

# Model-based clustering via linear cluster-weighted models

S. Ingrassia<sup>a,\*</sup>, S.C. Minotti<sup>b</sup>, A. Punzo<sup>a</sup>

<sup>a</sup>*Dipartimento di Economia e Impresa, Università di Catania,  
Corso Italia 55, 95129 Catania (Italy)*

<sup>b</sup>*Dipartimento di Statistica, Università di Milano-Bicocca (Italy)*

---

## Abstract

A novel family of twelve mixture models with random covariates, 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 the 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. Finally, a simulation study for evaluating the performance of the BIC and the ICL is presented.

### Keywords:

Cluster-weighted model, Mixture models with random covariates, Model-based clustering, Multivariate  $t$  distribution.

2000 MSC: 62H30, 62H99

---

## 1. Introduction

In direct applications of finite mixture models (see Titterton 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 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), Titterton et al. (1985), McLachlan & Peel (2000), and Frühwirth-Schnatter (2006).

This paper focuses on data arising from a real-valued random vector  $(Y, \mathbf{X}')' : \Omega \rightarrow \mathbb{R}^{d+1}$ , having joint density  $p(y, \mathbf{x})$ , where  $Y$  is the response variable and  $\mathbf{X}$  is the vector of covariates. Standard model-based clustering techniques assume that  $\Omega$  can be partitioned into  $G$  groups  $\Omega_1, \dots, \Omega_G$ . 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  $\mathbf{x}$  can be modeled by

$$Y = \mu(\mathbf{x}; \boldsymbol{\beta}_g) + \varepsilon_g = \beta_{0g} + \boldsymbol{\beta}'_{1g} \mathbf{x} + \varepsilon_g,$$

where  $\boldsymbol{\beta}_g = (\beta_{0g}, \boldsymbol{\beta}'_{1g})'$ ,  $\mu(\mathbf{x}; \boldsymbol{\beta}_g) = E(Y|\mathbf{X} = \mathbf{x}, \Omega_g)$  is the linear regression function and  $\varepsilon_g$  is the error variable, independent with respect to  $\mathbf{X}$ , with zero mean and finite constant variance  $\sigma_g^2$ ,  $g = 1, \dots, G$ . However, as highlighted in Hennig (2000), finite mixtures of linear regressions are inadequate for most of the applications because they assume *assignment independence*: the probability for a point  $(y, \mathbf{x}')'$  to be generated by one of the mixture components has

---

\*Corresponding author

Email addresses: s.ingrassia@unict.it (S. Ingrassia), simona.minotti@unimib.it (S.C. Minotti), antonio.punzo@unict.it (A. Punzo)

to be the same for all covariates values  $\mathbf{x}$ . In other words, the assignment of the data points to the clusters has to be independent of the covariates.

Here, differently from finite mixtures of linear regressions, we assume random covariates having a parametric specification. This allows for *assignment dependence*: the covariate distributions of the mixture components can also be distinct. In the framework of mixture models with random covariates, the cluster weighted model (CWM; Gershenfeld, 1997), with equation

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

also called saturated mixture regression model by Wedel (2002), constitutes a reference approach to model the joint density. In (1), normality of both  $p(y | \mathbf{x}, \Omega_g)$  and  $p(\mathbf{x} | \Omega_g)$  is commonly assumed (see, e.g., Gershenfeld, 1997 and Punzo, 2014). Alternatively, Ingrassia et al. (2012) propose also the use of the  $t$  distribution which provides, as other approaches (Punzo & McNicholas, 2013, 2014a,b), 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

$$p(y | \mathbf{x}, \Omega_g) = h_t(y | \mathbf{x}; \boldsymbol{\xi}_g, \zeta_g) = \frac{\Gamma\left(\frac{\zeta_g + 1}{2}\right)}{(\pi \zeta_g \sigma_g^2)^{\frac{1}{2}} \left\{1 + \delta[y, \mu(\mathbf{x}; \boldsymbol{\beta}_g); \sigma_g^2]\right\}^{\frac{\zeta_g + 1}{2}}} \quad (2)$$

and

$$p(\mathbf{x} | \Omega_g) = h_{td}(\mathbf{x}; \boldsymbol{\vartheta}_g, \nu_g) = \frac{\Gamma\left(\frac{\nu_g + d}{2}\right) |\boldsymbol{\Sigma}_g|^{-\frac{1}{2}}}{(\pi \nu_g)^{\frac{d}{2}} \left[1 + \delta(\mathbf{x}, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_g)\right]^{\frac{\nu_g + d}{2}}}, \quad (3)$$

with  $\boldsymbol{\xi}_g = \{\boldsymbol{\beta}_g, \sigma_g^2\}$ ,  $\boldsymbol{\vartheta}_g = \{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}$ ,  $\delta[y, \mu(\mathbf{x}; \boldsymbol{\beta}_g); \sigma_g^2] = [y - \mu(\mathbf{x}; \boldsymbol{\beta}_g)]^2 / \sigma_g^2$ , and  $\delta(\mathbf{x}, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_g) = (\mathbf{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g)$ . Thus, (2) is the density of a (generalized) univariate  $t$  distribution, with location parameter  $\mu(\mathbf{x}; \boldsymbol{\beta}_g)$ , scale parameter  $\sigma_g^2$ , and  $\zeta_g$  degrees of freedom, while (3) is the density of a multivariate  $t$  distribution with location parameter  $\boldsymbol{\mu}_g$ , inner product matrix  $\boldsymbol{\Sigma}_g$ , and  $\nu_g$  degrees of freedom. By substituting (2) and (3) into (1), we obtain the linear  $t$  CWM

$$p(y, \mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^G \pi_g h_t(y | \mathbf{x}; \boldsymbol{\xi}_g, \zeta_g) h_{td}(\mathbf{x}; \boldsymbol{\vartheta}_g, \nu_g), \quad (4)$$

where the set of all unknown parameters is denoted by  $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_G\}$ , with  $\boldsymbol{\psi}_g = \{\pi_g, \boldsymbol{\xi}_g, \zeta_g, \boldsymbol{\vartheta}_g, \nu_g\}$ . Recent developments in CWMs can be found in Punzo (2014), Punzo & McNicholas (2014a), Punzo & Ingrassia (2015a,b), Subedi et al. (2013, 2015), and Ingrassia et al. (2015).

In this paper, we introduce a family of twelve linear CWMs obtained from (4) by imposing convenient component distributional constraints. If  $\zeta_g, \nu_g \rightarrow \infty$ , the 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 us to infer if the group-structure of the data is due to the contribution of  $\mathbf{X}$ ,  $Y | \mathbf{X}$ , 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. In Section 7 some applications to real data are illustrated. In Section 8 simulations for a comparison between BIC and ICL are described. Finally, in Section 9, 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  $(y_1, \mathbf{x}'_1)', \dots, (y_N, \mathbf{x}'_N)'$  be a sample of size  $N$  from (4). Once  $\underline{\psi}$  is estimated (fixed), the posterior probability that the generic unit  $(y_n, \mathbf{x}'_n)'$ ,  $n = 1, \dots, N$ , comes from component  $\Omega_g$  is given by

$$\tau_{ng} = P(\Omega_g | y_n, \mathbf{x}_n; \underline{\psi}) = \frac{\pi_g h_t(y_n | \mathbf{x}_n; \xi_g, \zeta_g) h_{t_d}(\mathbf{x}_n; \vartheta_g, \nu_g)}{p(y_n, \mathbf{x}_n; \underline{\psi})}, \quad g = 1, \dots, G. \quad (5)$$

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

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

and

$$p(\mathbf{x}; \underline{\pi}, \underline{\vartheta}, \underline{\nu}) = \sum_{g=1}^G \pi_g h_{t_d}(\mathbf{x}; \vartheta_g, \nu_g), \quad (7)$$

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

**Proposition 1.** *Given  $\underline{\pi}$ ,  $\underline{\vartheta}$ , and  $\underline{\nu}$ , if  $h_t(y | \mathbf{x}; \xi_1, \zeta_1) = \dots = h_t(y | \mathbf{x}; \xi_G, \zeta_G) = h_t(y | \mathbf{x}; \xi, \zeta)$ , 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 for the linear  $t$  CWM in (4) can be written as

$$\tau_{ng} = \frac{\pi_g h_t(y_n | \mathbf{x}_n; \xi, \zeta) h_{t_d}(\mathbf{x}_n; \vartheta_g, \nu_g)}{\sum_{j=1}^G \pi_j h_t(y_n | \mathbf{x}_n; \xi, \zeta) h_{t_d}(\mathbf{x}_n; \vartheta_j, \nu_j)} = \frac{\pi_g h_{t_d}(\mathbf{x}_n; \vartheta_g, \nu_g)}{\sum_{j=1}^G \pi_j h_{t_d}(\mathbf{x}_n; \vartheta_j, \nu_j)},$$

which correspond to the posterior probabilities for model (7).  $\square$

**Proposition 2.** *Given  $\underline{\pi}$ ,  $\underline{\xi}$ , and  $\underline{\zeta}$ , if  $h_{t_d}(\mathbf{x}; \vartheta_1, \nu_1) = \dots = h_{t_d}(\mathbf{x}; \vartheta_G, \nu_G) = h_{t_d}(\mathbf{x}; \vartheta, \nu)$ , 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 for the linear  $t$  CWM in (4) can be written as

$$\tau_{ng} = \frac{\pi_g h_t(y_n | \mathbf{x}_n; \xi_g, \zeta_g) h_{t_d}(\mathbf{x}_n; \vartheta, \nu)}{\sum_{j=1}^G \pi_j h_t(y_n | \mathbf{x}_n; \xi_j, \zeta_j) h_{t_d}(\mathbf{x}_n; \vartheta, \nu)} = \frac{\pi_g h_t(y_n | \mathbf{x}_n; \xi_g, \zeta_g)}{\sum_{j=1}^G \pi_j h_t(y_n | \mathbf{x}_n; \xi_j, \zeta_j)},$$

which correspond to the posterior probabilities for model (6).  $\square$

Note that the results in Proposition 1 and 2 are not restricted to the  $t$  distribution; in fact, they can be easily extended to the general CWM in (1). Further, some results about the relation between linear Gaussian (or  $t$ ) CWMs and finite mixture of regressions are given in Ingrassia et al. (2012). Finally, it is important to underline that up to now there are no theoretical results on the identifiability for linear CWMs; however, since they can be seen as mixture models with random covariates, the results in Hennig (2000, Section 3, Model 2.a) can apply.

### 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  $\nu_g, \zeta_g \rightarrow \infty$  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  $\sigma_g^2$ ) 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  $Y|X, \Omega_g$  ( $N \equiv$ Normal and  $t \equiv t$ ), respectively, while the second two denote the distribution constraint between clusters ( $E \equiv$ Equal and  $V \equiv$ Variable) for  $X|\Omega_g$  and  $Y|X, \Omega_g$ , respectively.

Only two of the models given in Table 1,  $NN$ -VV and  $tt$ -VV, have been developed previously; the former corresponds to the linear Gaussian CWM of Gershensfeld (1997), while the latter coincides with the linear  $t$  CWM in 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  $G = 1$ , it results  $VV \equiv VE \equiv EV$  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 the most general model  $tt$ -VV. Details for all the other models are given in Appendix A.

