Multivariate cluster-weighted models based on seemingly unrelated linear regression

# Multivariate cluster-weighted models based on seemingly unrelated linear regression

Cecilia Diani, Giuliano Galimberti, Gabriele Soffritti*

**Abstract**

A class of cluster-weighted models for a vector of continuous random variables is proposed. This class provides an extension to cluster-weighted modelling of multivariate and correlated responses that let the researcher free to use a different vector of covariates for each response. The class also includes parsimonious models obtained by imposing suitable constraints on the component-covariance matrices of either the responses or the covariates. Conditions for model identifiability are illustrated and discussed. Maximum likelihood estimation is carried out by means of an expectation-conditional maximisation algorithm. The effectiveness and usefulness of the proposed models are shown through the analysis of simulated and real datasets.

*Keywords:* cluster analysis, ECM algorithm, Gaussian mixture model, multivariate linear regression, parsimonious model

*2010 MSC:* 62J05, 62H12, 62F12

## 1. Introduction

Cluster-weighted modelling is a flexible framework for data analysis introduced by Gershenfeld (1997) in which the joint distribution of a given random vector is modelled by assuming that this vector is composed of an outcome $\mathbf{Y}$

*Corresponding author
Department of Statistical Sciences, University of Bologna
via delle Belle Arti 41, 40126 Bologna, Italy
Tel.: +39 051 2098192
Fax: +39 051 2086242
Email address: gabriele.soffritti@unibo.it
ORCID: https://orcid.org/0000-0002-7575-892X

*This is the accepted version of an article published by Elsevier:*

*https://doi.org/10.1016/j.csda.2022.107451*

(response, dependent variable) and its explanatory variables $\mathbf{X}$ (covariates, predictors); in order to account for the possible presence of unknown clusters of observations, a finite mixture is embedded into the model. Thus, cluster-weighted models are useful to perform multivariate regression analysis with random covariates in the presence of unobserved heterogeneity. Such models play a prominent role when the sample observations come from several sub-populations, the distribution of the outcome as well as the effect of the covariates on the response change with the sub-populations and the covariates are not under the control of the researcher.

An intense research into cluster-weighted models has been carried out over the last decade. Ingrassia et al. (2012) and Ingrassia et al. (2014) have developed models for continuous variables under both Gaussian and Student $t$ mixture distributions. Solutions suitable for dealing with various types of responses are detailed in Punzo and Ingrassia (2013), Punzo and Ingrassia (2015), Ingrassia et al. (2015) and Di Mari et al. (2020). Models with non-linear relationships or many covariates have been proposed by Punzo (2014) and Subedi et al. (2013), respectively. Robustified solutions have been developed by Subedi et al. (2015) and Punzo and McNicholas (2017). As far as vectors of continuous random variables with a multivariate response are concerned, Dang et al. (2017) have developed a family of parsimonious Gaussian cluster-weighted models, where suitable constraints are imposed on the eigen-decomposition of the component-covariance matrices so as to mitigate the problem of a large number of model parameters when dealing with several variables. An underlying assumption in the family of parsimonious Gaussian cluster-weighted models introduced by Dang et al. (2017) is that all the covariates in the model affect each examined response. However, in some situations there may be prior information concerning the absence of certain covariates from the linear term employed in the prediction of a certain response, and different covariates may be expected to be relevant in the prediction of different responses, as in the seemingly unrelated regression context (Srivastava and Giles, 1987). This approach to multivariate regression has been extensively employed in the modelling of multivariate economic

data, where some given aspects of economic behaviour are typically assumed to depend on different economic variables according to a certain general theory. Classical examples can be found in White and Hewings (1982) and Giles and Hampton (1984), where multivariate regression models with different vectors of covariates were specified and estimated based on employment equations and Cobb-Douglas production functions in different geographical locations, respectively. Other fields in which the same approach has been successfully employed are medicine, food quality, tourism economics, quality of life and health (see, e.g., Keshavarzi et al., 2012; Cadavez and Henningsen, 2012; Keshavarzi et al., 2013; Disegna and Osti, 2016; Heidari et al., 2017). Other regression models for multivariate responses based on finite mixture models have been introduced by Soffritti and Galimberti (2011); Dang and McNicholas (2015); Galimberti et al. (2016). The `flexmix` package (Grün and Leisch, 2008) in the `R` environment (R Core Team, 2020) provides a general framework for the specification and estimation of finite mixtures of regression models.

This paper introduces a class of multivariate seemingly unrelated Gaussian linear cluster-weighted models. Models from this class are able to capture both the linear dependencies among responses and the linear effects of the covariates on the responses from sample observations coming from heterogeneous populations. Furthermore, with these models the researcher is enabled to specify a different vector of covariates for each response. The paper addresses the model identification and maximum likelihood (ML) estimation. This latter task is carried out by resorting to an expectation-conditional maximisation (ECM) algorithm. In order to keep the total number of parameters as low as possible, parsimonious models are included into the novel class, where parsimony is attained by constraining the component-covariance matrices using a parameterisation for such matrices which is based on their spectral decomposition (see, e.g., Celeux and Govaert, 1995). With this approach, fourteen different covariance structures are allowed for both the covariates and the responses. The usefulness and the great flexibility of the resulting model class is shown through two studies, based on the analysis of real datasets, aiming at determining the effect

of prices and promotional activities on sales of canned tuna and at evaluating the link between tourism flows and attendance at museums and monuments. The effectiveness of an approach based on the proposed model class in terms of parameter recovery and classification recovery is demonstrated through Monte Carlo studies.

The paper is organised as follows. Section 2.1 defines the novel class of cluster-weighted models. Section 2.2 shows how the models belonging to this class relate to some existing models. Information on model identifiability is provided in Section 2.3. Details about the ML estimation are given in Section 2.4 and the Appendices. The initialisation and convergence of the ECM algorithm and the issue of model selection are treated in Sections 2.5 and 2.6. Parsimonious models are introduced in Section 2.7. Results of the analyses of simulated and real datasets are summarised in Sections 3 and 4, respectively. Section 5 provides some concluding remarks.

## 2. Multivariate seemingly unrelated linear cluster-weighted analysis

### *2.1. Multivariate seemingly unrelated linear cluster-weighted models*

Following Dang et al. (2017), in a cluster-weighted model the random vectors $\mathbf{X}_i$ and $\mathbf{Y}_i$ containing the $P$ covariates and the $D$ responses for the $i$th observation, respectively, come from a population $\Omega$ which is assumed to be partitioned into $K$ disjoint groups $\Omega_1, \ldots, \Omega_K$. Thus, $\Omega = \Omega_1 \cup \ldots \cup \Omega_K$; furthermore, $\Omega_k \cap \Omega_{k'} = \emptyset \; \forall k \neq k'$. In the models proposed here both $\mathbf{X}_i$ and $\mathbf{Y}_i$ are continuous random vectors, $\mathbf{X}_i$ takes values in $\mathbb{R}^P$, $\mathbf{Y}_i$ takes values in $\mathbb{R}^D$ and the probability density function (p.d.f.) of $(\mathbf{X}_i, \mathbf{Y}_i)$ can be written as

$$f(\mathbf{x}_i, \mathbf{y}_i) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i|\Omega_k) f(\mathbf{y}_i|\mathbf{x}_i, \Omega_k), \tag{1}$$

where $\pi_k = \mathbb{P}(\Omega_k)$ is the mixing weight and represents the prior probability of the $k$th group, $f(\mathbf{x}_i|\Omega_k)$ is the p.d.f. of $\mathbf{X}_i$ given $\Omega_k$ and $f(\mathbf{y}_i|\mathbf{x}_i, \Omega_k)$ is the conditional p.d.f. of the response $\mathbf{Y}_i$ given the value $\mathbf{x}_i$ of the covariates $\mathbf{X}_i$ and the group $\Omega_k$. As far as the mixing weights are concerned, they are supposed to

be positive ($\pi_k > 0\ \forall k$); in addition, they have to sum to 1 ($\sum_{k=1}^{K} \pi_k = 1$). Here $\mathbf{X}_i|\Omega_k$ is assumed to follow a $P$-variate normal distribution with mean vector $\boldsymbol{\mu}_{\mathbf{X}_k}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}_k}$, $k = 1, \ldots, K$. Thus, the expected values, variances and covariances of $\mathbf{X}_i|\Omega_k$ are equal for all observations coming from group $\Omega_k$, while they are different for observations belonging to other groups. As far as $\mathbf{Y}_i|(\mathbf{X}_i = \mathbf{x}_i, \Omega_k)$ is concerned, its distribution is modelled using a $D$-variate normal distribution with conditional expected vector given by some linear transformation of $\mathbf{x}_i$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}_k}$. Thus, variances and covariances of $\mathbf{Y}_i|(\mathbf{X}_i = \mathbf{x}_i, \Omega_k)$ are equal for observations coming from the same group; the expected values of $\mathbf{Y}_i|(\mathbf{X}_i = \mathbf{x}_i, \Omega_k)$ for such observations vary with the observations. Furthermore, different correlation structures among both the covariates and the responses across the $K$ groups are assumed.

In order to describe how a cluster-weighted model with a different vector of covariates for each response can be obtained, the following additional notation is required. Suppose that only $P_d$ of the $P$ covariates ($P_d \leq P$) are considered to be relevant for the prediction of the $d$th response. Thus, let $\mathbf{x}_{id} = (x_{i,d_1}, x_{i,d_2}, \ldots, x_{i,d_{P_d}})'$ be the vector composed of the values of such $P_d$ covariates for the $i$th observation and $\mathbf{x}^*_{id} = (1, \mathbf{x}'_{id})'$. Furthermore, let $\boldsymbol{\beta}_{kd} = (\beta_{kd_1}, \beta_{kd_2}, \ldots, \beta_{kd_{P_d}})'$ be the $P_d$-dimensional vector of regression coefficients capturing the linear effect of these $P_d$ covariates on the $d$th response in the $k$th group, and $\boldsymbol{\beta}^*_{kd} = (\beta_{kd0}, \boldsymbol{\beta}'_{kd})'$. Then, the vector containing all linear effects on the $D$ responses in the $k$th group is given by $\boldsymbol{\beta}^*_k = (\boldsymbol{\beta}^{*\prime}_{k1}, \ldots, \boldsymbol{\beta}^{*\prime}_{kd}, \ldots, \boldsymbol{\beta}^{*\prime}_{kD})'$; the length of this vector is $(P^*+D)$, where $P^* = \sum_{d=1}^{D} P_d$. Finally, the following $(P^* + D) \times D$ partitioned matrix is required:

$$\mathcal{X}_i = \begin{bmatrix} \mathbf{x}^*_{i1} & \mathbf{0}_{P_1+1} & \ldots & \mathbf{0}_{P_1+1} \\ \mathbf{0}_{P_2+1} & \mathbf{x}^*_{i2} & \ldots & \mathbf{0}_{P_2+1} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{P_D+1} & \mathbf{0}_{P_D+1} & \ldots & \mathbf{x}^*_{iD} \end{bmatrix},$$

with $\mathbf{0}_{P_d+1}$ denoting the $(P_d + 1)$-dimensional null vector. With this notation,

the conditional expected vector of $\mathbf{Y}_i|(\mathbf{X}_i = \mathbf{x}_i, \Omega_k)$ is given by

$$\boldsymbol{\mu}_{\mathbf{Y}_k}(\mathbf{x}_i; \boldsymbol{\beta}_k^*) = \mathcal{X}_i' \boldsymbol{\beta}_k^* = (\mathbf{x}_{i1}^{*\prime} \boldsymbol{\beta}_{k1}^*, \ldots, \mathbf{x}_{id}^{*\prime} \boldsymbol{\beta}_{kd}^*, \ldots, \mathbf{x}_{iD}^{*\prime} \boldsymbol{\beta}_{kD}^*)'. \tag{2}$$

According to this equation, the conditional expected value of the $d$th response within the $k$th group is given by the linear term $\mathbf{x}_{id}^{*\prime} \boldsymbol{\beta}_{kd}^*$, which only depends on the $P_d$ covariates included in the vector $\mathbf{x}_{id}$. It is worth noting that the regression coefficients vary across groups, which means that the effect of the covariates on the responses changes with the groups. Embedding all these assumptions into model (1) leads to

$$f(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k \phi_P \left(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{X}_k}, \boldsymbol{\Sigma}_{\mathbf{X}_k}\right) \phi_D \left(\mathbf{y}_i | \mathbf{x}_i; \mathcal{X}_i' \boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}\right), \tag{3}$$

where $\phi_P$ ($\phi_D$) represents the p.d.f. of a $P$-variate ($D$-variate) Gaussian random vector, $\boldsymbol{\psi} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_{\mathbf{X}_1}, \ldots, \boldsymbol{\mu}_{\mathbf{X}_K}, \boldsymbol{\Sigma}_{\mathbf{X}_1}, \ldots, \boldsymbol{\Sigma}_{\mathbf{X}_K}, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*, \boldsymbol{\Sigma}_{\mathbf{Y}_1}, \ldots, \boldsymbol{\Sigma}_{\mathbf{Y}_K}\}$ denotes the set of all model parameters and $\boldsymbol{\Psi}$ is the parameter space. The number of free parameters in $\boldsymbol{\psi}$ is $K - 1 + K(P + P^* + D) + K[P(P + 1)/2 + D(D+1)/2]$, which is the sum of the unknown mixture weights, expected values, variances and covariances.

It is worth stressing that the model in equation (3) differs from the model proposed by Dang et al. (2017) because of a different definition of the linear term for the conditional expected value of $\mathbf{Y}_i|(\mathbf{X}_i = \mathbf{x}_i, \Omega_k)$. If all the $P$ covariates are considered to be relevant for the prediction of all responses, that is $\mathbf{x}_{id} = \mathbf{x}_i$ $\forall d$, then $\mathbf{x}_{id}^* = \mathbf{x}_i^*$ $\forall d$, where $\mathbf{x}_i^* = (1, \mathbf{x}_i')'$, and the following equality holds:

$$\mathcal{X}_i = \mathbf{I}_D \otimes \mathbf{x}_i^*,$$

where $\mathbf{I}_D$ is the identity matrix of order $D$ and $\otimes$ denotes the Kronecker product operator (see, e.g., Magnus and Neudecker, 1988). Then, equation (2) can be rewritten as

$$\boldsymbol{\mu}_{\mathbf{Y}_k}(\mathbf{x}_i; \boldsymbol{\beta}_k^*) = (\mathbf{I}_D \otimes \mathbf{x}_i^*)' \boldsymbol{\beta}_k^* = \mathbf{B}_k' \mathbf{x}_i^*, \ k = 1, \ldots, K,$$

where $\mathbf{B}_k = [\boldsymbol{\beta}_{k1}^* \cdots \boldsymbol{\beta}_{kd}^* \cdots \boldsymbol{\beta}_{kD}^*]$, thus leading to the multivariate Gaussian cluster-weighted model introduced by Dang et al. (2017). As illustrated in Section 1, seemingly unrelated regression models can be considered as multivariate

regression models in which prior information about the absence of certain covariates for the prediction of certain responses is explicitly taken into consideration (Srivastava and Giles, 1987). Thus, equation (3) can also be seen as a multivariate Gaussian cluster-weighted model in which some regression coefficients are constrained to be a priori equal to zero. To the best of the authors' knowledge, the inclusion of such constraints in the multivariate Gaussian cluster-weighted model framework has not been addressed yet.

### *2.2. Relationships with linear clusterwise regression models*

Under suitable conditions, it is possible to establish some relationships between the multivariate seemingly unrelated Gaussian linear cluster-weighted models just introduced and some Gaussian linear clusterwise regression models.

In Section 2.1 it has been highlighted that models (3) assume that $\mathbf{X}_i|\Omega_k \sim N_P(\boldsymbol{\mu}_{\mathbf{X}_k}, \boldsymbol{\Sigma}_{\mathbf{X}_k})$, for $k = 1, \ldots, K$. If the p.d.f of $\mathbf{X}_i|\Omega_k$ does not depend on group $\Omega_k$, i.e., $\phi_P\left(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{X}_k}, \boldsymbol{\Sigma}_{\mathbf{X}_k}\right) = \phi_P\left(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}}\right)$ for every $k = 1, \ldots, K$, then equation (3) can also be written as

$$f(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\psi}) = \phi_P\left(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}}\right) \sum_{k=1}^{K} \pi_k \phi_D\left(\mathbf{y}_i|\mathbf{x}_i; \mathcal{X}_i'\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}\right), \tag{4}$$

where

$$f(\mathbf{y}_i|\mathbf{x}_i; \tilde{\boldsymbol{\psi}}) = \sum_{k=1}^{K} \pi_k \phi_D\left(\mathbf{y}_i|\mathbf{x}_i; \mathcal{X}_i'\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}\right), \tag{5}$$

with $\tilde{\boldsymbol{\psi}} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*, \boldsymbol{\Sigma}_{\mathbf{Y}_1}, \ldots, \boldsymbol{\Sigma}_{\mathbf{Y}_K}\}$, is the seemingly unrelated Gaussian clusterwise linear regression model described in Galimberti and Soffritti (2020). This means that the assignment of the data points to the groups is independent of the covariates; such a condition is also known as assignment independence (see, e.g., Hennig, 2000). Furthermore, if the researcher sets $\mathbf{x}_{id} = \mathbf{x}_i$ $\forall d$ (i.e., all the $P$ covariates are assumed to be relevant for the prediction of all responses), then equation (5) leads to the traditional multivariate Gaussian clusterwise linear regression models (Jones and McLachlan, 1992). Thus, when in equation (3) the following conditions hold true: $\boldsymbol{\mu}_{\mathbf{X}_k} = \boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\Sigma}_{\mathbf{X}_k} = \boldsymbol{\Sigma}_{\mathbf{X}}$ for $k = 1, \ldots, K$, then the information about the $K$ disjoint groups $\Omega_1, \ldots, \Omega_K$

that compose the population $\Omega$ can be equivalently obtained either from the analysis of the conditional p.d.f. $f(\mathbf{y}_i|\mathbf{x}_i;\tilde{\boldsymbol{\psi}})$ through seemingly unrelated linear clusterwise models or from the analysis of the joint p.d.f. $f(\mathbf{x}_i,\mathbf{y}_i;\boldsymbol{\psi})$ through seemingly unrelated linear cluster-weighted models.

