Stat Comput (2016) 26:1025–1038
DOI 10.1007/s11222-015-9587-0

Using mixtures in seemingly unrelated linear regression models with non-normal errors

Giuliano Galimberti[1] · **Elena Scardovi**[1] · **Gabriele Soffritti**[1]

Received: 18 March 2014 / Accepted: 23 June 2015 / Published online: 5 July 2015

Abstract Seemingly unrelated linear regression models are introduced in which the distribution of the errors is a finite mixture of Gaussian distributions. Identifiability conditions are provided. The score vector and the Hessian matrix are derived. Parameter estimation is performed using the maximum likelihood method and an Expectation–Maximisation algorithm is developed. The usefulness of the proposed methods and a numerical evaluation of their properties are illustrated through the analysis of simulated and real datasets.

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

1 Introduction

"Seemingly unrelated regression equations" is an expression first used by Zellner (1962). It indicates a set of equations for modelling the dependence of D variables ($D \geq 1$) on one or more regressors in which the error terms in the different equations are allowed to be correlated and, thus, the equations should be jointly considered. The range of situations for which models composed of seemingly unrelated regression equations are appropriate is wide, including cross-section data, time-series data and repeated measures (see, e.g., Srivastava and Giles 1987; Park 1993). For example, these models can be used to study the effect of prices and promotional activities on sales for different brands of a given product. In particular, when $D = 2$ brands (A and B) are considered, the following system of equations can be defined:

$$\begin{cases} Y_{i1} = \beta_{01} + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \epsilon_{i1} \\ Y_{i2} = \beta_{02} + \beta_{21}x_{i3} + \beta_{22}x_{i4} + \epsilon_{i2} \end{cases} \quad i = 1, \ldots, I, \qquad (1)$$

where Y_{i1}, x_{i1} and x_{i2} are the log unit sale, the measure of a display activity and the log price, respectively, registered in week i for brand A; Y_{i2}, x_{i3} and x_{i4} provide the same information for brand B. In this situation, in order to account for a possible correlation between the error terms ϵ_{i1} and ϵ_{i2}, the linear regression models that compose system (1) should be jointly examined.

Seemingly unrelated regression models have been studied through many approaches. In Zellner (1962, 1963) feasible generalized least squares estimators are introduced and their properties are analysed. The maximum likelihood estimator from a Gaussian distribution for the error terms is investigated, for example, in Kmenta and Gilbert (1968), Oberhofer and Kmenta (1974), Magnus (1978), Park (1993). Further developments have been obtained by using bootstrap methods (see, e.g., Rocke 1989; Rilstone and Veall 1996) and a likelihood distributional analysis (Fraser et al. 2005). Many studies have been performed also in a Bayesian framework (see, e.g., Zellner 1971; Percy 1992; Ando and Zellner 2010; Zellner and Ando 2010a). Most of these methods have been developed under the assumption that the distribution of the

Electronic supplementary material The online version of this article (doi:10.1007/s11222-015-9587-0) contains supplementary material, which is available to authorized users.

✉ Giuliano Galimberti
giuliano.galimberti@unibo.it

Elena Scardovi
elena.scardovi2@unibo.it

Gabriele Soffritti
gabriele.soffritti@unibo.it

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

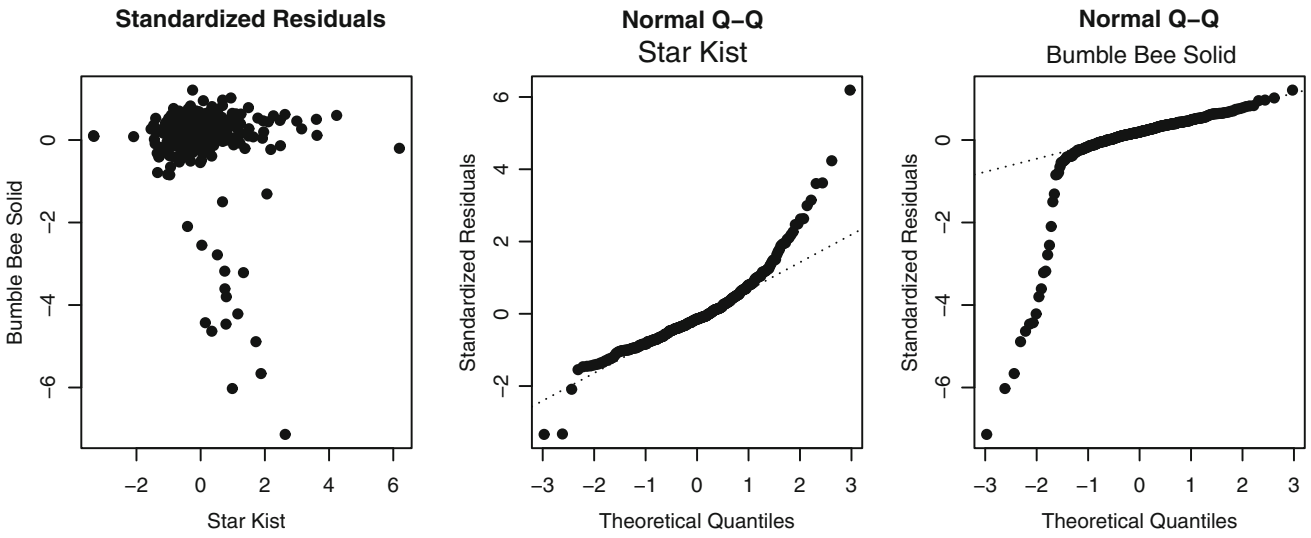

Fig. 1 Canned tuna dataset: scatterplot and Q–Q plots of the standardized residuals

error terms is Gaussian. Some of them are implemented in the `R` package `systemfit` (Henningsen and Hamann 2007).

As an example, the above-mentioned methods can be employed to estimate parameters of system (1) using data taken from Chevalier et al. (2003) and available within the `R` package `bayesm` (Rossi 2012). This dataset 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, together with a measure of the display activity and the log price of each brand. In particular, two brands are examined: Star Kist 6 oz. (brand A) and Bumble Bee Solid 6.12 oz. (brand B). Figure 1 shows the standardized residuals obtained from model (1); it is evident that they have a non-Gaussian behavior. In particular, both marginal distributions are skewed. Furthermore, a subset of observations is characterised by strongly negative residuals for the equation describing Bumble Bee sales.

The problem of dealing with non-Gaussian errors in seemingly unrelated regression models has already been tackled in the statistical literature. For example, properties of the feasible generalized least squares estimators under non-Gaussian errors are investigated in Srivastava and Maekawa (1995) and Kurata (1999). Solutions obtained using elliptical distributions are described in Ng (2002). The use of multivariate Student t errors is suggested in Kowalski et al. (1999) and Zellner and Ando (2010b).

The aim of this paper is to propose the use of Gaussian mixtures for modelling the error term distribution in a seemingly unrelated linear regression model. Finite mixtures represent a convenient and flexible framework for dealing with distributions of unknown shapes, as they can account for skewness, kurtosis and multimodality. They are widely employed in many areas of multivariate analysis, especially for model-based cluster analysis, discriminant analysis and multivariate density estimation (see, e.g., McLachlan and Peel 2000). Recently, finite mixtures of Gaussian and Student t distributions have been employed also in multiple and multivariate linear regression analysis (see, e.g., Bartolucci and Scaccia 2005; Soffritti and Galimberti 2011; Galimberti and Soffritti 2014) to handle non-normal error terms. In this context, the use of finite mixture has the advantage of capturing the effect of omitted nominal regressors from the model and obtaining robust estimates of the regression coefficients when the distribution of the error terms is non-normal.

The paper is organized as follows. Section 2 illustrates the theory behind the new methodology. Namely, the novel models are presented in Sect. 2.1. Theorem 1 provides conditions for the model identifiability (Sect. 2.2). The score vector and the Hessian matrix for the model parameter are reported in Sect. 2.3 (Theorems 2 and 3). Details about the maximum likelihood (ML) estimation through an Expectation–Maximisation (EM) algorithm are given in Sect. 2.4. Section 2.5 addresses model selection issues. Results obtained from the analysis of real datasets using the proposed approach are presented in Sect. 3, together with the results derived from models with multivariate Gaussian or Student t error terms. In particular, Sect. 3.1 describes the results for the above-mentioned canned tuna dataset. In Sect. 4 some concluding remarks are provided. Proofs of Theorems 2 and 3 and other technical results are in Appendix. Finally, further experimental results obtained from real and simulated datasets are provided in the online Supplementary material.

2 Seemingly unrelated regression models with a Gaussian mixture for the error terms

2.1 The general model

The novel model can be introduced as follows. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{id}, \ldots, Y_{iD})'$ be the vector of the D dependent variables for the ith observation, $i = 1, \ldots, I$. Furthermore, let $\mathbf{x}_{id}$ be the vector composed of the fixed values of the P_d regressors for the ith observation in the equation for the dth dependent variable, $d = 1, \ldots, D$. A seemingly unrelated regression model can be defined through the following system of equations:

$$\begin{cases} Y_{i1} = \beta_{01} + \mathbf{x}_{i1}'\boldsymbol{\beta}_1 + \epsilon_{i1} \\ \vdots \\ Y_{id} = \beta_{0d} + \mathbf{x}_{id}'\boldsymbol{\beta}_d + \epsilon_{id} \qquad i = 1, \ldots, I, \\ \vdots \\ Y_{iD} = \beta_{0D} + \mathbf{x}_{iD}'\boldsymbol{\beta}_D + \epsilon_{iD} \end{cases} \tag{2}$$

where β_{0d}, $\boldsymbol{\beta}_d$, and ϵ_{id} are the intercept, the regression coefficient vector and the error term for the ith observation in the equation for the dth dependent variable, respectively. Equation (2) can be written in compact form using the following matrix notation:

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \mathbf{X}_i'\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, I, \tag{3}$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0d}, \ldots, \beta_{0D})'$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_d', \ldots, \boldsymbol{\beta}_D'\right)'$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{id}, \ldots, \epsilon_{iD})'$, and $\mathbf{X}_i$ is the following $P \times D$ partitioned matrix:

$$\begin{bmatrix} \mathbf{x}_{i1} & \mathbf{0}_{P_1} & \cdots & \mathbf{0}_{P_1} \\ \mathbf{0}_{P_2} & \mathbf{x}_{i2} & \cdots & \mathbf{0}_{P_2} \\ \vdots & \vdots & & \vdots \\ \mathbf{0}_{P_D} & \mathbf{0}_{P_D} & \cdots & \mathbf{x}_{iD} \end{bmatrix}, \tag{4}$$

with $\mathbf{0}_{P_d}$ denoting the P_d-dimensional null vector and $P = \sum_{d=1}^{D} P_d$.

Remark 1 This definition of seemingly unrelated regression model differs from the one originally introduced by Zellner (1962); however, these two definitions are equivalent (see, for example, Park 1993). The choice of the model definition given in Eq. (3) is motivated by its analytical convenience in deriving some technical results described in this paper.

