Flexible Clustering with a Sparse Mixture of Generalized Hyperbolic Distributions

Alexa A. Sochaniwsky\textsuperscript{1} Michael B.P. Gallaugher

Yang Tang\textsuperscript{1} Paul D. McNicholas\textsuperscript{1}

\textsuperscript{1}Dept. of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada.

\textsuperscript{2}Department of Statistical Science, Baylor University, TX, USA

Abstract

Robust clustering of high-dimensional data is an important topic because clusters in real datasets are often heavy-tailed and/or asymmetric. Traditional approaches to model-based clustering often fail for high dimensional data, e.g., due to the number of free covariance parameters. A parametrization of the component scale matrices for the mixture of generalized hyperbolic distributions is proposed. This parameterization includes a penalty term in the likelihood. An analytically feasible expectation-maximization algorithm is developed by placing a gamma-lasso penalty constraining the concentration matrix. The proposed methodology is investigated through simulation studies and illustrated using two real datasets.

Keywords: Asymmetric clusters, flexible clustering, generalized hyperbolic distributions, GHD-GLS, penalized likelihood, sparse mixture models.

1 Introduction

In recent years, the use of finite mixture distributions to model heterogeneous data has undergone intensive development in numerous fields such as pattern recognition, cluster analysis,
and bioinformatics. Traditionally, Gaussian mixture models dominated the literature; however, when clusters are asymmetric and/or have heavier tails, using Gaussian mixture models tend to overestimate the number of clusters and the result can be clustering results that are not useful in practice (see, e.g., Franczak et al., 2014). Consider the data in Fig. 1 where two asymmetric clusters are generated from a $G = 2$ component generalized hyperbolic distribution (Browne and McNicholas, 2015). Gaussian mixtures are fitted to these data for $G = 1, \ldots, 6$ components and the Bayesian Information criterion (BIC; Schwarz, 1978) selects a $G = 3$ component model. Notably, the Gaussian components cannot be merged to return the correct clusters (Fig. 1). For reasons such as this, recent work on model-based clustering has focused on mixtures of non-elliptical distributions (e.g., Murray et al., 2020; Gallaugher et al., 2022; Dang et al., 2023; Tomarchio et al., 2023; McLaughlin et al., 2024). See McNicholas (2016b) for a review of model-based clustering.

Figure 1: Scatter plots of a two-component mixture of generalized hyperbolic distributions with colour used for true labels (left) and predicted labels from Gaussian model-based clustering (right).

In the case of high dimensional data, traditional methods in the area of asymmetric/flexible model-based clustering and classification can fail due to the number of free covariance parameters and, therefore, alternative techniques are sometimes needed. One such technique is to map the data to a (much) lower dimensional space. Mixtures of factor analyzers models with non-elliptical distributions took off in the last decade, including work on multivariate skew-t distributions (Murray et al., 2014a,b) and generalized hyper-
bolic distributions (Tortora et al., 2016). These methods work well with particular datasets; however, considering that this formulation of the skew-t distribution is a special and limiting case of the generalized hyperbolic distribution, they tend to work well on the same datasets. The notion of a joint generalized hyperbolic distribution, which accounts for cluster-specific subspaces, has also been considered (Tang et al., 2018).

Various parameterizations of the component covariance matrices $\Sigma_1, \ldots, \Sigma_G$ have been considered for dimension reduction in Gaussian mixture models (e.g., Banfield and Raftery, 1993; Celeux and Govaert, 1995; Bouveyron et al., 2007). Krishnamurthy (2011) consider a sparse covariance matrix for Gaussian mixture models by including a penalty term in the likelihood. Herein, the work of Krishnamurthy (2011) is extended by considering a mixture of generalized hyperbolic distributions. The method in Krishnamurthy (2011) also involves a Laplace prior on each element of the concentration matrix; herein, a gamma hyperprior is considered for the hyperparameter.

2 Background

2.1 Model-Based Clustering

Model-based clustering, using a finite mixture model, is a common clustering approach. A $G$-component finite mixture model assumes a random vector $X$ has density

$$f(x \mid \vartheta) = \sum_{g=1}^{G} \pi_g f_g(x \mid \theta_g),$$

where $\vartheta = (\pi_1, \pi_2, \ldots, \pi_G, \theta_1, \theta_2, \ldots, \theta_G)$, $f_g(\cdot)$ is the $g$th component density, and $\pi_g > 0$ is the $g$th mixing proportion such that $\sum_{g=1}^{G} \pi_g = 1$. Typically, the mixture model assumes the component densities to be of the same type for each component $g$, e.g., Gaussian, $t$-distribution, etc. McNicholas (2016a) traces the relationship between clustering and mixture models all the way back to Tiedeman (1955), with the earliest use of a mixture model for clustering presented in Wolfe (1965) who used a Gaussian mixture model. Other early work in the area of Gaussian mixture models can be found in Baum et al. (1970) and Scott and Symons (1971).
The mathematical tractability of the Gaussian mixture model has made it very popular in the literature; however, a Gaussian distribution may not be appropriate in the presence of asymmetry and/or heavy tails. To combat this issue, work in the area of non-Gaussian mixtures has become popular including distributions with parameterization for concentration such as the $t$ distribution \cite{Peel2000, Andrews2011, Andrews2012, Lin2014} and the power exponential distribution \cite{Dang2015}. Additionally, in the area of robust clustering, one could consider a contaminated approach \cite{Punzo2016}, a trimmed likelihood approach \cite{Garcia-Escudero2020}, or the OCLUST algorithm \cite{Clark2024}. There has also been significant work in the area of mixtures of skewed distributions such as the skew-$t$ distribution \cite{Lin2010, Vrbik2012, Vrbik2014, Lee2014, Murray2014a, Murray2014b}, the normal-inverse Gaussian distribution \cite{Karlis2009, O'Hagan2016, Fang2022}, the generalized hyperbolic distribution \cite{Browne2015, Tortora2016, Wei2019, Wei2020}, and the skewed power exponential distribution \cite{Dang2023}.

