Statistical Papers (2023) 64:883–921
https://doi.org/10.1007/s00362-022-01344-6

REGULAR ARTICLE

Seemingly unrelated clusterwise linear regression for contaminated data

Gabriele Perrone[1] · Gabriele Soffritti[1]

Received: 22 December 2021 / Revised: 5 July 2022 / Accepted: 10 July 2022 /
Published online: 6 August 2022
© The Author(s) 2022

Abstract

Clusterwise regression is an approach to regression analysis based on finite mixtures which is generally employed when sample observations come from a population composed of several unknown sub-populations. Whenever the response is continuous, Gaussian clusterwise linear regression models are usually employed. Such models have been recently robustified with respect to the possible presence of mild outliers in the sub-populations. However, in some fields of research, especially in the modelling of multivariate economic data or data from the social sciences, there may be prior information on the specific covariates to be considered in the linear term employed in the prediction of a certain response. As a consequence, covariates may not be the same for all responses. Thus, a novel class of multivariate Gaussian linear clusterwise regression models is proposed. This class provides an extension to mixture-based regression analysis for modelling multivariate and correlated responses in the presence of mild outliers that let the researcher free to use a different vector of covariates for each response. Details about the model identification and maximum likelihood estimation via an expectation-conditional maximisation algorithm are given. The performance of the new models is studied by simulation in comparison with other clusterwise linear regression models. A comparative evaluation of their effectiveness and usefulness is provided through the analysis of a real dataset.

Keywords Contaminated Gaussian distribution · ECM algorithm · Mild outlier · Mixture of regression models · Model-based cluster analysis · Seemingly unrelated regression

✉ Gabriele Soffritti
gabriele.soffritti@unibo.it

Gabriele Perrone
gabriele.perrone4@unibo.it

[1] Department of Statistical Sciences, Alma Mater Studiorum - University of Bologna, via delle Belle Arti 41,
40126 Bologna, Italy

Mathematics Subject Classification 62J05 · 62H12 · 62F12

1 Introduction

In multivariate regression analysis, when modelling the dependence of a random vector $\mathbf{Y} = (Y_1, \ldots, Y_m, \ldots, Y_M)'$ of M responses on a given vector $\mathbf{X} = (X_1, \ldots, X_p, \ldots, X_P)'$ of P predictors through a sample $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I)\}$ drawn from a certain population, the following sources of complexity could affect the data and make the prediction of the responses a task difficult to perform.

(a) With multivariate longitudinal data, time-series data or repeated measures, the M responses contained in $\mathbf{Y}$ are typically correlated. Furthermore, in analyses of economic data or data from the social sciences, it is not unusual that prior information about the phenomenon under study enables the analyst to specify a system of M regression equations (one equation for each response) in which certain regressors contained in $\mathbf{X}$ are absent from certain regression equations. This is especially true for multivariate economic data referring to general theories (i.e., investment equations, production functions) or applications dealing with the explanation of a certain economic activity (i.e., demand of petrol, employment) in different geographical locations (see, e.g., Giles and Hampton 1984; White and Hewings 1982; Zellner 1962). Further examples can be found also in other fields, such as medicine, food quality, tourism economics, quality of life and health (see, e.g., Cadavez and Hennningsen 2012; Disegna and Osti 2016; Heidari et al. 2017; Keshavarzi et al. 2012, 2013). A parametric framework able to take into consideration both multivariate correlated responses and systems of regression equations with equation-dependent vectors of predictors (i.e., vectors which do not necessarily contain the same predictors for all the responses) is given by the so-called seemingly unrelated regression approach (see, e.g., Park 1993; Srivastava and Giles 1987). In particular, in this approach the random disturbances associated with the M regression equations are allowed to be correlated with each other; hence, the variance-covariance matrix $\mathbf{\Sigma}$ of the resulting M-dimensional vector of the error terms will have a non-diagonal structure.

(b) In general, real data can often be characterised by the presence of atypical observations. In parametric regression analysis, such observations negatively impact on both the estimation of the regression coefficients and the prediction of the responses based on the classical procedures. Such procedures have been widely recognized to be extremely sensitive to even seemingly minor or negligible deviations from some conventional assumptions (see, e.g., Tukey 1960). Thus, when the data are contaminated by such observations, it is crucial that robust methods are employed (see, e.g., Maronna et al. 2006). Departures from the Gaussian distribution of the error terms in the regression model caused by some mildly atypical observations can be managed by simply resorting to heavy-tailed models for the conditional distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$. Those observations are also called small or mild outliers (see, e.g., Ritter 2015). Examples of robust methods against the presence of such outliers have been developed by Lange et al. (1989), Kibria and Haq (1999), Lachos et al. (2011); to this end, the multivariate t distribution or scale mixtures of Gaussian distributions have been exploited. Another model able to manage the possible presence of mild outliers in a dataset is the contaminated Gaussian distribution (see, e.g., Aitkin and Wilson 1980; Tukey 1960). This probabilistic model is defined as a mixture of two Gaussian distributions having the same expected mean values but different variances-covariances. Furthermore, the Gaussian distribution having the smallest mixing weight also has inflated variances-covariances and is employed to represent the mild outliers. Maximum likelihood (ML) estimation can be performed via an expectation-maximisation (EM) algorithm (see Aitkin and Wilson 1980; Dempster et al. 1977). Once such a model is fitted to the observed data, each sample observation can be classified as either typical or outlier using the maximum a posteriori probability (for further details see, e.g., Aitkin and Wilson 1980). With an approach based on the use of one of these distributions, robustness can be achieved without suppressing any observation from the sample $\mathcal{S}$.

(c) Sometimes the population from which the sample $\mathcal{S}$ comes from is composed of a certain number, say K, of sub-populations. Furthermore, when the information about the value of K and the specific sub-population each sample observation belongs to is not known, $\mathcal{S}$ is characterised by unobserved heterogeneity. If this source of heterogeneity affects the distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$, then a mixture of K different regression models (one for each sub-population) will describe the distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ in the population. This phenomenon can be experienced in many fields, such as economics, marketing, agriculture, education, human genomics, quantitative finance, social sciences and transport systems (see, e.g., Ding 2006; Dyer et al. 2012; Elhenawy et al. 2017; Fair and Jaffe 1972; Kamakura 1988; McDonald et al. 2016; Qin and Self 2006; Tashman and Frey 2009; Turner 2000; Van Horn et al. 2015). In this case, the sample $\mathcal{S}$ should be analysed in a regression framework able to detect both the number of sub-populations and their regression models. Methods for clusterwise regression analysis play a special role. They exploit clusterwise regression models, which are mixtures of K regression models (see, e.g., De Sarbo and Cron 1988; Depraetere and Vandebroek 2014; Frühwirth-Schnatter 2006; Hosmer 1974). In these models, the mixing weights can also be expressed as a function of some concomitant variables (Wedel 2002). With M continuous responses in vector $\mathbf{Y}$, multivariate Gaussian clusterwise linear regression models are generally employed (see, e.g., Jones and McLachlan 1992). If the P predictors are random and the source of heterogeneity mentioned above affects the distribution of $(\mathbf{X}, \mathbf{Y})$, then Gaussian cluster-weighted models should be employed (see, e.g., Dang et al. 2017).

Recently, Mazza and Punzo (2020) have introduced methods to perform Gaussian clusterwise linear regression analysis which are robust with respect to heavy-tailed departures from Gaussianity due to the presence of mild outliers in the data. By relying on contaminated Gaussian clusterwise linear regression models, their methods are able to produce a simultaneous clustering of the sample observations and the detection of mild outliers in a multivariate regression context. In this way, they allow to manage the sources of complexity *(b)* and *(c)*; they are also capable of explaining the correlation among responses. A limitation of an approach based on those models is that the same vector of regressors has to be employed for the prediction of all responses. Galimberti and Soffritti (2020) have developed models for Gaussian clusterwise linear regression which make use of seemingly unrelated regression equations. The methods based on these latter models are suitable for the analysis of data affected by complexities *(a)* and *(c)*; however, they are not insensitive to the possible presence of mild outliers in the K sub-populations. Based on all these considerations, multivariate seemingly unrelated clusterwise linear regression models for data contaminated by mild outliers are introduced here. They are obtained from the models described in Mazza and Punzo (2020) by modifying the definition of the linear terms in the M regression equations so that a different vector of regressors can be employed for each dependent variable. With these new models, the three sources of complexities mentioned above are jointly taken into consideration when predicting the responses in a multivariate linear regression framework. Thus, a more flexible approach for the analysis of linear dependencies in multivariate data is provided.

The key contributions of this paper are:

- the specification of a novel class of models able to jointly account for the sources of complexity *(a)*, *(b)* and *(c)* mentioned above;
- a comparison with some other linear clusterwise regression models;
- the description of conditions for the identifiability of the novel models;
- details about ML estimation via an expectation-conditional maximisation (ECM) algorithm (Meng and Rubin 1993);
- a treatment of the initialisation and convergence of the ECM algorithm and the issue of model selection;
- an investigation of the effectiveness of the new models, based on simulated datasets, in comparison with the models proposed by Galimberti and Soffritti (2020) and Mazza and Punzo (2020);
- an application to a study of the effects of prices and promotional activities on sales for two U.S. brands of canned tuna.

The remainder of this paper is organised as follows. The novel models are introduced in Sect. 2.1. Section 2.2 shows how they relate to some clusterwise linear regression models. Identifiability is treated in Sect. 2.3. Section 2.4 and Appendix A provide details on the ECM algorithm. Issues of algorithm initialisation, convergence criterion and model selection are discussed in Sects. 2.5 and 2.6 . Section 3 contains a summary of the experimental results obtained from the analysis of simulated data. The study of the effects of prices and promotional activities on U.S. canned tuna sales is presented in Sect. 4. Finally, in Sect. 5, some concluding remarks and ideas for future research are illustrated.

2 Seemingly unrelated contaminated Gaussian linear clusterwise regression analysis

2.1 Seemingly unrelated contaminated Gaussian linear clusterwise regression models

In order to introduce the new model, the following notation is required. Suppose that only P_m of the P covariates contained in $\mathbf{X}$ are considered to be relevant for the prediction of the response Y_m, where $P_m \leq P$. Thus, let $\mathbf{X}_m = (X_{m_1}, X_{m_2}, \ldots, X_{m_{P_m}})'$ be the vector composed of such P_m covariates, and let $\mathbf{X}_m^* = (1, \mathbf{X}_m')'$. Furthermore, let $\boldsymbol{\beta}_{km} = (\beta_{k,m_1}, \beta_{k,m_2}, \ldots, \beta_{k,m_{P_m}})'$ be the vector of the P_m regression coefficients capturing the linear effect of such covariates on the response Y_m in the kth sub-population, and $\boldsymbol{\beta}_{km}^* = (\beta_{0k,m}, \boldsymbol{\beta}_{km}')'$. Then, the vector containing all linear effects on the M responses in the kth sub-population can be obtained by stacking the M regression coefficient vectors specific for the kth sub-population one underneath the other; it can be denoted as $\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_{k1}^{*\prime}, \ldots, \boldsymbol{\beta}_{km}^{*\prime}, \ldots, \boldsymbol{\beta}_{kM}^{*\prime})'$ and its length is $P^* + M$, where $P^* = \sum_{m=1}^{M} P_m$. Finally, the following $(P^* + M) \times M$ partitioned matrix is required:

$$\tilde{\mathbf{X}}^* = \begin{bmatrix} \mathbf{X}_1^* & \mathbf{0}_{P_1+1} & \cdots & \mathbf{0}_{P_1+1} \\ \mathbf{0}_{P_2+1} & \mathbf{X}_2^* & \cdots & \mathbf{0}_{P_2+1} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{P_M+1} & \mathbf{0}_{P_M+1} & \cdots & \mathbf{X}_M^* \end{bmatrix},$$

where $\mathbf{0}_{P_m+1}$ denotes the $(P_m + 1)$-dimensional null vector.

The random vector $\mathbf{Y}$ follows a seemingly unrelated contaminated Gaussian linear clusterwise regression model of order K if the conditional probability density function (p.d.f.) of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ has the form

$$f(\mathbf{y}|\mathbf{x}; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k h(\mathbf{y}; \boldsymbol{\theta}_k), \mathbf{y} \in \mathbb{R}^M, \tag{1}$$

where π_k is the mixing weight of the kth sub-population, with $\pi_k > 0$ for $k = 1, \ldots, K$, and $\sum_{k=1}^{K} \pi_k = 1$; $h(\mathbf{y}; \boldsymbol{\theta}_k)$ is the contaminated Gaussian p.d.f. of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ in the kth sub-population, defined as follows:

$$h(\mathbf{y}; \boldsymbol{\theta}_k) = \alpha_k \phi_M\left(\mathbf{y}; \boldsymbol{\mu}_k(\mathbf{x}; \boldsymbol{\beta}_k^*), \boldsymbol{\Sigma}_k\right) + (1 - \alpha_k)\phi_M\left(\mathbf{y}; \boldsymbol{\mu}_k(\mathbf{x}; \boldsymbol{\beta}_k^*), \eta_k \boldsymbol{\Sigma}_k\right), \tag{2}$$

and $\phi_M(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the p.d.f. of an M-dimensional Gaussian distribution with expected mean vector $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. The term $\boldsymbol{\mu}_k(\mathbf{x}; \boldsymbol{\beta}_k^*)$ in Eq. (2) is the conditional expected value of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ in the kth sub-population; it is defined as follows:

$$\boldsymbol{\mu}_k(\mathbf{x}; \boldsymbol{\beta}_k^*) = \tilde{\mathbf{x}}^{*\prime} \boldsymbol{\beta}_k^* = \begin{bmatrix} \mathbf{x}_1^{*\prime} \boldsymbol{\beta}_{k1}^* \\ \vdots \\ \mathbf{x}_m^{*\prime} \boldsymbol{\beta}_{km}^* \\ \vdots \\ \mathbf{x}_M^{*\prime} \boldsymbol{\beta}_{kM}^* \end{bmatrix}, \tag{3}$$

where $\tilde{\mathbf{x}}^*$ denotes the realisation of $\tilde{\mathbf{X}}^*$ obtained when $\mathbf{X} = \mathbf{x}$. Thus, $\tilde{\mathbf{x}}^{*\prime} \boldsymbol{\beta}_k^*$ coincides with an M-dimensional vector whose mth element is a linear combination of the realisations of the P_m regressors selected for the prediction of Y_m with weights given by the elements of vector $\boldsymbol{\beta}_{km}^*$. Terms $\alpha_k \in (0, 1)$ and $\eta_k > 1$ are the weight of the typical observations in the kth sub-population and the factor contaminating the conditional variances and covariances of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ for the mild outliers in the kth sub-population, respectively. In robust statistics, it is generally assumed that at least half of the observations are typical; thus, it is also possible to consider $\alpha_k \in [0.5, 1)$. As a consequence of the constraint $\eta_k > 1$, η_k represents an inflation parameter for the elements of $\boldsymbol{\Sigma}_k$. $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_k, \alpha_k, \eta_k)$ is the parameter vector of model (2). The parameter vector of model (1) is given by $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k, \ldots, \boldsymbol{\psi}_K)$, where $\boldsymbol{\psi}_k = (\pi_k, \boldsymbol{\theta}_k)$; the number of free parameters in this vector is equal to $n_{\boldsymbol{\psi}} = 3K - 1 + K(P^* + M) + K\frac{M(M+1)}{2}$.

