Parsimonious Seemingly Unrelated Contaminated Normal Cluster-Weighted Models

Journal of Classification

Parsimonious seemingly unrelated contaminated normal cluster-weighted models

Gabriele Perrone · Gabriele Soffritti

Accepted: 05 December 2023 / Published (online): 08 January 2024

Abstract Normal cluster-weighted models constitute a modern approach to linear regression which simultaneously perform model-based cluster analysis and multivariate linear regression analysis with random quantitative regressors. Robustified models have been recently developed, based on the use of the contaminated normal distribution, which can manage the presence of mildly atypical observations. A more flexible class of contaminated normal linear cluster-weighted models is specified here, in which the researcher is free to use a different vector of regressors for each response. The novel class also includes parsimonious models, where parsimony is attained by imposing suitable constraints on the component-covariance matrices of either the responses or the regressors. Identifiability conditions are illustrated and discussed. An expectation-conditional maximisation algorithm is provided for the maximum likelihood estimation of the model parameters. The effectiveness and usefulness of the proposed models are shown through the analysis of simulated and real datasets.

Keywords Contaminated normal distribution · ECM algorithm · Mixture model · Model-based cluster analysis · Parsimonious model · Seemingly unrelated regression

Mathematics Subject Classification (2020) 62J05 · 62H12 · 62F12

G. Perrone
Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy.
E-mail: gabriele.perrone4@unibo.it
ORCID: https://orcid.org/0000-0001-5930-6205

G. Soffritti (corresponding author)
Department of Statistical Sciences, University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy.
Tel.: +39-51-2098193. Fax: +39-51-2086242
E-mail: gabriele.soffritti@unibo.it
ORCID: https://orcid.org/0000-0002-7575-892X

1 Introduction

In an era of rapid technological change, vast amounts of complex data are being generated in many fields. Eliciting information from these kinds of datasets represents a crucial challenge faced by scientists and researchers. In order to achieve this aim, advanced and flexible tools and methods are required. From a statistical point of view, problems in learning from data have been classified as either unsupervised or supervised (Hastie et al., 2009). This latter class typically involves the task of modelling the dependence of M responses $\mathbf{Y} = (Y_1, \ldots, Y_M)'$ on P given predictors $\mathbf{X} = (X_1, \ldots, X_P)'$ through multivariate regression techniques. In this setting, several issues can make the data analysis more complex.

The novel methods introduced in this paper have been devised so as to be specifically employed when all the variables in $\mathbf{Y}$ as well as in $\mathbf{X}$ are continuous and the following situations arise.

(I) Data contain measurements obtained without actively controlling or manipulating any of the variables to be analysed. This is typically true in several disciplines (i.e., sociology, economics, business, ecology and geology). For the analysis of such data, regression models should treat both $\mathbf{X}$ and $\mathbf{Y}$ as random vectors. Thus, the joint distribution of $(\mathbf{X}, \mathbf{Y})$ in a given population of an investigation, say G, is generally modelled using a probability density function (p.d.f.) $f(\mathbf{x}, \mathbf{y})$ specified to take account of the different role played by the responses and predictors in the analysis; that is: $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x})f(\mathbf{y}|\mathbf{x})$.

(II) The population G is heterogeneous, as it is composed of K disjoint and homogeneous sub-populations, say $G_1, \ldots, G_k, \ldots, G_K$, and the sample data available for the estimation of the regression model are $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I)\}$. This means that the information about the specific sub-population each sample observation belongs to is missing. Furthermore, this source of unobserved heterogeneity in the data affects the distribution of $(\mathbf{X}, \mathbf{Y})$.

(III) The data $\mathcal{S}$ are contaminated by the presence of mildly atypical observations; that is, observations that in some way deviate from the general pattern of the data (Maronna et al., 2006). More specifically, a mild outlier can be considered as an observation which is sampled from a model different or even far from the overall assumed model (see Ritter, 2015, p. 79). In this situation, a suitable model for the data $\mathcal{S}$ should be specified by the researcher so as to be flexible enough to accommodate all the sample observations, including the atypical ones. This typically happens by adopting a model able to identify the atypical observations and by employing an estimation method which automatically down-weights such observations. In a regression framework, an observation $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{S}$ can be an outlier either in the hyperplane defined by the regression model for $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ (vertical or regression outlier) or in the predictor space (leverage point) (see, e.g., Rousseeuw and Leroy, 2005). When $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{S}$ is both a regression outlier and a leverage point it will have a large influence on the estimation

of the regression coefficients; thus, it is considered a bad leverage point (Rousseeuw and Leroy, 2005).

(IV) The multivariate regression model specified by the researcher is composed of a system of M regression equations (one equation for each response) with equation-dependent vectors of predictors (i.e., vectors which do not necessarily contain the same predictors for all the responses). This means that certain regressors contained in $\mathbf{X}$ are absent from certain regression equations. This situation is not unusual in economics or social sciences, where different predictors may be expected to be relevant in the prediction of the M responses according to some general theory or prior information about the phenomenon. Furthermore, the M responses contained in $\mathbf{Y}$ are correlated. This latter feature is typically observed with multivariate longitudinal data, time-series data or repeated measures.

An approach able to properly model the distribution of $(\mathbf{X}, \mathbf{Y})$ in the presence of the unobserved source of heterogeneity illustrated in situation (II) relies on the cluster-weighted (CW) models (Gershenfeld, 1997). In this approach, the missing information about the memberships of the K sub-populations is modelled using a mixture of K different p.d.f.'s, where each one of these functions is specified by taking account of the different role played by $\mathbf{X}$ and $\mathbf{Y}$. This leads to the following mixture model for the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$:

$$f(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}|G_k) f(\mathbf{y}|\mathbf{x}, G_k), \ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{P+M}, \tag{1}$$

where $\pi_1, \ldots, \pi_K$ are positive mixing weights summing to one and representing the prior probabilities of the K sub-populations (i.e., $\mathbb{P}(G_k) = \pi_k$), $f(\mathbf{x}|G_k)$ is the conditional p.d.f. of $\mathbf{X}$ given G_k, and $f(\mathbf{y}|\mathbf{x}, G_k)$ is the conditional p.d.f. of $\mathbf{Y}$ given $\mathbf{x}$ and G_k. An eminent member of the class of CW models for real-valued responses and predictors is the normal CW (NCW hereafter) model (Ingrassia et al., 2012; Dang et al., 2017). In this model, normal distributions are employed for the p.d.f. of both $\mathbf{X}|G_k$ and $\mathbf{Y}|(\mathbf{X} = \mathbf{x}, G_k)$, for $k = 1, \ldots, K$. Thus, (1) becomes

$$f(\mathbf{x}, \mathbf{y}; \boldsymbol{\vartheta}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \phi(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta}_k' \mathbf{x}^*, \boldsymbol{\Xi}_k), \ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{P+M}, \tag{2}$$

where $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the p.d.f. of a normal random vector with expected value $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}_k' \mathbf{x}^* = \mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}, G_k)$, for $k = 1, \ldots, K$, $\mathbf{x}^* = (1, \mathbf{x}')'$, $\boldsymbol{\beta}_k \in \mathbb{R}^{(1+P)\times M}$ is a matrix of intercepts and regression coefficients, and $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K)$ is the vector of the model parameters, with $\boldsymbol{\vartheta}_k = (\pi_k, \boldsymbol{\vartheta}_{k\mathbf{x}}, \boldsymbol{\vartheta}_{k\mathbf{y}})$, $\boldsymbol{\vartheta}_{k\mathbf{x}} = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\vartheta}_{k\mathbf{y}} = (\boldsymbol{\beta}_k, \boldsymbol{\Xi}_k)$, for $k = 1, \ldots, K$. CW models which allow $\mathbf{X}|G_k$ and $\mathbf{Y}|(\mathbf{X} = \mathbf{x}, G_k)$ to be modelled using skewed distributions have been recently introduced (Gallaugher et al., 2022). A by-product of a regression analysis based on a CW model is a set of estimated posterior probabilities that each sample observation comes from the K sub-populations. Thus, a clustering of the I sample observations that

compose $\mathcal{S}$ can also be obtained, based on a rule that assigns an observation to the sub-population from which it has the highest posterior probability of coming. As a result, CW models allow one to simultaneously perform multivariate regression and cluster analysis.

Mildly atypical observations in the data mentioned in situation (III) cause departures from the normal distribution. A way to manage these departures is to resort to heavy-tailed models, such as the t distribution or the contaminated normal distribution (see, e.g., Tukey, 1960; Aitkin and Wilson, 1980). This latter distribution is defined as a mixture of two normal distributions having the same expected mean values but different variances-covariances; the normal distribution having the smallest mixing weight also has inflated variances-covariances and is employed to represent the mildly atypical observations. Multivariate regression models robust against the presence of such observations and also suitable for situations (I) and (II) have been obtained from (1) by specifying either a t distribution (Ingrassia et al., 2012, 2014; Subedi et al., 2015) or a contaminated normal distribution (Punzo and McNicholas, 2017) for both $\mathbf{X}|G_k$ and $\mathbf{Y}|(\mathbf{X}=\mathbf{x}, G_k)$, $k=1,\ldots,K$. The CW models proposed by Punzo and McNicholas (2017) are also called contaminated normal cluster-weighted (CNCW) models. By relying on such models, it is possible to produce a simultaneous clustering of the sample observations and the detection of both mild outliers and leverage points in a multivariate regression context with random regressors. A limitation of the CNCW models is that the same vector of predictors has to be employed for all the M responses. Furthermore, these models can only manage mild outliers. When $\mathcal{S}$ contain gross outliers, that are unpredictable observations which generally appear when data are automatically extracted from complex objects by a machine and, thus, cannot be modelled by any distribution (see Ritter, 2015, p. 80), the common practice is to suppress them using an automatic method, such as trimming (Cuesta-Albertos et al., 1997). An approach to robust estimation of a cluster-weighted model via trimming has been recently introduced (García-Escudero et al., 2017). Some developments are reported in Cappozzo et al. (2021, 2023). Such an approach embeds an implementation of high-breakdown robust methods (see, e.g., Rousseeuw and Van Driessen, 1999), thereby producing an estimator of the model parameter with desirable robustness properties (see, e.g., Hennig, 2004; Ruwet et al., 2013; Farcomeni and Punzo, 2020).

Multivariate correlated responses and the systems of regression equations with equation-dependent vectors of predictors illustrated in situation (IV) can be managed by resorting to the so-called seemingly unrelated regression approach (see, e.g., Srivastava and Giles, 1987; Park, 1993). This approach has been embedded into the specification of a class of NCW models by Diani et al. (2022), thus leading to seemingly unrelated normal cluster-weighted (SuNCW) models. Thus, the methods based on these latter models are suitable for jointly managing the situations (I), (II) and (IV). However, they are not insensitive to the possible presence of mild outliers and leverage points in the K sub-populations.

Based on all these considerations, a novel class of multivariate seemingly unrelated contaminated normal cluster-weighted (SuCNCW) models for the analysis of data containing mildly atypical observations either in the distribution of $\mathbf{X}|G_k$ or in the distribution of $\mathbf{Y}|(\mathbf{X}=\mathbf{x}, G_k)$, $k=1,\ldots,K$, are introduced. With these novel models, the four situations mentioned above are jointly managed when predicting the responses in a multivariate linear regression framework with random predictors. In particular, SuCNCW models can be considered a more flexible version of the CNCW models described in Punzo and McNicholas (2017), as the linear terms in the M regression equations of a SuCNCW model are defined so that a different vector of regressors can be employed for each dependent variable. In order to keep the total number of parameters as low as possible, the novel class also includes parsimonious SuCNCW models; parsimony is attained by parameterising the covariance matrices of both $\mathbf{X}|G_k$ and $\mathbf{Y}|(\mathbf{X}=\mathbf{x}, G_k)$, for $k=1,\ldots,K$, with their eigen-decomposition, and by imposing constraints on parts of the elements of this decomposition (see, e.g., Celeux and Govaert, 1995). This leads to a flexible approach for the analysis of linear dependencies in multivariate data.

In summary, this paper provides the following key contributions: new parsimonious SuCNCW models able to jointly manage the situations (I)–(IV) are introduced (see Section 2.1); the relationships between the proposed models and other mixture regression models are described (Section 2.2); conditions for the identifiability of the SuCNCW models are illustrated (Section 2.3); maximum likelihood (ML) estimation via an expectation-conditional maximisation (ECM) algorithm (Meng and Rubin, 1993) is detailed (Section 2.4); notes on the robustness of parameter estimation based on the proposed methodology are given (Section 2.5); strategies for the initialisation and convergence of the ECM algorithm as well as for model selection are presented (Section 2.6); information about how to perform variable selection for modelling $\mathbb{E}(\mathbf{Y}|\mathbf{X}=\mathbf{x}, G_k)$, $k=1,\ldots,K$, is reported (Section 2.7); the effectiveness of the new models in comparison with NCW, CNCW and SuNCW models is investigated through simulated datasets (Section 3); a study aiming at evaluating the link between tourism flows and attendance at museums and monuments in two Italian regions is carried out (Section 4). Some experimental results are reported in a separate document as supplementary material. Concluding remarks and ideas for future research are illustrated in Section 5.

2 Seemingly unrelated contaminated normal cluster-weighted analysis

2.1 Seemingly unrelated contaminated normal cluster-weighted models

The new class of SuCNCW models is introduced starting from the CNCW models illustrated by Punzo and McNicholas (2017). These latter models can be obtained by replacing the normal distributions for $\mathbf{X}|G_k$ and $\mathbf{Y}|(\mathbf{X}=\mathbf{x}, G_k)$

in (1) with the following contaminated normal distributions, respectively:

$$h(\mathbf{x};\boldsymbol{\theta}_{k\mathbf{x}}) = \alpha_k\phi\left(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right) + (1-\alpha_k)\phi\left(\mathbf{x};\boldsymbol{\mu}_k,\eta_k\boldsymbol{\Sigma}_k\right),\ \mathbf{x}\in\mathbb{R}^P,$$
$$h(\mathbf{y}|\mathbf{x};\tilde{\boldsymbol{\vartheta}}_{k\mathbf{y}}) = \tau_k\phi\left(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}_k,\boldsymbol{\Xi}_k\right) + (1-\tau_k)\phi\left(\mathbf{y}|\mathbf{x};\boldsymbol{\beta}_k,\lambda_k\boldsymbol{\Xi}_k\right),\ \mathbf{y}\in\mathbb{R}^M,$$

where $\boldsymbol{\theta}_{k\mathbf{x}} = (\boldsymbol{\vartheta}_{k\mathbf{x}},\alpha_k,\eta_k)$, $\tilde{\boldsymbol{\vartheta}}_{k\mathbf{y}} = (\boldsymbol{\vartheta}_{k\mathbf{y}},\tau_k,\lambda_k)$. Parameters $\alpha_k \in (0,1)$ and $\tau_k \in (0,1)$ represent the weights of the typical observations in the predictor space and the regression hyperplane, respectively, within the sub-population G_k. In robust statistics it is generally assumed that at least half of the observations are typical; thus, it is possible to require that $\alpha_k \in [0.5,1)$ and $\tau_k \in [0.5,1)$. Parameters $\eta_k > 1$ and $\lambda_k > 1$ determine the degree of the contamination in the normal distributions for $\mathbf{X}|G_k$ and $\mathbf{Y}|(\mathbf{X}=\mathbf{x},G_k)$; namely, η_k and λ_k control the increase in variability due to the presence of the leverage points and the mild outliers, respectively, within G_k. Thus, the random vector $(\mathbf{X},\mathbf{Y})$ follows a CNCW model of order K if its p.d.f. has the form

$$f(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = \sum_{k=1}^{K}\pi_k h(\mathbf{x};\boldsymbol{\theta}_{k\mathbf{x}})h(\mathbf{y}|\mathbf{x};\tilde{\boldsymbol{\vartheta}}_{k\mathbf{y}}),\ (\mathbf{x},\mathbf{y})\in\mathbb{R}^{P+M}, \tag{3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)$, with $\boldsymbol{\theta}_k = (\pi_k,\boldsymbol{\theta}_{k\mathbf{x}},\tilde{\boldsymbol{\vartheta}}_{k\mathbf{y}})$.