While there have been some attempts at incorporating shrinkage into model-based clustering \cite[e.g.,][]{Bhattacharya2014, Casa2022}, approaches for high-dimensional data have generally shied away from shrinkage. The work presented herein is a step towards addressing this gap.

### 2.2 Sparse Gaussian Mixture Models

Assume that a sample of random vectors $X_1, \ldots, X_n$ comes from a $p$-dimensional Gaussian population with $G$ subpopulations such that the $g$th component has mean $\mu_g$ and covariance matrix $\Sigma_g$. Krishnamurthy \cite{Krishnamurthy2011} uses a penalized observed log-likelihood of the form

$$L(\vartheta) = \sum_{i=1}^{n} \log \left( \sum_{g=1}^{G} \pi_g \phi(\theta_g) \right) + \sum_{g=1}^{G} \log C_g,$$

where $C_g = \Sigma_g^{-1}$ is the concentration matrix for component $g$ and $\phi(\cdot)$ denotes the multivariate Gaussian density. If it is assumed that $C_{gij} \sim \text{Laplace}(0, 1/\lambda)$, then the penalty term becomes $\sum_{g=1}^{G} \lambda ||C_g||_1$, where $|| \cdot ||_1$ is the sum of the absolute values of the entries of $C_g$. 

4
Parameter estimation for this model requires the use of the graphical lasso method (Fried- 
man et al., 2008) in conjunction with an expectation-maximization (EM) algorithm (Demp- 
ster et al., 1977). The graphical lasso is a method for the maximization of

\[ f(C) = \log \det(C) - \text{tr}(SC) - \rho ||C||_1, \]

where \( S \) is the empirical covariance matrix and \( \rho \) is a tuning parameter.

2.3 Generalized Hyperbolic Distribution

Before introducing the generalized hyperbolic distribution, we briefly discuss the generalized 
inverse Gaussian distribution. A random variable \( W \) follows a generalized inverse Gaussian 
distribution, denoted by GIG(\( a, b, \gamma \)), if its density function can be written as

\[ f(y|a, b, \gamma) = \left(\frac{a}{b}\right) \frac{\gamma}{2} w^{\gamma-1} \exp\left\{ -\frac{ay + b/w}{2} \right\}, \]

where \( a, b \in \mathbb{R}^+ \) and

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

is the modified Bessel function of the third kind with index \( \gamma \in \mathbb{R} \). Expectations of some 
functions of a GIG random variable have a mathematically tractable form, e.g.:

\[ \mathbb{E}(W) = \sqrt{\frac{b}{a}} \frac{K_{\gamma+1}(\sqrt{ab})}{K_\gamma(\sqrt{ab})}, \]  

(1)

\[ \mathbb{E}(1/W) = \sqrt{\frac{a}{b}} \frac{K_{\gamma+1}(\sqrt{ab})}{K_\gamma(\sqrt{ab})} - \frac{2\gamma}{b}, \]  

(2)

\[ \mathbb{E}(\log W) = \log \left( \sqrt{\frac{b}{a}} \right) + \frac{1}{K_\gamma(\sqrt{ab})} \frac{\partial}{\partial \gamma} K_\gamma(\sqrt{ab}). \]  

(3)

In their derivation of the generalized hyperbolic distribution, Brown and McNicholas (2015) 
rely on an alternative parameterization of the GIG with density given by

\[ g(w|\omega, \eta, \gamma) = \frac{(w/\eta)^{\gamma-1}}{2\eta K_\gamma(\omega)} \exp\left\{ -\frac{\omega}{2} \left( \frac{w}{\eta} + \frac{\eta}{w} \right) \right\}, \]

(4)

where \( \omega = \sqrt{ab} \) and \( \eta = \sqrt{b/a} \). For notational clarity, we will denote the parameterization 
given in (4) by I(\( \omega, \eta, \gamma \)).
The generalized hyperbolic distribution in [Browne and McNicholas (2015)] arises as a special case of a variance-mean mixture model. This representation assumes that the \( p \)-dimensional random vector \( X \) can be written as

\[
X = \mu + W\alpha + \sqrt{W}U,
\]

where \( \mu \) is a location parameter, \( \alpha \) is the skewness, \( U \sim \mathcal{N}(0, \Sigma) \) and \( W \sim \mathcal{I}(\omega, 1, \gamma) \). The resulting density of the generalized hyperbolic distribution is

\[
f_{GH}(x|\theta) = \exp \left\{ (x - \mu)'\Sigma^{-1}\alpha \right\} \left[ \frac{\omega + \delta(x, \mu; \Sigma)}{\omega + \alpha'\Sigma^{-1}\alpha} \right]^{p/2} \times K_{\gamma-p/2}(\sqrt{\omega + \alpha'\Sigma^{-1}\alpha}|\omega + \delta(x, \mu; \Sigma)|) \\
(2\pi)^{p/2}|\Sigma|^{1/2}K_{\gamma}(\omega),
\]

where \( \gamma \in \mathbb{R} \) is an index parameter, \( \omega > 0 \) is a concentration parameter, and \( \delta(x, \mu; \Sigma) = (x - \mu)'\Sigma^{-1}(x - \mu) \).