In summary, the conditional p.d.f. $f(\mathbf{y}|\mathbf{x}; \boldsymbol{\psi})$ in Eq. (1) can be interpreted as a weighted average (namely, a mixture) of K Gaussian regression models with weights π_k, $k = 1, \ldots, K$. The kth component of this mixture represents a multivariate seemingly unrelated contaminated Gaussian linear regression model with intercepts and regression coefficients $\boldsymbol{\beta}_k^*$, symmetric and positive definite covariance matrix $\boldsymbol{\Sigma}_k$, proportion of typical points α_k and inflation parameter η_k. Thanks to the non-diagonal structure of the variance-covariance matrices $\boldsymbol{\Sigma}_k, k = 1, \ldots, K$, the proposed model is able to account for correlated random disturbances within each of the K sub-populations associated with the mixture (1). Since the contaminated Gaussian distribution (2) is a mixture of two Gaussian linear regression models which are both associated with the kth component of the mixture in Eq. (1), the model defined by this latter equation can also be considered as a mixture of $2K$ seemingly unrelated Gaussian clusterwise linear regression models, whose components can be grouped into K pairs, each of which contains two Gaussian components having the same expected values and proportional covariance matrices.

2.2 Comparisons with other linear clusterwise regression models

When specific conditions are met, some special linear regression models can be obtained from model (1).

- If $M > 1$ and $\mathbf{X}_m = \mathbf{X}\ \forall m$ (the same vector of predictors is considered for all responses), the following equality holds: $\tilde{\mathbf{x}}^* = \mathbf{I}_M \otimes \mathbf{x}^*$, where $\mathbf{I}_M$ is the identity matrix of order M and $\otimes$ denotes the Kronecker product operator (see, e.g., Magnus and Neudecker 1988). Equation (3) can be rewritten as

$$\boldsymbol{\mu}_k(\mathbf{x}; \boldsymbol{\beta}_k^*) = \left(\mathbf{I}_M \otimes \mathbf{x}^*\right)' \boldsymbol{\beta}_k^* = \mathbf{B}_k'\mathbf{x}, \ k = 1, \ldots, K, \tag{4}$$

where $\mathbf{B}_k = \left[\boldsymbol{\beta}_{k1}^* \cdots \boldsymbol{\beta}_{km}^* \cdots \boldsymbol{\beta}_{kM}^*\right]$. Thus, Eq. (1) reduces to the mixture of multivariate contaminated Gaussian regression models introduced by Mazza and Punzo (2020).

- If $M > 1$, $\alpha_k \to 1$ and $\eta_k \to 1$ $\forall k$ (there is no contamination in the data), the resulting model coincides with the mixture of multivariate seemingly unrelated linear regressions described in Galimberti and Soffritti (2020).
- If $\alpha_k \to 1$, $\eta_k \to 1$ $\forall k$ and $\mathbf{X}_m = \mathbf{X}$ $\forall m$ (there is no contamination in the data and the same vector of predictors is considered for all responses), Eq. (1) reduces to a mixture of either univariate Gaussian linear regression models (see, e.g., De Sarbo and Cron 1988; De Veaux 1989; Quandt and Ramsey 1978) or multivariate Gaussian linear regression models (see Jones and McLachlan 1992).
- If $\alpha_k \to 1$, $\eta_k \to 1$ $\forall k$, $\mathbf{X}_m = \mathbf{X}$ $\forall m$ and $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}^*$ $\forall k$ (there is no contamination in the data, the same vector of predictors is considered for all responses and their effects are the same across all the sub-populations), the resulting model coincides with a linear regression model with error terms distributed according to a mixture of K either univariate Gaussian distributions (Bartolucci and Scaccia 2005) or multivariate Gaussian distributions (Soffritti and Galimberti 2011).
- If $M > 1$, $\alpha_k \to 1$, $\eta_k \to 1$ $\forall k$, $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}^*$ $\forall k$ (there is no contamination in the data and the effects of the predictors are the same across all the sub-populations), a multivariate seemingly unrelated linear regression model whose error terms are assumed to follow a Gaussian mixture model is obtained (Galimberti et al. 2016).

Seemingly unrelated regression models represent 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, Eq. (1) can also be seen as a mixture of multivariate contaminated Gaussian regression models 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 these latter models has not been addressed yet. Models obtained from Eq. (1) by embedding different constraints on the regression coefficients could also be employed in any practical application in which the relevant regressors for each response cannot be established from a priori information and, thus, the choice of the regressors to be used for the M responses is questionable. As it will be illustrated in Sect. 4, in such situations strategies based on a joint use of models (1) and variable selection techniques could be devised and employed.

2.3 Identifiability

A preliminary requirement for the consistency and other asymptotic properties of the ML estimator is represented by identifiability of the model parameters. Thus, before detailing ML estimation of $\boldsymbol{\psi}$, a discussion about identifiability of model (1) is provided here. Consider the class of models $\mathfrak{F} = \{\mathfrak{F}_K, K = 1, \ldots, K_{max}\}$, where $\mathfrak{F}_K =$

$\{f(\mathbf{y}|\mathbf{x};\boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$, $f(\mathbf{y}|\mathbf{x};\boldsymbol{\psi})$ is the p.d.f. of $\mathbf{Y}|\mathbf{X}=\mathbf{x}$ under the seemingly unrelated contaminated Gaussian linear clusterwise regression model of order K defined in (1) and K_{max} denotes the maximum order specified by the researcher for that model. This class is said to be identifiable if, for any two models M, $\tilde{M} \in \mathfrak{F}$ with parameters $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k, \ldots, \boldsymbol{\psi}_K)$ and $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\psi}}_1, \ldots, \tilde{\boldsymbol{\psi}}_k, \ldots, \tilde{\boldsymbol{\psi}}_{\tilde{K}})$, respectively,

$$\sum_{k=1}^{K} \pi_k h\left(\mathbf{y}; \boldsymbol{\theta}_k\right) = \sum_{k=1}^{\tilde{K}} \tilde{\pi}_k h\left(\mathbf{y}; \tilde{\boldsymbol{\theta}}_k\right) \; \forall\, \mathbf{y} \in \mathbb{R}^M$$

implies that $K = \tilde{K}$ and $\boldsymbol{\psi} = \tilde{\boldsymbol{\psi}}$.

Several types of non-identifiability can affect the model class $\mathfrak{F}$. A first type is due to invariance to relabeling the components of the mixture (also known as label-switching). Non-identifiability can also be caused by potential overfitting associated with empty components or equal components (see, e.g., Frühwirth-Schnatter 2006). Imposing suitable constraints on the parameter space $\boldsymbol{\Psi}$ can prevent such sources of non-identifiability for $\mathfrak{F}$. Another type of non-identifiability affecting this class is specifically associated with the use of finite mixtures in linear regression analysis with fixed covariates, which requires an additional constraint on the number of components of the mixture (1) (see Hennig 2000). Non-identifiability due to empty components is avoided by requiring the positivity of all the mixing weights π_k. Conditions specifically devised for ensuring identifiability of mixtures of contaminated Gaussian regression models are provided in Mazza and Punzo (2020). These results have been exploited in Theorem 1 to show that model (1) is identifiable if the parameters $(\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_k)$, $k = 1, \ldots, K$, are pairwise distinct and the order K is exceeded by the number of distinct $(P_m - 1)$-dimensional hyperplanes required to cover the covariates employed for the prediction of Y_m, for $m = 1, \ldots, M$. In order to state Theorem 1, the following notation is also required: $\|\cdot\|_F$ is the element-wise matrix 2-norm (also known as the Frobenious norm); $H^{P_m-1} = \{\mathbf{x}_m \in \mathbb{R}^{P_m} : \boldsymbol{\lambda}'\mathbf{x}_m = c, \boldsymbol{\lambda} \in \mathbb{R}^{P_m}, \boldsymbol{\lambda} \neq \mathbf{0}\}$ is a $(P_m - 1)$-dimensional hyperplane; J_m is the minimum number of such hyperplanes required to cover the covariates $\mathbf{x}_m$; $\mathcal{H}^{P_m-1}$ is the space of $(P_m - 1)$-dimensional hyperplanes of $\mathbb{R}^{P_m}$.

Theorem 1 *Let $M \in \mathfrak{F}$ and $\tilde{M} \in \mathfrak{F}$ be two models, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k, \ldots, \boldsymbol{\psi}_K)$ and $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\psi}}_1, \ldots, \tilde{\boldsymbol{\psi}}_k, \ldots, \tilde{\boldsymbol{\psi}}_{\tilde{K}})$ the corresponding parameters and, without loss of generality, $K \geq \tilde{K}$. If*

(C1) $K < J_m$ for $m = 1, \ldots, M$, where

$$J_m := \min\left\{q_m : \{\boldsymbol{x}_{im}, i \in \mathcal{I}_m\} \subseteq \bigcup_{b=1}^{q_m} H_b^{P_m-1} : H_b^{P_m-1} \in \mathcal{H}^{P_m-1}\right\},$$

with $\mathcal{I}_m$ being an index set associated with the distinct covariate points available for the prediction of Y_m, and

(C2) $k \neq l$, *with* $k, l \in \{1, \ldots, K\}$, *implies*

$$\|\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_l^*\|_F^2 + \|\boldsymbol{\Sigma}_k - a\boldsymbol{\Sigma}_l\|_F^2 \neq 0 \; \forall a > 0,$$

then the class $\mathfrak{F}$ *is identifiable.*

Conditions *(C1)* and *(C2)* are obtained from Mazza and Punzo (2020) after suitable modifications of similar conditions required for the identifiability of their mixtures of contaminated Gaussian regression models. In particular, condition *(C2)* results from a simple substitution of the vector $\boldsymbol{\beta}_k^*$ of model (1) for the matrix $\mathbf{B}_k$ introduced in Eq. (4) containing the intercepts and regression coefficients in the kth component of the regression mixture model developed by Mazza and Punzo (2020). The modifications involved in the definition of the condition *(C1)* derive from the fact that each $Y_m \in \mathbf{Y}$ may have its own covariates and, thus, M different restrictions on K have to be required, each one involving a (possibly) different minimum number of low-dimensional hyperplanes to cover those covariates. As a consequence, the proof of Theorem 1 can be obtained by exploiting the same arguments illustrated in Mazza and Punzo (2020) for the proof of their theorem about identifiability of mixtures of contaminated Gaussian regression models.

2.4 Maximum likelihood estimation

The ML estimation of the parameters $\boldsymbol{\psi}$ is carried out here for a fixed value of K. Given a sample $\mathcal{S}$ of I independent observations drawn from model (1), the model log-likelihood is equal to $\ell(\boldsymbol{\psi}) = \sum_{i=1}^{I} \ln \left(\sum_{k=1}^{K} \pi_k h\left(\mathbf{y}_i; \boldsymbol{\theta}_k\right)\right)$. Following Mazza and Punzo (2020), ML estimates $\hat{\boldsymbol{\psi}}$ can be computed by means of an ECM algorithm, which represents a variant of the EM algorithm usually employed for the computation of ML estimates from incomplete data. In the considered situation, the missing information is twofold. On the one hand, there is a classical source of incompleteness of any mixture model associated with the component memberships of the I sample observations. On the other hand, it is not known whether such observations are outliers with reference to any component or not. These two sources can be described by two different types of K-dimensional vectors. For the ith sample observation, they are given by $\mathbf{z}_i$ and $\mathbf{u}_i$, respectively: $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$, with $z_{ik} = 1$ if the ith observation comes from the kth component and $z_{ik} = 0$ otherwise; $\mathbf{u}_i = (u_{i1}, \ldots, u_{iK})'$, with $u_{ik} = 1$ if the ith observation is typical in the kth component and $u_{ik} = 0$ if it is an outlier, for $k = 1, \ldots, K$. Then, the set of complete data would be $\mathcal{S}_c = \{(\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1, \mathbf{u}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I, \mathbf{z}_I, \mathbf{u}_I)\}$, and the complete-data likelihood function is equal to

$$L_c(\boldsymbol{\psi}) = \prod_{i=1}^{I} \prod_{k=1}^{K} \Big\{ \pi_k \Big[\alpha_k \phi_M\Big(\mathbf{y}_i; \boldsymbol{\mu}_k(\mathbf{x}_i; \boldsymbol{\beta}_k^*), \boldsymbol{\Sigma}_k\Big)\Big]^{u_{ik}} \Big[(1-\alpha_k)\phi_M\Big(\mathbf{y}_i; \boldsymbol{\mu}_k(\mathbf{x}_i; \boldsymbol{\beta}_k^*), \eta_k \boldsymbol{\Sigma}_k\Big)\Big]^{1-u_{ik}} \Big\}^{z_{ik}}.$$

Thus, up to an additive constant, the complete-data log-likelihood function employed in the ECM algorithm for the computation of the parameter estimates can be expressed as follows:

$$\ell_c(\boldsymbol{\psi}) = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\Big[\ln \pi_k + u_{ik} \ln \alpha_k + (1-u_{ik}) \ln(1-\alpha_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \\ -\Big(\frac{M}{2}\ln \eta_k\Big)(1-u_{ik}) - \frac{1}{2}\Big(u_{ik} + \frac{1-u_{ik}}{\eta_k}\Big)\delta^2_{\boldsymbol{\Sigma}_k}\Big(\mathbf{y}_i, \boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}^*_k)\Big)\Big],$$

where

$$\delta^2_{\boldsymbol{\Sigma}_k}\left(\mathbf{y}_i, \boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}^*_k)\right) = (\mathbf{y}_i - \boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}^*_k))'\boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}^*_k)) \tag{5}$$

is the squared Mahalanobis distance between $\mathbf{y}_i$ and $\boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}^*_k)$ with respect to the matrix $\boldsymbol{\Sigma}_k$.