In the EM framework, the generic observation  $(y_n, \mathbf{x}_n)'$  is viewed as being incomplete; its complete counterpart is given by  $(y_n, \mathbf{x}_n', z_n', u_n, v_n)'$ , where  $z_n$  is the component-label vector in which  $z_{ng} = 1$  if  $(y_n, \mathbf{x}_n)'$  comes from the  $g$ th component ( $z_{ng} = 0$  otherwise), while  $u_n$  and  $v_n$  arise from the standard theory of the (multivariate)  $t$  distribution according to which

$$Y_n | \mathbf{x}_n, v_n, z_{ng} = 1 \stackrel{\text{i.i.d.}}{\sim} N\left(\mu(\mathbf{x}_n; \beta_g), \frac{\sigma_g^2}{v_n}\right) \quad (8)$$

$$V_n | z_{ng} = 1 \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}\left(\frac{\zeta_g}{2}, \frac{\zeta_g}{2}\right), \quad (9)$$

for  $n = 1, \dots, N$ , and

$$\mathbf{X}_n | u_n, z_{ng} = 1 \stackrel{\text{i.i.d.}}{\sim} N\left(\mu_g, \frac{\Sigma_g}{u_n}\right) \quad (10)$$

$$U_n | z_{ng} = 1 \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}\left(\frac{\nu_g}{2}, \frac{\nu_g}{2}\right), \quad (11)$$

for  $n = 1, \dots, N$ . 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

$$l_c(\boldsymbol{\psi}) = l_{1c}(\boldsymbol{\pi}) + l_{2c}(\boldsymbol{\xi}) + l_{3c}(\boldsymbol{\zeta}) + l_{4c}(\boldsymbol{\vartheta}) + l_{5c}(\boldsymbol{\gamma}), \quad (12)$$

Model Identifier	$X \Omega_g$		$Y \mathbf{x}, \Omega_g$		Number of free parameters				
	Density	Constraint	Density	Constraint	$X$	$Y \mathbf{x}$	weights		
$tt$ -VV	$t$	Variable	$t$	Variable	$G\left(d + \frac{d(d+1)}{2} + 1\right)$	+	$G(d+3)$	+	$G-1$
$tt$ -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	$d + \frac{d(d+1)}{2} + 1$	+	$G(d+3)$	+	$G-1$
$NN$ -VV	Normal	Variable	Normal	Variable	$G\left(d + \frac{d(d+1)}{2}\right)$	+	$G(d+2)$	+	$G-1$
$NN$ -VE	Normal	Variable	Normal	Equal	$G\left(d + \frac{d(d+1)}{2}\right)$	+	$d+2$	+	$G-1$
$NN$ -EV	Normal	Equal	Normal	Variable	$d + \frac{d(d+1)}{2}$	+	$G(d+2)$	+	$G-1$
$tN$ -VV	$t$	Variable	Normal	Variable	$G\left(d + \frac{d(d+1)}{2} + 1\right)$	+	$G(d+2)$	+	$G-1$
$tN$ -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	$d + \frac{d(d+1)}{2} + 1$	+	$G(d+2)$	+	$G-1$
$Nt$ -VV	Normal	Variable	$t$	Variable	$G\left(d + \frac{d(d+1)}{2}\right)$	+	$G(d+3)$	+	$G-1$
$Nt$ -VE	Normal	Variable	$t$	Equal	$G\left(d + \frac{d(d+1)}{2}\right)$	+	$d+3$	+	$G-1$
$Nt$ -EV	Normal	Equal	$t$	Variable	$d + \frac{d(d+1)}{2}$	+	$G(d+3)$	+	$G-1$

Table 1: Overview of linear CWMs. In “model identifier”, the first and second letters represent, respectively, the density of  $X|\Omega_g$  and  $Y|\mathbf{x}, \Omega_g$  (here  $N \equiv \text{Normal}$ ), while the third and fourth letters indicate, respectively, if  $h_{t_d}(\mathbf{x}; \boldsymbol{\theta}_g, \nu_g)$  and  $h_t(y|\mathbf{x}; \boldsymbol{\xi}_g, \zeta_g)$  are assumed to be Equal=E or Variable=V between groups.

where

$$\begin{aligned}
l_{1c}(\underline{\pi}) &= \sum_{n=1}^N \sum_{g=1}^G z_{ng} \ln \pi_g, \\
l_{2c}(\underline{\xi}) &= \frac{1}{2} \sum_{n=1}^N \sum_{g=1}^G z_{ng} \left\{ -\ln(2\pi) + \ln v_n - \ln \sigma_g^2 - v_n \delta \left[ y_n, \mu(\mathbf{x}_n; \boldsymbol{\beta}_g); \sigma_g^2 \right] \right\}, \\
l_{3c}(\underline{\zeta}) &= \sum_{n=1}^N \sum_{g=1}^G z_{ng} \left[ -\ln \Gamma\left(\frac{\zeta_g}{2}\right) + \frac{\zeta_g}{2} \ln \frac{\zeta_g}{2} + \frac{\zeta_g}{2} (\ln v_n - v_n) - \ln v_n \right], \\
l_{4c}(\underline{\boldsymbol{\theta}}) &= \frac{1}{2} \sum_{n=1}^N \sum_{g=1}^G z_{ng} \left[ -d \ln(2\pi) + d \ln u_n - \ln |\boldsymbol{\Sigma}_g| - u_n \delta(\mathbf{x}_n, \boldsymbol{\mu}_g; \boldsymbol{\Sigma}_g) \right]
\end{aligned}$$

and

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

#### 4.1. E-step

The E-step, on the  $(k+1)$ th iteration, requires the calculation of

$$Q(\underline{\boldsymbol{\psi}}; \underline{\boldsymbol{\psi}}^{(k)}) = E_{\underline{\boldsymbol{\psi}}^{(k)}} \left[ l_c(\underline{\boldsymbol{\psi}}) \mid (y_1, \mathbf{x}'_1)', \dots, (y_n, \mathbf{x}'_n)' \right]. \quad (13)$$

In order to do this, we need to calculate  $E_{\underline{\boldsymbol{\psi}}^{(k)}}(Z_{ng} \mid y_n, \mathbf{x}_n)$ ,  $E_{\underline{\boldsymbol{\psi}}^{(k)}}(V_n \mid y_n, \mathbf{x}_n, z_n)$ ,  $E_{\underline{\boldsymbol{\psi}}^{(k)}}(\tilde{V}_n \mid y_n, \mathbf{x}_n, z_n)$ ,  $E_{\underline{\boldsymbol{\psi}}^{(k)}}(U_n \mid \mathbf{x}_n, z_n)$ , and  $E_{\underline{\boldsymbol{\psi}}^{(k)}}(\tilde{U}_n \mid \mathbf{x}_n, z_n)$ , for  $n = 1, \dots, N$  and  $g = 1, \dots, G$ , where  $\tilde{U}_n = \ln U_n$  and  $\tilde{V}_n = \ln V_n$ . It follows that

$$\begin{aligned}
E_{\underline{\boldsymbol{\psi}}^{(k)}}(Z_{ng} \mid y_n, \mathbf{x}_n) &= \tau_{ng}^{(k)} \\
&= \frac{\pi_g^{(k)} h_t(y_n \mid \mathbf{x}_n; \boldsymbol{\xi}_g^{(k)}, \zeta_g^{(k)}) h_{td}(\mathbf{x}_n; \boldsymbol{\theta}_g^{(k)}, \nu_g^{(k)})}{p(y_n, \mathbf{x}_n; \underline{\boldsymbol{\psi}}^{(k)}),}
\end{aligned} \quad (14)$$

$$\begin{aligned}
E_{\underline{\boldsymbol{\psi}}^{(k)}}(V_n \mid y_n, \mathbf{x}_n, z_{ng} = 1) &= \nu_{ng}^{(k)} \\
&= \frac{\zeta_g^{(k)} + 1}{\zeta_g^{(k)} + \delta \left[ y_n, \mu(\mathbf{x}_n; \boldsymbol{\beta}_g^{(k)}); \sigma_g^{2(r)} \right]}
\end{aligned} \quad (15)$$

and

$$\begin{aligned}
E_{\underline{\boldsymbol{\psi}}^{(k)}}(U_n \mid y_n, \mathbf{x}_n, z_{ng} = 1) &= u_{ng}^{(k)} \\
&= \frac{\nu_g^{(k)} + d}{\nu_g^{(k)} + \delta(\mathbf{x}_n, \boldsymbol{\mu}_g^{(k)}; \boldsymbol{\Sigma}_g^{(k)}),}
\end{aligned} \quad (16)$$

where the expectations are affected (see the subscript) using the current fit  $\underline{\boldsymbol{\psi}}^{(k)}$  for  $\underline{\boldsymbol{\psi}}$  ( $n = 1, \dots, N$  and  $g = 1, \dots, G$ ). Regarding the last two expectations, from the standard theory on the gamma distribution, we have that

$$\begin{aligned}
E_{\underline{\boldsymbol{\psi}}^{(k)}}(\tilde{V}_n \mid y_n, \mathbf{x}_n, z_{ng} = 1) &= \tilde{\nu}_{ng}^{(k)} \\
&= \ln \nu_{ng}^{(k)} + \psi\left(\frac{\zeta_g^{(k)} + 1}{2}\right) - \ln\left(\frac{\zeta_g^{(k)} + 1}{2}\right)
\end{aligned} \quad (17)$$

and

$$\begin{aligned} E_{\psi^{(k)}}(\bar{U}_n | \mathbf{x}_n, z_{ng} = 1) &= \bar{u}_{ng}^{(k)} \\ &= \ln u_{ng}^{(k)} + \psi\left(\frac{v_g^{(k)} + d}{2}\right) - \ln\left(\frac{v_g^{(k)} + d}{2}\right), \end{aligned} \quad (18)$$

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

$$\mathcal{Q}(\underline{\psi}; \underline{\psi}^{(k)}) = \mathcal{Q}_1(\underline{\pi}; \underline{\psi}^{(k)}) + \mathcal{Q}_2(\underline{\xi}; \underline{\psi}^{(k)}) + \mathcal{Q}_3(\underline{\zeta}; \underline{\psi}^{(k)}) + \mathcal{Q}_4(\underline{\vartheta}; \underline{\psi}^{(k)}) + \mathcal{Q}_5(\underline{v}; \underline{\psi}^{(k)}), \quad (19)$$

where

$$\mathcal{Q}_1(\underline{\pi}; \underline{\psi}^{(k)}) = \sum_{n=1}^N \sum_{g=1}^G \tau_{ng}^{(k)} \ln \pi_g, \quad (20)$$

$$\mathcal{Q}_2(\underline{\xi}; \underline{\psi}^{(k)}) = \sum_{n=1}^N \sum_{g=1}^G \tau_{ng}^{(k)} \mathcal{Q}_{2n}(\underline{\xi}_g; \underline{\psi}^{(k)}), \quad (21)$$

$$\mathcal{Q}_3(\underline{\zeta}; \underline{\psi}^{(k)}) = \sum_{n=1}^N \sum_{g=1}^G \tau_{ng}^{(k)} \mathcal{Q}_{3n}(\zeta_g; \underline{\psi}^{(k)}), \quad (22)$$

$$\mathcal{Q}_4(\underline{\vartheta}; \underline{\psi}^{(k)}) = \sum_{n=1}^N \sum_{g=1}^G \tau_{ng}^{(k)} \mathcal{Q}_{4n}(\underline{\vartheta}_g; \underline{\psi}^{(k)}) \quad (23)$$

and

$$\mathcal{Q}_5(\underline{v}; \underline{\psi}^{(k)}) = \sum_{n=1}^N \sum_{g=1}^G \tau_{ng}^{(k)} \mathcal{Q}_{5n}(v_g; \underline{\psi}^{(k)}), \quad (24)$$

with

$$\mathcal{Q}_{2n}(\underline{\xi}_g; \underline{\psi}^{(k)}) = \frac{1}{2} \left\{ -\ln(2\pi) + \bar{v}_{ng}^{(k)} - \ln \sigma_g^2 - v_{ng} \delta[y_n, \mu(\mathbf{x}_n; \underline{\beta}_g); \sigma_g^2] \right\}$$

and

$$\mathcal{Q}_{4n}(\underline{\vartheta}_g; \underline{\psi}^{(k)}) = \frac{1}{2} \left[ -d \ln(2\pi) + d \bar{u}_{ng}^{(k)} - \ln |\underline{\Sigma}_g| - u_{ng} \delta(\mathbf{x}_n, \underline{\mu}_g; \underline{\Sigma}_g) \right],$$

and where, on ignoring terms not involving  $\zeta_g$  and  $v_g$ , respectively,

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

and

$$\mathcal{Q}_{5n}(v_g; \underline{\psi}^{(k)}) = -\ln \Gamma\left(\frac{v_g}{2}\right) + \frac{v_g}{2} \ln \frac{v_g}{2} + \frac{v_g}{2} \left[ \bar{u}_{ng}^{(k)} - \ln u_{ng}^{(k)} + \sum_{n=1}^N (\ln u_{ng}^{(k)} - u_{ng}^{(k)}) \right].$$

#### 4.2. M-step

On the M-step, at the  $(k+1)$ th iteration, it follows from (19) that  $\underline{\pi}^{(k+1)}$ ,  $\underline{\xi}^{(k+1)}$ ,  $\underline{\zeta}^{(k+1)}$ ,  $\underline{\vartheta}^{(k+1)}$ , and  $\underline{v}^{(k+1)}$  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  $\zeta_g^{(k+1)}$  and  $v_g^{(k+1)}$  need to be computed iteratively.

The updated estimates of the mixture weights are

$$\pi_g^{(k+1)} = \sum_{n=1}^N \tau_{ng}^{(k)} / n, \quad (25)$$

while those of  $\boldsymbol{\theta}_g$ ,  $g = 1, \dots, G$ , result

$$\boldsymbol{\mu}_g^{(k+1)} = \sum_{n=1}^N \tau_{ng}^{(k)} u_{ng}^{(k)} \mathbf{x}_n / \sum_{n=1}^N \tau_{ng}^{(k)} u_{ng}^{(k)} \quad (26)$$

and

$$\boldsymbol{\Sigma}_g^{(k+1)} = \sum_{n=1}^N \tau_{ng}^{(k)} u_{ng}^{(k)} (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)}) (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)})' / \sum_{n=1}^N \tau_{ng}^{(k)} u_{ng}^{(k)}, \quad (27)$$

where, as motivated for example in Shoham (2002), the true denominator  $\sum_n \tau_{ng}^{(k)}$  of (27) has been changed to yield a significantly faster convergence for the EM algorithm.

