Covariance matrix estimation of the maximum likelihood estimator in multivariate clusterwise linear regression

Noname manuscript No.
(will be inserted by the editor)

Covariance matrix estimation of the maximum likelihood estimator in multivariate clusterwise linear regression

Giuliano Galimberti, Lorenzo Nuzzi, Gabriele Soffritti

Received: date / Accepted: date

Abstract The expectation-maximisation algorithm is employed to perform maximum likelihood estimation in a wide range of situations, including regression analysis based on clusterwise regression models. A disadvantage of using this algorithm is that it is unable to provide an assessment of the sample variability of the maximum likelihood estimator. This inability is a consequence of the fact that the algorithm does not require deriving an analytical expression for the Hessian matrix, thus preventing from a direct evaluation of the asymptotic covariance matrix of the estimator. A solution to this problem when performing linear regression analysis through a multivariate Gaussian clusterwise regression model is developed. Two estimators of the asymptotic covariance matrix of the maximum likelihood estimator are proposed. In practical applications their use makes it possible to avoid resorting to bootstrap techniques and general purpose mathematical optimisers. The performances of these estimators are evaluated in analysing small simulated and real datasets; the obtained results illustrate their usefulness and effectiveness in practical applications. From a theoretical point of view, under suitable conditions, the proposed estimators are shown to be consistent.

Keywords EM algorithm · Gaussian mixture model · Hessian matrix · Sandwich estimator · Score vector

Mathematics Subject Classification (2010) MSC 62J99

1 Introduction

Switching regression, clusterwise regression and finite mixtures of regressions are terms used interchangeably to denote an approach to regression analy-

This is the accepted version of an article published by Springer: https://doi.org/10.1007/s10260-020-00523-9

sis in which the dependence of a $p \times 1$ random vector $\mathbf{Y}$ of responses on a $q \times 1$ vector $\mathbf{X}$ of predictors is described by means of a finite mixture (see, e.g., Frühwirth-Schnatter, 2006). Such an approach is mainly employed when sample observations come from populations composed of several subpopulations, each of which is characterised by a different regression model. Examples can be found in many fields, such as economics, marketing, agriculture, education, quantitative finance, social sciences and transport systems (see, e.g., Fair and Jaffe, 1972; Kamakura, 1988; Turner, 2000; Ding, 2006; Tashman and Frey, 2009; Dyer et al., 2012; Van Horn et al., 2015; McDonald et al., 2016; Elhenawy et al., 2017)). Mixtures of regression models naturally arise when a categorical or dummy predictor is omitted from a regression model (see, e.g., Hosmer, 1974). This approach can also be used in outlier detection or robust regression estimation; examples are provided by Aitkin and Tunnicliffe Wilson (1980), García-Escudero et al. (2010), García-Escudero et al. (2017) and Mazza and Punzo (2017).

When the p responses are absolutely continuous random variables, a finite mixture of Gaussian regressions is generally employed. This model has been extensively investigated. A detailed treatment of the identifiability when the response is univariate, both with random and fixed predictors, is given by Hennig (2000) (see also Frühwirth-Schnatter, 2006, section 8.2.2). Identifiability conditions for models with multivariate responses and random predictors under a finite Gaussian mixture for the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ can be found in Dang et al. (2017). Estimation has been developed in both the Bayesian and likelihood-based frameworks. With respect to this latter approach, model parameters are usually estimated through the maximum likelihood (ML) method by resorting to algorithms which are widely employed in incomplete-data problems: namely, the expectation-maximisation (EM) algorithm or some of its variants (see Faria and Soromenho, 2010, for a comparison in the special case of univariate response). In a clusterwise regression model the sample data $\{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, I\}$ are incomplete because the specific regression model 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 by-product of a linear regression analysis based on these methods is a set of estimated posterior probabilities that each sample observation come from the different regression models of the mixture. Thus, a by-product of this approach to linear regression analysis is a clustering of the I sample observations, based on a rule that assigns an observation to the regression model of the mixture from which it has the highest posterior probability of coming. Specific functions implementing the EM algorithm for a finite mixture of Gaussian regressions are available, for example, in some packages for the `R` software environment (R Core Team, 2019), such as `flexmix` (Grün and Leisch, 2008), `mixtools` (Benaglia et al., 2009) and `flexCWM` (Mazza et al., 2018).

A drawback of performing ML estimation through an EM algorithm is that this algorithm does not automatically produce an estimate of the covariance matrix of the ML estimator. The assessment of the sample variability of the parameter estimates in a clusterwise regression 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 predictor on a given response within any sub-population and tests for the significance of the difference between the effects of the same predictor on a given response in two different sub-populations. Examples of modern applications in which such methods could be fruitfully employed are the identification of significant differential effects of family environment on children's aggression (Dyer et al., 2012), children's trauma (Van Horn et al., 2015) and children's social skills (McDonald et al., 2016). Thus, additional computations are necessary to obtain an assessment of the sample variability of model parameter estimates. In a finite mixture model the asymptotic covariance matrix of the ML estimator is generally estimated by using the inverse of the observed information matrix obtained from the incomplete data log-likelihood (see, e.g., McLachlan and Peel, 2000, section 2.15). This matrix can be computed or approximated in several ways. One approach involves using the conditional moments of the gradient vector and the second-order derivative matrix of the complete data log-likelihood (Louis, 1982). In the case of independent and identically distributed observations only the conditional moments of the gradient vector are required, and an approximation to the observed information matrix is given by the so-called empirical observed information matrix (Meilijson, 1989). Another approach is based on a direct evaluation of the gradient vector and/or the second-order derivative matrix of the incomplete data log-likelihood (Boldea and Magnus, 2009). This second approach is analytically more complex than the first one because the incomplete data log-likelihood is given by a logarithm of a sum and, thus, its derivatives contain ratios. Furthermore, estimates of the standard errors of the ML estimator can be computed by bootstrap methods (see, e.g., Newton and Raftery, 1994; Basford et al., 1997; McLachlan and Peel, 2000). As far as the Gaussian clusterwise regression model is concerned, only a limited number of methods has been investigated. Namely, Arminger et al. (1999) describe an approximation of the observed information matrix which is based on the first derivatives of the complete data log-likelihood. The Louis' approach is developed in Turner (2000) when the response is univariate; the obtained solution is implemented in the `R` package `mixreg` (Turner, 2014). Packages `mixtools` and `flexmix` provide standard errors of the ML estimates that are computed by the parametric bootstrap and by resorting to a general purpose optimiser, respectively. As far as the intercepts and regression coefficients are concerned, approximated standard error estimates are also computed by the `flexCWM` package according to an approach in which a number of separate weighted linear regression analyses are carried out (one for each linear regression model of the mixture), where the sample observations are weighted with their estimated posterior probabilities of coming from the different regression models of the mixture.

In this paper the problem of assessing the sample variability of the parameter estimates in a Gaussian clusterwise linear regression model is investigated. Using arguments similar to the ones exploited in Boldea and Magnus (2009), two estimators of the asymptotic covariance matrix of the ML estimator are introduced: one is based on the second-order derivative matrix of the incomplete data log-likelihood; the other exploits a sandwich approach (see, e.g., White, 1982). The performances of the two estimators in the presence of small samples are studied by simulation and are compared to the performance of the parametric bootstrap-based estimator. Comparisons based on the analysis of two real datasets are also provided.

The key contributions of this paper are:

- the development of analytical expressions for the gradient vector (Theorem 1) and the second-order derivative matrix (Theorem 2) of the incomplete-data log-likelihood, which makes it feasible to compute the two asymptotic covariance matrix estimates of the ML estimator;
- a numerical evaluation of the performances of the proposed estimators in the presence of finite samples and their comparison with the parametric bootstrap-based estimator;
- some theoretical results about strong consistency of the two proposed estimators (Theorems 3 and 4).

It is important to note that the estimators examined in this paper have been previously studied under finite Gaussian mixture models (Boldea and Magnus, 2009). However, the treatment by Boldea and Magnus (2009) focuses on the estimation of the asymptotic covariance matrix without reporting any proof of the asymptotic properties considered here. Furthermore, since finite Gaussian mixture models do not account for dependences via covariates, whenever the main research interest is in capturing the effect of predictors on the responses, the solutions due to Boldea and Magnus (2009) cannot be obviously exploited. This latter task can be accomplished by means of the methods developed here.

The remainder of the paper is organised as follows. Section 2 contains the definition of multivariate Gaussian clusterwise linear regression model, the definition of the two estimators of the asymptotic covariance matrix of the ML estimator and the analytical expressions of the first and second-order derivative matrix of the incomplete data log-likelihood that are required for their computation. Appendices A, B, C and D provide details of the derivation of these analytical expressions. Section 3 summarises the experimental results of the simulation study. The comparison based on two real datasets is given in Section 4. The main asymptotic results are described in Section 5, along with the regularity conditions needed. The corresponding proofs are given in Appendices E, F and G. An additional document with supplementary material contains a detailed description of simulation study settings and tables with further results from the analyses described in Sections 3 and 4.

2 Finite mixtures of Gaussian regression models

2.1 Model specification and ML estimation

Let $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ be the $p \times 1$ random vector of absolutely continuous dependent variables and $\mathbf{X} = (X_1, \ldots, X_q)'$ the $q \times 1$ random vector of predictors, which may include continuous and/or dummy variables. A finite mixture of Gaussian regression models can be defined as follows.

Definition 1 *The random vector $\mathbf{Y}$ follows a finite mixture of Gaussian linear regression models of order K if the conditional density function of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ has the form*

$$f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi\left(\mathbf{y};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right),\ \mathbf{y} \in \mathbb{R}^p, \tag{1}$$

with

$$\boldsymbol{\mu}_k = \boldsymbol{\gamma}_k + \boldsymbol{\Pi}_k \mathbf{x},\ k = 1, \ldots, K, \tag{2}$$

$\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')' \in \boldsymbol{\Theta}$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K-1})'$ such that $\pi_k > 0$ $\forall k$, $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\theta}_k = \left(\boldsymbol{\gamma}_k', (\mathrm{vec}\left(\boldsymbol{\Pi}_k'\right))', (\mathrm{v}\left(\boldsymbol{\Sigma}_k\right))'\right)'$ and $\phi\left(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ denotes the p-dimensional Gaussian density function with expected mean value $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$.

The conditional density function $f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$ in equation (1) can be interpreted as a weighted average with weights π_k, $k = 1, \ldots, K$. The k-th component of this function represents a multivariate Gaussian linear regression model with a $p \times 1$ intercept vector $\boldsymbol{\gamma}_k$, a $p \times q$ regression coefficients matrix $\boldsymbol{\Pi}_k$ and a symmetric and positive definite covariance matrix $\boldsymbol{\Sigma}_k$. The definition of $\boldsymbol{\theta}_k$ involves the vec(.) and v(.) operators. Namely, vec($\mathbf{A}$) is the column vector obtained by stacking the columns of matrix $\mathbf{A}$ one underneath the other, and v($\mathbf{B}$) denotes the column vector obtained from vec($\mathbf{B}$) by eliminating all supradiagonal elements of a symmetric matrix $\mathbf{B}$ (thus, v($\mathbf{B}$) contains only the distinct elements of $\mathbf{B}$) (for more details see, e.g., Magnus and Neudecker, 1988). Hereafter the class of finite mixtures of Gaussian regression models just defined is denoted as $\mathfrak{F}_K = \{f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.

Similarly to any other finite mixture model, any changing in the labels used to distinguish the regression models that compose the mixture in equation (1) modifies the model parameter without changing the conditional density function of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$. This problem can be dealt with by imposing appropriate constraints on $\boldsymbol{\theta}$. A solution can be obtained by requiring that any two parameter vectors $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k'}$ differ in at least one element, which need not be the same for all the vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ (for further details see, e.g., Frühwirth-Schnatter, 2006, Section 1.3.3). A further problem with the regression mixture model (1) is non-identifiability due to potential overfitting (Frühwirth-Schnatter, 2006, Section 1.3.2). Conditions ensuring identifiability of models for clusterwise linear regression are provided in Hennig (2000). Hereafter it is assumed that those conditions are satisfied, so that the model class $\mathfrak{F}_K$ is identifiable.

Given a sample of I observations $(\mathbf{x}_1', \mathbf{y}_1')', \ldots, (\mathbf{x}_I', \mathbf{y}_I')'$, the incomplete data log-likelihood of the model (1) is equal to

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{I} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{I} \log \left(\sum_{k=1}^{K} \pi_k \phi \left(\boldsymbol{y}_i; \boldsymbol{\gamma}_k + \boldsymbol{\Pi}_k \boldsymbol{x}_i, \boldsymbol{\Sigma}_k \right) \right), \tag{3}$$

where $l_i(\boldsymbol{\theta}) = \log f(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\theta}) = \log \left(\sum_{k=1}^{K} \pi_k \phi \left(\boldsymbol{y}_i; \boldsymbol{\gamma}_k + \boldsymbol{\Pi}_k \boldsymbol{x}_i, \boldsymbol{\Sigma}_k \right) \right)$. The ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be computed using an EM algorithm (for details see Jones and McLachlan, 1992).

2.2 Covariance matrix estimation of the ML estimator

Two estimators of the covariance matrix $\mathrm{Cov}(\hat{\boldsymbol{\theta}})$ of the ML estimator of $\boldsymbol{\theta}$ are proposed. They are obtained by exploiting some general results concerning the ML estimator when the model is misspecified (White, 1982). In order to provide the definition of these estimators, some preliminaries have to be introduced. Let the model score vector and the Hessian matrix be $s(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and $H(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$, respectively. The score vector can also be expressed as $s(\boldsymbol{\theta}) = \sum_{i=1}^{I} s_i(\boldsymbol{\theta})$, where $s_i(\boldsymbol{\theta}) = \frac{\partial l_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. In a similar way, the Hessian matrix can be written as $H(\boldsymbol{\theta}) = \sum_{i=1}^{I} H_i(\boldsymbol{\theta})$, where $H_i(\boldsymbol{\theta}) = \frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$.

The two proposed estimators of $\mathrm{Cov}(\hat{\boldsymbol{\theta}})$ are defined as follows:

$$\widehat{\mathrm{Cov}}_1(\hat{\boldsymbol{\theta}}) = - \left(H(\hat{\boldsymbol{\theta}}) \right)^{-1}, \tag{4}$$

$$\widehat{\mathrm{Cov}}_2(\hat{\boldsymbol{\theta}}) = \left(H(\hat{\boldsymbol{\theta}}) \right)^{-1} \left(\sum_{i=1}^{I} s_i(\hat{\boldsymbol{\theta}}) \left(s_i(\hat{\boldsymbol{\theta}}) \right)' \right) \left(H(\hat{\boldsymbol{\theta}}) \right)^{-1}, \tag{5}$$

where $-H(\hat{\boldsymbol{\theta}})$ is the observed information matrix and $s_i(\hat{\boldsymbol{\theta}}) = \left. \frac{\partial l_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}$. The first estimator is computed from the Hessian matrix; the second estimator is based on the sandwich approach. In order to compute them, analytical expressions for $s(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ are needed. Here are details about how such expressions can be directly obtained from the incomplete-data log-likelihood $l(\boldsymbol{\theta})$ defined in equation (3).

