Model-based clustering via linear cluster-weighted models

S. Ingrassia\textsuperscript{a,*}, S.C. Minotti\textsuperscript{b}, A. Punzo\textsuperscript{a}

\textsuperscript{a}Dipartimento di Economia e Impresa, Università di Catania, Corso Italia 55, 95129 Catania (Italy)
\textsuperscript{b}Dipartimento di Statistica, Università di Milano-Bicocca (Italy)

Abstract

A novel family of twelve mixture models, all nested in the linear \( t \) cluster-weighted model (CWM), is introduced for model-based clustering. The linear \( t \) CWM was recently presented as a robust alternative to the better known linear Gaussian CWM. The proposed family of models provides a unified framework that also includes the linear Gaussian CWM as a special case. Maximum likelihood parameter estimation is carried out within the EM framework, and both the BIC and ICL are used for model selection. A simple and effective hierarchical random initialization is also proposed for the EM algorithm. The novel model-based clustering technique is illustrated in some applications to real data.

Keywords:
Cluster-weighted model, Mixture models, Model-based clustering, Multivariate \( t \) distribution.

2000 MSC: 62H30, 62H99

1. Introduction

In direct applications of finite mixture models (see Titterington et al., 1985, pp. 2–3), we assume that each mixture-component represents a group (or cluster) in the original data. The term “model-based clustering” has been used to

*Corresponding author

Email addresses: s.ingrasia@unict.it (S. Ingrassia), simona.minotti@unimib.it (S.C. Minotti), antonio.punzo@unict.it (A. Punzo)
describe the adoption of mixture models for clustering or, more often, to describe the use of a family of mixture models for clustering (see Fraley & Raftery 1998 and McLachlan & Basford 1988). An overview of mixture models is given in Everitt & Hand (1981), Titterington et al. (1983), McLachlan & Peel (2000), and Frühwirth-Schnatter (2006).

This paper focuses on data arising from a real-valued random vector \((Y, X')' : \Omega \rightarrow \mathbb{R}^{d+1}\), having joint density \(p(y, x)\). Standard model-based clustering techniques assume that \(\Omega\) can be partitioned into \(G\) groups \(\Omega_1, \ldots, \Omega_G\). As for finite mixtures of linear regressions (see, e.g., Leisch 2004 and Frühwirth-Schnatter 2006, Chapter 8) we assume that, for each \(\Omega_g\), the dependence of \(Y\) on \(x\) can be modeled by

\[
Y = \mu(x; \beta_g) + \varepsilon_g = \beta_{0g} + \beta_{1g}'x + \varepsilon_g,
\]

where \(\beta_g = (\beta_{0g}, \beta_{1g})'\), \(\mu(x; \beta_g) = E(Y | X = x, \Omega_g)\) is the linear regression function and \(\varepsilon_g\) is the error variable, independent with respect to \(X\), with zero mean and finite constant variance \(\sigma^2_g\), \(g = 1, \ldots, G\). In the mixture framework, the cluster weighted model (CWM; Gershenfeld, 1997), with equation

\[
p(y, x) = \sum_{g=1}^{G} \pi_g p(y, x|\Omega_g) = \sum_{g=1}^{G} \pi_g p(y|x, \Omega_g) p(x|\Omega_g), \quad (1)
\]

constitutes a reference approach to model the joint density. In (1), normality of both \(p(y|x, \Omega_g)\) and \(p(x|\Omega_g)\) is commonly assumed (Gershenfeld 1997, Schöner 2000, and Schöner & Gershenfeld 2001). Alternatively, Ingrassia et al. (2012) propose the use of the \(t\) distribution, which provides more robust fitting for groups of observations with longer than normal tails or noise data (see, e.g., Zellner 1976, Lange et al. 1989, Peel & McLachlan 2000, McLachlan & Peel 2000, Chapter 7, Chatzis & Varvarigou 2008 and Greselin & Ingrassia 2010).
In particular, the authors consider

$$p(y|\mathbf{x}, \Omega) = h_t(y|\mathbf{x}; \xi_g, \zeta_g) = \frac{\Gamma\left(\frac{\zeta_g + 1}{2}\right)}{(\pi \zeta_g \sigma_g^2)^{\frac{\zeta_g}{2}}} \left\{1 + \delta \left[y, \mu(\mathbf{x}; \beta_g); \sigma_g^2\right]\right\}^{\frac{\zeta_g}{2}}$$

(2)

$$p(\mathbf{x}|\Omega_g) = h_{td}(\mathbf{x}; \vartheta_g, \nu_g) = \frac{\Gamma\left(\frac{\nu_g + d}{2}\right)}{(\pi \nu_g)^{\frac{d}{2}} \left[1 + \delta (\mathbf{x}, \mu_g; \Sigma_g)\right]^{\frac{\nu_g + d}{2}}},$$

(3)

with $\xi_g = \{\beta_g, \sigma_g^2\}$, $\vartheta_g = \{\mu_g, \Sigma_g\}$, $\delta \left[y, \mu(\mathbf{x}; \beta_g); \sigma_g^2\right] = [y - \mu(\mathbf{x}; \beta_g)]^2 / \sigma_g^2$, and $\delta (\mathbf{x}, \mu_g; \Sigma_g) = (\mathbf{x} - \mu_g)^T \Sigma_g^{-1} (\mathbf{x} - \mu_g)$. Thus, (2) is the density of a (generalized) univariate $t$ distribution, with location parameter $\mu(\mathbf{x}; \beta_g)$, scale parameter $\sigma_g^2$, and $\zeta_g$ degrees of freedom, while (3) is the density of a multivariate $t$ distribution with location parameter $\mu_g$, inner product matrix $\Sigma_g$, and $\nu_g$ degrees of freedom. By substituting (2) and (3) into (1), we obtain the linear $t$ CWM

$$p(y, \mathbf{x}; \psi_{\text{CWM}}) = \sum_{g=1}^{G} \pi_g h_t(y|\mathbf{x}; \xi_g, \zeta_g) h_{td}(\mathbf{x}; \vartheta_g, \nu_g),$$

(4)

where the set of all unknown parameters is denoted by $\psi_{\text{CWM}} = \{\psi_1, \ldots, \psi_G\}$, with $\psi_g = \{\pi_g, \xi_g, \zeta_g, \vartheta_g, \nu_g\}$.

In this paper, we introduce a family of twelve linear CWMs obtained from (4) by imposing convenient component distributional constraints. If $\zeta_g, \nu_g \to \infty$, the more famous linear Gaussian (normal) CWM is obtained as a special case. The resulting models are easily interpretable and appropriate for describing various practical situations. In particular, they also allow one to infer if the group-structure of the data is due to the contribution of $\mathbf{X}$, $\mathbf{Y}|\mathbf{X}$, or both.

The paper is organized as follows. In Section 2 we recall model-based clustering according to the CW approach, and give some preliminary results. In Section 3 we introduce the novel family of models. Model fitting in the EM paradigm is presented in Section 4; related computational aspects are addressed in Section 5 and model selection is discussed in Section 6. Some applications to real data are presented in Section 7. In Section 8 we give a summary of the paper and some directions for further research.
2. Preliminary results for model-based clustering

This section recalls some basic ideas on model-based clustering according to the CWM approach and provides some preliminary results that will be useful for definition and justification of our family of models.

Let \((y_1, x_1)', \ldots, (y_N, x_N')\) be a sample of size \(N\) from (4). Once \(\psi\) is estimated (fixed), the posterior probability that the generic unit \((y_n, x_n')\), \(n = 1, \ldots, N\), comes from component \(\Omega_g\) is given by

\[
\tau_{ng} = P(\Omega_g | y_n, x_n; \psi) = \frac{\pi_g h_t(y | x; \xi_g, \zeta_g) h_t(x; \vartheta_g, \nu_g)}{p(y, x; \psi)} , \quad g = 1, \ldots, G.
\] (5)

These probabilities, which depend on both marginal and conditional densities, represent the basis for clustering and classification.