The hth iteration of the E-step in the ECM algorithm consists in calculating the conditional expectation of $l_c(\boldsymbol{\psi})$ on the basis of the current estimate $\boldsymbol{\psi}^{(h)}$ of the model parameters $\boldsymbol{\psi}$; up to an additive constant, this expected value can be expressed as follows:

$$Q\left(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)}\right) = \mathbb{E}_{\boldsymbol{\psi}^{(h)}}[l_c(\boldsymbol{\psi})] \\ = \sum_{i=1}^{I}\sum_{k=1}^{K} \hat{z}_{ik}^{(h)}\Big\{ \ln \pi_k^{(h)} + \hat{u}_{ik}^{(h)} \ln \alpha_k^{(h)} + (1-\hat{u}_{ik}^{(h)}) \ln(1-\alpha_k^{(h)}) + \\ + Q_i\Big(\boldsymbol{\beta}^*_k, \boldsymbol{\Sigma}_k|\boldsymbol{\psi}^{(h)}\Big)\Big\},$$

where

$$Q_i\Big(\boldsymbol{\beta}^*_k, \boldsymbol{\Sigma}_k|\boldsymbol{\psi}^{(h)}\Big) = -\frac{1}{2}\Big[\ln|\boldsymbol{\Sigma}_k^{(h)}| + M(1-\hat{u}_{ik}^{(h)})\ln \eta_k^{(h)} + \\ +\Big(\hat{u}_{ik}^{(h)} + \frac{1-\hat{u}_{ik}^{(h)}}{\eta_k^{(h)}}\Big)\delta^2_{\boldsymbol{\Sigma}_k^{(h)}}\left(\mathbf{y}_i, \boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}_k^{*(h)})\right)\Big],$$

$\hat{z}_{ik}^{(h)}$ and $\hat{u}_{ik}^{(h)}$ are the posterior probabilities (evaluated using $\boldsymbol{\psi}^{(h)}$) that the ith observation is generated from the kth component of the mixture (1) and that the ith observation is a typical point of such a component, respectively:

$$\hat{z}_{ik}^{(h)} = \mathbb{E}_{\boldsymbol{\psi}^{(h)}}[Z_{ik}|(\mathbf{x}_i, \mathbf{y}_i)] = \frac{\pi_k^{(h)} h\left(\mathbf{y}_i;\boldsymbol{\theta}_k^{(h)}\right)}{f\left(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\psi}^{(h)}\right)}, \tag{6}$$

$$\hat{u}_{ik}^{(h)} = \mathbb{E}_{\boldsymbol{\psi}^{(h)}}[U_{ik}|(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)] = \frac{\alpha_k^{(h)}\phi\Big(\mathbf{y}_i;\boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}_k^{*(h)}), \boldsymbol{\Sigma}_k^{(h)}\Big)}{h\left(\mathbf{y}_i;\boldsymbol{\theta}_k^{(h)}\right)}, \tag{7}$$

with $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})'$ denoting a K-dimensional multinomial random vector with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, and $U_{ik}|Z_{ik} = 1$ having a Bernoulli distribution with success probability of α_k.

As far as the conditional maximisation is concerned, the update of $\boldsymbol{\psi}^{(h)}$ is carried out by considering the following two parameter sub-vectors: $\boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{\beta}^*, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)'$, where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*)$, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. At the $(h+1)$th iteration of the ECM algorithm, $\boldsymbol{\gamma}^{(h)} = (\boldsymbol{\pi}^{(h)}, \boldsymbol{\beta}^{*(h)}, \boldsymbol{\Sigma}^{(h)}, \boldsymbol{\alpha}^{(h)})$ is updated through the maximisation of $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)})$ with respect to $\boldsymbol{\gamma}$ with $\boldsymbol{\eta}$ fixed at $\boldsymbol{\eta}^{(h)}$ (first CM step); then, the update of $\boldsymbol{\eta}^{(h)}$ is carried out by maximising $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)})$ with respect to $\boldsymbol{\eta}$ with $\boldsymbol{\gamma}$ fixed at $\boldsymbol{\gamma}^{(h+1)}$ (second CM step). The resulting updates of $\pi_k^{(h)}$, $\alpha_k^{(h)}$ and $\eta_k^{(h)}$ are:

$$\pi_k^{(h+1)} = \frac{1}{I}\sum_{i=1}^{I} \hat{z}_{ik}^{(h)},$$

$$\alpha_k^{(h+1)} = \frac{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)} \hat{u}_{ik}^{(h)}}{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}}, \tag{8}$$

$$\eta_k^{(h+1)} = \max\left\{1, \frac{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}(1-\hat{u}_{ik}^{(h)})\delta^2_{\boldsymbol{\Sigma}_k^{(h+1)}}\left(\mathbf{y}_i, \boldsymbol{\mu}_k(\mathbf{x}_i; \boldsymbol{\beta}_k^{*(h+1)})\right)}{M\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}(1-\hat{u}_{ik}^{(h)})}\right\}. \tag{9}$$

Such updates coincide with the ones reported in Mazza and Punzo (2020) for their model. Based on Eq. (9), it is possible to highlight that the update $\eta_k^{(h+1)}$ will be larger when the kth component is highly contaminated by the presence of outliers (i.e., when it is characterised by many observations with a small value of $\hat{u}_{ik}^{(h)}$ and a large value of the squared Mahalanobis distance from $\boldsymbol{\mu}_k(\mathbf{x}_i; \boldsymbol{\beta}_k^{*(h+1)})$). As far as the remaining parameters are concerned, their updates are (details are reported in the Appendix):

$$\boldsymbol{\beta}_k^{*(h+1)} = \left(\sum_{i=1}^{I} \hat{z}_{ik}^{(h)} \hat{w}_{ik}^{(h)} \tilde{\mathbf{x}}_i^* \boldsymbol{\Sigma}_k^{(h)^{-1}} \tilde{\mathbf{x}}_i^{*'}\right)^{-1} \left(\sum_{i=1}^{I} \hat{z}_{ik}^{(h)} \hat{w}_{ik}^{(h)} \tilde{\mathbf{x}}_i^* \boldsymbol{\Sigma}_k^{(h)^{-1}} \mathbf{y}_i\right), \tag{10}$$

$$\boldsymbol{\Sigma}_k^{(h+1)} = \frac{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)} \hat{w}_{ik}^{(h)} \left(\mathbf{y}_i - \tilde{\mathbf{x}}_i^{*'} \boldsymbol{\beta}_k^{*(h+1)}\right)\left(\mathbf{y}_i - \tilde{\mathbf{x}}_i^{*'} \boldsymbol{\beta}_k^{*(h+1)}\right)'}{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}}, \tag{11}$$

where

$$\hat{w}_{ik}^{(h)} = \hat{u}_{ik}^{(h)} + \frac{1-\hat{u}_{ik}^{(h)}}{\eta_k^{(h)}}. \tag{12}$$

It is worth noting that the matrix $\sum_{i=1}^{I} \hat{z}_{ik}^{(h)} \hat{w}_{ik}^{(h)} \tilde{\mathbf{x}}_i^* \boldsymbol{\Sigma}_k^{(h)^{-1}} \tilde{\mathbf{x}}_i^{*'}$ in (10) has to be nonsingular; otherwise, the update $\boldsymbol{\beta}_k^{*(h+1)}$ cannot be computed. Equation (10) also highlights that this update can be considered as a generalised least squares estimate with weights depending on $\hat{w}_{ik}^{(h)}$; this latter term also affects the update $\boldsymbol{\Sigma}_k^{(h+1)}$ in (11), which represents a weighted sum of squared residuals. Using such weights leads to a reduction in the effects of the outliers on the estimation of $\boldsymbol{\beta}_k^{*(h+1)}$; thus, this approach provides robust estimates of $\boldsymbol{\beta}_k^{*(h+1)}$, for $k = 1, \ldots K$. Furthermore, based on (12), sample observations with the highest posterior estimated probabilities of being generated from the kth component and of representing typical points in the kth component will have the largest impact on the updates of both the regression coefficients and covariances within that component.

Once the convergence is reached and the ML estimates $\hat{\boldsymbol{\psi}}$ are computed, by exploiting Eq. (6) the ECM algorithm provides estimates of the posterior probabilities $\mathbb{P}_{\hat{\boldsymbol{\psi}}}[Z_{ik} = 1|(\mathbf{x}_i, \mathbf{y}_i)] = \hat{z}_{ik}$, $i = 1, \ldots, I$, $k = 1, \ldots, K$. Such estimated probabilities can be employed to partition the I sample observations into K clusters, by assigning each observation to the component showing the highest posterior probability; for the ith observation:

$$\text{MAP}(\hat{z}_{ik}) = \begin{cases} 1 & \text{if } \max_h\{\hat{z}_{ih}\} \text{ occurs when } h = k; \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, Eq. (7) allows to compute the estimated posterior probabilities $\mathbb{P}_{\hat{\boldsymbol{\psi}}}[U_{ik} = 1|(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{z}}_i)] = \hat{u}_{ik}$, and an intra-cluster distinction between typical observations and mild outliers can be defined: the ith observation will be classified as an outlier of the hth cluster, where h is the label of the component for which $\text{MAP}(\hat{z}_{ik}) = 1$, if $\hat{u}_{ih} < 0.5$. From the ML estimates $\hat{\boldsymbol{\psi}}$ and Eq. (5) it is also possible to compute the estimated squared Mahalanobis distances $\hat{d}_{ik}^2 = \delta_{\hat{\boldsymbol{\Sigma}}_k}^2\left(\mathbf{y}_i, \hat{\boldsymbol{\mu}}_k(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_k^*)\right)$, $i = 1, \ldots, I$, $k = 1, \ldots, K$, which can be employed as multivariate measures of the outlyingness of the I sample observations with respect to the K clusters detected by the model. From the definition of the squared Mahalanobis distance given in Eq. (5) and the expressions for $\hat{u}_{ik}^{(h)}$ and $\hat{w}_{ik}^{(h)}$ reported in Eqs. (7) and (12), respectively, it is possible to express both $\hat{u}_{ik}$ and $\hat{w}_{ik}$ as decreasing functions of $\hat{d}_{ik}^2$ (see Mazza and Punzo 2020, for the explicit expressions). Thus, atypical observations could also be detected and studied by considering the values of $\hat{d}_{ik}^2$ $\forall (i, k) \in \{i \in \{1, \ldots, I\}, k : \text{MAP}(\hat{z}_{ik}) = 1\}$ and by focusing on the largest values obtained in this way (see McLachlan and Peel 2000, p. 232).

2.5 Technical details about the ECM algorithm

A crucial point of any EM-based algorithm is the choice of the starting values for the model parameters (i.e., $\boldsymbol{\psi}^{(0)}$). Multiple executions of the algorithm in association with multiple random initialisations or approaches based on non-random choices of either $\boldsymbol{\psi}^{(0)}$ or the missing information can provide a solution (see, e.g., Biernacki et al. 2003; Karlis and Xekalaki 2003). As far as the ECM algorithm described above is concerned, the initialisation technique illustrated in Mazza and Punzo (2020) could be modified so as to be employed also for model (1). This task would require setting the initial values

$\hat{z}_{ik}^{(0)}$, $i = 1, \ldots, I$, $k = 1, \ldots, K$, equal to the posterior probabilities from the EM algorithm for the estimation of the seemingly unrelated Gaussian clusterwise linear regression models, which are nested in model (1) when $\alpha_k \to 1^-$ and $\eta_k \to 1^+$, $k = 1, \ldots, K$; furthermore, $\hat{u}_{ik}^{(0)} = 0.999$, $i = 1, \ldots, I$, $k = 1, \ldots, K$. Another strategy for the initialisation of $\boldsymbol{\psi}$ which exploits the relationship between model (1) and seemingly unrelated Gaussian clusterwise linear regression models (see Sect. 2.2) could be composed of the following three steps. Firstly, a Gaussian mixture model with K components is fitted to the sample residuals of a seemingly unrelated linear regression model (Srivastava and Giles 1987); this allows to obtain the starting values $\pi_k^{(0)}$ and $\boldsymbol{\Sigma}_k^{(0)}$. Secondly, the starting values $\boldsymbol{\beta}_k^{*(0)}$ are obtained from the fitting of K different seemingly unrelated linear regression models, one for each cluster of the partition associated with the Gaussian mixture model considered in the previous step. Thirdly, $\alpha_k^{(0)}$ and $\eta_k^{(0)}$, $k = 1 \ldots, K$, are set equal to 0.999 and 1.001, respectively. Models involved in the first two steps can be estimated through the packages `mclust` (Scrucca et al. 2017) and `systemfit` (Henningsen and Hamann 2007) in the `R` environment (R Core Team 2021). In the analyses of Sects. 3 and 4 , the ECM algorithm has been initialised using this latter strategy. Furthermore, since $(1 - \alpha_k)$ in model (1) can be considered as the proportion of outliers in the kth sub-population, when this model is employed for outlier detection, a reasonable requirement is that in each cluster the number of typical observations cannot be smaller than the number of outliers, that is $\alpha_k \in [0.5, 1)$ $\forall k$. To guarantee this result, constraints on the estimation of α_k, $k = 1, \ldots, K$, have been included in the ECM algorithm; namely, Eq. (8) has been modified as follows: $\alpha_k^{(h+1)} = \max \left\{ 0.5, \frac{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)} \hat{u}_{ik}^{(h)}}{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}} \right\}$.

In order to avoid premature stops of the ECM algorithm 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, a convergence criterion based on the Aitken acceleration (Aitken 1926) has been adopted. It consists in stopping the algorithm when $|\ell_A^{(h+1)} - \ell(\boldsymbol{\psi}^{(h)})| < \epsilon$, where $0 < \epsilon < +\infty$, $\ell_A^{(h+1)}$ is $(h+1)$th Aitken accelerated estimate of the log-likelihood limit, and $\ell(\boldsymbol{\psi}^{(h)})$ is the incomplete log-likelihood evaluated at $\boldsymbol{\psi}^{(h)}$ (see, e.g., McNicholas 2010). Furthermore, a criterion based on a maximum number of iterations for the ECM algorithm has been employed. In the analyses of Sects. 3 and 4 , the maximum number of iterations and ϵ have been set equal to 500 and 10^{-6}, respectively. Furthermore, in order to circumvent the possible issue of unbounded likelihood associated with a degenerate model, the ECM algorithm has been developed by embedding some constraints on the eigenvalues of $\boldsymbol{\Sigma}_k^{(h)}$ for $k = 1, \ldots, K$. Namely, for all estimated covariance matrices, the ratio between the smallest and the largest eigenvalues is required to be not lower than 10^{-10}.