Regarding the updated estimates of  $\boldsymbol{\xi}_g$ ,  $g = 1, \dots, G$ , maximization of (21), after some algebra, yields

$$\boldsymbol{\beta}_{1g}^{(k+1)} = \left( \begin{array}{c} \frac{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n \mathbf{x}_n'}{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)}} - \frac{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n \sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n'}{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} \sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)}} \end{array} \right)^{-1} \cdot \left( \begin{array}{c} \frac{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} y_n \mathbf{x}_n}{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)}} - \frac{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} y_n \sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} \sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)}} \end{array} \right), \quad (28)$$

$$\beta_{0g}^{(k+1)} = \frac{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} y_n}{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)}} - \boldsymbol{\beta}_{1g}^{(k+1)'} \frac{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)}} \quad (29)$$

and

$$\sigma_g^{2(k+1)} = \sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)} [y_n - (\beta_{0g}^{(k+1)} + \boldsymbol{\beta}_{1g}^{(k+1)'} \mathbf{x}_n)]^2 / \sum_{n=1}^N \tau_{ng}^{(k)} v_{ng}^{(k)}, \quad (30)$$

where the denominator of (30) has been modified in line with what was explained for equation (27).

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

$$\sum_{n=1}^N \frac{\partial}{\partial \zeta_g} Q_{3n}(\zeta_g; \boldsymbol{\psi}^{(k)}) = 0 \quad (31)$$

and

$$\sum_{n=1}^N \frac{\partial}{\partial \nu_g} Q_{5n}(\nu_g; \boldsymbol{\psi}^{(k)}) = 0, \quad (32)$$

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

$$\begin{aligned} -\psi\left(\frac{\zeta_g}{2}\right) + \ln \frac{\zeta_g}{2} + 1 + \frac{1}{N_g^{(k)}} \sum_{n=1}^N \tau_{ng}^{(k)} (\ln v_{ng}^{(k)} - v_{ng}^{(k)}) + \\ \psi\left(\frac{\nu_g^{(k)} + 1}{2}\right) - \ln\left(\frac{\nu_g^{(k)} + 1}{2}\right) = 0 \end{aligned} \quad (33)$$

and

$$\begin{aligned}
& -\psi\left(\frac{\nu_g}{2}\right) + \ln \frac{\nu_g}{2} + 1 + \frac{1}{N_g^{(k)}} \sum_{n=1}^N \tau_{ng}^{(k)} (\ln u_{ng}^{(k)} - u_{ng}^{(k)}) + \\
& \psi\left(\frac{\nu_g^{(k)} + d}{2}\right) - \ln\left(\frac{\nu_g^{(k)} + d}{2}\right) = 0,
\end{aligned} \tag{34}$$

where  $N_g^{(k)} = \sum_n \tau_{ng}^{(k)}$ ,  $g = 1, \dots, G$ .

## 5. Computational issues and partition evaluation

This section presents some issues concerning practical implementation of the EM algorithm described in Section 4 (see also Appendix A).

### 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  $\psi^{(0)}$  (see, e.g., Bagnato & Punzo, 2013). An alternative approach, more natural in the authors' opinion, is to specify a value for  $\mathbf{z}_n^{(0)}$ ,  $n = 1, \dots, N$  (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 helps in obtaining the 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  $\mathbf{z}_n^{(0)}$  are randomly drawn from a multinomial distribution with probabilities  $(1/G, \dots, 1/G)$ . Once the EM-estimates  $\widehat{\tau}_{ng}^{NN-VE}$  and  $\widehat{\tau}_{ng}^{NN-EV}$  of the posterior probabilities have been obtained for these models, we can compute the maximum *a posteriori* (MAP) classification, say  $\text{MAP}(\widehat{\tau}_{ng}^{NN-VE}) = \widehat{z}_{ng}^{NN-VE}$  and  $\text{MAP}(\widehat{\tau}_{ng}^{NN-EV}) = \widehat{z}_{ng}^{NN-EV}$ , where

$$\text{MAP}(\widehat{\tau}_{ng}) = \widehat{z}_{ng} = \begin{cases} 1 & \text{if } \max_j \{\widehat{\tau}_{nj}\} \text{ occurs in component } g \\ 0 & \text{otherwise.} \end{cases}$$

Then, 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. Thus,  $\widehat{z}_{ng}^{NN-VE}$  is used to initialize the EM of both  $tN$ -VE and  $Nt$ -VE, obtaining  $\widehat{z}_{ng}^{tN-VE}$  and  $\widehat{z}_{ng}^{Nt-VE}$ , respectively, while  $\widehat{z}_{ng}^{NN-EV}$  is used to initialize the EM of both  $tN$ -EV and  $Nt$ -EV, leading to  $\widehat{z}_{ng}^{tN-EV}$  and  $\widehat{z}_{ng}^{Nt-EV}$ , 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. 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.

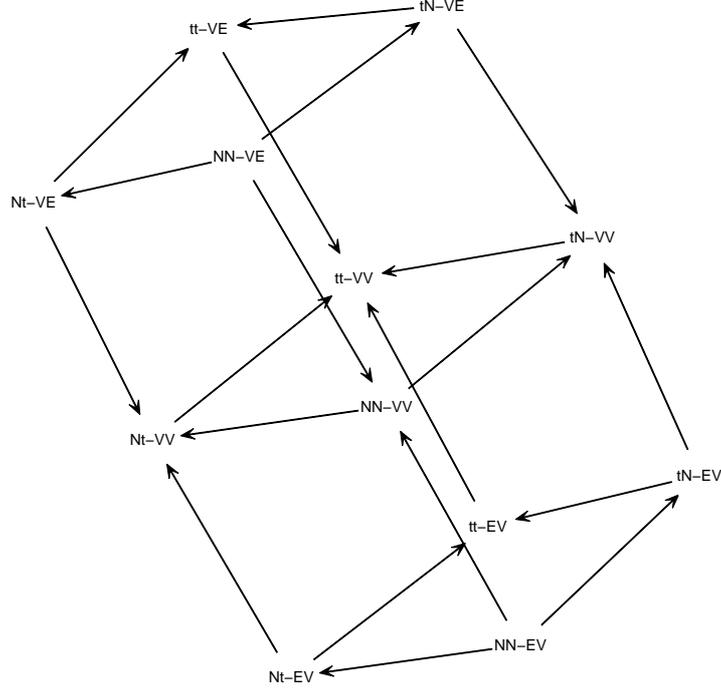


Figure 1: Relationships among the models in the hierarchical initialization strategy. Arrows are oriented from the model used for initialization to the model to be estimated.

### 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  $k + 1$  (Böhning et al., 1994) is given by

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

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

## 6. Model selection and clustering performance

In model-based clustering, model selection criteria are commonly used to choose the best model and to select the number of groups. Among them, we will adopt the Bayesian information criterion (BIC; Schwarz, 1978)

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

where  $\widehat{\psi}$  is the ML estimate of  $\psi$ ,  $l(\widehat{\psi})$  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), and the integrated completed likelihood (ICL; Biernacki et al., 2000) in the formulation given by Andrews & McNicholas (2011)

$$\text{ICL} \approx \text{BIC} + \sum_{n=1}^N \sum_{g=1}^G \text{MAP}(\widehat{\tau}_{ng}) \ln \widehat{\tau}_{ng}. \quad (35)$$

A different ICL definition is used by Baek & McLachlan (2011). The two definitions differ on whether or not it is the MAP of the fuzzy clustering in the first part of the entropy. It is not immediately clear from Biernacki et al. (2000) which definition is correct. We have chosen the formulation in (35) because it appears more widely adopted in literature (see, e.g., McNicholas & Murphy, 2008, 2010; McNicholas & Subedi, 2012).

In order to evaluate the clustering performance in cases in which the true classification is known, the adjusted Rand index (ARI; Hubert & Arabie, 1985), and the misclassification rate will be taken into account. We recall that 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:

- GENDER = gender of the respondent;
- HEIGHT = height of the respondent, measured in centimeters;
- WEIGHT = weight of the respondent, measured in kilograms;
- 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. The considered data are available at <http://www.economia.unict.it/punzo/>. 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 WEIGHT and HEIGHT, then on HEIGHT and 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: HEIGHT and 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 displays a bar plot of the HEIGHT variable, including the overall empirical marginal density as well as the empirical marginal densities, for  $G_M$  and  $G_F$ , weighted according to their sizes; bars are color-coded, using a gray scale, with respect 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  $G_F$ . In these terms note also that:

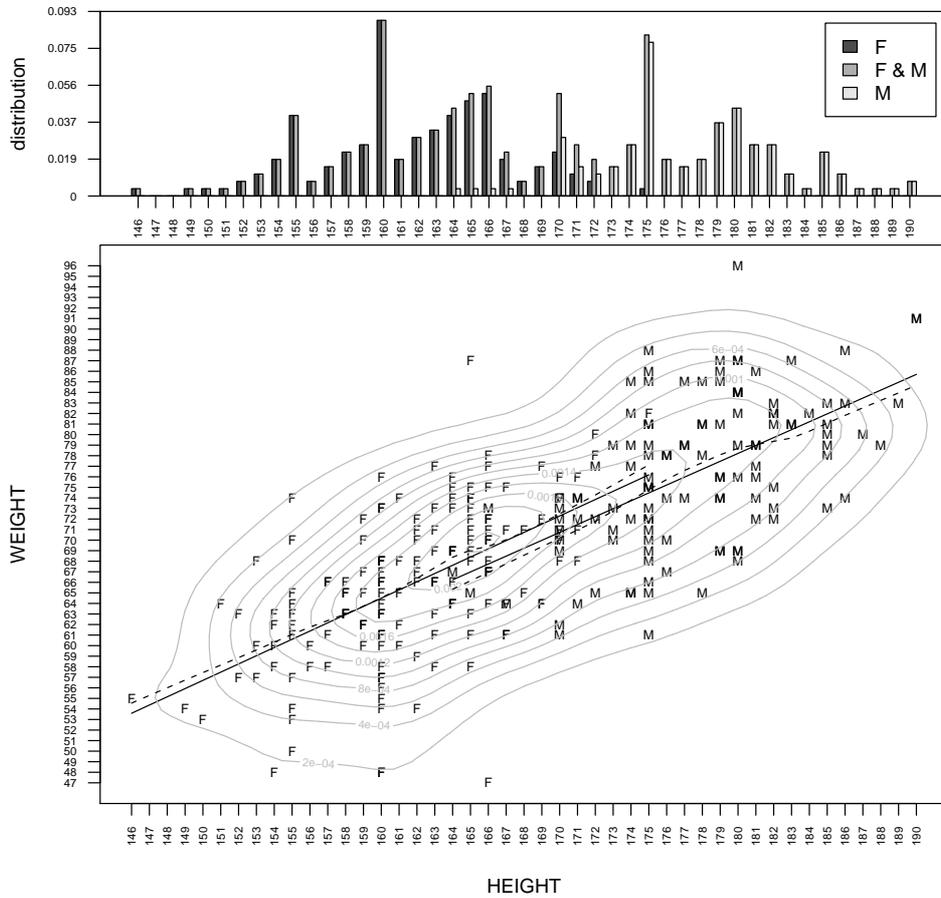


Figure 2: Student Data: CW-plot of HEIGHT and WEIGHT for 119 male, and 151 female, students (M denotes male and F female).