Furthermore, when the following conditions hold true: *i)* the conditional distribution of $Y_{id}|\mathbf{X}_i=\mathbf{x}_i$ changes with $K_d$ disjoint groups $\Omega_{d1},\ldots,\Omega_{dk_d},\ldots,\Omega_{dK_d}$ that compose the population $\Omega$ for $d=1,\ldots,D$; *ii)* these $D$ partitions of $\Omega$ associated with the $D$ responses are mutually independent (i.e., the population is characterised by $D$ independent cluster structures) (Galimberti and Soffritti, 2007); *iii)* the assignment independence condition holds true for each of these groupings, then the following model can be defined:

$$f(\mathbf{x}_i,\mathbf{y}_i;\boldsymbol{\psi})=\phi_P\left(\mathbf{x}_i;\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\Sigma}_{\mathbf{X}}\right)\prod_{d=1}^{D}\sum_{k_d=1}^{K_d}\pi_{k_d}\phi_1\left(y_{id}|\mathbf{x}_i;\mathbf{x}_{id}^{*\prime}\boldsymbol{\beta}_{k_dd}^{*},\sigma_{k_dd}^{2}\right), \qquad (6)$$

where $y_{id}$ is the $d$th element of $\mathbf{y}_i$, $\mathbf{x}_{id}^{*\prime}\boldsymbol{\beta}_{k_dd}^{*}$ and $\sigma_{k_dd}^{2}$ are the conditional expected value and the variance of $Y_{id}|\mathbf{X}_i=\mathbf{x}_i$ within the group $\Omega_{dk_d}$, respectively. Thus, under conditions *i)–iii)*, model (6) holds true and the information about the $D$ independent partitions of the sample observations should be obtained from $D$ univariate seemingly unrelated linear clusterwise regression models.

In the light of the relationships just illustrated, it is possible to conclude that multivariate seemingly unrelated linear cluster-weighted models will be more effective than multivariate seemingly unrelated linear clusterwise regression models when the assignment independence condition does not hold true. Furthermore, an analysis based on the proposed models should be carried out rather than $D$ separate analyses, based on $D$ univariate seemingly unrelated linear clusterwise regression models, whenever either the condition of $D$ independent cluster structures or the assignment independence condition do not hold for the examined population.

### *2.3. Model identifiability*

Identifiability is essential for parameter estimation and represents a preliminary requirement for the consistency and other asymptotic properties of the

ML estimator. Generally speaking, several types of non-identifiability can affect finite mixture models. A first type is due to invariance to relabeling the components (also known as label-switching). Furthermore, non-identifiability is caused by potential overfitting associated with empty components or equal components (see, e.g., Frühwirth-Schnatter, 2006, p. 15). Thus, identifiability of finite mixture models may be achieved after imposing suitable constraints on the parameter space. As far as multivariate Gaussian cluster-weighted models are concerned, conditions ensuring their identifiability have been defined by Dang et al. (2017). Those conditions can be easily modified in order to hold true also for the seemingly unrelated Gaussian linear cluster-weighted models defined according to equation (3).

The constraints to be imposed on the parameters are $\pi_k > 0 \ \forall \ k$ and $(\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}) \neq (\boldsymbol{\beta}_h^*, \boldsymbol{\Sigma}_{\mathbf{Y}_h})$ for $k \neq h$. These constraints make it possible to avoid the two types of non-identifiability illustrated above. Thus, in order to ensure identifiability, the following class of seemingly unrelated cluster-weighted models has to be considered:

$$\begin{aligned}\mathfrak{F} = \Big\{ f(\mathbf{x}, \mathbf{y}; \bar{\boldsymbol{\psi}}) : f(\mathbf{x}, \mathbf{y}; \bar{\boldsymbol{\psi}}) \quad &= \quad \sum_{k=1}^{K} \pi_k \phi_P \left(\mathbf{x}; \boldsymbol{\mu}_{\mathbf{X}_k}, \boldsymbol{\Sigma}_{\mathbf{X}_k}\right) \phi_D \left(\mathbf{y}|\mathbf{x}; \mathcal{X}'\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}\right), \\ &\quad (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{P+D}, \bar{\boldsymbol{\psi}} \in \bar{\boldsymbol{\Psi}}, K \in \mathbb{N} \Big\},\end{aligned}$$

where $\bar{\boldsymbol{\Psi}}$ is the constrained parameter space, defined as follows:

$$\bar{\boldsymbol{\Psi}} = \left\{ \bar{\boldsymbol{\psi}} \in \boldsymbol{\Psi} : \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1, (\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}) \neq (\boldsymbol{\beta}_h^*, \boldsymbol{\Sigma}_{\mathbf{Y}_h}) \text{ for } k \neq h \right\}.$$

An additional condition for the class $\mathfrak{F}$ to be identifiable is the existence of a set $\mathcal{W} \subseteq \mathbb{R}^P$ having probability equal to one according to the $P$-dimensional Gaussian distribution such that the following clusterwise regression model

$$\sum_{k=1}^{K} \phi_M \left(\mathbf{y}|\mathbf{x}; \mathcal{X}'\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}\right) \alpha_k(\mathbf{x}), \ \mathbf{y} \in \mathbb{R}^D,$$

is identifiable for each fixed $\mathbf{x} \in \mathcal{W}$, where $\alpha_1(\mathbf{x}), ..., \alpha_K(\mathbf{x})$ are positive weights summing to one for each $\mathbf{x} \in \mathcal{W}$. Under this condition, it is possible to prove that the class $\mathfrak{F}$ results to be identifiable in $\mathcal{W} \times \mathbb{R}^D$. The proof of this result can be

easily obtained from the proof of the analogous result for multivariate Gaussian cluster-weighted models (see Dang et al., 2017, Appendix A), by simply changing the linear term for the conditional expected value of $\mathbf{Y}|(\mathbf{X} = \mathbf{x}, \Omega_k)$.

### *2.4. Parameter estimation*

Given a sample $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I)\}$ of $I$ independent observations from model (3), ML estimation of the model parameters $\boldsymbol{\psi}$ can be carried out by means of an ECM algorithm developed under a general framework dealing with incomplete-data problems (Dempster et al., 1977; Meng and Rubin, 1993). The missing information is the specific component of the mixture from which the sample observations come from; such information can be described by the $K$-dimensional vectors $(\mathbf{z}_1, \ldots, \mathbf{z}_I)$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ with $z_{ik} = 1$ if the $i$th observation comes from the $k$th component and $z_{ik} = 0$ otherwise, for $k = 1, \ldots, K$. Then, the complete data would be $\mathcal{S}_c = \{(\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I, \mathbf{z}_I)\}$. Thus, the likelihood functions derived from the incomplete data and the complete data are

$$
\begin{aligned}
L(\boldsymbol{\psi}|\mathcal{S}) &= \prod_{i=1}^{I}\Big[\sum_{k=1}^{K}\pi_k\phi_P\left(\mathbf{x}_i;\boldsymbol{\mu}_{\mathbf{X}_k},\boldsymbol{\Sigma}_{\mathbf{X}_k}\right)\phi_D\left(\mathbf{y}_i|\mathbf{x}_i;\mathcal{X}_i'\boldsymbol{\beta}_k^*,\boldsymbol{\Sigma}_{\mathbf{Y}_k}\right)\Big], \\
L(\boldsymbol{\psi}|\mathcal{S}_c) &= \prod_{i=1}^{I}\prod_{k=1}^{K}\Big[\pi_k\phi_P\left(\mathbf{x}_i;\boldsymbol{\mu}_{\mathbf{X}_k},\boldsymbol{\Sigma}_{\mathbf{X}_k}\right)\phi_D\left(\mathbf{y}_i|\mathbf{x}_i;\mathcal{X}_i'\boldsymbol{\beta}_k^*,\boldsymbol{\Sigma}_{\mathbf{Y}_k}\right)\Big]^{z_{ik}},
\end{aligned}
$$

respectively; the complete-data log-likelihood function employed in the ECM algorithm for the computation of the parameter estimates is

$$
\begin{aligned}
\ell(\boldsymbol{\psi}|\mathcal{S}_c) &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\Big[\ln\pi_k + \ln\phi_P\left(\mathbf{x}_i;\boldsymbol{\mu}_{\mathbf{X}_k},\boldsymbol{\Sigma}_{\mathbf{X}_k}\right) \\
&\quad + \ln\phi_D\left(\mathbf{y}_i|\mathbf{x}_i;\mathcal{X}_i'\boldsymbol{\beta}_k^*,\boldsymbol{\Sigma}_{\mathbf{Y}_k}\right)\Big]. \qquad (7)
\end{aligned}
$$

The $h$th iteration of the E-step in the ECM algorithm consists in calculating the conditional expectation of the complete-data log-likelihood (7) on the basis of the current estimate $\hat{\boldsymbol{\psi}}^{(h)}$ of the model parameters $\boldsymbol{\psi}$:

$$
\begin{aligned}
\mathbb{E}_{\hat{\boldsymbol{\psi}}^{(h)}}\left[\ell(\boldsymbol{\psi}|\mathcal{S}_c)\right] &= \sum_{i=1}^{I}\sum_{k=1}^{K}\hat{\tau}_{ik}^{(h)}\Big[\ln\hat{\pi}_k^{(h)} + Q_1\Big(\boldsymbol{\mu}_{\mathbf{X}_k},\boldsymbol{\Sigma}_{\mathbf{X}_k}|\hat{\boldsymbol{\psi}}^{(h)}\Big) \\
&\quad + Q_2\Big(\boldsymbol{\beta}_k^*,\boldsymbol{\Sigma}_{\mathbf{Y}_k}|\hat{\boldsymbol{\psi}}^{(h)}\Big)\Big], \qquad (8)
\end{aligned}
$$

where

$$
\begin{aligned}
Q_1\Big(\boldsymbol{\mu}_{\mathbf{X}_k}, \boldsymbol{\Sigma}_{\mathbf{X}_k}|\hat{\boldsymbol{\psi}}^{(h)}\Big) &= \frac{1}{2}\Big[ - P\ln(2\pi) - \ln|\hat{\boldsymbol{\Sigma}}^{(h)}_{\mathbf{X}_k}| \\
&\quad -\Big(\mathbf{x}_i - \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}_k}\Big)' \hat{\boldsymbol{\Sigma}}^{(h)(-1)}_{\mathbf{X}_k}\Big(\mathbf{x}_i - \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}_k}\Big)\Big] , \\
Q_2\Big(\boldsymbol{\beta}^*_k, \boldsymbol{\Sigma}_{\mathbf{Y}_k}|\hat{\boldsymbol{\psi}}^{(h)}\Big) &= \frac{1}{2}\Big[ - D\ln(2\pi) - \ln|\hat{\boldsymbol{\Sigma}}^{(h)}_{\mathbf{Y}_k}| \\
&\quad -\Big(\mathbf{y}_i - \mathcal{X}_i'\hat{\boldsymbol{\beta}}^{*(h)}_k\Big)' \hat{\boldsymbol{\Sigma}}^{(h)(-1)}_{\mathbf{Y}_k}\Big(\mathbf{y}_i - \mathcal{X}_i'\hat{\boldsymbol{\beta}}^{*(h)}_k\Big)\Big],
\end{aligned}
$$

and $\hat{\tau}^{(h)}_{ik}$ provides the posterior probability (evaluated using the current estimate $\hat{\boldsymbol{\psi}}^{(h)}$) that $(\mathbf{x}_i, \mathbf{y}_i)$ is generated from the $k$th component of the mixture, that is

$$
\begin{aligned}
\hat{\tau}^{(h)}_{ik} &= \mathbb{E}_{\hat{\boldsymbol{\psi}}^{(h)}}[Z_{ik}|\mathbf{x}_i, \mathbf{y}_i] = \mathbb{P}_{\hat{\boldsymbol{\psi}}^{(h)}}\{Z_{ik} = 1|\mathbf{x}_i, \mathbf{y}_i\} \\
&= \frac{\hat{\pi}^{(h)}_k \phi_P\left(\mathbf{x}_i; \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}_k}, \hat{\boldsymbol{\Sigma}}^{(h)}_{\mathbf{X}_k}\right) \phi_D\left(\mathbf{y}_i|\mathbf{x}_i; \mathcal{X}_i'\hat{\boldsymbol{\beta}}^{*(h)}_k, \hat{\boldsymbol{\Sigma}}^{(h)}_{\mathbf{Y}_k}\right)}{\sum_{k'=1}^{K} \hat{\pi}^{(h)}_{k'} \phi_P\left(\mathbf{x}_i; \hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}_{k'}}, \hat{\boldsymbol{\Sigma}}^{(h)}_{\mathbf{X}_{k'}}\right) \phi_D\left(\mathbf{y}_i|\mathbf{x}_i; \mathcal{X}_i'\hat{\boldsymbol{\beta}}^{*(h)}_{k'}, \hat{\boldsymbol{\Sigma}}^{(h)}_{\mathbf{Y}_{k'}}\right)}. \qquad (9)
\end{aligned}
$$

The $(h+1)$th update of $\hat{\boldsymbol{\psi}}^{(h)}$ is obtained by a sequence of CM-steps involved in the ECM algorithm. These steps are meant to maximise the conditional expectation of $\ell(\boldsymbol{\psi}|\mathcal{S}_c)$ with respect to $\boldsymbol{\psi}$. This maximisation can be achieved by setting the first order derivatives of $\mathbb{E}[\ell(\boldsymbol{\psi}|\mathcal{S}_c)]$ equal to zero and then solving the resulting system of equations with respect to the parameters of interest. Since this expected value can be decomposed in a sum of three terms, each one depending on a specific set of parameters (see equation (8)), maximisation can be carried out separately for each set of parameters. The resulting updates of $\hat{\pi}^{(h)}_k$, $\hat{\boldsymbol{\mu}}^{(h)}_{\mathbf{X}_k}$, $\hat{\boldsymbol{\Sigma}}^{(h)}_{\mathbf{X}_k}$, $k = 1, \ldots, K$ are:

$$
\begin{aligned}
\hat{\pi}^{(h+1)}_k &= \frac{1}{I}\sum_{i=1}^{I} \hat{\tau}^{(h)}_{ik}, \qquad (10) \\
\hat{\boldsymbol{\mu}}^{(h+1)}_{\mathbf{X}_k} &= \frac{\sum_{i=1}^{I} \hat{\tau}^{(h)}_{ik}\mathbf{x}_i}{\sum_{i=1}^{I} \hat{\tau}^{(h)}_{ik}}, \qquad (11) \\
\hat{\boldsymbol{\Sigma}}^{(h+1)}_{\mathbf{X}_k} &= \frac{\sum_{i=1}^{I} \hat{\tau}^{(h)}_{ik}\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}^{(h+1)}_{\mathbf{X}_k}\right)\left(\mathbf{x}_i - \hat{\boldsymbol{\mu}}^{(h+1)}_{\mathbf{X}_k}\right)'}{\sum_{i=1}^{I} \hat{\tau}^{(h)}_{ik}}. \qquad (12)
\end{aligned}
$$

Such updates coincide with the ones reported in Dang et al. (2017). The CM-

steps to update the remaining parameters are (see Appendix A for a proof)

$$\hat{\boldsymbol{\beta}}_k^{*(h+1)} = \left[\sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \mathcal{X}_i \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h)}\right)^{-1} \mathcal{X}_i'\right]^{-1} \left[\sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \mathcal{X}_i \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h)}\right)^{-1} \mathbf{y}_i\right], \quad (13)$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h+1)} = \frac{\sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \left(\mathbf{y}_i - \mathcal{X}_i' \hat{\boldsymbol{\beta}}_k^{*(h+1)}\right) \left(\mathbf{y}_i - \mathcal{X}_i' \hat{\boldsymbol{\beta}}_k^{*(h+1)}\right)'}{\sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)}}. \quad (14)$$

It is worth noting that the matrix $\sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \mathcal{X}_i \left(\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h)}\right)^{-1} \mathcal{X}_i'$ has to be nonsingular in order for the update $\hat{\boldsymbol{\beta}}_k^{*(h+1)}$ in equation (13) to exist. In addition, Appendix B shows that equation (13) is equivalent to the expression reported in Dang et al. (2017) for the updates of the regression coefficient matrix when $\mathbf{x}_{id} = \mathbf{x}_i \ \forall d$. As a consequence, in this special case the ECM algorithm described in this section reduces to the EM algorithm described in Dang et al. (2017). Finally, once the convergence is reached, the ECM algorithm also provides estimates of the posterior probabilities according to equation (9), which can be used to partition the $I$ observations into $K$ clusters, by assigning each observation to the component showing the highest posterior probability.

