Multivariate response and parsimony for Gaussian cluster-weighted models

Utkarsh J. Dang, Antonio Punzo, Paul D. McNicholas, Salvatore Ingrassia, Ryan P. Browne

Department of Mathematics & Statistics, University of Guelph

Abstract

A family of parsimonious Gaussian cluster-weighted models (CWMs) is presented. This family concerns a multivariate extension to cluster-weighted modelling that can account for correlations between multivariate response. Parsimony is attained by constraining parts of an eigen-decomposition imposed on the component covariance matrices. A sufficient condition for identifiability is provided and an expectation-maximization algorithm is presented for parameter estimation. Model performance is investigated on both synthetic and classical real data sets and compared with some popular approaches. Finally, it is shown that accounting for linear dependencies in the presence of a linear regression structure offers better performance, vis-à-vis clustering, over existing methodologies.

1 Introduction

Mixture models have seen increasing use over the last decade or so with important applications in model-based clustering and classification. Various mixture-based methods have emerged for clustering multivariate data. Arguably, the most famous model-based clustering methodology is the Gaussian parsimonious clustering models (GPCMs; Celeux and Govaert, 1995) family, which is supported by the mclust (Fraley and Raftery, 1999), mixture (Browne and McNicholas, 2013), and Rmixmod (Lebret et al., 2012) packages for R (R Core Team, 2013). However, such models do not typically account for dependencies via covariates.

*Department of Biology, McMaster University, Hamilton, Ontario, Canada. E-mail: udang@mcmaster.ca.
†Department of Economics and Business, University of Catania, Catania, Italy.
‡Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada.
§Department of Economics and Business, University of Catania, Catania, Italy.
¶Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada.
When there is a clear regression relationship between some variables, important insight can be gained by accounting for functional dependencies between those variables. For such data, traditional model-based clustering methods that fail to incorporate such a relationship may not perform as well.

Some popular mixture-based methodologies that deal with regression data are finite mixtures of regression (FMR; DeSarbo and Cron, 1988), and finite mixtures of regression with concomitant variables (FMRC; Wedel, 2002), supported by the flexmix (Leisch, 2004; Grün et al., 2008) package for R. FMR only model the distribution of the response given the covariates, whereas FMRC also model the mixing weights of the components as a multinomial logistic model of some concomitant variables (which are often the covariate variables). However, these methodologies do not explicitly use the distribution of the covariates for clustering, i.e., the assignment of data points to clusters does not directly utilize any information from the distribution of the covariates.

A flexible framework for density estimation and clustering of data with local functional dependencies is represented by the cluster-weighted model (CWM; Gershenfeld, 1997), also called the saturated mixture regression model by Wedel (2002). Ingrassia et al. (2012) investigated CWMs in a general statistical mixture framework. The same paper presented theoretical and numerical properties, and discussed the performance of the model under both Gaussian and $t$ distributional assumptions (see also Ingrassia et al., 2014). As opposed to FMR and FMRC, CWMs allow for assignment dependence (cf. Hennig, 2000), i.e., the assignment of an observation to a cluster is also dependent on the distribution of the covariates. In such models, the component covariates’ distributions can be distinct. Some extensions of this methodology have previously dealt with non-linear local relationships (Punzo, 2014), high dimensional covariates (Subedi et al., 2013), and various response types (Ingrassia et al., 2014).

In this paper, a family of parsimonious Gaussian CWMs is presented. It concerns multivariate response, while CWMs have only dealt with univariate response so far. Note that parsimony is vital in real data applications and this is the first time that CWMs are being used with eigen-decomposed covariance structures. Regarding multivariate response in the FMR and FMRC framework, note that while using multivariate response is possible in flexmix, it currently does not account for correlated response variables, i.e., these models assume independence between the response variables. FMR models that deal with correlated response variables have been recently proposed (Soffritti and Galimberti, 2011; Galimberti and Soffritti, 2014) but these models do not decompose the covariance structure nor do they use information from the distribution of the covariates. Furthermore, these papers do not investigate FMRC models that deal with correlated response. Families of eigen-decomposed parsimonious FMR and FMRC models that can account for correlated response variables have been recently proposed (eFMR and eFMRC; Dang and McNicholas, 2013); however, these models do not take into account the distribution of the covariates.

For the proposed multivariate response CWM, parsimonious models are developed by constraining parts of an eigen-decomposition imposed on both (the response and the co-
variates) component covariance matrices. This family of parsimonious models is referred to as the eigen-decomposed multivariate response CWM (eMCWM). The Bayesian information criterion (BIC; Schwarz, 1978) and the integrated completed likelihood (ICL; Biernacki et al., 2000) are considered for model selection on this family. Comparisons to FMR, FMRC, eFMR, eFMRC, and GPCMs are made. Note that hereafter, flexmix FMR and FMRC models will simply be referred to as the FMR and FMRC models.

The rest of the paper is organized as follows. In Section 2, basic ideas on CWMs are summarized. In Section 3, we recall an eigen-decomposition of a covariance matrix. Identifiability is treated in Section 4. An EM algorithm for maximum likelihood parameter estimation is presented in Section 5. Moreover, issues of model selection, algorithm initialization, convergence criterion and performance assessment are also discussed. In Section 6, results of numerical studies based on both real and simulated data are presented. Finally, in Section 7, some conclusions and ideas for future research are discussed.

2 Cluster-weighted models

Multivariate correlated responses can be accounted for quite conveniently in a CWM framework. Let $X$ and $Y$ be random vectors defined on $\Omega$ with joint probability distribution $p(x, y)$. Here, the response vector $Y$ has values in $\mathbb{R}^d$ and the vector of covariates $X$ has values in $\mathbb{R}^p$. Let $\Omega$ be partitioned into $G$ disjoint groups, such that $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$. Then, in a CWM framework, the joint probability $p(x, y)$ can be decomposed as

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

where $p(y|x, \Omega_g)$ is the conditional density of the multivariate response $Y$ given the covariates $X$ and $\Omega_g$, $p(X|\Omega_g)$ is the probability density of $X$ given $\Omega_g$, and $\pi_g = p(\Omega_g)$ are the mixing weights, where $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$, $g = 1, \ldots, G$. In model (1), $X$ is assumed to be normally distributed with mean $\mu_{Xg}$ and covariance matrix $\Sigma_{Xg}$, and $Y|X = x$ is assumed to be normally distributed with conditional mean $\mu_Y(x|\beta_g)$, given by some linear transformation of $X$, and covariance matrix $\Sigma_{Yg}$, $g = 1, \ldots, G$. Here, $\mu_Y(x|\beta_g) = \beta_g' x^*$ is used where $\beta_g \in \mathbb{R}^{(1+p) \times d}$ and $x^* = (1, x)$. Then, model (1) can be rewritten as

$$p(x, y|\vartheta) = \sum_{g=1}^{G} \phi_d(y|x, \mu_Y(x, \beta_g), \Sigma_{Yg}) \phi_p(x, \mu_{Xg}, \Sigma_{Xg}) \pi_g,$$

where $\phi_d$ ($\phi_p$) represents the density of a $d$-variate ($p$-variate) Gaussian random vector and $\vartheta$ denotes the set of all parameters.
3 Parsimonious Models