The proposed model is based on the assumption that the I error terms are independent and identically distributed, and that

$$\boldsymbol{\epsilon}_i \sim \sum_{k=1}^{K} \pi_k N_D(\boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k), \quad i = 1, \ldots, I, \tag{5}$$

where π_k's are positive weights that sum to 1, the $\boldsymbol{\nu}_k$'s are D-dimensional mean vectors that satisfy the constraint $\sum_{k=1}^{K} \pi_k \boldsymbol{\nu}_k = \mathbf{0}_D$, the $\boldsymbol{\Sigma}_k$'s are $D \times D$ positive definite symmetric matrices and $N_D(\boldsymbol{\nu}_k, \boldsymbol{\Sigma}_k)$ denotes the D-dimensional Gaussian distribution with parameters $\boldsymbol{\nu}_k$ and $\boldsymbol{\Sigma}_k$.

Given Eqs. (3) and (5), the conditional probability density function (p.d.f.) of the D-dimensional random vector $\mathbf{Y}_i$ given $\mathbf{X}_i$ is

$$\sum_{k=1}^{K} \pi_k \phi_D(\mathbf{y}_i; \boldsymbol{\lambda}_k + \mathbf{X}_i'\boldsymbol{\beta}, \boldsymbol{\Sigma}_k), \quad \mathbf{y}_i \in \mathbb{R}^D, \quad i = 1, \ldots, I, \tag{6}$$

where $\phi_D(\mathbf{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the p.d.f. of the D-dimensional Gaussian distribution $N_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ evaluated at $\mathbf{y}_i$, and $\boldsymbol{\lambda}_k = \boldsymbol{\beta}_0 + \boldsymbol{\nu}_k$. Differently from the $\boldsymbol{\nu}_k$'s, the $\boldsymbol{\lambda}_k$'s are not subject to any constraint. For this reason, in this paper the attention is focused on the vector of the model parameters given by $\boldsymbol{\theta} = \left(\boldsymbol{\pi}', \boldsymbol{\beta}', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K'\right)'$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K-1})'$, $\boldsymbol{\theta}_k = \left(\boldsymbol{\lambda}_k', \mathrm{v}(\boldsymbol{\Sigma}_k)'\right)'$ for $k = 1, \ldots, K$, with $\mathrm{v}(\boldsymbol{\Sigma}_k)$ denoting the $\frac{1}{2}D(D+1)$-dimensional vector formed by stacking the columns of the lower triangular portion of $\boldsymbol{\Sigma}_k$ (see, e.g., Schott 2005).

Suppose that the ith observation was drawn from the kth component of the mixture. Then, the equation for such an observation would be

$$\mathbf{Y}_i = \boldsymbol{\lambda}_k + \mathbf{X}_i'\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_{ik}, \tag{7}$$

where $\tilde{\boldsymbol{\epsilon}}_{ik} \sim N_D(\mathbf{0}_D, \boldsymbol{\Sigma}_k)$. Thus, the model defined by Eq. (6) can be seen as a mixture of K seemingly unrelated linear regression models with Gaussian error terms. This property of the proposed model makes it possible to establish a link with finite mixtures of regression models (see, e.g., Frühwirth-Schnatter 2006). This kind of mixtures constitutes a flexible tool for the identification of K unknown sub-populations of observations (clusters), each of which is characterised by a specific relationship between the dependent variables and the regressors. In these models it is generally assumed that each component of the mixture is associated with a cluster. It is worth noting that, according to model (6), the K clusters of observations differ in the intercepts for the D dependent variables and in the covariance matrices for the error terms, while the regression coefficients are equal across clusters. Thus, the K unknown sub-populations can also be interpreted as the categories of an unobserved (and, thus, omitted) nominal regressor that affects both the conditional expected values and covariances of the dependent variables.

In the special case where $K = 1$, model (6) results in the classical seemingly unrelated regression model with Gaussian errors. If $\mathbf{x}_{id} = \mathbf{x}_i \ \forall d$ (the vectors of the regressors for the D equations coincide), the following equality holds:

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

where $\mathbf{I}_D$ is the identity matrix of order D and $\otimes$ denotes the Kronecker product operator (see, e.g., Schott 2005). Thus, Eq. (7) can be rewritten as

$$\mathbf{Y}_i = \boldsymbol{\lambda}_k + (\mathbf{I}_D \otimes \mathbf{x}_i)' \boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_{ik} = \boldsymbol{\lambda}_k + \boldsymbol{B}'\mathbf{x}_i + \tilde{\boldsymbol{\epsilon}}_{ik}, \tag{8}$$

where $\boldsymbol{B} = \left[\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_d \cdots \boldsymbol{\beta}_D\right]$. Equation (8) corresponds to the model described in Soffritti and Galimberti (2011). Furthermore, the model proposed by Bartolucci and Scaccia (2005) can be obtained when $D = 1$. Finally, if $P_d = 0$ $\forall d$, model (6) results in the mixture model with K Gaussian components (see, e.g., McLachlan and Peel 2000).

2.2 Model identifiability

As any finite mixture model, also model (6) is invariant under permutations of the labels of the K components (see, e.g., McLachlan and Peel 2000). For the proposed model, whose parameter is $\boldsymbol{\theta} = \left(\boldsymbol{\pi}', \boldsymbol{\beta}', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K'\right)'$, the following theorem holds:

Theorem 1 *Model (6) is identifiable, provided that, for $d = 1, \ldots, D$, vectors $\{\mathbf{x}_{id}, i = 1, \ldots, I\}$ do not lie on a common $(P_d - 1)$-dimensional hyperplane.*

Proof The identifiability condition described in Theorem 1 is a generalization of the usual condition for the identifiability of a multiple linear regression model. It is required in order to guarantee identifiability of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K$ that characterise the conditional expectations for the D dependent variables.

Furthermore, consider the joint conditional p.d.f. of a random sample $\mathbf{y}_1, \ldots, \mathbf{y}_I$ from the model (6), given the fixed values of the regressors contained in $\mathbf{X}_1, \ldots, \mathbf{X}_I$:

$$f(\mathbf{y}_1, \ldots, \mathbf{y}_I; \mathbf{X}_1, \ldots, \mathbf{X}_I, \boldsymbol{\theta}) = \prod_{i=1}^{I} \left[\sum_{k=1}^{K} \pi_k \phi_D\left(\mathbf{y}_i; \boldsymbol{\lambda}_k + \mathbf{X}_i'\boldsymbol{\beta}, \boldsymbol{\Sigma}_k\right)\right]. \tag{9}$$

Formula (9) can be re-written as follows:

$$f\left(\mathbf{y}_1, \ldots, \mathbf{y}_I; \mathbf{X}_1, \ldots, \mathbf{X}_I, \boldsymbol{\theta}\right) = \sum_{j=1}^{J} \pi_j \phi_{D \cdot I}\left(\mathbf{y}; \boldsymbol{\lambda}_j + \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_j\right), \tag{10}$$

where $J = K^I$, $\mathbf{y} = \left(\mathbf{y}_1', \ldots, \mathbf{y}_I'\right)'$, $\mathbf{X} = [\mathbf{X}_1 \ldots \mathbf{X}_I]'$, $\pi_j = \prod_{i=1}^{I} \pi_{k_i^{(j)}}$, $\boldsymbol{\lambda}_j = \left(\boldsymbol{\lambda}_{k_1^{(j)}}', \ldots, \boldsymbol{\lambda}_{k_I^{(j)}}'\right)'$, $\boldsymbol{\Sigma}_j = diag(\boldsymbol{\Sigma}_{k_1^{(j)}}, \ldots, \boldsymbol{\Sigma}_{k_I^{(j)}})$ is a block diagonal matrix, and $\mathbf{k}^{(j)} = \left(k_1^{(j)}, \ldots, k_I^{(j)}\right)'$ is the jth element of the set $A_{K,I} = \{(k_1, \ldots, k_I)' : k_i \in \{1, \ldots, K\},\ i = 1, \ldots, I\}$ containing the J arrangements of the first K positive integers amongst I with repetitions. The proof can be completed by showing that mixtures (10) are identifiable. The proof of this latter result can be found in Soffritti and Galimberti (2011).

2.3 Score vector and Hessian matrix

Given a random sample $\mathbf{y}_1, \ldots, \mathbf{y}_I$ from the model (6), the log-likelihood is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{I} \ln\left(\sum_{k=1}^{K} \pi_k \phi_D(\mathbf{y}_i; \boldsymbol{\lambda}_k + \mathbf{X}_i'\boldsymbol{\beta}, \boldsymbol{\Sigma}_k)\right). \tag{11}$$

The log-likelihood (11) can be used to derive the ML estimator of $\boldsymbol{\theta}$. Furthermore, Redner and Walker (1984) showed that, under suitable conditions, an estimate of the asymptotic variance of the ML estimator of the parameters in a finite mixture model can be obtained using the Hessian matrix. In order to obtain the score vector and the Hessian matrix the following notation is introduced. Let

$$f_{ki} = \frac{\pi_k}{(2\pi)^{D/2} \det(\boldsymbol{\Sigma}_k)^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)' \times \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)\right];$$

$\alpha_{ki} = \frac{f_{ki}}{\left(\sum_{l=1}^{K} f_{li}\right)}$; $\mathbf{a}_k = \frac{1}{\pi_k}\mathbf{e}_k$ for $k = 1, \ldots, K-1$ and $\mathbf{a}_K = -\frac{1}{\pi_K}\mathbf{1}_{(K-1)}$, where $\mathbf{e}_k$ is the kth column of $\mathbf{I}_{(K-1)}$ and $\mathbf{1}_{(K-1)}$ denotes the $(K-1)$-dimensional vector having each component equal to 1; $\mathbf{b}_{ki} = \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)$; $\mathbf{B}_{ki} = \boldsymbol{\Sigma}_k^{-1} - \mathbf{b}_{ki}\mathbf{b}_{ki}'$;

$$\mathbf{c}_{ki} = \begin{bmatrix} \mathbf{b}_{ki} \\ -\frac{1}{2}\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right) \end{bmatrix},$$

where $\mathbf{G}$ denotes the duplication matrix and $\mathrm{vec}(\mathbf{B}_{ki})$ denotes the vector formed by stacking the columns of the matrix $\mathbf{B}_{ki}$ one underneath the other (see, e.g., Schott 2005).