The following two propositions, which generalize some results given in Ingrassia et al. (2012), require the preliminary definition of

\[
p(y | x; \pi, \xi, \zeta) = \sum_{g=1}^{G} \pi_g h_t(y | x; \xi_g, \zeta_g)
\] (6)

\[
p(x; \pi, \vartheta, \nu) = \sum_{g=1}^{G} \pi_g h_t(x; \vartheta_g, \nu_g),
\] (7)

which respectively correspond to a finite mixture of linear \(t\) regressions and a finite mixture of multivariate \(t\) distributions \((\pi = \{\pi_1, \ldots, \pi_{G-1}\}, \xi = \{\xi_1, \ldots, \xi_G\}, \zeta = \{\zeta_1, \ldots, \zeta_G\}, \vartheta = \{\vartheta_1, \ldots, \vartheta_G\}, \text{and} \ \nu = \{\nu_1, \ldots, \nu_G\})).

**Proposition 1.** Given \(\pi, \vartheta, \text{and} \ \nu, \) if \(h_t(y | x; \xi_1, \zeta_1) = \cdots = h_t(y | x; \xi_G, \zeta_G) = h_t(y | x; \xi, \zeta)\), then models (4) and (7) generate the same posterior probabilities.

**Proof.** If the component conditional densities do not depend on \(\Omega_g\), then the
posterior probabilities from the linear t CWM in (4) can be written as

$$
\tau_{ng} = \frac{\pi_{g} h_{t}(y|x; \xi, \zeta) h_{t_{d}}(x; \vartheta_{g}, \nu_{g})}{\sum_{j=1}^{G} \pi_{j} h_{t}(y|x; \xi_{j}, \zeta_{j}) h_{t_{d}}(x; \vartheta_{j}, \nu_{j})} = \frac{\pi_{g} h_{t_{d}}(x; \vartheta_{g}, \nu_{g})}{\sum_{j=1}^{G} \pi_{j} h_{t_{d}}(x; \vartheta_{j}, \nu_{j})}
$$

which correspond to the posterior probabilities of model (7).

**Proposition 2.** Given $\pi$, $\xi$, and $\zeta$, if $h_{t_{d}}(x; \vartheta_{1}, \nu_{1}) = \cdots = h_{t_{d}}(x; \vartheta_{G}, \nu_{G}) = h_{t_{d}}(x; \vartheta, \nu)$, then models (4) and (6) generate the same posterior probabilities.

**Proof.** If the component marginal densities do not depend on $\Omega_{g}$, then the posterior probabilities from the linear t CWM in (4) can be written as

$$
\tau_{ng} = \frac{\pi_{g} h_{t}(y|x; \xi_{g}, \zeta_{g}) h_{t_{d}}(x; \vartheta_{g}, \nu_{g})}{\sum_{j=1}^{G} \pi_{j} h_{t}(y|x; \xi_{j}, \zeta_{j}) h_{t_{d}}(x; \vartheta_{j}, \nu_{j})} = \frac{\pi_{g} h_{t}(y|x; \xi_{g}, \zeta_{g})}{\sum_{j=1}^{G} \pi_{j} h_{t}(y|x; \xi_{j}, \zeta_{j})}
$$

which correspond to the posterior probabilities of model (6).

Note that the results of Proposition 1 and 2 are not restricted to the t distribution; in fact, they can be easily extended to the general CWM in (1).

3. **The family of linear CWMs**

This section introduces the novel family of mixture models obtained from the linear t CWM. In (4), let us consider:

- component conditional densities $h_{t}$ having the same parameters for all $\Omega_{g}$,
- component marginal densities $h_{t_{d}}$ having the same parameters for all $\Omega_{g}$,
• degrees of freedom $\zeta_g$ tending to infinity for each $\Omega_g$, and
• degrees of freedom $\nu_g$ tending to infinity for each $\Omega_g$.

By combining such constraints, we obtain twelve parsimonious and easily interpretable linear CWMs that are appropriate for describing various practical situations; they are schematically presented in Table 1 along with the number of parameters characterizing each component of the CW decomposition. For instance, if $\nu_g, \zeta_g \to \infty$ for each $\Omega_g$, we are assuming a normal distribution for the component conditional and marginal densities; furthermore, we can assume different linear models (in terms of $\beta_g$ and $\sigma^2_g$) in each cluster while keeping the density of $X$ equal between clusters. From a notational viewpoint, this leads to a linear CWM that we have simply denoted as $NN-EV$: the first two letters represent the distribution of $X|\Omega_g$ and $Y|X, \Omega_g$ ($N \equiv \text{Normal and } t \equiv t$), respectively, while the second two denote the distribution constraint between clusters ($E \equiv \text{Equal and } V \equiv \text{Variable}$) for $X|\Omega_g$ and $Y|X, \Omega_g$, respectively.

Only two of the models given in Table 1, $NN-\text{VV}$ and $tt-\text{VV}$, have been developed previously; the former corresponds to the linear Gaussian CWM of Gershenfeld (1997), while the latter coincides with the linear $t$ CWM of Ingrassia et al. (2012). Furthermore, in principle there are sixteen models arising from the combination of the aforementioned constraints; nevertheless, four of them — those which should be denoted as EE — do not make sense. Indeed, they lead to a single cluster regardless of the value of $G$. Finally, we remark that when $G = 1$, $VV \equiv VE \equiv EV$ regardless of the chosen distribution.

4. Estimation via the EM algorithm

The EM algorithm (Dempster et al., 1977) is the standard tool for maximum likelihood (ML) estimation of the parameters for mixture models. This section describes the EM algorithm for all the linear CWMs in Table 1.

In the EM framework, the generic observation $(y_n, x_n', z_n', u_n, w_n)'$ is viewed as being incomplete; its complete counterpart is given by $(y_n, x_n', z_n', u_n, w_n)'$, where $z_n$
| Model | $X|\Omega_g$ | $Y|x, \Omega_g$ | Number of free parameters |
|-------|---------------|----------------|--------------------------|
| Identifier | Density | Constraint | Density | Constraint | $X$ | $Y|x$ | weights |
| $tt$-VV | $t$ | Variable | $t$ | Variable | $G \left( d + \frac{d(d+1)}{2} + 1 \right)$ | + | $G \left( d + 3 \right)$ | + | $G - 1$ |
| $tt$-VE | $t$ | Variable | $t$ | Equal | $G \left( d + \frac{d(d+1)}{2} + 1 \right)$ | + | $d + 3$ | + | $G - 1$ |
| $tt$-EV | $t$ | Equal | $t$ | Variable | $d + \frac{d(d+1)}{2} + 1$ | + | $G \left( d + 3 \right)$ | + | $G - 1$ |
| $NN$-VV | Normal | Variable | Normal | Variable | $G \left( d + \frac{d(d+1)}{2} \right)$ | + | $G \left( d + 2 \right)$ | + | $G - 1$ |
| $NN$-VE | Normal | Variable | Normal | Equal | $G \left( d + \frac{d(d+1)}{2} \right)$ | + | $d + 2$ | + | $G - 1$ |
| $NN$-EV | Normal | Equal | Normal | Variable | $d + \frac{d(d+1)}{2}$ | + | $G \left( d + 2 \right)$ | + | $G - 1$ |
| $tN$-VV | $t$ | Variable | Normal | Variable | $G \left( d + \frac{d(d+1)}{2} + 1 \right)$ | + | $G \left( d + 2 \right)$ | + | $G - 1$ |
| $tN$-VE | $t$ | Variable | Normal | Equal | $G \left( d + \frac{d(d+1)}{2} + 1 \right)$ | + | $d + 2$ | + | $G - 1$ |
| $tN$-EV | $t$ | Equal | Normal | Variable | $d + \frac{d(d+1)}{2} + 1$ | + | $G \left( d + 2 \right)$ | + | $G - 1$ |
| $Nt$-VV | Normal | Variable | $t$ | Variable | $G \left( d + \frac{d(d+1)}{2} \right)$ | + | $G \left( d + 3 \right)$ | + | $G - 1$ |
| $Nt$-VE | Normal | Variable | $t$ | Equal | $G \left( d + \frac{d(d+1)}{2} \right)$ | + | $d + 3$ | + | $G - 1$ |
| $Nt$-EV | Normal | Equal | $t$ | Variable | $d + \frac{d(d+1)}{2}$ | + | $G \left( d + 3 \right)$ | + | $G - 1$ |

