Adv Data Anal Classif (2013) 7:5–40
DOI 10.1007/s11634-013-0124-8

REGULAR ARTICLE

Clustering and classification via cluster-weighted factor analyzers

Sanjeena Subedi · Antonio Punzo · Salvatore Ingrassia · Paul D. McNicholas

Received: 12 October 2012 / Revised: 15 January 2013 / Accepted: 23 January 2013 /
Published online: 7 March 2013

Abstract In model-based clustering and classification, the cluster-weighted model is a convenient approach when the random vector of interest is constituted by a response variable Y and by a vector $\boldsymbol{X}$ of p covariates. However, its applicability may be limited when p is high. To overcome this problem, this paper assumes a latent factor structure for $\boldsymbol{X}$ in each mixture component, under Gaussian assumptions. This leads to the cluster-weighted factor analyzers (CWFA) model. By imposing constraints on the variance of Y and the covariance matrix of $\boldsymbol{X}$, a novel family of sixteen CWFA models is introduced for model-based clustering and classification. The alternating expectation-conditional maximization algorithm, for maximum likelihood estimation of the parameters of all models in the family, is described; to initialize the algorithm, a 5-step hierarchical procedure is proposed, which uses the nested structures of the models within the family and thus guarantees the natural ranking among the sixteen likelihoods. Artificial and real data show that these models have very good clustering and classification performance and that the algorithm is able to recover the parameters very well.

S. Subedi · P. D. McNicholas
Department of Mathematics and Statistics,
University of Guelph, Guelph, ON N1G 2W1, Canada
e-mail: ssubedi@uoguelph.ca

P. D. McNicholas
e-mail: paul.mcnicholas@uoguelph.ca

A. Punzo · S. Ingrassia (✉)
Department of Economics and Business,
University of Catania, Corso Italia 55, 95129 Catania, Italy
e-mail: s.ingrassia@unict.it

A. Punzo
e-mail: antonio.punzo@unict.it

Keywords Cluster-weighted models · Factor analysis · Mixture models · Parsimonious models

Mathematics Subject Classification (2010) 62H30 · 62H25

1 Introduction

Mixture models have been used for clustering for at least fifty years (Wolfe 1963, 1970). Following the inception of the expectation-maximization (EM) algorithm (Dempster et al. 1977), parameter estimation became more manageable and applications of mixture models for clustering and classification became more common (see Titterington et al. 1985; McLachlan and Basford 1988 for examples). With the increasing availability of computational power, the popularity of mixture models has grown consistently since the mid-1990s, including notable work by Banfield and Raftery (1993), Celeux and Govaert (1995), Ghahramani and Hinton (1997), Tipping and Bishop (1999), McLachlan and Peel (2000a), Fraley and Raftery (2002), Dean et al. (2006), Bouveyron et al. (2007), McNicholas and Murphy (2008), McNicholas and Murphy (2010a), Karlis and Santourian (2009), Lin (2010), Scrucca (2010), Baek et al. (2010), Andrews et al. (2011), Browne et al. (2012), McNicholas and Subedi (2012), and Browne and McNicholas (2012), amongst others.

Consider data $(\boldsymbol{x}, y)$ that are realizations of the pair $(\boldsymbol{X}, Y)$ defined on some space Ω, where $Y \in \mathbb{R}$ is a response variable and $\boldsymbol{X} \in \mathbb{R}^p$ is a vector of covariates. Suppose that Ω can be partitioned into G groups, say $\Omega_1, \ldots, \Omega_G$. Let $p(\boldsymbol{x}, y)$ be the joint density of $(\boldsymbol{X}, Y)$. In this paper, we shall consider a mixture model having density of the form

$$p(\boldsymbol{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \phi\left(y|\boldsymbol{x}; m(\boldsymbol{x}; \boldsymbol{\beta}_g), \sigma_g^2\right) \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right), \tag{1}$$

where $\phi(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ denotes a p-variate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, and $\phi(y|\boldsymbol{x}; m(\boldsymbol{x}; \boldsymbol{\beta}_g), \sigma_g^2)$ denotes the (Gaussian) density of the conditional distribution of $Y|\boldsymbol{x}$ with mean $m(\boldsymbol{x}; \boldsymbol{\beta}_g) = \beta_{0g} + \boldsymbol{\beta}_{1g}'\boldsymbol{x}$, $\beta_{0g} \in \mathbb{R}$ and $\boldsymbol{\beta}_{1g} \in \mathbb{R}^p$, and variance σ_g^2. Model parameters are denoted by $\boldsymbol{\theta}$. The density in (1) defines the linear Gaussian cluster-weighted model (see, e.g., Gershenfeld 1997; Schöner 2000). Quite recently, the cluster-weighted model (CWM) has been developed under more general assumptions: Ingrassia et al. (2012a) consider t distributions, Ingrassia et al. (2013) introduce a family of twelve parsimonious linear t CWMs for model-based clustering, and Ingrassia et al. (2012b) propose CWMs with categorical responses. Finally, Punzo (2012) introduces the polynomial Gaussian CWM as a flexible tool for clustering and classification.

In the mosaic of work around the use of mixture models for clustering and classification, CWMs have their place in applications with random covariates. Indeed, differently from finite mixture of regressions (see, e.g., Leisch 2004; Frühwirth-Schnatter 2006), which are examples of mixture models with fixed covariates, the CWM allows

for assignment dependence: the covariate distributions for each group Ω_g can also be distinct. In clustering and classification terms, this means that $\boldsymbol{X}$ can directly affect the clustering results and this represents an advantage, for most applications, with respect to the fixed covariates approach (Hennig 2000).

However, the applicability of model (1) in high dimensional $\boldsymbol{X}$-spaces still remains a challenge. In particular, the number of parameters for this model is $(G-1)+G(p+2)+G[p+p(p+1)/2]$, of which $Gp\,(p+1)\,/2$ are used for the component (or group) covariance matrices $\boldsymbol{\Sigma}_g$ of $\boldsymbol{X}$, $g=1,\ldots,G$, and this increases quadratically with p. To overcome this issue, we assume a latent Gaussian factor structure for $\boldsymbol{X}$, in each mixture-component, which leads to the factor regression model (FRM) of Y on $\boldsymbol{x}$ (see West 2003; Wang et al. 2007, and Carvalho et al. 2008). The FRM assumes $\boldsymbol{\Sigma}_g=\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g'+\boldsymbol{\Psi}_g$, where the loading matrix is a $p\times q$ matrix of parameters typically with $q\ll p$ and the noise matrix $\boldsymbol{\Psi}_g$ is a diagonal matrix. The adoption of this group covariance structure in (1) leads to the linear Gaussian cluster-weighted factor analyzers (CWFA) model, which is characterized by $G\,[pq-q\,(q-1)\,/2]+Gp$ parameters for the group covariance matrices. The CWFA model follows the principle of the general form of mixtures of factor analyzers regarding $\boldsymbol{X}$. Mixtures of factor analyzers were introduced by Ghahramani and Hinton (1997) and further developed by Tipping and Bishop (1999) and McLachlan and Peel (2000b). More recent results have been also provided in Montanari and Viroli (2010, 2011).

Starting from the works of McNicholas and Murphy (2008), McNicholas (2010), Ingrassia et al. (2012b), and Ingrassia et al. (2013), a novel family of sixteen mixture models—obtained as special cases of the linear Gaussian CWFA by conveniently constraining the component variances of $Y|\boldsymbol{x}$ and $\boldsymbol{X}$—is introduced to facilitate parsimonious model-based clustering and classification in the defined paradigm. The novelty of this proposal is that it considers a family of models that use factor models in a regression context to effectively, and flexibly, reduce dimensionality.

The paper is organized as follows. Sect. 2 recalls the FRM; the linear Gaussian CWFA models are introduced in Sect. 3. Model fitting with the alternating expectation-conditional maximization (AECM) algorithm is presented in Sect. 4. Section 5 addresses computational details on some aspects of the AECM algorithm and discusses model selection and evaluation. Artificial and real data are considered in Sect. 6, and the paper concludes with discussion and suggestions for further work in Sect. 7.

2 The factor regression model

The factor analysis model (Spearman 1904; Bartlett 1953), for the p-dimensional variable $\boldsymbol{X}$, postulates that

$$\boldsymbol{X}=\boldsymbol{\mu}+\boldsymbol{\Lambda}\boldsymbol{U}+\boldsymbol{e}, \tag{2}$$

where $\boldsymbol{U}\sim N_q\left(\boldsymbol{0},\boldsymbol{I}_q\right)$ is a q-dimensional ($q\ll p$) vector of latent factors, $\boldsymbol{\Lambda}$ is a $p\times q$ matrix of factor loadings, and $\boldsymbol{e}\sim N_p\left(\boldsymbol{0},\boldsymbol{\Psi}\right)$, with $\boldsymbol{\Psi}=\text{diag}\left(\psi_1^2,\ldots,\psi_p^2\right)$, independent of $\boldsymbol{U}$. Then $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Lambda}\boldsymbol{\Lambda}'+\boldsymbol{\Psi})$ and, conditional on $\boldsymbol{u}$, results in $\boldsymbol{X}|\boldsymbol{u}\sim N_p\left(\boldsymbol{\mu}+\boldsymbol{\Lambda}\boldsymbol{u},\boldsymbol{\Psi}\right)$.

Model (2) can be considered jointly with the standard (linear) regression model $Y = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{X} + \varepsilon$ leading to the Factor Regression Model (FRM) (see West 2003; Wang et al. 2007, and Carvalho et al. 2008);

$$Y = \beta_0 + \boldsymbol{\beta}_1' (\boldsymbol{\mu} + \boldsymbol{\Lambda} \boldsymbol{U} + \boldsymbol{e}) + \varepsilon = \left(\beta_0 + \boldsymbol{\beta}_1' \boldsymbol{\mu}\right) + \boldsymbol{\beta}_1' \boldsymbol{\Lambda} \boldsymbol{U} + \left(\boldsymbol{\beta}_1' \boldsymbol{e} + \varepsilon\right),$$

where ε is assumed to be independent of $\boldsymbol{U}$ and $\boldsymbol{e}$. The mean and variance of Y are given by

$$\begin{aligned} \mathbb{E}(Y) &= \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{\mu} \\ \text{Var}(Y) &= \text{Var}\left(\boldsymbol{\beta}_1' \boldsymbol{\Lambda} \boldsymbol{U}\right) + \text{Var}\left(\boldsymbol{\beta}_1' \boldsymbol{e}\right) + \text{Var}(\varepsilon) \\ &= \boldsymbol{\beta}_1' \boldsymbol{\Lambda} \boldsymbol{\Lambda}' \boldsymbol{\beta}_1 + \boldsymbol{\beta}' \boldsymbol{\Psi} \boldsymbol{\beta}_1 + \sigma^2 = \boldsymbol{\beta}_1' \left(\boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \Psi\right) \boldsymbol{\beta}_1 + \sigma^2, \end{aligned}$$

respectively, and so $Y \sim N\left(\beta_0 + \boldsymbol{\beta}_1' \boldsymbol{\mu}, \boldsymbol{\beta}_1' \left(\boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \Psi\right) \boldsymbol{\beta}_1 + \sigma^2\right)$.

Consider the triplet $\left(Y, \boldsymbol{X}', \boldsymbol{U}'\right)'$. Its mean is given by

$$\mathbb{E} \begin{bmatrix} Y \\ \boldsymbol{X} \\ \boldsymbol{U} \end{bmatrix} = \begin{bmatrix} \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{\mu} \\ \boldsymbol{\mu} \\ \boldsymbol{0} \end{bmatrix},$$

and because $\text{Cov}(\boldsymbol{X}, Y) = (\boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \boldsymbol{\Psi}) \boldsymbol{\beta}_1$ and $\text{Cov}(\boldsymbol{U}, Y) = \boldsymbol{\Lambda}' \boldsymbol{\beta}_1$, it results that

$$\text{Cov} \begin{bmatrix} Y \\ \boldsymbol{X} \\ \boldsymbol{U} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1' \boldsymbol{\Sigma} \boldsymbol{\beta}_1 + \sigma^2 & \boldsymbol{\beta}_1' \boldsymbol{\Sigma} & \boldsymbol{\beta}_1' \boldsymbol{\Lambda} \\ \boldsymbol{\Sigma} \boldsymbol{\beta}_1 & \boldsymbol{\Sigma} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}' \boldsymbol{\beta}_1 & \boldsymbol{\Lambda}' & \boldsymbol{I}_q \end{bmatrix},$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. Now, we can write the joint density of $\left(Y, \boldsymbol{X}', \boldsymbol{U}'\right)'$ as

$$p(y, \boldsymbol{x}, \boldsymbol{u}) = \phi(y | \boldsymbol{x}, \boldsymbol{u}) \, \phi(\boldsymbol{x} | \boldsymbol{u}) \, \phi(\boldsymbol{u}). \tag{3}$$

Here, the distribution and related parameters for both $\boldsymbol{X}|\boldsymbol{u}$ and $\boldsymbol{U}$ are known. Thus, we only need to analyze the distribution of $Y|\boldsymbol{x}, \boldsymbol{u}$. Importantly, $\mathbb{E}(Y|\boldsymbol{x}, \boldsymbol{u}) = \mathbb{E}(Y|\boldsymbol{x})$ and $\text{Var}(Y|\boldsymbol{x}, \boldsymbol{u}) = \text{Var}(Y|\boldsymbol{x})$, and so $Y|\boldsymbol{x}, \boldsymbol{u} \sim N\left(\beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}, \sigma^2\right)$; mathematical details are given in Appendix A. This implies that $\phi(y|\boldsymbol{x}, \boldsymbol{u}) = \phi(y|\boldsymbol{x})$ and, therefore, Y is conditionally independent of $\boldsymbol{U}$ given $\boldsymbol{X} = \boldsymbol{x}$, so that (3) becomes

$$p(y, \boldsymbol{x}, \boldsymbol{u}) = \phi(y|\boldsymbol{x}) \, \phi(\boldsymbol{x}|\boldsymbol{u}) \, \phi(\boldsymbol{u}). \tag{4}$$

Similarly, $\boldsymbol{U}|y, \boldsymbol{x} \sim N\left(\boldsymbol{\gamma}(\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{I}_q - \boldsymbol{\gamma} \boldsymbol{\Lambda}\right)$, where $\boldsymbol{\gamma} = \boldsymbol{\Lambda}' \left(\boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \boldsymbol{\Psi}\right)^{-1}$, and thus $\boldsymbol{U}$ is conditionally independent of Y given $\boldsymbol{X} = \boldsymbol{x}$. Therefore,

$$\begin{aligned} \mathbb{E}[\boldsymbol{U}|\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}] &= \boldsymbol{\gamma}(\boldsymbol{x} - \boldsymbol{\mu}), \quad \text{and} \\ \mathbb{E}\left[\boldsymbol{U}\boldsymbol{U}'|\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}\right] &= \boldsymbol{I}_q - \boldsymbol{\gamma} \boldsymbol{\Lambda} + \boldsymbol{\gamma} (\boldsymbol{x} - \boldsymbol{\mu}) (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\gamma}'. \end{aligned}$$

3 The modelling framework

3.1 The general model

Assume that for each Ω_g, $g = 1, \ldots, G$, the pair $(\boldsymbol{X}, Y)$ satisfies a FRM, that is

$$Y = \beta_{0g} + \boldsymbol{\beta}'_{1g}\boldsymbol{X} + \varepsilon_g \quad \text{with} \quad \boldsymbol{X} = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \boldsymbol{U}_g + \boldsymbol{e}_g, \tag{5}$$

where $\boldsymbol{\Lambda}_g$ is a $p \times q$ matrix of factor loadings, $\boldsymbol{U}_g \sim N_q\left(\boldsymbol{0}, \boldsymbol{I}_q\right)$ is the vector of factors, $\boldsymbol{e}_g \sim N_p\left(\boldsymbol{0}, \boldsymbol{\Psi}_g\right)$ are the errors, $\boldsymbol{\Psi}_g = \text{diag}\left(\psi_{1g}, \ldots, \psi_{pg}\right)$, and $\varepsilon_g \sim N(0, \sigma_g^2)$. Then the linear Gaussian CWM in (1) can be extended in order to include the underlying factor structure (5) for $\boldsymbol{X}$. In particular, by recalling that Y is conditionally independent of $\boldsymbol{U}$ given $\boldsymbol{X} = \boldsymbol{x}$ in the generic Ω_g, we get

$$p\left(\boldsymbol{x}, y; \boldsymbol{\theta}\right) = \sum_{g=1}^{G} \pi_g \phi\left(y | \boldsymbol{x}; m\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right), \sigma_g^2\right) \phi\left(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}'_g + \boldsymbol{\Psi}_g\right), \tag{6}$$

where $\boldsymbol{\theta} = \left\{\pi_g, \boldsymbol{\beta}_g, \sigma_g^2, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g; g = 1, \ldots, G\right\}$. Model (6) is the linear Gaussian CWFA, which we shall refer to as the CWFA model herein.

