Multivariate cluster weighted models using skewed distributions

Michael P. B. Gallaugher · Salvatore D. Tomarchio · Paul D. McNicholas · Antonio Punzo

Received: 31 January 2021 / Revised: 2 October 2021 / Accepted: 23 October 2021 / Published online: 15 November 2021
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
Much work has been done in the area of the cluster weighted model (CWM), which extends the finite mixture of regression model to include modelling of the covariates. Although many types of distributions have been considered for both the response(s) and covariates, to our knowledge skewed distributions have not yet been considered in this paradigm. Herein, a family of 24 novel CWMs is considered which allows both the responses and covariates to be modelled using one of four skewed distributions (the generalized hyperbolic and three of its skewed special cases, i.e., the skew-t, the variance-gamma and the normal-inverse Gaussian distributions) or the normal distribution. Parameter estimation is performed using the expectation-maximization algorithm and both simulated and real data are used for illustration.

Keywords Mixture models · Cluster weighted models · Skewed distributions · Clustering

Mathematics Subject Classification 62H30 · 68T10

1 Introduction
Clustering is the process of finding underlying group structure in heterogeneous data. Although many methods exist for clustering, one of the most prevalent in the literature is model-based, and makes use of the $G$ component finite mixture model. The finite mixture model assumes that the joint probability density function (pdf) of a random
vector $V$, of dimension $d_V$, is

$$p(v; \theta) = \sum_{g=1}^{G} \pi_g f(v; \theta_g),$$

(1)

where $\pi_g > 0$ are the mixing proportions, with $\sum_{g=1}^{G} \pi_g = 1$, $f(\cdot)$ are the component pdfs parameterized by $\theta_g$, and $\theta$ contains all the parameters of the model. As discussed by McNicholas (2016a), the relationship between the finite mixture model and clustering was initially proposed by Tiedeman (1955). Some years after, Wolfe (1965) first utilized a Gaussian (or normal) mixture model for model-based clustering. Since then, there have been a myriad of contributions to this branch of the literature, mainly considering mixtures of non-normal distributions (a recent review is given by McNicholas 2016b). Some of these include mixtures of symmetric distributions which parameterize tail weight and may be useful for modelling data with outliers; examples are Peel and McLachlan (2000), Andrews and McNicholas (2011), Andrews and McNicholas (2012), Steane et al. (2012), Lin et al. (2014) and Dang et al. (2015). Additionally, many distributions that parameterize both skewness and tail weight have also been used in mixture models; see, for example, Karlis and Santourian 2009, Lin (2009), Lin (2010), Pyne et al. (2009), Wang et al. (2009), Frühwirth-Schnatter and Pyne (2010), Vrbik and McNicholas (2012), Vrbik and McNicholas (2014), Lee and McLachlan (2014), Murray et al. (2014a), Murray et al. (2014b), Browne and McNicholas (2015), Hung and Chang-Chien (2017), McNicholas et al. (2017) and Dang et al. (2019). The greater flexibility of these latter models address the flaw of the normal mixture model in overfitting the true number of components when used on skewed data.

One drawback of the non-normal mixture models mentioned thus far is that they do not typically account for dependencies via covariates. When there is a clear regression relationship between the variables, important insight can be gained by accounting for functional dependencies between them. In such scenarios, the finite mixture of regressions (FMRs; DeSarbo and Cron 1988) may be employed. As in traditional regression analysis, the FMR model assumes that the covariates are fixed, and therefore the distribution of the covariates is not taken into consideration when performing the cluster analysis. Indeed, such a model is also known as finite mixture of regression with fixed covariates. An extension of FMR is provided by the finite mixtures of regression models with concomitant variables (Dayton and Macready 1988), also known as mixture of experts (MoE: Murphy and Murphy 2020a), where the weights of the mixture functionally depend on a set of covariates and are usually modeled by a multinomial logistic distribution. However, even if the covariates are somehow incorporated into the clustering process, this model still belongs to the fixed covariates approach (Murphy and Murphy 2020a).

Unlike FMR and MoE, the cluster weighted model (CWM) offers a greater flexibility in that the distribution of the covariates is explicitly taken into account. First introduced by Gershenfeld (1997), it is also sometimes referred to as a finite mixture of regressions with random covariates. Over the years, several CWMs have been introduced in the literature; examples are Ingrassia et al. (2012), Ingrassia et al. (2014),...
Ingrassia et al. (2015), Punzo (2014), Punzo and Ingrassia (2015), Punzo and Ingrassia (2016), Ingrassia and Punzo (2016), Subedi et al. (2013), Subedi et al. (2015), Berta et al. (2016), Mazza et al. (2018), Punzo et al. (2021), Zarei et al. (2019), Di Mari et al. (2020), Galimberti and Soffritti (2020) and Počuča et al. (2020). However, all these CWMs consider a single response variable modelled by a univariate distribution. Examples of CWMs with multiple responses are provided by Dang et al. (2017), who use a multivariate normal (N) distribution for both the responses and the covariates, and (Punzo et al. 2018, 2021) in a hidden Markov model framework. Herein, we extend this branch of literature by proposing the use of multivariate skewed distributions for the responses, the covariates, or both. Specifically, the generalized hyperbolic (GH) and three of its skewed special cases; namely, the variance-gamma (VG), the skew-\(t\) (ST) and the normal-inverse Gaussian (NIG) distributions will be used. By also considering the N distribution, we introduce a family of 24 new CWMs that are flexible enough to cope with scenarios where both the responses and the covariates are skewed, or in which one of the two sets of variables is normally distributed and the other is skewed.

The remainder of this paper is laid out as follows. In Sect. 2, a detailed background is given. Section 3 presents the family of CWMs, including results about their identifiability; the illustration of the expectation-maximization (EM) algorithm for parameter estimation is postponed to Appendix B. Section 4 considers two analyses on simulated data, in which the parameter recovery and the classification performances for our models are evaluated. A comparison between FMRs, MoE models and CWMs is also discussed. Section 5 applies our CWMs, along with the above competing models, to two real datasets. A baseline clustering method concerning mixtures of multivariate generalized hyperbolic distributions (MGH-Ms) is also considered. This is done to show that accounting for an underlying "response-covariate" structure in the data can improve clustering performance. Lastly, we provide a summary and discuss possible avenues for future work in Sect. 6.

2 Background

In this section, we present some preliminaries that are used in the development of our model. In Sect. 2.1 we firstly recall the generalized inverse Gaussian (GIG) distribution, and then in Sect. 2.2 we present the four skewed distributions that will be used in this paper.

2.1 Generalized inverse Gaussian distribution

The GIG distribution (for a detailed analysis of its statistical properties, see Jorgensen (2012) plays an important role in the parameter estimation of the CWMs considered in this paper. A random variable \(W\) has a GIG distribution with parameters \(a > 0\), \(\lambda > 0\), \(\kappa > 0\), and \(\beta \in (-\infty, \infty)\),
$b > 0$, and $\lambda \in \mathbb{R}$, denoted herein by $\mathcal{GIG}(a, b, \lambda)$, if its pdf can be written as

$$h(w; a, b, \lambda) = \left(\frac{a}{b}\right)^{\frac{\lambda}{2}} \frac{w^{\lambda-1}}{2K_{\lambda}(\sqrt{ab})} \exp \left[-\frac{1}{2} \left(aw + \frac{b}{w}\right)\right],$$

where

$$K_{\lambda}(u) = \frac{1}{2} \int_{0}^{\infty} w^{\lambda-1} \exp \left\{-\frac{u}{2} \left(w + \frac{1}{w}\right)\right\} \, dw$$

is the modified Bessel function of the third kind with index $\lambda$. Expectations of some functions of a GIG random variable are mathematically tractable, e.g.:

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

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

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

An alternative parameterization of the GIG distribution is given by Browne and McNicholas (2015):

$$h(w; \omega, \eta, \lambda) = \left(\frac{w}{\eta}\right)^{\frac{\lambda}{2} - 1} \frac{1}{2\eta K_{\lambda}(\omega)} \exp \left[-\frac{\omega}{2} \left(\frac{w}{\eta} + \frac{\eta}{w}\right)\right],$$

where $\omega = \sqrt{ab}$, $\eta = \sqrt{b/a}$, and $\lambda$ are concentration, scale, and index parameters, respectively. For notational clarity, we will denote the GIG distribution parameterized as in (5) by $\mathcal{I}(\omega, \eta, \lambda)$.

### 2.2 Multivariate skewed distributions

Many skewed distributions may be derived by using a normal variance-mean mixture model. In a multivariate framework, this model assumes that a $d$-variate random vector $V$ can be written as

$$V = \mu + W\alpha + \sqrt{W}U,$$

where $\mu$ is a location parameter, $\alpha$ is a skewness parameter, $W$ is a positive random variable, and $U \sim \mathcal{N}_d(\mathbf{0}, \Sigma)$, where $\mathcal{N}_d(\mu, \Sigma)$ denotes a $d$-variate normal distribution with mean $\mu$ and covariance matrix $\Sigma$. Thus, we have that

$$V|W = w \sim \mathcal{N}_d(\mu + w\alpha, w\Sigma).$$
One of the most common distributions that can be obtained via (6) is the GH distribution. Specifically, a \( d \)-dimensional GH distribution, denoted by \( \mathcal{GH}_d(\mu, \alpha, \Sigma, \lambda, \omega) \), arises with \( W \sim \mathcal{I}(\omega, 1, \lambda) \), \( \omega > 0 \) and \( \lambda \in \mathbb{R} \), and the resulting pdf of \( V \) is

\[
f_{\text{GH}}(v; \mu, \alpha, \Sigma, \lambda, \omega) = \frac{\exp\left[ (v - \mu)'\Sigma^{-1}\alpha \right]}{(2\pi)^{d/2}|\Sigma|^{1/2}} K_{\lambda}(\omega) \times K_{\lambda-d/2}\left( \sqrt{[\rho(\alpha, \Sigma) + \omega][\delta(v; \mu, \Sigma) + \omega]} \right),
\]

(8)

where

\[
\delta(v; \mu, \Sigma) = (v - \mu)'\Sigma^{-1}(v - \mu) \quad \text{and} \quad \rho(\alpha, \Sigma) = \alpha'\Sigma^{-1}\alpha.
\]

It is important to notice that

\[
W|V = v \sim \mathcal{IG}(\rho(\alpha, \Sigma) + \omega, \delta(v; \mu, \Sigma) + \omega, \lambda - d/2).
\]

This information, together with (7), provides a hierarchical representation of the GH distribution, that is useful for random data generation and for the implementation of the EM algorithm discussed in Appendix B.