Theorem 2 *The score vector for the parameters of model (6) is composed of the sub-vectors $\frac{\partial}{\partial \boldsymbol{\pi}'} l\left(\boldsymbol{\theta}\right)$, $\frac{\partial}{\partial \boldsymbol{\beta}'} l\left(\boldsymbol{\theta}\right)$, $\frac{\partial}{\partial \boldsymbol{\theta}_1'} l\left(\boldsymbol{\theta}\right)$, $\ldots$, $\frac{\partial}{\partial \boldsymbol{\theta}_K'} l\left(\boldsymbol{\theta}\right)$, where*

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

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

$$\frac{\partial}{\partial \boldsymbol{\theta}_k} l(\boldsymbol{\theta}) = \sum_{i=1}^{I} \alpha_{ki} \mathbf{c}_{ki}, \quad k = 1, \ldots, K,$$

with $\bar{\mathbf{a}}_i = \sum_{k=1}^{K} \alpha_{ki} \mathbf{a}_k$ *and* $\bar{\mathbf{b}}_i = \sum_{k=1}^{K} \alpha_{ki} \mathbf{b}_{ki}$.

Theorem 3 *The Hessian matrix* $H(\boldsymbol{\theta})$ *for the parameters of model (6) is equal to*

$$\begin{bmatrix} \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial \boldsymbol{\pi} \partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta}) \\ \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial \boldsymbol{\theta}_1 \partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\pi}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) & \frac{\partial^2}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\theta}_1'} l(\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial \boldsymbol{\theta}_K \partial \boldsymbol{\theta}_K'} l(\boldsymbol{\theta}) \end{bmatrix}, \tag{12}$$

where

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

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

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

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} l(\boldsymbol{\theta}) = -\sum_{i=1}^{I} \mathbf{X}_i \left[\bar{\mathbf{B}}_i + \bar{\mathbf{b}}_i \bar{\mathbf{b}}_i' \right] \mathbf{X}_i',$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}_k'} l(\boldsymbol{\theta}) = -\sum_{i=1}^{I} \alpha_{ki} \mathbf{X}_i \left[\mathbf{F}_{ki} - \left(\mathbf{b}_{ki} - \bar{\mathbf{b}}_i \right) \mathbf{c}_{ki}' \right], \quad k = 1, \ldots, K,$$

$$\frac{\partial^2}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k'} l(\boldsymbol{\theta}) = -\sum_{i=1}^{I} \alpha_{ki} \left[\mathbf{C}_{ki} - (1 - \alpha_{ki}) \mathbf{c}_{ki} \mathbf{c}_{ki}' \right], \quad k = 1, \ldots, K,$$

$$\frac{\partial^2}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_h'} l(\boldsymbol{\theta}) = -\sum_{i=1}^{I} \alpha_{ki} \alpha_{hi} \mathbf{c}_{ki} \mathbf{c}_{hi}', \quad \forall k \neq h,$$

with $\bar{\mathbf{B}}_i = \sum_{k=1}^{K} \alpha_{ki} \left(\boldsymbol{\Sigma}_k^{-1} - \mathbf{b}_{ki} \mathbf{b}_{ki}' \right)$, $\mathbf{F}_{ki} = \left[\boldsymbol{\Sigma}_k^{-1} \left(\mathbf{b}_{ki}' \otimes \boldsymbol{\Sigma}_k^{-1} \right) \mathbf{G} \right]$ *and*

$$\mathbf{C}_{ki} = \begin{bmatrix} \boldsymbol{\Sigma}_k^{-1} & \left(\mathbf{b}_{ki}' \otimes \boldsymbol{\Sigma}_k^{-1} \right) \mathbf{G} \\ \mathbf{G}' \left(\mathbf{b}_{ki} \otimes \boldsymbol{\Sigma}_k^{-1} \right) & \frac{1}{2} \mathbf{G}' \left[\left(\boldsymbol{\Sigma}_k^{-1} - 2\mathbf{B}_{ki} \right) \otimes \boldsymbol{\Sigma}_k^{-1} \right] \mathbf{G} \end{bmatrix}.$$

Proofs of Theorems 2 and 3 are provided in Appendix 1 and 2, respectively.

Remark 2 After some suitable simplifications, Theorems 2 and 3 provide the score vector and the Hessian matrix also for the models introduced in Bartolucci and Scaccia (2005) and Soffritti and Galimberti (2011). Furthermore, they represent a generalization of Theorem 1 in Boldea and Magnus (2009).

2.4 An EM algorithm for maximum likelihood estimation

The score vector and the Hessian matrix described in Sect. 2.3 can be used to compute the ML estimates of the model parameter $\boldsymbol{\theta}$ through a Newton-Raphson algorithm for the maximisation of $l(\boldsymbol{\theta})$ in Eq. (11). However, the evaluation of the Hessian matrix at each iteration can be computationally expensive, especially with large samples. In order to avoid this problem, in this Section an EM algorithm is developed by resorting to the approach for incomplete-data problems (Dempster et al. 1977; McLachlan and Krishnan 2008). This approach is widely employed in finite mixture models, where the source of unobservable information is the specific component of the mixture model that generates each sample observation. Specifically, this unobservable information for the ith observation can be described by the K-dimensional vector $\mathbf{z}_i' = (z_{i1}, \ldots, z_{iK})$, where $z_{ik} = 1$ when $\mathbf{y}_i$ is generated from the kth component, and $z_{ik} = 0$ otherwise, for $k = 1, \ldots, K$. Thus, $\sum_{k=1}^{K} z_{ik} = 1$, $i = 1, \ldots, I$.

Consider the following hierarchical representation for $\mathbf{y}_i | \mathbf{X}_i$:

$$\mathbf{z}_i \sim mult(1, \pi_1, \ldots, \pi_K),$$
$$\mathbf{y}_i | (\mathbf{X}_i, z_{ik} = 1) \sim N_D \left(\boldsymbol{\lambda}_k + \mathbf{X}_i' \boldsymbol{\beta}, \boldsymbol{\Sigma}_k \right),$$

where $mult(1, \pi_1, \ldots, \pi_K)$ denotes the K-dimensional multinomial distribution with parameters $\pi_1, \ldots, \pi_K$, and assume that this representation independently holds for $i = 1, \ldots, I$. Then, the complete-data log-likelihood $l_c(\boldsymbol{\theta})$ of model (6) can be expressed as

$$l_c(\boldsymbol{\theta}) = \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \ln f_{ki}. \tag{13}$$

The first order differential of $l_c(\boldsymbol{\theta})$ is

$$\begin{aligned} \mathrm{d} l_c(\boldsymbol{\theta}) &= \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \mathrm{d} \ln f_{ki} \\ &= (\mathrm{d}\boldsymbol{\pi})' \sum_{k=1}^{K} z_{\cdot k} \mathbf{a}_k + (\mathrm{d}\boldsymbol{\beta})' \sum_{i=1}^{I} \sum_{k=1}^{K} z_{ik} \mathbf{X}_i \mathbf{b}_{ki} \\ &\quad + \sum_{k=1}^{K} (\mathrm{d}\boldsymbol{\theta}_k)' \sum_{i=1}^{I} z_{ik} \mathbf{c}_{ki} \end{aligned}$$

$$= (\mathrm{d}\boldsymbol{\pi})' \sum_{k=1}^{K} z_{\cdot k}\mathbf{a}_k + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\left[(\mathrm{d}\boldsymbol{\lambda}_k)' + (\mathrm{d}\boldsymbol{\beta})'\mathbf{X}_i\right]\mathbf{b}_{ki} \quad (14)$$

$$-\frac{1}{2}\sum_{k=1}^{K} \mathrm{d}\,(\mathrm{v}\boldsymbol{\Sigma}_k)'\mathbf{G}'\mathrm{vec}\left(\sum_{i=1}^{I} z_{ik}\mathbf{B}_{ki}\right) \quad (15)$$

where the second and third equalities are obtained using Eq. (34) in Appendix 1, and $z_{\cdot k} = \sum_{i=1}^{I} z_{ik}$.

To determine the solution of each M step of the EM algorithm, it is convenient to introduce the following notation. Let $\mathrm{d}l_{c2}$ and $\mathrm{d}l_{c3}$ denote the expressions in Eqs. (14) and (15), respectively. Let $\boldsymbol{\gamma} = \left(\boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_K', \boldsymbol{\beta}'\right)'$ be the $(D \cdot K + P)$-dimensional vector comprising the intercepts of all components and regression coefficients for all dependent variables. $\mathbf{O}_k$ is a matrix of dimension $(D \cdot K) \times D$ obtained by extracting the columns of the matrix $\mathbf{I}_{(D\cdot K)}$ from the $(1 + (k-1)\cdot D)$th to the $(D + (k-1)\cdot D)$th, for $k = 1, \ldots, K$. Furthermore, let $\mathbf{X}_{ki} = \begin{bmatrix}\mathbf{O}_k \\ \mathbf{X}_i\end{bmatrix}$; this is a matrix of dimension $(D \cdot K + P) \times D$ such that $\mathbf{X}_{ki}'\boldsymbol{\gamma} = \boldsymbol{\lambda}_k + \mathbf{X}_i'\boldsymbol{\beta}$ and $\mathbf{X}_{ki}'\mathrm{d}\boldsymbol{\gamma} = \mathrm{d}\boldsymbol{\lambda}_k + \mathbf{X}_i'\mathrm{d}\boldsymbol{\beta}$. Using this latter notation, the expressions of $\mathrm{d}l_{c2}$ and $\mathrm{d}l_{c3}$ in Eqs. (14) and (15) turn into

$$\begin{aligned}\mathrm{d}l_{c2} &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}(\mathrm{d}\boldsymbol{\gamma})'\mathbf{X}_{ki}\mathbf{b}_{ki} \\ &= (\mathrm{d}\boldsymbol{\gamma})'\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_i - \mathbf{X}_{ki}'\boldsymbol{\gamma}\right) \\ &= (\mathrm{d}\boldsymbol{\gamma})'\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{y}_i \\ &\quad - (\mathrm{d}\boldsymbol{\gamma})'\left(\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_{ki}'\right)\boldsymbol{\gamma}, \end{aligned} \quad (16)$$

$$\begin{aligned}\mathrm{d}l_{c3} &= -\frac{1}{2}\sum_{k=1}^{K}\mathrm{d}\,(\mathrm{v}\boldsymbol{\Sigma}_k)'\mathbf{G}'\mathrm{vec}\left(\sum_{i=1}^{I} z_{ik}\boldsymbol{\Sigma}_k^{-1} - \sum_{i=1}^{I} z_{ik}\mathbf{b}_{ki}\mathbf{b}_{ki}'\right) \\ &= -\frac{1}{2}\sum_{k=1}^{K}\mathrm{d}\,(\mathrm{v}\boldsymbol{\Sigma}_k)'\mathbf{G}'\mathrm{vec}\left(z_{\cdot k}\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\boldsymbol{\Sigma}_k^{-1}\right), \end{aligned} \quad (17)$$

where $\mathbf{S}_k = \sum_{i=1}^{I} z_{ik}\left(\mathbf{y}_i - \mathbf{X}_{ki}'\boldsymbol{\gamma}\right)\left(\mathbf{y}_i - \mathbf{X}_{ki}'\boldsymbol{\gamma}\right)'$. Using Eq. (17) and some properties of the vec operator (see, in particular, Schott 2005, Theorem 8.11) it is also possible to write

$$\begin{aligned}\mathrm{d}l_{c3} &= \frac{1}{2}\sum_{k=1}^{K}\mathrm{d}\,(\mathrm{v}\boldsymbol{\Sigma}_k)'\mathbf{G}'\mathrm{vec}\left[\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{S}_k - z_{\cdot k}\boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\right] \\ &= \frac{1}{2}\sum_{k=1}^{K}\mathrm{d}\,(\mathrm{v}\boldsymbol{\Sigma}_k)'\mathbf{G}'\left(\boldsymbol{\Sigma}_k^{-1}\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathbf{G}\mathrm{v}\left(\mathbf{S}_k - z_{\cdot k}\boldsymbol{\Sigma}_k\right). \end{aligned} \quad (18)$$