### 3 Methodology

#### 3.1 Overview

To extend the methodology in [Krishnamurthy (2011)]\(^{1}\), assume that \( \lambda_g \sim \text{Gamma}(s, r) \) and \( C_{gij}|\lambda_g \sim \text{Laplace}(0, 1/\lambda_g) \) for each \( g = 1, \ldots, G \), where \( C_g \) is the inverse of a scale matrix \( \Sigma_g \) and \( C_{gij} \) is the random variable corresponding to the element in the \( i \)th row and \( j \)th column of \( C_g \). The joint density of \( C_g \) and \( \lambda_g \) is given by

\[
f(C_g, \lambda_g) = \frac{r^s}{\Gamma(s)} \lambda_g^{s-1} \exp \{-r\lambda_g\} \prod_{i=1}^p \prod_{j=1}^p \frac{\lambda_g}{2} \exp \{-\lambda_g |c_{gij}|\}
\]

\[
= \frac{r^s}{\Gamma(s)2^p} \lambda_g^{s+p^2-1} \exp \{-\lambda_g (r + ||C_g||_1)\},
\]

and it can be shown that \( \lambda_g |c_g \sim \text{gamma}(s+p^2, r + ||C_g||_1) \) and that the marginal distribution of \( C_g \) is

\[
f(C_g) = \frac{r^s}{\Gamma(s)2^p} \left[ \frac{\Gamma(s + p^2)}{(r + ||C_g||_1)^{s+p^2}} \right].
\]

Note that Stirling’s formula ([De Moivre 1730]) is used to estimate the Gamma function in the marginal distribution.
Suppose we observe a random sample $x_1, \ldots, x_n$ from a $G$-component mixture of generalized hyperbolic distributions. Following Krishnamurthy (2011), the observed penalized log-likelihood is

$$L(\theta) = \sum_{i=1}^{N} \log \sum_{g=1}^{G} \pi_g f_{\text{GH}}(x_i | \mu_g, \alpha_g, \Sigma_g, \omega_g, \gamma_g) + \sum_{g=1}^{G} \log f(c_g).$$

### 3.2 Parameter Estimation

Define $z_{ig}$ so that $z_{ig} = 1$ if $x_i$ is in component $g$ and $z_{ig} = 0$ otherwise. Now, the complete-data comprise the $z_{ig}$ together with the latent $w_{ig}$ and the unknown $\lambda_g$ for $i = 1, \ldots, n$ and $g = 1, \ldots, G$, and the complete-data penalized log-likelihood is

$$L_C(\theta) = K + \frac{1}{2} \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log |C_g| + \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log h(w_{ig} | \omega_g, \gamma_g)

- \frac{1}{2} \text{tr} \left\{ \sum_{g=1}^{G} C_g \sum_{i=1}^{N} z_{ig} \left[ (1/w_{ig})(x_i - \mu_g)(x_i - \mu_g)' - (x_i - \mu_g)\alpha_g - \alpha_g(x_i - \mu_g)' + w_{ig} \alpha_g \alpha_g' \right] \right\}

- \sum_{g=1}^{G} \lambda_g ||C_g||_1,$

where $K$ is a constant with respect to the parameters. An EM algorithm is used to maximize the complete-data likelihood and an outline is given below.

**E-step** Update $\hat{z}_{ig}, a_{ig}, b_{ig}, c_{ig}, \hat{\lambda}_g$, where

$$\hat{z}_{ig} := \mathbb{E}[z_{ig} | x_i] = \frac{\hat{\pi}_g^{(t)} f_{\text{GH}}(X_i | \hat{\theta}_g^{(t)})}{\sum_{h=1}^{G} \hat{\pi}_h^{(t)} f_{\text{GH}}(X_i | \hat{\theta}_h^{(t)})},$$

$$a_{ig} := \mathbb{E}[W_{ig} | x_i, z_{ig} = 1], \quad b_{ig} := \mathbb{E}[1/W_{ig} | x_i, z_{ig} = 1], \quad c_{ig} := \mathbb{E}[\log W_{ig} | x_i, z_{ig} = 1],$$

and $\hat{\theta}_g^{(t)} = \{ \hat{\mu}_g^{(t)}, \hat{\alpha}_g^{(t)}, \hat{\omega}_g^{(t)}, \hat{\delta}_g^{(t)}, \hat{\Sigma}_g^{(t)} \}$. Fortunately, in the generalized hyperbolic case,

$$W_{ig} | x_i, z_{ig} = 1 \sim I(\omega_g + \alpha_g' C_g \alpha_g, \omega_g + \delta(x_i, \mu_g | C_g), \gamma_g - p/2),$$

and so $a_{ig}, b_{ig}$ and $c_{ig}$ can be calculated using (1)–(3). The update for $\lambda_g$ is

$$\mathbb{E}[\lambda_g | c_g] = \frac{s + p^2}{||c_g^{(t)}||_1 + r} =: \hat{\lambda}_g.$$

Hereafter, we use the notation $n_g = \sum_{i=1}^{n} \hat{z}_{ig}, \bar{a}_g = 1/n_g \sum_{i=1}^{n} \hat{z}_{ig} a_{ig}, \bar{b}_g = 1/n_g \sum_{i=1}^{n} \hat{z}_{ig} b_{ig}$ and $\bar{c}_g = 1/n_g \sum_{i=1}^{n} \hat{z}_{ig} c_{ig}$.
M-step  Update $\pi_g$, $\mu_g$, $\alpha_g$, $\omega_g$, $\gamma_g$ and $C_g$. The updates for all these parameters, except $C_g$, are identical to those given in Browne and McNicholas (2015) and are given by

$$
\hat{\pi}_g^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ig}, \quad \hat{\mu}_g^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}x_i(\hat{a}_g b_{ig} - 1)}{\sum_{i=1}^{n} \hat{z}_{ig}(\hat{a}_g b_{ig} - 1)}, \quad \hat{\alpha}_g^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}(\hat{b}_g - \hat{b}_{ig})}{\sum_{i=1}^{n} \hat{z}_{ig}(\hat{a}_g b_{ig} - 1)}.
$$