Note that we are using the GH parameterization of Browne and McNicholas (2015). Indeed, if the more classical parameterization given in McNeil et al. (2005) were used, the constraint \(|\Sigma| = 1\) would need to be imposed to ensure identifiability. However, as Browne and McNicholas (2015) point out, such a constraint would be prohibitively restrictive for model-based clustering and classification applications. Thus, Browne and McNicholas (2015) replace the constraint \(|\Sigma| = 1\) with the constraint \( \eta = 1 \) on the scale parameter \( \eta \) in (5).

The GH distribution is very flexible and contains many special cases known by specific names. Herein, we take a closer look at three of these distributions. Specifically, we consider:

- The \( d \)-dimensional VG distribution, denoted by \( \mathcal{VG}_d(\mu, \alpha, \Sigma, \psi) \), with pdf

\[
f_{\text{VG}}(v; \mu, \alpha, \Sigma, \psi) = \frac{2\psi^\psi \exp\left[ (v - \mu)'\Sigma^{-1}\alpha \right]}{(2\pi)^{d/2}|\Sigma|^{1/2} \psi(\psi)} \times K_{\psi-d/2}\left( \sqrt{[\rho(\alpha, \Sigma) + 2\psi][\delta(v; \mu, \Sigma)]} \right),
\]

(10)

with \( \psi > 0 \). The VG distribution can also be obtained via (6) when \( W \sim \mathcal{G}(\psi, \psi) \), where \( \mathcal{G}(a, b) \), \( a > 0 \) and \( b > 0 \), denotes the Gamma distribution with pdf

\[
h(w; a, b) = \frac{b^a}{\Gamma(a)} w^{a-1} \exp(-bw).
\]

To understand the relationship between the VG and GH distributions, see the appendix in Kim and Browne (2019). Note that, to obtain the parametrization
in (10), we need to replace \( \omega = 2\gamma \) with \( \omega = 2\gamma\lambda \) in the aforementioned appendix. Lastly, note that for the VG distribution we have

\[
W | V = v \sim \mathcal{IG} (\rho(\alpha, \Sigma) + 2\psi, \delta(v; \mu, \Sigma), \psi - d/2).
\]  

(11)

• The \( d_V \)-dimensional ST distribution, denoted by \( \text{ST}_d(\mu, \alpha, \Sigma, v) \), with pdf

\[
f_{\text{ST}}(v; \mu, \alpha, \Sigma, v) = \frac{2 (\frac{\nu}{2})^\nu \exp \left[ (v - \mu)' \Sigma^{-1} \alpha \right]}{(2\pi)^{d/2} |\Sigma|^\frac{d}{2} \Gamma(\frac{\nu}{2})} \times \frac{\nu + d}{2} \left( \sqrt{\rho(\alpha, \Sigma)} [\delta(v; \mu, \Sigma) + v] \right),
\]  

(12)

with \( \nu > 0 \). The ST distribution can also be obtained via (6) when \( W \sim \mathcal{IG}(v/2, v/2) \), where \( \mathcal{IG}(a, b) \), \( a > 0 \) and \( b > 0 \), is the inverse Gamma distribution with pdf

\[
h(w; a, b) = \frac{b^a}{\Gamma(a)} w^{a-1} \exp \left( -\frac{b}{w} \right).
\]

To analyze the relationship between the ST and GH distributions, see Appendix A. Furthermore, for the ST distribution we have

\[
W | V = v \sim \mathcal{IG} (\rho(\alpha, \Sigma), \delta(v; \mu, \Sigma) + v, -(v + d)/2).
\]  

(13)

• The \( d_V \)-dimensional NIG distribution, denoted by \( \mathcal{NIG}_d(\mu, \alpha, \Sigma, \kappa) \), with pdf

\[
f_{\text{NIG}}(v; \mu, \alpha, \Sigma, \kappa) = \frac{2 \exp \left\{ (v - \mu)' \Sigma^{-1} \alpha + \kappa \right\}}{(2\pi)^{d/2} |\Sigma|^\frac{d}{2} \Gamma(\frac{\nu}{2})} \times \frac{\nu + d}{2} \left( \sqrt{\rho(\alpha, \Sigma) + \kappa^2} [\delta(v; \mu, \Sigma) + 1] \right),
\]  

(14)

with \( \kappa > 0 \). The NIG distribution can also be obtained via (6) when \( W \sim \mathcal{IN}(1, \kappa) \), where \( \mathcal{IN}(a, b) \), \( a > 0 \) and \( b > 0 \), is the inverse Gaussian distribution with pdf

\[
h(w; a, b) = \frac{a}{\sqrt{2\pi}} \exp(ab) w^{\frac{3}{2}} \exp \left\{ -\frac{1}{2} \left( \frac{a^2}{w} + b^2w \right) \right\}.
\]

To understand the relationship between the NIG and GH distributions, we can easily see that \( W \sim \mathcal{IN}(1, \kappa) \) is a special case of \( W \sim \mathcal{IG}(\omega, 1, \lambda) \) when \( \omega = \kappa \) and \( \lambda = -1/2 \), so that the two distributions can be seen as nested. Lastly, note that for the NIG distribution we have

\[
W | V = v \sim \mathcal{IG} (\rho(\alpha, \Sigma) + \kappa^2, \delta(v; \mu, \Sigma) + 1, -(1 + d)/2).
\]  

(14)
3 Methodology

3.1 Cluster weighted models with skewed distributions

In many empirical studies the random vector of interest $V$ is composed by a random response vector $Y$, of dimension $d_Y$, and by a random covariate vector $X$, of dimension $d_X$, with $d_X + d_Y = d_V$; that is, $V = (X', Y')'$. Therefore, it is important to take into account the different role of $X$ and $Y$ in the definition of the mixture model (1). In the CWM framework we take advantage of the different role of the variables writing the joint pdf of $Y$ and $X$ as

$$p(x, y; \theta) = \sum_{g=1}^G \pi_g f(y|x; \theta_{Y|g}) f(x; \theta_{X|g}), \quad (15)$$

where $\theta = \{\pi_g, \theta_{X|g}, \theta_{Y|g}; g = 1, \ldots, G\}$ represents the set of all parameters. Hereafter, we will add a subscript to the parameters and the mixing variables to distinguish those related to $X$ from those referring to $Y$ thus avoiding the use of multiple symbols and letters and simplifying the notation.

The dependence of $Y$ on $X = x$ in the $g$-th mixture component is typically accounted for by allowing the mean or, more generally, some parameter in $\theta_{Y|g}$ (typically a location parameter, see e.g., Chen et al. (2014), Ferreira et al. (2015), Chamroukhi (2017) and Doğru and Arslan (2017) to depend on $x$ via some linear or nonlinear functional relationship. The possibility to specify different parametric models for either $f( y|x; \theta_{Y|g} )$ or $f( x; \theta_{X|g} )$, as well as different linear/nonlinear dependencies for $Y|X = x$ in each mixture component, makes the CWM a very flexible modelling approach. Moreover, in a clustering perspective via mixtures of regression models, the CWM has the advantage to allow for assignment dependence (Hennig 2000), in the sense that the component marginal distributions $f( x; \theta_{X|g} )$ of the covariates can also be distinct and they can affect the assignment of the data points $(x', y')'$ to the clusters. These advantages yielded a large amount of work about CWMs in the last decade, as discussed in Sect. 1. In particular, Dang et al. (2017) assume, for the $g$-th mixture component, a $d_X$-variate normal distribution for $X$, say $X \sim N_{d_X}(\mu_{X|g}, \Sigma_{X|g})$, and a $d_Y$-variate normal distribution for $Y|X = x$, say $Y|X = x \sim N_{d_Y}(\mu_{Y|g}(x; B_g), \Sigma_{Y|g})$, by assuming the linear relation $\mu_{Y|g}(x; B_g) = B'_g x^*$, where $B_g$ is a $(1 + d_X) \times d_Y$ matrix of regression coefficients and $x^* = (1, x')'$. This means that $\theta_{X|g} = \{\mu_{X|g}, \Sigma_{X|g}\}$ and $\theta_{Y|g} = \{B_g, \Sigma_{Y|g}\}$.

For the purposes of this paper, the densities $f( y|x; \theta_{Y|g} )$ and $f( x; \theta_{X|g} )$ in (15) can be any of the four multivariate skewed distributions presented in Sect. 2.2 or the multivariate normal distribution. Being members of the family of distributions defined by (6), in the $g$-th mixture component we have that

$$Y|x = B'_g x^* + W_{Y|g} \alpha_{Y|g} + \sqrt{W_{Y|g}} U_{Y|g}, \quad (16)$$

$$X = \mu_{X|g} + W_{X|g} \alpha_{X|g} + \sqrt{W_{X|g}} U_{X|g}. \quad (17)$$
In addition, \( f(y|x; \theta_{Y|g}) \) and \( f(x; \theta_{X|g}) \) need not be of the same type, thus creating
a family of 25 CWMs, 24 of which are herein introduced. For notational clarity, each
model will be labeled by separating with a dash the acronyms used for \( f(x; \theta_{X|g}) \)
and \( f(y|x; \theta_{Y|g}) \), respectively. For example, if we consider a CWM having a GH
distribution for \( X \) and a ST distribution for \( Y|X = x \), it will be referred to as GH-ST
CWM.

### 3.2 Identifiability

Maximum likelihood (ML) is the traditional approach to obtain estimates of the param-
eters of the CWM in (15). ML estimates can be computationally obtained by the
well-known EM algorithm (Dempster et al. 1977). A detailed presentation of the EM
algorithm for our CWMs is given in Appendix B. However, as a preliminary requisite
to parameter estimation, it is important to investigate identifiability of our models.
With this aim, we firstly need a definition of the nontrivial concept of identifiability
for CWMs, and we do that by using the notation by Titterington et al. (1985).