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: Student Data: 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	
$NN$	<b>-3726.197</b>	-3756.561	-3742.947	<b>-3750.466</b>	-3880.260	-3767.213	$NN$	0.750	0.008	0.750
$tN$	-3737.394	-3762.160	-3754.144	-3761.663	-3885.858	-3778.409	$tN$	0.750	0.008	0.750
$Nt$	-3731.795	-3766.517	-3749.642	-3756.064	-3869.845	-3773.484	$Nt$	0.750	0.005	<b>0.776</b>
$tt$	-3742.992	-3772.115	-3760.839	-3767.261	-3875.443	-3784.681	$tt$	0.750	0.005	<b>0.776</b>

$NN$ -VE (Gaussian marginal and conditional component densities and equal linear model between clusters) is the best model according to both BIC (-3726.197) and ICL (-3750.466). The corresponding CW-plot is displayed

in Figure 3. As for the ARI is concerned, in practice we have similar results for all models of type VE and VV. Thus, the group structure of the data is due to different intra-group distributions for the covariates, while the linear

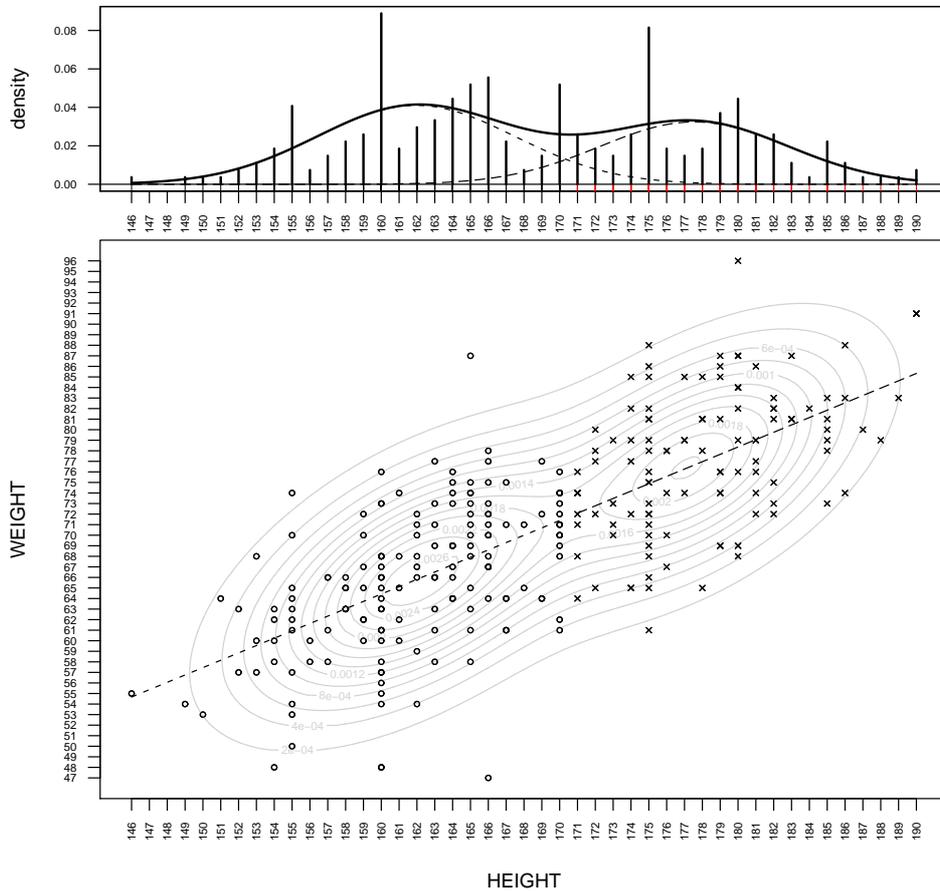


Figure 3: Student Data: CW-plot of HEIGHT and WEIGHT for  $NN$ -VE ( $G = 2$ ).

relationship is homogenous. In other words, this is a case of assignment dependence that a standard finite mixture of linear regressions is not able to represent. In order to show it empirically, we have also fitted a mixture of  $G = 2$  linear Gaussian regressions by means of the `flexmix` function of the R-package `flexmix` (Leisch, 2004). The group-conditional distribution of  $Y|X$  is Gaussian like in  $NN$ -VE. Figure 4 highlights that the mixture model with a fixed covariate is not able to recognize the group-structure of the data. This is also confirmed by an ARI value equal to 0.00288.

### 7.1.2. Second scenario: HEIGHT.F and HEIGHT

Figure 5 shows the CW-plot of HEIGHT.F and HEIGHT by considering the classification induced by GENDER. Although, also in this case, linearity between variables appear to be reasonable, the linear models for the two groups differ, especially in terms of intercept. Note also that, the  $F$ -test of homoscedasticity of the residuals in the two groups gives a  $p$ -value of 0.086 while the  $t$ -tests for equal slopes and equal intercepts provide practically null  $p$ -values.

As in Section 7.1.1, we fit the linear CWMs, with  $G = 2$ , ignoring the true classification induced by GENDER. The values of BIC, ICL, and ARI for the twelve models are given in Table 3. In this case, the best model is  $NN$ -EV (see also the corresponding CW-plot in Figure 6). The fitted model also appears to be a good compromise in terms of the ARI values of Table 3(c). Differently from the first scenario, here the group-structure is due to the different

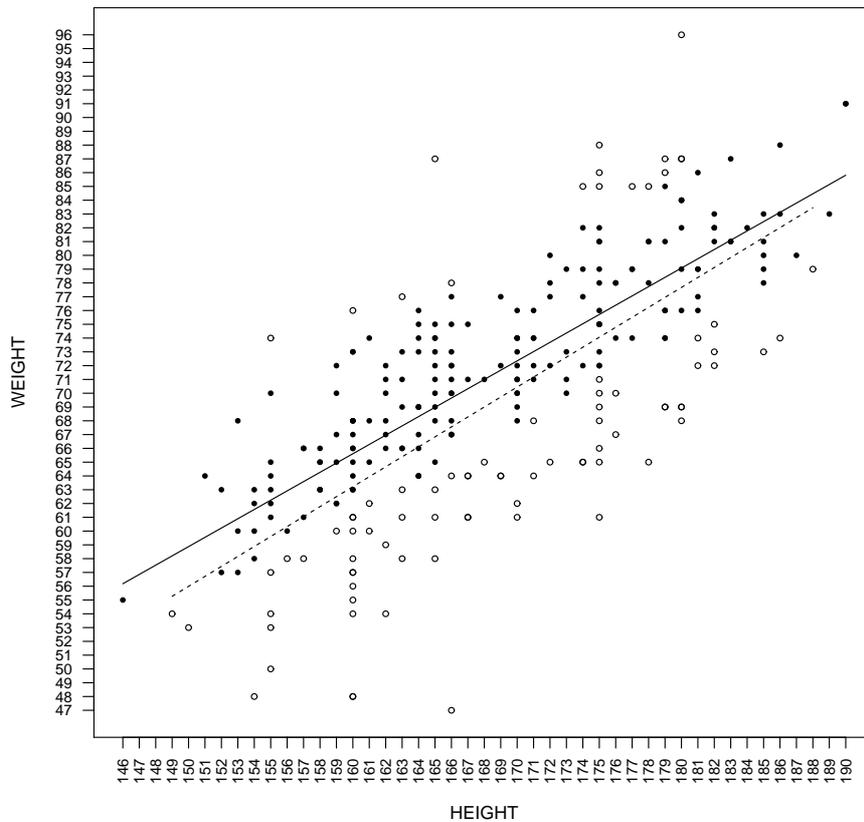


Figure 4: Student Data: Scatter plot of WEIGHT versus HEIGHT. The two types of lines and symbols displayed arise from the fit of a mixture of  $G = 2$  Gaussian regressions.

Table 3: Student Data: 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		
<i>NN</i>	-3726.339	<b>-3594.401</b>	-3601.955	<i>NN</i>	-3822.623	<b>-3597.252</b>	-3605.016	<i>NN</i>	0.009	0.898	<b>0.912</b>
<i>tN</i>	-3737.536	-3599.999	-3613.152	<i>tN</i>	-3833.820	-3602.850	-3616.212	<i>tN</i>	0.009	0.898	<b>0.912</b>
<i>Nt</i>	-3731.937	-3605.598	-3613.152	<i>Nt</i>	-3828.221	-3608.449	-3616.212	<i>Nt</i>	0.009	0.898	<b>0.912</b>
<i>tt</i>	-3743.134	-3611.196	-3624.348	<i>tt</i>	-3839.418	-3614.047	-3627.409	<i>tt</i>	0.009	0.898	<b>0.912</b>

intra-group linear models, while the distribution of the covariate is homogenous. This is an example of assignment independence which can be recognized by a simple mixture of  $G = 2$  linear (Gaussian) regressions too.

## 7.2. Tourist data

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 (2013) and are available at [http://www.robertocellini.it/doc/master\\_specializzazione/Cellini-Cuccia\\_ApEc2013\\_data1996-20](http://www.robertocellini.it/doc/master_specializzazione/Cellini-Cuccia_ApEc2013_data1996-20). The CW-plot of the labeled data (with respect to months) is shown in Figure 7. It is straightforward to note how the heterogeneity of the data reveals a clear group-structure. Figure 8 shows the values of the BIC and the ICL for the

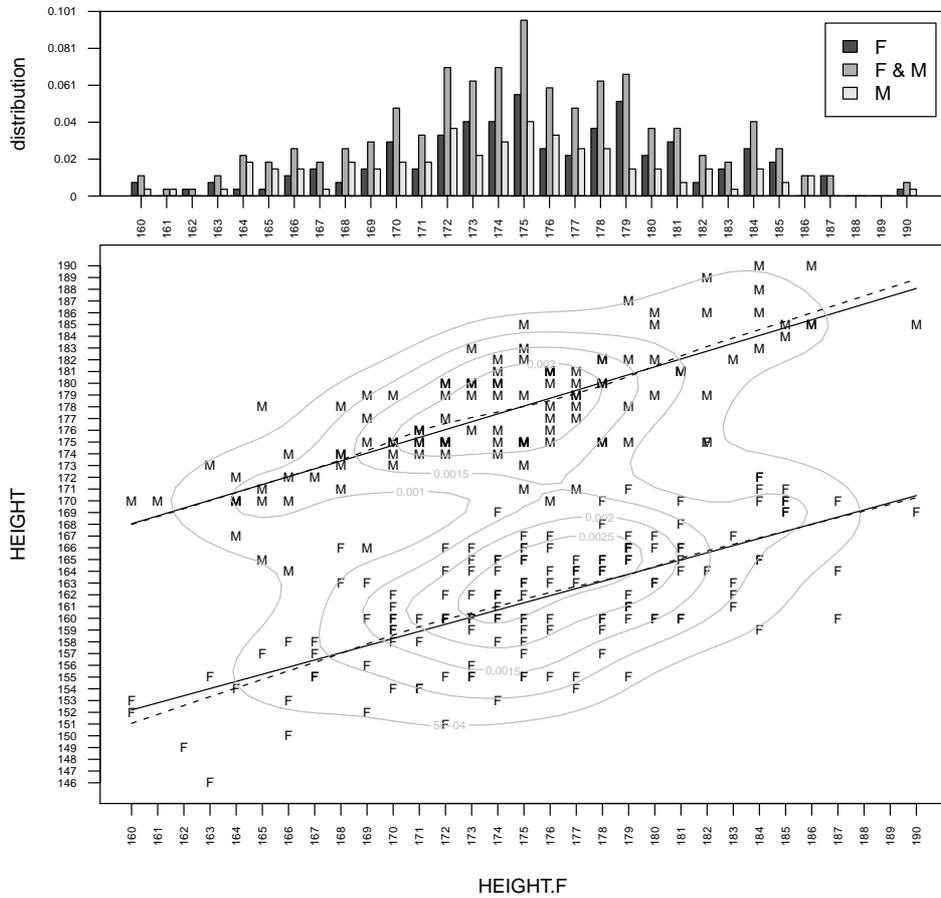


Figure 5: Student Data: CW-plot of HEIGHT and HEIGHT.F for 119 male and 151 female, students (M denotes male and F female).

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 9. 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

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

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

in November, which concern years 2006 and 2010). In particular, we have:

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

**Group 2** : units in June and September,

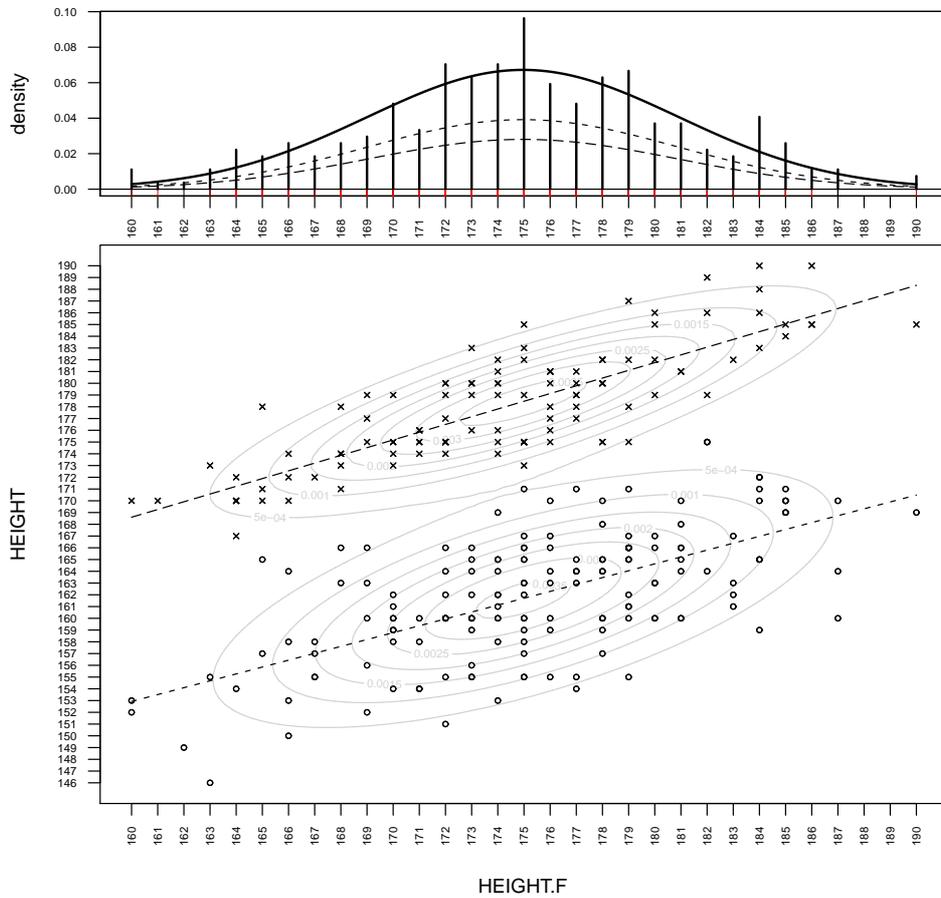


Figure 6: Student Data: CW-plot of HEIGHT.F and HEIGHT for NN-EV ( $G = 2$ ).