The updates for $\gamma_g$ and $\omega_g$ cannot be obtained in closed form and have to be updated using numerical techniques. The details are given in Browne and McNicholas (2015) and are given by

$$
\hat{\gamma}_g^{(t+1)} = \frac{1}{n} \left( \sum_{i=1}^{n} \hat{z}_{ig} \right) \left[ \frac{\partial}{\partial s} \log(K_{\gamma_g}(\hat{\omega}_g^{(t)})) \right]_{s=\gamma_g^{(t)}}^{-1},
$$

(5)

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

(6)

where the derivative of the Bessel function with respect to the index in (5) is calculated numerically and

$$
q(\gamma_g, \omega_g) = \sum_{i=1}^{N} z_{ig} \left[ \log(K_{\gamma_g}(\omega_g)) - \gamma_g \log w_{ig} - \frac{1}{2} \omega_g \left( w_{ig} + \frac{1}{w_{ig}} \right) \right].
$$

The partial derivatives in (6) are described in Browne and McNicholas (2015) and can be written as

$$
\frac{\partial}{\partial \omega} q(\gamma_g, \omega_g) = \frac{1}{2} \left[ R_{\gamma_g}(\omega_g) + R_{-\gamma_g}(\omega_g) - (\hat{a}_g + \hat{b}_g) \right],
$$

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

where $R_{\gamma_g}(\omega_g) = K_{\gamma_g+1}(\omega_g)/K_{\gamma_g}(\omega_g)$. The update for $\Sigma_g$ is calculated as follows. Using the graphical lasso method, find

$$
\hat{C}_g^{(t+1)} = \arg\min_{C} \left\{ -\log |C| + \text{tr}(CS_g) + \frac{\hat{\lambda}_g}{n_g} ||C||_1 \right\},
$$

where

$$
S_g = \frac{1}{n_g} \left\{ \sum_{i=1}^{N} \hat{z}_{ig} \left[ b_{ig}(x_i - \hat{\mu}_g^{(t+1)})(x_i - \hat{\mu}_g^{(t+1)})' - (x_i - \hat{\mu}_g^{(t+1)})(\hat{\alpha}_g^{(t+1)})' \right. \\
- \left. \hat{\alpha}_g^{(t+1)}(x_i - \hat{\mu}_g^{(t+1)})' + a_{ig} \hat{\alpha}_g^{(t+1)}(\hat{\alpha}_g^{(t+1)})' \right] \right\}.
$$

Then, the update for $\Sigma_g$ is $\hat{\Sigma}_g^{(t+1)} = \hat{C}_g^{(t+1)}$. Despite the use of numerical methods throughout the M-step, the monotonously of the likelihood is preserved. Hereafter, we will refer to this model as the GHD-GLS model.
3.3 Stopping Rule

McNicholas et al. (2010) show that it is possible for the likelihood to “plateau” and then increase again, and so an EM algorithm may be stopped prematurely if lack of progress in the likelihood is used as the stopping rule. An alternative is to use a stopping rule based on the Aitken acceleration Aitken (1926). The Aitken acceleration at iteration $t$ is

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

where $l(t)$ is the observed likelihood at iteration $t$. Böhning et al. (1994) and Lindsay (1995) consider the quantity

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

which is an estimate, at iteration $t + 1$, of the observed log-likelihood after many iterations. As in McNicholas et al. (2010), we terminate the algorithm when $l_{\infty}^{(k+1)} - l^{(k)} \in (0, \epsilon)$, where $\epsilon$ is small and positive.

3.4 Model Selection

In a typical clustering scenario, the number of groups is not known $a priori$ and, therefore, has to be selected using some criterion. Because the lasso penalty term shrinks the elements of the concentration matrices, there will almost certainly be some elements that are 0 and this needs to be considered when selecting the number of groups. One example in the literature that demonstrates an approach for dealing with this is lasso-penalized BIC (LPBIC; Bhattacharya and McNicholas, 2014). The method utilizes a quadratic approximation to the penalty term. However, this cannot be derived here because of the form of $f(C_g)$. Therefore, we propose using the BIC with an effective number of non-zero parameters. This method, however, requires a pre-specified cut-off value for determining which elements can be considered zero. In this paper, we use $10^{-5}$ as this cut-off value.
4 Simulation Studies

4.1 Overview

The performance of the GHD-GHL model is assessed in three ways. Experiment 1 (Section 4.2) investigates the sensitivity of the model to different values of the gamma hyperparameters, i.e., \((s, r)\), as well as the effectiveness of BIC in choosing the correct model. Experiment 2 (Section 4.3) is designed to assess the proposed sparse modelling approach through different dependency patterns among variables for each component. In Experiment 3 (Section 4.4), the proposed model is compared with the parsimonious Gaussian mixture models (PGMM; McNicholas and Murphy, 2008, 2010) from the \texttt{pgmm} package (McNicholas et al., 2023) for the R software (\texttt{R Core Team}, 2023), the mixture of generalized hyperbolic distributions (MGHD), and the mixture of generalized hyperbolic factor analyzers (MGHFA; Tortora et al., 2016) — the R package \texttt{mixGHD} (Tortora et al., 2021) is used to implement the latter two approaches. All methods are initialized using \texttt{k-means}. When the true classes are known, the performance of the GHD-GLS approach can be assessed using the adjusted Rand index (ARI; Hubert and Arabie, 1985). The ARI has expected value 0 under random classification and takes the value 1 under perfect class agreement.

4.2 Experiment 1

A total of 100 samples of each combination of \(n_g\) and \((s, r)\) is generated for \(G = 2\) from a 100 dimension MGHD model (i.e., \(p = 100\)) with \(\pi_1 = \pi_2 = 1/G\). The scale matrices are of the form \(\Sigma_g = \sigma_g^2 I\) to promote sparsity. An example of the simulated data for \(n_g = 500\) is given in Figure 2 where there is some overlap in the clusters. Four combinations of the location and rate parameters for the gamma hyperparameters are considered. Table 1 shows the BIC and ARI values averaged on the 100 samples for each pair \((s, r)\) as well as the number of times that the correct model is favoured by the BIC for each scenario. As shown in Table 1, the clustering results do not vary much for different values of the gamma hyperparameters. The BIC usually selects the correct model, and the ARI increases as the number of observations increases. Based on these results, we use \((s, r) = (1, 1)\) for the remaining analyses herein.
Figure 2: The first five variables for one of the simulated datasets from Experiment 1 with $n_g = 500$.