Let

\[
\mathcal{F} = \{ f(x, y; \theta_X, \theta_Y) = f(y|x; \theta_Y) f(x; \theta_X), \theta_X \in \Theta_X, \theta_Y \in \Theta_Y, \\
x \in \mathbb{R}^{d_X}, y \in \mathbb{R}^{d_Y} \} 
\]

be the class of \((d_X \times d_Y)\)-dimensional densities from which CWMs are to be formed.
The class \( \mathcal{P} \) of CWMs, i.e., of finite mixtures of \( \mathcal{F} \), is defined by

\[
\mathcal{P} = \left\{ p(x, y; \theta) = \sum_{g=1}^{G} \pi_g f(x, y; \theta_{X|g}, \theta_{Y|g}), \pi_g > 0, \sum_{g=1}^{G} \pi_g = 1, \\
f(x, y; \theta_{X|g}, \theta_{Y|g}) \in \mathcal{F}, \text{ all } G = 1, 2, \ldots, x \in \mathbb{R}^{d_X}, y \in \mathbb{R}^{d_Y} \right\} 
\]

so that \( \mathcal{P} \) is the convex hull of \( \mathcal{F} \). In (19), \( f(x, y; \theta_{X|1}, \theta_{Y|1}), \ldots, f(x, y; \theta_{X|G}, \theta_{Y|G}) \) are assumed to be distinct members of \( \mathcal{F} \). This assumption, as well as the posi-
tivity of the weights \( \pi_1, \ldots, \pi_G \), avoids nonidentifiability due to potential overfitting
(Crawford 1994). This happens because, without these assumptions, a mixture with \( G \)
components can be always written as a mixture with \( G+1 \) components where either one
component is empty or two components are equal; for details see Frühwirth-Schnatter
(2006, Section 1.3.2).

Based on these preliminaries, we formalize identifiability of \( \mathcal{P} \) in Definition 3.1.
Definition 3.1 (Identifiability) Let
\[ p(x, y; \theta) = \sum_{g=1}^{G} \pi_g f(x, y; \theta_X|g, \theta_Y|g) \]
and
\[ p(x, y; \tilde{\theta}) = \sum_{j=1}^{\tilde{G}} \tilde{\pi}_j f(x, y; \tilde{\theta}_X|j, \tilde{\theta}_Y|j) \]
be any two members of \( \mathcal{P} \) and suppose that \( p(x, y; \theta) \equiv p(x, y; \tilde{\theta}) \) if and only if \( G = \tilde{G} \) and we can order the summations such that \( \pi_g = \tilde{\pi}_g, \theta_X|g = \tilde{\theta}_X|g \) and \( \theta_Y|g = \tilde{\theta}_Y|g, g = 1, \ldots, G \). Then \( \mathcal{P} \) is identifiable.

Sufficient conditions for the identifiability of N-N CWMs and mixtures of GH distributions are given in Dang et al. (2017) and Browne and McNicholas (2015), respectively. Further useful results about identifiability of bivariate CWMs are provided by Frimpong et al. (2008). We will take advantage of all these results in Theorem 3.1 to show that the more flexible (less parsimonious) GH-GH CWM is identifiable provided that all \( \theta_Y|g \) are pairwise distinct and \( f(x; \theta_X|g) \) is not degenerate, \( g = 1, \ldots, G \).

Theorem 3.1 (Identifiability of the GH-GH CWM) If \( \theta_Y|g \neq \theta_Y|s \), with \( g \neq s \), and \( f_{GH}(x; \theta_X|g) \) is not degenerate, \( g, s = 1, \ldots, G \), then the class \( \mathcal{P}_{GH-GH} \) of finite mixtures of
\[ \mathcal{F}_{GH-GH} = \{ f_{GH-GH}(x, y; \theta_X, \theta_Y) = f_{GH}(y|x; \theta_Y) f_{GH}(x; \theta_X), \theta_X \in \Theta_X, \theta_Y \in \Theta_Y, x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y} \} \]
is identifiable.

Proof See Appendix B.

The results about identifiability of the N-N CWM given by Dang et al. (2017), as well as the fact that the VG, ST and NIG distributions are nested in the GH distribution (refer to Sect. 2.2), allow us to apply the sufficient condition of identifiability given in Theorem 3.1 to all of the CWMs illustrated in this paper.

4 Simulated data analyses

In this section, several aspects related to our models are analyzed. First, in Sect. 4.1 the parameter recovery and classification performance are analyzed under different scenarios. Then, in Sect. 4.2 our models are compared to FMRs and MoE models. A comparison with the N-N CWM is obtained as by-product. For a better comparability among the competing models, the corresponding fitting algorithms are all started by using the initialization strategy discussed in Appendix B.1. The capability of the Bayesian information criterion (BIC; Schwarz 1978) in detecting the data generating model (DGM) is also evaluated.
Because of the high number of CWMs introduced in this manuscript, we will focus our attention on four of the 24 novel CWMs. Specifically, we analyze four models that can cover the following different scenarios:

1. \( f(x; \theta_{X|g}) \) and \( f(y|x; \theta_{Y|g}) \) are pdfs of the same skewed type;
2. \( f(x; \theta_{X|g}) \) and \( f(y|x; \theta_{Y|g}) \) are pdfs of a different skewed type;
3. \( f(x; \theta_{X|g}) \) is skewed and \( f(y|x; \theta_{Y|g}) \) is normal;
4. \( f(x; \theta_{X|g}) \) is normal and \( f(y|x; \theta_{Y|g}) \) is skewed.

As illustrative examples, we consider the (1) GH-GH CWM, (2) VG-ST CWM, (3) NIG-N CWM and (4) N-ST CWM. Note that the models are chosen so that all the distributions considered in this manuscript are incorporated in some capacity. We consider the case with \( d_Y = 2, d_X = 3 \) and sample size \( N = 400 \). The parameters used to generate the data are displayed in Table 1.

For each of the four CWMs, 100 datasets are generated and the corresponding model is fitted with \( G = 2 \). The mean squared error (MSE) of the parameter estimates, over the 100 datasets, are then reported in Table 2 for Group 1 and in Table 3 for Group 2.

Overall, the MSEs assume small values, although there are notable exceptions for some specific parameters. First, we can see that the MSEs related to the GH-GH CWM are slightly worse than those of the other CWMs. This may be due to the larger number of model parameters as well as the greater estimation complexity of the GH
| Par. | GH-GH | VG-ST | NIG-N | N-ST |
|------|-------|-------|-------|------|
| $\pi_1$ | 0.001 | 0.001 | 0.001 | 0.001 |
| $\mu_{X|1}$ | $(0.089, 0.161, 0.132)'$ | $(0.020, 0.063, 0.038)'$ | $(0.014, 0.031, 0.029)'$ | $(0.006, 0.014, 0.010)'$ |
| $\alpha_{X|1}$ | $(0.106, 0.113, 0.091)'$ | $(0.029, 0.078, 0.046)'$ | $(0.039, 0.054, 0.046)'$ | — |
| $\Sigma_{X|1}$ | $(0.167 0.046 0.031)$ | $(0.020 0.016 0.011)$ | $(0.048 0.030 0.021)$ | $(0.001 0.012 0.009)$ |
| $\omega_{X|1}$ | 0.155 | — | — | — |
| $\lambda_{X|1}$ | 0.328 | — | — | — |
| $\psi_{X|1}$ | — | 0.414 | — | — |
| $\kappa_{X|1}$ | — | — | 0.018 | — |
| $B_1$ | $(0.205 0.753)$ | $(0.302 0.319)$ | $(0.102 0.072)$ | $(0.242 0.289)$ |
| $\alpha_{Y|1}$ | $(0.016, 0.171)'$ | $(0.015, 0.059)'$ | — | $(0.012, 0.052)'$ |
| $\Sigma_{Y|1}$ | $(0.201 0.015)$ | $(0.027 0.008)$ | $(0.014 0.004)$ | $(0.020 0.008)$ |
| $\omega_{Y|1}$ | 0.742 | — | — | — |
| $\lambda_{Y|1}$ | 1.014 | — | — | — |
| $\nu_{Y|1}$ | — | 1.271 | — | 1.233 |
Table 3  MSEs of the parameter estimates for Group 2 over 100 datasets for each CWM

| Par. | GH-GH          | VG-ST          | NIG-N          | N-ST          |
|------|----------------|----------------|----------------|---------------|
| \(\pi_2\) | 0.001          | 0.001          | 0.001          | 0.001         |
| \(\mu_{X|2}\) | \((0.100, 0.076, 0.074)'\) | \((0.020, 0.053, 0.029)'\) | \((0.011, 0.014, 0.015)'\) | \((0.006, 0.005, 0.005)'\) |
| \(\alpha_{X|2}\) | \((0.081, 0.080, 0.076)'\) | \((0.031, 0.057, 0.039)'\) | \((0.044, 0.053, 0.049)'\) | —             |
| \(\Sigma_{X|2}\) | \((0.028 0.171 0.029)\) | \((0.006 0.021 0.008)\) | \((0.009 0.067 0.015)\) | \((0.005 0.009 0.006)\) |
| \(\omega_{X|2}\) | 0.185          | —              | —              | —             |
| \(\lambda_{X|2}\) | 0.257          | —              | —              | —             |
| \(\psi_{X|2}\) | —              | 1.249          | —              | —             |
| \(\kappa_{X|2}\) | —              | —              | 0.029          | —             |
| \(B_2\) | \((0.922 0.347)\) | \((0.344 0.321)\) | \((0.139 0.113)\) | \((0.390 0.248)\) |
| \(\alpha_{Y|2}\) | \((0.04 0.004)\) | \((0.007 0.006)\) | \((0.003 0.002)\) | \((0.007 0.006)\) |
| \(\Sigma_{Y|2}\) | \((0.04 0.003)\) | \((0.007 0.007)\) | \((0.002 0.002)\) | \((0.007 0.006)\) |
| \(\omega_{Y|2}\) | \((0.145 0.019')\) | \((0.056 0.014')\) | —              | \((0.074 0.015')\) |
| \(\lambda_{Y|2}\) | 1.411          | —              | —              | —             |
| \(\psi_{Y|2}\) | 1.858          | —              | —              | —             |
| \(\nu_{Y|2}\) | —              | 1.521          | —              | 1.317         |
distribution (Aas and Hobæk Haff 2005; Göncü and Yang 2016). Conversely, the MSEs of the parameters related to the normal distribution in the NIG-N and N-ST CWMs are relatively smaller than in the other cases likely because of the reduced estimation complexity.

By looking at the MSEs of some specific parameters, we observe that those related to \( \omega, \lambda, \psi \) and \( \nu \) are noticeably higher than the others. This may be due to the fact the updates for these parameters cannot be obtained in closed form and that, in general, the parameters governing the tail weight are the most difficult to be estimated (Punzo and Bagnato 2021). Furthermore, the MSEs of the intercepts are higher than those of the other regression coefficients in each CWM. This might depend on how far the groups are in the \( x \)- and \( y \)-spaces, so that small differences in the estimated slopes can produce big differences in the estimates of the intercepts (Punzo and Ingrassia 2016).