Since the vector $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')'$, is composed of $K+1$ subvectors, $s(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ can be partitioned as follows:

$$s(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi}} \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} \\ \cdots \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_K} \end{pmatrix}, \; H(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}'} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi} \partial \boldsymbol{\theta}_1'} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi} \partial \boldsymbol{\theta}_K'} \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\pi}'} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_K'} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\pi}'} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\theta}_1'} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\theta}_K'} \end{pmatrix}.$$

The explicit forms of such quantities, as developed in this paper, require the introduction of some additional notation. Namely, let

$$f_{ki} = \pi_k \phi\left(\boldsymbol{y}_i; \boldsymbol{\gamma}_k + \boldsymbol{\Pi}_k \boldsymbol{x}_i, \boldsymbol{\Sigma}_k\right), \ \forall k, \forall i, \tag{6}$$

$$\alpha_{ki} = \frac{f_{ki}}{\left(\sum_{l=1}^{K} f_{li}\right)}, \ \forall k, \forall i, \tag{7}$$

$$\mathbf{a}_k = \frac{1}{\pi_k}\mathbf{e}_k, \ k = 1, \ldots, K-1, \tag{8}$$

$$\mathbf{a}_K = -\frac{1}{\pi_K}\mathbf{1}_{K-1},$$

where $\mathbf{e}_k$ denotes the k-th column of $\mathbf{I}_{K-1}$ (the identity matrix of order $K-1$) and $\mathbf{1}_{K-1}$ is the $(K-1) \times 1$ vector having each component equal to 1. The following vectors and matrices are also required:

$$\mathbf{b}_{ki} = \boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i), \ \forall k, \forall i, \tag{9}$$

$$\mathbf{B}_{ki} = \boldsymbol{\Sigma}_k^{-1} - \mathbf{b}_{ki}\mathbf{b}_{ki}', \ \forall k, \forall i. \tag{10}$$

Finally, let $\mathbf{G}$ be the duplication matrix, that is the unique matrix which transforms $\mathrm{v}(\mathbf{B})$ into $\mathrm{vec}(\mathbf{B})$ ($\mathbf{G}\mathrm{v}(\mathbf{B}) = \mathrm{vec}(\mathbf{B})$) (for more details see, e.g., Magnus and Neudecker, 1988).

The following theorems provide analytical expressions for $s(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$. Their proofs are available in Appendices A to D.

Theorem 1 *The subvectors that compose $s(\boldsymbol{\theta})$ are given by*

$$\frac{\partial l\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\pi}} = \sum_{i=1}^{I} \bar{\mathbf{a}}_i,$$

$$\frac{\partial l\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\gamma}_k} = \sum_{i=1}^{I} \alpha_{ki}\mathbf{b}_{ki}, \ \forall k,$$

$$\frac{\partial l\left(\boldsymbol{\theta}\right)}{\partial \mathrm{vec}(\boldsymbol{\Pi}_k')} = \sum_{i=1}^{I} \alpha_{ki}\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right), \ \forall k,$$

$$\frac{\partial l\left(\boldsymbol{\theta}\right)}{\partial \mathrm{v}(\boldsymbol{\Sigma}_k)} = -\frac{1}{2}\sum_{i=1}^{I} \alpha_{ki}\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right), \ \forall k,$$

where $\bar{\mathbf{a}}_i = \sum_{k=1}^{K} \alpha_{ki}\mathbf{a}_k$.

Theorem 2 *The submatrices that compose $H(\boldsymbol{\theta})$ are given by*

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}'} = -\sum_{i=1}^{I} \bar{\mathbf{a}}_i \bar{\mathbf{a}}_i',$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi} \partial \boldsymbol{\gamma}_k'} = \sum_{i=1}^{I} \alpha_{ki} \left(\mathbf{a}_k - \bar{\mathbf{a}}_i\right) \mathbf{b}_{ki}', \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi} \partial \left[\mathrm{vec}(\boldsymbol{\Pi}_k')\right]'} = \sum_{i=1}^{I} \alpha_{ki} \left(\mathbf{a}_k - \bar{\mathbf{a}}_i\right) \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right]', \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\pi} \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_k)\right]'} = -\frac{1}{2} \sum_{i=1}^{I} \alpha_{ki} \left(\mathbf{a}_k - \bar{\mathbf{a}}_i\right) \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]' \mathbf{G}, \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_k \partial \boldsymbol{\gamma}_k'} = -\sum_{i=1}^{I} \alpha_{ki} \left[\boldsymbol{\Sigma}_k^{-1} - (1-\alpha_{ki}) \mathbf{b}_{ki} \mathbf{b}_{ki}'\right], \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_k \partial \boldsymbol{\gamma}_l'} = -\sum_{i=1}^{I} \alpha_{ki} \alpha_{li} \mathbf{b}_{ki} \mathbf{b}_{li}', \ \forall k \neq l,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_k \partial \left[\mathrm{vec}(\boldsymbol{\Pi}_k')\right]'} = -\sum_{i=1}^{I} \alpha_{ki} \left\{\boldsymbol{\Sigma}_k^{-1} \otimes \mathbf{x}_i' + \right.$$
$$\left. - (1-\alpha_{ki}) \mathbf{b}_{ki} \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right]'\right\}, \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_k \partial \left[\mathrm{vec}(\boldsymbol{\Pi}_l')\right]'} = -\sum_{i=1}^{I} \alpha_{ki} \alpha_{li} \mathbf{b}_{ki} \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{li}'\right)\right]', \ \forall k \neq l,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_k \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_k)\right]'} = -\sum_{i=1}^{I} \alpha_{ki} \left\{\mathbf{b}_{ki}' \otimes \boldsymbol{\Sigma}_k^{-1} + \right.$$
$$\left. + \frac{1}{2} (1-\alpha_{ki}) \mathbf{b}_{ki} \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]'\right\} \mathbf{G}, \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_k \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_l)\right]'} = \frac{1}{2} \sum_{i=1}^{I} \alpha_{ki} \alpha_{li} \mathbf{b}_{ki} \left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]' \mathbf{G}, \ \forall k \neq l,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mathrm{vec}\left(\boldsymbol{\Pi}_k'\right) \partial \left[\mathrm{vec}\left(\boldsymbol{\Pi}_k'\right)\right]'} = -\sum_{i=1}^{I} \alpha_{ki} \left\{\boldsymbol{\Sigma}_k^{-1} \otimes \left(\mathbf{x}_i \mathbf{x}_i'\right) + \right.$$
$$\left. - (1-\alpha_{ki}) \, \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right]'\right\}, \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mathrm{vec}\left(\boldsymbol{\Pi}_k'\right) \partial \left[\mathrm{vec}\left(\boldsymbol{\Pi}_l'\right)\right]'} = -\sum_{i=1}^{I} \alpha_{ki} \alpha_{li} \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{li}'\right)\right]', \ \forall k \neq l,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mathrm{vec}\left(\boldsymbol{\Pi}_k'\right) \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_k)\right]'} = -\sum_{i=1}^{I} \alpha_{ki} \left\{\boldsymbol{\Sigma}_k^{-1} \otimes \left(\mathbf{x}_i \mathbf{b}_{ki}'\right) + \right.$$
$$\left. + \frac{1}{2} (1-\alpha_{ki}) \, \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]'\right\} \mathbf{G}, \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mathrm{vec}\left(\boldsymbol{\Pi}_k'\right) \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_l)\right]'} = \frac{1}{2}\sum_{i=1}^{I} \alpha_{ki}\alpha_{li} \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) \left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]' \mathbf{G}, \ \forall k \neq l,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mathrm{v}(\boldsymbol{\Sigma}_k) \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_k)\right]'} = -\frac{1}{2}\sum_{i=1}^{I} \alpha_{ki} \mathbf{G}' \left\{ \left(\boldsymbol{\Sigma}_k^{-1} - 2\mathbf{B}_{ki}\right)' \otimes \boldsymbol{\Sigma}_k^{-1} - \frac{1}{2}\left(1 - \alpha_{ki}\right) \mathrm{vec}\left(\mathbf{B}_{ki}\right) \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]' \right\} \mathbf{G}, \ \forall k,$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mathrm{v}(\boldsymbol{\Sigma}_k) \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_l)\right]'} = -\frac{1}{4}\sum_{i=1}^{I} \alpha_{ki}\alpha_{li} \mathbf{G}' \mathrm{vec}\left(\mathbf{B}_{ki}\right) \left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]' \mathbf{G}, \ \forall k \neq l.$$

3 Results from the analysis of simulated datasets

3.1 Settings, studies and scenarios

Two extensive Monte Carlo simulation studies are performed. In both studies, two scenarios are examined: correct model specification and model misspecification due to heteroscedasticity. In the first scenario of the first study the datasets are generated under models belonging to the class $\mathfrak{F}_2$ with $p = 2$ dependent variables and $q = 2$ predictors. Models from the class $\mathfrak{F}_3$ with $p = 2$ dependent variables and $q = 4$ predictors are employed to generate the datasets in the first scenario of the second study. The specific values of the model parameters $\boldsymbol{\theta}$ are reported in Section A of the supplementary material together with a detailed description of the procedure used to generate the datasets. Using each of those values, $R = 10000$ datasets (of size I) have been obtained. Assuming that the r-th dataset $\{(\mathbf{x}_{1r}', \mathbf{y}_{1r}')', \ldots, (\mathbf{x}_{Ir}', \mathbf{y}_{Ir}')'\}$ is generated from a model belonging to the true model class, the ML estimate $\hat{\boldsymbol{\theta}}_r$ of $\boldsymbol{\theta}$ is computed for $r = 1, \ldots, R$. Technical details about the ML estimation are reported in Section 3.2. Thus, in the first scenario of both studies the fitted models are correctly specified. In the second scenario the procedure for generating the datasets coincides with the one in the first scenario, but it makes use of matrices $\boldsymbol{\Sigma}_k$ that vary with the sample units. More specifically, variances and covariances are generated as a quadratic function of the sample values of the predictors (for more details see the supplementary material); thus, there is heteroscedasticity in the data. Models fitted to these datasets still assume that each sample data come from a model belonging to the classes $\mathfrak{F}_2$ and $\mathfrak{F}_3$ in the first and second studies, respectively. Thus, in the second scenario there is a form of model misspecification due to the heteroscedasticity. The total number of estimated parameters T is 19 in the first study, 41 in the second study.

In each examined scenario the above illustrated Monte Carlo experiments are performed twice, with samples of size $I = 250$ and $I = 500$. The R independent estimates of $\boldsymbol{\theta}$ computed in each scenario are used to approximate the true distribution of $\hat{\boldsymbol{\theta}}$. Namely, denoting the t-th element of $\hat{\boldsymbol{\theta}}_r$ as $\hat{\theta}_{rt}$, an

approximation of the true standard error of $\hat{\theta}_t$ is obtained as follows:

$$\mathrm{SE}(\hat{\theta}_t) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_{rt} - \bar{\theta}_t\right)^2}, \qquad \bar{\theta}_t = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_{rt}.$$

For $R_1 = 1000$ datasets obtained as described above, the estimated standard error of $\hat{\theta}_{rt}$ is computed using the parametric bootstrap and the two proposed information-based estimators. Specific `R` functions implementing these estimators have been written and employed. Namely, denoting the estimated standard error of $\hat{\theta}_{rt}$ as $\widehat{se}(\hat{\theta}_{rt})$, three different estimated standard errors of $\hat{\theta}_{rt}$ are obtained for $r = 1, \ldots, R_1$, $t = 1, \ldots, T$: $\widehat{se}_1(\hat{\theta}_{rt})$, $\widehat{se}_2(\hat{\theta}_{rt})$ and $\widehat{se}_B(\hat{\theta}_{rt})$, where $\widehat{se}_1(.)$ and $\widehat{se}_2(.)$ are the standard errors estimated using equations (4) and (5), respectively; $\widehat{se}_B(.)$ is the notation employed for the parametric-bootstrap estimated standard error. For each dataset, this latter estimation is given by the standard deviation of the parameter estimates over 100 bootstrap samples drawn from the estimated model. The performances of the three examined approaches for computing $\widehat{se}(\hat{\theta}_t)$ are evaluated on the basis of an estimate of their biases and (squared root) mean squared errors. Namely, for each approach the following quantities are computed:

$$\mathrm{BIAS}\left(\widehat{se}(\hat{\theta}_t)\right) = M_1\left(\widehat{se}(\hat{\theta}_t)\right) - \mathrm{SE}(\hat{\theta}_t),\ t = 1, \ldots, T,$$
$$\mathrm{RMSE}\left(\widehat{se}(\hat{\theta}_t)\right) = \sqrt{V\left(\widehat{se}(\hat{\theta}_t)\right) + \mathrm{BIAS}^2\left(\widehat{se}(\hat{\theta}_t)\right)},\ t = 1, \ldots, T,$$

where

$$M_1\left(\widehat{se}(\hat{\theta}_t)\right) = \frac{1}{R_1}\sum_{r=1}^{R_1}\widehat{se}(\hat{\theta}_{rt}), \qquad M_2\left(\widehat{se}(\hat{\theta}_t)\right) = \frac{1}{R_1}\sum_{r=1}^{R_1}\widehat{se}^2(\hat{\theta}_{rt}),$$
$$V\left(\widehat{se}(\hat{\theta}_t)\right) = M_2\left(\widehat{se}(\hat{\theta}_t)\right) - M_1^2\left(\widehat{se}(\hat{\theta}_t)\right).$$

A comparative evaluation of the three approaches for computing $\widehat{se}(\hat{\theta}_t)$ is also carried out through the coverage probabilities (CP) of confidence intervals based on the examined standard errors' estimates and the standard normal quantiles. Although biases, mean squared errors and coverage probabilities are estimated for all model parameters, in this section the comparison is focused on the regression coefficients (i.e. the parameters in $\boldsymbol{\Pi}_k$, $k = 1, \ldots, K$).