Difficulties with this ECM algorithm can arise when matrices $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}_k}^{(h+1)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h+1)}$ in equations (12) and (14) are singular or nearly singular. Another difficulty with ML estimation of Gaussian mixture models is the unboundedness of the likelihood function (see, e.g. Frühwirth-Schnatter, 2006, p. 173). A way to deal with these problems is to introduce suitable constraints on the parameter space $\boldsymbol{\Psi}$ and to perform the estimation under a constrained $\boldsymbol{\Psi}$ (see, e.g. Ingrassia and Rocci, 2011; Rocci et al., 2018). All the analyses illustrated in this paper have been carried out through an implementation of the proposed ECM algorithm, which also allows the estimation of the multivariate linear cluster-weighted models introduced by Dang et al. (2017), in the `R` environment. Such an implementation embeds suitable constraints on the eigenvalues of both $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}_k}^{(h+1)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h+1)}$ for $k = 1, \ldots, K$. Namely, following Dang et al. (2017), all estimated covariance matrices have been required to have eigenvalues greater than the conservative bound $10^{-20}$; furthermore, the ratio between the smallest and the largest eigenvalues of such matrices is required to be not lower than $10^{-10}$.

Finally, in order to avoid problems associated with the invariance of a mixture distribution to relabeling its components (see, e.g., Frühwirth-Schnatter, 2006, p. 15), the $K$ estimated components of the model (3) have been labeled according to the estimated prior probabilities taken in non-decreasing order.

### *2.5. Initialisation and convergence of the ECM algorithm*

A crucial point of any ECM algorithm is the choice of the starting values for the model parameters (i.e., $\hat{\boldsymbol{\psi}}^{(0)}$). An approach based on multiple random initialisations and multiple executions of the ECM algorithm could be adopted. Approaches based on non-random choices can be employed. A solution could be obtained by resorting to the following two-step strategy. In the first step a mixture of $K$ Gaussian models is estimated for the joint distribution all covariates and responses. This task can be carried out, for example, by resorting to the `mclust` package (Scrucca et al., 2017) for the `R` environment. The $K$ prior probabilities, mean vectors and covariance matrices for the predictors estimated in this way are used as $\hat{\pi}_k^{(0)}$, $\hat{\boldsymbol{\mu}}_{\mathbf{X}_k}^{(0)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}_k}^{(0)}$, for $k = 1, \ldots, K$. In the second step $\hat{\boldsymbol{\beta}}_k^{*(0)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(0)}$ are obtained from an estimate of the parameters of the conditional distribution of the responses given the predictors based on the fitting of a seemingly unrelated Gaussian linear regression model to the sample observations that have been assigned to the $k$th component of the mixture model estimated in the first step. The `R` package `systemfit` (Henningsen and Hamann, 2007) can be exploited to perform this task. Another way to obtain $\hat{\pi}_k^{(0)}$, $\hat{\boldsymbol{\mu}}_{\mathbf{X}_k}^{(0)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}_k}^{(0)}$, for $k = 1, \ldots, K$ could be based on the fitting of a mixture of $K$ Gaussian models for the marginal distribution of the covariates in the first step of the previous strategy while keeping the second step unchanged. In all analyses reported in this paper involving either models defined in equation (3) or models introduced by Dang et al. (2017), both these strategies have been simultaneously employed; thus, two different initialisations have been considered for each analysed dataset. Then, the ECM algorithm has been initialised with the strategy leading to the largest value of the incomplete log-likelihood.

In the `R` function employed for the parameter estimation in all analyses

summarised in this paper, the convergence of the ECM algorithm has been evaluated through a criterion based on the Aitken acceleration (Aitken, 1926) which consists in stopping the ECM algorithm when $\left| l_A^{(h+1)} - \ell(\hat{\boldsymbol{\psi}}^{(h)}|\mathcal{S}) \right| < \epsilon$, where $l_A^{(h+1)}$ is the $(h+1)$th Aitken accelerated estimate of the log-likelihood limit and $\ell(\hat{\boldsymbol{\psi}}^{(h)}|\mathcal{S})$ is the value of the incomplete log-likelihood at the $h$th iteration (see, e.g., McNicholas, 2010, for more details). Such criterion can avoid premature stops associated with the use of lack of progress stopping criteria, such as the one based on the difference between the log-likelihood values at two consecutive steps of the ECM algorithm. The maximum number of iterations for the ECM algorithm and the value for $\epsilon$ have been set equal to 500 and $10^{-8}$, respectively.

### *2.6. Model selection*

The ECM algorithm described in Section 2.4 performs the ML estimation for a given value of $K$. However, in most practical applications, the number of groups is not known and must be determined from the data $\mathcal{S}$. A common solution to this task is obtained by resorting to model selection criteria which allows to trade-off the fit (measured by $l_M(\hat{\boldsymbol{\psi}}|\mathcal{S})$, the maximum of the incomplete loglikelihood of model $M$) and complexity (given by $npar_M$, the number of free parameters in model $M$) (see, e.g., Frühwirth-Schnatter, 2006, subsections 4.4.2-4.4.3). In the context of Gaussian mixture models and Gaussian cluster-weighted models (see, e.g., Fraley and Raftery, 2002; Dang et al., 2017), the Bayesian Information Criterion ($BIC$) (Schwarz, 1978) has performed well and is commonly employed. It can be computed as follows: $BIC_M = -2l_M(\hat{\boldsymbol{\psi}}|\mathcal{S}) + npar_M \ln I$. Given a collection of competing fitted candidate models, the one that minimises $BIC_M$ is preferred. Model selection criteria that also consider the quality of the estimated partition of the sample observations represent another possible solution (see, e.g., Frühwirth-Schnatter, 2006, subsection 7.1.4).

### *2.7. Parsimonious models*

As the number of free parameters in equation (3) incresases quadratically with both the number of responses and the number of predictors, analyses

based on the proposed models can become unfeasible in practical applications. This problem can be overcome by introducing constraints on the elements of the covariance matrices $\boldsymbol{\Sigma}_{\mathbf{X}_k}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}_k}$ $(k = 1, \ldots, K)$ according to the approach illustrated in Celeux and Govaert (1995). In this approach, the following eigen-decomposition of the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}_k}$ has to be considered: $\boldsymbol{\Sigma}_{\mathbf{X}_k} = \alpha_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k'$, where $\alpha_k = |\boldsymbol{\Sigma}_{\mathbf{X}_k}|^{1/D}$, $\mathbf{A}_k$ is the diagonal matrix containing the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{X}_k}$ (normalised in such a way that $|\mathbf{A}_k| = 1$) and $\mathbf{D}_k$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}_k}$. Thus, variances and covariances in $\boldsymbol{\Sigma}_{\mathbf{X}_k}$ can be obtained from $\alpha_k$, $\mathbf{A}_k$ and $\mathbf{D}_k$. From a geometrical point of view, such parameters determine the volume, shape and orientation of the $k$th cluster of observations with respect to the predictors. By constraining one or more of these three parameters to be equal across components, 14 different covariance structures for the predictors in models (3) with $K > 1$ can be determined (see Celeux and Govaert, 1995, for more details). Additional information about these parameterisations can be found in Table 1. The application of the same approach to the covariance matrices $\boldsymbol{\Sigma}_{\mathbf{Y}_k}$, $k = 1, \ldots, K$ leads to a class of 196 different models for any given $K > 1$. Equations (12) and (14) represent the solutions for the model in which the covariance structures of both predictors and responses are fully unconstrained. For all other parsimonious models, the CM-step updates for the estimation of $\boldsymbol{\Sigma}_{\mathbf{X}_k}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}_k}$ in the ECM algorithm have to be modified; these modified updates can be computed either in closed form or using iterative procedures, depending on the specific parameterisation to be employed (see Celeux and Govaert, 1995, for more details). The CM-step updates $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}_k}^{(h+1)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h+1)}$ associated with the parameterisations `EVE` and `VVE` can be computed using the F-G algorithm (Flury and Gautschi, 1986) or one of its variants (see, e.g., Lin, 2014). Algorithms which are computationally feasible also in high-dimensional situations have been recently introduced (Browne and McNicholas, 2014a,b). All the experimental results illustrated here and concerning the `EVE` and `VVE` parameterisations have been obtained using the algorithms given in Browne and McNicholas (2014a). When $K = 1$, only three covariance structures for both responses and covariates are possible: diagonal with different

Table 1: Parsimonious parameterisations for the component-covariance matrices

| Acronym | Model | Distribution | Volume | Shape | Orientation |
|---|---|---|---|---|---|
| EEE | $\alpha\mathbf{DAD}'$ | Ellipsoidal | Equal | Equal | Equal |
| VVV | $\alpha_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k'$ | Ellipsoidal | Variable | Variable | Variable |
| EII | $\alpha\mathbf{I}$ | Spherical | Equal | Equal | – |
| VII | $\alpha_k\mathbf{I}$ | Spherical | Variable | Equal | – |
| EEI | $\alpha\mathbf{A}$ | Diagonal | Equal | Equal | – |
| VEI | $\alpha_k\mathbf{A}$ | Diagonal | Variable | Equal | – |
| EVI | $\alpha\mathbf{A}_k$ | Diagonal | Equal | Variable | – |
| VVI | $\alpha_k\mathbf{A}_k$ | Diagonal | Variable | Variable | – |
| EEV | $\alpha\mathbf{D}_k\mathbf{AD}_k'$ | Ellipsoidal | Equal | Equal | Variable |
| VEV | $\alpha_k\mathbf{D}_k\mathbf{AD}_k'$ | Ellipsoidal | Variable | Equal | Variable |
| EVE | $\alpha\mathbf{DA}_k\mathbf{D}'$ | Ellipsoidal | Equal | Variable | Equal |
| VVE | $\alpha_k\mathbf{DA}_k\mathbf{D}'$ | Ellipsoidal | Variable | Variable | Equal |
| VEE | $\alpha_k\mathbf{DAD}'$ | Ellipsoidal | Variable | Equal | Equal |
| EVV | $\alpha\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k'$ | Ellipsoidal | Equal | Variable | Variable |

entries (VI), diagonal with the same entries (EI) and fully unconstrained (VV). Thus, nine differentially parameterised one-component cluster-weighted models can be obtained.

## 3. Results from Monte Carlo studies

The main purpose of the studies based on simulated datasets illustrated here is to obtain an evaluation of the effectiveness of the proposed methodology in comparison with the approach introduced by Dang et al. (2017), where the same vector of covariates has to be employed for all responses. Thus, cluster-weighted models belonging to two different classes have been fitted to each simulated dataset: *i)* models in which all the $D$ responses are assumed to depend on all the $P$ examined covariates (i.e., the models proposed by Dang et al. (2017)); *ii)* models defined according to equation (3) in which each response has its

specific predictors. From now on, such models have been denoted as `CW` and `SuCW`, respectively. A hundred datasets of $I = 450$ independent observations have been randomly generated from model (3) with $D = 2$ responses, $P = 3$ predictors and $K = 3$ components in which the elements of the conditional expected vector (2) are defined as follows:

$$E(Y_{i1}|\mathbf{X}_i = \mathbf{x}_i, \Omega_k) = \beta_{k10} + \beta_{k11}x_{i1} + \beta_{k12}x_{i2}, \quad (15)$$

$$E(Y_{i2}|\mathbf{X}_i = \mathbf{x}_i, \Omega_k) = \beta_{k20} + \beta_{k21}x_{i1} + \beta_{k22}x_{i3}. \quad (16)$$

Thus, the model employed to generate the datasets assumes that the first response $Y_1$ depends on $X_1$ and $X_2$ while $Y_2$ depends on $X_1$ and $X_3$. Furthermore, the component-covariance structures of both the predictors and the responses are defined using the `VVV` parameterisation. The specific values of the parameters for the data-generating model are: $\pi_1 = 0.4$, $\pi_2 = 0.35$, $\pi_3 = 0.25$, $\boldsymbol{\mu}_{\mathbf{X}_1} = (0, 0, 0)'$, $\boldsymbol{\mu}_{\mathbf{X}_2} = (2, 4, -2)'$, $\boldsymbol{\mu}_{\mathbf{X}_3} = \boldsymbol{\mu}_{\mathbf{X}_2} + 2\epsilon \cdot \mathbf{1}_P$, where $\mathbf{1}_P$ is the $P \times 1$ vector having each element equal to 1, $\boldsymbol{\beta}_1^* = (-2, 0.75, 1, 1, 0.5, -2)'$, $\boldsymbol{\beta}_2^* = (0.5, 1.75, 0.25, 1, 1, 1)'$, $\boldsymbol{\beta}_3^* = \boldsymbol{\beta}_2^* + \epsilon \cdot \mathbf{1}_6$, $\boldsymbol{\Sigma}_{\mathbf{X}_1} = \begin{pmatrix} 1.72 & -0.18 & 0.27 \\ -0.18 & 1.89 & 0.27 \\ 0.27 & 0.27 & 2.89 \end{pmatrix}$, $\boldsymbol{\Sigma}_{\mathbf{X}_2} = \begin{pmatrix} 2.33 & -0.52 & -0.06 \\ -0.52 & 0.88 & -0.34 \\ -0.06 & -0.34 & 1.04 \end{pmatrix}$, $\boldsymbol{\Sigma}_{\mathbf{X}_3} = \boldsymbol{\Sigma}_{\mathbf{X}_2}$, $\boldsymbol{\Sigma}_{\mathbf{Y}_1} = \begin{pmatrix} 1.34 & 0.47 \\ 0.47 & 1.66 \end{pmatrix}$, $\boldsymbol{\Sigma}_{\mathbf{Y}_2} = \begin{pmatrix} 0.50 & 0.04 \\ 0.04 & 1.50 \end{pmatrix}$, $\boldsymbol{\Sigma}_{\mathbf{Y}_3} = \boldsymbol{\Sigma}_{\mathbf{Y}_2}$. Since the second and third components of the data-generating model only differ in the values of intercepts and regression coefficients and the expected values of the regressors, the separation between such components depends on $\epsilon$. The simulated datasets have been generated using the following values of $\epsilon$: 0.275, 0.3, 0.325, 0.350 and 0.375; this allows an evaluation of the performances of the approaches based on `SuCW` and `CW` models under different experimental levels of separation between those components. Figures 1 and 2 show the scatterplots for two simulated datasets obtained with $\epsilon = 0.3$ and $\epsilon = 0.375$, respectively.

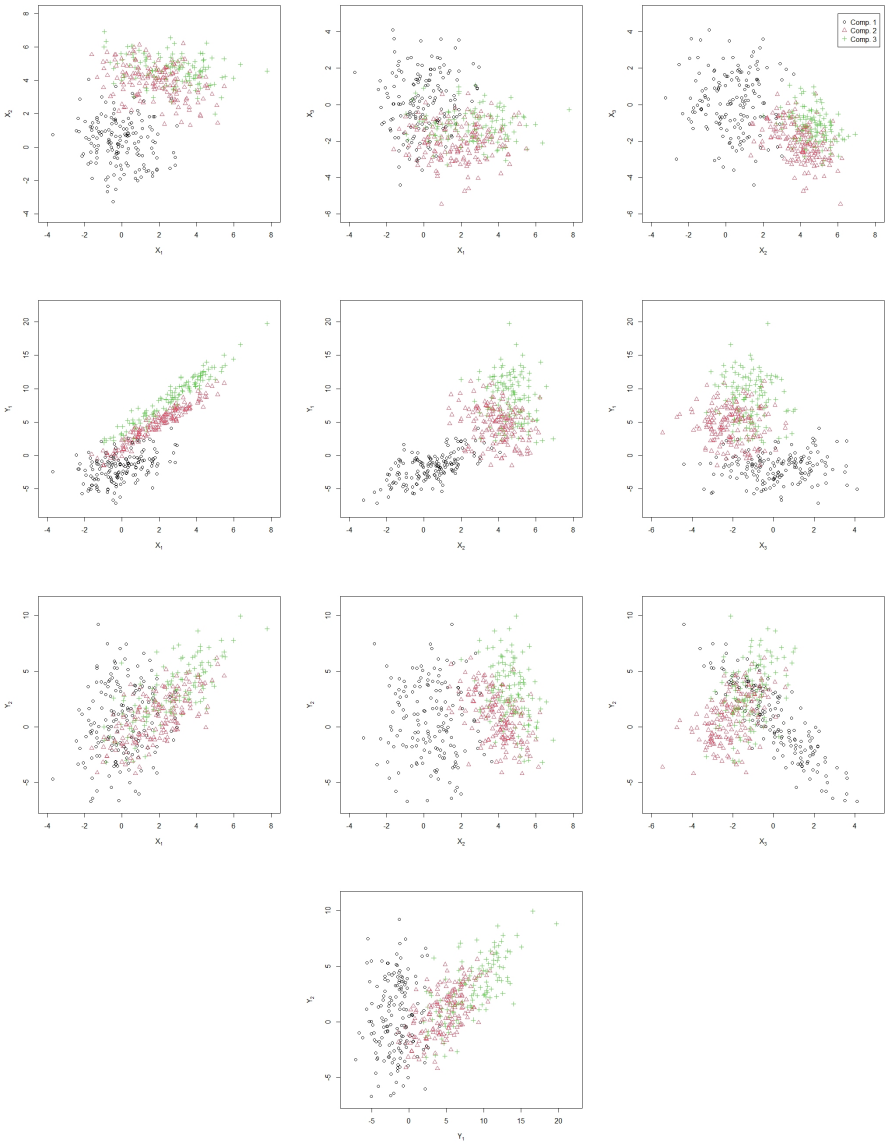


Figure 1: Bivariate scatterplots for pairs of variables in a simulated dataset, $\epsilon = 0.3$.