Finally, we assess the classification performance by fitting all our 24 CWMs for \( G \in \{1, 2, 3\} \) to each generated dataset. In detail, we use the adjusted Rand index (ARI; Hubert and Arabie 1985), which calculates the agreement between the true classification and the one predicted by the model. An ARI of 1 indicates perfect agreement between the two partitions, whereas the expected value of the ARI under random classification is 0. We report that, regardless of the DGM, the average ARI, computed over 100 datasets and by using the classifications produced by the best fitting models according to the BIC, is equal to 1.

### 4.2 Comparison between CWMs, FMRs and MoE models

For illustrative purposes, we generate data from CWMs based on the ST distribution, namely the ST-ST CWM, ST-N CWM and N-ST CWM (that are examples of the scenarios 1, 3 and 4 described in Sect. 4.1, respectively), as well as from the N-N CWM. For each of these four CWMs, 100 datasets are generated and our 24 CWMs, along with the N-N CWM, are fitted to each dataset. We also fit to the data the FMRs based on the distributions considered in this manuscript, as well as the unconstrained “full MoE model” discussed in Murphy and Murphy (2020a) and implemented by using the \texttt{MoEClust} package (Murphy and Murphy 2020b). All the models are fitted to the data for \( G \in \{1, 2, 3\} \). We set \( d_Y = 2, d_X = 3, N = 400 \) and the parameters displayed in Table 4 to generate the datasets. An example of generated dataset from each CWM is displayed in Fig. 1; here, it is clear that there is a grouping structure in the covariates.

We start by analyzing the results related to the CWMs. In detail, each sub-plot of Fig. 2 illustrates the number of times each \( G \) is chosen by the BIC for each model over the 100 datasets.

In Fig. 2a, we can see that when the data are generated by the ST-ST CWM, all the CWMs for which either \( f (x; \theta_X|g) \), \( f (y|x; \theta_Y|g) \), or both are assumed to be normal, problems arise in detecting the true number of groups in the data. As discussed in Sect. 1, mixture models based on the normal distribution have the tendency of overfitting the true number of components when used on skewed data. This is confirmed by our results, but it is also interesting to notice that this issue has a different magnitude depending on which one of \( f (x; \theta_X|g) \) or \( f (y|x; \theta_Y|g) \) is modelled
Table 4  Parameters used to generate the datasets from the considered CWMs

| CWM          | Par. Group 1 | Group 2 |
|--------------|--------------|---------|
| All          | $\pi_g$      | 0.50    | 0.50    |
| All          | $\mu_{X|g}$  | $(-2.50, 4.00, 3.00)$ | $(-2.50, -3.00, -3.00)$ |
| ST-ST, ST-N  | $\alpha_{X|g}$ | $(2.90, -0.50, -0.05)$ | $(2.30, -0.90, -0.35)$ |
| All          | $\Sigma_{X|g}$ | $(-0.50, 0.45, -0.75)$ | $(-0.90, 1.55, 0.25)$ |
| ST-ST, ST-N  | $\nu_{X|g}$  | 7       | 7       |
| All          | $B_g$        | $(-6.00, 1.00)$ | $(-10.00, 7.50)$ |
| ST-ST, N-ST  | $\alpha_{Y|g}$ | $(2.00, -2.50)$ | $(-1.00, 2.00)$ |
| All          | $\Sigma_{Y|g}$ | $(1.80, -0.30)$ | $(-0.30, 2.00)$ |
| ST-ST, N-ST  | $\nu_{Y|g}$  | 7       | 7       |

Fig. 1  Examples of pairwise scatter plots when the DGM is: a ST-ST CWM, b ST-N CWM, c N-ST CWM and d N-N CWM
using the normal pdf. Specifically, when $f(x; \theta_X|g)$ is assumed to be skewed and $f(y|x; \theta_Y|g)$ is assumed to be normal, most of the times $G = 2$ is still properly selected, although it is still not as accurate as the CWMs where both $f(x; \theta_X|g)$ and $f(y|x; \theta_Y|g)$ are assumed to be skewed. On the other hand, when $f(x; \theta_X|g)$ is normal and $f(y|x; \theta_Y|g)$ is skewed, $G = 3$ is nearly always chosen.

When the datasets are generated from a ST-N CWM, the only models having serious problems are those with normally distributed covariates in each mixture component, as shown in Fig. 2b. Because of their greater flexibility, all the CWMs assuming a skewed $f(y|x; \theta_Y|g)$ are able to accurately model symmetric data. The results for the N-ST CWM are displayed in Fig. 2c. Here, the only CWMs that present issues are those for which $f(y|x; \theta_Y|g)$ is normal. Lastly, when the data are generated by the N-N CWM, $G = 2$ is properly selected for all the CWMs, as shown in Fig. 2d.
Regarding the capability of the BIC in detecting the exact DGM, we observed that over the 100 datasets generated by the ST-ST and N-ST CWMs, correct model is detected 78 and 82 times, respectively. The occasions in which the BIC fails are due to a wrong distribution chosen for only one of the covariates or the conditional distribution of the responses. Under no circumstances are both distributions incorrectly chosen. When the ST-N CWM is considered, the BIC performance is even better than before, as it selects the correct model 99 times. Similarly, the DGM is selected 100 times when the data are generated by the N-N CWM.

The classification results of the CWMs are shown in Fig. 3. Specifically, by taking into account the BIC results of Fig. 2, for each CWM we report the average ARI computed over the 100 datasets generated by each DGM. Here, the models that have the lowest ARI values are those assuming normal covariates for the datasets generated from the ST-N and ST-ST models. All the other CWMs produce very good classifications for the four DGMs considered.

We now discuss the results related to the FMRs. As before, each sub-plot of Fig. 4 reports the number of times each $G$ is chosen by the BIC for each model over the 100 datasets. We can see that in all the situations illustrated in Fig. 4, when the skewed FMRs are considered, $G = 1$ is repeatedly selected. This has also an effect on the ARI values that are practically null. Conversely, when the N-FMRs is analyzed, it seems that $G = 1$ is mainly selected when $f(y|x; \theta y_{g})$ of the DGMs is modelled using the normal density (see Fig. 4b, d), whereas some group structures are detected when $f(y|x; \theta y_{g})$ of the DGMs is skewed (see Fig. 4a, c). Nevertheless, the produced classifications are meaningless since the ARI values are approximately equal to 0. Thus, despite the clear separation between the two groups, the FMRs approach is unable to correctly identify them.

Lastly, the results concerning the MoE models are reported in Table 5. Although it is a fixed covariates approach, the use of covariate information improves the clustering process when compared to the FMRs. This is particularly relevant when the DGMs are the ST-N CWM — with a higher average ARI then those from CWMs having a...
Multivariate cluster weighted models using skewed…

Fig. 4 Radar plots of the number of times, over 100 replications, each number of groups \( G \) is chosen by the BIC for the competing 5 FMRs. The DGMs are: a ST-ST CWM, b ST-N CWM, c N-ST CWM and d N-N CWM.

Table 5 Number of times, over 100 replications, each value of \( G \) is chosen by the BIC for the MoE models and for each DGM. The last columns reports the average ARI values

| DGM       | \( G = 1 \) | \( G = 2 \) | \( G = 3 \) | ARI |
|-----------|-------------|-------------|-------------|-----|
| ST-ST CWM | 1           | 36          | 63          | 0.64|
| ST-N CWM  | 0           | 97          | 3           | 0.96|
| N-ST CWM  | 1           | 49          | 50          | 0.38|
| N-N CWM   | 3           | 95          | 2           | 0.86|

normal \( f(x; \theta_{X|g}) \) — and the N-N CWM. However, and similarly to FMRs, when \( f(y|x; \theta_{Y|g}) \) of the DGMs is skewed the model performance deteriorates, with an incorrect number of groups chosen at least half of the time.
5 Real data applications

5.1 Overview

In this section, all the CWMs discussed herein are applied to two real datasets. For comparison purposes, FMRs, MoE models and MGH-Ms are fitted to the data too. Specifically, MGH-Ms are fitted on the merged data via the mixGHD package (Tortora et al. 2021).

5.2 Data

The first application considers the AIS dataset included in the ssn package (Azzalini 2020). It contains measurements of 102 male and 100 female athletes (i.e., $N = 202$ and $G = 2$) collected at the Australian Institute of Sport. The subset of seven variables, considered recently in the mixtures of regressions literature (Soffritti and Galimberti 2011; Dang et al. 2017) is analyzed. Specifically, we consider red cell count (RCC), white cell count (WCC), plasma ferritin concentration (FE), body mass index (BMI), sum of skin folds (SSF), body fat percentage (BFT), and lean body mass (LBM). As in Soffritti and Galimberti (2011) and Dang et al. (2017), the blood composition variables (RCC, WCC and FE) are selected as the response variables, while the biometrical variables (BMI, SSF, BFT and LBM) are the covariates. The pairwise scatter plots of the data, colored according to the true data classification, are given in Fig. 5. As we can see, many of the variables seem to present a skewed behavior, so that our distributions should be able to accurately model the data.
The second application considers ANVUR data which are handled by the Italian national agency for the evaluation of universities and research institutes (ANVUR; for more information about ANVUR and its primary role in the Italian university system, see https://www.anvur.it/en/homepage/). It contains measurements for 33 families of bachelor’s degrees and 42 families of master’s degrees (i.e., \( N = 75 \) and \( G = 2 \)). According to Italian law, each family represents a set of degrees sharing a similar topic and that are considered equivalent in legal value. The dataset consists of the following quantitative indicators measuring the percentage of: students that continued in the second year of a university study program (CNT), students who complete their studies within the normal duration of their study program (COM), students who obtained at least 40 course credits during their first year (CC), and teaching hours provided by professors having a permanent contract over the total teaching hours (TH). The variables related to the continuation or completion of studies (CNT and COM) are selected as the response variables, while those concerning the course credits and the teaching hours are the covariates (CC and TH). All the indicators refer to 2017 and are measured among the non-telematic Italian universities. The pairwise scatter plots of the data, colored according to the true data classification, are illustrated in Fig. 6.

Also in this case, there seems to be present a skewed behavior in the data, particularly when the TH variable is considered.