3.2 Technical details about ML estimation

As the `flexmix` package allows to estimate parameters of model (1) only when the K covariance matrices are diagonal, a more general function implementing the EM algorithm for the ML estimation has been developed in the `R` environment. In this function, the starting estimates of the model parameters are computed using the results obtained through the `flexmix` function of the `flexmix`

package. Thus, in the ML estimation process considered here, the p responses are initially assumed to be conditionally independent within each component of the mixture. As far as the convergence of the implemented EM algorithm is concerned, the following criteria have been exploited. The EM algorithm is stopped when the number of iterations reaches 500 or $|l_\infty^{(r+1)} - l^{(r)}| < 10^{-8}$, where $l^{(r)}$ is the log-likelihood value from iteration r, and $l_\infty^{(r+1)}$ is the asymptotic estimate of the log-likelihood at iteration $r+1$ (Dang and McNicholas, 2015). In order to avoid difficulties arising when the estimates of the K covariance matrices at any iteration of the EM algorithm are singular or nearly singular, in the EM algorithm implemented here all such estimated 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}.

3.3 Results from the first Monte Carlo study

The results obtained in the first scenario (correct specification) with $I = 250$ have been summarised in a graphical form in Figure 1. Each panel of this figure refers to a specific regression coefficient, denoted as $\boldsymbol{\Pi}_{k(jl)}$, with $k = 1, 2$, $j = 1, 2$ and $l = 1, 2$. The dashed line identifies the (approximated) true standard error of the corresponding ML estimate ($\mathrm{SE}(\hat{\boldsymbol{\Pi}}_{k(jl)})$), whereas the three black points represent the values of $M_1\left(\widehat{se}(\hat{\theta}_t)\right)$ obtained with the three estimators. Thus, this part of the plots makes it possible to graphically represent the bias associated with the use of each estimator. In order to provide a graphical representation of the root mean squared errors, the quantities $\mathrm{SE}(\hat{\boldsymbol{\Pi}}_{k(jl)}) \pm RMSE\left(\widehat{se}(\hat{\boldsymbol{\Pi}}_{k(jl)})\right)$ have been computed for each estimator and horizontal lines have been drawn at each of these values. These horizontal lines provide a summary description of the variability of each estimator around the true standard error. Table A of the supplementary material shows the numerical values used to draw Figure 1. It is worth noting that such values have been multiplied by 100 to facilitate presentation. Biases and root mean square errors are generally small. By focusing the attention on the bias, it is typically negative. As far as the regression coefficients are concerned, the lowest absolute bias in estimating the standard error is registered using the bootstrap approach; for the other model parameters (results are not shown) the best performance in terms of bias is obtained with the method based on the Hessian matrix. Moving the attention on the accuracy in estimating the standard error of the model parameters, the lowest root mean square errors are mostly obtained using the Hessian-based method. The bootstrap approach is slightly more accurate than the sandwich method in estimating the standard errors of the regression coefficients. The effective confidence levels computed using the three examined methods (Table 1) are very similar to one another and are also quite close to the nominal ones. In particular, only two null hypotheses of equality between the effective and the nominal confidence levels should be

Table 1: Coverage probability of the 90% and 95% confidence intervals for $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$, based on the three examined standard error estimators of $\hat{\boldsymbol{\theta}}$ in the first scenario (correct model specification) with $I = 250$. Coverage probabilities significantly differing from the corresponding nominal ones (two-tailed normal test with Bonferroni-corrected $\alpha = 0.00125$) are reported in italics.

	CP 90%			CP 95%		
θ_t	$\widehat{se}_B(\hat{\theta}_t)$	$\widehat{se}_1(\hat{\theta}_t)$	$\widehat{se}_2(\hat{\theta}_t)$	$\widehat{se}_B(\hat{\theta}_t)$	$\widehat{se}_1(\hat{\theta}_t)$	$\widehat{se}_2(\hat{\theta}_t)$
$\boldsymbol{\Pi}_{1(11)}$	0.913	0.917	0.911	0.957	0.960	0.952
$\boldsymbol{\Pi}_{1(12)}$	0.891	0.889	0.884	0.942	0.942	0.938
$\boldsymbol{\Pi}_{1(21)}$	0.884	0.883	0.888	0.938	0.936	0.931
$\boldsymbol{\Pi}_{1(22)}$	0.881	0.882	0.873	0.931	0.935	*0.927*
$\boldsymbol{\Pi}_{2(11)}$	0.877	0.876	0.875	0.936	0.938	*0.928*
$\boldsymbol{\Pi}_{2(12)}$	0.904	0.908	0.894	0.954	0.953	0.943
$\boldsymbol{\Pi}_{2(21)}$	0.888	0.892	0.878	0.950	0.948	0.934
$\boldsymbol{\Pi}_{2(22)}$	0.877	0.879	0.875	0.944	0.936	0.938

rejected, according to asymptotic two-tailed normal tests for a proportion at a Bonferroni-corrected 0.01/8=0.00125 significance level. Namely, these two null hypotheses refer to confidence intervals obtained using the sandwich approach (see nominal confidence levels in italics in Table 1). The Bonferroni correction is applied to account for multiple tests performed for each estimation method and each nominal confidence level. The results obtained with $I = 500$ (see Tables B and C in the supplementary material) are quite similar. Increasing the sample size leads to a reduction in the gap between the RMSE of the bootstrap and sandwich approaches.

When the fitted models assume homoscedasticity while data are heteroscedastic (second scenario), biases and root mean square errors increase for all estimated standard errors (see Figure 2 and Table D in the supplementary material). The lowest absolute biases are mostly obtained through the sandwich method. This method also provides the most accurate estimates of the standard error for the majority of the model parameters. As far as the confidence intervals for the regression coefficients are concerned (Table 2), they are conservative with all the examined estimators. It is worth mentioning that all effective coverage probabilities of the confidence intervals appear to be significantly different from the corresponding nominal one, according to asymptotic two-tailed normal tests for a proportion at a Bonferroni-corrected 0.01/8=0.00125 significance level. In particular, the effective coverage probabilities are remarkably lower than the nominal ones when using the bootstrap and the Hessian approaches. Such reduction is less evident with the sandwich approach, that allows to build intervals whose effective confidence levels are the closest to the corresponding nominal ones. Similar results have been obtained with $I = 500$ (see Tables E and F in the supplementary material).

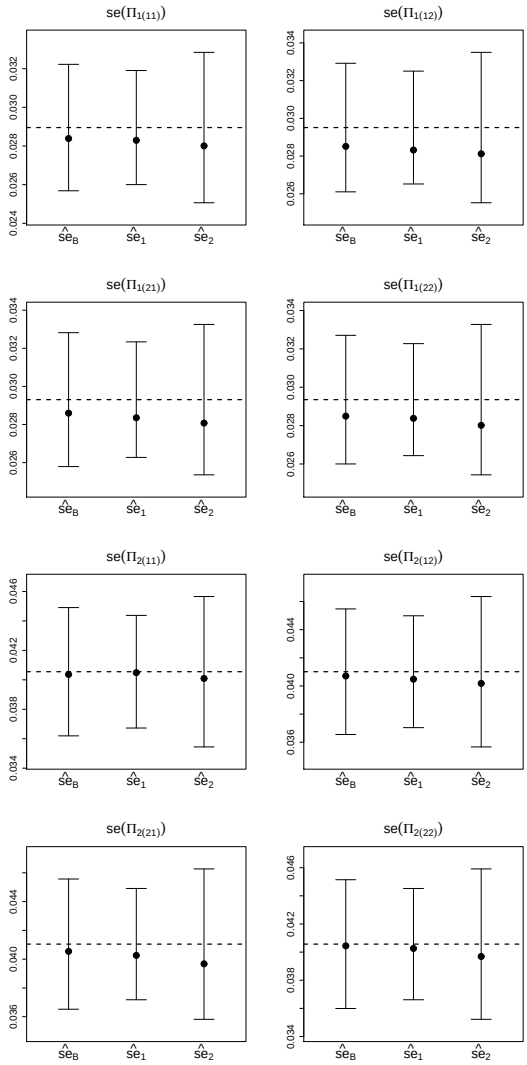

Fig. 1 Graphical representation of the biases and root mean square errors of the three examined estimators of SE($\hat{\boldsymbol{\Pi}}_{k(jl)}$) in the first scenario (correct model specification) with $I = 250$. The dashed line identifies SE($\hat{\boldsymbol{\Pi}}_{k(jl)}$). For each estimator, the black point represents the values of $M_1\left(\widehat{se}(\hat{\boldsymbol{\Pi}}_{k(jl)})\right)$ and the two wiskers are set to the values SE($\hat{\boldsymbol{\Pi}}_{k(jl)}$) $\pm$ $RMSE\left(\widehat{se}(\hat{\boldsymbol{\Pi}}_{k(jl)})\right)$.

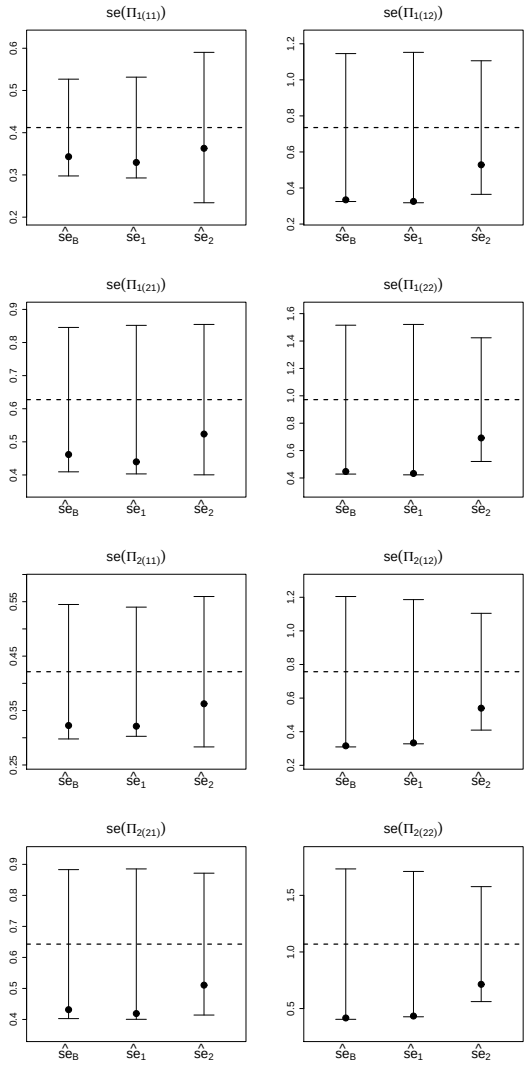

Fig. 2 Graphical representation of the biases and root mean square errors of the three examined estimators of $\mathrm{SE}(\hat{\boldsymbol{\Pi}}_{k(jl)})$ in the second scenario (misspecification due to heteroscedasticity) with $I = 250$. The dashed line identifies $\mathrm{SE}(\hat{\boldsymbol{\Pi}}_{k(jl)})$. For each estimator, the black point represents the values of $M_1\left(\widehat{se}(\hat{\boldsymbol{\Pi}}_{k(jl)})\right)$ and the two wiskers are set to the values $\mathrm{SE}(\hat{\boldsymbol{\Pi}}_{k(jl)}) \pm RMSE\left(\widehat{se}(\hat{\boldsymbol{\Pi}}_{k(jl)})\right)$.

Table 2: Coverage probability of the 90% and 95% confidence intervals for $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$, based on the three examined standard error estimators of $\hat{\boldsymbol{\theta}}$ in the second scenario (misspecification due to heteroscedasticity) with $I = 250$.Coverage probabilities significantly differing from the corresponding nominal ones (two-tailed normal test with Bonferroni-corrected $\alpha = 0.00125$) are reported in italics.

	CP 90%			CP 95%		
θ_t	$\widehat{se}_B(\hat{\theta}_t)$	$\widehat{se}_1(\hat{\theta}_t)$	$\widehat{se}_2(\hat{\theta}_t)$	$\widehat{se}_B(\hat{\theta}_t)$	$\widehat{se}_1(\hat{\theta}_t)$	$\widehat{se}_2(\hat{\theta}_t)$
$\boldsymbol{\Pi}_{1(11)}$	*0.823*	*0.808*	*0.825*	*0.878*	*0.874*	*0.891*
$\boldsymbol{\Pi}_{1(12)}$	*0.759*	*0.739*	*0.803*	*0.829*	*0.813*	*0.861*
$\boldsymbol{\Pi}_{1(21)}$	*0.557*	*0.538*	*0.746*	*0.638*	*0.626*	*0.815*
$\boldsymbol{\Pi}_{1(22)}$	*0.572*	*0.570*	*0.748*	*0.652*	*0.649*	*0.811*
$\boldsymbol{\Pi}_{2(11)}$	*0.754*	*0.753*	*0.792*	*0.832*	*0.836*	*0.854*
$\boldsymbol{\Pi}_{2(12)}$	*0.721*	*0.708*	*0.770*	*0.796*	*0.780*	*0.834*
$\boldsymbol{\Pi}_{2(21)}$	*0.422*	*0.437*	*0.633*	*0.495*	*0.515*	*0.718*
$\boldsymbol{\Pi}_{2(22)}$	*0.423*	*0.443*	*0.635*	*0.497*	*0.510*	*0.717*

3.4 Results from the second Monte Carlo study

The results from the second study are in line with those of the first study. In particular, with correctly specified models, the bootstrap-based estimator and the estimator based on the Hessian matrix have the lowest absolute biases and the lowest root mean square errors, respectively. Such estimators also register quite similar performances with respect to the coverage probabilities of the confidence intervals. When there is misspecification due to heteroscedasticity, the sandwich method has the lowest absolute biases and is also the most accurate; furthermore, it allows to compute confidence intervals whose effective levels are the closest to the nominal ones. All tables with these results are reported in Section B.2 of the supplementary material.

4 Results from the analysis of real datasets

4.1 Aphids data

The aphids dataset (Boiteau et al., 1998) provides information collected from 51 independent experiments performed to study the spread of a viral infection among potato plants. In the experiments, varying numbers of aphids were released in a flight chamber containing 81 tobacco plants (69 healthy, 12 infected) arranged in a 9×9 grid, and the number of healthy plants that resulted to be infected after an exposition of 24 hours was recorded. For each experiment, the dataset indicates the number of aphids released in the flight chamber (X_1) and the resulting number (out of a possible 69) of infected plants (Y_1). This dataset is available in the `R` package `mixreg` (Turner, 2014); the results of a linear regression analysis carried out through a mixture of two simple univariate Gaussian linear regression models are reported in Turner (2000).

Table 3: Parameter estimates $\hat{\boldsymbol{\theta}}$ and standard error estimates of $\hat{\boldsymbol{\theta}}$ in the aphids dataset.