For a single $q \times q$ covariance matrix, the number of free parameters increases quadratically with the dimensionality $q$. In model-based clustering, parsimony is usually necessary for real applications. Parsimony can be introduced by constraining parts of a particular decomposition of a covariance matrix (Celeux and Govaert, 1995; McNicholas et al., 2010). An eigen-decomposition of such a matrix (cf. Celeux and Govaert, 1995) yields

$$\Sigma_g = \lambda_g \Gamma_g \Delta_g \Gamma'_g,$$

for $g = 1, \ldots, G$, where $\lambda_g = |\Sigma_g|^{1/q}$ is a constant, $\Delta_g$ is a diagonal matrix with entries (sorted in decreasing order) proportional to the eigenvalues of $\Sigma_g$ with the constraint $|\Delta_g| = 1$, and $\Gamma_g$ is a $q \times q$ orthogonal matrix of the eigenvectors (ordered according to the eigenvalues) of $\Sigma_g$. Geometrically, $\lambda_g$ determines the volume, $\Delta_g$ the shape, and $\Gamma_g$ the orientation of the $g$th component. By constraining $\lambda_g$, $\Gamma_g$, and $\Delta_g$ to be equal or across groups, a family of fourteen models (Table 1) is obtained. This family can be further split in three subfamilies. Here, the EII and VII models belong to the spherical family, the models with an axis-aligned orientation belong to the diagonal family, while the rest of the models belong to the general family.

Table 1: Geometric interpretation and the number of free parameters in the eigen-decomposed covariance structures.

| Model | Volume | Shape | Orientation | $\Sigma_g$ | Free Cov. Parameters |
|-------|--------|-------|-------------|-----------|----------------------|
| EII   | Equal  | Spherical | -           | $\lambda I$ | 1                    |
| VII   | Variable | Spherical | -           | $\lambda_g I$ | $G$                 |
| EEI   | Equal  | Equal  | Axis-Aligned | $\lambda \Delta$ | $q$                  |
| VEE   | Equal  | Equal  | Equal  | $\lambda_g \Gamma \Delta \Gamma'$ | $q(q + 1)/2 + (G - 1)$ |
| EVE   | Equal  | Variable | Equal  | $\lambda \Gamma \Delta \Gamma'$ | $(q + 1)/2 + (G - 1)(q - 1)$ |
| EEV   | Equal  | Variable | Equal  | $\lambda_g \Delta \Gamma'$ | $Gq(q + 1)/2 - (G - 1)q$ |
| VEE   | Equal  | Variable | Equal  | $\lambda_g \Gamma \Delta \Gamma'$ | $q(q + 1)/2 + (G - 1)q$ |
| VEV   | Equal  | Variable | Variable | $\lambda_g \Gamma \Delta \Gamma'$ | $Gq(q + 1)/2 - (G - 1)q$ |
| EEV   | Equal  | Equal  | Variable | $\lambda \Gamma \Delta \Gamma'$ | $q(q + 1)/2 + (G - 1)(q - 1)$ |
| VVV   | Variable | Variable | Variable | $\lambda_g \Gamma \Delta \Gamma'$ | $Gq(q + 1)/2 - (G - 1)$ |

Here, the covariance matrices $\Sigma_{X_g}$ and $\Sigma_{Y_g}$ in (2) can be decomposed. Constraining $\lambda_g$, $\Gamma_g$, and $\Delta_g$ on these decompositions in Equation (2) leads to 14 different covariance parameters.
structures for both $X$ and $Y$, resulting in a total of $14 \times 14 = 196$ models. This will be referred to as the eMCWM family in the sequel. Note that for the purposes of notation, an eMCWM with a VEV covariance structure for $Y \mid X = x$ and a EII covariance structure for $X$ will be denoted as a VEV-EII model.

4 Identifiability

Identifiability is important for parameter inference and for the usual asymptotic theory to hold for maximum likelihood estimation of the model parameters (cf. Section 5). Identifiability of univariate and multivariate finite Gaussian mixture distributions has been proved in [Teicher (1963) and Yakowitz and Spragins (1968)], respectively, while general conditions for identifiability of mixtures of linear models can be found in [Hennig (2000)]. Identifiability for the generalized linear Gaussian CWM has recently been proved in [Ingrassia et al. (2014)]. Here, identification conditions are provided for the multivariate response (Gaussian) CWM defined in (2).

Generally speaking, identifiability for mixture models can be defined as follows. Consider a parametric class of density (probability) functions $F = \{ f(z \mid \psi) : z \in \mathcal{Z}, \psi \in \Psi \}$ and then the class of finite mixtures of functions in $F$,

$$
H = \left\{ h(z \mid \vartheta) : h(z \mid \vartheta) = \sum_{g=1}^{G} f(z \mid \psi_g) \pi_g, \text{ with } \pi_g > 0 \text{ and } \sum_{g=1}^{G} \pi_g = 1, \right. $$

$$
\left. f(\cdot \mid \psi_g) \in F, g = 1, \ldots, G, \psi_g \neq \psi_j \text{ for } g \neq j, G \in \mathbb{N}, z \in \mathcal{Z}, \vartheta \in \Theta \right\}.
$$

This class is identifiable if given two members

$$
h(z \mid \vartheta) = \sum_{g=1}^{G} f(z \mid \psi_g) \pi_g \quad \text{and} \quad h(z \mid \tilde{\vartheta}) = \sum_{s=1}^{\tilde{G}} f(z \mid \tilde{\psi}_s) \tilde{\pi}_s
$$

of $H$, then $h(z \mid \vartheta) = h(z \mid \tilde{\vartheta})$ implies that $G = \tilde{G}$ and for each $g \in \{1, \ldots, G\}$ there exists $s \in \{1, \ldots, \tilde{G}\}$ such that $\pi_g = \tilde{\pi}_s$ and $\psi_g = \tilde{\psi}_s$.

In Theorem 1, a sufficient identification condition is provided for the most general eMCWM (i.e., the VVV-VVV model). In particular, a sufficient condition for the identifiability
of the class is established

$$\mathcal{C} = \left\{ p(u, y \mid q) : p(u, y \mid q) = \sum_{g=1}^{G} \phi_d \left( \left. y \right| x, \mu_Y (x \mid \beta_g), \Sigma_Y \right) \phi_p \left( \left. x \right| \mu_X, \Sigma_X \right) \pi_g, \right\}$$

with $$\pi_g > 0, \sum_{g=1}^{G} \pi_g = 1, (\beta_g, \Sigma_Y) \neq (\beta_j, \Sigma_Y)$$ for $$g \neq j, (x, y) \in \mathbb{R}^{p+d},$$

$$\vartheta = \left\{ \beta_g, \Sigma_Y, \mu_X, \pi_g; g = 1, \ldots, G \right\} \in \Theta, G \in \mathbb{N}. \quad (3)$$

In the following theorem, sufficient conditions for $$\mathcal{C}$$ to be identifiable in $$\mathcal{X} \times \mathbb{R}^d$$ are provided, where $$\mathcal{X} \subseteq \mathbb{R}^p$$ is a set having probability one according to the multivariate Gaussian density $$\phi_p$$. In other words, it is proved that the class $$\mathcal{C}$$ is identifiable for almost all $$x \in \mathbb{R}^p$$ and for all $$y \in \mathbb{R}^d$$.

