

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

Yuhong Wei, Yang Tang and Paul D. McNicholas

Dept. of Mathematics & Statistics, McMaster University, Hamilton, Ontario, Canada.

Abstract

Robust clustering from incomplete data is an important topic because, in many practical situations, real data sets are heavy-tailed, asymmetric, and/or have arbitrary patterns of missing observations. Flexible methods and algorithms for model-based clustering are presented via mixture of the generalized hyperbolic distributions and its limiting case, the mixture of multivariate skew-t distributions. An analytically feasible EM algorithm is formulated for parameter estimation and imputation of missing values for mixture models employing missing at random mechanisms. The proposed methodologies are investigated through a simulation study with varying proportions of synthetic missing values and illustrated using a real dataset. Comparisons are made with those obtained from the traditional mixture of generalized hyperbolic distribution counterparts by filling in the missing data using the mean imputation method.

Keywords: Clustering; generalized hyperbolic; missing data; mixture models; skew-t.

1 Introduction

Finite mixture models are powerful and flexible tools for discovering unobserved heterogeneity in multivariate datasets. Assuming no prior knowledge of class labels, the application of finite mixture models 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). The multivariate t -distribution, being a heavy-tailed alternative to the multivariate Gaussian distribution, made (robust) mixture modelling based on mixtures of multivariate t -distributions the most natural extension (e.g., Peel and McLachlan, 2000; Andrews and McNicholas, 2011, 2012; Steane et al., 2012; Lin et al., 2014). In many practical situations, however, real world datasets exhibit clusters that are not just heavy tailed but also asymmetric; furthermore, clusters can also be asymmetric yet not heavy tailed. Over the few past years, much attention has been paid to non-Gaussian approaches to model-based clustering and classification, including work on multivariate skew- t distributions (e.g., Lin, 2010; Vrbik and McNicholas, 2012; Lee and McLachlan, 2014; Murray et al., 2014a,b, 2017b), shifted asymmetric Laplace distributions (Franczak et al., 2014), multivariate power exponential distributions (Dang et al., 2015), multivariate normal inverse Gaussian distributions (Karlis and Santourian, 2009; O’Hagan et al., 2016), generalized hyperbolic distributions (Browne and McNicholas, 2015; Morris and McNicholas, 2016; Tortora et al., 2016), and hidden truncation hyperbolic distributions (Murray et al., 2017a). A comprehensive review of model-based clustering work, up to and including some recent work on non-Gaussian mixtures, is given by McNicholas (2016b).

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. The maximum likelihood and Bayesian approaches are two common imputation paradigms for analyzing data with incomplete observations. Herein, the missing data mechanism is assumed to be missing at random (MAR), as per Rubin (1976) and Little and Rubin (1987), meaning that the probability that a variable is missing for a particular individual depends only on the observed data and not on the value of the missing variable. Note that missing completely at random (MCAR) is a special case of MAR. Under MAR, the missing data mechanisms are ignorable for methods using the maximum likelihood approach.

The maximum likelihood 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 max-

imum likelihood estimation using an expectation-maximization (EM) algorithm (Dempster et al., 1977) 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. Wang and Lin (2015) uses a formulation of the mixture of skew- t distributions for model-based clustering with missing data.

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 model selection. The chosen formulation of the (multivariate) generalized hyperbolic distribution (GHD) is that used by Browne and McNicholas (2015) and has formulations of several well-known distributions as special cases such as the multivariate skew- t , normal inverse Gaussian, variance-gamma, Laplace, and Gaussian distributions (cf. McNeil et al., 2005). 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 of (GPCMs; Banfield and Raftery, 1993; Celeux and Govaert, 1995); see Table 7 (Appendix A).

2 Background

2.1 Generalized Inverse Gaussian Distribution

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

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

where $\psi, \chi \in \mathbb{R}^+$, $\lambda \in \mathbb{R}$, and K_{λ} is the modified Bessel function of the third kind with index λ . Herein, we write $W \sim \text{GIG}(\lambda, \chi, \psi)$ to indicate that a random variable W has the GIG density as parameterized in (1). The GIG distribution has some attractive properties (Barndorff-Nielsen and Halgreen, 1977; Blæsild, 1978; Halgreen, 1979; Jørgensen, 1982),

including the tractability of the expectations:

$$\mathbb{E}[W^\alpha] = \left(\frac{\chi}{\psi}\right)^{\alpha/2} \frac{K_{\lambda+\alpha}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})}, \quad (2)$$

for $\alpha \in \mathbb{R}$, and

$$\mathbb{E}[\log W] = \log\left(\sqrt{\frac{\chi}{\psi}}\right) + \frac{\partial}{\partial \lambda} \log(K_\lambda(\sqrt{\psi\chi})). \quad (3)$$

Specifically, for $\alpha = 1$ and $\alpha = -1$, we have

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

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

$$f_{\mathcal{I}}(w \mid \lambda, \eta, \omega) = \frac{(w/\eta)^{\lambda-1}}{2\eta K_\lambda(\omega)} \exp\left\{-\frac{\omega}{2} \left(\frac{w}{\eta} + \frac{\eta}{w}\right)\right\} \quad (4)$$

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.

2.2 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 (1977) introduces the generalized hyperbolic distribution (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 by 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 param-

eters 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 generalized hyperbolic distribution with index parameter λ , concentration parameters χ and ψ , location vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$, and skewness vector $\boldsymbol{\alpha}$, denoted by $\mathbf{X} \sim \text{GH}_p(\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, if it can be represented by

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{U}, \quad (5)$$

where $\mathbf{U} \perp W$, $W \sim \text{GIG}(\lambda, \chi, \psi)$, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and the symbol \perp indicates independence. 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

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}{\psi + \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}} \right]^{\frac{\lambda - p/2}{2}} \frac{(\psi/\chi)^{\lambda/2} K_{\lambda - p/2} \left(\sqrt{(\chi + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}))(\psi + \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha})} \right)}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} K_\lambda(\sqrt{\chi\psi}) \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}\}}, \quad (6)$$

where $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between \mathbf{x} and $\boldsymbol{\mu}$, K_λ is the modified Bessel function of the third kind with index λ , 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

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu} + \mathbb{E}(W)\boldsymbol{\alpha} \quad \text{and} \quad \text{Var}(\mathbf{X}) = \mathbb{E}(W)\boldsymbol{\Sigma} + \text{Var}(W)\boldsymbol{\alpha}\boldsymbol{\alpha}^\top, \quad (7)$$

respectively, where $\mathbb{E}(W)$ and $\text{Var}(W)$ 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 λ , concentration parameter ω , 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

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\beta} + \sqrt{W}\mathbf{U}, \quad (8)$$

where $\mathbf{U} \perp W$, $W \sim \text{GIG}(\omega/\eta, \omega\eta, \lambda)$, with $\eta = 1$, and $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Under this parameter-

ization, the density of the generalized hyperbolic random vector \mathbf{X} is

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{\omega + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}{\omega + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}} \right]^{\frac{\lambda - p/2}{2}} \frac{K_{\lambda - p/2} \left(\sqrt{(\omega + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}))(\omega + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta})} \right)}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} K_\lambda(\omega) \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}\}}, \quad (9)$$

where $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})$ and $K_{\lambda - p/2}$ are as described earlier. We use this parameterization when we describe parameter estimation (cf. Section 3).

The following result shows an appealing closure property of the generalized hyperbolic distribution under affine transformation and conditioning as well as the formation of marginal distributions, which is useful for developing new methods presented later. Suppose that \mathbf{X} is a p -dimensional random vector having a generalized hyperbolic distribution as in (9), i.e., $\mathbf{X} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Assume that \mathbf{X} is partitioned as $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, where \mathbf{X}_1 takes values in \mathbb{R}^{d_1} and \mathbf{X}_2 in $\mathbb{R}^{d_2} = \mathbb{R}^{p-d_1}$, with

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

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

$$\mathbf{Y} \sim \text{GHD}_k(\lambda, \omega, \mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top, \mathbf{B}\boldsymbol{\beta}), \quad (10)$$

Proof. The result follows by substituting (8) into $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$. \square

Proposition 2. *The marginal distribution of \mathbf{X}_1 is a generalized hyperbolic distribution as in (9) with index parameter λ , concentration parameter ω , location vector $\boldsymbol{\mu}_1$, dispersion matrix $\boldsymbol{\Sigma}_{11}$, and skewness vector $\boldsymbol{\beta}_1$, i.e., $\mathbf{X}_1 \sim \text{GHD}_{d_1}(\lambda, \omega, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\beta}_1)$.*

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 λ, ω inherited from the mixing distribution $W \sim \mathcal{I}(\lambda, \eta = 1, \omega)$ remain the same under the affine transformation and marginal distribution. \square

Proposition 3. *The conditional distribution of \mathbf{X}_2 given $\mathbf{X}_1 = \mathbf{x}_1$ is a generalized hyperbolic*

distribution as in (6), 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

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

The proof of Proposition 3 is given in Appendix B.

2.3 The Multivariate Skew- t Distribution

There are several alternative formulations of multivariate skew- t distributions appearing in the literature (e.g., Branco and Dey, 2001; Sahu, Dey, and Branco, 2003; Murray, Browne, and McNicholas, 2014a; Lee and McLachlan, 2014). Lin and Lin (2011) develop a mixture of multivariate skew- t distributions incomplete data using the formulation of Sahu et al. (2003). Herein, the formulation of the multivariate skew- t distribution arising from the generalized hyperbolic distribution is used. This formulation of the multivariate skew- t distribution has been used by Murray et al. (2014a) to develop a mixture of skew- t factor analyzers model.

Following McNeil et al. (2005), a $p \times 1$ random vector \mathbf{X} is said to follow a multivariate skew- t distribution with degree of freedom parameter v , location vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$, and skewness vector $\boldsymbol{\beta}$, denoted by $\mathbf{X} \sim \text{ST}_p(v, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$, if it can be represented by

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\beta} + \sqrt{W}\mathbf{U}, \quad (11)$$

where $\mathbf{U} \perp W$, $W \sim \text{IG}(v/2, v/2)$, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\text{IG}(\cdot)$ denoting the inverse Gamma distribution. It follows that $\mathbf{X} \mid w \sim \mathcal{N}(\boldsymbol{\mu} + w\boldsymbol{\beta}, w\boldsymbol{\Sigma})$ and the pdf of the multivariate skew- t random vector \mathbf{X} is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \left[\frac{v + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})}{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}} \right]^{-\frac{v-p}{4}} \frac{v^{p/2} K_{(-v-p)/2} \left(\sqrt{(v + \delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma}))(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta})} \right)}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} \Gamma(v/2) 2^{v/2-1} \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}\}}. \quad (12)$$

This formulation of the multivariate skew- t distribution can be obtained as a special case of the generalized hyperbolic distribution by setting $\lambda = -v/2$ and $\chi = v$, and letting $\psi \rightarrow 0$. Similarly, this formulation of the multivariate skew- t distribution has a closed form under affine transformation and conditioning, and the formation of marginal distributions, which is

useful for developing new methods presented later. Suppose that \mathbf{X} is a p -dimensional random vector having the multivariate skew- t distribution as in (12), i.e., $\mathbf{X} \sim ST_p(v, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Assume that \mathbf{X} is partitioned as $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, where \mathbf{X}_1 takes values in \mathbb{R}^{d_1} and \mathbf{X}_2 in $\mathbb{R}^{d_2} = \mathbb{R}^{p-d_1}$, with

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

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

$$\mathbf{Y} \sim ST_k(v, \mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top, \mathbf{B}\boldsymbol{\beta}). \quad (13)$$

Proof. The proof follows easily by substituting (11) into $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$. \square

Proposition 5. *The marginal distribution of \mathbf{X}_1 is a multivariate skew- t distribution as in (12) with degree of freedom parameter v , location vector $\boldsymbol{\mu}_1$, dispersion matrix $\boldsymbol{\Sigma}_{11}$, and skewness vector $\boldsymbol{\beta}_1$, i.e., $\mathbf{X}_1 \sim ST_{d_1}(v, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\beta}_1)$.*

Proof. The proof follows easily by applying Proposition 4 and choosing $\mathbf{B} = [\mathbf{I}_{d_1}, \mathbf{0}]$ and $\mathbf{b} = \mathbf{0}$. The degree of freedom parameter v inherited from the mixing distribution $W \sim \text{IG}(v/2, v/2)$ remains invariant under affine transformation and marginal distribution. \square

Proposition 6. *The conditional distribution of \mathbf{X}_2 given $\mathbf{X}_1 = \mathbf{x}_1$ is a generalized hyperbolic distribution as in (6), 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*

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

The proof of Proposition 6 is similar to that for Proposition 3, hence is omitted. Similar results for Proposition 4, 5, and 6 have been obtained in Arellano-Valle and Genton (2010).