**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.

### 7.3. Crab data

The third application, based on the very popular crab data set of Campbell & Mahon (1974) on the genus *Leptograpsus*, has the aim of showing that the  $t$ -based linear CWMs ( $tN$ -VE,  $tN$ -EV,  $tN$ -VV,  $Nt$ -VE,  $Nt$ -EV,  $Nt$ -VV,  $tt$ -VE,  $tt$ -EV, and  $tt$ -VV) can provide more robust classification than the linear (completely) Gaussian ones ( $NN$ -VE,  $NN$ -EV, and  $NN$ -VV). Attention is focused on the sample of  $N = 100$  blue crabs, there being  $N_1 = 50$  males (group 1) and  $N_2 = 50$  females (group 2). Each specimen having  $p = 2$  measurements (in millimeters): the rear width ( $RW = Y$ ) and the length along the midline ( $CL = X$ ) of the carapace.

Following the scheme of McLachlan & Peel (2000, Section 7.8), some outliers were introduced by substituting the original value of  $y_{25}$  (11.9) with some atypical values (-15, -10, -5, and 0). This leads to four different “perturbed” data sets which are displayed in Figure 10.

Table 5 reports the number of misallocated observations for each of the twelve models and each perturbed version of the original data set. Estimates are obtained by directly using  $G = 2$ . The last two columns report the minimum

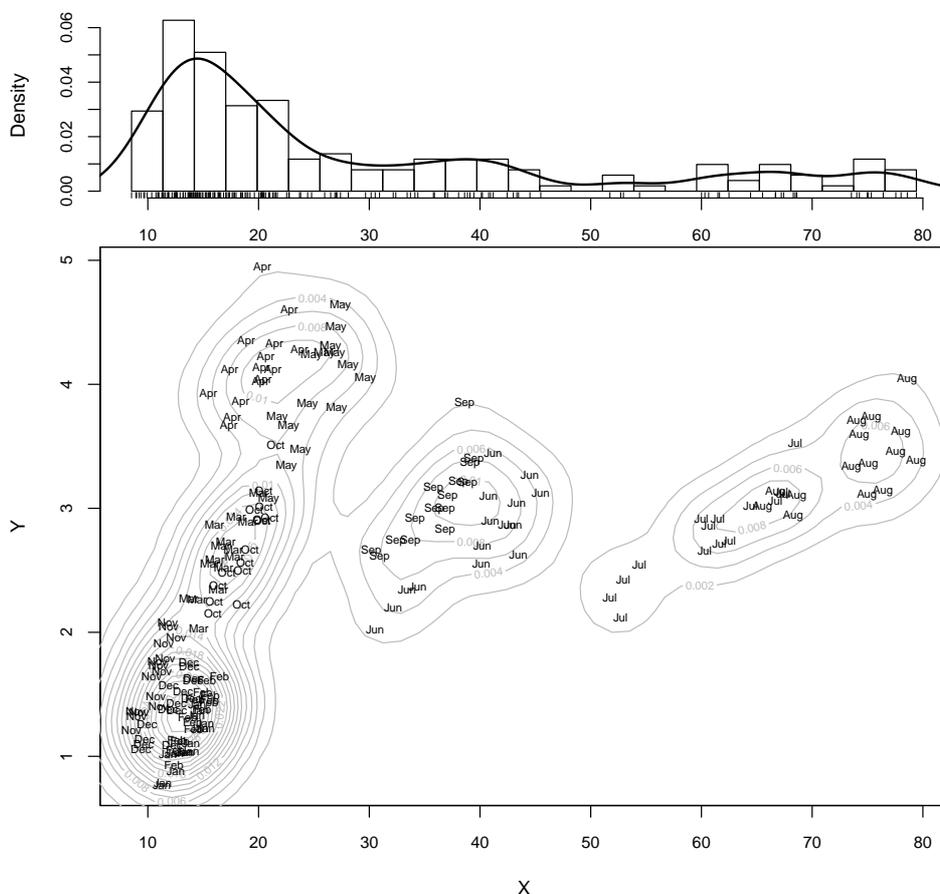


Figure 7: Tourist data: CW-plot of *tourist overnights* ( $X$ , in millions) and *attendance at museums and monuments* ( $Y$ , in millions) in Italy over the period from January 1996 to December 2010 ( $N = 180$ ). The univariate normal kernel density of  $X$  is superimposed on the histogram. The isodensities from a bivariate normal kernel density estimator are also visualized on the scatter plot. Month abbreviations are used as labels in the scatter plot.

$y_{25}$	linear Gaussian CWMs (A)			$t$ -based linear CWMs (B)									min(A)	min(B)
	$NN$ -VE	$NN$ -EV	$NN$ -VV	$tN$ -VE	$tN$ -EV	$tN$ -VV	$Nt$ -VE	$Nt$ -EV	$Nt$ -VV	$tt$ -VE	$tt$ -EV	$tt$ -VV		
-15	40	49	49	40	49	49	40	<b>16</b>	49	40	<b>16</b>	49	40	<b>16</b>
-10	40	49	50	40	49	50	40	<b>16</b>	25	40	<b>16</b>	25	40	<b>16</b>
-5	40	49	50	40	49	50	40	<b>13</b>	24	40	<b>13</b>	24	40	<b>13</b>
0	40	49	50	40	49	50	40	<b>13</b>	21	40	<b>13</b>	21	40	<b>13</b>

Table 5: Crab data: Comparison of the number of misallocated observations when fitting the family of linear CWMs on the sample of  $N = 100$  blue crabs. Bold numbers highlight the best results for each perturbed data set.

number of misallocated observations computed over the linear Gaussian CWMs and the  $t$ -based linear CWMs, respectively. From the bold numbers in Table 5 follows that some of the  $t$ -based linear CWMs, that is  $Nt$ -EV and  $tt$ -EV, are systematically more robust than the linear Gaussian CWMs (see also the results for  $Nt$ -VV and  $tt$ -VV). In particular, since the perturbations are inserted “vertically” on the  $Y$ -variable, the best performers have the  $t$  distribution for  $p(y|x, \Omega_g)$ ,  $g = 1, 2$ .

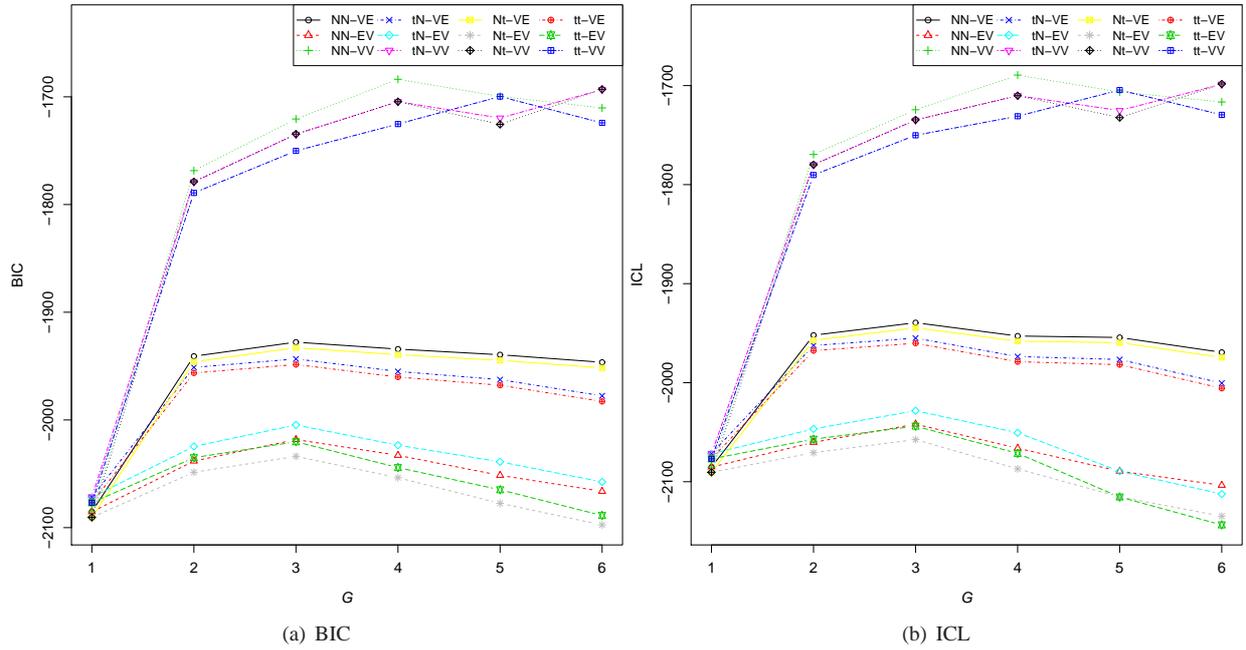


Figure 8: Tourist data: Values of the BIC and ICL ( $G = 1, \dots, 6$ ).

#### 7.4. *f.voles* data

The fourth application is based on the *f.voles* data set described in Flury (1997, Table 5.3.7) and available in the R-package *Flury*. This is an example with more than one covariate. Data refer to measurements on  $N = 86$  female voles from two species, *M. californicus* ( $N_1 = 45$ ) and *M. ochrogaster* ( $N_2 = 45$ ). Variables used here are: Species denoting the two species, Age measured in days, along with other six measurements related to skull (in units of 0.1 mm). The latter are named as in Airoidi & Hoffmann (1984):  $L_2$  = condylo-incisive length,  $L_9$  = length of incisive foramen,  $L_7$  = alveolar length of upper molar tooth row,  $B_3$  = zygomatic width,  $B_4$  = interorbital width, and  $H_1$  = skull height. The scatter plot matrix for grouped-data is shown in Figure 11.

The purpose of Airoidi & Hoffmann (1984) was to study age variability in *M. californicus* and *M. ochrogaster* and predict age on the basis of the skull measurements. In this study, we assume that data are unlabelled with respect to Species and compare the classification provided by the three approaches: the family of linear CWMs, mixtures of linear Gaussian regressions (estimated by the R-package *flexmix*), and parsimonious mixtures of Gaussian distributions (estimated using the R-package *mclust*; see Fraley et al., 2012, for details). For the first two classes of models, Age is the response variable  $Y$  and the  $d = 6$  skull measurements are the  $X$  variable. For parsimonious mixtures of Gaussian distributions, the vector  $(Y, X)'$  is considered as a whole. All the considered models have been fitted with  $G = 2$ .

