Journal of Classification (2021) 38:594–625
https://doi.org/10.1007/s00357-021-09390-9

Estimating the Covariance Matrix of the Maximum Likelihood Estimator Under Linear Cluster-Weighted Models

Gabriele Soffritti[1]

Accepted: 14 May 2021 / Published online: 31 July 2021

Abstract
In recent years, the research into cluster-weighted models has been intense. However, estimating the covariance matrix of the maximum likelihood estimator under a cluster-weighted model is still an open issue. Here, an approach is developed in which information-based estimators of such a covariance matrix are obtained from the incomplete data log-likelihood of the multivariate Gaussian linear cluster-weighted model. To this end, analytical expressions for the score vector and Hessian matrix are provided. Three estimators of the asymptotic covariance matrix of the maximum likelihood estimator, based on the score vector and Hessian matrix, are introduced. The performances of these estimators are numerically evaluated using simulated datasets in comparison with a bootstrap-based estimator; their usefulness is illustrated through a study aiming at evaluating the link between tourism flows and attendance at museums and monuments in two Italian regions.

Keywords Gaussian mixture model · Hessian matrix · Linear regression · Model-based cluster analysis · Sandwich estimator · Score vector

1 Introduction

Cluster-weighted models constitute an approach to regression analysis with random covariates in which supervised (regression) and unsupervised (model-based cluster analysis) learning methods are jointly exploited (Hastie et al., 2009). In this approach, a given random vector is assumed to be composed of an outcome $\mathbf{Y}$ (response, dependent variable) and its explanatory variables $\mathbf{X}$ (covariates, predictors). Furthermore, sample observations are allowed to come from a population composed of several unknown sub-populations. Finally, the joint distribution of the outcome and the covariates is modelled through a finite mixture model specified so as to account for a different effect of the covariates on the response within each sub-population. Thus, cluster-weighted models are useful to perform

✉ Gabriele Soffritti
gabriele.soffritti@unibo.it

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

model-based cluster analysis in a multivariate regression setting with random covariates and unobserved heterogeneity.

Since the introduction of this approach (Gershenfeld, 1997), the research into cluster-weighted models has been intense, especially in the last 8 years. Models for continuous variables under normal mixture models have been proposed by Ingrassia et al. (2012) and Dang et al. (2017). Robustified solutions have been developed by Ingrassia et al. (2014) and Punzo and McNicholas (2017), based on the use of Student t and contaminated normal mixture distributions, respectively. Punzo and Ingrassia (2013), Punzo and Ingrassia (2016), Ingrassia et al. (2015) and Di Mari et al. (2020) have introduced models for various types of responses. Models able to deal with non-linear relationships or many covariates have been proposed by Punzo (2014), Subedi et al. (2013) and Subedi et al. (2015).

By focusing the attention on Gaussian cluster-weighted models in which the effects of the covariates on the response within each sub-population are linear, model parameters are generally estimated through the maximum likelihood (ML) method by resorting to the expectation-maximisation (EM) algorithm (Dempster et al., 1977), which is widely employed in incomplete-data problems. In these models, the observed data $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, I\}$ are incomplete because the specific component density that generates the I sample observations is missing. This missing information is modelled through an unobserved variable coming from a pre-specified multinomial distribution and is added to the observed data so as to form the so-called complete data. Then, the ML estimate is computed from the maximisation of the complete data log-likelihood. A description of the EM algorithm for the linear Gaussian cluster-weighted model can be found in Dang et al. (2017). Specific functions implementing such algorithm for models with a univariate response are available in the package `flexCWM` (Mazza et al., 2018) for the `R` software environment (R Core Team, 2020).

A by-product of this estimating approach is a set of K estimated posterior probabilities that each sample observation comes from the K Gaussian distributions of the mixture. Thus, a by-product of fitting a linear Gaussian cluster-weighted model is a clustering of the I sample observations, based on a rule that assigns an observation to the distribution of the mixture to which it has the highest posterior probability of belonging. However, an estimating approach based on the use of an EM algorithm has the drawback of not automatically producing any estimate of the covariance matrix of the ML estimator. The assessment of the sample variability of the parameter estimates in a linear Gaussian cluster-weighted model is a necessary step in the subsequent development of inference methods for the model parameters, such as asymptotic confidence intervals, tests for the significance of the effect of any covariate on a given response within any sub-population and tests for the significance of the difference between the effects of the same covariate on a given response in two different sub-populations. Thus, additional computations are necessary to obtain an assessment of the sample variability of model parameter estimates. To the best of the author's knowledge, the only solution currently available for the linear Gaussian cluster-weighted models is implemented in the `flexCWM` package, where approximated standard errors are only provided for the intercepts and regression coefficients according to an approach in which a number of separate linear regression analyses with random covariates are carried out (one for each component of the mixture), and the sample observations are weighted with their estimated posterior probabilities of coming from the different components of the mixture. However, this approach only partially exploits the sample information about the parameters under a linear normal cluster-weighted model. Thus, other approaches could be investigated and detected. A solution can be obtained by resorting to bootstrap methods (see, e.g., Newton

& Raftery 1994; Basford et al. 1997; McLachlan & Peel 2000). However, the overall computational process associated with the use of bootstrap techniques can become particularly time-consuming and complex because of difficulties typically associated with the fitting of finite mixture models (e.g. label-switching problems, possible convergence failures of the EM algorithm on the bootstrap samples). These inconveniences could be avoided through an approach in which the observed information matrix is obtained from the incomplete data log-likelihood and employed to compute information-based estimators of the covariance matrix of the ML estimator (see e.g. McLachlan & Peel 2000). This task has been successfully carried out under normal mixture models (Boldea & Magnus, 2009) and clusterwise linear regression models (Galimberti et al., 2021).

In order to make it possible to properly assess both the variability of and the covariance between ML estimates of all the parameters under multivariate linear normal cluster-weighted models with a multivariate response, the gradient vector and second-order derivative matrix of the incomplete data log-likelihood for these models are explicitly derived here. Then, these results are used to obtain three estimators of the observed information matrix and the covariance matrix of the ML estimator. Properties of such estimators are numerically investigated using simulated datasets in comparison with the parametric bootstrap and the approach implemented in `flexCWM`. A numerical evaluation of the relationships between the estimators introduced here and those described by Boldea and Magnus (2009) is also provided. The practical usefulness of the proposed estimators is illustrated in a study aiming at evaluating the link between tourism flows and attendance at museums and monuments in two Italian regions.

The remainder of the paper is organised as follows. Section 2 provides the definition of multivariate Gaussian linear cluster-weighted model together with some quantities employed in the derivation of the score vector and the Hessian matrix. Section 3 describes the estimators of the information matrix and the covariance matrix of the ML estimator. Sections 4 and 5 summarize the main results obtained from the analysis of simulated and real datasets, respectively. The analytical expressions of the score vector and the Hessian matrix are reported in Appendix. Technical details and additional results from the analysis of simulated datasets can be found in a separate document as supplementary materials.

2 Score Vector and Hessian Matrix of Gaussian Linear Cluster-Weighted Models

Let $\mathbf{X} = (X_1, ..., X_p)'$ and $\mathbf{Y} = (Y_1, ..., Y_q)'$ be two continuous random vectors with joint probability density function (p.d.f.) $f(\mathbf{x}, \mathbf{y})$. The response vector $\mathbf{Y}$ and the covariate vector $\mathbf{X}$ take values in $\mathbb{R}^q$ and $\mathbb{R}^p$, respectively. Following Dang et al. (2017), $(\mathbf{X}', \mathbf{Y}')'$ follows a cluster-weighted model of order G if

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

where $\Omega_1, \ldots, \Omega_G$ denote the G unknown sub-populations ($\Omega_g \cap \Omega_j = \emptyset \ \forall g \neq j$), $\pi_g = \mathbb{P}(\Omega_g)$, $\pi_g > 0 \ \forall g$, $\sum_{g=1}^{G} \pi_g = 1$, $f(\mathbf{x}|\Omega_g)$ is the conditional p.d.f. of $\mathbf{X}$ given Ω_g, $f(\mathbf{y}|\mathbf{x}, \Omega_g)$ is the conditional p.d.f. of $\mathbf{Y}$ given $\mathbf{X}$ and Ω_g. A Gaussian linear cluster-weighted model is obtained from Eq. 1 by additionally assuming that the distributions of

$\mathbf{X}|\Omega_g$ and $\mathbf{Y}|(\mathbf{X}=\mathbf{x},\Omega_g)$ are Gaussian for $g=1,\ldots,G$ and the effect of $\mathbf{X}$ on $\mathbf{Y}$ for any Ω_g is linear. By embedding all these assumptions into model (1), $f(\mathbf{x},\mathbf{y})$ becomes

$$f(\mathbf{x},\mathbf{y};\boldsymbol{\vartheta})=\sum_{g=1}^{G}\pi_g\phi_p(\mathbf{x};\boldsymbol{\mu}_{\mathbf{X}_g},\boldsymbol{\Sigma}_{\mathbf{X}_g})\phi_q(\mathbf{y}|\mathbf{x};\mathbf{B}_g'\mathbf{x}^*,\boldsymbol{\Sigma}_{\mathbf{Y}_g}), \tag{2}$$

where $\phi_d(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ represents the p.d.f. of a normal d-dimensional random vector with expected value $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$,

$$\mathbf{B}_g'\mathbf{x}^*=\mathbb{E}(\mathbf{Y}|\mathbf{X}=\mathbf{x},\Omega_g),\ g=1,\ldots,G, \tag{3}$$

with $\mathbf{B}_g\in\mathbb{R}^{(1+p)\times q}$, $\mathbf{x}^*=(1,\mathbf{x}')'$, and $\boldsymbol{\vartheta}$ is the vector of the unknown parameters. It has been proved that linear Gaussian cluster-weighted models of order G define the same family of probability distributions generated by mixtures of G Gaussian models for the random vector $\mathbf{Z}=(\mathbf{X}',\mathbf{Y}')'$ (Ingrassia et al., 2012). However, it is important to stress that this latter type of mixtures cannot be employed to account for local linear dependencies between $\mathbf{X}$ and $\mathbf{Y}$.

The score vector and Hessian matrix of model (2) are derived by taking account of the fact that the weights $\pi_1,\ldots,\pi_G$ sum to one and the covariance matrices are symmetric. The first constraint is introduced in the maximization of the likelihood function by differentiating with respect to $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_{G-1})'$ and by setting $\pi_G=1-\pi_1-\cdots-\pi_{G-1}$. The constraints on the covariance matrices are dealt with by using the operators $\mathrm{vec}(\cdot)$, $\mathrm{v}(\cdot)$ and the duplication matrix. Namely, $\mathrm{vec}(\mathbf{B})$ is the column vector obtained by stacking the columns of matrix $\mathbf{B}$ one underneath the other. $\mathrm{v}(\boldsymbol{\Sigma})$ denotes the column vector obtained from $\mathrm{vec}(\boldsymbol{\Sigma})$ by eliminating all supradiagonal elements of a symmetric matrix $\boldsymbol{\Sigma}$ (thus, $\mathrm{v}(\boldsymbol{\Sigma})$ contains only the lower triangular part of $\boldsymbol{\Sigma}$). The duplication matrix $\mathbf{G}$ is the unique matrix which transforms $\mathrm{v}(\boldsymbol{\Sigma})$ into $\mathrm{vec}(\boldsymbol{\Sigma})$ ($\mathbf{G}\mathrm{v}(\boldsymbol{\Sigma})=\mathrm{vec}(\boldsymbol{\Sigma})$) (see e.g. Magnus & Neudecker 1988). Using this notation, the vector of the unknown parameters in model (2) can be denoted as $\boldsymbol{\vartheta}=(\boldsymbol{\pi}',\boldsymbol{\theta}_1',\ldots,\boldsymbol{\theta}_G')'$, where $\boldsymbol{\theta}_g=\left(\boldsymbol{\mu}_{\mathbf{X}_g}',\mathrm{v}\left(\boldsymbol{\Sigma}_{\mathbf{X}_g}\right)',\mathrm{vec}\left(\mathbf{B}_g\right)',\mathrm{v}\left(\boldsymbol{\Sigma}_{\mathbf{Y}_g}\right)'\right)'$.