3 MGHD with Incomplete Data

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

$$f_{\text{MGHD}}(\mathbf{x}_i \mid \boldsymbol{\Theta}) = \sum_{g=1}^G \pi_g f_{\text{GHD}}(\mathbf{x}_i \mid \lambda_g, \omega_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g), \quad (14)$$

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 (9).

To apply the MGHD model (14) in the clustering paradigm, introduce $\mathbf{z}_i = (z_{i1}, \dots, z_{iG})^\top$, where $z_{ig} = 1$ if observation i is in component g and $z_{ig} = 0$ otherwise. The corresponding random variable $\mathbf{Z}_i \sim \mathcal{M}(1; \pi_1, \dots, \pi_G)$, i.e., \mathbf{Z}_i follows a multinomial distribution with one trial and cell probabilities π_1, \dots, π_G .

A three-level hierarchical representation of the MGHD model (14) can be expressed by

$$\begin{aligned} \mathbf{X}_i \mid w_{ig}, z_{ig} = 1 &\sim \mathcal{N}(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g), \\ W_{ig} \mid z_{ig} = 1 &\sim \mathcal{I}(\lambda_g, \eta = 1, \omega_g), \\ \mathbf{Z}_i &\sim \mathcal{M}(1; \pi_1, \dots, \pi_G). \end{aligned} \quad (15)$$

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

$$l_c(\boldsymbol{\Theta}) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} [\log \pi_g + \log \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g) + \log h(w_{ig} \mid \lambda_g, \omega_g)]. \quad (16)$$

Browne and McNicholas (2015) present an EM algorithm for parameter estimation with the MGHD when there is no missing data in $\mathbf{x}_1, \dots, \mathbf{x}_n$. We are interested in parameter estimation for the MGHD model (14) when $\mathbf{x}_1, \dots, \mathbf{x}_n$ 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^o and \mathbf{x}_i^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., $\mathbf{x}_i = (\mathbf{x}_i^{o_i\top}, \mathbf{x}_i^{m_i\top})^\top$, but can be simplified for the sake of clarity.

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

$$\boldsymbol{\Sigma}_g = \begin{pmatrix} \boldsymbol{\Sigma}_{g,i}^{\text{oo}} & \boldsymbol{\Sigma}_{g,i}^{\text{om}} \\ \boldsymbol{\Sigma}_{g,i}^{\text{mo}} & \boldsymbol{\Sigma}_{g,i}^{\text{mm}} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_g^{-1} = \begin{pmatrix} (\boldsymbol{\Sigma}_{g,i}^{\text{oo}})^{-1} & (\boldsymbol{\Sigma}_{g,i}^{\text{om}})^{-1} \\ (\boldsymbol{\Sigma}_{g,i}^{\text{mo}})^{-1} & (\boldsymbol{\Sigma}_{g,i}^{\text{mm}})^{-1} \end{pmatrix}, \quad (17)$$

correspond to $\mathbf{x}_i = (\mathbf{x}_i^{\text{o}\top}, \mathbf{x}_i^{\text{m}\top})^\top$. As a result, in addition to the observed \mathbf{x}_i^{o} , the missing group membership z_{ig} , and the latent variable w_{ig} , the complete-data also include the missing data \mathbf{x}_i^{m} . In the framework of the EM algorithm, the missing data \mathbf{x}_i^{m} are considered to be random variables that are updated in each iteration. Hence, the complete-data log-likelihood (16) is rewritten as

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

Given (15), we establish the following:

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

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

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

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

$$\mathbf{X}_i^{\text{m}} \mid \mathbf{x}_i^{\text{o}}, z_{ig} = 1 \sim \text{GH}_{p-p_i^{\text{o}}} \left(\lambda_{g,i}^{\text{m|o}}, \chi_{g,i}^{\text{m|o}}, \psi_{g,i}^{\text{m|o}}, \boldsymbol{\mu}_{g,i}^{\text{m|o}}, \boldsymbol{\Sigma}_{g,i}^{\text{m|o}}, \boldsymbol{\beta}_{g,i}^{\text{m|o}} \right), \quad (18)$$

where

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

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

$$\mathbf{X}_i^m \mid \mathbf{x}_i^o, w_{ig}, z_{ig} = 1 \sim \mathcal{N}_{p-p_i^o}(\boldsymbol{\mu}_{g,i}^{m|o} + w_{ig}\boldsymbol{\beta}_{g,i}^{m|o}, w_{ig}\boldsymbol{\Sigma}_{g,i}^{m|o}). \quad (19)$$

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

$$W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1 \sim \text{GIG} \left(\omega_g + (\boldsymbol{\beta}_{g,i}^o)^\top (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\beta}_{g,i}^o, \omega_g + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^o \mid \boldsymbol{\Sigma}_{g,i}^{oo}), \lambda_g - \frac{p_i^o}{2} \right). \quad (20)$$

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

$$\begin{aligned} l_c(\boldsymbol{\Theta}) &= \sum_{i=1}^n \sum_{g=1}^G z_{ig} \log \pi_g + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[-\frac{p}{2} \log(2\pi) - \frac{p}{2} \log w_{ig} + \frac{1}{2} \log |\boldsymbol{\Sigma}_g^{-1}| \right] \\ &\quad - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\boldsymbol{\Sigma}_g^{-1} z_{ig} \frac{1}{w_{ig}} \begin{pmatrix} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)(\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top & (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)(\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \\ (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o) & (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)(\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \end{pmatrix} \right) \\ &\quad + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\boldsymbol{\Sigma}_g^{-1} z_{ig} \begin{pmatrix} \boldsymbol{\beta}_{g,i}^o \\ \boldsymbol{\beta}_{g,i}^m \end{pmatrix} \begin{pmatrix} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top & (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \end{pmatrix} \right) \\ &\quad + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\boldsymbol{\Sigma}_g^{-1} z_{ig} \begin{pmatrix} \mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o \\ \mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m \end{pmatrix} \begin{pmatrix} (\boldsymbol{\beta}_{g,i}^o)^\top & (\boldsymbol{\beta}_{g,i}^m)^\top \end{pmatrix} \right) - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} w_{ig} \boldsymbol{\beta}_{g,i}^\top \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\beta}_{g,i} \\ &\quad + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[(\lambda_g - 1) \log w_{ig} - \log(2K_{\lambda_g}(\omega_g)) - \frac{\omega_g}{2} \left(w_{ig} + \frac{1}{w_{ig}} \right) \right]. \end{aligned} \quad (21)$$

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

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:

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

Given (2), (3), and (20), we have the following expectations as to the latent variable W :

$$a_{ig}^{(k)} := \mathbb{E}(W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \sqrt{\frac{\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}{\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}}$$

$$\times \frac{K_{\lambda_g^{(k)} - \frac{p_i^o}{2} + 1} \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\omega_g^{(k)} + (\boldsymbol{\beta}_{g,i}^{o(k)})^\top (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}{K_{\lambda_g^{(k)} - \frac{p_i^o}{2}} \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\omega_g^{(k)} + (\boldsymbol{\beta}_{g,i}^{o(k)})^\top (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)},$$

$$b_{ig}^{(k)} := \mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)})$$

$$= -\frac{2\lambda_g^{(k)} - p_i^o}{\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})} + \sqrt{\frac{\omega_g^{(k)} + (\boldsymbol{\beta}_{g,i}^{o(k)})^\top (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}{\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}}$$

$$\times \frac{K_{\lambda_g^{(k)} - \frac{p_i^o}{2} + 1} \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\omega_g^{(k)} + (\boldsymbol{\beta}_{g,i}^{o(k)})^\top (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}{K_{\lambda_g^{(k)} - \frac{p_i^o}{2}} \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\omega_g^{(k)} + (\boldsymbol{\beta}_{g,i}^{o(k)})^\top (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)},$$

$$c_{ig}^{(k)} := \mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \log \left(\sqrt{\frac{\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}{\omega_g^{(k)} + (\boldsymbol{\beta}_{g,i}^{o(k)})^\top (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}} \right)$$

$$+ \frac{\partial}{\partial t} \log \left\{ K_t \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\omega_g^{(k)} + (\boldsymbol{\beta}_{g,i}^{o(k)})^\top (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right) \right\} \Big|_{t=(\lambda_g^{(k)} - \frac{p_i^o}{2})}.$$

For convenience, we use the following notation analogous to Browne and McNicholas (2015):

$$n_g^{(k)} = \sum_{i=1}^n \hat{z}_{ig}^{(k)}, \bar{a}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} a_{ig}^{(k)}, \bar{b}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} b_{ig}^{(k)}, \text{ and } \bar{c}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} c_{ig}^{(k)}.$$

For the actual missing data \mathbf{X}^m , we will also need the following expectations:

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

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

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

$$+ \boldsymbol{\mu}_{g,i}^{m|o(k)} (\boldsymbol{\beta}_{g,i}^{m|o(k)})^\top + \boldsymbol{\beta}_{g,i}^{m|o(k)} (\boldsymbol{\mu}_{g,i}^{m|o(k)})^\top + a_{ig}^{(k)} \boldsymbol{\beta}_{g,i}^{m|o(k)} (\boldsymbol{\beta}_{g,i}^{m|o(k)})^\top.$$

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:

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

where

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

and

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

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

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

The family of MGHD mixture models, with 14 parsimonious eigen-decomposed scaled covariance matrices corresponding to the famous GPCM family of models is proposed (see Appendix A for a brief discussion, including nomenclature). Details on the MST with

incomplete data are analogous to the MGHD with incomplete data and are provided in Appendix D.

4 Notes on Implementation

4.1 Initial values

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:

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

- 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 $\lambda_g^{(0)} = -1/2$ for the index and concentration parameters, which represents a special case of GHD (i.e., normal-inverse Gaussian) distribution, or set $v_g^{(0)} = 50$ 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.

