



ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE  
DELLA RICERCA

## Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

Modelling the role of variables in model-based cluster analysis

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Galimberti, G., Manisi, A., Soffritti, G. (2018). Modelling the role of variables in model-based cluster analysis. *STATISTICS AND COMPUTING*, 28(1), 145-169 [10.1007/s11222-017-9723-0].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/585288> since: 2017-05-09

*Published:*

DOI: <http://doi.org/10.1007/s11222-017-9723-0>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

**Galimberti, G., Manisi, A. & Soffritti, G. Modelling the role of variables in model-based cluster analysis. Stat Comput 28, 145–169 (2018).**  
<https://doi.org/10.1007/s11222-017-9723-0>

The final published version is available online at: <https://doi.org/10.1007/s11222-017-9723-0>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)*

***When citing, please refer to the published version.***

# Modelling the role of variables in model-based cluster analysis

Giuliano Galimberti · Annamaria Manisi · Gabriele Soffritti

Received: date / Accepted: date

**Abstract** In the framework of cluster analysis based on Gaussian mixture models it is usually assumed that all the variables provide information about the clustering of the sample units. Several variable selection procedures are available in order to detect the structure of interest for the clustering when this structure is contained in a variable sub-vector. Currently, in these procedures a variable is assumed to play one of (up to) three roles: *i*) informative, *ii*) uninformative and correlated with some informative variables, *iii*) uninformative and uncorrelated with any informative variable. A more general approach for modelling the role of a variable is proposed by taking into account the possibility that the variable vector provides information about more than one structure of interest for the clustering. This approach is developed by assuming that such information is given by non-overlapped and possibly correlated sub-vectors of variables; it is also assumed that the model for the variable vector is equal to a product of conditionally independent Gaussian mixture models (one for each variable sub-vector). Details about model identifiability, parameter estimation and model selection are provided. The usefulness and effectiveness of

the described methodology are illustrated using simulated and real datasets.

**Keywords** Clusterwise linear regression · EM algorithm · Gaussian mixture model · Genetic algorithm · Multiple cluster structure · Variable selection

## 1 Introduction

When a dataset is characterised by the presence of an unknown cluster structure (a grouping of a given set of observations into clusters), methods for cluster analysis are the correct unsupervised tool to extract such information from the data (Hastie *et al.* 2009). However there may be datasets in which such a structure is confined in a sub-space of the variable space. Since the effect of the presence of uninformative variables is a masking of the cluster structure (see, e.g., Gordon 1999, p. 23), the use of methods able to select the informative variables in a cluster analysis, such as those proposed by Fowlkes *et al.* (1988), Gnanadesikan *et al.* (1995), Brusco and Cradit (2001), Montanari and Lizzani (2001), Fraiman *et al.* (2008), Steinley and Brusco (2008a) and Witten and Tibshirani (2010), is crucial for a proper recovery of the unknown cluster structure from the observed data.

Another problem to be tackled when performing a cluster analysis is that datasets may be characterised by the presence of several independent unknown cluster structures, that is independent groupings of the same set of observations defined in different subspaces of the variable space. Since a classical assumption in cluster analysis is that one single cluster structure is contained in the data, most clustering methods can miss relevant information about the ways observations

---

G. Galimberti  
Department of Statistical Sciences, via delle Belle Arti 41,  
40126 Bologna, Italy  
E-mail: giuliano.galimberti@unibo.it

A. Manisi  
E-mail: annamaria.manisi@gmail.com

G. Soffritti  
Department of Statistical Sciences, via delle Belle Arti 41,  
40126 Bologna, Italy  
Tel.: +39-051-2098193  
Fax: +39-051-232153  
E-mail: gabriele.soffritti@unibo.it

are clustered. Methods able to avoid this drawback are due to Soffritti (2003), Friedman and Meulman (2004), Belitskaya-Levy (2006), Poon *et al.* (2013), Dang and Bailey (2015) and Liu *et al.* (2015), for example.

If the variables are continuous, model-based clustering methods that rely on Gaussian mixture models can be employed. In these models the joint probability density function (p.d.f.) of the variables is assumed to be a mixture of a finite number of Gaussian densities (one component density for each cluster) (see, e.g., McLachlan and Peel 2000; Melnykov and Maitra 2010). The number of components is generally chosen through model selection criteria, such as the Bayesian information criterion (*BIC*) (Schwarz 1978). In the framework of likelihood-based methods, parameters are usually estimated through the maximum likelihood (ML) method by resorting to the expectation-maximisation (EM) algorithm (Dempster *et al.* 1977). The clustering of the given set of observations is performed using a rule that assigns an observation to the component of the mixture to which it has the highest posterior probability of belonging. Gaussian mixture models with a parsimonious covariance structure can help controlling the model complexity when the number of variables is high. They can be obtained by reparameterising the component-covariance matrices according to their spectral decomposition and by imposing constraints on the resulting eigenvalues and/or eigenvectors over the mixture components (Banfield and Raftery 1993; Celeux and Govaert 1995). Another approach relies on the use of factor-analytic dimensionality reduction methods (see, e.g., McLachlan *et al.* 2003; McNicholas and Murphy 2008; Viroli 2010).

Methods that simultaneously select the informative variables and find the cluster structure, based on Gaussian mixture models, have been proposed by Dy and Brodley (2004), Law *et al.* (2004), Tadesse *et al.* (2005), Raftery and Dean (2006), Pan and Shen (2007), Xie *et al.* (2008), Wang and Zhu (2008), Galimberti *et al.* (2009), Maugis *et al.* (2009a), Maugis *et al.* (2009b), Zeng and Cheung (2009), Zhou *et al.* (2009), Guo *et al.* (2010) and Andrews and McNicholas (2014). Comparisons among some of these methods, based on analyses of simulated and real datasets, can be found in Steinley and Brusco (2008b), Witten and Tibshirani (2010), Celeux *et al.* (2011), Celeux *et al.* (2014) and Andrews and McNicholas (2014). In particular, the variable selection methods proposed by Raftery and Dean (2006) are based on a model in which the vector of the examined variables is assumed to be partitioned into two sub-vectors: one is composed of the informative variables and the other contains the uninformative ones. A greedy search algorithm is developed for partitioning

the variables into these two sub-vectors. In the steps of this algorithm, for each variable two different models are considered: *a*) a model in which the examined variable is jointly modelled with the informative variables through a Gaussian mixture with  $K$  components ( $K \geq 2$ ); *b*) a Gaussian linear regression model for the conditional distribution of the examined variable given the informative ones. These two models correspond to two different roles (informative and uninformative, respectively) for the examined variable. The selection of the informative variables is then recast as a sequence of model comparison problems, based on the *BIC*. Maugis *et al.* (2009a) exploit a similar approach by introducing an explicit formulation of the joint p.d.f. of the observed variables. In particular, they assume that this joint p.d.f. is equal to the product of a Gaussian mixture model with  $K$  components ( $K \geq 2$ ) for the distribution of the informative variables and a Gaussian linear regression model for the conditional distribution of the uninformative variables given the informative ones. This regression model allows the uninformative variables to be explained by only a subset of the informative ones. A more versatile model for the joint p.d.f. is proposed in Maugis *et al.* (2009b) by allowing some uninformative variables to be independent of all the informative ones. Local optima in the model spaces resulting from these two latter types of models are determined using a backward-stepwise algorithm. In order to overcome the well known drawbacks of stepwise techniques Scrucca (2016) suggests an approach based on the use of genetic algorithms. According to these latter methods, a variable can play one of (up to) three roles in cluster analysis. Overall, the methods just described make use of Gaussian linear regression models for performing variable selection in model-based cluster analysis; they embed a supervised learning process into an unsupervised one.

As far as the detection of multiple cluster structures is concerned, methods that make use of Gaussian mixtures are described in Galimberti and Soffritti (2007). In particular, these methods rely on a model in which the vector of the observed variables is assumed to be partitioned into independent sub-vectors, each of which provides information about an unknown cluster structure. Thus, the joint p.d.f. of the variable vector is given by a product of Gaussian mixture models, one for each variable sub-vector. According to this approach, each variable is relevant for detecting one specific structure of interest for the clustering. However, due to the assumption of independence among variable sub-vectors, such an approach is quite restrictive.

This paper illustrates an approach for discovering independent unknown cluster structures from possibly

correlated variable sub-vectors. Such a correlation is managed by replacing the assumption of independence among the variable sub-vectors with conditional independence assumptions and by resorting to multivariate linear regression models (Soffritti and Galimberti 2011) or seemingly unrelated linear regression models (Galimberti *et al.* 2015) in which the distribution of the error terms is modelled using a Gaussian mixture model. The model for the joint p.d.f. of the variable vector is specified so as to take into account also the possibility that some variables are uninformative. Namely, such uninformative variables can be correlated with some informative variables or uncorrelated with any informative variable. As a result, a general approach for modelling the role of variables in cluster analysis is obtained.

The paper is organised as follows. The basic idea of the proposed approach is introduced in Section 2. The theory is described in Section 3. Namely, the models used to develop the approach are presented in Sections 3.1, 3.4 and 3.6. A comparison with existing models is given in Section 3.2. A discussion about identifiability is provided in Section 3.3. Specifically, a theorem ensuring identifiability of models having the same partition of the variable vector is given in Section 3.3.1; the proof is reported in Appendix. Necessary conditions for identifiability of models having different partitions of the variable vector are illustrated in Section 3.3.2. Details on ML estimation and model selection are given in Sections 3.5 and 3.7, respectively. Sections 4 and 5 contain experimental results obtained from the analysis of real and simulated datasets, respectively. Concluding remarks are reported in Section 6. Additional experimental results are provided in the Supplementary Material together with a detailed description of three algorithms developed for exploring classes of models characterised by the presence of two unknown independent cluster structures.

## 2 An introductory example

The basic idea of the approach developed in this paper can be introduced through a simple example referring to the crabs dataset described in Campbell and Mahon (1974) and available in the package `MASS` (Venables and Ripley 2002) for the R software environment (R Core Team 2015). This dataset reports five morphological measurements (in mm) for 200 crabs of the species *Lepidograpsus variegatus*: frontal lobe size (FL), rear width (RW), carapace length (CL), carapace width (CW) and body depth (BD). Namely, the sample is composed of 50 crabs each of two colours (blue and orange) and both sexes. Detailed information about how the data were collected are reported in Campbell and Mahon

Table 1: Pearson correlation matrix in the crabs dataset.

	FL	RW	CL	CW	BD
FL	1.000	0.907	0.979	0.965	0.988
RW	0.907	1.000	0.893	0.900	0.889
CL	0.979	0.893	1.000	0.995	0.983
CW	0.965	0.900	0.995	1.000	0.968
BD	0.988	0.889	0.983	0.968	1.000

(1974) together with the results of a supervised analysis based on linear discriminant functions. Variables in the dataset are highly pairwise linear dependent (see Table 1).

Although the most natural way of analysing this dataset is through classification methods in the context of a supervised learning, several papers can be found in the statistical literature in which the crabs dataset is analysed according to the rules of an unsupervised process (i.e.: without accounting for the information about crabs' colour and sex). Some examples can be found in Raftery and Dean (2006), McNicholas *et al.* (2010), Viroli (2010), Andrews and McNicholas (2014), Scrucca and Raftery (2015), Malsiner-Walli *et al.* (2016) and Zhu and Melnykov (2016). The purpose of this type of analysis is to test novel clustering techniques and compare their performance with the one of other existing methods. In particular, these papers focus on the ability to recover the four classes that arise from the joint classification of the crabs based on their colour and sex. In line with these applications, the introductory example presented in this Section is set into an unsupervised learning process. Namely, in the first part of this example the interest is in detecting whether the crabs dataset is characterised by the presence of a cluster structure in the joint distribution of the five morphological measurements. Using the R package `mclust` (Fraley and Raftery 2002; Fraley *et al.* 2012) Gaussian mixture models with a number  $K$  of components from one to nine are fitted to the dataset. For each value of  $K$ , models with unconstrained and constrained covariance structures are estimated. Namely, the parsimonious models fitted to the data are obtained using the approach based on the eigen-decomposition of the component-covariance matrices (Banfield and Raftery 1993; Celeux and Govaert 1995). Note that all analyses are carried out with an `mclust` option `SVD` for the initialisation of the EM algorithm. This option transforms the variables through a singular value decomposition (Scrucca and Raftery 2015). In the analysis of the crabs dataset this option has been found to be the one that allows to select the model with the largest *BIC* (Scrucca and Raftery 2015). According to the *BIC* values, the best model is a parsimonious mixture of four components. Let this

Table 2: Classification of the crabs according to their colour and sex (BF = blue female, BM = blue male, OF = orange female, OM = orange male) and the cluster memberships estimated by the model  $M_1$  selected using `mclust`.