**Theorem 1.** Let $$\mathcal{C}$$ be the class defined in (3) and assume that there exists a set $$\mathcal{X} \subseteq \mathbb{R}^p$$ having probability equal to one according to the $$p$$-variate Gaussian distribution such that the mixture of regression models

$$\sum_{g=1}^{G} \phi_d \left( \left. y \right| x, \mu_Y (x \mid \beta_g), \Sigma_Y \right) \alpha_g (x), \quad y \in \mathbb{R}^d, \quad \alpha_1 (x), \ldots, \alpha_G (x)$$

is identifiable for each fixed $$x \in \mathcal{X}$$, where $$\alpha_1 (x), \ldots, \alpha_G (x)$$ are positive weights summing to one for each $$x \in \mathcal{X}$$. Then the class $$\mathcal{C}$$ is identifiable in $$\mathcal{X} \times \mathbb{R}^d$$.

**Proof.** The proof is given in Appendix A.

---

5 *Inference*

5.1 Parameter Estimation for eMCWM

Parameter estimation, via the EM algorithm of Dempster et al. (1977), is described here for the unconstrained VVV-VVV model from the eMCWM family. Let $$\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$$ be a sample of $$N$$ independent observations from model (2). Then, the incomplete-data likelihood function is

$$L (\vartheta \mid \mathcal{S}) = \prod_{i=1}^{N} p(x_i, y_i \mid q) = \prod_{i=1}^{N} \left[ \sum_{g=1}^{G} \phi_d \left( \left. y_i \right| \mu_Y (x_i, \beta_g), \Sigma_Y \right) \phi_p \left( \left. x_i \right| \mu_X, \Sigma_X \right) \pi_g \right].$$

Note that $$\mathcal{S}$$ is considered incomplete in the context of the EM algorithm. The complete-data are $$\mathcal{S}_c = \{(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)\},$$ where the missing variable $$z_i$$ is the component
Now, the corresponding complete-data likelihood is

\[ L_c(\vartheta|S_c) = \prod_{i=1}^{N} \prod_{g=1}^{G} \phi_d(y_i|\mu_Y(x_i, \beta_g), \Sigma_{Yg}) \phi_p(x_i|\mu_{Xg}, \Sigma_{Xg}) \pi_g]^{z_{ig}}. \]

The complete-data log-likelihood function can be decomposed as

\[ l_c(\vartheta|S_c) = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \left[ \log \phi_d(y_i|\mu_Y(x_i, \beta_g), \Sigma_{Yg}) + \log \phi_p(x_i|\mu_{Xg}, \Sigma_{Xg}) + \log \pi_g \right]. \]

The E-step involves calculating the expected complete-data log-likelihood

\[ Q(\vartheta|\vartheta^{(k)}) = \mathbb{E}_{\vartheta^{(k)}} \{ l_c(\vartheta|S_c) \} \]

\[ = \sum_{i=1}^{N} \sum_{g=1}^{G} \tau_{ig}^{(k)} \left[ Q_1(\beta_g, \Sigma_{Yg}|\vartheta^{(k)}) + Q_2(\mu_{Xg}, \Sigma_{Xg}|\vartheta^{(k)}) + \log \pi_g \right], \]

where

\[ \tau_{ig}^{(k)} = \mathbb{E}_{\vartheta^{(k)}} \{ Z_{ig}|x_i, y_i \} = \frac{\phi_d(y_i|\mu_Y(x_i|\beta_g^{(k)}), \Sigma_{Yg}) \phi_p(x_i|\mu_{Xg}, \Sigma_{Xg}) \pi_g}{\sum_{j=1}^{G} \phi_d(y_i|\mu_Y(x_i|\beta_j^{(k)}), \Sigma_{Yj}) \phi_p(x_i|\mu_{Xj}, \Sigma_{Xj}) \pi_j^{(k)}}, \]

provides the current value of \( z_{ig} \) on the \( k \)-th iteration and

\[ Q_1(\beta_g, \Sigma_{Yg}|\vartheta^{(k)}) = \frac{1}{2} \left[ -d \log (2\pi) - \log |\Sigma_{Yg}| - (y_i - \beta_g' x_i')' \Sigma_{Yg}^{-1}(y_i - \beta_g' x_i') \right], \]

\[ Q_2(\mu_{Xg}, \Sigma_{Xg}|\vartheta^{(k)}) = \frac{1}{2} \left[ -p \log (2\pi) - \log |\Sigma_{Xg}| - (x_i - \mu_{Xg})' \Sigma_{Xg}^{-1}(x_i - \mu_{Xg}) \right]. \]

The M-step on the \( (k+1) \)-th iteration of the EM algorithm involves the maximization of the conditional expectation of the complete-data log-likelihood with respect to \( \vartheta \). The update for \( \pi_g^{(k+1)} \) is

\[ \hat{\pi}_g^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} \tau_{ig}^{(k)}. \]  

The updates for \( \mu_{Xg}^{(k+1)} \) and \( \Sigma_{Xg}^{(k+1)} \), \( g = 1, \ldots, G \), are

\[ \hat{\mu}_{Xg}^{(k+1)} = \frac{\sum_{i=1}^{N} \tau_{ig}^{(k)} x_i}{\sum_{i=1}^{N} \tau_{ig}^{(k)}}, \]

\[ \hat{\Sigma}_{Xg}^{(k+1)} = \frac{\sum_{i=1}^{N} \tau_{ig}^{(k)} (x_i - \hat{\mu}_{Xg}^{(k+1)})' (x_i - \hat{\mu}_{Xg}^{(k+1)})}{\sum_{i=1}^{N} \tau_{ig}^{(k)}}. \]
These closed form updates can also be found in McLachlan and Peel (2000). The updates for $\beta_{g}^{(k+1)}$ and $\Sigma_{Yg}^{(k+1)}$ (see Appendix B for details), $g = 1, \ldots, G$, are

$$\hat{\beta}_{g}^{(k+1)'} = \left( \sum_{i=1}^{N} \tau_{ig}^{(k)} y_{i} x_{i}' \right) \left( \sum_{i=1}^{N} \tau_{ig}^{(k)} x_{i} x_{i}' \right)^{-1}$$ \hfill (8)

and

$$\hat{\Sigma}_{Yg}^{(k+1)} = \sum_{i=1}^{N} x_{i} \tau_{ig}^{(k)} \left( y_{i} - \hat{\beta}_{g}^{'} x_{i} \right) \left( y_{i} - \hat{\beta}_{g}^{'} x_{i} \right)' \sum_{i=1}^{N} x_{i} \tau_{ig}^{(k)}.$$ \hfill (9)

Equations (5) through (9) are the parameter updates for the VVV-FFF model of the eMCWM family. For the other models of this family, the M-step updates vary only with respect to the component covariance matrices $\Sigma_{Xg}$ and $\Sigma_{Yg}$; these updates are similar to those of the GPCM family of Celeux and Govaert (1995).

5.2 Model Selection

For choosing the “best” fitted model among a family of models, a likelihood-based model selection criteria is conventionally used, and for Gaussian mixture models, the BIC is the most popular. Even though mixture models generally do not satisfy the regularity conditions for the asymptotic approximation used in the development of the BIC (Keribin, 1998, 2000), it has performed quite well in practice and has been used extensively (Dasgupta and Raftery, 1998; Fraley and Raftery, 2002). The BIC can be calculated as

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

where $l(\hat{\vartheta})$ is the incomplete-data log-likelihood at the maximum likelihood estimates and $m$ is the number of free parameters. The ICL is another commonly used information criterion which additionally makes use of the estimated mean entropy, i.e., it takes into account the uncertainty of the classification of an observation to a component and can be computed as

$$\text{ICL} \approx \text{BIC} + \sum_{i=1}^{N} \sum_{g=1}^{G} \text{MAP}(\hat{z}_{ig}) \log \hat{z}_{ig}.$$  

Here, $\text{MAP}(\hat{z}_{ig})$ is the maximum a posteriori probability and equals 1 if $\max_{h}(\hat{z}_{ih})$, $h = 1, \ldots, G$, occurs at component $g$, and 0 otherwise.