In the family of linear CWMs, the two models providing the largest values for the BIC and the ICL were *NN-EV* and *NN-VE* (BIC: *NN-EV* =  $-3890.397$ , *NN-VE* =  $-3895.917$ ; ICL: *NN-VE* =  $-3896.143$ , *NN-EV* =  $-3902.788$ ). In particular, the two criteria selected a different model, although both the BIC and the ICL yielded quite close values for *NN-EV* and *NN-VE*. On the contrary, the resulting misclassification errors were very different: *NN-VE* (selected by the ICL) yielded a perfect classification, while *NN-EV* (selected by the BIC) yielded a misclassification error of 38.37%. A closer look to the membership probabilities showed that *NN-VE* led to a sharp classification (the entropy term in the ICL resulted 0.23), while the *NN-EV* led to a quite fuzzy classification (the entropy term resulted 12.39). We checked also the AIC for both models, and this agreed with ICL. Thus, *NN-VE* will be the only linear CWM considered hereafter. In the family of parsimonious mixtures of Gaussian distributions, the best model resulted *EEE* (homoscedastic group-covariance matrices; see Fraley et al., 2012 for details). Thus, we compared the performance of three Gaussian-based models whose classification results are reported in Table 6. The finite mixture of Gaussian regressions was the worse approach, reporting a misclassification rate of 0.40698. On the contrary, and surprisingly,

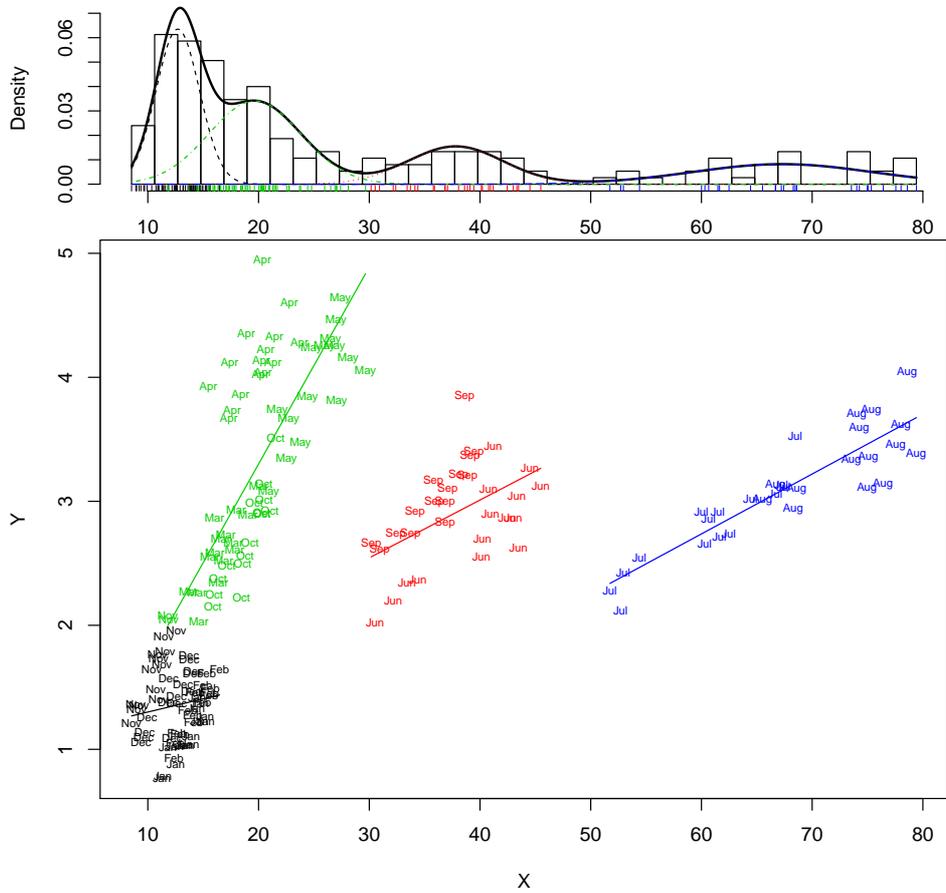


Figure 9: Tourist data: CW-plot of model  $NN-VV$  with  $G = 4$  components ( $X =$  “tourist overnights”, in millions, and  $Y =$  “attendance at museums and monuments”, in millions).

	CWM $NN-VE$	flexmix	mclust model EEE
ARI	1.00000	0.02430	0.90810
misclassification error	0.00000	0.40698	0.02326

Table 6: f.voies data: classification results using different mixture-based approaches ( $G = 2$ ).

our model  $NN-VE$  attains a perfect classification of the data (we remark the same optimal classification performance was obtained by all the “-VE” models in our family).

In conclusion, this is an example of “strong” assignment dependence where the group structure only depends by a different distribution of the covariates between the two groups (see also Proposition 1).

## 8. Comparing the BIC and the ICL

A simulation study is described for comparing the performance of the BIC and the ICL with regard to the proposed family of models. Five scenarios are presented where data are simulated according to the following models:  $NN-EV$ ,  $NN-VE$ ,  $NN-VV$ ,  $Nt-VE$ , and  $tN-EV$ . In each scenario, 50 data sets of size  $n = 400$  are simulated with:  $d = 1$ ,  $G = 2$ ,

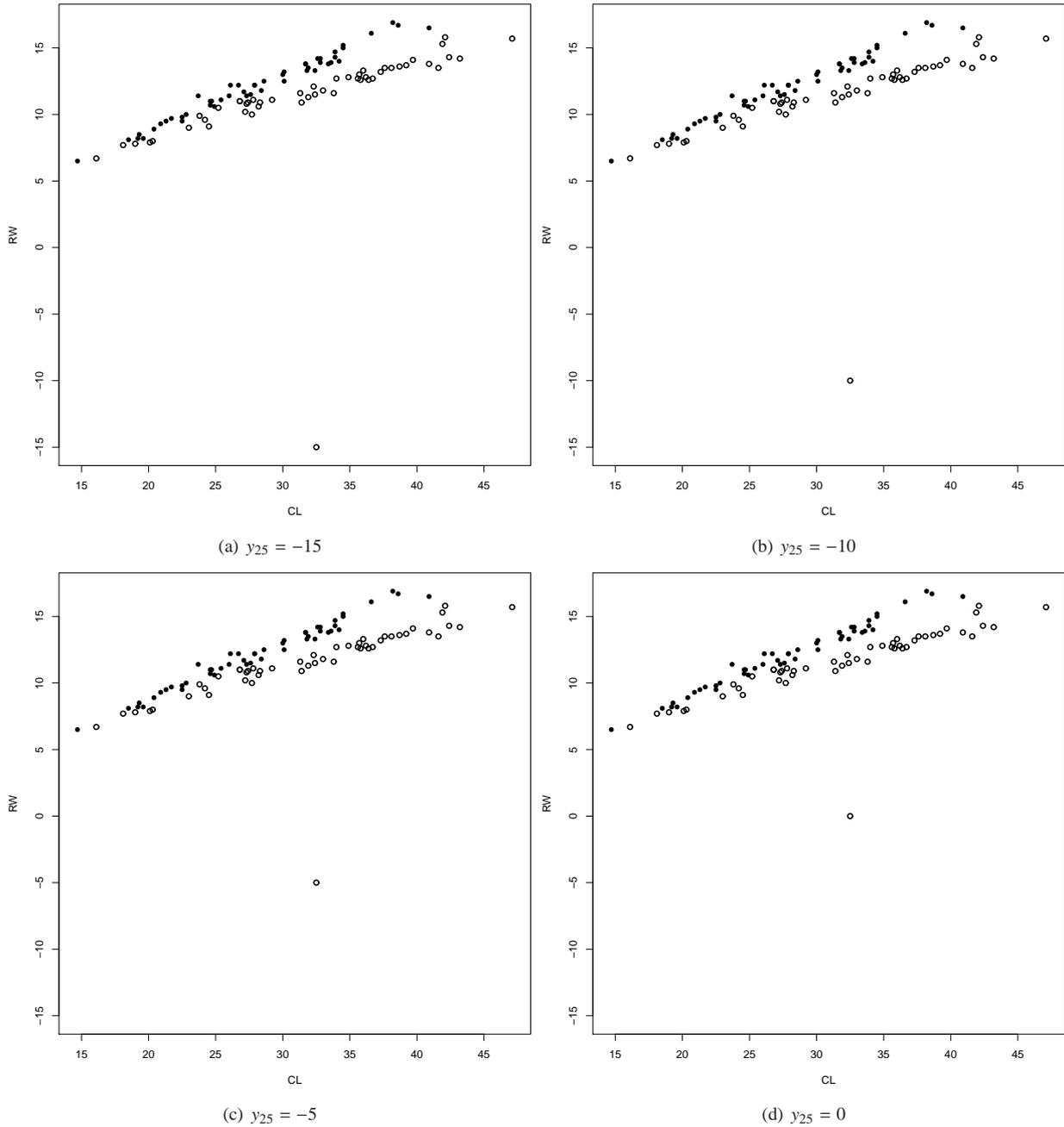


Figure 10: Scatter plots of the sample of  $N = 100$  blue crabs with different values for  $y_{25}$ . The variables are rear width (RW) and length along the midline (CL) of the carapace, for  $N_1 = 50$  males and  $N_2 = 50$  females ( $\circ$  denotes male and  $\bullet$  female).

and varying parameters. The choice of considering different parameters is made to avoid particular configurations which may favor one of the competitive model selection criteria.

In each replication, the generating (true) model is specified as follows:

- the mixture weight  $\pi_1$  is randomly generated by a uniform distribution on  $[0.2, 0.8]$ ;
- as the variable  $X$  is concerned, we refer to equation (3). Note that, we prefer to leave the matrix notation of the

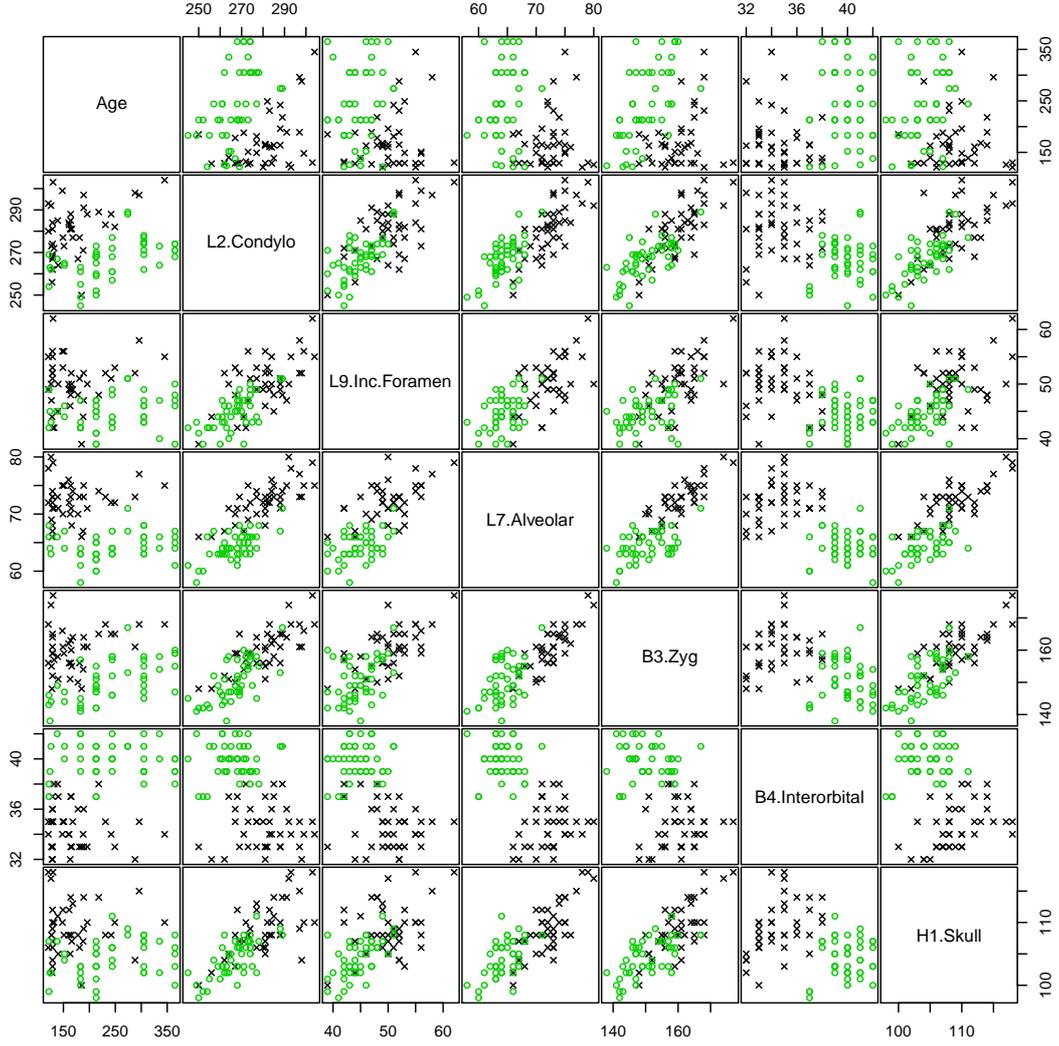


Figure 11: Scatter plot matrix of f.voles data (○ and × denote species *Microtus ochrogaster* and *M. californicus*, respectively).

parameters  $\mu_g$  and  $\Sigma_g$  even if, being  $d = 1$ , they are indeed scalar values. In particular