Suppose that the observed data $\mathcal{S}=\{(\mathbf{x}_i,\mathbf{y}_i),i=1,\ldots,I\}$ is composed of I independent and identically distributed observations. Then, the incomplete log-likelihood function under the model (2) is

$$l(\boldsymbol{\vartheta})=\sum_{i=1}^{I}\ln\left(\sum_{g=1}^{G}\pi_g\phi_p\left(\mathbf{x}_i;\boldsymbol{\mu}_{\mathbf{X}_g},\boldsymbol{\Sigma}_{\mathbf{X}_g}\right)\phi_q\left(\mathbf{y}_i|\mathbf{x}_i;\mathbf{B}_g'\mathbf{x}_i^*,\boldsymbol{\Sigma}_{\mathbf{Y}_g}\right)\right). \tag{4}$$

Each sample observation provides its own contribution to the gth term of the sum that defines the mixture (2). As far as the contribution of the ith observation is concerned, it is given by:

$$\begin{aligned}p_{gi}\ =\ &\pi_g(2\pi)^{-\frac{p}{2}}\det(\boldsymbol{\Sigma}_{\mathbf{X}_g})^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{X}_g})'\boldsymbol{\Sigma}_{\mathbf{X}_g}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{X}_g})\right]\\ &(2\pi)^{-\frac{q}{2}}\det(\boldsymbol{\Sigma}_{\mathbf{Y}_g})^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{y}_i-\mathbf{B}_g'\mathbf{x}_i^*)'\boldsymbol{\Sigma}_{\mathbf{Y}_g}^{-1}(\mathbf{y}_i-\mathbf{B}_g'\mathbf{x}_i^*)\right].\end{aligned} \tag{5}$$

By exploiting properties of the logarithm and trace, the following equality holds true:

$$\begin{aligned}\ln p_{gi}\ =\ &\ln\pi_g-\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln\det(\boldsymbol{\Sigma}_{\mathbf{X}_g})-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_{\mathbf{X}_g}^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{X}_g})(\mathbf{x}_i-\boldsymbol{\mu}_{\mathbf{X}_g})'\right]+\\ &-\frac{q}{2}\ln(2\pi)-\frac{1}{2}\ln\det(\boldsymbol{\Sigma}_{\mathbf{Y}_g})-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_{\mathbf{Y}_g}^{-1}(\mathbf{y}_i-\mathbf{B}_g'\mathbf{x}_i^*)(\mathbf{y}_i-\mathbf{B}_g'\mathbf{x}_i^*)'\right].\end{aligned} \tag{6}$$

The explicit forms of the score vector and Hessian matrix, as developed here, require the introduction of some additional notation. Namely, let

$$\alpha_{gi} = \frac{p_{gi}}{\sum_{j=1}^{G} p_{ji}}, \tag{7}$$

$$\mathbf{a}_g = \frac{1}{\pi_g}\mathbf{e}_g, \qquad g = 1, \ldots, G-1, \tag{8}$$

$$\mathbf{a}_G = -\frac{1}{\pi_G}\mathbf{1}_{G-1},$$

where $\mathbf{e}_g$ denotes the gth column of the identity matrix of order $G-1$ and $\mathbf{1}_{G-1}$ is the $(G-1)\times 1$ vector having each element equal to 1. The following quantities are also required:

$$\mathbf{o}_{gi} = \boldsymbol{\Sigma}_{\mathbf{Y}_g}^{-1}(\mathbf{y}_i - \mathbf{B}_g'\mathbf{x}_i^*), \tag{9}$$

$$\mathbf{O}_{gi} = \boldsymbol{\Sigma}_{\mathbf{Y}_g}^{-1} - \mathbf{o}_{gi}\mathbf{o}_{gi}', \tag{10}$$

$$\mathbf{f}_{gi} = \boldsymbol{\Sigma}_{\mathbf{X}_g}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{X}_g}), \tag{11}$$

$$\mathbf{F}_{gi} = \boldsymbol{\Sigma}_{\mathbf{X}_g}^{-1} - \mathbf{f}_{gi}\mathbf{f}_{gi}'. \tag{12}$$

The explicit forms of the score vector $S(\boldsymbol{\vartheta})$ and Hessian matrix $H(\boldsymbol{\vartheta})$ for a Gaussian linear cluster-weighted model are provided in Theorems 1 and 2 (see Appendix). Proofs can be found in the document with the supplementary materials.

3 Covariance Matrix Estimation of the ML Estimator

In the ML approach, the information matrix $\mathfrak{I}(\boldsymbol{\vartheta})$ plays a crucial role, as it is used to asymptotically estimate the covariance of the ML estimator of the model parameters. Under regularity conditions and if the model is correctly specified, $\mathfrak{I}(\boldsymbol{\vartheta})$ is given either by the covariance of the score function $\mathbb{E}\left(S(\boldsymbol{\vartheta})S(\boldsymbol{\vartheta})'\right)$ or the negative of the expected value of the Hessian matrix $-\mathbb{E}\left(H(\boldsymbol{\vartheta})\right)$. Clearly, an analytical evaluation of the expectations required to obtain $\mathfrak{I}(\boldsymbol{\vartheta})$ under model (2) is quite complex. By exploiting some asymptotic results concerning ML estimation (White, 1982), it is possible to obtain the following asymptotic estimators of $\mathfrak{I}(\boldsymbol{\vartheta})$:

$$\mathcal{I}_1 = \sum_{i=1}^{I} \mathbf{S}_i(\hat{\boldsymbol{\vartheta}})\mathbf{S}_i(\hat{\boldsymbol{\vartheta}})', \quad \mathcal{I}_2 = -\sum_{i=1}^{I} \mathbf{H}_i(\hat{\boldsymbol{\vartheta}}),$$

where $\mathbf{S}_i(\hat{\boldsymbol{\vartheta}})$ and $\mathbf{H}_i(\hat{\boldsymbol{\vartheta}})$ denote the contribution of the ith sample observation to the score function and Hessian matrix evaluated at the ML estimate $\hat{\boldsymbol{\vartheta}}$, respectively. They can be used to obtain the following asymptotic estimators of $\mathrm{Cov}(\hat{\boldsymbol{\vartheta}})$, the covariance matrix of $\hat{\boldsymbol{\vartheta}}$:

$$\widehat{\mathrm{Cov}}_1(\hat{\boldsymbol{\vartheta}}) = \mathcal{I}_1^{-1}, \tag{13}$$

$$\widehat{\mathrm{Cov}}_2(\hat{\boldsymbol{\vartheta}}) = \mathcal{I}_2^{-1}. \tag{14}$$

Under suitable conditions (see e.g. Newey & McFadden 1994; Ritter 2015, for a general discussion and some results specifically dealing with finite mixture models, respectively), $\widehat{\mathrm{Cov}}_1(\hat{\boldsymbol{\vartheta}})$ and $\widehat{\mathrm{Cov}}_2(\hat{\boldsymbol{\vartheta}})$ can be considered consistent estimators of $\mathrm{Cov}(\hat{\boldsymbol{\vartheta}})$ when the model

is correctly specified. By exploiting the so-called sandwich approach (see e.g. White 1980), the following robust estimator can also be employed:

$$\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}}) = \mathcal{I}_2^{-1}\mathcal{I}_1\mathcal{I}_2^{-1}. \tag{15}$$

It is worth noting that for the existence of the estimators $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\vartheta}})$ and $\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}})$ it is required that matrix $\mathcal{I}_2$ is positive definite and well conditioned. The same requirements have to be fulfilled by matrix $\mathcal{I}_1$ in order to guarantee that $\widehat{\text{Cov}}_1(\hat{\boldsymbol{\vartheta}})$ exists. With large-scale covariance matrices and small sample sizes, $\mathcal{I}_1$ and/or $\mathcal{I}_2$ could be ill-conditioned. These situations can be managed by resorting to procedures able to produce improved estimators of $\mathfrak{I}(\boldsymbol{\vartheta})$ from either $\mathcal{I}_1$ or $\mathcal{I}_2$. For example, the algorithm by Higham (1988) computes the nearest positive definite matrix of a given symmetric matrix by adjusting its eigenvalues. Other approaches which exploit techniques of variance reduction could also be adopted (see e.g. Schäfer and Strimmer 2005).

4 Experimental Results from Simulated Datasets

4.1 Numerical Study of the Properties of the Proposed Estimators

In order to evaluate the properties of $\widehat{\text{Cov}}_1(\hat{\boldsymbol{\vartheta}})$, $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\vartheta}})$ and $\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}})$ in comparison with the estimators based on the parametric bootstrap and the approach implemented in `flexCWM`, five Monte Carlo studies have been performed. In the first study, the artificial datasets have been generated under a model defined by Eqs. 2–3 where $G = 2$, $q = 1$ and $p = 2$. As far as the model parameters are concerned, the following values have been employed: $\pi_1 = 0.7$, $\pi_2 = 0.3$, $\boldsymbol{\Sigma}_{\mathbf{Y}_1} = 1.5$, $\boldsymbol{\Sigma}_{\mathbf{Y}_2} = 1$, $\mathbf{B}_1' = (5, 2, 2)$, $\mathbf{B}_2' = (1, -2, -2)$, $\boldsymbol{\mu}_{\mathbf{X}_1}' = (-2, -2)$, $\boldsymbol{\mu}_{\mathbf{X}_2}' = (2, 2)$, $\boldsymbol{\Sigma}_{\mathbf{X}_1} = \left(\begin{smallmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{smallmatrix}\right)$, $\boldsymbol{\Sigma}_{\mathbf{X}_2} = \left(\begin{smallmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{smallmatrix}\right)$. Such values have been chosen so as to produce two well-separated groups of observations (see the upper panel of Fig. 1, with the pairwise scatterplots of the variables X_1, X_2 and Y_1 for a sample of size $I = 500$ generated as just described). In this way, problems of label switching across simulations are less likely to occur. Furthermore, the ML estimates of $\boldsymbol{\vartheta}$ are expected to be accurate enough to let the analysis be focused on the different ways of estimating the standard error of $\hat{\boldsymbol{\vartheta}}$. Using these parameter values, $R = 10,000$ datasets (of size I) have been generated as follows:

1. For the rth dataset ($r = 1, \ldots, R$), a sample of size I is drawn from the standard p-dimensional normal distribution; this gives the vectors $\boldsymbol{\epsilon}_{1r}, \ldots, \boldsymbol{\epsilon}_{Ir}$;
2. For the rth dataset ($r = 1, \ldots, R$), a sample of size I is drawn from the standard q-dimensional normal distribution; this gives the vectors $\boldsymbol{\eta}_{1r}, \ldots, \boldsymbol{\eta}_{Ir}$;
3. For the rth dataset ($r = 1, \ldots, R$), a sample of size I is drawn from the Bernoulli distribution with parameter π_1; this produces the 0-1 vector $\mathbf{z}_r = (z_{1r}, \ldots, z_{Ir})'$;
4. For the ith observation ($i = 1, \ldots, I$) of the rth dataset, vectors $\mathbf{x}_{ir}$ and $\mathbf{y}_{ir}$ are obtained as follows:

$$\begin{aligned} \mathbf{x}_{ir} &= \boldsymbol{\mu}_{\mathbf{X}_1} + \mathbf{A}_{\mathbf{X}_1}\boldsymbol{\epsilon}_{ir}, \quad \mathbf{y}_{ir} = \mathbf{B}_1'\mathbf{x}_{ir} + \mathbf{A}_{\mathbf{Y}_1}\boldsymbol{\eta}_{ir} \text{ if } z_{ir} = 1, \\ \mathbf{x}_{ir} &= \boldsymbol{\mu}_{\mathbf{X}_2} + \mathbf{A}_{\mathbf{X}_2}\boldsymbol{\epsilon}_{ir}, \quad \mathbf{y}_{ir} = \mathbf{B}_2'\mathbf{x}_{ir} + \mathbf{A}_{\mathbf{Y}_2}\boldsymbol{\eta}_{ir} \text{ if } z_{ir} = 0, \end{aligned}$$

where $\mathbf{A}_{\mathbf{X}_g}$ and $\mathbf{A}_{\mathbf{Y}_g}$ are matrices obtained from the spectral decompositions of $\boldsymbol{\Sigma}_{\mathbf{X}_g}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}_g}$, respectively. Such matrices are constructed such that $\mathbf{A}_{\mathbf{X}_g}\mathbf{A}_{\mathbf{X}_g}' = \boldsymbol{\Sigma}_{\mathbf{X}_g}$, $\mathbf{A}_{\mathbf{Y}_g}\mathbf{A}_{\mathbf{Y}_g}' = \boldsymbol{\Sigma}_{\mathbf{Y}_g}$, for $g = 1, 2$.

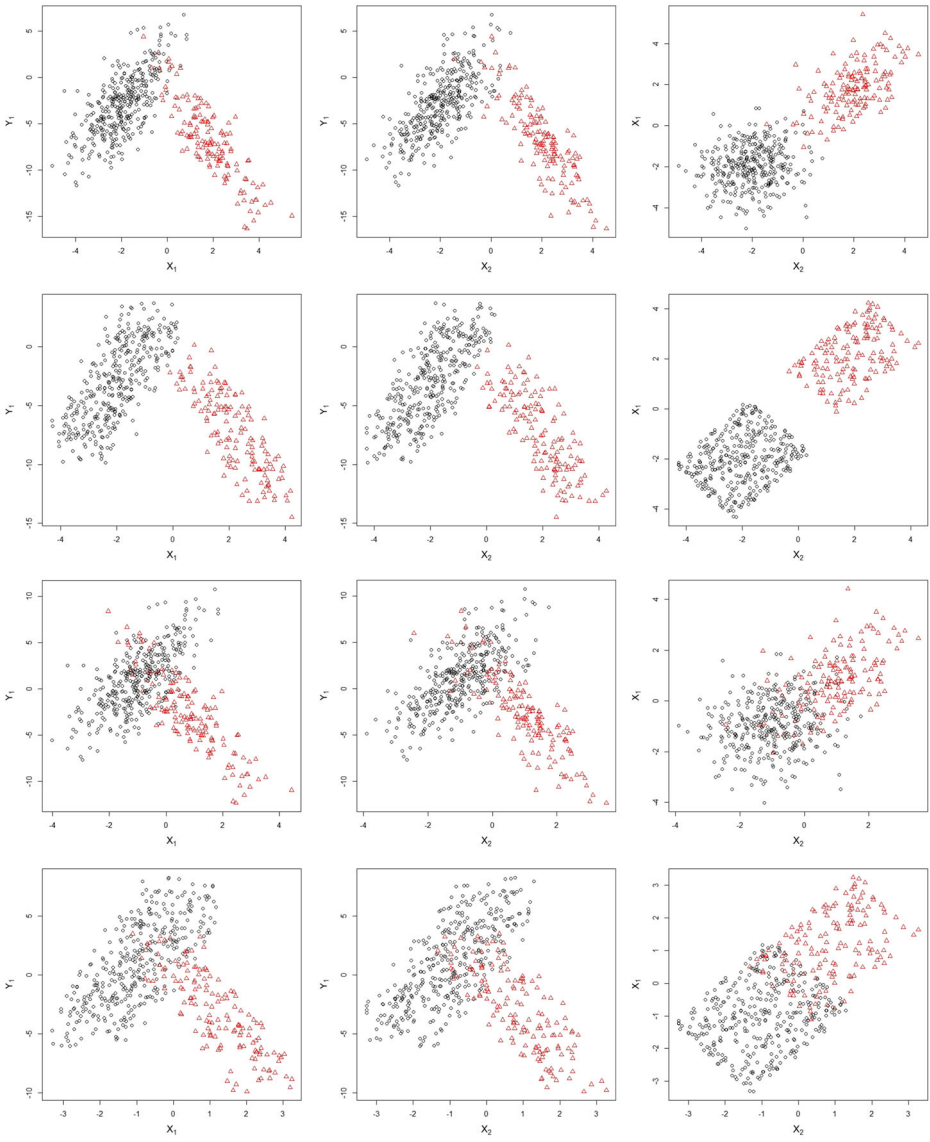

Fig. 1 Pairwise scatterplots of X_1, X_2 and Y_1 for four samples of size $I = 500$ generated in the first four studies. Upper and lower panels refer to the first and fourth studies, respectively; intermediate panels refer to the second and third ones. Black circles and red triangles correspond to $g = 1$ and $g = 2$, respectively