If only P_m of the P covariates ($P_m \leq P$) are known or assumed to be relevant for the prediction of Y_m ($m = 1,\ldots,M$), the linear predictor $\boldsymbol{\beta}_k'\mathbf{x}^*$ employed for modelling the conditional expected value $\mathbb{E}(\mathbf{Y}|\mathbf{X}=\mathbf{x},G_k)$ in (3) should be modified accordingly. To this end, let $\mathbf{X}_m = (\tilde{X}_1,\ldots,\tilde{X}_{P_m})'$ be the vector composed of such P_m covariates, and let $\boldsymbol{\beta}_{km} = (\beta_{km1},\ldots,\beta_{kmP_m})'$ be the vector of the P_m regression coefficients capturing the linear effect of $\mathbf{X}_m$ on Y_m in the kth sub-population. Furthermore, let $\mathbf{X}_m^* = (1,\mathbf{X}_m')'$ and $\boldsymbol{\beta}_{km}^* = (\beta_{km0},\boldsymbol{\beta}_{km}')'$. Then, $\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_{k1}^{*\prime},\ldots,\boldsymbol{\beta}_{km}^{*\prime},\ldots,\boldsymbol{\beta}_{kM}^{*\prime})'$ represents the (P^*+M)-dimensional vector containing all the linear effects of the relevant predictors on the M responses in the kth sub-population, where $P^* = \sum_{m=1}^{M}P_m$. Finally, the $(P^*+M)\times M$ design matrix is defined as follows:

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

where $\mathbf{0}_{P_m+1}$ represents the (P_m+1)-dimensional null vector. Using this additional notation, it is possible to obtain the following definition for the conditional expected value of $\mathbf{Y}|\mathbf{X}=\mathbf{x}$ in the kth sub-population:

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}=\mathbf{x},G_k) = \tilde{\mathbf{x}}^{*\prime}\boldsymbol{\beta}_k^* = \begin{bmatrix} \mathbf{x}_1^{*\prime}\ \boldsymbol{\beta}_{k1}^* \\ \vdots \\ \mathbf{x}_m^{*\prime}\ \boldsymbol{\beta}_{km}^* \\ \vdots \\ \mathbf{x}_M^{*\prime}\ \boldsymbol{\beta}_{kM}^* \end{bmatrix}, \tag{4}$$

where $\tilde{\mathbf{x}}^*$ is the realisation of the design matrix $\tilde{\mathbf{X}}^*$ obtained when $\mathbf{X} = \mathbf{x}$. The vector defined in (4) has length M; its mth element is given by a linear combination of the P_m regressors selected by the researcher for the prediction of Y_m whose coefficients are given by the elements of the vector $\boldsymbol{\beta}^*_{km}$. Thus, inserting the expression given in (4) into (3) leads to the new SuCNCW model. More formally, the random vector $(\mathbf{X}, \mathbf{Y})$ follows a SuCNCW model of order K if its p.d.f. has the form

$$f(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi}) = \sum_{k=1}^{K} \pi_k h\left(\mathbf{x}; \boldsymbol{\theta}_{k\mathbf{x}}\right) h\left(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{k\mathbf{y}}\right), \ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{P+M}, \tag{5}$$

where the vector of the model parameters is $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K)$, with $\boldsymbol{\psi}_k = (\pi_k, \boldsymbol{\theta}_{k\mathbf{x}}, \boldsymbol{\theta}_{k\mathbf{y}})$ and $\boldsymbol{\theta}_{k\mathbf{y}} = (\boldsymbol{\beta}^*_k, \boldsymbol{\Xi}_k, \tau_k, \lambda_k)$. From the comparison between $\boldsymbol{\psi}$ and the vector $\boldsymbol{\theta}$ with the parameters of (3) it is clear that a CNCW model of order K and a SuCNCW model of order K have the same parameters except for the K matrices containing the intercepts and regression coefficients. In (5) it is assumed that $\pi_k > 0$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. As far as the parameters α_k, η_k, τ_k and λ_k are concerned, the requirements coincide with those previously illustrated for (3). The number of free parameters in (5) is $n_{\boldsymbol{\psi}} = 5K - 1 + K(P + P^* + M) + K[\frac{P(P+1)}{2} + \frac{M(M+1)}{2}]$.

The typical properties of the CNCW model defined according to (3) (i.e., the ability to determine the membership of an observation $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{S}$ to a specific sub-population and to establish whether $(\mathbf{x}_i, \mathbf{y}_i)$ is an outlier in the regression hyperplane and/or in the predictor space in that sub-population) are inherited by the SuCNCW model given in (5). In addition, this latter model offers a more parsimonious specification of the linear term to be employed in the prediction of $\mathbf{Y}$ whenever it is known or assumed that certain covariates are not relevant for this task. Model (5) can also be considered as a CNCW model in which some regression coefficients are constrained to be a priori equal to zero. To the best of the authors' knowledge, including such constraints in the specification of a multivariate CNCW model has not been addressed yet.

In practical applications in which the analysis involves either many responses or many predictors, using (5) to perform the analysis can become unfeasible. This is a consequence of the fact that the number of free parameters $n_{\boldsymbol{\psi}}$ of a SuCNCW model increases quadratically both with M and with P. A way to manage this issue is to resort to the approach illustrated in Celeux and Govaert (1995). With this approach, a reparameterisation of (5) is obtained, in which the covariance matrices $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Xi}_k$, for $k = 1, \ldots, K$, are expressed in terms of their eigenvalues and eigenvectors; furthermore, the introduction of suitable constraints on such quantities allows to obtain parsimonious SuCNCW models. More specifically, let $\mathbf{A}_k$ be the diagonal matrix containing the eigenvalues of $\boldsymbol{\Sigma}_k$, normalised in such a way that $|\mathbf{A}_k| = 1$; let $\mathbf{D}_k$ be the matrix with the corresponding eigenvectors, and $\xi_k = |\boldsymbol{\Sigma}_k|^{1/D}$. By exploiting the eigen-decomposition $\boldsymbol{\Sigma}_k = \xi_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k$, variances and covariances in $\boldsymbol{\Sigma}_k$ can be obtained from ξ_k, $\mathbf{A}_k$ and $\mathbf{D}_k$, which control the volume, shape and orientation of the kth cluster of observations with respect to the predictors.

Table 1 Parameterisations of the component-covariance matrices.

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

Constraining ξ_k, $\mathbf{A}_k$ and $\mathbf{D}_k$ on this decomposition in (5) with $K > 1$ will lead to 14 different covariance structures for the predictors. Additional information about these parameterisations are reported in Table 1. When ξ_k, $\mathbf{A}_k$ and $\mathbf{D}_k$ are all variable across the K clusters (VVV acronym in Table 1), the resulting covariance structures of the predictors will be fully unconstrained. From the simultaneous application of the same decomposition to the covariance matrices $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Xi}_k$, for $k = 1, \ldots, K$, 196 differentially parameterised SuCNCW models of order K can be obtained, for any given $K > 1$. As far as the specification of models of order $K = 1$ is concerned, the possible covariance structures for both responses and covariates are: diagonal with different entries (VI), diagonal with the same entries (EI) and fully unconstrained (VV). Thus, when $K = 1$, only nine differentially parameterised models can be specified.

2.2 Comparisons with other mixture regression models

When specific conditions are met, some normal CW models can be obtained from (5).

- If $M > 1$, $P_m = P$ and $\mathbf{X}_m = \mathbf{X}$ $\forall m$ (the same vector of covariates is employed in the prediction of the M responses), the realisation of the design matrix $\tilde{\mathbf{X}}^*$ is equal to $\tilde{\mathbf{x}}^* = \mathbf{I}_M \otimes \mathbf{x}^*$, with $\mathbf{I}_M$ being the identity matrix of order M and $\otimes$ denoting the Kronecker product operator (see, e.g., Magnus and Neudecker, 1988). Thus, (4) becomes

$$\mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}, G_k) = \left(\mathbf{I}_M \otimes \mathbf{x}^*\right)' \boldsymbol{\beta}_k^* = \mathbf{B}_k'\mathbf{x}^*, \; k = 1, \ldots, K, \tag{6}$$

  where $\mathbf{B}_k = [\boldsymbol{\beta}_{k1}^* \cdots \boldsymbol{\beta}_{km}^* \cdots \boldsymbol{\beta}_{kM}^*]$. Thus, (5) reduces to the CNCW model (Punzo and McNicholas, 2017).
- If $M > 1$, $\alpha_k \to 1$, $\eta_k \to 1$, $\tau_k \to 1$ and $\lambda_k \to 1$ $\forall k$ (there is no contamination in the data), the model resulting from (5) coincides with the SuNCW model described in Diani et al. (2022).

– If $M > 1$, $P_m = P$ and $\mathbf{X}_m = \mathbf{X}$ $\forall m$, $\alpha_k \to 1$, $\eta_k \to 1$, $\tau_k \to 1$ and $\lambda_k \to 1$ $\forall k$ (there is no contamination in the data and the same vector of covariates is employed in the prediction of the M responses), (5) leads to the multivariate NCW model given in (2) (Dang et al., 2017).

As illustrated in Section 2.1, SuCNCW models assume that $\mathbf{X}|G_k$ follows a contaminated normal distribution with parameters $\boldsymbol{\theta}_{k\mathbf{x}} = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k, \eta_k)$, for $k = 1, \ldots, K$. However, for some datasets it may happen that the probability a point $(\mathbf{x}, \mathbf{y})$ belongs to one of the K distributions of the mixture (5) is the same for all covariate values $\mathbf{x}$. In that case, the assignment of the data points to the sub-populations is independent of the covariates. This condition is known as assignment independence (see, e.g., Hennig, 2000). This implies that the p.d.f of $\mathbf{X}|G_k$ does not depend on G_k, and $h(\mathbf{x}; \boldsymbol{\theta}_{k\mathbf{x}}) = h(\mathbf{x}; \boldsymbol{\theta})$ for every $k = 1, \ldots, K$, where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \eta)$. Thus, under the assignment independence condition, (5) becomes

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

where

$$f(\mathbf{y}|\mathbf{x}; \tilde{\boldsymbol{\psi}}) = \sum_{k=1}^{K} \pi_k h\left(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}_{k\mathbf{y}}\right), \ \mathbf{y} \in \mathbb{R}^{M}, \tag{7}$$

with $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\psi}}_1, \ldots, \tilde{\boldsymbol{\psi}}_K)$, $\tilde{\boldsymbol{\psi}}_k = (\pi_k, \boldsymbol{\theta}_{k\mathbf{y}})$, is the seemingly unrelated contaminated normal clusterwise regression model described in Perrone and Soffritti (2023). As a consequence, when in (5) the following conditions hold true: $\boldsymbol{\mu}_k = \boldsymbol{\mu}$, $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$, $\alpha_k = \alpha$ and $\eta_k = \eta$ for $k = 1, \ldots, K$, then the task of extracting the information about both the K disjoint sub-populations that compose the population G and the distinction between typical observations and mild outliers in the regression hyperplane within each sub-population can be equivalently carried out using either the conditional p.d.f. $f(\mathbf{y}|\mathbf{x}; \tilde{\boldsymbol{\psi}})$ through seemingly unrelated contaminated normal clusterwise models or the joint p.d.f. $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi})$ through SuCNCW models.

2.3 Identifiability

Since identifiability represents a regularity condition for the asymptotic theory to hold for the ML estimator, a discussion about identifiability of (5) is provided here. In particular, this discussion focuses on the class of models $\mathfrak{F} = \{\mathfrak{F}_K, K = 1, \ldots, K_{\max}\}$, with $\mathfrak{F}_K = \{f(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$, where $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\psi})$ is the p.d.f. of $(\mathbf{X}, \mathbf{Y})$ under the SuCNCW model of order K defined in (5) and $K_{\max}$ denotes the maximum order specified by the researcher for that model. This class is identifiable if, for any two members M, $\tilde{M} \in \mathfrak{F}$ with parameters $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k, \ldots, \boldsymbol{\psi}_K)$ and $\tilde{\boldsymbol{\psi}} = (\tilde{\boldsymbol{\psi}}_1, \ldots, \tilde{\boldsymbol{\psi}}_s, \ldots, \tilde{\boldsymbol{\psi}}_{\tilde{K}})$, respectively, the

equality

$$\sum_{k=1}^{K} \pi_k h\left(\mathbf{x};\boldsymbol{\theta}_{k\mathbf{x}}\right) h\left(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{k\mathbf{y}}\right) = \sum_{s=1}^{\tilde{K}} \tilde{\pi}_s h\left(\mathbf{x};\tilde{\boldsymbol{\theta}}_{s\mathbf{x}}\right) h\left(\mathbf{y}|\mathbf{x};\tilde{\boldsymbol{\theta}}_{s\mathbf{y}}\right) \;\forall\, (\mathbf{x},\mathbf{y}) \in \mathbb{R}^{P+M}$$

implies that $K = \tilde{K}$ and for each $k \in \{1,\ldots,K\}$ there exists $s \in \{1,\ldots,K\}$ such that $\pi_k = \tilde{\pi}_s$, $\boldsymbol{\theta}_{k\mathbf{x}} = \tilde{\boldsymbol{\theta}}_{s\mathbf{x}}$ and $\boldsymbol{\theta}_{k\mathbf{y}} = \tilde{\boldsymbol{\theta}}_{s\mathbf{y}}$.