2.6 Determining the value of *K*

As illustrated in Sect. 2.4, the ML estimation of $\boldsymbol{\psi}$ based on the ECM algorithm is carried out for a given number of mixture components. When this number is not known and has to be determined from the data $\mathcal{S}$, it is common practice to employ model selection criteria able to take account of different aspects which are considered relevant when evaluating the adequacy of a model (see, e.g., Depraetere and Vandebroek 2014; Frühwirth-Schnatter 2006). For example, the Bayesian Information Criterion (Schwarz 1978) provides a trade-off between the fit and the model complexity; it can be computed as follows:

$$BIC(K) = 2\ell(\hat{\boldsymbol{\psi}}) - n_{\boldsymbol{\psi}} \ln I.$$

Model selection criteria that also consider the uncertainty of the estimated partition of the sample observations could be employed. An example is represented by the integrated completed likelihood (Biernacki et al. 2000), which can be computed according to different ways of measuring the uncertainty of the estimated partition (see, e.g., Andrews and McNicholas 2011; Baek and McLachlan 2011):

$$ICL_1(K) = 2\ell(\hat{\boldsymbol{\psi}}) - n_{\boldsymbol{\psi}} \ln I + 2\sum_{i=1}^{I}\sum_{k=1}^{K} \mathrm{MAP}(\hat{z}_{ik}) \ln \hat{z}_{ik},$$

$$ICL_2(K) = 2\ell(\hat{\boldsymbol{\psi}}) - n_{\boldsymbol{\psi}} \ln I + 2\sum_{i=1}^{I}\sum_{k=1}^{K} \hat{z}_{ik} \ln \hat{z}_{ik}.$$

These latter criteria penalize complex models more severely than BIC because of the presence of an additional penalty, which represents the estimated mean entropy. Thus, when using these criteria in comparison with the BIC, one cluster should be less likely split into two different components. ICL_1 and ICL_2 differ on whether a soft (i.e., $\hat{z}_{ik}$) or hard (i.e., $\mathrm{MAP}(\hat{z}_{ik})$) clustering is considered in the estimation of the mean entropy. Higher values of these criteria indicate better-fit models; as it will be illustrated in Sect. 4, BIC, ICL_1 and ICL_2 can also be employed to select the predictors to be considered in the linear terms employed in the prediction of the M responses in model (1).

3 Results from Monte Carlo studies

3.1 Settings

This section focuses on the investigation of the effectiveness of models (1) (mixtures of contaminated seemingly unrelated Gaussian regressions, hereafter denoted as MCSG) in comparison with other approaches using simulated datasets. This task has been carried out in a multivariate setting with $M = 4$ responses, $P = 4$ covariates and datasets comprising $K = 3$ groups of observations. The additional models considered in the comparison are those described by Mazza and Punzo (2020) and Galimberti and Soffritti (2020). From now on, these latter models have been denoted as MCG (mixtures of contaminated Gaussian regressions) and MSG (mixtures of seemingly unrelated Gaussian regressions), respectively.

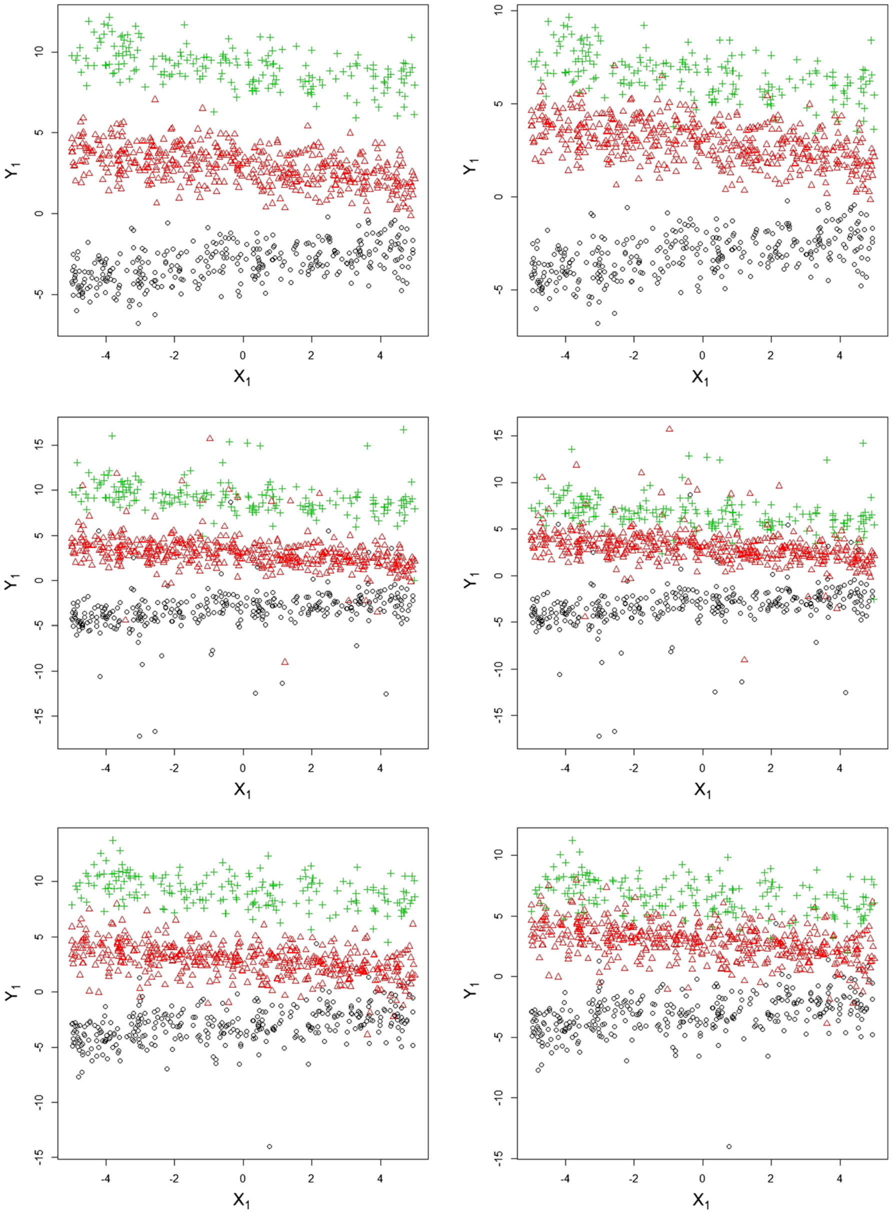

Fig. 1 Scatterplots of X_1 and Y_1 for samples of size $I = 1000$ generated from the first (upper panel), second (intermediate panel) and third (lower panel) data generation processes under higher ($\epsilon = 9$, left panels) and lower ($\epsilon = 6.5$, right panels) degree of separation. Black circle, red triangle and green plus correspond to $k = 1$, $k = 2$ and $k = 3$, respectively

The simulated datasets have been generated using three different data generation processes:

(a) MSG;
(b) MCSG with $\alpha_k = 0.9\ \forall k$, $\eta_1 = 40$, $\eta_2 = \eta_3 = 20$;
(c) mixtures of regression models with seemingly unrelated t-distributed errors (MSt), with $\nu_1 = \nu_2 = \nu_3 = 4$ degrees of freedom.

In all the regression models employed to generate the datasets, the response Y_m has been assumed to depend on X_m, for $m = 1, 2, 3, 4$; thus, $P_m = 1\ \forall m$. With each process, the following parameters have been employed: $\pi_1 = 0.3, \pi_2 = 0.5, \pi_3 = 0.2$, $\boldsymbol{\beta}_1^* = (-3, 0.2, -3, 0.2, -3, 0.2, -3, 0.2)'$, $\boldsymbol{\beta}_2^* = -\boldsymbol{\beta}_1^*$, $\boldsymbol{\beta}_3^* = (3 + \epsilon, -0.2, 3 + \epsilon, -0.2, 3 + \epsilon, -0.2, 3 + \epsilon, -0.2)$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.0 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1.0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1.0 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1.0 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1.00 & 0.75 & 0.75 & 0.75 \\ 0.75 & 1.00 & 0.75 & 0.75 \\ 0.75 & 0.75 & 1.00 & 0.75 \\ 0.75 & 0.75 & 0.75 & 1.00 \end{pmatrix}.$$

It is worth noting that the second and third components only differ in the intercepts of the four regression equations. Covariate values have been generated by a uniform distribution over the interval $(-5, 5)$. As concerns ϵ, two alternatives have been considered in order to produce two different degrees of separation between groups of observations: $\epsilon = 9$ (higher degree), $\epsilon = 6.5$ (lower degree). Figure 1 shows the scatterplots of the variables Y_1 and X_1 for a sample of size $I = 1000$ generated using the MSG (upper panel), MCSG (central panel) and MSt (lower panel) processes with $\epsilon = 9$ (on the left) and $\epsilon = 6.5$ (on the right). Due to the values of the regression coefficients employed to model the linear dependencies of Y_m and X_m across the three components, the scatterplots of Y_m and X_m for $m = 2, 3, 4$ are similar. Under each data generating process, 100 random samples of size I have been simulated for each ϵ. As far as the sample size is concerned, the following values have been examined: $I = 500, 1000$. Thus, the degree of separation and the sample size can be considered as experimental factors. This yields a total of 600 generated datasets for each I. The whole analysis has been run on an IBM x3750 M4 server with 4 Intel Xeon E5-4620 processors with 8 cores and 128GB RAM.

3.2 Results

A first analysis has been carried out where the MSG, MCG and MCSG models of order $K = 3$ have been fitted to each dataset. It is worth noting that the MCG models have been specified and estimated by assuming that each of the four responses depends on all covariates. Thus, using such models leads to non-parsimonious specifications for all the models that have generated the simulated datasets, as 12 regression coefficients for each component have been estimated although in fact they are equal to zero. The average execution times (over the 100 datasets with $I = 500$) for the MCSG models have ranged between 2.499 and 55.020 s, depending on the process and the specific value of ϵ employed to generate the datasets. Concerning the other two models, the minimum and maximum average execution times have resulted to be equal to 1.722 and 24.580 s with MSG models, 7.765 and 58.520 s with MCG models. It is worth noting that, since

Table 1 Estimation of α_k and η_k: averages and standard deviations of the estimates over 100 samples for the fitted MCG and MCSG models of order $K = 3$ ($I = 500$)

	MCG						MCSG					
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$
First process—high separation												
Average	0.987	0.981	0.983	1.030	1.043	1.017	0.985	0.989	0.989	1.008	1.000	1.000
SD	0.027	0.050	0.041	0.268	0.238	0.159	0.049	0.013	0.013	0.077	0.000	0.000
First process—low separation												
Average	0.967	0.959	0.963	1.060	1.063	1.167	0.980	0.983	0.980	1.033	1.031	1.027
SD	0.086	0.100	0.096	0.316	0.219	0.653	0.065	0.038	0.057	0.177	0.156	0.137
Second process—high separation												
Average	0.912	0.923	0.901	41.251	18.456	21.903	0.909	0.921	0.899	40.645	17.552	18.886
SD	0.040	0.036	0.056	12.373	10.450	15.307	0.043	0.056	0.060	11.610	10.622	8.109
Second process—low separation												
Average	0.910	0.937	0.895	39.388	16.137	19.439	0.906	0.940	0.892	39.748	14.666	16.705
SD	0.038	0.039	0.077	13.692	11.943	16.325	0.049	0.043	0.078	13.112	11.290	9.825
Third process—high separation												
Average	0.797	0.746	0.775	7.802	5.331	6.366	0.810	0.742	0.769	9.156	4.386	5.917
SD	0.155	0.149	0.161	8.739	6.755	5.281	0.149	0.146	0.171	12.476	2.044	4.457
Third process—low separation												
Average	0.784	0.761	0.790	7.602	5.672	7.025	0.802	0.751	0.782	7.838	4.444	5.852
SD	0.160	0.141	0.171	8.804	10.984	6.709	0.152	0.133	0.155	10.084	1.980	4.283

Table 2 Estimation of α_k and η_k: averages and standard deviations of the estimates over 100 samples for the fitted MCG and MCSG models with $K = 3$ ($I = 1000$)

	MCG						MCSG					
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$
First process—high separation												
Average	0.999	0.999	0.998	1.001	1.001	1.015	0.999	0.999	0.999	1.001	1.001	1.001
SD	0.000	0.000	0.012	0.000	0.000	0.145	0.000	0.000	0.000	0.000	0.000	0.000
First process—low separation												
Average	0.973	0.986	0.983	1.196	1.116	1.155	0.973	0.984	0.988	1.185	1.157	1.090
SD	0.080	0.061	0.064	0.444	0.260	0.450	0.079	0.062	0.050	0.394	0.343	0.272
Second process—high separation												
Average	0.901	0.909	0.901	41.061	19.267	18.435	0.903	0.912	0.897	40.358	18.882	18.850
SD	0.033	0.038	0.042	7.512	6.962	6.380	0.025	0.029	0.045	7.465	6.788	5.862
Second process—low separation												
Average	0.897	0.917	0.881	37.201	16.182	18.595	0.900	0.924	0.889	37.569	15.547	19.722
SD	0.049	0.052	0.092	11.199	9.403	11.045	0.033	0.040	0.071	8.730	9.090	12.401
Third process—high separation												
Average	0.860	0.713	0.789	9.317	5.433	6.132	0.849	0.727	0.789	7.846	4.344	6.132
SD	0.102	0.138	0.136	16.161	12.798	3.347	0.116	0.134	0.133	9.670	2.569	6.354
Third process—low separation												
Average	0.812	0.757	0.762	10.637	4.706	10.179	0.812	0.763	0.764	6.883	4.579	6.650
SD	0.122	0.124	0.155	33.859	2.597	25.126	0.124	0.113	0.140	8.518	3.895	8.262