- if the model assumes  $\mu_1 \neq \mu_2$ , then  $\mu_1$  and  $\mu_2$  are randomly generated by a standard normal distribution. If  $\mu_1 = \mu_2 = \mu$ , then  $\mu$  is drawn by a standard normal distribution;
- if the model assumes  $\Sigma_1 \neq \Sigma_2$ , then  $\Sigma_1$  and  $\Sigma_2$  are randomly generated by a  $\chi_1^2$  distribution. If  $\Sigma_1 = \Sigma_2 = \Sigma$ , then  $\Sigma$  is drawn by a  $\chi_1^2$  distribution;
- if  $p(x|\Omega_1)$  and  $p(x|\Omega_2)$  are assumed to be  $t$ ;
  - \* if the model assumes  $\nu_1 \neq \nu_2$ , then  $\nu_1$  and  $\nu_2$  are randomly generated by a uniform distribution on  $[2, 5]$ ;
  - \* if the model assumes  $\nu_1 = \nu_2 = \nu$ , then  $\nu$  is drawn by a uniform distribution on  $[2, 5]$ ;
- as the variable  $Y$  is concerned, by referring to equation (2), we have that
  - if the model assumes  $\beta_{01} \neq \beta_{02}$  and  $\beta_{11} \neq \beta_{12}$ , then  $\beta_{01}$  and  $\beta_{02}$  are randomly generated by a standard normal distribution while  $\beta_{11}$  and  $\beta_{12}$  are drawn from a uniform distribution on  $[-2, 2]$ . If  $\beta_{01} = \beta_{02} = \beta_0$

and  $\beta_{11} = \beta_{12} = \beta_1$ , then  $\beta_0$  is generated by a standard normal distribution and  $\beta_1$  is drawn from a uniform distribution on  $[-2, 2]$ ;

- if the model assumes  $\sigma_1^2 \neq \sigma_2^2$ , then  $\sigma_1^2$  and  $\sigma_2^2$  are randomly generated by a  $\chi_1^2$  distribution. If  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , then  $\sigma^2$  is generated by a  $\chi_1^2$  distribution;
- if  $p(y|\mathbf{x}, \Omega_1)$  and  $p(y|\mathbf{x}, \Omega_2)$  are assumed to be  $t$ 
  - \* if the model assumes  $\zeta_1 \neq \zeta_2$ , then  $\zeta_1$  and  $\zeta_2$  are randomly generated by a uniform distribution on  $[2, 5]$ ;
  - \* if the model assumes  $\zeta_1 = \zeta_2 = \zeta$ , then  $\zeta$  is drawn by a uniform distribution on  $[2, 5]$ .

The defined models guarantee various degrees of overlap between groups according to the generated parameters.

In each replication, the true model is adopted to generate the data set; thus, all the 12 models are fitted with  $G \in \{1, 2, 3\}$ , leading to a total of 36 fitted models. Table 7 and Table 8 show the results for the BIC and the ICL, respectively. Here, a value in position  $(i, j)$  has to be read as “number of times that the combination (model, number of groups) on column  $j$  is selected to fit the true model (with  $G = 2$ ) on row  $i$ ”. Bold numbers highlight the number of times that the pair (true model,  $G = 2$ ) is selected. Note that: the columns referred to models of type “ $tt$ ” are missing simply because they have never been selected, and the sum by row is greater than 50 because, when  $G = 1$  is selected, there is not difference between “-VV”, “-VE”, and “-EV” (see Section 3). By comparing the results in these tables, the BIC seems to perform better than the ICL. In particular, the ICL selects models with only one group a larger number of times than the BIC. This is probably induced by the scheme of definition of the true model that allows for groups with a strong overlap; thus, the entropy term of the ICL carries out a strong penalization which leads to the choice  $G = 1$ . Figure 12 and Figure 13 display two examples where this happens. From these examples we understand as it is difficult to establish the best model selection criterion; indeed, the ICL may be seen as better if the user actually does not want to separate two mixture components that are so similar that they do not constitute two different clusters in terms of interpretation. So, in general, it depends on the meaning of the data which criterion is better.

## 9. Conclusions and discussion

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 the proposed family to the model-based classification context. Moreover, the identifiability issue needs to be adequately addressed; a reference point is given by Hennig (2000). Finally, Section 8 presented first results to find out a suitable model selection criterion and motivates further research in this direction.

## Acknowledgements

The authors sincerely thank the Associate Editor and the referees for very helpful comments and valuable suggestions that have contributed to improving the quality of the manuscript.

## Appendix A. EM-constraints for parsimonious models

In the following we describe how to impose constraints on the EM algorithm, described in Section 4 for the most general model  $tt$ -VV, to obtain parameter estimates for all the other models in Table 1. To this end, the itemization given at the beginning of Section 3 will be considered as a benchmark scheme.

Fitted True	G	NN-EV			NN-VE			NN-VV			Nt-EV			Nt-VE			Nt-VV			tN-EV			tN-VE			tN-VV		
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
NN-EV	8	<b>40</b>	0	8	0	0	8	0	0	2	0	0	2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
NN-VE	5	0	0	5	<b>40</b>	0	5	0	0	4	0	0	4	0	0	4	0	0	0	0	0	0	0	0	1	0	0	0
NN-VV	0	0	0	0	0	0	0	<b>50</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nt-VE	0	1	0	0	0	0	0	0	0	5	0	0	5	<b>43</b>	1	5	0	0	0	0	0	0	0	0	0	0	0	0
tN-EV	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	<b>47</b>	0	2	0	0	2	0	0	

Table 7: Simulation results for the BIC. Values in the table show the number of times, over 50 replications, that the model, and number of groups, on the column are selected to fit the true model (with two groups) which appears in the corresponding row. Bold numbers highlight the largest number of times that the model selection criteria selects the true model.

Fitted True $G$	$NN$ -EV			$NN$ -VE			$NN$ -VV			$Nt$ -EV			$Nt$ -VE			$Nt$ -VV			$tN$ -EV			$tN$ -VE			$tN$ -VV					
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3			
$NN$ -EV	14	<b>27</b>	0	14	0	0	14	0	0	9	0	0	9	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$NN$ -VE	13	0	0	13	<b>24</b>	0	13	0	0	0	0	0	0	0	0	0	0	0	12	0	0	12	1	0	12	0	0	12	0	0
$NN$ -VV	0	0	0	0	0	0	0	<b>47</b>	0	1	0	0	1	0	0	1	0	0	2	0	0	2	0	0	2	0	0	2	0	0
$Nt$ -VE	0	0	0	0	0	0	0	0	0	15	0	0	15	<b>30</b>	1	15	0	0	2	2	0	2	0	0	2	0	0	2	0	0
$tN$ -EV	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	16	<b>33</b>	0	16	0	0	16	0	0	16	0	0

Table 8: Simulation results for the ICL. Values in the table show the number of times, over 50 replications, that the model, and number of groups, on the column are selected to fit the true model (with two groups) which appears in the corresponding row. Bold numbers highlight the number of times that the model selection criteria selects the true model.

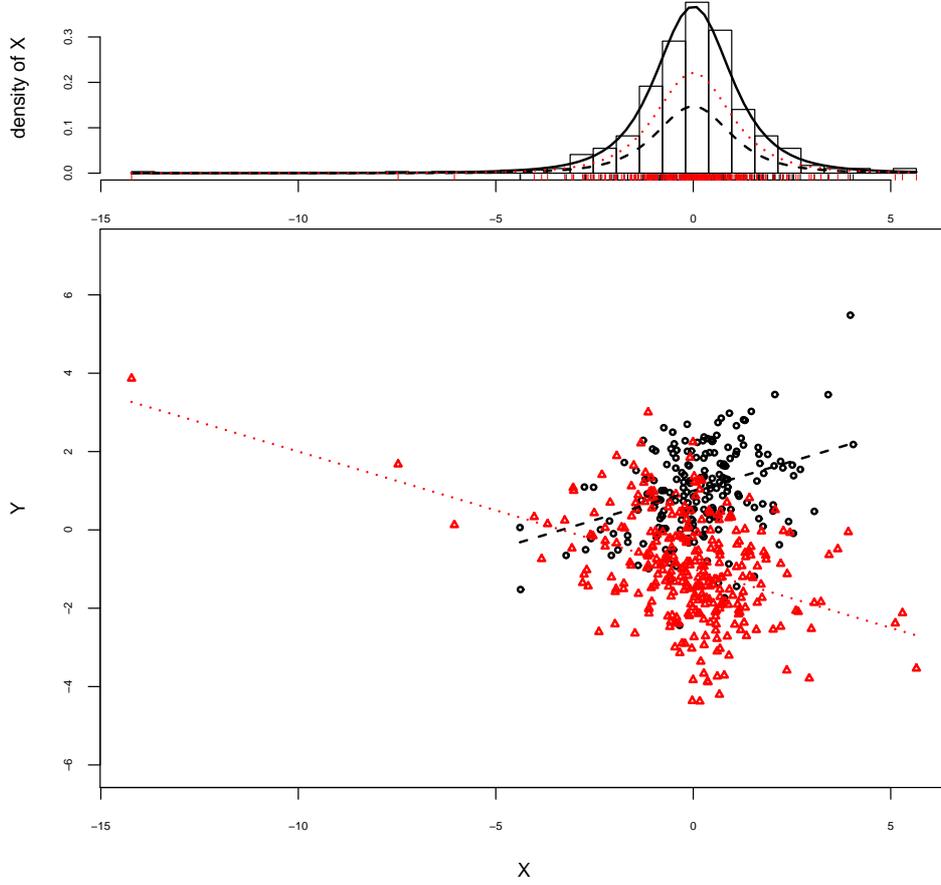


Figure 12: CW-plot of data randomly generated from a  $tN$ -EV model with  $G = 2$ .

#### Appendix A.1. Common $t$ for the component marginal densities

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

$$u_n^{(k)} = \frac{\nu^{(k)} + d}{\nu^{(k)} + \delta(\mathbf{x}_n, \boldsymbol{\mu}^{(k)}; \boldsymbol{\Sigma}^{(k)})} \quad (\text{A.1})$$

and

$$\widetilde{u}_n^{(k)} = \ln u_n^{(k)} + \psi\left(\frac{\nu^{(k)} + d}{2}\right) - \ln\left(\frac{\nu^{(k)} + d}{2}\right),$$

respectively. Furthermore, noting that  $\sum_g \tau_{ng} = 1$ , equations (23) and (24) can be rewritten as

$$Q_4(\boldsymbol{\vartheta}; \boldsymbol{\psi}^{(k)}) = \sum_{n=1}^N Q_{4n}(\boldsymbol{\vartheta}; \boldsymbol{\psi}^{(k)}) \quad (\text{A.2})$$

and

$$Q_5(\nu; \boldsymbol{\psi}^{(k)}) = \sum_{n=1}^N Q_{5n}(\nu; \boldsymbol{\psi}^{(k)}),$$

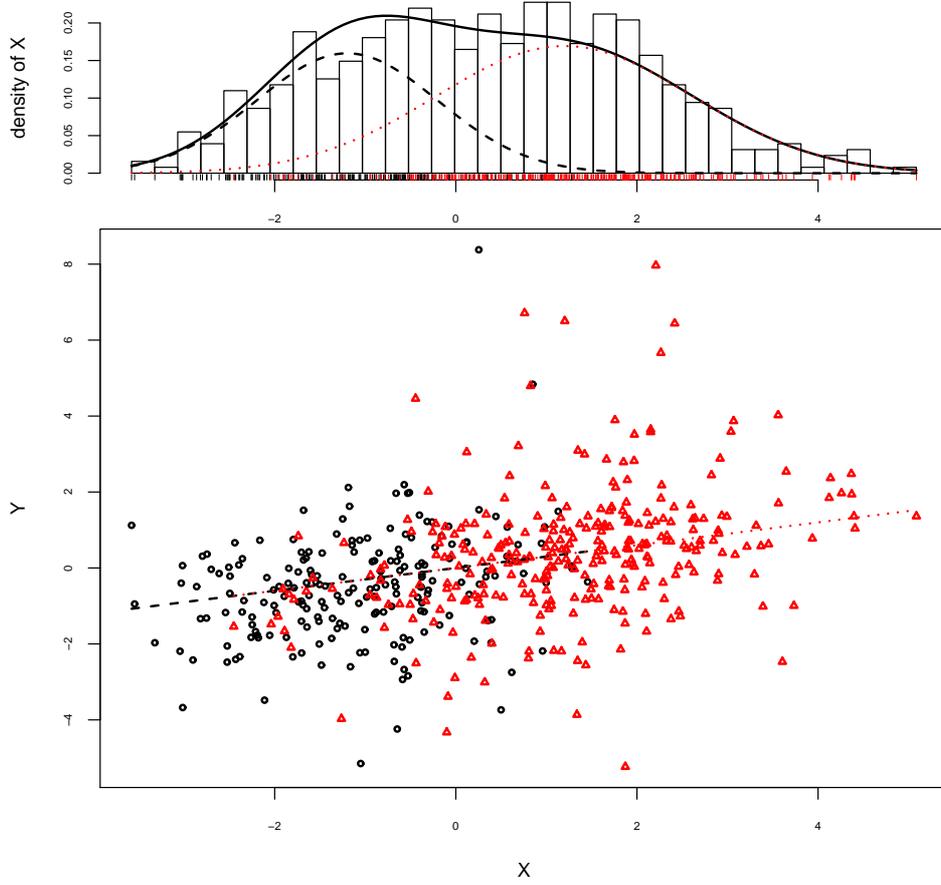


Figure 13: CW-plot of data randomly generated from a  $Nt$ -VE model with  $G = 2$ .