3.2 Parsimonious versions of the model

To introduce parsimony, we extend the linear Gaussian CWFA after the fashion of McNicholas and Murphy (2008) by allowing constraints across groups on σ_g^2, $\boldsymbol{\Lambda}_g$, and $\boldsymbol{\Psi}_g$, and on whether $\boldsymbol{\Psi}_g = \psi_g \boldsymbol{I}_p$ (isotropic assumption). The full range of possible constraints provides a family of sixteen different parsimonious CWFAs (Table 1).

Here, models are identified by a sequence of four letters. The letters refer to whether or not the constraints $\sigma_g^2 = \sigma^2$, $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, and $\boldsymbol{\Psi}_g = \psi_g \boldsymbol{I}_p$, respectively, are imposed. The constraints on the group covariances of $\boldsymbol{X}$ are in the spirit of McNicholas and Murphy (2008), while that on the group variances of Y are borrowed from Ingrassia et al. (2013). Each letter can be either C, if the corresponding constraint is applied, or U if the particular constraint is not applied. For example, model CUUC assumes equal Y variances between groups, unequal loading matrices, and unequal, but isotropic, noise.

3.3 Model-based classification

Suppose that m of the n observations in $\mathcal{S}$ are labeled. Within the model-based classification framework, we use all of the n observations to estimate the parameters in (6); the fitted model classifies each of the $n - m$ unlabeled observations through the corresponding maximum a posteriori probability (MAP). As a special case, if $m = 0$, we obtain the clustering scenario. Drawing on Hosmer Jr. (1973), Titterington et al. (1985,

Table 1 Parsimonious covariance structures derived from the CWFA model

Model ID	Y variance	Loading matrix	Error variance	Isotropic	Covariance parameters
UUUU	Unconstrained	Unconstrained	Unconstrained	Unconstrained	$G + G\left[pq - q\left(q-1\right)/2\right] + Gp$
UUUC	Unconstrained	Unconstrained	Unconstrained	Constrained	$G + G\left[pq - q\left(q-1\right)/2\right] + G$
UUCU	Unconstrained	Unconstrained	Constrained	Unconstrained	$G + G\left[pq - q\left(q-1\right)/2\right] + p$
UUCC	Unconstrained	Unconstrained	Constrained	Constrained	$G + G\left[pq - q\left(q-1\right)/2\right] + 1$
UCUU	Unconstrained	Constrained	Unconstrained	Unconstrained	$G + \left[pq - q\left(q-1\right)/2\right] + Gp$
UCUC	Unconstrained	Constrained	Unconstrained	Constrained	$G + \left[pq - q\left(q-1\right)/2\right] + G$
UCCU	Unconstrained	Constrained	Constrained	Unconstrained	$G + \left[pq - q\left(q-1\right)/2\right] + p$
UCCC	Unconstrained	Constrained	Constrained	Constrained	$G + \left[pq - q\left(q-1\right)/2\right] + 1$
CUUU	Constrained	Unconstrained	Unconstrained	Unconstrained	$1 + G\left[pq - q\left(q-1\right)/2\right] + Gp$
CUUC	Constrained	Unconstrained	Unconstrained	Constrained	$1 + G\left[pq - q\left(q-1\right)/2\right] + G$
CUCU	Constrained	Unconstrained	Constrained	Unconstrained	$1 + G\left[pq - q\left(q-1\right)/2\right] + p$
CUCC	Constrained	Unconstrained	Constrained	Constrained	$1 + G\left[pq - q\left(q-1\right)/2\right] + 1$
CCUU	Constrained	Constrained	Unconstrained	Unconstrained	$1 + \left[pq - q\left(q-1\right)/2\right] + Gp$
CCUC	Constrained	Constrained	Unconstrained	Constrained	$1 + \left[pq - q\left(q-1\right)/2\right] + G$
CCCU	Constrained	Constrained	Constrained	Unconstrained	$1 + \left[pq - q\left(q-1\right)/2\right] + p$
CCCC	Constrained	Constrained	Constrained	Constrained	$1 + \left[pq - q\left(q-1\right)/2\right] + 1$

Section 4.3.3) show that knowing the label of just a small proportion of observations a priori can lead to improved clustering performance.

Notationally, if the ith observation is labeled, denote with $\widetilde{\boldsymbol{z}}_i = (\widetilde{z}_{i1}, \ldots, \widetilde{z}_{iG})$ its component membership indicator. Then, arranging the data so that the first m observations are labeled, the complete-data likelihood becomes

$$L_c(\boldsymbol{\theta}) = \prod_{i=1}^{m}\prod_{g=1}^{G}\left[\pi_g \phi\left(y_i|\boldsymbol{x}_i; m\left(\boldsymbol{x};\boldsymbol{\beta}_g\right), \sigma_g^2\right)\phi\left(\boldsymbol{x}_i|\boldsymbol{u}_{ig};\boldsymbol{\mu}_g,\boldsymbol{\Lambda}_g,\boldsymbol{\Psi}_g\right)\phi\left(\boldsymbol{u}_{ig}\right)\right]^{\widetilde{z}_{ig}}$$
$$\times \prod_{j=m+1}^{n}\prod_{h=1}^{H}\left[\pi_h \phi\left(y_j|\boldsymbol{x}_j; m\left(\boldsymbol{x};\boldsymbol{\beta}_h\right), \sigma_h^2\right)\phi\left(\boldsymbol{x}_j|\boldsymbol{u}_{jh};\boldsymbol{\mu}_h,\boldsymbol{\Lambda}_h,\boldsymbol{\Psi}_h\right)\phi\left(\boldsymbol{u}_{jh}\right)\right]^{z_{jh}},$$

where $H \geq G$ (often, it is assumed that $H = G$). For notational convenience, in this paper we prefer to present the AECM algorithm in the model-based clustering paradigm (cf. Sect. 4). However, the extension to the model-based classification context is simply obtained by substituting the 'dynamic' (with respect to the iterations of the algorithm) $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ with the 'static' $\widetilde{\boldsymbol{z}}_1, \ldots, \widetilde{\boldsymbol{z}}_m$.

3.4 On identifiability

No theoretical results are currently available on the identifiability of CWMs; however, because they can be seen as mixture models with random covariates, the results in

Hennig (2000, Section 3, Model 2.a) can apply. With regard to the latent factor structure, the necessary conditions for the identifiability of factor analyzers are discussed by Bartholomew and Knott (1999).

4 Maximum likelihood estimation

4.1 The AECM algorithm

The AECM algorithm (Meng and van Dyk 1997) is used for fitting all models within the CWFA family defined in Sect. 1. The expectation-conditional maximization (ECM) algorithm proposed by Meng and Rubin (1993) replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps. The AECM algorithm is an extension of the ECM algorithm, where the specification of the complete data is allowed to be different on each CM step.

Let $\mathcal{S} = \{(\boldsymbol{x}_i, y_i)\,; i = 1, \ldots, n\}$ be a sample of size n from (6). In the EM framework, the generic observation $(\boldsymbol{x}_i, y_i)$ is viewed as being incomplete; its complete counterpart is given by $\left(\boldsymbol{x}_i, y_i, \boldsymbol{u}_{ig}, \boldsymbol{z}_i\right)$, where $\boldsymbol{z}_i$ is the component-label vector in which $z_{ig} = 1$ if $(\boldsymbol{x}_i, y_i)$ comes from Ω_g and $z_{ig} = 0$ otherwise. Then the complete-data likelihood, by considering the result in (4), can be written as

$$L_c\left(\boldsymbol{\theta}\right) = \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\pi_g \phi\left(y_i|\boldsymbol{x}_i; m\left(\boldsymbol{x}; \boldsymbol{\beta}_g\right), \sigma_g^2\right)\phi\left(\boldsymbol{x}_i|\boldsymbol{u}_i; \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g\right)\phi\left(\boldsymbol{u}_{ig}\right)\right]^{z_{ig}}.$$

The idea of the AECM algorithm is to partition $\boldsymbol{\theta}$, say $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right)'$, in such a way that the likelihood is easy to maximize for $\boldsymbol{\theta}_1$ given $\boldsymbol{\theta}_2$ and *vice versa*. For the application of the AECM algorithm to our CWFA family, one iteration consists of two cycles, with one E-step and one CM-step for each cycle. The two CM-steps correspond to the partition of $\boldsymbol{\theta}$ into the two subvectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Then, we can iterate between these two conditional maximizations until convergence. In the next two sections, we illustrate the two cycles for the UUUU model only. Details on the other models of the CWFA family are given in Appendix B.

4.2 First cycle

Here, $\boldsymbol{\theta}_1 = \left\{\pi_g, \boldsymbol{\beta}_g, \boldsymbol{\mu}_g, \sigma_g^2; g = 1, \ldots, G\right\}$, where the missing data are the unobserved group labels $\boldsymbol{z}_i, i = 1, \ldots, n$. The complete-data likelihood is

$$L_1\left(\boldsymbol{\theta}_1\right) = \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\pi_g \phi\left(y_i|\boldsymbol{x}_i; m\left(\boldsymbol{x}_i; \boldsymbol{\beta}_g\right), \sigma_g^2\right)\phi\left(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)\right]^{z_{ig}}.$$

Consider the complete-data log-likelihood

$$
\begin{aligned}
&l_{\mathrm{c1}}\left(\boldsymbol{\theta}_1\right) \\
&\quad=\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\left[\pi_g\phi\left(y_i|\boldsymbol{x}_i;\, m\left(\boldsymbol{x}_i;\boldsymbol{\beta}_g\right),\sigma_g^2\right)\phi\left(\boldsymbol{x}_i;\boldsymbol{\mu}_g,\boldsymbol{\Lambda}_g,\boldsymbol{\Psi}_g\right)\right] \\
&\quad=-\frac{n\left(p+1\right)}{2}\ln 2\pi-\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\sigma_g^2-\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\frac{\left(y_i-\beta_{0g}-\boldsymbol{\beta}_{1g}'\boldsymbol{x}_i\right)^2}{\sigma_g^2} \\
&\qquad-\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\left|\boldsymbol{\Sigma}_g\right|-\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_g\right)'\boldsymbol{\Sigma}_g^{-1}\left(\boldsymbol{x}_i-\boldsymbol{\mu}_g\right) \\
&\qquad+\sum_{g=1}^{G} n_g\ln\pi_g,
\end{aligned}
$$

where $n_g=\sum_{i=1}^{n} z_{ig}$. Because $\boldsymbol{\Sigma}_g=\boldsymbol{\Lambda}\boldsymbol{\Lambda}_g'+\boldsymbol{\Psi}_g$, we get

$$
\begin{aligned}
l_{\mathrm{c1}}\left(\boldsymbol{\theta}_1\right)=&-\frac{n\left(p+1\right)}{2}\ln 2\pi-\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\sigma_g^2 \\
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\frac{\left(y_i-\beta_{0g}-\boldsymbol{\beta}_{1g}'\boldsymbol{x}_i\right)^2}{\sigma_g^2}-\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\left|\boldsymbol{\Lambda}\boldsymbol{\Lambda}_g'+\boldsymbol{\Psi}_g\right| \\
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\mathrm{tr}\left\{\left(\boldsymbol{x}_i-\boldsymbol{\mu}_g\right)\left(\boldsymbol{x}_i-\boldsymbol{\mu}_g\right)'\left(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g'+\boldsymbol{\Psi}_g\right)^{-1}\right\} \\
&+\sum_{g=1}^{G} n_g\ln\pi_g.
\end{aligned}
$$

The E-step on the first cycle of the $(k+1)$st iteration requires the calculation of $Q_1\left(\boldsymbol{\theta}_1;\boldsymbol{\theta}^{(k)}\right)=\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[l_c\left(\boldsymbol{\theta}_1\right)|\mathcal{S}\right]$, which is the expected complete-data log-likelihood given the observed data and using the estimate $\boldsymbol{\theta}^{(k)}$ from the kth iteration. In practice, it requires calculating $\mathbb{E}_{\boldsymbol{\theta}^{(k)}}\left[Z_{ig}|\mathcal{S}\right]$; this step is achieved by replacing each z_{ig} by $z_{ig}^{(k+1)}$, where

$$
z_{ig}^{(k+1)}=\frac{\pi_j^{(k)}\phi\left(y_i|\boldsymbol{x}_i;\, m\left(\boldsymbol{x}_i;\boldsymbol{\beta}_g^{(k)}\right),\sigma_g^{2(k)}\right)\phi\left(\boldsymbol{x}_i|\boldsymbol{\mu}_g^{(k)},\boldsymbol{\Lambda}_g^{(k)},\boldsymbol{\Psi}_g^{(k)}\right)}{\sum_{j=1}^{G}\pi_j^{(k)}\phi\left(y_i|\boldsymbol{x}_i;\, m\left(\boldsymbol{x}_i;\boldsymbol{\beta}_j^{(k)}\right),\sigma_j^{2(k)}\right)\phi\left(\boldsymbol{x}_i|\boldsymbol{\mu}_j^{(k)},\boldsymbol{\Lambda}_j^{(k)},\boldsymbol{\Psi}_j^{(k)}\right)}.
$$

For the M-step, the maximization of this complete-data log-likelihood yields

$$\pi_g^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ig}^{(k+1)}$$

$$\boldsymbol{\mu}_g^{(k+1)} = \frac{1}{n_g}\sum_{i=1}^{n} z_{ig}^{(k+1)}\boldsymbol{x}_i$$

$$\boldsymbol{\beta}_{1g}^{(k+1)} = \left[\frac{1}{n_g}\sum_{i=1}^{n} z_{ig}^{(k+1)} y_i\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)}\right)\right]\left[\frac{1}{n_g}\sum_{i=1}^{n} z_{ig}^{(k+1)}\boldsymbol{x}_i'\boldsymbol{x}_i - \boldsymbol{\mu}_g^{'(k+1)}\boldsymbol{\mu}_g^{(k+1)}\right]^{-1}$$

$$\beta_{0g}^{(k+1)} = \frac{1}{n_g}\sum_{i=1}^{n} z_{ig}^{(k+1)} y_i - \boldsymbol{\beta}_{1g}^{'(k+1)}\boldsymbol{\mu}_g^{(k+1)}$$

$$\sigma_g^{2(k+1)} = \frac{1}{n_g}\sum_{i=1}^{n} z_{ig}^{(k+1)}\left\{y_i - \left(\beta_{0g}^{(k+1)} + \boldsymbol{\beta}_{1g}^{'(k+1)}\boldsymbol{x}_i\right)\right\}^2,$$

where $n_g^{(k+1)} = \sum_{i=1}^{n} z_{ig}^{(k+1)}$. Following the notation in McLachlan and Peel (2000a), we set $\boldsymbol{\theta}^{(k+1/2)} = \left\{\boldsymbol{\theta}_1^{(k+1)}, \boldsymbol{\theta}_2^{(k)}\right\}$.