Table 3 Bias and RMSE for the regression coefficients β_{km} under MSG, MCG and MCSG models of order $K = 3$ in the first process ($I = 500$)

	Bias			RMSE		
	MSG	MCG	MCSG	MSG	MCG	MCSG
High separation						
β_{11}	0.307	0.563	0.312	0.022	0.029	0.022
β_{12}	− 0.012	− 0.108	− 0.013	0.021	0.028	0.021
β_{13}	− 0.085	0.047	− 0.084	0.024	0.030	0.024
β_{14}	0.145	0.147	0.148	0.023	0.029	0.023
β_{21}	− 0.027	0.014	− 0.027	0.014	0.021	0.014
β_{22}	− 0.119	− 0.028	− 0.119	0.010	0.024	0.010
β_{23}	0.111	0.205	0.111	0.013	0.022	0.013
β_{24}	− 0.256	− 0.165	− 0.256	0.013	0.023	0.013
β_{31}	− 0.112	− 0.141	− 0.112	0.021	0.038	0.021
β_{32}	0.239	0.439	0.239	0.021	0.036	0.021
β_{33}	− 0.257	− 0.576	− 0.257	0.021	0.036	0.021
β_{34}	0.094	0.060	0.094	0.021	0.034	0.021
Low separation						
β_{11}	0.307	0.571	0.309	0.022	0.029	0.022
β_{12}	− 0.012	− 0.106	− 0.016	0.021	0.028	0.021
β_{13}	− 0.085	0.049	− 0.089	0.024	0.030	0.024
β_{14}	0.145	0.147	0.153	0.023	0.029	0.023
β_{21}	0.010	0.107	0.005	0.014	0.022	0.014
β_{22}	− 0.098	0.153	− 0.097	0.010	0.026	0.010
β_{23}	0.107	0.204	0.117	0.013	0.025	0.013
β_{24}	− 0.252	− 0.047	− 0.252	0.014	0.025	0.014
β_{31}	− 0.224	− 0.034	− 0.219	0.021	0.046	0.021
β_{32}	0.195	0.820	0.190	0.023	0.042	0.023
β_{33}	− 0.244	− 0.512	− 0.251	0.022	0.041	0.022
β_{34}	0.094	0.166	0.092	0.021	0.040	0.021

Biases have been multiplied by 100 to facilitate presentation

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 should be regarded as merely illustrative and can be reduced using more efficient implementations. In the first analysis, the performances of the three competing models have been evaluated with respect to the following aspects: *(i)* the estimation of the proportions of typical observations and the degrees of contamination (proper estimation of α_k and η_k); *(ii)* the ability to recover the true values of the unknown parameters (parameter recovery); *(iii)* the ability to recover the true partition of the sample observations (classification recovery). When evaluating properties of the parameter estimators using simulation studies under mixture models, there may be label switching issues. Several labeling methods have been proposed. For the models examined here, as in Bai et al. (2012),

Table 4 Bias and RMSE for the regression coefficients β_{km} under MSG, MCG and MCSG models of order $K = 3$ in the second process ($I = 500$)

	Bias			RMSE		
	MSG	MCG	MCSG	MSG	MCG	MCSG
High separation						
β_{11}	6.732	0.515	0.288	0.103	0.029	0.023
β_{12}	6.819	− 0.054	− 0.008	0.105	0.029	0.023
β_{13}	6.728	0.062	− 0.083	0.105	0.029	0.023
β_{14}	6.816	0.297	0.246	0.104	0.031	0.025
β_{21}	− 0.833	0.104	0.049	0.061	0.022	0.015
β_{22}	− 0.983	− 0.058	− 0.113	0.057	0.025	0.012
β_{23}	− 0.852	0.131	0.086	0.064	0.023	0.014
β_{24}	− 1.165	− 0.195	− 0.288	0.060	0.025	0.014
β_{31}	− 1.220	− 0.260	− 0.272	0.044	0.041	0.023
β_{32}	− 0.441	0.295	0.270	0.034	0.036	0.021
β_{33}	− 0.917	− 0.593	− 0.241	0.041	0.039	0.021
β_{34}	− 0.248	0.261	0.184	0.034	0.036	0.021
Low separation						
β_{11}	7.440	0.900	0.306	0.118	0.052	0.022
β_{12}	7.583	0.331	0.025	0.118	0.046	0.023
β_{13}	7.517	0.418	− 0.104	0.118	0.045	0.023
β_{14}	7.421	0.527	0.189	0.117	0.050	0.025
β_{21}	− 1.508	0.368	0.030	0.074	0.024	0.014
β_{22}	− 1.791	0.140	− 0.070	0.079	0.025	0.012
β_{23}	− 1.611	− 0.008	0.123	0.081	0.026	0.014
β_{24}	− 1.890	0.010	− 0.266	0.079	0.026	0.013
β_{31}	− 3.674	− 0.764	− 0.089	0.129	0.137	0.034
β_{32}	− 3.169	− 3.185	0.174	0.101	0.200	0.052
β_{33}	− 3.644	− 1.903	− 0.536	0.145	0.177	0.077
β_{34}	− 2.049	− 1.250	0.325	0.101	0.201	0.044

Biases have been multiplied by 100 to facilitate presentation

Yao et al. (2014) and Mazza and Punzo (2020), labels have been chosen by minimising the Euclidean distance to the true parameter values.

A second analysis has been carried out so as to obtain an evaluation of the three approaches without exploiting the knowledge of the true number of components. Thus, in addition to the models already examined in the first analysis, also models of order $K = 1, 2, 4, 5$ have been fitted to each dataset. All the obtained results have been employed to collect information on the following aspects: *(iv)* the capability to reach the best trade-off between the fit and model complexity; *(v)* the ability of BIC, ICL_1 and ICL_2 to detect the true value of K (comparison among information criteria).

Table 5 Bias and RMSE for the regression coefficients β_{km} under MSG, MCG and MCSG models of order $K = 3$ in the third process ($I = 500$)

	Bias			RMSE		
	MSG	MCG	MCSG	MSG	MCG	MCSG
High separation						
β_{11}	0.786	0.090	0.296	0.034	0.034	0.029
β_{12}	0.861	0.224	0.411	0.035	0.043	0.029
β_{13}	0.674	0.300	0.254	0.033	0.041	0.030
β_{14}	0.532	0.108	− 0.157	0.035	0.043	0.027
β_{21}	0.145	− 0.014	0.055	0.018	0.037	0.016
β_{22}	0.109	− 0.810	− 0.003	0.017	0.045	0.014
β_{23}	− 0.082	− 0.211	− 0.152	0.020	0.041	0.018
β_{24}	0.162	− 0.023	0.027	0.015	0.032	0.014
β_{31}	− 0.206	− 1.520	− 0.273	0.029	0.056	0.027
β_{32}	− 0.384	− 0.092	− 0.319	0.031	0.061	0.027
β_{33}	0.784	0.293	0.425	0.027	0.063	0.026
β_{34}	0.060	0.326	0.384	0.026	0.049	0.025
Low separation						
β_{11}	0.312	− 0.218	0.101	0.032	0.032	0.026
β_{12}	0.411	0.024	0.264	0.029	0.035	0.028
β_{13}	0.354	0.011	0.182	0.033	0.035	0.029
β_{14}	− 0.019	− 0.297	− 0.246	0.029	0.034	0.026
β_{21}	0.026	0.124	0.048	0.017	0.038	0.017
β_{22}	− 0.117	− 0.536	0.155	0.018	0.039	0.016
β_{23}	0.105	0.232	− 0.108	0.022	0.043	0.018
β_{24}	0.371	− 0.038	0.156	0.017	0.038	0.016
β_{31}	− 0.336	− 3.023	0.052	0.056	0.138	0.034
β_{32}	0.334	− 2.051	− 1.141	0.057	0.166	0.066
β_{33}	1.120	0.634	− 1.330	0.169	0.110	0.128
β_{34}	− 0.296	− 1.419	− 0.377	0.059	0.151	0.047

Biases have been multiplied by 100 to facilitate presentation

3.2.1 Estimation of α_k and η_k

The aspect *(i)* has been studied for the fitted MCG and MCSG models with $K = 3$. Under the first two data generation processes, the averages of the estimated proportions of good points ($\hat{\alpha}_k$) and the estimated inflation parameters ($\hat{\eta}_k$) are close to their true values under both MCG and MCSG models, regardless of the level of separation and the sample size (see the upper part of Tables 1 and 2). However, it is worth noting that slightly lower standard deviations of such estimates have been registered under the first process, thus giving an indication of a higher stability of the obtained estimates; furthermore, the estimation of η_1, η_2 and η_3 under the second process appears to be characterised by a certain instability, which results to reduce as the sample size I

Table 6 Bias and RMSE for the regression coefficients β_{km} under MSG, MCG and MCSG models of order $K = 3$ in the first process ($I = 1000$)

	Bias			RMSE		
	MSG	MCG	MCSG	MSG	MCG	MCSG
High separation						
β_{11}	0.162	0.128	0.162	0.016	0.020	0.016
β_{12}	− 0.066	0.009	− 0.066	0.017	0.022	0.017
β_{13}	0.127	0.478	0.127	0.015	0.020	0.015
β_{14}	0.070	0.084	0.070	0.017	0.020	0.017
β_{21}	− 0.126	− 0.314	− 0.126	0.008	0.014	0.008
β_{22}	− 0.042	− 0.080	− 0.042	0.008	0.015	0.008
β_{23}	0.081	0.077	0.081	0.010	0.016	0.010
β_{24}	− 0.057	0.080	− 0.057	0.008	0.014	0.008
β_{31}	0.075	0.161	0.075	0.014	0.025	0.014
β_{32}	− 0.153	− 0.073	− 0.153	0.015	0.026	0.015
β_{33}	0.091	0.158	0.091	0.015	0.024	0.015
β_{34}	− 0.124	− 0.452	− 0.124	0.014	0.025	0.014
Low separation						
β_{11}	0.159	0.122	0.161	0.016	0.020	0.016
β_{12}	− 0.065	0.012	− 0.060	0.017	0.022	0.017
β_{13}	0.129	0.474	0.127	0.015	0.020	0.015
β_{14}	0.070	0.077	0.073	0.017	0.020	0.017
β_{21}	− 0.008	0.276	− 0.008	0.009	0.015	0.009
β_{22}	− 0.008	− 0.045	− 0.007	0.009	0.016	0.009
β_{23}	0.059	− 0.071	0.056	0.010	0.016	0.010
β_{24}	0.028	− 0.149	0.031	0.008	0.016	0.008
β_{31}	− 0.034	− 0.027	− 0.032	0.014	0.028	0.014
β_{32}	− 0.067	− 0.248	− 0.067	0.014	0.031	0.014
β_{33}	0.069	− 0.238	0.070	0.016	0.031	0.016
β_{34}	− 0.031	0.013	− 0.030	0.015	0.031	0.015

Biases have been multiplied by 100 to facilitate presentation

increases using both MCG and MCSG models. As far as the results from the analyses of the datasets generated using the third process are concerned (lower part of Tables 1 and 2), the estimated values of α_k and η_k, $k = 1, 2, 3$, are far from 1, regardless of the values of ϵ and I. Thus, the departure from a four-dimensional Gaussian distribution for the errors of the regression model has been detected within each of the three mixture components of both MCG and MCSG models for both sample sizes. The standard deviations of $\hat{\eta}_k$, $k = 1, 2, 3$ are high, and this result holds true particularly with MCG models and $I = 1000$.

Table 7 Bias and RMSE for the regression coefficients β_{km} under MSG, MCG and MCSG models of order $K = 3$ in the second process ($I = 1000$)

	Bias			RMSE		
	MSG	MCG	MCSG	MSG	MCG	MCSG
High separation						
β_{11}	6.928	0.092	0.217	0.086	0.020	0.015
β_{12}	7.415	0.161	0.116	0.094	0.385	0.019
β_{13}	6.835	− 0.304	− 0.269	0.088	0.390	0.015
β_{14}	6.101	− 0.221	− 0.219	0.081	0.021	0.018
β_{21}	− 0.277	− 0.140	− 0.102	0.033	0.016	0.010
β_{22}	− 0.100	0.045	0.077	0.031	0.015	0.011
β_{23}	− 0.246	− 0.003	− 0.055	0.033	0.386	0.010
β_{24}	− 0.276	− 0.178	− 0.103	0.034	0.016	0.009
β_{31}	− 0.906	− 0.264	− 0.185	0.030	0.027	0.015
β_{32}	− 0.218	0.214	− 0.036	0.026	0.389	0.015
β_{33}	− 0.916	− 0.396	− 0.233	0.031	0.029	0.016
β_{34}	− 0.502	− 0.099	− 0.157	0.027	0.026	0.014
Low separation						
β_{11}	6.911	− 0.051	0.147	0.092	0.020	0.015
β_{12}	7.924	0.014	0.299	0.105	0.023	0.019
β_{13}	7.733	0.175	− 0.075	0.101	0.018	0.014
β_{14}	6.543	− 0.234	− 0.239	0.090	0.023	0.017
β_{21}	− 0.713	0.223	− 0.126	0.049	0.018	0.010
β_{22}	− 0.354	0.219	0.198	0.048	0.019	0.010
β_{23}	− 0.668	− 0.148	− 0.084	0.050	0.018	0.009
β_{24}	− 0.286	0.143	0.252	0.044	0.016	0.009
β_{31}	− 2.667	0.019	− 0.876	0.081	0.085	0.116
β_{32}	− 1.447	− 0.236	− 0.033	0.072	0.080	0.068
β_{33}	− 2.959	0.591	0.184	0.092	0.087	0.035
β_{34}	− 2.173	0.732	− 1.039	0.081	0.111	0.091

Biases have been multiplied by 100 to facilitate presentation

3.2.2 Parameter recovery

The evaluation of the aspect *(ii)* has been focused on the regression coefficients β_{km} and has been carried out by computing the following quantities:

$$\text{Bias}\left(\hat{\beta}_{km}\right) = \frac{\sum_{r=1}^{100} \hat{\beta}_{km}^{(r)}}{100} - \beta_{km}, \; k = 1, 2, 3, \; m = 1, 2, 3, 4,$$

$$\text{RMSE}\left(\hat{\beta}_{km}\right) = \sqrt{\frac{\sum_{r=1}^{100} \left(\beta_{km} - \hat{\beta}_{km}^{(r)}\right)^2}{100}}, \; k = 1, 2, 3, \; m = 1, 2, 3, 4,$$