Thus, the following alternative expression for $\mathrm{d}l_c(\boldsymbol{\theta})$ holds:

$$\begin{aligned}\mathrm{d}l_c(\boldsymbol{\theta}) &= (\mathrm{d}\boldsymbol{\pi})'\sum_{k=1}^{K} z_{\cdot k}\mathbf{a}_k + (\mathrm{d}\boldsymbol{\gamma})'\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{y}_i \\ &\quad - (\mathrm{d}\boldsymbol{\gamma})'\left(\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_{ki}'\right)\boldsymbol{\gamma} \\ &\quad + \frac{1}{2}\sum_{k=1}^{K}\mathrm{d}\,(\mathrm{v}\boldsymbol{\Sigma}_k)'\mathbf{G}'\left(\boldsymbol{\Sigma}_k^{-1}\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathbf{G}\mathrm{v}\left(\mathbf{S}_k - z_{\cdot k}\boldsymbol{\Sigma}_k\right). \end{aligned} \quad (19)$$

The first derivatives of $l_c(\boldsymbol{\theta})$ with respect to the parameters $\boldsymbol{\pi}$, $\boldsymbol{\gamma}$ and $\mathrm{v}\boldsymbol{\Sigma}_k$ $(k = 1, \ldots, K)$ are:

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\pi}}l_c(\boldsymbol{\theta}) &= \sum_{k=1}^{K} z_{\cdot k}\mathbf{a}_k, \\ \frac{\partial}{\partial\boldsymbol{\gamma}}l_c(\boldsymbol{\theta}) &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{y}_i \\ &\quad - \left(\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_{ki}'\right)\boldsymbol{\gamma}, \\ \frac{\partial}{\partial\,(\mathrm{v}\boldsymbol{\Sigma}_k)}l_c(\boldsymbol{\theta}) &= \frac{1}{2}\mathbf{G}'\left(\boldsymbol{\Sigma}_k^{-1}\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathbf{G}\mathrm{v}\left(\mathbf{S}_k - z_{\cdot k}\boldsymbol{\Sigma}_k\right), \\ &\quad k = 1, \ldots, K. \end{aligned}$$

In order to maximise $l_c(\boldsymbol{\theta})$ these derivatives are set equal to zero. By solving the resulting system of equations the following expressions are obtained:

$$\pi_k^* = z_{\cdot k}/I, \quad k = 1, \ldots, K, \quad (20)$$

and, provided that the matrix $\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_{ki}'$ is non-singular,

$$\boldsymbol{\gamma}^* = \left(\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_{ki}'\right)^{-1}\sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\mathbf{X}_{ki}\boldsymbol{\Sigma}_k^{-1}\mathbf{y}_i, \quad (21)$$

$$\boldsymbol{\Sigma}_k^* = z_{\cdot k}^{-1}\mathbf{S}_k, \quad k = 1, \ldots, K. \quad (22)$$

Using some initial value for $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(0)}$, the E-step on the $(r+1)$th iteration of the EM algorithm is effected by simply replacing z_{ik} by $E_{\boldsymbol{\theta}^{(r)}}(z_{ik}|\mathbf{y}_i, \mathbf{x}_i) = P_{\boldsymbol{\theta}^{(r)}}(z_{ik} = 1|\mathbf{y}_i, \mathbf{x}_i) = p_{ik}^{(r)}$, which is the posterior probability that $\mathbf{y}_i$ is generated from the kth component of the mixture. Namely:

$$p_{ik}^{(r)} = \frac{\pi_k^{(r)}\phi_D\left(\mathbf{y}_i; \boldsymbol{\lambda}_k^{(r)} + \mathbf{X}_i'\boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}_k^{(r)}\right)}{\sum_{h=1}^{K}\pi_h^{(r)}\phi_D\left(\mathbf{y}_i; \boldsymbol{\lambda}_h^{(r)} + \mathbf{X}_i'\boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}_h^{(r)}\right)}. \quad (23)$$

On the M-step at the $(r+1)$th iteration of the EM algorithm, the updated estimates of the model parameters $\pi_k^{(r+1)}$, $\boldsymbol{\gamma}^{(r+1)}$ and $\boldsymbol{\Sigma}_k^{(r+1)}$ are computed using Eqs. (20), (21) and (22), respectively, where z_{ik} is replaced by $p_{ik}^{(r)}$. As Eq. (21) depends on the $\boldsymbol{\Sigma}_k$'s and Eq. (22) depends on $\boldsymbol{\gamma}$, the updated estimates of such parameters at the $(r+1)$th iteration are obtained through an iterative process in which the estimate of $\boldsymbol{\gamma}$ is updated, given an estimate of the $\boldsymbol{\Sigma}_k$'s, and vice versa, until convergence.

Once the convergence is reached, in addition to the parameter estimates the EM algorithm also provides estimates of the posterior probabilities using Eq. (23). These estimated posterior probabilities can be used to partition the I observations into K clusters, by assigning each observation to the component showing the highest posterior probability.

A crucial point of any application of an EM algorithm is the choice of $\boldsymbol{\theta}^{(0)}$. In general, convergence with the EM algorithm is slow, and this problem can be worsened by a poor choice of $\boldsymbol{\theta}^{(0)}$ (see, e.g., McLachlan and Peel 2000). Furthermore, since in a mixture model the likelihood function usually has multiple maxima, different starting values for the model parameters can lead to different estimates. A way to deal with the latter problem is to consider multiple random initializations and then to select the parameter estimate leading to the largest likelihood value. For the model described in this paper, starting values for the EM algorithm can be chosen by adapting the strategies described in Galimberti and Soffritti (2014). $\boldsymbol{\beta}^{(0)}$ can be obtained by fitting the standard seemingly unrelated regression model; the sample residuals of this model can be used to derive starting values for the remaining parameters. For example, the estimates from the fitting of a Gaussian mixture model with K components to the sample residuals can be employed. Alternatively, a random initialisation can be obtained by randomly partitioning the sample residuals into K groups and by computing the corresponding relative group sizes, group mean vectors and group covariance matrices. Further details and a practical comparison of these two different approaches to initialising $\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ are provided in the online Supplementary material (see Sect. A).

2.5 Model selection

In the EM algorithm described in Sect. 2.4 the value of K is considered to be fixed and known. However, in most situations such value is not known and has to be estimated from the data. Several approaches dealing with this problem have been investigated in the framework of finite mixture models; most of them are model selection techniques which seek to find a parsimonious model that adequately describes the observed data (see, e.g., McLachlan and Peel 2000, chapter 6). An example is given by the Bayesian Information Criterion (Schwarz 1978):

$$BIC_M = 2 \max [l_M] - \text{npar}_M \log(I),$$

where $\max [l_M]$ is the maximum of the log-likelihood of a model M for the given sample of I observations, and npar_M is the number of unconstrained parameters to be estimated for that model. This criterion allows to trade-off the fit and parsimony of a given model: the greater the BIC, the better the model. In the following Sections, this criterion is used not only to establish the most appropriate value of K to be used in model (6) but also to compare models characterised by different error term distributions.

3 Experimental results

The usefulness and effectiveness of the methods described in Sect. 2 are illustrated through some examples based on real and simulated datasets. The main results obtained on two real datasets are summarised in the following Sections. Further results are reported in the online Supplementary material. In particular, a numerical evaluation of properties of the ML estimates is carried out on both simulated and real datasets, with a special emphasis on the regression coefficients (see Sects. B and C).

All analyses are performed in the `R` environment (R Core Team 2013). A specific function is used, that implements the ML estimation through the EM algorithm described in Sect. 2.4 and the calculation of the Hessian matrix defined in Theorem 3. For each examined dataset, parameters of model (6) are estimated with a value of K from 1 to K_{max} (the values of K_{max} used in the experiments are described in the following Sections). For the results illustrated in this Section, the starting values of the model parameters are obtained through a strategy that fits Gaussian mixture models to the sample residuals of the classical seemingly unrelated linear regression model. The package `mclust` (Fraley and Raftery 2002; Fraley et al. 2012) is used without imposing any restriction on the component-covariance matrices. 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 $r+1$ (McNicholas and Murphy 2008). The stopping rules for each M step are either when the mean Euclidean distance between two consecutive estimated vectors of the model parameters is lower than 10^{-8} or when the number of iterations reaches the maximum of 500. Estimates of the standard errors of the ML estimators of the regression coefficients are computed as the square root of the diagonal elements of $H(\hat{\boldsymbol{\theta}})^{-1}$ that refer to $\boldsymbol{\beta}$. Asymptotic confidence intervals for the regression coefficients are also

Table 1 Maximised log-likelihood and BIC value for the models fitted to the tuna dataset

Models	$l_M(\hat{\boldsymbol{\theta}})$	npar_M	BIC_M
$K=1$	−652.573	9	−1357.55
$K=2$	−290.753	15	−668.85
$K=3$	−251.536	21	−625.36
$K=4$	−245.083	27	−647.39
Student t	−333.064	10	−724.36

obtained by assuming an asymptotic normal distribution for the ML estimators.

Model (2) under the assumption of multivariate Student t error terms is also fitted. In order to perform this latter task, a modified version of the EM algorithm described in Lange et al. (1989) is developed and implemented in a specific R function. This function selects starting values for $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ by fitting univariate Student t linear regression models for each dependent variable. Starting values of all the other model parameters are obtained by fitting a multivariate Student t distribution to the residuals derived from the D univariate regression models. The EM stopping rules described above are exploited also in this second function.

3.1 Canned tuna dataset

In this example, in which $D=2$, $P_1=P_2=2$, the parameters of the system of Eq. (1) for the canned tuna brands Star Kist 6 oz. and Bumble Bee Solid 6.12 oz. (see Sect. 1) are estimated. The value of K_{max} used in this experiment is 4. Table 1 provides some model fitting results.

According to the BIC, the model with $K=3$ components provides the best description of the joint linear dependence of the log unit sales on the display activity and the log price for the two examined brands. The estimates of the prior probabilities of this model are 0.744, 0.195 and 0.061. Tables 2 and 3 report the estimates of the remaining parameters.