### 5.3 Results

In both applications, all the competing models are fitted with \( G \in \{1, 2, 3\} \) and the results are shown in Table 6. In detail, for each family of models (i.e., CWM, FMRs, MoE models and MGH-M) only the one selected by the BIC is reported.
When the AIS dataset is considered, the best CWM is the ST-ST with \( G = 2 \), which leads to an almost perfect classification given the high ARI value. Conversely, the best FMR model (that is based on the VG distribution) and MoE find \( G = 1 \) in the data. Notice that, even if the FMRs and MoE models were fitted directly with \( G = 2 \), the corresponding ARIs would be 0.02 (referred to the N FMR, that is the best FMR with \( G = 2 \)) and 0.14, respectively. A possible explanation of such behavior resides in the fact that the fixed covariates approaches look at heterogeneity only in \( Y | x \). As can be seen in Fig. 5, the two groups are not well separated with respect to \( Y | x \) and in some cases the linear relationship among the variables (see for example the subplot referring to RCC-LBM) is so similar for the two groups that fixed covariates approaches are not able to distinguish them. Conversely, when taking into account the distribution of the covariates, as done by the CWMs, we can gain additional information about the underlying group structure (see for example the subplot referring to the pair of variables BFT-LBM), that can lead to better clustering and classification performance.

As for the MGH-Ms, the best model according to the BIC has \( G = 2 \) components and an ARI of 0.88. Therefore, compared to the ST-ST CWM, the MGH-M produces a worse classification, suggesting that taking into account a regression structure might lead to better clustering performance.

When the ANVUR dataset is analyzed, we notice that the model producing the highest ARI is the VG-N CWM with \( G = 2 \) components. Unlike the previous dataset, where the best CWM had skewed distributions for both \( X \) and \( Y | x \) in each mixture component, here the normal distribution is chosen for the distribution of the responses. When the fixed covariate approaches are considered, the MoE models performs better than the best FMRs (that is based on the N distribution), and both models find \( G = 2 \) groups in the data. Thus, in this case the fixed covariates approaches detect a group structure in the data, probably helped by the greater heterogeneity in \( Y | x \) (see Fig. 6). Nevertheless, and despite that the two groups are not well separated in the covariate space (see the subplot referring to the pair of variables CC-TH), their classification is worse than the VG-N CWM. Therefore, even in such situations, the use of the covariates distribution can be useful in the clustering process. Finally, the MGH-Ms do not perform particularly well here, considering that they produce a classification that is worse than those of the VG-N CWM and MoE models. Thus, two out of three models that incorporate linear dependencies on covariates aided in better clustering performance.

| AIS          | ANVUR       |
|--------------|-------------|
| Model  | \( G \) | ARI | Model  | \( G \) | ARI |
| ST-ST CWM   | 2         | 0.92 | VG-N CWM | 2   | 0.84 |
| VG FMR      | 1         | 0.00 | N FMR   | 2   | 0.53 |
| MoE         | 1         | 0.00 | MoE     | 2   | 0.79 |
| MGH-Ms      | 2         | 0.88 | MGH-Ms  | 2   | 0.61 |
6 Conclusions

A novel family of 24 multivariate CWMs was introduced. Extending the completely unconstrained normal CWM of Dang et al. (2017), the distributions of responses and covariates were allowed to be skewed. For illustrative purposes, the following four skewed distributions were considered: the generalized hyperbolic, the skew-\(t\), the variance-gamma and the normal inverse Gaussian. Additionally, by also considering the normal distribution, our models were flexible enough to handle scenarios in which the covariates and the responses conditioned on the covariates are skewed, or in which one of the two sets of variables is normally distributed and the other one is skewed.

Identifiability conditions were discussed and an EM algorithm for parameter estimation was presented. The capability of recovering the parameters and the classification of the data generating model was tested in a simulation study. A comparison among the CWMs, the FMRs and MoE models was also investigated via simulated data. Specifically, it was shown that by ignoring the distribution of the covariates, FMRs may fail to recognize the underlying group structure in the data, despite the separation of the groups. The MoE models model provided better performance than FMRs, but it still produces worse results than the majority of CWMs.

All our CWMs, as well as the normal CWM, FMRs and MoE models were additionally fitted to two real datasets. MGH-Ms has been also considered as baseline clustering method. The results of both applications show that, in a regression setting, by explicitly considering the covariate distribution in the model formulation can help in detecting the true underlying group structure. Furthermore, when compared to MGH-Ms, the results suggest that taking into account the regression structure by use of linear dependencies might help in producing better clustering solutions.

Possible extensions of this work might be to consider a parsimonious structure for the covariance matrices, in the fashion of Dang et al. (2017), as well as constraining the parameters governing the tail behavior. Furthermore, a generalization of our models to the matrix-variate framework could be done by extending the matrix-normal CWM, recently introduced by Tomarchio et al. (2021), to include skewed matrix distributions (Gallaugher and McNicholas 2017, 2019).

Appendix

A Technical details on the ST distribution

In the fashion of Kim and Browne (2019), it is possible to show that the pdf in (12) can be obtained from the pdf in (8), by forcing \(\lambda\) and \(\omega\) to be a convenient function of \(\nu\), by letting \(\Sigma\) and \(\alpha\) to become large in a controlled way, and by letting \(\omega\) to become small in a controlled way. Specifically, let

\[
\theta = \left(\mu, \gamma^{-1}\alpha, \gamma^{-1}\Sigma, -\frac{\nu}{2}, \nu\gamma\right).
\]
where $\gamma > 0$ is a scaling factor. By substituting these parameter values into (8) we obtain

$$f_{GH}(v; \theta) = \frac{\exp[(v - \mu)'\Sigma^{-1}\alpha]}{(2\pi)^{d/2}|\gamma^{-1}\Sigma|^{d/2}K_{-\frac{d}{2}}(\gamma)} \left[\frac{\gamma\delta(v; \mu, \Sigma) + v\gamma}{\gamma^{-1}\rho(\alpha, \Sigma) + v\gamma}\right]^{-\frac{v+d}{4}} \times K_{-\frac{v+d}{2}}\left(\sqrt{\gamma^{-1}\rho(\alpha, \Sigma) + v\gamma}\right)\left[\gamma\delta(v; \mu, \Sigma) + v\gamma\right]^{-\frac{v+d}{4}} \times K_{-\frac{v+d}{2}}\left(\sqrt{\rho(\alpha, \Sigma) + v\gamma^2}\right)\left[\delta(v; \mu, \Sigma) + v\right],$$

which after some manipulation becomes

$$f_{GH}(v; \theta) = \frac{\gamma^{-\frac{v}{2}}}{K_{-\frac{v}{2}}(\gamma)} \frac{\exp[(v - \mu)'\Sigma^{-1}\alpha]}{(2\pi)^{d/2}|\Sigma|^{d/2}} \left[\frac{\delta(v; \mu, \Sigma) + v}{\rho(\alpha, \Sigma) + v\gamma^2}\right]^{-\frac{v+d}{4}} \times K_{-\frac{v+d}{2}}\left(\sqrt{\rho(\alpha, \Sigma) + v\gamma^2}\right)\left[\delta(v; \mu, \Sigma) + v\right].$$

Now, letting $\gamma \to 0$ and by using the following asymptotic relation

$$K_\lambda(x) \sim \Gamma(-\lambda)x^{-\lambda}$$

for $x \to 0$ and $\lambda < 0$,

we obtain

$$2\left(\frac{v}{2}\right)^{\frac{v}{2}} \frac{\exp[(v - \mu)'\Sigma^{-1}\alpha]}{(2\pi)^{d/2}|\Sigma|^{d/2}\Gamma(\frac{v}{2})} \left[\frac{\delta(v; \mu, \Sigma) + v}{\rho(\alpha, \Sigma)}\right]^{-\frac{v+d}{4}} \times K_{-\frac{v+d}{2}}\left(\sqrt{\rho(\alpha, \Sigma)\delta(v; \mu, \Sigma) + v}\right),$$

which is the density reported in (12).

### B Parameter estimation

Let $(x'_1, y'_1)'$, $\ldots$, $(x'_N, y'_N)'$ be a random sample of $N$ independent observations from (15). In the context of the EM algorithm, the random sample is considered incomplete. Specifically, we have two sources of incompleteness. The first source arises from the fact that, for each observation, we do not know its component membership; to govern this source, we use an indicator vector $z_i = (z_{i1}, \ldots, z_{iG})$, where $z_{ig} = 1$ if observation $i$ is in group $g$, and $z_{ig} = 0$ otherwise. The second source arises if $f(y|x; \theta_{Y|g})$ or $f(x; \theta_{X|g})$ are skewed; to govern this source, we need the latent variables $W_{Y|g}$ and $W_{X|g}$ introduced in (17).

Based on this source of incompleteness, we can write the complete-data log-likelihood in the following way

$$l(\vartheta) = l_1(\pi) + l_2(\theta_X) + l_3(\theta_Y),$$

(22)
where $\pi = (\pi_1, \ldots, \pi_G)'$, and

$$l_1(\pi) = \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} \log(\pi_g).$$

If $X$ in component $g, g = 1, \ldots, G$, follows one of the four skewed distributions,

$$l_2(\theta_X) = \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} \log[h(w_{ig}x; \phi_w_{X|g})] + C_X$$

$$- \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} [\log(|\Sigma_{X|g}|) + w_{ig}x' \alpha'_{g} \Sigma^{-1}_{X|g} \alpha_{X|g}$$

$$+ \frac{1}{w_{ig}X} (x_i - \mu_{X|g})' \Sigma^{-1}_{X|g} (x_i - \mu_{X|g})$$

$$- (x_i - \mu_{X|g})' \Sigma^{-1}_{X|g} \alpha_{X|g} - \alpha'_{g} \Sigma^{-1}_{X|g} (x_i - \mu_{X|g})],$$

where $h(w_{ig}x; \phi_w_{X|g})$ is the appropriate pdf for $W_{ig}X$ discussed in Sect. 2, with parameters notated as $\phi_w_{X|g}$, while $C_X$ is constant with respect to the parameters. On the other hand, if $X$ in component $g, g = 1, \ldots, G$, is normally distributed then

$$l_2(\theta_X) = C_X - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} [\log(|\Sigma_{X|g}|) + (x_i - \mu_{X|g})' \Sigma^{-1}_{X|g} (x_i - \mu_{X|g})].$$

Similarly, if $Y|x$ in component $g, g = 1, \ldots, G$, is distributed according to one of the four skewed distributions,

$$l_3(\alpha_Y) = \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} \log[h(w_{ig}y; \phi_w_{Y|g})] + C_Y$$

$$- \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} [\log(|\Sigma_{Y|g}|) + w_{ig}y' \alpha'_{g} \Sigma^{-1}_{Y|g} \alpha_{Y|g}$$

$$+ \frac{1}{w_{ig}Y} (y_i - B'_g x^*_i)' \Sigma^{-1}_{Y|g} (y_i - B'_g x^*_i)$$

$$- (y_i - B'_g x^*_i)' \Sigma^{-1}_{Y|g} \alpha_{Y|g} - \alpha'_{g} \Sigma^{-1}_{Y|g} (y_i - B'_g x^*_i)]$$

where $h(w_{ig}y; \phi_w_{Y|g})$ is the appropriate pdf for $W_{ig}Y$ discussed in Sect. 2, with parameters notationally compacted as $\phi_w_{Y|g}$, while $C_Y$ is constant with respect to the parameters. Conversely, if $Y|x$ in component $g, g = 1, \ldots, G$, is normally distributed,

$$l_3(\alpha_Y) = C_Y - \frac{1}{2} \sum_{g=1}^{G} \sum_{i=1}^{N} z_{ig} [\log(|\Sigma_{Y|g}|) + (x_i - \mu_{Y|g})' \Sigma^{-1}_{Y|g} (x_i - \mu_{Y|g})].$$
After initialization, the EM algorithm proceeds iterating the following two steps until convergence.