Table 1: Overview of linear CWMs. In “model identifier”, the first and second letters represent, respectively, the density of $X|\Omega_g$ and $Y|x, \Omega_g$ (here $N \equiv $Normal), while the third and fourth letters indicate, respectively, if $h_{td}(x; \varphi_g, \nu_g)$ and $h_t(y|x; \xi_g, \zeta_g)$ are assumed to be Equal$\equiv$E or Variable$\equiv$V between groups.
is the component-label vector in which \( z_{ng} = 1 \) if \( (y_n, x'_n)' \) comes from the \( j \)th component and \( z_{ng} = 0 \) otherwise. In other words, it is convenient to view the observation augmented by \( z_n \) as still being incomplete and introduce into the complete observation the additional missing values \( u_n \) and \( v_n \), which are defined so that \( z_{ng} = 1 \). In particular, from the standard theory of the (multivariate) \( t \) distribution, \( N \) independent draws from \( t (\mu (x_n; \beta_g), \sigma^2_g, \zeta_g) \) and \( t_d (\mu_g, \Sigma_g, \nu_g) \) can be respectively described, by compounding, as

\[
Y_n \mid x_n, v_n, z_{ng} = 1 \sim \text{i.i.d. } N (\mu (x_n; \beta_g), \frac{\sigma^2_g}{v_n})
\]

(8)

\[
V_n \mid z_{ng} = 1 \sim \text{i.i.d. } \Gamma (\frac{\zeta_g}{2}, \frac{\zeta_g}{2})
\]

(9)

\[
X_n \mid u_n, z_{ng} = 1 \sim \text{i.i.d. } N (\mu_g, \frac{\Sigma_u}{u_n})
\]

(10)

\[
U_n \mid z_{ng} = 1 \sim \text{i.i.d. } \Gamma (\nu_g, \frac{\nu_g}{2})
\]

(11)

for \( n = 1, \ldots, N \). Because of the conditional structure of the complete-data model given by distributions (8), (9), (10), and (11), the complete-data log-likelihood can be decomposed as

\[
l_c (\psi) = l_{1c} (\pi) + l_{2c} (\xi) + l_{3c} (\zeta) + l_{4c} (\vartheta) + l_{5c} (\nu),
\]

(12)

where

\[
l_{1c} (\pi) = \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} \ln \pi_g
\]

\[
l_{2c} (\xi) = \frac{1}{2} \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} \left\{ -\ln (2\pi) + \ln v_n - \ln \sigma^2_g - v_n \delta [y_n, \mu (x_n; \beta_g); \sigma^2_g] \right\}
\]

\[
l_{3c} (\zeta) = \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} \left[ -\ln \Gamma \left( \frac{\zeta_g}{2} \right) + \frac{\zeta_g}{2} \ln \frac{\zeta_g}{2} + \frac{\zeta_g}{2} (\ln v_n - v_n) - \ln v_n \right]
\]

\[
l_{4c} (\vartheta) = \frac{1}{2} \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} \left[ -d \ln (2\pi) + d \ln u_n - \ln |\Sigma_g| - u_n \delta (x_n, \mu_g; \Sigma) \right]
\]

\[
l_{5c} (\nu) = \sum_{n=1}^{N} \sum_{g=1}^{G} z_{ng} \left[ -\ln \Gamma \left( \frac{\nu_g}{2} \right) + \frac{\nu_g}{2} \ln \frac{\nu_g}{2} + \frac{\nu_g}{2} (\ln u_n - u_n) - \ln u_n \right].
\]
4.1. E-step

The E-step on the \((k+1)\)th iteration of the EM algorithm requires the calculation of

\[
Q(\psi; \tilde{\psi}^{(k)}) = E_{\psi^{(k)}} \left[ l_c(\psi) \left| (y_1, x'_1), \ldots, (y_n, x'_n) \right. \right].
\] (13)

In order to do this, we need to calculate

\[
E_{\psi^{(k)}}(Z_{ng} | y_n, x_n),
\]

\[
E_{\psi^{(k)}}(V_n | y_n, x_n, z_n),
\]

\[
E_{\psi^{(k)}}(\tilde{V}_n | y_n, x_n, z_n),
\]

\[
E_{\psi^{(k)}}(U_n | x_n, z_n),
\]

and

\[
E_{\psi^{(k)}}(\tilde{U}_n | x_n, z_n),
\]

for \(n = 1, \ldots, N\) and \(g = 1, \ldots, G\), where \(\tilde{U}_n = \ln U_n\) and \(\tilde{V}_n = \ln V_n\).

It follows that

\[
E_{\psi^{(k)}}(Z_{ng} | y_n, x_n) = \tau_{ng}^{(k)}
\]

\[
= \frac{\pi_g^{(k)} h_t(y_n | x_n; \zeta_g^{(k)}, \zeta_g^{(k)} h_t(x_n; \theta_g^{(k)}, \nu_g^{(k)})}{p(y_n, x_n; \psi^{(k)})},
\] (14)

which corresponds, in analogy to (5), to the posterior probability that \((y_n, x'_n)\) belongs to the \(j\)th component of the mixture, using the current fit \(\tilde{\psi}^{(k)}\) for \(\psi\) \((n = 1, \ldots, N\) and \(g = 1, \ldots, G\)).

Because the gamma distribution is the conjugate prior distribution for both \(U_n\) and \(V_n\), it is not difficult to show, respectively, that

\[
E_{\psi^{(k)}}(V_n | y_n, x_n, z_n = 1) = v_{ng}^{(k)}
\]

\[
= \frac{\zeta_j^{(k)} + 1}{\zeta_g^{(k)} + \delta(y_n, \mu_g^{(k)}; \sigma_g^{2(r)})},
\] (15)

and

\[
E_{\psi^{(k)}}(U_n | y_n, x_n, z_n = 1) = u_{ng}^{(k)}
\]

\[
= \frac{\nu_j^{(k)} + d}{\nu_g^{(k)} + \delta(x_n, \mu_g^{(k)}; \Sigma_g^{(k)})},
\] (16)

Regarding the last two expectations, from the standard theory on the gamma distribution, we have that

\[
E_{\psi^{(k)}}(\tilde{V}_n | y_n, x_n, z_n = 1) = \tilde{v}_{ng}^{(k)}
\]

\[
= \ln v_{ng}^{(k)} + \psi\left(\frac{\zeta_j^{(k)} + 1}{2}\right) - \psi\left(\frac{\zeta_j^{(k)} + 1}{2}\right).
\] (17)
\[ E_{\psi^{(k)}} \left( \tilde{U}_n | x_n, z_{ng} = 1 \right) = \tilde{u}_{ng}^{(k)} \]

\[ = \ln v_{ng}^{(k)} + \psi \left( \frac{\nu_j^{(k)} + d}{2} \right) - \ln \left( \frac{\nu_j^{(k)} + d}{2} \right) \]

(18)

where \( \psi(s) = [\partial \Gamma(s)/\partial s]/\Gamma(s) \) is the Digamma function.

Using the results from (14) to (17) to calculate (13), we have that

\[ Q \left( \psi; \psi^{(k)} \right) = Q_1 \left( \pi; \psi^{(k)} \right) + Q_2 \left( \xi; \psi^{(k)} \right) + Q_3 \left( \zeta; \psi^{(k)} \right) + Q_4 \left( \vartheta; \psi^{(k)} \right) + Q_5 \left( \nu; \psi^{(k)} \right) \]

(19)

where

\[ Q_1 \left( \pi; \psi^{(k)} \right) = \sum_{n=1}^{N} \sum_{g=1}^{G} \tau_{ng}^{(k)} \ln \pi_g, \]

(20)

\[ Q_2 \left( \xi; \psi^{(k)} \right) = \sum_{n=1}^{N} \sum_{g=1}^{G} \tau_{ng}^{(k)} Q_{2n} \left( \xi_g; \psi^{(k)} \right), \]

(21)

\[ Q_3 \left( \zeta; \psi^{(k)} \right) = \sum_{n=1}^{N} \sum_{g=1}^{G} \tau_{ng}^{(k)} Q_{3n} \left( \zeta_g; \psi^{(k)} \right), \]

(22)

\[ Q_4 \left( \vartheta; \psi^{(k)} \right) = \sum_{n=1}^{N} \sum_{g=1}^{G} \tau_{ng}^{(k)} Q_{4n} \left( \vartheta_g; \psi^{(k)} \right), \]

(23)

\[ Q_5 \left( \nu; \psi^{(k)} \right) = \sum_{n=1}^{N} \sum_{g=1}^{G} \tau_{ng}^{(k)} Q_{5n} \left( \nu_g; \psi^{(k)} \right), \]