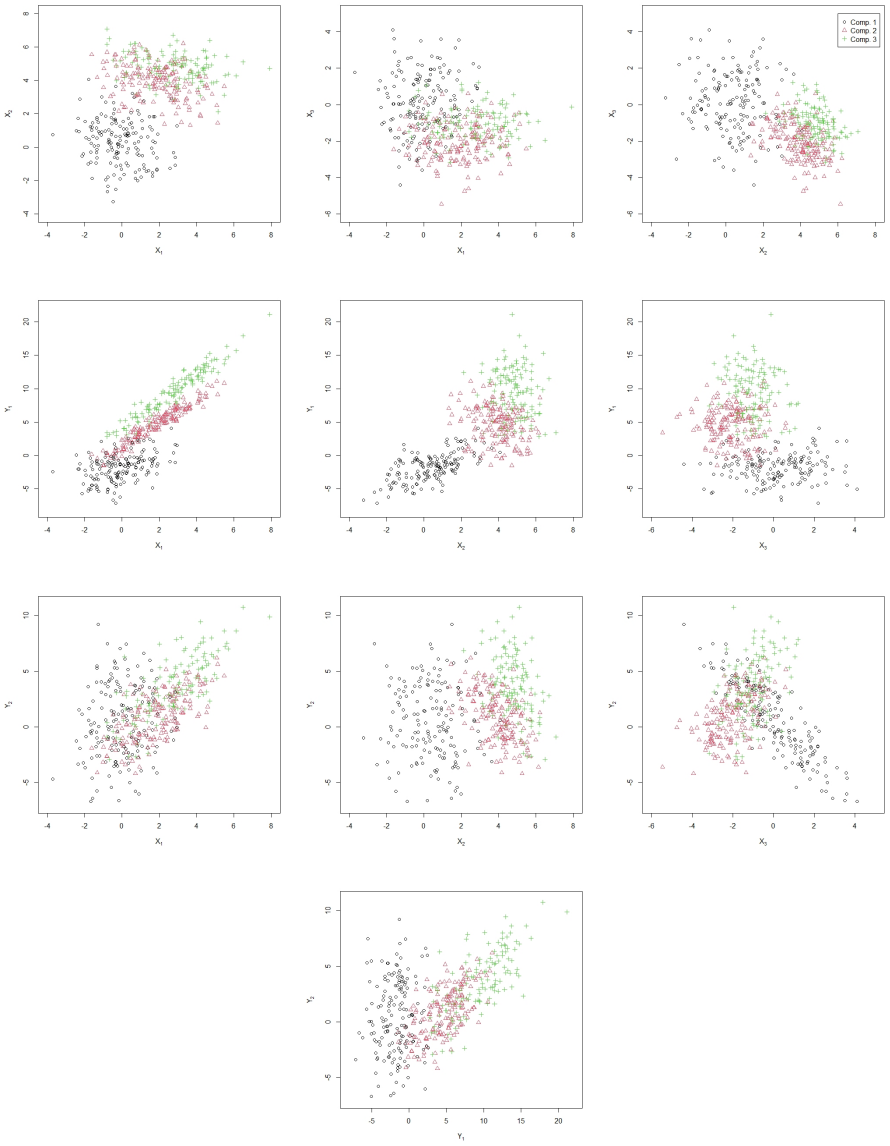


Figure 2: Bivariate scatterplots for pairs of variables in a simulated dataset, $\epsilon = 0.375$.

A first analysis has been carried out where the 196 SuCW and CW models with $K = 3$ components associated with all the parameterisations for the component-covariance structures of both the predictors and the responses have been fitted to each dataset. It is worth noting that using CW models leads to non-parsimonious specifications for such datasets, as six regression coefficients (two for each component) have been estimated although in fact they are equal to zero. The analysis has been run on an IBM x3750 M4 server with 4 Intel Xeon E5-4620 processors with 8 cores and 128GB RAM. The average execution times (over 100 datasets) for SuCW models have ranged between 2.698 and 35.309 seconds, depending on the specific combination of parameterisations for the component covariance matrices and the value of $\epsilon$. Concerning CW models, the minimum and maximum average execution times have resulted to be equal to 3.382 and 40.710 seconds, respectively. Since the implementation of the ECM algorithm has not been carried out with the goal of being efficient from a computational point of view, these CPU times are merely illustrative and can be reduced using more efficient implementations. For all the models fitted to any dataset, the value of $BIC$ has been computed and the models with the lowest $BIC$ within the two collections of fitted models have been selected. The 100 pairs of models selected as just illustrated, one for each simulated dataset, have been employed to compare the effectiveness of the two approaches. As expected, SuCW models have resulted to be preferable to CW ones. For each dataset $BIC_{\texttt{SuCW}} < BIC_{\texttt{CW}}$ for all the examined values of $\epsilon$ with the exception of two datasets when $\epsilon = 0.350$.

A further evaluation of the two approaches has been performed by examining their ability to recover the true values of the unknown parameters (i.e., parameter recovery). In particular, the attention has been focused on the bias and the root mean squared error (RMSE) for the regression coefficients in equations

(15) and (16). Namely, the following quantities have been computed

$$Bias\left(\hat{\beta}_{kdp}\right) = \left|\beta_{kdp} - \frac{\sum_{r=1}^{100}\hat{\beta}_{kdp}^{(r)}}{100}\right|,\ k=1,2,3,\ d=1,2,\ p=1,2,$$

$$RMSE\left(\hat{\beta}_{kdp}\right) = \sqrt{\frac{\sum_{r=1}^{100}\left(\beta_{kdp} - \hat{\beta}_{kdp}^{(r)}\right)^2}{100}},\ k=1,2,3,\ d=1,2,\ p=1,2,$$

where $\hat{\beta}_{kdp}^{(r)}$ is the ML estimate of $\beta_{kdp}$ obtained from the $r$th dataset ($r = 1,\ldots,100$). Note that `CW` models contain additional regression coefficients associated with the equation-specific irrelevant regressors. The bias and RMSE have been computed also for these additional coefficients, using 0 as their true value. Tables 2 and 3 report the values of bias and RMSE, respectively, obtained for each value of $\epsilon$. Overall, both approaches tend to provide acceptable results in terms of recovering the true values of the regression coefficients. This is evident for the parameters of the first component. As far as the second and third components are concerned, there seems to be a tendency for `SuCW` models to perform slightly better than `CW` models, especially considering the RMSE for low values of $\epsilon$. It is also worth noting that `CW` models appear to be capable of recognising the presence of irrelevant regressors, as the corresponding estimated regression coefficients are on average very close to 0. However, the RMSE of some of these estimates tend to be large, suggesting a a low precision in the estimation of the effect of some irrelevant regressors. This precision seems to improve as the separation among components increases.

The performance of the two approaches has also been evaluated by their ability to properly estimate the true classification of the sample observations (i.e., classification recovery). This task has been carried out by means of the adjusted Rand index ($ARI$) (Hubert and Arabie, 1985). Some summary statistics of this index (over the 100 datasets) for both approaches by the five examined levels of separation are reported in Table 4. These results show that the classification recovery associated with the use of both approaches increases with the level of separation between the second and third components (see the mean and median values of $ARI$ in Table 4); on the contrary, the interquartile range and

Table 2: Bias for the regression coefficients under `SuCW` and `CW` models in the first study.

| | $\epsilon = 0.275$ | | $\epsilon = 0.3$ | | $\epsilon = 0.325$ | | $\epsilon = 0.350$ | | $\epsilon = 0.375$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` |
| $\beta_{111} = 0.75$ | 0.005 | 0.003 | 0.005 | 0.004 | 0.006 | 0.005 | 0.007 | 0.005 | 0.006 | 0.005 |
| $\beta_{112} = 1$ | 0.003 | 0.001 | 0.002 | 0.003 | 0.003 | 0.003 | 0.002 | 0.005 | 0.002 | 0.003 |
| $\beta_{121} = 0.5$ | 0.010 | 0.013 | 0.005 | 0.007 | 0.005 | 0.008 | 0.002 | 0.006 | 0.005 | 0.008 |
| $\beta_{122} = -2$ | 0.000 | 0.005 | 0.002 | 0.004 | 0.002 | 0.004 | 0.002 | 0.004 | 0.002 | 0.004 |
| $\beta_{211} = 1.75$ | 0.027 | 0.090 | 0.011 | 0.023 | 0.001 | 0.010 | 0.037 | 0.002 | 0.003 | 0.002 |
| $\beta_{212} = 0.25$ | 0.067 | 0.160 | 0.023 | 0.040 | 0.009 | 0.022 | 0.010 | 0.023 | 0.000 | 0.002 |
| $\beta_{221} = 1$ | 0.028 | 0.039 | 0.019 | 0.011 | 0.013 | 0.009 | 0.007 | 0.008 | 0.009 | 0.007 |
| $\beta_{222} = 1$ | 0.004 | 0.026 | 0.028 | 0.042 | 0.027 | 0.023 | 0.027 | 0.053 | 0.028 | 0.028 |
| $\beta_{311} = 1.75 + \epsilon$ | 0.002 | 0.005 | 0.002 | 0.003 | 0.001 | 0.002 | 0.003 | 0.003 | 0.002 | 0.003 |
| $\beta_{312} = 0.25 + \epsilon$ | 0.052 | 0.100 | 0.014 | 0.034 | 0.010 | 0.016 | 0.012 | 0.016 | 0.004 | 0.017 |
| $\beta_{321} = 1 + \epsilon$ | 0.023 | 0.065 | 0.022 | 0.028 | 0.013 | 0.012 | 0.017 | 0.018 | 0.009 | 0.009 |
| $\beta_{322} = 1 + \epsilon$ | 0.004 | 0.020 | 0.013 | 0.015 | 0.016 | 0.007 | 0.009 | 0.011 | 0.020 | 0.021 |
| Irrelevant regressors | | | | | | | | | | |
| $\beta_{113} = 0$ | – | 0.007 | – | 0.003 | – | 0.003 | – | 0.001 | – | 0.003 |
| $\beta_{123} = 0$ | – | 0.022 | – | 0.021 | – | 0.021 | – | 0.022 | – | 0.021 |
| $\beta_{213} = 0$ | – | 0.060 | – | 0.016 | – | 0.015 | – | 0.006 | – | 0.002 |
| $\beta_{223} = 0$ | – | 0.001 | – | 0.011 | – | 0.005 | – | 0.004 | – | 0.003 |
| $\beta_{313} = 0$ | – | 0.067 | – | 0.040 | – | 0.028 | – | 0.023 | – | 0.019 |
| $\beta_{323} = 0$ | – | 0.082 | – | 0.011 | – | 0.002 | – | 0.007 | – | 0.003 |

Table 3: RMSE for the regression coefficients under `SuCW` and `CW` models in the first study.

| | $\epsilon = 0.275$ | | $\epsilon = 0.3$ | | $\epsilon = 0.325$ | | $\epsilon = 0.350$ | | $\epsilon = 0.375$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` |
| $\beta_{111} = 0.75$ | 0.058 | 0.059 | 0.058 | 0.059 | 0.059 | 0.059 | 0.060 | 0.060 | 0.058 | 0.060 |
| $\beta_{112} = 1$ | 0.075 | 0.075 | 0.068 | 0.069 | 0.069 | 0.069 | 0.069 | 0.069 | 0.068 | 0.069 |
| $\beta_{121} = 0.5$ | 0.096 | 0.094 | 0.075 | 0.077 | 0.075 | 0.077 | 0.077 | 0.078 | 0.075 | 0.077 |
| $\beta_{122} = -2$ | 0.055 | 0.063 | 0.054 | 0.057 | 0.054 | 0.057 | 0.054 | 0.057 | 0.054 | 0.057 |
| $\beta_{211} = 1.75$ | 0.264 | 0.374 | 0.087 | 0.121 | 0.058 | 0.088 | 0.487 | 0.135 | 0.045 | 0.045 |
| $\beta_{212} = 0.25$ | 0.182 | 0.316 | 0.116 | 0.166 | 0.090 | 0.127 | 0.131 | 0.142 | 0.073 | 0.074 |
| $\beta_{221} = 1$ | 0.211 | 0.339 | 0.096 | 0.120 | 0.080 | 0.105 | 0.111 | 0.117 | 0.067 | 0.073 |
| $\beta_{222} = 1$ | 0.437 | 0.548 | 0.126 | 0.181 | 0.113 | 0.130 | 0.440 | 0.341 | 0.103 | 0.111 |
| $\beta_{311} = 1.75 + \epsilon$ | 0.089 | 0.176 | 0.083 | 0.100 | 0.074 | 0.089 | 0.074 | 0.085 | 0.059 | 0.063 |
| $\beta_{312} = 0.25 + \epsilon$ | 0.185 | 0.252 | 0.130 | 0.167 | 0.117 | 0.152 | 0.125 | 0.153 | 0.104 | 0.127 |
| $\beta_{321} = 1 + \epsilon$ | 0.112 | 0.275 | 0.103 | 0.115 | 0.083 | 0.113 | 0.086 | 0.099 | 0.076 | 0.088 |
| $\beta_{322} = 1 + \epsilon$ | 0.149 | 0.197 | 0.131 | 0.150 | 0.125 | 0.176 | 0.132 | 0.145 | 0.119 | 0.135 |
| Irrelevant regressors | | | | | | | | | | |
| $\beta_{113} = 0$ | – | 0.057 | – | 0.048 | – | 0.048 | – | 0.052 | – | 0.048 |
| $\beta_{123} = 0$ | – | 0.082 | – | 0.081 | – | 0.081 | – | 0.080 | – | 0.080 |
| $\beta_{213} = 0$ | – | 0.179 | – | 0.106 | – | 0.099 | – | 0.081 | – | 0.062 |
| $\beta_{223} = 0$ | – | 0.214 | – | 0.200 | – | 0.156 | – | 0.126 | – | 0.125 |
| $\beta_{313} = 0$ | – | 0.353 | – | 0.139 | – | 0.108 | – | 0.101 | – | 0.089 |
| $\beta_{323} = 0$ | – | 0.294 | – | 0.163 | – | 0.172 | – | 0.162 | – | 0.163 |

Table 4: Summary statistics of the $ARI$ index under `SuCW` and `CW` models in the first study: mean, median, interquartile range (IQR) and standard deviation (SD). The p-values in the last row refer to the paired samples Wilcoxon test for the hypothesis of equality between $ARI$s for each $\epsilon$.

| | $\epsilon = 0.275$ | | $\epsilon = 0.3$ | | $\epsilon = 0.325$ | | $\epsilon = 0.350$ | | $\epsilon = 0.375$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` |
| Mean | 0.852 | 0.806 | 0.901 | 0.887 | 0.927 | 0.920 | 0.936 | 0.936 | 0.956 | 0.956 |
| Median | 0.878 | 0.862 | 0.911 | 0.901 | 0.930 | 0.929 | 0.945 | 0.947 | 0.959 | 0.960 |
| IQR | 0.062 | 0.192 | 0.035 | 0.043 | 0.030 | 0.034 | 0.026 | 0.028 | 0.020 | 0.023 |
| SD | 0.083 | 0.121 | 0.037 | 0.070 | 0.026 | 0.048 | 0.057 | 0.053 | 0.018 | 0.019 |
| p-value | $< 10^{-5}$ | | 0.0717 | | 0.573 | | 0.262 | | 0.935 | |

Table 5: Distributions of the number of components for the best `SuCW` and `CW` models in the second study.

| | $\epsilon = 0.275$ | | $\epsilon = 0.3$ | | $\epsilon = 0.325$ | | $\epsilon = 0.350$ | | $\epsilon = 0.375$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` |
| $K = 2$ | 48 | 99 | 7 | 84 | 1 | 39 | 1 | 10 | 0 | 1 |
| $K = 3$ | 52 | 1 | 93 | 16 | 99 | 61 | 98 | 90 | 100 | 99 |
| $K = 4$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

the standard deviation of $ARI$ seem to show a decreasing trend. Furthermore, `SuCW` models tend to be characterised by a greater ability to properly estimate the true classification of the sample observations for each examined value of $\epsilon$, even though the differences in terms of mean and median values of $ARI$ seem to vanish for larger values of $\epsilon$. This pattern is confirmed by the results of the paired samples Wilcoxon test, suggesting that the difference between the mean values of $ARI$ with the two approaches appears to be statistically significant only when the degree of separation is low (see the p-values in the last row of Table 4). This behaviour might be connected to the fact that, as the value of $\epsilon$ increases, not only the differences between the two approaches in recovering the actual values of the parameters tend to vanish, but also there is an improvement in the ability of `CW` models to recognise the presence of irrelevant regressors.

A second analysis has been carried out, where the performance of the two approaches has been evaluated without exploiting the knowledge of neither the number of components nor the parameterisation of the component-covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$ employed to generate the datasets. Thus, 597 different `SuCW` models have been estimated for each simulated dataset: 196 differentially parameterised models for each $K = 2, 3, 4$ and 9 models with $K = 1$. The same task has been carried out by employing `CW` models. Then, the best `SuCW` and `CW` models fitted to each dataset have been selected according to the $BIC$. Table 5 summarises the results of this procedure in terms of recovery of the true $K$. The impact of the value of $\epsilon$ on this aspect is evident. Generally speaking, the ability to select the correct value of $K$ improves as the separation increases. By focusing the attention on the distributions of the number of components for the best `CW` models fitted to the 100 datasets, it emerges that with such an approach the true number of components tends to be severely underestimated with the two lowest levels of separation ($\epsilon = 0.275, 3$). On the contrary, using `SuCW` models leads to the selection of the correct number of components for the majority of the simulated datasets with all levels of separation; furthermore, the proportion of datasets for which the selected `SuCW` model has three components increases quickly with $\epsilon$, reaching 93% when $\epsilon = 0.3$ and approaching nearly 100% for larger values of $\epsilon$.