4.3 Second cycle

Here, $\boldsymbol{\theta}_2 = \left\{\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g; g = 1, \ldots, G\right\}$, where the missing data are the unobserved group labels $\boldsymbol{z}_i$ and the latent factors $\boldsymbol{u}_{ig}$, $i = 1, \ldots, n$, and $g = 1, \ldots, G$. Therefore, the complete-data likelihood is

$$\begin{aligned} L_{c2}(\boldsymbol{\theta}_2) &= \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\phi\left(y_i|\boldsymbol{x}_i, \boldsymbol{u}_{ig}; m\left(\boldsymbol{x}_i; \boldsymbol{\beta}_g^{(k+1)}\right), \sigma_g^{2(k+1)}\right)\right. \\ &\quad \left.\times\,\phi\left(\boldsymbol{x}_i|\boldsymbol{u}_{ig}; \boldsymbol{\mu}_g^{(k+1)}, \boldsymbol{\Sigma}_g\right)\phi(\boldsymbol{u}_{ig})\pi_g^{(k+1)}\right]^{z_{ig}} \\ &= \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\phi\left(y_i|\boldsymbol{x}_i; m\left(\boldsymbol{x}_i; \boldsymbol{\beta}_g^{(k+1)}\right), \sigma_g^{2(k+1)}\right)\right. \\ &\quad \left.\times\,\phi\left(\boldsymbol{x}_i|\boldsymbol{u}_{ig}; \boldsymbol{\mu}_g^{(k+1)}, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g\right)\phi(\boldsymbol{u}_{ig})\pi_g^{(k+1)}\right]^{z_{ig}}, \end{aligned}$$

because Y is conditionally independent of $\boldsymbol{U}$ given $\boldsymbol{X} = \boldsymbol{x}$ and

$$\phi\left(\boldsymbol{x}_i|\boldsymbol{u}_{ig}; \boldsymbol{\mu}_g^{(k+1)}, \boldsymbol{\Psi}_g\right) = \frac{1}{\left|2\pi\boldsymbol{\Psi}_g\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)} - \boldsymbol{\Lambda}_g\boldsymbol{u}_{ig}\right)'\right.$$

$$\times \boldsymbol{\Psi}_g^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)} - \boldsymbol{\Lambda}_g \boldsymbol{u}_{ig}\right)\Big\}$$

$$\phi\left(\boldsymbol{u}_{ig}\right) = \frac{1}{(2\pi)^{q/2}} \exp\left\{-\frac{1}{2}\boldsymbol{u}_{ig}'\boldsymbol{u}_{ig}\right\}.$$

Hence, the complete-data log-likelihood is

$$\begin{aligned} l_{c2}\left(\boldsymbol{\theta}_2\right) = & -\frac{n\left(p+q+1\right)}{2}\ln\left(2\pi\right) - \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\sigma_g^{2(k+1)} + \\ & -\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\frac{\left(y_i - \beta_{0g}^{(k+1)} - \boldsymbol{\beta}_{1g}^{'(k+1)}\boldsymbol{x}_i\right)^2}{2\hat{\sigma}_g^2} \\ & + \sum_{g=1}^{G} n_g \ln\pi_g + \frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\left|\boldsymbol{\Psi}_g^{-1}\right| + \\ & -\frac{1}{2}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\mathrm{tr}\left\{\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)} - \boldsymbol{\Lambda}_g\boldsymbol{u}_{ig}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)} - \boldsymbol{\Lambda}_g\boldsymbol{u}_{ig}\right)'\boldsymbol{\Psi}_g^{-1}\right\}, \end{aligned}$$

where we set

$$\boldsymbol{S}_g^{(k+1)} = \frac{1}{n_g^{(k+1)}}\sum_{i=1}^{n} z_{ig}^{(k+1)}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)}\right)'.$$

The E-step on the second cycle of the $(k+1)$st iteration requires the calculation of $Q_2\left(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(k+1/2)}\right) = \mathbb{E}_{\boldsymbol{\theta}^{(k+1/2)}}\left[l_{c2}\left(\boldsymbol{\theta}_2\right)|\mathcal{S}\right]$. Therefore, we must calculate the following conditional expectations: $\mathbb{E}_{\boldsymbol{\theta}^{(k+1/2)}}\left(Z_{ig}|\mathcal{S}\right)$, $\mathbb{E}_{\boldsymbol{\theta}^{(k+1/2)}}\left(Z_{ig}\boldsymbol{U}_{ig}|\mathcal{S}\right)$, and $\mathbb{E}_{\boldsymbol{\theta}^{(k+1/2)}}\left(Z_{ig}\boldsymbol{U}_{ig}\boldsymbol{U}_{ig}'|\mathcal{S}\right)$. Based on (2), these are given by

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}^{(k+1/2)}}\left(Z_{ig}\boldsymbol{U}_{ig}|\mathcal{S}\right) &= z_{ig}^{(k+1)}\boldsymbol{\gamma}_g^{(k)}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_g^{(k+1)}\right) \\ \mathbb{E}_{\boldsymbol{\theta}^{(k+1/2)}}\left(Z_{ig}\boldsymbol{U}_{ig}\boldsymbol{U}_{ig}'|\mathcal{S}\right) &= z_{ig}^{(k+1)}\left\{\boldsymbol{I}_q - \boldsymbol{\gamma}_g^{(k)}\boldsymbol{\Lambda}_g^{(k)} + \boldsymbol{\gamma}_g^{(k)}\boldsymbol{S}_g\boldsymbol{\gamma}_g^{'(k)}\right\} = z_{ig}^{(k+1)}\boldsymbol{\Theta}_g^{(k)}, \end{aligned}$$

where

$$\boldsymbol{\gamma}_g^{(k)} = \boldsymbol{\Lambda}_g^{'(k)}\left(\boldsymbol{\Lambda}_g^{(k)}\boldsymbol{\Lambda}_g^{'(k)} + \boldsymbol{\Psi}_g^{(k)}\right)^{-1} \tag{7}$$

$$\boldsymbol{\Theta}_g^{(k)} = \boldsymbol{I}_q - \boldsymbol{\gamma}_g^{(k)}\boldsymbol{\Lambda}_g^{(k)} + \boldsymbol{\gamma}_g^{(k)}\boldsymbol{S}_g^{(k+1)}\boldsymbol{\gamma}_g^{'(k)}. \tag{8}$$

Thus, the gth term of the expected complete-data log-likelihood $Q_2\left(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(k+1/2)}\right)$ becomes

$$\begin{aligned} & Q_2\left(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g; \boldsymbol{\theta}^{(k+1/2)}\right) \\ & = \mathrm{C}(\boldsymbol{\theta}_1^{(k+1)}) + \frac{1}{2} n_g^{(k+1)} \ln |\boldsymbol{\Psi}_g^{-1}| - \frac{1}{2} n_g^{(k+1)} \mathrm{tr}\left\{\boldsymbol{S}_g^{(k+1)} \boldsymbol{\Psi}_g^{-1}\right\} \\ & \quad + n_g^{(k+1)} \mathrm{tr}\left\{\boldsymbol{\Lambda}_g \boldsymbol{\gamma}_g^{(k)} \boldsymbol{S}_g^{(k+1)} \boldsymbol{\Psi}_g^{-1}\right\} - \frac{1}{2} n_g^{(k+1)} \mathrm{tr}\left\{\boldsymbol{\Lambda}_g' \boldsymbol{\Psi}_g^{-1} \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)}\right\}, \end{aligned} \tag{9}$$

where $\mathrm{C}\left(\boldsymbol{\theta}_1^{(k+1)}\right)$ denotes the terms in (9) that do not depend on $\boldsymbol{\theta}_2$. Then (9) is maximized for $\left\{\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Psi}}\right\}$, satisfying

$$\frac{\partial Q_2}{\partial \boldsymbol{\Lambda}_g} = n_g^{(k+1)} \boldsymbol{\Psi}_g^{-1} \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g^{'(k)} - n_g^{(k+1)} \boldsymbol{\Psi}_g^{-1} \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)} = \mathbf{0}$$

$$\frac{\partial Q_2}{\partial \boldsymbol{\Psi}_g^{-1}} = \frac{1}{2} n_g^{(k+1)} \boldsymbol{\Psi}_g - \frac{1}{2} n_g^{(k+1)} \boldsymbol{S}_g^{(k+1)} + n_g^{(k+1)} \boldsymbol{S}_g^{'(k+1)} \boldsymbol{\gamma}_g^{'(k)} \boldsymbol{\Lambda}_g' - \frac{1}{2} n_g^{(k+1)} \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)} \boldsymbol{\Lambda}_g' = \mathbf{0}.$$

Therefore,

$$\boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g^{'(k)} - \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)} = \mathbf{0} \tag{10}$$

$$\boldsymbol{\Psi}_g - \boldsymbol{S}_g^{(k+1)} + 2\boldsymbol{S}_g^{'(k+1)} \boldsymbol{\gamma}_g^{'(k)} \boldsymbol{\Lambda}_g' - \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)} \boldsymbol{\Lambda}_g' = \mathbf{0}. \tag{11}$$

From (10), we get

$$\hat{\boldsymbol{\Lambda}}_g = \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g'^{(k)} \boldsymbol{\Theta}_g^{-1}, \tag{12}$$

and substituting in (11) we get

$$\boldsymbol{\Psi}_g - \boldsymbol{S}_g^{(k+1)} + 2\boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g'^{(k)} \left(\boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g'^{(k)} \boldsymbol{\Theta}_g^{-1}\right)' - \left(\mathbf{S}_g \hat{\boldsymbol{\gamma}}_g' \boldsymbol{\Theta}_g^{-1}\right) \boldsymbol{\Theta}_g \left(\mathbf{S}_g \hat{\boldsymbol{\gamma}}_g' \boldsymbol{\Theta}_g^{-1}\right)' = \mathbf{0},$$

which yields

$$\hat{\boldsymbol{\Psi}}_g = \mathrm{diag}\left\{\boldsymbol{S}_g^{(k+1)} - \hat{\boldsymbol{\Lambda}}_g \hat{\boldsymbol{\gamma}}_g \boldsymbol{S}_g^{(k+1)}\right\}. \tag{13}$$

Hence, the maximum likelihood estimates for $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are obtained by iteratively computing

$$\begin{aligned} \boldsymbol{\Lambda}_g^+ &= \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g' \boldsymbol{\Theta}_g^{-1} \\ \boldsymbol{\Psi}_g^+ &= \mathrm{diag}\left\{\boldsymbol{S}_g^{(k+1)} - \boldsymbol{\Lambda}_g^+ \boldsymbol{\gamma}_g \boldsymbol{S}_g^{(k+1)}\right\}, \end{aligned}$$

where the superscript $^+$ denotes the update estimate. Using (7) and (8), we get

$$\begin{aligned}\boldsymbol{\gamma}_g^+ &= \boldsymbol{\Lambda}_g'^+ \left(\boldsymbol{\Lambda}_g^+ \boldsymbol{\Lambda}_g'^+ + \boldsymbol{\Psi}_g^+\right)^{-1} \\ \boldsymbol{\Theta}_g^+ &= \boldsymbol{I}_q - \boldsymbol{\gamma}_g^+ \boldsymbol{\Lambda}_g^+ + \boldsymbol{\gamma}_g^+ \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g'^+. \end{aligned} \tag{14}$$

4.4 Outline of the algorithm

In summary, the procedure can be described as follows. For a given initial guess $\boldsymbol{\theta}^{(0)}$, on the $(k+1)$st iteration, the algorithm carries out the following steps for $g = 1, \ldots, G$:

1. Compute $\pi_g^{(k+1)}$, $\boldsymbol{\mu}_g^{(k+1)}$, $\boldsymbol{\beta}_g^{(k+1)}$, $\sigma_g^{2(k+1)}$;
2. Set $\boldsymbol{\Lambda}_g \leftarrow \boldsymbol{\Lambda}_g^{(k)}$ and $\boldsymbol{\Psi} \leftarrow \boldsymbol{\Psi}_g^{(k)}$, and compute $\boldsymbol{\gamma}_g$ and $\boldsymbol{\Theta}_g$;
3. Repeat the following steps until convergence on $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$:
   (a) Set $\boldsymbol{\Lambda}_g^+ \leftarrow \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g' \boldsymbol{\Theta}_g^{-1}$ and $\boldsymbol{\Psi}_g^+ \leftarrow \text{diag}\left\{\boldsymbol{S}_g^{(k+1)} - \boldsymbol{\Lambda}_g^+ \boldsymbol{\gamma}_g \boldsymbol{S}_g^{(k+1)}\right\}$;
   (b) Set $\boldsymbol{\gamma}_g^+ \leftarrow \boldsymbol{\Lambda}_g'^+ \left(\boldsymbol{\Lambda}_g^+ \boldsymbol{\Lambda}_g'^+ + \boldsymbol{\Psi}_g^+\right)^{-1}$ and $\boldsymbol{\Theta}_g^+ \leftarrow \boldsymbol{I}_q - \boldsymbol{\gamma}_g^+ \boldsymbol{\Lambda}_g^+ + \boldsymbol{\gamma}_g^+ \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g'^+$;
   (c) Set $\boldsymbol{\Lambda}_g \leftarrow \boldsymbol{\Lambda}_g^+$, $\boldsymbol{\Psi}_g \leftarrow \boldsymbol{\Psi}_g^+$, $\boldsymbol{\gamma}_g \leftarrow \boldsymbol{\gamma}_g^+$, and $\boldsymbol{\Theta}_g \leftarrow \boldsymbol{\Theta}_g^+$;
4. Set $\boldsymbol{\Lambda}_g^{(k+1)} \leftarrow \boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g^{(k+1)} \leftarrow \boldsymbol{\Psi}_g$.

4.5 AECM initialization: a 5-step procedure

The choice of starting values is a well known and important issue in EM-based algorithms. The standard approach consists of selecting a value for $\boldsymbol{\theta}^{(0)}$. An alternative method, more natural in the authors' opinion, consists of choosing a value for $\boldsymbol{z}_i^{(0)}$, $i = 1, \ldots, n$ (see McLachlan and Peel 2000a, p. 54). Within this approach, and due to the hierarchical structure of the CWFA family of parsimonious models, we propose a 5-step hierarchical initialization procedure.

For a fixed number of groups G, let $\boldsymbol{z}_i^{(0)}$, $i = 1, \ldots, n$, be the initial classification for the AECM algorithm, so that $z_{ig}^{(0)} \in \{0, 1\}$ and $\sum_g z_{ig}^{(0)} = 1$. The set $\left\{\boldsymbol{z}_i^{(0)}; i = 1, \ldots, n\right\}$ can be obtained either through some clustering procedure (here we consider the k-means method) or by random initialization, for example by sampling from a multinomial distribution with probabilities $(1/G, \ldots, 1/G)$. Then, at the first step of the procedure, the most constrained CCCC model is estimated from these starting values. At the second step, the resulting (AECM-estimated) $\hat{z}_{ig}$ are taken as the starting group membership labels to initialize the AECM-algorithm of the four models {UCCC, CUCC, CCUC, CCCU} obtained by relaxing one of the four constraints. At the third step, the AECM-algorithm for each of the six models {CCUU, CUCU, UCCU, CUUC, UCUC, UUCC} with two constraints is initialized using the $\hat{z}_{ig}$ from the previous step and the model with the highest likelihood. For example, to initialize CCUU we use the $\hat{z}_{ig}$ from the model having the highest likelihood between CCCU and CCUC. In this fashion, the initialization procedure continues according to the scheme displayed in Fig. 1, until the least constrained model UUUU is estimated at the fifth step.