θ_t	$\hat{\theta}_t$	$\widehat{se}_B(\hat{\theta}_t)$	$\widehat{se}_1(\hat{\theta}_t)$	$\widehat{se}_2(\hat{\theta}_t)$
π_1	0.5016	0.0930	0.0803	0.0796
$\boldsymbol{\gamma}_1$	3.4745	1.2961	1.0704	0.9922
$\boldsymbol{\Pi}_1$	0.0553	0.0074	0.0065	0.0073
$\boldsymbol{\Sigma}_1$	9.7051	2.7937	3.0131	2.4009
$\boldsymbol{\gamma}_2$	0.8586	0.7039	0.3678	0.2778
$\boldsymbol{\Pi}_2$	0.0024	0.0035	0.0025	0.0023
$\boldsymbol{\Sigma}_2$	1.2653	1.0607	0.4076	0.4179

Estimates of the standard error for the model parameter estimates, based on the second-order derivatives of the complete log-likelihood, are also provided.

Table 3 reports the parameter estimates (obtained through the `R` function summarised in Section 3.2) of the same model examined in Turner (2000) together with their standard errors estimated using both equations (4) and (5). For comparison purposes also parametric bootstrap-based estimated standard errors are provided; they are computed from 100 independent samples drawn from the mixture of two Gaussian linear regression models with parameters fixed at their estimated values. The estimates of the standard error obtained using the second order derivatives are in accordance to the finding reported in Turner (2000). The sandwich standard errors are somewhat smaller than the bootstrap ones for all parameter estimates; the same result holds for four of the seven parameters when comparing $\widehat{se}_2(\hat{\theta}_t)$ with $\widehat{se}_1(\hat{\theta}_t)$. The latter result may be linked to the fact that a mixture of two binomial logistic regressions would be a more plausible model for this dataset and, thus, the analysis suffers from problems of heteroschedasticity due to model misspecification. Approximated 95% confidence intervals for the model parameters are also computed by exploiting the asymptotic Gaussian distribution for the ML estimator (see Table O in the supplementary material). The three interval estimates of the regression coefficient $\boldsymbol{\Pi}_2$ contain the zero value, thus suggesting that the null hypothesis $H_0 : \boldsymbol{\Pi}_2 = 0$ should not be rejected at a 5% significance level. The bootstrap intervals for $\boldsymbol{\gamma}_2$ and $\boldsymbol{\Sigma}_2$ are much larger than the ones obtained using the two information-based estimation methods.

4.2 Canned tuna data

The canned tuna dataset (Chevalier *et al.*, 2003) contains the weekly sales for seven of the top 10 U.S. brands in the canned tuna product category for $I = 338$ weeks between September 1989 and May 1997. Furthermore, a measure of the display activity and the log price of each brand is provided. This dataset is available within the `R` package `bayesm` (Rossi, 2019). In this application, the focus is on the evaluation of the effect of prices and promotional activities on sales for two brands: Star Kist 6 oz. and Bumble Bee Solid 6.12 oz. The analysed dataset contains information about the following variables:

Table 4: Maximised log-likelihood and Bayesian information criterion of four Gaussian linear regression models fitted to the canned tuna dataset.

K	$l_I(\hat{\boldsymbol{\theta}})$	BIC
1	−646.7672	−1369.2340
2	−271.8119	−700.8461
3	−210.7231	−660.1911
4	−187.6005	−695.4686

$\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$ and $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$, where Y_{i1}, X_{i1} and X_{i2} are the log unit sale, a measure of the display activity and the log price, respectively, registered in week i for Star Kist 6 oz.; Y_{i2}, X_{i3} and X_{i4} provide the same information for Bumble Bee Solid 6.12 oz. This same dataset was previously analysed in Galimberti et al. (2016) by means of mixtures of constrained Gaussian linear regression models, in which the effects of both the display activity and the log price were assumed to be equal across the mixture regression models; it was also assumed that the display activity and the log price for a brand do not affect the sales of the other brand.

Mixtures of Gaussian linear regression models of order $K = 1, 2, 3, 4$ are estimated through the `R` function summarised in Section 3.2. Table 4 shows the values of the maximised log-likelihood and the Bayesian information criterion (BIC) (Schwarz, 1978) for the four estimated models, where $BIC = 2l(\hat{\boldsymbol{\theta}}) - npar \ln(I)$, with $l(\hat{\boldsymbol{\theta}})$ and $npar$ denoting the maximum value of $l(\boldsymbol{\theta})$ and number of estimated parameters in the model, respectively. Thus, according to this criterion, the best solution is obtained with a mixture of three Gaussian linear regression models. The same order was also obtained for the constrained Gaussian linear regression mixture selected in Galimberti et al. (2016).

The estimates of the regression coefficients for such a mixture are reported in the second column of Table 5. This table also reports the standard errors of $\hat{\boldsymbol{\theta}}$ estimated using the two solutions described in Section 2.2 and the parametric bootstrap with 100 bootstrap samples. The same information for $\boldsymbol{\pi}$, $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$, and $\boldsymbol{\gamma}_3$ is available in Table P of the supplementary material. For this dataset, the three standard error estimators lead to quite similar results. For ten out of the 24 regression coefficients, the null hypothesis $H_0 : \boldsymbol{\Pi}_{k(jl)} = 0$ should not be rejected according to all compared strategies (see Table Q in the supplementary material). Contrary to what was assumed in the constrained mixture regression model examined by Galimberti et al. (2016), the log price for the brand Star Kist 6 oz. (X_2) seems to have a significant positive effect on the log sales of the brand Bumble Bee Solid 6.12 oz. (Y_2) in two of the three regression models of the selected mixture (see the $14th$ and $22nd$ rows in Tables 5 and Q). Moreover, the effect of the display activity for the brand Star Kist 6 oz. (X_1) on the sales of the same brand (Y_1) appears to be significant only in the second regression model of the mixture (see the first, ninth and $17th$ rows in Tables 5 and Q). This latter result suggests that the assumption of equal effects of the display activity across all mixture regression models could be inadequate. Finally, an interesting feature emerging from both the

Table 5: Estimates of the regression coefficients and their standard error estimates in the canned tuna dataset.

θ_t	$\hat{\theta}_t$	$\widehat{se}_B(\hat{\theta}_t)$	$\widehat{se}_1(\hat{\theta}_t)$	$\widehat{se}_2(\hat{\theta}_t)$
$\boldsymbol{\Pi}_{1(11)}$	−0.2192	0.4098	0.3491	0.4076
$\boldsymbol{\Pi}_{1(12)}$	−3.5468	1.0798	1.0204	1.1287
$\boldsymbol{\Pi}_{1(13)}$	0.2991	0.4425	0.3585	0.3258
$\boldsymbol{\Pi}_{1(14)}$	1.0538	3.3366	3.2542	3.2486
$\boldsymbol{\Pi}_{1(21)}$	−0.2264	0.1274	0.0852	0.1281
$\boldsymbol{\Pi}_{1(22)}$	−0.2688	0.4328	0.2602	0.3718
$\boldsymbol{\Pi}_{1(23)}$	0.1251	0.2195	0.0677	0.0492
$\boldsymbol{\Pi}_{1(24)}$	−3.2157	0.8607	0.6439	0.4278
$\boldsymbol{\Pi}_{2(11)}$	0.2990	0.0620	0.0689	0.0756
$\boldsymbol{\Pi}_{2(12)}$	−3.0103	0.1756	0.2439	0.3627
$\boldsymbol{\Pi}_{2(13)}$	−0.2804	0.0670	0.0710	0.0735
$\boldsymbol{\Pi}_{2(14)}$	−1.8009	0.5659	0.7239	0.9461
$\boldsymbol{\Pi}_{2(21)}$	0.0929	0.0560	0.0566	0.0663
$\boldsymbol{\Pi}_{2(22)}$	0.4128	0.1433	0.1717	0.2004
$\boldsymbol{\Pi}_{2(23)}$	0.1017	0.0538	0.0526	0.0528
$\boldsymbol{\Pi}_{2(24)}$	−4.1043	0.4953	0.5204	0.7092
$\boldsymbol{\Pi}_{3(11)}$	0.0869	0.3035	0.3398	0.6105
$\boldsymbol{\Pi}_{3(12)}$	−4.9454	0.9617	0.4628	0.4033
$\boldsymbol{\Pi}_{3(13)}$	0.0978	0.3801	0.3455	0.5889
$\boldsymbol{\Pi}_{3(14)}$	3.1429	3.0822	3.7078	6.6238
$\boldsymbol{\Pi}_{3(21)}$	1.0053	0.6983	0.7519	0.7775
$\boldsymbol{\Pi}_{3(22)}$	4.2550	2.0007	1.0455	0.6998
$\boldsymbol{\Pi}_{3(23)}$	2.6237	0.7476	0.7042	0.8088
$\boldsymbol{\Pi}_{3(24)}$	−18.4834	7.1063	6.6339	7.2836

constrained and unconstrained linear regression mixtures is that the smallest cluster of weeks detected by these models comprises 17 consecutive weeks in which Bumble Bee Solid 6.12 oz. tuna sales registered a relatively low mean level. These weeks correspond to the period from mid-October 1990 to mid-February 1991, just after a worldwide boycott campaign promoted against Bumble Bee tuna because Bumble Bee was found to be buying yellow-fin tuna caught by dolphin-unsafe techniques (Baird and Quastel, 2011).

5 Asymptotic results

The consistency of the estimators developed in Section 2.2 can be easily proved when some assumptions and regularity conditions are introduced with respect to the joint distribution of the regressors and dependent variables. This allows to exploit some general results about the asymptotic behaviour of the ML estimator when the model is misspecified (White, 1982).

Let $\mathbf{Z}_i = (\mathbf{X}_i', \mathbf{Y}_i')'$. In this Section the following model for the joint density function of $\mathbf{Z}_i$ is considered:

$$h(\mathbf{z}_i; \boldsymbol{\psi}) = q(\mathbf{x}_i; \boldsymbol{\vartheta}) f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}), \ \mathbf{z}_i = (\mathbf{x}_i', \mathbf{y}_i')' \in \mathbb{R}^{G+D}, \tag{11}$$

for some $\boldsymbol{\psi} = (\boldsymbol{\vartheta}', \boldsymbol{\theta}')' \in \boldsymbol{\Psi} = \boldsymbol{\Upsilon} \times \boldsymbol{\Theta}$. The parametric function $q(\mathbf{x}; \boldsymbol{\vartheta}) = dQ(\mathbf{x}; \boldsymbol{\vartheta}) / d\mu$ represents the Radon-Nikodym density of $Q(\mathbf{x}; \boldsymbol{\vartheta})$, the marginal

distribution of $\mathbf{X}$, with $Q(\mathbf{x};\boldsymbol{\vartheta}) \in \mathcal{B} = \{Q(\mathbf{x};\boldsymbol{\vartheta}); \boldsymbol{\vartheta} \in \boldsymbol{\Upsilon}\}$. The conditional density function of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ is given in equation (1). Hereafter the class of finite mixtures of Gaussian linear regression models with random predictors just defined is denoted as $\mathfrak{F}_K^* = \{h(\mathbf{z};\boldsymbol{\psi}), \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$, where K is the order of the mixture model (1). Identifiability conditions for finite mixtures of univariate Gaussian linear regression models with random predictors are discussed in Hennig (2000), and can be easily extended to the multivariate scenario. In this Section it is assumed that such conditions are satisfied, in order to guarantee the identifiability of the model class $\mathfrak{F}_K^*$. Note that in this model, the parameter vector $\boldsymbol{\psi}$ is partitioned into two non-overlapping subvectors: the first subvector is related to the marginal distribution of $\mathbf{X}$, while the second one contains the parameters of the conditional distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$. This partition of the parameter vector $\boldsymbol{\psi}$ has an impact on the properties of $\hat{\boldsymbol{\theta}}$. Namely, this estimator results to be independent of the ML estimator of $\boldsymbol{\vartheta}$. Section 5.2 shows how to exploit this independence to derive some consistency results for the covariance matrix estimators described in Section 2.2, starting from some asymptotic results for the ML estimator of $\boldsymbol{\psi}$. This implies that $\boldsymbol{\vartheta}$ does not need to be estimated (unless it is of independent interest).

5.1 Regularity conditions

Here is the list of (standard) regularity conditions for the existence, consistency and asymptotic normality of the ML estimator of the parameter characterising the family $\mathfrak{F}_K^* = \{h(\mathbf{z};\boldsymbol{\psi}) = q(\mathbf{x};\boldsymbol{\vartheta})f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}), \boldsymbol{\psi} = (\boldsymbol{\vartheta},\boldsymbol{\theta}) \in \boldsymbol{\Psi} = \boldsymbol{\Upsilon} \times \boldsymbol{\Theta}\}$ of (possibly misspecified) finite mixtures of Gaussian regression models with random predictors. These conditions coincide with those introduced by White (1982) and are formulated with respect to random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_i, \ldots, \mathbf{Z}_I$ with $\mathbf{Z}_i = (\mathbf{X}_i, \mathbf{Y}_i)$ and by assuming that $\mathbf{Z}_i$ takes values on a measurable Euclidean space Ω $\forall i$ and that the model parameter $\boldsymbol{\psi}$ is a vector of length d. Matrices $\mathbf{A}(\boldsymbol{\psi})$ and $\mathbf{B}(\boldsymbol{\psi})$ in condition (C6) are defined as follows:

$$\mathbf{A}(\boldsymbol{\psi}) = \mathbb{E}\left(\frac{\partial^2 \log h(\mathbf{z}_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right),$$
$$\mathbf{B}(\boldsymbol{\psi}) = \mathbb{E}\left(\left(\frac{\partial \log h(\mathbf{z}_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right)\left(\frac{\partial \log h(\mathbf{z}_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right)'\right).$$

(C1) $\mathbf{Z}_1, \ldots, \mathbf{Z}_i, \ldots, \mathbf{Z}_I$ are independent and identically distributed random vectors with common joint distribution function G and measurable Radon-Nikodym density $g = dG/d\nu$.

(C2) The parametric family of distribution functions $H(\mathbf{z},\boldsymbol{\psi})$ specified for $\mathbf{Z}_i$ has Radon-Nikodym densities $h(\mathbf{z};\boldsymbol{\psi}) = dH(\mathbf{z},\boldsymbol{\psi})/d\mu$ that are equal to $q(\mathbf{x};\boldsymbol{\vartheta})f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$; these densities are measurable in $\mathbf{z}$ $\forall\boldsymbol{\psi} \in \bar{\boldsymbol{\Psi}}$, and continuous in $\boldsymbol{\psi}$ $\forall\mathbf{z} \in \Omega$, where $\bar{\boldsymbol{\Psi}} = \bar{\boldsymbol{\Upsilon}} \times \bar{\boldsymbol{\Theta}}$ is a compact subset of $\boldsymbol{\Psi} \subset \mathbb{R}^d$.