**E-Step.** The E-step requires the calculation of the conditional expectation of (22). Thus, we first need to calculate

\[
\hat{z}_{ig} = \frac{\hat{\pi}_g f \left( y_i | x_i; \hat{\theta}_{Y|g} \right) f \left( x_i; \hat{\theta}_{X|g} \right)}{\sum_{h=1}^{G} \hat{\pi}_h f \left( y_i | x_i; \hat{\theta}_{Y|h} \right) f \left( x_i; \hat{\theta}_{X|h} \right)},
\]

which corresponds to the posterior probability that the unlabeled observation \((X'_i, Y'_i)\)' belongs to the \(g\)th component of the CWM. In addition, if the distribution of \(X\) in component \(g\), \(g = 1, \ldots, G\), is skewed, the following values need to be updated:

\[
\hat{I}_{ig}^X := \mathbb{E}[W_{ig}^X | z_{ig} = 1, x_i, \hat{\phi}_{W_X|g}], \\
\hat{m}_{ig}^X := \mathbb{E}[1/W_{ig}^X | z_{ig} = 1, x_i, \hat{\phi}_{W_X|g}], \\
\hat{n}_{ig}^X := \mathbb{E}[\log(W_{ig}^X) | z_{ig} = 1, x_i, \hat{\phi}_{W_X|g}].
\]

If the distribution of \(Y|x\) in component \(g\), \(g = 1, \ldots, G\), is skewed, then the following values are also updated:

\[
\hat{I}_{ig}^Y := \mathbb{E}[W_{ig}^Y | z_{ig} = 1, y_i, x_i, \hat{\phi}_{W_Y|g}], \\
\hat{m}_{ig}^Y := \mathbb{E}[1/W_{ig}^Y | z_{ig} = 1, y_i, x_i, \hat{\phi}_{W_Y|g}], \\
\hat{n}_{ig}^Y := \mathbb{E}[\log(W_{ig}^Y) | z_{ig} = 1, y_i, x_i, \hat{\phi}_{W_Y|g}].
\]

These updates depend on which of the skewed distributions is considered. However, as shown in Sect. 2.2, the conditional latent variables are all GIG distributed. Therefore, all of the required expectations can be calculated using (2)–(4).

**M-Step.** The M-step involves the maximization of the conditional expectation of the complete-data log-likelihood, allowing all parameters to be updated. Specifically, the update for \(\pi_g\) is

\[
\hat{\pi}_g = \frac{1}{N} \sum_{i=1}^{N} \hat{z}_{ig}.
\]

The parameters related to the distribution of \(X\) in component \(g\), \(g = 1, \ldots, G\), are updated as follows. For skewed distributions, we have the following updates for \(\mu_{X|g}\) and \(\alpha_{X|g}\)
Instead, for the normal distribution, we have

\[
\hat{\mu}_{X|g} = \sum_{i=1}^{N} \frac{\hat{z}_{ig} x_i}{\sum_{i=1}^{N} \hat{z}_{ig} \hat{I}_{X|g} \hat{m}_{ig} X - T_g}
\]

and

\[
\hat{\alpha}_{X|g} = \sum_{i=1}^{N} \frac{\hat{z}_{ig} (\hat{m}_{X|g} - \hat{m}_{ig} X)}{\sum_{i=1}^{N} \hat{z}_{ig} \hat{I}_{X|g} \hat{m}_{ig} X - T_g}
\]

where \( T_g = \sum_{i=1}^{N} \hat{z}_{ig} \hat{I}_{X|g} = (1/T_g) \sum_{i=1}^{N} \hat{z}_{ig} \hat{I}_{ig} X \) and \( \hat{m}_{X|g} = (1/T_g) \sum_{i=1}^{N} \hat{z}_{ig} \hat{m}_{ig} X \).

The update for \( \Sigma_{X|g} \) is

\[
\hat{\Sigma}_{X|g} = \frac{1}{T_g} \sum_{i=1}^{G} \sum_{g=1}^{N} \hat{z}_{ig} \left[ \hat{m}_{ig} X (x_i - \hat{\mu}_{X|g}) (x_i - \hat{\mu}_{X|g})' \right]
\]

\[
- (x_i - \hat{\mu}_{X|g}) \hat{\alpha}_{X|g}' - \hat{\alpha}_{X|g} (x_i - \hat{\mu}_{X|g})'
\]

\[
+ \hat{I}_{ig} \hat{X}_{X|g} \hat{\alpha}_{X|g}'.
\]

Instead, for the normal distribution, we have

\[
\hat{\mu}_{X|g} = \frac{1}{T_g} \sum_{i=1}^{G} \sum_{g=1}^{N} \hat{z}_{ig} x_i, \quad \hat{\Sigma}_{X|g} = \frac{1}{T_g} \sum_{i=1}^{G} \sum_{g=1}^{N} \hat{z}_{ig} (x_i - \hat{\mu}_{X|g}) (x_i - \hat{\mu}_{X|g})'.
\]

The parameters related to the distribution of \( Y|x \) in component \( g, g = 1, \ldots, G \), are updated as follows. For skewed distributions the updates for \( B_g \) and \( \alpha_{Y|g} \) are

\[
\hat{B}_g = \hat{P}_g^{-1} R_g \quad \text{and} \quad \hat{\alpha}_{Y|g} = \frac{1}{T_g \hat{I}_{Y|g}} \left( \sum_{i=1}^{N} \hat{z}_{ig} y_i - R_g' \hat{P}_g^{-1} \sum_{i=1}^{N} \hat{z}_{ig} x_i' \right),
\]

where

\[
\hat{P}_g = \sum_{i=1}^{N} \hat{z}_{ig} \hat{m}_{ig} Y x_i x_i' - \frac{1}{T_g \hat{I}_{Y|g}} \left( \sum_{i=1}^{N} \hat{z}_{ig} x_i' \right) \left( \sum_{i=1}^{N} \hat{z}_{ig} x_i' \right)
\]

and

\[
\hat{R}_g = \sum_{i=1}^{N} \hat{z}_{ig} \hat{m}_{ig} Y x_i y_i' - \frac{1}{T_g \hat{I}_{Y|g}} \left( \sum_{i=1}^{N} \hat{z}_{ig} x_i' \right) \left( \sum_{i=1}^{N} \hat{z}_{ig} y_i' \right),
\]

with \( \hat{I}_{Y|g} = (1/T_g) \sum_{i=1}^{N} \hat{z}_{ig} \hat{I}_{ig} Y \). The update for \( \Sigma_{Y|g} \) is

\[
\hat{\Sigma}_{Y|g} = \frac{1}{T_g} \sum_{i=1}^{N} \hat{z}_{ig} \left[ \hat{m}_{ig} Y (y - \hat{B}_g x_i') (y - \hat{B}_g x_i')' \right]
\]

\[
- (y - \hat{B}_g x_i') \hat{\alpha}_{Y|g} y_i' - \hat{\alpha}_{Y|g} (y - \hat{B}_g x_i')' + \hat{I}_{ig} \hat{Y}_{Y|g} \hat{\alpha}_{Y|g}'.
\]
Conversely, in the case of a multivariate normal distribution, the updates for \( \mathbf{B}_g \) and \( \Sigma_{Y|g} \) are

\[
\hat{B}_g = \left( \sum_{i=1}^{N} \mathbf{z}_{ig} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i=1}^{N} \mathbf{z}_{ig} \mathbf{y}_i \mathbf{y}_i' \right)
\]

and

\[
\hat{\Sigma}_{Y|g} = \frac{1}{T_g} \sum_{i=1}^{N} \mathbf{z}_{ig} (\mathbf{y}_i - \hat{B}_g \mathbf{x}_i)(\mathbf{y}_i - \hat{B}_g \mathbf{x}_i)'.
\]

Finally, if either \( \mathbf{X} \) or \( \mathbf{Y}|\mathbf{x} \) in component \( g, g = 1, \ldots, G \), follows one of the skewed distributions, then there are the additional tailedness and, in the case of the GH distribution, the index parameters that need to be updated. The updates for each distribution are now given.

**Skew-t distribution**

In the case of the ST distribution, we need to update the degrees of freedom \( \nu_g \). This update cannot be obtained in closed form, and thus needs to be performed numerically. For the covariates the update for \( \nu_X|g \) is obtained by solving the equation

\[
\log \left( \frac{\nu_X|g}{2} \right) + 1 - \varphi \left( \frac{\nu_X|g}{2} \right) - \frac{1}{T_g} \sum_{i=1}^{N} \mathbf{z}_{ig} (\hat{m}_{ig} \mathbf{X} + \hat{n}_{ig} \mathbf{X}) = 0,
\]

where \( \varphi(\cdot) \) denotes the digamma function. When the responses are considered, the update for \( \nu_Y|g \) is obtained via (23), after the replacement of \( \nu_X|g, \hat{m}_{ig} \mathbf{X} \) and \( \hat{n}_{ig} \mathbf{X} \) with \( \nu_Y|g, \hat{m}_{ig} \mathbf{Y} \) and \( \hat{n}_{ig} \mathbf{Y} \), respectively.