In order to assess the possible consequences of a wrong choice of $K$ on the ability of `CW` models to recognise the presence of equation-specific irrelevant regressors, the biases of the estimates of the effects of these regressors have been computed for `CW` models with $K$ equal to 1, 2 and 4. According to the values reported in Table 6, it appears that the estimates of the regression coefficients for the irrelevant regressors can be severely biased when the number of components is lower than the true one. On the contrary, when the number of components exceeds the true $K$, the results are comparable with those obtained in the first analysis using models with $K = 3$ components (see the lower part of Table 2). It is also worth noting that for some coefficients the bias seems to show a trend which increases with the separation among components.

Table 6: Bias for the regression coefficients for equation-specific irrelevant regressors under `CW` models in the second study.

| | | $\epsilon = 0.275$ | $\epsilon = 0.3$ | $\epsilon = 0.325$ | $\epsilon = 0.350$ | $\epsilon = 0.375$ |
|---|---|---|---|---|---|---|
| $K = 1$ | $\beta_{113} = 0$ | 0.047 | 0.065 | 0.084 | 0.105 | 0.127 |
| | $\beta_{123} = 0$ | 0.525 | 0.504 | 0.483 | 0.462 | 0.440 |
| $K = 2$ | $\beta_{113} = 0$ | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| | $\beta_{123} = 0$ | 0.022 | 0.022 | 0.022 | 0.022 | 0.022 |
| | $\beta_{213} = 0$ | 0.377 | 0.428 | 0.479 | 0.529 | 0.578 |
| | $\beta_{223} = 0$ | 0.110 | 0.128 | 0.147 | 0.166 | 0.186 |
| $K = 4$ | $\beta_{113} = 0$ | 0.011 | 0.024 | 0.023 | 0.008 | 0.032 |
| | $\beta_{123} = 0$ | 0.034 | 0.047 | 0.043 | 0.035 | 0.027 |
| | $\beta_{213} = 0$ | 0.017 | 0.012 | 0.004 | 0.014 | 0.013 |
| | $\beta_{223} = 0$ | 0.095 | 0.015 | 0.011 | 0.093 | 0.065 |
| | $\beta_{313} = 0$ | 0.056 | 0.017 | 0.010 | 0.011 | 0.012 |
| | $\beta_{323} = 0$ | 0.009 | 0.056 | 0.004 | 0.011 | 0.004 |
| | $\beta_{413} = 0$ | 0.070 | 0.017 | 0.026 | 0.024 | 0.018 |
| | $\beta_{423} = 0$ | 0.057 | 0.024 | 0.013 | 0.013 | 0.000 |

As far as the classification recovery is concerned, the obtained results demonstrates that the ability to estimate the true classification of the sample observations with both approaches increases with $\epsilon$. However, the gap between the two approaches in terms of mean and median $ARI$ is quite large and statistically significant for the three smallest values of $\epsilon$ (see the Table 7). It is worth noting that the behaviour of the variability of the $ARI$ index is strictly related to the variability in the distribution of the optimal value of $K$ selected according to the $BIC$. In summary, the obtained results seem to suggest that the inclusion of the regressor $X_3$ in the equation (15) and the regressor $X_2$ in the equation (16) has a negative impact both on the choice of the correct number of components and on the reconstruction of the true classification of the sample observations. However, the consequences of including these irrelevant regressors seem to become negligible as the separation among components increases. A possible explanation of this behaviour could be related to the fact that the clustering task is eased when the components are well-separated. In such situations, even if `CW` models are non-parsimonious, they can lead to the correct choice of $K$. As a consequence, they are able to provide estimates for the regression coefficients of irrelevant regressors that are sufficiently close to zero, so that the inclusion of such regressors has little effect on the estimated posterior probabilities employed to classify the sample observations.

## 4. Results from the analysis of real data

Two real situations have been examined to evaluate the practical usefulness of `SuCW` models in comparison with `CW` models. For both these model classes, models have been estimated for $K$ from 1 to 9. For each of these values, all possible parsimonious `CW` and `SuCW` models have been fitted (see Section 2.7). Analyses of the examined real datasets have been carried out also through the clusterwise regression models described in Section 2.2. Namely, the comparison with models (5) allows to assess the adequacy of the assignment independence assumption. Furthermore, from the comparison with models (6) it is possi-

Table 7: Summary statistics of the *ARI* index for the best model `SuCW` and `CW` model in the second study: mean, median, interquartile range (IQR) and standard deviation (SD). The p-values in the last row refer to the paired samples Wilcoxon test for the hypothesis of equality between *ARI*s for each $\epsilon$.

| | $\epsilon = 0.275$ | | $\epsilon = 0.3$ | | $\epsilon = 0.325$ | | $\epsilon = 0.350$ | | $\epsilon = 0.375$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` | `SuCW` | `CW` |
| Mean | 0.766 | 0.644 | 0.885 | 0.687 | 0.924 | 0.820 | 0.939 | 0.916 | 0.956 | 0.953 |
| Median | 0.785 | 0.644 | 0.905 | 0.652 | 0.929 | 0.904 | 0.945 | 0.945 | 0.959 | 0.960 |
| IQR | 0.236 | 0.036 | 0.041 | 0.048 | 0.030 | 0.273 | 0.026 | 0.030 | 0.020 | 0.023 |
| SD | 0.120 | 0.034 | 0.071 | 0.103 | 0.035 | 0.138 | 0.038 | 0.087 | 0.018 | 0.036 |
| p-value | $< 10^{-10}$ | | $< 10^{-15}$ | | $< 10^{-8}$ | | 0.022 | | 0.851 | |

ble to establish whether fitting multivariate cluster-weighted models based on seemingly unrelated linear regression for $D$ responses leads to an improvement over an approach based on $D$ univariate seemingly unrelated linear clusterwise regression models. From now on, models (5) and (6) are denoted as `SuCR` and `uSuCR`, respectively. Models from equations (5) and (6) have been estimated also using the same vector of covariates for all responses (i.e., with $\mathbf{x}_{id} = \mathbf{x}_i$ $\forall d$); in the following, they are denoted as `CR` and `uCR`, respectively. All these clusterwise regression models have been fitted for a number of components from 1 to 9 through a specific function developed in the `R` environment which also allows the estimation of seemingly unrelated linear parsimonious clusterwise models (for more details see Galimberti and Soffritti, 2020). Parameters $\boldsymbol{\mu}_{\mathbf{X}}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$ of the Gaussian distribution for the covariates in models (5) and (6) have been estimated under three possible structures of $\boldsymbol{\Sigma}_{\mathbf{X}}$: fully unconstrained (`VV`), diagonal with $P$ unequal variances (`VI`) and diagonal with equal variances (`EI`). As far as the variances $\sigma_{k_d d}$, $k_d = 1, \ldots, K_d$, in the univariate clusterwise regression models are concerned, the estimation has been carried out under both an homoscedastic (`E`) and heteroscedastic (`V`) assumption.

### *4.1. Canned tuna sales in USA*

Data taken from Chevalier et al. (2003) and available within the `R` package `bayesm` (Rossi, 2012) provides information about seven of the top 10 U.S. brands in the canned tuna product category for $I = 338$ weeks between September 1989 and May 1997 (`tuna` dataset). The available information is the volume of weekly sales (`Move`), a measure of the display activity (`Nsale`) and the log price (`Lprice`) of each brand. Analyses illustrated here have been focused on $D = 2$ products: Bumble Bee Chunk 6.12 oz. (BBC) and Bumble Bee Solid 6.12 oz. (BBS). A previous study about the effect of prices and promotional activities on sales for these two products, based on clusterwise linear regression models (Galimberti and Soffritti, 2020), demonstrated that the effect of log price on log unit sales is not homogeneous during the examined period of time for both products. Furthermore, a search for the predictors to be employed in the two regression equations showed that models including only the log unit prices should be preferred. Thus, the analysis here has been focused on four variables: $\mathbf{Y} = (\texttt{Lmove BBC}, \texttt{Lmove BBS})'$, $\mathbf{X} = (\texttt{Lprice BBC}, \texttt{Lprice BBS})'$, where `Lmove` denotes the logarithm of `Move`. As typically happens with food prices, also prices of BBC and BBS appear to change according to an almost discrete grid of values (see the scatterplot on the left part of Figure 3). Although the Pearson's correlation coefficient between the two responses is low (0.1844), according to the Student's $t$ test the hypothesis of linear independence between `Lmove BBC` and `Lmove BBS` has to be rejected; `Lmove BBC` results to be negatively and strongly correlated with `Lprice BBC`; there is also a negative and significant linear dependence between `Lmove BBS` and the logarithm of the prices for both products (see Table 8).

By assuming that prices for each of the two examined products can only affect sales of the same product, `SuCW`, `SuCR` and `uSuCR` models have been specified by using `Lprice BBS` as regressor in the equation for `Lmove BBS`, `Lprice BBC` as regressor for `Lmove BBC`. Table 9 reports the models which best fit the `tuna` dataset according to the $BIC$ for each combination of the nine examined values of $K$ and each of the model types `SuCW` and `CW`. All these models have

Table 8: Pearson's correlation matrix (lower diagonal part) and p-values of the Student's $t$ test for the hypothesis of linear independence between variables (upper diagonal part) from the `tuna` dataset.

| | `Lmove BBC` | `Lmove BBS` | `Lprice BBC` | `Lprice BBS` |
|---|---|---|---|---|
| `Lmove BBC` | 1.0000 | 0.0007 | $< 10^{-67}$ | 0.2678 |
| `Lmove BBS` | 0.1844 | 1.0000 | 0.0011 | $< 10^{-8}$ |
| `Lprice BBC` | $-0.7727$ | $-0.1767$ | 1.0000 | 0.4420 |
| `Lprice BBS` | $-0.0604$ | $-0.3172$ | 0.0420 | 1.0000 |

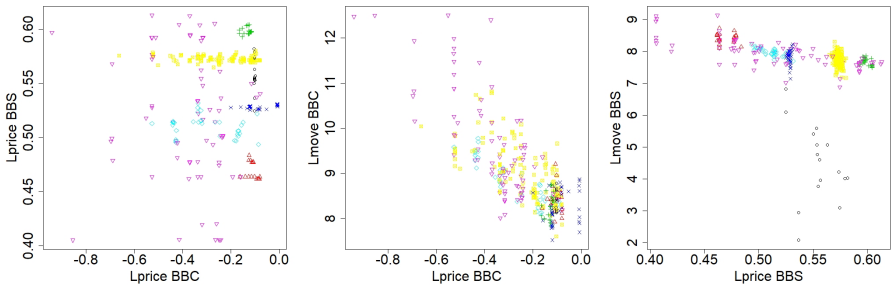


Figure 3: Scatterplots for three pairs of variables from the analysis of the `tuna` dataset. Observations are pictured with seven different colours and symbols according to the classification obtained from the best model.

Table 9: Best models fitted to the `tuna` dataset within some given model classes and their $BIC$ values. Columns $acr.X$ and $acr.Y$ report the acronyms of the parsimonious paramaterisations for the component-covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

| Model | $K$ | $acr.X$ | $acr.Y$ | $BIC_M$ | Model | $K$ | $acr.X$ | $acr.Y$ | $BIC_M$ |
|---|---|---|---|---|---|---|---|---|---|
| SuCW | 1 | EEI | EEI | −18.9 | CW | 1 | EEI | EEI | −18.1 |
| SuCW | 2 | VVE | EVV | −812.2 | CW | 2 | VVE | EVV | −794.2 |
| SuCW | 3 | VVI | EVV | −929.1 | CW | 3 | VEV | VVE | −922.3 |
| SuCW | 4 | VVE | VVE | −1195.0 | CW | 4 | VVE | VVE | −1157.4 |
| SuCW | 5 | VVI | VEV | −1282.0 | CW | 5 | VVI | VVE | −1267.1 |
| SuCW | 6 | VVI | VEV | −1355.2 | CW | 6 | VVI | VVE | −1333.7 |
| SuCW | 7 | VVI | VEV | −1389.8 | CW | 7 | VVI | VVE | −1331.4 |
| SuCW | 8 | VVI | VEV | −1387.2 | CW | 8 | VVI | VVE | −1341.3 |
| SuCW | 9 | VVI | VEV | −1371.1 | CW | 9 | VVI | VVI | −1326.4 |

been estimated within a limit of 237 iterations of the ECM algorithm. Figure 4 shows the values of the $BIC$ for the best `CW` and `SuCW` models by $K$. As far as the clusterwise regression models are concerned, Table 10 summarises some information about the best fitted models within each of the model classes `SuCR`, `CR`, `uSuCR` and `uCR` obtained from equations (5) and (6). Overall, it seems that the best trade-off between the fit and complexity can be obtained using the `SuCW` model with $K = 7$ clusters of weeks. The convergence of the ECM algorithm for the parameter estimation has been reached after 53 iterations. For the clusters detected by this model, the distributions of the two regressors are diagonal with variable volumes and shapes. As far as the joint conditional distributions of the two responses given the corresponding regressors are concerned, clusters are characterized by ellipsoidal distributions with variable volumes and orientations and equal shape.

The first cluster is composed of 16 consecutive weeks corresponding to the period from end-October 1990 to mid-February 1991 (see the additional information about this dataset available at the University of Chicago website http://research.chicagobooth.edu/kilts/marketing-databases/dominicks). In that

Table 10: Best models fitted to the `tuna` dataset within the model classes defined from equations (5) and (6) and their $BIC$ values.

| Model | Best fitted model | $BIC_M$ |
|---|---|---|
| $\phi_P\left(\mathbf{x}_i;\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\Sigma}_{\mathbf{X}}\right)$ | $M_a$: $acr.X =$ `VI` | $-1408.0$ |
| $\sum_{k=1}^{K}\pi_k\phi_D\left(\mathbf{y}_i\vert\mathbf{x}_i;\mathcal{X}_i'\boldsymbol{\beta}_k^*,\boldsymbol{\Sigma}_{\mathbf{Y}_k}\right)$ | $M_b$: $K=3$, $acr.Y =$ `EVV` | 652.5 |
| $\sum_{k=1}^{K}\pi_k\phi_D\left(\mathbf{y}_i\vert\mathbf{x}_i;\mathbf{B}_k'\mathbf{x}_i^*,\boldsymbol{\Sigma}_{\mathbf{Y}_k}\right)$ | $M_c$: $K=3$, $acr.Y =$ `VVE` | 683.4 |
| $\sum_{k=1}^{K}\pi_k\phi_1\left(y_{i1}\vert\mathbf{x}_i;\mathbf{x}_{i1}^{*\prime}\boldsymbol{\beta}_{k1}^*,\sigma_{k1}^2\right)$ | $M_d$: $K=2$, $acr.Y =$ `V` | 496.1 |
| $\sum_{k=1}^{K}\pi_k\phi_1\left(y_{i1}\vert\mathbf{x}_i;\boldsymbol{\beta}_{k1}^{*\prime}\mathbf{x}_i^*,\sigma_{k1}^2\right)$ | $M_e$: $K=2$, $acr.Y =$ `V` | 505.4 |
| $\sum_{k=1}^{K}\pi_k\phi_1\left(y_{i2}\vert\mathbf{x}_i;\mathbf{x}_{i2}^{*\prime}\boldsymbol{\beta}_{k2}^*,\sigma_{k2}^2\right)$ | $M_f$: $K=2$, $acr.Y =$ `V` | 162.0 |
| $\sum_{k=1}^{K}\pi_k\phi_1\left(y_{i2}\vert\mathbf{x}_i;\boldsymbol{\beta}_{k2}^{*\prime}\mathbf{x}_i^*,\sigma_{k2}^2\right)$ | $M_g$: $K=2$, $acr.Y =$ `V` | 164.7 |
| `SuCR` | $M_a$ and $M_b$ | $-755.5$ |
| `CR` | $M_a$ and $M_c$ | $-724.6$ |
| `uSuCR` | $M_a$, $M_d$ and $M_f$ | $-749.9$ |
| `uCR` | $M_a$, $M_e$ and $M_g$ | $-737.9$ |

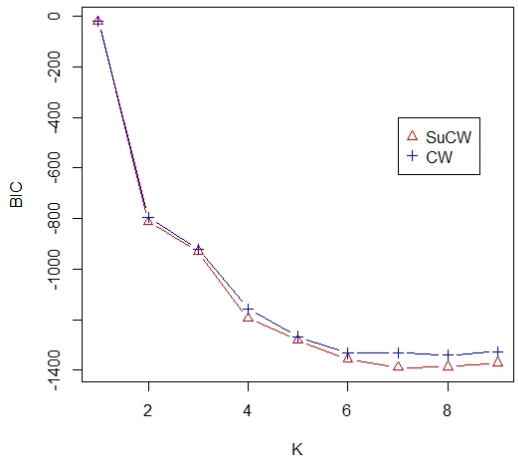


Figure 4: `Tuna` dataset: BIC values of the best `CW` and `SuCW` models by number of components.