4.3 Experiment 2

Two scenarios are considered with different structures for $\Sigma_g$ and $p = 20$ dimensions. Fig. 3 presents the heat maps of the true component covariance structures for the two scenarios. In each scenario, a set of 100 samples for each $n_g \in \{250, 500, 1000\}$ is generated from a three-component MGHD with the corresponding covariance structures. The GHD-GLS models are fitted for $G = 3$ with $k$-means starts. Figs. 4 and 5 show the averaged estimated component covariance matrices for each sample size in Scenarios 1 and 2, respectively. Overall, the GHD-GLS models show promising performance in recovering the underlying structures of
Table 1: Values of \(n_g\) and \((s, r)\) used in Experiment 1 together with a summary of the results from applying GHD-GLS, i.e., the percent of runs where the BIC selected \(G = 2\) as well as average BIC and ARI values.

| \((s, r)\)     | BIC\% | Avg. BIC      | Avg. ARI |
|----------------|-------|---------------|----------|
| \(n_g = 500\) | (1, 0.5) | 85 | -559763.1 | 0.9954 |
|                | (0.5, 0.5) | 85 | -559774.0 | 0.9954 |
|                | (1, 1)    | 88 | -582372.2 | 0.9953 |
|                | (0.5, 1)  | 87 | -582374.3 | 0.9953 |
| \(n_g = 750\) | (1, 0.5) | 62 | -849145.1 | 0.9959 |
|                | (0.5, 0.5) | 59 | -849144.4 | 0.9959 |
|                | (1, 1)    | 79 | -859274.5 | 0.9953 |
|                | (0.5, 1)  | 83 | -859274.1 | 0.9953 |
| \(n_g = 1000\)| (1, 0.5) | 100 | -1122732 | 0.9942 |
|                | (0.5, 0.5) | 100 | -1122737 | 0.9941 |
|                | (1, 1)    | 100 | -1127642 | 0.9936 |
|                | (0.5, 1)  | 100 | -1127641 | 0.9936 |

the component covariance matrices in both scenarios. As one would expect, the estimation becomes more accurate as the sample size grows.

### 4.4 Experiment 3

In this experiment, we compare our proposed approach with three comparator approaches: PGMM, MGHD, and MGHFA. PGMM is developed for high-dimensional symmetrical data whereas the MGHD and MGHFA approaches can model clusters with skewness and/or heavy tails. Two scenarios are considered: one where 100 samples are generated from the MGHD (Scenario 3) and another where 100 samples are generated from a mixture of Gaussian distributions (Scenario 4). Each combination of \(n_g \in \{500, 750, 1000\}\) and \(p \in \{20, 50, 100\}\) is considered for a three-component mixture model with the covariance structure in Fig 3(a).

Tables 2 and 3 show the averaged BIC and ARI over the 100 samples in Scenarios 3
Figure 3: Heat maps of the true component covariance matrices in the two scenarios from Experiment 2.

and 4, respectively. Tables 8 and 9 (Appendix A) show the average and standard deviation of the times that correspond to the results in Table 2 and 3, respectively. In Scenario 3, GHD-GLS performs the best overall but MGHD is the fastest. PGMM and MGHFA are effective for dimension reduction; however, the classification results are significantly worse than for GHD-GLS. In Scenario 4, unsurprisingly, PGMM performs the best among all four approaches because the samples are generated from Gaussian mixtures.
Figure 4: Heat maps of the estimated component covariance matrices in Scenario 1 for each sample size.

(a) $n_g = 250$

(b) $n_g = 500$

(c) $n_g = 1000$
Figure 5: Heat maps of the estimated component covariance matrices in Scenario 2 for each sample size.
Table 2: A comparison of the selected GHD-GLS, PGMM, MGHD and MGHFA in Scenario 3 (replications=100).

|       | GHD-GLS |         | PGMM |         | MGHD |         | MGHFA |         |
|-------|---------|---------|------|---------|------|---------|-------|---------|
|       | BIC     | ARI     | BIC  | ARI     | BIC  | ARI     | BIC   | ARI     |
| n_g=500 | -190781.9 | 0.870 | -186772.5 | 0.558 | -187942.0 | 0.867 | -186891.0 | 0.617 |
| p=20 | n_g=750 | -283592.8 | 0.885 | -279930.3 | 0.558 | -280032.0 | 0.883 | -279634.1 | 0.621 |
|       | n_g=1000 | -376203.6 | 0.890 | -373013.5 | 0.574 | -372002.3 | 0.888 | -372295.0 | 0.622 |
| n_g=500 | -499851.4 | 0.926 | -471148.5 | 0.551 | -487045.2 | 0.915 | -472169.9 | 0.534 |
| p=50 | n_g=750 | -731673.2 | 0.939 | -706102.2 | 0.550 | -719751.8 | 0.934 | -706460.8 | 0.526 |
|       | n_g=1000 | -964276.9 | 0.950 | -941169.3 | 0.550 | -952085.3 | 0.947 | -940735.6 | 0.516 |
| n_g=500 | -1095283 | 0.952 | -938231.1 | 0.530 | -1017441 | 0.859 | -940535.7 | 0.625 |
| p=100 | n_g=750 | -1572922 | 0.973 | -1406088 | 0.530 | -1483629 | 0.944 | -1407180 | 0.604 |
|       | n_g=1000 | -2045354 | 0.985 | -1873921 | 0.531 | -1948150 | 0.972 | -1873840 | 0.609 |