4.2 Model Selection and Stopping Criterion

In general, the number of mixture components G is not known *a priori*, and needs to be estimated from the data. Two widely used model selection techniques are the Bayesian information criterion (BIC; Schwarz, 1978) and the integrated completed likelihood (ICL; Biernacki et al., 2000), which are given respectively by

$$\text{BIC} = 2l(\mathbf{x}, \hat{\Theta}) - \rho \log(n) \quad \text{and} \quad \text{ICL} \approx \text{BIC} + 2 \sum_{i=1}^n \sum_{g=1}^G \text{MAP} \{ \hat{z}_{ig} \} \log(\hat{z}_{ig}),$$

where $l(\hat{\Theta})$ is the maximized log-likelihood evaluated at the maximum likelihood estimate $\hat{\Theta}$, ρ is the number of free parameters, n is the number of observations, \hat{z}_{ig} represents the estimated *a posteriori* probability that \mathbf{x}_i arises from the g th component, and MAP denotes the maximum *a posteriori* 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 bigger the BIC or ICL value, the better the fitted model.

The EM algorithm can be stopped iterations after the maximum number of iterations, or when the Aitken stopping criterion (Aitken, 1926) is satisfied. The Aitken acceleration at iteration k is

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

where $l^{(k)}$ is the log-likelihood at iterations k . This yields an asymptotic estimate of the log-likelihood at iteration $k + 1$:

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

(Böhning et al., 1994; Lindsay, 1995), and the EM algorithm is stopped when $l_{\infty}^{(k+1)} - l^{(k)} < \epsilon$, provided this difference is positive (McNicholas et al., 2010).

5 Numerical Examples

Studies based on both simulated and real datasets are used to compare the clustering performance of the proposed approach. Our proposed family of models for incomplete data is compared to multivariate t mixture with ML estimation in the presence of missing values

(Mt). BIC is used to select the model; models with higher values of BIC are preferable. The adjusted Rand index (ARI; Hubert and Arabie, 1985) is used to compare predicted classifications to true classes when applicable. The Rand index (Rand, 1971) is the ratio of pairwise agreements to total pairs, and the ARI corrects the Rand index to account for chance agreement. The ARI has expected value 0 under random classification and takes the value 1 for perfect class agreement. A detailed discussion of the ARI, and arguments in favour of its use, are given by Steinley (2004).

5.1 Simulation Studies

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 (11) and (8), respectively.

For each mixture component, $n_g = 200$ two-dimensional vectors \mathbf{x}_i are generated. The presumed parameters of Σ_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 $\beta_1 = (1, 1)^\top$ and $\beta_2 = (-1, -1)^\top$, the degrees of freedoms for the MST is $v_1 = 7$ and $v_2 = 5$, and the values of other parameters for the MGHD are $\omega_1 = \omega_2 = 6$ and $\lambda_1 = -1/2$ and $\lambda_2 = 1$.

The datasets considered in the simulation studies are summarized in Table 1 and examples are plotted in Figure 1. The datasets are overlapping, making this a relatively difficult clustering scenario even when the datasets are complete.

Artificial missing datasets are simulated by removing $n \times r$ elements from each column of the simulated samples through two different MAR patterns and the MCAR mechanism under three missing rates — $r = 0.05$ (low), $r = 0.15$ (moderate), and $r = 0.3$ (high) — while maintaining the condition that each observation has at least one observed attribute. For the MAR mechanism, data points in the first column are sorted in descending order. Column 2 is then divided into four equal blocks and, for each block, a specified number of

Table 1: Summary of simulated datasets.

Dataset	Distribution	Covariance structure (Σ_g)	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

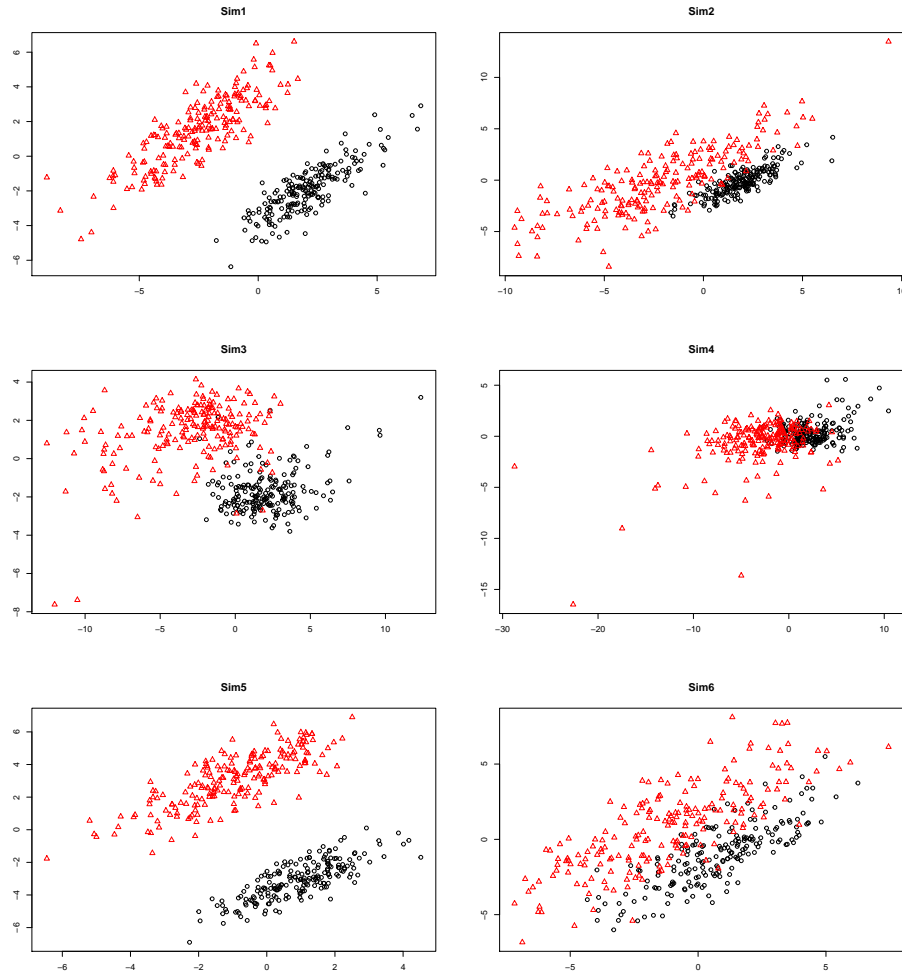


Figure 1: Exemplar scatter plots for simulated datasets, where colour and plotting symbol represent true labels (component membership).

elements (see Table 2) are removed at random. When $p = 1$, the second column is used.

Table 2: Number of missing observations for each pattern.

r	Pattern 1	Pattern 2
5%	(10,3,6,1)	(1,6,3,10)
15%	(30,9,18,3)	(3,18,9,30)
30%	(60,18,36,6)	(6,36,18,60)

First, we examine the ability of our proposed model to recover underlying parameters when the number of components and the covariance structure are correctly specified. These experiments comprise 100 replications per combination of missing pattern and missingness rate. The means of the parameter estimates with their associated standard deviations and bias are summarized in Table 8 and 9 (Appendix E). The means of most parameter estimates are close to the true values with small standard deviations when $r = 0.05$. The standard deviations increase as the missing rate increases, while at the same time, the average ARI slightly decreases. The means of estimated λ_1 and λ_2 in Sim1 are quite far from the true value because we obtain those estimates using an approximation to the Bessel function. In addition, there is no significant difference among the three missing patterns. Therefore, we use MCAR in the rest of the data examples.

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, and index parameters. We compute the average ARI for the parsimonious MGHD and MST models introduced here as well as Mt under the circumstances of unknown clusters ($G = 1, \dots, 4$). The detailed results are summarized in Table 10 (Appendix E). From Table 10, we observe the following:

- The average ARI decreases as the missing rate rises. As expected, overlapping components typically have lower ARI than the well-separated components. In addition, the average ARI considerably decreases when the missing rate reaches 30% ($r = 0.30$) for Sim2, Sim4 and Sim6.
- Our proposed parsimonious MGHD models for incomplete data perform significantly better than Mt . The family of MGHD models generally yields much higher ARI than its competitor parsimonious MST for incomplete data when the datasets are generated from a generalized hyperbolic distribution.

- 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 or Mt for incomplete data for datasets with overlapping mixtures.
- The BIC prefers MGHD over Mt in Sim5 and Sim6 where the data is generated from GMMs. We find that the samples are not necessarily symmetric, particularly with missing values. Figure 2 and 3 show exemplar scatter plots for data from Sim5 and Sim6 for $r = 0.10$. The Mt tends to overestimate the number of clusters, hence, has a lower averaged BIC.

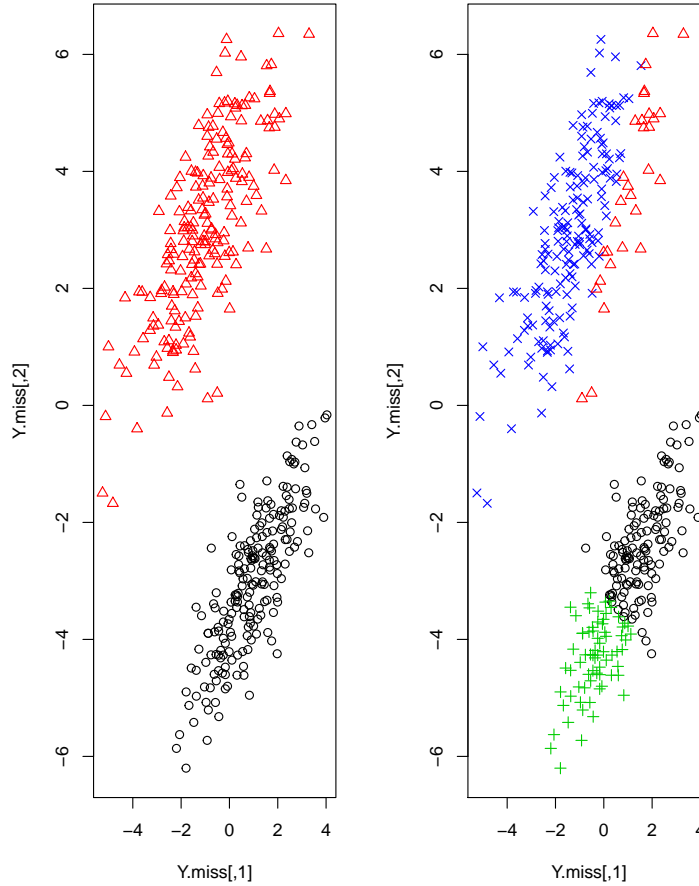


Figure 2: Exemplar scatter plots for Sim5, with true labels (left) and clustering results from the best Mt models (right), where colour and plotting symbol represent true (left) or predicted (right) class.