Cluster	Colour and sex			
	BF	BM	OF	OM
1	49	11	0	0
2	0	0	5	50
3	0	39	0	0
4	1	0	45	0
aRi	0.794			

model be denoted as  $M_1$ . Details about the specific constraints on the component-covariance matrices in model  $M_1$  are provided in Section 4.1. In order to measure the association between the cluster structure detected by model  $M_1$  and the classification of the crabs, the joint classification of the crabs based on colour and sex and the segmentation obtained from the model is examined, and the adjusted Rand index (aRi) (Hubert and Arabie 1985) is computed. From the obtained results (see Table 2) it emerges that such an association is high, and the clustering obtained from the selected model reproduces quite well the four classes of crabs defined from their colour and sex.

As described in Section 1, the presence of a cluster structure may be masked by uninformative variables. Thus, a further analysis is carried out through the R package `clustvarsel` (Scrucca and Raftery 2014) that implements the variable selection methods proposed by Raftery and Dean (2006). This second analysis is performed using the same initialisation of the EM algorithm mentioned above and with a number  $K$  of components from one to nine. The results suggest that only four morphological measurements are relevant for clustering the crabs. They are frontal lobe size, rear width, carapace width and body depth. The carapace length can be discarded. The best Gaussian mixture model fitted to the p.d.f. of the four selected measurements is a parsimonious mixture of four components with the same constraints on the covariance matrices already used in model  $M_1$  (details are reported in Section 4.1). Using this parsimonious model allows to obtain a partition of the crabs with an increased agreement with the partition based on colour and sex (see Table 3). The resulting model for the p.d.f. of the five measurements is denoted as  $M_2$  and is given by the product of the just described Gaussian mixture model for the joint distribution of (FL, RW, CW, BD) and a linear regression model for the conditional distribution of CL given (FL, RW, CW, BD).

Table 3: Classification of the crabs according to their colour and sex and the cluster membership estimated by the model  $M_2$  selected using `clustvarsel`.

Cluster	Colour and sex			
	BF	BM	OF	OM
1	50	10	0	0
2	0	0	3	50
3	0	40	0	0
4	0	0	47	0
aRi	0.840			

Given the drawbacks of the variable selection methods proposed by Raftery and Dean (2006), further analyses are carried out by resorting to two C++ softwares (`SelvarClust` and `SelvarClustIndep`) that incorporate algorithms for fitting and selecting the models described in Maugis *et al.* (2009a,b). They are available at <http://perso.math.univ-toulouse.fr/maugis>. Both softwares lead to the same result obtained by `clustvarsel`. According to these approaches, four morphological measurements (frontal lobe size, rear width, carapace width and body depth) play the role of informative variables, while the carapace length can be considered uninformative and correlated with all the informative variables.

Although these methods are able to recover the four classes of crabs, the obtained results do not provide an explicit information about a distinctive feature of the dataset, that is the independence between colour and sex. In order to capture this feature it is necessary to use unsupervised methods able to explicitly account for the presence of two independent sources of clustering without imposing any restriction on the sample correlation structure of the measurements. In order to achieve this goal, the idea developed in this paper is that the information about a first source of clustering is given by the marginal distribution of some morphological measurements, while the conditional distribution of some other measurements provides information about the second source of clustering. Namely, using methods illustrated in Section 3, the vector of the five measurements is split into two sub-vectors. For the joint marginal p.d.f. of (RW, CL) a parsimonious mixture of two Gaussian components is selected, while the joint conditional p.d.f. of (FL, CW, BD) given (RW, CL) is modelled using a parsimonious mixture of two Gaussian linear regression models with the same regression coefficients. Details about the parsimonious covariance structure of these two mixture models are provided in Section 4.1. The joint model for the five morphological measurements, given by the product of these two mixture models and denoted as  $M_3$ , allows to obtain two independent clusterings of the crabs. The first clustering, based on the analysis of rear width and carapace length,

Table 4: Comparison between the cluster structure detected in the joint marginal distribution of RW and CL (cluster structure I) and the classifications of crabs based on their colour and/or sex.

Cluster I	Colour and sex				Colour		Sex	
	BF	BM	OF	OM	B	O	F	M
1	50	7	50	3	57	53	100	10
2	0	43	0	47	47	43	0	90
aRi	0.400				-0.003		0.810	

Table 5: Comparison between the cluster structure detected in the conditional distribution of (FL, CW, BD) given (RW, CL) (cluster structure II) and the classifications of crabs based on their colour and/or sex.

Cluster II	Colour and sex				Colour		Sex	
	BF	BM	OF	OM	B	O	F	M
1	50	50	1	0	100	1	51	50
2	0	0	49	50	0	99	49	50
aRi	0.486				0.980		-0.005	

reproduces quite well the classification of crabs based on their sex (see Table 4). On the contrary, the clustering obtained by modelling the dependence of frontal lobe size, carapace width and body depth on rear width and carapace length using the above mentioned mixture of two Gaussian linear regression models is almost perfectly associated with the classification based on colour (see Table 5). Thus, according to model  $M_3$ , all measurements play a relevant role for clustering the crabs: RW and CL provide information about the classification based on colour, while FL, CW and BD (conditionally on the previous two measurements) are informative on the classification based on sex. This partition of the variable vector is consistent with the results reported in Campbell and Mahon (1974) and based on a linear discriminant analysis. Namely, a first canonical variable, which differentiates between the two species, is obtained as a contrast between the carapace width relative to the width of the front lip and the depth of the body. Furthermore, a second canonical variable resulting from a contrast between the rear width and the carapace length allows to identify males and females. However, it is worth noting that the methods proposed in this paper allow to discover this partition of the variables without exploiting the information about crabs' colour and sex. To the authors' knowledge no other unsupervised method is able to obtain this same result in reference to this benchmark dataset.

Model  $M_3$  provides information also about a third clustering of the crabs based on the joint p.d.f. of the five measurements. This clustering can be obtained by jointly examining the two cluster structures described above. Table 6 compares this third clustering with the

Table 6: Classification of the crabs according to their colour and sex and the segmentation based on the joint examination of the two independent cluster structures detected by model  $M_3$ .

Cluster I	Cluster II	Colour and sex			
		BF	BM	OF	OM
1	1	50	7	1	0
1	2	0	0	49	3
2	1	0	43	0	0
2	2	0	0	0	47
aRi	0.859				

classification based on colour and sex. According to the aRi, using model  $M_3$  leads to an improvement over models  $M_1$  and  $M_2$  in recovering this latter classification. Furthermore, a limitation of these latter models is that neither  $M_1$  nor  $M_2$  provide any explicit information about the existence of two independent sources of clustering (sex and colour). This limitation is overcome if model  $M_3$  is employed. Further comments and results obtained from the analysis of this dataset using methods illustrated in Section 3 are reported in Section 4.1.

### 3 Modelling the role of variables through Gaussian mixtures

#### 3.1 Models with two independent clusterings

Let  $\mathbf{X} = (X_1, \dots, X_L)$  be the random vector composed of  $L$  observed continuous variables to be used for clustering  $n$  sample units. Assume that the examined dataset is characterised by two unknown independent cluster structures  $S_1$  and  $S_2$ , that is two independent clusterings of the  $n$  sample units. Such structures can be modelled by assuming that two independent nominal latent variables  $Z_1$  and  $Z_2$  (with  $K_1$  and  $K_2$  categories, respectively) affect the probability distribution of  $\mathbf{X}$ . More specifically, assume that  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}, \mathbf{X}^U)$  denotes a splitting of  $\mathbf{X}$  into three non-overlapped sub-vectors, where  $\mathbf{X}^U$  can be empty. Namely,  $\mathbf{X}^{S_g}$  is the sub-vector containing  $L_g$  variables that provide information about the cluster structure  $S_g$  ( $g = 1, 2$ ), while the sub-vector  $\mathbf{X}^U$  is composed of  $L_U$  uninformative variables, with  $L_1 + L_2 + L_U = L$ .

The proposed model is based on the following set of assumptions.

- (A1) The marginal distribution of  $\mathbf{X}^{S_1}$  is affected by the latent variable  $Z_1$  and is given by a Gaussian mixture model with  $K_1$  components. Namely:

$$f(\mathbf{x}^{S_1}; \boldsymbol{\theta}_1) = \sum_{k_1=1}^{K_1} \pi_{k_1}^{(1)} \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)}), \quad (1)$$

where  $K_1 \geq 2$ ,  $\phi_{L_1}(\cdot; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)})$  is the p.d.f. of the  $L_1$ -dimensional normal distribution with mean vector  $\boldsymbol{\mu}_{k_1}^{(1)}$  and positive definite covariance matrix  $\boldsymbol{\Sigma}_{k_1}^{(1)}$ ,  $\boldsymbol{\theta}_1 = (\boldsymbol{\pi}_1, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ ,  $\boldsymbol{\pi}_1 = (\pi_1^{(1)}, \dots, \pi_{K_1}^{(1)})$ ,  $\boldsymbol{\mu}_1 = (\boldsymbol{\mu}_1^{(1)}, \dots, \boldsymbol{\mu}_{K_1}^{(1)})$  and  $\boldsymbol{\sigma}_1 = (\boldsymbol{\Sigma}_1^{(1)}, \dots, \boldsymbol{\Sigma}_{K_1}^{(1)})$ .

(A2) The sub-vector  $\mathbf{X}^{S_2}$  is assumed to be conditionally independent of  $Z_1$  given  $\mathbf{X}^{S_1}$ . Furthermore, the conditional p.d.f. of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$  is affected by the latent variable  $Z_2$  and is equal to

$$f(\mathbf{x}^{S_2} | \mathbf{x}^{S_1}; \boldsymbol{\theta}_2) = \sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} \phi_{L_2}(\mathbf{x}^{S_2}; \boldsymbol{\mu}_{k_2}^{(2)}, \boldsymbol{\Sigma}_{k_2}^{(2)}), \quad (2)$$

where  $K_2 \geq 2$  and

$$\boldsymbol{\mu}_{k_2}^{(2)} = \boldsymbol{\beta}_0 + \boldsymbol{\lambda}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \quad k_2 = 1, \dots, K_2, \quad (3)$$

with  $\boldsymbol{\beta}_0$  and  $\boldsymbol{\lambda}_{k_2}^{(2)}$  denoting  $L_2$ -dimensional vectors and  $\mathbf{B}_{21}$  representing a  $L_2 \times L_1$  matrix of regression coefficients. The condition defined by equation (3) is equivalent to assuming that the dependence of  $\mathbf{X}^{S_2}$  on  $\mathbf{X}^{S_1}$  is given by a multivariate linear regression model whose error terms follow a mixture of  $K_2$  Gaussian components. Specifically,

$$\mathbf{X}^{S_2} = \boldsymbol{\beta}_0 + \mathbf{B}_{21} \mathbf{x}^{S_1} + \boldsymbol{\epsilon}_2, \quad (4)$$

$$\boldsymbol{\epsilon}_2 \sim \sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} N_{L_2}(\boldsymbol{\lambda}_{k_2}^{(2)}, \boldsymbol{\Sigma}_{k_2}^{(2)}), \quad (5)$$

where  $N_{L_2}(\boldsymbol{\lambda}_{k_2}^{(2)}, \boldsymbol{\Sigma}_{k_2}^{(2)})$  denotes the  $L_2$ -dimensional normal distribution with mean vector  $\boldsymbol{\lambda}_{k_2}^{(2)}$  and positive definite covariance matrix  $\boldsymbol{\Sigma}_{k_2}^{(2)}$ . Thus, the conditional distribution of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$  is described by a mixture of  $K_2$  Gaussian linear regression models (see, e.g., Quandt and Ramsey 1978; De Sarbo and Cron 1988) with the constraint that the effect of  $\mathbf{X}^{S_1}$  on the expected value of  $\mathbf{X}^{S_2}$  is the same for all the  $K_2$  components of the mixture (2). In order to guarantee identifiability of the model defined by equations (4) and (5), it is necessary to require some constraints on  $\boldsymbol{\beta}_0$  or the  $\boldsymbol{\lambda}_k^{(2)}$ 's. Namely,  $\boldsymbol{\beta}_0 = \mathbf{0}$  or  $\sum_{k_2} \pi_{k_2}^{(2)} \boldsymbol{\lambda}_{k_2}^{(2)} = \mathbf{0}$ . This problem does not arise if model (3) is directly parameterised as follows:

$$\boldsymbol{\mu}_{k_2}^{(2)} = \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \quad k_2 = 1, \dots, K_2, \quad (6)$$

where  $\boldsymbol{\gamma}_{k_2}^{(2)} = \boldsymbol{\beta}_0 + \boldsymbol{\lambda}_{k_2}^{(2)}$ . Then,  $\boldsymbol{\theta}_2 = (\boldsymbol{\pi}_2, \boldsymbol{\gamma}_2, \mathbf{B}_{21}, \boldsymbol{\sigma}_2)$ ,  $\boldsymbol{\pi}_2 = (\pi_1^{(2)}, \dots, \pi_{K_2}^{(2)})$ ,  $\boldsymbol{\gamma}_2 = (\boldsymbol{\gamma}_1^{(2)}, \dots, \boldsymbol{\gamma}_{K_2}^{(2)})$ ,  $\boldsymbol{\sigma}_2 = (\boldsymbol{\Sigma}_1^{(2)}, \dots, \boldsymbol{\Sigma}_{K_2}^{(2)})$  (for further details see Soffritti and Galimberti 2011).