(24)

with

\[ Q_{2n} \left( \xi_g; \psi^{(k)} \right) = \frac{1}{2} \left\{ -\ln (2\pi) + \tilde{v}_{ng}^{(k)} - \ln \sigma_g^2 - v_{ng} \delta \left[ y_n, \mu \left( x_n; \beta_g \right); \sigma_g^2 \right] \right\} \]

and

\[ Q_{4n} \left( \vartheta_g; \psi^{(k)} \right) = \frac{1}{2} \left[ -d \ln (2\pi) + d \tilde{\alpha}_{ng}^{(k)} - \ln |\Sigma_g| - u_{ng} \delta \left( x_n, \mu_g; \Sigma_g \right) \right], \]

and where, on ignoring terms not involving \( \zeta_g \) and \( \nu_g \), respectively,

\[ Q_{3n} \left( \zeta_g; \psi^{(k)} \right) = -\ln \Gamma \left( \frac{\zeta_g}{2} \right) + \frac{\zeta_g}{2} \ln \frac{\zeta_g}{2} + \frac{\zeta_g}{2} \left[ \tilde{v}_{ng}^{(k)} - \ln v_{ng}^{(k)} + \sum_{n=1}^{N} \left( \ln v_{ng}^{(k)} - v_{ng}^{(k)} \right) \right] \]

and

\[ Q_{5n} \left( \nu_g; \psi^{(k)} \right) = -\ln \Gamma \left( \frac{\nu_g}{2} \right) + \frac{\nu_g}{2} \ln \frac{\nu_g}{2} + \frac{\nu_g}{2} \left[ \tilde{u}_{ng}^{(k)} - \ln u_{ng}^{(k)} + \sum_{n=1}^{N} \left( \ln u_{ng}^{(k)} - u_{ng}^{(k)} \right) \right]. \]
4.2. M-step

On the M-step at the \((k + 1)\)th iteration of the EM algorithm, it follows from (19) that \(\bar{\pi}^{(k+1)}, \bar{\xi}^{(k+1)}, \bar{\zeta}^{(k+1)}, \bar{\vartheta}^{(k+1)}\), and \(\bar{\nu}^{(k+1)}\) can be computed independently of each other, by separate consideration of (20), (21), (22), (23), and (24), respectively. The solutions for \(\pi_g^{(k+1)}, \xi_g^{(k+1)},\) and \(\vartheta_g^{(k+1)}\) exist in closed form. Only the updates \(\zeta^{(k+1)}_g\) and \(\nu^{(k+1)}_g\) need to be computed iteratively.

Regarding the mixture weights, maximization of \(Q_1(\bar{\pi}; \bar{\psi}^{(k)})\) in (20) with respect to \(\bar{\pi}\), subject to the constraints on those parameters, is obtained by maximizing the augmented function

\[
\sum_{n=1}^{N} \sum_{g=1}^{G} \tau_{ng}^{(k)} \ln \pi_g - \lambda \left( \sum_{g=1}^{G} \pi_g - 1 \right),
\]

where \(\lambda\) is a Lagrangian multiplier. Setting the derivative of equation (25) with respect to \(\pi_g\) equal to zero and solving for \(\pi_g\) yields

\[
\pi_g^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} / n.}{\sum_{n=1}^{N} \tau_{ng}^{(k)} / n.}
\]

With reference to the updated estimates of \(\vartheta_g, g = 1, \ldots, G\), maximization of (23) leads to

\[
\mu_g^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} u_{ng}^{(k)} x_n / \sum_{n=1}^{N} \tau_{ng}^{(k)} u_{ng}^{(k)}}{\sum_{n=1}^{N} \tau_{ng}^{(k)} u_{ng}^{(k)}}
\]

\[
\Sigma_g^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} u_{ng}^{(k)} (x_n - \mu_g^{(k+1)}) (x_n - \mu_g^{(k+1)}) / \sum_{n=1}^{N} \tau_{ng}^{(k)} u_{ng}^{(k)}}{\sum_{n=1}^{N} \tau_{ng}^{(k)} u_{ng}^{(k)}},
\]

where, as motivated in [Kent et al. (1994), Meng & van Dyk (1997), Liu (1997), Liu et al. (1998), and Shoham (2002), among others, the true denominator \(\sum_{n=1}^{N} \tau_{ng}^{(k)}\) of (25) has been changed to yield a significantly faster convergence for the EM algorithm.

Regarding the updated estimates of \(\xi_g, g = 1, \ldots, G\), maximization of (24),
after some algebra, yields

$$\beta^{(k+1)}_{1j} = \left( \sum_{n=1}^{N} \tau^{(k)}_{ng} v^{(k)}_{ng} x_n x'_n \right)^{-1} \left( \sum_{n=1}^{N} \tau^{(k)}_{ng} v^{(k)}_{ng} x_n \right) - \frac{1}{N} \sum_{n=1}^{N} \tau^{(k)}_{ng} v^{(k)}_{ng} \right)$$

$$\sigma^{2(k+1)}_g = \sum_{n=1}^{N} \tau^{(k)}_{ng} v^{(k)}_{ng} \left[ y_n - \left( \beta^{(k+1)}_{1j} + \beta^{(k+1)'}_{1j} x_n \right) \right]^2 \left/ \sum_{n=1}^{N} \tau^{(k)}_{ng} v^{(k)}_{ng} \right)$$

where the denominator of (31) has been modified in line with what was explained for equation (28).

As said before, because we are acting in the most general case in which the degrees of freedom $\zeta_g$ and $\nu_g$ are inferred from the data, we need to numerically solve the equations

$$\sum_{n=1}^{N} \frac{\partial}{\partial \zeta_g} Q_{3n} \left( \zeta_g; \psi^{(k)} \right) = 0$$

and

$$\sum_{n=1}^{N} \frac{\partial}{\partial \nu_g} Q_{5n} \left( \nu_g; \psi^{(k)} \right) = 0,$$

which correspond to finding $\zeta^{(k+1)}_g$ and $\nu^{(k+1)}_g$ as the respective solutions of

$$-\psi \left( \frac{\zeta_g}{2} \right) + \ln \frac{\zeta_g}{2} + 1 + \frac{1}{n_g} \sum_{n=1}^{N} \tau^{(k)}_{ng} \left( \ln v^{(k)}_{ng} - v^{(k)}_{ng} \right) +$$

$$\psi \left( \frac{\zeta_g + 1}{2} \right) - \ln \left( \frac{\zeta_g + 1}{2} \right) = 0$$
and

\[-\psi\left(\frac{\nu_g}{2}\right) + \ln \frac{\nu_g}{2} + 1 + \frac{1}{n_g} \sum_{n}^{N} \tau_{ng}^{(k)} \left(\ln u_{ng}^{(k)} - u_{ng}^{(k)}\right) + \psi\left(\frac{\nu_g^{(k)}}{2} + d\right) - \ln \left(\frac{\nu_g^{(k)} + d}{2}\right) = 0, \quad (35)\]

where \(n_g^{(k)} = \sum_{n}^{G} \tau_{ng}^{(k)}, \quad g = 1, \ldots, G.\)

4.3. Constraints for parsimonious models

In the following, we describe how to impose constraints on the EM algorithm described above in order to obtain parameter estimates for all the models in Table 1. To this end, the itemization given at the beginning of Section 3 will be considered as a benchmark scheme.

4.3.1. Common \(t\) for the component marginal densities

In the case when we constrain all the groups to have a common \(t\) distribution for \(X\), we have that \(\mu_1 = \cdots = \mu_G = \mu, \Sigma_1 = \cdots = \Sigma_G = \Sigma, \) and \(\nu_1 = \cdots = \nu_G = \nu.\) Thus, in the \((k + 1)\)th iteration of the EM algorithm, equations (16) and (18) must be replaced by

\[u_n^{(k)} = \frac{\nu_g^{(k)} + d}{\nu_g^{(k)} + d \left(x_n, \mu^{(k)}; \Sigma^{(k)}\right)}, \quad (36)\]

and

\[\tilde{u}_n^{(k)} = \ln u_n^{(k)} + \psi\left(\frac{\nu_g^{(k)} + d}{2}\right) - \ln \left(\frac{\nu_g^{(k)} + d}{2}\right), \quad (37)\]

respectively.