The model class $\mathfrak{F}$ is affected by several sources of non-identifiability. As any finite mixture model, (5) is invariant under relabelling the K distributions of the mixture (label switching). Another source is represented by potential overfitting associated with empty components or equal components of the mixture (see, e.g., Frühwirth-Schnatter, 2006, for further details). In order to prevent such sources of non-identifiability for $\mathfrak{F}$, some constraints have been imposed on the parameter space $\boldsymbol{\Psi}$. They have been obtained by suitably modifying the constraints described in Punzo and McNicholas (2017) for ensuring the identifiability of CNCW models. Namely, for (5), it is required that $\pi_k > 0\ \forall\ k$ and $(\boldsymbol{\beta}_k^*, \boldsymbol{\Xi}_k) \neq (\boldsymbol{\beta}_h^*, \boldsymbol{\Xi}_h)\ \forall\ k \neq h$. Thanks to these constraints, the two sources of non-identifiability due to empty components and equal components can be avoided. Thus, in order to ensure identifiability, the following restricted class of SuCNCW models is introduced:

$$\bar{\mathfrak{F}} = \Big\{ f(\mathbf{x},\mathbf{y};\bar{\boldsymbol{\psi}}) : f(\mathbf{x},\mathbf{y};\bar{\boldsymbol{\psi}}) = \sum_{k=1}^{K} \pi_k h\left(\mathbf{x};\boldsymbol{\theta}_{k\mathbf{x}}\right) h\left(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}_{k\mathbf{y}}\right), \\ (\mathbf{x},\mathbf{y}) \in \mathbb{R}^{P+M}, \bar{\boldsymbol{\psi}} \in \bar{\boldsymbol{\Psi}}, K \in \mathbb{N} \Big\},$$

where $\bar{\boldsymbol{\Psi}}$ is the following constrained parameter space:

$$\bar{\boldsymbol{\Psi}} = \left\{ \bar{\boldsymbol{\psi}} \in \boldsymbol{\Psi} : \pi_k > 0\ \forall\ k, \sum_{k=1}^{K} \pi_k = 1, (\boldsymbol{\beta}_k^*, \boldsymbol{\Xi}_k) \neq (\boldsymbol{\beta}_h^*, \boldsymbol{\Xi}_h)\ \forall\ k \neq h \right\}.$$

For the identifiability of the class $\bar{\mathfrak{F}}$ it is also required that there exists a set $\mathcal{X} \subseteq \mathbb{R}^P$ having probability equal to one according to the P-dimensional contaminated normal distribution such that the following mixture of contaminated normal regression models

$$\sum_{k=1}^{K} \pi_k(\mathbf{x}) h\left(\mathbf{y}|\mathbf{x};\tilde{\mathbf{x}}^{*\prime}\boldsymbol{\beta}_k^*, \boldsymbol{\Xi}_k\right),\ \mathbf{y} \in \mathbb{R}^M,$$

is identifiable for each fixed $\mathbf{x} \in \mathcal{X}$, where $\pi_1(\mathbf{x}), ..., \pi_K(\mathbf{x})$ are positive weights summing to one for each $\mathbf{x} \in \mathcal{W}$. Then, it is possible to prove that the class $\bar{\mathfrak{F}}$ is identifiable in $\mathcal{X} \times \mathbb{R}^M$. Such a proof can be easily obtained by exploiting the same arguments described in Punzo and McNicholas (2017, Appendix B) for the identifiability of CNCW models with the following modifications: *(i)* the linear term to be considered in the conditional expected value of $\mathbf{Y}|(\mathbf{X} =$

$\mathbf{x}, G_k)$ is $\tilde{\mathbf{x}}^{*\prime}\boldsymbol{\beta}_k^*$; *(ii)* the set of all covariate points to be employed to distinguish between different regression coefficients $\boldsymbol{\beta}_k^*$ by different values of $\tilde{\mathbf{x}}^{*\prime}\boldsymbol{\beta}_k^*$ is:

$$\mathcal{X} = \Big\{\mathbf{x} \in \mathbb{R}^P : \forall\ \mathbf{x}_m \in \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}, \forall\ k, h \in \{1, \ldots, K\} \text{ and } s, t \in \{1, \ldots, \tilde{K}\}\ ,$$
$$\tilde{\mathbf{x}}_m^{*\prime}\boldsymbol{\beta}_{km}^* = \tilde{\mathbf{x}}_m^{*\prime}\boldsymbol{\beta}_{hm}^* \Rightarrow \boldsymbol{\beta}_{km}^* = \boldsymbol{\beta}_{hm}^*,\ \tilde{\mathbf{x}}_m^{*\prime}\boldsymbol{\beta}_{km}^* = \tilde{\mathbf{x}}_m^{*\prime}\tilde{\boldsymbol{\beta}}_{sm}^* \Rightarrow \boldsymbol{\beta}_{km}^* = \tilde{\boldsymbol{\beta}}_{sm}^*\ ,$$
$$\tilde{\mathbf{x}}_m^{*\prime}\tilde{\boldsymbol{\beta}}_{sm}^* = \tilde{\mathbf{x}}_m^{*\prime}\tilde{\boldsymbol{\beta}}_{tm}^* \Rightarrow \tilde{\boldsymbol{\beta}}_{sm}^* = \tilde{\boldsymbol{\beta}}_{tm}^*\Big\}\ .$$

2.4 An ECM algorithm for ML estimation

Let $\mathcal{S} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I)\}$ be a sample of I independent observations drawn from (5). Under these conditions, the log-likelihood function can be written as $l(\boldsymbol{\psi}) = \sum_{i=1}^{I} \ln\left(\sum_{k=1}^{K} \pi_k h\left(\mathbf{x}_i; \boldsymbol{\theta}_{k\mathbf{x}}\right) h\left(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}_{k\mathbf{y}}\right)\right)$. Similarly to other finite mixture models and following Punzo and McNicholas (2017), ML estimation of $\boldsymbol{\psi}$ has been carried out for a fixed value of K under a general framework dealing with incomplete-data problems (Dempster et al., 1977; Meng and Rubin, 1993). In the considered situation, there are three different types of incompleteness in the data $\mathcal{S}$: *(i)* the missing information about the specific sub-populations from which the I sample observations come from; *(ii)* the missing information about whether such observations are leverage points with reference to any given G_k or not; *(iii)* the missing information about whether each observation is an outlier with reference to any given G_k or not. The first type is typical of any finite mixture model; the second and third types are specific for (5). Such information can be described using three different types of K-dimensional vectors. For the ith sample observation, they are given by $\mathbf{z}_i$, $\mathbf{v}_i$, $\mathbf{u}_i$. Namely, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$, with $z_{ik} = 1$ if the ith observation comes from the sub-population G_k and $z_{ik} = 0$ otherwise, for $k = 1, \ldots, K$; $\mathbf{v}_i = (v_{i1}, \ldots, v_{iK})'$, with $v_{ik} = 1$ if the ith observation is not a leverage point within the sub-population G_k and $v_{ik} = 0$ if it is a leverage point; $\mathbf{u}_i = (u_{i1}, \ldots, u_{iK})'$, with $u_{ik} = 1$ if the ith observation is typical within the sub-population G_k and $u_{ik} = 0$ if it is an outlier. Thus, the complete data would be $\mathcal{S}_c = \{(\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1, \mathbf{v}_1, \mathbf{u}_1), \ldots, (\mathbf{x}_I, \mathbf{y}_I, \mathbf{z}_I, \mathbf{v}_I, \mathbf{u}_I)\}$. Then, following Punzo and McNicholas (2017), to find the ML estimate $\hat{\boldsymbol{\psi}}$, an ECM algorithm (Meng and Rubin, 1993) has been developed. To this end, the complete-data likelihood function can be written as

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

and the complete-data log-likelihood function employed in the ECM algorithm for the computation of $\hat{\boldsymbol{\psi}}$ is equal to

$$\begin{aligned}\ell_c(\boldsymbol{\psi}) = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\Big[&\ln \pi_k + v_{ik}\ln\alpha_k + (1-v_{ik})\ln(1-\alpha_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \\ &-\Big(\frac{P}{2}\ln\eta_k\Big)(1-v_{ik}) - \frac{1}{2}\Big(v_{ik} + \frac{1-v_{ik}}{\eta_k}\Big)\delta^2_{\boldsymbol{\Sigma}_k}\Big(\mathbf{x}_i, \boldsymbol{\mu}_k\Big) + \\ &+ u_{ik}\ln\tau_k + (1-u_{ik})\ln(1-\tau_k) - \frac{1}{2}\ln|\boldsymbol{\Xi}_k| + \\ &-\Big(\frac{M}{2}\ln\lambda_k\Big)(1-u_{ik}) - \frac{1}{2}\Big(u_{ik} + \frac{1-u_{ik}}{\lambda_k}\Big)\delta^2_{\boldsymbol{\Xi}_k}\Big(\mathbf{y}_i, \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^*\Big)\Big],\end{aligned}$$

where

$$\delta^2_{\boldsymbol{\Sigma}_k}(\mathbf{x}_i, \boldsymbol{\mu}_k) = (\mathbf{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k), \tag{8}$$

$$\delta^2_{\boldsymbol{\Xi}_k}\left(\mathbf{y}_i, \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^*\right) = (\mathbf{y}_i - \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^*)'\boldsymbol{\Xi}_k^{-1}(\mathbf{y}_i - \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^*) \tag{9}$$

are squared Mahalanobis distances: the first is computed between $\mathbf{x}_i$ and $\boldsymbol{\mu}_k$ with respect to $\boldsymbol{\Sigma}_k$; the second is computed between $\mathbf{y}_i$ and $\tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^*$ with respect to $\boldsymbol{\Xi}_k$.

For the description of the ECM algorithm, it is convenient to introduce the following vectors: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*)$, $\boldsymbol{\Xi} = (\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_K)$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$. The ECM algorithm is an iterative sequence. At each iteration, an E-step is followed by two CM-steps. The first CM-step focuses on the parameter sub-vector $\boldsymbol{\psi}_a = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\beta}^*, \boldsymbol{\Xi}, \boldsymbol{\tau})$. The second CM-step involves the parameter sub-vector $\boldsymbol{\psi}_b = (\boldsymbol{\eta}, \boldsymbol{\lambda})$. Iterations are repeated until convergence.

On the hth iteration of the E-step, given the current estimate $\boldsymbol{\psi}^{(h)}$ of the model parameters $\boldsymbol{\psi}$, the conditional expectation of $l_c(\boldsymbol{\psi})$ has to be computed; up to an additive constant, it is equal to:

$$\begin{aligned}Q\left(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)}\right) &= \mathbb{E}_{\boldsymbol{\psi}^{(h)}}[\ell_c(\boldsymbol{\psi})] \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K}\hat{z}_{ik}^{(h)}\Big\{\ln\pi_k^{(h)} + \hat{v}_{ik}^{(h)}\ln\alpha_k^{(h)} + (1-\hat{v}_{ik}^{(h)})\ln(1-\alpha_k^{(h)}) + \\ &+ Q_{i1}\Big(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{\psi}^{(h)}\Big) + \hat{u}_{ik}^{(h)}\ln\tau_k^{(h)} + (1-\hat{u}_{ik}^{(h)})\ln(1-\tau_k^{(h)}) + \\ &+ Q_{i2}\Big(\boldsymbol{\beta}_k^*, \boldsymbol{\Xi}_k|\boldsymbol{\psi}^{(h)}\Big)\Big\},\end{aligned}$$

where

$$Q_{i1}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{\psi}^{(h)}\right) = -\frac{1}{2}\Big[\ln|\boldsymbol{\Sigma}_k^{(h)}| + P(1-\hat{v}_{ik}^{(h)})\ln\eta_k^{(h)} + \\ +\Big(\hat{v}_{ik}^{(h)} + \frac{1-\hat{v}_{ik}^{(h)}}{\eta_k^{(h)}}\Big)\delta^2_{\boldsymbol{\Sigma}_k^{(h)}}\left(\mathbf{x}_i, \boldsymbol{\mu}_k^{(h)}\right)\Big],$$

$$Q_{i2}\left(\boldsymbol{\beta}_k^*, \boldsymbol{\Xi}_k|\boldsymbol{\psi}^{(h)}\right) = -\frac{1}{2}\Big[\ln|\boldsymbol{\Xi}_k^{(h)}| + M(1-\hat{u}_{ik}^{(h)})\ln\lambda_k^{(h)} + \\ +\Big(\hat{u}_{ik}^{(h)} + \frac{1-\hat{u}_{ik}^{(h)}}{\lambda_k^{(h)}}\Big)\delta^2_{\boldsymbol{\Xi}_k^{(h)}}\left(\mathbf{y}_i, \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h)}\right)\Big],$$

with

$$\hat{z}_{ik}^{(h)} = \mathbb{E}_{\boldsymbol{\psi}^{(h)}}[Z_{ik}|(\mathbf{x}_i, \mathbf{y}_i)] = \frac{\pi_k^{(h)} h\left(\mathbf{x}_i; \boldsymbol{\theta}_{k\mathbf{x}}^{(h)}\right) h\left(\mathbf{y}_i; \boldsymbol{\theta}_{k\mathbf{y}}^{(h)}\right)}{f\left(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\psi}^{(h)}\right)}, \tag{10}$$

$$\hat{v}_{ik}^{(h)} = \mathbb{E}_{\boldsymbol{\psi}^{(h)}}[V_{ik}|(\mathbf{x}_i, \mathbf{z}_i)] = \frac{\alpha_k^{(h)}\phi\Big(\mathbf{x}_i; \boldsymbol{\mu}_k^{(h)}, \boldsymbol{\Sigma}_k^{(h)}\Big)}{h\left(\mathbf{x}_i; \boldsymbol{\theta}_{k\mathbf{x}}^{(h)}\right)}, \tag{11}$$

$$\hat{u}_{ik}^{(h)} = \mathbb{E}_{\boldsymbol{\psi}^{(h)}}[U_{ik}|(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)] = \frac{\tau_k^{(h)}\phi\Big(\mathbf{y}_i; \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h)}, \boldsymbol{\Xi}_k^{(h)}\Big)}{h\left(\mathbf{y}_i; \boldsymbol{\theta}_{k\mathbf{y}}^{(h)}\right)}. \tag{12}$$

The random vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})'$ denotes a K-dimensional multinomial random vector with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, $V_{ik}|Z_{ik} = 1$ and $U_{ik}|Z_{ik} = 1$ being two Bernoulli random variables with success probability of α_k and τ_k, respectively, for $i = 1, \ldots, I$ and $k = 1, \ldots, K$. Thus, $\hat{z}_{ik}^{(h)}$, $\hat{v}_{ik}^{(h)}$ and $\hat{u}_{ik}^{(h)}$ represent posterior probabilities (evaluated using $\boldsymbol{\psi}^{(h)}$) of the following three events: *(i)* the sample observation $(\mathbf{x}_i, \mathbf{y}_i)$ comes from the kth distribution in (5); *(ii)* $(\mathbf{x}_i, \mathbf{y}_i)$ is not a leverage point within such a distribution; *(iii)* $(\mathbf{x}_i, \mathbf{y}_i)$ is not an outlier within such a distribution.

At the first CM-step on the $(h+1)$th iteration of the ECM algorithm, the sub-vector $\boldsymbol{\psi}_a^{(h)}$ is updated through the maximisation of $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)})$ with respect to $\boldsymbol{\psi}_a$ with $\boldsymbol{\psi}_b$ fixed at $\boldsymbol{\psi}_b^{(h)}$. The resulting updates of $\pi_k^{(h)}$, $\alpha_k^{(h)}$, $\tau_k^{(h)}$,

$\boldsymbol{\mu}_k^{(h)}$ and $\boldsymbol{\Sigma}_k^{(h)}$ are:

$$\pi_k^{(h+1)} = \frac{1}{I}\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}, \tag{13}$$

$$\alpha_k^{(h+1)} = \frac{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}\hat{v}_{ik}^{(h)}}{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}}, \tag{14}$$

$$\tau_k^{(h+1)} = \frac{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}\hat{u}_{ik}^{(h)}}{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}}, \tag{15}$$

$$\boldsymbol{\mu}_k^{(h+1)} = \frac{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}\hat{w}_{1_{ik}}^{(h)}\mathbf{x}_i}{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}\hat{w}_{1_{ik}}^{(h)}}, \tag{16}$$

$$\boldsymbol{\Sigma}_k^{(h+1)} = \frac{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}\hat{w}_{1_{ik}}^{(h)}\left(\mathbf{x}_i-\boldsymbol{\mu}_k^{(h+1)}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_k^{(h+1)}\right)'}{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}}, \tag{17}$$

where

$$\hat{w}_{1_{ik}}^{(h)} = \hat{v}_{ik}^{(h)} + \frac{1-\hat{v}_{ik}^{(h)}}{\eta_k^{(h)}}. \tag{18}$$

Such updates coincide with the solutions obtained for the CNCW model (for further details see Punzo and McNicholas, 2017, Appendices C.1-C.4). As far as the remaining elements of the sub-vector $\boldsymbol{\psi}_a^{(h)}$ are concerned, their updates are:

$$\boldsymbol{\beta}_k^{*(h+1)} = \Big(\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}\hat{w}_{2_{ik}}^{(h)}\tilde{\mathbf{x}}_i^*\boldsymbol{\Xi}_k^{(h)^{-1}}\tilde{\mathbf{x}}_i^{*'}\Big)^{-1}\Big(\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}\hat{w}_{2_{ik}}^{(h)}\tilde{\mathbf{x}}_i^*\boldsymbol{\Xi}_k^{(h)^{-1}}\boldsymbol{y}_i\Big), \tag{19}$$

$$\boldsymbol{\Xi}_k^{(h+1)} = \frac{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}\hat{w}_{2_{ik}}^{(h)}\left(\mathbf{y}_i-\tilde{\mathbf{x}}_i^{*'}\boldsymbol{\beta}_k^{*(h+1)}\right)\left(\mathbf{y}_i-\tilde{\mathbf{x}}_i^{*'}\boldsymbol{\beta}_k^{*(h+1)}\right)'}{\sum_{i=1}^{I}\hat{z}_{ik}^{(h)}}, \tag{20}$$

where

$$\hat{w}_{2_{ik}}^{(h)} = \hat{u}_{ik}^{(h)} + \frac{1-\hat{u}_{ik}^{(h)}}{\lambda_k{}^{(h)}}. \tag{21}$$

The updates illustrated in (19)–(20) coincide with the ones obtained for the seemingly unrelated contaminated normal clusterwise regression models defined by (7) (further details can be found in Perrone and Soffritti, 2023, Appendix A).

