (GPCMs) of Banfield and Raftery (1993) and Celeux and Govaert (1995). Furthermore, we developed a unified approach to mixtures of generalized hyperbolic factor analyzers model for handling high-dimensional data in the presence of missing values as well as heavy-tailed and/or asymmetric clusters. Both of these models are demonstrated using simulated and real data and perform favourably compared to the mean imputation method.

# Chapter 2

# Background

## 2.1 Growth Mixture Models

Over the past two decades, growth mixture models (GMMs) have been widely used for the analysis of longitudinal data. Suppose a longitudinal study features $n$ subjects and $T$ time points or measurement occasions. For subject $i$ ($i = 1, \ldots, n$), let $\mathbf{y}_i$ be a $T \times 1$ vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ where $y_{it}$ represents the outcome on occasion $t$ ($t = 1, \ldots, T$), let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})'$ be an $m \times 1$ vector of observed time-invariant covariates, let $\boldsymbol{\eta}_i$ be a $q \times 1$ vector containing $q$ continuous latent variables, and let $\mathbf{C}_i$ be a $K \times 1$ vector consisting of $K$ class variables. Note that $\mathbf{C}_i = (C_{i1}, \ldots, C_{iK})'$ has a multinomial distribution, where $C_{ik} = 1$ if individual $i$ is in class $k$ and $C_{ik} = 0$ otherwise. The conventional GMM with Gaussian random effects can be represented using a hierarchical three-level formulation as follows.

At level 1 of the GMM, the continuous outcome variables $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are related

to the continuous latent variables $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$ via

$$\mathbf{Y}_i = \boldsymbol{\Lambda}_y \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \tag{2.1}$$

for $i = 1, \ldots, n$, $\boldsymbol{\epsilon}_i$ is a $T \times 1$ vector of residuals or measurement errors that is assumed to follow a multivariate Gaussian distribution $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_k)$, and $\boldsymbol{\Lambda}_y$ is a $T \times q$ design matrix consisting of factor loadings with each column corresponding to specific aspects of change. The matrix $\boldsymbol{\Lambda}_y$ and the vector $\boldsymbol{\eta}_i$ determine the growth trajectory of the model. For instance, when $q = 3$, $\boldsymbol{\eta}_i = (\eta_{0i}, \eta_{1i}, \eta_{2i})$, and $\boldsymbol{\Lambda}_y$ is a $T \times 3$ matrix. Assuming $a_t$ are age-related time scores ($t = 1, 2, \ldots, T$) centred at age $a_0$, then $\boldsymbol{\Lambda}_y$ is given by

$$\boldsymbol{\Lambda}_y = \begin{pmatrix} 1 & a_1 - a_0 & (a_1 - a_0)^2 \\ 1 & a_2 - a_0 & (a_2 - a_0)^2 \\ \vdots & \vdots & \vdots \\ 1 & a_{T-1} - a_0 & (a_{T-1} - a_0)^2 \end{pmatrix}.$$

At level 2 of the GMM, the continuous latent variables $\boldsymbol{\eta}$ are related to the latent categorical variables $\mathbf{c}$ and to the observed time-invariant covariate vector $\mathbf{x}$ by the relation

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + \boldsymbol{\zeta}_i, \tag{2.2}$$

where $\boldsymbol{\alpha}_k$ ($k = 1, \ldots, K$) denotes the intercept parameter for class $k$, $\boldsymbol{\zeta}_i$ is a $q$-dimensional vector of residuals assumed to follow a multivariate Gaussian distribution $\boldsymbol{\zeta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_k)$, and $\boldsymbol{\Gamma}_k$ is a $q \times m$ parameter matrix representing the effect of $\mathbf{x}$ on the latent continuous variables $\boldsymbol{\eta}$ and assumed to be different among classes. Note

that the level 2 errors $\boldsymbol{\zeta}_i$ are uncorrelated with the measurement errors $\boldsymbol{\epsilon}_i$. We may allow for class-specific effects $\boldsymbol{\Gamma}_k$ in (2.2) that are equal across classes.

By combining the first two levels of the GMM, we have

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.3}$$

where $\pi_k = \Pr(C_{ik} = 1)$ is the class probability or mixing proportions satisfying $0 < \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$, and $\phi(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a multivariate Gaussian density with mean $\boldsymbol{\mu}_k = \boldsymbol{\Lambda}_y(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i)$ and covariance matrix $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k$. Notice that the GMM in (2.3) assumes that class probability $\pi_k$ is constant for each class.

At level 3 of the GMM, we assume that the class probabilities are no longer constant, but depend on the observed covariates. In other words, we want to know how $\pi_k$ is related to an individual's background variables, such as gender and income. At this level, the categorical latent variables $\mathbf{C}_i$ represent membership of mixture components that are related to $\mathbf{x}$ through a multinomial logit regression for unordered categorical responses. Define $\pi_{ik} = \Pr(C_{ik} = 1 \mid \mathbf{x}_i)$, i.e., the probability that subject $i$ falls into the $k$th class depending on the covariates $\mathbf{x}_i$. Let $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iK})'$ and

$$\text{logit}(\boldsymbol{\pi}_i) = \left( \log\left(\frac{\pi_{i1}}{\pi_{iK}}\right), \log\left(\frac{\pi_{i2}}{\pi_{iK}}\right), \ldots, \log\left(\frac{\pi_{iK-1}}{\pi_{iK}}\right) \right)'.$$

Then,

$$\text{logit}(\boldsymbol{\pi}_i) = \boldsymbol{\alpha}_c + \boldsymbol{\Gamma}_c \mathbf{x}_i, \tag{2.4}$$

where $\boldsymbol{\alpha}_c$ is a $(K-1) \times 1$ parameter vector and $\boldsymbol{\Gamma}_c$ is a $(K-1) \times q$ parameter matrix.

By combining these three levels of the GMM, we have

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik} \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.5}$$

Note that the model is not a finite mixture model anymore because the class probability is not constant with respect to $i$ (cf. McLachlan and Peel, 2000).

## 2.2 Missing Data Mechanism

Unobserved or missing observations are frequently a hindrance in multivariate datasets and so developing mixture models that can accommodate incomplete data is an important issue in finite mixture modelling. The maximum likelihood and Bayesian approaches are two common imputation paradigms for analyzing data with incomplete observations. Little and Rubin (2002) posited three different missing data mechanisms that remain in use today: (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR). In the missing data literature, data are often partitioned into two parts: the observed data ($\mathbf{X}^o$) and the missing data ($\mathbf{X}^m$). In this context, the missing data mechanism can be elegantly described through relationships among $\mathbf{X}^o$, $\mathbf{X}^m$, and the 'cause' of data missingness. MACR is a process in which the cause of missingness is independent of both $\mathbf{X}^o$ and $\mathbf{X}^m$. For MAR, the cause of missingness is not related to $\mathbf{X}^m$, but may depend on $\mathbf{X}^o$. Note that MACR is a special case of MAR. If data missingness are related to $\mathbf{X}^m$ or some unobserved latent variables, then the missing data mechanism is NMAR. Throughout this thesis, the missing data mechanism is assumed to be missing at random (MAR), under which the missing data mechanisms are ignorable for methods

using the maximum likelihood approach.

## 2.3 Non-Elliptical Distributions

There are a variety of non-elliptical distributions in the literature (Lee and McLachlan, 2014). However, the focus of this thesis will be on two non-elliptical distributions that arise as part of a larger family with nice properties called the normal variance-mean mixture distributions (NVMMs; Barndorff-Nielsen et al., 1982; Gneiting, 1997), namely the generalized hyperbolic distribution (GHD) and multivariate skew-t distribution (GST). Formally, the $p$-dimensional random variable $\mathbf{X}$ is said to have a multivariate NVMM if its density can be written in the form

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{U}, \qquad \mathbf{U} \perp W \tag{2.6}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are parameter vectors in $\mathbb{R}^p$, $W \geq 0$ is a univariate random variable, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is a multivariate Gaussian distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$, and the symbol $\perp$ indicates independence. Note that different distributional forms of $W$ will lead to many well-known distributions, such as multivariate skew-t distributions, multivariate t-distributions, and multivariate Gaussian distributions. Further examples of normal variance-mean mixtures are given by McNeil et al. (2005) and McNicholas (2016a), among others.

### 2.3.1 Generalized Inverse Gaussian Distribution

The random variable $W \in \mathbb{R}^+$ is said to have a generalized inverse Gaussian (GIG) distribution, introduced by (Good, 1953), with parameters $\lambda$, $\chi$, and $\psi$ if its probability density function is given by

$$f_{\text{GIG}}(w \mid \lambda, \chi, \psi) = \frac{(\psi/\chi)^{\lambda/2} w^{\lambda-1}}{2K_\lambda(\sqrt{\psi\chi})} \exp\left\{-\frac{\psi w + \chi/w}{2}\right\}, \tag{2.7}$$

where $\psi, \chi \in \mathbb{R}^+, \lambda \in \mathbb{R}$, and $K_\lambda$ is the modified Bessel function of the third kind with index $\lambda$. Herein, we write $W \sim \text{GIG}(\lambda, \chi, \psi)$ to indicate that a random variable $W$ has the GIG density as parameterized in (2.7). The GIG distribution has some attractive properties (Barndorff-Nielsen and Halgreen, 1977a; Blæsild, 1978; Halgreen, 1979; Jørgensen, 1982), including the tractability of the expectations:

$$\begin{aligned}
\mathbb{E}[W] &= \sqrt{\frac{\chi}{\psi}} \frac{K_{\lambda+1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})}, \\
\mathbb{E}[1/W] &= \sqrt{\frac{\psi}{\chi}} \frac{K_{\lambda-1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})} = \sqrt{\frac{\psi}{\chi}} \frac{K_{\lambda+1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})} - \frac{2\lambda}{\chi}, \\
\mathbb{E}[\log W] &= \log\left(\sqrt{\frac{\chi}{\psi}}\right) + \frac{\partial}{\partial\lambda} \log(K_\lambda(\sqrt{\psi\chi})).
\end{aligned} \tag{2.8}$$

These tractable expected values lead to the development of a computationally efficient E-step of the EM algorithm and its extensions throughout this thesis.

Browne and McNicholas (2015) introduce another parameterization of the GIG distribution by setting $\omega = \sqrt{\psi\chi}$ and $\eta = \sqrt{\chi/\psi}$. Write $W \sim \mathcal{I}(\lambda, \eta, \omega)$; its density is given by

$$f_{\mathcal{I}}(w \mid \lambda, \eta, \omega) = \frac{(w/\eta)^{\lambda-1}}{2\eta K_\lambda(\omega)} \exp\left\{-\frac{\omega}{2}\left(\frac{w}{\eta} + \frac{\eta}{w}\right)\right\} \tag{2.9}$$

for $w > 0$, where $\eta \in \mathbb{R}^+$ is a scale parameter and $\omega \in \mathbb{R}^+$ is a concentration parameter. These two parameterizations of the GIG distribution are important ingredients for building the generalized hyperbolic distribution presented later.

### 2.3.2 Generalized Hyperbolic Distribution

Several alternative parameterizations of the GHD have appeared in the literature, e.g., Barndorff-Nielsen and Blæsild (1981), McNeil et al. (2005), and Browne and McNicholas (2015). Barndorff-Nielsen and Halgreen (1977b) introduces the GHD to model the distribution of the sand grain sizes and subsequent reports described its statistical properties (e.g., Barndorff-Nielsen, 1978; Barndorff-Nielsen and Blæsild, 1981). However, under this standard parameterization, the parameters of the mixing distribution are not invariant under affine transformations. An important innovation was made by McNeil et al. (2005), who gave a new parameterization of the GHD. Under this new parameterization, the linear transformation of GHD remains in the same sub-family characterized by the parameters of the mixing distribution. However, there is an identifiability issue arising under this parameterization. To solve this problem, Browne and McNicholas (2015) give an alternative parameterization.

Following McNeil et al. (2005), a $p \times 1$ random vector $\mathbf{X}$ is said to follow a GHD with index parameter $\lambda$, concentration parameters $\chi$ and $\psi$, location vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$, and skewness vector $\boldsymbol{\alpha}$, denoted by $\mathbf{X} \sim \mathrm{GH}_p(\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, if it can be represented by (2.6), where $W \sim \mathrm{GIG}(\lambda, \chi, \psi)$, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. It follows that $\mathbf{X} \mid w \sim \mathcal{N}(\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})$. So, the density of the generalized hyperbolic random

vector $\mathbf{X}$ is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}{\psi + \boldsymbol{\alpha}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}}\right]^{\frac{\lambda - p/2}{2}} \times \frac{(\psi/\chi)^{\lambda/2} K_{\lambda - p/2}\left(\sqrt{(\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}))(\psi + \boldsymbol{\alpha}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha})}\right)}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} K_\lambda(\sqrt{\chi\psi}) \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}\}}, \quad (2.10)$$

where $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}$, $K_\lambda$ is the modified Bessel function of the third kind with index $\lambda$, and $\boldsymbol{\vartheta} = (\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ denotes the model parameters. The mean and covariance matrix of $\mathbf{X}$ are

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu} + \mathbb{E}(W)\boldsymbol{\alpha} \quad \text{and} \quad \mathbb{V}\text{ar}(\mathbf{X}) = \mathbb{E}(W)\boldsymbol{\Sigma} + \mathbb{V}\text{ar}(W)\boldsymbol{\alpha}\boldsymbol{\alpha}^\intercal, \quad (2.11)$$

respectively, where $\mathbb{E}(W)$ and $\mathbb{V}\text{ar}(W)$ are the mean and variance of the random variable $W$, respectively.

Note that, in this parameterization, we need to hold $|\boldsymbol{\Sigma}| = 1$ to ensure identifiability. Using $|\boldsymbol{\Sigma}| = 1$ solves the identifiability problem but would be prohibitively restrictive for model-based clustering and classification applications. Hence, Browne and McNicholas (2015) develop a new parameterization of the GHD with index parameter $\lambda$, concentration parameter $\omega$, location vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$, and skewness vector $\boldsymbol{\beta} = \eta\boldsymbol{\alpha}$, denoted by $\mathbf{X} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Note that $\eta = 1$. This formulation is given by

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\beta} + \sqrt{W}\mathbf{U}, \quad (2.12)$$

where $W \sim \mathcal{I}(\lambda, 1, \omega)$ and $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Under this parameterization, the density of the generalized hyperbolic random vector $\mathbf{X}$ is

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\omega + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}{\omega + \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}}\right]^{\frac{\lambda - p/2}{2}} \times \frac{K_{\lambda - p/2}\left(\sqrt{(\omega + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}))(\omega + \boldsymbol{\beta}^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta})}\right)}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} K_\lambda(\omega) \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}\}}, \quad (2.13)$$

where $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})$ and $K_{\lambda - p/2}$ are as described earlier. This parameterization of the GHD, together with the following multivariate skew-t distribution, are used for model development throughout this thesis. Now, recalling that $W \sim \mathcal{I}(\omega, 1, \lambda)$ and that the unconditional distribution of $\mathbf{X}$ is generalized hyperbolic, Bayes' theorem gives $W \mid \mathbf{x} \sim \text{GIG}(\omega + \boldsymbol{\beta}^{'} \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}, \omega + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}), \lambda - p/2)$. This elegant result will be used to extend GMMs to the generalized hyperbolic distribution and handling incomplete data in Chapters 3 and 4, respectively.

### 2.3.3 The Multivariate Skew-$t$ Distribution

Several alternative formulations of the multivariate skew-t distribution have appeared in the literature (e.g., Branco and Dey, 2001; Sahu et al., 2003; Lee and McLachlan, 2014; McNeil et al., 2005). The formulation of the multivariate skew-t distribution used herein arises as a special and limiting case of the GHD by setting $\lambda = -\nu/2$ and $\chi = \nu$, and letting $\psi \to 0$. This formulation of the multivariate skew-t distribution has been used by Murray et al. (2014) to develop a mixture of skew-t factor analyzer models.

A $p$-dimensional skew-t random variable $\mathbf{X}$ has the density function

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{v + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}{\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}}\right]^{\frac{-\nu-p}{4}} \times \frac{\nu^{\nu/2} K_{(-\nu-p)/2}\left(\sqrt{(\nu + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}))(\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})}\right)}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma(\nu/2)2^{\nu/2-1}\exp\{-(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\}}. \quad (2.14)$$

where $\boldsymbol{\mu}$ is the location parameter, $\boldsymbol{\Sigma}$ is the scale parameter, $\boldsymbol{\beta}$ is the skew parameter, $\nu$ is the degree of freedom parameter, and $K_{(-\nu-p)/2}$ and $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})$ are as defined in (2.10). We write $\mathbf{X} \sim \text{GST}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \nu)$ to denote that the random variable $\mathbf{X}$ follows the skew-t distribution such that it has the density in (2.14). Now, $\mathbf{X} \sim \text{GST}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \nu)$ can be obtained through the relationship in (2.12) with $W \sim \text{IG}(\nu/2, \nu/2)$, where $\text{IG}(\cdot)$ denotes the inverse-gamma distribution. We have $\mathbf{X} \mid w \sim \mathcal{N}(\boldsymbol{\mu} + w\boldsymbol{\beta}, w\boldsymbol{\Sigma})$, and so, from Bayes's theorem, $W \mid \mathbf{x} \sim \text{GIG}(\boldsymbol{\beta}^{'}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}, \nu + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}), -(\nu+p)/2)$.

## 2.4 The EM algorithm and Extensions

### 2.4.1 The EM algorithm

As is typical within the field of finite mixture modelling, the expectation-maximization (EM) algorithm (Dempster et al., 1977) is used to find maximum likelihood (ML) estimates of the model parameters in the presence of incomplete data. Due to its construction, it is a natural and effective procedure for handling missing data problem, which can meet with both the actual and conceptual missing data. The two aspects of application of the EM algorithm are illustrated throughout this thesis. For instance, in Chapter 3, data are complete without actual missing values but treated as incomplete

by adding conceptual 'missing' data for latent variables, such as latent growth factors $\boldsymbol{\eta}$ and latent trajectory class variable $\mathbf{C}$. In Chapters 4, 5, and 6, data are incomplete with both actual and conceptual missing data.

The EM algorithm is based on the complete-data, which refers to the combination of the observed data and the unobserved data (i.e., the actual and conceptual missing data). The algorithm alternates between two processes: an expectation (E-) step and a maximization (M-) step. In each E-step, the expected value of the complete-data log-likelihood, namely the so-called $\mathcal{Q}$ function, is computed conditional on the observed data and the current parameter estimates. The E-step can be further simplified to calculate the conditional expectations of the actual missing or latent variables. In the M-step, the expected complete-data log-likelihood $\mathcal{Q}$ is maximized with respect to the model parameters.

### 2.4.2 The AECM algorithm

The alternating expectation conditional maximization (AECM) algorithm (Meng and Van Dyk, 1997) is an extension of the EM algorithm, or more precisely, is a modification of the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993). Specifically, the ECM algorithm is an extension of the EM algorithm, where the M-step is simplified by performing a sequence of analytically tractable, simpler, and faster conditional maximization (CM-) steps, and the AECM algorithm is an extension of the ECM algorithm where the specification of complete-data is allowed to be different at each cycle of the algorithm. Similar to the regular M-step, the CM-step will maximize the conditional expectation of its corresponding complete-data log-likelihood at each cycle. Please refer to McLachlan and Krishnan

(2008) for complete details as well as illustrative examples of the EM algorithm and its extensions.

### 2.4.3 Stopping Criteria

The EM algorithm and its extensions iteratively update the model parameters until some pre-specified criteria are satisfied. Two stopping criteria are lack of progress and the Aitken's acceleration-based criterion. The lack of progress approach is to stop the algorithm depending on the difference in successive observed log-likelihood values, i.e., the EM algorithm is stopped when

$$l^{(r+1)} - l^{(r)} < \epsilon$$

for a given small threshold $\epsilon$. As McNicholas et al. (2010) pointed out that the drawback of this criterion is that the algorithm would be stopped before reaching the global maximum in situations where there are jumps in the likelihood.

Alternatively, Aitken's acceleration-based criterion (Aitken, 1926) is the most popular criterion. The Aitken acceleration at iteration $r$ is

$$a^{(r)} = \frac{l^{(r+1)} - l^{(r)}}{l^{(r)} - l^{(r-1)}},$$

where $l^{(r)}$ is the log-likelihood value evaluated at iteration $(r)$. Following Böhning et al. (1994) and Lindsay (1995), the asymptotic estimate of the log-likelihood at iteration $r+1$ is

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

For all the algorithms developed in this thesis, we use the method recommended by McNicholas et al. (2010), which stops the algorithm when

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

for some small positive $\epsilon$. Note that there exist other stopping criteria based on Aitken's acceleration (Böhning et al., 1994; Lindsay, 1995).

## 2.5 Model Selection

One of the main objectives related to the application of finite mixture modelling is to select a best model from a set of candidate models. Generally speaking, selecting the best model include several facets: determination of the number of cluster or mixture components, when relevant, choosing number of latent variables and component covariance structure, among others. There are a variety of options for model selection criteria. The most popular criterion for this purpose is the Bayesian information criterion (BIC; Schwarz, 1978). The BIC is defined as

$$\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\Theta}}) - \rho \log(n), \tag{2.15}$$

where $\hat{\boldsymbol{\vartheta}}$ is the ML estimate of model parameters $\boldsymbol{\vartheta}$, $l(\mathbf{x}, \hat{\boldsymbol{\Theta}})$ is the maximized log-likelihood value, $\rho$ is the number of free parameters, and $n$ is the number of observations in the model. Empirical evidence (e.g., McNicholas and Murphy, 2008; Baek et al., 2010) have shown that the BIC performs well in choosing the number of clusters and the ideal number of latent variables.

However, the BIC can be unreliable or does not necessarily give the best model from a set of candidate models (Biernacki et al., 2000; Baek and McLachlan, 2011; Bhattacharya and McNicholas, 2014). Hence, alternatives such as the integrated completed likelihood (ICL; Biernacki et al., 2000) have been considered. The ICL can be calculated via

$$\text{ICL} \approx \text{BIC} + 2\sum_{i=1}^{n}\sum_{g=1}^{G}\text{MAP}(\hat{z}_{ig})\log\hat{z}_{ig}, \tag{2.16}$$

where $\hat{z}_{ig}$ is the estimated posterior probability that $\mathbf{x}_i$ arises from the $g$th component, and MAP denotes the *maximum a posterior* probability such that $\text{MAP}(\hat{z}_{ig}) = 1$ if $\max_g(\hat{z}_{ig})$ occurs in the $g$th component and $\text{MAP}(\hat{z}_{ig}) = 0$ otherwise. The part after the plus sign, known as the estimated mean entropy, reflects the uncertainty in the classification of observations into components.

Another option is to consider the approximated weight of evidence (AWE; Banfield and Raftery, 1993). The AWE is given by

$$\text{AWE} = \text{BIC} + 2EN(\mathbf{z}) - \rho(3 + \log n), \tag{2.17}$$

where $EN(\mathbf{z}) = \sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\log\hat{z}_{ig}$ is the entropy of the classification matrix with the $(i, g)$th entry being $\hat{z}_{ig}$. Clearly, the ICL and AWE penalize complex models more severely than the BIC, and thus tend to select more parsimonious models in practice. When defined as in (2.15), (2.16), and (2.17), the model with the largest value of those criteria is selected. Nevertheless, there is no optimal strategy with respect to which criterion is always the best, and a combined use of these criteria could be helpful in screening reasonable candidate models.

## 2.6 Comparing Partitions

Throughout this thesis, the misclassification rate and the adjusted Rand index (ARI; Hubert and Arabie, 1985) are mostly used to assess the classification performance of the proposed methods. For this purpose, we often consider datasets with known true class membership. However, these true class memberships are entirely hidden from our algorithms and not used to aid the clustering. Because the true class membership for the datasets are known *a priori*, the misclassification (error) rate (ERR) can simply be calculated as

$$\text{ERR} = \frac{\text{number of observations that were misclassified}}{\text{total number of observations}} \tag{2.18}$$

The ARI is a method based on pairwise agreement and a corrected form of Rand index Rand (RI; 1971) for taking into account the fact of some cases will be correctly classified due to chance. For a better understanding, Table 2.1, which is taken from McNicholas (2016a), summaries the pairwise agreements and disagreements.

Table 2.1: Cross-tabulation of pairs for two partitions, where row represent pairs of observations from one partition and columns represent pairs from another partition.

| | Same group | Different group |
|---|---|---|
| Same group | A | B |
| Different group | C | D |

The RI is the ratio of the number of pairwise agreements to total number of pairs (i.e., number of pairwise agreements plus number of pairwise disagreements) and can be expressed as

$$\text{RI} = \frac{A + D}{N}, \tag{2.19}$$

where $N = A + B + C + D$ is the total number of pairs. Its values range from 0 to 1, with 1 indicating perfect class agreement. However, an issue with the RI is that it may be difficult to interpret for small values of the RI due to the fact that its expected value is greater than 0 under random classification. The ARI is an improvement of the RI and takes the form

$$\text{ARI} = \frac{N(A+D) - [(A+B)(A+C)+(C+D)(B+D)]}{N^2 - [(A+B)(A+C)+(C+D)(B+D)]} \tag{2.20}$$

(cf. Steinley, 2004). The expected value of ARI under randomization is zero. When compared to the true classification, the ARI value of 1 corresponds to perfect classification, while a negative value of ARI indicates that the classifier is worse than randomly performing a classification.

# Chapter 3

# Extending Growth Mixture Models Using Continuous Non-Elliptical Distributions

## 3.1 Introduction

In this chapter, growth mixture models with continuous non-elliptical distributions are developed using two different distributions, namely the generalized hyperbolic distribution and the multivariate skew-t distribution. It is well known that the generalized hyperbolic distribution is perhaps the most flexible alternative to the Gaussian distribution in the literature (McNeil et al., 2005; McNicholas, 2016a), including many well-known distributions as its limiting or special cases.

As mentioned in Chapter 2, the majority of work on GMMs are based on the normality assumption. When the normality assumption is violated, GMMs are tend to yield non-consistent estimates and overextraction of the number of latent classes (Arminger

et al., 1999; Bauer and Curran, 2003a,b; Guerra-Peña and Steinley, 2016). Therefore, we introduce a non-elliptical approach that allows for skewness and heavy tails while also parameterizing location and scale. This approach is effective and mathematically elegant.

## 3.2 Methodology

### 3.2.1 GMM with the generalized hyperbolic distribution

As discussed in Section 2.1, conventional GMMs assume that the residuals $\boldsymbol{\epsilon}$ and $\boldsymbol{\zeta}$ have multivariate Gaussian distribution with zero means and within-class covariance matrices, respectively. We are interested in constructing a GMM with generalized hyperbolic distribution model errors, denoted by GHD-GMM. The generalized hyperbolic distribution can be represented as a normal mean-variance mixture, where the mixing weight has a GIG distribution. To this end, we introduce a latent continuous variable $W$ with $W_{ik} \mid c_{ik} = 1 \sim \mathcal{I}(\omega_k, 1, \lambda_k)$. Accordingly, conditional on $c_{ik}$ and $w_{ik}$, we assume that model errors $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\zeta}_i$ are non-centered Gaussian error terms with distinct covariance matrices:

$$\boldsymbol{\epsilon}_i \mid w_{ik}, c_{ik} = 1 \sim \mathcal{N}(w_i \boldsymbol{\beta}_{yk}, w_{ik} \boldsymbol{\Theta}_k), \tag{3.1}$$

$$\boldsymbol{\zeta}_i \mid w_{ik}, c_{ik} = 1 \sim \mathcal{N}(w_i \boldsymbol{\beta}_{\eta k}, w_{ik} \boldsymbol{\Psi}_k), \tag{3.2}$$

where $\boldsymbol{\Theta}_k$ is the diagonal covariance matrix for $\boldsymbol{\epsilon}_i$, and $\boldsymbol{\Psi}_k$ is the covariance matrix for $\boldsymbol{\zeta}_i$. The $T$-dimensional vector $\boldsymbol{\beta}_{yk}$ is a vector of skewness parameters, which we refer to as the skewness parameter for the measurement errors. The $q$-dimensional

vector $\boldsymbol{\beta}_{\eta k}$ is the vector of skew parameters for the continuous latent variables $\boldsymbol{\eta}_i$. Then, based on (2.2) and (3.1), the observed random variables $\mathbf{Y}_i$, conditional on $\boldsymbol{\eta}_i$, $c_{ik}$, and $w_{ik}$, follow a conditional Gaussian distribution of the form

$$\mathbf{Y}_i \mid \boldsymbol{\eta}_i, w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\boldsymbol{\Lambda}_y \boldsymbol{\eta}_i + w_i \boldsymbol{\beta}_{yk}, w_{ik} \boldsymbol{\Theta}_k). \tag{3.3}$$

And, based on (2.2) and (3.2),

$$\boldsymbol{\eta}_i \mid \mathbf{x}_i, w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik} \boldsymbol{\beta}_{\eta k}, w_{ik} \boldsymbol{\Psi}_k). \tag{3.4}$$

and, from the preceding equations, we have the conditional distribution

$$\mathbf{Y}_i \mid \mathbf{x}_i, w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_k + w_{ik}(\boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k} + \boldsymbol{\beta}_{yk}), w_{ik} \boldsymbol{\Sigma}_k), \tag{3.5}$$

where $\boldsymbol{\mu}_k = \boldsymbol{\Lambda}_y(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i)$ and $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k$. Recalling the elegant result explained in Section 2.3, we obtain the conditional distributions

$$\boldsymbol{\eta}_i \mid \mathbf{x}_i, c_{ik} = 1 \sim \text{GHD}_q(\lambda_k, \omega_k, \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i, \boldsymbol{\Psi}_k, \boldsymbol{\beta}_{\eta k}), \tag{3.6}$$

$$\mathbf{Y}_i \mid \mathbf{x}_i, c_{ik} = 1 \sim \text{GHD}_T(\lambda_k, \omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k} + \boldsymbol{\beta}_{yk}). \tag{3.7}$$

By combining the preceding setup and level 3 of the GMM from Section 2.1, we arrive at a GMM with density

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik} f_{\text{GHD},T}(\mathbf{y}_i; \lambda_k, \omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k} + \boldsymbol{\beta}_{yk}), \tag{3.8}$$

where $f_{\text{GHD},T}(\cdot)$ is the density of a $T$-dimensional random variable following a GHD

as given in (2.13). Note that the overall skewness for $\mathbf{Y}_i$ is $\boldsymbol{\Lambda}_y\boldsymbol{\beta}_{\eta k} + \boldsymbol{\beta}_{yk}$. In the above setup, the dependent observed variable $\mathbf{Y}_i$, the latent growth factors $\boldsymbol{\eta}_i$, and residual variables $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\zeta}_i$ all have generalized hyperbolic distributions. Note that the distribution of the covariates $\mathbf{x}_i$ is not modelled; please refer to Muthén and Asparouhov (2015) for detailed explanations.

### 3.2.2 GMM with the multivariate skew-t distribution

In this section, we are interested in extending the conventional GMM to have multivariate skew-t distribution model errors, denoted by GST-GMM. As in the case for the GHD, the multivariate skew-t distribution also has a convenient representation as a normal mean-variance mixture; this time, the weight has an inverse-gamma distribution. In analogous fashion to the GHD-GMM, a latent continuous random variable $W_{ik}$ is first introduced, where $W_{ik} \mid c_{ik} = 1 \sim \text{IG}(\nu_k/2, \nu_k/2)$. Accordingly, we assume that $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\zeta}_i$ are non-centred Gaussian error terms with their own covariance matrices as in (3.1) and (3.2), and $\mathbf{y}_i$ and $\boldsymbol{\eta}_i$ are conditionally normally distributed as in (3.3) and (3.4). From this characterization of the multivariate skew-t distribution, the following conditional distributions are obtained:

$$\boldsymbol{\eta}_i \mid \mathbf{x}_i, c_{ik} = 1 \sim \text{GST}_q(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i, \boldsymbol{\Psi}_k, \boldsymbol{\beta}_{\eta k}, \nu_k), \tag{3.9}$$

$$\mathbf{Y}_i \mid \mathbf{x}_i, c_{ik} = 1 \sim \text{GST}_T(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\beta}_{yk} + \boldsymbol{\Lambda}_y\boldsymbol{\beta}_{\eta k}, \nu_k), \tag{3.10}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are as described above and $\nu_k$ is a concentration parameter (i.e., the degrees of freedom). Similarly, we arrive at a GMM with a multivariate skew-t

distribution

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik} f_{\text{GST},T}(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\beta}_{yk} + \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k}, \nu_k), \tag{3.11}$$

where $\pi_{ik}$ is defined as in (2.4). In this setup, the random variable $\mathbf{Y}_i$, the latent growth factors $\boldsymbol{\eta}_i$, and the residual variables $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\zeta}_i$ all follow multivariate skew-t distributions.

### 3.2.3 Comments on the GHD-GMM and GST-GMM

In the preceding extensions of GMMs, recall that the overall skewness for $\mathbf{Y}_i$ is $\boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k} + \boldsymbol{\beta}_{yk}$, so there are a total of $T + q$ skewness parameters. Hence, the skewness parameters $\boldsymbol{\beta}_{yk}$ and $\boldsymbol{\beta}_{\eta k}$ are subject to identifiability issues, because no more than $T$ skewness parameters can be identified from the $T$-dimensional $\mathbf{Y}_i$. Therefore, two special formulations are considered in this chapter. The first formulation is where $\boldsymbol{\beta}_{yk} = \mathbf{0}$. In this formulation, the residuals for $\mathbf{Y}_i$ or the measurement errors are not skewed, i.e., $\boldsymbol{\epsilon}_i \mid w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\mathbf{0}, w_{ik}\boldsymbol{\Theta}_k)$. All of the skewness in the data is assumed to come from the distribution of latent factors. The second special formulation is the case where $\boldsymbol{\beta}_{\eta k} = \mathbf{0}$. In this formulation, the residuals for the latent factors $\boldsymbol{\eta}$ are symmetric, i.e., $\boldsymbol{\zeta}_i \mid w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\mathbf{0}, w_{ik}\boldsymbol{\Psi}_k)$. Accordingly, all of the skewness in the data is assumed to come from the residuals of $\mathbf{Y}_i$ or the measurement errors. In practice, we would want as much of the skewness as possible in the observed data $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ to be explained through the latent factors. There appears to be no optimal strategy with respect to which skewness parameter to estimate. Accordingly, four statistical models, differing with respect to the distributions of measurement errors

and random effects for the first two levels of the GMM, are employed and compared. These models are as follows:

- **Model I:** A model with independent multivariate generalized hyperbolic random effects and measurement errors while assuming all of the skewness in the data comes from the distribution of latent factors (i.e., GHD-GMM under $\boldsymbol{\beta}_{yk} = \mathbf{0}$).
- **Model II:** A model with independent multivariate generalized hyperbolic random effects and measurement errors while assuming all of the skewness in the data comes from the residuals of $\mathbf{Y}$ (i.e., GHD-GMM under $\boldsymbol{\beta}_{\eta k} = \mathbf{0}$).
- **Model III:** A model with independent multivariate skew-t random effects and measurement errors while assuming all of the skewness in the data comes from the distribution of latent factors (i.e., GST-GMM under $\boldsymbol{\beta}_{yk} = \mathbf{0}$).
- **Model IV:** A model with independent multivariate skew-t random effects and measurement errors while assuming all of the skewness in the data comes from the residuals of $\mathbf{Y}$ (i.e., GST-GMM under $\boldsymbol{\beta}_{\eta k} = \mathbf{0}$).

Take Model I (i.e., GHD-GMM under $\boldsymbol{\beta}_{yk} = \mathbf{0}$) as an example. For different trajectory classes, the parameters $\lambda_k, \omega_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_{\eta k}, \boldsymbol{\Theta}_k, \boldsymbol{\Psi}_k$, and $\boldsymbol{\Gamma}_k$ may be different across classes, or may be the same across the classes. By imposing constraints on all these parameters (different or the same across classes), we obtain a family of GHD-GMM models. In this chapter, we only consider two models. One model assumes that the parameters $\lambda_k, \omega_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_{\eta k}, \boldsymbol{\Theta}_k, \boldsymbol{\Psi}_k$, and $\boldsymbol{\Gamma}_k$ are different across classes, we call this model as the general model. The second model assumes that only the parameter $\boldsymbol{\alpha}_k$ is different across classes while all the other parameters are the same across classes,

i.e., $\lambda_k = \lambda$, $\omega_k = \omega$, $\boldsymbol{\beta}_{\eta k} = \boldsymbol{\beta}_\eta$, $\boldsymbol{\Theta}_k = \boldsymbol{\Theta}$, $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}$, and $\boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}$ for $k = 1, 2, \ldots, K$; we call this model the most constrained model.

## 3.3 Parameter Estimation

### 3.3.1 The EM algorithm for Model I

For our GHD-GMM under $\boldsymbol{\beta}_{yk} = \mathbf{0}$, the complete-data comprise the observed outcome data $\mathbf{y}_1, \ldots, \mathbf{y}_n$, the covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$, together with the latent categorical variables $\mathbf{c}_1, \ldots, \mathbf{c}_n$, the latent growth factors $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n$, and the latent $w_{ik}$. The observed-data log-likelihood can be expressed as follows:

$$\log\mathcal{L} = \sum_{i=1}^{n} \log p(\mathbf{y}_i \mid \mathbf{x}_i), \tag{3.12}$$

where

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik} f_{\text{GHD},T}(\mathbf{y}_i; \lambda_k, \omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k}) \tag{3.13}$$

and $\pi_{ik}$ is defined as in connection with (2.4).

Now, $\mathbf{Y}_i \mid \mathbf{x}_i, w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\boldsymbol{\mu}_k + w_{ik}\boldsymbol{\Lambda}_y\boldsymbol{\beta}_{\eta k}, w_{ik}\boldsymbol{\Sigma}_k)$ independently for $i = 1, \ldots, n$, $W_{ik} \mid c_{ik} = 1 \sim \mathcal{I}(\omega_k, 1, \tilde{\lambda}_k)$, and so, from Bayes's theorem, $W_{ik} \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1 \sim \text{GIG}(\psi_k, \chi_{ik}, \tilde{\lambda}_k)$ with $\psi_k = \omega_k + \boldsymbol{\beta}_{\eta k}^{'}\boldsymbol{\Lambda}_y^{'}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Lambda}_y\boldsymbol{\beta}_{\eta k}$, $\chi_{ik} = \omega_k + \delta(\mathbf{y}_i, \boldsymbol{\mu}_k \mid \boldsymbol{\Sigma}_k)$, and $\tilde{\lambda}_k = \lambda_k - T/2$. It follows that

$$\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\mathbf{V}_k({}^{-1}\boldsymbol{\Psi}_k(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i), w_{ik}\mathbf{V}_k), \tag{3.14}$$

where $\mathbf{V}_k = {}^{-1}\ ({}^{-1}\boldsymbol{\Psi}_k + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)$. The result in (3.14) is used to estimate the latent

growth factors $\boldsymbol{\eta}_i$, and a detailed proof thereof is given in Appendix A.1. Therefore, the complete-data likelihood is given by

$$\mathcal{L}_{\mathrm{c}}(\vartheta) = \prod_{i=1}^{n}\prod_{k=1}^{K}[\pi_{ik}\phi(\mathbf{y}_i \mid \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i, w_{ik}\boldsymbol{\Theta}_k)\phi(\boldsymbol{\eta}_i \mid \boldsymbol{\alpha}_k+\boldsymbol{\Gamma}_k\mathbf{x}_i+w_{ik}\boldsymbol{\beta}_{\eta k}, w_{ik}\boldsymbol{\Psi}_k)h(w_{ik} \mid \omega_k, \lambda_k)]^{c_{ik}},$$

with the same notation used previously, where $h(w_{ik} \mid \omega_k, \lambda_k)$ is the density of a GIG distribution in (2.9) with $\eta = 1$.

After some algebra, the complete-data log-likelihood is given by

$$\mathcal{L}_c(\vartheta \mid \mathbf{y}, \mathbf{x}) = \mathcal{L}_{1c}(\pi) + \mathcal{L}_{2c}(\boldsymbol{\Theta}_k) + \mathcal{L}_{2c}(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_{\eta k}, \boldsymbol{\Psi}_k, \boldsymbol{\Gamma}_k) + \mathcal{L}_{4c}(\boldsymbol{\lambda}, \boldsymbol{\omega}), \tag{3.15}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$, and

$$\begin{aligned}
\mathcal{L}_{1c} &= \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\log\pi_{ik},\\
\mathcal{L}_{2c} &= \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\left\{\frac{1}{2}\log|\boldsymbol{\Theta}_k^{-1}| - \frac{1}{2w_{ik}}\mathbf{y}_i^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i + \frac{1}{w_{ik}}\mathbf{y}_i^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y\boldsymbol{\eta}_i - \frac{1}{2w_{ik}}\boldsymbol{\eta}^{'}\boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y\boldsymbol{\eta}\right\} + C_1,\\
\mathcal{L}_{3c} &= \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\left\{\frac{1}{2}\log|\boldsymbol{\Psi}_k^{-1}| - \frac{1}{2w_{ik}}\boldsymbol{\eta}_i^{'}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\eta}_i + \frac{1}{w_{ik}}(\boldsymbol{\alpha}_k+\boldsymbol{\Gamma}_k\mathbf{x}_i)^{'}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\eta}_i + \boldsymbol{\beta}_{\eta k}^{'}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\eta}_i\right.\\
&\qquad -\frac{1}{2w_{ik}}(\boldsymbol{\alpha}_k+\boldsymbol{\Gamma}_k\mathbf{x}_i)^{'}\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k+\boldsymbol{\Gamma}_k\mathbf{x}_i) - (\boldsymbol{\alpha}_k+\boldsymbol{\Gamma}_k\mathbf{x}_i)^{'}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\eta k}\\
&\qquad \left.-\frac{w_{ik}}{2}\boldsymbol{\beta}_{\eta k}^{'}\boldsymbol{\Psi}_k^{-1}\boldsymbol{\beta}_{\eta k}\right\} + C_2,\\
\mathcal{L}_{4c} &= \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\left\{(\lambda_k-1)\log w_{ik} - \log K_{\lambda_k}(\omega_k) - \frac{\omega_k}{2}\left(w_{ik}+\frac{1}{w_{ik}}\right)\right\},
\end{aligned}$$

where $C_1$ and $C_2$ are constants with respect to model parameters.