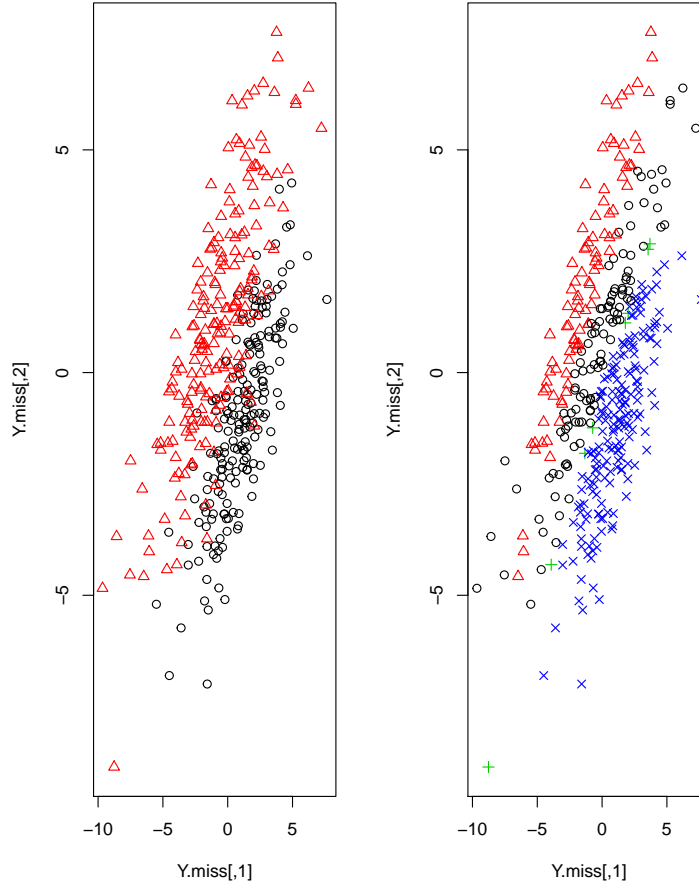


Figure 3: Exemplar scatter plots for Sim6, with true labels (left) and clustering results from the best Mt models (right), where colour and plotting symbol represent true (left) or predicted (right) class.

5.2 Breast Cancer Diagnostic Dataset

The breast cancer diagnostic data consists of ten real-valued features on 569 cases of breast tumours – 357 benign and 212 malignant. The mean, standard error, and “worst” or largest of these features were computed for each image, resulting in 30 attributes. This dataset is complete, so for illustration purposes we consider levels of missing data $r = 0.05$ and $r = 0.15$ by deleting observations through an MCAR mechanism while maintaining the condition that each observation has at least one observed attribute. The dataset is scaled prior to analysis.

The family of MGHD, MST and Mt models were fitted to these data for $G = 1, \dots, 4$.

We randomly assign each observation to one of the G groups and start with 20 random initializations of the algorithm, selecting the model with the maximum likelihood values. The key statistics of the best models for MGHD, MST and Mt are shown in Table 3. The results of this analysis show that the parsimonious MGHD outperforms the other models for all levels of missing data.

Table 3: A comparison of averaged BIC, ARI and the number of times (nt) when $G = 2$ is chosen among MGHD, MST, and Mt models on the tumour dataset with $G = 1, \dots, 4$.

	$r = 0.05$			$r = 0.15$		
	Avg.BIC	Avg.ARI	nt	Avg.BIC	Avg.ARI	nt
MGHD	12145	0.65	18	9654	0.58	16
MST	12661	0.55	15	10574	0.56	16
Mt	13605	0.47	10	11605	0.36	10

5.3 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 4. 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.

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. Lin (2014) perform the Mt 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%). The best model is the two-component MGHD model and $\Sigma_g = \text{EVE}$. Group 1 consists mainly of the non-diabetic patients and Group 2 consists mainly of the diabetic patients. We then fit the best model with 100 random initializations; Table 6 shows the key parameter estimates for this model as well as the corresponding standard errors. The standard errors of the model parameters

Table 4: 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(μ 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

Table 5: The BIC, ICL, selected Σ_g and the correct classification rate for our proposed approaches for clustering on the Pima Indian diabetes dataset.

	Σ_g	BIC	ICL	Accuracy
MGHD	EVE	-14016.95	-14053.61	69.11%
MST	VVI	-14109.1	-14186.1	62.37%

have been calculated using the bootstrap method described in Efron and Tibshirani (1986). The estimates for $\mu_g + \beta_g$ are quite similar to the parameter estimates presented in Wang and Lin (2015). The estimates for the skewness parameters indicate the presence of skewness in most of the variables.

6 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

Table 6: Summary of key model parameter estimates (standard errors) for the best chosen model (i.e., MGHd with $\Sigma_g = \text{EVE}$) for the Pima Indian diabetes dataset.

Parameter	$g = 1$	$g = 2$
μ_{1g}	-0.80 (0.11)	2.98 (1.78)
μ_{2g}	-0.97 (0.22)	1.35 (4.01)
μ_{3g}	-0.69 (0.14)	1.10 (2.65)
μ_{4g}	0.15 (0.08)	-0.50 (4.59)
μ_{5g}	-1.26 (1.73)	0.18 (0.25)
μ_{6g}	-0.66 (0.07)	0.57 (0.78)
μ_{7g}	-0.74 (0.12)	-2.67 (8.41)
μ_{8g}	-1.20 (0.31)	-2.01 (2.04)
β_{1g}	0.57 (0.05)	-2.18 (1.79)
β_{2g}	0.77 (0.47)	-0.92 (0.25)
β_{3g}	0.54 (0.40)	-0.78 (1.18)
β_{4g}	0.53 (0.31)	0.58 (0.38)
β_{5g}	0.11 (0.13)	0.11 (0.32)
β_{6g}	0.57 (0.16)	-0.37 (0.51)
β_{7g}	0.63 (0.18)	1.27 (0.46)
β_{8g}	0.87 (0.16)	2.91 (1.85)
ω_g	2.39 (1.81)	14.18 (6.83)
λ_g	0.02 (0.34)	-3.18 (4.60)
π_g	0.71 (0.09)	0.29 (0.10)

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.

In the present work, the missing data mechanism is assumed to be MAR. Future work will focus on a departure from this assumption. As a starting point, the behaviour of parameter estimates for models considered herein when we depart from the MAR assumption will be studied. Although we demonstrated the PMGHD and PMST approaches for clustering, they also can be applied for semi-supervised classification, discriminant analysis, and density estimation; furthermore, they could be used within the fractionally-supervised paradigm (Vrbik and McNicholas, 2015). 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. Finally, it will also be interesting to generalize all existing approaches to developing mixture of generalized hyperbolic factor analyzer models (Tortora et al., 2016), mixtures with hypercube contours (Franczak et al., 2015), and mixtures of multiple scaled generalized hyperbolic distributions for incomplete data (Tortora et al., 2017).

Acknowledgements This work was supported by an Ontario Graduate Scholarship (Wei), an Early Researcher Award from the Government of Ontario (McNicholas), and the Canada Research Chairs program (McNicholas).

References

- Aitken, A. C. (1926). On Bernoulli’s numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh* 46, 289–305.
- Andrews, J. L. and P. D. McNicholas (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing* 21(3), 361–373.
- Andrews, J. L. and P. D. McNicholas (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing* 22(5), 1021–1029.
- Arellano-Valle, R. and M. G. Genton (2010). Multivariate extended skew-t distributions and related families. *Metron* 68(3), 201–234.

- Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49(3), 803–821.
- Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 353(1674), 401–419.
- 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 (1977). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Probability Theory and Related Fields* 38(4), 309–311.
- 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*. Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.
- 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 and 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 (Methodological)*, 1–38.
- Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1(1), 54–75.
- 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.
- 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.
- Franczak, B. C., C. Tortora, R. P. Browne, and P. D. McNicholas (2015). Unsupervised learning via mixtures of skewed distributions with hypercube contours. *Pattern Recognition Letters* 58(1), 69–76.
- Ghahramani, Z. and M. I. Jordan (1994). Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems*. Citeseer.
- Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* 40(3-4), 237–264.
- 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*. Lecture Notes in Statistics. New York: Springer.

- Karlis, D. and A. Santourian (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing* 19(1), 73–83.
- 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.
- 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 and 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. and T.-C. Lin (2011). Robust statistical modelling using the multivariate skew t distribution with complete and incomplete data. *Statistical Modelling* 11(3), 253–277.
- Lin, T.-I., P. D. McNicholas, and H. J. Ho (2014). Capturing patterns via parsimonious t mixture models. *Statistics and Probability Letters* 88, 80–87.
- 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 (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
- Liu, C., D. B. Rubin, and Y. N. Wu (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* 85(4), 755–770.
- McLachlan, G. J., D. Peel, and R. Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis* 41(3), 379–388.
- McNeil, A., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ.

- McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Boca Raton: Chapman and Hall/CRC Press.
- McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification* *33*(3), 331–373.
- 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 and Data Analysis* *54*(3), 711–723.
- Morris, K. and P. D. McNicholas (2016). Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures. *Computational Statistics and Data Analysis* *97*, 133–150.
- Murray, P. M., R. P. Browne, and P. D. McNicholas (2014a). Mixtures of skew-factor analyzers. *Computational Statistics and Data Analysis* *77*, 326–335.
- Murray, P. M., R. B. Browne, and P. D. McNicholas (2017a). Hidden truncation hyperbolic distributions, finite mixtures thereof, and their application for clustering. *Journal of Multivariate Analysis* *161*, 141–156.
- Murray, P. M., R. B. Browne, and P. D. McNicholas (2017b). A mixture of SDB skew-t factor analyzers. *Econometrics and Statistics* *3*, 160–168.
- Murray, P. M., P. D. McNicholas, and R. P. Browne (2014b). A mixture of common skew-t factor analysers. *Stat* *3*(1), 68–82.
- 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 and Data Analysis* *93*, 18–30.
- Peel, D. and G. J. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* *10*(4), 339–348.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* *66*, 846–850.

- Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63(3), 581–592.
- 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.
- Steane, M. A., P. D. McNicholas, and R. Yada (2012). Model-based classification via mixtures of multivariate t-factor analyzers. *Communications in Statistics – Simulation and Computation* 41(4), 510–523.
- Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods* 9, 386–396.
- 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.
- Tortora, C., B. C. Franczak, R. P. Browne, and P. D. McNicholas (2017). A mixture of coalesced generalized hyperbolic distributions. arXiv preprint arXiv:1403.2332v7.
- 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.
- Vrbik, I. and P. D. McNicholas (2012). Analytic calculations for the EM algorithm for multivariate skew-t mixture models. *Statistics and Probability Letters* 82(6), 1169–1174.
- Vrbik, I. and P. D. McNicholas (2015). Fractionally-supervised classification. *Journal of Classification* 32(3), 359–381.
- Wang, W. L. and T.-I. Lin (2015). Robust model-based clustering via mixtures of skew-t distributions with missing information. *Advances in Data Analysis and Classification* 9(4), 423–445.
- 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.

A GPCM Family

Banfield and Raftery (1993) consider an eigen-decomposition of the component scale matrices (which is equivalent to the component covariance matrices for Gaussian mixtures), i.e.,

$$\mathbf{\Sigma}_g = \lambda_g \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g', \quad (22)$$

where $\lambda_g = |\mathbf{\Sigma}_g|^{1/p}$, $\mathbf{\Gamma}_g$ is the matrix of eigenvectors of $\mathbf{\Sigma}_g$, and $\mathbf{\Delta}_g$ is a diagonal matrix, such that $|\mathbf{\Delta}_g| = 1$, containing the normalized eigenvalues of $\mathbf{\Sigma}_g$ in decreasing order. Note that the columns of $\mathbf{\Gamma}_g$ are ordered to correspond to the elements of $\mathbf{\Delta}_g$. As Banfield and Raftery (1993) point out, the constituent elements of the decomposition in (22) can be viewed in the context of the geometry of the component, where λ_g represents the volume in p -space, $\mathbf{\Delta}_g$ the shape, and $\mathbf{\Gamma}_g$ the orientation. By imposing constraints on the elements of the decomposed covariance structure in (22), Celeux and Govaert (1995) introduce a family of GPCMs (Table 7).