Table 11: Estimated $\pi_k$, $\boldsymbol{\mu}_{\mathbf{X}_k}$ and $\boldsymbol{\beta}_k^*$ of the best model fitted to the `tuna` dataset. $\mathbf{a}[l]$ denotes the $l$th element of vector $\mathbf{a}$.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\hat{\pi}_k$ | 0.047 | 0.063 | 0.080 | 0.098 | 0.104 | 0.254 | 0.354 |
| $\hat{\boldsymbol{\mu}}_{\mathbf{X}_k}[1]$ | −0.103 | −0.106 | −0.130 | −0.085 | −0.297 | −0.366 | −0.245 |
| $\hat{\boldsymbol{\mu}}_{\mathbf{X}_k}[2]$ | 0.554 | 0.468 | 0.599 | 0.528 | 0.511 | 0.520 | 0.573 |
| $\hat{\boldsymbol{\beta}}_{k1}^*[1]$ | 7.946 | 8.506 | 8.925 | 8.203 | 7.844 | 7.751 | 8.239 |
| $\hat{\boldsymbol{\beta}}_{k1}^*[2]$ | −4.493 | 0.801 | 5.187 | −0.421 | −3.544 | −5.090 | −3.173 |
| $\hat{\boldsymbol{\beta}}_{k2}^*[1]$ | 17.051 | 8.820 | 16.503 | −4.040 | 13.142 | 10.504 | 20.766 |
| $\hat{\boldsymbol{\beta}}_{k2}^*[2]$ | −22.692 | −0.855 | −14.724 | 22.436 | −10.115 | −4.920 | −22.712 |

period a worldwide boycott campaign (promoted by the U.S. nongovernmental organisation Earth Island Institute) encouraged consumers not to buy Bumble Bee tuna because Bumble Bee was found to be buying yellow-fin tuna caught by dolphin-unsafe techniques (Baird and Quastel, 2011). The negative impact of such a campaign on Bumble Bee tuna sales appears to be evident for BBS (see the black points in the scatterplot of Figure 3 for this product). The mean prices of both products in the weeks of this cluster are quite high (see the first column in Table 11). Furthermore, prices of BBC in this cluster are highly homogeneous, as suggested by the low variance of `Lprice BBC` (not reported here). Finally, the effect of prices on sales in the same weeks is negative and particularly strong for BBS (see Table 11). The second cluster comprises 22 weeks (red points in the scatterplots of Figure 3), some of which are in close correspondence with Easter 1990 and 1991, Christmas 1993, Presidents day and Labor day 1994. They are mainly characterized by the lowest mean price of BBS and a negligible impact of prices on sales for both products. Furthermore, prices of BBS in such weeks result to be quite homogeneous. Cluster 3 is composed of 27 weeks (green points in Figure 3) with the highest mean price of BBS. In this cluster, the effects of prices on sales are negative for BBS and positive for BBC; furthermore, prices of both products are homogeneous. The special events corresponding to the weeks of cluster 3 are: Memorial days 1994 and 1995, 4th

of July 1994 and 1995, Halloween and Thanksgiving 1994. Cluster 4 is mainly composed of weeks from end-November 1995 to end-April 1997; two distinctive features of this cluster (34 weeks, dark blue points) are that it shows the highest mean price of BBC and highly homogeneous prices of BBS; furthermore, the estimated effect of prices on sales of BBS is positive and particularly strong. Labor day 1991, January 1992, Memorial days 1992 and 1993 are the events and periods associated with the weeks in cluster 5 (37 weeks, sky-blue points), which is characterized by intermediate mean prices and mild negative effects of prices on sales for both products. As far as clusters 6 and 7 are concerned, they contain 78 (purple points) and 124 (yellow points) weeks, respectively. The main distinctive feature of cluster 6 is that the variances of `Lprice BBC` and `Lprice BBS` (not reported here) are extremely large; furthermore, this cluster registers the lowest mean price of BBC. Cluster 6 mainly comprises weeks from mid-September 1991 to end-December 1991, January 1993, and the periods associated with Christmas 1992, Presidents day 1992 and 1993, Easter 1992, 1993 and 1995. Weeks belonging to cluster 7 are characterized by high and highly homogeneous prices of BBS; furthermore, the effect of prices on sales of BBS in these weeks is negative and particularly strong. In summary, by focusing the attention on the estimated regression coefficients of the seven clusters of weeks detected by the model, the main interesting findings are a clear evidence of differential effects of the log prices on the log unit sales for both products and the identification of two clusters in which such effects are positive for either BBS or BBC. The overall agreement between this partition and the one produced by the best `CW` model, which is composed of 8 clusters (see Table 12), is high ($ARI = 0.8293$): weeks have been classified in almost the same way by the two approaches; some exceptions mainly involve the sixth cluster of the partition illustrated above.

The comparison between these results and those produced from the best fitted linear clusterwise regression model (see the `SuCR` model in Table 10) shows that in the analysed dataset there is an additional source of heterogeneity over time, which appears to lie mainly in the prices of BBC tuna. Thus, when

Table 12: Cross-classification of the observations from the `tuna` dataset, based on the maximum posterior probabilities estimated from the best `CW` and `SuCW` fitted models. Labels for clusters reported in rows and colums refer to `CW` and `SuCW`, respectively.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 6 | 0 |
| 2 | 16 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 21 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 27 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 33 | 0 | 1 | 0 |
| 6 | 0 | 1 | 0 | 1 | 36 | 21 | 0 |
| 7 | 0 | 0 | 0 | 0 | 1 | 48 | 2 |
| 8 | 0 | 0 | 0 | 0 | 0 | 2 | 122 |

modelling the joint distribution of prices and sales for both products, more clusters have been detected (7 instead of 3). A further difference between the results obtained from these two approaches is that all the effects of log prices on the log unit sales for both products results to be negative within each cluster identified by the best linear clusterwise regression model. It is also worth noting that there is an almost perfect correspondence between one of the three clusters identified through the best linear clusterwise regression model and the first cluster described above (see Galimberti and Soffritti, 2020, for more details on the results obtained from the analysis of these data through the clusterwise regression approach).

As illustrated in Section 2.1, an underlying assumption of the best fitted model is that both $\mathbf{X}|\Omega_k$ and $\mathbf{Y}|(\mathbf{X}=\mathbf{x},\Omega_k)$ follow a multivariate normal distribution for $k=1,\ldots,K$. An evaluation of the adequacy of such an assumption for the examined dataset has been carried out by resorting to some measures of multivariate skewness and kurtosis (Mardia, 1970, 1974); by exploiting their asymptotic distribution derived under the the hypothesis of multivariate normality, those measures can also be employed as statistics for testing the hypothesis of multivariate normality. Namely, the function `mult.norm` of the `R`

Table 13: P-values of Mardia's skewness and kurtosis statistics for the residuals $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}_k}$ and $\mathbf{y} - \mathcal{X}'\boldsymbol{\beta}_k$, $k = 1, \ldots, K$, computed from the best model fitted to the `tuna` dataset.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}_k}$ | | | | | | | |
| skewness | 0.2654 | 0.1887 | 0.0124 | 0.0370 | 0.7230 | 0.0013 | 0.0000 |
| kurtosis | 0.9963 | 0.3052 | 0.6470 | 0.3134 | 0.0510 | 0.0777 | 0.0371 |
| $\mathbf{y} - \mathcal{X}'\boldsymbol{\beta}_k$ | | | | | | | |
| skewness | 0.0086 | 0.3526 | 0.7434 | 0.0829 | 0.8385 | 0.0002 | 0.0185 |
| kurtosis | 0.1308 | 0.5668 | 0.2121 | 0.3235 | 0.0483 | 0.0008 | 0.3927 |

package `QuantPsych` (Fletcher, 2012) has been employed to compute the values of such measures within each cluster detected by the best model from the estimated residuals $\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\mathbf{X}_k}$ and $\mathbf{y}_i - \mathcal{X}_i'\hat{\boldsymbol{\beta}}_k$ $\forall (i,k) \in \{(i,k), i \in \{1, \ldots, I\}, k = \arg\max_h\{\hat{\tau}_{ih}, h = 1, \ldots, K\}\}$; the p-values associated with the so obtained results are summarised in Table 13. Based on these findings, in the first five clusters the null hypothesis of multivariate normality should not be rejected at a Bonferroni-corrected $0.05/7 = 0.0071$ significance level. On the contrary, both types of residuals clearly deviate from the multivariate normality within the sixth cluster. As far as the seventh cluster is concerned, the null hypothesis should be rejected only for the residuals $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}_k}$.

### *4.2. Regional tourism in Italy*

In line with studies aiming at evaluating the link between tourism flows and attendance at museums and monuments (see, e.g., Cellini and Cuccia, 2013), the data analysed here provides information about tourist arrivals (denoted `Arriv`), tourist overnights (`Overn`) and visits to State museums, monuments and museum networks (`Visit`) with a monthly frequency over the period January 1999 to December 2017 in two Italian regions: Emilia Romagna (`ER`) and Veneto (`Ve`). Data concerning `Visit` has been obtained from the website of the Italian Min-

istry of Cultural Heritage[1]; the sources for `Arriv` and `Overn` are the websites of the two regional governments[2]. In this dataset the average stays (`AvStay`), computed as the ratio between `Overn` and `Arriv`, are also provided. Thus, the dataset is composed of $I = 228$ monthly observations for eight variables; from now on, it has been denoted as `RtI`. The goal of the analysis is to study the effect of the tourist arrivals and average stays on the visits to State museums, monuments and museum networks in Emilia-Romagna and Veneto. Thus, in this analysis $\mathbf{Y} = (\texttt{Visit ER}, \texttt{Visit Ve})'$, $\mathbf{X} = (\texttt{Arriv ER}, \texttt{AvStay ER}, \texttt{Arriv Ve}, \texttt{AvStay Ve})'$. The analysis has been performed using data in thousands. Figure 5 shows the bivariate scatterplots for pairs of regressors and pairs composed of one response and one regressor; month abbreviations are used as labels for the observations. Visits to to State museums, monuments and museum networks in the two regions result to be highly linearly dependent (see Table 14); high and positive pairwise correlations also characterise tourist arrivals and average stays in either region; the hypothesis of linear independence is not rejected between `Visit ER` and the average stays; the same result holds true also for `Visit Ve`.

Table 14: Pearson's correlation matrix (lower diagonal part) and p-values of the Student's $t$ test for the hypothesis of linear independence between variables (upper diagonal part) from the `RtI` dataset.

| | `Visit ER` | `Visit Ve` | `Arriv ER` | `AvStay ER` | `Arriv Ve` | `AvStay Ve` |
|---|---|---|---|---|---|---|
| `Visit ER` | 1.0000 | $< 10^{-66}$ | 0.0002 | 0.3684 | 0.0003 | 0.3879 |
| `Visit VE` | 0.8562 | 1.0000 | $< 10^{-8}$ | 0.7807 | $< 10^{-9}$ | 0.9572 |
| `Arriv ER` | 0.2421 | 0.3722 | 1.0000 | $< 10^{-53}$ | $< 10^{-166}$ | $< 10^{-47}$ |
| `AvStay ER` | −0.0598 | 0.0185 | 0.8081 | 1.0000 | $< 10^{-45}$ | $< 10^{-175}$ |
| `Arriv Ve` | 0.2394 | 0.4015 | 0.9826 | 0.7696 | 1.0000 | $< 10^{-40}$ |
| `AvStay Ve` | −0.0575 | −0.0036 | 0.7833 | 0.9856 | 0.7456 | 1.0000 |

[1] `http://www.statistica.beniculturali.it`.

[2] `https://statistica.regione.emilia-romagna.it/turismo`, `https://www.veneto.eu/web/area-operatori/statistiche`.

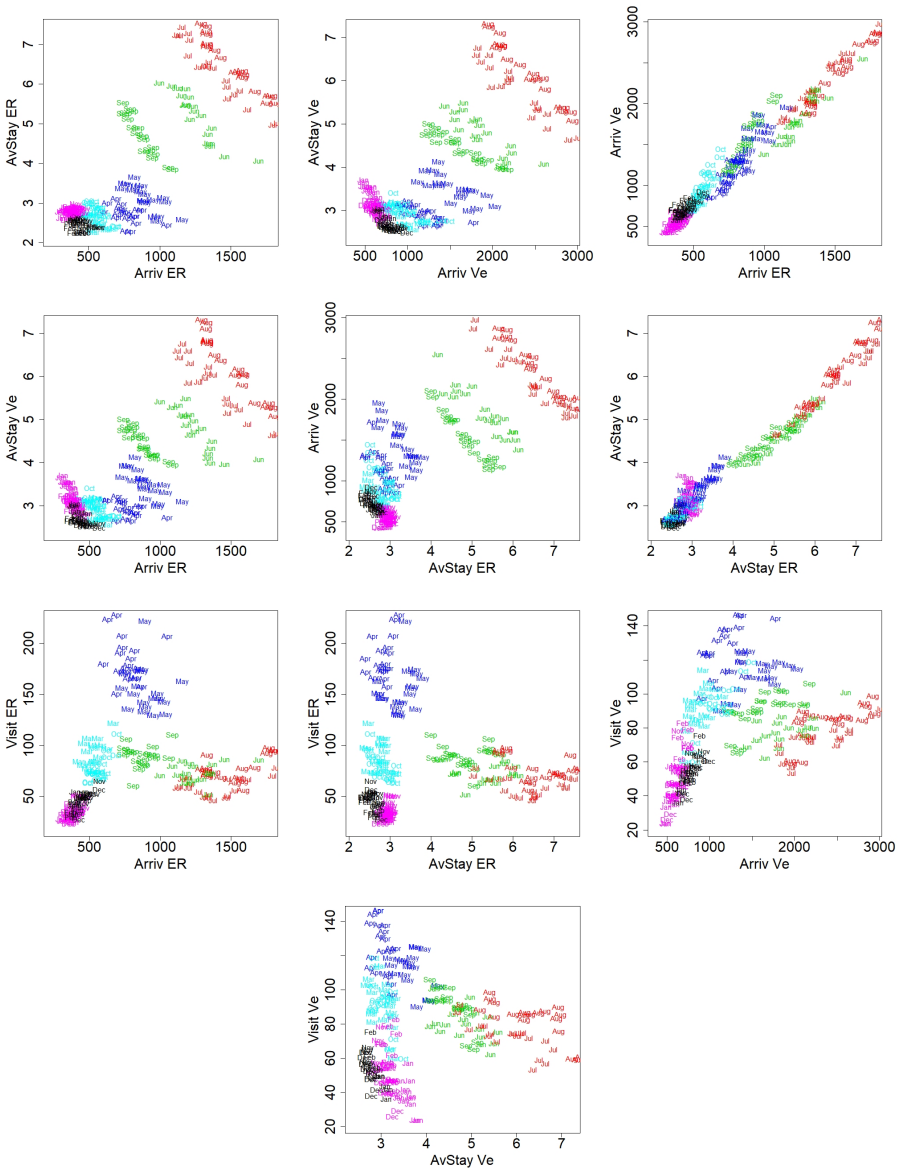


Figure 5: Bivariate scatterplots for pairs of variables in the analysis of the `RtI` dataset. Month abbreviations are used as labels. Observations are coloured according to the classification obtained from the best model.

Table 15: Best models fitted to the `RtI` dataset within some given model classes and their $BIC$ values. Columns $acr.X$ and $acr.Y$ report the acronyms of the parsimonious paramaterisations for the component-covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

| Model | $K$ | $acr.X$ | $acr.Y$ | $BIC_M$ | Model | $K$ | $acr.X$ | $acr.Y$ | $BIC_M$ |
|---|---|---|---|---|---|---|---|---|---|
| `SuCW` | 1 | `EEE` | `EEE` | 10948.6 | `CW` | 1 | `EEE` | `EEE` | 10965.7 |
| `SuCW` | 2 | `VVV` | `EEV` | 10223.8 | `CW` | 2 | `VVV` | `EEV` | 10146.2 |
| `SuCW` | 3 | `VVV` | `VVV` | 9991.5 | `CW` | 3 | `VVV` | `VVE` | 9934.6 |
| `SuCW` | 4 | `VVV` | `VEV` | 9898.8 | `CW` | 4 | `VVV` | `VEE` | 9886.6 |
| `SuCW` | 5 | `EVV` | `VEE` | 9822.3 | `CW` | 5 | `VVV` | `VII` | 9751.2 |
| `SuCW` | 6 | `VVV` | `VVI` | 9716.5 | `CW` | 6 | `VVV` | `VEE` | 9788.3 |
| `SuCW` | 7 | `VVV` | `VEV` | 9736.9 | `CW` | 7 | `VVV` | `VEE` | 9799.4 |
| `SuCW` | 8 | `EVV` | `VEV` | 9796.0 | `CW` | 8 | `EVV` | `VEE` | 9861.2 |
| `SuCW` | 9 | `VVV` | `VEV` | 9815.0 | `CW` | 9 | `VVV` | `VEE` | 9917.1 |

A first analysis has been performed by assuming that arrivals and average stays in each of the two regions can only affect attendance at museums and monuments of the same region. Thus, `SuCW` models have been specified by using `Arriv ER` and `AvStay ER` as regressors in the equation for `Visit ER`, `Arriv Ve` and `AvStay Ve` as regressors for `Visit Ve`. However, since Emilia-Romagna and Veneto are neighboring regions, arrivals and average stays in one region could also have an impact on the visits to State museums and monuments of the other region, hence the second analysis has been carried out through `CW` models. Table 15 provides information about the models which best fit the `RtI` dataset according to the $BIC$ for each combination of the nine examined values of $K$ and the two fitted model types. The convergence of the ECM algorithm for the estimation of these models has been reached within a limit of 161 iterations. Figure 6 shows the values of the $BIC$ for the best `CW` and `SuCW` models by $K$. Table 16 provides a summary of the results obtained from the best fitted models within each of the model classes defined by equations (5) and (6). Overall, the model with the best trade-off between the fit and complexity seems to be the `SuCW` model with $K = 6$ clusters of months. The ECM algorithm for the

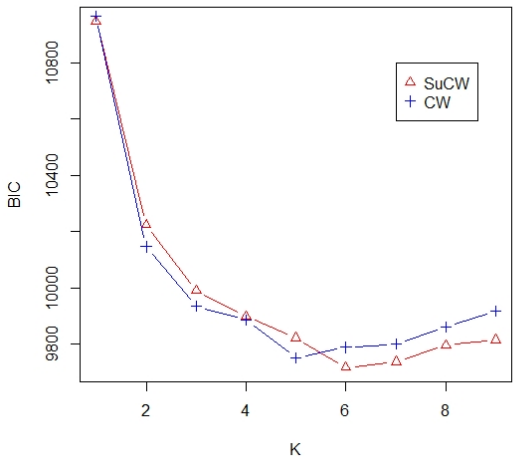


Figure 6: `RtI` dataset: BIC values of the best `CW` and `SuCW` models by number of components.

estimation of this model has reached the convergence after 47 iterations. For the resulting clusters, the four regressors have ellipsoidal distributions with variable volumes, shapes and orientations. As far as the joint conditional distributions of the two responses given the corresponding regressors are concerned, clusters show diagonal distributions with variable volumes and shapes, suggesting that `Visit Ve` and `Visit ER` are independent, conditionally on the regressors and cluster membership.