5.3 Initialization

The EM algorithm has been noted to be dependent on starting values. Singularities and convergence to local maxima are well documented, and Gaussian mixture models also have
unbounded likelihood surfaces (Titterington et al., 1985). Constraining eigenvalues can alleviate some of these issues (Ingrassia and Rocci, 2007; Browne et al., 2013). Alternatively, deterministic annealing (Zhou and Lange, 2010) can be employed to mitigate such issues. Initializing the EM algorithm multiple times using \( k \)-means (Hartigan and Wong, 1979) or random initializations and choosing the initial values of \( z_{ig} \) from the run picked by BIC can also help.

Here, the EM algorithm is initialized as follows. The EEE-EEE model is run 10 times for each \( G \): 9 times using a random initialization for the \( z_{ig} \), and once with a \( k \)-means initialization. From these models, the model with the highest BIC value is chosen; then, the associated MAP(\( \hat{z}_{ig} \)) is used to initialize the families of models. In our simulations, this procedure performed quite well.

### 5.4 Convergence Criterion

A common criterion to stop the EM algorithm is when the difference between the log-likelihood values on consecutive iterations is less than some \( \epsilon \). However, use of such lack of progress criteria might underestimate the correct value of the log-likelihood (McNicholas et al., 2010). Here, the Aitken stopping criterion (Aitken, 1926) is used to determine convergence. Basically, the Aitken acceleration procedure is used to compute an estimated asymptotic value. Based on this, a decision can be made regarding whether or not the algorithm has reached convergence, that is, whether or not the log-likelihood is sufficiently close to its estimated asymptotic value. See Böning et al. (1994), Lindsay (1995), and McNicholas et al. (2010) for details.

### 5.5 Performance Assessment

The adjusted Rand index (ARI; Hubert and Arabie, 1985) can be used to judge performance of a model relative to the true classification (when known). The ARI calculates the agreement between true and estimated classification by correcting the Rand index (Rand, 1971) to account for chance. Hence, an ARI of 1 corresponds to perfect clustering whereas an ARI of 0 implies that the results are no better than would be obtained by chance alone.

### 6 Experiments and Illustrations

The eMCWM family is implemented in R. Although the objective was to develop a mixture model that can incorporate linear dependencies on covariates, the models’ performance is also compared to the Gaussian parsimonious clustering models (GPCMs) family of Celeux and Govaert (1995). Note that GPCMs are constructed on different assumptions and are not meant to account for regression structures. However, GPCMs remain the most commonly used models in the literature for model-based clustering, and hence, a comparison to these models is relevant. As mentioned earlier, these models are supported by the mclust, mixture
and Rmixmod packages. In particular, mixture contains all 14 GPCMs and makes use of a majorization-minimization algorithm for the EVE and VVE models, which works better in higher dimensions than the Flury procedure used in Rmixmod (Browne and McNicholas, 2014). Note that for the M-step for the different covariance structures in Table 1, the mixture package is used. The eMCWM family is also compared to the eFMR and eFMRC families of Dang and McNicholas (2013). The flexmix FMR and FMRC algorithms allow for specification of a user-defined initialization matrix, and so, to facilitate comparison of the performance of the algorithms from the same starting values, all algorithms run on a specific data sample are initialized with the same set of MAP(\hat{\tau}_{ig}) values (Section 5.3). In Sections 6.1 and 6.2 some analyses respectively based on artificial and real data sets are presented.

6.1 Analysis on Simulated Data

Data, of size $N = 250$ and with $p = d = 2$, are generated from a two-component eMCWM (Figure 1). The data set will be referred to as Dataset1 hereafter. The component sizes are generated using a Bernoulli distribution with probability $\pi_1 = 0.35$. The responses are generated using a VEE covariance structure and the covariates using a VII covariance structure. Covariates are generated from a bivariate Gaussian distribution with mean $\boldsymbol{\mu}_{X1} = (3, 2.5)$ and $\boldsymbol{\mu}_{X2} = (1.1, -4)$ for components 1 and 2, respectively. The covariance matrices of the covariates for the two groups are $egin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $egin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, respectively. Under the VII decomposition, this corresponds to $\lambda_{X1} = 1$ and $\lambda_{X2} = 0.5$ for component 1 and component 2, respectively. The regression coefficient matrices used for the two groups are $egin{pmatrix} 2 & -0.5 & -2 & 1 \\ -1.5 & 2 \end{pmatrix}$ and $egin{pmatrix} 0 & 2 \\ -1 & 1 \end{pmatrix}$, respectively. Lastly, the error matrices for the two groups using a VEE covariance structure are $egin{pmatrix} 0.92 & 0.56 & 0.06 & 1.40 \\ 0.56 & 0.06 & 1.40 & 0.06 \end{pmatrix}$ and $egin{pmatrix} 1.725 & 1.050 \\ 1.050 & 2.625 \end{pmatrix}$, respectively. This corresponds to $\lambda_{Y1} = 0.8$, $\lambda_{Y2} = 1.5$, and $\Gamma_{Y1}\Delta_{Y1}\Gamma_{Y1}' = \Gamma_{Y2}\Delta_{Y2}\Gamma_{Y2}' = \begin{pmatrix} 1.15 & 0.70 \\ 0.70 & 1.75 \end{pmatrix}$. Due to the higher computational requirements of this family, 50 samples are generated. The eMCWM family is run using all 196 models for $G \in \{1, \ldots, 4\}$ resulting in a total of 784 models for each sample. Both the BIC and the ICL chose the same model each time. The VEE-VII model is chosen 43 out of 50 times. All 50 times, the selected model fit the correct number of components and resulted in perfect classification. The difference in BIC values between the chosen model and the VEE-VII model ranged between 0.16 and 7.42 with a median value of 0.70. Table 2 shows the estimated mean and range (in parentheses) of the estimated parameters for each component for the VEE-VII models. Clearly, the parameter estimates are quite close to the true values.

6.2 Analysis of Real Data Sets

6.2.1 Australian Institute of Sports Data

The Australian Institute of Sports (AIS) data (Cook and Weisberg, 1994) contains measurements on 202 athletes (100 female and 102 male) and is available in the R package sn.
A subset of 7 variables that has recently been used in the mixtures of regression literature (Soffritti and Galimberti, 2011) is analyzed here: red cell count (RCC), white cell count (WCC), plasma ferritin concentration (PFC), body mass index, sum of skin folds, body fat percentage, and lean body mass. The blood composition variables (RCC, WCC, and PFC) are selected as the response variables with the biometrical variables being the predictors. All algorithms are run for $G \in \{1, \ldots, 4\}$. Table 3 summarizes the results from running the eMCWM, eFMR, eFMRC, FMR, FMRC, and mixture algorithms.

Both the BIC and the ICL chose a two-component model with the VVI and VVE covariance structures for the response and covariates, respectively. This model yielded an ARI of 0.92. The estimated classification for the chosen eMCWM model is in Table 4. Note that the difference between the chosen eMCWM model and the model with the second best BIC value is quite small ($\approx 0.3$ BIC points). This latter model (VEI-VVE; 57 degrees of freedom) picked 2 components and resulted in an ARI of 0.87. The chosen (1 component) model from the eFMR and eFMRC families; and the FMR and FMRC models did not perform well. The mixture software, on the other hand, selected a three-component solution but the estimated classification does not reconcile as well with the known grouping.