By comparing the three components it emerges that the second component shows the highest value of the intercepts for both dependent variables (Table 2); this component is also highly homogeneous with respect to the log unit sales of Bumble Bee Solid 6.12 oz. A slightly lower estimated intercept and a very high estimated variance for the log unit sales of Bumble Bee Solid 6.12 oz. are the main specific features of the third component together with a negative estimated correlation between the log unit sales of the two brands.

As far as the effects of the regressors on the dependent variables are concerned (Table 3), they can be considered all significant (none of the 95 % asymptotic confidence intervals contains 0). The impact of the display activity on the log unit sales for Star Kist 6 oz. seems to be slightly higher than for Bumble Bee Solid 6.12 oz. The opposite result holds true for the effect of the log price.

Table 2 Estimates of parameters $\boldsymbol{\lambda}_k$ and $\boldsymbol{\Sigma}_k$ obtained from the best model fitted to the canned tuna dataset

	Y_1	Y_2
$\hat{\boldsymbol{\lambda}}_1'$	8.584	9.937
$\hat{\boldsymbol{\lambda}}_2'$	9.162	9.977
$\hat{\boldsymbol{\lambda}}_3'$	9.094	7.301
$\hat{\boldsymbol{\Sigma}}_1$	0.082	0.017
	0.265	0.050
$\hat{\boldsymbol{\Sigma}}_2$	0.730	0.021
	0.144	0.029
$\hat{\boldsymbol{\Sigma}}_3$	0.264	−0.528
	−0.623	2.724

Estimated correlation coefficients between dependent variables (in italics) are reported in the lower triangular parts of the three covariance matrices

Table 3 Estimates of the regression coefficients (RC) calculated from the best model fitted to the canned tuna dataset and their estimated SE

Dependent variable		Regressors	
Y_1		X_1	X_2
	RC	0.242	−3.176
	SE	0.067	0.213
	CI	(0.111, 0.373)	(−3.593, −2.759)
Y_2		X_3	X_4
	RC	0.123	−3.876
	SE	0.041	0.373
	CI	(0.043, 0.203)	(−4.607, −3.145)

The asymptotic CIs are computed at the 95 % level of confidence

The 338 weeks can be partitioned into three clusters by assigning each week to the component of the mixture that register the highest posterior probability. Most of the weeks (278) are assigned to the first cluster (CL1), while only 19 weeks are classified in the third cluster (CL3). Figure 2 shows the scatterplot of the log unit sales for the two brands in all weeks after removing the estimated effects of the two regressors. Weeks are labelled according to the cluster they are assigned. An interesting feature of the obtained classification emerging from this plot is that the third cluster is composed of weeks in which Bumble Bee Solid 6.12 oz. tuna sales register a relatively low mean level. Furthermore, it is also relevant to highlight that 17 out of the 19 weeks in the third cluster are consecutive from week 58 to week 74. According to additional information about the canned tuna dataset available at the University of Chicago website (http://research.chicagobooth.edu/kilts/marketing-databases/dominicks/), these weeks correspond to the period from mid-October 1990 to mid-February 1991.

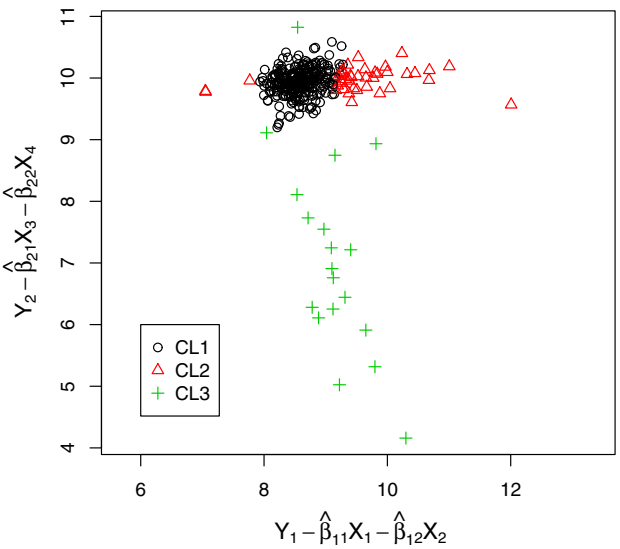

Fig. 2 Canned tuna dataset: scatterplot of the weekly log unit sales for the two brands after removing the estimated effects of the two regressors. Weeks are labelled according to their cluster membership

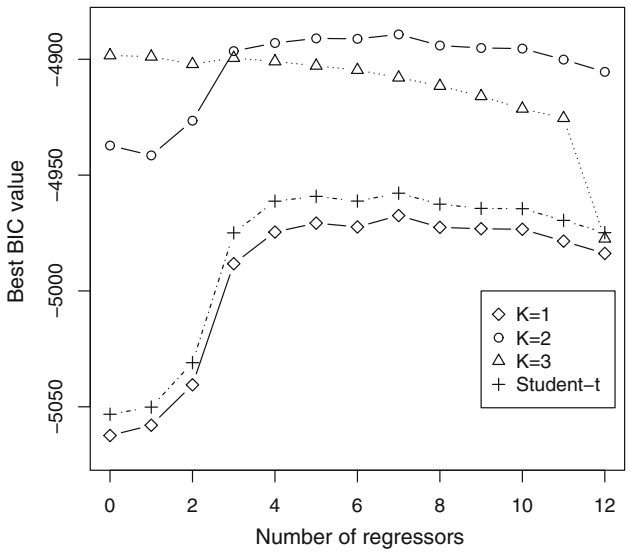

Fig. 3 AIS dataset: best BIC values by total number of regressors and number of components

It is worth noting that in that same period the U.S. non-governmental organization Earth Island Institute promoted a worldwide boycott campaign encouraging consumers not to buy Bumble Bee tuna because Bumble Bee was found to be buying yellow-fin tuna caught by dolphin-unsafe techniques (Baird and Quastel 2011).

3.2 AIS dataset

This Section describes some results obtained from the analysis of the Australian Institute of Sport (AIS) dataset (Cook and Weisberg 1994). Namely, the interest is focused on studying the joint linear dependence of $D = 4$ biometrical variables (Y_1: body mass index (BMI), Y_2: sum of skin folds (SSF), Y_3: percentage of body fat (PBF), Y_4: lean body mass (LBM)) on three variables providing information about blood composition (red cell count (RCC), white cell count (WCC), plasma ferritine concentration (PFC)). The same problem was investigated by Soffritti and Galimberti (2011) using multivariate linear regression models.

Seemingly unrelated linear regression models from Eq. (6) are estimated for $K = 1, 2, 3$. Since the choice of the regressors to be used for each biometrical variable may be questionable, for each value of K and each dependent variable an exhaustive search for the relevant regressors is performed. Thus, for each value of K, $2^{3 \cdot D} = 4096$ different regression models are fitted to the dataset. The same exhaustive search is carried out using seemingly unrelated regression models with multivariate Student t error terms. Thus, a total number of 16384 different models are examined. The total number P of regressors included in a model is between 0 and 12. The EM algorithm has failed due to the singularity of some matrices for two models when $K = 2$ and 40 models when $K = 3$.

In this situation, the BIC defined in Sect. 2.5 can be used not only to choose the best distribution for the error terms (multivariate Gaussian, Student t or Gaussian mixture in this study), but also the best subset of regressors for each equation in the system (2). Figure 3 shows the BIC values of the fitted models with the best trade-off (i.e., the maximum value of the BIC) among all the models having the same error distribution (Gaussian mixture with $K = 1, 2, 3$ and Student t) and the same number of regressors ($P = 0, \ldots, 12$). By comparing models having the same value of P it emerges that the best performance is obtained using Gaussian models with three components when the total number of regressors is low ($P = 0, 1, 2$); otherwise, Gaussian models with two components should be preferred. Thus, the introduction of a finite mixture for the distribution of the error terms allows to obtain a relevant improvement with respect to seemingly unrelated regression models with both Gaussian and Student t errors, for all P. Note that Student t models achieve a slightly better performance than classical models with Gaussian errors.

If models are compared by controlling the error distribution, $P = 7$ regressors should be used when $K = 1, 2$ and with Student t errors. Namely, for these three cases, the selected regressors for the equations of the dependent variables BMI, PBF and LBM are RCC and PFC; only RCC is selected as a relevant regressor for the equation of SSF. Hence, the numbers of regressors in the $D = 4$ equations associated with these three cases are $P_1 = P_3 = P_4 = 2$ and $P_2 = 1$. When $K = 3$, the best trade-off is obtained using a model without regressors ($P_1 = P_2 = P_3 = P_4 = 0$). Some results concerning these four latter models are illustrated in Table 4. Overall, according to the BIC the best

Table 4 Maximised log-likelihood and BIC value for the best models fitted to the AIS dataset

Models	P	$l_M(\hat{\boldsymbol{\theta}})$	npar_M	BIC_M
$K=1$	7	−2427.993	21	−4967.46
$K=2$	7	−2349.083	36	−4889.26
$K=3$	0	−2332.382	44	−4898.33
Student t	7	−2420.517	22	−4957.82

Table 5 Estimates of parameters $\boldsymbol{\lambda}_k$ and $\boldsymbol{\Sigma}_k$ obtained from the best model fitted to the AIS dataset

	BMI	SSF	PBF	LBM
$\hat{\boldsymbol{\lambda}}_1'$	10.04	86.57	23.19	−7.02
$\hat{\boldsymbol{\lambda}}_2'$	12.99	136.43	32.52	−4.88
$\hat{\boldsymbol{\Sigma}}_1$	3.96	5.14	−0.09	18.99
	0.198	169.94	31.21	2.63
	−0.017	*0.899*	7.10	−8.73
	0.810	*0.017*	*−0.278*	138.82
$\hat{\boldsymbol{\Sigma}}_2$	6.85	17.43	0.89	14.59
	0.244	744.38	107.03	−54.50
	0.080	*0.928*	17.88	−15.05
	0.681	*−0.244*	*−0.435*	67.07

Estimated correlation coefficients between dependent variables (in italics) are reported in the lower triangular parts of $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$

model is the one with $K=2$. In this model, the estimates of the parameters π_1 and π_2 are 0.619 and 0.381. Tables 5 and 6 report the estimates of the remaining parameters. Compared to the second component, the first component is characterised by lower values of the intercepts for all dependent variables and lower variances for BMI, SSF and PBF. Further differences between components concern some correlations (see the lower triangular parts of $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ in Table 5). The asymptotic confidence intervals for the regression coefficients are reported in Table 6, along with the estimated standard errors of the corresponding ML estimators. None of such confidence intervals contains the 0 value.

The best model can be used to assign each athlete to the component of the mixture that register the highest posterior probability, thus producing a partition of the sample into two clusters. Most of the athletes assigned to the second cluster are female (79.2 %), while 68.8 % of the athletes classified in the first cluster are male (Table 7). This classification of the athletes is statistically associated with athletes' gender ($\chi^2 = 43.96$, $p\,value = 3.36 \cdot 10^{-11}$). Thus, the omitted regressor captured by the selected model is strongly connected with athletes' gender. However, the partition discovered by the model has a misclassification rate equal to 0.27. Using multivariate regression models (by including RCC, WCC and PFC in each equation), Soffritti and Galimberti (2011) obtained a partition with a slightly lower misclassification rate (0.25). Note that this latter result was obtained by estimating a model with a larger number of parameters but a lower BIC value (−4905.44).