At the second CM-step on the $(h+1)$th iteration of the ECM algorithm, the update of $\boldsymbol{\psi}_b^{(h)}$ is obtained by maximising $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(h)})$ with respect to $\boldsymbol{\psi}_b$

with $\boldsymbol{\psi}_a$ fixed at $\boldsymbol{\psi}_a^{(h+1)}$. The resulting updates of $\eta_k^{(h)}$ and $\lambda_k^{(h)}$ are:

$$\eta_k^{(h+1)} = \max\left\{1, \frac{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}(1-\hat{v}_{ik}^{(h)})\delta^2_{\boldsymbol{\Sigma}_k^{(h+1)}}\left(\mathbf{x}_i, \boldsymbol{\mu}_k^{(h+1)}\right)}{P\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}(1-\hat{v}_{ik}^{(h)})}\right\}, \tag{22}$$

$$\lambda_k^{(h+1)} = \max\left\{1, \frac{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}(1-\hat{u}_{ik}^{(h)})\delta^2_{\boldsymbol{\Xi}_k^{(h+1)}}\left(\mathbf{y}_i, \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h+1)}\right)}{M\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}(1-\hat{u}_{ik}^{(h)})}\right\}. \tag{23}$$

Further details can be found in Punzo et al. (2018).

As far as the estimation of α_k and τ_k is concerned, (14) and (15) can be modified to guarantee that the estimated proportions of typical observations both in the regression hyperplane and in the predictor space within each cluster are at least 0.5. The modified equations are: $\alpha_k^{(h+1)} = \max\left\{0.5, \frac{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}\hat{v}_{ik}^{(h)}}{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}}\right\}$ and $\tau_k^{(h+1)} = \max\left\{0.5, \frac{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}\hat{u}_{ik}^{(h)}}{\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}}\right\}$, for $k = 1, \ldots, K$. The update $\boldsymbol{\beta}_k^{*(h+1)}$ can be computed only if the matrix $\sum_{i=1}^{I} \hat{z}_{ik}^{(h)}\hat{w}_{2_{ik}}^{(h)}\tilde{\mathbf{x}}_i^*\boldsymbol{\Xi}_k^{(h)^{-1}}\tilde{\mathbf{x}}_i^{*\prime}$ in (19) is non-singular. Furthermore, it is important to highlight that the updates of $\boldsymbol{\Sigma}_k^{(h)}$ and $\boldsymbol{\Xi}_k^{(h)}$ reported in (17) and (20) apply to the SuCNCW models with the `VVV` parameterisation for either $\boldsymbol{\Sigma}_k$ or $\boldsymbol{\Xi}_k$ (i.e., the fully unconstrained covariance structure of the predictors and responses). For the ML estimation of $\boldsymbol{\Sigma}_k$ or $\boldsymbol{\Xi}_k$ under any other SuCNCW model, the M step updates in the ECM algorithm can be computed either in closed form or using iterative procedures, depending on the specific parameterisation to be employed (see Celeux and Govaert, 1995, for more details). For the estimation of models obtained using the `EVE` and `VVE` parameterisations, it is possible to resort to some majorization-minimization algorithms which are computationally feasible in high-dimensional situations (Browne and McNicholas, 2014a,b).

The main result of the ECM algorithm is represented by the ML estimate $\hat{\boldsymbol{\psi}}$, that is the value of $\boldsymbol{\psi}^{(h)}$ at convergence. As a by-product, by (10)–(12) this algorithm also provides estimates of the following posterior probabilities: $\mathbb{P}_{\hat{\boldsymbol{\psi}}}[Z_{ik} = 1|(\mathbf{x}_i, \mathbf{y}_i)] = \hat{z}_{ik}$, $\mathbb{P}_{\hat{\boldsymbol{\psi}}}[V_{ik} = 1|(\mathbf{x}_i, \hat{\mathbf{z}}_i)] = \hat{v}_{ik}$ and $\mathbb{P}_{\hat{\boldsymbol{\psi}}}[U_{ik} = 1|(\mathbf{x}_i, \mathbf{y}_i, \hat{\mathbf{z}}_i)] = \hat{u}_{ik}$, for $i = 1, \ldots, I$ and $k = 1, \ldots, K$. Then, the I sample observations can be partitioned into K clusters using the maximum a posteriori 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, the estimates $\hat{v}_{ik}$ and $\hat{u}_{ik}$ can be employed to define four intra-cluster distinctions. Namely, if $\hat{v}_{ih} < 0.5$, where h is the label of the cluster for which $\text{MAP}(\hat{z}_{ik}) = 1$, the ith observation will be classified as a good leverage point for that cluster; if $\hat{u}_{ih} < 0.5$, then the ith observation will be classified as a mild outlier for the same cluster; if both conditions just mentioned are jointly fulfilled ($\hat{v}_{ih} < 0.5$ and $\hat{u}_{ih} < 0.5$), then the ith observation will be

classified as a bad leverage point for the hth cluster; finally, if $\hat{v}_{ih} \geq 0.5$ and $\hat{u}_{ih} \geq 0.5$, then the ith observation will be considered typical. The ML estimate $\hat{\boldsymbol{\psi}}$ can also be exploited in conjunction with (8) and (9) to compute the estimated squared Mahalanobis distances $\hat{d}^2_{ik\mathbf{x}} = \delta^2_{\hat{\boldsymbol{\Sigma}}_k}(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_k)$ and $\hat{d}^2_{ik\mathbf{y}} = \delta^2_{\hat{\boldsymbol{\Xi}}_k}\left(\mathbf{y}_i, \tilde{\mathbf{x}}_i^{*\prime}\hat{\boldsymbol{\beta}}_k^*\right)$, for $i = 1, \ldots, I$ and $k = 1, \ldots, K$, which can be interpreted as intra-cluster quantifications of the amount of deviations from the pattern of the observations assigned to any given cluster. Thus, a more detailed analysis of the leverage points and mild outliers could be carried out by considering the values of $\hat{d}^2_{ik\mathbf{x}}$ and $\hat{d}^2_{ik\mathbf{y}}$ $\forall (i,k) \in \{i \in \{1, \ldots, I\}, k : \text{MAP}(\hat{z}_{ik}) = 1\}$ and by focusing on the largest values obtained in this way (see McLachlan and Peel, 2000, p. 232).

2.5 Notes on the robustness of the parameter estimates

Equation (19) shows that the update $\boldsymbol{\beta}_k^{*(h+1)}$ can be seen as a generalised least squares estimate with weights depending on $\hat{z}_{ik}^{(h)}$ and $\hat{w}_{2_{ik}}^{(h)}$; such weights also affect the update $\boldsymbol{\Xi}_k^{(h+1)}$ in (20), which represents a weighted sum of squared residuals. From the definition of the squared Mahalanobis distance $\delta^2_{\boldsymbol{\Xi}_k}\left(\mathbf{y}_i, \tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^*\right)$ given in (9) and the expressions for $\hat{u}_{ik}^{(h)}$ and $\hat{w}_{2_{ik}}^{(h)}$ reported in (12) and (21), respectively, it is possible to write both $\hat{u}_{ik}$ and $\hat{w}_{2_{ik}}$ as decreasing functions of the estimated squared Mahalanobis distances $\hat{d}^2_{ik\mathbf{y}}$ illustrated in Section 2.4 (for the explicit expressions, see Punzo and McNicholas, 2017). As a consequence, sample observations with the highest posterior estimated probabilities of being generated from the kth distribution in (5) and of representing typical points in the regression hyperplane according to that distribution will have the largest impact on the updates of both the regression coefficients and covariances of $\mathbf{Y}|(\mathbf{X} = \mathbf{x}, G_k)$. For this reason, this approach provides robust estimates of $\boldsymbol{\beta}_k^*$ and $\boldsymbol{\Xi}_k$ for $k = 1, \ldots K$. In a similar way, the updates $\boldsymbol{\mu}_k^{(h+1)}$ and $\boldsymbol{\Sigma}_k^{(h+1)}$ given in (16) and (17) represent a weighted mean vector and a weighted covariance matrix, respectively, with weights depending on $\hat{z}_{ik}^{(h)}$ and $\hat{w}_{1_{ik}}^{(h)}$. Since $\hat{w}_{1_{ik}}^{(h)}$ shows the same structure of $\hat{w}_{2_{ik}}^{(h)}$ and the structure of the update for $\hat{v}_{ik}^{(h)}$ in (11) is the same as the one of the update for $\hat{u}_{ik}^{(h)}$ in (12), it is also possible to express both $\hat{v}_{ik}$ and $\hat{w}_{1_{ik}}$ as decreasing functions of the estimated squared Mahalanobis distances $\hat{d}^2_{ik\mathbf{x}}$ illustrated in Section 2.4. Thus, the term $\hat{w}_{1_{ik}}^{(h)}$ in (16) and (17) reduces the impact of the leverage points on the updates $\boldsymbol{\mu}_k^{(h+1)}$ and $\boldsymbol{\Sigma}_k^{(h+1)}$, which are robust solutions when estimating $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

It is also worth noting that, according to (22) and (23), the updates $\eta_k^{(h+1)}$ and $\lambda_k^{(h+1)}$ will be larger when the kth distribution of the mixture in (5) is highly contaminated by the presence of outliers and leverage points (i.e., when many observations show small values of $\hat{v}_{ik}^{(h)}$ and $\hat{u}_{ik}^{(h)}$ or, equivalently, large squared Mahalanobis distances from $\boldsymbol{\mu}_k^{(h+1)}$ and $\tilde{\mathbf{x}}_i^{*\prime}\boldsymbol{\beta}_k^{*(h+1)}$).

2.6 Technical details about ML estimation

Appropriately choosing the starting value $\boldsymbol{\psi}^{(0)}$ for the model parameters is paramount when estimating $\boldsymbol{\psi}$ using the ECM algorithm described in Section 2.4. To this end, strategies usually employed in finite mixture models (e.g., multiple executions of the algorithm using multiple random initialisations, approaches based on non-random choices of either $\boldsymbol{\psi}^{(0)}$ or the missing information) could be adopted (see, e.g., Biernacki et al., 2003; Karlis and Xekalaki, 2003, for more details). More specific initialisation strategies could be devised by resorting to the normal mixture model of order K for $(\mathbf{X}, \mathbf{Y})$. This latter model has been proved to represent a reparameterisation of the NCW (2) (see Ingrassia et al., 2012, for more details) which, in turn, is nested in the CNCW (3) when $\alpha_k \to 1^-$, $\tau_k \to 1^-$, $\eta_k \to 1^+$ and $\lambda_k \to 1^+$, for $k = 1, \ldots, K$. Thus, a first strategy for choosing the initial values $\hat{z}_{ik}^{(0)}$, for $i = 1, \ldots, I$ and $k = 1, \ldots, K$, is to set such quantities equal to the estimated posterior probabilities of the normal mixture model of order K for $(\mathbf{X}, \mathbf{Y})$. Furthermore, $\hat{v}_{ik}^{(0)}$ and $\hat{u}_{ik}^{(0)}$ could be set equal to 0.999 for $i = 1, \ldots, I$ and $k = 1, \ldots, K$. This strategy, which has been suggested and employed by Punzo and McNicholas (2017) for the ML estimation of (3), guarantees that the observed-data log-likelihood of the CNCW model will not be lower than the observed-data log-likelihood of the starting NCW model (see Punzo and McNicholas, 2017, for more details). In order to make the initialisation robust with respect to the presence of mildly atypical observations, the starting values $\hat{z}_{ik}^{(0)}$, $\hat{v}_{ik}^{(0)}$ and $\hat{u}_{ik}^{(0)}$, for $k = 1, \ldots, K$, could be jointly obtained from the fitting of a contaminated normal mixture model of order K for $(\mathbf{X}, \mathbf{Y})$.

In the analyses reported in Sections 3.2.1–3.2.6, the ECM algorithm has been initialised using a strategy composed of the following three steps. Firstly, the normal mixture model of order K for $(\mathbf{X}, \mathbf{Y})$ is estimated using the data $\mathcal{S}$. The resulting estimates of the mixing weights, the expected values and the variances-covariances of $\mathbf{X}$ are employed to obtain the starting values $\pi_k^{(0)}$, $\boldsymbol{\mu}_k^{(0)}$ and $\boldsymbol{\Sigma}_k^{(0)}$. Secondly, a seemingly unrelated linear regression model for $\mathbb{E}(\mathbf{Y}|\mathbf{X}_m = \mathbf{x}_m)$ is fitted to the subsample of $\mathcal{S}$ composed of the observations assigned to the kth cluster detected by the normal mixture model considered in the previous step (for $k = 1, \ldots, K$). The starting values $\boldsymbol{\beta}_k^{*(0)}$ and $\boldsymbol{\Xi}_k^{(0)}$ are given by the vector containing the estimated intercept and regression coefficients and the matrix with the variances and covariances of the sample residuals, respectively. Thirdly, $\alpha_k^{(0)}$ and $\tau_k^{(0)}$, for $k = 1 \ldots, K$, are set equal to 0.999; $\eta_k^{(0)}$ and $\lambda_k^{(0)}$ are set equal to 1.001. The results reported in Sections 3.2.7 and 4 have been obtained using a different initialisation strategy, in which $\boldsymbol{\mu}_k^{(0)}$, $\boldsymbol{\Sigma}_k^{(0)}$, $\alpha_k^{(0)}$ and $\tau_k^{(0)}$, for $k = 1 \ldots, K$, are set equal to the estimates of a contaminated normal mixture model of order K for $(\mathbf{X}, \mathbf{Y})$. Furthermore, the sample observations assigned to the kth cluster detected by this latter model are employed to estimate the seemingly unrelated linear regression model for $\mathbb{E}(\mathbf{Y}|\mathbf{X}_m = \mathbf{x}_m, G_k)$. The starting values $\boldsymbol{\psi}^{(0)}$ computed using

this latter strategy are expected to be robust with respect to the presence of mildly atypical observations. The packages `mclust` (Scrucca et al., 2017), `systemfit` (Henningsen and Hamann, 2007) and `ContaminatedMixt` (Punzo et al., 2018) in the `R` environment (R Core Team, 2022) have been employed to estimate the models involved in such strategies.