In the second study, the datasets have been obtained through the same procedure used in the first one except from the computation of vectors $\boldsymbol{\epsilon}_{ir}$ and $\boldsymbol{\eta}_{ir}$. Namely, a sample of size $I \cdot p$ is drawn from the uniform distribution in the interval (0,1) for the rth dataset ($r = 1, \ldots, R$); this produces a vector $\boldsymbol{\epsilon}_r^*$, whose elements are transformed as follows: $\epsilon_{jr} = \sqrt{12}(\epsilon_{jr}^* - 0.5)$, $j = 1, \ldots, I \cdot p$; the vector $\boldsymbol{\epsilon}_r$ resulting from this transformation has zero mean and unit variance; partitioning $\boldsymbol{\epsilon}_r$ into I $p-$dimensional vectors leads to $\boldsymbol{\epsilon}_{1r}, \ldots, \boldsymbol{\epsilon}_{Ir}$. The same process has been applied to obtain vectors $\boldsymbol{\eta}_{1r}, \ldots, \boldsymbol{\eta}_{Ir}$. The second panel of Fig. 1 provides

the pairwise scatterplots of X_1, X_2 and Y_1 for a sample of size $I = 500$ from this second study.

In the third and fourth studies, the datasets have been generated as in the first and second studies, respectively, but using the following model parameters: $\boldsymbol{\mu}_{\mathbf{X}_1} = (-1, -1)'$, $\boldsymbol{\mu}_{\mathbf{X}_2} = (1, 1)'$. This change in the values of $\boldsymbol{\mu}_{\mathbf{X}_1}$ and $\boldsymbol{\mu}_{\mathbf{X}_2}$ leads to overlapping groups of observations (see the pairwise scatterplots of X_1, X_2 and Y_1 for samples of size $I = 500$ in third and fourth panels of Fig. 1). The total number of model parameters in the first four studies is 19.

The fifth study has been carried out with the same settings of the first study but with $p = 8$ explanatory variables. The model parameters pertaining to $\mathbf{X}$ which have been employed to generate the datasets are as follows: $\boldsymbol{\mu}_{\mathbf{X}_1} = -2 \cdot \mathbf{1}_8$, $\boldsymbol{\mu}_{\mathbf{X}_2} = 2 \cdot \mathbf{1}_8$, $V(X_j|\Omega_g) = 1 \; \forall j$ for $g = 1, 2$, $\text{Cov}(X_j, X_h|\Omega_1) = 1 - \frac{|j-h|}{8} \; \forall j \neq h$, $\text{Cov}(X_j, X_h|\Omega_2) = 1 - \frac{|j-h|}{4} \; \forall j \neq h$. As far as the effects of the regressors on Y_1 are concerned, they have been set as follows: $\mathbf{B}_1 = (5, \boldsymbol{\mu}'_{\mathbf{X}_2})'$, $\mathbf{B}_2 = (1, \boldsymbol{\mu}'_{\mathbf{X}_1})'$. In this latter study, the total number of model parameters is 109.

In all studies, Monte Carlo experiments have been performed with two different sample sizes: $I = 250, 500$ in the first four studies, $I = 300, 500$ in the last study. In all experiments, it has been assumed that the rth dataset $\{(\mathbf{x}_{1r}, \mathbf{y}_{1r}), \ldots, (\mathbf{x}_{Ir}, \mathbf{y}_{Ir})\}$ is generated from a model defined by Eqs. 2–3 with $G = 2$. Thus, the maximum likelihood estimate $\hat{\boldsymbol{\vartheta}}_r$ of $\boldsymbol{\vartheta}$ has been computed for $r = 1, \ldots, R$ under such an assumption. Parameter estimation has been carried out through the EM algorithm as implemented in the function `cwm` of the package `flexCWM`. As far as the initialisation of the parameters is concerned, an option has been employed, which executes 5 independent runs of the k-means algorithm and picks the solution maximising the observed-data log-likelihood among these runs. The maximum number of iterations of the EM algorithm has been set equal to 400. A convergence criterion based on the Aitken acceleration has been used, with a threshold $\epsilon = 10^{-6}$ (for further details, see Mazza et al. 2018).

The R independent estimates of $\boldsymbol{\vartheta}$ are used to approximate the true distribution of $\hat{\boldsymbol{\vartheta}}$ and, in particular, the true standard errors of all the elements of $\hat{\boldsymbol{\vartheta}}$. Estimates of these standard errors have been computed using the three information-based estimators and the parametric bootstrap for $R_1 = 2000$ datasets obtained as described above. For each bootstrap estimate, 100 bootstrap samples have been employed. For the ML estimates of the model intercepts and regression coefficients, the standard errors estimated by the function `cwm` of the package `flexCWM` using the approach illustrated in the introduction have been included in the comparison. The performances of these strategies have been evaluated on the basis of an estimate of their biases and root mean squared errors (RMSE). A comparative evaluation of such approaches has been carried out also through the coverage probabilities (CP) of 90% and 95% confidence intervals based on the examined standard errors' estimates and the standard normal quantiles. In this latter comparison, the attention is focused on the expected mean values of the regressors (i.e. $\boldsymbol{\mu}_{\mathbf{X}_1}$ and $\boldsymbol{\mu}_{\mathbf{X}_2}$) and the regression coefficients (all the entries in the first column of $\mathbf{B}_1$ and $\mathbf{B}_2$ except the first one).

Tables 1, 2, 3 and 4 contain the biases and RMSEs for the first four Monte Carlo studies with samples of size 250. The same information for the last study and the sample size $I = 300$ can be found in Table 5. The corresponding values for the CPs are summarised in Tables 6, 7, 8, 9 and 10. In all the tables, the results obtained using the function `cwm` of the package `flexCWM`, the bootstrap and the estimators defined in Eqs. 13–15 are denoted as `cwm`, *Boot*, C_1, C_2 and C_3, respectively. From now on, $\mathbf{B}_g[j, k]$ is used to denote the element on the j-th row and k-th column of matrix $\mathbf{B}_g$; $\boldsymbol{\mu}_g[j]$ represents the j-th element

Table 1 Biases and root mean squared errors of the examined standard error estimators of $\hat{\vartheta}$ in the first study, $I = 250$

	BIAS					RMSE				
ϑ	cwm	*Boot*	C_1	C_2	C_3	cwm	*Boot*	C_1	C_2	C_3
π_1		0.031	0.050	0.038	0.040		0.217	0.096	0.090	0.092
$\mathbf{B}_1[1,1]$	−5.048	−2.829	−4.068	−2.958	−0.602	5.301	3.651	4.596	3.380	3.156
$\mathbf{B}_1[2,1]$	−1.766	−1.013	−1.426	−1.076	−0.256	1.874	1.335	1.641	1.232	1.088
$\mathbf{B}_1[3,1]$	−1.724	−0.958	−1.373	−1.032	−0.215	1.821	1.274	1.577	1.175	1.078
$\boldsymbol{\Sigma}_{\mathbf{Y}_1}$		−0.693	4.722	4.331	4.473		0.833	4.823	4.340	4.601
$\boldsymbol{\mu}_{\mathbf{X}_1}[1]$		−0.048	0.168	−0.047	−0.039		0.687	0.511	0.448	0.448
$\boldsymbol{\mu}_{\mathbf{X}_1}[2]$		0.010	0.240	0.026	0.033		0.697	0.521	0.432	0.434
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[1,1]$		−0.146	0.527	−0.078	−0.170		1.424	1.684	1.199	1.533
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[1,2]$		−0.085	0.408	−0.023	−0.070		1.419	1.650	1.197	1.525
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[2,2]$		−0.001	0.688	0.046	−0.080		1.417	1.741	1.198	1.525
$\mathbf{B}_2[1,1]$	−15.263	0.223	3.462	−0.630	−1.199	15.395	4.171	6.374	3.415	5.514
$\mathbf{B}_2[2,1]$	−6.046	0.301	1.871	−0.103	−0.445	6.105	1.793	2.934	1.416	2.255
$\mathbf{B}_2[3,1]$	−6.226	0.085	1.620	−0.300	−0.617	6.280	1.742	2.760	1.383	2.208
$\boldsymbol{\Sigma}_{\mathbf{Y}_2}$		0.040	10.541	8.444	8.227		0.781	10.828	8.486	8.545
$\boldsymbol{\mu}_{\mathbf{X}_2}[1]$		−0.080	0.632	−0.139	−0.076		1.480	1.567	1.244	1.323
$\boldsymbol{\mu}_{\mathbf{X}_2}[2]$		−0.104	0.624	−0.145	−0.087		1.496	1.574	1.256	1.327
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[1,1]$		−0.442	1.759	−0.419	−0.702		3.516	4.710	3.360	4.257
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[1,2]$		−0.468	1.125	−0.455	−0.548		3.519	4.512	3.364	4.234
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[2,2]$		−0.341	1.853	−0.352	−0.676		3.504	4.746	3.352	4.253

All entries have been multiplied by 100

Table 2 Biases and root mean squared errors of the examined standard error estimators of $\hat{\vartheta}$ in the second study, $I = 250$

	BIAS					RMSE				
ϑ	cwm	*Boot*	C_1	C_2	C_3	cwm	*Boot*	C_1	C_2	C_3
π_1		0.006	0.024	0.001	0.001		0.221	0.087	0.082	0.082
$\mathbf{B}_1[1,1]$	−4.220	−1.959	−4.021	−2.477	−0.184	4.367	2.827	4.231	2.721	1.737
$\mathbf{B}_1[2,1]$	−1.446	−0.683	−1.397	−0.813	0.050	1.514	1.034	1.499	0.935	0.701
$\mathbf{B}_1[3,1]$	−1.639	−0.863	−1.562	−1.008	−0.174	1.700	1.160	1.652	1.103	0.703
$\boldsymbol{\Sigma}_{\mathbf{Y}_1}$		1.722	12.779	6.679	3.074		1.780	12.815	6.683	3.094
$\boldsymbol{\mu}_{\mathbf{X}_1}[1]$		−0.130	0.019	−0.142	−0.141		0.647	0.385	0.392	0.392
$\boldsymbol{\mu}_{\mathbf{X}_1}[2]$		0.051	0.214	0.051	0.052		0.664	0.458	0.384	0.384
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[1,1]$		1.730	5.682	1.712	−0.128		2.126	6.030	1.967	0.678
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[1,2]$		2.793	7.498	2.761	−0.020		3.054	7.764	2.926	0.666
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[2,2]$		1.709	5.689	1.719	−0.123		2.109	6.036	1.973	0.677
$\mathbf{B}_2[1,1]$	−13.882	1.521	2.659	−0.015	−0.617	13.951	4.027	4.590	2.618	3.525
$\mathbf{B}_2[2,1]$	−6.010	0.290	0.950	−0.184	−0.413	6.042	1.595	1.918	1.187	1.622
$\mathbf{B}_2[3,1]$	−6.090	0.247	0.904	−0.255	−0.508	6.120	1.570	1.909	1.185	1.645
$\boldsymbol{\Sigma}_{\mathbf{Y}_2}$		2.767	21.632	11.065	5.659		2.868	21.835	11.095	5.791
$\boldsymbol{\mu}_{\mathbf{X}_2}[1]$		0.148	0.589	−0.041	−0.043		1.343	1.299	0.986	0.984
$\boldsymbol{\mu}_{\mathbf{X}_2}[2]$		0.164	0.585	−0.046	−0.048		1.320	1.270	0.964	0.964
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[1,1]$		3.549	10.810	3.160	−0.135		4.532	12.011	4.027	1.765
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[1,2]$		4.838	12.754	4.430	−0.002		5.599	13.788	5.084	1.760
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[2,2]$		3.518	10.708	3.151	−0.094		4.508	11.920	4.020	1.763

All entries have been multiplied by 100

Table 3 Biases and root mean squared errors of the examined standard error estimators of $\hat{\vartheta}$ in the third study, $I = 250$