Furthermore, noting that \(\sum_{g} \tau_{ng} = 1, \) equations (23) and (24) can be rewritten as

\[Q_4 \left(\varphi; \psi^{(k)}\right) = \sum_{n=1}^{N} Q_{4n} \left(\varphi; \psi^{(k)}\right) \quad (38)\]

and

\[Q_5 \left(\nu; \psi^{(k)}\right) = \sum_{n=1}^{N} Q_{5n} \left(\nu; \psi^{(k)}\right), \quad (39)\]
respectively, where

\[
Q_{4n}(\vartheta; \psi^{(k)}) = \frac{1}{2} \left[ -d \ln (2\pi) + d\bar{\nu}_n^{(k)} - \ln |\Sigma| - u_n \delta(x_n, \mu; \Sigma) \right]
\]

and

\[
Q_{5n}(\nu; \psi^{(k)}) = -\ln \Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2} \ln \frac{\nu}{2} + \frac{\nu}{2} \left[ \bar{\nu}_n^{(k)} - \ln u_n^{(k)} + \sum_{n=1}^{N} \left( \ln u_n^{(k)} - u_n^{(k)} \right) \right].
\]

Maximization of (38), with respect to \(\vartheta\), leads to

\[
\mu^{(k+1)} = \frac{\sum_{n=1}^{N} u_n^{(k)} x_n}{\sum_{n=1}^{N} u_n^{(k)}}
\]

(40)

and

\[
\Sigma^{(k+1)} = \frac{\sum_{n=1}^{N} u_n^{(k)} (x_n - \mu^{(k+1)}) (x_n - \mu^{(k+1)})'}{\sum_{n=1}^{N} u_n^{(k)}}.
\]

(41)

For the updating of \(\nu\), we need to numerically solve the equation

\[
\sum_{n=1}^{N} \frac{\partial}{\partial \nu} Q_{5n}(\nu; \psi^{(k)}) = 0,
\]

(42)

which corresponds to finding \(\nu^{(k+1)}\) as the solution of

\[
-\psi\left(\frac{\nu}{2}\right) + \ln \frac{\nu}{2} + 1 + \sum_{n=1}^{N} \left( \ln u_n^{(k)} - u_n^{(k)} \right) + \psi\left(\frac{\nu^{(k)} + d}{2}\right) - \ln \left(\frac{\nu^{(k)} + d}{2}\right) = 0.
\]

(43)

4.3.2. Common \(\mu\) for the component conditional densities

Similarly, in the case when we constrain all the groups to have a common \(\mu\) distribution for \(Y|x\), we have that \(\beta_{11} = \cdots = \beta_{1k} = \beta_1, \beta_{01} = \cdots = \beta_{0k} = \beta_0, \sigma_1^2 = \cdots = \sigma_G^2 = \sigma^2\), and \(\zeta_1 = \cdots = \zeta_G = \zeta\). Thus, in the \((k+1)\)th iteration of the EM algorithm, equations (15) and (17) must be replaced by

\[
v_n^{(k)} = \frac{\zeta^{(k)} + 1}{\zeta_0^{(k)} + \delta \left[ y_n, \mu(x_n; \beta^{(k)}); \sigma^2(r) \right]}
\]

(44)

and

\[
\bar{v}_n^{(k)} = \ln v_n^{(k)} + \psi\left(\frac{\zeta^{(k)} + 1}{2}\right) - \ln \left(\frac{\zeta^{(k)} + 1}{2}\right),
\]

(45)

respectively.
Also, equations (21) and (22) can be rewritten as

\[ Q_2 (\xi; \psi^{(k)}) = \sum_{n=1}^{N} Q_{2n} (\xi; \psi^{(k)}) \] (46)

and

\[ Q_3 (\zeta; \psi^{(k)}) = \sum_{n=1}^{N} Q_{3n} (\zeta; \psi^{(k)}) \] (47)

respectively, where

\[ Q_{2n} (\xi; \psi^{(k)}) = \frac{1}{2} \left\{ - \ln (2\pi) + \bar{v}_n^{(k)} - \ln \sigma^2 - v_n \delta [y_n, \mu (x_n; \beta); \sigma^2] \right\} \]

and

\[ Q_{3n} (\zeta; \psi^{(k)}) = - \ln \Gamma \left( \frac{\zeta}{2} \right) + \frac{\zeta}{2} \ln \frac{\zeta}{2} + \frac{\zeta}{2} \left[ \bar{v}_n^{(k)} - \ln v_n^{(k)} + \sum_{n=1}^{N} \left( \ln v_n^{(k)} - v_n \right) \right]. \]

Maximization of (46), with respect to \( \xi \), leads to the updates

\[
\begin{align*}
\beta_1^{(k+1)} &= \left( \sum_{n=1}^{N} v_n^{(k)} x_n x_n' \right)^{-1} - \beta_0^{(k+1)} x_n x_n', \\
\beta_0^{(k+1)} &= \sum_{n=1}^{N} v_n^{(k)} y_n - \beta_1^{(k+1)} \sum_{n=1}^{N} v_n^{(k)} x_n, \\
\sigma^2^{(k+1)} &= \frac{\left( \beta_0^{(k+1)} + \beta_1^{(k+1)} x_n \right)^2}{\sum_{n=1}^{N} v_n^{(k)}}. \quad (48)
\end{align*}
\]

For the updating of \( \zeta \), we need to numerically solve the equation

\[
\sum_{n=1}^{N} \frac{\partial}{\partial \nu} Q_{3n} (\zeta; \psi^{(k)}) = 0, \quad (51)
\]
which corresponds to finding $\zeta^{(k+1)}$ as the solution of

$$-\psi\left(\frac{\zeta}{2}\right) + \ln \frac{\zeta}{2} + 1 + \sum_{n=1}^{N} \left(\ln v_{n}^{(k)} - v_{n}^{(k)}\right) + \psi\left(\frac{\zeta^{(k)} + 1}{2}\right) - \ln \left(\frac{\zeta^{(k)} + 1}{2}\right) = 0.
$$

(52)

4.3.3. Normal component marginal densities

Furthermore, the normal case for the component distributions of $X$ can be obtained, as stated previously, as a limiting case when $\nu_{g} \to \infty$, $g = 1, \ldots, G$.

Then, in (16), $u_{ng}^{(k)} \to 1$. Substituting this value into (27) and (28), we obtain

$$
\mu_{g}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} x_{n}}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}
$$

(53)

and

$$
\Sigma_{g}^{(k+1)} = \frac{\sum_{n=1}^{N} \tau_{ng}^{(k)} (x_{n} - \mu_{g}^{(k+1)}) (x_{n} - \mu_{g}^{(k+1)})'}{\sum_{n=1}^{N} \tau_{ng}^{(k)}}.
$$

(54)

Naturally, in this case, we do not compute the additional $M$-step maximizing $Q_{5} (\nu; \psi_{\zeta}^{(k)})$ in (24). Accordingly, for the sub-case $\mu_{1} = \cdots = \mu_{G} = \mu$ and $\Sigma_{1} = \cdots = \Sigma_{G} = \Sigma$, in equation (36) we have $u_{n}^{(k)} \to 1$ and the updated estimates of $\mu$ and $\Sigma$ become

$$
\mu = \frac{1}{n} \sum_{n=1}^{N} x_{n}
$$

(55)

and

$$
\Sigma = \frac{1}{n} \sum_{n=1}^{N} (x_{n} - \mu) (x_{n} - \mu)',
$$

(56)

which, naturally, do not depend on the EM-iterations.

4.3.4. Normal component conditional densities

Finally, the normal case for the component distributions of $Y|X$ can be obtained as a limiting case when $\zeta_{g} \to \infty$, $g = 1, \ldots, G$. Then, in (15), $v_{ng}^{(k)} \to 1$. 