### 6.2.2 Iris Data

The algorithms are also run on the famous Iris data set (Anderson, 1935; Fisher, 1936). This data set, available in the datasets package as part of R, provides measurements on sepal length and width, and petal length and width for 50 flowers each from 3 species of Iris: setosa, versicolor, and virginica. The width measurements are taken to be the
Table 2: Parameter estimates (rounded off to 2 decimals) from 50 samples for the VEE-VII model for Dataset1.

| Parameter Estimate (range) |
|-----------------------------|
| $\pi_1$ 0.35 (0.30, 0.40) |
| $\pi_2$ 0.65 (0.60, 0.70) |
| $\mu_{X1}$ 3.00 (2.74, 3.27); 2.49 (2.25, 2.79) |
| $\mu_{X2}$ 1.10 (1.01, 1.22); -4.01 (-4.10, -3.90) |
| $\lambda_{X1}$ 0.98 (0.81, 1.20) |
| $\lambda_{X2}$ 0.51 (0.44, 0.60) |
| $\beta_1'$ $\begin{pmatrix} 1.97(0.82, 3.23) & -0.51(-0.80, -0.25) & -0.98(-1.22, -0.80) \\ -1.98(-3.14, -0.71) & 1.49(1.10, 1.81) & 2.00(1.78, 2.25) \end{pmatrix}$ |
| $\beta_2'$ $\begin{pmatrix} -0.09(-1.30, 1.28) & 2.22(1.89, 2.47) & -1.01(-1.31, -0.63) \\ 1.00(-0.43, 2.71) & 2.00(1.61, 2.37) & 1.51(1.13, 1.94) \end{pmatrix}$ |
| $\Sigma_{Y1}$ $\begin{pmatrix} 0.91(0.66, 1.13) & 0.55(0.33, 0.76) \\ 0.55(0.33, 0.76) & 1.39(1.01, 1.84) \end{pmatrix}$ |
| $\Sigma_{Y2}$ $\begin{pmatrix} 1.66(1.37, 2.01) & 1.01(0.72, 1.26) \\ 1.01(0.72, 1.26) & 2.55(2.00, 3.11) \end{pmatrix}$ |

response variables with the other variables as the covariates. The algorithms are run for $G \in \{1, \ldots, 4\}$ (Table 5). The selected eMCWM model is a 3 component model with an ARI of 0.90 (Table 6). Note that the difference between the chosen eMCWM model and the model with the second best BIC value is $\approx 1.25$ BIC points. This latter model selected a 2 component model (VVV-VVV; 29 degrees of freedom) with an ARI of 0.57. This model put together datapoints from versicolor, and virginica in one group. The model with the second best BIC value is the model that the ICL chose: a 2 component VVV model with an ARI of 0.53.

The eFMR and FMR models resulted in two-component models yielding poor ARI values. A two-component VVI model is selected from the eFMRC family with an ARI of 0.45. This model clusters setosa and virginica perfectly with observations from versicolor assigned to the other two clusters. Because the flexmix FMRC algorithm is, in essence, a VVI model, unsurprisingly, it also chose a two-component model with an ARI of 0.45. The GPCM family as implemented in mixture picked a 2 component model (ARI=0.57) with the data points from versicolor and virginica pooled together in one group (Table 6).

6.2.3 Crabs Data

The crabs data set contains five morphological measurements on 50 crabs each, of both sexes and colours (blue and orange) of the species Leptograpsus variegatus. These data were originally introduced in Campbell and Mahon (1974) and are available as part of the MASS
Table 3: Comparison of the performance of the models that are applied to the AIS data.

| Algorithm | Model | G | ARI | Parameters |
|-----------|-------|---|-----|------------|
| eMCWM     | VVI-VVE | 2 | 0.92 | 59         |
| eFMR      | VI    | 1 | 0.92 | 18         |
| eFMRC     | VI    | 1 | 0.92 | 18         |
| FMR       |       | 1 | 0.92 | 18         |
| FMRC      |       | 1 | 0.92 | 18         |
| mixture   | EVE   | 3 | 0.60 | 63         |

*Note that for a one-component model, the family comprises three covariance structures: VI (diagonal with different entries), EI (diagonal with same entries), and VV (full covariance matrix).

Table 4: Cross-tabulation of true and estimated classifications for two methods applied to the AIS data.

|          | eMCWM | mixture |
|----------|-------|---------|
|          | 1     | 2       |
|          |       |         |
| Female   | 99    | 1       |
|          |       |         |
| Male     | 3     | 99      |
|          |       |         |
|          | 83    | 14      |
|          | 3     |         |

The chosen eFMR model is a 2 component VVI model with an ARI of 0.40. Because the VVI model assumes independence between the response variables, that is equivalent to the `flexmix` FMR model and, unsurprisingly, the chosen FMR model is a two-component model with an ARI of 0.40 (Table 8). Note that the estimated classification from the selected 2 component eFMR model leads to good separation between the sexes of the crabs. If the class membership agreement is estimated based on only the sexes of the crabs, an ARI of 0.81 is achieved. The selected FMRC model fit three components (Table 8) with an
Table 5: Comparison of the performance of the models that are applied to the Iris data.

| Algorithm | Model   | $G$ | ARI | Parameters |
|-----------|---------|-----|-----|------------|
| eMCWM     | VEV-VEV | 3   | 0.90| 40         |
| eFMR      | VEI     | 2   | 0.14| 16         |
| eFMRC     | VVI     | 2   | 0.45| 19         |
| FMR       |         | 2   | 0.19| 17         |
| FMRC      |         | 2   | 0.45| 19         |
| mixture   | VEV     | 2   | 0.57| 26         |

Table 6: Cross-tabulation of true and estimated classifications for three methods applied to the Iris data.

|               | eMCWM | mixture | eFMRC |
|---------------|-------|---------|-------|
| setosa        | 1     | 50      | 50    |
| versicolor    | 2     | 50      | 31    |
| virginica     | 3     | 50      | 19    |

ARI of 0.69. It basically pools together the blue males and blue females while putting the orange males and females in different clusters. eFMRC picked a four-component model with an ARI of 0.83. mixture picked a four-component model (Table 8) with an ARI of 0.78. Note that even though the ARIs achieved from the selected eMCWM, eFMRC, and mixture models are close, the mixture model estimates many more parameters than the models that utilize linear dependencies between variables. Also, note that the performance of the eFMRC model should not be surprising. Recall that in an FMRC model, $\pi_g(x)$ are modelled by a multinomial logit model (DeSarbo and Cron, 1988). Anderson (1972) noted that this multinomial condition is satisfied if the covariate densities $p(x)$ are assumed to be multivariate Gaussian with the same covariance matrices, i.e., the EEE model for the covariates (cf. Ingrassia et al., 2012). Hence, the eFMRC EEE model should give similar clustering results as the eMCWM EEE-EEE model. As pointed out above, these models chose a four-component models with ARI values of 0.83 and 0.84, respectively.