(A3) The sub-vector  $\mathbf{X}^U$  is assumed to be conditionally independent of  $(Z_1, Z_2)$  given  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2})$ . Furthermore, the conditional distribution of  $\mathbf{X}^U$  given  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2})$  follows a Gaussian linear regression model. Specifically:

$$\begin{aligned} \mathbf{X}^U &= \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{x}^{S_1} + \mathbf{A}_2 \mathbf{x}^{S_2} + \boldsymbol{\epsilon}_U, \\ \boldsymbol{\epsilon}_U &\sim N_{L_U}(\mathbf{0}, \boldsymbol{\Sigma}_U), \end{aligned} \quad (7)$$

whose parameters are  $\boldsymbol{\theta}_U = (\boldsymbol{\alpha}_0, \mathbf{A}_1, \mathbf{A}_2, \boldsymbol{\Sigma}_U)$ , where  $\boldsymbol{\alpha}_0$  is a  $L_U$ -dimensional vector,  $\mathbf{A}_1$  and  $\mathbf{A}_2$  represent matrices of regression coefficients with dimensions  $L_U \times L_1$  and  $L_U \times L_2$ , respectively, and the covariance matrix  $\boldsymbol{\Sigma}_U$  is positive definite.

Thus, the joint p.d.f. of  $\mathbf{X}$  can be obtained as follows:

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\theta}) &= \sum_{k_1=1}^{K_1} \pi_{k_1}^{(1)} \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)}) \\ &\quad \times \sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} \phi_{L_2}(\mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2}^{(2)}) \\ &\quad \times \phi_{L_U}(\mathbf{x}^U; \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{x}^{S_1} + \mathbf{A}_2 \mathbf{x}^{S_2}, \boldsymbol{\Sigma}_U), \end{aligned} \quad (8)$$

where  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_U)$ .

If assumptions (A1) and (A2) hold, all the information about the first cluster structure is provided by the marginal distribution of the variable sub-vector  $\mathbf{X}^{S_1}$ . In particular, assuming that  $\mathbf{X}^{S_2}$  is conditionally independent of  $Z_1$  given  $\mathbf{X}^{S_1}$  ensures that, conditionally on  $\mathbf{X}^{S_1}$ ,  $\mathbf{X}^{S_2}$  does not provide any information about the first cluster structure. Assumption (A2) includes a specific dependence structure of  $\mathbf{X}^{S_2}$  on  $\mathbf{X}^{S_1}$  that allows to take into account the possible correlation between these two variable sub-vectors. Thus, the first cluster structure can be discovered by examining the marginal distribution of  $\mathbf{X}^{S_1}$  even if  $\mathbf{X}^{S_1}$  and  $\mathbf{X}^{S_2}$  are correlated. Assumption (A2) is analogous to the assumption introduced in Raftery and Dean (2006) and exploited in Maugis *et al.* (2009a,b) to perform variable selection in the presence of correlated variables; the main difference is that the conditional p.d.f. of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$  does not follow a Gaussian linear regression model but a mixture of  $K_2$  Gaussian linear regression models. Thus, assumption (A2) is adequate whenever the second cluster structure is hidden in the conditional distribution of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$ . Note that, by setting  $\mathbf{B}_{21} = \mathbf{0}$  in equation (6), this second structure is, in fact, defined in the marginal distribution of  $\mathbf{X}^{S_2}$ , thus leading to the model proposed by Galimberti and Soffritti (2007).

Both  $\mathbf{X}^{S_1}$  and  $\mathbf{X}^{S_2}$  represent vectors of informative variables. If, in addition,  $\mathbf{X}^U \neq \emptyset$ , model (8) also allows to perform variable selection. A graphical representation of model (8) is given in Figure 1.

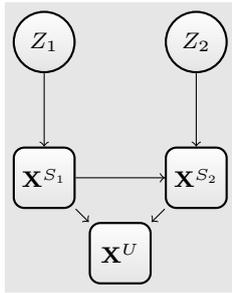


Fig. 1: Graphical representation of model (8). Circles and squares are used to denote latent variables and observed variables, respectively.

Equation (8) can be rewritten as follows:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \pi_{k_1}^{(1)} \pi_{k_2}^{(2)} \left[ \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)}) \right. \\ \times \phi_{L_2}(\mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2}^{(2)}) \\ \left. \times \phi_{L_U}(\mathbf{x}^U; \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{x}^{S_1} + \mathbf{A}_2 \mathbf{x}^{S_2}, \boldsymbol{\Sigma}_U) \right]. \quad (9)$$

Consider the Cartesian product between  $\{1, \dots, K_1\}$  and  $\{1, \dots, K_2\}$ , that are the sets of labels for the categories of  $Z_1$  and  $Z_2$ , respectively. Let the  $r$ -th element of such a set be denoted as  $c_r = (k_{1,r}, k_{2,r})$ ,  $r = 1, \dots, R$ , where  $R = K_1 \cdot K_2$ . Using this notation and exploiting properties of the Gaussian distribution (see, e.g., Anderson 2003), equation (9) can be rewritten as follows:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{r=1}^R \pi_r \phi_L(\mathbf{x}; \boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r), \quad (10)$$

where, for  $r = 1, \dots, R$ ,

$$\pi_r = \pi_{k_{1,r}}^{(1)} \pi_{k_{2,r}}^{(2)}, \quad (11)$$

$$\boldsymbol{\mu}_r = \mathbf{Q} \begin{pmatrix} \boldsymbol{\mu}_{k_{1,r}}^{(1)} \\ \boldsymbol{\gamma}_{k_{2,r}}^{(2)} \\ \boldsymbol{\alpha}_0 \end{pmatrix}, \quad (12)$$

$$\boldsymbol{\Sigma}_r = \mathbf{Q} \begin{bmatrix} \boldsymbol{\Sigma}_{k_{1,r}}^{(1)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{k_{2,r}}^{(2)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_U \end{bmatrix} \mathbf{Q}', \quad (13)$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{I}_{L_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_{21} & \mathbf{I}_{L_2} & \mathbf{0} \\ \mathbf{A}_1 + \mathbf{B}_{21} & \mathbf{A}_2 & \mathbf{I}_{L_U} \end{pmatrix},$$

with  $\mathbf{I}_L$  denoting the identity matrix of order  $L$ . Thus, the joint p.d.f. of  $\mathbf{X}$  in equation (8) is a Gaussian mixture of  $R$  components whose parameters are given in equations (11)-(13). As a consequence of above, assumptions (A1)-(A3) can also be interpreted as conditions that impose specific constraints on the parameters of a Gaussian mixture model for  $\mathbf{X}$ . These constraints imply a reduction in the number of free parameters. For example,

as far as the mixing proportions  $\pi_r$  in equation (11) are concerned, only  $(K_1 - 1)(K_2 - 1)$  free parameters have to be estimated instead of  $K_1 K_2 - 1$ . Since the matrix  $\mathbf{Q}$  is nonsingular, assumptions (A1)-(A3) also imply that all component-covariance matrices in the mixture model defined in equation (10) are positive definite. Furthermore, equation (10) highlights that the joint p.d.f. of  $\mathbf{X}$  is affected by a latent variable  $\tilde{Z}$  whose categories are obtained by considering the  $K_1 \cdot K_2$  combinations of categories of  $Z_1$  and  $Z_2$ . This latter result is due to the independence assumption between  $Z_1$  and  $Z_2$ .

Model (8) prevents the possible presence of uninformative variables that are independent of any other variable. Models that do not impose this restriction can be obtained as follows. Let  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}, \mathbf{X}^U, \mathbf{X}^I)$  be a splitting of  $\mathbf{X}$  into four non-overlapped sub-vectors, where the additional sub-vector  $\mathbf{X}^I$  can be empty. Specifically,  $\mathbf{X}^I$  contains  $L_I$  uninformative variables that are also independent of all the remaining variables. By assuming that  $\mathbf{X}^I$  has a marginal multivariate Gaussian distribution with mean vector  $\boldsymbol{\mu}_I$  and positive definite covariance matrix  $\boldsymbol{\Sigma}_I$ , the resulting p.d.f. of  $\mathbf{X}$  is defined as follows:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k_1=1}^{K_1} \pi_{k_1}^{(1)} \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)}) \\ \times \sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} \phi_{L_2}(\mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2}^{(2)}) \\ \times \phi_{L_U}(\mathbf{x}^U; \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{x}^{S_1} + \mathbf{A}_2 \mathbf{x}^{S_2}, \boldsymbol{\Sigma}_U) \\ \times \phi_{L_I}(\mathbf{x}^I; \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I), \quad (14)$$

where  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I)$ , with  $\boldsymbol{\theta}_I = (\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ . The graphical representation of the model obtained from equation (14) is given in Figure 2(a).

Whenever  $\mathbf{X}^U \neq \emptyset$ , both model (8) and model (14) force the uninformative variables in  $\mathbf{X}^U$  to linearly depend on all the informative variables (see equation (7)). This further restriction can be removed as follows. Let  $\mathbf{X}_U^{S_1}$  and  $\mathbf{X}_U^{S_2}$  denote the sub-vectors of  $\mathbf{X}^{S_1}$  and  $\mathbf{X}^{S_2}$ , respectively, useful to predict the uninformative variables  $\mathbf{X}^U$  according to a Gaussian linear regression model ( $\mathbf{X}_U^{S_1} \subseteq \mathbf{X}^{S_1}$ ,  $\mathbf{X}_U^{S_2} \subseteq \mathbf{X}^{S_2}$ ). Thus, equation (7) is modified as follows:

$$\mathbf{X}^U = \boldsymbol{\alpha}_0 + \mathbf{A}_{1U} \mathbf{x}_U^{S_1} + \mathbf{A}_{2U} \mathbf{x}_U^{S_2} + \boldsymbol{\epsilon}_U, \quad (15)$$

where  $\mathbf{A}_{1U}$  and  $\mathbf{A}_{2U}$  are matrices of regression coefficients both with  $L_U$  rows and a number of columns equal to the lengths of sub-vectors  $\mathbf{X}_U^{S_1}$  and  $\mathbf{X}_U^{S_2}$ , respectively. Combining equations (15) and (14) leads to

the following p.d.f. of  $\mathbf{X}$ :

$$\begin{aligned}
f(\mathbf{x}; \boldsymbol{\theta}) &= \sum_{k_1=1}^{K_1} \pi_{k_1}^{(1)} \phi_{L_1} \left( \mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)} \right) \\
&\times \sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} \phi_{L_2} \left( \mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2}^{(2)} \right) \\
&\times \phi_{L_U} \left( \mathbf{x}^U; \boldsymbol{\alpha}_0 + \mathbf{A}_{1U} \mathbf{x}_U^{S_1} + \mathbf{A}_{2U} \mathbf{x}_U^{S_2}, \boldsymbol{\Sigma}_U \right) \\
&\times \phi_{L_I} \left( \mathbf{x}^I; \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I \right), \tag{16}
\end{aligned}$$

where  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I)$ , with  $\boldsymbol{\theta}_U = (\boldsymbol{\alpha}_0, \mathbf{A}_{1U}, \mathbf{A}_{2U}, \boldsymbol{\Sigma}_U)$ . The graphical representation of the model obtained from equation (16) is given in Figure 2(b).