(C3) (a) $\mathbb{E}\left(\log g(\mathbf{Z}_i)\right)$ exists and $|\log h(\mathbf{z};\boldsymbol{\psi})| \leq m(\mathbf{z})\ \forall \boldsymbol{\psi} \in \bar{\boldsymbol{\Psi}}$, where m is integrable with respect to G; (b) $\mathbb{E}\left(\log \frac{g(\mathbf{Z})}{h(\mathbf{Z};\boldsymbol{\psi})}\right)$ has a unique minimum at $\breve{\boldsymbol{\psi}} \in \bar{\boldsymbol{\Psi}}$.

(C4) Derivatives $\frac{\partial \log h(\mathbf{z};\boldsymbol{\psi})}{\partial \psi_j}$, $j = 1, \ldots, d$, are measurable functions of $\mathbf{z}\ \forall \boldsymbol{\psi} \in \bar{\boldsymbol{\Psi}}$, and continuously differentiable functions of $\boldsymbol{\psi}\ \forall \mathbf{z} \in \Omega$.

(C5) Functions $|\frac{\partial^2 \log h(\mathbf{z};\boldsymbol{\psi})}{\partial \psi_j \partial \psi_{j'}}|$ and $\frac{\partial \log h(\mathbf{z};\boldsymbol{\psi})}{\partial \psi_j} \cdot \frac{\partial \log h(\mathbf{z};\boldsymbol{\psi})}{\partial \psi_{j'}}$, $j, j' = 1, \ldots, d$, are dominated by functions integrable with respect to $G\ \forall \mathbf{z} \in \Omega$ and $\forall \boldsymbol{\psi} \in \bar{\boldsymbol{\Psi}}$.

(C6) (a) $\breve{\boldsymbol{\psi}}$ is interior to $\bar{\boldsymbol{\Psi}}$; (b) the matrix $\mathbf{B}(\breve{\boldsymbol{\psi}})$ is nonsingular; (c) $\breve{\boldsymbol{\psi}}$ is a value for $\boldsymbol{\psi}$ such that the matrix $\mathbf{A}(\boldsymbol{\psi})$ has constant rank in some open neighborhood of $\breve{\boldsymbol{\psi}}$.

(C7) $|\partial[\frac{\partial h(\mathbf{z};\boldsymbol{\psi})}{\partial \psi_j} \cdot h(\mathbf{z};\boldsymbol{\psi})]/\partial \psi_{j'}|$, $j, j' = 1, \ldots, d$, are dominated by functions integrable with respect to $\nu\ \forall \boldsymbol{\psi} \in \bar{\boldsymbol{\Psi}}$, and the minimal support of $h(\mathbf{z};\boldsymbol{\psi})$ does not depend on $\boldsymbol{\psi}$.

Condition (C2) requires compactness of the parameter space. This restriction has been exploited in several papers providing asympotic results related to mixtures of regression models (see, e.g., Städler et al., 2010; Yao et al., 2011; Tang and Karunamuni, 2013). To guarantee compactness of $\bar{\boldsymbol{\Psi}}$ it is necessary that both $\bar{\boldsymbol{\Upsilon}}$ and $\bar{\boldsymbol{\Theta}}$ are compact. Similarly to Maugis et al. (2009), compactness of $\bar{\boldsymbol{\Theta}}$ can be achieved under the following constraints on the parameters of equation (1):

(co1) $(\pi_1, \ldots \pi_K)' \in \mathcal{P}(K)$, where $\mathcal{P}(K) = \{\mathbf{p} = (p_1, \ldots, p_K)' \in [0,1]^K : \sum_{k=1}^K p_k = 1\}$;

(co2) $\boldsymbol{\gamma}_k \in \mathcal{A}(\epsilon, p)\ \forall k$, where $\mathcal{A}(\epsilon, r) = \{\mathbf{a} \in \mathbb{R}^r : \|\mathbf{a}\| \leq \epsilon\}$, $0 < \epsilon < \infty$ and $\|\cdot\|$ is the Euclidean norm;

(co3) $\boldsymbol{\Pi}_k \in \mathcal{B}(\rho, p, q)\ \forall k$, where $\mathcal{B}(\rho, p, q) = \{\mathbf{B} \in \mathcal{M}_{p\times q} : |||\mathbf{B}||| \leq \rho\}$, with $\mathcal{M}_{p\times q}$ denoting the set of $p \times q$ matrices with real elements, $0 < \rho < \infty$ and $|||\cdot|||$ being the following matrix norm:

$$|||\mathbf{B}||| = \sup\{\|\mathbf{B}\mathbf{x}\| : \mathbf{x} \in \mathbb{R}^q \text{ with } \|\mathbf{x}\| = 1\}, \forall\ \mathbf{B} \in \mathcal{M}_{p\times q}.$$

(co4) $\boldsymbol{\Sigma}_k \in \mathcal{D}_p\ \forall k$, where $\mathcal{D}_p$ denotes the set of the $p \times p$ positive definite matrices with eigenvalues in $[a, b]$, with $0 < a < b < \infty$.

Condition (C3)(a) is required for the Kullback-Leibler information criterion $\mathbb{E}\left(\log \frac{g(\mathbf{Z})}{h(\mathbf{Z};\boldsymbol{\psi})}\right)$ to be well-defined. Condition (C3)(b) is equivalent to an identification condition that makes $\breve{\boldsymbol{\psi}}$ globally identifiable (Bowden, 1973). It is then necessary to guarantee that the model class $\mathcal{F}_K$ is identifiable. Clearly, identifiability of $\mathcal{F}_K$ requires that $\pi_k > 0\ \forall k$, so as to avoid empty components. In order to preserve compactness, Constraint (co1) should be modified as follows:

(co1′) if $K > 1$, then $(\pi_1, \ldots \pi_K)' \in \mathcal{P}(K, \delta_K)$, where $\mathcal{P}(K, \delta_K) = \{\mathbf{p} \in [\delta_K, 1 - \delta_K]^K : \sum_{k=1}^K p_k = 1\}$, $0 < \delta_K < 0.5$.

Furthermore, in order to prevent non-identifiability issues related to label-switching mentioned in Section 2.1, an ordering of the components can be explicitly imposed by introducing additional constraints on the parameters. Such ordering constraints can be applied either to the weights $\pi_1, \ldots, \pi_K$ or to a specific element of the vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$. For example, constraint (co1′) can be replaced with:

(co1″) if $K > 1$, then $(\pi_1, \ldots \pi_K)' \in \mathcal{P}(K, \delta_K, \eta_K)$, where $\mathcal{P}(K, \delta_K, \eta_K) = \{\mathbf{p} \in [\delta_K, 1-\delta_K]^K : p_k - p_{k-1} \geq \eta_K, k = 2, \ldots, K, \sum_{k=1}^K p_k = 1\}$, $0 < \delta_K < 0.5$, $0 < \eta_K < 0.5$.

Being an intersection between the compact set $[\delta_K, 1-\delta_K]^K$ and two closed sets, that is $\{\mathbf{p} \in R^K : p_k - p_{k-1} \geq \eta_K, k = 2, \ldots, K\}$ and $\{\mathbf{p} \in R^K : \sum_{k=1}^K p_k = 1\}$ also $\mathcal{P}(K, \delta_K, \eta_K)$ is a compact set. It is worth mentioning that the actual values of δ_K and η_K must decrease as K increases, in order to guarantee that $\mathcal{P}(K, \delta_K, \eta_K) \neq \emptyset$. The introduction of constraint (co1″) excludes all finite mixtures of Gaussian regression models with (at least) two components with equal weights. Such exclusion can be circumvented by replacing the ordering constraint on the weights with ordering constraints on any other component-specific parameter while preserving compactness of $\bar{\boldsymbol{\Theta}}$. For example, by focusing on the first element of the K intercept vectors, one could require that $\gamma_{k(11)} - \gamma_{k-1(11)} \geq \nu$, $k = 2, \ldots, K$, where $0 < \nu < \infty$.

Condition (C5) ensures that matrices $\mathbf{A}(\boldsymbol{\psi})$ and $\mathbf{B}(\boldsymbol{\psi})$ are continuous in $\boldsymbol{\psi}$ and a uniform law of large numbers can be applied to the following matrices:

$$\frac{1}{I}\sum_{i=1}^{I} \frac{\partial^2 \log h(\mathbf{z}_i; \boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \text{ and } \frac{1}{I}\sum_{i=1}^{I} \left(\frac{\partial \log h(\mathbf{z}_i; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \log h(\mathbf{z}_i; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right)'.$$

5.2 Consistency of $\widehat{\text{Cov}}_1(\hat{\boldsymbol{\theta}})$ and $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\theta}})$

The consistency results about $\widehat{\text{Cov}}_1(\hat{\boldsymbol{\theta}})$ and $\widehat{\text{Cov}}_2(\hat{\boldsymbol{\theta}})$ require some preliminaries. Let $(\mathbf{x}_1', \mathbf{y}_1')', \ldots, (\mathbf{x}_I', \mathbf{y}_I')'$ be I independent and identically distributed sample observations of $(\mathbf{X}', \mathbf{Y}')'$. Their joint density under the model defined by equations (1), (2) and (11) is equal to

$$\prod_{i=1}^{I} h(\mathbf{z}_i; \boldsymbol{\psi}) = \prod_{i=1}^{I} q(\mathbf{x}_i; \boldsymbol{\vartheta}) f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}).$$

Thus, the corresponding log-likelihood is equal to $l_I(\boldsymbol{\psi}) = \sum_{i=1}^I l_i(\boldsymbol{\psi})$, where $l_i(\boldsymbol{\psi}) = \log h(\mathbf{z}_i; \boldsymbol{\psi})$. This log-likelihood can also be expressed as $l_I(\boldsymbol{\vartheta}) + l_I(\boldsymbol{\theta})$, where $l_I(\boldsymbol{\vartheta}) = \sum_{i=1}^I l_i(\boldsymbol{\vartheta})$, with $l_i(\boldsymbol{\vartheta}) = \log q(\mathbf{x}_i; \boldsymbol{\vartheta})$ and $l_I(\boldsymbol{\theta})$ is defined according to equation (3). Thus, the score vector $s(\boldsymbol{\psi}) = \frac{\partial l_I(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$ can be expressed as $s(\boldsymbol{\psi}) = \sum_{i=1}^I s_i(\boldsymbol{\psi})$, where $s_i(\boldsymbol{\psi}) = \frac{\partial l_i(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$. In a similar way, the Hessian matrix $H(\boldsymbol{\psi}) = \frac{\partial^2 l_I(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}$ can be written as $H(\boldsymbol{\psi}) = \sum_{i=1}^I H_i(\boldsymbol{\psi})$, where

$H_i(\boldsymbol{\psi}) = \frac{\partial^2 l_i(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}$. Clearly, $s(\boldsymbol{\psi})$ and $H(\boldsymbol{\psi})$ are defined only when the partial derivatives exist. Furthermore, they can be partitioned as follows:

$$s(\boldsymbol{\psi}) = \begin{pmatrix} s(\boldsymbol{\vartheta}) \\ s(\boldsymbol{\theta}) \end{pmatrix}, \quad H(\boldsymbol{\psi}) = \begin{bmatrix} H(\boldsymbol{\vartheta}) & \mathbf{0} \\ \mathbf{0} & H(\boldsymbol{\theta}) \end{bmatrix},$$

where $s(\boldsymbol{\vartheta}) = \frac{\partial l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}$, $s(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, $H(\boldsymbol{\vartheta}) = \frac{\partial^2 l(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'}$ and $H(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$. As a consequence of above, $\hat{\boldsymbol{\psi}}_I = (\hat{\boldsymbol{\vartheta}}_I, \hat{\boldsymbol{\theta}}_I)$, the ML estimator of $\boldsymbol{\psi}$ based on I sample observations, can be obtained by a separate maximisation of $l_I(\boldsymbol{\vartheta})$ and $l_I(\boldsymbol{\theta})$. As far as $\boldsymbol{\vartheta}$ is concerned, its estimator will depend on the probability distribution specified for the predictors. As already mentioned, $\hat{\boldsymbol{\theta}}_I$ can be computed using an EM algorithm.

Let $g(\mathbf{z})$ be the true density function of $\mathbf{Z}$ and admit that $g(\mathbf{z})$ may not belong to $\mathfrak{F}_K^*$ for a given value of K. This means that model (11) is correctly specified if $g(\mathbf{z}) = h(\mathbf{z}; \boldsymbol{\psi}_0)$ for some $\boldsymbol{\psi}_0 = (\boldsymbol{\vartheta}_0', \boldsymbol{\theta}_0')' \in \boldsymbol{\Psi}$; otherwise it is not.

Under all the above mentioned conditions, the following proposition summarises some results concerning $\hat{\boldsymbol{\psi}}_I$ that hold true provided that the appropriate inverses and expectations (taken with respect to the true distribution) exist. A proof is reported in Appendix E.

Proposition 1

(a) Given the conditions (C1)-(C2) the existence of $\hat{\boldsymbol{\psi}}_I$ is ensured for all I.

(b) Given the conditions (C1)-(C3) the following convergence holds true:

$$\hat{\boldsymbol{\psi}}_I \xrightarrow{a.s.} \breve{\boldsymbol{\psi}}, \text{ with } \breve{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi} \in \boldsymbol{\Psi}}{\operatorname{argmin}} \mathbb{E}\left(\log \frac{g(\mathbf{Z})}{h(\mathbf{Z}; \boldsymbol{\psi})}\right).$$

(c) Given the conditions (C1)-(C6) the following convergences hold true:

$$\sqrt{I} \cdot \left(\hat{\boldsymbol{\psi}}_I - \breve{\boldsymbol{\psi}}\right) \xrightarrow{L} N(\boldsymbol{0}, \boldsymbol{C}(\breve{\boldsymbol{\psi}})), \tag{12}$$

$$I \cdot (H(\hat{\boldsymbol{\psi}}_I))^{-1} \left(\sum_{i=1}^{I} s_i(\hat{\boldsymbol{\psi}}_I) \left(s_i(\hat{\boldsymbol{\psi}}_I)\right)'\right) (H(\hat{\boldsymbol{\psi}}_I))^{-1} \xrightarrow{a.s.} \boldsymbol{C}(\breve{\boldsymbol{\psi}}) \text{ element by element}, \tag{13}$$

$$I \cdot H(\hat{\boldsymbol{\psi}}_I) \xrightarrow{a.s.} \mathbb{E}(H_i(\breve{\boldsymbol{\psi}})) \text{ element by element}, \tag{14}$$

where $\boldsymbol{C}(\breve{\boldsymbol{\psi}}) = \boldsymbol{C}(\boldsymbol{\psi})|_{\boldsymbol{\psi}=\breve{\boldsymbol{\psi}}}$, with

$$\boldsymbol{C}(\boldsymbol{\psi}) = (\mathbb{E}(H_i(\boldsymbol{\psi})))^{-1} (\mathbb{E}(s_i(\boldsymbol{\psi}) s_i(\boldsymbol{\psi})')) (\mathbb{E}(H_i(\boldsymbol{\psi})))^{-1},$$

$H(\hat{\boldsymbol{\psi}}_I) = H(\boldsymbol{\psi})|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_I}$ and $s_i(\hat{\boldsymbol{\psi}}_I) = s_i(\boldsymbol{\psi})|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_I}$.