The ECM algorithm is stopped using either a convergence criterion which exploits the Aitken acceleration (Aitken, 1926) or when a pre-specified maximum number of iterations has been reached. The convergence criterion is based on the computation of the quantity $|\ell_A^{(h+1)} - \ell(\boldsymbol{\psi}^{(h)})|$, where $\ell_A^{(h+1)}$ 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). Iterations are stopped when this quantity is lower than a positive and finite tolerance threshold tol. The analyses reported in Sections 3 and 4 have been carried out with $tol = 10^{-4}$ and 500 as the maximum number of iterations. Finally, some constraints on the eigenvalues of $\boldsymbol{\Sigma}_k^{(h)}$ and $\boldsymbol{\Xi}_k^{(h)}$ $(k = 1, \ldots, K)$ have been embedded in the ECM algorithm to avoid the issue of a unbounded likelihood caused by a degenerate model. Namely, following Dang et al. (2017), all eigenvalues have been required to be greater than 10^{-20}.

Since the ECM algorithm returns an estimate of $\boldsymbol{\psi}$ for a given value of K, in any practical application in which this number is not known, it has to be determined from the data $\mathcal{S}$. This task is typically carried out by resorting to model selection criteria, such as the Bayesian information criterion (BIC) (Schwarz, 1978) or the integrated completed likelihood (ICL) (Biernacki et al., 2000). They can be computed as follows:

$$\begin{aligned}
\text{BIC} &= 2\ell(\hat{\boldsymbol{\psi}}) - n_{\boldsymbol{\psi}} \ln I, \\
\text{ICL} &= 2\ell(\hat{\boldsymbol{\psi}}) - n_{\boldsymbol{\psi}} \ln I + 2 \sum_{i=1}^{I} \sum_{k=1}^{K} \text{MAP}(\hat{z}_{ik}) \ln \hat{z}_{ik}.
\end{aligned}$$

Higher values of these criteria indicate better-fit models. The BIC evaluates the adequacy of a model by taking account of the trade-off between the fit and the model complexity. In the computation of the ICL, an additional penalty accounting for the uncertainty of the estimated partition is considered (see, e.g., Andrews and McNicholas, 2011; Baek and McLachlan, 2011). Such a penalty is based on a hard (i.e., $\text{MAP}(\hat{z}_{ik})$) clustering of the sample observations. As a consequence, the ICL can penalize complex models more severely than BIC; furthermore, it should less likely split one cluster into two different components. This latter feature is consistent with the fact that the ICL has been proposed as a criterion able to select the model which shows the greatest evidence of clustering (Biernacki et al., 2000). In contrast, selecting the number of components which leads to a good approximation to the density is the aspect which the BIC mainly focuses on (Baudry et al., 2010).

2.7 Feature selection

SuCNCW analyses based on (5) can be carried out for given vectors $\mathbf{X}_m = (\tilde{X}_1, \ldots, \tilde{X}_{P_m})'$, $m = 1, \ldots, M$. Thus, it is necessary that the researcher identifies the covariates which are relevant for the prediction of each response. In some situations, this task can be performed by exploiting prior information or a certain general theory about the phenomenon. Examples are represented by employment equations (White and Hewings, 1982), Cobb-Douglas production functions (Giles and Hampton, 1984), prediction of carcass compositions (Cadavez and Hennningsen, 2012) and tourists' expenditure behaviour (Disegna and Osti, 2016). However, there may also be situations in which the specific regressors to be used for the prediction of each response are questionable. In this latter case, the analysis of a given dataset should be carried out by examining a class of SuCNCW models in which $\mathbf{X}_m \in \mathcal{R}$ for $m = 1, \ldots, M$, where $\mathcal{R} = \{\emptyset, \ldots, \mathbf{X}\}$ contains all possible subsets of regressors which can be considered for modelling the conditional expected value $\mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}, G_k)$ according to (4). Once such models are fitted to the dataset, the specific covariates to be included in the vectors $\mathbf{X}_m = (\tilde{X}_1, \ldots, \tilde{X}_{P_m})'$, $m = 1, \ldots, M$, can be determined on the basis of suitable statistical criteria (i.e., a model selection criterion). However, an exhaustive search of all the possible subsets of regressors for the M responses can be unfeasible in practical applications, particularly in the presence of complex and high-dimensional datasets. To manage this issue, strategies that perform variable selection in a multivariate regression framework can be employed (see, e.g., Miller, 1991). Stepwise selection techniques represent a common approach; however, strategies based on a local exploration of $\mathcal{R}$, in which variables are included and/or excluded one-by-one, can produce poor results, especially when there are many covariates. This drawback can be overcome by resorting to stochastic iterative algorithms. In particular, complex problems admitting a vast number of possible solutions can be solved through evolutionary algorithms by exploiting principles and operators of the biological evolution of a species (see, e.g., Goldberg, 1989). Namely, mutation is a random alteration of a gene in a chromosome; crossover is a random process of genome recombination that applies to pairs of chromosomes; selection is a random process in which chromosomes are chosen from a generation for later breeding, depending on their fitness for the evolution of the species. Some applications in statistics are illustrated in Chatterjee et al. (1996). Genetic algorithms for subset selection in model-based clustering are described in Scrucca (2016) and Galimberti et al. (2018).

In order to perform covariate selection through a genetic algorithm in a SuCNCW model, the covariates to be included in the linear predictor for $\mathbb{E}(\mathbf{Y}|\mathbf{X} = \mathbf{x}, G_k)$ can be described by resorting to a binary encoding scheme. In the presence of P covariates, this approach leads to an ordered sequence of P genes for each response. Each gene can take on a value of either 0 or 1, where 0/1 represents the exclusion/inclusion of a certain covariate for the prediction of Y_m. The fitness of the SuCNCW model resulting from such a selection can be measured by means of the BIC or the ICL. In general, an execution of a

genetic algorithm starts from an initial population of N randomly generated chromosomes. Then, an iterative evolution process is performed, based on the three operators mentioned above (selection, crossover, mutation), with the goal of generating novel populations composed of chromosomes characterised by improved fitness values. The chromosomes of the resulting novel generation are assigned their fitness, and the evolution process repeats. Usually, the algorithm stops when a maximum number of populations ($d_{\max}$) has been generated. N and $d_{\max}$ represent tuning parameters of the algorithm which have to be fixed in advance by the researcher. Values of $d_{\max}$ and N should be chosen to ensure a proper exploration of $\mathcal{R}$ and the given model space.

3 Simulation studies

3.1 Settings

The task of investigating the effectiveness of SuCNCW models in comparison with NCW, CNCW and SuNCW models has been carried out in a multivariate setting with $M = 2$ responses, $P = 3$ covariates and simulated datasets comprising observations randomly sampled from $K = 3$ different distributions.

All the models employed to generate the datasets have been specified within the seemingly unrelated approach. More specifically, the response Y_1 has been assumed to linearly depend on X_1 and X_2, while the assumption for Y_2 is that it linearly depends on X_1 and X_3. Thus, $\mathbf{X}_1 = (X_1, X_2)'$, $\mathbf{X}_2 = (X_1, X_3)'$, and (4) reduces to:

$$\mathbb{E}(Y_1|\mathbf{X} = \mathbf{x}, G_k) = \mathbf{x}_1^{*\prime}\boldsymbol{\beta}_{k1}^* = \beta_{k10} + \beta_{k11}x_1 + \beta_{k12}x_2,$$
$$\mathbb{E}(Y_2|\mathbf{X} = \mathbf{x}, G_k) = \mathbf{x}_2^{*\prime}\boldsymbol{\beta}_{k2}^* = \beta_{k20} + \beta_{k21}x_1 + \beta_{k22}x_3.$$

As far as the data generation processes are concerned, models belonging to the following classes have been employed:

I SuNCW;
II SuCNCW with $\alpha_k = 0.95,\ \eta_k = 5, \tau_k = 0.9,\ \lambda_k = 10\ \forall k$;
III Student-t CW models with $\nu_1 = \nu_2 = \nu_3 = 4$ degrees of freedom;
IV SuNCW with 1% of the points $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$, with $z_{ik} = 1$, randomly substituted by outliers with coordinates $(\mathbf{x}_i, \tilde{\mathbf{y}}_i)$, where $\tilde{y}_{im}$, for $m = 1, 2$, is generated from a uniform distribution over the interval $(-11, -9)$;
V SuNCW with 1% of the points $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$, with $z_{ik} = 1$, randomly substituted by good leverage points with coordinates $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)$, where $\tilde{x}_{ip}$, for $p = 1, 2, 3$, is generated from a uniform distribution over the interval $(-7, -5)$, $\tilde{y}_{i1} = \beta_{110} + \beta_{111}\tilde{x}_{i1} + \beta_{112}\tilde{x}_{i2}$, and $\tilde{y}_{i2} = \beta_{120} + \beta_{121}\tilde{x}_{i1} + \beta_{122}\tilde{x}_{i3}$;
VI SuNCW with 1% of the points $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$, with $z_{ik} = 1$, randomly substituted by bad leverage points with coordinates $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)$, where $\tilde{x}_{ip}$, for $p = 1, 2, 3$, is generated from a uniform distribution over the interval $(-7, -5)$, and $\tilde{y}_{im}$, for $m = 1, 2$, is generated from a uniform distribution over the interval $(-11, -9)$.

All of these processes share the following common parameters for the data generation: $\pi_1 = 0.4$, $\pi_2 = 0.35$, $\pi_3 = 0.25$, $\boldsymbol{\mu}_1 = (0, 0, 0)'$, $\boldsymbol{\mu}_2 = (2, 4, -2)'$, $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_2 + 2\epsilon \cdot \mathbf{1}_P$, where $\mathbf{1}_P$ is the $P \times 1$ vector having each element equal to 1, $\boldsymbol{\beta}_1^* = (-2, 0.75, 1, 1, 0.5, -2)'$, $\boldsymbol{\beta}_2^* = (0.5, 1.75, 0.25, 1, 1, 1)'$, $\boldsymbol{\beta}_3^* = \boldsymbol{\beta}_2^* + \epsilon \cdot \mathbf{1}_6$,

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.72 & -0.18 & 0.27 \\ -0.18 & 1.89 & 0.27 \\ 0.27 & 0.27 & 2.89 \end{pmatrix}, \ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2.33 & -0.52 & -0.06 \\ -0.52 & 0.88 & -0.34 \\ -0.06 & -0.34 & 1.04 \end{pmatrix}, \ \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_2,$$

$$\boldsymbol{\Xi}_1 = \begin{pmatrix} 1.34 & 0.47 \\ 0.47 & 1.66 \end{pmatrix}, \ \boldsymbol{\Xi}_2 = \begin{pmatrix} 0.50 & 0.04 \\ 0.04 & 1.50 \end{pmatrix}, \ \boldsymbol{\Xi}_3 = \boldsymbol{\Xi}_2.$$

Thus, the covariance structures of both the predictors and the responses within the three groups have been obtained using the `VVV` parameterisation. Since the difference between the parameters $(\boldsymbol{\theta}_{k\mathbf{x}}, \boldsymbol{\theta}_{k\mathbf{y}})$ for $k = 2, 3$ only depends on ϵ, different values of ϵ can be chosen so as to determine different degrees of separation between the second and third groups of sample observations.

Under each process mentioned above, 100 different datasets have been generated considering the sample size ($I = 500, 1000$) and the degree of separation ($\epsilon = 0.35, 0.55$) as experimental factors. Overall, 2400 different datasets have been generated. The whole analysis has been run by employing an IBM x3750 M4 server with 4 Intel Xeon E5-4620 processors with 8 cores and 128GB RAM.

It is worth noting that, since NCW models of order K can be seen as nested in the CNCW models of order K and the latter models can be seen as a special case of SuCNCW models of order K, when data come from a NCW model of order K, the performances of these three types of models are expected to be similar (see the simulation study in Punzo and McNicholas, 2017).

3.2 Results

The comparative study of the effectiveness of the four model classes has been structured into three parts. In the first part, SuNCW, NCW, CNCW and SuCNCW models of order $K = 3$ with the `VVV` parameterisation for both $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Xi}_k$ $(k = 1, 2, 3)$ have been fitted to each dataset generated according to the processes I, II and III. As far as NCW and CNCW models are concerned, each response has been assumed to linearly depend on all the covariates; namely:

$$\mathbb{E}(Y_1|\mathbf{X} = \mathbf{x}, G_k) = \beta_{k10} + \beta_{k11}x_1 + \beta_{k12}x_2 + \beta_{k13}x_3,$$
$$\mathbb{E}(Y_2|\mathbf{X} = \mathbf{x}, G_k) = \beta_{k20} + \beta_{k21}x_1 + \beta_{k22}x_3 + \beta_{k23}x_2.$$

With this specification, the fitted NCW and CNCW models are not parsimonious: for each k, six regression coefficients have been estimated although, in fact, only four of them are different from zero. As far as the time elapsed between the start and completion of the parameter estimation is concerned, fitting a SuCNCW model has required - on average over the 100 datasets with $I = 500$ - between 1.069 and 72.999 seconds, depending on the data generation process and the value of ϵ. The minimum and maximum average execution

times have resulted to be equal to 0.991 and 14.195 seconds with SuNCW models, 0.995 and 16.341 seconds with NCW models, 1.034 and 78.343 seconds with CNCW models. However, it is important to note that the ECM algorithm has not been implemented with the goal of being efficient from a computational point of view. Thus, more efficient implementations could greatly reduce these illustrative CPU times. In the first part of this study, the comparison among the competing models has been carried out by examining their performances with reference to the following three aspects: *(i)* the estimation of the proportions of typical observations and the degrees of contamination both in the predictor space (proper estimation of α_k and η_k) and in the regression hyperplane (proper estimation of τ_k and λ_k); *(ii)* the ability to recover the true values of the unknown parameters (parameter recovery); *(iii)* the capability to recover the true partition of the sample observations (classification recovery). Aspect *(i)* has been studied only for the fitted CNCW and SuCNCW models. The evaluation of *(ii)* focused on the regression coefficients. In order to prevent the effects of label switching issues on the evaluation of these aspects, the components of the mixtures involved in each fitted model have been labelled by minimising the Euclidean distance to the true parameter values (see, e.g., Bai et al., 2012; Yao, 2014; Punzo and McNicholas, 2017).

In the second part, the study aims at evaluating the performances of the four model classes without exploiting the knowledge of the true value of K. Thus, also SuNCW, NCW, CNCW and SuCNCW models of order $K = 1, 2, 4$ with the `VVV` parameterisation for $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Xi}_k$ have been fitted to each dataset generated using I, II and III. The results obtained for all the examined values of K have been employed to study the following aspects: *(iv)* the capability to reach the best trade-off between the fit and model complexity; *(v)* the ability of BIC and ICL to detect the true value of K; *(vi)* a further evaluation of the classification recovery.

The third part is focused on the performance of CNCW and SuCNCW models in detecting regression outliers, good leverage points and bad leverage points (aspect *(vii)*). To this end, CNCW and SuCNCW models of order $K = 3$ with the `VVV` parameterisation for $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Xi}_k$ have been fitted to each dataset generated using IV, V and VI. Some experimental results are summarised in Tables A–L (see supplementary material).