	BIAS					RMSE				
ϑ	cwm	*Boot*	C_1	C_2	C_3	cwm	*Boot*	C_1	C_2	C_3
π_1		−0.000	0.044	0.029	0.054		0.243	0.131	0.128	0.156
$\mathbf{B}_1[1, 1]$	−3.552	−1.794	−2.556	−1.668	0.061	3.674	2.347	2.873	1.982	2.091
$\mathbf{B}_1[2, 1]$	−1.977	−1.054	−1.468	−1.061	−0.118	2.079	1.551	1.697	1.256	1.214
$\mathbf{B}_1[3, 1]$	−1.902	−0.950	−1.362	−0.985	−0.062	1.998	1.482	1.589	1.174	1.197
$\mathbf{\Sigma}_{\mathbf{Y}_1}$		−0.759	4.911	4.647	5.113		0.906	5.024	4.681	5.343
$\boldsymbol{\mu}_{\mathbf{X}_1}[1]$		−0.042	0.196	−0.031	−0.015		0.749	0.531	0.456	0.462
$\boldsymbol{\mu}_{\mathbf{X}_1}[2]$		0.005	0.258	0.030	0.043		0.753	0.544	0.446	0.453
$\mathbf{\Sigma}_{\mathbf{X}_1}[1, 1]$		−0.141	0.551	−0.064	−0.150		1.438	1.712	1.215	1.539
$\mathbf{\Sigma}_{\mathbf{X}_1}[1, 2]$		−0.061	0.448	0.005	−0.041		1.432	1.681	1.213	1.533
$\mathbf{\Sigma}_{\mathbf{X}_1}[2, 2]$		0.027	0.736	0.081	−0.040		1.431	1.780	1.216	1.533
$\mathbf{B}_2[1, 1]$	−12.209	−0.143	1.969	−0.536	−0.266	12.282	3.176	4.327	3.009	5.175
$\mathbf{B}_2[2, 1]$	−7.068	0.144	1.863	−0.261	−0.442	7.126	2.091	3.175	1.749	2.918
$\mathbf{B}_2[3, 1]$	−7.042	0.159	1.828	−0.268	−0.449	7.096	2.064	3.101	1.675	2.758
$\mathbf{\Sigma}_{\mathbf{Y}_2}$		−0.070	11.005	8.879	9.018		0.844	11.323	8.947	9.464
$\boldsymbol{\mu}_{\mathbf{X}_2}[1]$		−0.344	0.749	−0.195	−0.018		1.689	1.894	1.463	1.814
$\boldsymbol{\mu}_{\mathbf{X}_2}[2]$		−0.425	0.701	−0.271	−0.123		1.697	1.879	1.455	1.720
$\mathbf{\Sigma}_{\mathbf{X}_2}[1, 1]$		−0.779	2.310	−0.486	−0.818		3.659	5.263	3.470	4.814
$\mathbf{\Sigma}_{\mathbf{X}_2}[1, 2]$		−0.629	1.846	−0.366	−0.593		3.630	5.077	3.455	4.781
$\mathbf{\Sigma}_{\mathbf{X}_2}[2, 2]$		−0.539	2.619	−0.255	−0.664		3.616	5.406	3.445	4.791

All entries have been multiplied by 100

Table 4 Biases and root mean squared errors of the examined standard error estimators of $\hat{\vartheta}$ in the fourth study, $I = 250$

	BIAS					RMSE				
ϑ	cwm	*Boot*	C_1	C_2	C_3	cwm	*Boot*	C_1	C_2	C_3
π_1		−0.080	0.054	−0.014	−0.012		0.258	0.164	0.135	0.141
$\mathbf{B}_1[1,1]$	−9.960	−8.445	−9.492	−7.788	−5.135	9.984	8.543	9.538	7.861	5.506
$\mathbf{B}_1[2,1]$	−2.944	−2.158	−2.820	−2.069	−0.925	2.981	2.303	2.878	2.143	1.316
$\mathbf{B}_1[3,1]$	−2.854	−2.042	−2.706	−1.984	−0.880	2.892	2.198	2.767	2.058	1.259
$\boldsymbol{\Sigma}_{\mathbf{Y}_1}$		1.270	12.324	6.503	3.441		1.361	12.365	6.517	3.528
$\boldsymbol{\mu}_{\mathbf{X}_1}[1]$		−0.288	−0.040	−0.221	−0.211		0.707	0.401	0.436	0.435
$\boldsymbol{\mu}_{\mathbf{X}_1}[2]$		−0.195	0.064	−0.119	−0.110		0.694	0.428	0.411	0.415
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[1,1]$		1.098	5.222	1.147	−0.695		1.649	5.596	1.491	0.958
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[1,2]$		1.696	6.562	1.705	−1.108		2.096	6.864	1.953	1.290
$\boldsymbol{\Sigma}_{\mathbf{X}_1}[2,2]$		1.045	5.187	1.115	−0.725		1.615	5.563	1.466	0.981
$\mathbf{B}_2[1,1]$	−13.889	−0.999	0.779	−1.230	−1.327	13.932	3.227	3.449	3.060	4.382
$\mathbf{B}_2[2,1]$	−9.282	−1.819	−0.835	−2.155	−2.318	9.309	2.621	2.264	2.656	3.180
$\mathbf{B}_2[3,1]$	−9.201	−1.693	−0.689	−2.053	−2.255	9.227	2.505	2.210	2.560	3.130
$\boldsymbol{\Sigma}_{\mathbf{Y}_2}$		2.759	22.773	11.819	6.563		2.882	23.022	11.865	6.753
$\boldsymbol{\mu}_{\mathbf{X}_2}[1]$		−0.600	0.441	−0.375	−0.206		1.666	1.754	1.437	1.641
$\boldsymbol{\mu}_{\mathbf{X}_2}[2]$		−0.654	0.391	−0.425	−0.255		1.668	1.682	1.415	1.541
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[1,1]$		3.047	12.617	3.316	0.036		4.361	14.157	4.448	2.594
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[1,2]$		5.083	15.362	5.038	0.118		5.964	16.650	5.845	2.596
$\boldsymbol{\Sigma}_{\mathbf{X}_2}[2,2]$		2.905	12.415	3.206	−0.034		4.262	13.978	4.366	2.594

All entries have been multiplied by 100

Table 5 Biases and root mean squared errors of the examined standard error estimators of $\hat{\vartheta}$ in the fifth study, $I = 300$

	BIAS					RMSE				
ϑ	cwm	*Boot*	C_1	C_2	C_3	cwm	*Boot*	C_1	C_2	C_3
π_1		0.003	0.037	0.011	0.011		0.196	0.079	0.068	0.068
$\mathbf{B}_1[1, 1]$	−4.292	−0.898	−0.286	−2.723	−0.828	4.518	19.941	2.232	3.005	2.445
$\mathbf{B}_1[-1, 1]$	−4.002	−2.017	−0.147	−2.543	−0.767	4.227	3.121	2.114	2.823	2.368
$\boldsymbol{\Sigma}_{\mathbf{Y}_1}$		−0.403	5.809	3.905	4.025		4.250	5.891	3.910	4.115
$\boldsymbol{\mu}_{\mathbf{X}_1}$		0.008	1.117	−0.014	−0.014		0.851	1.228	0.359	0.359
$\mathrm{v}(\boldsymbol{\Sigma}_{\mathbf{X}_1})$		0.098	1.922	0.032	−0.050		2.505	2.727	1.587	1.791
$\mathbf{B}_2[1, 1]$	−30.889	2.842	54.228	−2.217	−4.911	31.112	20.522	61.001	6.438	10.367
$\mathbf{B}_2[-1, 1]$	−10.633	0.797	18.699	−0.645	−1.557	10.714	3.544	20.958	2.297	3.645
$\boldsymbol{\Sigma}_{\mathbf{Y}_2}$		0.056	19.834	7.571	7.339		3.695	20.658	7.603	7.592
$\boldsymbol{\mu}_{\mathbf{X}_2}$		−0.037	6.633	−0.141	−0.141		1.848	7.429	0.942	0.942
$\mathrm{v}(\boldsymbol{\Sigma}_{\mathbf{X}_2})$		0.010	10.143	−0.093	−0.344		3.884	12.418	3.120	3.485

All entries have been multiplied by 100

Table 6 Coverage probability of the 90% and 95% confidence intervals for $\boldsymbol{\mu}_{\mathbf{X}_g}$, $\mathbf{B}_g[2, 1]$ and $\mathbf{B}_g[3, 1]$ based on the standard error estimators of $\hat{\boldsymbol{\vartheta}}$ in the first study, $I = 250$

	CP 90%					CP 95%				
	cwm	*Boot*	C_1	C_2	C_3	cwm	*Boot*	C_1	C_2	C_3
$\boldsymbol{\mu}_{\mathbf{X}_1}[1]$		0.896	0.909	0.899	0.900		0.948	0.955	0.949	0.949
$\boldsymbol{\mu}_{\mathbf{X}_1}[2]$		0.901	0.913	0.905	0.905		0.946	0.956	0.951	0.951
$\boldsymbol{\mu}_{\mathbf{X}_2}[1]$		0.886	0.909	0.886	0.889		0.940	0.957	0.941	0.941
$\boldsymbol{\mu}_{\mathbf{X}_2}[2]$		0.885	0.903	0.885	0.883		0.941	0.959	0.941	0.941
$\mathbf{B}_1[2, 1]$	*0.818*	*0.856*	*0.838*	*0.859*	0.889	*0.890*	*0.918*	*0.908*	*0.923*	0.941
$\mathbf{B}_1[3, 1]$	*0.822*	*0.858*	*0.843*	*0.861*	0.893	*0.897*	*0.920*	*0.907*	*0.921*	0.942
$\mathbf{B}_2[2, 1]$	*0.616*	0.914	*0.938*	0.900	0.888	*0.705*	0.961	*0.972*	0.954	0.941
$\mathbf{B}_2[3, 1]$	*0.627*	0.906	*0.934*	0.895	*0.875*	*0.705*	0.953	*0.971*	0.949	*0.929*

Entries in italics denote effective coverage probabilities significantly different from the corresponding nominal ones (two-tailed normal tests, $\alpha = 0.00125$)

Table 7 Coverage probability of the 90% and 95% confidence intervals for $\boldsymbol{\mu}_{\mathbf{X}_g}$, $\mathbf{B}_g[2, 1]$ and $\mathbf{B}_g[3, 1]$ based on the standard error estimators of $\hat{\boldsymbol{\vartheta}}$ in the second study, $I = 250$

	CP 90%					CP 95%				
	`cwm`	*Boot*	C_1	C_2	C_3	`cwm`	*Boot*	C_1	C_2	C_3
$\boldsymbol{\mu}_{\mathbf{X}_1}[1]$		0.888	0.895	0.890	0.890		0.944	0.949	0.943	0.943
$\boldsymbol{\mu}_{\mathbf{X}_1}[2]$		0.913	0.920	0.910	0.910		0.956	0.964	0.958	0.958
$\boldsymbol{\mu}_{\mathbf{X}_2}[1]$		0.900	0.919	0.900	0.898		0.942	0.958	0.943	0.942
$\boldsymbol{\mu}_{\mathbf{X}_2}[2]$		0.893	0.909	0.893	0.893		0.939	0.950	0.938	0.938
$\mathbf{B}_1[2, 1]$	*0.839*	*0.872*	*0.846*	*0.872*	0.903	*0.904*	*0.931*	*0.908*	*0.935*	0.951
$\mathbf{B}_1[3, 1]$	*0.835*	*0.863*	*0.834*	*0.866*	0.895	*0.899*	*0.920*	*0.903*	*0.919*	0.941
$\mathbf{B}_2[2, 1]$	*0.630*	0.905	*0.930*	0.896	0.885	*0.709*	0.941	*0.970*	0.937	*0.930*
$\mathbf{B}_2[3, 1]$	*0.618*	0.905	*0.932*	0.891	*0.876*	*0.696*	0.954	*0.977*	0.946	*0.934*

Entries in italics denote effective coverage probabilities significantly different from the corresponding nominal ones (two-tailed normal tests, $\alpha = 0.00125$)

Table 8 Coverage probability of the 90% and 95% confidence intervals for $\boldsymbol{\mu}_{\mathbf{X}_g}$, $\mathbf{B}_g[2, 1]$ and $\mathbf{B}_g[3, 1]$ based on the standard error estimators of $\hat{\boldsymbol{\vartheta}}$ in the third study, $I = 250$

	CP 90%					CP 95%				
	cwm	*Boot*	C_1	C_2	C_3	cwm	*Boot*	C_1	C_2	C_3
$\boldsymbol{\mu}_{\mathbf{X}_1}[1]$		0.898	0.911	0.899	0.899		0.944	0.954	0.947	0.946
$\boldsymbol{\mu}_{\mathbf{X}_1}[2]$		0.897	0.912	0.902	0.903		0.947	0.954	0.948	0.948
$\boldsymbol{\mu}_{\mathbf{X}_2}[1]$		0.881	0.911	0.887	0.885		*0.932*	0.951	0.935	0.938
$\boldsymbol{\mu}_{\mathbf{X}_2}[2]$		*0.878*	0.910	0.885	0.887		0.939	0.954	0.936	0.941
$\mathbf{B}_1[2, 1]$	*0.805*	*0.853*	*0.837*	*0.857*	0.892	*0.882*	*0.917*	*0.905*	*0.921*	0.949
$\mathbf{B}_1[3, 1]$	*0.821*	*0.871*	*0.846*	*0.870*	0.894	*0.892*	*0.920*	*0.914*	*0.924*	0.943
$\mathbf{B}_2[2, 1]$	*0.574*	0.906	*0.939*	0.898	0.882	*0.665*	0.952	*0.972*	0.947	0.935
$\mathbf{B}_2[3, 1]$	*0.590*	0.904	*0.938*	0.898	0.880	*0.685*	0.953	*0.972*	0.950	0.942

Entries in italics denote effective coverage probabilities significantly different from the corresponding nominal ones (two-tailed normal tests, $\alpha = 0.00125$)

Table 9 Coverage probability of the 90% and 95% confidence intervals for $\boldsymbol{\mu}_{\mathbf{X}_g}$, $\mathbf{B}_g[2, 1]$ and $\mathbf{B}_g[3, 1]$ based on the standard error estimators of $\hat{\boldsymbol{\vartheta}}$ in the fourth study, $I = 250$