Table 6 Estimates of the regression coefficients (RC) calculated from the best model fitted to the AIS dataset and their estimated SEs

Dependent variable	Regressors	
	RCC	PFC
BMI		
RC	2.286	0.01310
SE	0.339	0.00300
CI	(1.621, 2.950)	(0.007, 0.019)
SSF		
RC	−7.746	–
SE	2.783	–
CI	(−13.200, −2.292)	–
PBF		
RC	−2.724	−0.00500
SE	0.565	0.00186
CI	(−3.832, −1.616)	(−0.009, −0.001)
LBM		
RC	14.211	0.05232
SE	1.649	0.01511
CI	(10.979, 17.442)	(0.023, 0.082)

The asymptotic CIs are computed at the 95 % level of confidence

Table 7 AIS dataset: joint classification of the athletes according to gender and cluster membership estimated by the best model

Cluster	Gender		
	Female	Male	
1	39	86	125
2	61	16	77
	100	102	202

4 Concluding remarks

In this paper, multivariate Gaussian mixtures are used to model the error terms in seemingly unrelated linear regressions. This allows to exploit the flexibility of mixtures for dealing with non-Gaussian errors. In particular, the resulting models are able to handle asymmetric and heavy-tailed errors and to detect and capture the effect of relevant nominal regressors omitted from the model. Furthermore, by setting the number of components equal to one or by constraining all the equations to have the same regressors, some solutions already described in the statistical literature can be obtained as special cases.

The approach described in this paper can be extended by considering other parametric families for the mixture components. For example, similarly to Galimberti and Soffritti (2014), mixtures of multivariate Student t distributions could

be used to define a more general class of models, which contains model (6) as a limiting case.

Parsimonious seemingly unrelated linear regression models can be obtained by introducing some constraints on the component covariance matrices $\boldsymbol{\Sigma}_k$'s, based on the spectral decomposition (see, e.g., Banfield and Raftery 1993; Celeux and Govaert 1995; McLachlan et al. 2003; McNicholas and Murphy 2008). Such models could provide a good fit for some datasets by using a lower number of parameters; they could be useful especially in the presence of a large number of dependent variables.

In Sect. 3 the BIC is used to select the relevant regressors in each equation as well as the number of mixture components. The use of this criterion can be motivated on the basis of both theoretical and practical results (see, e.g., Cutler and Windham 1994; Keribin 2000; Ray and Lindsay 2008; Maugis et al. 2009a, b). Clearly, other model selection criteria could be used, such as the ICL (Biernacki et al. 2000), which additionally takes into account the uncertainty of the classification of the sample units to the mixture components.

Some computational issues could arise when using the models proposed in this paper. For example, when the number of candidate regressors is large, an exhaustive search for the relevant regressors for each equation could be unfeasible. A possible solution could be obtained by resorting to stochastic search techniques, such as genetic algorithms (see, e.g., Chatterjee et al. 1996). As far as the EM algorithm is concerned, different initialisation strategies may be considered and evaluated (see, e.g., Biernacki et al. 2003; Melnykov and Melnykov 2012). Although these issues are not the main focus of this paper, they could deserve further investigation.

Appendix 1: Proof of Theorem 2

The proof is based on the computation of the first order differential of $l(\boldsymbol{\theta})$. The model log-likelihood in Eq. (11) can be expressed as $l(\boldsymbol{\theta}) = \sum_{i=1}^{I} \ln\left(\sum_{k=1}^{K} f_{ki}\right)$. Thus, the first differential of $l(\boldsymbol{\theta})$ is

$$\mathrm{d}l(\boldsymbol{\theta}) = \sum_{i=1}^{I} \mathrm{d}\ln\left(\sum_{k=1}^{K} f_{ki}\right) = \sum_{i=1}^{I}\left(\sum_{k=1}^{K} \alpha_{ki}\,\mathrm{d}\ln f_{ki}\right). \quad (24)$$

Up to an additive constant, $\ln f_{ki}$ is equal to

$$\ln \pi_k - \frac{1}{2}\ln\det(\boldsymbol{\Sigma}_k) - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)'\right],$$

and

$$\mathrm{d}\ln f_{ki} = \mathrm{d}\ln\pi_k + \mathrm{d}_{ki1} + \mathrm{d}_{ki2} + \mathrm{d}_{ki3}, \quad (25)$$

where

$$\mathrm{d}_{ki1} = -\frac{1}{2}\mathrm{d}\left(\ln\det(\boldsymbol{\Sigma}_k)\right), \quad (26)$$

$$\mathrm{d}_{ki2} = -\frac{1}{2}\mathrm{tr}\left[\mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right)\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)'\right], \quad (27)$$

$$\mathrm{d}_{ki3} = -\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\left(\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)'\right)\right]. \quad (28)$$

The four terms in Eq. (25) can be re-expressed as follows:

$$\mathrm{d}\ln\pi_k = (\mathrm{d}\boldsymbol{\pi})'\mathbf{a}_k, \quad (29)$$

$$\mathrm{d}_{ki1} = -\frac{1}{2}\mathrm{tr}\left[(\mathrm{d}\boldsymbol{\Sigma}_k)\,\boldsymbol{\Sigma}_k^{-1}\right], \quad (30)$$

$$\mathrm{d}_{ki2} = \frac{1}{2}\mathrm{tr}\left[(\mathrm{d}\boldsymbol{\Sigma}_k)\,\mathbf{b}_{ki}\mathbf{b}_{ki}'\right], \quad (31)$$

$$\mathrm{d}_{ki3} = (\mathrm{d}\boldsymbol{\lambda}_k)'\mathbf{b}_{ki} + (\mathrm{d}\boldsymbol{\beta})'\mathbf{X}_i\mathbf{b}_{ki}, \quad (32)$$

where Eqs. (30)–(32) are obtained by exploiting some results from matrix derivatives (Magnus and Neudecker 1988, pp. 182–183; Schott 2005, pp. 292,293,361). Since the sum of d_{ki1} and d_{ki2} results in

$$\mathrm{d}_{ki1} + \mathrm{d}_{ki2} = -\frac{1}{2}\mathrm{d}\left(\mathrm{v}\boldsymbol{\Sigma}_k\right)'\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right), \quad (33)$$

(see Schott 2005, pp. 293,313,356,374) inserting Eqs. (29), (32) and (33) in Eq. (25) leads to

$$\begin{aligned}\mathrm{d}\ln f_{ki} &= (\mathrm{d}\boldsymbol{\pi})'\mathbf{a}_k + (\mathrm{d}\boldsymbol{\beta})'\mathbf{X}_i\mathbf{b}_{ki} + (\mathrm{d}\boldsymbol{\lambda}_k)'\mathbf{b}_{ki} \\ &\quad -\frac{1}{2}\mathrm{d}\left(\mathrm{v}\boldsymbol{\Sigma}_k\right)'\mathbf{G}'\mathrm{vec}\left(\mathbf{B}_{ki}\right) \\ &= (\mathrm{d}\boldsymbol{\pi})'\mathbf{a}_k + (\mathrm{d}\boldsymbol{\beta})'\mathbf{X}_i\mathbf{b}_{ki} + (\mathrm{d}\boldsymbol{\theta}_k)'\mathbf{c}_{ki}. \end{aligned} \quad (34)$$

Using Eqs. (24) and (34), $\mathrm{d}l(\boldsymbol{\theta})$ can be expressed as

$$\begin{aligned}\mathrm{d}l(\boldsymbol{\theta}) &= (\mathrm{d}\boldsymbol{\pi})'\sum_{i=1}^{I}\sum_{k=1}^{K}\alpha_{ki}\mathbf{a}_k + (\mathrm{d}\boldsymbol{\beta})'\sum_{i=1}^{I}\mathbf{X}_i\sum_{k=1}^{K}\alpha_{ki}\mathbf{b}_{ki} \\ &\quad + \sum_{k=1}^{K}(\mathrm{d}\boldsymbol{\theta}_k)'\sum_{i=1}^{I}\alpha_{ki}\mathbf{c}_{ki}, \end{aligned} \quad (35)$$

thus proving the theorem.

Appendix 2: Proof of Theorem 3

The proof is based on the computation of the second order differential of $l(\boldsymbol{\theta})$:

$$\mathrm{d}^2 l(\boldsymbol{\theta}) = \sum_{i=1}^{I}\mathrm{d}^2\ln\left(\sum_{k=1}^{K} f_{ki}\right), \quad (36)$$

where

$$\mathrm{d}^2 \ln\left(\sum_{k=1}^{K} f_{ki}\right) = \sum_{k=1}^{K} \alpha_{ki} \mathrm{d}^2 \ln f_{ki} + \sum_{k=1}^{K} \alpha_{ki} \left(\mathrm{d} \ln f_{ki}\right)^2 - \left(\sum_{k=1}^{K} \alpha_{ki} \mathrm{d} \ln f_{ki}\right)^2 \tag{37}$$

(see Boldea and Magnus 2009, Appendix).