3.2.1 Estimation of α_k, τ_k, η_k, λ_k

When the datasets only contain typical observations (process I), the averages of the estimated proportions of good points ($\hat{\alpha}_k$ and $\hat{\tau}_k$) and the estimated inflation parameters ($\hat{\eta}_k$ and $\hat{\lambda}_k$) are close to 1. In the presence of datasets with contaminated observations generated according to II, the estimates of such parameters are, on average, close to their true values. These results hold true under both CNCW and SuCNCW models, regardless of the level of separation and the sample size (see the upper and central parts of Tables A–D). Thus, the proportions of good points and the inflation parameters appear to

be properly estimated using either model. However, in II slightly higher standard deviations have been registered for the estimated inflation parameters, especially for λ_k. These latter results seem to highlight that the estimation of the inflation parameters is characterised by a certain instability under both CNCW and SuCNCW models. This phenomenon seems to reduce as the sample size increases. Finally, with the contaminated datasets generated according to III, the mean values of $\hat{\alpha}_k$, $\hat{\tau}_k$, $\hat{\eta}_k$ and $\hat{\lambda}_k$ for $k = 1, 2, 3$ are all quite far from 1, regardless of the values of ϵ and I (see the lower part of Tables A–D). Thus, CNCW and SuCNCW models have been able to detect the departure from a normal distribution for both $\mathbf{X}|G_k$ and $\mathbf{Y}|(\mathbf{X} = \mathbf{x}, G_k)$, for $k = 1, 2, 3$, due to the use of the Student-t distribution in III.

3.2.2 Parameter recovery

To evaluate *(ii)* with respect to the regression coefficients β_{kmp}, the following quantity has been computed:

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

where $\hat{\beta}_{kmp}^{(r)}$ is the ML estimate of β_{kmp} obtained from the rth dataset ($r = 1, \ldots, 100$). Since NCW and CNCW models also contain some regression coefficients associated with irrelevant regressors, the RMSE has been computed for these additional coefficients, using 0 as their true value.

Under process I, SuNCW and SuCNCW models give similar results in terms of the ability to recover the regression coefficients with both sample sizes and both degrees of separation (see their RMSEs in Tables E and F). Since all the parameters able to capture the possible presence of mildly atypical observations in SuCNCW have been properly estimated (see *(i)*), the inclusion of these parameters when the analysed datasets do not contain atypical observations does not show any relevant impact on the recovery of the true β_{kmp}. On the contrary, if irrelevant predictors are included in both regression equations (i.e., using NCW and CNCW models), a slight increase in the RMSEs of some regression coefficients is observed when $I = 500$, and this is especially true for $\epsilon = 0.35$. However, such an effect almost disappears when $I = 1000$. With the contaminated datasets generated using process II, as expected, SuCNCW models show the best performance with both sample sizes and both degrees of separation (see Tables G and H). With this process, the accuracy of CNCW models seem to be slightly higher than that of NCW and SuNCW models for the majority of the regression coefficients. Under process III, the lowest RMSEs are still obtained using the SuCNCW model with all the examined experimental situations (see Tables I and J). Furthermore, thanks to their effectiveness in detecting the non-normality of the distributions of $\mathbf{X}|G_k$ and $\mathbf{Y}|(\mathbf{X} = \mathbf{x}, G_k)$ for $k = 1, 2, 3$, CNCW models generally perform slightly better than NCW and SuNCW models. However, it is worth noting that, with the

lowest values of I and ϵ, the RMSEs obtained using the SuNCW model are slightly lower than those registered with CNCW models for the majority of the regression coefficients. As far as the irrelevant regressors are concerned, NCW and CNCW models appear to be equally capable of recognising their presence in the analysis of uncontaminated datasets, as the corresponding estimated regression coefficients are on average quite close to 0. However, when the data are contaminated, large values of the RMSE have been registered for the estimates of the regression coefficients associated with the irrelevant regressors in the second and third cluster. This latter result is particularly evident when the separation between these two clusters is low. Furthermore, the precision of CNCW models in the estimation of the effect of most irrelevant regressors using contaminated datasets results to be higher than that of NCW models.

3.2.3 Classification recovery

The study of *(iii)* has required an evaluation of the agreement between the partitions of the sample units detected by the four types of models and the true partition. To this end, the adjusted Rand index (ARI) (Hubert and Arabie, 1985) has been employed. Average values and standard deviations of this index (over the 100 datasets) for the four model classes under the three data generation processes by the examined levels of the two experimental factors are reported in Table 2. When the analysed datasets do not contain atypical observations, the classification recovery of all model classes is almost perfect with both levels of separation and both sample sizes (ARI ≥ 0.945). The results obtained under processes II and III show that the classification recovery associated with the use of all models increases with the level of separation between the second and third components for each value of I; it also increases with the sample size for each value of ϵ. With datasets generated using these processes, SuCNCW models are characterised by the greatest ability to properly estimate the true classification of the sample observations for each examined level of the two experimental factors. The partitions obtained from SuCNCW models also show a good agreement with the true partitions ($0.804 \leq \text{ARI} \leq 0.966$). Among the other three model classes, CNCW models outperform both SuNCW and NCW models. For these two latter models, the classification recovery appears to be markedly lower, especially with the lowest level of separation ($0.639 \leq \text{ARI} \leq 0.663$ with $I = 500$, $0.665 \leq \text{ARI} \leq 0.678$ with $I = 1000$).

3.2.4 Trade-off between fit and complexity

In order to study aspect *(iv)*, for each dataset and each model class, the models of order $\hat{K}_{\text{IC}}$ have been selected, where IC denotes an information criterion ($\text{IC} \in \{\text{BIC}, \text{ICL}\}$) and $\hat{K}_{\text{IC}} = \arg\max \text{IC}(K)$ for $K \in \{1, 2, 3, 4\}$. Then, for each information criterion and each dataset, the four values of $\text{IC}(\hat{K}_{\text{IC}})$ associated with the four examined model classes have been compared, and the

Table 2 Classification recovery of the fitted SuNCW, NCW, CNCW and SuCNCW models with $K = 3$: average values (standard deviations) of the ARI index over 100 samples.

I	Process	ϵ	SuNCW	NCW	CNCW	SuCNCW
500	I	0.55	0.988 (0.009)	0.988 (0.009)	0.988 (0.009)	0.988 (0.009)
	I	0.35	0.949 (0.017)	0.945 (0.037)	0.945 (0.037)	0.949 (0.018)
	II	0.55	0.921 (0.089)	0.915 (0.093)	0.939 (0.078)	0.954 (0.039)
	II	0.35	0.799 (0.119)	0.770 (0.123)	0.848 (0.112)	0.882 (0.077)
	III	0.55	0.848 (0.141)	0.838 (0.146)	0.911 (0.085)	0.923 (0.060)
	III	0.35	0.663 (0.108)	0.639 (0.095)	0.744 (0.131)	0.804 (0.108)
1000	I	0.55	0.988 (0.005)	0.988 (0.006)	0.988 (0.006)	0.988 (0.005)
	I	0.35	0.954 (0.010)	0.953 (0.010)	0.953 (0.010)	0.954 (0.010)
	II	0.55	0.938 (0.060)	0.935 (0.060)	0.962 (0.038)	0.966 (0.010)
	II	0.35	0.805 (0.127)	0.781 (0.123)	0.858 (0.119)	0.892 (0.079)
	III	0.55	0.855 (0.145)	0.850 (0.146)	0.930 (0.051)	0.938 (0.015)
	III	0.35	0.678 (0.122)	0.665 (0.110)	0.804 (0.116)	0.845 (0.080)

model with the highest IC value has been selected as the most adequate fitted model. Table 3 provides the frequency distribution of the models selected in this way by each IC for each data generation process and each value of ϵ and I. As expected, SuNCW models have almost always been selected as the most adequate for the analysis of uncontaminated datasets. With datasets containing atypical observations generated through processes II and III, best trade-off between fit and complexity is generally obtained by the fitted SuCNCW models. Such results hold true regardless of the level of separation and the information criterion employed to perform model selection.

3.2.5 Comparison among information criteria

Information on aspect *(v)* has been obtained by evaluating the number of times each value of K has been selected by each examined criterion. When the analysed datasets do not contain atypical observations and the level of separation between the second and third cluster is high (process I, $\epsilon = 0.55$), the presence of three clusters is (almost) always recognised by both information criteria regardless of the fitted model and the sample size (see the upper part of Tables K and L). If the level of separation is reduced ($\epsilon = 0.35$), the ability of the BIC to correctly detect the presence of three clusters remains good regardless of the fitted model only with the largest sample size. When $I = 500$, the true order of the generated datasets is slightly underestimated by the BIC when CNCW models are employed. With datasets generated using I, the ICL shows a clear preference for $K = 3$ components regardless of the model type with both values of ϵ but only when the sample size is $I = 1000$. Otherwise, the true value of K is almost always properly estimated by this criterion as long as models embedding the information on the relevant regressors are fitted (e.g., SuNCW and SuCNCW). With the other two examined models, the true number of clusters appears to be underestimated, and this is especially true of CNCW models.

Table 3 Trade-off between fit and compexity: number of selections over 100 samples of SuNCW, NCW, CNCW and SuCNCW models, based on the highest BIC and ICL.

I	IC	Process	ϵ	SuNCW	NCW	CNCW	SuCNCW
500	BIC	I	0.55	100	0	0	0
		I	0.35	100	0	0	0
		II	0.55	0	1	1	98
		II	0.35	1	2	2	95
		III	0.55	2	1	1	96
		III	0.35	0	0	8	92
	ICL	I	0.55	100	0	0	0
		I	0.35	99	1	0	0
		II	0.55	0	1	1	98
		II	0.35	1	2	2	95
		III	0.55	2	2	1	95
		III	0.35	2	0	6	92
1000	BIC	I	0.55	100	0	0	0
		I	0.35	100	0	0	0
		II	0.55	0	0	0	100
		II	0.35	3	1	4	92
		III	0.55	0	0	1	99
		III	0.35	0	0	8	92
	ICL	I	0.55	100	0	0	0
		I	0.35	99	1	0	0
		II	0.55	0	0	0	100
		II	0.35	4	1	6	89
		III	0.55	0	0	1	99
		III	0.35	0	0	6	94

Under processes II and III, when SuNCW and NCW models are fitted to the data, the BIC shows a tendency to select $K = 4$ (outliers are typically accommodated using an additional cluster), regardless of the level of separation (see Mazza and Punzo, 2020). This tendency appears to be more evident when the sample size is large. The ICL shows the same behaviour in association with SuNCW models (for both levels of separation) and NCW models (for $\epsilon = 0.55$). On the contrary, with SuCNCW models, BIC and ICL correctly identify three clusters for the majority of the simulated datasets, regardless of the sample size and the degree of separation. When these two criteria are employed in the selection of CNCW models, the true value of K is properly estimated provided that the degree of separation is high or the sample size is large; otherwise, the order of CNCW models is generally underestimated.

3.2.6 Classification recovery (without exploiting the knowledge of K)

In order to study aspect *(vi)*, for each generated dataset the ARI index has been computed between the partitions of the sample units detected by the fitted models showing the highest BIC value under each competing model class and the true partition. In general, the resulting average values of the ARI index (see Table 4) are quite similar to the ones obtained by exploiting the knowledge of the true K (see Table 2). Obviously, whenever the value of K

Table 4 Classification recovery of the fitted SuNCW, NCW, CNCW and SuCNCW models with the highest BIC: average values (standard deviations) of the ARI index over 100 samples.

I	Process	ϵ	SuNCW	NCW	CNCW	SuCNCW
500	I	0.55	0.987 (0.012)	0.987 (0.011)	0.987 (0.011)	0.987 (0.012)
	I	0.35	0.945 (0.035)	0.932 (0.068)	0.891 (0.119)	0.940 (0.055)
	II	0.55	0.891 (0.048)	0.890 (0.059)	0.952 (0.038)	0.957 (0.019)
	II	0.35	0.805 (0.091)	0.767 (0.111)	0.791 (0.134)	0.886 (0.068)
	III	0.55	0.870 (0.071)	0.868 (0.078)	0.921 (0.050)	0.923 (0.043)
	III	0.35	0.749 (0.102)	0.682 (0.108)	0.711 (0.123)	0.803 (0.106)
1000	I	0.55	0.988 (0.005)	0.988 (0.006)	0.988 (0.006)	0.988 (0.005)
	I	0.35	0.952 (0.015)	0.951 (0.015)	0.951 (0.146)	0.952 (0.015)
	II	0.55	0.885 (0.034)	0.885 (0.034)	0.965 (0.011)	0.966 (0.010)
	II	0.35	0.827 (0.077)	0.819 (0.085)	0.861 (0.109)	0.898 (0.061)
	III	0.55	0.880 (0.021)	0.879 (0.023)	0.934 (0.037)	0.938 (0.016)
	III	0.35	0.769 (0.102)	0.737 (0.122)	0.834 (0.083)	0.865 (0.032)

determined according to the BIC is equal to the true K, the ability to recover the true classification coincides with the one evaluated in Section 3.2.3. Thus, in general, using the BIC to estimate the value of K seems to have a negligible impact on the classification recovery of SuCNCW models. The impact on the performance of the other three model types is more evident, especially for SuNCW and NCW models. More specifically, in the presence of datasets with contaminated observations generated according to II and III, SuNCW and NCW models of order $\hat{K}_{\text{BIC}}$ show a slight increase in the ability to estimate the true classification of the sample observations in comparison with the same models of order 3. A possible explanation of this behaviour could be related to the fact that such models are not able to properly account for contaminated observations; thus, according to the BIC, SuNCW and NCW models of order 4 should be preferred.

3.2.7 Ability to detect regression outliers, good and bad leverage points

For the evaluation of aspect *(vii)*, comparisons between CNCW models and SuCNCW models are based on both the proportion of atypical observations that are correctly classified as atypical (true positive rate, TPR) and the proportion of typical points incorrectly classified as atypical (false positive rate, FPR). Such rates have been computed using the intra-cluster classifications calculated using $\hat{v}_{ik}$ and $\hat{u}_{ik}$ (see Section 2.4). Table 5 reports these rates for the data generation processes IV–VI.

In terms of FPRs, both types of models show almost optimal results with any data generation process for each examined level of the two experimental factors ($0.000 \leq \text{FPR} \leq 0.010$). Thus, under the examined circumstances it is extremely rare to classify typical observations as regression outliers, good or bad leverage points by using SuCNCW and CNCW models. As far as TPRs are concerned, optimal results are associated with the processes IV and V.

Table 5 Values of TPRs and FPRs (rates across 100 replications) for processes IV–VI.

			$I = 500$		$I = 1000$	
Process	ϵ	Model	TPR	FPR	TPR	FPR
IV	0.55	SuCNCW	1.000	0.035	0.997	0.010
		CNCW	1.000	0.035	0.997	0.010
	0.35	SuCNCW	1.000	0.020	0.997	0.009
		CNCW	0.990	0.020	0.997	0.009
V	0.55	SuCNCW	1.000	0.051	1.000	0.007
		CNCW	1.000	0.051	1.000	0.007
	0.35	SuCNCW	1.000	0.035	1.000	0.007
		CNCW	1.000	0.035	1.000	0.007
VI	0.55	SuCNCW	0.429	0.000	0.505	0.001
		CNCW	0.418	0.000	0.427	0.001
	0.35	SuCNCW	0.524	0.001	0.604	0.000
		CNCW	0.488	0.001	0.517	0.000

This means that SuCNCW and CNCW models appear to be equally capable of detecting regression outliers and good leverage points. In terms of ability to correctly classify bad leverage points, a slightly better performance is registered with SuCNCW models for each examined level of the two experimental factors. However, it is also worth noting that, in general, both models only partly succeed in recognising bad leverage points ($0.429 \leq \text{TPR} \leq 0.604$ using SuCNCW models; $0.418 \leq \text{TPR} \leq 0.517$ with CNCW models); greater difficulties are registered with small samples and high separation between the second and third cluster.