Table 8 Bias and RMSE for the regression coefficients β_{km} under MSG, MCG and MCSG models of order $K = 3$ in the third process ($I = 1000$)

	Bias			RMSE		
	MSG	MCG	MCSG	MSG	MCG	MCSG
High separation						
β_{11}	0.325	− 0.128	− 0.022	0.022	0.027	0.019
β_{12}	0.412	0.057	− 0.011	0.024	0.026	0.022
β_{13}	0.686	0.268	0.160	0.022	0.025	0.019
β_{14}	0.326	− 0.049	0.091	0.027	0.028	0.024
β_{21}	0.006	− 0.199	− 0.027	0.011	0.020	0.011
β_{22}	0.217	0.330	0.035	0.012	0.020	0.011
β_{23}	− 0.011	− 0.280	− 0.131	0.012	0.019	0.011
β_{24}	− 0.324	− 0.406	− 0.233	0.013	0.018	0.012
β_{31}	− 0.049	0.125	− 0.083	0.021	0.033	0.019
β_{32}	0.118	0.154	0.003	0.018	0.032	0.017
β_{33}	− 0.170	0.052	− 0.190	0.020	0.036	0.018
β_{34}	− 0.271	− 0.516	− 0.251	0.020	0.033	0.018
Low separation						
β_{11}	0.197	0.035	0.052	0.022	0.028	0.018
β_{12}	− 0.075	− 0.289	− 0.160	0.021	0.038	0.019
β_{13}	0.540	0.430	0.407	0.023	0.028	0.020
β_{14}	0.257	0.081	0.130	0.019	0.027	0.018
β_{21}	0.084	0.140	0.063	0.013	0.023	0.012
β_{22}	0.137	− 0.142	− 0.049	0.013	0.026	0.011
β_{23}	0.140	0.279	0.213	0.014	0.021	0.012
β_{24}	− 0.143	− 0.130	− 0.117	0.012	0.024	0.012
β_{31}	− 0.911	− 1.273	0.050	0.057	0.104	0.019
β_{32}	− 1.822	− 2.135	0.061	0.085	0.162	0.021
β_{33}	− 1.087	− 1.037	0.041	0.077	0.107	0.021
β_{34}	− 0.408	− 0.881	0.156	0.069	0.083	0.022

Biases have been multiplied by 100 to facilitate presentation

where $\hat{\beta}_{km}^{(r)}$ is the ML estimate of β_{km} obtained from the rth dataset ($r = 1, \ldots, 100$) using models of order $K = 3$. With $I = 500$ and under the first data generating process (Table 3), MSG and MCSG models show the same performance in terms of recovering the true values of the regression coefficients with both degrees of separation. The good performance of MCSG models is consistent with the proper estimation of α_k and η_k associated with these models under the first process (see the previous aspect). On the contrary, the inclusion of irrelevant predictors in the four regression equations (MCG models) leads to a slight increase in the RMSEs. With contaminated datasets of size $I = 500$, as expected, the lowest (absolute) biases and RMSEs are obtained using the MCSG model (see Table 4); there also seems to be a tendency for MCG models to perform slightly better than MSG models for the majority of the regression

Table 9 Classification recovery of the fitted MSG, MCG and MCSG models of order $K = 3$: average values (standard deviations) of the ARI index over 100 samples ($I = 500$)

Process	ϵ	MSG	MCG	MCSG
I	9	0.999 (0.003)	0.999 (0.003)	0.999 (0.003)
I	6.5	0.946 (0.018)	0.937 (0.028)	0.946 (0.018)
II	9	0.818 (0.024)	0.911 (0.027)	0.910 (0.031)
II	6.5	0.723 (0.094)	0.806 (0.100)	0.821 (0.087)
III	9	0.931 (0.033)	0.936 (0.037)	0.937 (0.040)
III	6.5	0.721 (0.147)	0.745 (0.145)	0.776 (0.129)

coefficients. When the datasets are generated with $I = 500$ and according to the third process, the highest accuracy in the estimation of the regression coefficients is obtained using MCSG models (see Table 5). It is also worth noting that, in spite of their ability to detect a departure from the Gaussian distribution within each component, MCG models show the lowest accuracy. Similar results have been obtained with $I = 1000$ (see Tables 6, 7 and 8).

3.2.3 Classification recovery

To obtain information on the aspect *(iii)*, the partitions of the sample units associated with the models of order $K = 3$ under each competing model class have been compared with the true partition; the agreement with this latter partition has been measured by resorting to the adjusted Rand index (ARI) (Hubert and Arabie 1985). When the datasets are generated using the first process and the highest level of separation (see the upper part of Tables 9 and 10), an almost perfect classification recovery ($ARI = 0.999$) is obtained by each of the three models regardless of the sample size. When the level of separation is low ($\epsilon = 6.5$), a slight decrease in the ability to recover the true partition of the sample observations is registered for all models and, in particular, for the MCG ones when $I = 500$ ($ARI = 0.937$). When there are outliers in the data and $\epsilon = 9$, the best performance is obtained using either MCG models or MCSG models with both sample sizes ($ARI = 0.91$); these latter models slightly outperform MCG models when $\epsilon = 6.5$. As far as MSG models are concerned, due to their inability to manage the presence of mild outliers in the data, the classification recovery appears to be markedly lower, especially with the lowest level of separation ($ARI = 0.723$ with $I = 500$, $ARI = 0.716$ with $I = 1000$). Under the third process and the highest level of separation, good performances are obtained by all models with both sample sizes ($ARI > 0.93$). When the level of separation is reduced, a general decrease in the capability to reconstruct the true partition is registered; MCSG models appear to be less affected by this tendency, regardless of the sample size.

3.2.4 Trade-off between fit and complexity

In order to study the aspect *(iv)*, for each dataset and each model class, the models of order $\hat{K}_{IC}$ have been selected, where IC denotes an information criterion ($IC \in$

Table 10 Classification recovery of the fitted MSG, MCG and MCSG models of order $K = 3$: average values (standard deviations) of the ARI index over 100 samples ($I = 1000$)

Process	ϵ	MSG	MCG	MCSG
I	9	0.999 (0.002)	0.999 (0.002)	0.999 (0.002)
I	6.5	0.951 (0.011)	0.949 (0.012)	0.951 (0.011)
II	9	0.803 (0.015)	0.914 (0.023)	0.916 (0.021)
II	6.5	0.716 (0.088)	0.823 (0.092)	0.831 (0.082)
III	9	0.941 (0.016)	0.943 (0.013)	0.944 (0.014)
III	6.5	0.706 (0.147)	0.814 (0.095)	0.814 (0.102)

$\{BIC, ICL_1, ICL_2\}$) and $\hat{K}_{IC} = \arg\max IC(K)$ for $K \in \{1, 2, 3, 4, 5\}$. Then, the average values of the 100 resulting values of $BIC(\hat{K}_{BIC})$, $ICL_1(\hat{K}_{ICL_1})$ and $ICL_2(\hat{K}_{ICL_2})$ have been computed within the three model classes. As expected, when datasets of $I = 500$ observations are generated without outliers (first process), the best trade-off between the fit and model complexity is reached by MSG models, regardless of the level of separation and the criterion employed to select the best model (see the upper part of Table 11). With these datasets, MCSG models slightly outperform MCG models. When there are outliers in the data (second process) or the error terms of the K regression models have tails heavier than the Gaussian ones (third process), MCSG shows the best performance in terms of capability to reach the best trade-off between fit and complexity, regardless of the level of separation and the criterion employed to select the best model (see the lower part of Table 11). Interestingly, when the outliers are generated using a MCSG model (second process), MSG models slightly outperform MCG models, regardless of the value of ϵ. Similar conclusions can be drawn also from the results obtained when $I = 1000$ (see Table 12).

3.2.5 Comparison among information criteria

As far as the aspect *(v)* is concerned, the attention has been focused on the number of times each value of K has been selected by each examined criterion. With datasets generated using the first process and the highest level of separation, all the examined information criteria always recognize the presence of three clusters, regardless of the fitted model and the sample size (see the upper part of Tables 13 and 14). If the level of separation is reduced ($\epsilon = 6.5$), the BIC still tends to correctly identify the presence of three clusters regardless of the fitted model only with the largest sample size. If $I = 500$, the same tendency is slightly weaker with MSG and MCSG models; the order of the models employed to generate the datasets is always underestimated by the BIC when MCG models are employed. ICL_1 and ICL_2 show a clear preference for $K = 3$ components only when models embedding the information on the relevant regressors (e.g., MSG and MCSG) are employed and the sample size is $I = 1000$. Otherwise, they generally underestimate the true number of clusters. Under the second process, when MSG models are fitted to the data, all the examined information criteria show a clear tendency to select $K = 4$ components an additional component accommodating

Table 11 Average values of $BIC(\hat{K}_{BIC})$, $ICL_1(\hat{K}_{ICL_1})$ and $ICL_2(\hat{K}_{ICL_2})$ over 100 samples ($I = 500$)

$BIC(\hat{K}_{BIC})$			$ICL_1(\hat{K}_{ICL_1})$			$ICL_2(\hat{K}_{ICL_2})$		
MSG	MCG	MCSG	MSG	MCG	MCSG	MSG	MCG	MCSG
First process—high separation								
– 5776	– 6002	– 5807	– 5776	– 6003	– 5808	– 5777	– 6003	– 5809
First process—low separation								
– 5731	– 5894	– 5753	– 5740	– 5895	– 5759	– 5748	– 5898	– 5764
Second process—high separation								
– 6650	– 6756	– 6558	– 6657	– 6776	– 6577	– 6674	– 6802	– 6603
Second process—low separation								
– 6601	– 6667	– 6508	– 6621	– 6682	– 6531	– 6655	– 6702	– 6555
Third process—high separation								
– 6979	– 7065	– 6886	– 6995	– 7076	– 6898	– 7012	– 7096	– 6919
Third process—low separation								
– 6866	– 6895	– 6775	– 6892	– 6906	– 6787	– 6906	– 6925	– 6806

Table 12 Average values of $BIC(\hat{K}_{BIC})$, $ICL_1(\hat{K}_{ICL_1})$ and $ICL_2(\hat{K}_{ICL_2})$ over 100 samples ($I = 1000$)

$BIC(\hat{K}_{BIC})$			$ICL_1(\hat{K}_{ICL_1})$			$ICL_2(\hat{K}_{ICL_2})$		
MSG	MCG	MCSG	MSG	MCG	MCSG	MSG	MCG	MCSG
First process—high separation								
−11,298	− 11,552	− 11,339	− 11,299	− 11,553	− 11,340	− 11,300	− 11,554	− 11,341
First process—low separation								
− 11,217	− 11,469	− 11,257	− 11,253	− 11,492	− 11,293	− 11,296	− 11,507	− 11,334
Second process—high separation								
− 13,116	− 13,202	− 13,000	− 13,131	− 13,251	− 13,049	− 13,167	− 13,313	− 13,111
Second process—low separation								
− 12,989	− 13,107	− 12,923	− 13,039	− 13,159	− 13,002	− 13,119	− 13,209	− 13,070
Third process—high separation								
− 13,699	− 13,760	− 13,541	− 13,773	− 13,786	− 13,568	− 13,833	− 13,829	− 13,611
Third process—low separation								
− 13,495	− 13,510	− 13,346	− 13,611	− 13,536	− 13,401	− 13,681	− 13,575	− 13,444

Table 13 Comparison among information criteria: number of selections over 100 samples for MSG, MCG and MCSG models of order $K \in \{1, 2, 3, 4, 5\}$ ($I = 500$)

K	$BIC(K)$			$ICL_1(K)$			$ICL_2(K)$		
	MSG	MCG	MCSG	MSG	MCG	MCSG	MSG	MCG	MCSG
First process—high separation									
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	100	100	100	100	100	100	100	100	100
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
First process—low separation									
1	0	0	0	0	0	0	0	0	0
2	25	100	51	52	100	72	76	100	85
3	75	0	49	48	0	28	24	0	15
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
Second process—high separation									
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	100	98	0	100	98	0	100	99
4	99	0	2	99	0	2	99	0	1
5	1	0	0	1	0	0	1	0	0
Second process—low separation									
1	0	0	0	0	0	0	0	0	0
2	0	99	50	0	99	75	0	100	94
3	11	1	50	15	1	25	19	0	6
4	89	0	0	85	0	0	81	0	0
5	0	0	0	0	0	0	0	0	0
Third process—high separation									
1	0	0	0	0	0	0	0	0	0
2	0	2	0	0	2	1	0	2	1
3	52	98	99	70	98	98	77	98	96
4	39	0	1	25	0	1	22	0	3
5	9	0	0	5	0	0	1	0	0
Third process—low separation									
1	0	0	0	0	0	0	0	0	0
2	40	100	89	82	100	100	93	100	100
3	24	0	11	7	0	0	4	0	0
4	27	0	0	10	0	0	3	0	0
5	9	0	0	1	0	0	0	0	0

Table 14 Comparison among information criteria: number of selections over 100 samples for MSG, MCG and MCSG models of order $K \in \{1, 2, 3, 4, 5\}$ ($I = 1000$)

K	$BIC(K)$			$ICL_1(K)$			$ICL_2(K)$		
	MSG	MCG	MCSG	MSG	MCG	MCSG	MSG	MCG	MCSG
First process—high separation									
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	100	100	100	100	100	100	100	100	100
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
First process—low separation									
1	0	0	0	0	0	0	0	0	0
2	0	13	0	0	49	0	17	84	24
3	100	87	100	100	51	100	83	16	76
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
Second process—high separation									
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	99	100	0	99	100	0	100	100
4	100	1	0	100	1	0	100	0	0
5	0	0	0	0	0	0	0	0	0
Second process—low separation									
1	0	0	0	0	0	0	0	0	0
2	0	19	4	0	80	17	0	93	68
3	0	81	91	1	20	81	8	7	31
4	100	0	5	99	0	2	92	0	1
5	0	0	0	0	0	0	0	0	0
Third process—high separation									
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	14	100	98	48	100	99	69	100	99
4	69	0	2	49	0	1	31	0	1
5	17	0	0	3	0	0	0	0	0
Third process—low separation									
1	0	0	0	0	0	0	0	0	0
2	1	88	12	44	100	88	81	100	100
3	19	12	87	13	0	12	10	0	0
4	67	0	1	40	0	0	9	0	0
5	13	0	0	3	0	0	0	0	0

outliers is typically selected), regardless of the level of separation and the sample size (see also Mazza and Punzo 2020). On the contrary, with both MCG and MCSG models, the three criteria almost always correctly identify three components, regardless of the sample size, provided that the degree of separation is high. When $\epsilon = 6.5$, the same result is obtained by the BIC in association with MCG and MCSG models and by ICL_1 in association with MCSG models only with the largest sample size; otherwise, due to both a low separation between two clusters and a low sample size, the examined criteria generally underestimate the true value of K. This behaviour is particularly evident when the selection of K is based on ICL_2. A possible explanation for this is that the penalty employed by ICL_2 (a function of the uncertainty of the estimated posterior probabilities $\hat{z}_{ik}$) is the most severe and is also expected to be particularly large whenever the analysed dataset contains true clusters which are not well separated. When the datasets are generated using the third process and the smallest sample size, the obtained results show that, if $\epsilon = 9$, the three criteria generally detect the true value of K (see the lower part of Table 13). This tendency appears to be stronger when MCG and MCSG models are employed. These results hold true also with $I = 1000$ except when MSG models are fitted to the data and K is selected using either the BIC or the ICL_1; in these latter situations the true K is overestimated. On the contrary, when the degree of separation is low, models of order $K = 2$ are generally selected from each examined model class according to ICL_1 and ICL_2, regardless of the sample size. Also this result could be due to the role played by the penalties employed by these two latter criteria in the presence of true clusters which are not well separated. As far as the BIC is concerned, it allows to detect the true number of components only when MCSG models are fitted to samples of size $I = 1000$. It also shows a tendency to underestimate the true K both with MCSG models fitted to smaller samples and with MCG models regardless of the sample size. Finally, a slight preference with MSG models of order $K = 2$ and $K = 4$ emerges in association with samples of size $I = 500$ and $I = 1000$, respectively.