Four clusters are perfectly related to some months (see Table 17). They are:
cluster 2: observations in July and August;
cluster 3: observations in June and September;
cluster 4: observations in April and May;
cluster 5: observations in March and October.
As far as the months from November to February are concerned, observations from January 1999 to November 2010 and those of January 2011 and February 2012 have been assigned to cluster 6; cluster 1 comprises all the remaining observations in such months. The obtained cluster structure clearly reflects

seasonal patterns characterising tourism flows. Observations in cluster 2 (July and August) are characterized by the highest mean values of tourist arrivals and average stays in both regions, followed by those in cluster 3 (June and September) and cluster 4 (April and May) (see Table 18). From the comparison between clusters 1 and 6 it emerges that wintertime tourism flows have changed in both regions, showing an increase in the mean number of arrivals and a decrease in the mean number of stays in recent years (cluster 1). In all clusters, Veneto is characterised by mean values of both regressors which are higher than those of Emilia-Romagna except for the average stays from June to September. As far as the estimated regression coefficients are concerned (see Table 18), the first interesting finding is that the effects of both the tourist arrivals and the average stays on the number of visits result to be not homogeneous during the examined period of time. In both regions, such effects are positive in July and August; in Emilia-Romagna, this result also holds true in the months belonging to cluster 6. In the other clusters of months the effect of tourist arrivals are generally positive in both regions, while the average number of stays seem to have a negative impact on the number of visits. This latter impact in Veneto appears to be stronger than that in Emilia-Romagna in April, May, June and September; the opposite result holds true for all the other months.

The comparison between this partition and the one based on the maximum posterior probabilities estimated from the best `CW` fitted model (see Table 19) suggests that they are quite similar ($ARI = 0.8014$); the main difference is that according to the approach based on `CW` models all the observations in the months from November to February should be grouped into the same cluster. This latter result mainly depends on the fact that, in the best model fitted to the `RtI` dataset within the class of `CW` models with $K = 6$, the effects of both `Arriv Ve` on `Visit ER` and `Arriv ER` on `Visit Ve` in two clusters have been estimated to be quite similar (detailed results are not reported) and, thus, a better trade-off between the fit and the complexity is reached by the best fitted `CW` model with $K = 5$. Furthermore, the comparison between the results obtained through `SuCW` models and those produced from linear clusterwise regression analyses (see Table

Table 16: Best models fitted to the `RtI` dataset within the model classes defined from equations (5) and (6) and their $BIC$ values.

| Model | Best fitted model | $BIC_M$ |
|---|---|---|
| $\phi_P\left(\mathbf{x}_i;\boldsymbol{\mu}_{\mathbf{X}},\boldsymbol{\Sigma}_{\mathbf{X}}\right)$ | $M_a$: $acr.X =$ `VV` | 6808.1 |
| $\sum_{k=1}^{K}\pi_k\phi_D\left(\mathbf{y}_i\vert\mathbf{x}_i;\mathcal{X}_i'\boldsymbol{\beta}_k^*,\boldsymbol{\Sigma}_{\mathbf{Y}_k}\right)$ | $M_b$: $K=4$, $acr.Y =$ `VEV` | 3943.3 |
| $\sum_{k=1}^{K}\pi_k\phi_D\left(\mathbf{y}_i\vert\mathbf{x}_i;\mathbf{B}_k'\mathbf{x}_i^*,\boldsymbol{\Sigma}_{\mathbf{Y}_k}\right)$ | $M_c$: $K=3$, $acr.Y =$ `VEV` | 3920.2 |
| $\sum_{k=1}^{K}\pi_k\phi_1\left(y_{i1}\vert\mathbf{x}_i;\mathbf{x}_{i1}^{*\prime}\boldsymbol{\beta}_{k1}^*,\sigma_{k1}^2\right)$ | $M_d$: $K=4$, $acr.Y =$ `V` | 2166.6 |
| $\sum_{k=1}^{K}\pi_k\phi_1\left(y_{i1}\vert\mathbf{x}_i;\boldsymbol{\beta}_{k1}^{*\prime}\mathbf{x}_i^*,\sigma_{k1}^2\right)$ | $M_e$: $K=3$, $acr.Y =$ `V` | 2183.4 |
| $\sum_{k=1}^{K}\pi_k\phi_1\left(y_{i2}\vert\mathbf{x}_i;\mathbf{x}_{i2}^{*\prime}\boldsymbol{\beta}_{k2}^*,\sigma_{k2}^2\right)$ | $M_f$: $K=4$, $acr.Y =$ `E` | 1987.2 |
| $\sum_{k=1}^{K}\pi_k\phi_1\left(y_{i2}\vert\mathbf{x}_i;\boldsymbol{\beta}_{k2}^{*\prime}\mathbf{x}_i^*,\sigma_{k2}^2\right)$ | $M_g$: $K=2$, $acr.Y =$ `V` | 1987.1 |
| `SuCR` | $M_a$ and $M_b$ | 10751.4 |
| `CR` | $M_a$ and $M_c$ | 10728.3 |
| `uSuCR` | $M_a$, $M_d$ and $M_f$ | 10962.0 |
| `uCR` | $M_a$, $M_e$ and $M_g$ | 10978.6 |

Table 17: Cross-classification of the observations from the `RtI` dataset, based on their variable time identified by month and maximum posterior probability estimated from the best fitted model.

| $k$ | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 8 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 19 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 0 | 19 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 19 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 0 | 0 |
| 6 | 13 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 11 |

Table 18: Estimated $\pi_k$, $\boldsymbol{\mu}_{\mathbf{X}_k}$ and $\boldsymbol{\beta}_k^*$ of the best model fitted to the `RtI` dataset. $\mathbf{a}[l]$ denotes the $l$th element of vector $\mathbf{a}$.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\hat{\pi}_k$ | 0.121 | 0.167 | 0.167 | 0.167 | 0.167 | 0.213 |
| $\hat{\boldsymbol{\mu}}_{\mathbf{X}_k}[1]$ | 377.1 | 1389.8 | 1001.0 | 767.7 | 487.4 | 316.5 |
| $\hat{\boldsymbol{\mu}}_{\mathbf{X}_k}[2]$ | 2.350 | 6.304 | 4.847 | 2.823 | 2.572 | 2.694 |
| $\hat{\boldsymbol{\mu}}_{\mathbf{X}_k}[3]$ | 665.5 | 2248.1 | 1639.0 | 1251.5 | 867.7 | 502.5 |
| $\hat{\boldsymbol{\mu}}_{\mathbf{X}_k}[4]$ | 2.588 | 5.894 | 4.511 | 3.119 | 2.836 | 3.039 |
| $\hat{\boldsymbol{\beta}}_{k1}^*[1]$ | 34.063 | −263.846 | 143.648 | 278.219 | 163.170 | −68.964 |
| $\hat{\boldsymbol{\beta}}_{k1}^*[2]$ | 0.117 | 0.116 | −0.028 | −0.095 | 0.008 | 0.090 |
| $\hat{\boldsymbol{\beta}}_{k1}^*[3]$ | −15.948 | 26.302 | −7.994 | −14.227 | −34.080 | 25.994 |
| $\hat{\boldsymbol{\beta}}_{k2}^*[1]$ | 29.502 | −43.971 | 183.051 | 179.415 | 106.489 | −6.196 |
| $\hat{\boldsymbol{\beta}}_{k2}^*[2]$ | 0.071 | 0.035 | −0.006 | 0.003 | 0.032 | 0.132 |
| $\hat{\boldsymbol{\beta}}_{k2}^*[3]$ | −9.889 | 6.876 | −20.067 | −21.422 | −16.044 | −4.297 |

16) demonstrates that there is some clear evidence of seasonal heterogeneity not only in attendance at museums and monuments but also in tourism flows. Finally, a joint analysis for the two examined Italian regions based on seemingly unrelated cluster-weighted models results to be more effective than two separate linear clusterwise regression analyses.

As in the previous application, the `mult.norm` function of the `R` package `QuantPsych` has been employed to obtain an evaluation of the adequacy of the normality assumption within each cluster detected by the best model, based on the p-values of Mardia's measures of multivariate skewness and kurtosis computed from the estimated residuals $\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{\mathbf{X}_k}$ and $\mathbf{y}_i - \mathcal{X}_i'\hat{\boldsymbol{\beta}}_k$ $\forall (i,k) \in \{(i,k), i \in \{1,\ldots,I\}, k = \arg\max_h\{\hat{\tau}_{ih}, h = 1,\ldots,K\}\}$ (see Table 20. The obtained results suggest that the null hypothesis of multivariate normality should not be rejected at a Bonferroni-corrected $0.05/6 = 0.0083$ significance level in any cluster.

Table 19: Cross-classification of the observations from the `RtI` dataset, based on the maximum posterior probabilities estimated from the best `CW` and `SuCW` fitted models. Labels for clusters reported in rows and colums refer to `CW` and `SuCW`, respectively.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 34 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 38 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 38 | 0 |
| 4 | 0 | 38 | 4 | 0 | 0 | 0 |
| 5 | 27 | 0 | 0 | 0 | 0 | 49 |

Table 20: P-values of Mardia's skewness and kurtosis statistics for the residuals $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}_k}$ and $\mathbf{y} - \mathcal{X}'\boldsymbol{\beta}_k$, $k = 1, \ldots, K$, computed from the best model fitted to the `RtI` dataset.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}_k}$ | | | | | | |
| skewness | 0.4513 | 0.3886 | 0.1150 | 0.4504 | 0.0422 | 0.0317 |
| kurtosis | 0.0171 | 0.0375 | 0.6161 | 0.4693 | 0.4039 | 0.8658 |
| $\mathbf{y} - \mathcal{X}'\boldsymbol{\beta}_k$ | | | | | | |
| skewness | 0.6053 | 0.6263 | 0.5742 | 0.2526 | 0.6386 | 0.0212 |
| kurtosis | 0.0369 | 0.3261 | 0.9268 | 0.3922 | 0.1747 | 0.7168 |

## 5. Conclusions

The proposed multivariate seemingly unrelated Gaussian linear cluster-weighted models can account for heterogeneous regression data with multivariate correlated responses, each one depending on its own set of covariates. This latter feature represents the main novelty of the models proposed here in reference with the ones introduced by Dang et al. (2017), thus leading to a more flexible modelling of data in applications where prior information concerning the absence of certain covariates from the linear term employed in the prediction of a certain response has to be conveyed into the model, and different covariates are expected to be relevant in the prediction of different responses. The distribution of the covariates is also explicitly incorporated in the model formulation. The resulting approach encompasses the models introduced by Dang et al. (2017) as well as other Gaussian mixture-based linear regression models with random covariates. Details about identifiability, ML estimation and model selection have been provided. Furthermore, models with a reduced number of variance-covariance parameters have been specified. The comparisons among some cluster-weighted models and clusterwise linear regression models based on the analyses of the `tuna` and `RtI` datasets have highlighted the effectiveness of the proposed models in detecting the presence of unobserved heterogeneity; such models have been proved to be useful also to establish the relevance of a multivariate regression analysis and the inadequacy of the assignment independence assumption in both applications. From the Monte Carlo studies it appears that including irrelevant regressors in a cluster-weighted model can lead to a wrong choice of the number of components and a sub-optimal reconstruction of the true classification of the sample observations, especially when the components are not well-separated. The approach introduced here is able to avoid some drawbacks due to the presence of irrelevant regressors in a multivariate Gaussian linear cluster-weighted model. This happens because the proposed models are multivariate Gaussian linear cluster-weighted models in which some regression coefficients are set a priori equal to zero. Thus, the proposed approach also

represents a framework for multivariate linear cluster-weighted analysis under such constraints.

As far as the development of inferential methods for the parameters of the proposed models is concerned, an assessment of the sample variability of the parameter estimates is required. Since the ECM algorithm does not automatically produce any estimate of the covariance matrix of the ML estimator, additional computations are necessary. To this end, several approaches commonly employed under finite mixture models could be exploited (see, e.g., McLachlan and Peel, 2000). For example, estimates of the asymptotic covariance matrix of the ML estimator can be computed through an approach which is based on the gradient vector and the second-order derivative matrix of the incomplete data log-likelihood, and makes also use of a sandwich estimator. This approach has been successfully applied to Gaussian mixture models (Boldea and Magnus, 2009), $t$ mixture models (Wang and Lin, 2016), clusterwise Gaussian linear regression models (Galimberti et al., 2021) and Gaussian linear cluster-weighted models (Soffritti, 2021). In addition, given the critical role played by the initialisation in any ECM algorithm, further investigation might be needed in order to confirm the encouraging results described in Section 3. In particular, this additional investigation should focus on the performance of the proposed strategy in presence of high dimensional data.

Another crucial aspect associated with the adoption of the proposed models in practical applications is the assessment of their adequacy. For finite mixtures of linear regression models with a univariate response and fixed, concomitant or random covariates, Ingrassia and Punzo (2020) have recently introduced some indices able to measure the association between the response variable and the latent groups, the model goodness-of-fit, and the proportion of the total variation in the response which remains unexplained by the fitted model. Local and overall coefficients of determination have also been described. After suitable modifications, those indices could also be employed to assess the adequacy of the multivariate cluster-weighted models based on seemingly underlated linear regression illustrated here.

Multivariate seemingly unrelated linear cluster-weighted analyses based on the proposed models implicitly require that the researcher has prior information on the specific covariates that have to be included in the linear term employed in the prediction of each response in the model. In practical applications in which the choice of the regressors to be used for different responses is questionable, the relevant regressors for each response can be detected through strategies (e.g., stepwise techniques, genetic algorithms) that allow to perform variable selection in a multivariate regression framework. To this end, the optimal model for the given dataset should be determined from a model class which also includes the cluster-weighted models based on seemingly unrelated linear regression illustrated in this paper.

Finally, it is worth noting that the new multivariate seemingly unrelated cluster-weigthed models described here have been specified under the following assumptions: *i)* the joint conditional distribution of the $P$ covariates given the group $\Omega_k$ is Gaussian $\forall k$; *ii)* the joint conditional distribution of the $D$ responses given the covariates and the group $\Omega_k$ is Gaussian $\forall k$; *iii)* the conditional expected value of the $D$ responses given the covariates and the group $\Omega_k$ is a linear transformation of the covariates $\forall k$. These assumptions could be relaxed by resorting to the approaches developed by Punzo (2014), Punzo and McNicholas (2017), Gallaugher et al. (2021) or Sahin and Czado (2021) so as to obtain multivariate seemingly unrelated cluster-weigthed models which could be more effectively employed in the analysis of real datasets composed of unknown clusters of observations characterised by skewed distributions, outliers or non-linear relationships.

## Appendix A. Derivation of $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}^{*(h+1)}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h+1)}$

The CM-steps to update the estimates of the model parameters $\boldsymbol{\beta}_k^*$ and $\boldsymbol{\Sigma}_{\mathbf{Y}_k}$ at the $(h+1)$th iteration in the ECM algorithm, as illustrated in equations (13)

and (14), can be obtained as follows.