16
Substituting this value into (29) and (30), we obtain

\[
\beta_{1j}^{(k+1)} = \left( \sum_{n=1}^{N} \tau_{ng}^{(k)} x_n x_n' \right) - \left( \sum_{n=1}^{N} \tau_{ng}^{(k)} x_n \sum_{n=1}^{N} \tau_{ng}^{(k)} x_n' \right)^{-1} \left( \sum_{n=1}^{N} \tau_{ng}^{(k)} y_n x_n \sum_{n=1}^{N} \tau_{ng}^{(k)} x_n' \right)
\]

\[
\beta_{0j}^{(k+1)} = \sum_{n=1}^{N} \tau_{ng}^{(k)} y_n x_n - \beta_{1j}^{(k+1)}\left( \sum_{n=1}^{N} \tau_{ng}^{(k)} x_n \right)
\]

\[
\sigma_{\theta}^{2(k+1)} = \sum_{n=1}^{N} \tau_{ng} \left[ y_n - \left( \beta_{0j}^{(k+1)} + \beta_{1j}^{(k+1)} x_n \right) \right]^2 / \sum_{n=1}^{N} \tau_{ng}.
\]

In this case, we again do not compute the additional M-step maximizing \(Q_\beta(\zeta; \psi^{(k)})\) in (22). Accordingly, for the sub-case \(\beta_{11} = \cdots = \beta_{1k} = \beta_1, \beta_{01} = \cdots = \beta_{0k} = \beta_0,\) and \(\sigma_{\theta}^{2} = \cdots = \sigma_{\theta}^{2} = \sigma^2,\) in equation (14) we have \(v_n^{(k)} \to 1\) and the updated estimates of \(\beta_1, \beta_0,\) and \(\sigma^2\) become

\[
\beta_1 = \left( \frac{1}{n} \sum_{n=1}^{N} x_n x_n' - \frac{1}{n^2} \sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n' \right)^{-1} \left( \frac{1}{n} \sum_{n=1}^{N} y_n x_n - \frac{1}{n^2} \sum_{n=1}^{N} y_n \sum_{n=1}^{N} x_n \right)
\]

\[
\beta_0 = \frac{1}{n} \sum_{n=1}^{N} y_n - \frac{1}{n} \beta_1' \sum_{n=1}^{N} x_n
\]

\[
\sigma^2 = \frac{1}{n} \sum_{n=1}^{N} \left[ y_n - \left( \beta_0 + \beta_1' x_n \right) \right]^2,
\]

which do not depend on the EM-iterations.

5. Computational issues and partition evaluation

This section presents some issues concerning practical implementation of the EM algorithm described in Section 4.
5.1. Estimating the degrees of freedom

Code for all of the analyses presented herein was written in the R computing environment (R Development Core Team, 2011) and a numerical search for the estimates of the degrees of freedom was carried out using the uniroot command in the stats package. This command is based on the Fortran subroutine zeroin described by Brent (1973). In order to expedite convergence, the range of values for \( \nu_g, \zeta_g, \nu, \) and \( \zeta \) was restricted to \( (2, 200] \). Previous work in the context of model-based clustering (see Andrews & McNicholas, 2011) and some experiments whose results are not reported here suggest that these restrictions do not hamper classification performance and show that the upper limit of 200 does not thwart the recovery of an underlying normal structure.

5.2. EM initialization

It is well known that the choice of starting values represents an important issue in the EM algorithm. The standard initialization consists of selecting a value for \( \psi^{(0)} \). An alternative approach, more natural in the authors’ opinion, is to specify a value for \( z_n^{(0)}, n = 1, \ldots, N \) (see McLachlan & Peel, 2000, p. 54). Within this approach, and due to the structure of our family of linear CWMs, we propose a random-hierarchical initialization procedure that allows for a guaranteed natural ranking among the likelihoods.

For a fixed \( G \), we start by considering NN-VE and NN-EV, because the former is nested in all of the VE-models, the latter is nested in all of the EV models, and both are nested in all of the VV-models. For NN-VE and NN-EV only, a random initialization is repeated 10 times, from different random positions, and the solution maximizing the likelihood among these 10 runs is selected. Note that, as underlined by Andrews et al. (2011), mixtures based on the multivariate \( t \) distribution are more sensitive to bad starting values than their Gaussian counterparts. Thus, by considering random initialization only for the above models of type NN, we prevent the possible failure of the algorithm due to poor starting values for models of type \( Nt, tN, \) and \( tt \). In each run, the
$N$ vectors $z_n^{(0)}$ are randomly drawn from a multinomial distribution with probabilities $(1/G, \ldots, 1/G)$. Once the EM-estimates of the posterior probabilities $\tau_{ng}$ have been obtained for these models, say $\hat{\tau}_{NG-VE}^{NN-VE}$ and $\hat{\tau}_{NG-EV}^{NN-EV}$, the hierarchical initialization procedure proceeds according to the scheme in Figure 1, where each arrow is directed from the model used for initialization to the model to be estimated.

Figure 1: Relationships among the models in the hierarchical initialization strategy. Arrows are oriented from the model used for initialization to the model to be estimated.

Thus, $\hat{\tau}_{NG}^{NN-VE}$ is used to initialize the EM of both $tN-VE$ and $Nt-VE$, obtaining $\hat{\tau}_{ng}^{tN-VE}$ and $\hat{\tau}_{ng}^{Nt-VE}$, respectively, while $\hat{\tau}_{ng}^{NN-EV}$ is used to initialize the EM of both $tN-EV$ and $Nt-EV$, leading to $\hat{\tau}_{ng}^{tN-EV}$ and $\hat{\tau}_{ng}^{Nt-EV}$, respectively. Also, following the same principle, the model between $NN-VE$ and $NN-EV$ leading to the maximum likelihood is used to initialize the EM for $NN-VV$. 

19
Without going into further details on this hierarchical procedure, in the last step the model between $Nt$-VV, $tN$-VV, $tt$-VE, and $tt$-EV leading to the maximum likelihood is used to initialize the EM of $tt$-VV.

5.3. Convergence criterion

The Aitken acceleration procedure (Aitken, 1926) is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the EM algorithm. Based on this estimate, a decision can be made regarding whether or not the algorithm has reached convergence; that is, whether or not the log-likelihood is sufficiently close to its estimated asymptotic value. The Aitken acceleration at iteration $k$ is given by

\[ a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}} \]

where $l^{(k+1)}$, $l^{(k)}$, and $l^{(k-1)}$ are the log-likelihood values from iterations $k + 1$, $k$, and $k - 1$, respectively. Then, the asymptotic estimate of the log-likelihood at iteration $k + 1$ (Böhm et al., 1994) is given by

\[ l^{(k+1)} = l^{(k)} + \frac{1}{1 - a^{(k)}} \left( l^{(k+1)} - l^{(k)} \right) \]

In the analyses in Section 7, we follow McNicholas (2010) and stop our algorithms when $l^{(k+1)} - l^{(k)} < \epsilon$, with $\epsilon = 0.05$.

6. Model selection and performance

Once we have fitted data according to all models described in Table 1, we need to select a good model. For this purpose, model selection criteria are usually taken into account. In the following, we focus on the Bayesian information criterion (BIC) and the integrated completed likelihood (ICL). Note that the term “model selection” has a twofold meaning in model-based clustering. First, it means the choice of the parametric structure, i.e., the selection of the best member of the family of models in Table 1; second, it means the selection of the number $G$ of mixture components.
6.1. The Bayesian information criterion (BIC)

A classical approach is to select the model that maximizes the integrated likelihood. The integrated likelihood is usually approximated by the BIC (Schwarz, 1978), which is the most commonly used selection technique in model-based clustering (Fraley & Raftery, 2002; McNicholas & Murphy, 2008, 2010). The use of the BIC in mixture model selection was proposed by Dasgupta & Raftery (1998) based on an approximation of Bayes factors (Kass & Raftery, 1995). The BIC is given by

\[ \text{BIC} = 2l(\hat{\psi}) - m \ln N, \]

where \( \hat{\psi} \) is the ML estimate of \( \psi \), \( l(\hat{\psi}) \) is the maximized observed-data log-likelihood, and \( m \) is the overall number of free parameters in the model (see the last three columns in Table 1). In the mixture framework, the regularity conditions in Leroux (1992), Kass & Raftery (1995), Kass & Wasserman (1995), and Keribin (2000) are not generally satisfied (see Biernacki et al. 2000, McNicholas & Murphy 2008, and McNicholas 2011) and there is a lack of theoretical justification for the BIC approximation. Simulations experiments (see Roeder & Wasserman 1997, and Fraley & Raftery 1998, 2002) provide practical evidence that the BIC performs well as a model selection criterion for mixture models; however, in some cases the model selected by the BIC does not necessarily give the best predicted classifications (Andrews & McNicholas, 2011). Moreover, the integrated likelihood could show a tendency towards assigning multiple mixture components to what is really just one cluster (Biernacki et al., 2000).