4 Results from the analysis of canned tuna sales

The practical usefulness and effectiveness of the proposed models have been evaluated through the analysis of a dataset containing the volume of weekly sales (`Move`) for seven of the top 10 U.S. brands in the canned tuna product category for $I = 338$ weeks between September 1989 and May 1997 (Chevalier et al. 2003). Measures of the display activity (`Nsale`) and the log price (`Lprice`) of each brand in each week are also available. This dataset is included in the `R` package `bayesm` (Rossi 2012). The analysis here considers two products: Star Kist 6 oz. (SK) and Bumble Bee Solid 6.12 oz. (BBS). In order to study the dependence of canned tuna sales on prices and promotional activites for these two brands, the analysis has been carried out starting from the following vectors of variables: $\mathbf{Y} = (Y_1 =$ `Lmove SK`, $Y_2 =$ `Lmove BBS`), $\mathbf{X} = (X_1 =$ `Nsale SK`, $X_2 =$ `Lprice SK`, $X_3 =$ `Nsale BBS`, $X_4 =$ `Lprice BBS`), where `Lmove` denotes the logarithm of `Move`; thus, $M = 2$ and $P = 4$. Previous studies focused on other brands are illustrated in Galimberti et al. (2016) and Galimberti and Soffritti (2020).

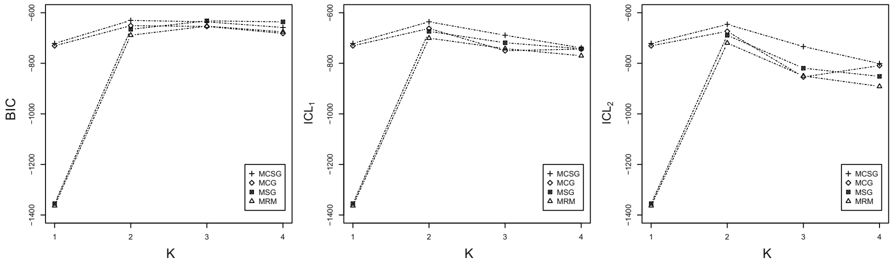

Fig. 2 Values of BIC, ICL_1 and ICL_2 for the best MCG, MCSG, MSG and MRM models by number of components in the analysis of tuna sales

Table 15 Maximised log-likelihood and values of BIC, ICL_1 and ICL_2 for six models selected from the classes MCSG, MCG, MSG and MRM in the analysis of tuna sales

Class	K	$\mathbf{X}_1$	$\mathbf{X}_2$	$\ell(\hat{\boldsymbol{\psi}})$	n_ψ	BIC	ICL_1	ICL_2
MCSG	2	X_1, X_2	X_2, X_3, X_4	– 242.5	25	– 630.5	– 636.0	– 646.1
MCG	2	X_2, X_3, X_4	X_2, X_3, X_4	– 247.0	27	– 651.1	– 662.3	– 673.5
MSG	2	X_1, X_2	X_3, X_4	– 277.5	19	– 665.6	– 673.8	– 689.2
MRM	2	X_2, X_4	X_2, X_4	– 289.2	19	– 688.9	– 700.5	– 719.9
MSG	3	X_2	X_3, X_4	– 240.4	26	– 632.2	– 737.4	– 865.7
MRM	3	X_2, X_3, X_4	X_2, X_3, X_4	– 224.6	35	– 653.0	– 750.0	– 877.9

The analysis has been carried out through MSG, MCG and MCSG models. The additional class comprising mixtures of linear Gaussian regression models (Jones and McLachlan 1992) has been included in the comparison; the notation employed for this model class is MRM. Models from each of these four classes have been estimated for $K \in \{1, 2, 3, 4\}$. Furthermore, since prices and promotional activities for one product could have an impact on the sales of the other product, models from MSG and MCSG classes have been specified and fitted by considering all possible sub-vectors of $\mathbf{X}$ as vectors $\mathbf{X}_m$, $m = 1, 2$, for each K. Thus, the analysis has also included an exhaustive search of the relevant regressors for both `Lmove SK` and `Lmove BBS`. For each K, $2^{P \cdot M} = 256$ different mixtures of regression models have been estimated either with contamination or without contamination; the overall number of estimated models is 2048. It is worth noting that none of the models employed in this analysis explicitly accounts for serial dependencies that may characterise this dataset.

Figure 2 shows the values of BIC, ICL_1 and ICL_2 for the fitted MCSG, MSG, MCG and MRM models which maximise each of these model selection criteria by K. An analysis based on a single linear regression model without contamination (MSG and MRM models with $K = 1$) is clearly inadequate according to all criteria. The best trade-off among the fit, the model complexity and the uncertainty of the estimated partition of the weeks is reached by models of order $K = 2$ for each of the four examined model classes. If model selection is only based on the fit and the model complexity, the best MCSG and MCG models still have $K = 2$ components, while MSG and MRM models of order $K = 3$ should be preferred.

Table 16 Parameter estimates of the overall best model for the analysis of tuna sales

$\hat{\boldsymbol{\psi}}$	$k = 1$	$k = 2$
$\hat{\pi}_k$	0.062	0.938
$\hat{\alpha}_k$	0.827	0.829
$\hat{\eta}_k$	13.44	6.80
$\hat{\boldsymbol{\beta}}'^{*}_{k1}$	(8.86, 0.59, – 4.68)	(8.65, 0.27, – 3.11)
$\hat{\boldsymbol{\beta}}'^{*}_{k2}$	(15.09, 3.91, 2.77, – 17.84)	(9.98, 0.25, 0.12, – 3.82)
$\hat{\boldsymbol{\Sigma}}_k$	$\begin{pmatrix} 0.043 & -0.022 \\ -0.022 & 0.126 \end{pmatrix}$	$\begin{pmatrix} 0.118 & 0.011 \\ 0.011 & 0.028 \end{pmatrix}$

Table 15 reports more detailed information about the six models which best fit the analysed dataset according to the three model selection criteria over the five examined values of K within each model class. All the examined criteria select a seemingly unrelated contaminated Gaussian linear clusterwise regression model of order $K = 2$ as the overall best model for studying the effect of prices and promotional activities on sales for the two brands. In this model, the log unit sales of SK canned tuna are regressed on the log prices and the promotional activities of the same brand; as far as the regressors for the BBS log unit sales are concerned, the selected regressors are the log prices of both brands and the promotional activities of BBS. From the parameter estimates (see Table 16) it emerges that the analysed dataset is characterised both by heterogeneity over time and by the presence of atypical observations. This latter feature seems to characterise the two clusters of weeks detected by the model almost in the same way (the estimated weights of the typical observations are $\hat{\alpha}_1 = 0.827$ and $\hat{\alpha}_2 = 0.829$); however, the strength of the contaminating effect on the conditional variances and covariances of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ results to be stronger in the first cluster, where the estimated inflation parameter for the elements of $\boldsymbol{\Sigma}_1$ is larger ($\hat{\eta}_1 = 13.44$). Heterogeneity over time appears to emerge both in some effects of the selected regressors and in the conditional expected variances and covariances of log sales for the typical observations. From the estimates of the regression equation for `Lmove SK` it emerges that sales of SK canned tuna are negatively affected by prices and positively affected by promotional activities of the same brand within both clusters detected by the model. However, the estimated effects of these two variables in the first cluster result to be stronger than those in the second cluster. Similar results have been obtained with reference to the regression equation for `Lmove BBS`, from which it also emerges that the log prices of SK canned tuna positively affect the log unit sales of the other brand, especially in the first cluster of weeks. As far as the estimated conditional variances and covariances are concerned, typical weeks in the first cluster appear to be characterised by values of `Lmove SK` which are more homogeneous than those of `Lmove BBC`; the opposite holds true for the typical weeks belonging to the second cluster. Heterogeneity over time appears to emerge also in the correlation between log sales of SK and BBS products, which is slightly positive (0.191) within the largest cluster of weeks, while a mild negative correlation (– 0.299) between `Lmove SK` and `Lmove BBC` is estimated in the weeks belonging to the first cluster.

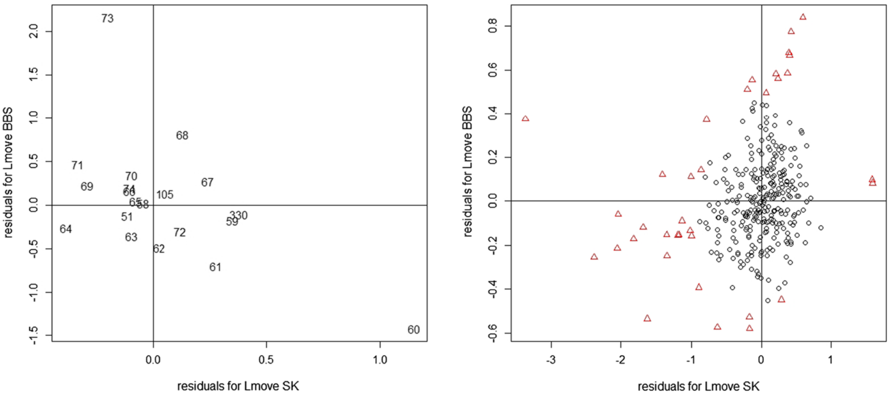

Fig. 3 Scatterplots of the estimated residuals for the weeks assigned to the first (left) and second (right) clusters detected by the overall best model for the analysis of tuna sales. Points of the first scatterplot are labelled with the number of the corresponding weeks. Black circle and red triangle in the second scatterplot correspond to typical and outlying weeks, respectively

The first cluster determined according to the highest estimated posterior probabilities of the selected model is composed of 20 weeks; 17 of these weeks are consecutive (from week no. 58 to week no. 74) and correspond to a period (from mid-October 1990 to mid-February 1991) characterised by a worldwide boycott campaign encouraging 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 selected model seems to suggest that such events may be one of the sources of the unobserved heterogeneity detected by the analysis. The fact that the estimated effects of all the selected regressors on the log prices of both products are stronger in the first cluster of weeks and weaker in the second cluster could be associated with those events. According to the rule for the intra-class distinction between typical observations and mild outliers illustrated in Sect. 2.4, some weeks have been classified as mild outliers within both clusters. As far as the first cluster is concerned, this has happened for week no. 60 (immediately after Halloween 1990) and week no. 73 (2 weeks immediately before Presidents day 1999). For these weeks, the estimated squared Mahalanobis distances $\hat{d}_{i1}^2$, equal to 36.68 and 37.82, respectively, appear to be extremely higher than those of the other 18 weeks of the same cluster, which are comprised between 0.05 and 7.05. From the estimated sample residuals $\mathbf{y}_i - \hat{\boldsymbol{\mu}}_1(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_1^*)$ for the 20 weeks belonging to the first cluster (see the scatterplot on the left side of Fig. 3) it emerges that week no. 60 noticeably deviates from the other weeks because log unit sales of SK tuna are slightly lower than the predicted value, while an opposite result characterises the log unit sales of BBS tuna. On the contrary, the selected model identifies week no. 73 as a mild outlier mainly because of a large overestimation of the sales of BBS tuna. Among the 318 weeks of the second cluster, 35 have resulted to be mild outliers, most of which are associated with holidays and special events that took place between September 1989 and mid-October 1990 or between mid-February and May 1997. The scatterplot with the estimated sample residuals $\mathbf{y}_i - \hat{\boldsymbol{\mu}}_2(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_2^*)$ for all the weeks of the second cluster (see the right side of Fig. 3) shows that, for the majority of the 35 mild outlying weeks, the reason for the outlyingness detected by the model has been an overestimation or an underestimation of the sales for either brands. The values of the estimated distances $\hat{d}_{i2}^2$ for the weeks that have been classified as typical are between 0.003 and 7.993; the minimum and maximum of the same distances for the outlying weeks are 8.20 and 114.95, respectively.