In the E-step, we compute the conditional exception of $\mathcal{L}_c(\boldsymbol{\vartheta} \mid \mathbf{y}, \mathbf{x})$ given in (3.15), denoted $\mathcal{Q}$. First, let $p_{ik}$ denote the probability that the $i$th observation belongs to

the $k$th component of the mixture, and is updated by

$$p_{ik} := \mathbb{E}\left[C_{ik} \mid \mathbf{y}_i, \mathbf{x}_i\right] = \frac{\pi_{ik} f_{\mathrm{GHD},T}(\mathbf{y}_i; \lambda_k, \omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k})}{\sum_{l=1}^{K} \pi_{il} f_{\mathrm{GHD},T}(\mathbf{y}_i; \lambda_l, \omega_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta l})}.$$

The resulting E-step quantiles regarding the latent growth factors $\boldsymbol{\eta}$ and the latent continuous variable $W$ are given by

$$E_{1ik} := \mathbb{E}\left[W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \sqrt{\frac{\chi_{ik}}{\psi_k}} \frac{K_{\tilde{\lambda}_k+1}(\sqrt{\psi_k \chi_{ik}})}{K_{\tilde{\lambda}_k}(\sqrt{\psi_k \chi_{ik}})},$$

$$E_{2ik} := \mathbb{E}\left[1/W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \sqrt{\frac{\psi_k}{\chi_{ik}}} \frac{K_{\tilde{\lambda}_k+1}(\sqrt{\psi_k \chi_{ik}})}{K_{\tilde{\lambda}_k}(\sqrt{\psi_k \chi_{ik}})} - \frac{2\tilde{\lambda}_k}{\chi_{ik}},$$

$$E_{3ik} := \mathbb{E}\left[\log W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \log\left(\sqrt{\frac{\chi_{ik}}{\psi_k}}\right) + \frac{1}{K_{\tilde{\lambda}_k}(\sqrt{\psi_k \chi_{ik}})} \frac{\partial}{\partial \tilde{\lambda}_k} K_{\tilde{\lambda}_k}(\sqrt{\psi_k \chi_{ik}}),$$

$$E_{4ik} := \mathbb{E}\left[\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1\right] = \mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + E_{1ik}\boldsymbol{\beta}_{\eta k}) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i),$$

$$E_{5ik} := \mathbb{E}\left[(1/W_{ik})\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1\right] = E_{2ik}\mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i) + \mathbf{V}_k \boldsymbol{\Psi}_k^{-1} \boldsymbol{\beta}_{\eta k},$$

$$\begin{aligned} E_{6ik} := \mathbb{E}[(1/W_{ik})\boldsymbol{\eta}_i \boldsymbol{\eta}_i^{'} \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1] &= \mathbf{V}_k + \mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i)\boldsymbol{\beta}_{\eta k} \boldsymbol{\Psi}_k^{-1} \mathbf{V}_k \\ &+ E_{2ik}\mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i)(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i)^{'} \mathbf{V}_k \\ &+ \mathbf{V}_k \boldsymbol{\Psi}_k^{-1} \boldsymbol{\beta}_{\eta k}^{'} (\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i)^{'} \mathbf{V}_k + E_{1ik}\mathbf{V}_k \boldsymbol{\Psi}_k^{-1} \boldsymbol{\beta}_{\eta k} \boldsymbol{\beta}_{\eta k}^{'} \boldsymbol{\Psi}_k^{-1} \mathbf{V}_k, \end{aligned}$$

and $\psi_k$, $\chi_{ik}$, and $\tilde{\lambda}_k$ are as previously defined. These attractive closed forms for $E_{1ik}$, $E_{2ik}$, and $E_{3ik}$ exist because $W_{ik} \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1 \sim \mathrm{GIG}(\psi_k, \chi_{ik}, \tilde{\lambda}_k)$, and so we can use the formulae as in (2.8). The existence of these attractive closed forms for $E_{4ik}$, $E_{5ik}$, and $E_{6ik}$ is due to the conditional Gaussian distribution of $\boldsymbol{\eta}$ as in (3.14).

In the M-step, we maximize $\mathcal{Q}$ with respect to the model parameters to get the

updates. With respect to the parameters $\boldsymbol{\alpha}_c$ and $\boldsymbol{\Gamma}_c$, the M-step maximizes

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\log\pi_{ik}, \tag{3.16}$$

which may be viewed as a multinomial logistic regression with fractional observations $p_{ik}$. The parameters $\omega_k$ and $\lambda_k$ are estimated by maximizing the following function

$$q_k(\omega_k,\lambda_k) = -\log K_{\lambda_k}(\omega_k) + (\lambda_k - 1)\bar{d}_k - \frac{\omega_k}{2}(\bar{a}_k + \bar{b}_k), \tag{3.17}$$

where $n_k = \sum_{i=1}^{n} p_{ik}$, $\bar{a}_k = \frac{1}{n_k}\sum_{i=1}^{n} p_{ik}E_{1ik}$, $\bar{b}_k = \frac{1}{n_k}\sum_{i=1}^{n} p_{ik}E_{2ik}$, and $\bar{d}_k = \frac{1}{n_k}\sum_{i=1}^{n} p_{ik}E_{3ik}$. The associated updates are

$$\hat{\lambda}_k = \bar{c}_k\hat{\lambda}_k^{\text{prev}}\left[\frac{\partial}{\partial t}\log K_t(\hat{\omega}_k^{\text{prev}})\mid_{t=\hat{\lambda}_k^{\text{prev}}}\right]^{-1},$$
$$\hat{\omega}_k = \hat{\omega}_k^{\text{prev}} - \left[\frac{\partial}{\partial t}q_k(t,\hat{\lambda}_k)\Big|_{t=\hat{\omega}_k^{\text{prev}}}\right]\left[\frac{\partial^2}{\partial t^2}q_k(t,\hat{\lambda}_k)\Big|_{t=\hat{\omega}_k^{\text{prev}}}\right]^{-1},$$

where the superscript "prev" means the previous estimate — refer to Browne and

McNicholas (2015) for further details. Finally, we get the updates of the other parameters in the model:

$$\hat{\boldsymbol{\Theta}}_k = \text{diag}\Bigg(\frac{\sum_{i=1}^{n} p_{ik}(E_{2ik}\mathbf{y}_i\mathbf{y}_i^{'} - \mathbf{y}_i E_{5ik}^{'}\boldsymbol{\Lambda}_y^{'} - \boldsymbol{\Lambda}_y E_{5ik}\mathbf{y}_i^{'} + \boldsymbol{\Lambda}_y E_{6ik}\boldsymbol{\Lambda}_y^{'})}{n_k}\Bigg),$$

$$\hat{\boldsymbol{\Gamma}}_k = \sum_{i=1}^{n} p_{ik}\left(E_{5ik}\mathbf{x}_i^{'} - E_{2ik}\hat{\boldsymbol{\alpha}}_k\mathbf{x}_i^{'} - \hat{\boldsymbol{\beta}}_k\mathbf{x}_i^{'}\right)\left(\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}E_{2ik}\mathbf{x}_i\mathbf{x}_i^{'}\right)^{-1},$$

$$\hat{\boldsymbol{\alpha}}_k = \frac{\bar{a}_k\sum_{i=1}^{n} p_{ik}(E_{5ik} - E_{2ik}\hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i) - \sum_{i=1}^{n} p_{ik}E_{4ik} + \sum_{i=1}^{n} p_{ik}\hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i}{n_k(\bar{a}_k\bar{b}_k - 1)},$$

$$\hat{\boldsymbol{\beta}}_{\eta k} = \frac{\bar{b}_k\sum_{i=1}^{n} p_{ik}(E_{4ik} - \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i) - \sum_{i=1}^{n} p_{ik}E_{5ik} + \sum_{i=1}^{n} p_{ik}E_{2ik}\hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i}{n_k(\bar{b}_k\bar{b}_k - 1)},$$

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_k = \frac{1}{n_k}\sum_{i=1}^{n} p_{ik}\Big(&E_{6ik} - \hat{\boldsymbol{\beta}}_k E_{4ik}^{'} - E_{5ik}(\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)^{'} - E_{4ik}\hat{\boldsymbol{\beta}}_k^{'} - (\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)E_{5ik}^{'} \\ &+ E_{2ik}(\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)(\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)^{'} + (\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)\hat{\boldsymbol{\beta}}_{\eta k}^{'} + \hat{\boldsymbol{\beta}}_{\eta k}(\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)^{'} + E_{1ik}\hat{\boldsymbol{\beta}}_{\eta k}\hat{\boldsymbol{\beta}}_{\eta k}^{'}\Big).\end{aligned}$$

### 3.3.2 The EM algorithm for Model III

Similarly, for our Model III, parameter estimation is carried out within the EM algorithm framework. Suppose we observe the outcome $\mathbf{y}_i$ and the covariates $\mathbf{x}_i$ from a GMM with skew-t random effects as in (3.11) but with $\boldsymbol{\beta}_{yk} = \mathbf{0}$. There are three sources of unobserved data: the latent categorical variables $\mathbf{c}_i$, the latent growth factors $\boldsymbol{\eta}_i$, and the latent $w_{ik}$. The complete-data log-likelihood can be expressed as follows:

$$\begin{aligned}\mathcal{L}_c(\vartheta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\big[&\log\pi_{ik} + \log\phi(\mathbf{y}_i \mid \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i, w_{ik}\boldsymbol{\Theta}_k) \\ &+ \log\phi(\boldsymbol{\eta}_i \mid \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}, w_{ik}\boldsymbol{\Psi}_k) + \log f(w_{ik} \mid \nu_k/2, \nu_k/2)\big],\end{aligned}$$

where $\pi_{ik}$ is defined as above in (2.4) and $f(w_{ik} \mid \nu_k/2, \nu_k/2)$ is the density of the inverse Gamma distribution.

The E-step requires the computation of the expected value of the complete-data log-likelihood. Note that $W_{ik} \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1 \sim \text{GIG}(\psi_k^*, \chi_{ik}^*, \lambda_k^*)$ with $\psi_k^* = \boldsymbol{\beta}_{\eta k}^{'} \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k}$, $\chi_{ik}^* = \nu_k + \delta(\mathbf{y}_i, \boldsymbol{\mu}_k \mid \boldsymbol{\Sigma}_k)$, and $\lambda_k^* = -(\lambda_k + T)/2$. So, we have convenient forms for the following expected values:

$$E_{1ik}^* := \mathbb{E}\left[W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \sqrt{\frac{\chi_{ik}^*}{\psi_k^*}} \frac{K_{\lambda_k^*+1}(\sqrt{\psi_k^* \chi_{ik}^*})}{K_{\lambda_k^*}(\sqrt{\psi_k^* \chi_{ik}^*})},$$

$$E_{2ik}^* := \mathbb{E}\left[1/W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \sqrt{\frac{\psi_k^*}{\chi_{ik}^*}} \frac{K_{\lambda_k^*+1}(\sqrt{\psi_k^* \chi_{ik}^*})}{K_{\lambda_k^*}(\sqrt{\psi_k^* \chi_{ik}^*})} - \frac{2\lambda_k^*}{\chi_{ik}^*},$$

$$E_{3ik}^* := \mathbb{E}\left[\log W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \log\left(\sqrt{\frac{\chi_{ik}^*}{\psi_k^*}}\right) + \frac{1}{K_{\lambda_k^*}(\sqrt{\psi_k^* \chi_{ik}^*})} \frac{\partial}{\partial \lambda_k^*} K_{\lambda_k^*}(\sqrt{\psi_k^* \chi_{ik}^*}).$$

We also need the expected value of the class membership, i.e.,

$$\tau_{ik} := \mathbb{E}\left[C_{ik} \mid \mathbf{y}_i, \mathbf{x}_i\right] = \frac{\pi_{ik} f_{\text{GHD},T}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta k}, v_k)}{\sum_{l=1}^{K} \pi_{il} f_{\text{GHD},T}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_{\eta l}, v_l)},$$

as well as the following conditional expectations, which are similar to those derived in the E-step of parameter estimation for the GHD-GMM:

$$E_{4ik}^* := \mathbb{E}\left[\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1\right] = \mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + E_{1ik}^* \boldsymbol{\beta}_{\eta k}) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i),$$

$$E_{5ik}^* := \mathbb{E}\left[(1/W_{ik})\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1\right] = E_{2ik}^* \mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i) + \mathbf{V}_k \boldsymbol{\Psi}_k^{-1} \boldsymbol{\beta}_{\eta k},$$

$$\begin{aligned} E_{6ik}^* := \mathbb{E}[(1/W_{ik})\boldsymbol{\eta}_i \boldsymbol{\eta}_i^{'} \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1] &= \mathbf{V}_k + \mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i) \boldsymbol{\beta}_{\eta k} \boldsymbol{\Psi}_k^{-1} \mathbf{V}_k \\ &+ E_{2ik}^* \mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i)(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i)^{'} \mathbf{V}_k, \\ &+ \mathbf{V}_k \boldsymbol{\Psi}_k^{-1} \boldsymbol{\beta}_{\eta k}^{'} (\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'} \boldsymbol{\Theta}_k^{-1} \mathbf{y}_i)^{'} \mathbf{V}_k + E_{1ik}^* \mathbf{V}_k \boldsymbol{\Psi}_k^{-1} \boldsymbol{\beta}_{\eta k} \boldsymbol{\beta}_{\eta k}^{'} \boldsymbol{\Psi}_k^{-1} \mathbf{V}_k. \end{aligned}$$

The M-step requires the computation of the parameter updates to maximize the expected value of the complete-data log-likelihood. In this step, the parameter updates for $\boldsymbol{\alpha}_c, \boldsymbol{\Gamma}_c, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_{\eta k}, \boldsymbol{\Theta}_k, \boldsymbol{\Psi}_k$ and $\boldsymbol{\Gamma}_k$ are obtained in closed form and are similar to those derived in the M-step of parameter estimation for the GHD-GMM, hence are omitted here. We solve the equation

$$\log\left(\frac{\nu_k}{2}\right) + 1 - \varphi\left(\frac{\nu_k}{2}\right) - \frac{1}{n_k}\sum_{i=1}^{n} \tau_{ik}\left(E^*_{3ik} + E^*_{2ik}\right) = 0 \tag{3.18}$$

for $\nu_k$, numerically, where $n_k = \sum_{i=1}^{n} \tau_{ik}$. Parameter estimation for Models II and IV is outlined in Appendix A.2.

## 3.4 Illustrations

### 3.4.1 Alcoholic consumption data from the National Longitudinal Survey of Youth

The National Longitudinal Survey of Youth (NLSY) is a longitudinal study conducted by the United States Bureau of Labor Statistics with the goal of understanding the interaction between labor force participation, education, and health behaviors in children and adolescents. The sample for this study was a cohort of children who were between the ages of 12 and 17 when first interviewed in 1997. The data of interest were gathered each year between 1997 and 2011 and again in 2013 (15 total possible interviews). Each respondent provided a number between 0 and 98 that represents the number of alcoholic drinks they typically consume on a given day on which they are drinking. Tracking heterogeneity in trajectories of alcohol consumption has been

a frequent topic in the social and biomedical sciences with many objectives: to determine the effect of excessive drinking over time, design interventions that can prevent excessive drinking, or model the interplay between consumption of different drugs. Because we are interested in modeling drinking behaviour over the life span, the data are shifted from representing year of interview to age. To minimize the effect of missing data, individuals who were interviewed between the ages of 16 and 19 were used for the following analyses. In this analysis, the time invariant covariates are not considered.

We implement the Gaussian GMM via `Mplus` Version 7.3 (Muthén and Muthén, 2012). Our proposed GHD-GMM and GST-GMM are implemented in `R` and run with $K$ ranging from 1 to 10 until the best model is obtained under each scenario. Table 3.1 shows the results of fitting all of the models as aforementioned for a varying number of latent classes. The BIC values show that more than eight classes are needed with the conventional GMM, two are needed with constrained Models I and IV, and three are needed for all of the other models. The BIC values for the GST-GMM and GHD-GMM are better than the BIC for the Gaussian GMM. Notably, the BIC values for the GHD-GMM do not always improve on that for the GST-GMM. Among all fitted models, the three-cluster general GST-GMM under $\boldsymbol{\beta}_{yk} = \mathbf{0}$ (i.e., general Model III) is preferable according to the BIC. It is worth mentioning that, even though the skew-t distribution is a special case of the generalized hyperbolic distribution, the GST-GMM seems to be useful in addition to the GHD-GMM.

The best-fitting model, the three-cluster skew-t, breaks the data into three groups. The first group, comprising 56% of the population, begins with low-moderate drinking ($< 1$ drink per drinking day), slightly increases during adolescence, and by age 19

Table 3.1: Results of fitting Gaussian, GST, and GHD GMMs for consumption data from the National Longitudinal Survey of Youth.

| | GMM-Gaussian (constrained) | | | GMM-Gaussian (general) | | |
|---|---|---|---|---|---|---|
| Classes | Log-likelihood | Free paras | BIC | Log-likelihood | Free paras | BIC |
| 1 | -14983.42 | 9 | -30030.27 | -14983.42 | 9 | -30030.27 |
| 2 | -14623.41 | 12 | -29331.40 | -12671.88 | 19 | -25477.69 |
| 3 | -14330.00 | 15 | -28765.72 | -12233.95 | 29 | -24672.31 |
| 4 | -14182.66 | 18 | -28492.19 | -12119.21 | 39 | -24513.30 |
| 5 | -14076.42 | 21 | -28300.85 | -12027.60 | 49 | -24400.58 |
| 6 | -14015.58 | 24 | -28200.32 | -11950.97 | 59 | -24317.78 |
| 7 | -13980.78 | 27 | -28151.86 | -11906.09 | 69 | -24298.53 |
| 8 | -13937.17 | 30 | -28085.80 | -11870.44 | 79 | -24297.69 |
| 9 | -13916.40 | 33 | -28065.38 | | | |
| | GHD-GMM (Model I, constrained) | | | GHD-GMM (Model I, general) | | |
| Classes | Log-likelihood | Free paras | BIC | Log-likelihood | Free paras | BIC |
| 1 | -12403.20 | 13 | -24898.19 | -12403.28 | 13 | -24898.19 |
| 2 | -12315.75 | 16 | -24744.27 | -12119.53 | 27 | -24429.36 |
| 3 | -12315.50 | 19 | -24764.91 | -11958.92 | 41 | -24206.82 |
| 4 | | | | -11953.84 | 55 | -24295.33 |
| | GHD-GMM (Model II, constrained) | | | GHD-GMM (Model II, general) | | |
| Classes | Log-likelihood | Free paras | BIC | Log-likelihood | Free paras | BIC |
| 1 | -12399.68 | 15 | -24883.94 | -12399.68 | 15 | -24905.09 |
| 2 | -12312.27 | 18 | -24737.32 | -12166.47 | 31 | -24551.45 |
| 3 | -12288.26 | 21 | -24717.49 | -12002.12 | 47 | -24335.51 |
| 4 | -12287.98 | 24 | -24745.12 | -11956.47 | 63 | -24356.99 |
| | GST-GMM (Model III, constrained) | | | GST-GMM (Model III, general) | | |
| Classes | Log-likelihood | Free paras | BIC | Log-likelihood | Free paras | BIC |
| 1 | -12421.85 | 12 | -24928.28 | -12421.92 | 12 | -24928.42 |
| 2 | -12352.31 | 15 | -24810.34 | -12151.6 | 25 | -24479.41 |
| 3 | -12340.61 | 18 | -24808.10 | -11966.84 | 38 | -24201.52 |
| 4 | -12348.28 | 21 | -24844.58 | -11925.67 | 51 | -24210.82 |
| | GST-GMM (Model IV, constrained) | | | GST-GMM (Model IV, general) | | |
| Classes | Log-likelihood | Free paras | BIC | Log-likelihood | Free paras | BIC |
| 1 | -12418.18 | 14 | -24935.05 | -12418.19 | 14 | -24935.05 |
| 2 | -12348.01 | 17 | -24748.06 | -12118.12 | 29 | -24440.64 |
| 3 | -12347.48 | 20 | -24756.18 | -11990.60 | 44 | -24291.33 |
| 4 | | | | -11938.08 | 59 | -24292.00 |

the average drinks per drinking day is at about 1. These can be considered "consistent low" drinkers. Although the intercept for this class is heavily positively skewed

(intercept skewness = 2.59), the slope is not (slope skewness = 0.03), which indicates that the individual slopes are nearly normally distributed around the class slope of 0.21. The second class, comprising 23% of the population, are what will be called the "decreasing" drinkers. This class has an intercept of around five drinks per drinking day (a drinking binge) and ends at about 3 drinks per drinking day (just below the amount considered a drinking binge).[1] The intercept is again positively skewed (intercept skewness = 2.90) but the slope is negatively skewed (slope skewness = $-0.78$), suggesting that individuals in this class decrease their consumption quickly over the period of adolescence. The third class, comprising 20% of the population, will be called the "increasing moderate" drinkers. Their initial level of drinking is around 2.87 drinks per drinking day (less than a binge) and this increases during adolescence, ending at age 19 around 7 drinks per drinking day (far above a drinking binge). Both the slope and intercept are slightly positively skewed (intercept skewness = 0.48, slope skewness = 0.41).

These results suggest that, during adolescence, which is typically a time when alcohol consumption is initiated, individuals will have different reactions to the exposure to alcohol given their previous experience – those individuals who are low drinkers will tend to continue to be low drinkers, those who have already consumed alcohol heavily will begin to taper back to safe levels (alluding to these individuals "knowing their limits" when it comes to alcohol), and those who are only at moderate levels tend to increase to heavy drinking. This model can be useful because it indicates which 15-year-olds should be the target of interventions if the goal is to prevent

[1]The World Health Organization defines heavy episodic drinking (also called a drinking "binge") as the consumption of 60 or more grams of alcohol on one occasion (`www.who.int/gho/alcohol/consumption_patterns/heavy_episodic_drinkers_text/en/`), which is about four standard drinks (`www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/what-standard-drink`).

heavy drinking in late adolescence. Although the high drinkers may appear to be the most likely to develop problems related to alcohol, they may "grow out" of their alcohol consumption; therefore, it may be better to focus efforts on the 15-year-olds that only drink at moderate levels.

Other models tend to find a similar developmental pattern. For instance, the three-class GHD model (which has a very similar BIC to the three-class skew-t model) demonstrates a similar pattern in the clusters: one consistent low, one high but decreasing, and one moderate and increasing. This suggests that the same pattern endures regardless of the distributional assumptions. However, the cluster proportions differ slightly (58%, 24%, and 17%, for the low, high/decreasing, and moderate/increasing classes, respectively), which seems to suggest that the GHD model classifies more individuals into the "low" class than the skew-t model. If the goal of the analysis is to identify groups to target for interventions for the prevention of alcoholism, the proportions found in the skew-t model might be preferred as they create population groups that are larger. Therefore, interventions targeting this group may have a greater impact on the population than those targeting a smaller group.

### 3.4.2 Simulation Studies

In addition to the real data application of our proposed model, simulation studies are carried out to further illustrate the performance of the proposed GST-GMM and GST-GMM models in recovering the underlying model parameters and the clustering ability. We use the relationship between the generalized hyperbolic distribution and the Gaussian distribution (cf. Section 2.3) to generate GHD-GMM and GST-GMM data. Data are generated in a number of scenarios: linear and quadratic GMMs with

different distributions of the measurement errors and random effects, resulting in four distinct simulation experiments. We analyzed the data in several different ways: using a Gaussian distribution via `Mplus`, using the proposed constrained and general GST-GMM and GHD-GMM models under $\boldsymbol{\beta}_{yk} = \mathbf{0}$, and using the proposed constrained and general GST-GMM and GHD-GMM models under $\boldsymbol{\beta}_{\eta k} = \mathbf{0}$. The purpose of analyzing the simulated data in this way is to compare our proposed GHD-GMM and GST-GMM models with the Gaussian GMM developed by Muthén and colleagues, which dominates the literature on GMMs.

In the first simulation experiment, the dataset is generated by a two-class general GHD-GMM under $\boldsymbol{\beta}_{\eta k}$ with linear growth and five time points ($n_1 = n_2 = 400$). In the second simulation experiment, the dataset is generated by a two-class general GHD-GMM under $\boldsymbol{\beta}_{yk}$ with quadratic growth and thirty time points ($n_1 = n_2 = 200$). In the third simulation experiment, the simulated dataset is generated by a three-class general GST-GMM under $\boldsymbol{\beta}_{\eta k}$ with linear growth and eight time points ($n_1 = n_2 = n_3 = 500$). In the fourth simulation experiment, the dataset is generated by a two-class general GST-GMM under $\boldsymbol{\beta}_{yk}$ with quadratic growth and twenty time points ($n_1 = n_2 = 500$). Individual trajectories for these four simulation experiments are plotted in Figure 3.1.

First, we evaluate the ability of our proposed model to recover underlying parameters. To this end, 100 datasets are generated for each of the four simulation experiments. True values and the means of the parameter estimates with their associated standard deviations are summarized in Tables 3.2–3.5. The results for each of the four simulation experiments show that the means of parameter estimates are close to the true values with small standard deviations; hence, our proposed approach

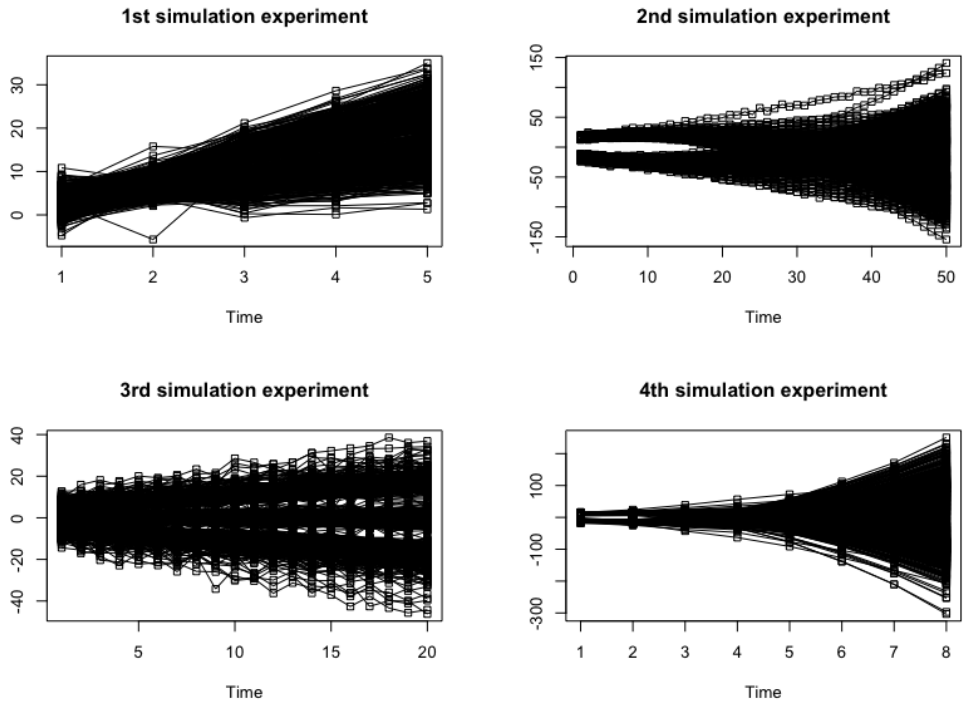


Figure 3.1: Individual observation trajectories plots for the four simulation experiments.

is effective at parameter recovery.

Second, we compare the two formulations for extending GMMs. One hundred datasets are generated for the four simulation experiments above and analyzed using GMMs developed herein. The means of the BIC, the ARI, and the misclassification rates (ERR) are summarized in Table 3.6. As anticipated, the best models obtained are those with underlying true data structures. There is, however, one disadvantage with Model II and Model IV. In terms of model complexity, Model II has $K(T-q)$ more parameters than Model I, and Model IV has $K(T-q)$ more parameters than Model III. Hence, Models II and IV need larger sample sizes as small class sizes

Table 3.2: Key model parameters as well as means and standard deviations of the associated parameter estimations from the 100 runs for the first simulation experiment (Model II).

| | True values | Means | Standard deviations |
|---|---|---|---|
| $\boldsymbol{\alpha}_1$ | $(4, 5)'$ | $(4.00, 5.00)'$ | $(0.11, 0.08)'$ |
| $\boldsymbol{\alpha}_2$ | $(2, 3)'$ | $(1.99, 2.98)'$ | $(0.18, 0.11)'$ |
| $\boldsymbol{\beta}_1$ | $(1, -1, 1, 1, 1)'$ | $(0.94, -0.92, 0.93, 0.94, 0.93)'$ | $(0.35, 0.33, 0.39, 0.44, 0.51)'$ |
| $\boldsymbol{\beta}_2$ | $(-1, 1, -1, -1, -1)'$ | $(-0.77, 0.83, -0.72, -0.68, -0.67)'$ | $(0.27, 0.36, 0.33, 0.43, 0.55)$ |
| $\lambda_1$ | -1 | -0.7 | 0.73 |
| $\lambda_2$ | -2 | -1.05 | 0.85 |
| $\omega_1$ | 2 | 2.02 | 0.32 |
| $\omega_2$ | 3 | 3.11 | 0.58 |
| $\boldsymbol{\Psi}_1$ | $\begin{bmatrix} 1 & 0 \\ 0 & 0.7 \end{bmatrix}$ | $\begin{bmatrix} 0.92 & 0.00 \\ 0.00 & 0.65 \end{bmatrix}$ | $\begin{bmatrix} 0.34 & 0.08 \\ 0.08 & 0.22 \end{bmatrix}$ |
| $\boldsymbol{\Psi}_2$ | $\begin{bmatrix} 1.5 & 0 \\ 0 & 0.8 \end{bmatrix}$ | $\begin{bmatrix} 1.20 & -0.01 \\ -0.01 & 0.63 \end{bmatrix}$ | $\begin{bmatrix} 0.36 & 0.08 \\ 0.08 & 0.18 \end{bmatrix}$ |

Table 3.3: Key model parameters as well as means and standard deviations of the associated parameter estimations from the 100 runs for the second simulation experiment (Model I).

| | True values | Means | Standard deviations |
|---|---|---|---|
| $\boldsymbol{\alpha}_1$ | $(15, 8, -6)'$ | $(14.98, 8.00, -6.00)'$ | $(0.10, 0.09, 0.15)'$ |
| $\boldsymbol{\alpha}_2$ | $(-14, -10, 6)'$ | $(-14.06, -9.96, 6.02)'$ | $(0.24, 0.20, 0.16)'$ |
| $\boldsymbol{\beta}_1$ | $(1, 1, 1)'$ | $(1.05, 1.01, 1)'$ | $(0.35, 0.32, 0.39)'$ |
| $\boldsymbol{\beta}_2$ | $(-1, -1, -1)'$ | $(-0.89, -0.95, -0.95)'$ | $(0.24, 0.25, 0.27)'$ |
| $\lambda_1$ | -1 | -0.52 | 0.70 |
| $\lambda_2$ | 2 | 1.05 | 1.28 |
| $\omega_1$ | 2 | 2.02 | 0.31 |
| $\omega_2$ | 3 | 2.92 | 0.51 |
| $\boldsymbol{\Psi}_1$ | $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ | $\begin{bmatrix} 0.99 & 0.00 & 0.00 \\ 0.00 & 0.70 & 0.00 \\ 0.00 & 0.00 & 1.99 \end{bmatrix}$ | $\begin{bmatrix} 0.33 & 0.07 & 0.12 \\ 0.07 & 0.21 & 0.11 \\ 0.12 & 0.11 & 0.62 \end{bmatrix}$ |
| $\boldsymbol{\Psi}_2$ | $\begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.9 \end{bmatrix}$ | $\begin{bmatrix} 1.38 & 0.00 & 0.01 \\ 0.00 & 0.74 & -0.01 \\ 0.01 & -0.01 & 0.84 \end{bmatrix}$ | $\begin{bmatrix} 0.39 & 0.08 & 0.08 \\ 0.08 & 0.21 & 0.08 \\ 0.08 & 0.08 & 0.21 \end{bmatrix}$ |

can create problems, such as singularity of the covariance matrix and slow or non-convergence of the EM algorithm.

Table 3.4: Key model parameters as well as means and standard deviations of the associated parameter estimations from the 100 runs for the third simulation experiment (Model IV).

| | True values | Means | Standard deviations |
|---|---|---|---|
| $\boldsymbol{\alpha}_1$ | $(4, 5)'$ | $(4.01, 4.98)'$ | $(0.11, 0.10)'$ |
| $\boldsymbol{\alpha}_2$ | $(0, 0)'$ | $(-0.01, 0.01)'$ | $(0.07, 0.08)'$ |
| $\boldsymbol{\alpha}_3$ | $(-4, -5)'$ | $(-3.91, -4.9)'$ | $(0.81, 1.01)'$ |
| $\boldsymbol{\beta}_1$ | $(1, 1)'$ | $(1.00, 1.01)'$ | $(0.10, 0.09)'$ |
| $\boldsymbol{\beta}_2$ | $(0, 0)'$ | $(-0.01, -0.02)'$ | $(0.17, 0.19)'$ |
| $\boldsymbol{\beta}_3$ | $(-1, -1)'$ | $(-0.99, -0.98)$ | $(0.24, 0.22)'$ |
| $\nu_1$ | 7 | 7.09 | 0.61 |
| $\nu_2$ | 5 | 4.97 | 0.41 |
| $\nu_3$ | 6 | 6.08 | 0.50 |
| $\boldsymbol{\Psi}_1$ | $\begin{bmatrix} 1 & 0 \\ 0 & 0.7 \end{bmatrix}$ | $\begin{bmatrix} 1.00 & 0.01 \\ 0.01 & 0.68 \end{bmatrix}$ | $\begin{bmatrix} 0.07 & 0.05 \\ 0.05 & 0.07 \end{bmatrix}$ |
| $\boldsymbol{\Psi}_2$ | $\begin{bmatrix} 0.7 & 0 \\ 0 & 0.6 \end{bmatrix}$ | $\begin{bmatrix} 0.72 & 0.03 \\ 0.03 & 0.64 \end{bmatrix}$ | $\begin{bmatrix} 0.28 & 0.32 \\ 0.32 & 0.40 \end{bmatrix}$ |
| $\boldsymbol{\Psi}_3$ | $\begin{bmatrix} 1.5 & 0 \\ 0 & 0.8 \end{bmatrix}$ | $\begin{bmatrix} 1.36 & 0.00 \\ 0.00 & 0.76 \end{bmatrix}$ | $\begin{bmatrix} 0.27 & 0.07 \\ 0.07 & 0.08 \end{bmatrix}$ |

Table 3.5: Key model parameters as well as means and standard deviations of the associated parameter estimations from the 100 runs for the fourth simulation experiment (Model III).

| | True values | Means | Standard deviations |
|---|---|---|---|
| $\boldsymbol{\alpha}_1$ | $(8, 7, -3)'$ | $(7.95, 7.01, -2.93)'$ | $(0.18, 0.11, 0.05)'$ |
| $\boldsymbol{\alpha}_2$ | $(-8, -7, 3)'$ | $(-7.98, -6.99, 2.93)'$ | $(0.21, 0.12, 0.05)'$ |
| $\boldsymbol{\beta}_1$ | $(1, 1, 1, 1,$ $1, 1, 1, 1)'$ | $(1.04, 0.98, 0.78, 0.44,$ $-0.02, -0.62, -1.35, -2.21)'$ | $(0.16, 0.17, 0.18, 0.18,$ $0.18, 0.18, 0.22, 0.32)'$ |
| $\boldsymbol{\beta}_2$ | $(-1, -1, -1, -1,$ $-1, -1, -1, -1)'$ | $(-1.02, -0.96, -0.78, -0.46,$ $-0.01, 0.56, 1.26, 2.09)'$ | $(0.17, 0.15, 0.16, 0.16,$ $0.18, 0.22, 0.29, 0.40)'$ |
| $\nu_1$ | 7 | 7.43 | 0.80 |
| $\nu_2$ | 6 | 6.27 | 0.59 |
| $\boldsymbol{\Psi}_1$ | $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0 & 0.8 \end{bmatrix}$ | $\begin{bmatrix} 0.99 & 0.00 & 0.01 \\ 0.00 & 0.71 & 0.00 \\ 0.01 & 0.00 & 0.80 \end{bmatrix}$ | $\begin{bmatrix} 0.11 & 0.07 & 0.05 \\ 0.07 & 0.06 & 0.04 \\ 0.05 & 0.04 & 0.06 \end{bmatrix}$ |
| $\boldsymbol{\Psi}_2$ | $\begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.9 \end{bmatrix}$ | $\begin{bmatrix} 1.49 & 0.01 & 0.01 \\ 0.01 & 0.79 & 0.01 \\ 0.01 & 0.01 & 0.90 \end{bmatrix}$ | $\begin{bmatrix} 0.20 & 0.10 & 0.08 \\ 0.10 & 0.09 & 0.04 \\ 0.08 & 0.04 & 0.07 \end{bmatrix}$ |

Table 3.6: Comparison of results — including average BIC, ARI, and EER values — for Models I–IV.

| | Model I | | | | Model II | | | |
|---|---|---|---|---|---|---|---|---|
| | Free paras | BIC | ARI | ERR | Free paras | BIC | ARI | ERR |
| 1st simulation (model II) | 29 | –16018.99 | 0.92 | 0.02 | 35 | –15366.59 | 0.97 | 0.01 |
| 2nd simulation (model I) | 89 | –69895.48 | 1.00 | 0.00 | 143 | –70507.17 | 1.00 | 0.00 |
| | Model IV | | | | Model III | | | |
| | Free paras | BIC | ARI | ERR | Free paras | BIC | ARI | ERR |
| 3rd simulation (model IV) | 50 | –101669.80 | 1.00 | 0.00 | 68 | –101587.60 | 1.00 | 0.00 |
| 4th simulation (model III) | 67 | –37271.48 | 1.00 | 0.00 | 101 | –37265.55 | 1.00 | 0.00 |

Finally, we compare our proposed Models I and III with the Gaussian GMMs (via `Mplus`). Herein, 100 datasets are generated, as described before, for the second simulation experiment and analyzed with different distributions: the generalized hyperbolic distribution via our proposed Model I, the multivariate skew-t distribution via our proposed Model III, and Gaussian GMMs via the `Mplus` software. Table 3.7 summarizes the percentage of the replications favoured by the BIC when analyzing those 100 generated datasets for 1–6 latent classes (note that 6 latent classes were never selected, cf. Table 3.7). It is not surprising that the Gaussian GMMs overestimate the number of classes, pointing to five classes 33% of the time and to four classes 67% of the time. Because the true data are generated from the generalized hyperbolic distribution, the Gaussian GMMs need to extract more latent classes to capture the observed variable distribution. The generalized hyperbolic and skew-t GMMs (i.e., the proposed Models I and III) correctly choose the number of classes 100% of the time. It is noteworthy to mention that the best model, based on the BIC, is a two-class generalized hyperbolic GMM consistently. Typical analysis results for

fitting the Gaussian, the multivariate skew-t, and the generalized hyperbolic GMMs to the second simulation experiment are given in Table 3.8.

Table 3.7: Percent preferred by the BIC when analyzing the second simulation experiment with Model I, Model III, and GMM along with number of classes.

| | Number of classes | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Model I | 0 | 100 | 0 | 0 | 0 |
| Model III | 0 | 100 | 0 | 0 | 0 |
| GMM | 0 | 0 | 0 | 67 | 33 |

Table 3.8: Results of fitting the Gaussian, GST, and GHD to the second simulation experiment.

| Model | $K$ | Free paras | Log-likelihood | BIC | ARI | ERR |
|---|---|---|---|---|---|---|
| Model I | 1 | 64 | −35261.30 | −70906.06 | 0 | 0.50 |
| | 2 | 129 | −34283.40 | −69339.69 | 1 | 0 |
| | 3 | 194 | −34231.13 | −69624.60 | 0.86 | 0.08 |
| Model III | 1 | 63 | −35295.15 | −70967.76 | 0 | 0.5 |
| | 2 | 127 | −34292.80 | −69346.20 | 1 | 0 |
| | 3 | 191 | −34240.11 | −69624.59 | 0.86 | 0.09 |
| GMM | 1 | 59 | −38474.99 | −77303.47 | 0 | 0.50 |
| | 2 | 119 | −35524.70 | −71762.39 | 0.77 | 0.06 |
| | 3 | 179 | −34872.72 | −70817.91 | 0.51 | 0.29 |
| | 4 | 239 | −34544.54 | −70521.04 | 0.52 | 0.39 |
| | 5 | 299 | −34356.91 | −70505.26 | 0.46 | 0.45 |
| | 6 | 359 | −34252.98 | −70656.90 | 0.44 | 0.46 |

## 3.5 Discussion

We have introduced novel GHD-GMM and GST-GMM models, which are extensions of the GMMs introduced by Verbeke and Lesaffre (1996) to the generalized hyperbolic and skew-t distributions, respectively, to facilitate heavier tails or asymmetry.