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_k^{*\prime}} Q\Big(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)}\Big) &= \frac{\partial}{\partial \boldsymbol{\beta}_k^{*\prime}} \sum_{i=1}^{I}\sum_{k=1}^{K} \hat{\tau}_{ik}^{(h)} Q_2\Big(\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}|\boldsymbol{\psi}^{(h)}\Big) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}_k^{*\prime}} \Big[ -\frac{1}{2} \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \Big(\mathbf{y}_i - \mathcal{X}_i'\boldsymbol{\beta}_k^*\Big)' \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{(h)(-1)} \Big(\mathbf{y}_i - \mathcal{X}_i'\boldsymbol{\beta}_k^*\Big)\Big] \\
&= -\frac{1}{2} \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \frac{\partial}{\partial \boldsymbol{\beta}_k^{*\prime}} \Big( -2\mathbf{y}_i' \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{(h)(-1)} \mathcal{X}_i'\boldsymbol{\beta}_k^* + \boldsymbol{\beta}_k^{*\prime} \mathcal{X}_i' \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{(h)(-1)} \mathcal{X}_i'\boldsymbol{\beta}_k^*\Big) \\
&= -\frac{1}{2} \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \Big( -2\mathbf{y}_i' \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{(h)(-1)} \mathcal{X}_i' + 2\boldsymbol{\beta}_k^{*\prime} \mathcal{X}_i \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{(h)(-1)} \mathcal{X}_i'\Big) \\
&= \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \mathbf{y}_i' \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{(h)(-1)} \mathcal{X}_i' - \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \boldsymbol{\beta}_k^{*\prime} \mathcal{X}_i \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{(h)(-1)} \mathcal{X}_i' \,. \qquad \text{(A.1)}
\end{aligned}
$$

Setting (A.1) equal to the null vector, $\boldsymbol{\Sigma}_{\mathbf{Y}_k}^{(h)}$ equal to $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{(h)}$ and solving the so obtained system with respect to $\boldsymbol{\beta}_k^*$ leads to the solution reported in equation (13).

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{-1}} Q\Big(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)}\Big) &= \frac{\partial}{\partial \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{-1}} \sum_{i=1}^{I}\sum_{k=1}^{K} \hat{\tau}_{ik}^{(h)} Q_2\Big(\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}|\boldsymbol{\psi}^{(h)}\Big) \\
&= \frac{\partial}{\partial \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{-1}} \Bigg\{ \frac{1}{2} \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \Big[\ln|\boldsymbol{\Sigma}_{\mathbf{Y}_k}^{-1}| \\
&\quad - \Big(\mathbf{y}_i - \mathcal{X}_i'\boldsymbol{\beta}_k^{*(h+1)}\Big)' \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{-1} \Big(\mathbf{y}_i - \mathcal{X}_i'\boldsymbol{\beta}_k^{*(h+1)}\Big)\Big]\Bigg\} \\
&= \frac{1}{2} \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \frac{\partial}{\partial \boldsymbol{\Sigma}_{\mathbf{Y}_k}^{-1}} \Big[\ln|\boldsymbol{\Sigma}_{\mathbf{Y}_k}^{-1}| \\
&\quad - tr\Big(\boldsymbol{\Sigma}_{\mathbf{Y}_k}^{-1} (\mathbf{y}_i - \mathcal{X}_i'\boldsymbol{\beta}_k^{*(h+1)})(\mathbf{y}_i - \mathcal{X}_i'\boldsymbol{\beta}_k^{*(h+1)})'\Big)\Big] \\
&= \frac{1}{2} \Bigg[ \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} \boldsymbol{\Sigma}_{\mathbf{Y}_k} - \sum_{i=1}^{I} \hat{\tau}_{ik}^{(h)} (\mathbf{y}_i - \mathcal{X}_i'\boldsymbol{\beta}_k^{*(h+1)})(\mathbf{y}_i - \mathcal{X}_i'\boldsymbol{\beta}_k^{*(h+1)})' \Bigg] \qquad \text{(A.2)}
\end{aligned}
$$

where the second and third equalities are obtained using properties of trace and transpose and differentiation rules of functions of matrices. Setting (A.2) equal to the null matrix, $\boldsymbol{\beta}_k^{*(h+1)}$ equal to $\hat{\boldsymbol{\beta}}_k^{*(h+1)}$ and solving the resulting system with respect to $\boldsymbol{\Sigma}_{\mathbf{Y}_k}$ gives the update in equation (14).

## Appendix B. Expression of $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}^{*(\boldsymbol{h+1})}$ when $\mathbf{x}_{\boldsymbol{id}} = \mathbf{x}_{\boldsymbol{i}}\ \forall \boldsymbol{d}$

Similarly to Park (1993), equation (13) can be rewritten as

$$\hat{\boldsymbol{\beta}}_k^* = \left\{ \mathcal{X} \left[ \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \otimes \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \right] \mathcal{X}' \right\}^{-1} \mathcal{X} \left[ \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \otimes \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \right] \mathbf{y}, \tag{B.1}$$

where the superscripts $(h)$ and $(h+1)$ have been dropped to ease notation, $\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_I]$ is a $(P^* + D) \times (D \cdot I)$ matrix, $\operatorname{diag}(\hat{\boldsymbol{\tau}}_k)$ is a diagonal matrix whose diagonal elements are the values $\hat{\tau}_{ik}$ $(i = 1, \ldots, I)$ and $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_I')'$. Consider now the vectors $\mathbf{v}_d = (y_{1d}, y_{2d}, \ldots, y_{Id})'$, containing the values of the $d$th response on the $I$ observations $(d = 1, \ldots, D)$, and the vector $\mathbf{v} = (\mathbf{v}_1', \ldots, \mathbf{v}_D')'$. It is evident that $\mathbf{v}$ and $\mathbf{y}$ contain the same values but in a different order. As shown in Park (1993), by exchanging the rows of the identity matrix of order $D \cdot I$, it is possible to define a matrix $\mathbf{L}$ such that

$$\mathbf{L}\mathbf{L}' = \mathbf{L}'\mathbf{L} = \mathbf{I}_{D \cdot I}$$

and

$$\mathbf{L}\mathbf{y} = \mathbf{v}.$$

Matrix $\mathbf{L}$ can also be used to reorder the columns of $\mathcal{X}$ and the rows and columns of $\left[ \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \otimes \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \right]$. Namely,

$$\mathcal{X}\mathbf{L}' = \mathcal{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0}_{(P_1+1)\times I} & \cdots & \mathbf{0}_{(P_1+1)\times I} \\ \mathbf{0}_{(P_2+1)\times I} & \mathbf{Z}_2 & \cdots & \mathbf{0}_{(P_2+1)\times I} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{(P_D+1)\times I} & \mathbf{0}_{(P_D+1)\times I} & \cdots & \mathbf{Z}_D \end{bmatrix},$$

where $\mathbf{Z}_d = [\mathbf{x}_{1d}^*, \mathbf{x}_{2d}^*, \cdots, \mathbf{x}_{Id}^*]$ is a $(P_d + 1) \times I$ matrix $(d = 1, \ldots, D)$, and

$$\mathbf{L} \left[ \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \otimes \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \right] \mathbf{L}' = \left[ \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \otimes \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \right].$$

Thus, an equivalent expression for $\hat{\boldsymbol{\beta}}_k^*$ is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_k^* &= \left\{ \mathcal{X}\mathbf{L}'\mathbf{L} \left[ \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \otimes \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \right] \mathbf{L}'\mathbf{L}\mathcal{X}' \right\}^{-1} \mathcal{X}\mathbf{L}'\mathbf{L} \left[ \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \otimes \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \right] \mathbf{L}'\mathbf{L}\mathbf{y} \\ &= \left\{ \mathcal{Z} \left[ \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \otimes \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \right] \mathcal{Z}' \right\}^{-1} \mathcal{Z} \left[ \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}^{-1} \otimes \operatorname{diag}(\hat{\boldsymbol{\tau}}_k) \right] \mathbf{v}. \end{aligned} \tag{B.2}$$

If $\mathbf{x}_{id} = \mathbf{x}_i \ \forall d$, then $\mathbf{Z}_d = [\mathbf{x}_1^*, \mathbf{x}_2^*, \cdots, \mathbf{x}_I^*] = \mathbf{Z} \ \forall d$ and

$$\mathcal{Z} = \mathbf{I}_D \otimes \mathbf{Z}. \tag{B.3}$$

By exploiting equation (B.3) and the properties of the Kronecker product (see, e.g., Magnus and Neudecker, 1988), equation (B.2) can be simplified as follows:

$$\hat{\boldsymbol{\beta}}_k^* = \left\{ \mathbf{I}_D \otimes \left[ \mathbf{Z} \ \mathrm{diag}\left(\hat{\boldsymbol{\tau}}_k\right) \mathbf{Z}' \right]^{-1} \mathbf{Z} \ \mathrm{diag}\left(\hat{\boldsymbol{\tau}}_k\right) \right\} \mathbf{v}. \tag{B.4}$$

Firstly, note that equation (B.4) does not depend on $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_k}$. Furthermore, the matrix $\left\{ \mathbf{I}_D \otimes \left[ \mathbf{Z} \ \mathrm{diag}\left(\hat{\boldsymbol{\tau}}_k\right) \mathbf{Z}' \right]^{-1} \mathbf{Z} \ \mathrm{diag}\left(\hat{\boldsymbol{\tau}}_k\right) \right\}$ has a block-diagonal structure. By coupling it with the structure of the vector $\mathbf{v}$, the following expression for the vector $\hat{\boldsymbol{\beta}}_{kd}^*$ containing the estimated coefficients associated with the $d$th response in the $k$th group can be obtained:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{kd}^* &= \left[ \mathbf{Z} \ \mathrm{diag}\left(\hat{\boldsymbol{\tau}}_k\right) \mathbf{Z}' \right]^{-1} \mathbf{Z} \ \mathrm{diag}\left(\hat{\boldsymbol{\tau}}_k\right) \mathbf{v}_d \\ &= \left( \sum_{i=1}^{I} \hat{\tau}_{ik} \mathbf{x}_i^* \mathbf{x}_i^{*\prime} \right)^{-1} \left( \sum_{i=1}^{I} \hat{\tau}_{ik} \mathbf{x}_i^* y_{id} \right), \quad d = 1, \ldots, D. \end{aligned} \tag{B.5}$$

Apart from differences related to notation, it can be noticed that equation (B.5) coincides with the $d$th row of the matrix defined in equation (8) in Dang et al. (2017).

## References


Aitken AC (1926) A series formula for the roots of algebraic and transcendental equations. Proc. R. Soc. Edinb. 45:14–22

Baird IG, Quastel N (2011) Dolphin-safe tuna from California to Thailand: localisms in environmental certification of global commodity networks. Ann. Assoc. Am. Geogr. 101:337–355

Browne RP, McNicholas PD (2014a) Estimating common principal components in high dimensions. Adv. Data Anal. Classif. 8:217–226

Browne RP, McNicholas PD (2014b) Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. Stat. Comput. 24:203–210

Boldea O, Magnus JR (2009) Maximum likelihood estimation of the multivariate normal mixture model. J. Am. Stat. Assoc. 104:1539–1549

Cadavez VAP, Henningsen A (2012) The use of seemingly unrelated regression (SUR) to predict the carcass composition of lambs. Meat Sci. 92:548–553

Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recognit. 28:781–793

Cellini R, Cuccia T (2013) Museum and monument attendance and tourism flow: a time series analysis approach. Appl. Econ. 45:3473–3482

Chevalier JA, Kashyap AK, Rossi PE (2003) Why don't prices rise during periods of peak demand? Evidence from scanner data. Am. Econ. Rev. 93:15–37

Dang UJ, McNicholas PD (2015) Families of parsimonious finite mixtures of regression models. In: Morlini I, Minerva T, Vichi M (eds) Advances in statistical models for data analysis. Springer, Cham, pp 73–84

Dang UJ, Punzo A, McNicholas PD, Ingrassia S, Browne RP (2017) Multivariate response and parsimony for Gaussian cluster-weighted models. J. Classif. 34:4–34

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood for incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39:1–22

Di Mari R, Bakk Z, Punzo A (2020) A random-covariate approach for distal outcome prediction with latent class analysis. Struct. Equ. Model. 27:351–368

Disegna M, Osti L (2016) Tourists' expenditure behaviour: the influence of satisfaction and the dependence of spending categories. Tour. Econ. 22:5–30

Fletcher TD (2012) `QuantPsyc`: Quantitative psychology tools. `R` package version 1.5. `URL http://CRAN.R-project.org/package=QuantPsyc`

Flury BN, Gautschi W (1986) An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. J. Sci. Statist. Comput. 7:169–184

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc. 97:611–631

Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, New York

Galimberti G., Soffritti G. (2007) Model-based methods to identify multiple cluster structures in a data set. Comput. Stat. Data Anal. 52:520–536

Galimberti G, Scardovi E, Soffritti G (2016) Using mixtures in seemingly unrelated linear regression models with non-normal errors. Stat. Comput. 26:1025–1038

Galimberti G, Nuzzi L, Soffritti G (2021) Covariance matrix estimation of the maximum likelihood estimation in multivariate clusterwise linear regression. Stat. Methods Appl. 30:235–268

Galimberti G, Soffritti G (2020) Seemingly unrelated clusterwise linear regression. Adv. Data Anal. Classif. 14:235–260

Gallaugher MPB, Tomarchio SD, McNicholas PD, Punzo A (2021) Multivariate cluster weighted models using skewed distributions. Adv. Data Anal. Classif. https://doi.org/10.1007/s11634-021-00480-5

Gershenfeld N (1997) Nonlinear inference and cluster-weighted modeling. Ann. N. Y. Acad. Sci. 808:18–24

Giles S, Hampton P (1984) Regional production relationships during the industrialization of New Zealand, 1935–1948. Reg. Sci. 24:519–533

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

Heidari S., Keshavarz S., Mirahmadizadeh A (2017) Application of seemingly unrelated regression (SUR) in determination of risk factors of fatigue and general health among the employees of petrochemical companies. J. Health Sci. Surveillance Sys. 5:1–8

Hennig C (2000) Identifiability of models for clusterwise linear regression. J. Classif. 17:273–296

Henningsen A, Hamann JD (2007) `systemfit`: a package for estimating systems of simultaneous equations in `R`. J. Stat. Softw. 23(4):1–40

Hubert L, Arabie P (1985) Comparing partitions. J. Classif. 2:193–218

Ingrassia S, Minotti SC, Vittadini G (2012) Local statistical modeling via a cluster-weighted approach with elliptical distributions. J. Classif. 29:363–401

Ingrassia S, Minotti SC, Punzo A (2014) Model-based clustering via linear cluster-weighted models. Comput. Stat. Data Anal. 71:159–182

Ingrassia S, Punzo A (2020) Cluster validation for mixtures of regressions via the total sum of squares decomposition. J. Classif. 37:526–547

Ingrassia S, Punzo A, Vittadini G, Minotti SC (2015) The generalized linear mixed cluster-weighted model. J. Classif. 32:85–113

Ingrassia S, Rocci R (2011) Degeneracy of the EM algorithm for the MLE of multivariate Gaussian mixtures and dynamic constraints. Comput. Stat. Data Anal. 55:1715–1725

Jones PN, McLachlan GJ (1992) Fitting finite mixture models in a regression context. Aust. J. Stat. 34:233–240

Keshavarzi S, Ayatollahi SMT, Zare N, Pakfetrat M (2012) Application of seemingly unrelated regression in medical data with intermittently observed time-dependent covariates. Comput. Math. Methods Med. 821643

Keshavarzi S, Ayatollahi SMT, Zare N, Sharif F (2013) Quality of life of child-bearing age women and its associated factors: an application of seemingly unrelated regression (SUR) models. Qual. Life Res. 22:1255–1263

Lin TI (2014) Learning from incomplete data via parameterized $t$ mixture models through eigenvalue decomposition. Comput. Stat. Data Anal. 71:183–195

Magnus JR, Neudecker H (1988) Matrix differential calculus with applications in statistics and econometrics. John Wiley & Sons, New York

Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57:519–530

Mardia, K.V (1974) Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. Sankhya, 36:115–128

McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York

McNicholas PD (2010) Model-based classification using latent Gaussian mixture models. J. Stat. Plan. Inference 140:1175–1181

Meng X, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80:267–278

Park T (1993) Equivalence of maximum likelihood estimation and iterative two-stage estimation for seemingly unrelated regression models. Comm. Statist.: Theory Meth. 22:2285–2296

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

Punzo A, Ingrassia S (2013) On the use of the generalized linear exponential cluster-weighted model to assess local linear independence in bivariate data. QdS - J. Methodol. Appl. Stat. 15:131–144

Punzo A, Ingrassia S (2015) Clustering bivariate mixed-type data via the cluster-weighted model. Comput. Stat. 31:989–1013

Punzo A, McNicholas PD (2017) Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. J. Classif. 34:249–293

R Core Team (2020) `R`: a language and environment for statistical computing. `R` Foundation for Statistical Computing, Vienna, Austria. `URL http://www.R-project.org`

Rocci R, Gattone SA, Di Mari R (2018) A data driven equivariant approach to constrained Gaussian mixture modeling. Adv. Data Anal. Classif. 12:235–260

Rossi PE (2012) `bayesm`: Bayesian inference for marketing/micro-econometrics. `R` package version 2.2-5. `URL http://CRAN.R-project.org/package=bayesm`

Sahin Ö, Czado C (2021) Vine copula mixture models and clustering for non-Gaussian data. Econ. Stat. https://doi.org/10.1016/j.ecosta.2021.08.011

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

Scrucca L, Fop M, Murphy TB, Raftery AE (2017) `mclust 5`: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal 8/1:205–223

Soffritti G (2021) Estimating the covariance matrix of the maximum likelihood estimator under linear cluster-weighted models. J. Classif. 38:594–625

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

Srivastava VK, Giles DEA (1987) Seemingly unrelated regression equations models. Marcel Dekker, New York

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

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

Wang WL, Lin TI (2016) Maximum likelihood inference for the multivariate t mixture model. J. Multivar. Anal. 149:54–64

White EN, Hewings GJD (1982) Space-time employment modelling: some results using seemingly unrelated regression estimators. J. Reg. Sci. 22:283–302