4 Analysis of regional tourism in Italy

The aim of the analysis summarised here is to evaluate the link between tourism flows and attendance at museums and monuments in two Italian regions: Emilia-Romagna (`ER`) and Veneto (`Ve`). To this end, data concerning tourist arrivals and tourist overnights have been downloaded from the websites of Emilia-Romagna[1] and Veneto[2] regional governments; the source of the data on visits to State museums, monuments and museum networks is the website of the Italian Ministry of Cultural Heritage[3]. The resulting dataset on regional tourism in Italy (from now on, `rtI` dataset) is composed of $I = 276$ monthly observations (from January 1999 to December 2021) for the following variables: tourist arrivals (denoted `Arriv`, in thousands), tourist average stays (`AvStay`, computed as the ratio between tourist overnights and tourist arrivals) and visits to State museums, monuments and museum networks (`Visit`) in Emilia-Romagna and Veneto. Because of the goal of the analysis, these variables have been partitioned as follows: $\mathbf{Y} = (Y_1 =$ `Visit ER`, $Y_2 =$ `Visit Ve`$)'$, $\mathbf{X} = (X_1 =$`Arriv ER`, $X_2 =$ `Arriv Ve`, $X_3 =$`AvStay ER`, $X_4 =$ `AvStay`

[1] `https://statistica.regione.emilia-romagna.it/turismo`

[2] `https://www.veneto.eu/web/area-operatori/statistiche`

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

Ve)′. Thus, $M = 2$ and $P = 4$. Figures 1 and 2 show the bivariate scatterplots for variables $(X_p, X_{p'})$, $p, p' = 1, \ldots, P$, $p \neq p'$ and (Y_m, X_p), $m = 1, \ldots, M$, $p = 1, \ldots, P$, respectively; month abbreviations are used as labels for the observations. It is worth noting that between 9 March 2020 and 3 May 2020 a national lockdown was imposed in Italy because of the Covid-19 pandemic. Containment measures on the second wave of the pandemic were adopted by the Italian Parliament between 13 October 2020 and 31 January 2021. Such measures had a dramatic impact on attendance at museums and monuments (e.g., no visits were registered in April 2020, December 2020 and January 2021) and tourism flows (e.g., in April 2020, tourists stayed on average 15.4 days in Emilia-Romagna, 19.8 days in Veneto). A preliminary evaluation of the pairwise linear dependencies for such variables has been carried out (see Table 6). Visits to State museums, monuments and museum networks in the two regions result to be highly linearly dependent; high and positive pairwise correlations also characterise tourist arrivals and average stays in either region; low negative correlations are observed between `Visit` and `AvStay` for both regions.

Since Emilia-Romagna and Veneto are neighbouring regions, tourist arrivals and average stays in one region could have an impact on the visits to State museums and monuments of the other region. Thus, in order to properly select the vectors of regressors $\mathbf{X}_1$ and $\mathbf{X}_2$ for the prediction of `Visit ER` and `Visit Ve`, the analysis of the `rtI` dataset has been carried out by fitting contaminated and uncontaminated NCW models of order K, for $K = 1, \ldots, 9$, in which, for each K, $\mathbf{X}_m$ has been allowed to vary within $\mathcal{R}$ for $m = 1, 2$. To this end, a genetic algorithm has been implemented in `R` through the package `GA` (Scrucca, 2013). More specifically, a binary encoding has been employed to distinguish between excluded and included covariates. It is worth noting that, when the algorithm generates a chromosome with the same sequence of P genes (i.e., the same selection of covariates) for the two responses, either an NCW model or a CNCW model is specified. Two additional integer codes have been inserted in each chromosome so that the stochastic search has also included the parameterisations of the covariance matrices $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Xi}_k$, for $k = 1, \ldots, K$ (see Table 1). As far as the fitness measure is concerned, the BIC has been employed. For each model class and each value of K, 12 independent executions of such a genetic algorithm have been performed, one for each combination of the following values for the tuning parameters: $N = 200, 300, 400, 500$; $d_{\max} = 30, 40, 50$. For each type of model, the total number of models that have been examined is about 400000.

Within each of the two types of model classes, the best trade-off between the fit and the model complexity is reached by models of order $K = 5$. More detailed information about these models can be found in Table 7. The overall best trade-off is reached by a SuCNCW model in which the two sub-vectors of $\mathbf{X}$ employed to define the design matrix are $\mathbf{X}_1$ = (`Arriv.ER`, `Arriv.Ve`, `AvStay.ER`) and $\mathbf{X}_2$ =(`Arriv.Ve`, `AvStay.ER`, `AvStay.Ve`). The distributions of the four regressors in the five clusters of months determined by this latter model according to the rule of the maximum a posteriori probability are

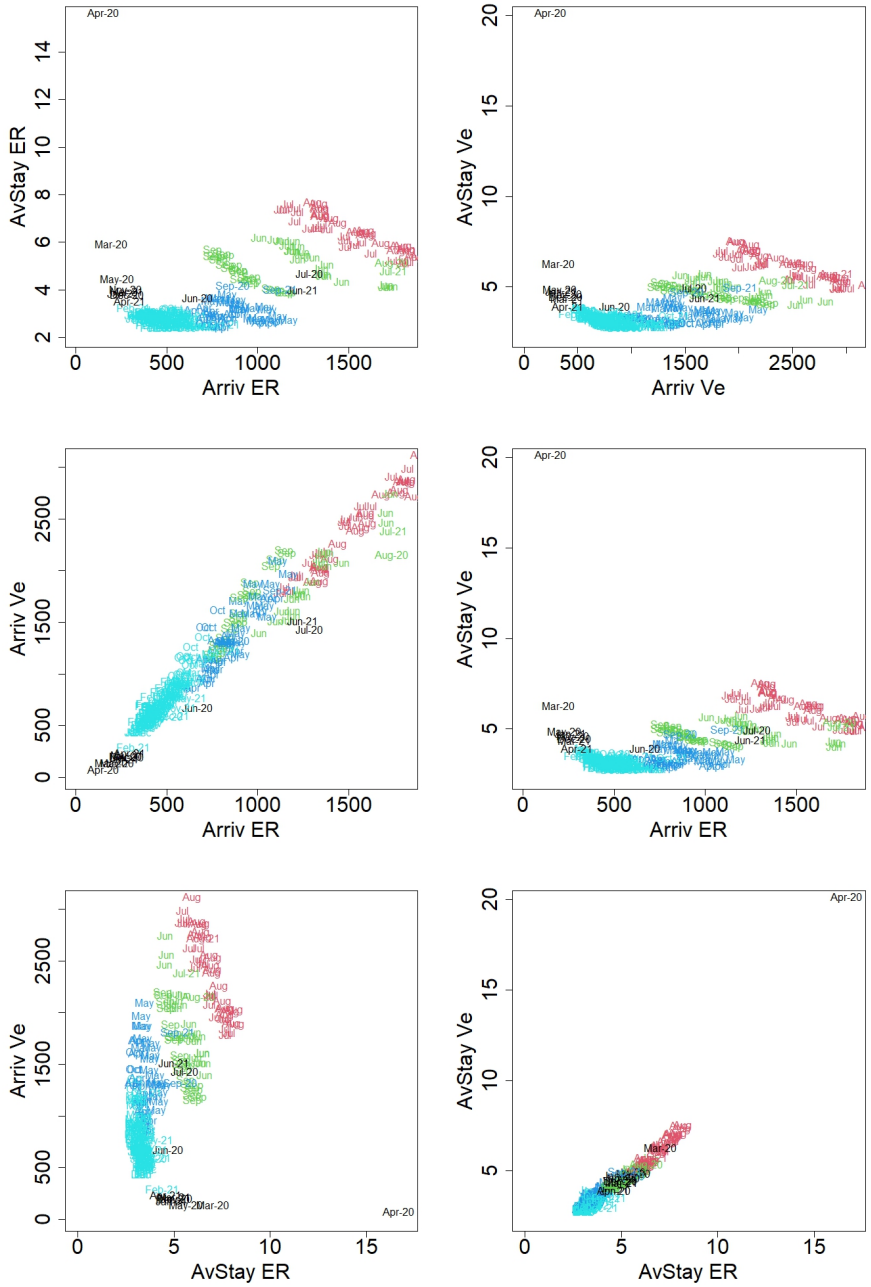

Fig. 1 Scatterplots for pairs of regressors in the analysis of the `rtI` dataset. Month abbreviations are used as labels. Observations are coloured according to the classification obtained from the best fitted model.

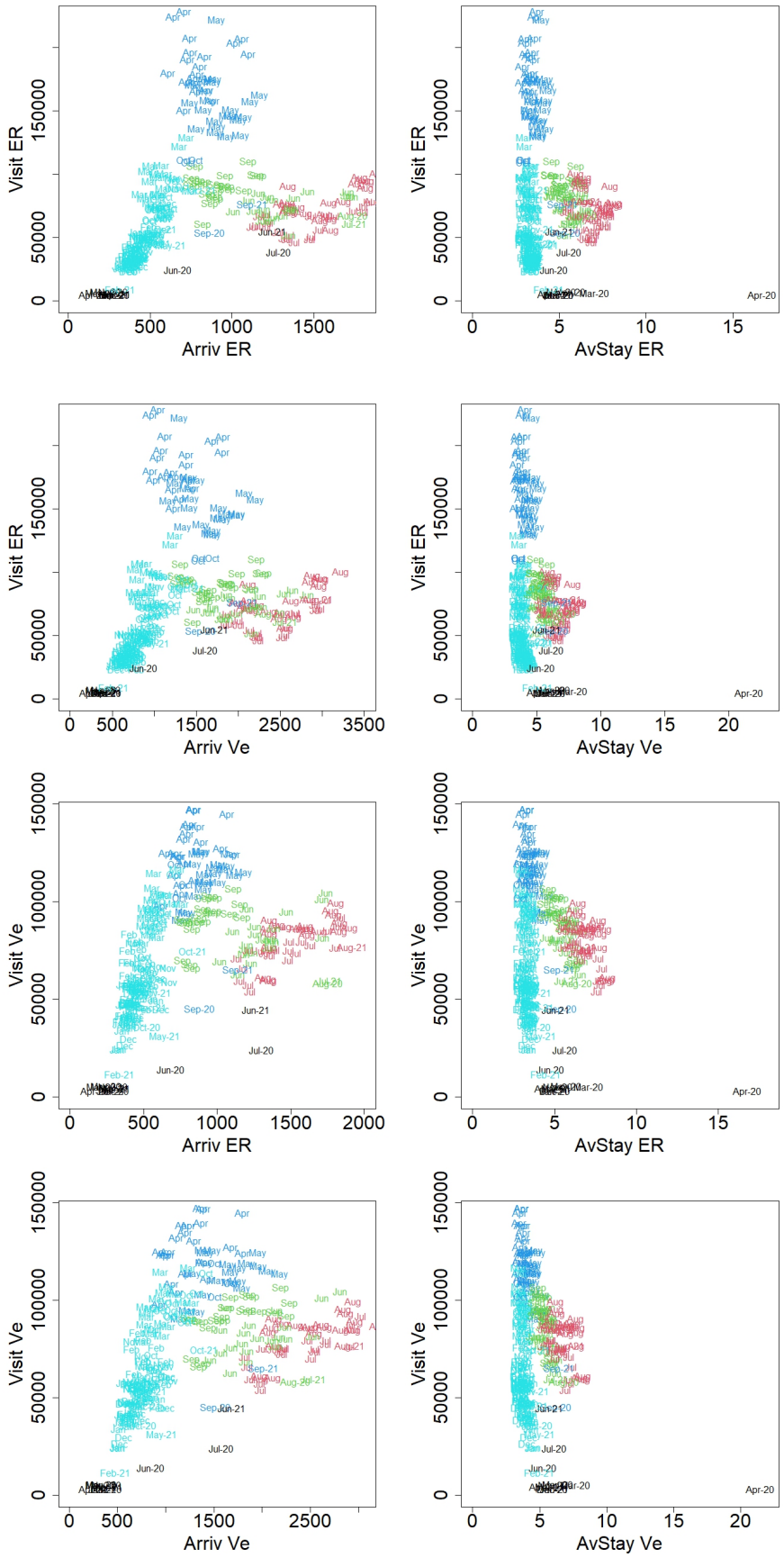

Fig. 2 Scatterplots for pairs (response, regressor) in the analysis of the `rtI` dataset. Month abbreviations are used as labels. Observations are coloured according to the classification obtained from the best fitted model.

Table 6 Pearson's correlation matrix (lower diagonal part) and p-values of the Student's t test (upper diagonal part) for the hypotheses of pairwise linear independence between the six variables from the `rtI` dataset.

	`Visit.ER`	`Visit.Ve`	`Arriv.ER`	`Arriv.Ve`	`AvStay.ER`	`AvStay.Ve`
`Visit.ER`	1.0000	$< 10^{-79}$	$< 10^{-6}$	$< 10^{-6}$	0.0676	0.0322
`Visit.Ve`	0.8562	1.0000	$< 10^{-12}$	$< 10^{-16}$	0.3288	0.0386
`Arriv.ER`	0.2913	0.4128	1.0000	$< 10^{-194}$	$< 10^{-29}$	$< 10^{-18}$
`Arriv.Ve`	0.3097	0.4653	0.9804	1.0000	$< 10^{-24}$	$< 10^{-13}$
`AvStay.ER`	−0.1102	−0.0590	0.6137	0.5744	1.0000	$< 10^{-153}$
`AvStay.Ve`	−0.1290	−0.1246	0.4759	0.4409	0.9607	1.0000

Table 7 Information on the best fitted models using the genetic algorithm from the contaminated (C) and uncontaminated (U) model classes in the analysis of the `rtI` dataset.

Model	K	acr.X	acr.Y	$\mathbf{X}_1$	$\mathbf{X}_2$	$\ell(\hat{\boldsymbol{\psi}})$	$n_{\boldsymbol{\psi}}$	BIC
C	5	`EVV`	`VEE`	(X_1, X_2, X_3)	(X_2, X_3, X_4)	−9571.7	137	−19913.4
U	5	`VVV`	`VEV`	(X_1, X_3)	(X_2, X_3)	−9653.8	115	−19953.9

ellipsoidal with variable shapes and orientations and equal volume; as far as the joint conditional distributions of the two responses given the regressors are concerned, clusters are characterized by equal shapes and orientations and variable volumes. The obtained estimates of $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and $\boldsymbol{\beta}^*$ are reported in Table 8. Table 9 shows the sizes of the five clusters of months together with the within-cluster sizes of the four categories detected by the values of $\hat{u}_{ik}$ and $\hat{v}_{ik}$ (typical observations: $\hat{u}_{ik} \geq 0.5$ and $\hat{v}_{ik} \geq 0.5$; regression outliers: $\hat{u}_{ik} < 0.5$ and $\hat{v}_{ik} \geq 0.5$; good leverage points: $\hat{u}_{ik} \geq 0.5$ and $\hat{v}_{ik} < 0.5$; bad leverage points: $\hat{u}_{ik} < 0.5$ and $\hat{v}_{ik} < 0.5$). From an overall inspection of the parameter estimates it emerges that both the joint distribution of tourist arrivals and average stays and the conditional distribution of visits to museums and monuments of the two regions are affected by a source of unobserved heterogeneity over time. Furthermore, the values of $\hat{\alpha}_k$, $\hat{\eta}_k$, $\hat{\tau}_k$ and $\hat{\lambda}_k$, for $k = 1, \ldots, 5$, suggest that the analysed dataset is also contaminated by the presence of leverage points and regression outliers. As far as the obtained cluster structure is concerned, it clearly reflects both seasonal patterns characterising tourism flows and the effects of the Covid-19 pandemic on tourism in Italy (see Table 10).