Updates are derived for parameter estimation within the EM algorithm framework in the model-based clustering context, which is made feasible by the fact that the generalized hyperbolic distribution can be represented as a normal mean-variance mixture, where the weight follows a GIG distribution. In our extension GMMs, four models were considered (GHD-GMM and GST-GMM under $\boldsymbol{\beta}_{yk} = \mathbf{0}$, GHD-GMM and GST-GMM under $\boldsymbol{\beta}_{\eta k} = \mathbf{0}$) and their performance was compared using simulated and real data. In terms of interpretation, GHD-GMM under $\boldsymbol{\beta}_{\eta k} = \mathbf{0}$ is better than GHD-GMM under $\boldsymbol{\beta}_{yk} = \mathbf{0}$ because the skewness parameters are in the data space and the interpretation of the skewness parameters is clear. However, naturally, models with a minimal number of parameters would be preferable. Hence, in terms of model complexity, GHD-GMM under $\boldsymbol{\beta}_{yk} = \mathbf{0}$ is preferable to GHD-GMM under $\boldsymbol{\beta}_{\eta k} = \mathbf{0}$, because the former model has $K(T - q)$ fewer parameters than the latter.

We believe that this kind of mixture modeling approach for longitudinal data is important in many biostatistical and psychological applications, allowing accurate inference of model parameters and class membership probabilities while adjusting for heterogeneity, heavy tails, and skewness in the data. The proposed GHD-GMM and GST-GMM models have several advantages over Gaussian GMMs. The proposed GHD-GMM, which includes the multivariate skew-t, variance-gamma distribution, multivariate Gaussian distributions, etc., as special or limiting cases, provides flexibility to handle a broader range of multivariate longitudinal data.

The models proposed herein can be further developed in various ways. First, for the first level of the GMM, only $q$th order polynomial equations are considered, and so kernel regressions or non-linear regressions could be incorporated into the model. Second, Bayesian mixture modeling may offer researchers an alternative way to handle

clustering of longitudinal data due to the popularity and advances in Markov chain Monte Carlo techniques. Finally, it is also worthwhile to consider more general parametric distributions of measurement errors and random effects, such as the coalesced generalized hyperbolic distribution and the multiple scaled generalized hyperbolic distribution (Tortora et al., 2014, 2016).

# Chapter 4

# Growth Mixture Model Analysis with Continuous Non-Elliptical Random Effects for Incomplete Data

## 4.1 Introduction

In this chapter, we detail the methodological development of growth mixture models with continuous non-elliptical random effects designed for analyzing longitudinal data in the presence of arbitrary missing values and is split into 4 sections. Section 4.2 describes the growth mixture models with continuous non-elliptical random effects for incomplete data, in Section 4.3 outlines the ML estimation via an expectation-maximization algorithm, in Section 4.4 both simulated and real data analyses are

used to illustrated our approach and in Section 4.5 the chapter gives a discussion of the work.

Before proceeding to the growth mixture model with non-elliptical distribution with missing information, let us clarify some notations used in the following sections. As we mentioned in Section 2.2, in the missing data literature, the data are often partitioned into two parts: the observed component ($\mathbf{Y}_i^o$) and the missing component ($\mathbf{Y}_i^m$) with dimensions $T_i^o \times 1$ and $T_i^m \times 1$, respectively, where $T_i^o + T_i^m = T$. For a fully observed data point $\mathbf{Y}_i$, $\mathbf{Y}_i^m$ does not exist. To facilitate computation, following Finkbeiner (1979) and Lin et al. (2006), we introduce two missingness indicator matrices, denoted by $\mathbf{O}_i$ ($T_i^o \times T$) and $\mathbf{M}_i$ ($T_i^m \times T$). For each $i = 1, \ldots, n$, $\mathbf{Y}_i^o$ and $\mathbf{Y}_i^m$ are related to $\mathbf{Y}_i$ by $\mathbf{Y}_i^o = \mathbf{O}_i\mathbf{Y}_i$ and $\mathbf{Y}_i^m = \mathbf{M}_i\mathbf{Y}_i$, respectively. Specifically, $\mathbf{O}_i$ and $\mathbf{M}_i$ can be formed by extracting from a $T$-dimensional identity matrix $\mathbf{I}_T$ corresponding to the respective row positions of $\mathbf{Y}_i^o$ and $\mathbf{Y}_i^m$ in $\mathbf{Y}_i$. It is easy to verify that $\mathbf{Y}_i = \mathbf{O}_i'\mathbf{Y}_i^o + \mathbf{M}_i'\mathbf{Y}_i^m$ and $\mathbf{O}_i'\mathbf{O}_i + \mathbf{M}_i'\mathbf{M}_i = \mathbf{I}_T$.

## 4.2 Model Description

In the previous chapter, we developed four statistical models when extending the GMMs to have non-elliptical distributions. We compared the two formulations for extending GMMs and demonstrated that GMMs with non-elliptical random effects (i.e., Model I and Model III) are preferable to GMMs with non-elliptical measurement errors (i.e., Model II and Model IV). Therefore, we only consider to generalize GMMs with non-elliptical random effects to accommodate missing values. For a better notational convenience, we rewrite the GMMs with non-elliptical random effects via GHD and GST herein. The growth mixture model with generalized hyperbolic

random effects (abbreviated as GHD-GMM) can be written as

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik} f_{\text{GHD},T}(\mathbf{y}_i; \lambda_k, \omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_k), \tag{4.1}$$

where $\boldsymbol{\mu}_k = \boldsymbol{\Lambda}_y(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i)$ and $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k$. Similary, the growth mixture model with multivariate skew-t random effects (abbreviated as GST-GMM) can be written as

$$p(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik} f_{\text{GST},T}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Lambda}_y \boldsymbol{\beta}_k, \nu_k), \tag{4.2}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are as defined above.

### 4.2.1 GHD-GMM with missing infromation

Apply the above standard set-up to a $K$-component GHD-GMM as defined in (4.1), then we obtain the marginal distribution of the observed data $\mathbf{y}_i^o$ is

$$p(\mathbf{y}_i^o \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik} f_{\text{GHD},T_i^o}(\mathbf{y}_i^o; \lambda_k, \omega_k, \boldsymbol{\mu}_k^o, \boldsymbol{\Sigma}_k^{oo}, \boldsymbol{\beta}_k^o), \tag{4.3}$$

where $\boldsymbol{\mu}_k^o = \mathbf{O}_i \boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k^{oo} = \mathbf{O}_i \boldsymbol{\Sigma}_k \mathbf{O}_i^{'}$, $\boldsymbol{\beta}_k^o = \mathbf{O}_i \boldsymbol{\Lambda}_y \boldsymbol{\beta}_k$, and the observed-data log-likelihood function becomes

$$l_o(\boldsymbol{\vartheta} \mid \mathbf{y}_i^o, \mathbf{x}_i) = \sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_{ik} f_{\text{GHD},T_i^o}(\mathbf{y}_i^o; \lambda_k, \omega_k, \boldsymbol{\mu}_k^o, \boldsymbol{\Sigma}_k^{oo}, \boldsymbol{\beta}_k^o) \right), \tag{4.4}$$

where $\boldsymbol{\vartheta} = (\pi_{ik}, \lambda_k, \omega_k, \boldsymbol{\alpha}_c, \boldsymbol{\Gamma}_c, \boldsymbol{\alpha}_k, \boldsymbol{\Gamma}_k, \boldsymbol{\beta}_k, \boldsymbol{\Theta}_k, \boldsymbol{\Psi}_k)$ is the vector of parameters. However, it is very difficult to obtain ML estimates by directly working on the maximization of $l_o(\boldsymbol{\vartheta} \mid \mathbf{y}^o, \mathbf{x}_i)$. To compute the ML estimates of unknown parameters involved

in (4.3), we make use of the EM algorithm (Dempster et al., 1977).

To facilitate the ML estimation of the GHD-GMM model with missing information via the EM algorithm, we will augment the observed data $\mathbf{y}_i^o$ and $\mathbf{x}_i$ with the truly missing data $\mathbf{y}_i^m$ together with the unobserved or latent variables $\boldsymbol{\eta}_i, \mathbf{c}_i$, and $W_i$. let $\mathbf{c}_i = (c_{i1}, \ldots, c_{iK})$ for $i = 1, \ldots, n$ be the set of allocation vectors or class membership indicators, where the component membership $c_{ik} = 1$ if $\mathbf{y}_i$ belongs to the $k$th component and $c_{ik} = 0$ otherwise. It follows that $\mathbf{c}_i$ follows a multinomial distribution with one trial and cell probabilities $\pi_{i1}, \ldots, \pi_{iK}$ in connection with (2.4).

Subsequently, in this 'new' complete-data framework, we establish the following proposition, which is essential to evaluate the required conditional expectations in the E-step of the EM algorithm described in Section 4.3.

**Proposition 4.2.1.** *Given* (2.1), (2.2), $w_i \mid c_{ik} = 1 \sim \mathcal{I}(\lambda_k, \omega_k)$, *and let the* $\mathbf{y}_i^o$ *and* $\mathbf{y}_i^m$ *be the observed and missing components of* $\mathbf{y}_i$, *respectively. we have the following conditional distributions:*

*a. The conditional distribution of* $\mathbf{y}_i^o$ *given* $\mathbf{x}_i$, $W_i$, *and* $c_{ik} = 1$ *is*

$$\mathbf{y}_i^o \mid \mathbf{x}_i, w_i, c_{ik} = 1 \sim \mathcal{N}\left(\boldsymbol{\mu}_k^o + w_i \mathbf{O}_i \boldsymbol{\Lambda}_y \boldsymbol{\beta}_i, w_i \mathbf{O}_i \boldsymbol{\Sigma}_k \mathbf{O}_i^{'}\right).$$

*b. The conditional distribution of* $W$ *given* $\mathbf{Y}_i^o$, $\mathbf{x}_i$, *and* $c_{ik} = 1$ *is*

$$W_i \mid \mathbf{Y}_i^o, \mathbf{x}_i, c_{ik} = 1 \sim GIG\left(\lambda_{ik}^*, \chi_{ik}^*, \psi_{ik}^*\right),$$

*where* $\lambda_{ik}^* = \lambda_k - \frac{T_i^o}{2}$, $\chi_{ik}^* = \omega_k + \delta(\mathbf{y}_i^o, \boldsymbol{\mu}_k^o \mid \boldsymbol{\Sigma}_k^{oo})$, $\psi_{ik}^* = \omega_k + \boldsymbol{\beta}_k^{'} \boldsymbol{\Lambda}_y^{'} \mathbf{S}_{ik}^{oo} \boldsymbol{\Lambda}_y \boldsymbol{\beta}_k$, *and* $\mathbf{S}_{ik}^{oo} = \mathbf{O}_i^{'}(\mathbf{O}_i \boldsymbol{\Sigma}_k^{oo} \mathbf{O}_i)^{-1} \mathbf{O}_i$.

*c. The conditional distribution of* $\mathbf{y}_i^m$ *given* $\mathbf{y}_i^o$, $\mathbf{x}_i$, $W_i$, *and* $c_{ik} = 1$ *is*

$$\mathbf{y}_i^m \mid \mathbf{y}_i^o, \mathbf{x}_i, w_i, c_{ik} = 1 \sim \mathcal{N}\Big(\mathbf{M}_i(\boldsymbol{\mu}_k + w_i\boldsymbol{\Lambda}_y\boldsymbol{\beta}_k) + \boldsymbol{\Sigma}_k\mathbf{S}_{ik}^{oo}(\mathbf{y}_i - \boldsymbol{\mu}_k - w_i\boldsymbol{\Lambda}_y\boldsymbol{\beta}_k), w_i\mathbf{M}_i(\mathbf{I}_T - \boldsymbol{\Sigma}_k\mathbf{S}_{ik}^{oo})\boldsymbol{\Sigma}_k\mathbf{M}_i^{'}\Big)$$

*d. The conditional distribution of* $\boldsymbol{\eta}_i$ *given* $\mathbf{y}_i^o$, $\mathbf{x}_i$, $W_i$, *and* $c_{ik} = 1$ *is*

$$\boldsymbol{\eta}_i \mid \mathbf{y}_i^o, \mathbf{x}_i, W_i, c_{ik} = 1 \sim \mathcal{N}\left(\mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i + w_i\boldsymbol{\beta}_k) + \boldsymbol{\Lambda}_y^{'}\mathbf{T}_{ik}^{oo}\mathbf{y}_i), w_i\mathbf{V}_k\right),$$

*where* $\mathbf{V}_k = (\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y\mathbf{T}_{ik}^{oo}\boldsymbol{\Lambda}_y)^{-1}$ *and* $\mathbf{T}_{ik}^{oo} = \mathbf{O}_i^{'}(\mathbf{O}_i\boldsymbol{\Theta}_k^{oo}\mathbf{O}_i)^{-1}\mathbf{O}_i$.

*e. The conditional distribution of* $\mathbf{y}_i^m$ *given* $\mathbf{y}_i^o$, $\boldsymbol{\eta}_i$, $W_i$, *and* $c_{ik} = 1$ *is*

$$\mathbf{y}_i^m \mid \mathbf{y}_i^o, \boldsymbol{\eta}_i, w_i, c_{ik} = 1 \sim \mathcal{N}\Big(\mathbf{M}_i\boldsymbol{\Lambda}_y\boldsymbol{\eta}_i + \mathbf{M}_i\boldsymbol{\Theta}_k\mathbf{T}_{ik}^{oo}(\mathbf{y}_i - \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i), w_i\mathbf{M}_i(\mathbf{I}_T - \boldsymbol{\Theta}_k\mathbf{T}_{ik}^{oo})\boldsymbol{\Theta}_k\mathbf{M}_i^{'}\Big),$$

*where* $\mathbf{T}_{ik}^{oo}$ *is as defined above .*

*Proof.* The proof is based on statistical properties of the multivariate Gaussian distribution and it is is straightforward, hence is omitted here. □

### 4.2.2 GST-GMM with missing information

Taking an analogous fashion to the GHD-GMM with missing information, for the GST-GMM with missing information, we obtain the marginal distribution of the

observed data $\mathbf{y}_i^o$ is

$$p(\mathbf{y}_i^o \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik} f_{\text{GST},T_i^o}(\mathbf{y}_i^o; \nu_k, \boldsymbol{\mu}_k^o, \boldsymbol{\Sigma}_k^{oo}, \boldsymbol{\beta}_k^o), \tag{4.5}$$

where $\boldsymbol{\mu}_k^o$, $\boldsymbol{\Sigma}_k^{oo}$, and $\boldsymbol{\beta}_k^o$ are as defined above in connection with (4.3). The observed-data log-likelihood function becomes

$$l_o(\boldsymbol{\vartheta} \mid \mathbf{y}_i^o, \mathbf{x}_i) = \sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_{ik} f_{\text{GST},T_i^o}(\mathbf{y}_i^o; \nu_k, \boldsymbol{\mu}_k^o, \boldsymbol{\Sigma}_k^{oo}, \boldsymbol{\beta}_k^o) \right), \tag{4.6}$$

where $\boldsymbol{\vartheta} = (\pi_{ik}, \nu_k, \boldsymbol{\alpha}_c, \boldsymbol{\Gamma}_c, \boldsymbol{\alpha}_k, \boldsymbol{\Gamma}_k, \boldsymbol{\beta}_k, \boldsymbol{\Theta}_k, \boldsymbol{\Psi}_k)$ is the vector of parameters in the case of the GST-GMM with missing information. Similarly, after augmenting the observed data $\mathbf{y}_i^o$ and $\mathbf{x}_i$ with the missing variables $\mathbf{y}_i^m, \boldsymbol{\eta}_i, \mathbf{c}_i$, and $W_i$, we obtain the following proposition. The other distributions in the proposition 2 are the same as in the GHD-GMM case, hence is omitted here.

**Proposition 4.2.2.** *Given* (2.1), (2.2), *and* $w_{ik} \sim \mathcal{I}(\frac{\nu_k}{2}, \frac{\nu_k}{2})$, *we obtain the conditional distribution of* $W$ *given* $\mathbf{y}_i^o$, $\mathbf{x}_i$, *and* $c_{ik} = 1$ *as*

$$W_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik} = 1 \sim GIG(\lambda_{ik}^\star, \chi_{ik}^\star, \psi_{ik}^\star),$$

*where* $\lambda_{ik}^\star = -\frac{\lambda_k + T_i^o}{2}$, $\chi_{ik}^\star = \nu_k + \delta(\mathbf{y}_i^o, \boldsymbol{\mu}_k^o \mid \boldsymbol{\Sigma}_k^{oo})$, $\psi_{ik}^\star = \boldsymbol{\beta}_k^{'} \boldsymbol{\Lambda}_y^{'} \mathbf{S}_{ik}^{oo} \boldsymbol{\Lambda}_y \boldsymbol{\beta}_k$, *and* $\mathbf{S}_{ik}^{oo}$ *is as defined above.*

## 4.3 Parameter estimation

### 4.3.1 The EM algorithm for GHD-GMM with missing information

For the GHD-GMM model with missing information, as mentioned previously, the complete-data are composed of the observed data $(\mathbf{y}_i^o, \mathbf{x}_i)$ and the missing or unobserved data $(\mathbf{y}_i^m, \boldsymbol{\eta}_i, \mathbf{c}_i, w_i)$. Hence, the complete-data log-likelihood function is

$$l_c(\boldsymbol{\vartheta} \,|\mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{x}_i, \boldsymbol{\eta}_i, \mathbf{c}_i, w_i) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\Big[\log\pi_{ik} + \log\phi(\mathbf{y}_i \mid \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i, w_{ik}\boldsymbol{\Theta}_k) + \log\phi(\boldsymbol{\eta}_i \mid \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}, w_{ik}\boldsymbol{\Psi}_k) + \log h(w_{ik} \mid \omega_k, \lambda_k)\Big]. \tag{4.7}$$

The E-step involves calculating the so-called $\mathcal{Q}$ function, which is the conditional expectation of the complete-data log-likelihood as in Equation (4.7) given the observed data $\mathbf{y}_i^o$ and $\mathbf{x}_i$ and the current estimates of $\hat{\boldsymbol{\vartheta}}^{(r)}$. First, the conditional expected value of $c_{ik}$ given the observed data is given by

$$p_{ik}^{(r)} := \mathbb{E}(c_{ik} \mid \mathbf{y}_i^o, \mathbf{x}_i, \hat{\boldsymbol{\vartheta}}^{(r)}) = \frac{\hat{\pi}_{ik} f_{\text{GHD},T_i^o}(\hat{\lambda}_k^{(r)}, \hat{\omega}_k^{(r)}, \hat{\boldsymbol{\mu}}_k^{o(r)}, \hat{\boldsymbol{\Sigma}}_k^{oo(r)}, \hat{\boldsymbol{\beta}}_k^{o(r)})}{\sum_{l=1}^{K} \hat{\pi}_{il} f_{\text{GHD},T_i^o}(\hat{\lambda}_l^{(r)}, \hat{\omega}_l^{(r)}, \hat{\boldsymbol{\mu}}_l^{o(r)}, \hat{\boldsymbol{\Sigma}}_l^{oo(r)}, \hat{\boldsymbol{\beta}}_l^{o(r)})}, \tag{4.8}$$

where $\hat{\boldsymbol{\mu}}_k^{o(r)}$, $\hat{\boldsymbol{\beta}}_k^{o(r)}$, and $\hat{\boldsymbol{\Sigma}}_k^{oo(r)}$ are $\boldsymbol{\mu}_k^o$, $\boldsymbol{\beta}_k^o$, and $\boldsymbol{\Sigma}_k^{oo}$ evaluated at $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}^{(r)}$, respectively. It denotes the posterior probability of $i$th observation $\mathbf{y}_i^o$ belonging to the $k$th component at the $r$th iteration. From proposition 1b, we have convenient closed forms

for the following conditional expectations:

$$a_{ik}^{(r)} := \mathbb{E}\left[w_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = \sqrt{\frac{\hat{\chi}_{ik}^{*(r)}}{\hat{\psi}_{ik}^{*(r)}}} \frac{K_{\hat{\lambda}_{ik}^{*(r)}+1}(\sqrt{\hat{\psi}_{ik}^{*(r)}\hat{\chi}_{ik}^{*(r)}})}{K_{\hat{\lambda}_{ik}^{*(r)}}(\sqrt{\hat{\psi}_{ik}^{*(r)}\hat{\chi}_{ik}^{*(r)}})},$$

$$b_{ik}^{(r)} := \mathbb{E}\left[1/w_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = \sqrt{\frac{\hat{\psi}_{ik}^{*(r)}}{\hat{\chi}_{ik}^{*(r)}}} \frac{K_{\hat{\lambda}_{ik}^{*(r)}+1}(\sqrt{\hat{\psi}_{ik}^{*(r)}\hat{\chi}_{ik}^{*(r)}})}{K_{\hat{\lambda}_{ik}^{*(r)}}(\sqrt{\hat{\psi}_{ik}^{*(r)}\hat{\chi}_{ik}^{*(r)}})} - \frac{2\hat{\lambda}_{ik}^{*(r)}}{\hat{\chi}_{ik}^{*(r)}},$$

$$d_{ik}^{(r)} := \mathbb{E}\left[\log w_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = \log(\sqrt{\frac{\hat{\chi}_{ik}^{*(r)}}{\hat{\psi}_{ik}^{*(r)}}}) + \frac{1}{K_{\hat{\lambda}_{ik}^{*(r)}}(\sqrt{\hat{\psi}_{ik}^{*(r)}\hat{\chi}_{ik}^{*(r)}})} \frac{\partial}{\partial \hat{\lambda}_{ik}^{*(r)}} K_{\hat{\lambda}_{ik}^{*(r)}}(\sqrt{\hat{\psi}_{ik}^{*(r)}\hat{\chi}_{ik}^{*(r)}}),$$

where $\hat{\chi}_{ik}^{*(r)}$, $\hat{\psi}_{ik}^{*(r)}$, $\hat{\lambda}_{ik}^{*(r)}$ are $\chi_{ik}^{*}$, $\psi_{ik}^{*}$, $\lambda_{ik}^{*}$ are evaluated at $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}^{(r)}$, respectively. Inspired by the key idea of the EM algorithm, we impute the missing values $\mathbf{y}_i^m$ to yield a complete data set at each iteration. To this end, for the actual missing values $\mathbf{y}_i^m$ and the latent continuous growth factors $\boldsymbol{\eta}_i$, based on proposition 1 (c, d, and e),

the following conditional expectations are also needed,

$$
\begin{aligned}
E_{1ik}^{(r)} &:= \mathbb{E}\left[1/w_i\mathbf{y}_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = b_{ik}^{(r)}\hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\mathbf{y}_i \\
&\qquad + b_{ik}^{(r)}\left(\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\right)\hat{\boldsymbol{\mu}}_k^{(r)} + \left(\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\right)\boldsymbol{\Lambda}_y\hat{\boldsymbol{\beta}}_k^{(r)}, \\
E_{2ik}^{(r)} &:= \mathbb{E}\left[1/w_i\mathbf{y}_i\mathbf{y}_i^{'} \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = \left(\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\right)\hat{\boldsymbol{\Sigma}}_k^{(r)} \\
&\quad + a_{ik}^{(r)}\left(\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\right)\boldsymbol{\Lambda}_y\hat{\boldsymbol{\beta}}_k^{(r)}\hat{\boldsymbol{\beta}}_k^{'(r)}\boldsymbol{\Lambda}_y^{'}\left(\mathbf{I}_T - \hat{\mathbf{S}}_{ik}^{oo(r)}\hat{\boldsymbol{\Sigma}}_k^{(r)}\right) \\
&\quad + \left(\hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\mathbf{y}_i + (\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)})\hat{\boldsymbol{\mu}}_k^{(r)}\right)\hat{\boldsymbol{\beta}}_k^{'(r)}\boldsymbol{\Lambda}_y^{'}\left(\mathbf{I}_T - \hat{\mathbf{S}}_{ik}^{oo(r)}\hat{\boldsymbol{\Sigma}}_k^{(r)}\right) \\
&\quad + \left(\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\right)\boldsymbol{\Lambda}_y\hat{\boldsymbol{\beta}}_k^{(r)}\left(\hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\mathbf{y}_i + (\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)})\hat{\boldsymbol{\mu}}_k^{(r)}\right)^{'} \\
&\quad + b_{ik}^{(r)}\left(\hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\mathbf{y}_i + (\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)})\hat{\boldsymbol{\mu}}_k^{(r)}\right)\left(\hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)}\mathbf{y}_i + (\mathbf{I}_T - \hat{\boldsymbol{\Sigma}}_k^{(r)}\hat{\mathbf{S}}_{ik}^{oo(r)})\hat{\boldsymbol{\mu}}_k^{(r)}\right)^{'}, \\
E_{3ik}^{(r)} &:= \mathbb{E}\left[\boldsymbol{\eta}_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = \hat{\mathbf{V}}_k^{(r)}\left(\hat{\boldsymbol{\Psi}}_k^{-1(r)}(\hat{\boldsymbol{\alpha}}_k^{(r)} + \hat{\boldsymbol{\Gamma}}_k^{(r)}\mathbf{x}_i + a_{ik}^{(r)}\hat{\boldsymbol{\beta}}_k^{(r)}) + \boldsymbol{\Lambda}_y^{'}\hat{\mathbf{T}}_{ik}^{oo(r)}\mathbf{y}_i\right), \\
E_{4ik}^{(r)} &:= \mathbb{E}\left[1/w_{ik}\boldsymbol{\eta}_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = \hat{\mathbf{V}}_k^{(r)}\left(b_{ik}^{(r)}\hat{\boldsymbol{\mu}}_k^{(r)} + \hat{\boldsymbol{\Psi}}_k^{-1(r)}\hat{\boldsymbol{\beta}}_k^{(r)} + b_{ik}^{(r)}\boldsymbol{\Lambda}_y^{'}\hat{\mathbf{T}}_{ik}^{oo(r)}\mathbf{y}_i\right), \\
E_{5ik}^{(r)} &:= \mathbb{E}\left[1/w_{ik}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{'} \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = \hat{\mathbf{V}}_k^{(r)} + a_{ik}^{(r)}\hat{\mathbf{V}}_k^{(r)}\hat{\boldsymbol{\Psi}}_k^{-1(r)}\hat{\boldsymbol{\beta}}_k^{(r)}\hat{\boldsymbol{\beta}}_k^{'(r)}\hat{\boldsymbol{\Psi}}_k^{-1(r)}\hat{\mathbf{V}}_k^{(r)} \\
&\qquad + \hat{\mathbf{V}}_k^{(r)}\hat{\boldsymbol{\Psi}}_k^{-1(r)}\hat{\boldsymbol{\beta}}_k^{(r)}\left(\hat{\boldsymbol{\mu}}_k^{(r)} + \boldsymbol{\Lambda}_y^{'}\hat{\mathbf{T}}_{ik}^{oo(r)}\mathbf{y}_i\right)^{'}\hat{\mathbf{V}}_k^{(r)} \\
&\qquad + \hat{\mathbf{V}}_k^{(r)}\left(\hat{\boldsymbol{\mu}}_k^{(r)} + \boldsymbol{\Lambda}_y^{'}\hat{\mathbf{T}}_{ik}^{oo(r)}\mathbf{y}_i\right)\hat{\boldsymbol{\beta}}_k^{'(r)}\hat{\boldsymbol{\Psi}}_k^{-1(r)}\hat{\mathbf{V}}_k^{(r)} \\
&\qquad + b_{ik}^{(r)}\hat{\mathbf{V}}_k^{(r)}\left(\hat{\boldsymbol{\mu}}_k^{(r)} + \boldsymbol{\Lambda}_y^{'}\hat{\mathbf{T}}_{ik}^{oo(r)}\mathbf{y}_i\right)\left(\hat{\boldsymbol{\mu}}_k^{(r)} + \boldsymbol{\Lambda}_y^{'}\hat{\mathbf{T}}_{ik}^{oo(r)}\mathbf{y}_i\right)^{'}\hat{\mathbf{V}}_k^{(r)}, \\
E_{6ik}^{(r)} &:= \mathbb{E}\left[1/w_{ik}\mathbf{y}_i\boldsymbol{\eta}_i^{'} \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right] = \hat{\boldsymbol{\Theta}}_k^{(r)}\hat{\mathbf{T}}_{ik}^{oo(r)}\mathbf{y}_i E_{4ik}^{'(r)} + \left(\mathbf{I}_T - \hat{\boldsymbol{\Theta}}_k^{(r)}\hat{\mathbf{T}}_{ik}^{oo(r)}\right)\boldsymbol{\Lambda}_y E_{5ik}^{(r)},
\end{aligned}
$$

where $\hat{\boldsymbol{\mu}}_k^{(r)}, \hat{\boldsymbol{\alpha}}_k^{(r)}, \hat{\boldsymbol{\Gamma}}_k^{(r)}, \hat{\boldsymbol{\Sigma}}_k^{(r)}, \hat{\boldsymbol{\Psi}}_k^{(r)}, \mathbf{S}_{ik}^{oo(r)}, \hat{\mathbf{T}}_{ik}^{oo(r)}$, and $\hat{\mathbf{V}}_k^{(r)}$ are corresponding themselves evaluated at $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}^{(r)}$, respectively. The proof of these conditional expectations follows directly from the law of iterative expectations and the use of $\mathbf{O}_i^{'}\mathbf{O}_i(\mathbf{I}_T - \boldsymbol{\Sigma}_k\mathbf{S}_{ig}^{oo}) = \mathbf{0}$ and $\mathbf{O}_i^{'}\mathbf{O}_i(\mathbf{I}_T - \boldsymbol{\Theta}_k\mathbf{T}_{ig}^{oo}) = \mathbf{0}$.

After forming the $\mathcal{Q}$ function, the M-step involves maximizing such $\mathcal{Q}$ function

with respect to model parameters $\boldsymbol{\vartheta}$. For notational convenience, let $n_k^{(r)} = \sum_{i=1}^n p_{ik}^{(r)}$, $\bar{a}_k^{(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^n p_{ik}^{(r)} a_{ik}^{(r)}$, $\bar{b}_k^{(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^n p_{ik}^{(r)} b_{ik}^{(r)}$, and $\bar{d}_k^{(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^n p_{ik}^{(r)} d_{ik}^{(r)}$. In summary, the resulting M-step can be implemented as follows:

(i) Update the parameters $\hat{\boldsymbol{\alpha}}_c^{(r)}$ and $\hat{\boldsymbol{\Gamma}}_c^{(r)}$ by maximizing

$$\sum_{i=1}^n \sum_{k=1}^K p_{ik}^{(r)} \log \pi_{ik}$$

with respect to $\boldsymbol{\alpha}_c$ and $\boldsymbol{\Gamma}_c$, which can be seen as a multinomial logistic regression with fractional observations $p_{ik}^{(r)}$.

(ii) Update the measurement errors $\hat{\boldsymbol{\Theta}}_k^{(r)}$ by differentiating the $\mathcal{Q}$ function with respect to $\boldsymbol{\Theta}_k$, which gives

$$\hat{\boldsymbol{\Theta}}_k^{(r+1)} = \frac{1}{n_k^{(r)}} \operatorname{diag} \left( \sum_{i=1}^n p_{ik}^{(r)} (E_{2ik}^{(r)} - E_{6ik}^{(r)} \boldsymbol{\Lambda}_y' - \boldsymbol{\Lambda}_y E_{6ik}^{\prime(r)} + \boldsymbol{\Lambda}_y E_{5ik}^{(r)} \boldsymbol{\Lambda}_y') \right),$$

where $\operatorname{diag}(\cdot)$ means a diagonal matrix constructed by extracting the main diagonal elements of a square matrix.

(iii) Update the regression coefficients $\hat{\boldsymbol{\Gamma}}_k^{(r)}$ by differentiating the $\mathcal{Q}$ function with respect to $\boldsymbol{\Gamma}_k$, which gives

$$\hat{\boldsymbol{\Gamma}}_k^{(r+1)} = \sum_{i=1}^n p_{ik}^{(r)} \left( E_{4ik}^{(r)} \mathbf{x}_i' - b_{ik}^{(r)} \hat{\boldsymbol{\alpha}}_k^{(r)} \mathbf{x}_i' - \hat{\boldsymbol{\beta}}_k^{(r)} \mathbf{x}_i^i \right) \left[ \sum_{i=1}^n p_{ik}^{(r)} b_{ik}^{(r)} \mathbf{x}_i \mathbf{x}_i' \right]^{-1}.$$

(iv) Update the intercept and skewness parameters $\hat{\boldsymbol{\alpha}}_k^{(r)}$ and $\hat{\boldsymbol{\beta}}_k^{(r)}$ by differentiating

the $\mathcal{Q}$ function with respect to $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$, respectively, which leads to

$$\hat{\boldsymbol{\alpha}}_k^{(r+1)} = \frac{\sum_{i=1}^n p_{ik}^{(r)} \left( E_{4ik}^{(r)} - b_{ik}^{(r)} \hat{\boldsymbol{\Gamma}}_k^{(r+1)} \mathbf{x}_i - \hat{\boldsymbol{\beta}}_k^{(r)} \right)}{\sum_{i=1}^n p_{ik}^{(r)} b_{ik}^{(r)}},$$

and

$$\hat{\boldsymbol{\beta}}_k^{(r+1)} = \frac{\sum_{i=1}^n p_{ik}^{(r)} \left( E_{3ik} - \hat{\boldsymbol{\alpha}}_k^{(r+1)} - \hat{\boldsymbol{\Gamma}}_k^{(r+1)} \mathbf{x}_i \right)}{\sum_{i=1}^n p_{ik}^{(r)} a_{ik}^{(r)}}.$$

(v) Update the model errors $\hat{\boldsymbol{\Psi}}_k^{(r)}$ by differentiating the $\mathcal{Q}$ function with respect to $\boldsymbol{\Psi}_k$, which gives

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_k^{(r+1)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^n p_{ik}^{(r)} \Bigg( & E_{5ik}^{(r)} - (E_{4ik}^{(r)} - \hat{\boldsymbol{\beta}}_k^{(r+1)})(\hat{\boldsymbol{\alpha}}_k^{(r+1)} + \hat{\boldsymbol{\Gamma}}_k^{(r+1)} \mathbf{x}_i)' - E_{3ik}^{(r)} \hat{\boldsymbol{\beta}}_k^{'(r+1)} \\ & - \hat{\boldsymbol{\beta}}_k^{(r+1)} E_{3ik}^{'(r)} - (\hat{\boldsymbol{\alpha}}_k^{(r+1)} + \hat{\boldsymbol{\Gamma}}_k^{(r+1)} \mathbf{x}_i)(E_{4ik}^{(r)} - \hat{\boldsymbol{\beta}}_k^{(r+1)})' \\ & + b_{ik}^{(r)} (\hat{\boldsymbol{\alpha}}_k^{(r+1)} \hat{\boldsymbol{\Gamma}}_k^{(r+1)} \mathbf{x}_i)(\hat{\boldsymbol{\alpha}}_k^{(r+1)} + \hat{\boldsymbol{\Gamma}}_k^{(r+1)} \mathbf{x}_i)' + a_{ik}^{(r)} \hat{\boldsymbol{\beta}}_k^{(r+1)} \hat{\boldsymbol{\beta}}_k^{'(r+1)} \Bigg)\end{aligned}$$

(vi) Update the index and concentrate parameters $\hat{\lambda}_k$ and $\hat{\omega}_k^{(r)}$ by maximizing the following function

$$q_k(\lambda_k, \omega_k) = -\log K_{\lambda_k}(\omega_k) + (\lambda_k - 1)\bar{d}_k^{(r)} - \frac{\omega_k}{2}(\bar{a}_k^{(r)} + \bar{b}_k^{(r)}).$$

This leads to

$$\hat{\lambda}_k^{(r+1)} = \bar{d}_k^{(r)} \hat{\lambda}_k^{(r)} \left[ \frac{\partial}{\partial t} \log K_t(\hat{\omega}_k^{(r)}) \mid_{t=\hat{\lambda}_k^{(r)}} \right]^{-1},$$

and

$$\hat{\omega}_k^{(r+1)} = \hat{\omega}_k^{(r)} - \left[ \frac{\partial}{\partial t} q_k(t, \hat{\lambda}_k^{(r+1)}) \mid_{t=\hat{\omega}_k^{(r)}} \right] \left[ \frac{\partial^2}{\partial t^2} q_k(t, \hat{\lambda}_k^{(r+1)}) \mid_{t=\hat{\omega}_k^{(r)}} \right]^{-1}.$$

These updates are analogous to those used by Browne and McNicholas (2015).

### 4.3.2 The EM algorithm for GST-GMM with missing information

Analogous to the GHD-GMM with missing information, for the GST-GMM with missing information, the complete-data consist of the observed data $(\mathbf{y}_i^o, \mathbf{x}_i)$ and the missing data $(\mathbf{y}_i^m, \boldsymbol{\eta}_i, \mathbf{c}_i, w_i)$. Accordingly, the complete-data log-likelihood function is

$$l_c(\boldsymbol{\vartheta} \,|\mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{x}_i, \boldsymbol{\eta}_i, \mathbf{c}_i, w_i) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}\Big[\log\pi_{ik} + \log\phi(\mathbf{y}_i \mid \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i, w_{ik}\boldsymbol{\Theta}_k) + \log\phi(\boldsymbol{\eta}_i \mid \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}, w_{ik}\boldsymbol{\Psi}_k) + \log h(w_{ik} \mid \frac{\nu_k}{2}, \frac{\nu_k}{2})\Big]. \tag{4.9}$$

The E-step involves the calculation of the expected value of the class membership $c_{ik}$, i.e.,

$$\tau_{ik}^{(r)} := \mathbb{E}(c_{ik} \mid \mathbf{y}_i^o, \mathbf{x}_i, \hat{\boldsymbol{\vartheta}}^{(r)}) = \frac{\hat{\pi}_{ik} f_{\mathrm{GST},T_i^o}(\hat{\nu}_k^{(r)}, \hat{\boldsymbol{\mu}}_k^{o(r)}, \hat{\boldsymbol{\Sigma}}_k^{oo(r)}, \hat{\boldsymbol{\beta}}_k^{o(r)})}{\sum_{l=1}^{K} \hat{\pi}_{il} f_{\mathrm{GST},T_i^o}(\hat{\nu}_l^{(r)}, \hat{\boldsymbol{\mu}}_l^{o(r)}, \hat{\boldsymbol{\Sigma}}_l^{oo(r)}, \hat{\boldsymbol{\beta}}_l^{o(r)})}, \tag{4.10}$$

as well as the following conditional expectations:

$$A_{ik}^{(r)} := \mathbb{E}\left[w_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right], \qquad B_{ik}^{(r)} := \mathbb{E}\left[1/w_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right],$$
$$D_{ik}^{(r)} := \mathbb{E}\left[\log w_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right], \qquad E_{1ik}^{\star(r)} := \mathbb{E}\left[1/w_i\mathbf{y}_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right],$$
$$E_{2ik}^{\star(r)} := \mathbb{E}\left[1/w_i\mathbf{y}_i\mathbf{y}_i^{'} \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right], \qquad E_{3ik}^{\star(r)} := \mathbb{E}\left[\boldsymbol{\eta}_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right],$$
$$E_{4ik}^{\star(r)} := \mathbb{E}\left[1/w_{ik}\boldsymbol{\eta}_i \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right], \qquad E_{5ik}^{\star(r)} := \mathbb{E}\left[1/w_{ik}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{'} \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right],$$
$$E_{6ik}^{\star(r)} := \mathbb{E}\left[1/w_{ik}\mathbf{y}_i\boldsymbol{\eta}_i^{'} \mid \mathbf{y}_i^o, \mathbf{x}_i, c_{ik}=1\right],$$

which are similar to those used in the case of the GHD-GMM with missing information, hence are omitted here. The M-step proceeds in analogy to that for the GHD-GMM with missing information, except that the degrees of freedom parameter $\nu_k$ is updated here instead of $\lambda_k$ and $\omega_k$. We update the degree of freedom parameter $\hat{\nu}_k^{(r)}$ by solving the root of

$$\log\left(\frac{\nu_k}{2}\right) + 1 - \varphi\left(\frac{\nu_k}{2}\right) - \frac{1}{n_k^{(r)}}\sum_{i=1}^{n}\tau_{ik}^{(r)}\left(D_{ik}^{(r)} + B_{ik}^{(r)}\right) = 0,$$

where $n_k^{(r)} = \sum_{i=1}^{n}\tau_{ik}^{(r)}$ and $\varphi(\cdot)$ is the digamma function. The `uniroot` function from R is employed to carry out the root finding of the above equation. In both cases of GHD-GMM and GST-GMM with missing information, the E- and M-step of the EM algorithm are iterated repeatedly until a Aitken acceleration-based criterion is satisfied.