**Generalized hyperbolic distribution**

For the GH distribution, we would update \( \lambda_g \) and \( \omega_g \). These updates are derived from Browne and McNicholas (2015), and rely on the log convexity of \( K_s(t) \) in both \( s \) and \( t \) (Baricz 2010). For notational purposes in this section, the superscript “prev” is used to distinguish the previous update from the current one. The resulting updates, when \( \mathbf{X} \) is considered, are

\[
\hat{\lambda}_X|g = \overline{\nu}_X|g \hat{\lambda}_X|g \left[ \frac{\partial}{\partial s} \log(K_s(\hat{\omega}_X|g)) \right]_{s = \hat{\omega}_X|g}^{\hat{\omega}_X|g}^{-1},
\]

\[
\hat{\omega}_X|g = \hat{\omega}_X|g - \left[ \frac{\partial}{\partial s} q(\hat{\lambda}_X|g, s) \right]_{s = \hat{\lambda}_X|g}^{\hat{\lambda}_X|g} \left[ \frac{\partial^2}{\partial s^2} q(\hat{\lambda}_X|g, s) \right]_{s = \hat{\lambda}_X|g}^{\hat{\lambda}_X|g}^{-1},
\]

where \( q(\cdot) \) denotes the digamma function. The responses are handled similarly.
where the derivative in (24) is calculated numerically,

\[ q(\lambda_{X|g}, \omega_{X|g}) = \sum_{i=1}^{N} z_{ig} \left[ \log(K_{X|g}(\omega_{X|g})) - \lambda_{X|g} \overline{n}_{X|g} - \frac{1}{2} \omega_{X|g} (\overline{l}_{X|g} + \overline{m}_{X|g}) \right], \]

and \( \overline{n}_{X|g} = (1/T_{g}) \sum_{i=1}^{N} \hat{z}_{ig} \hat{n}_{ig} x. \) When \( Y \) is considered, \( \lambda_{X|g}, \omega_{X|g}, \overline{l}_{X|g}, \overline{m}_{X|g}, \) and \( \overline{n}_{X|g} \) are replaced with \( \lambda_{Y|g}, \omega_{Y|g}, \overline{l}_{Y|g}, \overline{m}_{Y|g}, \) and \( \overline{n}_{Y|g}, \) respectively, where \( \overline{m}_{Y|g} = (1/T_{g}) \sum_{i=1}^{N} \hat{z}_{ig} \hat{m}_{ig} y \) and \( \overline{n}_{Y|g} = (1/T_{g}) \sum_{i=1}^{N} \hat{z}_{ig} \hat{n}_{ig} y. \)

**Variance-gamma distribution**

For the VG distribution, the update for \( \psi_{g} \) cannot be obtained in closed form. When \( X \) is considered, this update is obtained by solving the equation

\[ \log \psi_{X|g} + 1 - \varphi(\psi_{X|g}) + \overline{n}_{X|g} - \overline{l}_{X|g} = 0. \]  

Clearly, when \( Y \) is considered, \( \psi_{X|g}, \overline{n}_{X|g} \) and \( \overline{l}_{X|g} \) are replaced with \( \psi_{Y|g}, \overline{n}_{Y|g} \) and \( \overline{l}_{Y|g}, \) respectively.

**Normal inverse Gaussian distribution**

The NIG distribution is the only having a closed form expression for its tailedness parameter. In detail, when we consider the covariates, the update of \( \kappa_{X|g} \) is

\[ \hat{k}_{X|g} = \frac{1}{\overline{l}_{X|g}}. \]

If the responses are considered, we replace \( \kappa_{X|g} \) and \( \overline{l}_{X|g} \) with \( \kappa_{Y|g} \) and \( \overline{l}_{Y|g}, \) respectively.

**B.1 Initialization of the algorithm**

To initialize the EM algorithm, we followed the approach discussed in Dang et al. (2017). Specifically, the \( z_{ig} \) are initialized in two different ways: 10 times using a random soft initialization and once with a \( k \)-means (hard) initialization. Therefore, for each \( G \), the algorithms are run 11 times until convergence, and the solution producing the highest log-likelihood value is chosen. Notice that, for the \( k \)-means initialization, the initial \( z_{ig} \) are selected from the best \( k \)-means clustering results from 10 random starting values, and it is implemented by using the \texttt{kmeans()} function of the R statistical software (R Core Team 2019).
C Proof of Theorem 3.1

Proof Suppose that

\[ p_{GH-GH}(x, y; \vartheta) = p_{GH-GH}(x, y; \tilde{\vartheta}). \]  \hspace{1cm} (27)

Integrating out \( y \) from each side of (27) yields an equality on the marginal distribution of \( X \), i.e.,

\[ \sum_{g=1}^{G} \pi_g f_{GH}(x; \theta_{X|g}) = \sum_{j=1}^{\tilde{G}} \tilde{\pi}_j f_{GH}(x; \tilde{\theta}_{X|j}) \]

\[ p_{GH}(x; \pi, \theta_X) = p_{GH}(x; \tilde{\pi}, \tilde{\theta}_X). \]  \hspace{1cm} (28)

where \( \theta_X = \{ \theta_{X|g}; g = 1, \ldots, G \} \), \( \tilde{\theta}_X = \{ \tilde{\theta}_{X|j}; j = 1, \ldots, \tilde{G} \} \), \( \pi = \{ \pi_g; g = 1, \ldots, G \} \) and \( \tilde{\pi} = \{ \tilde{\pi}_j; j = 1, \ldots, \tilde{G} \} \). Dividing the left-hand (right-hand) side of (27) by the left-hand (right-hand) side of (28) leads to

\[ \sum_{g=1}^{\tilde{G}} \pi_g f_{GH}(x; \theta_{X|g}) \frac{p_{GH}(x; \pi, \theta_X)}{\tilde{\pi}_j f_{GH}(x; \tilde{\theta}_{X|j})} = \sum_{j=1}^{\tilde{G}} \tilde{\pi}_j f_{GH}(y|x; \tilde{\theta}_Y|j) \frac{p_{GH}(y|x; \theta_Y)}{p_{GH}(x; \pi, \theta_X)} \]

\[ p_{GH}(y|x; \vartheta) = p_{GH}(y|x; \tilde{\vartheta}). \]  \hspace{1cm} (29)

For each fixed value of \( x \), \( p_{GH}(y|x; \vartheta) \) and \( p_{GH}(y|x; \tilde{\vartheta}) \) are mixtures of \( dY \)-variate GH distributions for \( Y \) (see Browne and McNicholas 2015).

Now, recall from Sect. 3.1 that the location parameter \( \mu_{Y|g} \) of the \( dY \)-variate GH distribution of \( Y \) in the \( g \)th mixture component is related to the covariates \( X \), through the regression coefficients \( B_g \) by the relation \( B_g' x^* \), \( g = 1, \ldots, G \). Define the set of all covariate points \( x \) which can be used to distinguish different regression coefficients \( B_g \) by different values of \( B_g' x^* \), i.e.

\[ \mathcal{X} := \{ x \in \mathbb{R}^{dx} : \forall g, s \in \{1, \ldots, G\} \text{ and } j, t \in \{1, \ldots, \tilde{G}\}, \]

\[ B_g' x^* = B_s' x^* \rightarrow B_g = B_s, \]

\[ B_g' x^* = \tilde{B}_j' x^* \rightarrow B_g = \tilde{B}_j, \]

\[ \tilde{B}_j' x^* = \tilde{B}_t' x^* \rightarrow \tilde{B}_j = \tilde{B}_t \}. \]

Note that \( \mathcal{X} \) is complement of a finite union of hyperplanes of \( \mathbb{R}^{dx} \). Therefore,

\[ \int_{\mathcal{X}} p_{GH}(x; \pi, \theta_X) dx = 1. \]
For $x \in \mathcal{X}$, all $\left\{ B^\prime_g x^*, \Sigma Y_{[g]}, \alpha Y_{[g]}, \lambda Y_{[g]}, \omega Y_{[g]} \right\}$, $g = 1, \ldots, G$, are pairwise distinct because all $\left\{ B_g, \Sigma Y_{[g]}, \alpha Y_{[g]}, \lambda Y_{[g]}, \omega Y_{[g]} \right\}$, $g = 1, \ldots, G$, are pairwise distinct for the hypothesis of the theorem. As mentioned above, for each fixed value of $x$, $p_{GH}(y|x; \vartheta)$ is a mixture of $d_Y$-variate GH distributions, which being identifiable (Browne and McNicholas 2015) implies that $G = \tilde{G}$ and that, for each $g \in \{1, \ldots, G\}$, there exists a $j \in \{1, \ldots, G\}$ such that

$$B_g = \tilde{B}_j, \quad \Sigma Y_{[g]} = \tilde{\Sigma}_Y j, \quad \alpha Y_{[g]} = \tilde{\alpha}_Y j, \quad \lambda Y_{[g]} = \tilde{\lambda}_Y j, \quad \omega Y_{[g]} = \tilde{\omega}_Y j$$

and

$$\frac{\pi_g f_{GH}(x; \theta X_{[g]})}{p_{GH}(x; \pi, \theta X)} = \frac{\tilde{\pi}_j f_{GH}(x; \tilde{\theta} X_{[j]})}{p_{GH}(x; \tilde{\pi}, \tilde{\theta} X)}. \quad (30)$$

Now, based on (28), the equality in (30) simplifies to

$$\pi_g f_{GH}(x; \theta X_{[g]}) = \tilde{\pi}_j f_{GH}(x; \tilde{\theta} X_{[j]}), \quad \forall \ x \in \mathcal{X}. \quad (31)$$

Integrating (31) over $x \in \mathcal{X}$ yields $\pi_g = \tilde{\pi}_j$. Therefore, the condition (31) further simplifies as

$$f_{GH}(x; \theta X_{[g]}) = f_{GH}(x; \tilde{\theta} X_{[j]}), \quad \forall \ x \in \mathcal{X}.$$}

The equalities $\mu_{X_{[g]}} = \tilde{\mu}_{X_{[j]}}, \Sigma X_{[g]} = \tilde{\Sigma}_X j, \alpha X_{[g]} = \tilde{\alpha}_X j, \lambda X_{[g]} = \tilde{\lambda}_X j$, and $\omega X_{[g]} = \tilde{\omega}_X j$ simply arise from the identifiability of the $d_X$-variate GH distribution, and this completes the proof. 