respectively, where

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

and

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

Maximization of (A.2), with respect to  $\boldsymbol{\vartheta}$ , leads to

$$\boldsymbol{\mu}^{(k+1)} = \frac{\sum_{n=1}^N u_n^{(k)} \mathbf{x}_n}{\sum_{n=1}^N u_n^{(k)}}$$

and

$$\boldsymbol{\Sigma}^{(k+1)} = \frac{\sum_{n=1}^N u_n^{(k)} (\mathbf{x}_n - \boldsymbol{\mu}^{(k+1)}) (\mathbf{x}_n - \boldsymbol{\mu}^{(k+1)})'}{\sum_{n=1}^N u_n^{(k)}}.$$

For the updating of  $v$ , we need to numerically solve the equation

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

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

$$-\psi\left(\frac{v}{2}\right) + \ln \frac{v}{2} + 1 + \sum_{n=1}^N (\ln u_n^{(k)} - u_n^{(k)}) + \psi\left(\frac{v^{(k)} + d}{2}\right) - \ln\left(\frac{v^{(k)} + d}{2}\right) = 0. \quad (\text{A.3})$$

*Appendix A.2. Common  $t$  for the component conditional densities*

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

$$v_n^{(k)} = \frac{\zeta^{(k)} + 1}{\zeta_g^{(k)} + \delta [y_n, \mu(\mathbf{x}_n; \boldsymbol{\beta}^{(k)}); \sigma^{2(r)}]} \quad (\text{A.4})$$

and

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

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

$$\mathcal{Q}_2(\boldsymbol{\xi}; \boldsymbol{\psi}^{(k)}) = \sum_{n=1}^N \mathcal{Q}_{2n}(\boldsymbol{\xi}; \boldsymbol{\psi}^{(k)}) \quad (\text{A.5})$$

and

$$\mathcal{Q}_3(\zeta; \boldsymbol{\psi}^{(k)}) = \sum_{n=1}^N \mathcal{Q}_{3n}(\zeta; \boldsymbol{\psi}^{(k)}),$$

respectively, where

$$\mathcal{Q}_{2n}(\boldsymbol{\xi}; \boldsymbol{\psi}^{(k)}) = \frac{1}{2} \left\{ -\ln(2\pi) + \tilde{v}_n^{(k)} - \ln \sigma^2 - v_n \delta [y_n, \mu(\mathbf{x}_n; \boldsymbol{\beta}); \sigma^2] \right\}$$

and

$$\mathcal{Q}_{3n}(\zeta; \boldsymbol{\psi}^{(k)}) = -\ln \Gamma\left(\frac{\zeta}{2}\right) + \frac{\zeta}{2} \ln \frac{\zeta}{2} + \frac{\zeta}{2} \left[ \tilde{v}_n^{(k)} - \ln v_n^{(k)} + \sum_{n=1}^N (\ln v_n^{(k)} - v_n^{(k)}) \right].$$

Maximization of (A.5), with respect to  $\boldsymbol{\xi}$ , leads to the updates

$$\begin{aligned} \boldsymbol{\beta}_1^{(k+1)} &= \left( \frac{\sum_{n=1}^N v_n^{(k)} \mathbf{x}_n \mathbf{x}_n'}{\sum_{n=1}^N v_n^{(k)}} - \frac{\sum_{n=1}^N v_n^{(k)} \mathbf{x}_n \sum_{n=1}^N v_n^{(k)} \mathbf{x}_n'}{\sum_{n=1}^N v_n^{(k)} \sum_{n=1}^N v_n^{(k)}} \right)^{-1} \\ &\quad \cdot \left( \frac{\sum_{n=1}^N v_n^{(k)} y_n \mathbf{x}_n}{\sum_{n=1}^N v_n^{(k)}} - \frac{\sum_{n=1}^N v_n^{(k)} y_n \sum_{n=1}^N v_n^{(k)} \mathbf{x}_n}{\sum_{n=1}^N v_n^{(k)} \sum_{n=1}^N v_n^{(k)}} \right), \\ \beta_0^{(k+1)} &= \frac{\sum_{n=1}^N v_n^{(k)} y_n}{\sum_{n=1}^N v_n^{(k)}} - \beta_1^{(k+1)'} \frac{\sum_{n=1}^N v_n^{(k)} \mathbf{x}_n}{\sum_{n=1}^N v_n^{(k)}} \end{aligned}$$

and

$$\sigma^{2(k+1)} = \sum_{n=1}^N v_n^{(k)} \left[ y_n - \left( \beta_0^{(k+1)} + \beta_1^{(k+1)'} \mathbf{x}_n \right) \right]^2 / \sum_{n=1}^N v_n^{(k)}.$$

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

$$\sum_{n=1}^N \frac{\partial}{\partial \zeta} Q_{3n}(\zeta; \underline{\psi}^{(k)}) = 0,$$

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

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

### Appendix A.3. Normal component marginal densities

The normal case for the component distributions of  $\mathbf{X}$  can be obtained, as stated previously, as a limiting case when  $v_g \rightarrow \infty$ ,  $g = 1, \dots, G$ . Then, in (16),  $u_{ng}^{(k)} \rightarrow 1$ . Substituting this value into (26) and (27), we obtain

$$\boldsymbol{\mu}_g^{(k+1)} = \sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{x}_n / \sum_{n=1}^N \tau_{ng}^{(k)}$$

and

$$\boldsymbol{\Sigma}_g^{(k+1)} = \sum_{n=1}^N \tau_{ng}^{(k)} (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)}) (\mathbf{x}_n - \boldsymbol{\mu}_g^{(k+1)})' / \sum_{n=1}^N \tau_{ng}^{(k)}.$$

Naturally, in this case, we do not compute the additional  $M$ -step maximizing  $Q_5(\underline{v}; \underline{\psi}^{(k)})$  in (24). Accordingly, for the sub-case  $\boldsymbol{\mu}_1 = \dots = \boldsymbol{\mu}_G = \boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_G = \boldsymbol{\Sigma}$ , in equation (A.1) we have  $u_n^{(k)} \rightarrow 1$  and the updated estimates of  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  become

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{n=1}^N \mathbf{x}_n$$

and

$$\boldsymbol{\Sigma} = \frac{1}{n} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}) (\mathbf{x}_n - \boldsymbol{\mu})',$$

which do not depend on the EM-iterations.

### Appendix A.4. Normal component conditional densities

The normal case for the component distributions of  $Y|\mathbf{X}$  can be obtained as a limiting case when  $\zeta_g \rightarrow \infty$ ,  $g = 1, \dots, G$ . Then, in (15),  $v_{ng}^{(k)} \rightarrow 1$ . Substituting this value into (28) and (29), we obtain

$$\boldsymbol{\beta}_{1g}^{(k+1)} = \begin{pmatrix} \frac{\sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{x}_n \mathbf{x}_n'}{\sum_{n=1}^N \tau_{ng}^{(k)}} - \frac{\sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{x}_n \sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{x}_n'}{\sum_{n=1}^N \tau_{ng}^{(k)} \sum_{n=1}^N \tau_{ng}^{(k)}} \\ \frac{\sum_{n=1}^N \tau_{ng}^{(k)} y_n \mathbf{x}_n}{\sum_{n=1}^N \tau_{ng}^{(k)}} - \frac{\sum_{n=1}^N \tau_{ng}^{(k)} y_n \sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^N \tau_{ng}^{(k)} \sum_{n=1}^N \tau_{ng}^{(k)}} \end{pmatrix}^{-1}.$$

$$\beta_{0g}^{(k+1)} = \frac{\sum_{n=1}^N \tau_{ng}^{(k)} y_n}{\sum_{n=1}^N \tau_{ng}^{(k)}} - \beta_{1g}^{(k+1)'} \frac{\sum_{n=1}^N \tau_{ng}^{(k)} \mathbf{x}_n}{\sum_{n=1}^N \tau_{ng}^{(k)}}$$

and

$$\sigma_g^{2(k+1)} = \sum_{n=1}^N \tau_{ng}^{(k)} \left[ y_n - (\beta_{0g}^{(k+1)} + \beta_{1g}^{(k+1)'} \mathbf{x}_n) \right]^2 / \sum_{n=1}^N \tau_{ng}^{(k)}.$$

We again do not compute the additional  $M$ -step maximizing  $Q_3(\zeta; \psi^{(k)})$  in (22). Accordingly, for the sub-case  $\beta_{11} = \dots = \beta_{1G} = \beta_1$ ,  $\beta_{01} = \dots = \beta_{0G} = \beta_0$ , and  $\sigma_1^2 = \dots = \sigma_G^2 = \sigma^2$ , in equation (A.4) we have  $v_n^{(k)} \rightarrow 1$  and the updated estimates of  $\beta_1$ ,  $\beta_0$ , and  $\sigma^2$  become

$$\beta_1 = \left( \frac{1}{n} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n' - \frac{1}{n^2} \sum_{n=1}^N \mathbf{x}_n \sum_{n=1}^N \mathbf{x}_n' \right)^{-1} \left( \frac{1}{n} \sum_{n=1}^N y_n \mathbf{x}_n - \frac{1}{n^2} \sum_{n=1}^N y_n \sum_{n=1}^N \mathbf{x}_n \right),$$

$$\beta_0 = \frac{1}{n} \sum_{n=1}^N y_n - \frac{1}{n} \beta_1' \sum_{n=1}^N \mathbf{x}_n$$

and

$$\sigma^2 = \frac{1}{n} \sum_{n=1}^N [y_n - (\beta_0 + \beta_1' \mathbf{x}_n)]^2,$$

which do not depend on the EM-iterations.

## References

- Airoldi, J. P., & Hoffmann, R. S. (1984). *Age variation in voles (Microtus californicus, M. ochrogaster) and its significance for systematic studies*. Occasional papers of the Museum of Natural History 111 University of Kansas Lawrence, KS.
- 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.
- Baek, J., & McLachlan, G. (2011). Mixtures of common  $t$ -factor analyzers for clustering high-dimensional microarray data. *Bioinformatics*, 27, 1269–1276.
- Bagnato, L., & Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the  $k$ -bumps algorithm. *Computational Statistics*, 28, 1571–1597.
- 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.
- Böhning, 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.
- Campbell, N. A., & Mahon, R. J. (1974). A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. *Australian Journal of Zoology*, 22, 417–425.
- Cellini, R., & Cuccia, T. (2013). Museum and monument attendance and tourism flow: A time series analysis approach. *Applied Economics*, 45, 3473–3482.
- 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.
- 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.
- Flury, B. (1997). *A first course in multivariate statistics*. New York: Springer.
- 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., Murphy, T. B., & Scrucca, L. (2012). *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. Technical report 597 Department of Statistics, University of Washington Seattle, Washington, USA.
- Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.
- Gershfeld, 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.
- Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17, 273–296.
- 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, 363–401.
- Ingrassia, S., Punzo, A., & Vittadini, G. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, 32.
- 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.
- 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., & Murphy, T. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, 18, 285–296.
- McNicholas, P., & Murphy, T. (2010). Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, 26, 2705–2712.
- 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.
- Peel, D., & McLachlan, G. (2000). Robust mixture modelling using the  $t$  distribution. *Statistics and Computing*, 10, 339–348.
- Punzo, A. (2014). Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. *Statistical Modelling*, 14, 257–291.
- Punzo, A., & Ingrassia, S. (2015a). Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, .
- Punzo, A., & Ingrassia, S. (2015b). Parsimonious generalized linear Gaussian cluster-weighted models. In I. Morlini, T. Minerva, & M. Vichi (Eds.), *Advances in Statistical Models for Data Analysis Studies in Classification, Data Analysis and Knowledge Organization*. Switzerland: Springer International Publishing. Forthcoming.
- Punzo, A., & McNicholas, P. D. (2013). *Robust Clustering via Parsimonious Mixtures of Contaminated Gaussian Distributions*. arXiv.org e-print 1305.4669 available at: <http://arxiv.org/abs/1305.4669>.
- Punzo, A., & McNicholas, P. D. (2014a). *Robust Clustering in Regression Analysis via the Contaminated Gaussian Cluster-Weighted Model*. arXiv.org e-print 1409.6019 available at: <http://arxiv.org/abs/1409.6019>.
- Punzo, A., & McNicholas, P. D. (2014b). *Robust High-Dimensional Modeling with the Contaminated Gaussian Distribution*. arXiv.org e-print 1408.2128 available at: <http://arxiv.org/abs/1408.2128>.
- R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- 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.
- Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, 7, 5–40.
- Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P. D. (2015). Cluster-weighted  $t$ -factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications*, 24.
- Titterton, 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.
- Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica*, 56, 362–375.
- 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.