### 4.3.3 Estimation of random effects and imputation of missing values

When the algorithm achieves convergence, we can estimate random effects, namely growth factor scores $\boldsymbol{\eta}_i$, and obtain the imputation of the missing values $\mathbf{y}_i^m$. Specifically, let $\hat{\boldsymbol{\eta}}_{ik}$ be the estimated conditional expectation of growth factor scores corresponding to $\mathbf{y}_i^o$ for the $k$th class, which can be computed by substituting $\boldsymbol{\vartheta}$ with the ML estimate $\hat{\boldsymbol{\vartheta}}$ into $E_{3ir}^{(r)}$ or $E_{3ir}^{\star(r)}$ depending on presumed underlying distribution for the random effects. Moreover, the estimated random effects scores corresponding to $\mathbf{y}_i^o$ can be viewed as

$$\hat{\boldsymbol{\eta}}_i = \sum_{k=1}^{K} \hat{z}_{ik} \hat{\boldsymbol{\eta}}_{ik},$$

where $\hat{z}_{ik}$, calculated through (4.8) or (4.10) evaluated at ML estimates $\hat{\boldsymbol{\vartheta}}$, denotes the posterior probability of $\mathbf{y}_i^o$ belonging to the $k$th component of the GMM. Next, we perform the imputation of missing values $\mathbf{y}_i^m$ via the conditional expectation method. According to proportion 1c,

$$\mathbf{y}_{ik}^m := \mathbb{E}\left[\mathbf{y}_i^m \mid \mathbf{y}_i^o, \mathbf{x}_i, w_{ik}, c_{ik} = 1\right] = \mathbf{M}_i(\boldsymbol{\mu}_k + w_{ik}\boldsymbol{\Lambda}_y\boldsymbol{\beta}_k) + \boldsymbol{\Sigma}_k\mathbf{S}_{ik}^{oo}(\mathbf{y}_i - \boldsymbol{\mu}_k - w_{ik}\boldsymbol{\Lambda}_y\boldsymbol{\beta}_k). \tag{4.11}$$

Therefore, the imputation of missing values $\mathbf{y}_i^m$ can be defined as

$$\hat{\mathbf{y}}_i^m = \sum_{k=1}^{K} \hat{z}_{ik} \hat{\mathbf{y}}_{ik}^m,$$

where $\hat{z}_{ik}$ is as defined above. Note that $w_{ik}$ in (4.11) are evaluated via $\hat{a}_{ik}$ or $\hat{A}_{ik}$ at ML estimates $\hat{\boldsymbol{\vartheta}}$ depending on the presumed underlying distribution of the random effects.

Specifically, if the random effects follow the generalized hyperbolic distribution, then $\hat{a}_{ik}$ is used; if the random effects follow the multivariate skew-t distribution, then $\hat{A}_{ik}$ is used.

## 4.4 Illustration

### 4.4.1 Simulation Studies

To assess the performance of our proposed models (i.e., GHD-GMM and GST-GMM) with varying proportions of missing values, we performed two simulation studies in this section. In simulation studies, we were interested in both the number of selected classes in terms of the Bayesian information cirterion (BIC; Schwarz, 1978) , as well as the classification performance using the adjusted rand index (ARI; Hubert and Arabie, 1985) and the misclassification rate (ERR).

We utilize the normal variance-mean mixture as given in (2.6) to generate GHD-GMM and GST-GMM data. In the first simulation experiment, we consider two classes coming from a five time points linear GMM with generalized hyperbolic random effects. In the second simulation experiment, we simulate two classes from a eight time points quadratic GMM with multivariate skew-t distribution. In each simulation, the GHD-GMM and GST-GMM data are generated with a total of 30 replicates and two different sample sizes, i.e., $n_k = 250$ and $n_k = 500$. Table 4.1 lists the true model parameters for the simulated data. In Figure 4.1, we show the individual's trajectory plots for a typical simulated GHD-GMM and GST-GMM dataset with sample size $n_k = 250$. The individual's trajectory plot for the simulated GHD-GMM dataset suggests a considerable amount for heterogeneity, i.e., the two classes

are well separated. While we see that there is very little separation between the two classes for the simulated GST-GMM dataset.

Table 4.1: True model parameters for the simulated data

| | Trajectory class 1 | Trajectory class 2 |
|---|---|---|
| GHD-GMM | $\lambda_1 = -5$<br>$\omega_1 = 2$<br>$\boldsymbol{\alpha}_1 = (4, 5)^{'}$<br>$\boldsymbol{\beta}_1 = (1, -1)^{'}$<br>$\boldsymbol{\Theta}_1 = \text{diag}(1, 2, 3, 4, 5)$<br>$\boldsymbol{\Psi}_1 = \begin{bmatrix} 2.18 & 1.07 \\ 1.07 & 3.35 \end{bmatrix}$ | $\lambda_2 = -6$<br>$\omega_2 = 3$<br>$\boldsymbol{\alpha}_2 = (1, 2)^{'}$<br>$\boldsymbol{\beta}_2 = (-1, 1)^{'}$<br>$\boldsymbol{\Theta}_2 = \text{diag}(1.5, 2.5, 3.5, 4.5, 5.5)$<br>$\boldsymbol{\Psi}_2 = \begin{bmatrix} 1.51 & -0.18 \\ -0.18 & 1.37 \end{bmatrix}$ |
| GST-GMM | $\nu_1 = 4$<br>$\boldsymbol{\alpha}_1 = (4, -5, 3)^{'}$<br>$\boldsymbol{\beta}_1 = (1, 1, 1)^{'}$<br>$\boldsymbol{\Theta}_1 = \text{diag}(1, 2.4, 3.7, 4.3, 5.2, 6.5, 7.6, 8.7)$<br>$\boldsymbol{\Psi}_1 = \begin{bmatrix} 2.11 & -0.01 & -0.22 \\ -0.01 & 2.09 & -0.10 \\ 0.22 & -0.10 & 3.71 \end{bmatrix}$ | $\nu_2 = 5$<br>$\boldsymbol{\alpha}_2 = (-2, 3, -4)^{'}$<br>$\boldsymbol{\beta}_2 = (-1, -1, -1)^{'}$<br>$\boldsymbol{\Theta}_2 = \text{diag}(1, 2.1, 3.1, 4.1, 5.1, 6.1, 7.1, 8.3)$<br>$\boldsymbol{\Psi}_2 = \begin{bmatrix} 4.08 & -0.66 & -0.06 \\ -0.66 & 3.32 & -0.10 \\ -0.06 & -0.10 & 4.65 \end{bmatrix}$ |

The simulated datasets are complete under each scenario considered, so for illustration purposes we remove entries at random under missing rates of low (5% and 10%), moderate (20%), and relatively high (30%). Throughout this chapter, the proposed models (i.e., GHD-GMM and GST-GMM) and their Gaussian GMM counterpart were used to carried out parameter estimation. Our proposed models are implemented in R and Gaussian GMM is carried out by `Mplus` Version 7.3. As stated in previous chapter, a family of the GHD-GMMs or GST-GMMs can be obtained by imposing the constraints on the model parameters $\boldsymbol{\vartheta}$ (equal or different across classes). We only consider the general model (all the model parameters are different across classes) and the most constrained model (only $\boldsymbol{\alpha}_k$ is different across classes) herein.

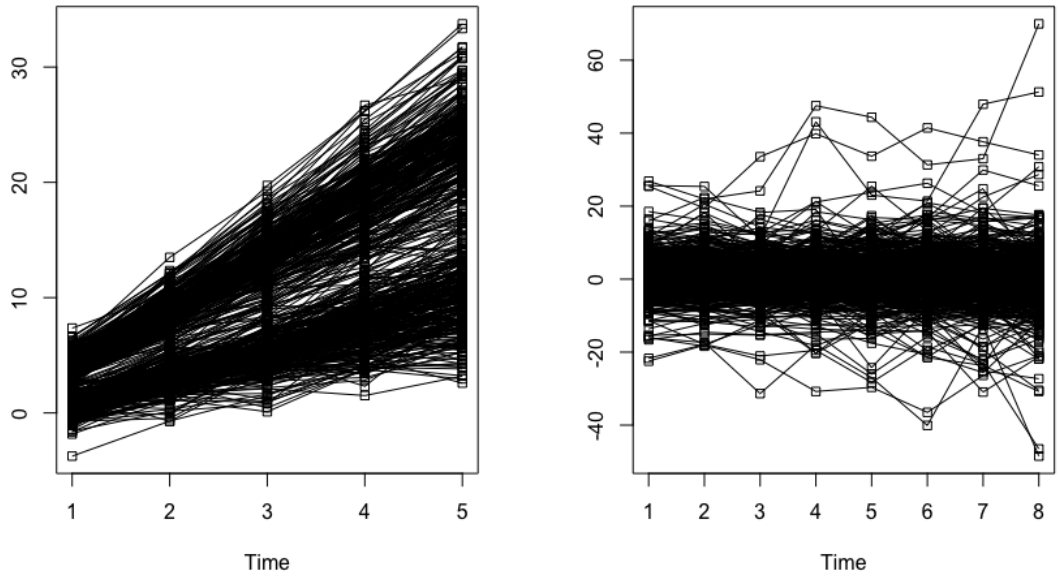


Figure 4.1: Individual's trajectories plots for a typical simulated GHD-GMM and GST-GMM dataset with $n_k = 250$

In Table 4.2 and 4.3, we show the number of times each latent classes was selected, as well as the average ARI and the average ERR and their corresponding associated standard deviations. Overall, it can be seen that our proposed GHD-GMM with missing information outperforms the GST-GMM and the Gaussian GMM counterpart in selecting the correct number of latent classes with a very high average ARI ($\overline{\text{ARI}}$) and low average ERR ($\overline{\text{ERR}}$). When the true underlying classes are well separated (e.g., first simulation), all the models perform well but the GHD-GMM performs best in terms of ARI and ERR. When the true underlying classes are overlapping (e.g., second simulation), the GHD-GMM and GST-GMM leads to much higher ARI and much lower ERR than their Gaussian GMM counterpart. The Gaussian GMM usually overestimte the true trajectory classes unless the underlying true classes are well separated. When the true underlying classes are overlapping, the increase of

missing rate generally decrease the ARI and increase the ERR.

Table 4.2: The number of classes selected by the BIC for the first simulation experiment with two different sample sizes under different missing rates (r)

| r | Fitted Models | Number of classes | | | | | $\overline{\text{ARI}}$ (std. dev.) | $\overline{\text{ERR}}$ (std. dev.) |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | | |
| $n_k = 250$ | | | | | | | | |
| | GMM-Normal | 0 | 0 | 22 | 8 | 0 | 0.955(0.040) | 0.017(0.021) |
| 5% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.971(0.019) | 0.007(0.005) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.971(0.019) | 0.007(0.005) |
| | GMM-Normal | 0 | 1 | 18 | 8 | 3 | 0.957(0.042) | 0.015(0.022) |
| 10% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.973(0.014) | 0.007(0.003) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.973(0.015) | 0.007(0.004) |
| | GMM-Normal | 0 | 9 | 19 | 2 | 0 | 0.968(0.018) | 0.009(0.007) |
| 20% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.972(0.015) | 0.007(0.004) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.971(0.015) | 0.007(0.004) |
| | GMM-Normal | 0 | 10 | 20 | 0 | 0 | 0.967(0.018) | 0.008(0.005) |
| 30% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.971(0.016) | 0.007(0.004) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.967(0.019) | 0.008(0.005) |
| $n_k = 500$ | | | | | | | | |
| | GMM-Normal | 0 | 10 | 20 | 0 | 0 | 0.895(0.067) | 0.055(0.051) |
| 5% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.972(0.011) | 0.007(0.003) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.972(0.011) | 0.007(0.003) |
| | GMM-Normal | 0 | 17 | 13 | 0 | 0 | 0.926(0.0616) | 0.035(0.040) |
| 10% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.970(0.010) | 0.007(0.003) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.970(0.010) | 0.007(0.003) |
| | GMM-Normal | 0 | 18 | 12 | 0 | 0 | 0.923(0.068) | 0.036(0.039) |
| 20% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.969(0.011) | 0.008(0.003) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.969(0.011) | 0.007(0.003) |
| | GMM-Normal | 0 | 22 | 8 | 0 | 0 | 0.939(0.052) | 0.025(0.036) |
| 30% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.963(0.014) | 0.009(0.003) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.963(0.014) | 0.009(0.003) |

Table 4.3: The number of classes selected by the BIC for the second simulation experiment with two different sample sizes under different missing rates (r)

| r | Fitted Models | Number of classes | | | | | $\overline{\text{ARI}}$ (std. dev.) | $\overline{\text{ERR}}$ (std. dev.) |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | | |
| $n_k = 250$ | | | | | | | | |
| | GMM-Normal | 0 | 20 | 10 | 0 | 0 | 0.501(0.167) | 0.272(0.088) |
| 5% | GST-GMM | 1 | 28 | 1 | 0 | 0 | 0.895(0.171) | 0.035(0.088) |
| | GHD-GMM | 1 | 29 | 0 | 0 | 0 | 0.891(0.173) | 0.036(0.088) |
| | GMM-Normal | 0 | 25 | 5 | 0 | 0 | 0.436(0.218) | 0.302(0.099) |
| 10% | GST-GMM | 3 | 27 | 0 | 0 | 0 | 0.832(0.284) | 0.067(0.147) |
| | GHD-GMM | 2 | 28 | 0 | 0 | 0 | 0.865(0.237) | 0.051(0.122) |
| | GMM-Normal | 0 | 28 | 2 | 0 | 0 | 0.416(0.298) | 0.282(0.131) |
| 20% | GST-GMM | 2 | 27 | 1 | 0 | 0 | 0.854(0.233) | 0.054(0.121) |
| | GHD-GMM | 1 | 29 | 0 | 0 | 0 | 0.882(0.174) | 0.038(0.088) |
| | GMM-Normal | 0 | 30 | 0 | 0 | 0 | 0.311(0.272) | 0.282(0.146) |
| 30% | GST-GMM | 1 | 29 | 0 | 0 | 0 | 0.876(0.184) | 0.040(0.090) |
| | GHD-GMM | 0 | 30 | 0 | 0 | 0 | 0.905(0.079) | 0.025(0.023) |
| $n_k = 500$ | | | | | | | | |
| | GMM-Normal | 0 | 10 | 20 | 0 | 0 | 0.423(0.110) | 0.352(0.069) |
| 5% | GST-GMM | 0 | 29 | 1 | 0 | | 0.932(0.017) | 0.017(0.004) |
| | GHD-GMM | 2 | 27 | 1 | 0 | 0 | 0.867(0.241) | 0.051(0.124) |
| | GMM-Normal | 0 | 15 | 15 | 0 | 0 | 0.474(0.086) | 0.320(0.070) |
| 10% | GST-GMM | 1 | 29 | 0 | 0 | | 0.897(0.170) | 0.035(0.088) |
| | GHD-GMM | 0 | 29 | 1 | 0 | 0 | 0.927(0.017) | 0.019(0.006) |
| | GMM-Normal | 0 | 18 | 12 | 0 | 0 | 0.474(0.093) | 0.334(0.086) |
| 20% | GST-GMM | 0 | 30 | 0 | 0 | 0 | 0.920(0.015) | 0.021(0.004) |
| | GHD-GMM | 1 | 29 | 0 | 0 | 0 | 0.883(0.170) | 0.038(0.088) |
| | GMM-normal | 0 | 22 | 8 | 0 | 0 | 0.518(0.167) | 0.271(0.108) |
| 30% | GST-GMM | 2 | 28 | 0 | 0 | 0 | 0.851(0.232) | 0.054(0.121) |
| | GHD-GMM | 1 | 29 | 0 | 0 | 0 | 0.881(0.168) | 0.039(0.087) |

### 4.4.2 Body mass index (BMI) from the National Longitudinal Survey of Youth (NLSY)

In this section, we illustrate the application of our proposed non-elliptical GMMs with missing information through the analysis of the BMI development over the ages 12-23

years using the data from the National Longitudinal Survey of Youth 1997 (NLSY), a nationally-representative survey conducted yearly by the United States Department of Labor's Bureau of Labor Statistics. The NLSY began in 1997 with youths between the ages of 12 and 16 years old and continued to assess them annually. In addition to labor participation, the NLSY also collected health-related data, including Body Mass Index (BMI). In this Section, we only consider the black women ($n = 1160$).

Table 4.4 depicts the summary statistics for the data along with the missing data information. Overall, the means of the BMI increased over time. The missing data rates range from 13.97% to 94.48%, and there are 227 patterns of missing data. Figure 4.2 shows boxplots of the twelve attributes of BMI from NLSY ages 12-23 for black women. From this figure, it can be seen that the distributions of many attributes exhibit heavier tail weight than normal distributions, indicating that the assumption of normality is not reasonable for this dataset.

Table 4.4: Summary statistics for BMI development ages 12-23 from NLSY

| Variables | Mean | Std. | Missing data (count) | Missing data (percentage) |
|---|---|---|---|---|
| bmi12 | 21.70 | 4.38 | 1036 | 89.31 |
| bmi13 | 22.79 | 4.83 | 833 | 71.81 |
| bmi14 | 23.57 | 5.31 | 615 | 53.02 |
| bmi15 | 23.84 | 5.21 | 453 | 39.05 |
| bmi16 | 24.45 | 5.53 | 222 | 19.14 |
| bmi17 | 25.06 | 5.92 | 162 | 13.97 |
| bmi18 | 25.68 | 6.09 | 217 | 18.71 |
| bmi19 | 26.24 | 6.58 | 390 | 33.62 |
| bmi20 | 26.70 | 6.83 | 565 | 48.71 |
| bmi21 | 26.85 | 6.68 | 762 | 65.69 |
| bmi22 | 27.71 | 7.27 | 916 | 78.97 |
| bmi23 | 28.55 | 7.06 | 1096 | 94.48 |

We analyze the BMI data using our proposed models and their Gaussian GMM

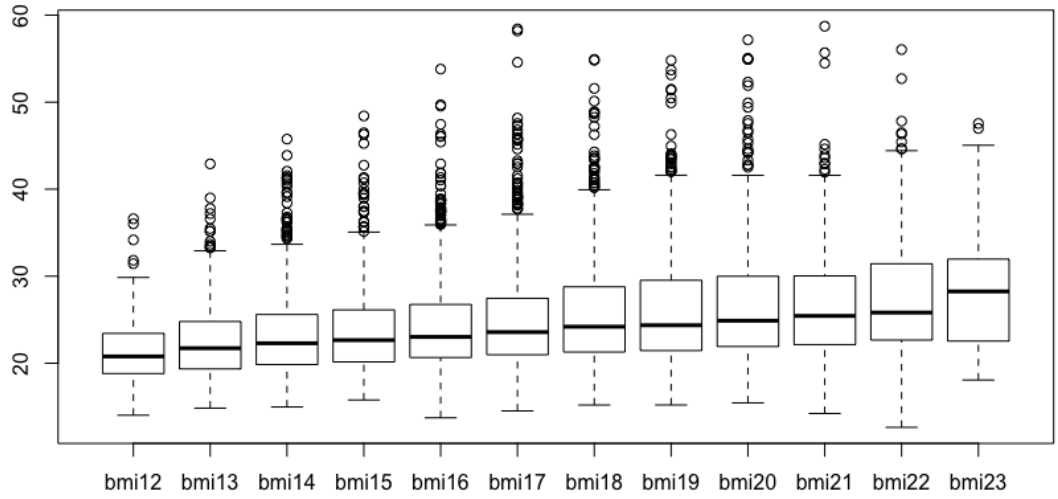


Figure 4.2: Boxplots for the twelve attributes of BMI ages 12-23 from NLSY

counterpart with the latent classes ranging from 1 to 6 until the best model is selected under each scenario considered. Like in the simulation studies, we only consider the general and most constrained models among the families of the GMMs herein. Table 4.5 shows the results of fitting general and constrained normal, skew-t and generalized hyperbolic GMMs for varying number of latent classes. In light of the BIC, the results show that all GHD-GMM models are superior to the GST-GMM models and the Gaussian GMMs. The best one is the general GHD-GMM with two latent classes. The resulting ML estimates of the key model parameters are shown in Table 4.6. Not surprisingly, the Gaussian GMM select more than two latent classes because the data attributes exhibit skewness.

Table 4.5: Results of fitting general and most constrained normal, skew-t and generalized hyperbolic GMMs for BMI development from NLSY

| | GMM-Normal(general) | | | GMM-Normal(constrained) | | |
|---|---|---|---|---|---|---|
| Classes | Loglikelihood | Free paras | BIC | Loglikelihood | Free paras | BIC |
| 1 | -17012.734 | 21 | -34173.647 | -17012.734 | 21 | -34173.647 |
| 2 | -15737.978 | 43 | -31775.371 | -16851.028 | 25 | -33878.461 |
| 3 | -15542.482 | 65 | -31543.615 | -16769.795 | 29 | -33744.220 |
| 4 | -15462.996 | 87 | -31539.878 | -16720.005 | 33 | -33672.863 |
| 5 | | | | -16685.476 | 37 | -33632.506 |
| 6 | | | | -16684.621 | 41 | -33658.546 |

| | GST-GMM(general) | | | GST-GMM(constrained) | | |
|---|---|---|---|---|---|---|
| Classes | Loglikelihood | Free paras | BIC | Loglikelihood | Free paras | BIC |
| 1 | -15730.611 | 25 | -31637.626 | -15744.836 | 25 | -31666.077 |
| 2 | -15512.380 | 51 | -31384.625 | -15703.949 | 29 | -31612.526 |
| 3 | -15443.812 | 77 | -31430.949 | -15709.897 | 33 | -31652.647 |

| | GHD-GMM(general) | | | GHD-GMM(constrained) | | |
|---|---|---|---|---|---|---|
| Classes | Loglikelihood | Free paras | BIC | Loglikelihood | Free paras | BIC |
| 1 | -15607.208 | 26 | -31397.876 | -15607.208 | 26 | -31397.876 |
| 2 | -15471.416 | 53 | -31316.808 | -15582.629 | 30 | -31376.944 |
| 3 | -15416.920 | 80 | -31398.334 | -15574.508 | 34 | -31388.926 |

Table 4.6: The estimated key model parameters of the two-class general GHD-GMM for BMI from NLSY

| Class 1 | Class 2 |
|---|---|
| $\hat{\lambda}_1 = 0.64$ | $\hat{\lambda}_2 = -0.02$ |
| $\hat{\omega}_1 = 0.88$ | $\hat{\omega}_2 = 0.79$ |
| $\hat{\boldsymbol{\alpha}}_1 = c(9.48, 1.70, 9.48)^{'}$ | $\hat{\boldsymbol{\alpha}}_2 = c(10.83, 6.99, 10.83)^{'}$ |
| $\hat{\boldsymbol{\beta}}_1 = c(1.14, 0.73, 1.14)^{'}$ | $\hat{\boldsymbol{\beta}}_2 = c(1.58, 2.31, 1.58)^{'}$ |

## 4.5 Discussion

In this chapter we have proposed the GHD-GMM and GST-GMM with arbitrary patterns of missing values, which allows the analysts to fit longitudinal data in the simultaneous presence of asymmetry, heavy tail weights, and missing values. We have

developed a computationally tractable AECM algorithm for carrying out ML estimation based on nice statistical properties of the normal variance-mean mixture. Rather than deleting or filling in incomplete cases, ML treats the missing data as latent or unobserved random variables to be updated at each iteration until convergence. The computation procedure for the imputation of the missing values and the estimation of the random effects are easy to implement once the ML estimates are achieved.

An advantage of the GHD and GST distributions for the random effects is their propensity for accommodating asymmetry and heavier tail weight than its normal distribution counterpart for random effects. Numerical results illustrated from simulated and real data indicate that our proposed models compares favourably to the conventional GMM counterpart when the normality assumption is violated. Even when the data are truly normally distributed, our proposed GHD-GMM with missing values could be used to check the reproducibility of a normal GMM solution due to the flexibility of the generalized hyperbolic distribution.

Looking forward, there are a number of extensions that would benefit from future research. Our proposed models are only applicable to single outcome longitudinal data, so a natural extension is to accommodate multi-outcome longitudinal data. We can also generalize our proposed models by replacing the polynomial regression at level 1 of the GMMs with other functional models, such as the spline and wavelet bases. Finally, a joint modelling of the time-to-event data and longitudinal data would be challenging and worthwhile extension by adding the standard Cox proportional hazard or accelerated failure time survival models.

# Chapter 5

# Mixtures of Generalized Hyperbolic Distributions and Mixtures of Skew-t Distributions for Model-Based Clustering with Incomplete Data

## 5.1 Introduction

As discussed in Chapter 1, more attention has been paid to develop mixture models that can accommodate incomplete data in model-based clustering. The ML approach to clustering incomplete data has been well studied and is often used, particularly for Gaussian mixture models (e.g., Ghahramani and Jordan, 1994; Lin et al., 2006;

Browne et al., 2013). Wang et al. (2004) present a framework of ML estimation using an EM algorithm to fit a mixture of multivariate $t$-distributions with arbitrary missing data patterns, which was generalized by Lin et al. (2009) to efficient supervised learning via the parameter expanded (PX-EM) algorithm (Liu et al., 1998) through two auxiliary indicator matrices. Lin (2014) further develops a family of multivariate-$t$ mixture models with 14 eigen-decomposed scale matrices in the presence of missing data through a computationally flexible EM algorithm by incorporating two auxiliary indicator matrices.

In this chapter, we consider fitting mixtures of generalized hyperbolic distributions (MGHD) and mixtures of multivariate skew-t distributions (MST) with missing information. In each case, an EM algorithm is used for parameter estimation. In addition to considering missing data, we develop families of MGHD and MST mixture models, each with 14 parsimonious eigen-decomposed scale matrices corresponding to the famous Gaussian parsimonious clustering models (GPCMs) of Banfield and Raftery (1993) and Celeux and Govaert (1995).

## 5.2 Statistical Properties of the GHD and GST

Before outlining the details for this algorithm, we first present some distributional properties of the GHD and GST, which is useful for developing the parameter estimation presented in Section 5.3.

The following result shows an appealing closure property of the GHD and GST under affine transformation and conditioning as well as the formation of marginal distributions. Suppose that $\mathbf{X}$ is a $p$-dimensional random vector having a GHD as in (2.13), i.e., $\mathbf{X} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Assume that $\mathbf{X}$ is partitioned as $\mathbf{X} =$

$(\mathbf{X}_1^\intercal, \mathbf{X}_2^\intercal)^\intercal$, where $\mathbf{X}_1$ takes values in $\mathbb{R}^{d_1}$ and $\mathbf{X}_2$ in $\mathbb{R}^{d_1} = \mathbb{R}^{p-d_1}$, with

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\beta}$ have similar partitions. Furthermore, $\boldsymbol{\Sigma}_{11}$ is $d_1 \times d_1$ and $\boldsymbol{\Sigma}_{22}$ is $d_2 \times d_2$.

**Proposition 5.2.1.** *Affine transformation of the GHD. If* $\mathbf{X} \sim GHD_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ *and* $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$ *where* $\mathbf{B} \in \mathbb{R}^{k \times p}$ *and* $\mathbf{b} \in \mathbb{R}^p$*, then*

$$\mathbf{X} \sim GHD_k(\lambda, \omega, \mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\intercal, \mathbf{B}\boldsymbol{\beta}), \tag{5.1}$$

*Proof.* The result follows by substituting (2.12) into $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$. □

**Proposition 5.2.2.** *The marginal distribution of* $\mathbf{X}_1$ *is a GHD as in* (2.13) *with index parameter* $\lambda$*, concentration parameter* $\omega$*, location vector* $\boldsymbol{\mu}_1$*, dispersion matrix* $\boldsymbol{\Sigma}_{11}$*, and skewness vector* $\boldsymbol{\beta}_1$*, i.e.,* $\mathbf{X}_1 \sim GHD_{d_1}(\lambda, \omega, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\beta}_1)$*.*

*Proof.* The result follows by applying Proposition 1 and choosing $\mathbf{B} = [\mathbf{I}_{d_1}, \mathbf{0}]$ and $\mathbf{b} = \mathbf{0}$. The parameters $\lambda, \omega$ inherited from the mixing distribution $W \sim \mathcal{I}(\lambda, \eta = 1, \omega)$ remain the same under the affine transformation and marginal distribution. □

**Proposition 5.2.3.** *The conditional distribution of* $\mathbf{X}_2$ *given* $\mathbf{X}_1 = \mathbf{x}_1$ *is a GHD as*

*in* (2.10)*, i.e.,* $\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1 \sim GH_{d_2}(\lambda_{2|1}, \chi_{2|1}, \psi_{2|1}, \boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}, \boldsymbol{\beta}_{2|1})$*, where*

$$\begin{aligned}
\lambda_{2|1} &= \lambda - \frac{d_1}{2}, & \chi_{2|1} &= \omega + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\intercal \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1),\\
\psi_{2|1} &= \omega + \boldsymbol{\beta}_1^\intercal \boldsymbol{\Sigma}_{11}^\intercal \boldsymbol{\beta}, & \boldsymbol{\mu}_{2|1} &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^\intercal \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1),\\
\boldsymbol{\Sigma}_{2|1} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\intercal \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}, & \boldsymbol{\beta}_{2|1} &= \boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\intercal \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1.
\end{aligned}$$

The proof of Proposition 5.2.3 is given in Appendix B.2.

Similarly, suppose that $\mathbf{X}$ is a $p$-dimensional random vector having the multivariate skew-t distribution as in (2.14), i.e., $\mathbf{X} \sim \mathrm{GST}_p(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Assume that $\mathbf{X}$ is partitioned as $\mathbf{X} = (\mathbf{X}_1^\intercal, \mathbf{X}_2^\intercal)^\intercal$, where $\mathbf{X}_1$ takes values in $\mathbb{R}^{d_1}$ and $\mathbf{X}_2$ in $\mathbb{R}^{d_1} = \mathbb{R}^{p-d_1}$, with

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\beta}$ have similar partitions. Furthermore, $\boldsymbol{\Sigma}_{11}$ is $d_1 \times d_1$ and $\boldsymbol{\Sigma}_{22}$ is $d_2 \times d_2$.

**Proposition 5.2.4.** *Affine transformation of the multivariate skew-t distribution. If* $\mathbf{X} \sim GST_p(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ *and* $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$*, where* $\mathbf{B} \in \mathbb{R}^{k \times p}$ *and* $\mathbf{b} \in \mathbb{R}^p$*, then*

$$\mathbf{X} \sim GST_k(\nu, \mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\intercal, \mathbf{B}\boldsymbol{\beta}). \tag{5.2}$$

*Proof.* The proof is similar to Proposition 5.2.1, hence is omitted here. □

**Proposition 5.2.5.** *The marginal distribution of* $\mathbf{X}_1$ *is a multivariate skew-t distribution as in* (2.14) *with degree of freedom parameter* $\nu$*, location vector* $\boldsymbol{\mu}_1$*, dispersion*

*matrix* $\boldsymbol{\Sigma}_{11}$*, and skewness vector* $\boldsymbol{\beta}_1$*, i.e.,* $\mathbf{X}_1 \sim GST_{d_1}(\nu, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\beta}_1)$.

*Proof.* The proof follows easily by applying Proposition 5.2.4 and choosing $\mathbf{B} = [\mathbf{I}_{d_1}, \mathbf{0}]$ and $\mathbf{b} = \mathbf{0}$. The degree of freedom parameter $\nu$ inherited from the mixing distribution $W \sim \text{IG}(\nu/2, \nu/2)$ remains invariant under affine transformation and marginal distribution. □

**Proposition 5.2.6.** *The conditional distribution of* $\mathbf{X}_2$ *given* $\mathbf{X}_1 = \mathbf{x}_1$ *is a GHD as in* (2.10)*, i.e.,* $\mathbf{X}_2 \mid \mathbf{x}_1 \sim GH_{d_2}(\lambda_{2|1}, \chi_{2|1}, \psi_{2|1}, \boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}, \boldsymbol{\beta}_{2|1})$*, where*

$$\begin{aligned}
\lambda_{2|1} &= -(\nu + d_1)/2, & \chi_{2|1} &= \nu + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\intercal \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \\
\psi_{2|1} &= \boldsymbol{\beta}_1^\intercal \boldsymbol{\Sigma}_{11}^\intercal \boldsymbol{\beta}, & \boldsymbol{\mu}_{2|1} &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^\intercal \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \\
\boldsymbol{\Sigma}_{2|1} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\intercal \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}, & \boldsymbol{\beta}_{2|1} &= \boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\intercal \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1.
\end{aligned}$$

The proof of Proposition 5.2.6 is similar to that for Proposition 5.2.3, hence is omitted.

## 5.3 Methodology

### 5.3.1 MGHD with Incomplete Data

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be $p$-dimensional random variables arising from a heterogeneous population with $G$ disjoint MGHD subpopulations. That is, each $\mathbf{X}_i$ has the density

$$f_{\text{MGHD}}(\mathbf{x}_i \mid \boldsymbol{\Theta}) = \sum_{g=1}^{G} \pi_g f_{\text{GHD}}(\mathbf{x}_i \mid \lambda_g, \omega_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g), \tag{5.3}$$

where $\pi_g > 0$, such that $\sum_{g=1}^{G} \pi_g = 1$ are the mixing proportions, $\boldsymbol{\Theta}$ denotes the model parameters, and $f_{\text{GHD}}(\mathbf{x}_i \mid \lambda_g, \omega_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g)$ is the GHD density defined in (2.13).

To apply the MGHD model (5.3) in the clustering paradigm, introduce $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ig})^{\intercal}$, where $Z_{ig} = 1$ if observation $i$ is in component $g$ and $Z_{ig} = 0$ otherwise. We have $\mathbf{Z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_G)$, i.e., $\mathbf{Z}_i$ follows a multinomial distribution with one trial and cell probabilities $\pi_1, \ldots, \pi_G$.

A three-level hierarchical representation of the MGHD model (5.3) can be expressed by

$$\begin{aligned} \mathbf{X}_i \mid (w_{ig}, Z_{ig} = 1) &\sim \mathcal{N}(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g), \\ W_{ig} \mid (Z_{ig} = 1) &\sim \mathcal{I}(\lambda_g, \eta = 1, \omega_g), \\ \mathbf{Z}_i &\sim \mathcal{M}(1; \pi_1, \ldots, \pi_G). \end{aligned} \tag{5.4}$$

The complete-data consist of the observed $\mathbf{x}_i$ together with the missing group membership $z_{ig}$ and the latent $w_{ig}$, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$, and the complete-data log-likelihood is given by

$$l_{\text{c}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\log(\pi_g) + \log(\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g)) + \log(h(w_{ig} \mid \lambda_g, \omega_g))\right]. \tag{5.5}$$

Browne and McNicholas (2015) present an EM algorithm for parameter estimation with the MGHD when there is no missing data in $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We are interested in parameter estimation for the MGHD model (5.3) when $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are partially observed with arbitrary missing patterns. The missing data mechanism is assumed to be MAR. Assume now that we split $\mathbf{x}_i$ into two components, $\mathbf{x}_i^{\text{o}}$ and $\mathbf{x}_i^{\text{m}}$ that denote

the observed and missing components of $\mathbf{x}_i$, respectively. In general, each data vector $\mathbf{x}_i$ may have a different pattern of missing features, i.e., $\mathbf{x}_i = (\mathbf{x}_i^{\mathrm{o}_i\intercal}, \mathbf{x}_i^{\mathrm{m}_i\intercal})^\intercal$, but can be simplified for the sake of clarity.

For each $\mathbf{x}_i = (\mathbf{x}_i^{\mathrm{o}\intercal}, \mathbf{x}_i^{\mathrm{m}\intercal})^\intercal$, partition the vector mean $\boldsymbol{\mu}_g = (\boldsymbol{\mu}_{g,i}^{\mathrm{o}\intercal}, \boldsymbol{\mu}_{g,i}^{\mathrm{m}\intercal})^\intercal$, where $\boldsymbol{\mu}_{g,i}^{\mathrm{o}}$ and $\boldsymbol{\mu}_{g,i}^{\mathrm{m}}$ denote the sub-vectors of $\boldsymbol{\mu}_g$ matching the observed and missing components of $\mathbf{x}_i$, respectively. Similarly, the skewness vector is $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g,i}^{\mathrm{o}\intercal}, \boldsymbol{\beta}_{g,i}^{\mathrm{m}\intercal})^\intercal$ and the covariance matrix $\boldsymbol{\Sigma}_g$ as

$$\boldsymbol{\Sigma}_g = \begin{pmatrix} \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}} & \boldsymbol{\Sigma}_{g,i}^{\mathrm{om}} \\ \boldsymbol{\Sigma}_{g,i}^{\mathrm{mo}} & \boldsymbol{\Sigma}_{g,i}^{mm} \end{pmatrix} \text{ and } \quad \boldsymbol{\Sigma}_g^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{g,i}^{-1,\mathrm{oo}} & \boldsymbol{\Sigma}_{g,i}^{-1,\mathrm{om}} \\ \boldsymbol{\Sigma}_{g,i}^{-1,\mathrm{mo}} & \boldsymbol{\Sigma}_{g,i}^{-1,\mathrm{mm}} \end{pmatrix}, \tag{5.6}$$

correspond to $\mathbf{x}_i = (\mathbf{x}_i^{\mathrm{o}\intercal}, \mathbf{x}_i^{\mathrm{m}\intercal})^\intercal$. As a result, in addition to the observed $\mathbf{x}_i^{\mathrm{o}}$, the missing group membership $z_{ig}$, and the latent variable $w_{ig}$, the complete-data also include the missing data $\mathbf{x}_i^{\mathrm{m}}$. In the framework of the EM algorithm, the missing data $\mathbf{x}_i^{\mathrm{m}}$ are considered to be random variables that are updated in each iteration. Hence, the complete-data log-likelihood (5.5) is rewritten as

$$l_{\mathrm{c}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\big[\log \pi_g + \log \phi(\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}_i^{\mathrm{m}} \mid \boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g) + \log h_{\mathcal{I}}(w_{ig} \mid \lambda_g, \omega_g)\big].$$

Given (5.4), we establish the following:

- The marginal distribution of $\mathbf{X}_i^{\mathrm{o}}$ given is

$$\mathbf{X}_i^{\mathrm{o}} \sim \sum_{g=1}^{G} \pi_g f_{\mathrm{GHD},p_i^{\mathrm{o}}}(\lambda_g, \omega_g, \boldsymbol{\mu}_{g,i}^{\mathrm{o}}, \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}}, \boldsymbol{\beta}_{g,i}^{\mathrm{o}}),$$

  where $p_i^{\mathrm{o}}$ is the dimension corresponding to the observed component $\mathbf{x}_i^{\mathrm{o}}$, which

should be exactly written as $p_i^{\mathrm{o}_i}$ but here is simplified.

- The conditional distribution of $\mathbf{X}_i^{\mathrm{m}}$ given $\mathbf{x}_i^{\mathrm{o}}$ and $Z_{ig} = 1$, according to Proposition 3, is

$$\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig} = 1 \sim \mathrm{GH}_{p-p_i^{\mathrm{o}}}\left(\lambda_{g,i}^{\mathrm{m|o}}, \chi_{g,i}^{\mathrm{m|o}}, \psi_{g,i}^{\mathrm{m|o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{m|o}}, \boldsymbol{\Sigma}_{g,i}^{\mathrm{m|o}}, \boldsymbol{\beta}_{g,i}^{\mathrm{m|o}}\right), \tag{5.7}$$

where

$$\begin{aligned}
\lambda_{g,i}^{\mathrm{m|o}} &= \lambda_g - \frac{p_i^{\mathrm{o}}}{2}, & \chi_{g,i}^{\mathrm{m|o}} &= \omega_g + (\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}})^{\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}}),\\
\psi_g^{\mathrm{m|o}} &= \omega_g + \boldsymbol{\beta}_{g,i}^{\mathrm{o}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}}, & \boldsymbol{\mu}_{g,i}^{\mathrm{m|o}} &= \boldsymbol{\mu}_g^{\mathrm{m}} + \boldsymbol{\Sigma}_{g,i}^{\mathrm{om}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}}),\\
\boldsymbol{\Sigma}_{g,i}^{\mathrm{m|o}} &= \boldsymbol{\Sigma}_{g,i}^{\mathrm{mm}} - \boldsymbol{\Sigma}_{g,i}^{\mathrm{om}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}\boldsymbol{\Sigma}_{g,i}^{\mathrm{om}}, & \boldsymbol{\beta}_{g,i}^{\mathrm{m|o}} &= \boldsymbol{\beta}_{g,i}^{\mathrm{m}} - \boldsymbol{\Sigma}_{g,i}^{\mathrm{om}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}}.
\end{aligned}$$

- The conditional distribution of $\mathbf{X}_i^{\mathrm{m}}$ given $\mathbf{x}_i^{\mathrm{o}}, w_{ig}$, and $Z_{ig} = 1$ is

$$\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, w_{ig}, Z_{ig} = 1 \sim \mathcal{N}_{p-p_i^{\mathrm{o}}}(\boldsymbol{\mu}_{g,i}^{\mathrm{m|o}} + w_{ig}\boldsymbol{\beta}_{g,i}^{\mathrm{m|o}}, w_{ig}\boldsymbol{\Sigma}_{g,i}^{\mathrm{m|o}}). \tag{5.8}$$

- The conditional distribution of $W_i$ given $\mathbf{x}_i^{\mathrm{o}}$ and $Z_{ig} = 1$ is

$$W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig} = 1 \sim \mathrm{GIG}\left(\omega_g + \boldsymbol{\beta}_{g,i}^{\mathrm{o}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}}, \omega_g + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}}), \lambda_g - \frac{p_i^{\mathrm{o}}}{2}\right). \tag{5.9}$$

After a little algebra, we get the complete data log-likelihood function is

$$
\begin{aligned}
l_{\mathrm{c}}(\boldsymbol{\Theta}) = & \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[-\frac{p}{2}\log(2\pi) - \frac{p}{2}\log w_{ig} + \frac{1}{2}\log|\boldsymbol{\Sigma}_g^{-1}|\right] \\
& - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1} z_{ig}\frac{1}{w_{ig}}\begin{pmatrix} (\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}})(\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}})^{\intercal} & (\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}})(\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})^{\intercal} \\ (\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})^{\intercal}(\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}}) & (\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})(\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})^{\intercal} \end{pmatrix}\right) \\
& + \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1} z_{ig}\begin{pmatrix}\boldsymbol{\beta}_{g,i}^{\mathrm{o}}\\ \boldsymbol{\beta}_{g,i}^{\mathrm{m}}\end{pmatrix}\begin{pmatrix}(\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}})^{\intercal} & (\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})^{\intercal}\end{pmatrix}\right) \\
& + \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1} z_{ig}\begin{pmatrix}\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}}\\ \mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_{g,i}^{\mathrm{o}\intercal} & \boldsymbol{\beta}_{g,i}^{\mathrm{m}\intercal}\end{pmatrix}\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}w_{ig}\boldsymbol{\beta}_{g,i}^{\intercal}\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\beta}_{g,i} \\
& + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[(\lambda_g-1)\log w_{ig} - \log(2K_{\lambda_g}(\omega_g)) - \frac{\omega_g}{2}\left(w_{ig}+\frac{1}{w_{ig}}\right)\right]. \qquad (5.10)
\end{aligned}
$$

On the $k$th iteration of the E-step, the expected value of the complete data log-likelihood is computed given the observed data $\mathbf{x}_1^{\mathrm{o}},\ldots,\mathbf{x}_n^{\mathrm{o}}$ and the current parameter updates $\boldsymbol{\Theta}^{(k)}$. That is, we need to compute $\mathbb{E}(Z_{ig} \mid \mathbf{x}_i^{\mathrm{o}};\boldsymbol{\Theta}^{(k)})$, $\mathbb{E}(W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1;\boldsymbol{\Theta}^{(k)})$, $\mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1;\boldsymbol{\Theta}^{(k)})$, $\mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1;\boldsymbol{\Theta}^{(k)})$, $\mathbb{E}(\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1, w_i;\boldsymbol{\Theta}^{(k)})$, and $\mathbb{E}(\mathbf{X}_i^{\mathrm{m}}\mathbf{X}_i^{\mathrm{m}\intercal} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1, w_i;\boldsymbol{\Theta}^{(k)})$.