	CP 90%					CP 95%				
	`cwm`	*Boot*	C_1	C_2	C_3	`cwm`	*Boot*	C_1	C_2	C_3
$\boldsymbol{\mu}_{\mathbf{X}_1}[1]$		0.884	0.901	0.892	0.892		0.938	0.948	0.940	0.940
$\boldsymbol{\mu}_{\mathbf{X}_1}[2]$		0.894	0.910	0.903	0.904		0.947	0.959	0.949	0.949
$\boldsymbol{\mu}_{\mathbf{X}_2}[1]$		*0.866*	0.892	*0.876*	0.885		*0.922*	0.940	*0.928*	0.941
$\boldsymbol{\mu}_{\mathbf{X}_2}[2]$		*0.853*	0.886	*0.866*	*0.871*		*0.920*	0.937	*0.926*	*0.932*
$\mathbf{B}_1[2, 1]$	*0.820*	*0.857*	*0.828*	*0.866*	0.903	*0.895*	*0.920*	*0.903*	*0.923*	0.948
$\mathbf{B}_1[3, 1]$	*0.810*	*0.848*	*0.815*	*0.854*	0.893	*0.882*	*0.904*	*0.890*	*0.913*	0.940
$\mathbf{B}_2[2, 1]$	*0.580*	0.894	*0.925*	0.887	0.881	*0.660*	0.944	*0.973*	0.938	*0.922*
$\mathbf{B}_2[3, 1]$	*0.566*	0.898	*0.930*	0.891	*0.874*	*0.648*	0.947	*0.976*	0.945	*0.926*

Entries in italics denote effective coverage probabilities significantly different from the corresponding nominal ones (two-tailed normal tests, $\alpha = 0.00125$)

Table 10 Coverage probability of the 90% and 95% confidence intervals for $\boldsymbol{\mu}_{\mathbf{X}_g}$ and $\mathbf{B}_g[-1, 1]$ based on the standard error estimators of $\hat{\boldsymbol{\vartheta}}$ in the fifth study, $I = 300$

	CP 90%					CP 95%				
	`cwm`	*Boot*	C_1	C_2	C_3	`cwm`	*Boot*	C_1	C_2	C_3
$\boldsymbol{\mu}_{\mathbf{X}_1}$		0.891	0.939	0.893	0.893		0.944	0.975	0.946	0.946
$\boldsymbol{\mu}_{\mathbf{X}_2}$		0.897	0.988	0.898	0.898		0.947	0.997	0.947	0.947
$\mathbf{B}_1[-1, 1]$	0.831	0.869	0.898	0.862	0.888	0.897	0.924	0.947	0.920	0.940
$\mathbf{B}_2[-1, 1]$	0.616	0.906	0.995	0.889	0.869	0.699	0.955	0.999	0.945	0.926

of $\boldsymbol{\mu}_g$. Due to the large number of parameters pertaining to the regressors in the fifth study, Table 5 provides the mean values of the biases and RMSEs of the ML estimates over the elements of each of the following vectors of model parameters: $\boldsymbol{\mu}_{\mathbf{X}_g}$, $\mathrm{v}(\boldsymbol{\Sigma}_{\mathbf{X}_g})$, $\mathbf{B}_g[-1, 1]$ for $g = 1, 2$, where $\mathbf{B}_g[-1, 1]$ is the vector obtained by dropping the first element from the first column of $\mathbf{B}_g$. Thus, $\mathbf{B}_g[-1, 1]$ comprises the regression coefficients of the p covariates on Y_1 given Ω_g. In a similar way, Table 10 contains the mean values of the CPs of 90% and 95% confidence intervals over $\boldsymbol{\mu}_{\mathbf{X}_g}$ and $\mathbf{B}_g[-1, 1]$ for $g = 1, 2$.

Under the experimental conditions considered in the first Monte Carlo study, biases are generally small for all the estimated standard errors (see Table 1). The overall best performance in terms of accuracy seems to be achieved by means of the estimator $\widehat{\mathrm{Cov}}_2(\hat{\boldsymbol{\vartheta}})$. The bootstrap approach appears to provide the most precise estimates of the standard errors of $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_g}$. The sandwich method is slightly more accurate than the bootstrap approach in estimating the standard errors of the ML estimates of the expected values of the regressors; the opposite result holds true when dealing with the estimation of the standard errors of $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}_g}$. The highest root mean square errors are mostly obtained using either the function `cwm` of the package `flexCWM` or the estimator $\widehat{\mathrm{Cov}}_1(\hat{\boldsymbol{\vartheta}})$, which are therefore not recommended. These results confirm both the best performance of an estimator based on the Hessian and the poor performance of an estimator based on the gradient vector under correctly specified models registered in a study dealing with multivariate normal mixture models (Boldea & Magnus, 2009). It is also worth noting that the accuracy of the approach implemented in `flexCWM` sharply deteriorates when the ML estimates of the intercept and regression coefficients of the second group ($\mathbf{B}_2$) are considered. As far as the effective confidence levels for the parameters $\boldsymbol{\mu}_{\mathbf{X}_g}$, $\mathbf{B}_g[2, 1]$ and $\mathbf{B}_g[3, 1]$ are concerned (Table 6), the obtained results are generally similar to one another and quite close to the nominal confidence levels for all the examined methods except the one implemented in `flexCWM`. With this latter method, the effective confidence levels for the regression coefficients clearly deviate from the nominal ones, especially in the second group of observations. All these results have been employed to run tests for the hypotheses of equality between effective and nominal confidence levels. This task has been carried out through asymptotic two-tailed normal tests for a proportion at a 0.00125 significance level (the Bonferroni correction $0.01/8$ has been adopted to account for multiple tests performed for each estimation method and each nominal confidence level). All the effective CPs of the confidence intervals for the model regression coefficients obtained using both the estimator based on the gradient and the approach implemented in `flexCWM` appear to be significantly different from the corresponding nominal ones (see the entries in italics in Table 6). As far as the results from the bootstrap-based and Hessian-based estimators are concerned, the null hypothesis of equality between effective and nominal confidence levels should be rejected for two regression coefficients at both examined confidence levels; the same null hypothesis has to be rejected for only one regression coefficient when using the the sandwich approach.

In the second Monte Carlo study, a substantial increase in the biases of the estimated standard errors of $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}_g}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_g}$ has been registered with all the examined estimators except for the sandwich method (Table 2); this latter method is also the most accurate. Using a Gaussian cluster-weighted model for the analysis of datasets generated under a uniform cluster-weighted model seems to have a little impact on the confidence intervals for both the expected values of the regressors and the regression coefficients (Table 7).

When the data are obtained from two overlapping groups of observations drawn from Gaussian distributions (third study), the resulting biases and RMSEs (see Table 3) are quite similar to the ones from the first study; the main effect of the reduction in the separation

of the two groups is a general slight increase in the RMSEs with all the examined methods. The best accuracy is still achieved by the estimator $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\vartheta}})$ for all model parameters, with the exception of $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_g}$, whose standard errors are more accurately estimated by the bootstrap approach. As far as the effect of this reduction on the effective CPs of the confidence intervals is concerned (Table 8), the most remarkable result is an increase in the gap between nominal and effective CPs associated with the use of the approach implemented in `flexCWM` for the regression coefficients in the second group of observations.

When the two overlapping groups of observations are generated from the uniform distribution (fourth study), both biases and RMSEs of $\hat{\mathbf{B}}_1$ results remarkably increased with all the examined methods (see Table 4). However, it is worth noting that the lowest of such increases has always been associated with the use of $\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}})$, which is also the estimator with the best accuracy for the ML estimate of variances and covariances of the regressors in both groups and the majority of the model parameter estimates. Furthermore, the sandwich estimator shows the best performance in terms of effective CPs that are not significantly different from the nominal ones (Table 9).

In the presence of datasets generated under a Gaussian cluster-weighted model with $p = 8$ regressors (fifth study), the most remarkable effects of a larger number of covariates on the performance of the examined estimators with samples of size $I = 300$ appear to be (see Table 5 in comparison with Table 1) a sharp decrease in the accuracy of the estimates of the standard errors produced by the method based on the gradient for all the parameters of the second group of observations and a deterioration in the performances of the bootstrap approach in reference to all the parameters of the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ and Ω_g, especially $\mathbf{B}_g[1, 1]$, for $g = 1, 2$. As far as the methods $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\vartheta}})$ and $\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}})$ are concerned, biases and RMSEs result to be quite similar to the ones from the first study for all parameter estimates except the intercepts for both groups. Thus, the best overall accuracy is still achieved by the estimator $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\vartheta}})$.

The results from the five studies with samples of size 500 (see Tables A–J in the separate document with the supplementary materials) are generally in line with those just described. It is worth noting that using a larger sample size leads to a reduction in the RMSEs for all the examined estimators. With datasets containing two separated groups of observations (first and second studies), all the effective CPs of the confidence intervals obtained using the sandwich approach appear to be not significantly different from the corresponding nominal ones. When overlapping groups are considered (third and fourth studies), the estimator $\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}})$ has produced confidence intervals whose effective levels are the closest to the nominal ones. Thus, overall, the obtained results show the robustness of the sandwich method.

4.2 A Comparison with Some Estimators Under Normal Mixtures

As already mentioned in the "Introduction", three information-based estimators of the covariance matrix of the ML estimator for finite normal mixture models were developed by Boldea and Magnus (2009): two of them are based on the gradient vector and the Hessian matrix of the incomplete log-likelihood under a normal mixture model; the third estimator exploits the sandwich approach. From now on, these three estimators will be denoted as BM_1, BM_2 and BM_3, respectively. Furthermore, it has been already highlighted that finite mixtures of Gaussian distributions and linear Gaussian cluster-weighted models define the same family of probability distributions (Ingrassia et al., 2012). Thus, it could be interesting to obtain an evaluation of the relationships between the estimators described in Section 3 and the estimators developed by Boldea and Magnus (2009). This task has been numerically

performed by means of two additional simulation studies. The model employed to generate the simulated datasets is:

$$f(x, y; \boldsymbol{\psi}) = \sum_{g=1}^{G} \pi_g \phi_2(x, y; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{16}$$

where $p = q = 1$, $G = 2$, $\boldsymbol{\psi} = (\pi_1, \boldsymbol{\psi}_1', \boldsymbol{\psi}_2')'$, $\boldsymbol{\psi}_1 = (\boldsymbol{\mu}_1', \mathrm{v}(\boldsymbol{\Sigma}_1)')'$, $\boldsymbol{\psi}_1 = (\boldsymbol{\mu}_2', \mathrm{v}(\boldsymbol{\Sigma}_2)')'$, $\pi_1 = 0.7$, $\pi_2 = 0.3$, $\boldsymbol{\mu}_1' = (0, 0)$, $\boldsymbol{\mu}_2' = (\epsilon, \epsilon)$, $\boldsymbol{\Sigma}_1 = \left(\begin{smallmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{smallmatrix}\right)$, $\boldsymbol{\Sigma}_2 = \left(\begin{smallmatrix} 2.0 & 1.0 \\ 1.0 & 2.0 \end{smallmatrix}\right)$. The two studies have been carried out using 5 and 10 as values of ϵ so as to obtain two different levels of separation between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In each study, 100 datasets have been generated for each of two sample sizes: $I = 100, 250$. The `R` packages `mclust` (Scrucca et al., 2016) and `flexCWM` have been employed to compute the ML estimates of $\boldsymbol{\psi}$ in model (16) with $G = 2$ and the ML estimates of $\boldsymbol{\vartheta}$ in model (2) with $G = 2$, respectively. Furthermore, the standard errors of $\hat{\boldsymbol{\vartheta}}$ have been estimated using Eqs. 13–15. As far as $\hat{\boldsymbol{\psi}}$ is concerned, estimated standard errors have been computed according to the solutions described in Boldea and Magnus (2009).

Models (16) and (2) are characterised by the same value of π_1. Furthermore, as illustrated by Ingrassia et al. (2012), some elements in $\boldsymbol{\vartheta}$ coincide with some elements in $\boldsymbol{\psi}$; namely, $\boldsymbol{\mu}_{\mathbf{X}_1}[1] = \boldsymbol{\mu}_1[1]$, $\boldsymbol{\Sigma}_{\mathbf{X}_1}[1, 1] = \boldsymbol{\Sigma}_1[1, 1]$, $\boldsymbol{\mu}_{\mathbf{X}_2}[1] = \boldsymbol{\mu}_2[1]$, $\boldsymbol{\Sigma}_{\mathbf{X}_2}[1, 1] = \boldsymbol{\Sigma}_2[1, 1]$. Thus, the comparison between the estimators described in Section 3 and the estimators developed by Boldea and Magnus (2009) has been focused on the following two subvectors of model parameters: $\bar{\boldsymbol{\vartheta}} = (\pi_1, \boldsymbol{\mu}_{\mathbf{X}_1}[1], \boldsymbol{\Sigma}_{\mathbf{X}_1}[1, 1], \boldsymbol{\mu}_{\mathbf{X}_2}[1], \boldsymbol{\Sigma}_{\mathbf{X}_2}[1, 1])$, $\bar{\boldsymbol{\psi}} = (\pi_1, \boldsymbol{\mu}_1[1], \boldsymbol{\Sigma}_1[1, 1], \boldsymbol{\mu}_2[1], \boldsymbol{\Sigma}_2[1, 1])$. Let $se_m(\hat{\bar{\boldsymbol{\vartheta}}}_r[j])$ be the standard error of the j-th element of $\hat{\bar{\boldsymbol{\vartheta}}}$ computed using the estimator C_m and the r-th dataset. Furthermore, let $se_m(\hat{\bar{\boldsymbol{\psi}}}_r[j])$ be the standard error of the j-th element of $\hat{\bar{\boldsymbol{\psi}}}$ obtained from the estimator BM_m. In order to compare such estimated standard errors, the following differences have been computed: $d_{rm}(j) = se_m(\hat{\bar{\boldsymbol{\psi}}}_r[j]) - se_m(\hat{\bar{\boldsymbol{\vartheta}}}_r[j])$, $j = 1, \ldots, 5$, $m = 1, 2, 3$, $r = 1, \ldots, 100$. The results obtained for $\epsilon = 5$ and $\epsilon = 10$ with samples of size $I = 100$ are graphically represented in Figs. 2 and 3, respectively. The distributions of the differences $d_{rm}(j)$ for almost all the examined parameters appear to be centred around 0 and highly homogeneous, thus highlighting a general equivalence between the standard errors resulting from the two models. This result holds true especially for the estimators based on the Hessian matrix and the sandwich approach ($m = 2, 3$) when the separation between the two groups is larger ($\epsilon = 10$), and for the estimator based on the gradient vector ($m = 1$) when the separation is low ($\epsilon = 5$). However, it is also worth noting that with both levels of separation the differences in the standard errors of $\hat{\pi}_1$ computed using the two estimators based on the gradient vector show a median value slightly below 0 and a distribution with negative skewness. Similar results have been obtained with samples of size $I = 250$ (see Figures A and B in the supplementary material).