More general models can be defined. For example, in equation (4) the variables in the sub-vector  $\mathbf{X}^{S_2}$  are forced to linearly depend on all the variables belonging to  $\mathbf{X}^{S_1}$ . This restriction can be removed as follows. Let  $\mathbf{X}_{S_2}^{S_1}$  denote the sub-vector of  $\mathbf{X}^{S_1}$  composed of the predictors of  $\mathbf{X}^{S_2}$  ( $\mathbf{X}_{S_2}^{S_1} \subseteq \mathbf{X}^{S_1}$ ). Then,  $\boldsymbol{\mu}_{k_2}^{(2)}$  in equation (6) becomes

$$\boldsymbol{\mu}_{k_2}^{(2)} = \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21}^* \mathbf{x}_{S_2}^{S_1}, \quad k_2 = 1, \dots, K_2, \tag{17}$$

where the number of columns of matrix  $\mathbf{B}_{21}^*$  coincides with the length of  $\mathbf{X}_{S_2}^{S_1}$ . If, in addition,  $L_2 \geq 2$ , it is possible to define a model in which a different sub-vector of  $\mathbf{X}^{S_1}$  can be employed for predicting each variable in  $\mathbf{X}^{S_2}$ . Such a model can be specified as follows. Let  $\mathbf{X}^{S_2}[l]$  be the  $l$ -th variable of  $\mathbf{X}^{S_2}$  and  $\mathbf{X}_l^{S_1}$  be the sub-vector of  $\mathbf{X}^{S_1}$  containing the predictors of  $\mathbf{X}^{S_2}[l]$ . Then,  $\boldsymbol{\mu}_{k_2}^{(2)}$  can be obtained as follows:

$$\boldsymbol{\mu}_{k_2}^{(2)}[l] = \boldsymbol{\gamma}_{k_2}^{(2)}[l] + (\mathbf{x}_l^{S_1})' \boldsymbol{\beta}_l, \quad l = 1, \dots, L_2, \tag{18}$$

where the notation  $\mathbf{a}[l]$  is used to denote the  $l$ -th element of vector  $\mathbf{a}$  and  $\boldsymbol{\beta}_l$  is the vector of the regression coefficients of  $\mathbf{X}_l^{S_1}$  on  $\mathbf{X}^{S_2}[l]$ . The joint model for the vector  $\mathbf{X}^{S_2}$  resulting from equation (18) represents a seemingly unrelated linear regression model with a Gaussian mixture for the errors (for further details see Galimberti *et al.* 2015).

### 3.2 Comparison with Gaussian mixture models for clustering and regression

Some approaches for unsupervised analysis introduced in Sections 1 and 2 can be obtained from models (8) and (16) by admitting that the variable sub-vector  $\mathbf{X}^{S_2}$  can be empty.

- If  $\mathbf{X}^{S_2} = \mathbf{X}^U = \emptyset$ , model (8) reduces to a Gaussian mixture model for the joint distribution of  $\mathbf{X}$ , thus leading to the approach to cluster analysis based on Gaussian mixture models.

- If  $\mathbf{X}^{S_2} = \emptyset$  and  $\mathbf{X}^U \neq \emptyset$ , equation (8) defines a model with only one cluster structure hidden in the variable sub-space associated with the sub-vector  $\mathbf{X}^{S_1}$ . Specifically, vectors  $\mathbf{X}^{S_1}$  and  $\mathbf{X}^U$  contain informative and uninformative variables, respectively. This model represents the tool for model-based cluster analysis with variable selection according to the approach developed by Raftery and Dean (2006).
- When  $\mathbf{X}^{S_2} = \emptyset$ ,  $\mathbf{X}^U \neq \emptyset$  and  $\mathbf{X}^I = \emptyset$ , equation (16) gives the model developed by Maugis *et al.* (2009a).
- When  $\mathbf{X}^{S_2} = \emptyset$ ,  $\mathbf{X}^U \neq \emptyset$  and  $\mathbf{X}^I \neq \emptyset$ , model (16) reduces to the model described in Maugis *et al.* (2009b).

Furthermore, if  $K_1$  and  $K_2$  in equations (1)-(6) are allowed to be equal to 1 and  $\mathbf{X}^{S_2} \neq \emptyset$ ,  $\mathbf{X}^U = \emptyset$ , model (8) can result in two types of linear regression models.

- When  $K_1 = 1$  and  $K_2 > 1$ , model (8) reduces to a mixture of Gaussian linear regression models with the same regression coefficients (Soffritti and Galimberti 2011) and Gaussian random predictors. Such a model assumes that in the dataset there is only one latent clustering of the sample units and that this clustering is hidden in the conditional distribution of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$ .
- When  $K_1 = K_2 = 1$ , model (8) reduces to a Gaussian linear regression model with Gaussian random predictors. According to this model no cluster structure is defined either in the marginal distribution of  $\mathbf{X}^{S_1}$  or in the conditional distribution of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$ .

Thus, a general framework for model-based clustering, linear regression with Gaussian random predictors and multiple cluster structure detection that encompasses all the methods described so far can be obtained from the models illustrated in Section 3.1 by simply admitting that  $\mathbf{X}^{S_2}$  can be empty and allowing  $K_1$  and  $K_2$  to be equal to one.

### 3.3 Identifiability

Some results concerning identifiability are developed for two classes of models obtained from equation (14). The first class is composed of models that have the same splitting of  $\mathbf{X}$  but differ in the values of  $K_1$  and/or  $K_2$ . The second class is obtained by examining models that can also have different splittings of  $\mathbf{X}$ . A theoretical result on the identifiability in the first class under mild and simple conditions is provided in Section 3.3.1. Section 3.3.2 contains a description of six situations that generate non-identifiability in the second model class;

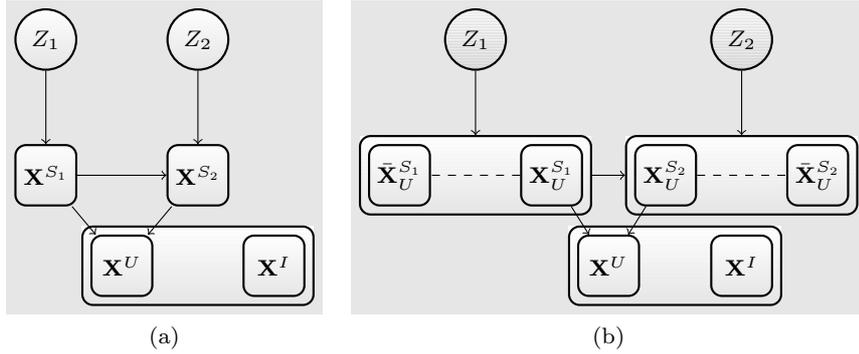


Fig. 2: Graphical representations of two models obtained from equation (16), where  $\bar{\mathbf{X}}_U^{S_1} = \mathbf{X}^{S_1} \setminus \mathbf{X}_U^{S_1}$ ,  $\bar{\mathbf{X}}_U^{S_2} = \mathbf{X}^{S_2} \setminus \mathbf{X}_U^{S_2}$ .

it also provides five necessary conditions for the identifiability of this class.

Due to the richness of the class of models that can be obtained from equation (14), simple and general sufficient conditions for identifiability can hardly be derived. Thus, it is important to remark that the identifiability is generally not ensured.

### 3.3.1 Models with the same splitting of $\mathbf{X}$

The model class examined in this Section is composed of models obtained from equation (14) by keeping  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}, \mathbf{X}^U, \mathbf{X}^I)$  equal to a given splitting of  $\mathbf{X}$  while varying the values of  $K_1$  and  $K_2$ . Namely, models with values of  $K_g \in \{2, \dots, K_{gmax}\}$ ,  $g = 1, 2$ , are admitted, where  $K_{1max}$  and  $K_{2max}$  denote the maximum number of components specified by the researcher for the mixtures defined in equations (1) and (2), respectively. Let the resulting model class be denoted as  $\mathcal{M}$ . The generic element of  $\mathcal{M}$  is  $M = (K_1, K_2)$ . For each  $M \in \mathcal{M}$  the parameterised p.d.f's are  $f(\cdot | \boldsymbol{\theta}_M)$ , with  $\boldsymbol{\theta}_M = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I)$ . The corresponding parameter space is  $\mathcal{Q}_M$ .

Under mild conditions, the identifiability of the model class  $\mathcal{M}$  is ensured by the following theorem.

**Theorem 1** *Let  $\Theta_M$  be a subset of  $\mathcal{Q}_M$  whose elements  $\boldsymbol{\theta}_M$  fulfil the following conditions:*

(I1) *Conditions for  $\boldsymbol{\theta}_1$ :*

$$\begin{aligned} \pi_{k_1}^{(1)} &> 0 \quad \forall k_1; \\ \left( \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)} \right) &\neq \left( \boldsymbol{\mu}_{k'_1}^{(1)}, \boldsymbol{\Sigma}_{k'_1}^{(1)} \right) \quad \forall k_1 \neq k'_1; \end{aligned}$$

(I2) *Conditions for  $\boldsymbol{\theta}_2$ :*

$$\begin{aligned} \pi_{k_2}^{(2)} &> 0 \quad \forall k_2; \\ \left( \boldsymbol{\gamma}_{k_2}^{(2)}, \boldsymbol{\Sigma}_{k_2}^{(2)} \right) &\neq \left( \boldsymbol{\gamma}_{k'_2}^{(2)}, \boldsymbol{\Sigma}_{k'_2}^{(2)} \right) \quad \forall k_2 \neq k'_2. \end{aligned}$$

Let  $M = (K_1, K_2)$  and  $M^* = (K_1^*, K_2^*)$  be two models. If  $\boldsymbol{\theta}_M \in \Theta_M$  and  $\boldsymbol{\theta}_{M^*}^* \in \Theta_{M^*}$  exist such that  $f(\mathbf{x}; \boldsymbol{\theta}_M) = f(\mathbf{x}; \boldsymbol{\theta}_{M^*}^*) \quad \forall \mathbf{x} \in \mathbb{R}^L$ , then  $M = M^*$  and  $\boldsymbol{\theta}_M = \boldsymbol{\theta}_{M^*}^*$  (up to permutations of the mixture components of the marginal p.d.f. of  $\mathbf{X}^{S_1}$  and of the mixture components of the conditional p.d.f. of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$ ).

The constraints on the parameter space described by the conditions (I1) and (I2) avoid non-identifiability due to both empty components and equal components in the mixtures defined in equations (1) and (2). The same types of constraint are used to avoid non-identifiability of a finite mixture distribution due to potential overfitting (see, e.g., Frühwirth-Schnatter 2006, p. 19). The non-identifiability due to invariance to relabeling the components is common to all finite mixture models (McLachlan and Peel 2000). A proof of Theorem 1 is provided in Appendix.

### 3.3.2 Models with different splittings of $\mathbf{X}$

In order to define the second class of models it is convenient to use the following notation. Let a generic model  $M = (K_1, K_2)$  introduced in Section 3.3.1 be denoted as  $M_{(S_1, S_2, U, I)}^{(2)}(K_1, K_2)$ , and let the corresponding model class be  $\mathcal{M}_{(S_1, S_2, U, I)}^{(2)}$ , where  $(S_1, S_2, U, I)$  is the splitting of the variable index set  $\mathcal{I} = \{1, \dots, L\}$  associated with the splitting  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}, \mathbf{X}^U, \mathbf{X}^I)$  of the variable vector  $\mathbf{X}$  employed to specify all models belonging to the class  $\mathcal{M}_{(S_1, S_2, U, I)}^{(2)}$ .