The first cluster detected by the best fitted model contains most of the months after the Italian national lockdown. More specifically, they are March – July 2020, November 2020 – January 2021, March – April 2021 and June 2021; they are depicted in black in the scatterplots of Figures 1 and 2. Eight of these months can be considered typical (see the first row of Table 9). April 2020 has been classified as a good leverage point. As the scatterplot on the left in the upper panel of Figure A shows (see supplementary material), July 2020 and June 2021 are bad leverage points (see also the estimated squared Mahalanobis

Table 8 Estimated $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$, $\boldsymbol{\tau}$, $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and $\boldsymbol{\beta}^*$ of the overall best fitted model for the `rtI` dataset.

k	1	2	3	4	5
$\hat{\pi}_k$	0.040	0.156	0.159	0.169	0.476
$\hat{\alpha}_k$	0.727	0.958	0.500	0.500	0.991
$\hat{\eta}_k$	60.846	4.368	3.973	2.060	1.001
$\hat{\tau}_k$	0.811	0.990	0.983	0.942	0.627
$\hat{\lambda}_k$	159.767	1.001	1.001	3.166	12.522
$\hat{\mu}_{k1}$	169.826	1424.830	993.681	782.446	399.909
$\hat{\mu}_{k2}$	188.254	2309.745	1664.209	1301.921	676.582
$\hat{\mu}_{k3}$	3.897	6.188	4.739	2.801	2.541
$\hat{\mu}_{k4}$	4.320	5.785	4.450	3.122	2.840
$\hat{\beta}_{k10}$	−4212.878	−267481.600	160643.063	244965.197	−44834.551
$\hat{\beta}_{k11}$	−41.167	134.701	−18.850	228.213	134.282
$\hat{\beta}_{k12}$	72.611	−8.944	−5.526	−138.167	26.713
$\hat{\beta}_{k13}$	278.338	26093.850	−11362.030	−29971.231	7403.148
$\hat{\beta}_{k20}$	−3476.399	8312.470	199205.040	166642.784	−78867.123
$\hat{\beta}_{k21}$	19.243	22.919	−12.065	13.813	95.057
$\hat{\beta}_{k22}$	3224.954	−11628.520	−6864.809	45146.485	40638.694
$\hat{\beta}_{k23}$	−2334.827	14889.680	−14264.491	−63002.821	−11195.437

Table 9 Sizes of the five clusters of months from the `rtI` dataset detected by the overall best model and their within-cluster distributions into four categories, based on $\hat{u}_{ik}$ and $\hat{v}_{ik}$.

Cluster k	typical	outlier	good leverage	bad leverage	cluster size
1	8	0	1	2	11
2	42	0	1	0	43
3	25	0	19	0	44
4	25	0	21	1	47
5	92	39	0	0	131

Table 10 Cross-classification of the observations from the `rtI` dataset, based on their variable time identified by month and maximum posterior probability estimated from the best fitted model.

k	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	1	0	2	2	1	2	1	0	0	0	1	1
2	0	0	0	0	0	0	21	22	0	0	0	0
3	0	0	0	0	0	21	1	1	21	0	0	0
4	0	0	0	21	21	0	0	0	2	3	0	0
5	22	23	21	0	1	0	0	0	0	20	22	22

distances $\hat{d}^2_{i1_{\mathbf{y}}}$ in the first column of Table 11). Due to the lockdown and the containment measures on the second wave of the Covid-19 pandemic, this is the cluster with the lowest estimated mean values of tourist arrivals in both regions (see the first column in Table 8). For the same reasons, the estimated mean values of `AvStay ER` and `AvStay Ve` for this cluster are slightly higher than the overall mean values, which are 3.60 days in Emilia-Romagna and 3.72

Table 11 Minimum and maximum values of the estimated squared Mahalanobis distances $\hat{d}^2_{ik\mathbf{y}}$ within the five clusters of months for the `rtI` dataset, by the categories: typical observations/outliers.

k	1	2	3	4	5
Typical observations					
$\min\{\hat{d}^2_{ik\mathbf{y}}\}$	0.01	0.00	0.04	0.07	0.02
$\max\{\hat{d}^2_{ik\mathbf{y}}\}$	7.20	7.59	7.80	7.05	6.06
Outliers					
$\min\{\hat{d}^2_{ik\mathbf{y}}\}$	189.81	-	-	14.70	6.86
$\max\{\hat{d}^2_{ik\mathbf{y}}\}$	475.80	-	-	14.70	80.17

days in Veneto. This cluster is also characterised by positive estimated effects of tourist arrivals in Veneto and average stays in Emilia-Romagna on visits to museums and monuments in both regions.

The second cluster detected by the best fitted model contains 43 observations, which refer to July (years 1999-2019) and August (years 1999-2019 and 2021). They are depicted in red in the scatterplots of Figures 1 and 2. August 2021 has been classified as a good leverage point (see the second row of Table 9). As far as the possible presence of regression outliers in the distribution of $\mathbf{Y}|(\mathbf{X} = \mathbf{x}, G_2)$ is concerned, all the months belonging to this cluster have been identified as typical. This latter result is also evident from the estimated sample residuals $\mathbf{y}_i - \tilde{\mathbf{x}}_i^* \hat{\boldsymbol{\beta}}_2^*$ for the months of this cluster (see the scatterplot on the right in the upper panel of Figure A). A further proof is given by low values of $\hat{d}^2_{i2\mathbf{y}}$ (see the second column in Table 11). The months which belong to this cluster are also characterized by the highest mean values of tourist arrivals and average stays in both regions; furthermore, the estimated effects of tourist arrivals and average stays in either region on visits to museums and monuments in the same region result to be positive (see the second column of Table 8).

The 44 observations belonging to the third cluster refer to June (years 1999-2019), September (years 1999-2019), August 2020 and July 2021 (green-coloured months in Figures 1 and 2). Almost half of these months have been classified as good leverage points (see the third row of Table 9). Similarly to the previous cluster, the estimated sample residuals $\mathbf{y}_i - \tilde{\mathbf{x}}_i^* \hat{\boldsymbol{\beta}}_3^*$ (see the scatterplot on the left in the central panel of Figure A) and the low values of $\hat{d}^2_{i3\mathbf{y}}$ (see the third column in Table 11) prove the absence of outlying months. Observations belonging to this cluster are characterized by high mean values of tourist arrivals and average stays in both regions; as far as the regression coefficients are concerned, an interesting finding is that all the estimated effects of the covariates appear to be negative (third column of Table 8).

Cluster 4 comprises the 47 dark blue observations depicted in Figures 1 and 2, which refer to April (years 1999-2019), May (years 1999-2019), October (years 2017-2019) and September (years 2020-2021). 21 of these months have

been identified as good leverage points. As the scatterplot on the right in the central panel of Figure A shows, September 2020 represents a bad leverage point; the estimate of the squared Mahalanobis distance $\hat{d}^2_{i4_{\mathbf{y}}}$ of this point is 14.70 (see Table 11). The estimated mean values of tourist arrivals for this cluster in both regions are slightly higher than the overall mean values, which are 728.25 in Emilia-Romagna and 1197.09 in Veneto); the opposite result holds true for the estimated mean values of `AvStay ER` and `AvStay.Ve`. Furthermore, this cluster is mainly characterised by the highest positive effect of tourist arrivals in Emilia-Romagna on the number of visits in the same region and the highest positive effect of average stays in Emilia-Romagna on the number of visits in Veneto.

As far as cluster 5 is concerned, it contains 131 months (see the sky-blue labels in Figures 1 and 2): January (years 1999-2020), February (all the examined years), March (years 1999-2019), May 2021, October (years 1999-2016, 2020-2021), November (all the examined years except 2020) and December (all the examined years except 2020). It is characterised by the absence of leverage points. However, 39 months of this cluster have been classified as mild outliers (see the red triangles in the scatterplot in the bottom panel of Figure A). It is worth noting that, for these 39 months, the estimated squared Mahalanobis squared distances $\hat{d}^2_{i5_{\mathbf{y}}}$ are all larger than those computed for the other weeks of the same cluster (see the fifth column in Table 11). A distinctive feature of cluster 5 is the lowest estimated mean value of `AvStay` in both regions. This cluster is also characterised by the highest positive effect of tourist arrivals in Veneto on the number of visits in the same region and a strong positive effect of average stays in Emilia-Romagna on the number of visits in Veneto.

5 Conclusions

The SuCNCW models introduced here perform robust clustering in multivariate linear regression analysis with correlated responses and random regressors for datasets characterised by unobserved heterogeneity and mildly atypical observations. They can also be employed to identify outliers, and good and bad leverage points within each detected cluster. Compared to the models introduced in Punzo and McNicholas (2017), the main novelty of these models is that a different vector of regressors is considered for each response. Thanks to this feature, the data analyst is enabled to convey prior information concerning the absence of certain regressors from the linear term employed in the prediction of a certain response in any application in which different regressors are expected to be relevant in the prediction of different responses. SuCNCW models with a reduced number of variance-covariance parameters have also been specified; they can be more effectively employed when the analysis involves either many responses or many regressors. Furthermore, since SuCNCW models encompass other normal mixture-based linear regression models with random regressors (Dang et al., 2017; Punzo and McNicholas, 2017; Diani et

al., 2022), they represent a flexible approach for simultaneous robust clustering and detection of mildly atypical observations in linear regression analysis.

Monte Carlo studies have shown that either the inclusion of irrelevant regressors in a cluster-weighted model or the presence of mildly atypical observations in the data can negatively affect the reconstruction of both the true classification and true parameter values, especially when the clusters of observations are not well-separated. Furthermore, they can have a negative impact on the choice of the order K of a CW model. The obtained results have demonstrated that such difficulties can be managed by resorting to SuCNCW models. In practical applications in which the regressors to be considered in the linear predictor of each response have to be determined from the data, an approach based on a joint use of SuCNCW models and techniques for variable selection (e.g., genetic algorithms, stepwise strategies) can also allow to identify the relevant predictors for each regression equation.

A disadvantage of using an ECM algorithm to perform ML estimation of SuCNCW models is that it does not provide a direct assessment of the sample variability of the ML estimates. To this end, approaches commonly employed under finite mixture models could be employed (see, e.g., McLachlan and Peel, 2000). Some of them rely on the gradient vector, the second-order derivative matrix of the incomplete data log-likelihood, or a robust sandwich-type estimation procedure. Solutions have been developed for normal mixture models (Boldea and Magnus, 2009), t mixture models (Wang and Lin, 2016), clusterwise normal regression models (Galimberti et al., 2021) and normal cluster-weighted models (Soffritti, 2021).

Another disadvantage of the methods illustrated in this paper is that they produce parameter estimates which are expected to be sensitive to grossly atypical observations. More specifically, even a single extreme observation can have harmful effects, thereby leading to the breakdown of at least one component estimate. Embedding an approach based on trimming in the estimation of a SuCNCW model (see, e.g., García-Escudero et al., 2017; Farcomeni and Punzo, 2020) can manage this issue. A further avenue of future research could be the development of SuCNCW models accounting for censored and missing responses (see, e.g., Lin and Wang, 2022, 2023).

Data availability and other statements

- The sources of the real-world data supporting the findings of the study reported in Section 4 are the websites of Emilia-Romagna[4] and Veneto[5] regional governments (tourist arrivals and overnights) and the website of the Italian Ministry of Cultural Heritage[6] (visits to State museums, monuments and museum networks).
- The authors have no relevant interests to disclose.

[4] `https://statistica.regione.emilia-romagna.it/turismo`

[5] `https://www.veneto.eu/web/area-operatori/statistiche`

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

– The authors have no conflicts of interest to disclose.
– The authors are grateful to three anonymous reviewers for their constructive comments and valuable suggestions.

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

Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R (2010) Combining mixture components for clustering. J Comput Graph Stat 19(2): 332–353

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

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

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

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

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

Cappozzo A, García-Escudero LA, Greselin F, Mayo-Iscar A (2021) Parameter choice, stability and validity for robust cluster weighted modeling. Stats 4: 602–615

Cappozzo A, García-Escudero LA, Greselin F, Mayo-Iscar A (2023) Graphical and computational tools to guide parameter choice for the cluster weighted robust model. J Comput Graph Stat https://doi.org/10.1080/10618600.2022.2154218

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

Chatterjee S, Laudato M, Lynch LA (1996) Genetic algorithms and their statistical applications: an introduction. Comput Stat Data Anal 22: 633–651

Cuesta-Albertos JA, Gordaliza A, Matran C (1997) Trimmed k means: an attempt to robustify quantizers. Ann Stat 25(2): 553–576

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

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

Diani C, Galimberti G, Soffritti G (2022) Multivariate cluster-weighted models based on seemingly unrelated linear regression. Comput Stat Data Anal 171: 107451

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

Farcomeni A, Punzo A (2020) Robust model-based clustering with mild and gross outliers. Test 29: 989–1007

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

Gallaugher MPB, Tomarchio SD, McNicholas PD, Punzo A (2022) Multivariate cluster weighted models using skewed distributions. Adv Data Anal Classif 16: 93–124

Galimberti G, Manisi A, Soffritti G (2018) Modelling the role of variables in model-based cluster analysis. Stat Comput 28(1): 145–169

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

García-Escudero LA, Gordaliza A, Greselin F, Ingrassia S, Mayo-Iscar A (2017) Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. Stat Comput 27: 377–402

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

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

Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading

Hastie, Tibshirani, Friedman (2009) The elements of statistical learning. Second edition. Springer, New York

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

Hennig C (2004) Breakdown points for maximum likelihood estimators of location-scale mixtures. Ann Stat 32: 1313–1340

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

Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1): 193–218
Ingrassia S, Minotti SC, Vittadini G (2012) Local statistical modeling via a cluster-weighted approach with elliptical distributions. J Classif 29(3): 363–401
Ingrassia S, Minotti SC, Punzo A (2014) Model-based clustering via linear cluster-weighted models. Comput Stat Data Anal 71: 159–182
Karlis D, Xekalaki E (2003) Choosing initial values for the EM algorithm for finite mixtures. Comput Stat Data Anal 41(3–4): 577–590
Lin T-I, Wang W-L (2022) Multivariate linear mixed models with censored and nonignorable missing outcomes, with application to AIDS studies. Biom J 64, 1325–1339
Lin T-I, Wang W-L (2023) Flexible modeling of multiple nonlinear longitudinal trajectories with censored and non-ignorable missing outcomes. Stat Methods Med Res 32(3): 593–608
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 Papers 61(2): 787–822
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
Miller AJ (1991) Subset selection in regression (2nd ed.). Chapman and Hall, Boca Raton
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
Perrone G, Soffritti G (2023) Seemingly unrelated clusterwise linear regression for contaminated data. Stat Papers 64: 883–921
Punzo A, McNicholas PD (2017) Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. J Classif 34(2): 249–293
Punzo A, Mazza A, McNicholas, PD (2018) `ContaminatedMixt`: an `R` package for fitting parsimonious mixtures of multivariate contaminated normal distributions. J Stat Softw 85(10): 1–25
R Core Team (2022) `R`: a language and environment for statistical computing. `R` Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
Ritter G (2015) Robust cluster analysis and variable selection. Chapman and Hall, Boca Raton
Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3): 212–223
Rousseeuw PJ, Leroy AM (2005) Robust regression and outlier detection. Wiley, New York

Ruwet C, García-Escudero LA, Gordaliza A, Mayo-Iscar A (2013) On the breakdown behavior of the tclust clustering procedure. Test 22(3): 466–487

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

Scrucca L (2013) `GA`: a package for genetic algorithms in `R`. J Stat Softw 53(4): 1–37

Scrucca L (2016) Genetic algorithms for subset selection in model-based clustering. In: Celebi ME, Aydin K (eds) Unsupervised learning algorithms. Springer, Berlin, pp 55–70

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

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

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

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, California, pp 448–485

Wang W-L, Lin T-I (2016) Maximum likelihood inference for the multivariate t mixture model. J Multivar Anal 149: 54–64

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

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