5 Analysing Regional Tourism Data in Italy

Similar to other studies (see e.g. Cellini & Cuccia 2013), the analysis summarised here aims at evaluating the link between tourism flows and attendance at museums and monuments, with a focus on two Italian regions: Emilia Romagna (`ER`) and Veneto (`Ve`). For both regions, three variables have been examined: tourist arrivals (denoted `Arriv`),

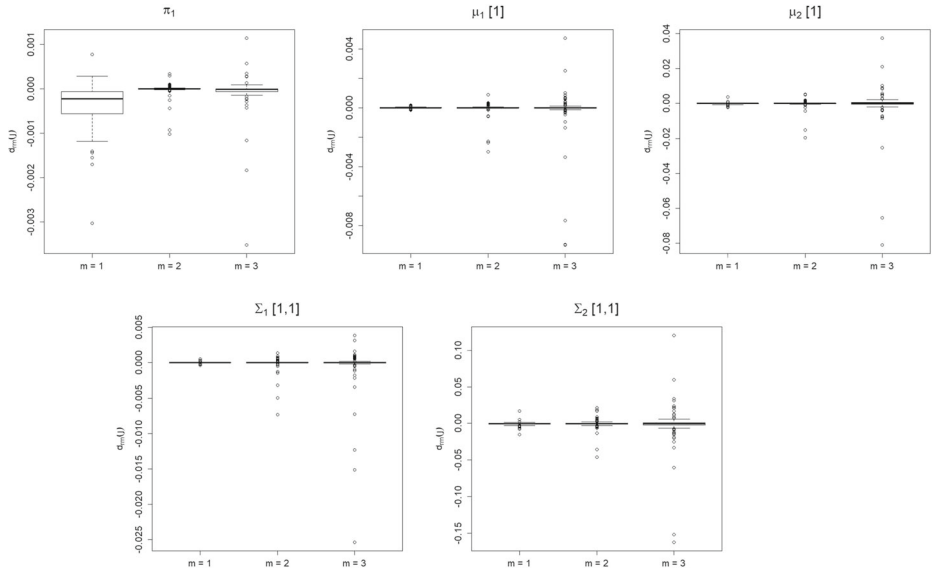

Fig. 2 Boxplots of the differences $d_{rm}(j)$ with samples of size $I = 100$ and $\epsilon = 5$

tourist overnights (`Overn`) and visits to State museums, monuments and museum networks (`Visit`). Measurements for such variables are available with a monthly frequency over the period January 1999 to December 2017. The source for `Visit` is the website of the Italian Ministry of Cultural Heritage (http://www.statistica.beniculturali.it). Data on `Arriv` and

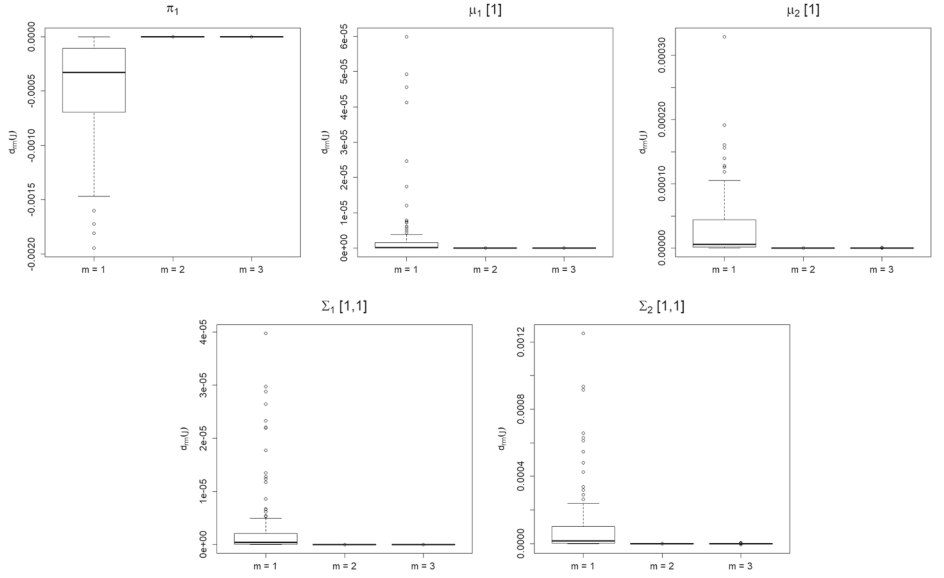

Fig. 3 Boxplots of the differences $d_{rm}(j)$ with samples of size $I = 100$ and $\epsilon = 10$

Table 11 Maximised log-likelihood and Bayesian information criterion of eight cluster-weighted models fitted to the tourism dataset

G	$l(\hat{\boldsymbol{\vartheta}})$	BIC
1	−8596.886	17,340.36
2	−8056.517	16,411.65
3	−7803.084	16,056.80
4	−7692.981	15,988.62
5	−7593.141	15,940.96
6	−7539.561	15,985.82
7	−7456.020	15,970.76
8	−7401.414	16,013.57

`Overn` have been obtained from the websites of Emilia-Romagna[1] and Veneto[2] regional governments. Overall, the analysed dataset is composed of $I = 228$ monthly observations for six variables. Because of the goal of the analysis, these variables have been partitioned as follows: $\mathbf{Y} = (\texttt{Visit ER}, \texttt{Visit Ve})'$, $\mathbf{X} = (\texttt{Arriv ER}, \texttt{Overn ER}, \texttt{Arriv Ve}, \texttt{Overn Ve})'$. The analysed data are expressed in thousands.

Models obtained from Eqs. 2–3 have been fitted to the dataset for G from 1 to 8. To this end, a function written for the `R` software environment which implements an EM algorithm for the ML estimation of multivariate Gaussian cluster-weighted linear models has been employed. In order to prevent problems due to the presence of singular or nearly singular matrices during the iterations, all covariance matrices have been required to have eigenvalues greater than 10^{-20}; furthermore, the ratio between the smallest and the largest eigenvalues of such matrices is required to be not lower than 10^{-10}. Model parameters are initialised according to a two-step strategy. In the first step, the joint distribution of the covariates and responses is estimated using a mixture of G Gaussian models through the `mclust` package. This produces the required starting values for the G weights, mean vectors and covariance matrices for the predictors. In the second step, the initialisation of $\mathbf{B}_g$ and $\boldsymbol{\Sigma}_{\mathbf{Y}_g}$ is obtained from the fitting of a Gaussian linear regression model to the sample of observations that have been assigned to the gth component of the mixture model estimated in the first step. The `R` package `systemfit` (Henningsen & Hamann, 2007) has been exploited to perform this task. The maximum number of iterations of the EM algorithm has been set equal to 500. A convergence criterion based on the Aitken acceleration has been used, with a threshold $\epsilon = 10^{-8}$.

Table 11 shows the values of the maximised log-likelihood ($l(\hat{\boldsymbol{\vartheta}})$) and the Bayesian information criterion (BIC) (Schwarz, 1978) for the eight fitted models, where $BIC = 2l(\hat{\boldsymbol{\vartheta}}) - npar \ln(I)$, and $npar$ denotes the number of model parameters. According to this criterion, the model with the best trade-off between the fit and complexity seems to be the one with $G = 5$ clusters of months. With this model, there is a perfect correspondence between some clusters and some months (see Table 12): cluster 2 only contains observations in April and May; cluster 4 only contains observations in June, July and August; cluster 5 only contains observations in January, February, November and December. As far as the remaining months are concerned, observations in September for the years 2005–2017 have been assigned to cluster 1; cluster 3 comprises the remaining observations in September together with all the observations in March and October. The obtained cluster structure

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

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

Table 12 Cross-classification of the observations from the tourism dataset, based on their variable time identified by month and maximum posterior probability estimated from the cluster-weighted model with $G = 5$

g	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0	0	0	0	0	0	0	0	13	0	0	0
2	0	0	0	19	19	0	0	0	0	0	0	0
3	0	0	19	0	0	0	0	0	6	19	0	0
4	0	0	0	0	0	19	19	19	0	0	0	0
5	19	19	0	0	0	0	0	0	0	0	19	19

clearly reflects seasonal patterns characterising tourism flows. Observations in cluster 4 (June, July and August) are characterized by the highest mean values of tourist arrivals and overnights in both regions, followed by those in cluster 1 (September 2005–2017), cluster 2 (April and May), cluster 3 (March, October and September 1999–2004) and cluster 4 (from November to February) (see the $\hat{\boldsymbol{\mu}}_g[j]$ in Table 13). In all clusters, Veneto is characterised by mean values of both regressors which are always higher than those of Emilia-Romagna. During the examined period of time, there seems to be heterogeneity also on the effects of the tourist arrivals and overnights on the number of visits in both regions (see the $\hat{\mathbf{B}}_g[j, k]$ in Table 13). Such effects do not result to be always positive. Furthermore, an increase in the tourist arrivals and overnights in one region does not necessarily have a positive impact on the number of visits to State museums, monuments and museum networks of the other region.

Estimates of the standard errors for the parameter estimates of the selected cluster-weighted model have been computed by the boostrap approach, using 100 bootstrap samples generated from the selected model. Furthermore, $\text{Cov}(\hat{\boldsymbol{\vartheta}})$ has been estimated by resorting to Eqs. 13–15. The algorithm by Higham (1988), as implemented in the `R` package `corpcor`,

Table 13 Estimated π_g, $\boldsymbol{\mu}_{\mathbf{X}_g}$ and $\mathbf{B}_g$ of the cluster-weighted model with $G = 5$ fitted to the tourism dataset

g	1	2	3	4	5
$\hat{\pi}_g$	0.057	0.168	0.192	0.250	0.333
$\boldsymbol{\mu}_g[1]$	848.1	766.9	515.6	1327.0	338.5
$\boldsymbol{\mu}_g[2]$	3654.3	2158.4	1573.2	7765.1	864.7
$\boldsymbol{\mu}_g[3]$	1634.4	1250.6	908.2	2095.6	561.6
$\boldsymbol{\mu}_g[4]$	6867.7	3906.2	2852.5	11420.1	1587.0
$\mathbf{B}_g[1, 1]$	47.771	197.894	5.204	−0.408	−2.183
$\mathbf{B}_g[2, 1]$	−0.173	0.229	0.447	0.138	0.171
$\mathbf{B}_g[3, 1]$	0.021	−0.037	−0.089	0.008	−0.020
$\mathbf{B}_g[4, 1]$	0.141	−0.124	−0.215	−0.061	0.013
$\mathbf{B}_g[5, 1]$	−0.018	0.007	0.063	−0.005	−0.008
$\mathbf{B}_g[1, 2]$	95.694	85.555	23.956	36.059	−19.050
$\mathbf{B}_g[2, 2]$	−0.121	0.002	0.181	0.132	−0.425
$\mathbf{B}_g[3, 2]$	0.022	0.025	−0.007	−0.010	0.141
$\mathbf{B}_g[4, 2]$	0.114	0.071	0.020	−0.053	0.172
$\mathbf{B}_g[5, 2]$	−0.024	−0.029	−0.013	0.005	−0.004