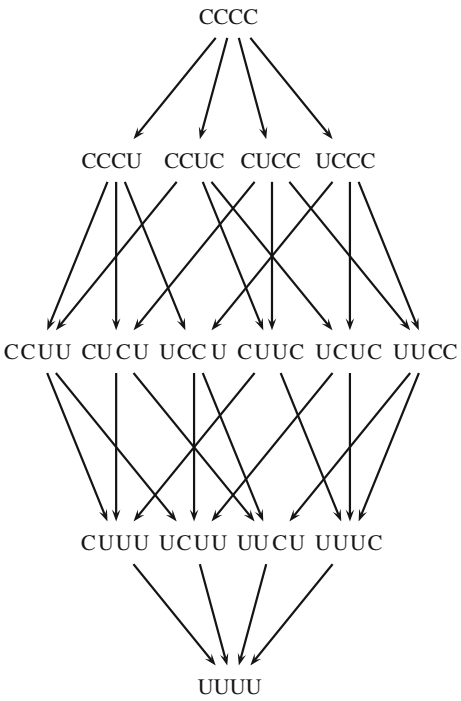

Fig. 1 Relationships among the models in the 5-step hierarchical initialization procedure. *Arrows* are oriented from the model used to initialize towards the model to be estimated

For all of the models in the CWFA family, in analogy with McNicholas and Murphy (2008), the initial values for the elements of $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\Psi}_g$ are generated from the eigen-decomposition of $\boldsymbol{S}_g$ as follows. The $\boldsymbol{S}_g$ are computed based on the values of $z_{ig}^{(0)}$. The eigen-decomposition of each $\boldsymbol{S}_g$ is obtained using the Householder reduction and the QL method (details given by Press et al. 1992). Then the initial values of the elements of $\boldsymbol{\Lambda}_g$ are set as $\lambda_{ij} = \sqrt{d_j}\rho_{ij}$, where d_j is the jth largest eigenvalue of $\boldsymbol{S}_g$ and ρ_{ij} is the ith element of the eigenvector corresponding to the jth largest eigenvalue of $\boldsymbol{S}_g$, where $i \in \{1, 2, \ldots, d\}$ and $j \in \{1, 2, \ldots, q\}$. The $\boldsymbol{\Psi}_g$ are then initialized as $\boldsymbol{\Psi}_g = \text{diag}\left(\boldsymbol{S}_g - \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g'\right)$.

4.6 Convergence criterion

The Aitken acceleration procedure (Aitken 1926) is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the AECM algorithm. Based on this estimate, a decision is made about whether the algorithm has reached convergence, i.e., whether 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 $k+1$ is

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

Böhning et al. (1994). In the analyses in Section 6, we stop our algorithms when $l_{\infty}^{(k+1)} - l^{(k)} < \epsilon$ (Böhning et al. 1994; Lindsay 1995). Note that we use $\epsilon = 0.05$ for the analyses herein.

5 Model selection and performance assessment

5.1 Model selection

The CWFA model, in addition to $\boldsymbol{\theta}$, is also characterized by the number of latent factors q and by the number of mixture components G. So far, these quantities have been treated as a priori fixed. Nevertheless, the estimation of these is required, for practical purposes, when choosing a relevant model.

For model-based clustering and classification, several model selection criteria are used, such as the Bayesian information criterion (BIC; Schwarz 1978), the integrated completed likelihood (ICL; Biernacki et al. 2000), and the Akaike information criterion (AIC; Sakamoto et al. 1983). Among these, the BIC is the most predominant in the literature and is given by

$$\text{BIC} = 2l(\hat{\boldsymbol{\theta}}) - \eta \ln n,$$

where $l(\hat{\boldsymbol{\theta}})$ is the maximized log-likelihood and η is the number of free parameters. This is the model selection criterion used in the analyses of Sect. 6.

5.2 Adjusted rand index

Although the data analyses of Sect. 6 are mainly conducted as clustering examples, the true classifications are actually known for these data. In these examples, the adjusted Rand index (ARI; Hubert and Arabie 1985) is used to measure class agreement. The Rand index (RI; Rand 1971) is based on pairwise comparisons and is obtained by dividing the number of pairwise 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 ARI corrects the RI to account for agreement by chance: a value of '1' indicates perfect agreement, '0' is expected under random classification, and negative values indicate a classification that is worse than would be expected by guessing.

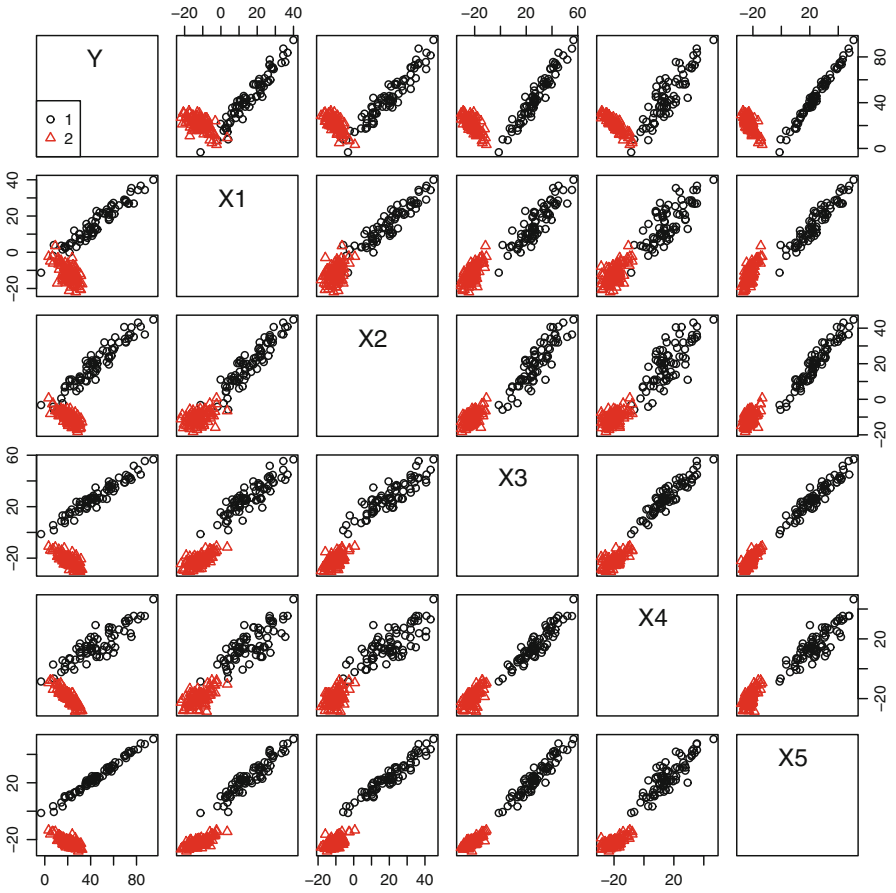

Fig. 2 Scatterplot matrix of the simulated data for Example 1

6 Data analyses

This section presents the application of the CWFA family of models to both artificial and real data sets. Code for the AECM algorithm, described in this paper, was written in the R computing environment (R Development Core Team 2012).

6.1 Simulated data

6.1.1 Example 1

The first data set consists of a sample of size $n = 175$ drawn from model UUCU with $G = 2$, $n_1 = 75$, $n_2 = 100$, $p = 5$, and $q = 2$ (see Fig. 2 for details). The parameters used for the simulation of the data are given in Table 2 (see Appendix C.1 for details on the covariance matrices $\boldsymbol{\Sigma}_g$, $g = 1, \ldots, G$).

Table 2 True and estimated parameters for the simulated data of Example 1

(a) Means of $\boldsymbol{X}$

	$\boldsymbol{\mu}_g$					$\hat{\boldsymbol{\mu}}_g$				
g	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
1	14.00	18.00	25.00	14.00	22.00	15.88	19.94	27.48	15.81	23.93
2	-12.00	-10.00	-22.00	-20.00	-22.00	-11.95	-10.36	-22.00	-19.67	-22.03

(b) Slopes

	$\boldsymbol{\beta}_{1g}$					$\hat{\boldsymbol{\beta}}_{1g}$				
g	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
1	0.47	0.02	0.42	0.03	0.87	0.50	0.03	0.46	0.02	0.81
2	-0.02	-0.63	-0.05	-0.85	-0.03	-0.04	-0.57	-0.01	-0.85	-0.18

(c) Conditional std. deviations

g	σ_g	$\hat{\sigma}_g$
1	2.00	1.24
2	4.00	3.79

(d) Intercepts

g	β_{0g}	$\hat{\beta}_{0g}$
1	4.50	4.34
2	-4.20	-6.35

All 16 CWFA models were fitted to the data for $G \in \{1, 2, 3\}$ and $q \in \{1, 2\}$, resulting in 96 different models. As noted above (Sect. 4.5), initialization of the $\boldsymbol{z}_i$, $i = 1, \ldots, n$, for the most constrained model (CCCC), and for each combination (G, q), was done using the k-means algorithm according to the `kmeans` function of the R package `stats`. The remaining 15 models, for each combination (G, q), were initialized using the 5-step hierarchical initialization procedure described in Sect. 4.5. The BIC values for all models were computed and the model with the largest BIC value was selected as the best. In this example, the model corresponding to the largest BIC value (−5,845.997) was a $G = 2$ component UUCU model with $q = 2$ latent factors, the same as the model used to generate the data. The selected model gave perfect classification and the estimated parameters were very close to the parameters used for data simulation (see Table 2 and Appendix C.1).

Figure 3 shows the BIC values of the top 10 models sorted in increasing order. The horizontal dotted line separates the models with a BIC value within 1 % of the maximum (over all 96 models) BIC value (hereafter simply referred to as the '1% line'). As mentioned earlier, the model with the largest BIC was UUCU (with $G = 2$ and $q = 2$). The subsequent two models, those above the 1 % line, were UUUU with $G = 2$ and $q = 2$ (BIC equal to −5,867.006) and CUCU with $G = 2$ and $q = 2$ (BIC equal to −5,869.839). These two models are structurally very close to the true UUCU model and also yielded perfect classification. It should also be noted that most of the models with high BIC values have $G = 2$ and $q = 2$.

6.1.2 Example 2

For the second data set, a sample of size $n = 235$ was drawn from the CUUC model with $G = 3$ groups (with $n_1 = 75$, $n_2 = 100$, and $n_3 = 60$) and $q = 2$ latent factors (see Fig. 4).

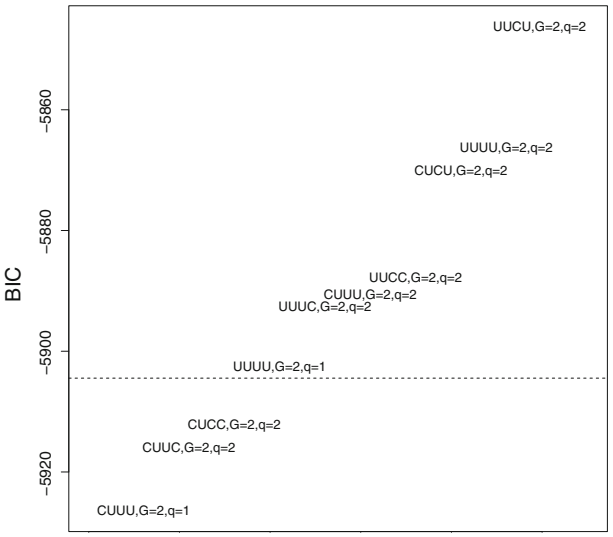

Fig. 3 BIC values of the top 10 models, sorted in increasing order, for Example 1

All 16 CWFA models were fitted to the data for $G \in \{1, 2, 3, 4\}$ and $q \in \{1, 2\}$, resulting in 128 different models. The algorithm was initialized in the same way as for Example 2. The model with the highest BIC (−6,579.116) was CUUC with $G = 3$ and $q = 2$, resulting in perfect classification. The estimated parameters of this model were very close to the true ones (Table 3 and Appendix C.2).

The plot of BIC values is omitted for the sake of space. Besides the true model, we underline that the other three models above the 1% line are UUUC (BIC = −6,583.692), CUUU (BIC = −6,637.222), and UUUU (BIC = −6,641.798), all with $G = 3$ and $q = 2$. Thus, these models are congruent, with respect to the true one, in terms of G and q. Moreover, they have a similar covariance structure to the true one (CUUC) and yielded perfect classification.

6.1.3 Example 3

A third simulated data set, of dimension $p + 1 = 11$, was generated from the CCUU model with $G = 2$ groups (with $n_1 = 75$ and $n_2 = 100$) and $q = 4$ latent factors. All 16 CWFA models were fitted to the data for $G \in \{1, 2, 3\}$ and $q \in \{1, 2, 3, 4, 5\}$, resulting in 240 different models. The algorithm was initialized in the same way as for Example 1. The model with the highest BIC (−10,190.23) was CCUU with $G = 2$ and $q = 4$, resulting in perfect classification. The estimated parameters of this model were very close to the true ones (Table 4).

As before, the plot of BIC values is omitted for the sake of space. The next three models with the highest BIC were UCUU (BIC $= -10, 192.989, q = 4$), CCUU (BIC $= -10, 195.264, q = 5$), and UCUU (BIC $= -10, 198.027, q = 5$), all with

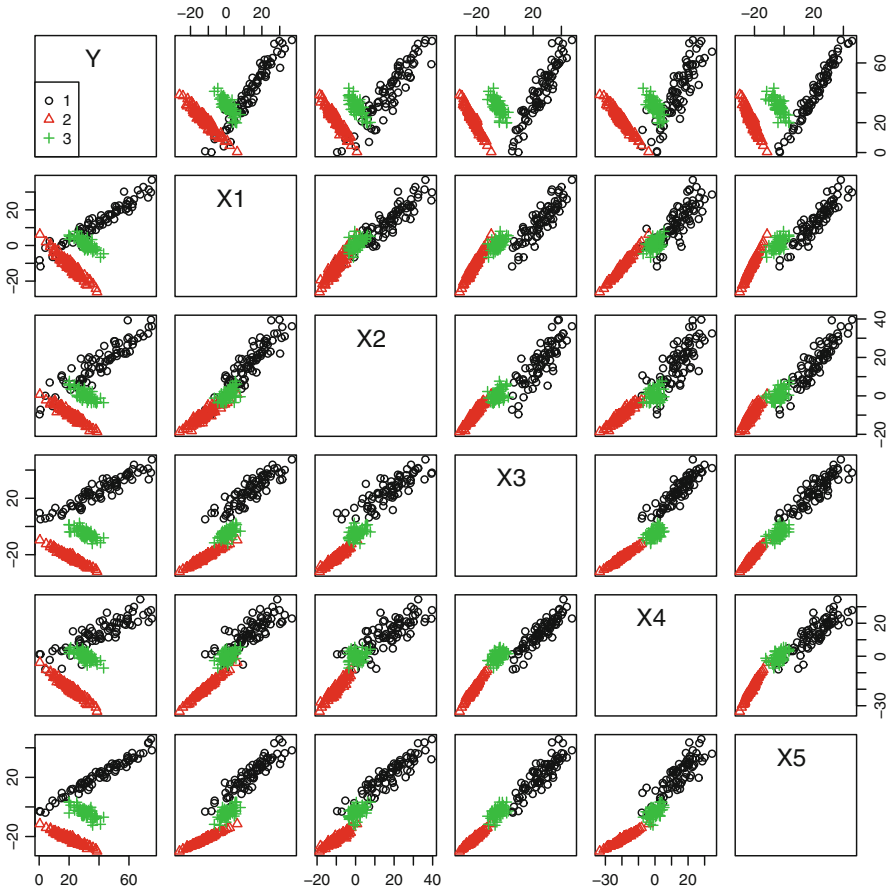

Fig. 4 Scatterplot matrix of the simulated data for Example 2

$G = 2$. All of these models had two components and a constrained loading matrix, and yielded perfect classification.