References

Aas K, Hobæk Haff I (2005) NIG and skew student’s t: two special cases of the generalised hyperbolic distribution. Appl Res Dev Res Rep
Andrews JL, McNicholas PD (2011) Extending mixtures of multivariate t-factor analyzers. Stat Comput 21(3):361–373
Andrews JL, McNicholas PD (2012) Model-based clustering, classification, and discriminant analysis via mixtures of multivariate $t$-distributions: the $t$ EIGEN family. Stat Comput 22(5):1021–1029
Azzalini A (2020) The R package sn: the skew-normal and related distributions such as the skew-$t$ (version 1.6-1). Università di Padova, Italia. http://azzalini.stat.unipd.it/SN
Baricz Á (2010) Turán type inequalities for some probability density functions. Stud Sci Math Hung 47(2):175–189
Berta P, Ingrassia S, Punzo A, Vittadini G (2016) Multilevel cluster-weighted models for the evaluation of hospitals. METRON 74(3):275–292
Browne RP, McNicholas PD (2015) A mixture of generalized hyperbolic distributions. Can J Stat 43(2):176–198
Chamroukhi F (2017) Skew t mixture of experts. Neurocomputing 266:390–408
Chen L, Pourahmadi M, Maadooliat M (2014) Regularized multivariate regression models with skew-t error distributions. J Stat Plan Inference 149:125–139
Crawford SL (1994) An application of the Laplace method to finite mixture distributions. J Am Stat Assoc 89(425):259–267
Dang UJ, Browne RP, McNicholas PD (2015) Mixtures of multivariate power exponential distributions. Biometrics 71(4):1081–1089
Dang UJ, Punzo A, McNicholas PD, Ingrassia S, Browne RP (2017) Multivariate response and parsimony for Gaussian cluster-weighted models. J Classif 34(1):4–34
Dang UJ, Gallaugher MP, Browne RP, McNicholas PD (2019) Model-based clustering and classification using mixtures of multivariate skewed power exponential distributions. arXiv preprint arXiv:1907.01938
Dayton CM, Macready GB (1988) Concomitant-variable latent-class models. J Am Stat Assoc 83(401):173–178
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39(1):1–38
DeSarbo WS, Cron WL (1988) A maximum likelihood methodology for clusterwise linear regression. J Classif 5(2):249–282
Di Mari R, Bakk Z, Punzo A (2020) A random-covariate approach for distal outcome prediction with latent class analysis. Struct Equ Model 27(3):351–368
Doğru FZ, Arslan O (2017) Parameter estimation for mixtures of skew Laplace normal distributions and application in mixture regression modeling. Commun Stat Theory Methods 46(21):10879–10896
Ferreira CS, Lachos VH, Bolfarine H (2015) Inference and diagnostics in skew scale mixtures of normal regression models. J Stat Comput Simul 85(3):517–537
Frimpong EY, Gage TB, Stratton H (2008) Identifiability of bivariate mixtures: an application to infant mortality models. PhD thesis, Citeseer
Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, New York
Frühwirth-Schnatter S, Pyne S (2010) Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. Biostatistics 11(2):317–336
Galimberti G, Soffritti G (2020) A note on the consistency of the maximum likelihood estimator under multivariate linear cluster-weighted models. Stat Probab Lett 157:1089630
Gallaugher MPB, McNicholas PD (2017) A matrix variate skew-t distribution. Stat 6(1):160–170
Gallaugher MPB, McNicholas PD (2019) Three skewed matrix variate distributions. Statist Probab Lett 145:103–109
Gershenfeld N (1997) Nonlinear inference and cluster-weighted modeling. Ann N Y Acad Sci 808(1):18–24
Göncü A, Yang H (2016) Variance-gamma and normal-inverse Gaussian models: goodness-of-fit to Chinese high-frequency index returns. North Am J Econ Finance 36:279–292
Hennig C (2000) Identifiability of models for clusterwise linear regression. J Classif 17(2):273–296
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
Hung W-L, Chang-Chien S-J (2017) Learning-based EM algorithm for normal-inverse Gaussian mixture model with application to extrasolar planets. J Appl Stat 44(6):978–999
Ingrassia S, Minotti SC, Vittadini G (2012) Local statistical modeling via the cluster-weighted approach with elliptical distributions. J Classif 29(3):363–401
Ingrassia S, Minotti SC, Punzo A (2014) Model-based clustering via linear cluster-weighted models. Comput Stat Data Anal 71:159–182
Ingrassia S, Punzo A, Vittadini G, Minotti SC (2015) The generalized linear mixed cluster-weighted model. J Classif 32(1):85–113
Ingrassia S, Punzo A (2016) Decision boundaries for mixtures of regressions. J Korean Stat Soc 45(2):295–306
Jorgensen B (2012) Statistical properties of the generalized inverse Gaussian distribution, vol 9. Springer, New York
Karlis D, Santourian A (2009) Model-based clustering with non-elliptically contoured distributions. Stat Comput 19(1):73–83
Kim N-H, Browne R (2019) Subspace clustering for the finite mixture of generalized hyperbolic distributions. Adv Data Anal Classif 13(3):641–661
Lee S, McLachlan GJ (2014) Finite mixtures of multivariate skew t-distributions: some recent and new results. Stat Comput 24:181–202
Lin TI (2009) Maximum likelihood estimation for multivariate skew normal mixture models. J Multivar Anal 100(2):257–265
Lin TI (2010) Robust mixture modeling using multivariate skew t distributions. Stat Comput 20(3):343–356
Lin T, McNicholas PD, Hsiu JH (2014) Capturing patterns via parsimonious t mixture models. Statist Probab Lett 88:80–87
Multivariate cluster weighted models using skewed… 123

Mazza A, Punzo A, Ingrassia S (2018) flexCWM: a flexible framework for cluster-weighted models. J Stat Softw 86(2):1–30

McNeil AJ, Frey R, Embrechts P (2005) Quantitative risk management: concepts, techniques and tools. Princeton University Press, Princeton

McNicholas PD (2016a) Mixture model-based classification. Chapman & Hall/CRC Press, Boca Raton

McNicholas PD (2016b) Model-based clustering. J Classif 33(3):331–373

McNicholas SM, McNicholas PD, Browne RP (2017) A mixture of variance-gamma factor analyzers. In: Ahmed SE (ed) Big and complex data analysis, contributions to statistics. Springer, Cham, pp 369–385

Murphy K, Murphy TB (2020) NoEClust: Gaussian parsimonious clustering models with covariates and a noise component. R package version 1.3.3. https://cran.r-project.org/package=NoEClust

Murray SM, McNicholas PD, Browne RB (2014a) Mixtures of skew-t factor analyzers. Comput Stat Data Anal 77:326–335

Murray SM, McNicholas PD, Browne RB (2014b) A mixture of common skew-t factor analyzers. Stat 3(1):68–82

Peel D, McLachlan GJ (2000) Robust mixture modelling using the t distribution. Stat Comput 10(4):339–348

Počuča N, Jevtić P, McNicholas PD, Miljkovic T (2020) Modeling frequency and severity of claims with the zero-inflated generalized cluster-weighted models. Math Econ Insur

Punzo A (2014) Flexible mixture modelling with the polynomial Gaussian cluster-weighted model. Stat Model 14(3):257–291

Punzo A, Ingrassia S (2015) Parsimonious generalized linear Gaussian cluster-weighted models. In: Morlini I, Minerwa T, Vichi M (eds) Advances in statistical models for data analysis, studies in classification, data analysis and knowledge organization. Springer, Switzerland, pp 201–209

Punzo A, Ingrassia S (2016) Clustering bivariate mixed-type data via the cluster-weighted model. Comput Statist 31(3):989–1013

Punzo A, Bagnato L (2021) The multivariate tail-inflated normal distribution and its application in finance. J Stat Comput Simul 91(1):1–36

Punzo A, Ingrassia S, Maruotti A (2018) Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. Stat Med 37(19):2797–2808

Punzo A, Ingrassia S, Maruotti A (2021) Multivariate hidden Markov regression models: random covariates and heavy-tailed distributions. Stat Pap 62(3):1519–1555

Pyne S, Hu X, Wang K, Rossin E, Lin T-I, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA et al (2009) Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci 106(21):8519–8524

R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Soffritti G, Galimberti G (2011) Multivariate linear regression with non-normal errors: a solution based on mixture models. Stat Comput 21(4):523–536

Steane MA, McNicholas PD, Yada R (2012) Model-based classification via mixtures of multivariate t-factor analyzers. Commun Stat Simul Comput 41(4):510–523

Subedi S, Punzo A, Ingrassia S, McNicholas PD (2013) Clustering and classification via cluster-weighted factor analyzers. Adv Data Anal Classif 7(1):5–40

Subedi S, Punzo A, Ingrassia S, McNicholas PD (2015) Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. Stat Methods Appl 24(4):623–649

Tiedeman DV (1955) On the study of types. In: Sells SB (ed) Symposium on pattern analysis. Air University, U.S.A.F. School of Aviation Medicine, Randolph Field, Texas

Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York

Tomarchio SD, McNicholas PD, Punzo A (2021) Matrix normal cluster-weighted models. J Classif 38(3)

Tortora C, Browne RP, ElSherbiny A, Franczak BC, McNicholas PD (2021) Model-based clustering, classification, and discriminant analysis using the generalized hyperbolic distribution: MixGHD R package. J Stat Softw 98(3):1–24

Vrbik I, McNicholas PD (2012) Analytic calculations for the EM algorithm for multivariate skew-t mixture models. Statist Probab Lett 82(6):1169–1174
Vrbik I, McNicholas PD (2014) Parsimonious skew mixture models for model-based clustering and classification. Comput Stat Data Anal 71:196–210
Wang K, Ng SK, McLachlan GJ (2009) Multivariate skew t mixture models: applications to fluorescence-activated cell sorting data. In: Digital image computing: techniques and applications. IEEE, pp 526–531
Wolfe JH (1965) A computer program for the maximum likelihood analysis of types, technical bulletin. U.S. Naval Personnel Research Activity, pp. 65–15
Zarei S, Mohammadpour A, Ingrassia S, Punzo A (2019) On the use of the sub-Gaussian $\alpha$-stable distribution in the cluster-weighted model. Iran J Sci Technol Trans A Sci 43(3):1059–1069

Publisher’s Note  Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.