Since $(\mathrm{d} \ln f_{ki})^2 = (\mathrm{d} \ln f_{ki})\,(\mathrm{d} \ln f_{ki})'$, using Eq. (34) it results that

$$\begin{aligned}(\mathrm{d} \ln f_{ki})^2 = {} & (\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} + (\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \mathbf{b}_{ki}' \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ & + (\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \mathbf{c}_{ki}' \mathrm{d}\boldsymbol{\theta}_k \\ & + (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki} \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} + (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki} \mathbf{b}_{ki}' \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ & + (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \mathbf{b}_{ki} \mathbf{c}_{ki}' \mathrm{d}\boldsymbol{\theta}_k \\ & + (\mathrm{d}\boldsymbol{\theta}_k)' \mathbf{c}_{ki} \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} + (\mathrm{d}\boldsymbol{\theta}_k)' \mathbf{c}_{ki} \mathbf{b}_{ki}' \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ & + (\mathrm{d}\boldsymbol{\theta}_k)' \mathbf{c}_{ki} \mathbf{c}_{ki}' \mathrm{d}\boldsymbol{\theta}_k . \end{aligned} \tag{38}$$

Similarly,

$$\begin{aligned}\left(\sum_{k=1}^{K} \alpha_{ki} \mathrm{d} \ln f_{ki}\right)^2 &= \left(\sum_{k=1}^{K} \alpha_{ki} \mathrm{d} \ln f_{ki}\right)\left(\sum_{k=1}^{K} \alpha_{ki} \mathrm{d} \ln f_{ki}\right)' \\ &= (\mathrm{d}\boldsymbol{\pi})' \bar{\mathbf{a}}_i \bar{\mathbf{a}}_i' \mathrm{d}\boldsymbol{\pi} + (\mathrm{d}\boldsymbol{\pi})' \bar{\mathbf{a}}_i \bar{\mathbf{b}}_i' \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ &\quad + (\mathrm{d}\boldsymbol{\pi})' \bar{\mathbf{a}}_i \sum_{k=1}^{K} \alpha_{ki} \mathbf{c}_{ki}' \mathrm{d}\boldsymbol{\theta}_k \\ &\quad + (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \bar{\mathbf{b}}_i \bar{\mathbf{a}}_i' \mathrm{d}\boldsymbol{\pi} + (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \bar{\mathbf{b}}_i \bar{\mathbf{b}}_i' \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ &\quad + (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \bar{\mathbf{b}}_i \sum_{k=1}^{K} \alpha_{ki} \mathbf{c}_{ki}' \mathrm{d}\boldsymbol{\theta}_k \\ &\quad + \left[\sum_{k=1}^{K} (\mathrm{d}\boldsymbol{\theta}_k)' \alpha_{ki} \mathbf{c}_{ki}\right] \bar{\mathbf{a}}_i' \mathrm{d}\boldsymbol{\pi} \\ &\quad + \left[\sum_{k=1}^{K} (\mathrm{d}\boldsymbol{\theta}_k)' \alpha_{ki} \mathbf{c}_{ki}\right] \bar{\mathbf{b}}_i' \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ &\quad + \sum_{k=1}^{K} \sum_{h=1}^{K} (\mathrm{d}\boldsymbol{\theta}_k)' \alpha_{ki} \alpha_{hi} \mathbf{c}_{ki} \mathbf{c}_{hi}' \mathrm{d}\boldsymbol{\theta}_l . \end{aligned} \tag{39}$$

Furthermore,

$$\begin{aligned}\mathrm{d}^2 \ln f_{ki} = {} & -(\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} - (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \boldsymbol{\Sigma}_k^{-1} \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ & - (\mathrm{d}\boldsymbol{\theta}_k)' \mathbf{F}_{ki}' \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ & - (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \mathbf{F}_{ki} \mathrm{d}\boldsymbol{\theta}_k - (\mathrm{d}\boldsymbol{\theta}_k)' \mathbf{C}_{ki} \mathrm{d}\boldsymbol{\theta}_k \end{aligned} \tag{40}$$

(see Appendix 3). From Eqs. (37), (38), (39) and (42) and by grouping together the common factors it follows that

$$\begin{aligned}\mathrm{d}^2 \ln\left(\sum_{k=1}^{K} f_{ki}\right) = {} & -(\mathrm{d}\boldsymbol{\pi})' \bar{\mathbf{a}}_i \bar{\mathbf{a}}_i' \mathrm{d}\boldsymbol{\pi} \\ & + (\mathrm{d}\boldsymbol{\pi})' \left[\left(\sum_{k=1}^{K} \alpha_{ki} \mathbf{a}_k \mathbf{b}_{ki}'\right) - \bar{\mathbf{a}}_i \bar{\mathbf{b}}_i'\right] \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ & + (\mathrm{d}\boldsymbol{\pi})' \left[\sum_{k=1}^{K} \alpha_{ki} \left(\mathbf{a}_k - \bar{\mathbf{a}}_i\right) \mathbf{c}_{ki}' \mathrm{d}\boldsymbol{\theta}_k\right] \\ & + (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \left[\left(\sum_{k=1}^{K} \alpha_{ki} \mathbf{b}_{ki} \mathbf{a}_k'\right) - \bar{\mathbf{b}}_i \bar{\mathbf{a}}_i'\right] \mathrm{d}\boldsymbol{\pi} \\ & - (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \left[\bar{\mathbf{B}}_i + \bar{\mathbf{b}}_i \bar{\mathbf{b}}_i'\right] \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ & - (\mathrm{d}\boldsymbol{\beta})' \mathbf{X}_i \left\{\sum_{k=1}^{K} \alpha_{ki} \left[\mathbf{F}_{ki} - \left(\mathbf{b}_{ki} - \bar{\mathbf{b}}_i\right) \mathbf{c}_{ki}'\right] \mathrm{d}\boldsymbol{\theta}_k\right\} \\ & + \left[\sum_{k=1}^{K} (\mathrm{d}\boldsymbol{\theta}_k)' \alpha_{ki} \mathbf{c}_{ki} \left(\mathbf{a}_k' - \bar{\mathbf{a}}_i'\right)\right] \mathrm{d}\boldsymbol{\pi} \\ & - \left\{\sum_{k=1}^{K} (\mathrm{d}\boldsymbol{\theta}_k)' \alpha_{ki} \left[\mathbf{F}_{ki}' - \mathbf{c}_{ki} \left(\mathbf{b}_{ki}' - \bar{\mathbf{b}}_i'\right)\right]\right\} \mathbf{X}_i' \mathrm{d}\boldsymbol{\beta} \\ & - \sum_{k=1}^{K} (\mathrm{d}\boldsymbol{\theta}_k)' \alpha_{ki} \left[\mathbf{C}_{ki} - \mathbf{c}_{ki} \mathbf{c}_{ki}'\right] \mathrm{d}\boldsymbol{\theta}_k \\ & - \sum_{k=1}^{K} \sum_{h=1}^{K} \left[(\mathrm{d}\boldsymbol{\theta}_k)' \alpha_{ki} \alpha_{hi} \mathbf{c}_{ki} \mathbf{c}_{hi}' \mathrm{d}\boldsymbol{\theta}_h\right] . \end{aligned} \tag{41}$$

Inserting Eq. (41) in Eq. (36) completes the proof.

Appendix 3: Second order differential of ln f_{ki}

Using Eq. (25) the second order differential of $\ln f_{ki}$ can be expressed as

$$\mathrm{d}^2 \ln f_{ki} = \mathrm{d}^2 \ln \pi_k + \mathrm{d}\left(\mathrm{d}_{ki1}\right) + \mathrm{d}\left(\mathrm{d}_{ki2}\right) + \mathrm{d}\left(\mathrm{d}_{ki3}\right) . \tag{42}$$

From Eq. (29) it follows that

$$\mathrm{d}^2 \ln \pi_k = -(\mathrm{d}\boldsymbol{\pi})' \mathbf{a}_k \mathbf{a}_k' \mathrm{d}\boldsymbol{\pi} . \tag{43}$$

The second term in Eq. (42) is equal to

$$\begin{aligned}\mathrm{d}\left(\mathrm{d}_{ki1}\right) &= -\frac{1}{2} \mathrm{tr}\left[\mathrm{d}\boldsymbol{\Sigma}_k \left(\mathrm{d}\boldsymbol{\Sigma}_k^{-1}\right)\right] \\ &= \frac{1}{2} \mathrm{tr}\left[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1}\right] . \end{aligned} \tag{44}$$

The third term that composes $\mathrm{d}^2 \ln f_{ki}$ results to be

$$\begin{aligned}\mathrm{d}\left(\mathrm{d}_{ki2}\right) = {} & \frac{1}{2} \mathrm{tr}\left[\mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right) \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta}\right) \right. \\ & \left. \times \left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta}\right)'\right] \\ & + \frac{1}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}_k^{-1} \left(\mathrm{d}\boldsymbol{\Sigma}_k\right) \mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right) \left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i' \boldsymbol{\beta}\right)\right. \end{aligned}$$

$$\times \left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)'\Big]$$
$$+\frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Sigma}_k^{-1}\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\left(\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)\right.$$
$$\left.\times \left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)'\right)\Big].$$

By exploiting some properties of the trace of a square matrix (see, e.g., Schott 2005), $\mathrm{d}\left(\mathrm{d}_{ki2}\right)$ can also be expressed as

$$\begin{aligned}\mathrm{d}\left(\mathrm{d}_{ki2}\right) &= \mathrm{tr}\Big[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right)\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)\\ &\quad\times\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}_k^{-1}\Big]\\ &\quad+\frac{1}{2}\mathrm{tr}\Big[\boldsymbol{\Sigma}_k^{-1}\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\left(\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)\right.\\ &\quad\left.\times\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)'\right)\Big],\end{aligned}$$

and using two theorems about the vec and trace operators (Schott 2005, Theorems 8.9 and 8.12) it follows that

$$\begin{aligned}\mathrm{d}\left(\mathrm{d}_{ki2}\right) &= \mathrm{tr}\Big[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right)\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)\\ &\quad\times\left(\mathbf{y}_i - \boldsymbol{\lambda}_k - \mathbf{X}_i'\boldsymbol{\beta}\right)'\boldsymbol{\Sigma}_k^{-1}\Big]\\ &\quad-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_k\right)\\ &\quad-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_k\right).\end{aligned}\tag{45}$$

From Eqs. (44) and (45) it follows that

$$\begin{aligned}\mathrm{d}\left(\mathrm{d}_{ki1}\right)+\mathrm{d}\left(\mathrm{d}_{ki2}\right) &= \frac{1}{2}\mathrm{tr}\Big[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\Big]\\ &\quad-\mathrm{tr}\Big[\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\mathbf{b}_{ki}\mathbf{b}_{ki}'\Big]\\ &\quad-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_k\right)\\ &\quad-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_k\right)\\ &= \frac{1}{2}\mathrm{tr}\Big\{\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\boldsymbol{\Sigma}_k^{-1}\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\Big[\boldsymbol{\Sigma}_k^{-1}\\ &\quad+\boldsymbol{\Sigma}_k^{-1}-\boldsymbol{\Sigma}_k^{-1}-2\mathbf{b}_{ki}\mathbf{b}_{ki}'\Big]\Big\}\\ &\quad-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_k\right)\\ &\quad-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_k\right)\\ &= -\frac{1}{2}\mathrm{vec}\left(\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)'\right)'\\ &\quad\times\Big[\left(\boldsymbol{\Sigma}_k^{-1}-2\mathbf{B}_{ki}\right)'\otimes\boldsymbol{\Sigma}_k^{-1}\Big]\mathrm{vec}\left(\mathrm{d}\boldsymbol{\Sigma}_k\right)\\ &\quad-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_k\right)\\ &\quad-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\left(\mathrm{vec}\,\boldsymbol{\Sigma}_k\right)\\ &= -\frac{1}{2}\mathrm{d}\left(\mathrm{v}\boldsymbol{\Sigma}_k\right)'\mathbf{G}'\Big[\left(\boldsymbol{\Sigma}_k^{-1}-2\mathbf{B}_{ki}\right)\otimes\boldsymbol{\Sigma}_k^{-1}\Big]\\ &\quad\times\mathbf{G}\mathrm{d}\left(\mathrm{v}\boldsymbol{\Sigma}_k\right)\\ &\quad-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathbf{G}\mathrm{d}\left(\mathrm{v}\boldsymbol{\Sigma}_k\right)\\ &\quad-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\left(\mathbf{b}_{ki}'\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathbf{G}\mathrm{d}\left(\mathrm{v}\boldsymbol{\Sigma}_k\right),\end{aligned}\tag{46}$$

where the third and fourth equalities are obtained using some properties of the vec operator (see, Schott 2005, p. 294).