7 Discussion

A novel family of CWMs that can account for heterogeneous regression data with multivariate correlated response is presented. The distribution of the covariates is also explicitly incorporated in the likelihood, which to the authors’ knowledge is also novel in multivari-
Table 7: Comparison of the performance of the models that are applied to the crabs data.

| Algorithm | Model   | $G$ | ARI | Parameters |
|-----------|---------|-----|-----|------------|
| eMCWM     | EEE-EVE | 4   | 0.82| 59         |
| eFMR      | VVI     | 2   | 0.40| 25         |
| eFMRC     | EEE     | 4   | 0.83| 51         |
| FMRC      | 2       | 0.40|     | 25         |
| FMRC      | 3       | 0.69|     | 42         |
| mixture   | EEV     | 4   | 0.78| 68         |

Table 8: Cross-tabulation of true and estimated classifications for four methods applied to the crabs data.

|          | eMCWM | eFMRC | mixture | eFMR |
|----------|-------|-------|---------|------|
|          | 1 2 3 4 | 1 2 3 4 | 1 2 3 4 | 1 2 |
| BM       | 39 11 | 39 11 | 38 12  | 46 4 |
| BF       | 50    | 50    | 49 1    | 4 46 |
| OM       | 50    | 50    | 50      | 50   |
| OF       | 4 46  | 3 47  | 5 45    | 2 48 |

“B”, “O”, “M”, and “F” refer to blue, orange, male, and female, respectively.

ate response regression methodologies. This allows for imposition of an eigen-decomposed structure separately on both the responses and the covariates’ component covariance matrices. Hence, eMCWM can handle data where $X$ and $Y|x$ might have different covariance structures. For this family, identification conditions are also provided.

The eMCWM family is quite parsimonious. Note that the completely unconstrained GPCM model (VVV) fits the same number of parameters as the completely unconstrained eMCWM model (VVV-VVV). When the eMCWM is applied to both simulated and real data, the BIC and the ICL are in agreement with each other for all but the iris data. For the iris data, the ICL picked a 2 component VVV-VVV model. The family’s performance is investigated on simulated and real benchmark data. More generally, in comparison to the FMR and FMRC models, eMCWM performed better because it explicitly used the distribution of the covariates. Not using that information lead to estimated clusters (from the eFMR and eFMRC, and the flexmix FMR and FMRC models) that did not agree with the observed grouping of the data. In comparison to the GPCMs, taking into account the regression structure by use of linear dependencies aided in better clustering performance on some benchmark data sets. As implemented in flexmix, FMR only models the distribution of the (assumed independent) $Y|x$, while FMRC models both the distribution of (assumed
independent) \( Y|x \) and a logistic model of the covariates, respectively. GPCMs and their ilk (e.g., [Andrews and McNicholas, 2012] [Vrbik and McNicholas, 2014]) do not account for linear dependencies and rely on modeling the data directly with an appropriate distribution. However, the eMCWM family models both linear dependencies and the distribution of the covariates. This results in better clustering performance when there is a clear regression relationship between the variables. Numerically, the EM algorithm is quite stable. However, to prevent fitting issues for the eMCWM family, the component sizes are computed before each M-step and a preset minimum size of the clusters is used (cf. Celeux and Diebolt, 1988 Grün et al., 2008).

The framework presented lends itself to a straightforward extension to model-based classification (e.g., McNicholas, 2010; Andrews et al., 2011) and discriminant analysis (Hastie and Tibshirani, 1996). In the case of the eMCWM family, currently, each response variable is regressed individually on a common set of predictor variables. This can be extended to take advantage of correlations between the response variables to improve predictive accuracy, in the fashion of the ‘curds and whey’ method (Breiman and Friedman, 1997). Here, the Gaussian distribution is used for both the distribution of the covariates and the response for the eMCWM family. For heavier tailed data, more robust distributions like the multivariate \( t \) distribution may be employed. The use of continuous distributions may be restrictive and more work needs to be done on incorporating mixed type data for both response and covariates.

Acknowledgements

This work is supported by an Alexander Graham Bell Canada Graduate Scholarship and Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Conflict of Interest
The authors have declared no conflict of interest.

Appendix

A Proof of Theorem 1

Proof. The proof builds upon results given in [Hennig, 2000] and [Ingrassia et al., 2014] . Consider the class of models defined in (3) and prove that the equality

\[
\sum_{g=1}^{G} \phi_d(y|x, \mu_Y(x|\beta_g), \Sigma_{Yg})\phi_p(x|\mu_{Xg}, \Sigma_{Xg})\pi_g = \sum_{s=1}^{\tilde{G}} \phi_d(y|x, \mu_Y(x|\tilde{\beta}_s), \tilde{\Sigma}_{Ys})\phi_p(x|\tilde{\mu}_{Xs}, \tilde{\Sigma}_{Xs})\tilde{\pi}_s
\]

(10)
holds for almost all \( x \in \mathbb{R}^p \) and for all \( y \in \mathbb{R}^d \) if and only if \( G = \tilde{G} \) and for each \( g \in \{1, \ldots, G\} \) there exists \( s \in \{1, \ldots, G\} \) such that \( \beta_g = \tilde{\beta}_s \), \( \lambda_g = \tilde{\lambda}_s \), \( \mu_g = \tilde{\mu}_s \), \( \Sigma_g = \tilde{\Sigma}_s \) and \( \pi_g = \tilde{\pi}_s \).

Integrating each side of (10) over \( \mathbb{R}^d \) yields

\[
\sum_{g=1}^{G} \phi_p(x|\mu_{Xg}, \Sigma_{Xg}) \pi_g = \sum_{s=1}^{\tilde{G}} \phi_p(x|\tilde{\mu}_s, \tilde{\Sigma}_s) \tilde{\pi}_s
\]

(11)

Let us set

\[
p(x|\mu_X, \Sigma_X, \pi) = \sum_{g=1}^{G} \phi_p(x|\mu_{Xg}, \Sigma_{Xg}) \pi_g \quad \text{and} \quad p(u|\tilde{\mu}, \tilde{\Sigma}) = \sum_{s=1}^{\tilde{G}} \phi_p(x|\tilde{\mu}_s, \tilde{\Sigma}_s) \tilde{\pi}_s,
\]

where \( \mu_X = \{\mu_{Xg}; \ g = 1, \ldots, G\} \), \( \Sigma_X = \{\Sigma_{Xg}; \ g = 1, \ldots, G\} \) and \( \pi = \{\pi_g; \ g = 1, \ldots, G\} \). Analogous notation applies for \( \tilde{\mu}_X, \tilde{\Sigma}_X \) and \( \tilde{\pi} \). Afterwards, based on Bayes’ theorem,

\[
p(\Omega_g|x, \mu_X, \Sigma_X, \pi) = \frac{\phi(x|\mu_{Xg}, \Sigma_{Xg}) \pi_g}{p(x|\mu_X, \Sigma_X, \pi)}, \quad g = 1, \ldots, G
\]

(12)

this allows to rewrite model (2) as

\[
p(x, y|\vartheta) = p(x|\mu_X, \Sigma_X, \pi) \sum_{g=1}^{G} \phi_d(y|x, \mu_Y(x|\beta_g), \Sigma_{Yg}) p(\Omega_g|x, \mu_X, \Sigma_X, \pi)
\]

\[
= p(u|\mu_X, \Sigma_X, \pi) p(y|x, \vartheta),
\]

(13)

where

\[
p(y|x, \vartheta) = \sum_{g=1}^{G} \phi_d(y|x, \mu_Y(x|\beta_g), \Sigma_{Yg}) p(\Omega_g|x, \mu_X, \Sigma_X, \pi), \quad y \in \mathbb{R}^d.
\]

(14)

Now, the class of models defined by (14) for almost all \( x \in \mathbb{R}^p \), if the equality

\[
\sum_{g=1}^{G} \phi_d(y|x, \mu_Y(x|\beta_g), \Sigma_{Yg}) p(\Omega_g|u, \mu_X, \Sigma_X, \pi) = \sum_{s=1}^{\tilde{G}} \phi_d(y|x, \mu_Y(x|\tilde{\beta}_s), \tilde{\Sigma}_{Ys}) p(\Omega_s|u, \tilde{\mu}_s, \tilde{\Sigma}_X, \tilde{\pi})
\]

implies \( G = \tilde{G} \) and for each \( g \in \{1, \ldots, G\} \) there exists \( s \in \{1, \ldots, G\} \) such that \( \beta_g = \tilde{\beta}_s \), \( \Sigma_{Yg} = \tilde{\Sigma}_{Ys} \), \( \mu_{Xg} = \tilde{\mu}_s \), \( \Sigma_{Xg} = \tilde{\Sigma}_s \) and \( \pi_g = \tilde{\pi}_s \).