5 Conclusions

A new family of seemingly unrelated clusterwise linear regression models for possibly contaminated data has been introduced. Such models can account for heterogeneous regression data with mild outliers and multivariate correlated responses, each one depending on its own vector of covariates. This latter feature represents the main novelty of the models proposed here in reference with the ones described in Mazza and Punzo (2020). The new family encompasses several other types of Gaussian mixture-based linear regression models previously proposed in the literature. It also provides a more flexible framework for modelling data in applications where sample observations could be atypical and different covariates are expected to be relevant in the prediction of different responses, based on some prior information to be conveyed in the analysis. The new family could be made more flexible by exploiting the approach illustrated in Celeux and Govaert (1995), which allows to introduce constraints on the elements of the covariance matrices $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$, so that models with a lower number of variances and covariances of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ in the K sub-populations are obtained. Monte Carlo studies have shown that the choice of the number of components and the reconstruction of the true classification of the sample observations can be negatively affected by the inclusion of irrelevant regressors in a clusterwise linear regression model, especially with overlapping clusters of observations. Whenever the choice of the regressors to be considered in the specification of the linear predictor of each response is questionable, models introduced here can be employed in conjunction with techniques for variable selection (e.g., genetic algorithms, stepwise strategies) in a multivariate regression setting in order to detect the relevant predictors for each regression equation. Since the ECM algorithm for the ML estimation of the model parameters does not automatically produce any estimate of the covariance matrix of the ML estimator, additional computations are necessary to obtain an assessment of the sample variability of model parameter estimates. This task could be carried out by means of some approaches commonly employed under finite mixture models (see, e.g., McLachlan and Peel 2000). We are currently developing an extension of the methods proposed herein to some mixtures of Gaussian linear regression models with random covariates (Punzo and McNicholas 2017). Another avenue of future research is represented by the study of seemingly unrelated clusterwise regression models explicitly accounting for contaminated data and space/time-dependent observations.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. This study has been funded by the University of Bologna, Italy.

Data availability The datasets generated and analysed during the current study are available from the corresponding author upon request. The real dataset on canned tuna sales which has been analysed on Sect. 4 is freely downloadable from the `R` package `bayesm` (Rossi 2012).

Code Availability The `R` code developed by the authors for the implementation of the ECM algorithm illustrated in Sects. 2.4–2.6 is available from the corresponding author upon request.

Declarations

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix A: Update of $\boldsymbol{\beta}_k^*$ and $\boldsymbol{\Sigma}_k$

The updates of the model parameters $\boldsymbol{\beta}_k^*$ and $\boldsymbol{\Sigma}_k$ at the $(h+1)$th first CM-step of the ECM algorithm, as illustrated in Eqs. (10) and (11), can be obtained as follows.

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}_k^{*\prime}} Q\Big(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)}\Big) &= \frac{\partial}{\partial \boldsymbol{\beta}_k^{*\prime}} \sum_{i=1}^{I}\sum_{k=1}^{K} \hat{z}_{ik}^{(h)} Q_i\Big(\boldsymbol{\beta}_k^*, \boldsymbol{\Sigma}_k|\boldsymbol{\psi}^{(h)}\Big)= \\ &= \frac{\partial}{\partial \boldsymbol{\beta}_k^{*\prime}} \sum_{i=1}^{I}\sum_{k=1}^{K} \frac{\hat{z}_{ik}^{(h)}}{2} \\ &\quad \Big[-\ln|\boldsymbol{\Sigma}_k| - M(1-\hat{u}_{ik}^{(h)})\ln\hat{\eta}_k^{(h)} - \hat{w}_{ik}^{(h)}\delta^2_{\boldsymbol{\Sigma}_k}(\mathbf{y}_i, \boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}_k^*))\Big].\end{aligned} \tag{13}$$

Focusing on the squared Mahalanobis distance $\delta^2_{\boldsymbol{\Sigma}_k}(\mathbf{y}_i, \boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}_k^*))$ and using properties of trace and transpose, it follows that

$$\begin{aligned}\delta^2_{\boldsymbol{\Sigma}_k}(\mathbf{y}_i, \boldsymbol{\mu}_k(\mathbf{x}_i;\boldsymbol{\beta}_k^*)) &= \boldsymbol{y}_i'\boldsymbol{\Sigma}_k^{-1}\boldsymbol{y}_i - \boldsymbol{y}_i'\boldsymbol{\Sigma}_k^{-1}\tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_k^{*\prime}\tilde{\mathbf{x}}_i^*\boldsymbol{\Sigma}_k^{-1}\boldsymbol{y}_i + \boldsymbol{\beta}_k^{*\prime}\tilde{\mathbf{x}}_i^*\boldsymbol{\Sigma}_k^{-1}\tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^* \\ &= \boldsymbol{y}_i'\boldsymbol{\Sigma}_k^{-1}\boldsymbol{y}_i - 2tr(\boldsymbol{\beta}_k^{*\prime}\tilde{\mathbf{x}}_i^*\boldsymbol{\Sigma}_k^{-1}\boldsymbol{y}_i) + \boldsymbol{\beta}_k^{*\prime}\tilde{\mathbf{x}}_i^*\boldsymbol{\Sigma}_k^{-1}\tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^*.\end{aligned} \tag{14}$$

Deriving (14) respect to $\boldsymbol{\beta}_k^{*\prime}$ and then replacing the so obtained result in (13) leads to

$$\frac{\partial}{\partial \boldsymbol{\beta}_k^{*\prime}} Q\Big(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)}\Big) = \sum_{i=1}^{I} -\frac{\hat{z}_{ik}^{(h)}}{2}\hat{w}_{ik}^{(h)}\Big(-2\boldsymbol{y}_i'\boldsymbol{\Sigma}_k^{-1}\tilde{\mathbf{x}}_i^{*\prime} + 2\boldsymbol{\beta}_k^{*\prime}\tilde{\mathbf{x}}_i^*\boldsymbol{\Sigma}_k^{-1}\tilde{\mathbf{x}}_i^{*\prime}\Big)$$

$$= \sum_{i=1}^{I} \hat{z}_{ik}^{(h)} \hat{w}_{ik}^{(h)} \boldsymbol{y}_i' \boldsymbol{\Sigma}_k^{-1} \tilde{\mathbf{x}}_i^{*\prime} - \sum_{i=1}^{I} \hat{z}_{ik}^{(h)} \hat{w}_{ik}^{(h)} \boldsymbol{\beta}_k^{*\prime} \tilde{\mathbf{x}}_i^{*} \boldsymbol{\Sigma}_k^{-1} \tilde{\mathbf{x}}_i^{*\prime}. \tag{15}$$

Setting (15) equal to the null vector, solving the so obtained system with respect to $\boldsymbol{\beta}_k^{*\prime}$ and using properties of transpose results in the solution reported in Eq. (10). Finally,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} Q\left(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)}\right) &= \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \sum_{i=1}^{I} \sum_{k=1}^{K} \hat{z}_{ik}^{(h)} Q_i\left(\boldsymbol{\beta}_k^{*}, \boldsymbol{\Sigma}_k|\boldsymbol{\psi}^{(h)}\right) \\
&= \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \sum_{i=1}^{I} \sum_{k=1}^{K} \frac{\hat{z}_{ik}^{(h)}}{2}\Big[-\ln|\boldsymbol{\Sigma}_k| - M(1-u_{ik})\ln\eta_k + \\
&\quad - \hat{w}_{ik}^{(h)}\left(\boldsymbol{y}_i - \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h+1)}\right)' \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{y}_i - \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h+1)}\right)\Big] \\
&= \frac{\partial}{\partial \boldsymbol{\Sigma}_k^{-1}} \sum_{i=1}^{I} \sum_{k=1}^{K} \frac{\hat{z}_{ik}^{(h)}}{2}\Big[\ln|\boldsymbol{\Sigma}_k^{-1}| - M(1-u_{ik})\ln\eta_k + \\
&\quad - \hat{w}_{ik}^{(h)} tr\left(\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_i - \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h+1)})(\boldsymbol{y}_i - \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h+1)})'\right)\Big]. \\
&= \sum_{i=1}^{I} \frac{\hat{z}_{ik}^{(h)}}{2}\Big[\boldsymbol{\Sigma}_k - \hat{w}_{ik}^{(h)}\left(\boldsymbol{y}_i - \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h+1)})(\boldsymbol{y}_i - \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h+1)}\right)'\Big],
\end{aligned} \tag{16}$$

where the second and third equalities are obtained using properties of trace and transpose and differentiation rules of functions of matrices. Setting (16) equal to the null matrix and solving the resulting system with respect to $\boldsymbol{\Sigma}_k$ gives the update in Eq. (11).

References

Aitken AC (1926) A series formula for the roots of algebraic and transcendental equations. Proc R Soc Edinb 45(1):14–22

Aitkin M, Wilson TG (1980) Mixture models, outliers, and the EM algorithm. Technometrics 22(3):325–331

Andrews JL, McNicholas PD (2011) Extending mixtures of multivariate t-factor analyzers. Stat Comput 21(3):361–373

Baek J, McLachlan GJ (2011) Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. Bioinformatics 27(9):1269–1276

Bai X, Yao W, Boyer JE (2012) Robust fitting of mixture regression models. Comput Stat Data Anal 56(7):2347–2359

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(2):337–355

Bartolucci F, Scaccia L (2005) The use of mixtures for dealing with non-normal regression errors. Comput Stat Data Anal 48(4):821–834

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Anal Mach Intell 22(7):719–725

Biernacki C, Celeux G, Govaert G (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput Stat Data Anal 41(3–4):561–575

Cadavez VAP, Hennningsen A (2012) The use of seemingly unrelated regression (SUR) to predict the carcass composition of lambs. Meat Sci 92(4):548–553

Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recognit 28(5):781–793

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(1):15–37

Dang UJ, Punzo A, McNicholas PD, Ingrassia S, Browne RP (2017) Multivariate response and parsimony for Gaussian cluster-weighted models. J Classif 34(1):4–34

De Sarbo WS, Cron WL (1988) A maximum likelihood methodology for clusterwise linear regression. J Classif 5(2):249–282

De Veaux RD (1989) Mixtures of linear regressions. Comput Stat Data Anal 8(3):227–245

Dempster A, Laird N, Rubin D (1977) Maximum likelihood for incomplete data via the EM algorithm. J R Stat Soc 39(1):1–38

Depraetere N, Vandebroek M (2014) Order selection in finite mixtures of linear regressions. Stat Pap 55(3):871–911

Ding C (2006) Using regression mixture analysis in educational research. Pract Assess Res Eval 11(1):1–11

Disegna M, Osti L (2016) Tourists' expenditure behaviour: the influence of satisfaction and the dependence of spending categories. Tour Econ 22(1):5–30

Dyer WJ, Pleck J, McBride B (2012) Using mixture regression to identify varying effects: a demonstration with paternal incarceration. J Marriage Fam 74(5):1129–1148

Elhenawy M, Rakha H, Chen H (2017) An automatic traffic congestion identification algorithm based on mixture of linear regressions. In: Helfert M, Klein C, Donnellan B, Gusikhin O (eds) Smart cities, green technologies, and intelligent transport systems. Springer, Cham, pp 242–256

Fair RC, Jaffe DM (1972) Methods of estimation for markets in disequilibrium. Econometrica 40:497–514

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

Galimberti G, Scardovi E, Soffritti G (2016) Using mixtures in seemingly unrelated linear regression models with non-normal errors. Stat Comput 26(5):1025–1038

Galimberti G, Soffritti G (2020) Seemingly unrelated clusterwise linear regression. Adv Data Anal Classif 14(2):235–260

Giles S, Hampton P (1984) Regional production relationships during the industrialization of New Zealand, 1935–1948. Reg Sci 24(4):519–532

Heidari S, Keshavarzi 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 Surveill Syst 5(4):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

Hosmer DW (1974) Maximum likelihood estimates of the parameters of a mixture of two regression lines. Commun Stat Theory Methods 3(10):995–1006

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218

Jones PN, McLachlan GJ (1992) Fitting finite mixture models in a regression context. Aust J Stat 34(2):233–240

Kamakura W (1988) A least squares procedure for benefit segmentation with conjoint experiments. J Mark Res 25(2):157–167

Karlis D, Xekalaki E (2003) Choosing initial values for the EM algorithm for finite mixtures. Comput Stat Data Anal 41(3–4):577–590

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 2012:821643

Keshavarzi S, Ayatollahi SMT, Zare N, Sharif F (2013) Quality of life of childbearing age women and its associated factors: an application of seemingly unrelated regression (SUR) models. Qual Life Res 22(6):1255–1263

Kibria BMG, Haq MS (1999) The multivariate linear model with multivariate t and intra-class covariance structure. Stat Pap 40(3):263–276

Lachos VH, Angolini T, Abanto-Valle CA (2011) On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. Stat Pap 52(3):567–590

Lange KL, Little RJA, Taylor JMG (1989) Robust statistical modeling using the t distribution. J Am Stat Assoc 84(408):881–896

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

Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, Chichester

Mazza A, Punzo A (2020) Mixtures of multivariate contaminated normal regression models. Stat Pap 61(2):787–822

McDonald SE, Shin S, Corona R et al (2016) Children exposed to intimate partner violence: identifying differential effects of family environment on children's trauma and psychopathology symptoms through regression mixture models. Child Abus Negl 58:1–11

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(5):1175–1181

Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80(2):267–278

Park T (1993) Equivalence of maximum likelihood estimation and iterative two-stage estimation for seemingly unrelated regression models. Commun Stat Theory Methods 22(8):2285–2296

Punzo A, McNicholas PD (2017) Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. J Classif 34(2):249–293

Qin LX, Self SG (2006) The clustering of regression models method with applications in gene expression data. Biometrics 62(2):526–533

Quandt RE, Ramsey JB (1978) Estimating mixtures of normal distributions and switching regressions. J Am Stat Assoc 73(364):730–738

R Core Team (2021) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Ritter G (2015) Robust cluster analysis and variable selection. Chapman and Hall, Boca Raton

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

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

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

Soffritti G, Galimberti G (2011) Multivariate linear regression with non-normal errors: a solution based on mixture models. Stat Comput 21(4):523–536

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

Tashman A, Frey RJ (2009) Modeling risk in arbitrage strategies using finite mixtures. Quant Finance 9(5):495–503

Tukey JW (1960) A survey of sampling from contaminated distributions. In: Olkin I (ed) Contributions to probability and statistics: essays in honor of Harold Hotelling, Stanford studies in mathematics and statistics. Stanford University Press, Redwood City, pp 448–485

Turner TR (2000) Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. Appl Stat 49(3):371–384

Van Horn ML, Jaki T, Masyn K et al (2015) Evaluating differential effects using regression interactions and regression mixture models. Educ Psychol Meas 75(4):677–714

Wedel M (2002) Concomitant variables in finite mixture models. Stat Neerl 56(3):362–375

White EN, Hewings GJD (1982) Space-time employment modelling: some results using seemingly unrelated regression estimators. J Reg Sci 22(3):283–302

Yao W, Wei Y, Yu C (2014) Robust mixture regression using the t-distribution. Comput Stat Data Anal 71:116–127

Zellner A (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J Am Stat Assoc 57(298):348–368

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.