Table 7: The nomenclature and scale matrix structure for each member of the GPCM family.

Nomenclature	Volume	Shape	Orientation	$\mathbf{\Sigma}_g$
EII	Equal	Spherical		$\lambda \mathbf{I}$
VII	Variable	Spherical		$\lambda_g \mathbf{I}$
EEI	Equal	Equal	Axis-Aligned	$\lambda \mathbf{\Delta}$
VEI	Variable	Equal	Axis-Aligned	$\lambda_g \mathbf{\Delta}$
EVI	Equal	Variable	Axis-Aligned	$\lambda \mathbf{\Delta}_g$
VVI	Variable	Variable	Axis-Aligned	$\lambda_g \mathbf{\Delta}_g$
EEE	Equal	Equal	Equal	$\lambda \mathbf{\Gamma} \mathbf{\Delta} \mathbf{\Gamma}'$
VEE	Variable	Equal	Equal	$\lambda_g \mathbf{\Gamma} \mathbf{\Delta} \mathbf{\Gamma}'$
EVE	Equal	Variable	Equal	$\lambda \mathbf{\Gamma} \mathbf{\Delta}_g \mathbf{\Gamma}'$
EEV	Equal	Equal	Variable	$\lambda \mathbf{\Gamma}_g \mathbf{\Delta} \mathbf{\Gamma}_g'$
VVE	Variable	Variable	Equal	$\lambda_g \mathbf{\Gamma} \mathbf{\Delta}_g \mathbf{\Gamma}'$
VEV	Variable	Equal	Variable	$\lambda_g \mathbf{\Gamma}_g \mathbf{\Delta} \mathbf{\Gamma}_g'$
EVV	Equal	Variable	Variable	$\lambda \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g'$
VVV	Variable	Variable	Variable	$\lambda_g \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g'$

B Some Useful 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{X} in Propositions 3 and 6.

Consider a partitioned random vector \mathbf{X} of p -dimension that follows the pdf as in (9) with

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad (23)$$

where \mathbf{X}_1 and \mathbf{X}_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{X} . Sometimes, it is more convenient to work with the inverse of dispersion matrix $\boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^\top)^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \\ -(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_{11}^{-1} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{pmatrix}. \quad (24)$$

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

$$\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Sigma}_{11})\det(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \quad (25)$$

C Outline of Proof of Proposition 3

Here, we derive the conditional density of \mathbf{X}_2 given that $\mathbf{X}_1 = \mathbf{x}_1$ if \mathbf{X}_1 and \mathbf{X}_2 are jointly generalized hyperbolic distributed, i.e., $\mathbf{X} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ with the partition in Appendix A. 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 A are heavily used in the course of the derivations. The conditional density is given by

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

where we combine (9) and Proposition 2. For the moment, we focus on the linear form and quadratic form in which \mathbf{x} enters the pdf in (9). Inserting the partition of $\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$ in (23) and the

inverse of dispersion matrix Σ^{-1} (24) into the quadratic form yields

$$\begin{aligned}
\delta(\mathbf{x}, \boldsymbol{\mu} \mid \Sigma) &= (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \begin{pmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top & (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \\
&= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^\top)^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
&\quad - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} \Sigma_{12}^\top \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
&\quad - (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&\quad + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
&\quad + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} \Sigma_{12}^\top \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
&\quad - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} \Sigma_{12}^\top \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
&\quad - (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1} \Sigma_{12} (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&\quad + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
&\quad + (\mathbf{x}_2 - \boldsymbol{\mu}_2 - \Sigma_{12}^\top \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1))^\top (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2 - \Sigma_{12}^\top \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)) \\
&= \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 \mid \Sigma_{11}) + \delta(\mathbf{x}_2, \boldsymbol{\mu}_{2|1} \mid \Sigma_{2|1}), \tag{26}
\end{aligned}$$

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

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

$$\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} \boldsymbol{\beta} &= \begin{pmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top & (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \\
&= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1} \boldsymbol{\beta}_1 + (\mathbf{x}_2 - \boldsymbol{\mu}_2 - \Sigma_{12}^\top \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1))^\top (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\boldsymbol{\beta}_2 - \Sigma_{12}^\top \Sigma_{11}^{-1} \boldsymbol{\beta}_1) \\
&= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1} \boldsymbol{\beta}_1 + (\mathbf{x}_2 - \boldsymbol{\mu}_{2|1})^\top \Sigma_{2|1}^{-1} \boldsymbol{\beta}_{2|1}, \tag{27}
\end{aligned}$$

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

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

$$\begin{aligned}
\boldsymbol{\beta}^\top \Sigma^{-1} \boldsymbol{\beta} &= \begin{pmatrix} \boldsymbol{\beta}_1^\top & \boldsymbol{\beta}_2^\top \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \\
&= \boldsymbol{\beta}_1^\top \Sigma_{11}^{-1} \boldsymbol{\beta}_1 + (\boldsymbol{\beta}_2 - \Sigma_{12}^\top \Sigma_{11}^{-1} \boldsymbol{\beta}_1)^\top (\Sigma_{22} - \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12})^{-1} (\boldsymbol{\beta}_2 - \Sigma_{12}^\top \Sigma_{11}^{-1} \boldsymbol{\beta}_1) \\
&= \boldsymbol{\beta}_1^\top \Sigma_{11}^{-1} \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2|1}^\top \Sigma_{2|1}^{-1} \boldsymbol{\beta}_{2|1}. \tag{28}
\end{aligned}$$

Finally, we substitute (25), (26), (27), and (28), and $p = d_1 + d_2$ into the conditional density,

and after some simple linear algebra, we obtain

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

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

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

Comparison with (6) reveals that this is a generalized hyperbolic distribution in the parameterization of McNeil et al. (2005) with

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

D MST with Incomplete Data

Analogous to the MGHD model (14), the MST model takes the density

$$f_{\text{MST}}(\mathbf{X}_i | \boldsymbol{\Theta}) = \sum_{g=1}^G \pi_g f_{\text{ST}}(\mathbf{X}_i | v_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g), \quad (29)$$

where $\boldsymbol{\Theta} = (\pi, \mathbf{v}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g)$ with $\mathbf{v}_g = (v_1, \dots, v_g)$ and $\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$, and $\boldsymbol{\beta}_g$ are as defined above. By introducing the group membership variables $\mathbf{Z}_i \sim \mathcal{M}(1; \pi_1, \dots, \pi_G)$, convenient three-layer

hierarchical representations are given by

$$\begin{aligned}\mathbf{X}_i \mid w_{ig}, z_{ig} = 1 &\sim \mathcal{N}(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g) \\ W_{ig} \mid z_{ig} = 1 &\sim \text{IG}(v_g/2, v_g/2). \\ \mathbf{Z}_i &\sim \mathcal{M}(1; \pi_1, \dots, \pi_G)\end{aligned}\tag{30}$$

Assume that the matrix $\mathbf{X} = (\mathbf{X}^{\text{o}\top}, \mathbf{X}^{\text{m}\top})^\top$ contains missing data. For each $\mathbf{x}_i = (\mathbf{x}_i^{\text{o}\top}, \mathbf{x}_i^{\text{m}\top})^\top$, we write $\boldsymbol{\mu}_g = (\boldsymbol{\mu}_{g,i}^{\text{o}\top}, \boldsymbol{\mu}_{g,i}^{\text{m}\top})^\top$, $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g,i}^{\text{o}\top}, \boldsymbol{\beta}_{g,i}^{\text{m}\top})^\top$, and finally the g th dispersion matrix $\boldsymbol{\Sigma}_g$ is partitioned as in (17). Hence, based on (30), we have the following conditional distributions:

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

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

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

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

$$\mathbf{X}_i^{\text{m}} \mid \mathbf{x}_i^{\text{o}}, z_{ig} = 1 \sim \text{GH}_{p-p_i^{\text{o}}}(\lambda_{g,i}^{\text{m|o}}, \chi_{g,i}^{\text{m|o}}, \psi_{g,i}^{\text{m|o}}, \boldsymbol{\mu}_{g,i}^{\text{m|o}}, \boldsymbol{\Sigma}_{g,i}^{\text{m|o}}, \boldsymbol{\beta}_{g,i}^{\text{m|o}}),\tag{31}$$

where

$$\begin{aligned}\lambda_{g,i}^{\text{m|o}} &= -\frac{v_g + p_i^{\text{o}}}{2}, & \psi_{g,i}^{\text{m|o}} &= v_g + (\mathbf{x}_i^{\text{o}} - \boldsymbol{\mu}_{g,i}^{\text{o}})^\top (\boldsymbol{\Sigma}_{g,i}^{\text{oo}})^{-1} (\mathbf{x}_i^{\text{o}} - \boldsymbol{\mu}_{g,i}^{\text{o}}), \\ \psi_{g,i}^{\text{m|o}} &= \boldsymbol{\beta}_{g,i}^{\text{o}\top} (\boldsymbol{\Sigma}_{g,i}^{\text{oo}})^{-1} \boldsymbol{\beta}_{g,i}^{\text{o}}, & \boldsymbol{\mu}_{g,i}^{\text{m|o}} &= \boldsymbol{\mu}_{g,i}^{\text{m}} + (\boldsymbol{\Sigma}_{g,i}^{\text{om}})^\top (\boldsymbol{\Sigma}_{g,i}^{\text{oo}})^{-1} (\mathbf{x}_i^{\text{o}} - \boldsymbol{\mu}_{g,i}^{\text{o}}), \\ \boldsymbol{\Sigma}_{g,i}^{\text{m|o}} &= \boldsymbol{\Sigma}_{g,i}^{\text{mm}} - (\boldsymbol{\Sigma}_{g,i}^{\text{om}})^\top (\boldsymbol{\Sigma}_{g,i}^{\text{oo}})^{-1} \boldsymbol{\Sigma}_{g,i}^{\text{om}}, & \boldsymbol{\beta}_{g,i}^{\text{m|o}} &= \boldsymbol{\beta}_{g,i}^{\text{m}} - (\boldsymbol{\Sigma}_{g,i}^{\text{om}})^\top (\boldsymbol{\Sigma}_{g,i}^{\text{oo}})^{-1} \boldsymbol{\beta}_{g,i}^{\text{o}}.\end{aligned}$$

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

$$\mathbf{X}_i^{\text{m}} \mid \mathbf{x}_i^{\text{o}}, w_{ig}, z_{ig} = 1 \sim \mathcal{N}_{p-p_i^{\text{o}}}(\boldsymbol{\mu}_{g,i}^{\text{m|o}} + w_{ig}\boldsymbol{\beta}_{g,i}^{\text{m|o}}, w_{ig}\boldsymbol{\Sigma}_{g,i}^{\text{m|o}}).\tag{32}$$