Recall from Section 2 that the expected value \( \mu_Y \) of \( Y|\Omega_g \) is related to the covariates \( X \) through the relation \( \mu_Y = \beta_{0g} + \beta'_{1g} x \), \( g = 1, \ldots, G \). Let us introduce the set

\[
\mathcal{X} = \{ x \in \mathbb{R}^p : \text{for each } g, j \in \{1, \ldots, G\}, \text{ and } s, t \in \{1, \ldots, \tilde{G}\} : \beta'_{g} x^* = \beta'_{j} x^* \Rightarrow \beta_g = \beta_j, \tilde{\beta}'_{s} x^* = \tilde{\beta}'_{t} x^* \Rightarrow \beta_g = \tilde{\beta}_s, \tilde{\beta}'_{s} x^* = \tilde{\beta}'_{t} x^* \Rightarrow \beta_{g} = \tilde{\beta}_t \}.
\]
According to (3), \((\beta_g, \Sigma_{Yg}) \neq (\beta_j, \Sigma_{Yj}), g \neq j\); thus, it follows that the quantities \((\beta'_g x^*, \Sigma_{Yg})\), 
g = 1, \ldots, G, are pairwise distinct for all \(x \in \mathcal{X}\) (indeed, the complement of \(\mathcal{X}\), i.e. \(\mathbb{R}^p \setminus \mathcal{X}\), is formed by a finite set of hyperplanes of \(\mathbb{R}^p\) and thus \(\mathbb{R}^p \setminus \mathcal{X}\) has null measure).

For any fixed \(x \in \mathcal{X}\), according to (12), \(\{p(\Omega_1| x, \mu_X, \Sigma_X, \pi), \ldots, p(\Omega_G| x, \mu_X, \Sigma_X, \pi)\}\) and \(\{p(\Omega_1| x, \tilde{\mu}_X, \tilde{\Sigma}_X, \tilde{\pi}), \ldots, p(\Omega_G| x, \tilde{\mu}_X, \tilde{\Sigma}_X, \tilde{\pi})\}\) are sets of positive numbers summing to one. It follows that, for each \(x \in \mathcal{X}\), the density \(p(y| u, \vartheta)\) given in (14) is a mixture of distributions of kind (4) and then it is identifiable, due to the assumptions of the theorem. Thus \(G = \tilde{G}\) and there exists \(s \in \{1, \ldots, G\}\) such that

\[
\beta_g = \tilde{\beta}_s, \quad \Sigma_{Yg} = \tilde{\Sigma}_Y, \quad \text{and} \quad p(\Omega_g| x, \mu_X, \Sigma_X, \pi) = p(\Omega_s| x, \tilde{\mu}_X, \tilde{\Sigma}_X, \tilde{\pi}).
\]

Moreover, since \(p(\Omega_g| x, \mu_X, \Sigma_X, \pi)\) and \(p(\Omega_s| x, \tilde{\mu}_X, \tilde{\Sigma}_X, \tilde{\pi})\) are defined according to (12), from (15) and (11), we get:

\[
\pi_g = \int_{\mathcal{X}} \pi_g \phi_p(x| \mu_{Xg}, \Sigma_{Xg}) dx = \int_{\mathcal{X}} \left( \sum_{g=1}^{G} \phi_p(x| \mu_{Xg}, \Sigma_{Xg}) \right) \pi_g dx = \int_{\mathcal{X}} p(\Omega_g| x, \mu_X, \Sigma_X, \pi) \left( \sum_{g=1}^{G} \phi_p(x| \mu_{Xg}, \Sigma_{Xg}) \pi_g \right) dx = \pi_s.
\]

Thus, for the same pair \((g, s)\) in (15), it results \(\pi_g = \pi_s\). Finally, we get

\[
\phi_p(x| \mu_{Xg}, \Sigma_g) = \frac{p(\Omega_g| x, \mu_{Xg}, \Sigma_{Xg}, \pi)}{\pi_g} \sum_{g=1}^{G} \phi_p(x| \mu_{Xg}, \Sigma_{Xg}) \pi_g = \frac{p(\Omega_s| x, \tilde{\mu}_X, \tilde{\Sigma}_X, \tilde{\pi})}{\pi_s} \sum_{s=1}^{\tilde{G}} \phi(u| \tilde{\mu}_{Xs}, \tilde{\Sigma}_{Xs}) \pi_s = \phi(u| \tilde{\mu}_{Xs}, \tilde{\Sigma}_{Xs}).
\]
From the identifiability of Gaussian distributions, again for the same pair \((g,s)\) in [15], it follows that
\[
\mu_{X_g} = \tilde{\mu}_{X_s} \quad \text{and} \quad \Sigma_{X_g} = \tilde{\Sigma}_{X_s},
\]
and this completes the proof. \(\square\)

## B M-step derivation

For the estimate of the regression coefficients \(\hat{\beta}^{(k+1)}_g\), \(g = 1, \ldots, G\):
\[
\sum_{i=1}^{N} \sum_{g=1}^{G} \tau_{ig}^{(k)} \frac{\partial Q_1(\beta_g, \Sigma_{Y_g} \mid \varphi^{(k)})}{\partial \beta'_g} = 0',
\]
which implies
\[
\frac{\partial}{\partial \beta'_g} \left\{ \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{-\tau_{ig}^{(k)}}{2} \left[ (y_i - \beta'_g x^*_i)' \Sigma_{Y_g}^{-1} (y_i - \beta'_g x^*_i) \right] \right\} = 0',
\]
yielding
\[
\frac{\partial}{\partial \beta'_g} \left\{ \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{-\tau_{ig}^{(k)}}{2} \left[ -y_i' \Sigma_{Y_g}^{-1} \beta'_g x^*_i - x^*_i' \beta_g \Sigma_{Y_g}^{-1} y_i + x^*_i' \beta_g \Sigma_{Y_g}^{-1} \beta'_g x^*_i \right] \right\} = 0'.
\]
Using properties of trace and transpose, we get
\[
\frac{\partial}{\partial \beta'_g} \left\{ \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{-\tau_{ig}^{(k)}}{2} \left[ \text{tr} \left( y_i' \Sigma_{Y_g}^{-1} \beta'_g x^*_i \right) + \text{tr} \left( x^*_i' \beta_g \Sigma_{Y_g}^{-1} y_i \right) - \text{tr} \left( x^*_i' \beta_g \Sigma_{Y_g}^{-1} \beta'_g x^*_i \right) \right] \right\} = 0'.
\]
Taking the derivative, we obtain
\[
\sum_{i=1}^{N} \frac{\tau_{ig}^{(k)}}{2} \left\{ \Sigma_{Y_g}^{-1} y_i x^*_i + \Sigma_{Y_g}^{-1} y_i x^*_i - \left[ \left( \Sigma_{Y_g}^{-1} \right)' \beta'_g x^*_i x^*_i + \Sigma_{Y_g}^{-1} \beta'_g x^*_i x^*_i \right] \right\} = 0',
\]
and finally
\[
\hat{\beta}_{g}^{(k+1)'} = \left( \sum_{i=1}^{N} \tau_{i}^{(k)} y_{i} x_{i}^{*} \right)' \left( \sum_{i=1}^{N} \tau_{i}^{(k)} x_{i}^{*} x_{i}^{*} \right)^{-1}.
\]