Table 3: A comparison of the selected GHD-GLS, PGMM, MGHD and MGHFA in Scenario 4 (replications=100).

|       | GHD-GLS |         | PGMM |         | MGHD |         | MGHFA |         |
|-------|---------|---------|------|---------|------|---------|-------|---------|
|       | BIC     | ARI     | BIC  | ARI     | BIC  | ARI     | BIC   | ARI     |
| n_g=500 | -170483.1 | 0.851 | -164229.7 | 1 | -167143.2 | 0.852 | -165129.1 | 0.994 |
| p=20 | n_g=750 | -253281.6 | 0.823 | -245958.0 | 1 | -249065.9 | 0.825 | -246998.7 | 0.990 |
|       | n_g=1000 | -335039.2 | 0.867 | -327611.4 | 1 | -330268.1 | 0.869 | -328582.3 | 1 |
| n_g=500 | -450749.6 | 0.830 | -417357.2 | 1 | -437503.0 | 0.817 | -419967.3 | 0.990 |
| p=50 | n_g=750 | -664606.7 | 0.782 | -625218.5 | 1 | -646100.1 | 0.772 | -628190.7 | 0.978 |
|       | n_g=1000 | -876314.8 | 0.757 | -833059.7 | 1 | -854186.0 | 0.750 | -835963.5 | 0.989 |
| n_g=500 | -982765.7 | 0.819 | -832882.6 | 1 | -919609.7 | 0.789 | -837742.0 | 0.995 |
| p=100 | n_g=750 | -1449128 | 0.829 | -1247655 | 1 | -1337731 | 0.817 | -1253036 | 1 |
|       | n_g=1000 | -1882055 | 0.806 | -1662259 | 1 | -1754512 | 0.795 | -1667926 | 1 |

16
5 Real Data Analyses

5.1 Overview

The proposed GHD-GLS approach is compared with two methods based on generalized hyperbolic distributions, i.e., MGHD and MGHFA, on real data. The Movehub quality of life data (Section 5.2) is selected to demonstrate the interpretability of our proposed approach, and the breast cancer diagnostic data (Section 5.3) is used because of its popularity as a benchmark dataset within the literature.

5.2 Movehub Quality of Life

The Movehub quality of life data consist of five key metrics for 216 cities (i.e., \( n = 216 \) and \( p = 5 \)): purchase power, health care, pollution, quality of life, and crime rate. An overall rating for a city is given considering all five metrics. The data are available online at [www.movehub.com](http://www.movehub.com). The GHD-GLS, MGHD, and MGHFA approaches are fitted to these data for \( G = 2, 3, 4 \) and, for MGHFA, for \( q = 1, 2, 3 \). The minimum BIC occurs at \( G = 2 \) for MGHD and MGHFA, and at \( G = 3 \) for GHD-GLS. A cross-tabulation of the predicted classifications against true classes for the three approaches is shown in Table 4. Table 5 presents the mean and standard deviation of the five metrics as well as the overall rating for each group.

| GHD-GLS | MGHD | MGHFA |
|---------|------|-------|
|         | Group 1 | Group 2 | Group 1 | Group 2 |
| Group 1 | 84     | 4      | 77      | 11      |
| Group 2 | 1      | 83     | 0       | 84      |
| Group 3 | 44     | 0      | 34      | 10      |

Group 1 consists of cities with lower ratings in purchase power, health care and quality of life, while having higher ratings in pollution and crime rate. The overall rating for the
Table 5: Mean and standard deviation of the five key metrics for each component of the three approaches.

| Group 1 | GHD-GLS | MGHD | MGHFA |
|---------|---------|------|-------|
|         | Mean    | Std. dev. | Mean | Std. dev. | Mean | Std. dev. |
| Overall Rating | 83.45 | 4.01 | 83.90 | 3.92 | 83.86 | 3.50 |
| Purchase Power | 58.14 | 12.51 | 60.90 | 12.08 | 62.70 | 11.92 |
| Health Care | 71.75 | 11.07 | 70.94 | 11.36 | 71.85 | 10.14 |
| Pollution | 23.07 | 10.10 | 38.48 | 23.89 | 36.57 | 23.20 |
| Quality of Life | 76.13 | 9.61 | 74.85 | 11.09 | 77.57 | 8.47 |
| Crime Rate | 35.21 | 14.37 | 38.55 | 16.46 | 36.53 | 15.37 |

| Group 2 | GHD-GLS | MGHD | MGHFA |
|---------|---------|------|-------|
|         | Mean    | Std. dev. | Mean | Std. dev. | Mean | Std. dev. |
| Overall Rating | 73.39 | 4.20 | 73.41 | 4.04 | 75.25 | 6.00 |
| Purchase Power | 24.95 | 8.67 | 25.10 | 8.42 | 29.33 | 12.26 |
| Health Care | 59.52 | 16.10 | 59.76 | 15.88 | 60.72 | 16.02 |
| Pollution | 56.82 | 23.87 | 55.26 | 24.28 | 54.40 | 24.42 |
| Quality of Life | 37.05 | 13.81 | 37.97 | 14.41 | 41.41 | 15.75 |
| Crime Rate | 45.88 | 15.51 | 45.48 | 15.54 | 46.43 | 16.01 |

| Group 3 | GHD-GLS | MGHD | MGHFA |
|---------|---------|------|-------|
|         | Mean    | Std. dev. | Mean | Std. dev. | Mean | Std. dev. |
| Overall Rating | 84.13 | 4.27 | NA | NA | NA | NA |
| Purchase Power | 64.25 | 12.53 | NA | NA | NA | NA |
| Health Care | 69.03 | 11.53 | NA | NA | NA | NA |
| Pollution | 67.47 | 11.76 | NA | NA | NA | NA |
| Quality of Life | 71.52 | 13.12 | NA | NA | NA | NA |
| Crime Rate | 44.93 | 18.31 | NA | NA | NA | NA |
cities in Group 1 is lower when compared to cities in Groups 2 and 3. Fig. 6 shows the sparse correlation structures among the five variables differ across groups which can only be found using the GHD-GLS approach. In particular, Group 1 is characterized by the relation between quality of life and purchase power, Group 2 is characterized by the relation between quality of life, crime rate, health care and purchase power, and Group 3 is characterized by the relation between quality of life and crime rate.