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

$$W_{ig} \mid \mathbf{x}_i^{\text{o}}, z_{ig} = 1 \sim \text{GIG}\left(\boldsymbol{\beta}_{g,i}^{\text{o}\top} (\boldsymbol{\Sigma}_{g,i}^{\text{oo}})^{-1} \boldsymbol{\beta}_{g,i}^{\text{o}}, v_g + \delta(\mathbf{x}_i^{\text{o}}, \boldsymbol{\mu}_{g,i}^{\text{o}} \mid \boldsymbol{\Sigma}_{g,i}^{\text{oo}}), -\frac{v_g + p_i^{\text{o}}}{2}\right).\tag{33}$$

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^m , for $i = 1, \dots, n$ and $g = 1, \dots, G$. Again, the complete data log-likelihood function is given by

$$l_c(\Theta) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} [\log \pi_g + \log \phi(\mathbf{x}_i^o, \mathbf{x}_i^m \mid \boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g) + \log f_{\text{IG}}(w_{ig} \mid v_g/2, v_g/2)]. \quad (34)$$

Furthermore, one can simplify (34) to

$$\begin{aligned} l_c(\Theta) = & \sum_{i=1}^n \sum_{g=1}^G z_{ig} \log \pi_g + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[-\frac{p}{2} \log(2\pi) - \frac{p}{2} \log w_{ig} + \frac{1}{2} \log |\boldsymbol{\Sigma}_g^{-1}| \right] \\ & - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\boldsymbol{\Sigma}_g^{-1} z_{ig} \frac{1}{w_{ig}} \begin{pmatrix} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)(\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top & (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)(\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \\ (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o) & (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)(\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \end{pmatrix} \right) \\ & + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\boldsymbol{\Sigma}_g^{-1} z_{ig} \begin{pmatrix} \boldsymbol{\beta}_{g,i}^o \\ \boldsymbol{\beta}_{g,i}^m \end{pmatrix} \begin{pmatrix} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top & (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \end{pmatrix} \right) \\ & + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\boldsymbol{\Sigma}_g^{-1} z_{ig} \begin{pmatrix} \mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o \\ \mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{g,i}^{o\top} & \boldsymbol{\beta}_{g,i}^{m\top} \end{pmatrix} \right) - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} w_{ig} \boldsymbol{\beta}_{g,i}^\top \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\beta}_{g,i} \\ & + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[\frac{v_g}{2} \log \left(\frac{v_g}{2} \right) - \log \Gamma \left(\frac{v_g}{2} \right) - \left(\frac{v_g}{2} + 1 \right) \log w_{ig} - \frac{v_g}{2w_{ig}} \right]. \end{aligned} \quad (35)$$

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}^o and the current parameter updates $\Theta^{(k)}$. 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

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

Given the observed data \mathbf{x}^o , the current parameter updates $\Theta^{(k)}$, and conditional distributions

(31) and (33), taking expectations for (35) leads to the following expectation updates in the E-step:

$$\begin{aligned}
A_{ig}^{(k)} &:= \mathbb{E}(W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \sqrt{\frac{v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}{\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}}} \\
&\quad \times \frac{K_{-(v_g^{(k)} + p_i^0)/2+1} \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}{K_{-(v_g^{(k)} + p_i^0)/2} \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}}, \\
B_{ig}^{(k)} &:= \mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) \\
&= \frac{v_g^{(k)} + p_i^0}{v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})} + \sqrt{\frac{\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}{v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}} \\
&\quad \times \frac{K_{-(v_g^{(k)} + p_i^0)/2+1} \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}{K_{-(v_g^{(k)} + p_i^0)/2} \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}, \\
C_{ig}^{(k)} &:= \mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \log \left(\sqrt{\frac{v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}{\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}}} \right) \\
&\quad + \frac{\partial}{\partial t} \log \left\{ K_t \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right) \right\} \Big|_{t=-(v_g^{(k)} + p_i^0)/2}, \\
\hat{\mathbf{x}}_{ig}^{m(k)} &:= \mathbb{E}(\mathbf{X}_i^m \mid \mathbf{x}_i^o, z_{ig} = 1) = \boldsymbol{\mu}_{g,i}^{m|o(k)} + A_{ig}^{(k)} \boldsymbol{\beta}_{g,i}^{m|o(k)}, \\
\hat{\mathbf{x}}_{ig}^{m(k)} &:= \mathbb{E}((1/W_i) \mathbf{X}_i^m \mid \mathbf{x}_i^o, z_{ig} = 1) = B_{ig}^{(k)} \boldsymbol{\mu}_{g,i}^{m|o(k)} + \boldsymbol{\beta}_{g,i}^{m|o(k)}, \\
\hat{\mathbf{x}}_{ig}^{m(k)} &:= \mathbb{E}((1/w_i) \mathbf{X}_i^m \mathbf{X}_i^{m\top} \mid \mathbf{x}_i^o, z_{ig} = 1) = \boldsymbol{\Sigma}_{g,i}^{m|o(k)} + B_{ig}^{(k)} \boldsymbol{\mu}_{g,i}^{m|o(k)} (\boldsymbol{\mu}_{g,i}^{m|o(k)})^\top \\
&\quad + \boldsymbol{\mu}_{g,i}^{m|o(k)} (\boldsymbol{\beta}_{g,i}^{m|o(k)})^\top + \boldsymbol{\beta}_{g,i}^{m|o(k)} (\boldsymbol{\mu}_{g,i}^{m|o(k)})^\top + A_{ig}^{(k)} \boldsymbol{\beta}_{g,i}^{m|o(k)} (\boldsymbol{\beta}_{g,i}^{m|o(k)})^\top.
\end{aligned}$$

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:

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

where

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

where

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

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

$$\log \left(\frac{v_g^{(k+1)}}{2} \right) + 1 - \varphi \left(\frac{v_g^{(k+1)}}{2} \right) - \frac{1}{n_g^{(k)}} \sum_{i=1}^n \tau_{ig} (C_{ig}^{(k)} + B_{ig}^{(k)}) = 0, \quad (36)$$

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

E Results from Simulation Studies

The results from the simulation studies are summarized in Tables 8, 9 and 10.

Table 8: Key model parameters as well as means, standard deviations and bias of the associated parameter estimations from the 100 runs for the first simulation experiment.