6.2. The integrated completed likelihood (ICL)

In an attempt to focus model selection on clusters rather than mixture components, Biernacki et al. (2000) proposed the ICL as an alternative to the BIC. The ICL essentially penalizes the BIC for estimated mean entropy. In practice, an approximate ICL is used and this is given by

\[ \text{ICL} \approx \text{BIC} + \sum_{n=1}^{N} \sum_{g=1}^{G} \text{MAP}(\hat{\tau}_{ng}) \ln \hat{\tau}_{ng}, \]

(60)
where $\hat{\tau}_{ng}$ is the estimated \textit{a posteriori} expected value of $Z_{ng}$ and

$$\text{MAP}(\hat{\tau}_{ng}) = \begin{cases} 1 & \text{if } \max_g \{\hat{\tau}_{ng}\} \text{ occurs in component } g \\ 0 & \text{otherwise,} \end{cases}$$

is the maximum \textit{a posteriori} classification given $\hat{\tau}_{ng}$. The estimated mean entropy reflects the uncertainty in the classification of observation $n$ into component $g$, thereby “punishing mixture components that are more spread out” (McNicholas & Subedi, 2012, p. 1117). Therefore, the ICL should favor well-separated clusters compared to the BIC.

Simulation studies by Biernacki et al. (2000) show that, from the practical point of view, the ICL seems to give an answer to the possible tendency of the BIC to overestimate the number of clusters when the model at hand does not fit the data well. However, similar to the BIC, the model selected by the ICL does not necessarily give the most accurate estimated classifications (McNicholas, 2011). In our analyses (Section 7), both the BIC and the ICL are used for model selection.

### 6.3. The Rand index and the adjusted Rand index

In order to evaluate the model performance in cases in which the true classification is known, the adjusted Rand index (ARI; Hubert & Arabie, 1985) is often taken into account as a measure of class agreement. The original Rand index (RI; Rand, 1971) is based on pairwise comparisons and is obtained by dividing the number of pair agreements (observations that should be in the same group and are, plus those that should not be in the same group and are not) by the total number of pairs. The RI assumes values between 0 and 1, where 0 indicates no pairwise agreement between the MAP classification and true group membership and 1 indicates perfect agreement. One criticism of the RI is that its expected value is greater than 0, making smaller values difficult to interpret. The ARI corrects the RI for chance by allowing for the possibility that classification performed randomly will correctly classify some observations. Thus, the
ARI has an expected value of 0 and perfect classification would result in a value equal to 1.

7. Applications to real data

This section illustrates some real data applications of the family of linear CWMs defined in Section 3.

7.1. Student Data

The first application concerns data coming from a survey of \( N = 270 \) students attending a statistics course at the Department of Economics and Business of the University of Catania in the academic year 2011/2012. The questionnaire included seven items, but the analysis we present below only concerns the following subset of variables:

- \( \text{GENDER} \) = gender of the respondent;
- \( \text{HEIGHT} \) = height of the respondent, measured in centimeters;
- \( \text{WEIGHT} \) = weight of the respondent, measured in kilograms;
- \( \text{HEIGHT.F} \) = height of respondent’s father, measured in centimeters.

There are \( G = 2 \) groups of respondents with respect to the GENDER variable: \( N_M = 119 \) males and \( N_F = 151 \) females. In the following, the two groups will be simply referred to as \( G_M \) and \( G_F \), respectively. Moreover, we shall focus first on the joint distributions of \( \text{WEIGHT} \) and \( \text{HEIGHT} \), then on \( \text{HEIGHT} \) and \( \text{HEIGHT.F} \). In both scenarios, data will be assumed unlabeled with respect to GENDER. However, the true labels will be useful for evaluating the quality of the obtained clustering.

7.1.1. First scenario: \( \text{HEIGHT} \) and \( \text{WEIGHT} \)

Figure 2 concerns the observed labeled data. This graphical representation will be simply referred to as the CW-plot. The top of Figure 2 includes a barplot of the \( \text{HEIGHT} \) variable, including the overall empirical marginal density as well
as the empirical marginal densities for $G_M$ and $G_F$; bars are color-coded, using a gray scale, according to the GENDER variable. We remark that many students tend to approximate their height to “classical” values, such as 155, 160, 170, 175, and so on. For classification purposes, the variable HEIGHT separates the two groups quite well. The bottom of Figure 2 is a scatter plot of HEIGHT and WEIGHT, where male and female students are labeled with $M$ and $F$, respectively. We give the isodensities of a bivariate normal kernel estimator as computed by the function `bkde2D` of the R-package `KernSmooth` (see, e.g., Wand & Jones, 1995). The plot also shows the functional dependence of WEIGHT on HEIGHT separately for $G_M$ and $G_F$; the solid lines concern the linear regression models while the dashed ones arise from a locally-weighted polynomial regression computed using the `lowess` function of the R-package `stats` (see Cleveland, 1979, for details). A simple visual comparison
between solid and dashed lines justifies the linearity assumption of WEIGHT on HEIGHT, underlying the linear CWMs of the proposed family. Moreover, the regression lines in Figure 2 seem to indicate that these models have the same parameters in \( G_M \) and \( G_F \). This conjecture is also statistically confirmed:

1. the \( t \)-test for equal slopes provides a \( p \)-value of 0.147,
2. the \( t \)-test for equal intercepts provides a \( p \)-value of 0.364, and
3. the F-test of homoscedasticity of residuals in the two groups provides a \( p \)-value of 0.992.

Now, let us ignore the true classification induced by GENDER and fit the data according to the linear CWMs in Table 1 by using the true value \( G = 2 \). Table 2 lists the values of the BIC, ICL, and ARI for the twelve models.

Table 2: Values of the BIC, ICL, and ARI (\( G = 2 \)). Bold numbers highlight the best model for each criterion/index.

|       | (a) BIC          | (b) ICL          | (c) ARI          |
|-------|------------------|------------------|------------------|
|       | VE   | EV   | VV   | VE   | EV   | VV   | VE   | EV   | VV   |
| \( N \) | \-3726.197  | \-3756.561 | \-3742.947  | \-3750.466  | \-3880.260 | \-3767.213  | \-3767.261  | \-3875.443 | \-3874.681 |
| \( t \)N | \-3737.394  | \-3762.160 | \-3754.144  | \-3761.063  | \-3885.858 | \-3778.409  | \-3767.261  | \-3875.443 | \-3874.681 |
| \( Nt \) | \-3731.795  | \-3766.517 | \-3749.642  | \-3756.064  | \-3869.845 | \-3773.484  | \-3767.261  | \-3875.443 | \-3874.681 |
| \( tt \) | \-3742.992  | \-3772.115 | \-3768.839  | \-3767.261  | \-3875.443 | \-3874.681  | \-3767.261  | \-3875.443 | \-3874.681 |

\( NN-VE \) (Gaussian marginal and conditional component densities and equal linear model between clusters) is the best model according to both the BIC (\-3726.197) and the ICL (\-3750.466). The corresponding CW-plot is displayed in Figure 3. As for the ARI, in practice we have similar results for all models of type VE and VV. This is an example in which the group structure of the data is due to different intra-group marginal distributions for the independent variable, while the linear relationship is homogenous. This is the typical situation in which a finite mixture of linear regressions is not able to recognize the real group-structure in the data (see Ingrassia et al., 2012, for details).
7.1.2. Second scenario: \( \text{HEIGHT.F and HEIGHT} \)

Figure 4 shows the CW-plot of \( \text{HEIGHT.F and HEIGHT} \) by considering the classification induced by \( \text{GENDER} \). Also in this case, the hypothesis of a linear relationship between the variables (solid lines) appear to be reasonable. However, the models for the two groups differ. Indeed, although the \( F \)-test of homoscedasticity of the residuals in the two groups gives a \( p \)-value of 0.086, the \( t \)-tests for equal slopes and equal intercepts provide practically null \( p \)-values. Finally, we carried out data modeling according to linear CWM (ignoring the true clustering induced by \( \text{GENDER} \)). The values for the BIC, ICL, and ARI for the twelve models are given in Table 3. In this case, the best model is \( \text{NN-EV} \) (see also the corresponding CW-plot in Figure 5). The fitted model also appears to be a good compromise in terms of the ARI values of Table 3. Differently from the first scenario, this is an example in which the group-structure is determined
Figure 4: Student Data: CW-plot of HEIGHT and HEIGHT.F for 119 male and 151 female students (M denotes male and F female).