(d) Given the conditions (C1)-(C7) and if $g(\mathbf{z}) = h(\mathbf{z}; \boldsymbol{\psi}_0)$ for some $\boldsymbol{\psi}_0 \in \boldsymbol{\Psi}$, then $\breve{\boldsymbol{\psi}} = \boldsymbol{\psi}_0$, $\mathbb{E}(H_i(\boldsymbol{\psi}_0)) = -\mathbb{E}(s_i(\boldsymbol{\psi}_0) s_i(\boldsymbol{\psi}_0)')$ and

$$\boldsymbol{C}(\boldsymbol{\psi}_0) = -\left(\mathbb{E}(H_i(\boldsymbol{\psi}_0))\right)^{-1} = \left(\mathbb{E}(s_i(\boldsymbol{\psi}_0) s_i(\boldsymbol{\psi}_0)')\right)^{-1}, \tag{15}$$

where $-\mathbb{E}(H_i(\boldsymbol{\psi}_0))$ represents the Fisher's information matrix evaluated at $\boldsymbol{\psi}_0$.

Part *(b)* of Proposition 1 states that, since $g(\mathbf{z})$ may not belong to $\mathfrak{F}_K^*$, $\hat{\boldsymbol{\psi}}_I$ is in fact a strongly consistent estimator of the parameter vector $\breve{\boldsymbol{\psi}} = (\breve{\boldsymbol{\vartheta}}', \breve{\boldsymbol{\theta}}')' \in \boldsymbol{\Psi}$ which minimises $\mathbb{E}\left(\log \frac{g(\mathbf{Z})}{h(\mathbf{Z};\boldsymbol{\psi})}\right)$, that is the Kullback-Leibler information criterion (Kullback and Leibler, 1951); this is a consequence of the fact that, when $g(\mathbf{z}) \notin \mathfrak{F}_K$, $l_I(\boldsymbol{\psi})/I$ represents a natural estimation of $\mathbb{E}\left(\log h(\mathbf{Z};\boldsymbol{\psi})\right)$ (Akaike, 1973). If $g(\mathbf{z}) = h(\mathbf{z};\boldsymbol{\psi}_0)$ then $\mathbb{E}\left(\log \frac{g(\mathbf{Z})}{h(\mathbf{Z};\boldsymbol{\psi})}\right)$ will attain its minimum at $\breve{\boldsymbol{\psi}} = \boldsymbol{\psi}_0$ and $\hat{\boldsymbol{\psi}}_I$ will be a strongly consistent estimator of $\boldsymbol{\psi}_0$. According to the results described in the parts *(c)* and *(d)* of the proposition, when the model is correctly specified and all the considered regularity conditions are met, the classical equivalence of the Hessian and outer product forms of the Fisher's information matrix holds true; both forms can be employed to obtain a consistent estimation of the asymptotic covariance matrix of $\hat{\boldsymbol{\psi}}_I$.

Finally, the asymptotic properties of $\widehat{\mathrm{Cov}}_1(\hat{\boldsymbol{\theta}})$ and $\widehat{\mathrm{Cov}}_2(\hat{\boldsymbol{\theta}})$ can be easily investigated by exploiting the block-diagonal structure of matrices $H(\boldsymbol{\psi})$, $\mathbb{E}(H_i(\boldsymbol{\psi}))$ and $\mathbf{C}(\boldsymbol{\psi})$. Their block-diagonal structure implies that the corresponding submatrices related to the parameters $\boldsymbol{\theta}$ do not depend on the parameters of the marginal distribution of $\mathbf{X}$, as these submatrices involve only first and second partial derivatives of the conditional distribution of $\mathbf{Y}|\mathbf{X} = \mathbf{x}$. In particular, consistency of $\widehat{\mathrm{Cov}}_1(\hat{\boldsymbol{\theta}})$ and $\widehat{\mathrm{Cov}}_2(\hat{\boldsymbol{\theta}})$ is given by following theorems. Their proofs are reported in Appendix F and Appendix G.

Theorem 3 *Given the conditions (C1)-(C7) and if* $g(\mathbf{z}) = h(\mathbf{z};\boldsymbol{\psi}_0)$ *for some* $\boldsymbol{\psi}_0 = (\boldsymbol{\vartheta}_0', \boldsymbol{\theta}_0')' \in \boldsymbol{\Psi}$

$$I \cdot \widehat{Cov}_1(\hat{\boldsymbol{\theta}}_I) \xrightarrow{a.s.} -\left(\mathbb{E}(H_i(\boldsymbol{\theta}_0))\right)^{-1}, \tag{16}$$

where $-\mathbb{E}(H_i(\boldsymbol{\theta}_0))$ *is the submatrix of* $-\mathbb{E}(H_i(\boldsymbol{\psi}_0))$ *that refers to* $\boldsymbol{\theta}_0$.

Theorem 4 *Given the conditions (C1)-(C6)*

$$I \cdot \widehat{Cov}_2(\hat{\boldsymbol{\theta}}_I) \xrightarrow{a.s.} \boldsymbol{C}(\breve{\boldsymbol{\theta}}), \tag{17}$$

where $\boldsymbol{C}(\breve{\boldsymbol{\theta}})$ *is the submatrix of* $\boldsymbol{C}(\breve{\boldsymbol{\psi}})$ *that refers to* $\breve{\boldsymbol{\theta}}$.

Thus, the estimator of $Cov(\hat{\boldsymbol{\theta}}_I)$ defined in equation (5), based on the sandwich approach, is strongly consistent when the model is misspecified; the estimator computed from the Hessian matrix given in equation (4) results to be consistent under the classical assumption of a correctly specified model.

6 Conclusions

In this paper formulae for computing the score vector and Hessian matrix of the incomplete log-likelihood for multivariate mixtures of Gaussian linear regression models are provided. These formulae are used to define two estimators of the covariance matrix of the ML estimator. From a practical point of view

the main advantage in using the proposed estimators is to avoid resorting to bootstrap-based methods and general purpose optimisers. Another advantage associated with an analytical evaluation of the Hessian matrix is that it can be used as a diagnostic tool to detect convergence failures of the EM algorithm. Using the suggested information-based estimators also allows to compute both approximated confidence intervals and Wald statistics for testing hypotheses about model parameters. With small samples and correctly specified models, the estimator based on the observed Hessian-based information matrix and the parametric bootstrap-based estimator show similar performances in terms of biases, root mean square errors and coverage probabilities of confidence intervals. Thus, in correctly specified models, the main benefits associated with the estimator based on the observed Hessian-based information matrix over the bootstrap are a gain in the time required to compute the standard errors, a simplification of the overall computational process and a protection against possible convergence failures of the EM algorithm on the bootstrap replicates. When models are characterised by misspecification due to heteroscedasticity, the estimator based on the sandwich matrix is to be preferred: it not only has the smallest absolute bias and the smallest root mean squared error but also produces effective confidence levels that are the closest to the desired nominal ones.

The clusterwise regression models examined in this paper allow to perform multivariate linear regression analysis with fixed predictors in the presence of unobserved heterogeneity. However, when the predictors are not under the control of the experimenter, clusterwise regression models with random predictors should be employed (Hennig, 2000; Wedel, 2002). In particular, multivariate linear cluster-weighted models (see, e.g., Dang et al., 2017) are able to capture both the linear dependencies among responses and the linear effects of the predictors on the responses from sample observations coming from unknown heterogeneous populations, each of which is characterised by a different joint distribution of $(\mathbf{X}', \mathbf{Y}')'$. We are currently in the process of extending the covariance matrix estimators of the ML estimator described in this paper under multivariate Gaussian linear cluster-weighted models. Further developments could be obtained under models in which the predictors may also affect the membership to the unknown heterogeneous populations (see, e.g., Dayton and Macready, 1988; Ingrassia and Punzo, 2016; Lamont et al., 2016).

References

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

Aitkin M, Rubin DB (1985) Estimation and hypothesis testing in finite mixture models. J R Stat Soc Ser B 47: 67–75

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds), Second International Symposium on Information Theory, Akademiai Kiado, pp 267–281

Arminger G, Stein P, Wittenberg J (1999) Mixtures of conditional mean and covariance structure models. Psychometrika 64: 475–494

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

Basford KE, Greenway DR, McLachlan GJ, Peel D (1997) Standard errors of fitted means under normal mixture models. Comput Stat 12: 1–17

Benaglia T, Chauveau D, Hunter DR, Young D (2009) `mixtools`: an `R` package for analyzing finite mixture models. J Stat Softw 32(6): 1–29

Boiteau G, Singh M, Singh RP, Tai GCC, Turner TR (1998) Rate of spread of PVY-n by alate Myzus persicae (Sulzer) from infected to healthy plants under laboratory conditions. Potato Research 41: 335 – 344

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

Bowden R (1973) The theory of parametric identification. Econometrica 41: 1069–1074

Chevalier, J. A., A. K. Kashyap and P. E. Rossi (2003). Why don't prices rise during periods of peak demand? Evidence from scanner data. Am Econ Rev 93: 15–37

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

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

Dayton CM, Macready GB (1988) Concomitant-variable latent-class models. J Am Stat Assoc 83: 173–178

Ding C (2006) Using regression mixture analysis in educational research. Practical Assessment Research & Evaluation 11: 1–11

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

Elhenawy M, Rakha H, Chen H (2017) An automatic traffic congestion identification algorithm based on mixture of linear regressions. In: Helfert M, Klein C, Donnellan B, Gusikhin O (eds) Smart Cities, Green Technologies, and Intelligent Transport Systems, Springer, Cham, pp 242–256

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

Faria S, Soromenho G (2010) Fitting mixtures of linear regressions. J Stat Comput Simul 80: 201–225

Frühwirth-Schnatter, S. (2006). Finite Mixture and Markov Switching Models. Springer, New York

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

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

García-Escudero LA, Gordaliza A, Mayo-Iscar A, San Martín R (2010) Robust clusterwise linear regression through trimming. Comput Stat Data Anal 54(12): 3057–3069

Grün B, Leisch F (2008) `FlexMix` version 2: finite mixtures with concomitant variables and varying and constant parameters. J Stat Softw 28(4): 1–35

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

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

Ingrassia S, Punzo A (2016) Decision boundaries for mixtures of regressions. J Korean Stat Soc 45: 295–306

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

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

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22: 79–86

Lamont AE, Vermunt JK, Van Horn ML (2016) Regression mixture models: Does modeling the covariance between independent variables and latent classes improve the results? Multivar Behav Res 51: 35–52

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J R Stat Soc B 44: 226–233

Magnus JR, Neudecker H (1988) Matrix Differential Calculus with Applications in Statistics and Econometrics. John Wiley & Sons, New York

Maugis C, Celeux G, Martin-Magniette ML (2009) Variable selection for clustering with Gaussian mixture models. Biometrics 65:701–709

Mazza A, Punzo A (2017) Mixtures of multivariate contaminated normal regression models. Stat Papers `DOI: 10.1007/s00362-017-0964-y`

Mazza A, Punzo A, Ingrassia S (2018) `flexCWM`: A flexible framework for cluster-weighted models. J Stat Softw 86(2): 1–30

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

McLachlan GJ, Peel D (2000) Finite Mixture Models. Wiley, New York

Meilijson I (1989) A fast improvement to the EM algorithm on its own terms. J R Stat Soc B 51: 127–138

Newton MA, Raftery AE (1994) Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). J R Stat Soc B 56: 3–48

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

Rossi, P. (2019). bayesm: Bayesian inference for marketing/micro-econometrics. R package version 3.1-4. URL https://CRAN.R-project.org/package=bayesm"

Schott, J. R. (2005). *Matrix Analysis for Statistics*. Second edition. John Wiley & Sons, New York

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

Städler N, Bühlmann P, van de Geer S (2010) ℓ1-penalization for mixture regression models. Test, 19: 209–256

Tang Q, Karunamuni RJ (2013) Minimum distance estimation in a finite mixture regression model. J Multivar Anal, 120: 185–204

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

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

Turner TR (2014) mixreg: Functions to fit mixtures of regressions. URL http://CRAN.R-project.org/package=mixreg. Accessed 11 January 2019

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

Venables WN, Ripley BD (2002) Modern Applied Statistics with S. Fourth edition. Springer, Ney York

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

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50: 1–25

Yao F, Fu Y, Lee TCM (2011) Functional mixture regression. Biostatistics, 12: 341–353

A Proof of Theorem 1

Using equations (3) and (6) it is possible to write $l(\boldsymbol{\theta}) = \sum_{i=1}^{I} l_i(\boldsymbol{\theta})$, where $l_i(\boldsymbol{\theta}) = \log(\sum_{k=1}^{K} f_{ki})$. By exploiting the result given by equation (A.1) in Boldea and Magnus (2009), the first order differential of $l(\boldsymbol{\theta})$ is equal to

$$\mathrm{d}l(\boldsymbol{\theta}) = \sum_{i=1}^{I} \mathrm{d}l_i(\boldsymbol{\theta}) = \sum_{i=1}^{I} \left(\sum_{k=1}^{K} \alpha_{ki} \mathrm{d} \log f_{ki} \right), \tag{18}$$

where α_{ki} is defined in equation (7). $\mathrm{d} \log f_{ki}$, the first order differential of $\log f_{ki}$, is equal to (see Appendix C)

$$\begin{aligned} \mathrm{d}\log f_{ki} = {} & (\mathrm{d}\boldsymbol{\pi})' \, \mathbf{a}_k + (\mathrm{d}\boldsymbol{\gamma}_k)' \, \mathbf{b}_{ki} + \left[\mathrm{dvec}(\boldsymbol{\Pi}_k') \right]' \mathrm{vec} \left(\mathbf{x}_i \mathbf{b}_{ki}' \right) + \\ & - \frac{1}{2} \left[\mathrm{dv}(\boldsymbol{\Sigma}_k) \right]' \mathbf{G}' \mathrm{vec} \left(\mathbf{B}_{ki} \right), \end{aligned} \tag{19}$$

where $\boldsymbol{a}_k$, $\mathbf{b}_{ki}$ and $\mathbf{B}_{ki}$ are defined in equations (8), (9) and (10), respectively.