Models characterised by different splittings  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}, \mathbf{X}^U, \mathbf{X}^I)$  of  $\mathbf{X}$  and different values of  $K_1, K_2$  can be obtained by admitting that the splitting  $(S_1, S_2, U, I)$  used to define the model class  $\mathcal{M}_{(S_1, S_2, U, I)}^{(2)}$  can vary. Namely, let  $\mathcal{V}$  be the family of the splittings of the variable index set  $\mathcal{I}$  into four elements (where the third and fourth elements can be equal to the empty set):

$\mathcal{V} = \{(S_1, S_2, U, I); (S_1, S_2, U, I) \in \mathcal{F}^4, S_g \neq \emptyset \forall g, S_1 \cap S_2 = \emptyset, S_1 \cap U = \emptyset, S_2 \cap U = \emptyset, I = \mathcal{I} \setminus S_1 \setminus S_2 \setminus U\}$ , where  $\mathcal{F}$  denotes the family of subsets of  $\mathcal{I}$ . Then, the wider model class is defined as

$$\mathcal{M}^{(2)} = \bigcup_{(S_1, S_2, U, I) \in \mathcal{V}} \mathcal{M}_{(S_1, S_2, U, I)}^{(2)}.$$

Hereafter the generic element of this class is denoted as  $M^{(2)} = (S_1, S_2, U, I, K_1, K_2)$ . For each model  $M^{(2)} \in \mathcal{M}^{(2)}$  the parameterised densities are  $f(\cdot | \boldsymbol{\theta}_{M^{(2)}})$ , with  $\boldsymbol{\theta}_{M^{(2)}} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I)$ . The corresponding parameter space is denoted by  $\mathcal{Q}_{M^{(2)}}$ .

Let  $(S_1, S_2, U, I, K_1, K_2)$  and  $(S_1^*, S_2^*, U^*, I^*, K_1^*, K_2^*)$  be any two members of  $\mathcal{M}^{(2)}$ , and let  $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I)$  and  $(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_U^*, \boldsymbol{\theta}_I^*)$  be their parameters. The model class  $\mathcal{M}^{(2)}$  is identifiable for  $\boldsymbol{\theta}_{M^{(2)}} \in \mathcal{Q}_{M^{(2)}}$  if  $f(\mathbf{x} | \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I) = f(\mathbf{x} | \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_U^*, \boldsymbol{\theta}_I^*)$  if and only if  $S_1 = S_1^*$ ,  $S_2 = S_2^*$ ,  $U = U^*$ ,  $I = I^*$ ,  $K_1 = K_1^*$ ,  $K_2 = K_2^*$  and it is possible to permute the component labels so that  $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^*$ ,  $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^*$ ,  $\boldsymbol{\theta}_U = \boldsymbol{\theta}_U^*$  and  $\boldsymbol{\theta}_I = \boldsymbol{\theta}_I^*$ .

The following remarks show six different types of situation that can prevent the identifiability of  $\mathcal{M}^{(2)}$  for  $\boldsymbol{\theta}_{M^{(2)}} \in \mathcal{Q}_{M^{(2)}}$  from being ensured. All types are associated with a moving of a variable sub-vector from an element of  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}, \mathbf{X}^U, \mathbf{X}^I)$  to another without changing the joint p.d.f. of  $\mathbf{X}$ .

*Remark 1 Moving variables from  $\mathbf{X}^{S_2}$  to  $\mathbf{X}^U$  or to  $\mathbf{X}^I$*   
Consider a given model  $M^{(2)} = (S_1, S_2, U, I, K_1, K_2) \in \mathcal{M}^{(2)}$ . Let  $s$  be a nonempty subset strictly included into  $S_2$  (thus, composed of  $L_s$  elements, with  $1 \leq L_s < L_2$ ), and  $\bar{s}$  be its complement in  $S_2$ , with the remaining  $L_2 - L_s = L_{\bar{s}}$  elements. Then the variable sub-vector  $\mathbf{X}^{S_2}$  and the model parameters  $\boldsymbol{\gamma}_{k_2}^{(2)}$ ,  $\boldsymbol{\Sigma}_{k_2}^{(2)}$ ,  $\mathbf{B}_{21}$  and  $\mathbf{A}_2$  are partitioned as follows:

$$\begin{aligned} \mathbf{X}^{S_2} &= \begin{pmatrix} \mathbf{X}^s \\ \mathbf{X}^{\bar{s}} \end{pmatrix}, \boldsymbol{\gamma}_{k_2}^{(2)} = \begin{pmatrix} \boldsymbol{\gamma}_{k_2, s} \\ \boldsymbol{\gamma}_{k_2, \bar{s}} \end{pmatrix}, \\ \boldsymbol{\Sigma}_{k_2}^{(2)} &= \begin{bmatrix} \boldsymbol{\Sigma}_{k_2, ss} & \boldsymbol{\Sigma}_{k_2, s\bar{s}} \\ \boldsymbol{\Sigma}'_{k_2, s\bar{s}} & \boldsymbol{\Sigma}_{k_2, \bar{s}\bar{s}} \end{bmatrix}, \mathbf{B}_{21} = \begin{bmatrix} \mathbf{B}_{s1} \\ \mathbf{B}_{\bar{s}1} \end{bmatrix}, \\ \mathbf{A}_2 &= [\mathbf{A}_{2s} \ \mathbf{A}_{2\bar{s}}]. \end{aligned}$$

By exploiting results concerning the Gaussian distribution, the  $k_2$ -th Gaussian density of the mixture for  $\mathbf{X}^{S_2} | \mathbf{X}^{S_1}$  in equation (8) can be decomposed as

$$\begin{aligned} \phi_{L_2}(\mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2}^{(2)}) &= \\ \phi_{L_s}(\mathbf{x}^s; \boldsymbol{\gamma}_{k_2, s} + \mathbf{B}_{s1} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2, ss}) & \\ \times \phi_{L_{\bar{s}}}(\mathbf{x}^{\bar{s}}; \boldsymbol{\mu}_{k_2, \bar{s} | s} + \boldsymbol{\Sigma}_{k_2, \bar{s} | s} \mathbf{x}^s, \boldsymbol{\Sigma}_{k_2, \bar{s}\bar{s} | s}), & \quad (19) \end{aligned}$$

where

$$\boldsymbol{\mu}_{k_2, \bar{s} | s} = \boldsymbol{\gamma}_{k_2, \bar{s}} + \mathbf{B}_{\bar{s}1} \mathbf{x}^{S_1} - \boldsymbol{\Sigma}_{k_2, \bar{s} | s} (\boldsymbol{\gamma}_{k_2, s} + \mathbf{B}_{s1} \mathbf{x}^{S_1}), \quad (20)$$

$$\boldsymbol{\Sigma}_{k_2, \bar{s} | s} = \boldsymbol{\Sigma}'_{k_2, s\bar{s}} \boldsymbol{\Sigma}_{k_2, ss}^{-1}, \quad (21)$$

$$\boldsymbol{\Sigma}_{k_2, \bar{s}\bar{s} | s} = \boldsymbol{\Sigma}_{k_2, \bar{s}\bar{s}} - \boldsymbol{\Sigma}'_{k_2, s\bar{s}} \boldsymbol{\Sigma}_{k_2, ss}^{-1} \boldsymbol{\Sigma}_{k_2, s\bar{s}}. \quad (22)$$

If the parameters in equations (20)-(22) are identical for  $k_2 = 1, \dots, K_2$  then the regression density of  $\mathbf{X}^{\bar{s}}$  on  $\mathbf{X}^s$  can be factorised from the Gaussian mixture for  $\mathbf{X}^{S_2} | \mathbf{X}^{S_1}$  and regrouped with the regression density of  $\mathbf{X}^U$  on  $\mathbf{X}^{S_1} \cup \mathbf{X}^{S_2}$ . Thus, a model  $(S_1, s, U \cup \bar{s}, I, K_1, K_2) \in \mathcal{M}^{(2)}$  is obtained that is different from  $M^{(2)}$  but cannot be distinguished from  $M^{(2)}$  by the joint p.d.f. of  $\mathbf{X}$ .

Furthermore, if the parameters in equations (20)-(22) are identical for  $k_2 = 1, \dots, K_2$ ,  $\mathbf{A}_{2\bar{s}} = \mathbf{0}$ ,  $\mathbf{B}_{\bar{s}1} = \mathbf{0}$  and  $\boldsymbol{\Sigma}_{k_2, s\bar{s}} = \mathbf{0} \forall k_2$ , then  $\mathbf{X}^{\bar{s}}$  is independent of  $\mathbf{X}^{S_1}$ ,  $\mathbf{X}^s$  and  $\mathbf{X}^U$ . Thus, the marginal density of  $\mathbf{X}^{\bar{s}}$  can be factorised from the Gaussian mixture for the conditional p.d.f. of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$  and regrouped with the marginal density of  $\mathbf{X}^I$ . This second situation leads to another model  $(S_1, s, U, I \cup \bar{s}, K_1, K_2) \in \mathcal{M}^{(2)}$  that is different from  $M^{(2)}$  but is undistinguishable from  $M^{(2)}$  by the joint p.d.f. of  $\mathbf{X}$ .

*Remark 2 Moving variables from  $\mathbf{X}^{S_1}$  to  $\mathbf{X}^U$  or to  $\mathbf{X}^I$*   
Similarly to the situation illustrated in *Remark 1*, suppose that  $t$  is a nonempty subset strictly included into  $S_1$  (thus, composed of  $L_t$  elements, with  $1 \leq L_t < L_1$ ), and  $\bar{t}$  is its complement in  $S_1$ , with the remaining  $L_1 - L_t = L_{\bar{t}}$  elements. The variable sub-vector  $\mathbf{X}^{S_1}$  and the model parameters  $\boldsymbol{\mu}_{k_1}^{(1)}$ ,  $\boldsymbol{\Sigma}_{k_1}^{(2)}$ ,  $\mathbf{B}_{21}$  and  $\mathbf{A}_1$  are partitioned accordingly:

$$\begin{aligned} \mathbf{X}^{S_1} &= \begin{pmatrix} \mathbf{X}^t \\ \mathbf{X}^{\bar{t}} \end{pmatrix}, \boldsymbol{\mu}_{k_1}^{(1)} = \begin{pmatrix} \boldsymbol{\mu}_{k_1, t} \\ \boldsymbol{\mu}_{k_1, \bar{t}} \end{pmatrix}, \\ \boldsymbol{\Sigma}_{k_1}^{(1)} &= \begin{bmatrix} \boldsymbol{\Sigma}_{k_1, tt} & \boldsymbol{\Sigma}_{k_1, t\bar{t}} \\ \boldsymbol{\Sigma}'_{k_1, t\bar{t}} & \boldsymbol{\Sigma}_{k_1, \bar{t}\bar{t}} \end{bmatrix}, \mathbf{B}_{21} = [\mathbf{B}_{2t} \ \mathbf{B}_{2\bar{t}}], \\ \mathbf{A}_1 &= [\mathbf{A}_{1t} \ \mathbf{A}_{1\bar{t}}]. \end{aligned} \quad (23)$$

Similarly to equation (19), the following decomposition holds for the  $k_1$ -th Gaussian density of the mixture for  $\mathbf{X}^{S_1}$  in equation (8):

$$\begin{aligned} \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)}) &= \phi_{L_t}(\mathbf{x}^t; \boldsymbol{\mu}_{k_1, t}, \boldsymbol{\Sigma}_{k_1, tt}) \\ \times \phi_{L_{\bar{t}}}(\mathbf{x}^{\bar{t}}; \boldsymbol{\mu}_{k_1, \bar{t} | t} + \boldsymbol{\Sigma}_{k_1, \bar{t} | t} \mathbf{x}^t, \boldsymbol{\Sigma}_{k_1, \bar{t}\bar{t} | t}), & \quad (24) \end{aligned}$$

where

$$\boldsymbol{\mu}_{k_1, \bar{t} | t} = \boldsymbol{\mu}_{k_1, \bar{t}} - \boldsymbol{\Sigma}_{k_1, \bar{t} | t} \boldsymbol{\mu}_{k_1, t}, \quad (25)$$

$$\boldsymbol{\Sigma}_{k_1, \bar{t} | t} = \boldsymbol{\Sigma}'_{k_1, t\bar{t}} \boldsymbol{\Sigma}_{k_1, tt}^{-1}, \quad (26)$$

$$\boldsymbol{\Sigma}_{k_1, \bar{t}\bar{t} | t} = \boldsymbol{\Sigma}_{k_1, \bar{t}\bar{t}} - \boldsymbol{\Sigma}'_{k_1, t\bar{t}} \boldsymbol{\Sigma}_{k_1, tt}^{-1} \boldsymbol{\Sigma}_{k_1, t\bar{t}}. \quad (27)$$

If the parameters in equations (25)-(27) are identical for  $k_1 = 1, \dots, K_1$  and  $\mathbf{B}_{2\bar{t}} = \mathbf{0}$  then the regression density of  $\mathbf{X}^{\bar{t}}$  on  $\mathbf{X}^t$  can be factorised from the Gaussian mixture for  $\mathbf{X}^{S_1}$  and regrouped with the regression density of  $\mathbf{X}^U$  on  $\mathbf{X}^{S_1} \cup \mathbf{X}^{S_2}$ . Thus, a model  $(t, S_2, U \cup \bar{t}, K_1, K_2) \in \mathcal{M}^{(2)}$  is obtained that is different from  $M^{(2)}$  but is undistinguishable from  $M^{(2)}$  by the joint p.d.f. of  $\mathbf{X}$ .