From Eq. (32) it is possible to write

$$\begin{aligned}\mathrm{d}\left(\mathrm{d}_{ki3}\right) &= \left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\mathrm{d}\mathbf{b}_{ki}+\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\mathrm{d}\mathbf{b}_{ki}\\ &= -\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\left(\boldsymbol{\Sigma}_k\right)\mathbf{b}_{ki}\\ &\quad-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\boldsymbol{\lambda}_k-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_i'\mathrm{d}\boldsymbol{\beta}\\ &\quad-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\left(\boldsymbol{\Sigma}_k\right)\mathbf{b}_{ki}-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\boldsymbol{\lambda}_k\\ &\quad-\left(\mathrm{d}\boldsymbol{\beta}\right)^{\top}\mathbf{X}_i\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_i'\mathrm{d}\boldsymbol{\beta}\\ &= -\mathrm{d}\left(\mathrm{v}\boldsymbol{\Sigma}_k\right)'\mathbf{G}'\left(\mathbf{b}_{ki}\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathrm{d}\boldsymbol{\lambda}_k-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\boldsymbol{\lambda}_k\\ &\quad-\left(\mathrm{d}\boldsymbol{\lambda}_k\right)'\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_i'\mathrm{d}\boldsymbol{\beta}-\mathrm{d}\left(\mathrm{v}\boldsymbol{\Sigma}_k\right)'\mathbf{G}'\left(\mathbf{b}_{ki}\otimes\boldsymbol{\Sigma}_k^{-1}\right)\mathbf{X}_i'\mathrm{d}\boldsymbol{\beta}\\ &\quad-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\boldsymbol{\lambda}_k-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_i'\mathrm{d}\boldsymbol{\beta},\end{aligned}\tag{47}$$

where the third equality results from the same theorems about the vec and trace operators employed above and the second equality is obtained using the following expression for $\mathrm{d}\mathbf{b}_{ki}$:

$$\begin{aligned}\mathrm{d}\mathbf{b}_{ki} &= \mathrm{d}\left(\boldsymbol{\Sigma}_k^{-1}\right)\left(\mathbf{y}_i-\boldsymbol{\lambda}_k-\mathbf{X}_i'\boldsymbol{\beta}\right)+\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\left(\mathbf{y}_i-\boldsymbol{\lambda}_k-\mathbf{X}_i'\boldsymbol{\beta}\right)\\ &= -\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\left(\boldsymbol{\Sigma}_k\right)\mathbf{b}_{ki}-\boldsymbol{\Sigma}_k^{-1}\mathrm{d}\boldsymbol{\lambda}_k-\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_i'\mathrm{d}\boldsymbol{\beta}.\end{aligned}$$

Inserting Eqs. (43), (46) and (47) in Eq. (42) and using the definitions of $\boldsymbol{\theta}_k$, $\mathbf{F}_{ki}$ and $\mathbf{C}_{ki}$ introduced in Sect. 2.3 results in the following expression for $\mathrm{d}^2\ln f_{ki}$:

$$\begin{aligned}\mathrm{d}^2\ln f_{ki} &= -\left(\mathrm{d}\boldsymbol{\pi}\right)'\mathbf{a}_k\mathbf{a}_k'\mathrm{d}\boldsymbol{\pi}-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\boldsymbol{\Sigma}_k^{-1}\mathbf{X}_i'\mathrm{d}\boldsymbol{\beta}\\ &\quad-\left(\mathrm{d}\boldsymbol{\theta}_k\right)'\mathbf{F}_{ki}'\mathbf{X}_i'\mathrm{d}\boldsymbol{\beta}-\left(\mathrm{d}\boldsymbol{\beta}\right)'\mathbf{X}_i\mathbf{F}_{ki}\mathrm{d}\boldsymbol{\theta}_k\\ &\quad-\left(\mathrm{d}\boldsymbol{\theta}_k\right)'\mathbf{C}_{ki}\mathrm{d}\boldsymbol{\theta}_k.\end{aligned}$$

References

Ando, T., Zellner, A.: Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques. Bayesian Anal. **5**, 65–96 (2010)

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

Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49**, 803–821 (1993)

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

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

Biernacki, C., Celeux, G., Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate

Gaussian mixture models. Comput. Stat. Data Anal. **41**, 561–575 (2003)
Boldea, O., Magnus, J.R.: Maximum likelihood estimation of the multivariate normal mixture model. J. Am. Stat. Assoc. **104**, 1539–1549 (2009)
Chatterjee, S., Laudato, M., Lynch, L.A.: Genetic algorithms and their statistical applications: an introduction. Comput. Stat. Data Anal. **22**, 633–651 (1996)
Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**, 781–793 (1995)
Chevalier, J.A., Kashyap, A.K., Rossi, P.E.: Why don't prices rise during periods of peak demand? Evidence from scanner data. Am. Econ. Rev. **93**, 15–37 (2003)
Cook, R.D., Weisberg, S.: An Introduction to Regression Graphics. Wiley, New York (1994)
Cutler, A., Windham, M.P.: Information-based validity functionals for mixture analysis. In: Bozdogan, H. (ed.) Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, pp. 149–170. Kluwer Academic, Dordrecht (1994)
Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood for incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39**, 1–22 (1977)
Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc. **97**, 611–631 (2002)
Fraley, C., Raftery, A.E., Murphy, T.B., Scrucca, L.: `mclust` version 4 for `R`: normal mixture modeling for model-based clustering, classification, and density estimation. Technical report no. 597, Department of Statistics, University of Washington (2012)
Fraser, D.A.S., Rekkas, M., Wong, A.: Highly accurate likelihood analysis for the seemingly unrelated regression problem. J. Econom. **127**, 17–33 (2005)
Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models. Springer, New York (2006)
Galimberti, G., Soffritti, G.: A multivariate linear regression analysis using finite mixtures of *t* distributions. Comput. Stat. Data Anal. **71**, 138–150 (2014)
Henningsen, A., Hamann, J.D.: `systemfit`: a package for estimating systems of simultaneous equations in `R`. J. Stat. Softw. **23**(4), 1–40 (2007)
Keribin, C.: Consistent estimation of the order of mixture models. Sankhyā **62**, 49–66 (2000)
Kmenta, J., Gilbert, R.: Small sample properties of alternative estimators of seemingly unrelated regressions. J. Am. Stat. Assoc. **63**, 1180–1200 (1968)
Kowalski, J., Mendoza-Blanco, J.R., Tu, X.M., Gleser, L.J.: On the difference in inference and prediction between the joint and independent *t*-error models for seemingly unrelated regressions. Commun. Stat. Theory **28**, 2119–2140 (1999)
Kurata, H.: On the efficiencies of several generalized least squares estimators in a seemingly unrelated regression model and a heteroscedastic model. J. Multivar. Anal. **70**, 86–94 (1999)
Lange, K.L., Little, R.J.A., Taylor, J.M.G.: Robust statistical modeling using the *t* distribution. J. Am. Stat. Assoc. **84**, 881–896 (1989)
Magnus, J.R.: Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. J. Econom. **7**, 281–312 (1978)
Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, Chichester (1988)
Maugis, C., Celeux, G., Martin-Magniette, M.-L.: Variable selection in model-based clustering: a general variable role modeling. Comput. Stat. Data Anal. **53**, 3872–3882 (2009a)
Maugis, C., Celeux, G., Martin-Magniette, M.-L.: Variable selection for clustering with Gaussian mixture models. Biometrics **65**, 707–709 (2009b)
McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, Chichester (2008)
McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, Chichester (2000)
McLachlan, G.J., Peel, D., Bean, R.W.: Modelling high-dimensional data by mixtures of factor analyzers. Comput. Stat. Data Anal. **41**, 379–388 (2003)
McNicholas, P.D., Murphy, T.B.: Parsimonious Gaussian mixture models. Stat. Comput. **18**, 285–296 (2008)
Melnykov, V., Melnykov, I.: Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. Comput. Stat. Data Anal. **56**, 1381–1395 (2012)
Ng, V.M.: Robust Bayesian inference for seemingly unrelated regressions with elliptical errors. J. Multivar. Anal. **83**, 409–414 (2002)
Oberhofer, W., Kmenta, J.: A general procedure for obtaining maximum likelihood estimates in generalized regression models. Econometrica **42**, 579–590 (1974)
Park, T.: Equivalence of maximum likelihood estimation and iterative two-stage estimation for seemingly unrelated regression models. Commun. Stat. Theory **22**, 2285–2296 (1993)
Percy, D.F.: Predictions for seemingly unrelated regression. J. R. Stat. Soc. Ser. B **54**, 243–252 (1992)
Ray, S., Lindsay, B.G.: Model selection in high dimensions: a quadratic-risk-based approach. J. R. Stat. Soc. Ser. B **70**, 95–118 (2008)
R Core Team: `R`: A language and environment for statistical computing. `R` Foundation for Statistical Computing, Vienna. URL http://www.R-project.org/ (2014)
Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. **26**, 195–239 (1984)
Rilstone, P., Veall, M.: Using bootstrapped confidence intervals for improved inferences with seemingly unrelated regression equations. Econom. Theory **12**, 569–580 (1996)
Rocke, D.: Bootstrap Bartlett adjustment in seemingly unrelated regression. J. Am. Stat. Assoc. **84**, 598–601 (1989)
Rossi, P.E.: `bayesm`: Bayesian inference for marketing/microeconometrics. `R` package version 2.2-5. URL http://CRAN.R-project.org/package=bayesm (2012)
Schott, J.R.: Matrix Analysis for Statistics, 2nd edn. Wiley, New York (2005)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
Soffritti, G., Galimberti, G.: Multivariate linear regression with non-normal errors: a solution based on mixture models. Stat. Comput. **21**, 523–536 (2011)
Srivastava, V.K., Giles, D.E.A.: Seemingly Unrelated Regression Equations Models. Marcel Dekker, New York (1987)
Srivastava, V.K., Maekawa, K.: Efficiency properties of feasible generalized least squares estimators in SURE models under non-normal disturbances. J. Econom. **66**, 99–121 (1995)
Zellner, A.: An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. J. Am. Stat. Assoc. **57**, 348–368 (1962)
Zellner, A.: Estimators for seemingly unrelated regression equations: some exact finite sample results. J. Am. Stat. Assoc. **58**, 977–992 (1963)
Zellner, A.: An Introduction to Bayesian Inference in Econometrics. Wiley, New York (1971)
Zellner, A., Ando, T.: A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model. J. Econom. **159**, 33–45 (2010a)
Zellner, A., Ando, T.: Bayesian and non-Bayesian analysis of the seemingly unrelated regression model with Student *t* errors, and its application for forecasting. Int. J. Forecast. **26**, 413–434 (2010b)