Furthermore, for the estimate of the covariance matrix \( \Sigma_{Y_{g}}^{(k+1)} \), \( g = 1, \ldots, G \):
\[
\sum_{i=1}^{N} \tau_{i}^{(k)} \frac{\partial Q_{1} \left( \beta_{g}^{(k+1)}, \Sigma_{Y_{g}} | \theta^{(k)} \right)}{\partial \Sigma_{Y_{g}}^{-1}} = 0',
\]
leading to
\[
\frac{\partial}{\partial \Sigma_{Y_{g}}^{-1}} \left\{ \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{\tau_{i}^{(k)}}{2} \left[ \log |\Sigma_{Y_{g}}| - \text{tr} \left( \Sigma_{Y_{g}}^{-1} \left( \mathbf{y}_{i} - \beta_{g}^{(k+1)'} x_{i}^{*} \right) \left( \mathbf{y}_{i} - \beta_{g}^{(k+1)'} x_{i}^{*} \right)' \right) \right] \right\} = 0',
\]
\[
\frac{\partial}{\partial \Sigma_{Y_{g}}^{-1}} \left\{ \sum_{i=1}^{N} \sum_{g=1}^{G} \frac{\tau_{i}^{(k)}}{2} \left[ \log |\Sigma_{Y_{g}}| - \text{tr} \left( \Sigma_{Y_{g}}^{-1} \left( \mathbf{y}_{i} - \beta_{g}^{(k+1)'} x_{i}^{*} \right) \left( \mathbf{y}_{i} - \beta_{g}^{(k+1)'} x_{i}^{*} \right)' \right) \right] \right\} = 0'.
\]
Taking the derivative we get
\[
\sum_{i=1}^{N} \frac{\tau_{i}^{(k)}}{2} \left\{ (\Sigma_{Y_{g}}^{-1})^{-1'} - \left[ \left( \mathbf{y}_{i} - \beta_{g}^{(k+1)'} x_{i}^{*} \right) \left( \mathbf{y}_{i} - \beta_{g}^{(k+1)'} x_{i}^{*} \right)' \right] \right\} = 0',
\]
and this results in
\[
\Sigma_{Y_{g}}^{(k+1)} = \frac{\sum_{i=1}^{N} \tau_{i}^{(k)} \left( \mathbf{y}_{i} - \hat{\beta}_{g}^{(k+1)'} x_{i}^{*} \right) \left( \mathbf{y}_{i} - \hat{\beta}_{g}^{(k+1)'} x_{i}^{*} \right)'}{\sum_{i=1}^{N} \tau_{i}^{(k)}}.
\]

References

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

Anderson, E. (1935). The Irises of the Gaspe peninsula. Bulletin of the American Iris Society 59, 2–5.
Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika* 59(1), 19–35.

Andrews, J. L. and P. D. McNicholas (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate $t$-distributions. *Statistics and Computing* 22(5), 1021–1029.

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

Azzalini, A. (2013). *R package sn*: The skew-normal and skew-$t$ distributions (version 0.4-18). Università di Padova, Italia.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(7), 719–725.

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* 46(2), 373–388.

Breiman, L. and J. H. Friedman (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59(1), 3–54.

Browne, R. P. and P. D. McNicholas (2013). *mixture*: Mixture Models for Clustering and Classification. R package version 1.0.

Browne, R. P. and P. D. McNicholas (2014). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification* 8(2), 217–226.

Browne, R. P., S. Subedi, and P. McNicholas (2013). Constrained optimization for a subset of the Gaussian parsimonious clustering models. *arXiv preprint arXiv:1306.5824*.

Campbell, N. A. and R. J. Mahon (1974). A multivariate study of variation in two species of rock crab of the genus Leptograpsus. *Australian Journal of Zoology* 22(3), 417–425.

Celeux, G. and J. Diebolt (1988). A random imputation principle: the stochastic EM algorithm. In *Rapports de Recherche 901*. INRIA.

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

Cook, R. D. and S. Weisberg (1994). *An introduction to regression graphics*. John Wiley & Sons.
Dang, U. J. and P. D. McNicholas (2013). Families of parsimonious finite mixtures of regression models. arXiv preprint arXiv:1312.0518v1.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* 93(441), 294–302.

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

DeSarbo, W. S. and W. L. Cron (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* 5(2), 249–282.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2), 179–188.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631.

Galimberti, G. and G. Soffritti (2014). A multivariate linear regression analysis using finite mixtures of t distributions. *Computational Statistics & Data Analysis* 71, 138–150.

Gershenfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences* 808(1), 18–24.

Grünn, B., F. Leisch, et al. (2008). Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software* 28(4), 1–35.

Hartigan, J. A. and M. A. Wong (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28(1), 100–108.

Hastie, T. and R. Tibshirani (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58(1), 155–176.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification* 17(2), 273–296.

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

Ingrassia, S., S. C. Minotti, and A. Punzo (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics and Data Analysis* 71, 159–182.
Ingrassia, S., A. Punzo, and G. Vittadini (2014). The Generalized Linear Mixed Cluster-Weighted Model. *Journal of Classification*, forthcoming.

Ingrassia, S., S. C. Minotti, and G. Vittadini (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of Classification* 29(3), 363–401.

Ingrassia, S. and R. Rocci (2007). Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis* 51(11), 5339–5351.

Keribin, C. (1998). Estimation consistante de l’ordre de modèles de mélange. *Comptes Rendus de l’Académie des Sciences-Series I-Mathematics* 326(2), 243–248.

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

Lebret, R., S. Iovleff, and A. Longeville (2012). *Rmixmod*: Mixture Modelling Package. R package version 1.0.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* 11(8), 1–18.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–163.

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

McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference* 140(5), 1175–1181.

McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics & Data Analysis* 54(3), 711–723.

Punzo, A. (2014). Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. *Statistical Modelling* 14(3), 257–291.

R Core Team (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336), 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
Soffritti, G. and G. Galimberti (2011). Multivariate linear regression with non-normal errors: a solution based on mixture models. *Statistics and Computing* 21(4), 523–536.

Subedi, S., A. Punzo, S. Ingrassia, and P. McNicholas (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification* 7(1), 5–40.

Teicher, H. (1963). Identifiability of finite mixtures. *Annals of Mathematical Statistics* 34(4), 1265–1269.

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

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

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

Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica* 56(3), 362–375.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *Annals of Mathematical Statistics* 39(1), 209–214.

Zhou, H. and K. L. Lange (2010). On the bumpy road to the dominant mode. *Scandinavian Journal of Statistics* 37(4), 612–631.