The predicted classification from our approach agrees with MGHD on 167 cities and disagrees on 49 cities. Out of those 49 cities, GHD-GLS distinguishes 44 of those cities as a third group. Similarly, the predicted classification from our approach agrees with MGHFA on 161 cities and disagrees on 55 cities where GHD-GLS distinguishes 34 of those cities as a third group. The remaining 21 cities are placed into Group 2 where they belong to Group 1 or 3 using GHD-GLS.

5.3 Breast Cancer Diagnostic Dataset

The breast cancer diagnostic data was first used in Street et al. (1993). Ten real-valued features on 569 cases of breast tumours are reported: 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. The GHD-GLS, MGHD and MGHFA methods are fitted for $G = 2, 3, 4$ and, for MGHFA, for $q = 1, 2, 3$. A summary of the best models from the GHD-GLS, MGHD and MGHFA approaches is given in Table 6. The GHD-GLS and MGHD approaches give the correct number of components, i.e., $G = 2$. The GHD-GLS approach yields the best ARI among the three approaches (ARI = 0.77) and misclassifies only 34 out of 569 observations (Table 7).

6 Discussion

The GHD-GLS approach for flexible clustering of high-dimensional data was developed based on a mixture of generalized hyperbolic distributions with a penalty term in the likelihood constraining the component-specific concentration matrices. This allows the association structure of the variables to vary across the mixture components. The gamma-lasso penalty
Figure 6: The correlation structures among the five variables for each group for the quality of life data using the GHD-GLS approach.

used herein enabled the development of an analytically feasible EM algorithm. The BIC with effective number of non-zero parameters was used for model selection. Three simulation studies were carried out to illustrate the proposed GHD-GLS approach and compare with PGMM, MGHD and MGHFA. The GHD-GLS approach was also applied to two real datasets and its performance was compared to the MGHD and MGHFA approaches. In the case of the Movehub dataset, the dimensionality was low and the GHD-GLS approach identified three groups, splitting apart the first group that the MGHD approach discovered. Moreover, the
Table 6: A comparison of the selected GHD-GLS and two different approaches on the tumour data.

|      | G   | BIC      | ARI |
|------|-----|----------|-----|
| GHD-GLS | 2  | −36997.49 | 0.77 |
| MGHD  | 2  | −1523.13  | 0.66 |
| MGHFA | 4  | −25325.46 | 0.31 |

Table 7: Cross-tabulation of true versus predicted (A,B) classifications from the selected GHD-GLS for the tumour data (ARI = 0.77).

|      | A   | B   |
|------|-----|-----|
| Malignant | 352 | 5   |
| Benign   | 29  | 183 |

GHD-GLS approach was able to find different sparse correlation structures among variables, leading to a simpler interpretation of the clustering results. When fitted to the breast cancer diagnostic data, which is often used for benchmarking, the GHD-GLS approach gave superior classification performance when compared to the chosen MGHD and MGHFA. Future work will consider extension of the GHD-GLS approach to the matrix-variate paradigm, where it will be interesting to compare its performance to other approaches (e.g. Gallaugher and McNicholas, 2020).

Acknowledgements

This work was supported by a Vanier Canada Graduate Scholarship, the Canada Research Chairs program, an E.W.R. Steacie Memorial Fellowship, and a Dorothy Killam Fellowship.
References

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

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: The tEIGEN family. *Statistics and Computing* 22(5), 1021–1029.

Andrews, J. L., P. D. McNicholas, and S. Subedi (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics and Data Analysis* 55(1), 520–529.

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

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

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

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, 373–388.

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

Browne, R. P. and P. D. McNicholas (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics* 43(2), 176–198.
Casa, A., A. Cappozzo, and M. Fop (2022). Group-wise shrinkage estimation in penalized model-based clustering. *Journal of Classification* 39(3), 648–674.

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

Clark, K.M. and P.D. McNicholas (2024). Finding outliers in Gaussian model-based clustering. *Journal of Classification* 41(3).

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

Dang, U. J., M. P. B. Gallaugher, R. P. Browne, and P. D. McNicholas (2023). Model-based clustering and classification using mixtures of multivariate skewed power exponential distributions. *Journal of Classification* 40(1), 145–167.

De Moivre, A. (1730). *Miscellanea Analytica*. London: Tonson and Watts.

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

Fang, Y., D. Karlis, and S. Subedi (2022). Infinite mixtures of multivariate normal-inverse Gaussian distributions for clustering of skewed data. *Journal of Classification* 39(3), 510–552.

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.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3), 432–441.

Gallaugher, M. P. B. and P. D. McNicholas (2020). Mixtures of skewed matrix variate bilinear factor analyzers. *Advances in Data Analysis and Classification* 14(2), 415–434.
Gallaugher, M. P. B., S. D. Tomarchio, P. D. McNicholas, and A. Punzo (2022). Model-based clustering via skewed matrix-variate cluster-weighted models. *Journal of Statistical Computation and Simulation 31*(2), 413–421.

Garcia-Escudero, L. A, A. Mayo-Iscar, and M. Riani (2020). Model-based clustering with determinant-and-shape constraint. *Statistics and Computing 30*, 1363–1380.

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

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

Krishnamurthy, A. (2011). High-dimensional clustering with sparse Gaussian mixture models. Unpublished manuscript.