Furthermore, if the parameters in equations (25)-(27) are identical for  $k_1 = 1, \dots, K_1$ ,  $\mathbf{B}_{2\bar{t}} = \mathbf{0}$ ,  $\mathbf{A}_{1\bar{t}} = \mathbf{0}$  and  $\boldsymbol{\Sigma}_{k_1, t\bar{t}} = \mathbf{0} \forall k_1$ , then  $\mathbf{X}^{\bar{t}}$  is independent of  $\mathbf{X}^t$ ,  $\mathbf{X}^{S_2}$  and  $\mathbf{X}^U$ . Thus, the marginal density of  $\mathbf{X}^{\bar{t}}$  can be factorised from the Gaussian mixture for the marginal p.d.f. of  $\mathbf{X}^{S_1}$  and regrouped with the marginal density of  $\mathbf{X}^I$ . In this way another model  $(t, S_2, U, I \cup \bar{t}, K_1, K_2) \in \mathcal{M}^{(2)}$  is obtained that is different from  $M^{(2)}$  but cannot be distinguished from  $M^{(2)}$  by the joint p.d.f. of  $\mathbf{X}$ .

*Remark 3 Moving variables from  $\mathbf{X}^{S_1}$  to  $\mathbf{X}^{S_2}$*

Consider now a model  $M^{(2)} = (S_1, S_2, U, I, K_1, K_2) \in \mathcal{M}^{(2)}$  such that  $K_1$  is a composite number. Suppose that the latent variable  $Z_1$  can be re-expressed in terms of a vector  $(\tilde{Z}_1, \tilde{Z}_2)$ , where  $\tilde{Z}_1$  and  $\tilde{Z}_2$  are two independent nominal latent variables with  $H$  and  $J$  categories, respectively, with  $H$  and  $J$  denoting positive integers greater than one such that  $H \cdot J = K_1$ . Namely,  $Z_1$  is composed of the  $H \cdot J$  combinations of categories of  $\tilde{Z}_1$  and  $\tilde{Z}_2$ , and each category of  $Z_1$  corresponds to an element of the Cartesian product between  $\{1, \dots, H\}$  and  $\{1, \dots, J\}$ , that are the sets of labels for the categories of  $\tilde{Z}_1$  and  $\tilde{Z}_2$ , respectively. Thus,  $\pi_{k_1}^{(1)} = \pi_{h_{k_1}} \pi_{j_{k_1}}$ , where  $(h_{k_1}, j_{k_1})$  is the element of the Cartesian product between  $\{1, \dots, H\}$  and  $\{1, \dots, J\}$  that corresponds to the  $k_1$ -th category of  $Z_1$ .

Furthermore, suppose that for the model parameters  $\boldsymbol{\mu}_{k_1}^{(1)}$ ,  $\boldsymbol{\Sigma}_{k_1}^{(1)}$ , partitioned as illustrated in the *Remark 2*, the following equalities hold:

$$\begin{aligned} \boldsymbol{\Sigma}'_{k_1, t\bar{t}} \boldsymbol{\Sigma}_{k_1, tt}^{-1} &= \mathbf{B}_{\bar{t}t} \quad \forall k_1, \\ \boldsymbol{\mu}_{k_1, t} &= \boldsymbol{\mu}_{h_{k_1}}, \quad \boldsymbol{\Sigma}_{k_1, tt} = \boldsymbol{\Sigma}_{h_{k_1}} \quad \forall k_1 \in I_{h_{k_1}}, \\ \boldsymbol{\mu}_{k_1, \bar{t}} &= \boldsymbol{\mu}_{j_{k_1}}, \quad \boldsymbol{\Sigma}_{k_1, \bar{t}\bar{t}} = \boldsymbol{\Sigma}_{j_{k_1}} \quad \forall k_1 \in I_{j_{k_1}}, \end{aligned}$$

where  $\mathbf{B}_{\bar{t}t}$  is a  $L_{\bar{t}} \times L_t$  matrix of regression coefficients,  $\boldsymbol{\mu}_{h_{k_1}}$  and  $\boldsymbol{\Sigma}_{h_{k_1}}$  are a  $L_t \times 1$  vector and a  $L_t \times L_t$  matrix specific to the  $h_{k_1}$ -th category of  $\tilde{Z}_1$ ,  $\boldsymbol{\mu}_{j_{k_1}}$  and  $\boldsymbol{\Sigma}_{j_{k_1}}$  denote a  $L_{\bar{t}} \times 1$  vector and a  $L_{\bar{t}} \times L_{\bar{t}}$  matrix specific to the  $j_{k_1}$ -th category of  $\tilde{Z}_2$ ,  $I_{h_{k_1}} = \{k_1 = (h_{k_1}, j_{k_1}) : j_{k_1} = 1, \dots, J\}$  and  $I_{j_{k_1}} = \{k_1 = (h_{k_1}, j_{k_1}) : h_{k_1} = 1, \dots, H\}$ . Then the marginal distribution of  $\mathbf{X}^t$  is only affected by the latent variable  $\tilde{Z}_1$  and the sub-vector  $\mathbf{X}^{\bar{t}}$  is conditionally independent of  $\tilde{Z}_1$  given  $\mathbf{X}^t$ . As a consequence of above, the Gaussian mixture for  $\mathbf{X}^{S_1}$  in

equation (8) can be decomposed as

$$\begin{aligned} \sum_{k_1=1}^{K_1} \pi_{k_1}^{(1)} \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)}) &= \\ \sum_{h_{k_1}=1}^H \pi_{h_{k_1}} \phi_{L_t}(\mathbf{x}^t; \boldsymbol{\mu}_{h_{k_1}}, \boldsymbol{\Sigma}_{h_{k_1}}) & \\ \times \sum_{j_{k_1}=1}^J \pi_{j_{k_1}} \phi_{L_{\bar{t}}}(\mathbf{x}^{\bar{t}}; \boldsymbol{\mu}_{j_{k_1}} + \mathbf{B}_{\bar{t}t} \mathbf{x}^t, \boldsymbol{\Sigma}_{j_{k_1}}). & \quad (28) \end{aligned}$$

Regrouping the second term in the right part of equation (28) with the Gaussian mixture of  $\mathbf{X}^{S_2}$  on  $\mathbf{X}^{S_1}$  leads to the following Gaussian mixture for the p.d.f. of the variable sub-vector  $\mathbf{X}^{\bar{t}} \cup \mathbf{X}^{S_2}$  given  $\mathbf{X}^t$ :

$$\sum_{m=1}^M \pi_m \phi_{L_{\bar{t}}+L_2}(\mathbf{x}^{\bar{t}} \cup \mathbf{x}^{S_2}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m),$$

where  $M = J \cdot K_2$ ,  $\pi_m = \pi_{k_2}^{(2)} \pi_{j_{k_1}}$ ,

$$\begin{aligned} \boldsymbol{\mu}_m &= \begin{pmatrix} \boldsymbol{\mu}_{j_{k_1}} + \mathbf{B}_{\bar{t}t} \mathbf{x}^t \\ \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{2\bar{t}} \boldsymbol{\mu}_{j_{k_1}} + \mathbf{B}_{2t} \mathbf{x}^t \end{pmatrix}, \\ \boldsymbol{\Sigma}_m &= \begin{bmatrix} \boldsymbol{\Sigma}_{j_{k_1}} & \boldsymbol{\Sigma}_{j_{k_1}} \mathbf{B}'_{2\bar{t}} \\ \mathbf{B}_{2\bar{t}} \boldsymbol{\Sigma}_{j_{k_1}} & \boldsymbol{\Sigma}_{k_2}^{(2)} + \mathbf{B}_{2\bar{t}} \boldsymbol{\Sigma}_{j_{k_1}} \mathbf{B}'_{2\bar{t}} \end{bmatrix}. \end{aligned}$$

Thus, a model  $(t, \bar{t} \cup S_2, U, I, H, J \cdot K_2) \in \mathcal{M}^{(2)}$  is obtained that is different from  $M^{(2)}$  but cannot be distinguished from  $M^{(2)}$  by the joint p.d.f. of  $\mathbf{X}$ .

*Remark 4 Moving variables from  $\mathbf{X}^{S_2}$  to  $\mathbf{X}^{S_1}$*

Consider now a model  $M^{(2)} = (S_1, S_2, U, I, K_1, K_2) \in \mathcal{M}^{(2)}$  such that  $K_2$  is a composite number. Similarly to the situation illustrated in *Remark 3*, suppose that the latent variable  $Z_2$  can be described in terms of a vector composed of two independent nominal latent variables  $\tilde{Z}_3$  and  $\tilde{Z}_4$  with  $P$  and  $Q$  categories, respectively, where  $P > 1$ ,  $Q > 1$  and  $P \cdot Q = K_2$ . Thus,  $\pi_{k_2}^{(2)} = \pi_{p_{k_2}} \pi_{q_{k_2}}$ , where  $(p_{k_2}, q_{k_2})$  is the element of the Cartesian product between  $\{1, \dots, P\}$  and  $\{1, \dots, Q\}$  that corresponds to the  $k_2$ -th category of  $Z_2$ .

Suppose also that for the model parameters  $\boldsymbol{\gamma}_{k_2}^{(2)}$ ,  $\boldsymbol{\Sigma}_{k_2}^{(2)}$ , partitioned as illustrated in the *Remark 1*, the following equalities hold:

$$\begin{aligned} \boldsymbol{\Sigma}'_{k_2, s\bar{s}} \boldsymbol{\Sigma}_{k_2, ss}^{-1} &= \mathbf{B}_{\bar{s}s} \quad \forall k_2, \\ \boldsymbol{\gamma}_{k_2, s} &= \boldsymbol{\gamma}_{p_{k_2}}, \quad \boldsymbol{\Sigma}_{k_2, ss} = \boldsymbol{\Sigma}_{p_{k_2}} \quad \forall k_2 \in I_{p_{k_2}}, \\ \boldsymbol{\gamma}_{k_2, \bar{s}} &= \boldsymbol{\gamma}_{q_{k_2}} + \mathbf{B}_{\bar{s}1} \mathbf{x}^{S_1} + \mathbf{B}_{\bar{s}s} \mathbf{x}^s, \quad \boldsymbol{\Sigma}_{k_2, \bar{s}\bar{s}} = \boldsymbol{\Sigma}_{q_{k_2}} \quad \forall k_2 \in I_{q_{k_2}}, \end{aligned}$$

where  $\mathbf{B}_{\bar{s}s}$  is a  $L_{\bar{s}} \times L_s$  matrix of regression coefficients,  $\boldsymbol{\gamma}_{p_{k_2}}$  and  $\boldsymbol{\Sigma}_{p_{k_2}}$  are a  $L_s \times 1$  vector and a  $L_s \times L_s$  matrix specific to the  $p_{k_2}$ -th category of  $\tilde{Z}_3$ ,  $\boldsymbol{\gamma}_{q_{k_2}}$  and  $\boldsymbol{\Sigma}_{q_{k_2}}$  denote a  $L_{\bar{s}} \times 1$  vector and a  $L_{\bar{s}} \times L_{\bar{s}}$  matrix specific to the