Sim1 using MGHd									
$r = 0.05$									
	Pattern 1			Pattern 2			MCAR		
	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias
μ_1	(0.70, -3.30)'	(1.06, 0.94)'	(-0.30, -0.30)'	(0.45, -3.40)'	(1.23, 1.08)'	(-0.55, -0.40)'	(0.64, -3.22)'	(0.86, 0.79)'	(-0.36, -0.22)'
μ_2	(-0.74, 3.47)'	(2.11, 2.49)'	(0.26, 0.47)'	(-0.57, 3.40)'	(2.62, 2.17)'	(0.43, 0.41)'	(-0.64, 3.42)'	(2.54, 2.20)'	(0.36, 0.42)'
β_1	(1.59, 1.59)'	(1.30, 1.16)'	(0.59, 0.59)'	(1.91, 1.74)'	(1.49, 1.35)'	(0.91, 0.74)'	(1.67, 1.50)'	(1.06, 0.98)'	(0.67, 0.50)'
β_2	(-1.73, -1.98)'	(2.50, 2.92)'	(-0.73, -0.98)'	(-1.96, -1.91)'	(3.10, 2.56)'	(-0.96, -0.91)'	(-1.86, -1.94)'	(2.32, 2.09)'	(-0.86, -0.94)'
$\mu_1 + \beta_1$	(2.29, -1.71)'	(0.26, 0.25)'	(0.29, 0.29)'	(2.36, -1.66)'	(0.32, 0.32)'	(0.36, 0.34)'	(2.31, -1.71)'	(0.25, 0.26)'	(0.31, 0.29)'
$\mu_2 + \beta_2$	(-2.47, 1.48)'	(0.44, 0.47)'	(-0.47, -0.52)'	(-2.54, 1.50)'	(0.51, 0.43)'	(-0.54, -0.50)'	(-2.50, 1.48)'	(0.50, 0.54)'	(-0.50, -0.52)'
Σ_1	$\begin{bmatrix} 1.88 & 1.51 \\ 1.51 & 1.90 \end{bmatrix}$	$\begin{bmatrix} 0.32 & 0.27 \\ 0.27 & 0.30 \end{bmatrix}$	$\begin{bmatrix} 0.21 & 0.18 \\ 0.18 & 0.23 \end{bmatrix}$	$\begin{bmatrix} 1.96 & 1.57 \\ 1.57 & 1.98 \end{bmatrix}$	$\begin{bmatrix} 0.34 & 0.28 \\ 0.28 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 0.29 & 0.24 \\ 0.24 & 0.31 \end{bmatrix}$	$\begin{bmatrix} 1.95 & 1.57 \\ 1.57 & 1.97 \end{bmatrix}$	$\begin{bmatrix} 0.36 & 0.30 \\ 0.30 & 0.34 \end{bmatrix}$	$\begin{bmatrix} 0.28 & 0.25 \\ 0.25 & 0.30 \end{bmatrix}$
Σ_2	$\begin{bmatrix} 4.42 & 3.55 \\ 3.55 & 4.48 \end{bmatrix}$	$\begin{bmatrix} 0.66 & 0.58 \\ 0.58 & 0.68 \end{bmatrix}$	$\begin{bmatrix} 1.09 & 0.88 \\ 0.88 & 1.15 \end{bmatrix}$	$\begin{bmatrix} 4.38 & 3.53 \\ 3.53 & 4.43 \end{bmatrix}$	$\begin{bmatrix} 0.76 & 0.66 \\ 0.66 & 0.76 \end{bmatrix}$	$\begin{bmatrix} 1.05 & 0.86 \\ 0.86 & 1.10 \end{bmatrix}$	$\begin{bmatrix} 4.43 & 3.56 \\ 3.56 & 4.50 \end{bmatrix}$	$\begin{bmatrix} 0.68 & 0.58 \\ 0.58 & 0.68 \end{bmatrix}$	$\begin{bmatrix} 1.10 & 0.89 \\ 0.89 & 1.17 \end{bmatrix}$
λ_1	-2.26	1.27	-1.76	-2.70	1.70	-2.20	-2.51	1.26	-2.01
λ_2	2.93	1.40	1.93	2.79	1.40	1.79	2.88	1.40	1.88
π_1	0.50	0.00	0.00	0.50	0.01	0.00	0.50	0.01	0.00
π_2	0.50	0.00	0.00	0.50	0.01	0.00	0.50	0.01	0.00
ARI	0.96	0.02	0.00	0.96	0.02	0.00	0.95	0.09	0.00
$r = 0.15$									
	Pattern 1			Pattern 2			MCAR		
	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias
μ_1	(0.81, -3.14)'	(1.14, 1.14)'	(-0.19, -0.14)'	(0.72, -3.24)'	(1.10, 0.89)'	(-0.28, -0.24)'	(0.75, -3.23)'	(1.24, 1.11)'	(-0.25, -0.23)'
μ_2	(-0.69, 3.45)'	(2.58, 2.25)'	(0.31, 0.45)'	(-0.49, 3.47)'	(2.90, 3.00)'	(0.51, 0.47)'	(-0.67, 3.34)'	(1.95, 1.78)'	(0.33, 0.34)'
β_1	(1.46, 1.41)'	(1.40, 1.41)'	(0.46, 0.41)'	(1.38, 1.11)'	(1.49, 1.35)'	(0.38, 0.11)'	(1.59, 1.50)'	(1.59, 1.50)'	(0.59, 0.50)'
β_2	(-1.77, -1.95)'	(3.01, 2.61)'	(-0.77, -0.95)'	(-2.06, -1.99)'	(3.40, 3.50)'	(-1.06, -0.99)'	(-1.72, -1.80)'	(3.00, 2.62)'	(-0.72, -0.80)'
$\mu_1 + \beta_1$	(2.28, -1.72)'	(0.31, 0.32)'	(0.28, 0.28)'	(2.29, -1.69)'	(0.31, 0.27)'	(0.29, 0.31)'	(2.33, -1.61)'	(0.45, 0.39)'	(0.33, 0.39)'
$\mu_2 + \beta_2$	(-2.47, 1.51)'	(0.48, 0.40)'	(-0.47, -0.49)'	(-2.55, 1.48)'	(0.53, 0.58)'	(-0.55, -0.52)'	(-2.39, 1.54)'	(0.45, 0.50)'	(-0.39, -0.48)'
Σ_1	$\begin{bmatrix} 1.94 & 1.55 \\ 1.55 & 1.95 \end{bmatrix}$	$\begin{bmatrix} 0.38 & 0.34 \\ 0.34 & 0.38 \end{bmatrix}$	$\begin{bmatrix} 0.27 & 0.22 \\ 0.22 & 0.28 \end{bmatrix}$	$\begin{bmatrix} 1.90 & 1.51 \\ 1.51 & 1.90 \end{bmatrix}$	$\begin{bmatrix} 0.36 & 0.30 \\ 0.30 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 0.23 & 0.18 \\ 0.18 & 0.23 \end{bmatrix}$	$\begin{bmatrix} 1.91 & 1.51 \\ 1.51 & 1.93 \end{bmatrix}$	$\begin{bmatrix} 0.44 & 0.33 \\ 0.33 & 0.43 \end{bmatrix}$	$\begin{bmatrix} 0.24 & 0.18 \\ 0.18 & 0.26 \end{bmatrix}$
Σ_2	$\begin{bmatrix} 4.28 & 3.41 \\ 3.41 & 4.28 \end{bmatrix}$	$\begin{bmatrix} 0.72 & 0.62 \\ 0.62 & 0.69 \end{bmatrix}$	$\begin{bmatrix} 0.95 & 0.74 \\ 0.74 & 0.95 \end{bmatrix}$	$\begin{bmatrix} 4.30 & 3.43 \\ 3.43 & 4.33 \end{bmatrix}$	$\begin{bmatrix} 0.67 & 0.60 \\ 0.60 & 0.70 \end{bmatrix}$	$\begin{bmatrix} 0.74 & 0.76 \\ 0.76 & 0.77 \end{bmatrix}$	$\begin{bmatrix} 4.42 & 3.55 \\ 3.55 & 4.50 \end{bmatrix}$	$\begin{bmatrix} 0.73 & 0.68 \\ 0.68 & 0.75 \end{bmatrix}$	$\begin{bmatrix} 1.09 & 0.88 \\ 0.88 & 1.17 \end{bmatrix}$
λ_1	-2.51	1.38	-2.01	-2.47	1.67	-1.98	-3.14	1.77	-2.64
λ_2	3.12	1.76	2.12	3.24	1.43	2.24	2.52	1.44	1.52
π_1	0.50	0.01	0.00	0.50	0.01	0.00	0.50	0.01	0.00
π_2	0.50	0.01	0.00	0.50	0.01	0.00	0.50	0.01	0.00
ARI	0.93	0.02	0.00	0.93	0.02	0.00	0.93	0.09	0.00
$r = 0.30$									
	Pattern 1			Pattern 2			MCAR		
	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias
μ_1	(0.65, -3.30)'	(1.20, 1.22)'	(-0.35, -0.30)'	(0.53, -3.44)'	(1.23, 1.08)'	(-0.47, -0.44)'	(0.39, -3.49)'	(1.48, 1.53)'	(-0.61, -0.49)'
μ_2	(-0.58, 3.27)'	(1.96, 2.17)'	(0.42, 0.27)'	(-0.00, 4.09)'	(2.62, 2.17)'	(1.1, 0.9)'	(-0.81, 3.25)'	(1.89, 1.95)'	(0.19, 0.25)'
β_1	(1.72, 1.65)'	(1.52, 1.55)'	(0.72, 0.65)'	(1.82, 1.78)'	(1.49, 1.35)'	(0.82, 0.78)'	(1.95, 1.91)'	(1.81, 1.79)'	(0.95, 0.91)'
β_2	(-1.89, -1.72)'	(2.35, 2.57)'	(-0.89, -0.72)'	(-2.63, -2.75)'	(3.10, 2.56)'	(-1.63, -1.75)'	(-1.62, -1.78)'	(2.26, 2.22)'	(-0.62, -0.78)'
$\mu_1 + \beta_1$	(2.37, -1.65)'	(0.41, 0.42)'	(0.37, 0.35)'	(2.36, -1.65)'	(0.32, 0.32)'	(0.36, 0.35)'	(2.35, -1.58)'	(0.52, 0.47)'	(0.35, 0.42)'
$\mu_2 + \beta_2$	(-2.47, 1.55)'	(0.45, 0.45)'	(-0.47, -0.45)'	(-2.63, 1.34)'	(0.51, 0.43)'	(-0.65, -0.66)'	(-2.42, 1.47)'	(0.59, 0.50)'	(-0.42, -0.53)'
Σ_1	$\begin{bmatrix} 2.00 & 1.60 \\ 1.60 & 2.00 \end{bmatrix}$	$\begin{bmatrix} 0.41 & 0.35 \\ 0.35 & 0.38 \end{bmatrix}$	$\begin{bmatrix} 0.33 & 0.27 \\ 0.27 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 1.90 & 1.51 \\ 1.51 & 1.90 \end{bmatrix}$	$\begin{bmatrix} 0.34 & 0.28 \\ 0.28 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 0.23 & 0.18 \\ 0.18 & 0.23 \end{bmatrix}$	$\begin{bmatrix} 2.00 & 1.60 \\ 1.60 & 1.98 \end{bmatrix}$	$\begin{bmatrix} 0.56 & 0.48 \\ 0.48 & 0.53 \end{bmatrix}$	$\begin{bmatrix} 0.33 & 0.27 \\ 0.27 & 0.31 \end{bmatrix}$
Σ_2	$\begin{bmatrix} 4.44 & 3.56 \\ 3.56 & 4.45 \end{bmatrix}$	$\begin{bmatrix} 0.79 & 0.70 \\ 0.70 & 0.76 \end{bmatrix}$	$\begin{bmatrix} 1.11 & 0.89 \\ 0.89 & 1.12 \end{bmatrix}$	$\begin{bmatrix} 4.33 & 3.45 \\ 3.45 & 4.33 \end{bmatrix}$	$\begin{bmatrix} 0.76 & 0.66 \\ 0.66 & 0.76 \end{bmatrix}$	$\begin{bmatrix} 1.00 & 0.78 \\ 0.78 & 1.00 \end{bmatrix}$	$\begin{bmatrix} 4.37 & 3.49 \\ 3.49 & 4.32 \end{bmatrix}$	$\begin{bmatrix} 0.96 & 0.85 \\ 0.85 & 0.88 \end{bmatrix}$	$\begin{bmatrix} 1.04 & 0.82 \\ 0.82 & 0.99 \end{bmatrix}$
λ_1	-2.26	1.34	-1.76	-2.73	1.70	-2.23	-2.61	1.90	-2.11
λ_2	2.83	1.57	1.83	2.74	1.40	1.74	2.57	1.72	1.37
π_1	0.50	0.01	0.00	0.50	0.01	0.00	0.50	0.01	0.00
π_2	0.50	0.01	0.00	0.50	0.01	0.00	0.50	0.01	0.00
ARI	0.90	0.03	0.00	0.90	0.03	0.00	0.90	0.09	0.00

Table 9: Key model parameters as well as means, standard deviations and bias of the associated parameter estimations from the 100 runs for the first simulation experiment.