Inserting equation (19) in equation (18) gives

$$\begin{aligned}\mathrm{d}l(\boldsymbol{\theta}) = (\mathrm{d}\boldsymbol{\pi})' \sum_{i=1}^{I}\sum_{k=1}^{K} \alpha_{ki}\mathbf{a}_k + \sum_{k=1}^{K} (\mathrm{d}\boldsymbol{\gamma}_k)' \sum_{i=1}^{I} \alpha_{ki}\mathbf{b}_{ki} + \\ + \sum_{k=1}^{K} \left[\mathrm{dvec}(\boldsymbol{\Pi}_k')\right]' \sum_{i=1}^{I} \alpha_{ki}\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right) + \\ -\frac{1}{2}\sum_{k=1}^{K} \left[\mathrm{dv}(\boldsymbol{\Sigma}_k)\right]' \sum_{i=1}^{I} \alpha_{ki}\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right).\end{aligned}$$

Taking the derivatives with respect to $\boldsymbol{\pi}$, $\boldsymbol{\gamma}_k$, $\mathrm{vec}(\boldsymbol{\Pi}_k')$ and $\mathrm{v}(\boldsymbol{\Sigma}_k)$ completes the proof.

B Proof of Theorem 2

The second order differential of $l(\boldsymbol{\theta})$ is given by

$$\mathrm{d}^2 l\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{I} \mathrm{d}^2 l_i\left(\boldsymbol{\theta}\right),$$

where

$$\mathrm{d}^2 l_i\left(\boldsymbol{\theta}\right) = \sum_{k=1}^{K} \alpha_{ki}\mathrm{d}^2 \log f_{ki} + \sum_{k=1}^{K} \alpha_{ki}\left(\mathrm{d}\log f_{ki}\right)^2 - \left(\sum_{k=1}^{K} \alpha_{ki}\mathrm{d}\log f_{ki}\right)^2 \tag{20}$$

(see equation (A.2) in Boldea and Magnus, 2009).

Equation (36) gives an expression for $\mathrm{d}^2\log f_{ki}$, the second order differential of $\log f_{ki}$. Furthermore, expressions for $\left(\mathrm{d}\log f_{ki}\right)^2$ and $\left(\sum_{k=1}^{K} \alpha_{ki}\mathrm{d}\log f_{ki}\right)^2$ can be obtained, after some algebra, by noting that:

$$\begin{aligned}\left(\mathrm{d}\log f_{ki}\right)^2 &= \left(\mathrm{d}\log f_{ki}\right)'\left(\mathrm{d}\log f_{ki}\right), \\ \left(\sum_{k=1}^{K} \alpha_{ki}\mathrm{d}\log f_{ki}\right)^2 &= \left(\sum_{k=1}^{K} \alpha_{ki}\mathrm{d}\log f_{ki}\right)'\left(\sum_{k=1}^{K} \alpha_{ki}\mathrm{d}\log f_{ki}\right),\end{aligned}$$

and by exploiting the result for $\mathrm{d}\log f_{ki}$ given in equation (19). This results in:

$$
\begin{aligned}
(\mathrm{d}\log f_{ki})^2 = {} & (\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} + (\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \mathbf{b}_{ki}' \mathrm{d}\boldsymbol{\gamma}_k + \\
& + (\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right]' \left(\mathrm{dvec}(\boldsymbol{\Pi}_k')\right) + \\
& - \frac{1}{2} (\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]' \mathbf{G} \mathrm{dv}(\boldsymbol{\Sigma}_k) + (\mathrm{d}\boldsymbol{\gamma}_k)' \mathbf{b}_{ki} \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} + \\
& + (\mathrm{d}\boldsymbol{\gamma}_k)' \mathbf{b}_{ki} \mathbf{b}_{ki}' \mathrm{d}\boldsymbol{\gamma}_k + \\
& + (\mathrm{d}\boldsymbol{\gamma}_k)' \mathbf{b}_{ki} \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right]' \mathrm{dvec}(\boldsymbol{\Pi}_k') + \\
& - \frac{1}{2} (\mathrm{d}\boldsymbol{\gamma}_k)' \mathbf{b}_{ki} \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]' \mathbf{G} \mathrm{dv}(\boldsymbol{\Sigma}_k) + \\
& + \left[\mathrm{dvec}\boldsymbol{\Pi}_k'\right]' \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} + \\
& + \left[\mathrm{dvec}(\boldsymbol{\Pi}_k')\right]' \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) \mathbf{b}_{ki}' \mathrm{d}\boldsymbol{\gamma}_k + \\
& + \left[\mathrm{dvec}(\boldsymbol{\Pi}_k')\right]' \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right]' \mathrm{dvec}(\boldsymbol{\Pi}_k') + \\
& - \frac{1}{2} \left[\mathrm{dvec}(\boldsymbol{\Pi}_k')\right]' \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]' \mathbf{G} \mathrm{dv}(\boldsymbol{\Sigma}_k) + \\
& - \frac{1}{2} \left[\mathrm{dv}(\boldsymbol{\Sigma}_k)\right]' \mathbf{G}' \mathrm{vec}\left(\mathbf{B}_{ki}\right) \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} + \\
& - \frac{1}{2} \left[\mathrm{dv}(\boldsymbol{\Sigma}_k)\right]' \mathbf{G}' \mathrm{vec}\left(\mathbf{B}_{ki}\right) \mathbf{b}_{ki}' \mathrm{d}\boldsymbol{\gamma}_k + \\
& - \frac{1}{2} \left[\mathrm{dv}(\boldsymbol{\Sigma}_k)\right]' \mathbf{G}' \mathrm{vec}\left(\mathbf{B}_{ki}\right) \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right]' \mathrm{dvec}(\boldsymbol{\Pi}_k') + \\
& + \frac{1}{4} \left[\mathrm{dv}(\boldsymbol{\Sigma}_k)\right]' \mathbf{G}' \mathrm{vec}\left(\mathbf{B}_{ki}\right) \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]' \mathbf{G} \mathrm{dv}(\boldsymbol{\Sigma}_k), \qquad (21)
\end{aligned}
$$

$$
\begin{aligned}
\left(\sum_{k=1}^{K} \alpha_{ki} \mathrm{d}\ln f_{ki}\right)^2 = {} & (\mathrm{d}\boldsymbol{\pi})' \bar{\mathbf{a}}_i \bar{\mathbf{a}}_i' \mathrm{d}\boldsymbol{\pi} + (\mathrm{d}\boldsymbol{\pi})' \bar{\mathbf{a}}_i \left(\sum_{k=1}^{K} \alpha_{ki} \mathbf{b}_{ki}' \mathrm{d}\boldsymbol{\gamma}_k\right) + \\
& + (\mathrm{d}\boldsymbol{\pi})' \bar{\mathbf{a}}_i \left\{\sum_{k=1}^{K} \alpha_{ki} \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right]' \left[\mathrm{dvec}(\boldsymbol{\Pi}_k')\right]\right\} + \\
& - \frac{1}{2} (\mathrm{d}\boldsymbol{\pi})' \bar{\mathbf{a}}_i \left\{\sum_{k=1}^{K} \alpha_{ki} \left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]' \mathbf{G} \mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right\} + \\
& + \left[\sum_{k=1}^{K} (\mathrm{d}\boldsymbol{\gamma}_k)' \alpha_{ki} \mathbf{b}_{ki}\right] \bar{\mathbf{a}}_i' \mathrm{d}\boldsymbol{\pi} + \sum_{k=1}^{K} \sum_{l=1}^{K} (\mathrm{d}\boldsymbol{\gamma}_k)' \alpha_{ki} \alpha_{li} \mathbf{b}_{ki} \mathbf{b}_{li}' \mathrm{d}\boldsymbol{\gamma}_l + \\
& + \sum_{k=1}^{K} \sum_{l=1}^{K} (\mathrm{d}\boldsymbol{\gamma}_k)' \alpha_{ki} \alpha_{li} \mathbf{b}_{ki} \left[\mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{li}'\right)\right]' \mathrm{dvec}(\boldsymbol{\Pi}_l') + \\
& - \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} (\mathrm{d}\boldsymbol{\gamma}_k)' \alpha_{ki} \alpha_{li} \mathbf{b}_{ki} \left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]' \mathbf{G} \mathrm{dv}(\boldsymbol{\Sigma}_l) +
\end{aligned}
$$

$$
\begin{aligned}
&+\left[\sum_{k=1}^{K}\left[\mathrm{dvec}(\boldsymbol{\Pi}_k')\right]'\alpha_{ki}\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\right]\bar{\mathbf{a}}_i'\mathrm{d}\boldsymbol{\pi}+\\
&+\sum_{k=1}^{K}\sum_{l=1}^{K}\left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]'\alpha_{ki}\alpha_{li}\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\mathbf{b}_{li}'\mathrm{d}\boldsymbol{\gamma}_l+\\
&+\sum_{k=1}^{K}\sum_{l=1}^{K}\left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]'\alpha_{ki}\alpha_{li}\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\left[\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{li}'\right)\right]'\mathrm{dvec}(\boldsymbol{\Pi}_l')+\\
&-\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}\left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]'\alpha_{ki}\alpha_{li}\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]'\mathbf{G}\mathrm{dv}(\boldsymbol{\Sigma}_l)+\\
&-\frac{1}{2}\left[\sum_{k=1}^{K}\left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]'\alpha_{ki}\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]\bar{\mathbf{a}}_i'\mathrm{d}\boldsymbol{\pi}+\\
&-\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}\left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]'\alpha_{ki}\alpha_{li}\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right)\mathbf{b}_{li}'\mathrm{d}\boldsymbol{\gamma}_l+\\
&-\frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}\left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]'\alpha_{ki}\alpha_{li}\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right)\left[\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{li}'\right)\right]'\mathrm{dvec}\left(\boldsymbol{\Pi}_l'\right)+\\
&+\frac{1}{4}\sum_{k=1}^{K}\sum_{l=1}^{K}\left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]'\alpha_{ki}\alpha_{li}\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right)\left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]'\mathbf{G}\mathrm{dv}\left(\boldsymbol{\Sigma}_l\right). \qquad (22)
\end{aligned}
$$

Inserting equations (36), (21) and (22) in equation (20) and taking the second order derivatives leads to

$$
\begin{aligned}
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\pi}\partial\boldsymbol{\pi}'} &= -\bar{\mathbf{a}}_i\bar{\mathbf{a}}_i',\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\pi}\partial\boldsymbol{\gamma}_k'} &= \alpha_{ki}\left(\mathbf{a}_k-\bar{\mathbf{a}}_i\right)\mathbf{b}_{ki}'\ \forall k,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\pi}\partial\left[\mathrm{vec}(\boldsymbol{\Pi}_k')\right]'} &= \alpha_{ki}\left(\mathbf{a}_k-\bar{\mathbf{a}}_i\right)\left[\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\right]'\ \forall k,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\pi}\partial\left[\mathrm{v}(\boldsymbol{\Sigma}_k)\right]'} &= -\frac{1}{2}\alpha_{ki}\left(\mathbf{a}_k-\bar{\mathbf{a}}_i\right)\left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]'\mathbf{G}\ \forall k,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_k\partial\boldsymbol{\gamma}_k'} &= -\alpha_{ki}\left[\boldsymbol{\Sigma}_k^{-1}-\left(1-\alpha_{ki}\right)\mathbf{b}_{ki}\mathbf{b}_{ki}'\right]\ \forall k,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_k\partial\boldsymbol{\gamma}_l'} &= -\alpha_{ki}\alpha_{li}\mathbf{b}_{ki}\mathbf{b}_{li}'\ \forall k\neq l,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_k\partial\left[\mathrm{vec}(\boldsymbol{\Pi}_k')\right]'} &= -\alpha_{ki}\left[\boldsymbol{\Sigma}_k^{-1}\otimes\mathbf{x}_i'-\left(1-\alpha_{ki}\right)\mathbf{b}_{ki}\left[\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\right]'\right]\ \forall k,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_k\partial\left[\mathrm{vec}(\boldsymbol{\Pi}_l')\right]'} &= -\alpha_{ki}\alpha_{li}\mathbf{b}_{ki}\left[\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{li}'\right)\right]'\ \forall k\neq l,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_k\partial\left[\mathrm{v}(\boldsymbol{\Sigma}_k)\right]'} &= -\alpha_{ki}\left[\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)+\frac{1}{2}\left(1-\alpha_{ki}\right)\mathbf{b}_{ki}\left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]'\right]\mathbf{G}\ \forall k,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}_k\partial\left[\mathrm{v}(\boldsymbol{\Sigma}_l)\right]'} &= \frac{1}{2}\alpha_{ki}\alpha_{li}\mathbf{b}_{ki}\left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]'\mathbf{G}\ \forall k\neq l,\\
\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\mathrm{vec}\left(\boldsymbol{\Pi}_k'\right)\partial\left[\mathrm{vec}(\boldsymbol{\Pi}_l')\right]'} &= -\alpha_{ki}\alpha_{li}\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\left[\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{li}'\right)\right]'\ \forall k\neq l,
\end{aligned}
$$

$$\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial \mathrm{vec}\left(\boldsymbol{\Pi}_k'\right) \partial \left[\mathrm{vec}(\boldsymbol{\Pi}_k')\right]'} = -\alpha_{ki}\left[\left(\boldsymbol{\Sigma}_k^{-1} \otimes (\mathbf{x}_i\mathbf{x}_i^\top)\right) + \right.$$
$$\left. - (1-\alpha_{ki})\,\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}^\top\right)\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}^\top\right)^\top\right] \ \forall k,$$
$$\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial \mathrm{vec}\left(\boldsymbol{\Pi}_k'\right) \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_k)\right]'} = -\alpha_{ki}\left[\left(\boldsymbol{\Sigma}_k^{-1} \otimes (\mathbf{x}_i\mathbf{b}_{ki}')\right)\right.$$
$$\left. + \frac{1}{2}(1-\alpha_{ki})\,\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]'\right]\mathbf{G} \ \forall k,$$
$$\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial \mathrm{vec}\left(\boldsymbol{\Pi}_k'\right) \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_l)\right]'} = \frac{1}{2}\alpha_{ki}\alpha_{li}\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]'\mathbf{G} \ \forall k \neq l,$$
$$\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial \mathrm{v}(\boldsymbol{\Sigma}_k) \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_k)\right]'} = -\frac{1}{2}\alpha_{ki}\mathbf{G}'\left[\left(\boldsymbol{\Sigma}_k^{-1} - 2\mathbf{B}_{ki}\right)' \otimes \boldsymbol{\Sigma}_k^{-1} + \right.$$
$$\left. - \frac{1}{2}(1-\alpha_{ki})\,\mathrm{vec}\left(\mathbf{B}_{ki}\right)\left[\mathrm{vec}\left(\mathbf{B}_{ki}\right)\right]'\right]\mathbf{G} \ \forall k,$$
$$\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial \mathrm{v}(\boldsymbol{\Sigma}_k) \partial \left[\mathrm{v}(\boldsymbol{\Sigma}_l)\right]'} = -\frac{1}{4}\alpha_{ki}\alpha_{li}\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right)\left[\mathrm{vec}\left(\mathbf{B}_{li}\right)\right]'\mathbf{G} \ \forall k \neq l.$$

Summing the contributions for the I observations completes the proof.