First, let $\hat{z}_{ig}^{(k)}$ denote the *a posteriori* probability that $i$-th observation belongs to the $g$-th component of the mixture, based on the observed data:

$$
\hat{z}_{ig}^{(k)} := \mathbb{E}(Z_{ig} \mid \mathbf{x}_i^{\mathrm{o}},\boldsymbol{\Theta}^{(k)}) = \frac{\pi_g^{(k)} f_{\mathrm{GHD},p_i^{\mathrm{o}}}(\mathbf{x}_i^{\mathrm{o}};\lambda_g^{(k)},\omega_g^{(k)},\boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)},\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)},\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}{\sum_{l=1}^{G}\pi_l^{(k)} f_{\mathrm{GHD},p_i^{\mathrm{o}}}(\mathbf{x}_i^{\mathrm{o}};\lambda_l^{(k)},\omega_l^{(k)},\boldsymbol{\mu}_{l,i}^{\mathrm{o}(k)},\boldsymbol{\Sigma}_{l,i}^{\mathrm{oo}(k)},\boldsymbol{\beta}_{l,i}^{\mathrm{o}(k)})}.
$$

Given (2.8) and (5.9), we have the following expectations as to the latent variable $W$:

$$a_{ig}^{(k)} := \mathbb{E}(W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \sqrt{\frac{\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})}{\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)}}} \times \frac{K_{\lambda_g^{(k)} - \frac{p_i^0}{2}+1}\left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)}{K_{\lambda_g^{(k)} - \frac{p_i^0}{2}}\left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)},$$

$$\begin{aligned} b_{ig}^{(k)} &:= \mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) \\ &= -\frac{2\lambda_g^{(k)} - p_i^{\mathrm{o}}}{\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})} + \sqrt{\frac{\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)}}{\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})}} \\ &\quad \times \frac{K_{\lambda_g^{(k)} - \frac{p_i^0}{2}+1}\left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)}{K_{\lambda_g^{(k)} - \frac{p_i^0}{2}}\left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)}, \end{aligned}$$

$$\begin{aligned} c_{ig}^{(k)} &:= \mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \log\left(\sqrt{\frac{\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})}{\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)}}}\right) \\ &\quad + \frac{\partial}{\partial t}\log\left\{K_t\left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)\right\}\Bigg|_{t=(\lambda_g^{(k)} - \frac{p_i^{\mathrm{o}}}{2})}. \end{aligned}$$

For convenience, we use the following notation analogous to Browne and McNicholas (2015): $n_g^{(k)} = \sum_{i=1}^n \hat{z}_{ig}^{(k)}$, $\bar{a}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} a_{ig}^{(k)}$, $\bar{b}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} b_{ig}^{(k)}$, and $\bar{c}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} c_{ig}^{(k)}$. For the actual missing data $\mathbf{X}^{\mathrm{m}}$, we will also need the

following expectations:

$$\begin{aligned}
\hat{\mathbf{x}}_{ig}^{\mathrm{m}(k)} &:= \mathbb{E}(\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig}=1) = \boldsymbol{\mu}_{g,i}^{\mathrm{m}|\mathrm{o}(k)} + a_{ig}^{(k)}\boldsymbol{\beta}_{g,i}^{\mathrm{m}|\mathrm{o}(k)},\\
\tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} &:= \mathbb{E}((1/W_i)\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig}=1) = b_{ig}^{(k)}\boldsymbol{\mu}_{g,i}^{\mathrm{m}|\mathrm{o}(k)} + \boldsymbol{\beta}_{g,i}^{\mathrm{m}|\mathrm{o}(k)},\\
\tilde{\tilde{\mathbf{x}}}_{ig}^{\mathrm{m}(k)} &:= \mathbb{E}((1/W_i)\mathbf{X}_i^{\mathrm{m}}\mathbf{X}_i^{\mathrm{m}\intercal} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig}=1) = \boldsymbol{\Sigma}_{g,i}^{\mathrm{m}|\mathrm{o}(k)} + b_{ig}^{(k)}\boldsymbol{\mu}_{g,i}^{\mathrm{m}|\mathrm{o}(k)}(\boldsymbol{\mu}_{g,i}^{\mathrm{m}|\mathrm{o}(k)})^{\intercal}\\
&\quad + \boldsymbol{\mu}_{g,i}^{\mathrm{m}|\mathrm{o}(k)}(\boldsymbol{\beta}_{g,i}^{\mathrm{m}|\mathrm{o}(k)})^{\intercal} + \boldsymbol{\beta}_{g,i}^{\mathrm{m}|\mathrm{o}(k)}(\boldsymbol{\mu}_{g,i}^{\mathrm{m}|\mathrm{o}(k)})^{\intercal} + a_{ig}^{(k)}\boldsymbol{\beta}_{g,i}^{\mathrm{m}|\mathrm{o}(k)}(\boldsymbol{\beta}_{g,i}^{\mathrm{m}|\mathrm{o}(k)})^{\intercal}.
\end{aligned}$$

On the $k$-th iteration of the M-step, the expected value of the complete data log-likelihood is maximized to get the updates for the parameter estimates as follows:

$$\begin{aligned}
\pi_g^{(k+1)} &= \frac{n_g^{(k)}}{n},\\
\boldsymbol{\mu}_g^{(k+1)} &= \frac{1}{\sum_{i=1}^n \hat{z}_{ig}^{(k)}(\bar{a}_g^{(k)}b_{ig}^{(k)}-1)}\sum_{i=1}^n \hat{z}_{ig}^{(k)}\begin{pmatrix}(\bar{a}_g^{(k)}b_{ig}^{(k)}-1)\mathbf{x}_i^{\mathrm{o}}\\ \bar{a}_g^{(k)}\tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} - \hat{\mathbf{x}}_{ig}^{\mathrm{m}(k)}\end{pmatrix},\\
\boldsymbol{\beta}_g^{(k+1)} &= \frac{1}{\sum_{i=1}^n \hat{z}_{ig}^{(k)}(\bar{a}_g^{(k)}b_{ig}^{(k)}-1)}\sum_{i=1}^n \hat{z}_{ig}^{(k)}\begin{pmatrix}(\bar{b}_g^{(k)}-b_{ig}^{(k)})\mathbf{x}_i^{\mathrm{o}}\\ \bar{b}_g^{(k)}\hat{\mathbf{x}}_{ig}^{\mathrm{m}(k)} - \tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)}\end{pmatrix},\\
\boldsymbol{\Sigma}_g^{(k+1)} &= \frac{1}{n_g^{(k)}}\sum_{i=1}^n \hat{z}_{ig}^{(k)}\boldsymbol{\Sigma}_{ig}^{(k+1)} - (\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g^{(k+1)})\boldsymbol{\beta}_g^{(k+1)\intercal} - \boldsymbol{\beta}_g^{(k+1)}(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g^{(k+1)})^{\intercal} + \bar{a}_g^{(k)}\boldsymbol{\beta}_g^{(k+1)}\boldsymbol{\beta}_g^{(k+1)\intercal},
\end{aligned}$$

where

$$\begin{aligned}
\bar{\mathbf{x}}_g &= \frac{1}{n_g^{(k)}}\sum_{i=1}^n \hat{z}_{ig}^{(k)}\begin{pmatrix}\mathbf{x}_i^{\mathrm{o}}\\ \hat{\mathbf{x}}_{ig}^{\mathrm{m}(k)}\end{pmatrix},\\
\boldsymbol{\Sigma}_{ig}^{(k+1)} &= \begin{pmatrix} b_{ig}^{(k)}(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_g^{\mathrm{o}(k+1)})(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_g^{\mathrm{o}(k+1)})^{\intercal} & (\mathbf{x}_i^{\mathrm{o}} - \hat{\boldsymbol{\mu}}_g^{\mathrm{o}(k+1)})(\tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} - b_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)})^{\intercal}\\ (\tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} - b_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)})(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_g^{\mathrm{o}(k+1)})^{\intercal} & \mathbf{k}_{ig}^{\mathrm{m}(k+1)}\end{pmatrix},
\end{aligned}$$

where

$$\mathbf{k}_{ig}^{\mathrm{m}(k+1)} = \tilde{\tilde{\mathbf{x}}}_{ig}^{\mathrm{m}(k)} - \tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)\intercal} - \hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)}\tilde{\mathbf{x}}_i^{\mathrm{m}(k)\intercal} + b_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)\intercal}.$$

Finally, the estimates of $\lambda_g^{(k+1)}$ and $\omega_g^{(k+1)}$ are given as solutions to maximize the function

$$q_g(\lambda_g, \omega_g) = -\log(K_{\lambda_g}(\omega_g)) + (\lambda_g - 1)\bar{c}_g - \frac{\omega_g}{2}(\bar{a}_g + \bar{b}_g),$$

and the associated updates are

$$\lambda_g^{(k+1)} = \bar{c}_g^{(k)}\lambda_g^{(k)}\left[\frac{\partial}{\partial\lambda_g^{(k)}}\log\left(K_{\lambda_g^{(k)}}(\omega_g^{(k)})\right)\right]^{-1},$$
$$\omega_g^{(k+1)} = \omega_g^{(k)} - \left[\frac{\partial}{\partial\omega_g^{(k)}}q_g(\lambda_g^{(k+1)}, \omega_g^{(k)})\right]\left[\frac{\partial^2}{\partial\omega_g^{2(k)}}q_g(\lambda_g^{(k+1)}, \omega_g^{(k)})\right]^{-1}.$$

### 5.3.2 MST with Incomplete Data

Analogous to the MGHD model (5.3), the MST model takes the density

$$f_{\mathrm{MST}}(\mathbf{X}_i \mid \boldsymbol{\Theta}) = \sum_{g=1}^{G}\pi_g f_{\mathrm{GST}}(\mathbf{X}_i \mid \nu_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g), \tag{5.11}$$

where $\boldsymbol{\Theta} = (\pi, \mathbf{v}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g)$ with $\mathbf{v}_g = (\nu_1, \ldots, \nu_g)$ and $\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$, and $\boldsymbol{\beta}_g$ are as defined above. By introducing the group membership variables $\mathbf{Z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_G)$,

convenient three-layer hierarchical representations are given by

$$\begin{aligned} \mathbf{X}_i \mid (w_{ig}, Z_{ig} = 1) &\sim \mathcal{N}(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g) \\ W_{ig} \mid (Z_{ig} = 1) &\sim \mathrm{IG}(\nu_g/2, \nu_g/2). \\ \mathbf{Z}_i &\sim \mathcal{M}(1; \pi_1, \ldots, \pi_G) \end{aligned} \tag{5.12}$$

Assume that the matrix $\mathbf{X} = (\mathbf{X}^{\mathrm{o}\intercal}, \mathbf{X}^{\mathrm{m}\intercal})^\intercal$ contains missing data. For each $\mathbf{x}_i = (\mathbf{x}_i^{\mathrm{o}\intercal}, \mathbf{x}_i^{\mathrm{m}\intercal})^\intercal$, we write $\boldsymbol{\mu}_g = (\boldsymbol{\mu}_{g,i}^{\mathrm{o}\intercal}, \boldsymbol{\mu}_{g,i}^{\mathrm{m}\intercal})^\intercal$, $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g,i}^{\mathrm{o}\intercal}, \boldsymbol{\beta}_{g,i}^{\mathrm{m}\intercal})^\intercal$, and finally the $g$th dispersion matrix $\boldsymbol{\Sigma}_g$ is partitioned as in (5.6). Hence, based on (5.12), we have the following conditional distributions:

- The marginal distribution of $\mathbf{X}_i^{\mathrm{o}}$ is

$$\mathbf{X}_i^{\mathrm{o}} \sim \sum_{g=1}^{G} \pi_g f_{\mathrm{GST}, p_i^{\mathrm{o}}}(\lambda_g, \omega_g, \boldsymbol{\mu}_{g,i}^{\mathrm{o}}, \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}}, \boldsymbol{\beta}_{g,i}^{\mathrm{o}}),$$

  where $p_i^{\mathrm{o}}$ is the dimension corresponding to the observed component $\mathbf{x}_i^{\mathrm{o}}$, which should be exactly written as $p_i^{\mathrm{o}_i}$ but here is simplified.

- The conditional distribution of $\mathbf{X}_i^{\mathrm{m}}$ given $\mathbf{x}_i^{\mathrm{o}}$ and $Z_{ig} = 1$, according to Proposition 6, is

$$\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig} = 1 \sim \mathrm{GH}_{p-p_i^{\mathrm{o}}}(\lambda_{g,i}^{\mathrm{m|o}}, \chi_{g,i}^{\mathrm{m|o}}, \psi_{g,i}^{\mathrm{m|o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{m|o}}, \boldsymbol{\Sigma}_{g,i}^{\mathrm{m|o}}, \boldsymbol{\beta}_{g,i}^{\mathrm{m|o}}), \tag{5.13}$$

where

$$\lambda_{g,i}^{\mathrm{m|o}} = -\frac{\nu_g + p_i^{\mathrm{o}}}{2}, \qquad \psi_{g,i}^{\mathrm{m|o}} = \nu_g + (\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}})^{\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}}),$$
$$\psi_{g,i}^{\mathrm{m|o}} = \boldsymbol{\beta}_{g,i}^{\mathrm{o}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}}, \qquad \boldsymbol{\mu}_{g,i}^{\mathrm{m|o}} = \boldsymbol{\mu}_{g,i}^{\mathrm{m}} + \boldsymbol{\Sigma}_{g,i}^{\mathrm{om}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}}),$$
$$\boldsymbol{\Sigma}_{g,i}^{\mathrm{m|o}} = \boldsymbol{\Sigma}_{g,i}^{\mathrm{mm}} - \boldsymbol{\Sigma}_{g,i}^{\mathrm{om}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}\boldsymbol{\Sigma}_{g,i}^{\mathrm{om}}, \quad \boldsymbol{\beta}_{g,i}^{\mathrm{m|o}} = \boldsymbol{\beta}_{g,i}^{\mathrm{m}} - \boldsymbol{\Sigma}_{g,i}^{\mathrm{om}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}}.$$

- The conditional distribution of $\mathbf{X}_i^{\mathrm{m}}$ given $\mathbf{x}_i^{\mathrm{o}}, w_{ig}$, and $Z_{ig} = 1$ is

$$\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, w_{ig}, Z_{ig} = 1 \sim \mathcal{N}_{p-p_i^{\mathrm{o}}}(\boldsymbol{\mu}_{g,i}^{\mathrm{m|o}} + w_{ig}\boldsymbol{\beta}_{g,i}^{\mathrm{m|o}}, w_{ig}\boldsymbol{\Sigma}_{g,i}^{\mathrm{m|o}}). \tag{5.14}$$

- The conditional distribution of $W_i$ given $\mathbf{x}_i^{\mathrm{o}}$ and $Z_{ig} = 1$ is

$$W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig} = 1 \sim \mathrm{GIG}\left(\boldsymbol{\beta}_{g,i}^{\mathrm{o}\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}}, \nu_g + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}}), -\frac{\nu_g + p_i^{\mathrm{o}}}{2}\right). \tag{5.15}$$

As in the case of the MGHD model with incomplete data, the complete data consists of the observed $\mathbf{x}_i$, the missing group membership $z_{ig}$, the latent $w_{ig}$, as well as the actual missing data $\mathbf{x}_i^{\mathrm{m}}$, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$. Again, the complete data log-likelihood function is given by

$$l_{\mathrm{c}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\log\pi_g + \log\phi(\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}_i^{\mathrm{m}} \mid \boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g)\right.$$
$$\left. + \log f_{\mathrm{IG}}(w_{ig} \mid \nu_g/2, \nu_g/2)\right]. \tag{5.16}$$

Furthermore, one can simplify (5.16) to

$$
\begin{aligned}
l_{\mathrm{c}}(\boldsymbol{\Theta}) = & \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[-\frac{p}{2}\log(2\pi) - \frac{p}{2}\log w_{ig} + \frac{1}{2}\log|\boldsymbol{\Sigma}_g^{-1}|\right] \\
& - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1} z_{ig}\frac{1}{w_{ig}}\begin{pmatrix} (\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}})(\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}})^{\intercal} & (\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}})(\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})^{\intercal} \\ (\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})^{\intercal}(\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}}) & (\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})(\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})^{\intercal} \end{pmatrix}\right) \\
& + \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1} z_{ig}\begin{pmatrix}\boldsymbol{\beta}_{g,i}^{\mathrm{o}}\\ \boldsymbol{\beta}_{g,i}^{\mathrm{m}}\end{pmatrix}\begin{pmatrix}(\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}})^{\intercal} & (\mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}})^{\intercal}\end{pmatrix}\right) \\
& + \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1} z_{ig}\begin{pmatrix}\mathbf{x}_i^{\mathrm{o}}-\boldsymbol{\mu}_{g,i}^{\mathrm{o}}\\ \mathbf{x}_i^{\mathrm{m}}-\boldsymbol{\mu}_{g,i}^{\mathrm{m}}\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_{g,i}^{\mathrm{o}\intercal} & \boldsymbol{\beta}_{g,i}^{\mathrm{m}\intercal}\end{pmatrix}\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}w_{ig}\boldsymbol{\beta}_{g,i}^{\intercal}\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\beta}_{g,i} \\
& + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\frac{\nu_g}{2}\log\left(\frac{\nu_g}{2}\right) - \log\Gamma\left(\frac{\nu_g}{2}\right) - \left(\frac{\nu_g}{2}+1\right)\log w_{ig} - \frac{\nu_g}{2w_{ig}}\right]. \qquad (5.17)
\end{aligned}
$$

On the $k$th iteration of the E-step, the expected value of the complete-data log-likelihood is computed given the observed data $\mathbf{X}^{\mathrm{o}}$ and the current parameter updates $\boldsymbol{\Theta}^{(k)}$. Denote by $\tau_{ig}^{(k)}$ the *a posteriori* probability that the $i$th observation belongs to the $g$th component of the mixture. Specifically, it can be calculated as

$$
\tau_{ig}^{(k)} := \mathbb{E}(Z_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\Theta}^{(k)}) = \frac{\pi_g^{(k)} f_{\mathrm{GST},p_i^{\mathrm{o}}}(\mathbf{x}_i^{\mathrm{o}}; \nu_g^{(k)}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)}, \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}, \boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}{\sum_{l=1}^{G}\pi_l^{(k)} f_{\mathrm{GST},p_i^{\mathrm{o}}}(\mathbf{x}_i^{\mathrm{o}}; v_l^{(k)}, \boldsymbol{\mu}_{l,i}^{\mathrm{o}(k)}, \boldsymbol{\Sigma}_{l,i}^{\mathrm{oo}(k)}, \boldsymbol{\beta}_{l,i}^{\mathrm{o}(k)})}.
$$

Given the observed data $\mathbf{x}^{\mathrm{o}}$, the current parameter updates $\boldsymbol{\Theta}^{(k)}$, and conditional distributions (5.13) and (5.15), taking expectations for (5.17) leads to the following expectation updates in the E-step:

$$A_{ig}^{(k)} := \mathbb{E}(W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \sqrt{\frac{\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})}{\boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)}}}$$
$$\times \frac{K_{-(\nu_g^{(k)}+p_i^0)/2+1}\left(\sqrt{(\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)}{K_{-(\nu_g^{(k)}+p_i^0)/2}\left(\sqrt{(\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)},$$

$$B_{ig}^{(k)} := \mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1; \boldsymbol{\Theta}^{(k)})$$
$$= \frac{\nu_g^{(k)} + p_i^{\mathrm{o}}}{\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})} + \sqrt{\frac{\boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)}}{\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})}}$$
$$\times \frac{K_{-(\nu_g^{(k)}+p_i^0)/2+1}\left(\sqrt{(\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)}{K_{-(\nu_g^{(k)}+p_i^0)/2}\left(\sqrt{(\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)},$$

$$C_{ig}^{(k)} := \mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \log\left(\sqrt{\frac{\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})}{\boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)}}}\right)$$
$$+ \frac{\partial}{\partial t}\log\left\{K_t\left(\sqrt{(\nu_g^{(k)} + \delta(\mathbf{x}_i^{\mathrm{o}}, \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)}))(\boldsymbol{\beta}_{g,i}^{o(k)\intercal}(\boldsymbol{\Sigma}_{g,i}^{\mathrm{oo}(k)})^{-1}\boldsymbol{\beta}_{g,i}^{\mathrm{o}(k)})}\right)\right\}\Bigg|_{t=-(\nu_g^{(k)}+p_i^{\mathrm{o}})/2},$$

$$\hat{\mathbf{x}}_{ig}^{\mathrm{m}(k)} := \mathbb{E}(\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig} = 1) = \boldsymbol{\mu}_{g,i}^{\mathrm{m|o}(k)} + A_{ig}^{(k)}\boldsymbol{\beta}_{g,i}^{\mathrm{m|o}(k)},$$

$$\tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} := \mathbb{E}((1/W_i)\mathbf{X}_i^{\mathrm{m}} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig} = 1) = B_{ig}^{(k)}\boldsymbol{\mu}_{g,i}^{\mathrm{m|o}(k)} + \boldsymbol{\beta}_{g,i}^{\mathrm{m|o}(k)},$$

$$\tilde{\tilde{\mathbf{x}}}_{ig}^{\mathrm{m}(k)} := \mathbb{E}((1/w_i)\mathbf{X}_i^{\mathrm{m}}\mathbf{X}_i^{m\intercal} \mid \mathbf{x}_i^{\mathrm{o}}, Z_{ig} = 1) = \boldsymbol{\Sigma}_{g,i}^{\mathrm{m|o}(k)} + B_{ig}^{(k)}\boldsymbol{\mu}_{g,i}^{\mathrm{m|o}(k)}(\boldsymbol{\mu}_{g,i}^{\mathrm{m|o}(k)})^{\intercal}$$
$$+ \boldsymbol{\mu}_{g,i}^{\mathrm{m|o}(k)}(\boldsymbol{\beta}_{g,i}^{\mathrm{m|o}(k)})^{\intercal} + \boldsymbol{\beta}_{g,i}^{\mathrm{m|o}(k)}(\boldsymbol{\mu}_{g,i}^{\mathrm{m|o}(k)})^{\intercal} + A_{ig}^{(k)}\boldsymbol{\beta}_{g,i}^{\mathrm{m|o}(k)}(\boldsymbol{\beta}_{g,i}^{\mathrm{m|o}(k)})^{\intercal}.$$

For convenience, let $n_g^{(k)} = \sum_{i=1}^n \tau_{ig}^{(k)}$, $\bar{A}_g^{(k)} = 1/n_g^{(k)}\sum_{i=1}^n \tau_{ig}^{(k)}A_{ig}^{(k)}$, $\bar{B}_g^{(k)} = 1/n_g^{(k)}\sum_{i=1}^n \tau_{ig}^{(k)}B_{ig}^{(k)}$, and $\bar{C}_g^{(k)} = 1/n_g^{(k)}\sum_{i=1}^n \tau_{ig}^{(k)}C_{ig}^{(k)}$. On the $k$th iteration of the M-step, we get updates for the parameter estimates of the mixture as follows:

$$\pi_g^{(k+1)} = \frac{n_g^{(k)}}{n},$$

$$\boldsymbol{\mu}_g^{(k+1)} = \frac{1}{\sum_{i=1}^n \hat{\tau}_{ig}^{(k)}(\bar{A}_g^{(k)}B_{ig}^{(k)} - 1)} \sum_{i=1}^n \hat{\tau}_{ig}^{(k)} \begin{pmatrix} (\bar{A}_g^{(k)}B_{ig}^{(k)} - 1)\mathbf{x}_i^{\mathrm{o}} \\ \bar{A}_g^{(k)}\tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} - \hat{\mathbf{x}}_{ig}^{\mathrm{m}(k)} \end{pmatrix},$$

$$\boldsymbol{\beta}_g^{(k+1)} = \frac{1}{\sum_{i=1}^n \hat{\tau}_{ig}^{(k)}(\bar{A}_g^{(k)}B_{ig}^{(k)} - 1)} \sum_{i=1}^n \hat{\tau}_{ig}^{(k)} \begin{pmatrix} (\bar{B}_g^{(k)} - B_{ig}^{(k)})\mathbf{x}_i^{\mathrm{o}} \\ \bar{B}_g^{(k)}\hat{\mathbf{x}}_{ig}^{\mathrm{m}(k)} - \tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} \end{pmatrix}$$

$$\boldsymbol{\Sigma}_g^{(k+1)} = \frac{1}{n_g^{(k)}} \sum_{i=1}^n \hat{\tau}_{ig}^{(k)}\boldsymbol{\Sigma}_{ig}^{(k+1)} - (\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g^{(k+1)})\boldsymbol{\beta}_g^{(k+1)\intercal} - \boldsymbol{\beta}_g^{(k+1)}(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g^{(k+1)})^{\intercal} + \bar{A}_g^{(k)}\boldsymbol{\beta}_g^{(k+1)}\boldsymbol{\beta}_g^{(k+1)\intercal},$$

where

$$\bar{\mathbf{x}}_g = \frac{1}{n_g^{(k)}} \sum_{i=1}^n \hat{\tau}_{ig}^{(k)} \begin{pmatrix} \mathbf{x}_i^{\mathrm{o}} \\ \hat{\mathbf{x}}_{ig}^{\mathrm{m}(k)} \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{ig}^{(k+1)} = \begin{pmatrix} B_{ig}^{(k)}(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k+1)})(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k+1)})^{\intercal} & (\mathbf{x}_i^{\mathrm{o}} - \hat{\boldsymbol{\mu}}_g^{\mathrm{o}(k+1)})(\tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} - B_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)})^{\intercal} \\ (\tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)} - B_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)})(\mathbf{x}_i^{\mathrm{o}} - \boldsymbol{\mu}_{g,i}^{\mathrm{o}(k+1)})^{\intercal} & \mathbf{k}_{ig}^{\mathrm{m}(k+1)} \end{pmatrix},$$

where

$$\mathbf{k}_{ig}^{\mathrm{m}(k+1)} = \tilde{\tilde{\mathbf{x}}}_{ig}^{\mathrm{m}(k+1)} - \tilde{\mathbf{x}}_{ig}^{\mathrm{m}(k)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)T} - \hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)}\tilde{\mathbf{x}}_i^{\mathrm{m}(k)\intercal} + B_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)}\hat{\boldsymbol{\mu}}_g^{\mathrm{m}(k+1)\intercal}.$$

Finally, as for the degree of freedom parameter $\nu_g$, the update does not exist in closed form. The update $\nu_g^{(k+1)}$ is the solution of

$$\log\left(\frac{\nu_g^{(k+1)}}{2}\right) + 1 - \varphi\left(\frac{\nu_g^{(k+1)}}{2}\right) - \frac{1}{n_g^{(k)}} \sum_{i=1}^n \tau_{ig}(C_{ig}^{(k)} + B_{ig}^{(k)}) = 0, \tag{5.18}$$

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

### 5.3.3 Notes on Implementation

It is well known that the EM algorithm can be heavily dependent on the initial values; indeed, good initial values of parameter estimates may speed up convergence. In this study, the following procedure for automatically generating initial values is used, unless otherwise specified.

- Fill in the missing values based on the mean imputation method.
- Perform $k$-means clustering and use the resulting clustering membership to initialize the *a posteriori* probability $\hat{z}_{ig}^0$. Accordingly, the initial values for the model parameters are then given by:

$$\hat{\pi}_g^{(0)} = \frac{\sum_{i=1}^n \hat{z}_{ig}^0}{n}, \quad \hat{\boldsymbol{\mu}}_g^{(0)} = \frac{\sum_{i=1}^n \hat{z}_{ig}^0 \mathbf{x}_i}{\sum_{i=1}^n \hat{z}_{ig}^0}, \quad \hat{\boldsymbol{\Sigma}}_g^{(0)} = \frac{\sum_{i=1}^n \hat{z}_{ig}^0 (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g^{(0)})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g^{(0)})^\intercal}{\sum_{i=1}^n \hat{z}_{ig}^0}.$$

- Set the skewness parameter $\boldsymbol{\beta}_g^{(0)}$ to be close to zero for symmetric data.
- When applicable, we set $\omega_g^{(0)} = 1$ and $\lambda_g^{(0)} = -0.5$ for the index and concentration parameters and set $\nu_g^{(0)} = 50$ for the near-normality assumption.

To enhance the computational efficiency of the EM algorithm, we update the parameters per missing pattern instead of per individual. We suggest rearranging $\mathbf{X}$ according to unique patterns of the missing data. The procedure can be implemented as follows:

- Build a binary $n$ by $p$ indicator matrix $\mathbf{R} = [r_{ij}]$, with each entry $r_{ij} = 1$ if $\mathbf{X}_{ij}$ is missing and $r_{ij} = 0$ otherwise;

- Find all unique missing patterns; and
- Update parameters per missing pattern instead of per individual.

## 5.4 Numerical Examples

Studies based on both simulated and real datasets are used to compare the clustering performance of the proposed approach. The simulated datasets are each two-component mixtures: a mixture of Gaussian distributions (GMM) with a general VEE covariance structure, a mixture of skew-t distributions (MST) with a diagonal VEI covariance structure, and a mixture of generalized hyperbolic distributions (MGHD) with a general VEE covariance structure. The GMM datasets are generated via the R function `rmvnorm` from the `mvtnorm` package for R, and the MST and MGHD datasets are generated using R code based on the stochastic representations in (2.12).

For each mixture component, $n_g = 200$ two-dimensional vectors $\mathbf{x}_i$ are generated. The presumed parameters of $\boldsymbol{\Sigma}_g$ $(g = 1, 2)$ for the VEE and VEI models are the same as those considered in Celeux and Govaert (1995) and Lin (2014). Each mixture component is centred on a different point giving well-separated and overlapping mixtures. Where applicable, the skewness parameters are $\boldsymbol{\beta}_1 = (1, 1)^\intercal$ and $\boldsymbol{\beta}_2 = (-1, -1)^\intercal$, the degrees of freedoms for the MST is $v_1 = v_2 = 7$, and the values of other parameters for the MGHD are $\omega_1 = \omega_2 = 4$ and $\lambda_1 = \lambda_2 = 6$. For each scenario, we create artificially incomplete datasets by removing data through an MAR mechanism from the simulated samples under missing rates $r$ ranging from 5% to 30% while maintaining the condition that each observation has at least one observed attribute. Then our proposed model for incomplete data is compared to the MGHD and MST for complete

data once missing data have been 'filled-in' with the sample mean of the associated attribute, via the mean imputation method. The misclassification rate ERR and the adjusted Rand index (ARI; Hubert and Arabie, 1985) are used to compare predicted classifications with true classes.

### 5.4.1 Simulation Studies

The datasets considered in the simulation studies are summarized in Table 5.1 and plotted in Figure 5.1. The datasets are overlapping, making this a relatively difficult clustering scenario even when the datasets are complete.

Table 5.1: Summary of simulated datasets

| Dataset | Distribution | Covariance structure | Separation between components |
|---|---|---|---|
| Sim1 | MGHD | VEE | well-separated |
| Sim2 | MGHD | VEE | overlapping |
| Sim3 | MST | VEI | well-separated |
| Sim4 | MST | VEI | overlapping |
| Sim5 | GMM | VEE | well-separated |
| Sim6 | GMM | VEE | overlapping |

First, we undertook a simulation study similar to those of Celeux and Govaert (1995) and Lin (2014) to investigate the classification performance of the MGHD VEE and MST VEI models with synthetic missing values ($r = 5\%, 30\%$). These two models discussed in this experiment are compared for the six simplest cases among a family of fourteen models. Simulations were run with a total of 80 replicates for the first four simulated datasets. The detailed numerical results are summarized in Tables 5.2 and 5.3, including the average misclassification rates with the associated standard deviations in parentheses. The following phenomena are observed, which are similar to results obtained by Lin (2014):

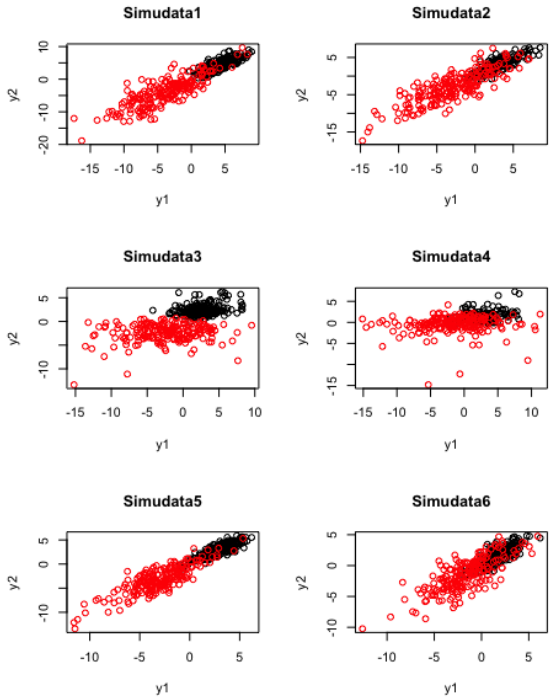


Figure 5.1: Exemplar scatter plots for simulated datasets.

- The average misclassification rate increases as the missing rate rises.
- The overlapping components typically have a higher misclassification rate than well-separated components.
- The bolded numbers indicate that the best results are generally associated with the true covariance structure.
- The standard deviations increase with missing data rate and the degree of component overlap.

As another illustration, we explore the flexibility of the MGHD model for incomplete data and study the performance of the BIC for model selection. As mentioned in the introduction, the GHD is a flexible distribution with skewness, concentration,

Table 5.2: Misclassification rates and associated standard deviations for each model fitted in Sim1, Sim2, Sim3, and Sim4 when $r = 5\%$.

| | EII | VII | EEI | VEI | EEE | VEE |
|---|---|---|---|---|---|---|
| Sim1 | 0.0670 | 0.0538 | 0.0717 | 0.0546 | 0.0557 | **0.0531** |
| | (0.0421) | (0.0125) | (0.0529) | (0.0134) | (0.0210) | (0.0131) |
| Sim2 | 0.1424 | 0.1214 | 0.1261 | 0.1223 | 0.1535 | **0.1210** |
| | (0.0577) | (0.0330) | (0.0308) | (0.0204) | (0.0640) | (0.0214) |
| Sim3 | 0.0763 | 0.0295 | 0.0509 | **0.0186** | 0.0541 | 0.0189 |
| | (0.0215) | (0.0117) | (0.0268) | (0.0062) | (0.0298) | (0.0062) |
| Sim4 | 0.3050 | 0.2019 | 0.3316 | **0.1907** | 0.3306 | 0.2425 |
| | (0.0631) | (0.0458) | (0.0752) | (0.0353) | (0.0582) | (0.1004) |

Table 5.3: Misclassification rates and associated standard deviations for each model fitted to Sim1, Sim2, Sim3, and Sim4 when $r = 30\%$.

| | EII | VII | EEI | VEI | EEE | VEE |
|---|---|---|---|---|---|---|
| Sim1 | 0.0784 | 0.0646 | 0.0853 | 0.0671 | 0.0692 | **0.0601** |
| | (0.0416) | (0.0192) | (0.0498) | (0.0203) | (0.0297) | (0.0305) |
| Sim2 | 0.1666 | 0.1520 | 0.1462 | 0.1427 | 0.1821 | **0.1425** |
| | (0.0607) | (0.0597) | (0.0475) | (0.0418) | (0.0662) | (0.0467) |
| Sim3 | 0.1092 | 0.0799 | 0.0936 | **0.0709** | 0.0936 | 0.0723 |
| | (0.0222) | (0.0257) | (0.0195) | (0.0128) | (0.0257) | (0.0112) |
| Sim4 | 0.3555 | 0.2826 | 0.3442 | **0.2759** | 0.3589 | 0.3019 |
| | (0.0702) | (0.0691) | (0.0804) | (0.0639) | (0.0638) | (0.0779) |

and index parameters. The six simulated datasets in Table 5.1 with missing rates ranging from 5 to 30% were generated under an MAR mechanism with 20 replicates for each dataset. The parsimonious MGHD and MST models introduced here are fitted to the simulated data and compared to mean imputation method with the number of components fixed to $G = 2$ and also with number of components selected from $G = 1, \ldots, 4$. We compute the average misclassification rates, the average ARI, and their associated standard deviations for the best model selected based on the BIC. The detailed results are summarized in Tables 5.4 and 5.5. The lowest misclassification rates and the highest ARI are highlighted. From these tables, we observe

the following:

- The average misclassification rate increases as the missing rate rises. As expected, overlapping components typically have higher misclassification rates than the well-separated components. In addition, the fit of the parsimonious MGHD and MST models to each simulated dataset does not considerably decrease as the missing data rate rises.

- Our proposed parsimonious MGHD and MST models for incomplete data perform significantly better than their counterparts parsimonious MGHD and MST model with mean imputation method (MI/MGHD, MI/MST). In addition, our proposed parsimonious MGHD generally yields much lower misclassification rates than its competitor parsimonious MST for incomplete data when the datasets are generated from generalized hyperbolic distribution, and lower or closer misclassification rates when the datasets are generated from the skew-t or Gaussian distribution.

- Our proposed parsimonious MGHD for incomplete data generally yields similar misclassification rates under circumstances of both known clusters and unknown clusters, while its competitor parsimonious MST generally yields poorer misclassification rates with unknown clusters. This is because the BIC always finds the true number of clusters when using the MGHD for incomplete data, but tends to overestimate the number of clusters when using the MST for incomplete data for datasets with overlapping mixtures.