Sim3 using MST									
$r = 0.05$									
	Pattern 1			Pattern 2			MCAR		
	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias
μ_1	(1.05, -3.18)'	(0.47, 0.36)'	(0.05, -0.18)'	(0.98, -3.12)'	(0.48, 0.35)'	(-0.02, -0.12)'	(1.00, -3.14)'	(0.47, 0.36)'	(0.00, -0.14)'
μ_2	(-0.80, 3.13)'	(0.58, 0.38)'	(0.20, 0.13)'	(-0.78, 3.22)'	(0.58, 0.40)'	(0.22, 0.22)'	(-0.77, 3.21)'	(0.58, 0.43)'	(0.23, 0.21)'
β_1	(0.93, 1.25)'	(0.48, 0.39)'	(-0.07, 0.25)'	(0.95, 1.23)'	(0.45, 0.35)'	(-0.05, 0.23)'	(0.96, 1.20)'	(0.45, 0.35)'	(-0.04, 0.20)'
β_2	(-1.06, -1.18)'	(0.56, 0.38)'	(-0.06, -0.18)'	(-1.16, -1.19)'	(0.55, 0.42)'	(-0.16, -0.19)'	(-1.13, -1.26)'	(0.60, 0.47)'	(-0.13, -0.26)'
$\mu_1 + \beta_1$	(1.98, -1.93)'	(0.16, 0.08)'	(-0.02, 0.07)'	(1.93, -1.89)'	(0.12, 0.07)'	(-0.07, 0.11)'	(1.96, -1.93)'	(0.12, 0.08)'	(-0.04, 0.07)'
$\mu_2 + \beta_2$	(-1.87, 1.94)'	(0.23, 0.10)'	(0.13, -0.06)'	(-1.94, 2.03)'	(0.22, 0.10)'	(0.06, 0.03)'	(-1.90, 1.95)'	(0.26, 0.11)'	(0.10, -0.05)'
Σ_1	$\begin{bmatrix} 3.36 & 0 \\ 0 & 0.34 \end{bmatrix}$	$\begin{bmatrix} 0.46 & 0 \\ 0 & 0.08 \end{bmatrix}$	$\begin{bmatrix} 0.36 & 0 \\ 0 & 0.01 \end{bmatrix}$	$\begin{bmatrix} 3.33 & 0 \\ 0 & 0.34 \end{bmatrix}$	$\begin{bmatrix} 0.42 & 0 \\ 0 & 0.08 \end{bmatrix}$	$\begin{bmatrix} 0.33 & 0 \\ 0 & 0.01 \end{bmatrix}$	$\begin{bmatrix} 3.32 & 0 \\ 0 & 0.35 \end{bmatrix}$	$\begin{bmatrix} 0.52 & 0 \\ 0 & 0.09 \end{bmatrix}$	$\begin{bmatrix} 0.32 & 0 \\ 0 & 0.02 \end{bmatrix}$
Σ_2	$\begin{bmatrix} 6.57 & 0 \\ 0 & 0.66 \end{bmatrix}$	$\begin{bmatrix} 1.02 & 0 \\ 0 & 0.14 \end{bmatrix}$	$\begin{bmatrix} 0.57 & 0 \\ 0 & -0.01 \end{bmatrix}$	$\begin{bmatrix} 6.58 & 0 \\ 0 & 0.67 \end{bmatrix}$	$\begin{bmatrix} 1.16 & 0 \\ 0 & 0.15 \end{bmatrix}$	$\begin{bmatrix} 0.58 & 0 \\ 0 & 0.00 \end{bmatrix}$	$\begin{bmatrix} 6.51 & 0 \\ 0 & 0.67 \end{bmatrix}$	$\begin{bmatrix} 1.14 & 0 \\ 0 & 0.15 \end{bmatrix}$	$\begin{bmatrix} 0.51 & 0 \\ 0 & 0.00 \end{bmatrix}$
ν_1	8.25	3.18	1.25	8.14	3.01	1.14	8.00	2.77	1.00
ν_2	5.89	1.98	0.89	5.79	1.40	0.79	6.35	2.48	1.35
π_1	0.50	0.02	0.00	0.50	0.01	0.00	0.50	0.02	0.00
π_2	0.50	0.02	0.00	0.50	0.01	0.00	0.50	0.02	0.00
ARI	0.81	0.03	0.00	0.96	0.02	0.00	0.81	0.09	0.00
$r = 0.15$									
	Pattern 1			Pattern 2			MCAR		
	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias
μ_1	(0.98, -3.26)'	(0.65, 0.43)'	(-0.02, -0.26)'	(0.96, -3.26)'	(0.59, 0.43)'	(-0.04, -0.26)'	(0.84, -3.26)'	(0.59, 0.52)'	(-0.16, -0.26)'
μ_2	(-0.76, 3.22)'	(0.50, 0.42)'	(0.24, 0.22)'	(-0.89, 3.21)'	(0.47, 0.37)'	(0.11, 0.21)'	(-0.73, 3.25)'	(0.64, 0.60)'	(0.27, 0.25)'
β_1	(0.99, 1.33)'	(0.66, 0.45)'	(-0.01, 0.33)'	(0.98, 1.32)'	(0.55, 0.44)'	(-0.02, 0.32)'	(1.08, 1.34)'	(0.89, 0.60)'	(0.08, 0.34)'
β_2	(-1.09, -1.28)'	(0.49, 0.44)'	(-0.09, -0.28)'	(-1.02, -1.27)'	(0.48, 0.38)'	(-0.02, -0.27)'	(-1.12, -1.29)'	(0.46, 0.52)'	(-0.12, -0.29)'
$\mu_1 + \beta_1$	(1.98, -1.93)'	(0.14, 0.08)'	(-0.02, 0.07)'	(1.94, -1.94)'	(0.15, 0.08)'	(-0.06, 0.06)'	(1.92, -1.92)'	(0.25, 0.20)'	(-0.08, 0.08)'
$\mu_2 + \beta_2$	(-1.86, 1.94)'	(0.22, 0.10)'	(0.14, -0.06)'	(-1.91, 1.93)'	(0.22, 0.11)'	(0.09, 0.07)'	(-1.86, 1.95)'	(0.32, 0.22)'	(0.14, -0.05)'
Σ_1	$\begin{bmatrix} 3.35 & 0 \\ 0 & 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.51 & 0 \\ 0 & 0.08 \end{bmatrix}$	$\begin{bmatrix} 0.35 & 0 \\ 0 & -0.01 \end{bmatrix}$	$\begin{bmatrix} 3.38 & 0 \\ 0 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 0.55 & 0 \\ 0 & 0.01 \end{bmatrix}$	$\begin{bmatrix} 0.38 & 0 \\ 0 & 0.00 \end{bmatrix}$	$\begin{bmatrix} 3.36 & 0 \\ 0 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 0.61 & 0 \\ 0 & 0.09 \end{bmatrix}$	$\begin{bmatrix} 0.36 & 0 \\ 0 & 0.00 \end{bmatrix}$
Σ_2	$\begin{bmatrix} 6.84 & 0 \\ 0 & 0.65 \end{bmatrix}$	$\begin{bmatrix} 1.26 & 0 \\ 0 & 0.15 \end{bmatrix}$	$\begin{bmatrix} 0.84 & 0 \\ 0 & -0.02 \end{bmatrix}$	$\begin{bmatrix} 6.79 & 0 \\ 0 & 0.65 \end{bmatrix}$	$\begin{bmatrix} 1.13 & 0 \\ 0 & 0.17 \end{bmatrix}$	$\begin{bmatrix} 0.79 & 0 \\ 0 & -0.02 \end{bmatrix}$	$\begin{bmatrix} 6.55 & 0 \\ 0 & 0.63 \end{bmatrix}$	$\begin{bmatrix} 1.44 & 0 \\ 0 & 0.17 \end{bmatrix}$	$\begin{bmatrix} 0.55 & 0 \\ 0 & -0.04 \end{bmatrix}$
ν_1	8.65	4.21	1.65	9.27	5.90	2.27	9.37	6.46	2.37
ν_2	6.35	1.91	1.35	6.13	1.69	1.13	6.57	2.88	1.57
π_1	0.50	0.01	0.00	0.50	0.01	0.00	0.50	0.02	0.00
π_2	0.50	0.01	0.00	0.50	0.01	0.00	0.50	0.02	0.00
ARI	0.78	0.04	0.00	0.78	0.04	0.00	0.74	0.04	0.00
$r = 0.30$									
	Pattern 1			Pattern 2			MCAR		
	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias	Mean	Std. dev.	Bias
μ_1	(0.95, -3.30)'	(0.53, 0.47)'	(-0.05, -0.30)'	(0.96, -3.20)'	(0.52, 0.42)'	(-0.04, -0.20)'	(0.74, -3.21)'	(0.50, 0.47)'	(-0.26, -0.21)'
μ_2	(-0.78, 3.27)'	(0.58, 0.41)'	(0.22, 0.27)'	(-0.72, 3.22)'	(0.60, 0.38)'	(0.28, 0.22)'	(-0.45, 3.24)'	(1.18, 0.51)'	(0.55, 0.24)'
β_1	(1.05, 1.27)'	(0.51, 0.47)'	(0.05, 0.27)'	(0.96, 1.25)'	(0.50, 0.43)'	(-0.04, 0.25)'	(1.10, 1.29)'	(0.93, 0.49)'	(0.10, 0.29)'
β_2	(-1.19, -1.27)'	(1.35, 0.57)'	(-0.19, -0.27)'	(-1.16, -1.30)'	(0.55, 0.43)'	(-0.16, -0.30)'	(-1.31, -1.32)'	(1.18, 0.54)'	(-0.31, -0.32)'
$\mu_1 + \beta_1$	(2.00, -2.03)'	(0.12, 0.09)'	(0.00, -0.03)'	(1.92, -1.94)'	(0.17, 0.09)'	(-0.08, 0.06)'	(1.84, -1.92)'	(0.19, 0.10)'	(-0.16, 0.08)'
$\mu_2 + \beta_2$	(-1.97, 2.00)'	(0.25, 0.11)'	(0.03, 0.00)'	(-1.87, 1.92)'	(0.25, 0.13)'	(0.13, -0.08)'	(-1.76, 1.92)'	(0.29, 0.14)'	(0.24, -0.08)'
Σ_1	$\begin{bmatrix} 3.26 & 0 \\ 0 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 0.54 & 0 \\ 0 & 0.13 \end{bmatrix}$	$\begin{bmatrix} 0.26 & 0 \\ 0 & 0.00 \end{bmatrix}$	$\begin{bmatrix} 3.26 & 0 \\ 0 & 0.33 \end{bmatrix}$	$\begin{bmatrix} 0.52 & 0 \\ 0 & 0.10 \end{bmatrix}$	$\begin{bmatrix} 0.26 & 0 \\ 0 & 0.00 \end{bmatrix}$	$\begin{bmatrix} 3.24 & 0 \\ 0 & 0.36 \end{bmatrix}$	$\begin{bmatrix} 0.61 & 0 \\ 0 & 0.14 \end{bmatrix}$	$\begin{bmatrix} 0.24 & 0 \\ 0 & 0.03 \end{bmatrix}$
Σ_2	$\begin{bmatrix} 6.53 & 0 \\ 0 & 0.67 \end{bmatrix}$	$\begin{bmatrix} 1.53 & 0 \\ 0 & 0.20 \end{bmatrix}$	$\begin{bmatrix} 0.53 & 0 \\ 0 & 0.00 \end{bmatrix}$	$\begin{bmatrix} 6.65 & 0 \\ 0 & 0.67 \end{bmatrix}$	$\begin{bmatrix} 1.23 & 0 \\ 0 & 0.18 \end{bmatrix}$	$\begin{bmatrix} 0.65 & 0 \\ 0 & 0.00 \end{bmatrix}$	$\begin{bmatrix} 6.35 & 0 \\ 0 & 0.67 \end{bmatrix}$	$\begin{bmatrix} 1.63 & 0 \\ 0 & 0.20 \end{bmatrix}$	$\begin{bmatrix} 0.35 & 0 \\ 0 & 0.00 \end{bmatrix}$
ν_1	8.56	4.12	1.56	8.42	3.42	1.42	9.19	4.71	2.19
ν_2	6.83	2.57	1.83	6.34	2.02	1.34	7.26	4.37	2.26
π_1	0.50	0.01	0.00	0.50	0.02	0.00	0.50	0.02	0.00
π_2	0.50	0.01	0.00	0.50	0.02	0.00	0.50	0.02	0.00
ARI	0.75	0.05	0.00	0.76	0.05	0.00	0.75	0.05	0.00

Table 10: A comparsion of average BIC and ARI between MGHD, MST, and Mt models (replications=100) with $G = 1, \dots, 4$.

		MGHD		MST		Mt	
		BIC	ARI	BIC	ARI	BIC	ARI
Sim1	$r = 0.05$	-1534	0.95	-1644	0.88	-1663	0.75
	$r = 0.15$	-1412	0.87	-1517	0.82	-1559	0.69
	$r = 0.30$	-1230	0.74	-1301	0.69	-1396	0.60
Sim2	$r = 0.05$	-1647	0.73	-1683	0.64	-1823	0.59
	$r = 0.15$	-1435	0.62	-1538	0.52	-1677	0.48
	$r = 0.30$	-1201	0.46	-1266	0.36	-1463	0.36
Sim3	$r = 0.05$	-1667	0.82	-1689	0.76	-1789	0.64
	$r = 0.15$	-1517	0.76	-1502	0.66	-1622	0.63
	$r = 0.30$	-1203	0.70	-1264	0.60	-1410	0.48
Sim4	$r = 0.05$	-1546	0.72	-1608	0.41	-1849	0.33
	$r = 0.15$	-1333	0.60	-1440	0.37	-1727	0.27
	$r = 0.30$	-1142	0.12	-1171	0.23	-1385	0.20
Sim5	$r = 0.05$	-1507	0.94	-1613	0.74	-1619	0.88
	$r = 0.15$	-1366	0.85	-1507	0.66	-1450	0.78
	$r = 0.30$	-1193	0.71	-1340	0.59	-1247	0.64
Sim6	$r = 0.05$	-1356	0.68	-1445	0.40	-1614	0.38
	$r = 0.15$	-1262	0.58	-1389	0.38	-1522	0.35
	$r = 0.30$	-1130	0.40	-1263	0.28	-1385	0.29