C First order differential of $\log f_{ki}$

Up to an additive constant, $\log f_{ki}$ in equation (18) is equal to

$$\log \pi_k - \frac{1}{2}\log\det\left(\boldsymbol{\Sigma}_k\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k\boldsymbol{x}_i\right)\left(\boldsymbol{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k\boldsymbol{x}_i\right)'\right).$$

Thus, its first order differential results to be equal to

$$\mathrm{dlog}\, f_{ki} = \mathrm{d}_{k0} + \mathrm{d}_{ki1} + \mathrm{d}_{ki2} + \mathrm{d}_{ki3}, \tag{23}$$

where

$$\mathrm{d}_{k0} = \mathrm{dlog}\,\pi_k = \left(\mathrm{d}\boldsymbol{\pi}\right)'\boldsymbol{a}_k, \tag{24}$$
$$\mathrm{d}_{ki1} = -\frac{1}{2}\mathrm{d}\left(\log\det\left(\boldsymbol{\Sigma}_k\right)\right),$$
$$\mathrm{d}_{ki2} = -\frac{1}{2}\mathrm{tr}\left[\mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right)\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k\mathbf{x}_i\right)\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k\mathbf{x}_i\right)'\right],$$
$$\mathrm{d}_{ki3} = -\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k\mathbf{x}_i\right)\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k\mathbf{x}_i\right)'\right].$$

Using Corollary 9.1.1 and Theorem 1.3 in Schott (2005), it results that

$$\mathrm{d}_{ki1} = -\frac{1}{2}\mathrm{tr}\left[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\right]. \tag{25}$$

Furthermore, since $\mathrm{d}(\boldsymbol{\Sigma}_k^{-1}) = -\boldsymbol{\Sigma}_k^{-1}\mathrm{d}(\boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1}$ (see, e.g., Magnus and Neudecker, 1988, p. 183), it is possible to write

$$\mathrm{d}_{ki2} = \frac{1}{2}\mathrm{tr}\left[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\mathbf{b}_{ki}\mathbf{b}_{ki}^\top\right]. \tag{26}$$

By exploiting Theorem 8.10 in Schott (2005), some results for the differential of matrix functions (see, e.g., Magnus and Neudecker, 1988, p. 182) and some properties of the vector operator (see, e.g., Schott, 2005, pages 313 and 356), we find

$$\mathrm{d}_{ki1} + \mathrm{d}_{ki2} = -\frac{1}{2}\left[\mathrm{dv}(\boldsymbol{\Sigma}_k)\right]'\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right), \tag{27}$$
$$\mathrm{d}_{ki3} = \left(\mathrm{d}\boldsymbol{\gamma}_k\right)'\mathbf{b}_{ki} + \left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]'\mathrm{vec}\left(\mathbf{x}_i\mathbf{b}_{ki}'\right). \tag{28}$$

Substituting equations (24), (27) and (28) in (23) leads to

$$\begin{aligned}\mathrm{dlog}\, f_{ki} = {} & (\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k + (\mathrm{d}\boldsymbol{\gamma}_k)' \mathbf{b}_{ki} + \left[\mathrm{dvec}(\boldsymbol{\Pi}_k')\right]' \mathrm{vec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right) + \\ & - \frac{1}{2} \left[\mathrm{dv}(\boldsymbol{\Sigma}_k)\right]' \mathbf{G}' \mathrm{vec}\left(\mathbf{B}_{ki}\right).\end{aligned}$$

D Second order differential of log f_{ki}

Using equation (6), the second order differential of $\log f_{ki}$ can be expressed as

$$\mathrm{d}^2 \log f_{ki} = \mathrm{d}^2 \log \boldsymbol{\pi}_k + \mathrm{d}^2 \log \phi\left(\mathbf{y}_i; \boldsymbol{\gamma}_k + \boldsymbol{\Pi}_k \mathbf{x}_i, \boldsymbol{\Sigma}_k\right), \tag{29}$$

where

$$\begin{aligned}\mathrm{d}^2 \log \pi_k &= -\left(\mathrm{d}\boldsymbol{\pi}\right)' \mathbf{a}_k \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi}, \\ \mathrm{d}^2 \log \phi\left(\mathbf{y}_i; \boldsymbol{\gamma}_k + \boldsymbol{\Pi}_k \mathbf{x}_i, \boldsymbol{\Sigma}_k\right) &= \mathrm{d}(\mathrm{d}_{ki1}) + \mathrm{d}(\mathrm{d}_{ki2}) + \mathrm{d}(\mathrm{d}_{ki3}),\end{aligned} \tag{30}$$

and d_{ki1}, d_{ki2} and d_{ki3} are defined in equations (25), (26), and (28), respectively. Thus, it is possible to write

$$\mathrm{d}\left(\mathrm{d}_{ki1}\right) = -\frac{1}{2} \mathrm{tr}\left[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\left(\mathrm{d}\boldsymbol{\Sigma}_k^{-1}\right)\right] = \frac{1}{2} \mathrm{tr}\left[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1}\right]. \tag{31}$$

where the last equation holds because of the rule for the differential of the inverse of a nonsingular matrix (see, e.g., Magnus and Neudecker, 1988, page 183). Furthermore,

$$\begin{aligned}\mathrm{d}\left(\mathrm{d}_{ki2}\right) &= \mathrm{d}\left(\frac{1}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)'\right]\right) \\ &= \frac{1}{2} \mathrm{tr}\left[\mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right) \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)'\right] + \\ &+ \frac{1}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right) \left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)'\right] + \\ &+ \frac{1}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1} \mathrm{d}\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)'\right] \\ &= -\mathrm{tr}\left[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)\left(\mathbf{y}_i - \boldsymbol{\gamma}_k - \boldsymbol{\Pi}_k \mathbf{x}_i\right)' \boldsymbol{\Sigma}_k^{-1}\right] + \\ &- \left(\mathrm{d}\boldsymbol{\gamma}_k\right)' \left(\mathbf{b}_{ki}' \otimes \boldsymbol{\Sigma}_k^{-1}\right) \mathrm{dvec}\left(\boldsymbol{\Sigma}_k\right) + \\ &- \left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]' \left[\boldsymbol{\Sigma}_k^{-1} \otimes \left(\mathbf{x}_i \mathbf{b}_{ki}'\right)\right] \mathrm{dvec}\left(\boldsymbol{\Sigma}_k\right),\end{aligned} \tag{32}$$

where the last equation is obtained using some properties of the trace and vec operators (see, e.g., Schott, 2005, Theorems 8.9, 8.10 and 8.11). As far as $\mathrm{d}(\mathrm{d}_{ki3})$ is concerned, since

$$\mathrm{d}\left(\mathrm{d}_{ki3}\right) = \left(\mathrm{d}\boldsymbol{\gamma}_k\right)' \mathrm{d}\mathbf{b}_{ki} + \left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right] \mathrm{dvec}\left(\mathbf{x}_i \mathbf{b}_{ki}'\right), \tag{33}$$

an expression for $\mathrm{d}\mathbf{b}_{ki}$ is required. This results to be

$$\mathrm{d}\left(\mathbf{b}_{ki}\right) = -\boldsymbol{\Sigma}_k^{-1} \mathrm{d}\left(\boldsymbol{\Sigma}_k\right) \mathbf{b}_{ki} - \boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\gamma}_k\right) - \boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\Pi}_k\right) \mathbf{x}_i. \tag{34}$$

Substituting equation (34) in (33) and using some properties of the vec operator and the Kronecker product leads to the following result:

$$\begin{aligned}\mathrm{d}\left(\mathrm{d}_{ki3}\right) = {} & -\left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]' \mathbf{G}' \left(\mathbf{b}_{ki} \otimes \boldsymbol{\Sigma}_k^{-1}\right) \mathrm{d}\boldsymbol{\gamma}_k - \left(\mathrm{d}\boldsymbol{\gamma}_k\right)' \boldsymbol{\Sigma}_k^{-1} \mathrm{d}\boldsymbol{\gamma}_k + \\ & - \left(\mathrm{d}\boldsymbol{\gamma}_k\right)' \left[\boldsymbol{\Sigma}_k^{-1} \otimes \mathbf{x}_i'\right] \mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right) - \left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]' \mathbf{G}' \left[\boldsymbol{\Sigma}_k^{-1} \otimes \left(\mathbf{b}_{ki} \mathbf{x}_{ki}'\right)\right] \mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right) + \\ & - \left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]' \left(\boldsymbol{\Sigma}_k^{-1} \otimes \mathbf{x}_i\right) \mathrm{d}\boldsymbol{\gamma}_k - \left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]' \left[\boldsymbol{\Sigma}_k^{-1} \otimes \left(\mathbf{x}_i \mathbf{x}_i'\right)\right] \mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right).\end{aligned} \tag{35}$$

By inserting equations (30), (31), (32) and (35) in (29) and after some algebra, the following expression for the second order differential of $\log f_{ki}$ is obtained:

$$\begin{aligned}
\mathrm{d}^2\log f_{ki} = & -(\mathrm{d}\boldsymbol{\pi})'\,\mathbf{a}_k\mathbf{a}_k'\mathrm{d}\boldsymbol{\pi} - \frac{1}{2}\left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]'\mathbf{G}'\left[\left(\boldsymbol{\Sigma}_k^{-1}-2\mathbf{B}_{ki}\right)'\otimes\boldsymbol{\Sigma}_k^{-1}\right]\mathbf{G}\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right) + \\
& -(\mathrm{d}\boldsymbol{\gamma}_k)'\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathbf{G}\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right) - \left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]'\left[\boldsymbol{\Sigma}_k^{-1}\otimes\left(\mathbf{x}_i\mathbf{b}_{ki}'\right)\right]\mathbf{G}\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right) + \\
& -\left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]'\mathbf{G}'\left(\mathbf{b}_{ki}\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\boldsymbol{\gamma}_k - (\mathrm{d}\boldsymbol{\gamma}_k)'\,\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\boldsymbol{\gamma}_k + \\
& -(\mathrm{d}\boldsymbol{\gamma}_k)'\left[\boldsymbol{\Sigma}_k^{-1}\otimes\mathbf{x}_i'\right]\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right) - \left[\mathrm{dv}\left(\boldsymbol{\Sigma}_k\right)\right]'\boldsymbol{G}'\left[\boldsymbol{\Sigma}_k^{-1}\otimes\left(\mathbf{b}_{ki}\mathbf{x}_{ki}'\right)\right]\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right) + \\
& -\left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]'\left(\boldsymbol{\Sigma}_k^{-1}\otimes\mathbf{x}_i\right)\mathrm{d}\boldsymbol{\gamma}_k + \\
& -\left[\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right)\right]'\left[\boldsymbol{\Sigma}_k^{-1}\otimes\left(\mathbf{x}_i\mathbf{x}_i'\right)\right]\mathrm{dvec}\left(\boldsymbol{\Pi}_k'\right). \qquad (36)
\end{aligned}$$

E Proof of Proposition 1

The results given in parts *(a)*, *(b)*, *(c)* and *(d)* follow immediately from the Theorems 2.1, 2.2, 3.2 and 3.3 of White (1982), respectively.

F Proof of Theorem 3

Let

$$\mathbf{C}_I(\boldsymbol{\psi}) = I\cdot\left(H(\boldsymbol{\psi})\right)^{-1}\left(\sum_{i=1}^{I}s_i(\boldsymbol{\psi})s_i(\boldsymbol{\psi})'\right)\left(H(\boldsymbol{\psi})\right)^{-1}.$$

According to the model properties, matrices $H(\boldsymbol{\psi})$, $\mathbb{E}(H_i(\boldsymbol{\psi}))$, $\mathbf{C}(\boldsymbol{\psi})$ and $\mathbf{C}_I(\boldsymbol{\psi})$ have a block-diagonal structure. Specifically:

$$\begin{aligned}
\mathbb{E}(H_i(\boldsymbol{\psi})) &= \begin{bmatrix}\mathbb{E}(H_i(\boldsymbol{\vartheta})) & \mathbf{0}\\ \mathbf{0} & \mathbb{E}(H_i(\boldsymbol{\theta}))\end{bmatrix},\\
\mathbf{C}(\boldsymbol{\psi}) &= \begin{bmatrix}\mathbf{C}(\boldsymbol{\vartheta}) & \mathbf{0}\\ \mathbf{0} & \mathbf{C}(\boldsymbol{\theta})\end{bmatrix},\\
\mathbf{C}_I(\boldsymbol{\psi}) &= \begin{bmatrix}\mathbf{C}_I(\boldsymbol{\vartheta}) & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_I(\boldsymbol{\theta})\end{bmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{C}_I(\boldsymbol{\vartheta}) &= I\cdot\left(H(\boldsymbol{\vartheta})\right)^{-1}\left(\sum_{i=1}^{I}s_i(\boldsymbol{\vartheta})s_i(\boldsymbol{\vartheta})'\right)\left(H(\boldsymbol{\vartheta})\right)^{-1},\\
\mathbf{C}_I(\boldsymbol{\theta}) &= I\cdot\left(H(\boldsymbol{\theta})\right)^{-1}\left(\sum_{i=1}^{I}s_i(\boldsymbol{\theta})s_i(\boldsymbol{\theta})'\right)\left(H(\boldsymbol{\theta})\right)^{-1},\\
\mathbf{C}(\boldsymbol{\vartheta}) &= \left(\mathbb{E}(H_i(\boldsymbol{\vartheta}))\right)^{-1}\mathbb{E}\left(s_i(\boldsymbol{\vartheta})s_i(\boldsymbol{\vartheta})'\right)\left(\mathbb{E}(H_i(\boldsymbol{\vartheta}))\right)^{-1},\\
\mathbf{C}(\boldsymbol{\theta}) &= \left(\mathbb{E}(H_i(\boldsymbol{\theta}))\right)^{-1}\mathbb{E}\left(s_i(\boldsymbol{\theta})s_i(\boldsymbol{\theta})'\right)\left(\mathbb{E}(H_i(\boldsymbol{\theta}))\right)^{-1},
\end{aligned}$$

with $s_i(\boldsymbol{\vartheta}) = \frac{\partial l_i(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}}$ and $s_i(\boldsymbol{\theta}) = \frac{\partial l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$.

Thus, the result given in equation (16) follows from equations (12), (14) and (15).

G Proof of Theorem 4

Since the matrices $H(\boldsymbol{\psi})$, $\mathbb{E}(H_i(\boldsymbol{\psi}))$, $\mathbf{C}(\boldsymbol{\psi})$ and $\mathbf{C}_I(\boldsymbol{\psi})$ have a block-diagonal structure, the result in equation (17) follows immediately from equations (12) and (13).