6.1.4 Example 4

A fourth simulated data set, of dimension $p + 1 = 21$, was generated from the UUCC model with $G = 2$ groups (with $n_1 = 120$ and $n_2 = 100$) and $q = 5$ latent factors.

All 16 CWFA models were fitted to the data for $G \in \{1, 2, 3\}$ and $q \in \{1, 2, 3, 4, 5, 6\}$, resulting in 288 different models. The algorithm was initialized in the same way as for Example 1. The model with the highest BIC (−24,199.57) was UUCC with $G = 2$ and $q = 5$, resulting in perfect classification. The estimated parameters of this model were very close to the true ones (Table 5).

The plot of BIC values is omitted for the sake of space. The next three models with the highest BIC were UUUC (BIC = −24204.955), CUCC (BIC = −24,213.742), and CUUC (BIC = 24,219.126), all with $G = 2$ and $q = 5$.

Table 3 True and estimated parameters for the simulated data of Example 2

(a) Means of $\boldsymbol{X}$

	$\boldsymbol{\mu}_g$					$\hat{\boldsymbol{\mu}}_g$				
g	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
1	0.00	0.00	-5.00	0.00	-4.00	0.82	0.48	-5.09	-0.21	-3.75
2	14.00	18.00	25.00	14.00	22.00	13.64	17.44	25.44	14.25	21.44
3	-12.00	-10.00	-22.00	-20.00	-22.00	-12.33	-10.22	-22.25	-20.24	-22.21

(b) Slopes

	$\boldsymbol{\beta}_{1g}$					$\hat{\boldsymbol{\beta}}_{1g}$				
g	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
1	-0.41	-0.87	-0.22	-0.62	-0.06	-0.34	-0.82	-0.32	-0.66	-0.09
2	0.47	0.02	0.42	0.03	0.87	0.51	0.00	0.38	0.05	0.84
3	-0.02	-0.63	-0.05	-0.85	-0.03	-0.04	-0.68	-0.36	-0.44	-0.18

(c) Conditional std. deviations

g	σ_g	$\hat{\sigma}_g$
1	2.00	2.30
2	2.00	2.30
3	2.00	2.30

(d) Intercepts

g	β_{0g}	$\hat{\beta}_{0g}$
1	30.00	29.39
2	4.50	5.31
3	-4.20	-6.69

Table 4 True and estimated parameters for the simulated data of Example 3

(a) Means of $\boldsymbol{X}$

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$\boldsymbol{\mu}_1$	-12.00	-10.00	-22.00	-20.00	-22.00	-10.00	-12.00	-17.00	-19.00	-10.00
$\hat{\boldsymbol{\mu}}_1$	-11.31	-9.23	-21.46	-19.35	-21.33	-9.29	-11.32	-16.14	-18.12	-9.31
$\boldsymbol{\mu}_2$	14.00	18.00	25.00	14.00	22.00	13.00	14.00	12.00	20.00	10.00
$\hat{\boldsymbol{\mu}}_2$	15.21	19.44	26.82	15.18	22.47	14.59	16.88	14.06	20.32	11.61

(b) Slopes

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$\boldsymbol{\beta}_1$	-0.12	-0.86	-0.59	-0.33	-0.64	-0.19	-0.36	-0.39	-0.22	-0.65
$\hat{\boldsymbol{\beta}}_1$	-0.14	-0.73	-0.48	-0.44	-0.66	-0.02	-0.49	-0.47	-0.29	-0.64
$\boldsymbol{\beta}_2$	0.86	0.04	0.60	0.42	0.35	0.89	0.82	0.17	0.59	0.31
$\hat{\boldsymbol{\beta}}_2$	0.86	0.06	0.52	0.46	0.33	0.86	0.88	0.21	0.63	0.24

(c) Conditional std. deviations

g	σ_g	$\hat{\sigma}_g$
1	2.00	1.53
2	2.00	1.53

(d) Intercepts

g	β_{0g}	$\hat{\beta}_{0g}$
1	4.50	5.23
2	-4.20	-5.42

Table 5 True and estimated parameters for the simulated data of Example 4

(a) Means of $\boldsymbol{X}$

	X									
$\boldsymbol{\mu}_1$	-12.00	-12.00	-12.00	-12.00	-12.00	-10.00	-10.00	-10.00	-10.00	-10.00
	-8.00	-8.00	-8.00	-8.00	-8.00	-10.00	-10.00	-10.00	-10.00	-10.00
$\hat{\boldsymbol{\mu}}_1$	-11.59	-12.00	-11.57	-11.94	-11.69	-9.67	-9.94	-9.79	-9.32	-9.74
	-7.68	-7.52	-8.33	-7.73	-7.73	-10.30	-10.06	-9.78	-9.49	-9.85
$\boldsymbol{\mu}_2$	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00
$\hat{\boldsymbol{\mu}}_2$	10.56	10.61	11.33	10.56	10.63	10.25	10.59	10.65	10.12	9.93
	10.72	10.50	10.18	9.83	10.82	10.60	10.80	10.87	11.05	10.20

(b) Slopes

	X									
$\boldsymbol{\beta}_1$	-0.74	-0.66	-0.64	-0.02	-0.54	-0.06	-0.78	-1.00	-0.65	-0.15
	-0.23	-0.46	-0.71	-0.69	-0.31	-0.30	-0.29	-0.54	-0.45	-0.64
$\hat{\boldsymbol{\beta}}_1$	-0.85	-0.68	-0.60	-0.00	-0.53	-0.00	-0.78	-0.91	-0.59	-0.09
	-0.15	-0.53	-0.84	-0.92	-0.23	-0.33	-0.36	-0.55	-0.42	-0.56
$\boldsymbol{\beta}_2$	0.77	0.75	0.32	0.59	0.75	0.24	0.54	0.63	0.30	0.20
	0.02	0.22	0.85	0.28	0.02	0.16	0.22	0.66	0.32	0.88
$\hat{\boldsymbol{\beta}}_2$	0.84	0.80	0.35	0.62	0.72	0.39	0.47	0.54	0.18	0.18
	0.13	0.11	0.77	0.34	0.08	0.13	0.21	0.61	0.36	0.93

(c) Conditional std. deviations

g	σ_g	$\hat{\sigma}_g$
1	4.20	3.99
2	2.00	1.75

(d) Intercepts

g	β_{0g}	$\hat{\beta}_{0g}$
1	-4.20	-4.36
2	4.50	4.11

6.2 The f.voles data set

In addition to the simulated data analyses of Sect. 6.1, the CWFA family was also applied to a real data set for both clustering and classification. The `f.voles` data set, detailed in Flury (1997, Table 5.3.7) and available in the `Flury` package for R, consists of measurements of female voles from two species, *M. californicus* and *M. ochrogaster*. The data consist of 86 observations for which we have a binary variable Species denoting the species (45 *Microtus ochrogaster* and 41 *M. californicus*), a variable Age measured in days, and six remaining variables related to skull measurements. The names of the variables are the same as in the original analysis of this data set by Airoldi and 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. All of the variables related to the skull are measured in units of 0.1 mm.

The purpose of Airoldi and Hoffmann (1984) was to study age variation in *M. californicus* and *M. ochrogaster* and to predict age on the basis of the skull measurements.

Table 6 Clustering of `f.voles` data using six different approaches

(a) CWFA: CCCU model ($q = 1$)

True \ Est.	1	2	3
Ochrogaster	24	21	–
Californicus	–	–	41

(b) PGMM: CCUU model ($q = 1$)

True \ Est.	1	2	3
Ochrogaster	34	9	2
Californicus	–	–	41

(c) FMA: $q = 2$

True \ Est.	1	2
Ochrogaster	43	2
Californicus	–	41

(d) FMR

True \ Est.	1	2	3	4	5
Ochrogaster	14	5	6	15	5
Californicus	10	3	8	9	11

(e) FMRC

True \ Est.	1	2
Ochrogaster	15	30
Californicus	3	38

(f) MCLUST: EEE model

True \ Est.	1	2
Ochrogaster	43	2
Californicus	–	41

The best model as indicated by the BIC is noted at the top of each sub-table

For our purpose, we assume that data are unlabelled with respect to Species and that our interest is in evaluating clustering and classification using the CWFA family models as well as comparing the algorithm with some well-established mixture model-based techniques. Therefore, Age can be considered the natural Y variable and the $p = 6$ skull measurements can be considered as the $\boldsymbol{X}$ variable for the CWFA framework.

6.2.1 Clustering

All sixteen linear Gaussian CWFA models were fitted—assuming no known group memberships—for $G \in \{2, \ldots, 5\}$ components and $q \in \{1, 2, 3\}$ latent factors, resulting in a total of 192 different models. The model with the largest BIC value was CCCU with $G = 3$ and $q = 1$, with a BIC of $-3{,}837.698$ and an ARI of 0.72. Table 6 displays the clustering results from this model.

Table 6 also shows the clustering results of the following model-based clustering approaches applied to the vector $(\boldsymbol{X}, Y)$:

PGMM: parsimonious latent Gaussian mixture models as described in McNicholas and Murphy (2008, 2010b); McNicholas (2010), and McNicholas et al. (2010), and estimated via the `pgmmEM` function of the R package `pgmm` (McNicholas et al. 2011);

FMA: factor mixture analysis as described in Montanari and Viroli (2010, 2011), and implemented via the `fma` function of the R package `FactMixtAnalysis`;

Table 7 Estimated parameters for the chosen CWFA model applied to the `f.voles` data

(a) Slopes ($\hat{\boldsymbol{\beta}}_{1g}$)

g	X_1	X_2	X_3	X_4	X_5	X_6
1	4.75	2.58	-15.60	14.04	1.05	-13.50
2	-0.62	-2.64	8.30	5.95	-24.41	0.70
3	3.43	-2.27	-5.77	3.09	0.37	-5.47

(b) Conditional std. deviations

g	$\hat{\sigma}$
$\{1, 2, 3\}$	2524.61

(c) Intercepts

g	$\hat{\beta}_{0g}$
1	-851.65
2	-11.53
3	-174.02

FMR: finite mixtures of linear Gaussian regressions as described, among many others, in DeSarbo and Cron (1988), and estimated via the `stepFlexmix` function of the R package `flexmix` (Leisch 2004);
FMRC: finite mixtures of linear Gaussian regressions with concomitants as described in Grün and Leisch (2008) and estimated via the `stepFlexmix` function of the R package `flexmix`; and
MCLUST: parsimonious mixtures of Gaussian distributions as described in Banfield and Raftery (1993); Celeux and Govaert (1995), and Fraley and Raftery (2002), and estimated via the `Mclust` function of the R package `mclust` (see Fraley et al. 2012, for details).

In all cases, we use the range $G \in \{2, \ldots, 5\}$ for mixture components and values $q \in \{1, 2, 3\}$ for the number of latent factors where relevant (i.e., PGMM and FMA). The best model is selected using the BIC. Finally, note that the pair $(\boldsymbol{X}, Y)$ is used as a unique input in MCLUST and FMA. Furthermore, as regards the former, all ten available covariance structures in the package `mclust` are considered while, *ceteris paribus* with the other approaches, no further covariates are considered for FMA. As seen from Table 6, *M. californicus* was classified correctly using the four approaches: CWFA, PGMM, FMA, and MCLUST. *M. ochrogaster* was classified into two sub-clusters using CWFA and PGMM while FMA and MCLUST classified it into one cluster. However, the CWFA approach had no misclassifications between the two species but PGMM, FMA, and MCLUST misclassified two *M. ochrogaster* as *M. californicus*. On these data, the other two approaches, FMR and FMRC, do not show a good clustering performance; poor results were obtained for FMR in particular.

For completeness, we give estimated parameters for the chosen CWFA model (CCCU, $q = 1$, $G = 3$) in Table 7.

The plot of BIC values is omitted for the sake of space. In Table 8 we list the five models which attained the largest BIC values. Notably, the first four models were characterized by $G = 3$ components and the subsequent four by $G = 2$.

Table 8 BIC values of the top 5 models, sorted in decreasing order, for the `f.voles` data

Model	G	q	BIC
CCCU	3	1	−3,837.698
UCCU	3	1	−3,839.322
UCCU	3	2	−3,848.574
CCUU	3	1	−3,851.072
CCUU	2	1	−3,852.178

Airoldi and Hoffmann (1984) mention that some unexplained geographic variation may exist among the voles. However, no covariate was available with such information. Hence, we opted for the scatter plot matrix to evaluate the presence of sub-clusters (see Fig. 5). Here, the scatter plot of the variables B_3 versus B_4 shows the presence of distinct sub-clusters for *M. ochrogaster*, which supports our results attained using CWFA modelling.

6.2.2 Classification

A subset of observations, consisting of 50 % of the data, was randomly selected and these observations were assumed to have known component membership. To allow for the unobserved sub-cluster noted in the clustering application of Sect. 6.2.1, we ran the algorithm for $G = 2, 3$ and $q = 1, 2, 3$. The best model (CCUU with $G = 2$ and $q = 1$) selected by the BIC (−3,843.482) gave perfect classification, as we can see from Table 9a.

We also ran the classification assuming that the data are actually comprised of three known groups. Therefore, using the classification observed by clustering, we also ran the classification algorithm with 50 % known (i.e., labelled) and 50 % unknown (i.e., unlabelled). To further allow for the unobserved sub-cluster, we ran the algorithm for $G \in \{3, 4\}$ and $q \in \{1, 2, 3\}$. The model selected using the BIC was CCCU with $G = 3$ and $q = 1$, with a BIC value of −3,837.383. Even though the BIC value observed using the classification approach (with $G = 3$ known groups) was very close to the BIC value using clustering, the sub-clusters do not have precisely the same classification using the two approaches. This could be a consequence of the classification of borderline observations among the sub-clusters using maximum *a posteriori* probability. However, the BIC value for the classification using three known groups was higher than the BIC value using two known groups, which again suggests the presence of sub-clusters.