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

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

Lin, T.-I., P. D. McNicholas, and J. H. Hsiu (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.

McLaughlin, P., B. C. Franczak, and A. B. Kashlak (2024). Unsupervised classification with a family of parsimonious contaminated shifted asymmetric Laplace mixtures. *Journal of Classification*. In press.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Boca Raton: Chapman & 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. ElSherbiny, K. R. Jampani, A. F. McDaid, and L. Banks (2023). pgmm: Parsimonious gaussian mixture models. R package version 1.2.7.

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.

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

Murray, P. M., R. P. Browne, and P. D. McNicholas (2020). Mixtures of hidden truncation hyperbolic factor analyzers. *Journal of Classification 37*(2), 366–379.

Murray, P. M., P. D. McNicholas, and R. B. Browne (2014b). A mixture of common skew-t factor analyzers. *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.

Punzo, A. and P.D. McNicholas (2024). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal 58*(6), 1506–1537.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464.

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

Street, N. W., W. H. Wolberg, and O. L. Mangasarian (1993). Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology*, Volume 1905, San Jose, pp. 861—870.

Tang, Y., R. P. Browne and P. D. McNicholas (2018). Flexible clustering of high-dimensional data via mixtures of joint generalized hyperbolic distributions. *Stat* 7(1), e177.

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.

Tomarchio, S. D., L. Bagnato, and A. Punzo (2023). Model-based clustering using a new multivariate skew distribution. *Advances in Data Analysis and Classification*. In press.

Tortora, C., R. P. Browne, A. ElSherbiny, B. C. Franczak, and P. D. McNicholas (2021). Model-based clustering, classification, and discriminant analysis using the generalized hyperbolic distribution: Mixghd r package. *Journal of Statistical Software* 98, 3.

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

Wei, Y., Y. Tang, and P. D. McNicholas (2019). Mixtures of generalized hyperbolic distributions and mixtures of skew-t distributions for model-based clustering with incomplete data. *Computational Statistics and Data Analysis* 130, 18–41.
Wei, Y., Y. Tang, and P. D. McNicholas (2020). Flexible high-dimensional unsupervised learning with missing data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(3), 610–621.

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

## A Timing Comparisons for Scenarios 3 and 4

Table 8: A comparison of the running times, in seconds, of the selected GHD-GLS, PGMM, MGHD and MGHFA in Scenario 3 (replications=100).

|       | GHD-GLS |   |   |   | PGMM |   |   |   | MGHD |   |   |   | MGHFA |   |   |
|-------|---------|---|---|---|------|---|---|---|------|---|---|---|-------|---|---|
|       | Time    | sd| Time| sd| Time| sd| Time| sd| Time| sd| Time| sd|
|       |         |   |     |   |     |   |     |   |     |   |     |   |
| $n_g = 500$ |  |   |  |   |     |   |     |   |     |   |     |   |
| $p = 20$  | 50.60   | 16.23 | 21.07 | 5.16 | 13.80 | 0.57 | 186.17 | 43.75 |
| $n_g = 750$ | 45.91 | 26.00 | 33.47 | 8.37 | 21.79 | 0.30 | 271.46 | 65.61 |
| $n_g = 1000$ | 25.15 | 16.87 | 45.00 | 10.07 | 30.20 | 1.82 | 362.52 | 88.93 |
| $n_g = 500$ | 122.62 | 59.11 | 56.00 | 6.90 | 40.40 | 1.28 | 377.46 | 87.20 |
| $p = 50$  | 307.93 | 106.36 | 89.03 | 9.06 | 62.79 | 2.96 | 554.12 | 139.84 |
| $n_g = 1000$ | 479.83 | 125.80 | 124.27 | 12.15 | 67.83 | 10.54 | 806.58 | 204.25 |
| $n_g = 500$ | 134.63 | 68.53 | 203.64 | 22.62 | 72.16 | 0.57 | 1160.90 | 295.96 |
| $p = 100$  | 300.48 | 72.89 | 324.11 | 26.33 | 105.78 | 0.81 | 1814.17 | 461.93 |
| $n_g = 1000$ | 638.71 | 120.19 | 442.53 | 36.75 | 140.65 | 1.40 | 2800.92 | 811.66 |
Table 9: A comparison of the running times, in seconds, of the selected GHD-GLS, PGMM, MGHD and MGHFA in Scenario 4 (replications=100).

|       | GHD-GLS |     | PGMM |     | MGHD |     | MGHFA |     |
|-------|---------|-----|------|-----|------|-----|-------|-----|
|       | Time    | sd  | Time | sd  | Time | sd  | Time  | sd  |
| n_g = 500 | 65.88  | 42.69| 4.63 | 2.68| 15.23 | 0.12| 101.79 | 3.48|
| p = 20 | n_g = 750 | 73.46 | 59.01| 7.79 | 2.49| 22.29 | 0.11| 151.09 | 4.61|
|       | n_g = 1000 | 62.34| 47.53| 10.70 | 0.62| 29.34 | 0.10| 194.36 | 7.39|
| n_g = 500 | 174.17 | 135.97| 14.60 | 0.75| 32.46 | 0.15| 218.03 | 13.94|
| p = 50 | n_g = 750 | 253.29 | 221.71| 22.98 | 0.42| 47.85 | 0.18| 326.04 | 13.91|
|       | n_g = 1000 | 353.61 | 327.37| 31.95 | 0.57| 63.53 | 0.36| 431.49 | 24.05|
| n_g = 500 | 64.14 | 31.86| 51.88 | 3.74| 72.00 | 0.72| 544.87 | 60.68|
| p = 100 | n_g = 750 | 27.27 | 59.15| 73.81 | 2.62| 105.66 | 0.92| 802.73 | 69.51|
|       | n_g = 1000 | 18.72 | 17.19| 100.78 | 1.46| 139.69 | 1.67| 1040.10 | 96.63|