Table 5.4: A comparison of average misclassification rates and ARI between MGHD, MST, and MI models with standard deviations in parentheses (replications=20) with $G = 2$.

| | | MGHD | | MST | | MI/MGHD | | MI/MST | |
|---|---|---|---|---|---|---|---|---|---|
| | r | ERR | ARI | ERR | ARI | ERR | ARI | ERR | ARI |
| Sim1 | 5% | **0.0539** | **0.7963** | 0.1346 | 0.5362 | 0.111 | 0.6039 | 0.1309 | 0.5482 |
| | | (0.016) | (0.0553) | (0.0293) | (0.0868) | (0.0259) | (0.0807) | (0.0339) | (0.0972) |
| | 10% | **0.0503** | **0.8093** | 0.1344 | 0.3355 | 0.1684 | 0.3355 | 0.233 | 0.3355 |
| | | (0.0135) | (0.0484) | (0.0325) | (0.1934) | (0.0302) | (0.1934) | (0.1167) | (0.1934) |
| | 20% | **0.0641** | **0.761** | 0.1241 | 0.2686 | 0.2821 | 0.2686 | 0.3385 | 0.1516 |
| | | (0.0208) | (0.0701) | (0.0409) | (0.2561) | (0.1452) | (0.2561) | (0.1135) | (0.1546) |
| | 30% | **0.0684** | **0.7478** | 0.113 | 0.602 | 0.1035 | 0.6606 | 0.3338 | 0.2374 |
| | | (0.0289) | (0.0933) | (0.0318) | (0.1001) | (0.0925) | (0.1884) | (0.1836) | (0.2969) |
| Sim2 | 5% | **0.1095** | **0.6133** | 0.1676 | 0.4532 | 0.1998 | 0.3699 | 0.195 | 0.386 |
| | | (0.0335) | (0.0961) | (0.0575) | (0.1382) | (0.053) | (0.1175) | (0.0635) | (0.132) |
| | 10% | **0.1114** | **0.4481** | 0.1694 | 0.4481 | 0.2621 | 0.2422 | 0.2893 | 0.2116 |
| | | (0.0462) | (0.146) | (0.0563) | (0.146) | (0.0678) | (0.1088) | (0.0966) | (0.1632) |
| | 20% | **0.1244** | **0.5662** | 0.1786 | 0.4232 | 0.1853 | 0.4342 | 0.2924 | 0.2084 |
| | | (0.0274) | (0.0786) | (0.0546) | (0.1225) | (0.1014) | (0.1698) | (0.0993) | (0.1561) |
| | 30% | **0.1244** | **0.5667** | 0.172 | 0.436 | 0.1293 | 0.5536 | 0.2616 | 0.3147 |
| | | (0.0297) | (0.0874) | (0.0426) | (0.11) | (0.0356) | (0.0897) | (0.1529) | (0.2266) |
| Sim3 | 5% | **0.0208** | **0.9186** | 0.0454 | 0.9186 | 0.0349 | 0.8651 | 0.0938 | 0.6913 |
| | | (0.0049) | (0.0187) | (0.0288) | (0.0187) | (0.0045) | (0.0167) | (0.0915) | (0.1774) |
| | 10% | **0.0304** | **0.882** | 0.0531 | 0.8014 | 0.0611 | 0.7703 | 0.1384 | 0.5736 |
| | | (0.0054) | (0.0204) | (0.0286) | (0.1006) | (0.0103) | (0.0363) | (0.1163) | (0.2459) |
| | 20% | **0.0497** | **0.8131** | 0.0689 | 0.7373 | 0.1461 | 0.5516 | 0.3017 | 0.2199 |
| | | (0.01) | (0.0365) | (0.0272) | (0.0971) | (0.1122) | (0.1941) | (0.1261) | (0.2232) |
| | 30% | **0.0674** | 0.6472 | 0.1076 | **0.7483** | 0.2808 | 0.292 | 0.4618 | 0.009 |
| | | (0.0091) | (0.1719) | (0.0921) | (0.0315) | (0.1631) | (0.2823) | (0.037) | (0.0169) |
| Sim4 | 5% | **0.191** | **0.3883** | 0.2891 | 0.1997 | 0.2065 | 0.3553 | 0.2968 | 0.3553 |
| | | (0.0454) | (0.1057) | (0.0789) | (0.135) | (0.0566) | (0.1238) | (0.1084) | (0.1238) |
| | 10% | 0.293 | 0.2248 | 0.3025 | 0.1745 | **0.2543** | **0.2756** | 0.3133 | 0.1789 |
| | | (0.1201) | (0.19) | (0.073) | (0.1055) | (0.0965) | (0.1586) | (0.1041) | (0.1505) |
| | 20% | **0.272** | **0.2403** | 0.3004 | 0.1896 | 0.3101 | 0.1953 | 0.317 | 0.1748 |
| | | (0.0942) | (0.1339) | (0.0917) | (0.1353) | (0.1175) | (0.1531) | (0.1056) | (0.1315) |
| | 30% | **0.2748** | **0.2138** | 0.3241 | 0.1447 | 0.415 | 0.0605 | 0.4699 | 0.0114 |
| | | (0.0575) | (0.1005) | (0.0776) | (0.0958) | (0.0939) | (0.112) | (0.051) | (0.0502) |
| Sim5 | 5% | **0.0776** | **0.7146** | 0.1155 | 0.5965 | 0.1448 | 0.5151 | 0.118 | 0.5855 |
| | | (0.0214) | (0.0714) | (0.0399) | (0.1201) | (0.0549) | (0.1374) | (0.0274) | (0.0836) |
| | 10% | **0.0783** | **0.7149** | 0.1214 | 0.5782 | 0.1816 | 0.4129 | 0.1665 | 0.4489 |
| | | (0.0328) | (0.1067) | (0.0388) | (0.1181) | (0.0483) | (0.1173) | (0.0377) | (0.0954) |
| | 20% | **0.0836** | **0.6982** | 0.1124 | 0.6065 | 0.2556 | 0.3462 | 0.2638 | 0.2605 |
| | | (0.0378) | (0.1204) | (0.0411) | (0.1272) | (0.169) | (0.2948) | (0.1011) | (0.1796) |
| | 30% | **0.101** | **0.6447** | 0.0986 | 0.6469 | 0.1441 | 0.5864 | 0.2673 | 0.3536 |
| | | (0.0478) | (0.145) | (0.0298) | (0.0946) | (0.1458) | (0.2624) | (0.1903) | (0.3161) |
| Sim6 | 5% | 0.2235 | 0.3136 | **0.2199** | **0.3312** | 0.2749 | 0.226 | 0.2469 | 0.2761 |
| | | (0.0493) | (0.0996) | (0.0704) | (0.1132) | (0.0806) | (0.1357) | (0.075) | (0.1141) |
| | 10% | 0.2439 | 0.2853 | **0.2384** | **0.3019** | 0.2813 | 0.2219 | 0.2784 | 0.2227 |
| | | (0.08) | (0.1459) | (0.0882) | (0.1451) | (0.0916) | (0.1348) | (0.0854) | (0.1338) |
| | 20% | **0.2518** | **0.2548** | 0.3039 | 0.19 | 0.4419 | 0.0216 | 0.4409 | 0.0182 |
| | | (0.0508) | (0.0966) | (0.0997) | (0.1257) | (0.0517) | (0.0377) | (0.0416) | (0.0347) |
| | 30% | 0.2495 | 0.2709 | **0.241** | **0.2755** | 0.2145 | 0.3292 | 0.2975 | 0.1987 |
| | | (0.0749) | (0.1285) | (0.0477) | (0.0935) | (0.0355) | (0.0778) | (0.0975) | (0.1153) |

Table 5.5: A comparsion of average misclassification rates and ARI between MGHD, MST, and MI models with standard deviations in parentheses (replications=20) with $G = 1, \ldots, 4$.

| | | MGHD | | MST | | MI/MGHD | | MI/MST | |
|---|---|---|---|---|---|---|---|---|---|
| Datasets | r | ERR | ARI | ERR | ARI | ERR | ARI | ERR | ARI |
| Sim1 | 5% | **0.0608** | **0.7744** | 0.0688 | 0.7712 | 0.1206 | 0.5935 | 0.1185 | 0.6069 |
| | | (0.0292) | (0.0925) | (0.0557) | (0.0998) | (0.0302) | (0.0874) | (0.0341) | (0.098) |
| | 10% | **0.0578** | **0.7823** | 0.2769 | 0.4558 | 0.1879 | 0.5029 | 0.2325 | 0.4794 |
| | | (0.0116) | (0.0412) | (0.0895) | (0.2147) | (0.0392) | (0.109) | (0.0882) | (0.139) |
| | 20% | **0.0674** | **0.7523** | 0.2311 | 0.5615 | 0.3108 | 0.2975 | 0.2963 | 0.3703 |
| | | (0.0335) | (0.1082) | (0.0604) | (0.1052) | (0.0552) | (0.1387) | (0.0541) | (0.1209) |
| | 30% | **0.0746** | **0.7267** | 0.2369 | 0.5605 | 0.4265 | 0.2461 | 0.3936 | 0.2825 |
| | | (0.0309) | (0.099) | (0.0576) | (0.075) | (0.1155) | (0.1705) | (0.0824) | (0.1374) |
| Sim2 | 5% | **0.1114** | **0.6092** | 0.3174 | 0.3703 | 0.1769 | 0.4482 | 0.3348 | 0.3444 |
| | | (0.0398) | (0.1061) | (0.0936) | (0.1716) | (0.0534) | (0.1331) | (0.0986) | (0.1437) |
| | 10% | **0.1188** | **0.5873** | 0.4068 | 0.4068 | 0.3018 | 0.3018 | 0.3281 | 0.3281 |
| | | (0.0425) | (0.1126) | (0.1431) | (0.1431) | (0.1263) | (0.1263) | (0.1149) | (0.1149) |
| | 20% | **0.1240** | **0.5722** | 0.3103 | 0.4056 | 0.3153 | 0.3081 | 0.3354 | 0.294 |
| | | (0.0444) | (0.1225) | (0.095) | (0.1076) | (0.0531) | (0.1011) | (0.0648) | (0.1311) |
| | 30% | **0.1319** | **0.5482** | 0.3036 | 0.386 | 0.476 | 0.1319 | 0.4586 | 0.1723 |
| | | (0.0437) | (0.1121) | (0.0577) | (0.069) | (0.0869) | (0.1268) | (0.086) | (0.1109) |
| Sim3 | 5% | **0.0198** | **0.9227** | 0.2155 | 0.6765 | 0.0335 | 0.8765 | 0.2608 | 0.5412 |
| | | (0.0055) | (0.0211) | (0.0774) | (0.1619) | (0.0075) | (0.0284) | (0.108) | (0.2765) |
| | 10% | **0.0556** | **0.8316** | 0.2386 | 0.5548 | 0.0878 | 0.7565 | 0.2969 | 0.3988 |
| | | (0.1049) | (0.1979) | (0.1161) | (0.2822) | (0.0367) | (0.0503) | (0.1256) | (0.2891) |
| | 20% | **0.0744** | **0.7629** | 0.246 | 0.3157 | 0.3251 | 0.3157 | 0.2673 | 0.4629 |
| | | (0.1010) | (0.1855) | (0.0793) | (0.2448) | (0.1154) | (0.2448) | (0.0673) | (0.1664) |
| | 30% | **0.0769** | **0.7162** | 0.2473 | 0.0904 | 0.4741 | 0.0904 | 0.5329 | 0.09 |
| | | (0.0141) | (0.0476) | (0.0938) | (0.1201) | (0.0726) | (0.1201) | (0.1012) | (0.1422) |
| Sim4 | 5% | 0.2419 | 0.3074 | 0.441 | 0.1751 | **0.2066** | **0.355** | 0.3004 | 0.1875 |
| | | (0.1054) | (0.1699) | (0.0632) | (0.1035) | (0.055) | (0.1256) | (0.1066) | (0.0965) |
| | 10% | 0.3004 | 0.2011 | 0.4401 | 0.1519 | **0.2518** | **0.2938** | 0.4048 | 0.1842 |
| | | (0.1066) | (0.1593) | (0.0608) | (0.075) | (0.0826) | (0.1387) | (0.0689) | (0.0911) |
| | 20% | **0.2703** | **0.2375** | 0.4359 | 0.1306 | 0.4323 | 0.0829 | 0.4395 | 0.0782 |
| | | (0.0859) | (0.1266) | (0.0743) | (0.0812) | (0.0589) | (0.0857) | (0.0515) | (0.0964) |
| | 30% | **0.3101** | **0.1691** | 0.4356 | 0.0975 | 0.5004 | 0.0101 | 0.6058 | 0.0043 |
| | | (0.0836) | (0.1088) | (0.0589) | (0.0535) | (0.0532) | (0.0246) | (0.0241) | (0.0169) |
| Sim5 | 5% | **0.0575** | **0.7844** | 0.2533 | 0.5515 | 0.127 | 0.5994 | 0.2596 | 0.5138 |
| | | (0.0214) | (0.0748) | (0.0596) | (0.1117) | (0.0397) | (0.1127) | (0.052) | (0.0749) |
| | 10% | **0.072** | **0.7346** | 0.2545 | 0.5235 | 0.1986 | 0.4692 | 0.2403 | 0.4556 |
| | | (0.0257) | (0.0872) | (0.0879) | (0.1428) | (0.0476) | (0.1087) | (0.0646) | (0.0896) |
| | 20% | **0.0766** | **0.7239** | 0.2459 | 0.5493 | 0.3454 | 0.2477 | 0.2975 | 0.3767 |
| | | (0.0445) | (0.1317) | (0.0894) | (0.1419) | (0.0808) | (0.1929) | (0.065) | (0.1543) |
| | 30% | **0.1064** | **0.6268** | 0.2395 | 0.5281 | 0.3915 | 0.2961 | 0.3983 | 0.2692 |
| | | (0.0455) | (0.1366) | (0.0739) | (0.1067) | (0.0912) | (0.1328) | (0.0705) | (0.1217) |
| Sim6 | 5% | **0.2211** | **0.3198** | 0.436 | 0.1709 | 0.2685 | 0.2415 | 0.4105 | 0.1983 |
| | | (0.0515) | (0.1062) | (0.0846) | (0.1114) | (0.0822) | (0.1381) | (0.0712) | (0.077) |
| | 10% | **0.2573** | **0.2515** | 0.4155 | 0.1921 | 0.3305 | 0.1857 | 0.4068 | 0.176 |
| | | (0.0672) | (0.1093) | (0.0693) | (0.0857) | (0.058) | (0.0935) | (0.1079) | (0.0958) |
| | 20% | **0.2501** | **0.2613** | 0.3865 | 0.1818 | 0.5618 | 0.0216 | 0.5625 | 0.006 |
| | | (0.0588) | (0.1072) | (0.0655) | (0.0869) | (0.0497) | (0.0377) | (0.0433) | (0.0146) |
| | 30% | **0.2597** | **0.2442** | 0.4205 | 0.1773 | 0.4992 | 0.0634 | 0.4923 | 0.1094 |
| | | (0.0626) | (0.1064) | (0.1032) | (0.0745) | (0.0903) | (0.0811) | (0.0607) | (0.0676) |

### 5.4.2 Italian Wine Data

In this first experiment, we apply our proposed parsimonious MGHD and MST models to the well-known Italian wine dataset, which includes thirteen chemical attributes of $n = 178$ Italian wines from Barolo (59), Grignolino (71), and Barbera (48) grape cultivars, which are treated as three intrinsic clusters. This dataset is available in the `gclus` package (Hurley, 2004) for R. This dataset is complete, so for illustration purposes we consider various levels of missing data ranging from 5 to 30% by deleting observations through an MAR mechanism. The dataset is scaled prior to analysis. The number of components is fixed at $G = 3$, then data are analyzed using our proposed parsimonious MGHD and MST models for incomplete data and their counterparts with mean imputation. The results of this analysis (Table 5.6) show that the parsimonious MGHD outperforms the other models for all levels of missing data.

Table 5.6: Misclassification rate and ARI values for our proposed approaches and using mean imputation for clustering on the wine dataset with different levels of missing rates ($r$).

| | MGHD | | MST | | MI/MGHD | | MI/MST | |
|---|---|---|---|---|---|---|---|---|
| $r$ | ERR | ARI | ERR | ARI | ERR | ARI | ERR | ARI |
| 5% | 0.0506 | 0.8465 | 0.0730 | 0.7844 | 0.0562 | 0.8222 | 0.0618 | 0.0618 |
| 10% | 0.1180 | 0.6779 | 0.1517 | 0.6052 | 0.1292 | 0.6455 | 0.1573 | 0.5929 |
| 20% | 0.3539 | 0.4128 | 0.3427 | 0.4645 | 0.3989 | 0.3456 | 0.3764 | 0.3367 |
| 30% | 0.3596 | 0.4280 | 0.3620 | 0.4073 | 0.3820 | 0.3327 | 0.3820 | 0.3327 |

### 5.4.3 Pima Indians Diabetes Data

Data on the diabetes status of 768 patients is obtained from the UCI Machine Learning data repository. The data include information on eight attributes, in which the attribute of number of times pregnant is treated as continuous variable because its

range is from 0 to 14. These data are a popular benchmark dataset for clustering for truly missing values, as 376 of the observations have at least one attribute missing. The data are overlapping and the numerous missing observations make clustering difficult. The detailed description of the attributes and their associated missing rates are summarized in Table 5.7. The dataset features 268 patients with a diabetes diagnosis and 500 without, and these are treated as two clusters. Again, this dataset is scaled prior to the analysis.

Table 5.7: A description of Pima Indian diabetes dataset

| | No. missing values | Sample mean | Sample std. dev. |
|---|---|---|---|
| Number of times pregnant | 0 | 3.85 | 3.37 |
| Plasma glucose concentration | 5 | 120.89 | 31.97 |
| Diastolic blood pressure (mm Hg) | 35 | 69.11 | 19.36 |
| Triceps skin fold thickness (mm) | 227 | 20.54 | 15.95 |
| 2-hour serum insulin(mu U/mL) | 374 | 79.80 | 115.24 |
| Body mass index | 11 | 31.99 | 7.88 |
| Diabetes pedigree function | 0 | 0.47 | 0.33 |
| Age (years) | 0 | 33.24 | 11.76 |

Because there are two known clusters, we fix $G = 2$ and compare the BIC and ICL values for 14 covariance structures of our proposed parsimonious MGHD and MST models. The clustering results are summarized in Table 5.8. Lin (2014) perform a comparable cluster analysis on these via a $t$ mixture model and matches the true cluster labels with 66.7% accuracy. Compared to Lin (2014), our proposed parsimonious MGHD model for incomplete data gives a higher accuracy rate (69.11%).

Table 5.8: Misclassification rate and ARI values for our proposed approaches for clustering on the Pima Indian diabetes dataset.

| | Structure | BIC | ICL | ERR | Accuracy |
|---|---|---|---|---|---|
| MGHD | EEE | -14016.95 | -14053.61 | 0.3089 | 69.11% |
| MST | VVI | -14109.1 | -14186.1 | 0.3763 | 62.37% |

## 5.5 Discussion

Approaches for clustering incomplete data where clusters may be heavy tailed and/or asymmetric is introduced, based on MGHD and MST. There approaches were further extended to parsimonious families of MGHD and MST models via eigen-decomposition of the component scale matrices. The BIC and ICL were used for model selection. It is well known that the BIC can tend to overestimate the number of clusters in practice; however, the results presented herein show that this overestimation can sometimes be mitigated via a more flexible component density such as the MGHD. An EM algorithm was developed to fit the MGHD and MST models to incomplete data, and later implemented in R. It is worth mentioning that our approaches are also applicable in situations with no missing data; and so we have MGHD and MST analogues of the models of Celeux and Govaert (1995). Our MGHD and MST models were applied to real and simulated heterogeneous datasets for clustering in the presence of missing values, and the PMGHD family performed favourably when compared to the PMST family as well as the MGHD and MST approaches with mean imputation.

Going forward, the PMGHD and PMST approaches for clustering with missing values can easily be extended to model-based classification, discriminant analysis, and density estimation. Furthermore, Bayesian analysis via a Gibbs sampler is another popular approach to handle missing data in multivariate datasets (e.g., Lin et al., 2009), so a fully Bayesian treatment will be considered as an alternative to the EM algorithm for parameter estimation.

# Chapter 6

# Flexible High-Dimensional Unsupervised Learning with Missing Data

## 6.1 Introduction

Recently, more attention has been paid to the analysis of heterogeneous high-dimensional data involving different patterns of missing values. However, many model-based clustering techniques, such as the commonly used mixtures factor analyzers (MFA) and mixtures of t-factor analyzers (MtFA) approaches, require complete data for statistical analysis. Naturally, this has led to the development of models for clustering high-dimensional data with missing values, such as the mixture of common factor analyzers (MCFA) model with missing values (Wang, 2013) and the mixture of common-t factor analyzers (MCtFA) with missing values (Wang, 2015).

In this chapter, we aim to develop a unified approach to the mixtures of generalized hyperbolic factor analyzers (MGHFA) for handling high-dimensional data in the presence of missing values as well as heavy-tailed and/or asymmetric clusters (Section 6.2). Maximum likelihood estimates for the MGHFA model with missing values are worked out via a variant of the expectation-maximization algorithm (EM; Dempster et al., 1977) (Section 6.3). To ease the computational burden, two auxiliary permutation matrices are introduced as in Lin et al. (2006). As a by-product, the proposed procedure provides a conditional predictor to impute the missing values and a classifier to cluster partially observed vectors. In Section 6.4, the methodology is illustrated through simulated data with varying proportions of artificially missing values and a real ozone dataset with truly missing values. Finally, some concluding remarks are given in Section 6.5.

## 6.2 Methodology

### 6.2.1 The MFA and MGHFA Models

Out of consideration for completeness, we briefly outline the MFA and MGHFA models herein. The main idea behind MFA is to reduce the number of parameters in the specification of the component-covariance matrices. Given $n$ independent $p$-dimensional continuous variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$, which come independently from a heterogeneous population with $G$ subgroups, the MFA can be written as

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \boldsymbol{\epsilon}_{ig} \tag{6.1}$$

with probability $\pi_g$, for $i = 1, \ldots, n$ and $g = 1, \ldots, G$, where $\boldsymbol{\mu}_g$ is a $p \times 1$ vector of component central location, $\boldsymbol{\Lambda}_g$ is a $p \times q$ matrix of factor loadings, $\mathbf{U}_{ig} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ is a $q \times 1$ vector of latent factors, and $\boldsymbol{\epsilon}_{ig} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$ is a $p \times 1$ vector of errors with $\boldsymbol{\Psi}_g = \text{diag}(\psi_{g1}, \ldots, \psi_{gp})$. Note that the $\mathbf{U}_{ig}$ are independently distributed and are independent of the $\boldsymbol{\epsilon}_{ig}$, which are also independently distributed. Under this model, the marginal distribution of $\mathbf{X}_i$ from the $g$th component is $\mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Psi}_g)$.

Tortora et al. (2016) consider an MGHFA model, where

$$\mathbf{X}_i = \boldsymbol{\mu}_g + W_{ig}\boldsymbol{\beta}_g + \sqrt{W_{ig}}(\boldsymbol{\Lambda}_g\mathbf{U}_{ig} + \boldsymbol{\epsilon}_{ig}) \tag{6.2}$$

with probability $\pi_g$, where $W_{ig} \sim \mathcal{I}(\lambda_g, \eta = 1, \omega_g)$, $\mathbf{U}_{ig} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$, and $\boldsymbol{\epsilon}_{ig} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$. Note that $\mathbf{U}_{ig}$ and $\boldsymbol{\epsilon}_{ig}$ satisfy the same independence relationships as for the MFA model. It follows that $\mathbf{X}_i \mid w_{ig} \sim \mathcal{N}(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Psi}_g))$. Then, they arrive at the MGHFA model with density

$$g(\mathbf{x} \mid \pi_1, \ldots, \pi_g, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_g) = \sum_{g=1}^{G} \pi_g f_{\text{GHD}}(\mathbf{x} \mid \lambda_g, \omega_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g), \tag{6.3}$$

where $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Psi}_g$.

Typically, denote which component each $\mathbf{X}_i$ belongs to, it is convenient to introduce $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$, where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iG})$ with $z_{ig} = 1$ if $\mathbf{x}_i$ belongs to the $g$th component and $Z_{ig} = 0$ otherwise. It follows that $\mathbf{Z}_i$ follows a multinomial distribution with one trial and cell probabilities $\pi_1, \ldots, \pi_G$, denoted by $\mathbf{Z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_G)$. According to (6.2), a four-level hierarchical representation of MGHFA models can be

formulated as

$$\mathbf{X}_i \mid (w_{ig}, \mathbf{u}_{ig}, z_{ig} = 1) \sim \mathcal{N}(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g + \boldsymbol{\Lambda}_g \mathbf{u}_{ig}, w_{ig}\boldsymbol{\Psi}_g), \tag{6.4}$$

$$\mathbf{U}_{ig} \mid (w_{ig}, z_{ig} = 1) \sim \mathcal{N}(\mathbf{0}, w_{ig}\mathbf{I}_q), \tag{6.5}$$

$$W_{ig} \mid (z_{ig} = 1) \sim \mathcal{I}(\lambda_g, \eta = 1, \omega_g), \tag{6.6}$$

$$\mathbf{Z}_i \sim \mathcal{M}(1; \pi_g, \ldots, \pi_G). \tag{6.7}$$

### 6.2.2 The MGHFA Model With Missing Information

To set up estimating equations for the MGHFA model allowing for missing information, $\mathbf{X}_i$ is partitioned into the observed component $\mathbf{X}_i^{\mathrm{o}}$ and the missing component $\mathbf{X}_i^{\mathrm{m}}$ with dimensions $p_i^{\mathrm{o}} \times 1$ and $p_i^{\mathrm{m}} \times 1$, respectively, where $p_i^{\mathrm{o}} + p_i^{\mathrm{m}} = p$. To facilitate computation, following Lin et al. (2006), two missingness indicator matrices are also introduced, denoted by $\mathbf{O}_i$ ($p_i^{\mathrm{o}} \times p$) and $\mathbf{M}_i$ ($p_i^{\mathrm{m}} \times p$), which can be extracted from a $p$-dimensional identity matrix $\mathbf{I}_p$ corresponding to the respective row positions of $\mathbf{X}_i^{\mathrm{o}}$ and $\mathbf{X}_i^{\mathrm{m}}$ in $\mathbf{X}_i$ such that $\mathbf{X}_i^{\mathrm{o}} = \mathbf{O}_i\mathbf{X}_i$ and $\mathbf{X}_i^{\mathrm{m}} = \mathbf{M}_i\mathbf{X}_i$. It is not difficult to verify that $\mathbf{X}_i = \mathbf{O}_i'\mathbf{X}_i^{\mathrm{o}} + \mathbf{M}_i'\mathbf{X}_i^{\mathrm{m}}$ and $\mathbf{O}_i'\mathbf{O}_i + \mathbf{M}_i'\mathbf{M}_i = \mathbf{I}_p$. Now, some important consequences are summarized in the following proposition, which is useful for evaluating the required conditional expectation in the E-step of the algorithm described in the next section.

**Proposition 6.2.1.** *Following from the MGHFA model* (6.2) *and the hierarchical representations* (6.4)–(6.7)*, we have:*

*a. The conditional distribution of* $\mathbf{X}_i^o$ *given* $w_{ig}$ *and* $z_{ig} = 1$ *is*

$$\mathbf{X}_i^o \mid w_{ig}, z_{ig} = 1 \sim \mathcal{N}_{p_i^o}(\boldsymbol{\mu}_{ig}^o + w_{ig}\boldsymbol{\beta}_{ig}^o, w_{ig}\boldsymbol{\Sigma}_{ig}^{oo}), \tag{6.8}$$

*where* $\boldsymbol{\mu}_{ig}^{o} = \mathbf{O}_i\boldsymbol{\mu}_g$, $\boldsymbol{\beta}_{ig}^{o} = \mathbf{O}_i\boldsymbol{\beta}_g$, *and* $\boldsymbol{\Sigma}_{ig}^{oo} = \mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^{'}$.

b. *The marginal distribution of the observed component* $\mathbf{X}_i^o$ *is*

$$g(\mathbf{x}_i^o) = \sum_{g=1}^{G} \pi_g f_{p_i^o, GHD}(\mathbf{x} \mid \lambda_g, \omega_g, \boldsymbol{\mu}_{ig}^o, \boldsymbol{\Sigma}_{ig}^{oo}, \boldsymbol{\beta}_{ig}^o), \tag{6.9}$$

*where* $p_i^o$ *is the dimension corresponding to the observed component* $\mathbf{x}_i^o$.

c. *The conditional distribution of* $\mathbf{X}_i^m$ *given* $\mathbf{x}_i^o$, $w_{ig}$, *and* $z_{ig} = 1$ *is*

$$\mathbf{X}_i^m \mid \mathbf{x}_i^o, w_{ig}, z_{ig} = 1 \sim \mathcal{N}_{p_i^o}(\boldsymbol{\zeta}_{ig}^{m\cdot o}, w_{ig}\boldsymbol{\Sigma}_{ig}^{m\cdot o}), \tag{6.10}$$

*where*

$$\begin{aligned}
\boldsymbol{\zeta}_{ig}^{m\cdot o} &= \mathbf{M}_i\left(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g + \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{oo}(\mathbf{x}_i - \boldsymbol{\mu}_g - w_{ig}\boldsymbol{\beta}_g)\right), \\
\boldsymbol{\Sigma}_{ig}^{m\cdot o} &= \mathbf{M}_i(\mathbf{I}_p - \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{oo})\boldsymbol{\Sigma}_g\mathbf{M}_i^{'}, \\
\mathbf{S}_{ig}^{oo} &= \mathbf{O}_i^{'}(\mathbf{O}_i\boldsymbol{\Sigma}_g\mathbf{O}_i^{'})^{-1}\mathbf{O}_i.
\end{aligned}$$

d. *The conditional distribution of* $W_{ig}$ *given* $\mathbf{x}_i^o$ *and* $z_{ig} = 1$ *is*

$$W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1 \sim GIG(\lambda_{ig}^{\star}, \chi_{ig}^{\star}, \psi_{ig}^{\star}), \tag{6.11}$$

*where* $\psi_{ig}^{\star} = \omega_g + \boldsymbol{\beta}_g\mathbf{S}_{ig}^{oo}\boldsymbol{\beta}_g^{'}$, $\chi_{ig}^{\star} = \omega_g + (\mathbf{x}_i - \boldsymbol{\mu}_g)^{'}\mathbf{S}_{ig}^{oo}(\mathbf{x}_i - \boldsymbol{\mu}_g)$, *and* $\lambda_{ig}^{\star} = \lambda_g - \frac{p_i^o}{2}$.

e. *The conditional distribution of* $\mathbf{X}_i^o$ *given* $w_{ig}$, $\mathbf{u}_{ig}$, *and* $z_{ig} = 1$ *is*

$$\mathbf{X}_i^o \mid w_{ig}, \mathbf{u}_{ig}, z_{ig} = 1 \sim \mathcal{N}_{p_i^o}(\boldsymbol{\zeta}_{ig}^o, w_{ig}\boldsymbol{\Psi}_{ig}^{oo}), \tag{6.12}$$

*where* $\boldsymbol{\zeta}_{ig}^{o} = \boldsymbol{\mu}_{ig}^{o} + w_{ig}\boldsymbol{\beta}_{ig}^{o} + \mathbf{O}_i\boldsymbol{\Lambda}_g\mathbf{u}_{ig}$ *and* $\boldsymbol{\Psi}_{ig}^{oo} = \mathbf{O}_i\boldsymbol{\Psi}_g\mathbf{O}_i^{'}$.

*f. The conditional distribution of* $\mathbf{X}_i^m$ *given* $\mathbf{x}_i^o$, $w_{ig}$, $\mathbf{u}_{ig}$, *and* $z_{ig} = 1$ *is*

$$\mathbf{X}_i^m \mid \mathbf{x}_i^o, w_{ig}, \mathbf{u}_{ig}, z_{ig} = 1 \sim \mathcal{N}(\boldsymbol{\gamma}_{ig}^{m\cdot o}, w_{ig}\boldsymbol{\Psi}_{ig}^{m\cdot o}), \tag{6.13}$$

*where*

$$\begin{aligned}
\boldsymbol{\gamma}_{ig}^{m\cdot o} &= \mathbf{M}_i(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g + \boldsymbol{\Lambda}_g\mathbf{u}_{ig} + \boldsymbol{\Psi}_g\mathbf{T}_{ig}^{oo}(\mathbf{x}_i - \boldsymbol{\mu}_g - w_{ig}\boldsymbol{\beta}_g - \boldsymbol{\Lambda}_g\mathbf{u}_{ig})),\\
\boldsymbol{\Psi}_{ig}^{m\cdot o} &= \mathbf{M}_i(\mathbf{I}_p - \boldsymbol{\Psi}_g\mathbf{T}_{ig}^{oo})\boldsymbol{\Psi}_g\mathbf{M}_i^{'},\\
\mathbf{T}_{ig}^{oo} &= \mathbf{O}_i^{'}(\mathbf{O}_i\boldsymbol{\Psi}_g\mathbf{O}_i^{'})^{-1}\mathbf{O}_i.
\end{aligned}$$

*g. The conditional distribution of* $\mathbf{U}_{ig}$ *given* $\mathbf{x}_i^o$, $w_{ig}$, *and* $z_{ig} = 1$ *is*

$$\mathbf{U}_{ig} \mid \mathbf{x}_i^o, w_{ig}, z_{ig} = 1 \sim \mathcal{N}\left(\boldsymbol{\alpha}_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g - w_{ig}\boldsymbol{\beta}_g), w_{ig}(\mathbf{I}_q - \boldsymbol{\alpha}_{ig}\boldsymbol{\Lambda}_g)\right), \tag{6.14}$$

*where* $\boldsymbol{\alpha}_{ig} = \boldsymbol{\Lambda}_g^{'}\mathbf{S}_{ig}^{oo}$.

The proof of Proposition 1 is straightforward and hence omitted.

## 6.3 Computational Techniques

### 6.3.1 Learning via the AECM Algorithm

To compute the maximum likelihood estimates for the parameters of MGHFA model with partially observed data, we adopt a modification of the expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993), namely the alternating

ECM (AECM) algorithm (Meng and Van Dyk, 1997). In our MGHFA models with missing information, the complete-data is composed of the observed data $\mathbf{x}_i^{\mathrm{o}}$ as well as the missing data $\mathbf{x}_i^{\mathrm{m}}$, the missing labels $z_{ig}$, the latent $w_{ig}$, and the latent factors $u_{ig}$.

For this application of the AECM algorithm to our MGHFA model with missing information, one iteration consists of two cycles, with one E-step and five CM-steps in the first cycle and one E-step and two CM-steps in the second cycle. In the first cycle of the algorithm, we update the mixing proportions $\pi_g$, the component means $\boldsymbol{\mu}_g$, the skewness $\boldsymbol{\beta}_g$, the concentration parameters $\omega_g$, and the index parameters $\lambda_g$. In the second cycle of the algorithm, we update the factor loadings matrices $\boldsymbol{\Lambda}_g$ and the error covariance matrices $\boldsymbol{\Psi}_g$.

In the first cycle of the AECM algorithm, when estimating $\pi_g$, $\lambda_g$, $\omega_g$, $\boldsymbol{\mu}_g$, and $\boldsymbol{\beta}_g$, the complete-data consist of the observed $\mathbf{x}_i^{\mathrm{o}}$, the missing $\mathbf{x}_i^{\mathrm{m}}$, the labels $z_{ig}$, and the latent $w_{ig}$. Hence, the complete-data log-likelihood is

$$\log L_1 = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\log \pi_g + \log \phi\left(\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}_i^{\mathrm{m}} \mid \boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g\right) + \log h(w_{ig} \mid \omega_g, \lambda_g)\right].$$

In the E-step of the first cycle, in order to compute the expected value of the complete-data log-likelihood $\log L_1$, we need to compute $\mathbb{E}(Z_{ig} \mid \mathbf{x}_i^{\mathrm{o}})$, $\mathbb{E}(W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$, $\mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$, $\mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$, $\mathbb{E}(\mathbf{X}_i \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$, $\mathbb{E}((1/W_{ig})\mathbf{X}_i \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$, and $\mathbb{E}((1/W_{ig})\mathbf{X}_i\mathbf{X}_i' \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$.

As usual, the expected value of the $Z_{ig}$ is given by

$$\mathbb{E}(Z_{ig} \mid \mathbf{x}_i^{\mathrm{o}}) = \frac{\pi_g f_{\mathrm{GHD}}(\mathbf{x}_i^{\mathrm{o}} \mid \lambda_g, \omega_g, \boldsymbol{\mu}_{ig}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ig}^{\mathrm{oo}}, \boldsymbol{\beta}_{ig}^{\mathrm{o}})}{\sum_h^G \pi_h f_{\mathrm{GHD}}(\mathbf{x}_i^{\mathrm{o}} \mid \lambda_h, \omega_h, \boldsymbol{\mu}_{ih}^{\mathrm{o}}, \boldsymbol{\Sigma}_{ih}^{\mathrm{o}}, \boldsymbol{\beta}_{ih}^{\mathrm{o}})} =: \hat{z}_{ig}. \tag{6.15}$$

Let $a_{ig} = \mathbb{E}(W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$, $b_{ig} = \mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$, and $c_{ig} = \mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1)$, which are implicit functions of parameters and can be evaluated directly by applying Proposition 1(d) and (2.8).

Recall that $\mathbf{X}_i = \mathbf{O}_i'\mathbf{X}_i^{\mathrm{o}} + \mathbf{M}_i'\mathbf{X}_i^{\mathrm{m}}$ and $\mathbf{O}_i'\mathbf{O}_i + M_i'\mathbf{M}_i = \mathbf{I}_p$. These simply lead to $\mathbf{O}_i'\mathbf{O}_i(\mathbf{I}_p - \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}}) = \mathbf{0}$. Then, based on Proposition 1(c), the following conditional expectations are obtained:

$$\mathbb{E}(\mathbf{X}_i \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1) = \boldsymbol{\mu}_g + a_{ig}\boldsymbol{\beta}_g + \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}}(\mathbf{x}_i - \boldsymbol{\mu}_g - a_{ig}\boldsymbol{\beta}_g) =: E_{1ig},$$

$$\mathbb{E}((1/W_{ig})\mathbf{X}_i \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1) = b_{ig}\boldsymbol{\mu}_g + \boldsymbol{\beta}_g + \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}}(b_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g) - \boldsymbol{\beta}_g) =: E_{2ig},$$

$$\begin{aligned}\mathbb{E}((1/W_{ig})\mathbf{X}_i\mathbf{X}_i' \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig} = 1; \hat{\boldsymbol{\Theta}}) &= (\mathbf{I}_p - \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}})\boldsymbol{\Sigma}_g + (\mathbf{I}_p - \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}})(b_{ig}\boldsymbol{\mu}_g\mathbf{x}_i' + \boldsymbol{\beta}_g\mathbf{x}_i')\mathbf{S}_{ig}^{\mathrm{oo}}\boldsymbol{\Sigma}_g \\ &+ (\mathbf{I}_p - \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}})(b_{ig}\boldsymbol{\mu}_g\boldsymbol{\mu}_g' + \boldsymbol{\mu}_g\boldsymbol{\beta}_g' + \boldsymbol{\beta}_g\boldsymbol{\mu}_g' + a_{ig}\boldsymbol{\beta}_g\boldsymbol{\beta}_g')(\mathbf{I}_p - \mathbf{S}_{ig}^{\mathrm{oo}}\boldsymbol{\Sigma}_g) + b_{ig}\boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}}\mathbf{x}_i\mathbf{x}_i'\mathbf{S}_{ig}^{\mathrm{oo}}\boldsymbol{\Sigma}_g \\ &+ \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}}(b_{ig}\mathbf{x}_i\boldsymbol{\mu}_g' + \mathbf{x}_i\boldsymbol{\beta}_g')(\mathbf{I}_p - \mathbf{S}_{ig}^{\mathrm{oo}}\boldsymbol{\Sigma}_g) =: E_{3ig}.\end{aligned}$$

After the expected value $Q_1$ of the complete-data log-likelihood ($\log L_1$) is formed, maximizing $Q_1$ with respect to $\pi_g$, $\boldsymbol{\mu}_g$, and $\boldsymbol{\beta}_g$ gives rise to the parameter updates

$$\hat{\pi}_g = \frac{n_g}{n}, \quad \hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^n \hat{z}_{ig}(\bar{a}_g E_{2ig} - E_{1ig})}{\sum_{i=1}^n \hat{z}_{ig}(b_{ig}\bar{a}_g - 1)}, \quad \text{and} \quad \hat{\boldsymbol{\beta}}_g = \frac{\sum_{i=1}^n \hat{z}_{ig}(\bar{b}_g E_{1ig} - E_{2ig})}{\sum_{i=1}^n \hat{z}_{ig}(b_{ig}\bar{a}_g - 1)},$$

respectively, where $n_g = \sum_{i=1}^n \hat{z}_{ig}$, $\bar{a}_g = 1/n_g \sum_{i=1}^n \hat{z}_{ig}a_{ig}$, $\bar{b}_g = 1/n_g \sum_{i=1}^n \hat{z}_{ig}b_{ig}$, and $\bar{c}_g = 1/n_g \sum_{i=1}^n \hat{z}_{ig}c_{ig}$. The estimates of the parameters $\omega_g$ and $\lambda_g$ are given as solutions to maximize the following function:

$$q_g(\lambda_g, \omega_g) = -\log K_{\lambda_g}(\omega_g) + (\lambda_g - 1)\bar{c}_g - \frac{\omega_g}{2}(\bar{a}_g + \bar{b}_g), \tag{6.16}$$

and the associated updates are

$$\hat{\lambda}_g = \bar{c}_g \hat{\lambda}_g^{\text{prev}} \left[ \frac{\partial}{\partial \hat{\lambda}_g^{\text{prev}}} \log K_{\hat{\lambda}_g^{\text{prev}}} \left( \hat{\omega}_g^{\text{prev}} \right) \right]^{-1},$$

$$\hat{\omega}_g = \hat{\omega}_g^{\text{prev}} - \left[ \frac{\partial}{\partial \hat{\omega}_g^{\text{prev}}} q_g \left( \hat{\omega}_g^{\text{prev}}, \hat{\lambda}_g \right) \right] \left[ \frac{\partial^2}{\partial (\hat{\omega}_g^{\text{prev}})^2} q_g \left( \hat{\omega}_g^{\text{prev}}, \hat{\lambda}_g \right) \right]^{-1},$$

where the superscript 'prev' denotes the previous estimate. Note that these are analogous to the updates given by Browne and McNicholas (2015).

In the second cycle of the AECM algorithm, when estimating $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$, the complete-data include the observed data $\mathbf{x}_i^{\text{o}}$, the missing data $\mathbf{x}_i^{\text{m}}$, the group labels $z_{ig}$, the latent $w_{ig}$, and the latent factors $\mathbf{u}_{ig}$. The complete-data log-likelihood can be written

$$\begin{aligned}
\log L_2 &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \big[ \log \pi_g + \log \phi \left( \mathbf{x}_i^{\text{o}}, \mathbf{x}_i^{\text{m}} \mid \boldsymbol{\mu}_g + w_{ig} \boldsymbol{\beta}_g + \boldsymbol{\Lambda}_g \mathbf{u}_{ig}, w_{ig} \boldsymbol{\Psi}_g \right) \\
&\qquad + \log \phi(\mathbf{u}_{ig} \mid \mathbf{0}, w_{ig} \mathbf{I}_q) + \log h(w_{ig} \mid \omega_g, \lambda_g) \big], \\
&= C + \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log |\boldsymbol{\Psi}_g^{-1}| \\
&\quad - \frac{1}{2} \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \Bigg[ \text{tr} \left( \frac{1}{w_{ig}} (\mathbf{x}_i \mathbf{x}_i^{'} - \mathbf{x}_i \boldsymbol{\mu}_g^{'} - \boldsymbol{\mu}_g \mathbf{x}_i^{'} + \boldsymbol{\mu}_g \boldsymbol{\mu}_g^{'}) \boldsymbol{\Psi}_g^{-1} \right) \\
&\quad - 2\text{tr} \left( \boldsymbol{\beta}_g (\mathbf{x}_i - \boldsymbol{\mu}_g)^{'} \boldsymbol{\Psi}_g^{-1} \right) + \text{tr} \left( w_{ig} \boldsymbol{\beta}_g \boldsymbol{\beta}_g^{'} \boldsymbol{\Psi}_g^{-1} \right) - 2\text{tr} \left( \frac{1}{w_{ig}} \boldsymbol{\Psi}_g^{-1} \boldsymbol{\Lambda}_g \mathbf{u}_{ig} \mathbf{x}_i^{'} \right) \\
&\quad + 2\text{tr} \left( \frac{1}{w_{ig}} \boldsymbol{\mu}_g^{'} \boldsymbol{\Psi}_g^{'} \boldsymbol{\Lambda}_g \mathbf{u}_{ig} \right) + 2\text{tr} \left( \boldsymbol{\beta}_g^{'} \boldsymbol{\Psi}_g^{-1} \boldsymbol{\Lambda}_g \mathbf{u}_{ig} \right) + \text{tr} \left( \frac{1}{w_{ig}} \boldsymbol{\Lambda}_g \mathbf{u}_{ig} \mathbf{u}_{ig}^{'} \boldsymbol{\Lambda}_g^{'} \boldsymbol{\Psi}_g^{-1} \right) \Bigg],
\end{aligned}$$

where $C$ is constant with respect to the parameters $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$. In the E-step of the second cycle, in order to compute the expected value of the complete-data log-likelihood $\log L_2$, in addtion to the same conditional expectations from the E-step

of the first cycle, we will also need to compute $\mathbb{E}(\mathbf{U}_{ig} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1)$, $\mathbb{E}((1/W_{ig})\mathbf{U}_i \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1)$, $\mathbb{E}((1/W_{ig})\mathbf{U}_i\mathbf{U}_i^{'} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1)$, and $\mathbb{E}((1/W_{ig})\mathbf{U}_i\mathbf{x}_i^{'} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1)$.