Table 14 $\hat{\mathbf{B}}_g[j,k]$, estimated standard errors, z_{gjk} values and p-values obtained using the bootstrap (columns 5–7) and Eq. 13 (columns 8–10)

g	j	k	$\hat{\mathbf{B}}_g[j,k]$	se($\hat{\mathbf{B}}_g[j,k]$)	z_{gjk}	p-value	se($\hat{\mathbf{B}}_g[j,k]$)	z_{gjk}	p-value
1	2	1	−0.173	0.318	−0.545	0.586	2.090	−0.083	0.934
1	3	1	0.021	0.078	0.262	0.793	0.305	0.067	0.946
1	4	1	0.141	0.172	0.820	0.412	0.648	0.217	0.828
1	5	1	−0.018	0.055	−0.328	0.743	0.133	−0.134	0.893
1	2	2	−0.121	0.330	−0.366	0.715	1.472	−0.082	0.935
1	3	2	0.022	0.080	0.278	0.781	0.196	0.113	0.910
1	4	2	0.114	0.175	0.652	0.515	0.474	0.240	0.810
1	5	2	−0.024	0.045	−0.538	0.590	0.053	−0.457	0.648
2	2	1	0.229	0.167	1.373	0.170	0.328	0.698	0.485
2	3	1	−0.037	0.044	−0.827	0.408	0.101	−0.364	0.716
2	4	1	−0.124	0.111	−1.121	0.262	0.196	−0.633	0.527
2	5	1	0.007	0.030	0.216	0.829	0.052	0.125	0.901
2	2	2	0.002	0.178	0.009	0.993	0.136	0.012	0.991
2	3	2	0.025	0.047	0.532	0.595	0.029	0.848	0.396
2	4	2	0.071	0.071	1.002	0.316	0.072	0.993	0.321
2	5	2	−0.029	0.021	−1.395	0.163	0.016	−1.868	0.062
3	2	1	0.447	0.148	3.023	0.003	0.168	2.663	0.008
3	3	1	−0.089	0.040	−2.224	0.026	0.037	−2.437	0.015
3	4	1	−0.215	0.094	−2.282	0.022	0.083	−2.570	0.010
3	5	1	0.063	0.028	2.299	0.022	0.028	2.270	0.023
3	2	2	0.181	0.195	0.927	0.354	0.145	1.251	0.211
3	3	2	−0.007	0.047	−0.154	0.878	0.036	−0.204	0.838
3	4	2	0.020	0.074	0.265	0.791	0.076	0.260	0.795
3	5	2	−0.013	0.018	−0.710	0.478	0.028	−0.456	0.648
4	2	1	0.138	0.088	1.565	0.118	0.058	2.384	0.017
4	3	1	0.008	0.024	0.343	0.732	0.011	0.778	0.436
4	4	1	−0.061	0.055	−1.100	0.271	0.035	−1.748	0.080
4	5	1	−0.005	0.013	−0.363	0.717	0.006	−0.737	0.461
4	2	2	0.132	0.200	0.657	0.511	0.032	4.130	0.000
4	3	2	−0.010	0.054	−0.195	0.846	0.008	−1.350	0.177
4	4	2	−0.053	0.084	−0.635	0.526	0.018	−2.941	0.003
4	5	2	0.005	0.015	0.337	0.736	0.004	1.155	0.248
5	2	1	0.171	0.082	2.078	0.038	0.069	2.489	0.013
5	3	1	−0.019	0.019	−1.017	0.309	0.018	−1.056	0.291
5	4	1	0.013	0.049	0.265	0.791	0.033	0.385	0.701
5	5	1	−0.008	0.008	−0.972	0.331	0.012	−0.647	0.517
5	2	2	−0.425	0.260	−1.635	0.102	0.108	−3.925	0.000
5	3	2	0.141	0.070	2.026	0.043	0.029	4.919	0.000
5	4	2	0.171	0.094	1.824	0.068	0.049	3.472	0.001
5	5	2	−0.004	0.013	−0.340	0.734	0.016	−0.280	0.779

Table 15 $\hat{\mathbf{B}}_g[j,k]$, estimated standard errors, z_{gjk} values and p-values obtained using Eqs. 14 (columns 5–7) and 15 (columns 8–10)

g	j	k	$\hat{\mathbf{B}}_g[j,k]$	se($\hat{\mathbf{B}}_g[j,k]$)	z_{gjk}	p-value	se($\hat{\mathbf{B}}_g[j,k]$)	z_{gjk}	p-value
1	2	1	−0.173	0.141	−1.234	0.217	0.157	−1.106	0.269
1	3	1	0.021	0.029	0.712	0.477	0.026	0.794	0.427
1	4	1	0.141	0.061	2.319	0.020	0.070	2.012	0.044
1	5	1	−0.018	0.015	−1.191	0.234	0.012	−1.478	0.139
1	2	2	−0.121	0.077	−1.570	0.116	0.070	−1.719	0.086
1	3	2	0.022	0.016	1.404	0.160	0.013	1.761	0.078
1	4	2	0.114	0.033	3.442	0.001	0.029	3.909	0.000
1	5	2	−0.024	0.008	−2.975	0.003	0.006	−4.254	0.000
2	2	1	0.229	0.162	1.411	0.158	0.148	1.545	0.122
2	3	1	−0.037	0.045	−0.817	0.414	0.037	−0.986	0.324
2	4	1	−0.124	0.098	−1.261	0.207	0.091	−1.366	0.172
2	5	1	0.007	0.024	0.268	0.789	0.022	0.302	0.763
2	2	2	0.002	0.090	0.017	0.986	0.091	0.017	0.986
2	3	2	0.025	0.025	1.004	0.315	0.027	0.916	0.360
2	4	2	0.071	0.055	1.302	0.193	0.057	1.254	0.210
2	5	2	−0.029	0.014	−2.146	0.032	0.015	−1.944	0.052
3	2	1	0.447	0.107	4.164	0.000	0.125	3.570	0.000
3	3	1	−0.089	0.027	−3.245	0.001	0.029	−3.110	0.002
3	4	1	−0.215	0.058	−3.691	0.000	0.060	−3.568	0.000
3	5	1	0.063	0.022	2.889	0.004	0.022	2.865	0.004
3	2	2	0.181	0.094	1.927	0.054	0.108	1.683	0.092
3	3	2	−0.007	0.025	−0.296	0.767	0.024	−0.306	0.759
3	4	2	0.020	0.052	0.381	0.704	0.052	0.379	0.705
3	5	2	−0.013	0.020	−0.660	0.509	0.019	−0.679	0.497
4	2	1	0.138	0.034	3.994	0.000	0.025	5.423	0.000
4	3	1	0.008	0.007	1.135	0.256	0.007	1.104	0.270
4	4	1	−0.061	0.021	−2.858	0.004	0.016	−3.751	0.000
4	5	1	−0.005	0.004	−1.051	0.293	0.004	−1.056	0.291
4	2	2	0.132	0.026	5.083	0.000	0.027	4.954	0.000
4	3	2	−0.010	0.005	−1.914	0.056	0.005	−2.015	0.044
4	4	2	−0.053	0.016	−3.346	0.001	0.017	−3.052	0.002
4	5	2	0.005	0.003	1.527	0.127	0.003	1.511	0.131
5	2	1	0.171	0.055	3.132	0.002	0.056	3.069	0.002
5	3	1	−0.019	0.016	−1.240	0.215	0.017	−1.156	0.247
5	4	1	0.013	0.024	0.540	0.589	0.022	0.576	0.564
5	5	1	−0.008	0.008	−0.920	0.357	0.008	−1.016	0.310
5	2	2	−0.425	0.075	−5.695	0.000	0.075	−5.684	0.000
5	3	2	0.141	0.022	6.561	0.000	0.022	6.271	0.000
5	4	2	0.171	0.033	5.261	0.000	0.028	6.032	0.000
5	5	2	−0.004	0.012	−0.379	0.704	0.011	−0.391	0.696

has been employed to adjust the eigenvalues of $\mathcal{I}_1$ and $\mathcal{I}_2$ so as to obtain their nearest positive definite matrices. Then, the estimated standard errors have been employed to run tests for the hypotheses $H_0 : \mathbf{B}_g[j,k] = 0$ for $j = 2,\ldots,p$, $k = 1,\ldots,q$, $g = 1,\ldots,G$. Such tests have been run under an asymptotic normal distribution for the z_{gjk} statistics, where $z_{gjk} = \frac{\hat{\mathbf{B}}_g[j,k]}{se(\hat{\mathbf{B}}_g[j,k])}$, with $se(\hat{\mathbf{B}}_g[j,k])$ denoting the estimated standard error of $\hat{\mathbf{B}}_g[j,k]$. Table 14 summarises the results obtained using the parametric bootstrap-based estimates and the score-based estimates; the results derived from the use of the other two estimators are reported in Table 15. According to all the examined methods and using $\alpha = 0.05$, the four examined regressors show a significant linear effect on the visits to State museums, monuments and museum networks in Emilia Romagna over the period 1999 to 2017 in March, October and over the period 1999 to 2004 in September (cluster 3); furthermore, in the central months of the winters from 1999 to 2017 (cluster 5), there are positive significant effects of `Arriv` on `Visit` in Emilia Romagna and of `Overn` on `Visit` in Veneto. According to the estimator based on $\mathcal{I}_1$, five additional regression coefficients (three within cluster 4, two within cluster 5) result to be significantly different from zero; the same conclusion is obtained from the use of $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\vartheta}})$ and $\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}})$. The results obtained using the three methods illustrated in Section 3 suggest that tourist arrivals in Emilia Romagna have a positive significant effect on the visits to State museums, monuments and museum networks both in Emilia Romagna and in Veneto in June, July and August. `Arriv ER` has a significant and negative effect on `Visit Ve` in January, February, November and December. Furthermore, `Arriv Ve` seems to significantly affect `Visit Ve`, with a positive effect in January, February, November and December and a negative effect in June, July and August. The tests based on the estimators $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\vartheta}})$ and $\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}})$ lead to a rejection of H_0 for further five regression coefficients, three of which concern the effect of tourist arrivals and overnights in Veneto in September for the years 2005–2017 (cluster 1). As far as cluster 2 is concerned, all regression coefficients seem to be not significantly different from 0 according to all approaches except for the effect of `Overn Ve` on `Visit Ve`, which results to be significant only when standard errors are estimated from the Hessian matrix.

By exploiting the results contained in the non-diagonal elements of matrices $\widehat{\text{Cov}}_1(\hat{\boldsymbol{\vartheta}})$, $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\vartheta}})$ and $\widehat{\text{Cov}}_3(\hat{\boldsymbol{\vartheta}})$, it is also possible to run tests for the significance of the difference between the effects of two different predictors on the same response in a given cluster or tests for the significance of the difference between the effects of the same predictor on the same response in two different clusters. For any given pair of regression coefficients in the model, the null hypothesis can be expressed as follows: $H_0 : \mathbf{B}_{g_1}[j_1,k_1] = \mathbf{B}_{g_2}[j_2,k_2]$. Three illustrative examples of hypotheses for the analysis of the tourism dataset are summarised in Tables 16 and 17, where $\hat{\delta}_{(g_1,j_1,k_1)(g_2,j_2,k_2)} = \hat{\mathbf{B}}_{g_1}[j_1,k_1] - \hat{\mathbf{B}}_{g_2}[j_2,k_2]$. These examples have been obtained from the following questions:

Table 16 Values of $\hat{\delta}_{(g_1,j_1,k_1)(g_2,j_2,k_2)}$ for testing three examples of $H_0 : \mathbf{B}_{g_1}[j_1,k_1] = \mathbf{B}_{g_2}[j_2,k_2]$ in the tourism data

Example	(g_1, j_1, k_1)	(g_2, j_2, k_2)	$\hat{\delta}_{(g_1,j_1,k_1)(g_2,j_2,k_2)}$
1	(4, 2, 1)	(4, 2, 2)	0.0061
2	(5, 2, 1)	(5, 4, 2)	−0.0007
3	(3, 2, 1)	(5, 2, 1)	0.2757

Table 17 Estimated standard errors (se) of $\hat{\delta}_{(g_1,j_1,k_1)(g_2,j_2,k_2)}$, z values and p-values obtained using Eqs. 13 (columns 2–4), 14 (columns 5–7) and 15 (columns 8–10) for testing three examples of $H_0 : \mathbf{B}_{g_1}[j_1, k_1] = \mathbf{B}_{g_2}[j_2, k_2]$ in the tourism data

Example	se	z	p-value	se	z	p-value	se	z	p-value
1	0.057	0.107	0.915	0.039	0.155	0.877	0.034	0.177	0.859
2	0.101	−0.007	0.995	0.079	−0.008	0.993	0.079	−0.008	0.993
3	0.181	1.522	0.128	0.120	2.292	0.022	0.137	2.015	0.044

(a) In June, July and August (cluster 4), do tourist arrivals in Emilia Romagna have a different effect on `Visit ER` and `Visit Ve` (first example)?
(b) In January, February, November and December (cluster 5), is the effect of tourist arrivals in Emilia Romagna on `Visit ER` different from the effect of tourist arrivals in Veneto on `Visit Ve` (second example)?
(c) Is the effect of tourist arrivals in Emilia Romagna on `Visit ER` in January, February, November and December (cluster 5) different from the one in March, September 1999–2004 and October (cluster 3) (third example)?
For each of these illustrations, Table 17 summarises the results of the z tests run using the three estimated covariance matrices (the non-diagonal elements are not reported). For the first two examples, the compared methods lead to results which are all in favour of the null hypothesis. In the third illustration, the null hypothesis should be rejected ($\alpha = 0.05$) when variances and covariances are estimated using both the Hessian and sandwich approaches.

6 Conclusions

Three information-based estimators of the asymptotic covariance matrix of the ML estimator under multivariate Gaussian linear cluster-weighted models have been illustrated. For their computation, formulae for the score vector and Hessian matrix of the incomplete log-likelihood have been derived. Properties of these estimators have been numerically evaluated using simulated samples in comparison with the parametric bootstrap-based estimator. For the ML estimates of the model intercepts and regression coefficients, the comparison has included an approach implemented in the package `flexCWM` in which estimated standard errors are computed by fitting G separate linear weighted regression models using the estimated posterior probabilities as weights. With correctly specified models, the most accurate estimator of the standard error of the ML estimator is the one based on the Hessian matrix. When Gaussian cluster-weighted models are fitted to datasets generated from a uniform distribution, the best accuracy is achieved with the sandwich estimator. Overall, the obtained results show the robustness of this latter method. Through these information-based estimators, the tasks of computing approximated confidence intervals and running tests concerning pairs of parameters can be easily carried out, as illustrated through a study aiming at evaluating the link between tourism flows and attendance at museums and monuments in two Italian regions. Asymptotic properties of the estimators introduced here could also be studied from a theoretical point of view. For example, suitable regularity conditions can be defined so as to provide a general assessment of their consistency (see for example Galimberti et al. (2020) for a similar study in the context of clusterwise regression models).