Table 3: Values of the BIC, ICL, and ARI ($G = 2$). Bold numbers highlight the best model for each criterion/index.

|          | VE   | EV   | VV   |
|----------|------|------|------|
| NN       | -3726.339 | -3594.401 | -3601.955 |
| tN       | -3737.536 | -3599.999 | -3613.152 |
| Nt       | -3731.937 | -3605.598 | -3613.152 |
| tt       | -3743.134 | -3611.196 | -3624.348 |

|          | VE   | EV   | VV   |
|----------|------|------|------|
| NN       | -3822.623 | -3597.252 | -3605.016 |
| tN       | -3833.820 | -3602.850 | -3616.212 |
| Nt       | -3828.221 | -3608.449 | -3616.212 |
| tt       | -3839.418 | -3614.047 | -3627.409 |

|          | VE   | EV   | VV   |
|----------|------|------|------|
| NN       | 0.009 | 0.898 | 0.912 |
| tN       | 0.009 | 0.898 | 0.912 |
| Nt       | 0.009 | 0.898 | 0.912 |
| tt       | 0.009 | 0.998 | 0.912 |

by the different intra-group linear models, while the marginal distribution of the independent variable is homogenous.

7.2. Museum & monument attendance and tourism flow

The second application focuses on $N = 180$ monthly data (tourism data) concerning tourist overnights ($X$, data in millions) and attendance at museums.
and monuments ($Y$, data in millions) in Italy over the 15-year period spanning from January 1996 to December 2010. These data have been recently analyzed by Cellini & Cuccia (forthcoming). The CW-plot of the labeled data (with respect to months) is shown in Figure 5. It is straightforward to note how the heterogeneity of the data reveals a clear group-structure. Figure 5 shows the values of the BIC and the ICL for the models in the proposed family of linear CWMs with $G$ ranging from 1 to 6. Both criteria (BIC=$-1683.727$ and ICL=$-1689.386$) suggest the $NN$-$VV$, with $G = 4$ components, displayed in Figure 8.

Here, it is interesting to analyze the relationship between the obtained clusters – characterized by 4 different slopes – and the time-covariate (months). (see Table 4). The four clusters, arising from the $NN$-$VV$, are almost perfectly related to the months (except for two units in November, which concern years 2006 and 2010). In particular, we have:
Figure 6: Tourist data: CW-plot of tourist overnights ($X$, in millions) and attendance at museums and monuments ($Y$, in millions) in Italy over the period from January 1996 to December 2010 ($N = 180$). The univariate normal kernel density of $X$ is superimposed on the histogram. The isodensities from a bivariate normal kernel density estimator are also visualized on the scatter plot. Month abbreviations are used as labels in the scatter plot.

Table 4: Relation between the $G = 4$ clusters, obtained with the fitted NN-VV, and the variable time identified by month.

| group | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1     | 15  | 15  | 0   | 0   | 0   | 0   | 0   | 0   | 15  | 13  | 15  |
| 2     | 0   | 0   | 0   | 0   | 15  | 0   | 0   | 15  | 0   | 0   | 0   | 0   |
| 3     | 0   | 0   | 15  | 15  | 15  | 0   | 0   | 0   | 15  | 2   | 0   |
| 4     | 0   | 0   | 0   | 0   | 0   | 15  | 15  | 0   | 0   | 0   | 0   | 0   |

Table 4: Relation between the $G = 4$ clusters, obtained with the fitted NN-VV, and the variable time identified by month.

**Group 1**: units from November to February,

**Group 2**: units in June and September,
Figure 7: Tourist data: Values of the BIC and ICL ($G = 1, \ldots, 6$).

Figure 8: Tourist data: CW-plot of model $NN-VV$ with $G = 4$ components ($X =$ “tourist overnights”, in millions, and $Y =$ “attendance at museums and monuments”, in millions).
Group 3: units in March, April, May, and October, and

Group 4: units in July and August.

This is an example in which the group structure of the data is due to differences both in the intra-group marginal distributions and the linear models.

8. Concluding remarks

In this paper, a novel family of twelve linear cluster-weighted models was presented. Such a family represents a flexible and powerful tool for model-based clustering. Maximum likelihood parameter estimation was performed according to the EM algorithm and model selection was accomplished using both the BIC and ICL. Many computational aspects were illustrated and a simple, but very effective, hierarchical random initialization method was introduced. Model-based clustering, using the proposed family, was appreciated on the grounds of some applications to real data. Here, it is interesting to note how the data set related to the survey of students in Section 7.1 justifies and motivates the search for a model in the proposed family. Future work will involve the extension of such a family to the model-based classification context.

References

Aitken, A. (1926). On Bernoulli’s numerical solution of algebraic equations. In *Proceedings of the Royal Society of Edinburgh* (pp. 289–305). Volume 46.

Andrews, J., & McNicholas, P. (2011). Extending mixtures of multivariate $t$-factor analyzers. *Statistics and Computing*, 21, 361–373.

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

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.
Böhmig, D., Dietz, E., Schaub, R., Schlattmann, P., & Lindsay, B. (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.

Brent, R. (1973). *Algorithms for minimization without derivatives*. New Jersey: Prentice Hall.

Cellini, R., & Cuccia, T. (forthcoming). Museum and monument attendance and tourism flow: a time series approach. *Applied Economics, .., ..*

Chatzis, S., & Varvarigou, T. (2008). Robust fuzzy clustering using mixtures of Student’s-t distributions. *Pattern Recognition Letters, 29*, 1901–1905.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836.

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

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

Everitt, B., & Hand, D. J. (1981). *Finite mixture distributions*. Chapman & Hall.

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

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*, 611–631.
Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.

Gershenfeld, N. (1997). Non linear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, 808, 18–24.

Greselin, F., & Ingrassia, S. (2010). Constrained monotone EM algorithms for mixtures of multivariate $t$ distributions. *Statistics and Computing*, 20, 9–22.

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

Ingrassia, S., Minotti, S. C., & Vittadini, G. (2012). Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, 29, ...–...

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Kass, R., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928–934.

Kent, J., Tyler, D., & Vardi, Y. (1994). A curious likelihood identity for the multivariate $t$-distribution. *Communications in Statistics. Simulation and Computation*, 23, 441–453.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya. The Indian Journal of Statistics. Series A*, 62, 49–66.

Lange, K. L., Little, R. J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the $t$ distribution. *Journal of the American Statistical Association*, 84, 881–896.

Leisch, F. (2004). *FlexMix*: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11, 1–18.
Leroux, B. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics, 20*, 1350–1360.

Liu, C. (1997). MI estimation of the multivariate t distribution and the EM algorithm. *Journal of Multivariate Analysis, 63*, 296–312.

Liu, C., Rubin, D., & Wu, Y. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika, 85*, 755–770.

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

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

McNicholas, P. (2010). Model-based classification using latent gaussian mixture models. *Journal of Statistical Planning and Inference, 140*, 1175–1181.

McNicholas, P. (2011). On model-based clustering, classification, and discriminant analysis. *Journal of the Iranian Statistical Society, 10*, 181–199.

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

McNicholas, P., & Murphy, T. (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics, 38*, 153–168.

McNicholas, P., & Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference, 142*, 1114–1127.

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

Peel, D., & McLachlan, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing, 10*, 339–348.
Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846–850.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association, 92*, 894–902.

Schöner, B. (2000). *Probabilistic Characterization and Synthesis of Complex Data Driven Systems*. Ph.D. Thesis, MIT.

Schöner, B., & Gershenfeld, N. (2001). Cluster weighted modeling: Probabilistic time series prediction, characterization, and synthesis. In A. Mees (Ed.), *Nonlinear Dynamics and Statistics*. Boston: Birkhauser.

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

Shoham, S. (2002). Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. *Pattern Recognition, 35*, 1127–1142.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons.

Wand, M., & Jones, M. (1995). *Kernel smoothing* volume 60 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate student- $t$ error terms. *Journal of the American Statistical Association, 71*, 400–405.