Recall that $\mathbf{X}_i = \mathbf{O}_i^{'}\mathbf{X}_i^{\mathrm{o}} + \mathbf{M}_i^{'}\mathbf{X}_i^{\mathrm{m}}$ and $\mathbf{O}_i^{'}\mathbf{O}_i + M_i^{'}\mathbf{M}_i = \mathbf{I}_p$. These simply give rise to $\mathbf{O}_i^{'}\mathbf{O}_i(\mathbf{I}_p - \boldsymbol{\Sigma}_g\mathbf{S}_{ig}^{\mathrm{oo}}) = \mathbf{0}$ and $\mathbf{O}_i^{'}\mathbf{O}_i(\mathbf{I}_p - \boldsymbol{\Psi}_g\mathbf{T}_{ig}^{\mathrm{oo}}) = \mathbf{0}$. Then, based on Proposition 1f and 1g, we obtain the following conditional expectations:

$$
\begin{aligned}
&\mathbb{E}(\mathbf{U}_i \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1) = \boldsymbol{\alpha}_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g - a_{ig}\boldsymbol{\beta}_g) =: E_{4ig},\\
&\mathbb{E}((1/W_{ig})\mathbf{U}_i \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1) = \boldsymbol{\alpha}_{ig}(b_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g) - \boldsymbol{\beta}_g) =: E_{5ig},\\
&\mathbb{E}((1/W_{ig})\mathbf{U}_i\mathbf{U}_i^{'} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1) = \mathbf{I}_q - \boldsymbol{\alpha}_{ig}\boldsymbol{\Lambda}_g + b_{ig}\boldsymbol{\alpha}_{ig}(\mathbf{x}_i - \boldsymbol{\mu}_g)(\mathbf{x}_i - \boldsymbol{\mu}_g)^{'}\boldsymbol{\alpha}_{ig}^{'}\\
&\qquad\qquad + a_{ig}\boldsymbol{\alpha}_{ig}\boldsymbol{\beta}_g\boldsymbol{\beta}_g^{'}\boldsymbol{\alpha}_{ig}^{'} - \boldsymbol{\alpha}_{ig}\left((\mathbf{x}_i - \boldsymbol{\mu}_g)\boldsymbol{\beta}_g^{'} + \boldsymbol{\beta}_g(\mathbf{x}_i - \boldsymbol{\mu}_g)^{'}\right)\boldsymbol{\alpha}_{ig}^{'} =: E_{6ig},\\
&\mathbb{E}((1/W_{ig})\mathbf{U}_i\mathbf{X}_i^{'} \mid \mathbf{x}_i^{\mathrm{o}}, z_{ig}=1) = E_{5ig}\mathbf{x}_i^{'}\mathbf{T}_{ig}^{\mathrm{oo}}\boldsymbol{\Psi}_g + E_{5ig}\boldsymbol{\mu}_g^{'}(\mathbf{I}_p - \mathbf{T}_{ig}^{\mathrm{oo}}\boldsymbol{\Psi}_g)\\
&\qquad\qquad + E_{4ig}(\mathbf{I}_p - \mathbf{T}_{ig}^{\mathrm{oo}}\boldsymbol{\Psi}_g) + E_{6ig}\boldsymbol{\Lambda}_g^{'}(\mathbf{I}_p - \mathbf{T}_{ig}^{\mathrm{oo}}\boldsymbol{\Psi}_g) =: E_{7ig}.
\end{aligned}
$$

Therefore, it follows that the expected value of the complete-data log-likelihood ($\log L_2$) evaluated with $z_{ig} = \hat{z}_{ig}$, $\boldsymbol{\mu}_g = \hat{\boldsymbol{\mu}}_g$, and $\boldsymbol{\beta}_g = \hat{\boldsymbol{\beta}}_g$ is of the form

$$
\begin{aligned}
Q_2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\log|\boldsymbol{\Psi}_g^{-1}| - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}\Big[&\mathrm{tr}\left((E_{3ig} - E_{2ig}\hat{\boldsymbol{\mu}}_g^{'} - \hat{\boldsymbol{\mu}}_g E_{2ig}^{'} + b_{ig}\hat{\boldsymbol{\mu}}_g\hat{\boldsymbol{\mu}}_g^{'})\boldsymbol{\Psi}_g^{-1}\right)\\
&- 2\mathrm{tr}\left(\hat{\boldsymbol{\beta}}_g(E_{1ig} - \hat{\boldsymbol{\mu}}_g)^{'}\boldsymbol{\Psi}_g^{-1}\right) + \mathrm{tr}\left(a_{ig}\hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\beta}}_g^{'}\boldsymbol{\Psi}_g^{-1}\right) - 2\mathrm{tr}\left(\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g E_{7ig}\right)\\
&+ 2\mathrm{tr}\left(\hat{\boldsymbol{\mu}}_g^{'}\boldsymbol{\Psi}_g^{'}\boldsymbol{\Lambda}_g E_{5ig}\right) + 2\mathrm{tr}\left(\hat{\boldsymbol{\beta}}_g^{'}\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}_g E_{4ig}\right) + \mathrm{tr}\left(\boldsymbol{\Lambda}_g E_{6ig}\boldsymbol{\Lambda}_g^{'}\boldsymbol{\Psi}_g^{-1}\right)\Big],
\end{aligned}
$$

where the constant $C$ is omitted. Differentiating $Q_2$ with respect to $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ and

solving the first derivative equalling to zero give rise to their associated updates:

$$\hat{\boldsymbol{\Lambda}}_g = \Bigg[\sum_{i=1}^{n} \hat{z}_{ig}\left(E'_{7ig} - \hat{\boldsymbol{\mu}}_g E'_{5ig} - \hat{\boldsymbol{\beta}}_g E'_{4ig}\right)\Bigg]\Bigg[\sum_{i=1}^{n} \hat{z}_{ig} E_{6ig}\Bigg]^{-1},$$

$$\hat{\boldsymbol{\Psi}}_g = \frac{1}{n_g}\sum_{i=1}^{n} \hat{z}_{ig}\Big[E_{3ig} - E_{2ig}\hat{\boldsymbol{\mu}}_g^{'} - \hat{\boldsymbol{\mu}}_g E'_{2ig} + b_{ig}\hat{\boldsymbol{\mu}}_g\hat{\boldsymbol{\mu}}_g^{'} - 2\hat{\boldsymbol{\beta}}_g(E_{1ig} - \hat{\boldsymbol{\mu}}_g)^{'} + a_{ig}\hat{\boldsymbol{\beta}}_g\hat{\boldsymbol{\beta}}_g^{'} - 2\hat{\boldsymbol{\Lambda}}_g E_{7ig} + 2\hat{\boldsymbol{\Lambda}}_g E_{5ig}\hat{\boldsymbol{\mu}}_g^{'} + 2\hat{\boldsymbol{\Lambda}}_g E_{4ig}\hat{\boldsymbol{\beta}}_g^{'} + \hat{\boldsymbol{\Lambda}}_g E_{6ig}\hat{\boldsymbol{\Lambda}}_g^{'}\Big].$$

The AECM algorithm iteratively updates the parameters until the Aitken acceleration based criterion is satisfied. Unless otherwise specified, the default value of $\epsilon$ is $10^{-5}$ in later numerical examples (Section 6.4).

### 6.3.2 Imputation of Missing Data

When the convergence is achieved, we obtain the maximum likelihood estimates of the parameters denoted by $\hat{\boldsymbol{\Theta}} = (\hat{\pi}_g, \hat{\lambda}_g, \hat{\omega}_g, \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\beta}}_g, \hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Psi}}_g, g = 1, \ldots, G)$. Therefore, the posterior probability of group membership for each observation at convergence can be estimated by

$$\hat{z}_{ig}^{\star} = \mathrm{P}(z_{ig} = 1 \mid \mathbf{x}_i^{\mathrm{o}}; \hat{\boldsymbol{\Theta}}) = \frac{\hat{\pi}_g f_{\mathrm{GHD}}(\mathbf{x}_i^{\mathrm{o}} \mid \hat{\lambda}_g, \hat{\omega}_g, \hat{\boldsymbol{\mu}}_{ig}^{\mathrm{o}}, \hat{\boldsymbol{\Sigma}}_{ig}^{\mathrm{oo}}, \hat{\boldsymbol{\beta}}_{ig}^{\mathrm{o}})}{\sum_h^G \hat{\pi}_h f_{\mathrm{GHD}}(\mathbf{x}_i^{\mathrm{o}} \mid \hat{\lambda}_h, \hat{\omega}_h, \hat{\boldsymbol{\mu}}_{ih}^{\mathrm{o}}, \hat{\boldsymbol{\Sigma}}_{ih}^{\mathrm{o}}, \hat{\boldsymbol{\beta}}_{ih}^{\mathrm{o}})}. \tag{6.17}$$

The resulting $\hat{z}_{ig}^{\star}$ can be used to cluster observations into groups based on the maximum *a posteriori* (MAP) probabilities. Specifically, $\mathrm{MAP}(\hat{z}_{ig}^{\star}) = 1$ if $\max_g(\hat{z}_{ig}^{\star})$ occurs in component $g$ and $\mathrm{MAP}(\hat{z}_{ig}^{\star}) = 0$ otherwise.

For analyzing incomplete data, it is important to fill in the missing data with plausible values. We implement the imputation of the missing values based on the conditional mean method. That is, by substituting the maximum likelihood estimates

$\hat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\beta}}_g, \hat{\boldsymbol{\Lambda}}_g, \hat{\boldsymbol{\Psi}}_g, g = 1, \ldots, G)$, it leads to a predictor of $\mathbf{x}_i^{\mathrm{m}}$ given by

$$\mathbf{M}_i \sum_{g=1}^{G} \hat{z}_{ig}^{\star}(\hat{\boldsymbol{\mu}}_g + a_{ig}\hat{\boldsymbol{\beta}}_g + \hat{\boldsymbol{\Sigma}}_g \hat{\mathbf{S}}_{ig}^{\mathrm{oo}}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g - a_{ig}\hat{\boldsymbol{\beta}}_g)). \tag{6.18}$$

### 6.3.3 Notes on implementation

Like any EM-type iterative algorithm, the AECM algorithm may suffer from computational problems such as slow convergence or even failure to converge. Often, good initial parameter values may speed up the convergence or lead to the attainment of a global optimum. To try to overcome such computational difficulties, we recommend a simple procedure to automatically obtain a set of suitable initial values for the AECM algorithm, as listed below.

1. Perform mean imputation to fill in the missing values for each attribute separately, i.e., the missing value $\mathbf{x}_{ip}^{\mathrm{m}}$ for the $i$th observation on the $p$th attribute was imputed by the sample mean of the observed values of the corresponding variable.

2. Perform the $k$-means clustering to initialize the zero-one membership label $\hat{z}_{ig}^{(0)}$. Accordingly, the initial values for the model parameters are then

$$\hat{\pi}_g^{(0)} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{(0)}}{n}, \quad \hat{\boldsymbol{\mu}}_g^{(0)} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{(0)} \mathbf{x}_i}{\sum_{i=1}^{n} \hat{z}_{ig}^{(0)}}, \quad \hat{\boldsymbol{\Sigma}}_g^{(0)} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{(0)} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g^{(0)})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g^{(0)})'}{\sum_{i=1}^{n} \hat{z}_{ig}^{(0)}}.$$

3. Generate the initial values for $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ via the eigen-decomposition of $\hat{\boldsymbol{\Sigma}}_g^{(0)}$ as follows:

   (a) The initial values of the $j$th column of $\boldsymbol{\Lambda}_g$ were set as $\gamma_j^{(0)} = \sqrt{d_j}\rho_j$,

where $d_j$ is the $j$th largest eigenvalue of $\hat{\boldsymbol{\Sigma}}_g^{(0)}$ and $\rho_j$ is the $j$th eigenvector corresponding to the $j$th largest eigenvalue of $\hat{\boldsymbol{\Sigma}}_g^{(0)}$ for $j \in \{1, \ldots, q\}$.

(b) The $\boldsymbol{\Psi}_g$ is then initialized as $\hat{\boldsymbol{\Psi}}_g^{(0)} = \text{diag}(\hat{\boldsymbol{\Sigma}}_g^{(0)} - \hat{\boldsymbol{\Lambda}}_g^{(0)} \hat{\boldsymbol{\Lambda}}_g^{(0)\prime})$.

4. Set the skewness parameter $\hat{\boldsymbol{\beta}}_g^{(0)} \approx \mathbf{0}$ for the near asymmetric assumption and set the index parameter $\hat{\lambda}_g^{(0)} = 1$ and the concentration parameter $\hat{\omega}_g^{(0)} = -0.5$ for simplicity.

To select an appropriate MGHFA model with missing information in terms of the number of mixture components $G$ and the number of latent factors $q$, we adopt two widely used model selection criteria: the Bayesian information criterion (BIC; Schwarz, 1978) and the approximated weight of evidence (AWE; Banfield and Raftery, 1993).

## 6.4 Numerical Examples

### 6.4.1 Simulation Studies

To examine the performance of the MGHFA model with missing values as defined above, we compared our proposed procedure to the existing mean imputation approach and the MSTFA model with missing values. The EM algorithm for learning the MGHFA and MSTFA models with missing values has been implemented in R (R Core Team, 2016) as `MGHFAMISS` and `MSTFAMISS`, respectively. A two-step procedure is considered. First, the missing values are imputed according to mean imputation, where the missing values are replaced by their unconditional means. Next, the model

parameters are estimated based on the "completed" data using some existing clustering methods found in R, namely:

- Parsimonious Gaussian mixture models (PGMM; McNicholas and Murphy, 2008): model-based clustering using Gaussian mixtures of factor analyzers. We use the function `pgmmEM` via the R package `pgmm` (McNicholas et al., 2015) to derive the results. For the purpose of comparison, the covariance structure is set to be UUU, i.e., we fit the MFA model.

- MGHFA (Tortora et al., 2016): model-based clustering using mixtures of generalized hyperbolic factor analyzers. The function `MGHFA` via the R package `MixGHD` (Tortora et al., 2015) is used to derive the results.

The samples were generated from a three-component MGHFA model with a bivariate normal factor ($q = 2$) under two different sizes, i.e., $n_g = 100$ and $n_g = 200$, respectively. Specifically, the data $\mathbf{x}_i$ were generated from

$$\mathbf{X}_i = \boldsymbol{\mu}_g + W_{ig}\boldsymbol{\beta}_g + \sqrt{W_{ig}}(\boldsymbol{\Lambda}_g \mathbf{U}_{ig} + \boldsymbol{\epsilon}_{ig}) \tag{6.19}$$

with probability $\pi_g$, where $\mathbf{U}_{ig}$ and $\boldsymbol{\epsilon}_{ig}$ satisfy distributional assumptions as in (6.2) and $g \in \{1, 2, 3\}$. The model parameters are given in Table 6.1. Synthetic missing datasets are simulated by deleting at random from the generated data under missing rates ranging from 5 to 30%. Figure 6.1 depicts a scatterplot of the simulated data and its underlying clustering structure for one of the simulated datasets.

For comparison, group memberships were initialized using the $k$-means clustering unless otherwise specified. The clustering experiments comprise 30 replications per combination of sample size and missingness rate. The performance assessments in

Table 6.1: True model parameters for the simulated data.

| Component 1 | Component 2 | Component 3 |
|---|---|---|
| $\lambda_1 = 5$ | $\lambda_2 = 3$ | $\lambda_3 = 4$ |
| $\omega_1 = 3$ | $\omega_2 = 6$ | $\omega_3 = 6$ |
| $\boldsymbol{\mu}_1 = (3, 3, 3, 3, 3, 3)'$ | $\boldsymbol{\mu}_2 = (0, 0, 0, 0, 0, 0)'$ | $\boldsymbol{\mu}_3 = (-3, -3, -3, -3, -3, -3)'$ |
| $\boldsymbol{\beta}_1 = (1, 1, -1, 1, -1, 1)$ | $\boldsymbol{\beta}_2 = (-1, 1, 1, 1, 1, -1, -1)'$ | $\boldsymbol{\beta}_3 = (1, -1, 1, -1, 1, -1)'$ |
| $\boldsymbol{\Lambda}_1 = \begin{pmatrix} -0.6 & -0.1 \\ 0.1 & -0.5 \\ -0.8 & 0.8 \\ -0.6 & -0.4 \\ 0.1 & -0.4 \\ 0.8 & -0.2 \end{pmatrix}$ | $\boldsymbol{\Lambda}_2 = \begin{pmatrix} -0.5 & -0.9 \\ 0.4 & 1.0 \\ -0.5 & -0.2 \\ -0.4 & 0.4 \\ 0.5 & 0.3 \\ -0.8 & 0.9 \end{pmatrix}$ | $\boldsymbol{\Lambda}_3 = \begin{pmatrix} 0.7 & -0.4 \\ 0.8 & 0.0 \\ -0.2 & 0.9 \\ -0.3 & 0.4 \\ 0.3 & 0.7 \\ -0.8 & 0.1 \end{pmatrix}$ |
| $\boldsymbol{\Psi}_1 = 2\mathbf{I}_6$ | $\boldsymbol{\Psi}_2 = \mathbf{I}_6$ | $\boldsymbol{\Psi}_3 = \mathbf{I}_6$ |

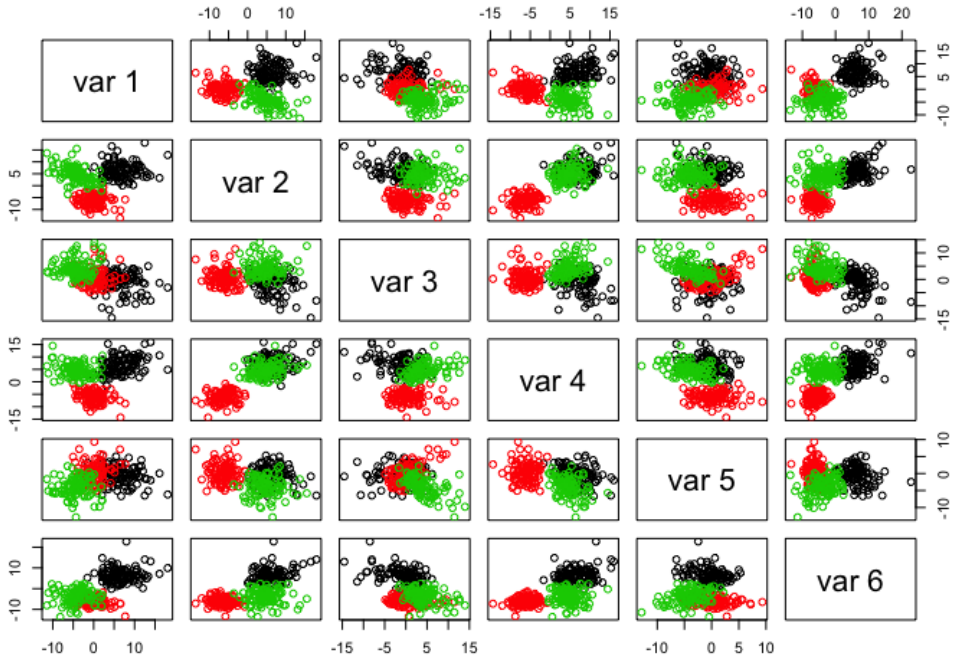


Figure 6.1: Scatterplot of one of the simulated datasets, where colours reflect true class

terms of classification are evaluated through the adjusted Rand index (ARI; Hubert

and Arabie, 1985) and misclassification rate (ERR). In this study, we fit the simulated data using PGMM with mean imputation (MI-PGMM), MGHFA with mean imputation (MI-MGHFA), MSTFAMISS, and MGHFAMISS models with $G = 3$ and $q = 2$.

Tables 6.2 and 6.3 report the mean of the BIC, AWE, ARI, and ERR together with their corresponding standard deviations (Std. Dev.) under each combination considered. Moreover, the frequencies (Freq.) supported by the BIC and AWE are also recorded. Not surprisingly, the results indicate that the best model based on the BIC and AWE is an MGHFAMISS model. At low levels of missingness, all methods perform well but the MGHFAMISS model performs best in terms of the ARI. At high levels of missingness, the MGHFAMISS model leads to much bigger BIC and AWE values as well as much higher ARI and much lower ERR values than those resulting from the MI-PGMM and MI-MGHFA models. Most of the time, the MSTFAMISS model gives slightly inferior results when compared to the best model.

Next, the predictive accuracy of the imputation of missing values is explored. The empirical discrepancy measure for imputed values is simply

$$\text{MSE} = \frac{1}{n^*} \sum_{i=1}^{n} (\mathbf{x}_i^{\text{m}} - \hat{\mathbf{x}}_i^{\text{m}})'(\mathbf{x}_i^{\text{m}} - \hat{\mathbf{x}}_i^{\text{m}}),$$

where $n^* = \sum_{i=n}^{n}(p - p_i^{\text{o}})$ is the number of missing values. Table 6.4 shows the mean MSE together with its standard deviations. The MGHFAMISS and MSTFAMISS models substantially outperform MI for all cases. Furthermore, the MGHFAMISS imputation demonstrates superiority for the reconstruction of missing values in data with the presence of longer tails and asymmetry when compared to the MSTFAMISS imputation.

### 6.4.2 Italian Wine Data

In addition to the simulated data experiments, our MGHFA model with missing data are applied to real data. In this first experiment, we apply our proposed MGHFA model with missing values to the well-known Italian wine data previously analyzed in Chapter 5. First, the wine data are standardized prior to analysis using the default `scale` function in R. Then, we modify the normalized wine data by adding seventeen noisy attributes, which are irrelevant for clustering purposes, to the original attributes. The noise attributes are generated from an independent uniform distribution in the interval $(-1, 1)$. These two datasets (i.e., original wine data and modified wine data) are complete, so for illustration purposes we remove entries through an MAR mechanism to obtain approximately 5, 10, 20, and 30 percent overall missingness.

To compare the BIC and the AWE with respect to choosing the number of latent factors, the MGHFAMISS model with $g = 3$ and $q = 1, \ldots, 7$ are applied for parameter estimation. Simulations were run with a total of thirty replications under each scenario considered. Table 6.5 summarizes the frequencies of each of the candidate models preferred by the BIC and the AWE for the original and modified wine data under various missing rates. Not surprisingly, the AWE tends to select models with a smaller number of factors than does the BIC. Table 6.6 lists the mean ARI and the mean ERR together with their corresponding standard deviations under each scenario considered. As anticipated, as the missingness rates increase the ARI values and the ERR values generally decrease and increase, respectively. Adding noisy variables leads to a slight worsening of the classification assessments.

### 6.4.3 Ozone Level Detection Data

To further demonstrate the proposed methodology, ozone level detection data with truly missing values are analyzed herein. The dataset, available from the UCI Machine Learning Repository (Lichman, 2013), was originally collected by Zhang et al. (2006) for the Houston, Galveston, and Briazoria (HGB) area from several databases within two major federal data warehouses and one local database for air quality control. These are, respectively, the United States Environmental Protection Agency (EPA) Air Quality System (AQS) and National Climate Data Center (NCDC) from the federal government and Continuous Ambient Monitoring Stations (CAMS) operated by the Texas Commission on Environmental Quality (TCEQ). There are two ground ozone level datasets: one is the one hour peak set, the other is the eight hour peak set, and both consist of at least 2500 observations with 72 continuous features containing various measures of air pollutant and meteorological information for the HGB area. As stated by Zhang and Fan (2008), forecasting ozone days is challenging because the dataset (a) is sparse, (b) contains a large number of irrelevant features (only about 10 out of 72 features have been verified by environmental scientists to be useful and relevant), and (c) is skewed and has a lot of missing values. The one hour ozone data feature 73 ozone days versus 2463 normal days and the eight hour ozone data feature 160 ozone days versus 2374 normal days. Both datasets contain 8.2% missing values. The status of whether a day is an ozone day or normal day was recorded for each observation, and is naturally used as the true class variable. These datasets have been previously analyzed by Zhang and Fan (2008) and Wang (2013). Wang (2013) analyzed these datasets using a mixture of common factor analyzers (MCFA) with missing values.

Before performing the fitting, we scale the partially observed dataset using the default `scale` function in R. Following Wang (2013), we fit a two-component MGHFAMISS model with $q = 1, \ldots, 60$. The largest number of latent factors is chosen such that the relationship

$$(p - q)^2 > (p + q)$$

is satisfied (Lawley and Maxwell, 1962).

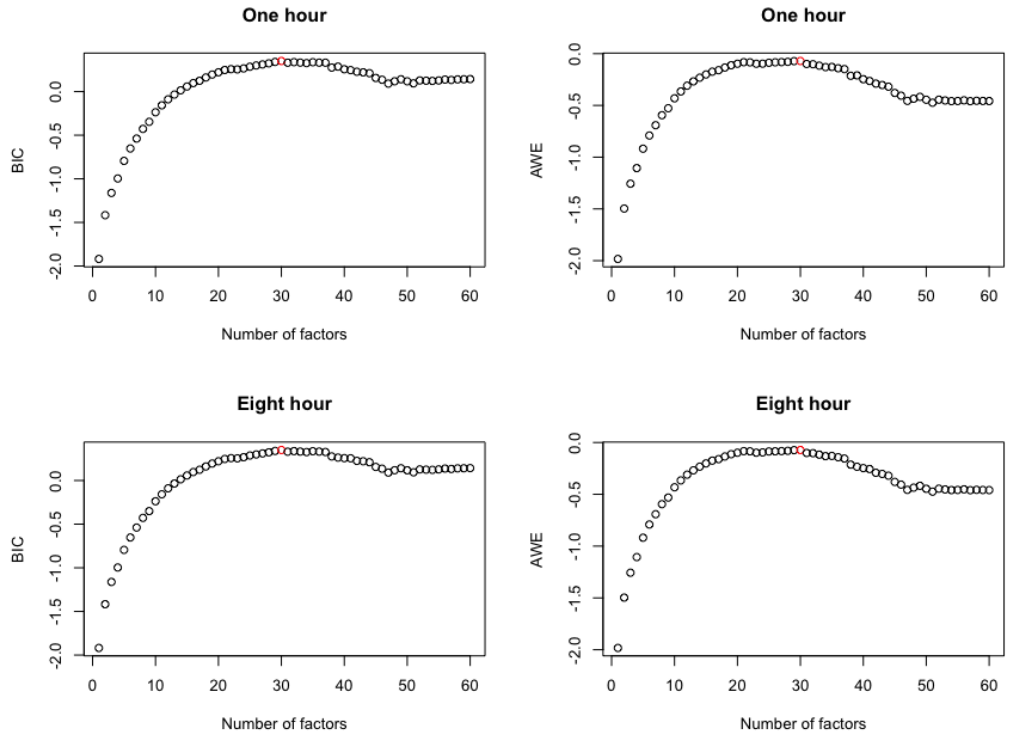


Figure 6.2: Plot of BIC and AWE values versus number of latent factors $q$ for the MGHFAMISS models fitted to the one hour and eight hour ozone data

Considering a plot of the BIC and AWE values versus the number of latent factors for the MGHFAMISS model (Figure 6.2), the BIC and the AWE clearly and coincidentally prefer $q = 30$ for both the one hour and eight hour ozone data. The

best model reported by Wang (2013) had an associated $q = 43$ and $q = 44$ based on the BIC for one hour and eight hour ozone data, respectively, and $q = 34$ based on the AWE for both datasets. Zhang and Fan (2008) stated that there are a larger number of irrelevant features for both datasets, so that our proposed MGHFA model with missing values outperforms the MCFA model with missing values in terms of choosing a smaller number of latent factors.

Furthermore, the correct classification rate, calculated from one minus the misclassification rate, lies in the range from 50.9% to 73.2% and from 51.2% to 74.0% for one hour and eight hour ozone data, respectively. Even though the classification accuracy is not very high, it is superior to the maximum correct classification rate of 72.5% reported by Wang (2013). Notably, they show their result is superior to that of the GMIX imputation (Lin et al., 2006) and the `mclust` (Fraley et al., 2012) methods. Consequently, our best MGHFAMISS model outperforms a variety of popular clustering methods for these two ozone datasets.

## 6.5 Discussion

The MGHFA model has been extended to accommodate complex missing patterns for high-dimensional data with heavy tails and strong asymmetry. By borrowing the attractive features of the GIG distribution, we developed an efficient and elegant parameter estimation for the MGHFA model with missing values within an AECM framework. To simplify matrix manipulations, two auxiliary permutation matrices were incorporated in the procedure. The proposed AECM algorithm can simultaneously take into account the missing values and clustering purpose. The analysis of simulated and real data reveal that the proposed method is quite effective for

the reconstruction of the missing values and outperforms other competing models in terms of clustering purpose when data contain missing information and clusters exhibit non-normal features such as asymmetry and/or heavy tails.

There are computational challenges that must be addressed when fitting the MGHFA model with missing information. Most particularly, the AECM algorithm requires the imputation of missing values on each iteration of the algorithm and, as the number of missing values become large, this task becomes increasingly time consuming. Implementing this approach in parallel would help to ease this computational burden. A families of parsimonious models could be obtained by considering a generalized hyperbolic analogue to the PGMM models of McNicholas and Murphy (2008, 2010). Future work will also include investigation of alternatives to the AECM algorithm for parameter estimation via Bayesian analysis to handling missing values (Utsugi and Kumagai, 2001; Lin et al., 2004, 2009). We will also consider alternatives to the BIC and the AWE for selecting the number of latent factors $q$, such as the LASSO-penalized BIC introduced by Bhattacharya and McNicholas (2014) for mixture model selection.

Table 6.2: Simulation results based on 30 replications ($n_g = 100$)

| Criteria | | MI-PGMM | MI-MGHFA | MSTFAMISS | MGHFAMISS |
|---|---|---|---|---|---|
| r=5% | | | | | |
| BIC | Mean | -10030.2598 | -10051.2803 | -9494.0017 | -9488.5104 |
| | Std. Dev. | 73.0375 | 67.7079 | 67.7139 | 67.2659 |
| | Freq. | 0 | 0 | 0 | 30 |
| AWE | Mean | | -10887.1058 | -10326.9726 | -10320.9896 |
| | Std. Dev. | | 67.8401 | 68.6834 | 68.0657 |
| | Freq. | | 0 | 0 | 30 |
| ARI | Mean | 0.9552 | 0.9580 | 0.9834 | 0.9827 |
| | Std. Dev. | 0.0303 | 0.0297 | 0.0138 | 0.0145 |
| ERR | Mean | 0.0151 | 0.0142 | 0.0056 | 0.0058 |
| | Std. Dev. | 0.0103 | 0.0102 | 0.0047 | 0.0049 |
| r=10% | | | | | |
| BIC | Mean | -10102.5544 | -10118.6320 | -9049.5149 | -9045.1046 |
| | Std. Dev. | 79.4933 | 77.7768 | 66.7636 | 66.5819 |
| | Freq. | 0 | 0 | 1 | 29 |
| AWE | Mean | | -10961.5987 | -9888.2528 | -9883.1796 |
| | Std. Dev. | | 79.5014 | 67.9787 | 67.7636 |
| | Freq. | | 0 | 1 | 29 |
| ARI | Mean | 0.8934 | 0.9439 | 0.9640 | 0.9659 |
| | Std. Dev. | 0.0767 | 0.0245 | 0.0276 | 0.0241 |
| ERR | Mean | 0.0382 | 0.0190 | 0.0122 | 0.0116 |
| | Std. Dev. | 0.0307 | 0.0084 | 0.0096 | 0.0083 |
| r=20% | | | | | |
| BIC | Mean | -10037.2565 | -10171.5710 | -8194.6957 | -8192.0231 |
| | Std. Dev. | 701.7855 | 74.0607 | 60.1604 | 59.9971 |
| | Freq. | 1 | 0 | 5 | 24 |
| AWE | Mean | | -11034.8195 | -9044.9730 | -9041.6883 |
| | Std. Dev. | | 74.4998 | 61.0852 | 60.9230 |
| | Freq. | 0 | 0 | 5 | 25 |
| ARI | Mean | 0.7218 | 0.8424 | 0.9452 | 0.9458 |
| | Std. Dev. | 0.1405 | 0.0813 | 0.0341 | 0.0338 |
| ERR | Mean | 0.1211 | 0.0567 | 0.0187 | 0.0184 |
| | Std. Dev. | 0.0926 | 0.0342 | 0.0123 | 0.0122 |
| r=30% | | | | | |
| BIC | Mean | -8681.0944 | -10043.6238 | -7277.6321 | -7275.3765 |
| | Std. Dev. | 1544.2082 | 94.3739 | 51.4588 | 50.9416 |
| | Freq. | 3 | 0 | 5 | 22 |
| AWE | Mean | | -10928.0620 | -8146.3752 | -8142.4176 |
| | Std. Dev. | | 97.5731 | 50.0552 | 49.9651 |
| | Freq. | | 0 | 3 | 27 |
| ARI | Mean | 0.4935 | 0.6584 | 0.8952 | 0.8970 |
| | Std. Dev. | 0.1482 | 0.1558 | 0.0441 | 0.0372 |
| ERR | Mean | 0.2377 | 0.1451 | 0.0366 | 0.0357 |
| | Std. Dev. | 0.1114 | 0.0993 | 0.0171 | 0.0135 |

Table 6.3: Simulation results based on 30 replications ($n_g = 200$)

| Criteria | | MI-PGMM | MI-MGHFA | MSTFAMISS | MGHFAMISS |
|---|---|---|---|---|---|
| r=5% | | | | | |
| BIC | Mean | -19786.2100 | -19718.5600 | -18592.1700 | -18584.0800 |
| | Std. Dev. | 112.4260 | 107.0459 | 113.7804 | 112.3613 |
| | Freq. | 0 | 0 | 1 | 29 |
| AWE | Mean | | -20627.2800 | -19496.2000 | -19487.3800 |
| | Std. Dev. | | 107.5092 | 114.4635 | 113.1241 |
| | Freq. | | 0 | 0 | 30 |
| ARI | Mean | 0.9646 | 0.9806 | 0.9884 | 0.9882 |
| | Std. Dev. | 0.0134 | 0.0096 | 0.0071 | 0.0070 |
| ERR | Mean | 0.0119 | 0.0065 | 0.0039 | 0.0039 |
| | Std. Dev. | 0.0045 | 0.0032 | 0.0024 | 0.0023 |
| r=10% | | | | | |
| BIC | Mean | -19943.2000 | -19862.7200 | -17717.8100 | -17709.8400 |
| | Std. Dev. | 97.1792 | 96.5181 | 98.4372 | 98.3385 |
| | Freq. | 0 | 0 | 0 | 30 |
| AWE | Mean | | -20789.8800 | -18631.8800 | -18623.1100 |
| | Std. Dev. | | 100.6342 | 100.6766 | 100.5329 |
| | Freq. | | 0 | 0 | 30 |
| ARI | Mean | 0.9308 | 0.9544 | 0.9796 | 0.9796 |
| | Std. Dev. | 0.0301 | 0.0270 | 0.0098 | 0.0101 |
| ERR | Mean | 0.2361 | 0.0154 | 0.0068 | 0.0068 |
| | Std. Dev. | 0.0107 | 0.0095 | 0.0033 | 0.0034 |
| r=20% | | | | | |
| BIC | Mean | -20034.3000 | -19942.5600 | -15987.7100 | -15987.7100 |
| | Std. Dev. | 110.1910 | 104.0624 | 104.0624 | 84.4214 |
| | Freq. | 0 | 0 | 0 | 30 |
| AWE | Mean | | -20911.7400 | -16927.9300 | -16920.6700 |
| | Std. Dev. | | 106.3821 | 86.8494 | 87.4063 |
| | Freq. | | 0 | 0 | 30 |
| ARI | Mean | 0.7950 | 0.8864 | 0.9490 | 0.9494 |
| | Std. Dev. | 0.1104 | 0.0319 | 0.0154 | 0.0142 |
| ERR | Mean | 0.0774 | 0.0391 | 0.0391 | 0.0171 |
| | Std. Dev. | 0.0551 | 0.0115 | 0.0053 | 0.0053 |
| r=30% | | | | | |
| BIC | Mean | -15275.3400 | -19714.8400 | -14214.8300 | -14209.6400 |
| | Std. Dev. | 848.4015 | 399.6284 | 85.8753 | 85.2652 |
| | Freq. | 3 | 0 | 1 | 26 |
| AWE | Mean | | -20736.8500 | -15188.8800 | -15182.1600 |
| | Std. Dev. | | 422.2199 | 89.8860 | 89.1632 |
| | Freq. | | 0 | 1 | 29 |
| ARI | Mean | 0.4288 | 0.7105 | 0.9074 | 0.9082 |
| | Std. Dev. | 0.1493 | 0.1307 | 0.0225 | 0.0236 |
| ERR | Mean | 0.3066 | 0.1178 | 0.0318 | 0.0316 |
| | Std. Dev. | 0.1221 | 0.0870 | 0.0080 | 0.0083 |

Table 6.4: Imputation performance for MI-PGMM, MI-MGHFA, MGHFAMISS, and MSTFAMISS models under various missing rates ($r$).

| | | MSE | | | |
|---|---|---|---|---|---|
| $r$ | | MI-PGMM | MI-MGHFA | MGHFAMISS | MGHFAMISS |
| | $n_g = 100$ | | | | |
| 5% | Mean | 28.9713 | 28.9713 | 9.5301 | 9.4876 |
| | Std. Dev. | 4.4879 | 4.4879 | 2.2601 | 2.318 |
| 10% | Mean | 29.3071 | 29.3071 | 9.7587 | 9.757 |
| | Std. Dev. | 3.7476 | 3.7476 | 1.2783 | 1.2892 |
| 20% | Mean | 28.314 | 28.314 | 10.6201 | 10.5251 |
| | Std. Dev. | 2.5078 | 2.5078 | 1.519 | 1.4792 |
| 30% | Mean | 28.5495 | 28.5495 | 11.5292 | 11.4841 |
| | Std. Dev. | 1.7307 | 1.7307 | 1.1275 | 1.0765 |
| | $n_g = 200$ | | | | |
| 5% | Mean | 28.3152 | 28.3152 | 8.7326 | 8.7032 |
| | Std. Dev. | 2.9408 | 2.9408 | 1.4740 | 1.4329 |
| 10% | Mean | 29.0423 | 29.0423 | 8.9467 | 8.9232 |
| | Std. Dev. | 2.1778 | 2.1778 | 0.8855 | 0.8818 |
| 20% | Mean | 28.0596 | 28.0596 | 9.6844 | 9.6442 |
| | Std. Dev. | 1.6919 | 1.6919 | 0.9224 | 0.9260 |
| 30% | Mean | 28.5089 | 28.5089 | 10.9109 | 10.8826 |
| | Std. Dev. | 1.2845 | 1.2845 | 0.7926 | 0.7965 |

Table 6.5: The frequencies of each of the MGHFAMISS models with $q = 1, \ldots, 7$ preferred by the BIC and AWE for the original and modified wine data under various missingness rates.

| | Original wine data | | | | | | | | Modified wine data | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5% | | 10% | | 20% | | 30% | | 5% | | 10% | | 20% | | 30% | |
| $q$ | BIC | AWE | BIC | AWE | BIC | AWE | BIC | AWE | BIC | AWE | BIC | AWE | BIC | AWE | BIC | AWE |
| 1 | 16 | 30 | 24 | 30 | 29 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 | 30 |
| 2 | 14 | 0 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4–7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 6.6: The ARI and ERR values for each of the MGHFAMISS models with $q = 1, \ldots, 7$ for the original and modified wine data under various missingness rates.

| | 5% | | 10% | | 20% | | 30% | |
|---|---|---|---|---|---|---|---|---|
| $q$ | ARI | ERR | ARI | ERR | ARI | ERR | ARI | ERR |
| | Original wine data | | | | | | | |
| 1 | 0.825 | 0.059 | 0.811 | 0.064 | 0.766 | 0.082 | 0.75 | 0.089 |
| | (0.059) | (0.021) | (0.062) | (0.023) | (0.07) | (0.027) | (0.082) | (0.033) |
| 2 | 0.906 | 0.031 | 0.876 | 0.041 | 0.805 | 0.066 | 0.78 | 0.076 |
| | (0.073) | (0.025) | (0.062) | (0.021) | (0.069) | (0.026) | (0.066) | (0.025) |
| 3 | 0.894 | 0.04 | 0.818 | 0.073 | 0.796 | 0.07 | 0.779 | 0.076 |
| | (0.099) | (0.061) | (0.142) | (0.087) | (0.089) | (0.034) | (0.072) | (0.028) |
| 4 | 0.879 | 0.04 | 0.827 | 0.058 | 0.773 | 0.084 | 0.746 | 0.098 |
| | (0.072) | (0.025) | (0.061) | (0.022) | (0.109) | (0.063) | (0.133) | (0.08) |
| 5 | 0.856 | 0.048 | 0.79 | 0.083 | 0.773 | 0.078 | 0.757 | 0.088 |
| | (0.084) | (0.029) | (0.12) | (0.082) | (0.073) | (0.028) | (0.101) | (0.058) |
| 6 | 0.837 | 0.06 | 0.805 | 0.072 | 0.781 | 0.075 | 0.745 | 0.093 |
| | (0.118) | (0.066) | (0.107) | (0.064) | (0.073) | (0.028) | (0.092) | (0.055) |
| 7 | 0.851 | 0.05 | 0.818 | 0.062 | 0.77 | 0.079 | 0.756 | 0.085 |
| | (0.076) | (0.027) | (0.072) | (0.026) | (0.075) | (0.029) | (0.062) | (0.024) |
| | Modified wine data | | | | | | | |
| 1 | 0.874 | 0.042 | 0.817 | 0.063 | 0.719 | 0.08 | 0.721 | 0.074 |
| | (0.057) | (0.021) | (0.051) | (0.019) | (0.221) | (0.068) | (0.207) | (0.032) |
| 2 | 0.897 | 0.034 | 0.869 | 0.043 | 0.756 | 0.06 | 0.717 | 0.075 |
| | (0.067) | (0.023) | (0.043) | (0.015) | (0.217) | (0.03) | (0.202) | (0.029) |
| 3 | 0.905 | 0.031 | 0.864 | 0.045 | 0.749 | 0.063 | 0.708 | 0.078 |
| | (0.055) | (0.019) | (0.045) | (0.016) | (0.213) | (0.028) | (0.199) | (0.029) |
| 4 | 0.894 | 0.035 | 0.837 | 0.059 | 0.741 | 0.066 | 0.709 | 0.078 |
| | (0.047) | (0.016) | (0.075) | (0.045) | (0.211) | (0.029) | (0.202) | (0.031) |
| 5 | 0.885 | 0.038 | 0.841 | 0.053 | 0.742 | 0.066 | 0.71 | 0.077 |
| | (0.047) | (0.016) | (0.047) | (0.017) | (0.211) | (0.03) | (0.2) | (0.029) |
| 6 | 0.871 | 0.043 | 0.834 | 0.056 | 0.749 | 0.063 | 0.704 | 0.08 |
| | (0.04) | (0.014) | (0.055) | (0.02) | (0.212) | (0.028) | (0.197) | (0.028) |
| 7 | 0.872 | 0.042 | 0.836 | 0.055 | 0.748 | 0.063 | 0.699 | 0.081 |
| | (0.045) | (0.016) | (0.048) | (0.017) | (0.211) | (0.027) | (0.195) | (0.027) |

# Chapter 7

# Conclusions

## 7.1 Summary

Finite mixture models continue to grow in prominence in the literature of modelling, especially used as tools for modelling heterogeneity. The work presented in this thesis has focused on the development and implementation of two topics in finite mixture modelling: extending growth mixture models and handling missing data in model-based clustering, using finite mixture of non-elliptical distributions. This work is a significant contribution to the body of literature on growth mixture models with non-elliptical distributions and model-based clustering with incomplete data.