Appendix

Theorem 1 *The score vector $\boldsymbol{S}(\boldsymbol{\vartheta})$ for a cluster-weighted model of order G contains the following subvectors:*

$$\frac{\partial l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\pi}} = \sum_{i=1}^{I} \bar{\boldsymbol{a}}_i, \quad \frac{\partial l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\mu}_{X_g}} = \sum_{i=1}^{I} \alpha_{gi} \boldsymbol{f}_{gi} \; \forall g, \quad \frac{\partial l(\boldsymbol{\vartheta})}{\partial \mathrm{v}(\boldsymbol{\Sigma}_{X_g})} = \frac{1}{2} \sum_{i=1}^{I} \alpha_{gi} \boldsymbol{J}' \mathrm{vec}(\boldsymbol{F}_{gi}) \; \forall g,$$

$$\frac{\partial l(\boldsymbol{\vartheta})}{\partial \mathrm{vec}(\boldsymbol{B}_g)} = \sum_{i=1}^{I} \alpha_{gi} \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{gi}') \; \forall g, \quad \frac{\partial l(\boldsymbol{\vartheta})}{\partial \mathrm{v}(\boldsymbol{\Sigma}_{Y_g})} = -\frac{1}{2} \sum_{i=1}^{I} \alpha_{gi} \boldsymbol{L}' \mathrm{vec}(\boldsymbol{O}_{gi}) \; \forall g,$$

with $\bar{\boldsymbol{a}}_i = \sum_{g=1}^{G} \alpha_{gi} \boldsymbol{a}_g$, $\boldsymbol{L}$ and $\boldsymbol{J}$ denoting duplication matrices with dimensions $q^2 \times \frac{q(q+1)}{2}$ and $p^2 \times \frac{p(p+1)}{2}$, respectively.

Theorem 2 *The Hessian matrix $\boldsymbol{H}(\boldsymbol{\vartheta})$ for a cluster-weighted model of order G contains the following submatrices:*

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}'} = -\sum_{i=1}^{I} \bar{\boldsymbol{a}}_i \bar{\boldsymbol{a}}_i', \quad \frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\pi} \partial (\mathrm{vec}(\boldsymbol{B}_g))'} = \sum_{i=1}^{I} \alpha_{gi} (\boldsymbol{a}_g - \bar{\boldsymbol{a}}_i) \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{gi}')' \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\pi} \partial (\mathrm{v}(\boldsymbol{\Sigma}_{Y_g}))'} = -\frac{1}{2} \sum_{i=1}^{I} \alpha_{gi} (\boldsymbol{a}_g - \bar{\boldsymbol{a}}_i) \mathrm{vec}(\boldsymbol{O}_{gi})' \boldsymbol{L} \; \forall g, \quad \frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\pi} \partial \boldsymbol{\mu}_{X_g}'}$$

$$= \sum_{i=1}^{I} \alpha_{gi} (\boldsymbol{a}_g - \bar{\boldsymbol{a}}_i) \boldsymbol{f}_{gi}' \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\pi} \partial (\mathrm{v}(\boldsymbol{\Sigma}_{X_g}))'} = -\frac{1}{2} \sum_{i=1}^{I} \alpha_{gi} (\boldsymbol{a}_g - \bar{\boldsymbol{a}}_i) \mathrm{vec}(\boldsymbol{F}_{gi})' \boldsymbol{J} \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \mathrm{vec}(\boldsymbol{B}_g) \partial (\mathrm{vec}(\boldsymbol{B}_g))'} = -\sum_{i=1}^{I} \alpha_{gi} \Big[\left(\boldsymbol{\Sigma}_{Y_g}^{-1} \otimes (\boldsymbol{x}_i^* \boldsymbol{x}_i'^*) \right)$$

$$-(1-\alpha_{gi}) \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{gi}') \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{gi}')' \Big] \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \mathrm{vec}(\boldsymbol{B}_g) \partial (\mathrm{vec}(\boldsymbol{B}_j))'} = -\sum_{i=1}^{I} \alpha_{gi} \alpha_{ji} \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{gi}') \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{ji}')' \; \forall g \neq j,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \mathrm{vec}(\boldsymbol{B}_g) \partial (\mathrm{v}(\boldsymbol{\Sigma}_{Y_g}))'} = -\sum_{i=1}^{I} \alpha_{gi} \Big[\left(\boldsymbol{\Sigma}_{Y_g}^{-1} \otimes (\boldsymbol{x}_i^* \boldsymbol{o}_{gi}') \right)$$

$$+\frac{1}{2}(1-\alpha_{gi}) \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{gi}') \mathrm{vec}(\boldsymbol{O}_{gi})' \Big] \boldsymbol{L} \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \mathrm{vec}(\boldsymbol{B}_g) \partial (\mathrm{v}(\boldsymbol{\Sigma}_{Y_j}))'} = \frac{1}{2} \sum_{i=1}^{I} \alpha_{gi} \alpha_{ji} \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{gi}') \mathrm{vec}(\boldsymbol{O}_{ji})' \boldsymbol{L} \; \forall g \neq j,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \mathrm{vec}(\boldsymbol{B}_g) \partial \boldsymbol{\mu}_{X_g}'} = \sum_{i=1}^{I} \alpha_{gi} (1-\alpha_{gi}) \mathrm{vec}(\boldsymbol{x}_i^* \boldsymbol{o}_{gi}') \boldsymbol{f}_{gi}' \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{vec}(\boldsymbol{B}_g)\partial \boldsymbol{\mu}'_{\boldsymbol{X}_j}} = -\sum_{i=1}^{I} \alpha_{gi}\alpha_{ji}\text{vec}(\boldsymbol{x}_i^*\boldsymbol{o}'_{gi})\boldsymbol{f}'_{ji} \; \forall g \neq j,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{vec}(\boldsymbol{B}_g)\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_g}))'} = -\frac{1}{2}\sum_{i=1}^{I} \alpha_{gi}(1-\alpha_{gi})\text{vec}(\boldsymbol{x}_i^*\boldsymbol{o}'_{gi})\text{vec}(\boldsymbol{F}_{gi})'\boldsymbol{J} \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{vec}(\boldsymbol{B}_g)\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_j}))'} = \frac{1}{2}\sum_{i=1}^{I} \alpha_{gi}\alpha_{ji}\text{vec}(\boldsymbol{x}_i^*\boldsymbol{o}'_{gi})\text{vec}(\boldsymbol{F}_{ji})'\boldsymbol{J} \; \forall g \neq j,$$

$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{v}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_g})\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_g}))'} &= -\frac{1}{2}\sum_{i=1}^{I} \alpha_{gi}\boldsymbol{L}'\Big[(\boldsymbol{\Sigma}_{\boldsymbol{Y}_g}^{-1} - 2\boldsymbol{O}_{gi})' \otimes \boldsymbol{\Sigma}_{\boldsymbol{Y}_g}^{-1} \\ &\quad -\frac{1}{2}(1-\alpha_{gi})\text{vec}(\boldsymbol{O}_{gi})\text{vec}(\boldsymbol{O}_{gi})'\Big]\boldsymbol{L} \; \forall g,\end{aligned}$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{v}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_g})\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_j}))'} = -\frac{1}{4}\sum_{i=1}^{I} \alpha_{gi}\alpha_{ji}\boldsymbol{L}'\text{vec}(\boldsymbol{O}_{gi})\text{vec}(\boldsymbol{O}_{ji})'\boldsymbol{L} \; \forall g \neq j,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{v}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_g})\partial \boldsymbol{\mu}'_{\boldsymbol{X}_g}} = -\frac{1}{2}\sum_{i=1}^{I} \alpha_{gi}(1-\alpha_{gi})\boldsymbol{L}'\text{vec}(\boldsymbol{O}_{gi})\boldsymbol{f}'_{gi} \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{v}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_g})\partial \boldsymbol{\mu}'_{\boldsymbol{X}_j}} = \frac{1}{2}\sum_{i=1}^{I} \alpha_{gi}\alpha_{ji}\boldsymbol{L}'\text{vec}(\boldsymbol{O}_{gi})\boldsymbol{f}'_{ji} \; \forall g \neq j,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{v}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_g})\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_g}))'} = \frac{1}{4}\sum_{i=1}^{I} \alpha_{gi}(1-\alpha_{gi})\boldsymbol{L}'\text{vec}(\boldsymbol{O}_{gi})\text{vec}(\boldsymbol{F}_{gi})'\boldsymbol{J} \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{v}(\boldsymbol{\Sigma}_{\boldsymbol{Y}_g})\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_j}))'} = -\frac{1}{4}\sum_{i=1}^{I} \alpha_{gi}\alpha_{ji}\boldsymbol{L}'\text{vec}(\boldsymbol{O}_{gi})\text{vec}(\boldsymbol{F}_{ji})'\boldsymbol{J} \; \forall g \neq j,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\mu}_{\boldsymbol{X}_g}\partial \boldsymbol{\mu}'_{\boldsymbol{X}_g}} = -\sum_{i=1}^{I} \alpha_{gi}\left[\boldsymbol{\Sigma}_{\boldsymbol{X}_g}^{-1} - (1-\alpha_{gi})\boldsymbol{f}_{gi}\boldsymbol{f}'_{gi}\right] \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\mu}_{\boldsymbol{X}_g}\partial \boldsymbol{\mu}'_{\boldsymbol{X}_j}} = -\sum_{i=1}^{I} \alpha_{gi}\alpha_{ji}\boldsymbol{f}_{gi}\boldsymbol{f}'_{ji} \; \forall g \neq j,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\mu}_{\boldsymbol{X}_g}\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_g}))'} = -\sum_{i=1}^{I} \alpha_{gi}\left[(\boldsymbol{f}'_{gi} \otimes \boldsymbol{\Sigma}_{\boldsymbol{X}_g}^{-1}) + \frac{1}{2}(1-\alpha_{gi})\boldsymbol{f}_{gi}\text{vec}(\boldsymbol{F}_{gi})'\right]\boldsymbol{J} \; \forall g,$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\mu}_{\boldsymbol{X}_g}\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_j}))'} = \frac{1}{2}\sum_{i=1}^{I} \alpha_{gi}\alpha_{ji}\boldsymbol{f}_{gi}\text{vec}(\boldsymbol{F}_{ji})'\boldsymbol{J} \; \forall g \neq j,$$

$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_g})\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_g}))'} &= -\frac{1}{2}\sum_{i=1}^{I} \alpha_{gi}\boldsymbol{J}'\Big[(\boldsymbol{\Sigma}_{\boldsymbol{X}_g}^{-1} - 2\boldsymbol{F}_{gi})' \otimes \boldsymbol{\Sigma}_{\boldsymbol{X}_g}^{-1} \\ &\quad -\frac{1}{2}(1-\alpha_{gi})\text{vec}(\boldsymbol{F}_{gi})\text{vec}(\boldsymbol{F}_{gi})'\Big]\boldsymbol{J} \; \forall g,\end{aligned}$$

$$\frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_g})\partial (\text{v}(\boldsymbol{\Sigma}_{\boldsymbol{X}_j}))'} = -\frac{1}{4}\sum_{i=1}^{I} \alpha_{gi}\alpha_{ji}\boldsymbol{J}'\text{vec}(\boldsymbol{F}_{gi})\text{vec}(\boldsymbol{F}_{ji})'\boldsymbol{J} \; \forall g \neq j.$$

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement.

References

Basford, K. E., Greenway, D. R., McLachlan, G. J., & Peel, D (1997). Standard errors of fitted component means of normal mixtures. *Computational Statistics*, *12*, 1–17.

Boldea, O., & Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, *104*, 1539–1549.

Cellini, R., & Cuccia, T. (2013). Museum and monument attendance and tourism flow: A time series analysis approach. *Applied Economics*, *45*, 3473–3482.

Dang, U., Punzo, A., McNicholas, P., Ingrassia, S., & Browne, R. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, *34*, 4–34.

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

Di Mari, R., Bakk, Z., & Punzo, A. (2020). A random-covariate approach for distal outcome prediction with latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(3), 351–368.

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

Galimberti, G., Nuzzi, L., & Soffritti, G. (2021). Covariance matrix estimation of the maximum likelihood estimation in multivariate clusterwise linear regression. *Statistical Methods & Applications*, *30*, 235–268.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction*, 2nd edn. New York: Springer.

Henningsen, A., & Hamann, J. D. (2007). systemfit: A package for estimating systems of simultaneous equations in R. *Journal of Statistical Software*, *23*(4), 1–40.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, *103*, 103–118.

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

Ingrassia, S., Minotti, S. C., & Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, *71*, 159–182.

Ingrassia, S., Punzo, A., Vittadini, G., & Minotti, S.C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, *32*, 85–113.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics*. New York: Wiley.

Mazza, A., Punzo, A., & Ingrassia, S. (2018). flexCWM: A flexible framework for cluster-weighted models. *Journal of Statistical Software*, *86*(2), 1–30.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of econometrics* (Volume 4, Chapter 36, pp. 2111–2245).

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society: Series B*, *56*, 3–48.

Punzo, A., & Ingrassia, S. (2013). On the use of the generalized linear exponential cluster-weighted model to assess local linear independence in bivariate data. *QdS - Journal of Methodological and Applied Statistics*, *15*, 131–144.

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

Punzo, A., & Ingrassia, S. (2016). Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, *31*, 989–1013.
Punzo, A., & McNicholas, P. D. (2017). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification*, *34*, 249–293.
R Core Team. (2020). *R: a language and environment for statistical computing Vienna*. Austria: R Foundation for Statistical Computing.
Ritter, G. (2015). *Robust cluster analysis and variable selection*. Boca Raton: CRC Press.
Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), Article, 32.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A.E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8/1*, 205–223.
Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P.D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, *7*(1), 5–40.
Subedi, S., Punzo, A., Ingrassia, S., & McNicholas, P.D. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications*, *24*, 623–649.
White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.