7 Conclusions, discussion, and future work

In this paper, we introduced a novel family of 16 parsimonious CWFA models. They are linear Gaussian cluster-weighted models in which a latent factor structure is assumed for the explanatory random vector in each mixture component. The parsimonious ver-

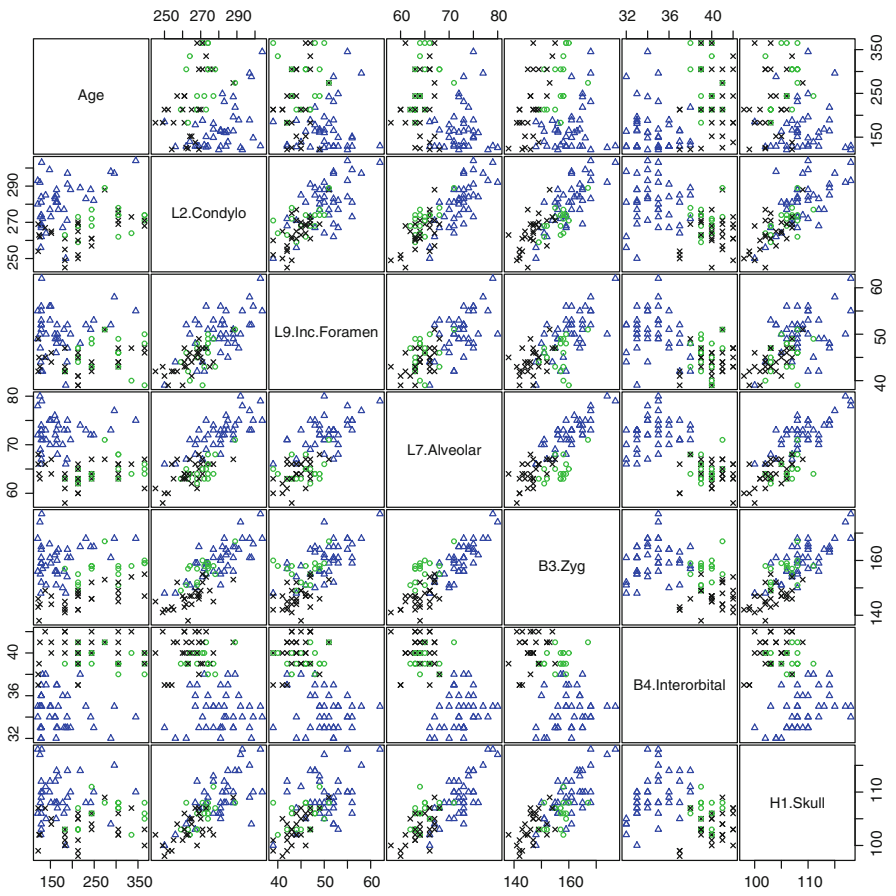

Fig. 5 Scatterplot matrix of `f.voles` data showing the classification observed from CWFA modelling using the clustering framework, where *times symbol* and *open circle* indicate sub-clusters of the *M. ochrogaster* species and *triangle* indicates the *M. californicus* species

sions are obtained by combining all of the constraints described in McNicholas and Murphy (2008) with one of the constraints illustrated in Ingrassia et al. (2013). Due to the introduction of a latent factor structure, the parameters are linear in dimensionality as opposed to the traditional linear Gaussian CWM where the parameters grow quadratically; therefore, our approach is more suitable for modelling complex high dimensional data.The AECM algorithm was used for maximum likelihood estimation of the model parameters. Being based on the EM algorithm, it is very sensitive to starting values due to presence of multiple local maxima. To overcome this problem, we proposed a 5-step hierarchical initialization procedure that utilizes the nested structures of the models within the CWFA family. Because these models have a hierarchical/nested structure, this initialization procedure guarantees a natural ranking on the likelihoods of the models in our family.Using artificial and real data, we demonstrated that these models give very good clustering performance and that the AECM algorithms used were able to recover the parameters very well.

Table 9 Classification of `f.voles` data assuming that 50 % of the observations have known group membership.

True	Est.		
	1	2	
a) 2 known groups			
ochrogaster	45	–	
californicus	–	41	
b) 3 known groups			
ochrogaster	28	17	–
californicus	–	–	41

Also, while the BIC was able to identify the correct model in our simulations, the choice of a convenient model selection criterion for these models is still an open question. Some future work will be devoted to the search for good model selection criteria for these models. Finally, we assumed that the number of factors was the same across groups, which might be too restrictive. However, assuming otherwise also increases the number of models that need to be fitted, resulting in an additional computational burden. Approaches such as variational Bayes approximations might be useful for significantly reducing the number of models that need to be fitted.

Acknowledgments The authors sincerely thank the Associate Editor and the referees for helpful comments and valuable suggestions that have contributed to improving the quality of the manuscript. The work of Subedi and McNicholas was partly supported by an Early Researcher Award from the Ontario Ministry of Research and Innovation.

Appendix A: The conditional distribution of $Y|\boldsymbol{x}, \boldsymbol{u}$

To compute the distribution of $Y|\boldsymbol{x}, \boldsymbol{u}$, we begin by recalling that if $\boldsymbol{Z} \sim N_q(\boldsymbol{m}, \boldsymbol{\Gamma})$ is a random vector with values in $\mathbb{R}^q$ and if $\boldsymbol{Z}$ is partitioned as $\boldsymbol{Z} = \left(\boldsymbol{Z}_1', \boldsymbol{Z}_2'\right)'$, where $\boldsymbol{Z}_1$ takes values in $\mathbb{R}^{q_1}$ and $\boldsymbol{Z}_2$ in $\mathbb{R}^{q_2} = \mathbb{R}^{q-q_1}$, then we can write

$$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix}.$$

Now, because $\boldsymbol{Z}$ has a multivariate normal distribution, $\boldsymbol{Z}_1|\boldsymbol{Z}_2 = \boldsymbol{z}_2$ and $\boldsymbol{Z}_2$ are statistically independent with $\boldsymbol{Z}_1|\boldsymbol{Z}_2 = \boldsymbol{z}_2 \sim N_{q_1}\left(\boldsymbol{m}_{1|2}, \boldsymbol{\Gamma}_{1|2}\right)$ and $\boldsymbol{Z}_2 \sim N_{q_2}(\boldsymbol{m}_2, \boldsymbol{\Gamma}_{22})$, where

$$\boldsymbol{m}_{1|2} = \boldsymbol{m}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}(\boldsymbol{z}_2 - \boldsymbol{m}_2) \quad \text{and} \quad \boldsymbol{\Gamma}_{1|2} = \boldsymbol{\Gamma}_{11} - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\Gamma}_{21}. \tag{15}$$

Therefore, setting $\boldsymbol{Z} = \left(\boldsymbol{Z}_1', \boldsymbol{Z}_2'\right)'$, where $\boldsymbol{Z}_1' = Y$ and $\boldsymbol{Z}_2 = \left(\boldsymbol{X}', \boldsymbol{U}'\right)'$, gives $\boldsymbol{m}_1 = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{\mu}$ and $\boldsymbol{m}_2 = \left(\boldsymbol{\mu}', \boldsymbol{0}'\right)'$, with the elements in $\boldsymbol{\Gamma}$ given by

$$\boldsymbol{\Gamma}_{11} = \boldsymbol{\beta}_1'\boldsymbol{\Sigma}\boldsymbol{\beta}_1 + \sigma^2, \quad \boldsymbol{\Gamma}_{22} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}' & \boldsymbol{I}_q \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\Gamma}_{12} = \begin{bmatrix} \boldsymbol{\beta}_1'\boldsymbol{\Sigma} & \boldsymbol{\beta}_1'\boldsymbol{\Lambda} \end{bmatrix}.$$

It follows that $Y|\boldsymbol{x}, \boldsymbol{u}$ is Gaussian with mean $\boldsymbol{m}_{y|\boldsymbol{x},\boldsymbol{u}} = \mathbb{E}\left(Y|\boldsymbol{x}, \boldsymbol{u}\right)$ and variance $\sigma^2_{y|\boldsymbol{x},\boldsymbol{u}} = \text{Var}\left(Y|\boldsymbol{x}, \boldsymbol{u}\right)$, in accordance with the formulae in (15). Because the inverse matrix of $\boldsymbol{\Gamma}_{22}$ is required in (15), the following formula for the inverse of a partitioned matrix is utilized:

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\left(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\right)^{-1} & \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \end{bmatrix}.$$

Again, writing $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, we have

$$\boldsymbol{\Gamma}_{22}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}' & \boldsymbol{I}_q \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Psi}^{-1} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\left(\boldsymbol{I}_q - \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1} \\ -\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1} & \left(\boldsymbol{I}_q - \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1} \end{bmatrix}.$$

Moreover, according to the Woodbury identity (Woodbury 1950),

$$\boldsymbol{\Sigma}^{-1} = (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1} = \boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_q + \Lambda'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}.$$

Now,

$$\boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1} = \begin{bmatrix} \boldsymbol{\beta}_1'\boldsymbol{\Sigma} & \boldsymbol{\beta}_1'\boldsymbol{\Lambda} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Psi}^{-1} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\left(\boldsymbol{I}_q - \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right)^{-1} \\ -\boldsymbol{\Lambda}\boldsymbol{\Psi}^{-1} & (\boldsymbol{I}_q - \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1' & 0 \end{bmatrix}.$$

Finally, according to (15), we have

$$\boldsymbol{m}_{y|\boldsymbol{x},\boldsymbol{u}} = \boldsymbol{m}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\left[\boldsymbol{z}_2 - \boldsymbol{m}_2\right] = \left(\beta_0 + \boldsymbol{\beta}_1'\boldsymbol{\mu}\right) + \begin{bmatrix} \boldsymbol{\beta}_1' & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} - \boldsymbol{\mu} \\ \boldsymbol{u} - \boldsymbol{0} \end{bmatrix} = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{x},$$

$$\sigma^2_{y|\boldsymbol{x},\boldsymbol{u}} = \boldsymbol{\Gamma}_{11} - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\Gamma}_{21} = \boldsymbol{\beta}_1'\boldsymbol{\Sigma}\boldsymbol{\beta}_1 + \sigma^2 - \begin{bmatrix} \boldsymbol{\beta}_1' & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}\boldsymbol{\beta}_1 \\ \boldsymbol{\Lambda}\boldsymbol{\beta}_1 \end{bmatrix} = \sigma^2.$$

Appendix B: Details on the AECM algorithm for the parsimonious models

This appendix details the AECM algorithm for the models summarized in Table 1.

B.1 Constraint on the Y variable

In all of the models whose identifier starts with 'C', that is the models in which the error variance terms σ_g^2 (of the response variable Y) are constrained to be equal across

groups, i.e., $\sigma_g^2 = \sigma^2$ for $g = 1, \ldots, G$, the common variance σ^2 at the $(k+1)$th iteration of the algorithm is computed as

$$\sigma^{2(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}^{(k+1)}\left\{y_i - \left(\beta_{0g}^{(k+1)} + \boldsymbol{\beta}_{1g}^{'(k+1)}\boldsymbol{x}_i\right)\right\}^2.$$

B.2 constraints on the $\boldsymbol{X}$ variable

With respect to the $\boldsymbol{X}$ variable, as explained in Sect. 3.2, we considered the following constraints on $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$: (i) equal loading matrices $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, (ii) equal error variance $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, and (iii) isotropic assumption: $\boldsymbol{\Psi}_g = \psi_g \boldsymbol{I}_p$. In such cases, the gth term of the expected complete-data log-likelihood $Q_2\left(\boldsymbol{\theta}_2; \boldsymbol{\theta}^{(k+1/2)}\right)$, and then the estimates (12) and (13) in Sect. 4.3, are computed as follows.

B.2.1 Isotropic assumption: $\boldsymbol{\Psi}_g = \psi_g \boldsymbol{I}_p$

In this case, Eq. (9) becomes

$$\begin{aligned} Q_2\left(\boldsymbol{\Lambda}_g, \psi_g; \boldsymbol{\theta}^{(k+1/2)}\right) = {} & \mathrm{C}\left(\boldsymbol{\theta}_1^{(k+1)}\right) + \frac{1}{2}n_g^{(k+1)} \ln|\psi_g^{-1}\boldsymbol{I}_p| - \frac{1}{2}n_g^{(k+1)}\psi_g^{-1}\mathrm{tr}\left\{\boldsymbol{S}_g^{(k+1)}\right\} \\ & + n_g^{(k+1)}\psi_g^{-1}\mathrm{tr}\left\{\boldsymbol{\gamma}_g^{(k)}\boldsymbol{S}_g^{(k+1)}\boldsymbol{\Lambda}_g\right\} - \frac{1}{2}n_g^{(k+1)}\psi_g^{-1}\mathrm{tr} \\ & \times \left\{\boldsymbol{\Lambda}_g\boldsymbol{\Theta}_g^{(k)}\boldsymbol{\Lambda}_g'\right\}, \end{aligned}$$

yielding

$$\frac{\partial Q_2}{\partial \psi_g^{-1}} = \frac{1}{2}n_g^{(k+1)}\left[p\psi_g - \mathrm{tr}\left\{\boldsymbol{S}_g^{(k+1)}\right\} + 2\mathrm{tr}\left\{\boldsymbol{\gamma}_g^{(k)}\boldsymbol{S}_g^{(k+1)}\boldsymbol{\Lambda}_g\right\} - \mathrm{tr}\left\{\boldsymbol{\Lambda}_g\boldsymbol{\Theta}_g^{(k)}\boldsymbol{\Lambda}_g'\right\}\right].$$

Then the estimated ψ_g is attained for $\hat{\psi}_g$, satisfying

$$\frac{\partial Q_2}{\partial \psi_g^{-1}} = 0 \Rightarrow p\psi_g - \mathrm{tr}\left\{\boldsymbol{S}_g^{(k+1)}\right\} + 2\mathrm{tr}\left\{\boldsymbol{\gamma}_g^{(k)}\boldsymbol{S}_g^{(k+1)}\boldsymbol{\Lambda}_g\right\} - \mathrm{tr}\left\{\boldsymbol{\Lambda}_g\boldsymbol{\Theta}_g^{(k)}\boldsymbol{\Lambda}_g'\right\} = 0.$$

Thus, according to (12), for $\boldsymbol{\Lambda}_g = \hat{\boldsymbol{\Lambda}}_g = \boldsymbol{S}_g^{(k+1)}\boldsymbol{\gamma}_g'^{(k)}\boldsymbol{\Theta}_g^{-1}$ we get $\mathrm{tr}\left\{\boldsymbol{\Lambda}_g\boldsymbol{\Theta}_g^{(k)}\boldsymbol{\Lambda}_g'\right\} = \mathrm{tr}\left\{\boldsymbol{\gamma}_g^{(k)}\boldsymbol{S}_g^{(k+1)}\boldsymbol{\Lambda}_g\right\}$ and, finally, $\hat{\psi}_g = \frac{1}{p}\mathrm{tr}\left\{\boldsymbol{S}_g^{(k+1)} - \hat{\boldsymbol{\Lambda}}_g\boldsymbol{\gamma}_g^{(k)}\boldsymbol{S}_g^{(k+1)}\right\}$. Thus,

$$\begin{aligned} \psi_g^{+} &= \frac{1}{p}\mathrm{tr}\left\{\boldsymbol{S}_g^{(k+1)} - \boldsymbol{\Lambda}_g\boldsymbol{\gamma}_g^{+}\boldsymbol{S}_g^{(k+1)}\right\} \\ \boldsymbol{\gamma}_g^{+} &= \boldsymbol{\Lambda}_g'\left(\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g^{+}\boldsymbol{I}_p\right)^{-1}, \end{aligned} \tag{16}$$

with $\boldsymbol{\Theta}_g^+$ computed according to (14).