In Chapters 3 and 4, we put forth two cases for substantial departure from Gaussian growth mixture model paradigm, namely growth mixture models with non-elliptical distributions via the generalized hyperbolic distribution and multivariate skew-t distribution (i.e., GHD-GMM and GST-GMM), for complete and incomplete data. Unlike existing skew-t growth mixture models (Lu and Huang, 2014; Muthén and Asparouhov, 2015), our approach is elegant and computationally straightforward.

Our proposed models perform favourably or equivalently, depending on whether the data are normal or non-normal, when compared to the Gaussian GMM counterpart. In the presence of heterogeneity, heavy tails, and skewness in longitudinal data, the proposed method can fit the data considerably better than Gaussian GMM reducing the risk of extracting latent classes that are merely due to non-normality of the outcomes. When the data are normal, the proposed GHD-GMM can be used to check the reproducibility of a Gaussian GMM solution due to the flexibility of the generalized hyperbolic distribution.

In Chapter 5, we proposed the approaches for clustering incomplete data when clusters may be asymmetric and/or heavy tailed, using mixtures of generalized hyperbolic distributions (MGHD) and mixtures of multivariate skew-t distributions (MST). In addition to considering missing data, to introduce parsimony, we also developed families of MGHD and MST mixture models, each with 14 parsimonious eigen-decomposed scale matrices corresponding to the famous Gaussian parsimonious clustering models (GPCMs) of Banfield and Raftery (1993) and Celeux and Govaert (1995). In Chapter 6, we presented a flexible latent variable approach for clustering high-dimensional data with missingness via mixtures of generalized hyperbolic factor analyzers (MGHFA). In each chapter, analytically feasible EM algorithm or its extensions are formulated for parameter estimation and imputation of missing values for mixture models are also investigated under missing at random mechanisms. The proposed methodologies investigated through simulation study with varying proportions of synthetic missing values and illustrated using real datasets.

## 7.2 Future Work

### 7.2.1 Alternatives to the EM algorithm

All of the models developed in this thesis make use of the EM algorithm or its extensions for parameter estimation. One major drawback to the EM approach is its slow rate of convergence and, in some cases, failure to converge. Common work-arounds to this problem involve multiple random starts or good initial values. Future work will investigate alternative methods such as Bayesian analysis for parameter estimation (e.g., Teschendorff et al., 2005; McGrory and Titterington, 2007; Subedi and McNicholas, 2014).

### 7.2.2 Not Missing At Random (NMAR)

All of the models developed in this thesis to tackle missing data are under the MAR assumption, which are often referred to as ignorable missingness mechanism because the parameters that govern the missingness are separable from the parameters that govern the data. Although the MAR assumption is often reasonable, there are situations where this assumption is not achievable. Hence, it becomes necessary to model the missingness mechanism that may contain information about the parameters of the complete-data population. Therefore, future work focusing on NMAR missing data mechanism would be beneficial.

### 7.2.3 Improvement to the Computational Efficiency

We have demonstrated that all of the models proposed in this thesis are effective in clustering. However, there are computational challenges that must be addressed when

fitting models with missing information. Notably, in Chapters 4, 5, and 6, the EM algorithm and its extensions require the imputation of missing values on each iteration of the algorithm and, as the dimension of the data and the number of missing values become large, this task becomes increasingly time consuming. Implementing this approach in parallel would help to ease this computational burden (cf. McNicholas and Murphy, 2010). There is also work to be done on the alternatives to the R programme language such as C and Python for implementing these models.

# Appendix A

# Details Required for GMMs with Non-Elliptical Distributions

## A.1 Distribution of $\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1$

Herein, we give the detailed derivation of the conditional distribution of $\boldsymbol{\eta}_i$ given $\mathbf{y}_i, \mathbf{x}_i, w_{ik}$, and $c_{ik} = 1$, which facilitate the computation of the conditional expectations in the E-step of the EM algorithm. It also serves as a way to estimate the growth factor scores. The joint distribution of $\boldsymbol{\eta}_i$ and $\mathbf{Y}_i$ given $\mathbf{x}_i, w_{ik}$, and $c_{ik} = 1$ is given by

$$\begin{pmatrix} \boldsymbol{\eta}_i \\ \mathbf{Y}_i \end{pmatrix} \Bigg| \, \mathbf{x}_i, w_{ik}, c_{ik} = 1 \sim \mathcal{N}\Bigg( \begin{pmatrix} \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k} \\ \boldsymbol{\Lambda}_y(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}) \end{pmatrix}, \begin{pmatrix} w_{ik}\boldsymbol{\Psi}_k & w_{ik}\boldsymbol{\Psi}_k\boldsymbol{\Lambda}_y^{'} \\ w_{ik}\boldsymbol{\Lambda}_y\boldsymbol{\Psi}_k & w_{ik}\boldsymbol{\Sigma}_k \end{pmatrix} \Bigg),$$

where $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_y\boldsymbol{\Psi}_k\boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k$.

According to the properties of the conditional distribution for multivariate normal

variables, the conditional distribution of $\boldsymbol{\eta}_i$ conditional on $\mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1$ is also a multivariate normal distribution with

$$\begin{aligned}\mathbb{E}(\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1) &= \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k} \\ &\quad + \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'}(\boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k)^{-1}(\mathbf{y}_i - \boldsymbol{\Lambda}_y(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k})), \\ \mathbb{V}\mathrm{ar}(\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1) &= w_{ik}\boldsymbol{\Psi}_k - w_{ik}\boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'}(\boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k)^{-1}\boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k.\end{aligned}$$

According to the Woodbury matrix identity (Woodbury, 1950), the covariance matrix for the latent variable $\boldsymbol{\eta}_i$ can be simplified to

$$\mathbb{V}\mathrm{ar}(\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1) = w_{ik}(\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)^{-1}.$$

Next, let us simplify the expectation of the latent variable $\boldsymbol{\eta}$:

$$\begin{aligned}&\mathbb{E}(\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1) \\ &= (\mathbf{I}_T - \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'}(\boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k)^{-1}\boldsymbol{\Lambda}_y)(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}) + \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'}(\boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k)^{-1}\mathbf{y}_i, \\ &= (\boldsymbol{\Psi}_k - \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'}(\boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k)^{-1}\boldsymbol{\Lambda}_y \boldsymbol{\Psi}_k)\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}) \\ &\quad + \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'}(\boldsymbol{\Theta}_k^{-1} - \boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y(\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)^{-1}\boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1})\mathbf{y}_i, \\ &= (\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)^{-1}\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}) \\ &\quad + (\boldsymbol{\Psi}_k(\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y) - \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)(\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)^{-1}\boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i, \\ &= (\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)^{-1}\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}) + (\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)^{-1}\boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i, \\ &= (\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)^{-1}(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{\eta k}) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i).\end{aligned}$$

Finally, we obtain the conditional distribution

$$\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + w_{ik}\boldsymbol{\beta}_{yk}) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i), w_{ik}\mathbf{V}_k),$$

where $\mathbf{V}_k = (\boldsymbol{\Psi}_k^{-1} + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y)^{-1}$.

## A.2 The EM algorithm for Model II and IV

The EM algorithm for Model II was employed for parameter estimation in an analogous fashion to the algorithm for Model I described in Section 3.3.1. The complete-data comprise the observed outcomes $\mathbf{y}_i$ and covariates $\mathbf{x}_i$, the class membership labels $c_{ik}$, the latent factors $\boldsymbol{\eta}_i$, and the latent variable $w_{ik}$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. Therefore, the complete-data log-likelihood is

$$\begin{aligned} l_c(\vartheta) = \sum_{i=1}^{n}\sum_{k=1}^{K} c_{ik}[ & \log\pi_{ik} + \log\phi(\mathbf{y}_i \mid \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i + w_{ik}\boldsymbol{\beta}_{yk}, w_{ik}\boldsymbol{\Theta}_k) \\ & + \log\phi(\boldsymbol{\eta}_i \mid \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i, w_{ik}\boldsymbol{\Psi}_k) + \log h(w_{ik} \mid \omega_k, \lambda_k)]. \end{aligned}$$

The E-step requires the computation of the conditional expectations regarding the latent factors $\boldsymbol{\eta}_i$ and the latent variable $W_{ik}$. Under this formulation,

$$\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, w_{ik}, c_{ik} = 1 \sim \mathcal{N}(\mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}(\mathbf{y}_i - w_{ik}\boldsymbol{\beta}_{yk}), w_{ik}\mathbf{V}_k),$$

and the conditional distribution of latent variable $W_{ik}$ given $\mathbf{y}_i, \mathbf{x}_i$, and $c_{ik} = 1$ is given by

$$W_{ik} \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1 \sim \text{GIG}(\psi_k^{\star}, \chi_{ik}, \tilde{\lambda}_k),$$

with $\psi_k^\star = \omega_k + \boldsymbol{\beta}_{yk}^{'}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_{yk}$, $\chi_{ik} = \omega_k + \delta(\mathbf{y}_i, \boldsymbol{\mu}_k \mid \boldsymbol{\Sigma}_k)$, $\tilde{\lambda}_k = \lambda_k - T/2$, where $\boldsymbol{\mu}_k = \boldsymbol{\Lambda}_y(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i)$, and $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_y\boldsymbol{\Psi}_k\boldsymbol{\Lambda}_y^{'} + \boldsymbol{\Theta}_k$.

Therefore, we have convenient forms for the following conditional expectations:

$$E_{1ik}^\star := \mathbb{E}\left[W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \sqrt{\frac{\chi_{ik}}{\psi_k^\star}}\frac{K_{\tilde{\lambda}_k+1}(\sqrt{\psi_k^\star\chi_{ik}})}{K_{\tilde{\lambda}_k}(\sqrt{\psi_k^\star\chi_{ik}})},$$

$$E_{2ik}^\star := \mathbb{E}\left[1/W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \sqrt{\frac{\psi_k^\star}{\chi_{ik}}}\frac{K_{\tilde{\lambda}_k+1}(\sqrt{\psi_k^\star\chi_{ik}})}{K_{\tilde{\lambda}_k}(\sqrt{\psi_k^\star\chi_{ik}})} - \frac{2\tilde{\lambda}_k}{\chi_{ik}},$$

$$E_{3ik}^\star := \mathbb{E}\left[\log W_i \mid \mathbf{x}_i, \mathbf{y}_i, c_{ik} = 1\right] = \log\left(\sqrt{\frac{\chi_{ik}}{\psi_k^\star}}\right) + \frac{1}{K_{\tilde{\lambda}_k}(\sqrt{\psi_k^\star\chi_{ik}})}\frac{\partial}{\partial\tilde{\lambda}_k}K_{\tilde{\lambda}_k}(\sqrt{\psi_k^\star\chi_{ik}}),$$

$$E_{4ik}^\star := \mathbb{E}\left[\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1\right] = \mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}(\mathbf{y}_i - E_{1ik}^\star\boldsymbol{\beta}_{yk})),$$

$$E_{5ik}^\star := \mathbb{E}\left[(1/W_{ik})\boldsymbol{\eta}_i \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1\right] = \mathbf{V}_k(E_{2ik}^\star(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i) - \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\beta}_{yk}),$$

$$\begin{aligned}E_{6ik}^\star := \mathbb{E}[(1/W_{ik})\boldsymbol{\eta}_i\boldsymbol{\eta}_i^{'} \mid \mathbf{y}_i, \mathbf{x}_i, c_{ik} = 1] &= \mathbf{V}_k - \mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i)\boldsymbol{\beta}_{yk}^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y\mathbf{V}_k \\ &+ E_{2ik}^\star\mathbf{V}_k(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i)(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i)^{'}\mathbf{V}_k, \\ &- \mathbf{V}_k\boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\beta}_{yk}^{'}(\boldsymbol{\Psi}_k^{-1}(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k\mathbf{x}_i) + \boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\mathbf{y}_i)^{'}\mathbf{V}_k + E_{1ik}^\star\mathbf{V}_k\boldsymbol{\Lambda}_y^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\beta}_{yk}\boldsymbol{\beta}_{yk}^{'}\boldsymbol{\Theta}_k^{-1}\boldsymbol{\Lambda}_y\mathbf{V}_k.\end{aligned}$$

At each E-step, the values of $E_{1ik}^\star$ to $E_{6ik}^\star$ are updated. We also update the value of the class membership variable $c_{ik}$ using

$$p_{ik}^\star := \frac{\pi_{ik}f_{\text{GHDT}}(\mathbf{y}_i; \tilde{\lambda}_k, \omega_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\beta}_{yk})}{\sum_{l=1}^K \pi_{il}f_{\text{GHDT}}(\mathbf{y}_i; \tilde{\lambda}_l, \omega_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \boldsymbol{\beta}_{yl})}.$$

At each M-step, the following model parameters are obtained by maximizing the conditional expected value of $l_c(\vartheta)$ and are updated sequentially. The updates for $\boldsymbol{\pi}_{ik}$, $\boldsymbol{\alpha}_c$, $\boldsymbol{\Gamma}_c$, $\tilde{\lambda}_k$, and $\omega_k$ are similar to those used in Section 3.3.1. We update the

skewness parameter $\boldsymbol{\beta}_{yk}$ using

$$\hat{\boldsymbol{\beta}}_{yk} = \frac{\sum_{i=1}^{n} p_{ik}^{\star}(\mathbf{y}_i - \boldsymbol{\Lambda}_y E_{4ik}^{\star})}{\sum_{i=1}^{n} p_{ik}^{\star} E_{1ik}^{\star}},$$

and the measurement error $\boldsymbol{\Theta}_k$ using

$$\begin{aligned}\hat{\boldsymbol{\Theta}}_k = \frac{1}{n_k}\sum_{i=1}^{n} p_{ik}^{\star}\Big(& E_{2ik}^{\star}\mathbf{y}_i\mathbf{y}_i^{'} - \mathbf{y}_i E_{5ik}^{\star'}\boldsymbol{\Lambda}_y^{'} - \mathbf{y}_i\hat{\boldsymbol{\beta}}_{yk}^{'} - \boldsymbol{\Lambda}_y E_{5ik}^{\star}\mathbf{y}_i^{'} + \boldsymbol{\Lambda}_y E_{6ik}^{\star}\boldsymbol{\Lambda}_y^{'} \\ & + \boldsymbol{\Lambda}_y E_{4ik}^{\star}\hat{\boldsymbol{\beta}}_{yk}^{'} - \hat{\boldsymbol{\beta}}_{yk}\mathbf{y}_i^{'} + \hat{\boldsymbol{\beta}}_{yk}E_{4ik}^{\star'}\boldsymbol{\Lambda}_y^{'} + E_{1ik}^{\star}\hat{\boldsymbol{\beta}}_{yk}\hat{\boldsymbol{\beta}}_{yk}^{'}\Big),\end{aligned}$$

where $n_k = \sum_{i=1}^{n} p_{ik}^{\star}$. We update $\boldsymbol{\Gamma}_k$, $\boldsymbol{\alpha}_k$, and $\boldsymbol{\Psi}_k$ sequentially using

$$\begin{aligned}\hat{\boldsymbol{\Gamma}}_k &= \left\{\sum_{i=1}^{n} p_{ik}^{\star}(E_{5ik}^{\star} - E_{2ik}^{\star}\hat{\boldsymbol{\alpha}}_k)\mathbf{x}_i^{'}\right\}\left\{\sum_{i=1}^{n} p_{ik}^{\star}E_{2ik}^{\star}\mathbf{x}_i\mathbf{x}_i^{'}\right\}^{-1}, \\ \hat{\boldsymbol{\alpha}}_k &= \frac{\sum_{i=1}^{n} p_{ik}^{\star}(E_{5ik}^{\star} - E_{2ik}^{\star}\hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)}{\sum_{i=1}^{n} p_{ik}^{\star}E_{2ik}^{\star}}, \\ \hat{\boldsymbol{\Psi}}_k &= \frac{1}{n_k}\sum_{i=1}^{n} p_{ik}^{\star}\Big(E_{6ik}^{\star} - E_{5ik}^{\star}(\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)^{'} - (\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)E_{5ik}^{\star'} \\ &\qquad + E_{2ik}^{\star}(\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)(\hat{\boldsymbol{\alpha}}_k + \hat{\boldsymbol{\Gamma}}_k\mathbf{x}_i)^{'}\Big).\end{aligned}$$

The parameter estimation for the model IV is similar to that for model II, hence, is omitted here.

# Appendix B

# Details Pertaining to MGHD and MST with Incomplete Data

## B.1 Some Matrix Computations

We here present some useful matrix computation results that are employed in the derivation of the conditional pdf of a partitioned generalized hyperbolic and multivariate skew-t random vector $\mathbf{Y}$ in Propositions 5.2.3 and 5.2.6.

Consider a partitioned random vector $\mathbf{Y}$ of $p$-dimension that follows the pdf as in Equation (2.13) with

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \tag{B.1}$$

where $\mathbf{Y}_1$ and $\mathbf{Y}_2$ have dimensions $d_1$ and $d_2 = p - d_1$, respectively. The mean, skewness and dispersion matrix are composed of blocks of appropriate dimensions

as partitions of $\mathbf{Y}$. Sometimes, it is more convenient to work with the inverse of dispersion matrix $\boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^{\intercal})^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \\ -(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{pmatrix}. \quad \text{(B.2)}$$

Furthermore, we have for the determinant of $\boldsymbol{\Sigma}$:

$$\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Sigma}_{11})\det(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \quad \text{(B.3)}$$

## B.2 Outline of Proof of Proposition 5.2.3

Here, we derive the conditional density of $\mathbf{Y}_2$ given that $\mathbf{Y}_1 = \mathbf{y}_1$ if $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are jointly generalized hyperbolic distributed, i.e., $\mathbf{Y} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ with the partition in Appendix B.1. Although basic probability theory indicates that the conditional pdf is a ratio of the joint and marginal pdfs, the expression takes a very complicated form. The results from Appendix B.1 are heavily used in the course of the derivations. The conditional density is given by

$$\begin{aligned} f_{\mathbf{Y}_2|\mathbf{Y}_1}(\mathbf{y}_2 \mid \mathbf{y}_1) &= \frac{f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2)}{f_{\mathbf{Y}_1}(\mathbf{y}_1)} \\ &= \frac{\left[\frac{\omega+\delta(\mathbf{y},\boldsymbol{\mu}|\boldsymbol{\Sigma})}{\omega+\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}}\right]^{\frac{\lambda-p/2}{2}} \frac{K_{\lambda-p/2}\left(\sqrt{(\omega+\delta(\mathbf{y},\boldsymbol{\mu}|\boldsymbol{\Sigma}))(\omega+\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})}\right)}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}K_{\lambda}(\omega)\exp\{-(\mathbf{y}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\}}}{\left[\frac{\omega+\delta(\mathbf{y}_1,\boldsymbol{\mu}_1|\boldsymbol{\Sigma}_{11})}{\omega+\boldsymbol{\beta}_1^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1}\right]^{\frac{\lambda-d_1/2}{2}} \frac{K_{\lambda-d_1/2}\left(\sqrt{(\omega+\delta(\mathbf{y}_1,\boldsymbol{\mu}_1|\boldsymbol{\Sigma}_{11}))(\omega+\boldsymbol{\beta}_1^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1)}\right)}{(2\pi)^{d_1/2}|\boldsymbol{\Sigma}_{11}|^{1/2}K_{\lambda}(\omega)\exp\{-(\mathbf{y}_1-\boldsymbol{\mu}_1)^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1\}}}, \end{aligned}$$

where we combine (2.13) and Proposition 5.2.2. For the moment, we focus on the linear form and quadratic form in which $\mathbf{y}$ enters the pdf in (2.13). Inserting the

partition of $\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$ in (B.1) and the inverse of dispersion matrix $\boldsymbol{\Sigma}^{-1}$ (B.2) into the quadratic form yields

$$
\begin{aligned}
\delta(\mathbf{y}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}) &= (\mathbf{y}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \\
&= \begin{pmatrix} (\mathbf{y}_1-\boldsymbol{\mu}_1)^{\intercal} & (\mathbf{y}_2-\boldsymbol{\mu}_2)^{\intercal} \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \mathbf{y}_1-\boldsymbol{\mu}_1 \\ \mathbf{y}_2-\boldsymbol{\mu}_2 \end{pmatrix} \\
&= (\mathbf{y}_1-\boldsymbol{\mu}_1)^{\intercal}(\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^{\intercal})^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1) \\
&\quad -(\mathbf{y}_2-\boldsymbol{\mu}_2)^{\intercal}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1) \\
&\quad -(\mathbf{y}_1-\boldsymbol{\mu}_1)^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2) \\
&\quad +(\mathbf{y}_2-\boldsymbol{\mu}_2)^{\intercal}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2) \\
&= (\mathbf{y}_1-\boldsymbol{\mu}_1)^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1) \\
&\quad +(\mathbf{y}_1-\boldsymbol{\mu}_1)^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1) \\
&\quad -(\mathbf{y}_2-\boldsymbol{\mu}_2)^{\intercal}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1) \\
&\quad -(\mathbf{y}_1-\boldsymbol{\mu}_1)^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2) \\
&\quad +(\mathbf{y}_2-\boldsymbol{\mu}_2)^{\intercal}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2) \\
&= (\mathbf{y}_1-\boldsymbol{\mu}_1)^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1) \\
&\quad +(\mathbf{y}_2-\boldsymbol{\mu}_2-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1))^{\intercal}(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1)) \\
&= \delta(\mathbf{y}_1, \boldsymbol{\mu}_1 \mid \boldsymbol{\Sigma}_{11}) + \delta(\mathbf{y}_2, \boldsymbol{\mu}_{2|1} \mid \boldsymbol{\Sigma}_{2|1}), \qquad \text{(B.4)}
\end{aligned}
$$

where $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1)$ and $\boldsymbol{\Sigma}_{2|1} = (\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}$.

Similarly, inserting into the linear form, following the same algebra as above, yields

$$\begin{aligned}(\mathbf{y}-\boldsymbol{\mu})^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} &= \begin{pmatrix}(\mathbf{y}_1-\boldsymbol{\mu}_1)^\intercal & (\mathbf{y}_2-\boldsymbol{\mu}_2)^\intercal\end{pmatrix}\boldsymbol{\Sigma}^{-1}\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\\ &= (\mathbf{y}_1-\boldsymbol{\mu}_1)^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1\\ &\quad + (\mathbf{y}_2-\boldsymbol{\mu}_2-\boldsymbol{\Sigma}_{12}^\intercal\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1))^\intercal(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\boldsymbol{\beta}_2-\boldsymbol{\Sigma}_{12}^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1)\\ &= (\mathbf{y}_1-\boldsymbol{\mu}_1)^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1+(\mathbf{y}_2-\boldsymbol{\mu}_{2|1})^\intercal\boldsymbol{\Sigma}_{2|1}^{-1}\boldsymbol{\beta}_{2|1}, \end{aligned}\tag{B.5}$$

where $\boldsymbol{\mu}_{2|1}$ and $\boldsymbol{\Sigma}_{2|1}$ are as described above, and $\boldsymbol{\beta}_{2|1}=\boldsymbol{\beta}_2-\boldsymbol{\Sigma}_{12}^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1$.

Furthermore, we investigate the term $\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$, we obtain

$$\begin{aligned}\boldsymbol{\beta}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} &= \begin{pmatrix}\boldsymbol{\beta}_1^\intercal & \boldsymbol{\beta}_2^\intercal\end{pmatrix}\boldsymbol{\Sigma}^{-1}\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\\ &= \boldsymbol{\beta}_1^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1+(\boldsymbol{\beta}_2-\boldsymbol{\Sigma}_{12}^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1)^\intercal(\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\boldsymbol{\beta}_2-\boldsymbol{\Sigma}_{12}^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1)\\ &= \boldsymbol{\beta}_1^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2|1}^\intercal\boldsymbol{\Sigma}_{2|1}\boldsymbol{\beta}_{2|1}. \end{aligned}\tag{B.6}$$

Finally, we substitute (B.3), (B.4), (B.5), and (B.6), and $p=d_1+d_2$ into the conditional density, and after some simple linear algebra, we obtain

$$f_{\mathbf{Y}_2|\mathbf{Y}_1}(\mathbf{y}_2\mid\mathbf{y}_1)=\frac{\left(\frac{\omega+\delta(\mathbf{y}_1,\boldsymbol{\mu}_1|\boldsymbol{\Sigma}_{11})+\delta(\mathbf{y}_2,\boldsymbol{\mu}_{2|1}|\boldsymbol{\Sigma}_{2|1})}{\omega+\boldsymbol{\beta}_1^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2|1}^\intercal\boldsymbol{\Sigma}_{2|1}\boldsymbol{\beta}_{2|1}}\right)^{\frac{\lambda-\frac{d_1}{2}-\frac{d_2}{2}}{2}}\left[\frac{\omega+\boldsymbol{\beta}_1^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1}{\omega+\delta(\mathbf{y}_1,\boldsymbol{\mu}_1|\boldsymbol{\Sigma}_{11})}\right]^{\frac{\lambda-d_1/2}{2}}}{(2\pi)^{\frac{d_2}{2}}|\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{12}^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}|^{\frac{1}{2}}}\times$$
$$\frac{K_{\lambda-\frac{d_1}{2}-\frac{d_2}{2}}\left(\sqrt{(\omega+\delta(\mathbf{y}_1,\boldsymbol{\mu}_1\mid\boldsymbol{\Sigma}_{11})+\delta(\mathbf{y}_2,\boldsymbol{\mu}_{2|1}\mid\boldsymbol{\Sigma}_{2|1}))(\omega+\boldsymbol{\beta}_1^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1+\boldsymbol{\beta}_{2|1}^\intercal\boldsymbol{\Sigma}_{2|1}\boldsymbol{\beta}_{2|1})}\right)}{K_{\lambda-\frac{d_1}{2}}\left(\sqrt{(\omega+\delta(\mathbf{y}_1,\boldsymbol{\mu}_1\mid\boldsymbol{\Sigma}_{11}))(\omega+\boldsymbol{\beta}_1^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1)}\right)\exp(-(\mathbf{y}_2-\boldsymbol{\mu}_{2|1})^\intercal\boldsymbol{\Sigma}_{2|1}^{-1}\boldsymbol{\beta}_{2|1})}.$$

Set $\lambda_{2|1}=\lambda-\frac{d_1}{2}$, $\chi_{2|1}=\omega+\delta(\mathbf{y}_1,\boldsymbol{\mu}_1\mid\boldsymbol{\Sigma}_{11})$, and $\psi_{2|1}=\omega+\boldsymbol{\beta}_1^\intercal\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1$, then we obtain

$$f_{\mathbf{Y}_2|\mathbf{Y}_1}(\mathbf{y}_2 \mid \mathbf{y}_1) = \left[\frac{\chi_{2|1} + \delta(\mathbf{y}_2, \boldsymbol{\mu}_{2|1} \mid \boldsymbol{\Sigma}_{2|1})}{\psi_{2|1} + \boldsymbol{\beta}_{2|1}^{\intercal}\boldsymbol{\Sigma}_{2|1}\boldsymbol{\beta}_{2|1}}\right]^{\frac{\lambda_{2|1}-\frac{d_2}{2}}{2}} \times \frac{\left(\frac{\psi_{2|1}}{\chi_{2|1}}\right)^{\frac{\lambda_{2|1}}{2}} K_{\lambda_{2|1}-\frac{d_2}{2}}\left(\sqrt{(\psi_{2|1} + \boldsymbol{\beta}_{2|1}^{\intercal}\boldsymbol{\Sigma}_{2|1}\boldsymbol{\beta}_{2|1})(\chi_{2|1} + \delta(\mathbf{y}_2, \boldsymbol{\mu}_{2|1} \mid \boldsymbol{\Sigma}_{2|1}))}\right)}{(2\pi)^{\frac{d_2}{2}}|\boldsymbol{\Sigma}_{2|1}|^{\frac{1}{2}} K_{\lambda_{2|1}}(\sqrt{\chi_{2|1}\psi_{2|1}})\exp(-(\mathbf{y}_2 - \boldsymbol{\mu}_{2|1})^{\intercal}\boldsymbol{\Sigma}_{2|1}^{-1}\boldsymbol{\beta}_{2|1})}.$$

Comparison with (2.10) reveals that this is a GHD in the parameterization of McNeil et al. (2005) with

$$\lambda_{2|1} = \lambda - \frac{d_1}{2}, \qquad \chi_{2|1} = \omega + (\mathbf{y}_1 - \boldsymbol{\mu}_1)^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1),$$

$$\psi_{2|1} = \omega + \boldsymbol{\beta}_1^{\intercal}\boldsymbol{\Sigma}_{11}\boldsymbol{\beta}, \qquad \boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1),$$

$$\boldsymbol{\Sigma}_{2|1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}, \qquad \boldsymbol{\beta}_{2|1} = \boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^{\intercal}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\beta}_1.$$

# Bibliography

Aitken, A. C. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh 45*(1), 14–22.

Arminger, G., P. Stein, and J. Wittenberg (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrika 64*(4), 475–494.

Baek, J. and G. J. McLachlan (2011). Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics 27*(9), 1269–1276.

Baek, J., G. J. McLachlan, and L. K. Flack (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*(7), 1298–1309.

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

Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics 5*(3), 151–157.

Barndorff-Nielsen, O. and P. Blæsild (1981). Hyperbolic distributions and ramifications: Contributions to theory and application. In C. Taillie, G. Patil, and

B. Baldessari (Eds.), *Statistical Distributions in Scientific Work*, Volume 79 of *NATO Advanced Study Institutes Series*, pp. 19–44. Springer Netherlands.

Barndorff-Nielsen, O. and C. Halgreen (1977a). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Probability Theory and Related Fields 38*(4), 309–311.

Barndorff-Nielsen, O. and C. Halgreen (1977b). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Probability Theory and Related Fields 38*(4), 309–311.

Barndorff-Nielsen, O., J. Kent, and M. Sørensen (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review/Revue Internationale de Statistique 50*(2), 145–159.

Bauer, D. J. and P. J. Curran (2003a). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods 8*(3), 338–363.

Bauer, D. J. and P. J. Curran (2003b). Overextraction of latent trajectory classes: Much ado about nothing? Reply to Rindskopf (2003), Muthén (2003), and Cudeck and Henly (2003). *Psychological Methods 8*, 384–393.

Bhattacharya, S. and P. D. McNicholas (2014). A LASSO-penalized BIC for mixture model selection. *Advances in Data Analysis and Classification 8*(1), 45–61.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(7), 719–725.

Blæsild, P. (1978). The shape of the generalized inverse Gaussian and hyperbolic distributions. Research Report 37, Department of Theoretical Statistics, Aarhus University, Denmark.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay (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*(2), 373–388.

Bouveyron, C., S. Girard, and C. Schmid (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis 52*(1), 502–519.

Branco, M. D. and D. K. Dey (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis 79*(1), 99 – 113.

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

Browne, R. P., P. D. McNicholas, and C. J. Findlay (2013). A partial EM algorithm for clustering white breads. *arXiv preprint arXiv:1302.6625*.

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

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

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

Everitt, B. S. and D. J. Hand (1981). *Finite Mixture Distributions*. Monographs on Applied Probability and Statistics, London: Chapman and Hall.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika 44*(4), 409–420.

Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal 41*(8), 578–588.

Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca (2012). mclust Version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report No. 597, Department of Statistics, University of Washington, Seattle,WA.

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

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag.

Ghahramani, Z. and M. I. Jordan (1994). Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems*. Citeseer.

Gneiting, T. (1997). Normal scale mixtures and dual probability densities. *Journal of Statistical Computation and Simulation 59*(4), 375–384.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40*(3-4), 237–264.

Guerra-Peña, K. and D. Steinley (2016). Extracting spurious latent classes in growth mixture modeling with nonnormal errors. *Educational and Psychological Measurement 76*(6), 933–953.

Halgreen, C. (1979). Self-decomposability of the generalized inverse Gaussian and hyperbolic distributions. *Probability Theory and Related Fields 47*(1), 13–17.

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

Hurley, C. B. (2004). Clustering visualizations of multidimensional data. *Journal of Computational and Graphical Statistics 13*(4), 788–806.

Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*. New York: Springer-Verlag.

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

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Lawley, D. and A. Maxwell (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician) 12*(3), 209–229.

Lee, S. and G. J. McLachlan (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing 24*(2), 181–202.

Lichman, M. (2013). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.

Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing 20*(3), 343–356.

Lin, T.-I. (2014). Learning from incomplete data via parameterized t mixture models through eigenvalue decomposition. *Computational Statistics & Data Analysis 71*, 183–195.

Lin, T. I., H. J. Ho, and C. L. Chen (2009). Analysis of multivariate skew normal models with incomplete data. *Journal of Multivariate Analysis 100*(10), 2337–2351.

Lin, T.-I., H. J. Ho, and P. S. Shen (2009). Computationally efficient learning of multivariate t mixture models with missing information. *Computational Statistics 24*(3), 375–392.

Lin, T. I., J. C. Lee, and H. J. Ho (2006). On fast supervised learning for normal mixture models with missing information. *Pattern Recognition 39*(6), 1177–1187.

Lin, T. I., J. C. Lee, and H. F. Ni (2004). Bayesian analysis of mixture modelling using the multivariate t distribution. *Statistics and Computing 14*(2), 119–130.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, Volume 5. California: Institute of Mathematical Statistics: Hayward.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. Hoboken, NJ: J Wiley & Sons.

Liu, C., D. B. Rubin, and Y. N. Wu (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika 85*(4), 755–770.

Lu, X. and Y. Huang (2014). Bayesian analysis of nonlinear mixed-effects mixture models for longitudinal data with heterogeneity and skewness. *Statistics in Medicine 33*(16), 2830–2849.

McGrory, C. A. and D. Titterington (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis 51*(11), 5352–5367.

McLachlan, G. and T. Krishnan (2008). *The EM algorithm and Extensions*. John Wiley & Sons.

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

McLachlan, G. J. and K. E. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker Inc.

McLachlan, G. J., D. Peel, and R. Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis 41*(3), 379–388.

McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Boca Raton: CRC Press, Taylor & Francis Group.

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

McNicholas, P. D., A. ElSherbiny, A. F. McDaid, and T. B. Murphy (2015). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.2.

McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing 18*(3), 285–296.

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics 26*(21), 2705–2712.

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

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika 80*(2), 267–278.

Meng, X.-L. and D. Van Dyk (1997). The EM algorithm–An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(3), 511–567.

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

Muthén, B. (2001a). Latent variable mixture modeling. In G. A. Marcoulides and R. E. Schumacker (Eds.), *New Developments and Techniques in Structural Equation Modeling*, pp. 1–33. Lawrence Erlbaum Associates, New Jersey, London.

Muthén, B. (2001b). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class–latent growth modeling. In L. M. Collins and A. G. Sayer (Eds.), *New Methods for the Analysis of Change*, pp. 291–322. APA, Washington, D.C.

Muthén, B. and T. Asparouhov (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs (Eds.), *Longitudinal Data Analysis*, pp. 143–165. CRC press, London, New York.

Muthén, B. and T. Asparouhov (2015). Growth mixture modeling with non-normal distributions. *Statistics in Medicine 34*(6), 1041–1058.

Muthén, B. and L. K. Muthén (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research 24*(6), 882–891.

Muthén, B. and K. Shedden (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics 55*(2), 463–469.

Muthén, B. O. and L. K. Muthén (1998–2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén and Muthén.

Nagin, D. S. (1999, Jun). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods 4*(2), 139–157.

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

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Sahu, S. K., D. K. Dey, and M. D. Branco (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics 31*(2), 129–150.

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

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods 9*(3), 386–396.

Subedi, S. and P. D. McNicholas (2014). Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions. *Advances in Data Analysis and Classification 8*(2), 167–193.

Teschendorff, A. E., Y. Wang, N. L. Barbosa-Morais, J. D. Brenton, and C. Caldas (2005). A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics 21*(13), 3025–3033.

Tiedeman, D. V. (1955). On the study of types. In S. B. Sells (Ed.), *Symposium on Pattern Analysis*. Randolph Field, Texas: Air University, U.S.A.F. School of Aviation Medicine.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley & Sons.

Tortora, C., R. P. Browne, B. C. Franczak, and P. D. McNicholas (2015). *MixGHD: Model Based Clustering, Classification and Discriminant Analysis Using the Mixture of Generalized Hyperbolic Distributions*. R package version 1.8.

Tortora, C., B. C. Franczak, R. P. Browne, and P. D. McNicholas (2014). Mixtures of multiple scaled generalized hyperbolic distributions. *arXiv preprint arXiv:1403.2332*.

Tortora, C., P. D. McNicholas, and R. P. Browne (2016). A mixture of generalized hyperbolic factor analyzers. *Advances in Data Analysis and Classification 10*(4), 423–440.

Utsugi, A. and T. Kumagai (2001). Bayesian analysis of mixtures of factor analyzers. *Neural Computation 13*(5), 993–1002.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*(433), 217–221.

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

Wang, H. X., Q. B. Zhang, B. Luo, and S. Wei (2004). Robust mixture modelling using multivariate t-distribution with missing information. *Pattern Recognition Letters 25*(6), 701–710.

Wang, W.-L. (2013). Mixtures of common factor analyzers for high-dimensional data with missing information. *Journal of Multivariate Analysis 117*, 120 –133.

Wang, W.-L. (2015). Mixtures of common-t factor analyzers for modeling high-dimensional data with missing values. *Computational Statistics & Data Analysis 83*, 223–235.

Woodbury, M. A. (1950). *Inverting modified matrices*. Statistical Research Group, Memorandum Report 42. Princeton University, Princeton, New Jersey.

Zhang, K. and W. Fan (2008). Forecasting skewed biased stochastic ozone days: Analyses, solutions and beyond. *Knowledge and Information Systems 14*(3), 299–326.

Zhang, K., W. Fan, X. Yuan, I. Davidson, and X. Li (2006). Forecasting skewed biased stochastic ozone days: Analyses and solutions. In *Proceedings of the Sixth International Conference on Data Mining*, pp. 753–764. IEEE.