$q_{k_2}$ -th category of  $\tilde{Z}_4$ ,  $I_{p_{k_2}} = \{k_2 = (p_{k_2}, q_{k_2}) : q_{k_2} = 1, \dots, Q\}$  and  $I_{q_{k_2}} = \{k_2 = (p_{k_2}, q_{k_2}) : p_{k_2} = 1, \dots, P\}$ . Then the conditional distribution of  $\mathbf{X}^s$  given  $\mathbf{X}^{S_1}$  is only affected by the latent variable  $\tilde{Z}_3$ , the sub-vector  $\mathbf{X}^{\bar{s}}$  is conditionally independent of  $\tilde{Z}_3$  given  $(\mathbf{X}^{S_1}, \mathbf{X}^s)$  and the Gaussian mixture for  $\mathbf{X}^{S_2} | \mathbf{X}^{S_1}$  in equation (8) can be decomposed as

$$\begin{aligned} & \sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} \phi_{L_2} \left( \mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2}^{(2)} \right) = \\ & \sum_{p_{k_2}=1}^P \pi_{p_{k_2}} \phi_{L_s} \left( \mathbf{x}^s; \boldsymbol{\gamma}_{p_{k_2}} + \mathbf{B}_{s1} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{p_{k_2}} \right) \\ & \times \sum_{q_{k_2}=1}^Q \pi_{q_{k_2}} \phi_{L_{\bar{s}}} \left( \mathbf{x}^{\bar{s}}; \boldsymbol{\gamma}_{q_{k_2}} + \mathbf{B}_{\bar{s}1} \mathbf{x}^{S_1} + \mathbf{B}_{\bar{s}s} \mathbf{x}^s, \boldsymbol{\Sigma}_{q_{k_2}} \right). \end{aligned} \quad (29)$$

By regrouping the first term in the right part of equation (29) with the Gaussian mixture of  $\mathbf{X}^{S_1}$  the following Gaussian mixture for the p.d.f. of the variable sub-vector  $\mathbf{X}^{S_1} \cup \mathbf{X}^s$  is obtained:

$$\sum_{a=1}^A \pi_a \phi_{L_1+L_s} \left( \mathbf{x}^{S_1} \cup \mathbf{x}^s; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a \right),$$

where  $A = K_1 \cdot P$ ,  $\pi_a = \pi_{k_1}^{(1)} \pi_{p_{k_2}}$ ,

$$\begin{aligned} \boldsymbol{\mu}_a &= \begin{pmatrix} \boldsymbol{\mu}_{k_1}^{(1)} \\ \boldsymbol{\gamma}_{p_{k_2}} + \mathbf{B}_{s1} \boldsymbol{\mu}_{k_1}^{(1)} \end{pmatrix}, \\ \boldsymbol{\Sigma}_a &= \begin{bmatrix} \boldsymbol{\Sigma}_{k_1}^{(1)} & \boldsymbol{\Sigma}_{k_1}^{(1)} \mathbf{B}'_{s1} \\ \mathbf{B}_{s1} \boldsymbol{\Sigma}_{k_1}^{(1)} & \boldsymbol{\Sigma}_{p_{k_2}} + \mathbf{B}_{s1} \boldsymbol{\Sigma}_{k_1}^{(1)} \mathbf{B}'_{s1} \end{bmatrix}. \end{aligned}$$

Thus, a model  $(S_1 \cup s, \bar{s}, U, I, K_1 \cdot P, Q) \in \mathcal{M}^{(2)}$  is obtained that is different from  $M^{(2)}$  but cannot be distinguished from  $M^{(2)}$  by the joint p.d.f. of  $\mathbf{X}$ .

*Remark 5 Switching the order of  $\mathbf{X}^{S_1}$  and  $\mathbf{X}^{S_2}$*   
Consider a model  $M^{(2)} = (S_1, S_2, U, I, K_1, K_2) \in \mathcal{M}^{(2)}$  such that  $\mathbf{B}_{21} = \mathbf{0}$ . For such a model the Gaussian mixture of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$  in equation (14) becomes

$$\sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} \phi_{L_2} \left( \mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{(2)}, \boldsymbol{\Sigma}_{k_2}^{(2)} \right).$$

In this situation the model  $(S_2, S_1, U, I, K_2, K_1)$  belongs to the class  $\mathcal{M}^{(2)}$ , it is different from  $M^{(2)}$  and is also undistinguishable from  $M^{(2)}$  by the joint p.d.f. of  $\mathbf{X}$ .

*Remark 6 Moving variables from  $\mathbf{X}^U$  to  $\mathbf{X}^I$*   
Consider a given model  $M^{(2)} = (S_1, S_2, U, I, K_1, K_2) \in \mathcal{M}^{(2)}$  such that  $U \neq \emptyset$  and  $L_U \geq 2$ . Let  $v$  a nonempty subset strictly included into  $U$  (thus, composed of  $L_v$

elements, with  $1 \leq L_v < L_U$ ), and let  $\bar{v}$  be its complement in  $U$ , with the remaining  $L_U - L_v = L_{\bar{v}}$  elements. Then, the variable sub-vector  $\mathbf{X}^U$  and the model parameters  $\boldsymbol{\alpha}_0$ ,  $\mathbf{A}_1$ ,  $\mathbf{A}_2$  and  $\boldsymbol{\Sigma}_U$  are partitioned accordingly:

$$\begin{aligned} \mathbf{X}^U &= \begin{pmatrix} \mathbf{X}^v \\ \mathbf{X}^{\bar{v}} \end{pmatrix}, \boldsymbol{\alpha}_0 = \begin{pmatrix} \boldsymbol{\alpha}_{0,v} \\ \boldsymbol{\alpha}_{0,\bar{v}} \end{pmatrix}, \mathbf{A}_1 = \begin{pmatrix} \mathbf{A}_{1,v} \\ \mathbf{A}_{1,\bar{v}} \end{pmatrix}, \\ \mathbf{A}_2 &= \begin{pmatrix} \mathbf{A}_{2,v} \\ \mathbf{A}_{2,\bar{v}} \end{pmatrix}, \boldsymbol{\Sigma}_U = \begin{bmatrix} \boldsymbol{\Sigma}_{vv} & \boldsymbol{\Sigma}_{v\bar{v}} \\ \boldsymbol{\Sigma}'_{v\bar{v}} & \boldsymbol{\Sigma}_{\bar{v}\bar{v}} \end{bmatrix}. \end{aligned}$$

According to some properties of the Gaussian distribution the conditional p.d.f. of  $\mathbf{X}^U$  given  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2})$  in equation (14) can be decomposed as follows:

$$\begin{aligned} & \phi_{L_U} \left( \mathbf{x}^U; \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{x}^{S_1} + \mathbf{A}_2 \mathbf{x}^{S_2}, \boldsymbol{\Sigma}_U \right) \\ &= \phi_{L_v} \left( \mathbf{x}^v; \boldsymbol{\alpha}_{0,v} + \mathbf{A}_{1,v} \mathbf{x}^{S_1} + \mathbf{A}_{2,v} \mathbf{x}^{S_2}, \boldsymbol{\Sigma}_{vv} \right) \\ & \times \phi_{L_{\bar{v}}} \left( \mathbf{x}^{\bar{v}}; \boldsymbol{\mu}_{\bar{v}|v} + \boldsymbol{\Sigma}_{\bar{v}|v} \mathbf{x}^v, \boldsymbol{\Sigma}_{\bar{v}\bar{v}|v} \right), \end{aligned} \quad (30)$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\bar{v}|v} &= \boldsymbol{\alpha}_{0,\bar{v}} + \mathbf{A}_{1,\bar{v}} \mathbf{x}^{S_1} + \mathbf{A}_{2,\bar{v}} \mathbf{x}^{S_2} \\ & - \boldsymbol{\Sigma}_{\bar{v}|v} (\boldsymbol{\alpha}_{0,v} + \mathbf{A}_{1,v} \mathbf{x}^{S_1} + \mathbf{A}_{2,v} \mathbf{x}^{S_2}), \end{aligned} \quad (31)$$

$$\boldsymbol{\Sigma}_{\bar{v}|v} = \boldsymbol{\Sigma}'_{v\bar{v}} \boldsymbol{\Sigma}_{vv}^{-1}, \quad (32)$$

$$\boldsymbol{\Sigma}_{\bar{v}\bar{v}|v} = \boldsymbol{\Sigma}_{\bar{v}\bar{v}} - \boldsymbol{\Sigma}'_{v\bar{v}} \boldsymbol{\Sigma}_{vv}^{-1} \boldsymbol{\Sigma}_{v\bar{v}}. \quad (33)$$

If  $\mathbf{A}_{2\bar{v}} = \mathbf{0}$ ,  $\mathbf{A}_{1\bar{v}} = \mathbf{0}$  and  $\boldsymbol{\Sigma}_{v\bar{v}} = \mathbf{0}$ , then  $\mathbf{X}^{\bar{v}}$  is independent of  $\mathbf{X}^{S_1}$ ,  $\mathbf{X}^{S_2}$  and  $\mathbf{X}^v$ . Thus, the marginal p.d.f. of  $\mathbf{X}^{\bar{v}}$  can be grouped together with the marginal p.d.f. of  $\mathbf{X}^I$ . This leads to a model  $(S_1, S_2, v, I \cup \bar{v}, K_1, K_2) \in \mathcal{M}^{(2)}$  that is different from  $M^{(2)}$  but cannot be distinguished from  $M^{(2)}$  by the joint p.d.f. of  $\mathbf{X}$ .

In order to avoid that the model class  $\mathcal{M}^{(2)}$  contains the above illustrated non distinct elements it is necessary for the model parameters to fulfil the following conditions.

(C1) Conditions for  $\boldsymbol{\gamma}_2, \boldsymbol{\sigma}_2$ :

$$\begin{aligned} & \forall s \subsetneq S_2 \exists (k_2, k'_2), 1 \leq k_2 < k'_2 \leq K_2 : \\ & \boldsymbol{\mu}_{k_2, \bar{s}|s} \neq \boldsymbol{\mu}_{k'_2, \bar{s}|s} \text{ or } \boldsymbol{\Sigma}_{k_2, \bar{s}|s} \neq \boldsymbol{\Sigma}_{k'_2, \bar{s}|s} \\ & \text{ or } \boldsymbol{\Sigma}_{k_2, \bar{s}\bar{s}|s} \neq \boldsymbol{\Sigma}_{k'_2, \bar{s}\bar{s}|s}. \end{aligned}$$

(C2) Conditions for  $\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1, \mathbf{B}_{21}$ :

$$\begin{aligned} & \forall t \subsetneq S_1 \mathbf{B}_{2\bar{t}} \neq \mathbf{0} \text{ or} \\ & \forall t \subsetneq S_1 \exists (k_1, k'_1), 1 \leq k_1 < k'_1 \leq K_1 : \\ & \boldsymbol{\mu}_{k_1, \bar{t}|t} \neq \boldsymbol{\mu}_{k'_1, \bar{t}|t} \text{ or } \boldsymbol{\Sigma}_{k_1, \bar{t}|t} \neq \boldsymbol{\Sigma}_{k'_1, \bar{t}|t} \\ & \text{ or } \boldsymbol{\Sigma}_{k_1, \bar{t}\bar{t}|t} \neq \boldsymbol{\Sigma}_{k'_1, \bar{t}\bar{t}|t}. \end{aligned}$$

(C3) Additional conditions for  $\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1$  when  $K_1$  is a composite number:

$$\forall t \subsetneq S_1, \forall k_1 \in \{1, \dots, K_1\} \exists I_{t,k_1} : \forall k'_1 \in I_{t,k_1} \\ \boldsymbol{\mu}_{k_1, \bar{t}|t} \neq \boldsymbol{\mu}_{k'_1, \bar{t}|t} \text{ or } \boldsymbol{\Sigma}_{k_1, \bar{t}|t} \neq \boldsymbol{\Sigma}_{k'_1, \bar{t}|t},$$

where  $I_{t,k_1} \subsetneq \{1, \dots, K_1\} \setminus \{k_1\}$  :  
 $|I_{t,k_1}| \geq \max(H, J) - 1, \forall (H, J) : H \cdot J = K_1$ .

(C4) Additional conditions for  $\boldsymbol{\mu}_2, \boldsymbol{\sigma}_2$  when  $K_2$  is a composite number:

$$\forall s \subsetneq S_2, \forall k_2 \in \{1, \dots, K_2\} \exists I_{s,k_2} : \forall k'_2 \in I_{s,k_2} \\ \boldsymbol{\mu}_{k_2, \bar{s}|s} \neq \boldsymbol{\mu}_{k'_2, \bar{s}|s} \text{ or } \boldsymbol{\Sigma}_{k_2, \bar{s}|s} \neq \boldsymbol{\Sigma}_{k'_2, \bar{s}|s},$$

where  $I_{s,k_2} \subsetneq \{1, \dots, K_2\} \setminus \{k_2\}$  :  
 $|I_{s,k_2}| \geq \max(P, Q) - 1, \forall (P, Q) : P \cdot Q = K_2$ .