B.2.2 Equal error variance: $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$

In this case, from Eq. (9), we have

$$Q_2\left(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}; \boldsymbol{\theta}^{(k+1/2)}\right) = \mathrm{C}(\boldsymbol{\theta}_1^{(k+1)}) - \frac{1}{2} n_g^{(k+1)} \ln|\boldsymbol{\Psi}| - \frac{1}{2} n_g^{(k+1)} \mathrm{tr}\left\{\boldsymbol{S}_g^{(k+1)} \boldsymbol{\Psi}^{-1}\right\} + n_g^{(k+1)} \mathrm{tr}\left\{\boldsymbol{\Lambda}_g \boldsymbol{\gamma}_g^{(k)} \boldsymbol{S}_g^{(k+1)} \boldsymbol{\Psi}^{-1}\right\} - \frac{1}{2} n_g^{(k+1)} \mathrm{tr} \times \left\{\boldsymbol{\Lambda}_g' \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)}\right\},$$

yielding

$$\frac{\partial Q_2\left(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}; \boldsymbol{\theta}^{(k+1/2)}\right)}{\partial \boldsymbol{\Psi}^{-1}} = \frac{1}{2} n_g^{(k+1)} \boldsymbol{\Psi} - \frac{1}{2} n_g^{(k+1)} \boldsymbol{S}_g^{(k+1)} + n_g^{(k+1)} \boldsymbol{S}_g^{'(k+1)} \boldsymbol{\gamma}_g^{'(k)} \boldsymbol{\Lambda}_g' - \frac{1}{2} n_g^{(k+1)} \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)} \boldsymbol{\Lambda}_g'.$$

Then the estimated $\hat{\boldsymbol{\Psi}}$ is obtained by satisfying

$$\sum_{g=1}^{G} \frac{\partial Q_2\left(\boldsymbol{\Lambda}_g, \boldsymbol{\Psi}; \boldsymbol{\theta}^{(k+1/2)}\right)}{\partial \boldsymbol{\Psi}^{-1}} = \mathbf{0},$$

that is

$$\frac{n}{2}\boldsymbol{\Psi} - \frac{1}{2}\sum_{g=1}^{G} n_g^{(k+1)} \boldsymbol{S}_g^{(k+1)} + \sum_{g=1}^{G} n_g^{(k+1)} \boldsymbol{S}_g^{'(k+1)} \boldsymbol{\gamma}_g^{'(k)} \boldsymbol{\Lambda}_g' - \frac{1}{2}\sum_{g=1}^{G} n_g^{(k+1)} \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)} \boldsymbol{\Lambda}_g' = \mathbf{0},$$

which can be simplified as

$$\frac{n}{2}\boldsymbol{\Psi} - \frac{1}{2}\sum_{g=1}^{G} n_g^{(k+1)} \left[\boldsymbol{S}_g^{(k+1)} + 2\boldsymbol{S}_g^{'(k+1)} \boldsymbol{\gamma}_g^{'(k)} \boldsymbol{\Lambda}_g' - \boldsymbol{\Lambda}_g \boldsymbol{\Theta}_g^{(k)} \boldsymbol{\Lambda}_g'\right] = \mathbf{0},$$

with $\sum_{g=1}^{G} n_g^{(k+1)} = n$. Again, according to (12), for $\boldsymbol{\Lambda}_g = \hat{\boldsymbol{\Lambda}}_g = \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g^{'(k)} \boldsymbol{\Theta}_g^{-1}$ we get $\hat{\boldsymbol{\Lambda}}_g \boldsymbol{\Theta}_g^{(k)} \hat{\boldsymbol{\Lambda}}_g' = \hat{\boldsymbol{\Lambda}}_g \boldsymbol{\gamma}_g^{(k)} \boldsymbol{S}_g^{(k+1)}$ and, afterwards,

$$\hat{\boldsymbol{\Psi}} = \sum_{g=1}^{G} \frac{n_g}{n} \text{diag}\left\{\boldsymbol{S}_g^{(k+1)} - \hat{\boldsymbol{\Lambda}}_g \boldsymbol{\gamma}_g'^{(k)} \boldsymbol{S}_g^{(k+1)}\right\}$$
$$= \sum_{g=1}^{G} \pi_g^{(k+1)} \text{diag}\left\{\boldsymbol{S}_g^{(k+1)} - \hat{\boldsymbol{\Lambda}}_g \boldsymbol{\gamma}_g^{(k)} \boldsymbol{S}_g^{(k+1)}\right\}. \quad (17)$$

Thus,

$$\boldsymbol{\Psi}^{+} = \sum_{g=1}^{G} \pi_g^{(k+1)} \text{diag}\left\{\boldsymbol{S}_g^{(k+1)} - \boldsymbol{\Lambda}_g^{+} \boldsymbol{\gamma}_g \boldsymbol{S}_g^{(k+1)}\right\}, \quad (18)$$
$$\boldsymbol{\gamma}_g^{+} = \boldsymbol{\Lambda}_g' \left(\boldsymbol{\Lambda}_g^{+} \boldsymbol{\Lambda}_g'^{+} + \boldsymbol{\Psi}^{+}\right)^{-1},$$

where $\boldsymbol{\Theta}_g^{+}$ is computed according to (14).

B.2.3 Equal loading matrices: $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$

In this case, Eq. (9) can be written as

$$Q_2\left(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g; \boldsymbol{\theta}^{(k+1/2)}\right) = \text{C}(\boldsymbol{\theta}_1^{(k+1)}) + \frac{1}{2} n_g^{(k+1)} \ln|\boldsymbol{\Psi}_g^{-1}| - \frac{1}{2} n_g^{(k+1)} \text{tr}\left\{\boldsymbol{S}_g^{(k+1)} \boldsymbol{\Psi}_g^{-1}\right\}$$
$$+ n_g^{(k+1)} \text{tr}\left\{\boldsymbol{\Lambda} \boldsymbol{\gamma}_g^{(k)} \boldsymbol{S}_g^{(k+1)} \boldsymbol{\Psi}_g^{-1}\right\} - \frac{1}{2} n_g^{(k+1)} \text{tr}\left\{\boldsymbol{\Lambda}' \boldsymbol{\Psi}_g^{-1} \boldsymbol{\Lambda} \boldsymbol{\Theta}_g^{(k)}\right\},$$

yielding

$$\frac{\partial Q_2\left(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g; \boldsymbol{\theta}^{(k+1/2)}\right)}{\partial \boldsymbol{\Lambda}} = n_g^{(k+1)} \boldsymbol{\Psi}_g^{-1} \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g'^{(k)} - n_g^{(k+1)} \boldsymbol{\Psi}_g^{-1} \boldsymbol{\Lambda} \boldsymbol{\Theta}_g^{(k)} = \boldsymbol{0}.$$

Then the estimated $\hat{\boldsymbol{\Lambda}}$ is obtained by solving

$$\sum_{g=1}^{G} \frac{\partial Q_2\left(\boldsymbol{\Lambda}, \boldsymbol{\Psi}_g; \boldsymbol{\theta}^{(k+1/2)}\right)}{\partial \boldsymbol{\Lambda}} = \sum_{g=1}^{G} n_g^{(k+1)} \boldsymbol{\Psi}_g^{-1} \left[\boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g'^{(k)} - \boldsymbol{\Lambda} \boldsymbol{\Theta}_g^{(k)}\right] = \boldsymbol{0}, \quad (19)$$

with $\boldsymbol{\gamma}_g^{(k)} = \boldsymbol{\Lambda}'^{(k)} \left(\boldsymbol{\Lambda}^{(k)} \boldsymbol{\Lambda}'^{(k)} + \boldsymbol{\Psi}_g^{(k)}\right)^{-1}$. In this case, the loading matrix cannot be solved directly and must be solved in a row-by-row manner as suggested by

McNicholas and Murphy (2008). Therefore,

$$\lambda_i^+ = \mathbf{r}_i \left(\sum_{g=1}^{G} \frac{n_g}{\psi_{g(i)}} \boldsymbol{\Theta}_g \right)^{-1} \tag{20}$$

$$\boldsymbol{\gamma}_g^+ = \boldsymbol{\Lambda}' \left(\boldsymbol{\Lambda}^+ \boldsymbol{\Lambda}'^+ + \boldsymbol{\Psi}_g^+ \right)^{-1} \tag{21}$$

$$\boldsymbol{\Theta}_g^+ = \boldsymbol{I}_q - \boldsymbol{\gamma}_g^+ \boldsymbol{\Lambda}^+ + \boldsymbol{\gamma}_g^+ \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g'^+, \tag{22}$$

where λ_i^+ is the ith row of the matrix $\boldsymbol{\Lambda}^+$, $\psi_{g(i)}$ is the ith diagonal element of $\boldsymbol{\Psi}_g$, and $\mathbf{r}_i$ represents the ith row of the matrix $\sum_{g=1}^{G} n_g^{(k+1)} \left(\boldsymbol{\Psi}_g' \right)^{-1} \boldsymbol{S}_g^{(k+1)}$.

B.2.4 Further details

A further schematization is here given without considering the constraint on the Y variable. Thus, with reference to the model identifier, we will only refer to the last three letters.

Models ended by UUU: no constraint is assumed.
Models ended by UUC: $\boldsymbol{\Psi}_g = \psi_g \boldsymbol{I}_p$, where the parameter ψ_g is updated according to (16).
Models ended by UCU: $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, where the matrix $\boldsymbol{\Psi}$ is updated according to (18).
Models ended by UCC: $\boldsymbol{\Psi}_g = \psi \boldsymbol{I}_p$. By combining (16) and (18) we obtain

$$\begin{aligned} \hat{\psi} &= \frac{1}{p} \sum_{g=1}^{G} \frac{n_g^{(k+1)}}{n} \mathrm{tr} \left\{ \boldsymbol{S}_g^{(k+1)} - \hat{\boldsymbol{\Lambda}}_g \boldsymbol{\gamma}_g^{(k)} \boldsymbol{S}_g^{(k+1)} \right\} \\ &= \frac{1}{p} \sum_{g=1}^{G} \hat{\pi}_g^{(k+1)} \mathrm{tr} \left\{ \boldsymbol{S}_g^{(k+1)} - \hat{\boldsymbol{\Lambda}}_g \boldsymbol{\gamma}_g^{(k)} \boldsymbol{S}_g^{(k+1)} \right\}. \end{aligned} \tag{23}$$

Thus, $\psi^+ = (1/p) \sum_{g=1}^{G} \pi_g^{(k+1)} \mathrm{tr} \left\{ \boldsymbol{S}_g^{(k+1)} - \boldsymbol{\Lambda}_g^+ \boldsymbol{\gamma}_g \boldsymbol{S}_g^{(k+1)} \right\}$ and $\boldsymbol{\gamma}_g^+ = \boldsymbol{\Lambda}_g'^+ \left(\boldsymbol{\Lambda}_g^+ \boldsymbol{\Lambda}_g'^+ + \psi^+ \boldsymbol{I}_p \right)^{-1}$, with $\boldsymbol{\Theta}_g^+$ computed according to (14).
Models ended by CUU: $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$, where the matrix $\boldsymbol{\Lambda}$ is updated according to (20). In this case, $\boldsymbol{\Psi}_g$ is estimated directly from (11) and thus $\boldsymbol{\Psi}_g^+ = \mathrm{diag} \left\{ \boldsymbol{S}_g^{(k+1)} - 2\boldsymbol{\Lambda}^+ \boldsymbol{\gamma}_g \boldsymbol{S}_g^{(k+1)} + \boldsymbol{\Lambda}^+ \boldsymbol{\Theta}_g \boldsymbol{\Lambda}'^+ \right\}$, with $\boldsymbol{\gamma}_g^+$ and $\boldsymbol{\Theta}_g^+$ computed according to (21) and (22), respectively.

Models ended by CUC: $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_g = \psi_g \boldsymbol{I}_p$. In this case, Equation (19), for $\boldsymbol{\Psi}_g = \psi_g \boldsymbol{I}_p$, yields

$$\sum_{g=1}^{G} \frac{\partial Q_2\left(\boldsymbol{\Lambda}, \psi_g; \boldsymbol{\theta}^{(k+1/2)}\right)}{\partial \boldsymbol{\Lambda}} = \sum_{g=1}^{G} n_g^{(k+1)} \psi_g^{-1} \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g^{'(k)} - \sum_{g=1}^{G} n_g^{(k+1)} \psi_g^{-1} \boldsymbol{\Theta}_g^{(k)} = \boldsymbol{0},$$

and afterwards

$$\hat{\boldsymbol{\Lambda}} = \left(\sum_{g=1}^{G} \frac{n_g^{(k+1)}}{\psi_g^{-1}} \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g^{'(k)} \right) \left(\sum_{g=1}^{G} \frac{n_g^{(k+1)}}{\psi_g^{-1}} \boldsymbol{\Lambda} \right)^{-1},$$

with $\boldsymbol{\gamma}_g^{(k)} = \boldsymbol{\Lambda}^{'(k)} \left(\boldsymbol{\Lambda}^{(k)} \boldsymbol{\Lambda}^{'(k)} + \psi_g^{(k)} \boldsymbol{I}_p \right)^{-1}$. Moreover, from

$$\frac{\partial Q_2\left(\boldsymbol{\Lambda}, \psi_g; \boldsymbol{\theta}^{(k+1/2)}\right)}{\partial \psi_g^{-1}} = \frac{p}{2}\psi_g - \frac{n_g^{(k+1)}}{2} \left[\mathrm{tr}\left\{ \boldsymbol{S}_g^{(k+1)} \right\} - 2\mathrm{tr}\left\{ \boldsymbol{S}_g^{'(k+1)} \boldsymbol{\gamma}_g^{'(k)} \boldsymbol{\Lambda}' \right\} + \mathrm{tr}\left\{ \boldsymbol{\Lambda} \boldsymbol{\Theta}_g^{(k+1)} \boldsymbol{\Lambda}' \right\} \right] = 0$$

we get $\hat{\psi}_g = (1/p)\mathrm{tr}\left\{ \boldsymbol{S}_g^{(k+1)} - 2\hat{\boldsymbol{\Lambda}} \boldsymbol{\gamma}_g^{'(k)} \boldsymbol{S}_g + \hat{\boldsymbol{\Lambda}} \boldsymbol{\Theta}_g \hat{\boldsymbol{\Lambda}}' \right\}$. Thus,

$$\begin{aligned} \boldsymbol{\Lambda}^{+} &= \left(\sum_{g=1}^{G} \frac{n_g^{(k+1)}}{\psi_g^{-1}} \boldsymbol{S}_g^{(k+1)} \boldsymbol{\gamma}_g' \right) \left(\sum_{g=1}^{G} \frac{n_g^{(k+1)}}{\psi_g^{-1}} \boldsymbol{\Lambda} \right)^{-1} \\ \psi_g^{+} &= \frac{1}{p} \mathrm{tr}\left\{ \boldsymbol{S}_g^{(k+1)} - 2\, \boldsymbol{\Lambda}^{+} \boldsymbol{\gamma}_g' \boldsymbol{S}_g + \boldsymbol{\Lambda}^{+} \boldsymbol{\Theta} \boldsymbol{\Lambda}'^{+} \right\} \\ \boldsymbol{\gamma}_g^{+} &= \boldsymbol{\Lambda}'^{+} \left(\boldsymbol{\Lambda}^{+} \boldsymbol{\Lambda}'^{+} + \psi_g^{+} \boldsymbol{I}_p \right)^{-1}, \end{aligned}$$

with $\boldsymbol{\Theta}_g^{+}$ computed according to (22).

Models ended by CCU: $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$, so that $\boldsymbol{\gamma}^{(k)} = \boldsymbol{\Lambda}'^{(k)} \left(\boldsymbol{\Lambda}^{(k)}\boldsymbol{\Lambda}^{(k)} + \boldsymbol{\Psi}^{(k)}\right)^{-1}$. Setting $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$ in (19), we get

$$\begin{aligned}\sum_{g=1}^{G} \frac{\partial Q_2\left(\boldsymbol{\Lambda}, \boldsymbol{\Psi}; \boldsymbol{\theta}^{(k+1/2)}\right)}{\partial \boldsymbol{\Lambda}} &= \sum_{g=1}^{G} n_g^{(k+1)} \boldsymbol{\Psi}^{-1}\left[\boldsymbol{S}_g^{(k+1)}\boldsymbol{\gamma}'^{(k)} - \boldsymbol{\Lambda}\boldsymbol{\Theta}_g^{(k)}\right] \\ &= \boldsymbol{\Psi}^{-1}\left[\boldsymbol{\gamma}'^{(k)} \sum_{g=1}^{G} n_g^{(k+1)} \boldsymbol{S}_g^{(k+1)} - \boldsymbol{\Lambda} \sum_{g=1}^{G} n_g^{(k+1)} \boldsymbol{\Theta}_g^{(k)}\right] \\ &= \boldsymbol{\Psi}^{-1}\left[\boldsymbol{\gamma}'^{(k)} \boldsymbol{S}^{(k+1)} - \boldsymbol{\Lambda}\boldsymbol{\Theta}^{(k)}\right] = \mathbf{0},\end{aligned}$$

where $\boldsymbol{S}^{(k+1)} = \sum_{g=1}^{G} \pi_g^{(k+1)} \boldsymbol{S}_g^{(k+1)}$ and $\boldsymbol{\Theta}^{(k)} = \sum_{g=1}^{G} \pi_g^{(k+1)} \boldsymbol{\Theta}_g^{(k)} = \boldsymbol{I}_q - \boldsymbol{\gamma}^{(k)}\boldsymbol{\Lambda}^{(k)} + \boldsymbol{\gamma}^{(k)}\boldsymbol{S}^{(k+1)}\boldsymbol{\gamma}'^{(k)}$. Thus,

$$\hat{\boldsymbol{\Lambda}} = \boldsymbol{S}^{(k+1)}\boldsymbol{\gamma}'^{(k)} \left(\boldsymbol{\Theta}^{(k)}\right)^{-1}. \tag{24}$$

Moreover, setting $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ in (17), we get $\hat{\boldsymbol{\Psi}} = \text{diag}\left\{\boldsymbol{S}^{(k+1)} - \hat{\boldsymbol{\Lambda}}\boldsymbol{\gamma}^{(k)}\boldsymbol{S}^{(k+1)}\right\}$. Hence,

$$\begin{aligned}\boldsymbol{\Lambda}^{+} &= \boldsymbol{S}^{(k+1)}\boldsymbol{\gamma}'\boldsymbol{\Theta}^{-1} \\ \boldsymbol{\Psi}^{+} &= \text{diag}\left\{\boldsymbol{S}^{(k+1)} - \boldsymbol{\Lambda}^{+}\boldsymbol{\gamma}\boldsymbol{S}^{(k+1)}\right\} \\ \boldsymbol{\gamma}_g^{+} &= \boldsymbol{\Lambda}'^{+}\left(\boldsymbol{\Lambda}^{+}\boldsymbol{\Lambda}'^{+} + \boldsymbol{\Psi}^{+}\right)^{-1},\end{aligned} \tag{25}$$

with $\boldsymbol{\Theta}_g^{+}$ computed according to (22).

Models ended by CCC: $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}_g = \psi \boldsymbol{I}_p$, so that $\boldsymbol{\gamma}^{(k)} = \boldsymbol{\Lambda}'^{(k)}\left(\boldsymbol{\Lambda}^{(k)}\boldsymbol{\Lambda}'^{(k)} + \psi^{(k)}\right)^{-1}$. Here, the estimated loading matrix is again (24), while the isotropic term obtained from (23) for $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ is $\hat{\psi} = (1/p)\text{tr}\left\{\boldsymbol{S}^{(k+1)} - \hat{\boldsymbol{\Lambda}}\boldsymbol{\gamma}^{(k)}\boldsymbol{S}^{(k+1)}\right\}$, with $\boldsymbol{\gamma}_g^{(k)} = \boldsymbol{\Lambda}_g'^{(k)}\left(\boldsymbol{\Lambda}_g^{(k)}\boldsymbol{\Lambda}_g'^{(k)} + \psi^{(k)}\boldsymbol{I}_p\right)^{-1}$. Hence, $\psi^{+} = (1/p)\text{tr}\left\{\boldsymbol{S}^{(k+1)} - \boldsymbol{\Lambda}^{+}\boldsymbol{\gamma}\boldsymbol{S}^{(k+1)}\right\}$ and $\boldsymbol{\gamma}^{+} = \boldsymbol{\Lambda}'^{+}\left(\boldsymbol{\Lambda}^{+}\boldsymbol{\Lambda}'^{+} + \psi^{+}\boldsymbol{I}_p\right)^{-1}$, with $\boldsymbol{\Lambda}^{+}$ and $\boldsymbol{\Theta}_g^{+}$ computed according to (25) and (22), respectively.

Appendix C: True and estimated covariance matrices of Sect. 6.1

Because the loading matrices are not unique, for the simulated data of Examples 1 and 2 we limit the attention to a comparison, for each $g = 1, \ldots, G$, of true and estimated covariance matrices.

C.1 Example 1

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 103.36 & 103.07 & 101.37 & 79.41 & 105.66 \\ 103.08 & 119.39 & 110.23 & 85.97 & 115.47 \\ 101.37 & 110.23 & 129.77 & 106.08 & 118.50 \\ 79.41 & 85.97 & 106.08 & 101.46 & 95.21 \\ 105.66 & 115.47 & 118.50 & 95.21 & 121.63 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 107.59 & 114.55 & 110.42 & 87.29 & 114.43 \\ 114.55 & 139.40 & 127.06 & 100.09 & 132.06 \\ 110.42 & 127.06 & 146.31 & 122.92 & 134.12 \\ 87.29 & 100.09 & 122.92 & 117.97 & 110.09 \\ 114.43 & 132.06 & 134.12 & 110.09 & 135.66 \end{bmatrix},$$

and

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 34.25 & 15.16 & 17.81 & 22.39 & 14.62 \\ 15.16 & 17.01 & 11.42 & 13.98 & 8.95 \\ 17.81 & 11.42 & 17.62 & 16.12 & 10.45 \\ 22.39 & 13.98 & 16.12 & 28.11 & 13.11 \\ 14.62 & 8.95 & 10.45 & 13.11 & 10.19 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 22.16 & 7.44 & 13.71 & 12.89 & 10.12 \\ 7.44 & 11.25 & 7.59 & 8.05 & 5.48 \\ 13.71 & 7.59 & 18.83 & 13.53 & 10.13 \\ 12.89 & 8.05 & 13.53 & 22.00 & 9.41 \\ 10.12 & 5.48 & 10.13 & 9.41 & 8.63 \end{bmatrix}.$$

C.2 Example 2

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 10.41 & 3.61 & 4.07 & 4.48 & 5.71 \\ 3.61 & 7.83 & 2.88 & 3.18 & 4.03 \\ 4.07 & 2.88 & 8.67 & 3.81 & 4.64 \\ 4.48 & 3.18 & 3.81 & 9.61 & 5.17 \\ 5.71 & 4.04 & 4.64 & 5.17 & 11.73 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 8.86 & 3.89 & 5.06 & 3.84 & 5.72 \\ 3.89 & 7.23 & 3.59 & 1.79 & 4.04 \\ 5.06 & 3.59 & 8.44 & 3.85 & 5.50 \\ 3.84 & 1.79 & 3.85 & 7.74 & 4.38 \\ 5.72 & 4.04 & 5.50 & 4.38 & 9.81 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 103.36 & 103.07 & 101.37 & 79.41 & 105.66 \\ 103.08 & 122.1 & 110.23 & 85.97 & 115.47 \\ 101.37 & 110.23 & 134.33 & 106.08 & 118.50 \\ 79.41 & 85.97 & 106.08 & 102.73 & 95.21 \\ 105.66 & 115.47 & 118.50 & 95.21 & 129.21 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 106.17 & 100.46 & 93.18 & 73.81 & 105.01 \\ 100.46 & 113.71 & 92.97 & 72.22 & 107.88 \\ 93.18 & 92.97 & 108.25 & 83.08 & 102.36 \\ 73.81 & 72.22 & 83.08 & 80.09 & 81.85 \\ 105.01 & 107.88 & 102.36 & 81.85 & 122.59 \end{bmatrix}.$$

and

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 25.19 & 15.16 & 17.81 & 22.39 & 14.62 \\ 15.16 & 10.67 & 11.42 & 13.98 & 8.95 \\ 17.81 & 11.42 & 13.12 & 16.12 & 10.45 \\ 22.39 & 13.98 & 16.12 & 20.31 & 13.11 \\ 14.62 & 8.95 & 10.45 & 13.11 & 8.70 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_3 = \begin{bmatrix} 32.47 & 19.91 & 23.06 & 28.78 & 18.80 \\ 19.91 & 14.10 & 14.96 & 18.25 & 11.66 \\ 23.06 & 14.96 & 16.95 & 20.77 & 13.45 \\ 28.78 & 18.25 & 20.77 & 25.95 & 16.77 \\ 18.80 & 11.66 & 13.45 & 16.77 & 11.10 \end{bmatrix}.$$

References

Airoldi, J, Hoffmann R (1984) Age variation in voles (Microtus californicus, M. ochrogaster) and its significance for systematic studies, Occasional papers of the Museum of Natural History, vol 111. University of Kansas, Lawrence

Aitken AC (1926) On Bernoulli's numerical solution of algebraic equations. Proc Royal Soc Edinburgh 46:289–305

Andrews JL, McNicholas PD, Subedi S (2011) Model-based classification via mixtures of multivariate t-distributions. Comput Stat Data Anal 55(1):520–529

Baek J, McLachlan GJ, Flack LK (2010) Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data. IEEE Trans Pattern Anal Mach Intell 32:1298–1309

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49(3):803–821

Bartholomew DJ, Knott M (1999) Latent variable models and factor analysis. In: Kendall's library of statistics, vol 7, 2nd edn. Edward Arnold, London

Bartlett M (1953) Factor analysis in psychology as a statistician sees it. In: Uppsala symposium on psychological factor analysis, Number 3 in Nordisk Psykologi's Monograph Series, Uppsala, Sweden, pp 23–34. Almquist and Wiksell, Uppsala

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell 22(7):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 Inst Stat Math 46(2):373–388

Bouveyron C, Girard S, Schmid C (2007) High-dimensional data clustering. Comput Stat Data Anal 52(1):502–519

Browne RP, McNicholas PD (2012) Model-based clustering, classification, and discriminant analysis of data with mixed type. J Stat Plann Infer 142(11):2976–2984
Browne RP, McNicholas PD, Sparling MD (2012) Model-based learning using a mixture of mixtures of Gaussian and uniform distributions. IEEE Trans Pattern Anal Mach Intell 34(4):814–817
Carvalho C, Chang J, Lucas J, Nevins J, Wang Q, West M (2008) High-dimensional sparse factor modeling: applications in gene expression genomics. J Am Stat Assoc 103(484):1438–1456
Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recogn 28(5):781–793
Dean N, Murphy TB, Downey G (2006) Using unlabelled data to update classification rules with applications in food authenticity studies. J Royal Stat Soc Ser C 55(1):1–14
Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc Ser B 39(1):1–38
DeSarbo W, Cron W (1988) A maximum likelihood methodology for clusterwise linear regression. J Classif 5(2):249–282
Flury B (1997) A first course in multivariate statistics. Springer, New York
Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631
Fraley C, Raftery AE, Murphy TB, 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. Springer, New York
Gershenfeld N (1997) Nonlinear inference and cluster-weighted modeling. Ann New York Acad Sci 808(1):18–24
Ghahramani Z, Hinton G (1997) The EM algorithm for factor analyzers. Technical report CRG-TR-96-1, University of Toronto, Toronto
Grün B, Leisch F (2008) Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw 28(4):1–35
Hennig C (2000) Identifiablity of models for clusterwise linear regression. J Classif 17(2):273–296
Hosmer D Jr (1973) A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics 29(4):761–770
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
Ingrassia S, Minotti SC, Punzo A (2013) Model-based clustering via linear cluster-weighted models. DOI:10.1016/j.csda.2013.02.012 Computational Statistics and Data Analysis
Ingrassia, S, Minotti SC, Punzo A, Vittadini G (2012a) Generalized linear cluster-weighted models. eprint arXiv: 1211.1171, http://arxiv.org/abs/1211.1171
Ingrassia S, Minotti SC, Vittadini G (2012b) Local statistical modeling via the cluster-weighted approach with elliptical distributions. J Classif 29(3):363–401
Karlis D, Santourian A (2009) Model-based clustering with non-elliptically contoured distributions. Stat Comput 19(1):73–83
Leisch F (2004) Flexmix: a general framework for finite mixture models and latent class regression in R. J Stat Softw 11(8):1–18
Lin T-I (2010) Robust mixture modeling using multivariate skew t distributions. Stat Comput 20:343–356
Lindsay BG (1995) Mixture models: theory, geometry and applications. In: NSF-CBMS regional conference series in probability and statistics, vol. 5. Institute of Mathematical Statistics, Hayward
McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York
McLachlan GJ, Peel D (2000a) Finite mixture models. Wiley, New York
McLachlan GJ, D Peel (2000b) Mixtures of factor analyzers. In: Proceedings of the seventh international conference on machine learning, pp 599–606. Morgan Kaufmann, San Francisco.
McNicholas PD (2010) Model-based classification using latent Gaussian mixture models. J Stat Plann Infer 140(5):1175–1181
McNicholas PD, Jampani KR, McDaid AF, Murphy TB, Banks L (2011) PGMM: Parsimonious Gaussian Mixture Models. R package version 1.0.
McNicholas PD, Murphy TB (2008) Parsimonious Gaussian mixture models. Stat Comput 18(3):285–296
McNicholas PD, Murphy TB (2010a) Model-based clustering of longitudinal data. Can J Stat 38(1):153–168
McNicholas PD, Murphy TB (2010b) Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics 26(21):2705–2712

McNicholas PD, Murphy TB, McDaid AF, Frost D (2010) Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. Comput Stat Data Anal 54(3):711–723
McNicholas PD, Subedi S (2012) Clustering gene expression time course data using mixtures of multivariate t-distributions. J Stat Plann Infer 142(5):1114–1127
Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80(2):267–278
Meng XL, van Dyk D (1997) The EM algorithm: an old folk-song sung to a fast new tune. J Royal Stat Soc Ser B (Stat Methodol) 59(3):511–567
Montanari A, Viroli C (2010) Heteroscedastic factor mixture analysis. Stat Modell 10(4):441–460
Montanari A, Viroli C (2011) Dimensionally reduced mixtures of regression models. J Stat Plann Infer 141(5):1744–1752
Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computation, 2nd edn. Cambridge University Press, Cambridge
Punzo, A (2012) Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. eprint arXiv: 1207.0939, http://arxiv.org/abs/1207.0939
Rand W (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850
Sakamoto Y, Ishiguro M, Kitagawa G (1983) Akaike information criterion statistics. Reidel, Boston
Schöner B (2000) Probabilistic characterization and synthesis of complex data driven systems. Ph. D. thesis, MIT
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Scrucca L (2010) Dimension reduction for model-based clustering. Stat Comput 20(4):471–484
R Development Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101
Tipping TE, Bishop CM (1999) Mixtures of probabilistic principal component analysers. Neural Comput 11(2):443–482
Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York
Wang Q, Carvalho C, Lucas J, West M (2007) BFRM: Bayesian factor regression modelling. Bull Int Soc Bayesian Anal 14(2):4–5
West M (2003) Bayesian factor regression models in the "large p, small n" paradigm. In: Bernardo J, Bayarri M, Berger J, Dawid A, Heckerman D, Smith A, West M (eds) Bayesian statistics, vol 7. Oxford University Press, Oxford, pp 723–732
Wolfe JH (1963) Object cluster analysis of social areas. Master's thesis, University of California, Berkeley
Wolfe JH (1970) Pattern clustering by multivariate mixture analysis. Multivariate Behav Res 5(3):329–350
Woodbury MA (1950) Inverting modified matrices. Statistical Research Group, Memo. Rep. no. 42. Princeton University, Princeton