(C5) Conditions for  $\mathbf{A}_1, \mathbf{A}_2, \boldsymbol{\Sigma}_U$ :

$$\forall v \subsetneq U \mathbf{A}_{1\bar{v}} \neq \mathbf{0} \text{ or } \mathbf{A}_{2\bar{v}} \neq \mathbf{0} \text{ or } \boldsymbol{\Sigma}_{vv} \neq \mathbf{0}.$$

The identifiability conditions (C2) and (C5) for the parameters of the model class  $\mathcal{M}^{(2)}$  are of the same type of conditions already introduced for the parameters of the model classes described in Maugis *et al.* (2009a,b). The other conditions are specific for the models introduced in this paper. The illustrated conditions also allow to avoid identifiability issues associated with movements of variables in directions that are opposite to the ones considered in *Remarks 1, 2* and *6*.

### 3.4 Models with more than two independent clusterings

Models (8), (14) and (16) can be modified so as to admit that the examined dataset is characterised by  $G$  unknown cluster structures, where  $G$  can be greater than two. These structures can be modelled by assuming that  $G$  independent nominal latent variables affect the probability distribution of  $\mathbf{X}$ . Namely, assume that  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}, \dots, \mathbf{X}^{S_G}, \mathbf{X}^U, \mathbf{X}^I)$  is a splitting of  $\mathbf{X}$  into  $G+2$  non-overlapped sub-vectors.  $\mathbf{X}^{S_g}$  is the sub-vector containing  $L_g$  variables that provide information about the cluster structure  $S_g$  ( $g = 1, \dots, G$ ). As described in Section 3.1, the sub-vectors  $\mathbf{X}^U$  and  $\mathbf{X}^I$  are composed of  $L_U$  and  $L_I$  uninformative variables, respectively ( $\sum_g L_g + L_U + L_I = L$ ). For the marginal distribution of  $\mathbf{X}^{S_1}$  the assumption (A1) used in Section 3.1 still holds. The remaining assumptions are modified as follows.

(A2\*) For  $g = 2, \dots, G$ , the sub-vector  $\mathbf{X}^{S_g}$  is assumed to be conditionally independent of  $(Z_1, \dots, Z_{g-1})$  given  $(\mathbf{X}^{S_1}, \dots, \mathbf{X}^{S_{g-1}})$ ; furthermore, the conditional p.d.f. of  $\mathbf{X}^{S_g}$  given  $(\mathbf{X}^{S_1}, \dots, \mathbf{X}^{S_{g-1}})$  is affected by a

latent variable  $Z_g$  (with  $K_g$  categories) and is equal to

$$f(\mathbf{x}^{S_g} | (\mathbf{x}^{S_1}, \dots, \mathbf{x}^{S_{g-1}}); \boldsymbol{\theta}_g) = \\ \sum_{k_g=1}^{K_g} \pi_{k_g}^{(g)} \phi_{L_g}(\mathbf{x}^{S_g}; \boldsymbol{\mu}_{k_g}^{(g)}, \boldsymbol{\Sigma}_{k_g}^{(g)}), \quad (34)$$

where  $K_g \geq 2$ ,

$$\boldsymbol{\mu}_{k_g}^{(g)} = \boldsymbol{\gamma}_{k_g}^{(g)} + \sum_{h=1}^{g-1} \mathbf{B}_{gh} \mathbf{x}^{S_h},$$

$\mathbf{B}_{gh}$  is a  $L_g \times L_h$  matrix of regression coefficients,  $\boldsymbol{\theta}_g = (\boldsymbol{\pi}_g, \boldsymbol{\gamma}_g, \mathbf{B}_{g1}, \dots, \mathbf{B}_{g,g-1}, \boldsymbol{\sigma}_g)$ ,  $\boldsymbol{\pi}_g = (\pi_1^{(g)}, \dots, \pi_{K_g}^{(g)})$ ,  $\boldsymbol{\gamma}_g = (\gamma_1^{(g)}, \dots, \gamma_{K_g}^{(g)})$ ,  $\boldsymbol{\sigma}_g = (\boldsymbol{\Sigma}_1^{(g)}, \dots, \boldsymbol{\Sigma}_{K_g}^{(g)})$ .

Thus, the dependence of  $\mathbf{X}^{S_g}$  on  $(\mathbf{X}^{S_1}, \dots, \mathbf{X}^{S_{g-1}})$  is described by a multivariate linear regression model whose error terms follow a mixture of  $K_g$  Gaussian components, with the constraint that the effect of  $\mathbf{X}^{S_h}$ ,  $h = 1, \dots, g-1$ , on  $\mathbf{X}^{S_g}$  is the same for all the  $K_g$  components of the mixture (34).

(A3\*) The sub-vector  $\mathbf{X}^U$  is assumed to be conditionally independent of  $(Z_1, \dots, Z_G)$  given  $(\mathbf{X}^{S_1}, \dots, \mathbf{X}^{S_G})$ ; furthermore, the conditional distribution of  $\mathbf{X}^U$  given  $(\mathbf{X}^{S_1}, \dots, \mathbf{X}^{S_G})$  follows a Gaussian linear regression model; namely:

$$\mathbf{X}^U = \boldsymbol{\alpha}_0 + \sum_{g=1}^G \mathbf{A}_g \mathbf{x}^{S_g} + \boldsymbol{\epsilon}_U. \quad (35)$$

As far as  $\mathbf{X}^I$  is concerned, it is assumed that it is independent of any other variable in  $\mathbf{X}$  and has a marginal distribution given by a multivariate Gaussian model with mean vector  $\boldsymbol{\mu}_I$  and positive definite covariance matrix  $\boldsymbol{\Sigma}_I$ . Thus, the joint p.d.f. of  $\mathbf{X}$  can be obtained as follows:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k_1=1}^{K_1} \pi_{k_1}^{(1)} \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)}) \\ \times \prod_{g=2}^G \left[ \sum_{k_g=1}^{K_g} \pi_{k_g}^{(g)} \phi_{L_g}(\mathbf{x}^{S_g}; \boldsymbol{\mu}_{k_g}^{(g)}, \boldsymbol{\Sigma}_{k_g}^{(g)}) \right] \\ \times \phi_{L_U} \left( \mathbf{x}^U; \boldsymbol{\alpha}_0 + \sum_{g=1}^G \mathbf{A}_g \mathbf{x}^{S_g}, \boldsymbol{\Sigma}_U \right) \\ \times \phi_{L_I}(\mathbf{x}^I; \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I), \quad (36)$$

where  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I)$ , with  $\boldsymbol{\theta}_U = (\boldsymbol{\alpha}_0, \mathbf{A}_1, \dots, \mathbf{A}_G, \boldsymbol{\Sigma}_U)$ .

Using a notation similar to the one adopted in Section 3.3.2, the class of models resulting from equation (36)

can be denoted as  $\mathcal{M}^{(G)}$ ; the generic model belonging to  $\mathcal{M}^{(G)}$  is  $M^{(G)} = (S_1, \dots, S_G, U, I, K_1, \dots, K_G)$  and the parameterised densities are  $f(\cdot | \boldsymbol{\theta}_{M^{(G)}})$ , with  $\boldsymbol{\theta}_{M^{(G)}} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I) \in \mathcal{Q}_{M^{(G)}}$ . Theorem 1 can be generalised so as to hold for any subclass of  $\mathcal{M}^{(G)}$  composed of models having the same splitting of  $\mathbf{X}$ . As far as the the identifiability of  $\mathcal{M}^{(G)}$  for  $\boldsymbol{\theta}_{M^{(G)}} \in \mathcal{Q}_{M^{(G)}}$  is concerned, it cannot be ensured. The conditions (C1)-(C5) illustrated in Section 3.3.2 can be extended to prevent distinct models in the class  $\mathcal{M}^{(G)}$  from being undistinguishable by the joint p.d.f. of  $\mathbf{X}$ .

Similarly to the general models defined in Section 3.1 by removing restrictions from model (14), more general models can be defined also from model (36). Although the formal specification of such models is omitted for the ease of presentation, they can be obtained through a generalisation of the equations (15)-(18).

### 3.5 Parameter estimation

For a given model  $M^{(G)} = (S_1, \dots, S_G, U, I, K_1, \dots, K_G) \in \mathcal{M}^{(G)}$ , whose parameters are  $\boldsymbol{\theta}_{M^{(G)}} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_G, \boldsymbol{\theta}_U, \boldsymbol{\theta}_I) \in \mathcal{Q}_{M^{(G)}}$ , the estimation can be performed using the ML method. Consider a random sample  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n)$ . According to equation (36) the log-likelihood of model  $M^{(G)}$  can be written as

$$l(\boldsymbol{\theta}_{M^{(G)}}) = l_1(\boldsymbol{\theta}_1) + \sum_{g=2}^G l_g(\boldsymbol{\theta}_g) + l_U(\boldsymbol{\theta}_U) + l_I(\boldsymbol{\theta}_I), \quad (37)$$

where

$$l_1(\boldsymbol{\theta}_1) = \sum_{i=1}^n \ln \left[ \sum_{k_1=1}^{K_1} \pi_{k_1}^{(1)} \phi_{L_1} \left( \mathbf{x}_i^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)} \right) \right],$$

$$l_g(\boldsymbol{\theta}_g) = \sum_{i=1}^n \ln \left[ \sum_{k_g=1}^{K_g} \pi_{k_g}^{(g)} \phi_{L_g} \left( \mathbf{x}_i^{S_g}; \boldsymbol{\mu}_{i,k_g}^{(g)}, \boldsymbol{\Sigma}_{k_g}^{(g)} \right) \right],$$

$$l_U(\boldsymbol{\theta}_U) = \sum_{i=1}^n \ln \left[ \phi_{L_U} \left( \mathbf{x}_i^U; \boldsymbol{\alpha}_0 + \sum_{g=1}^G \mathbf{A}_g \mathbf{x}_i^{S_g}, \boldsymbol{\Sigma}_U \right) \right]$$

and

$$l_I(\boldsymbol{\theta}_I) = \sum_{i=1}^n \ln \left[ \phi_{L_I} \left( \mathbf{x}_i^I; \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I \right) \right],$$

with

$$\boldsymbol{\mu}_{i,k_g}^{(g)} = \boldsymbol{\gamma}_{k_g}^{(g)} + \sum_{h=1}^{g-1} \mathbf{B}_{gh} \mathbf{x}_i^{S_h}.$$

Equation (37) shows that  $l(\boldsymbol{\theta}_{M^{(G)}})$  is composed of  $G + 2$  parts, each of which only depends on a sub-vector of  $\boldsymbol{\theta}_{M^{(G)}}$ . Thus,  $\hat{\boldsymbol{\theta}}_{M^{(G)}} = (\hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_G, \hat{\boldsymbol{\theta}}_U, \hat{\boldsymbol{\theta}}_I)$ ,

the ML estimate of  $\boldsymbol{\theta}_{M^{(G)}}$ , can be obtained by a separate maximization of the  $G + 2$  parts. More specifically,  $\hat{\boldsymbol{\theta}}_1$  can be computed by maximizing  $l_1(\boldsymbol{\theta}_1)$  through the EM algorithm for a Gaussian mixture model (see, e.g., McLachlan and Peel 2000, chapter 3). As far as the ML estimation of  $\boldsymbol{\theta}_g$  is concerned ( $g = 2, \dots, G$ ), it can be carried out by maximizing the corresponding  $l_g(\boldsymbol{\theta}_g)$ . Soffritti and Galimberti (2011) and Galimberti *et al.* (2015) provide two EM algorithms for computing  $\hat{\boldsymbol{\theta}}_g$ . The EM algorithm described in the latter paper is also suitable to deal with models resulting from equation (18). Finally,  $\hat{\boldsymbol{\theta}}_U$  and  $\hat{\boldsymbol{\theta}}_I$  can be computed using the ML solution for a multivariate linear regression model with Gaussian error terms and a multivariate Gaussian model, respectively (see, e.g., Srivastava 2002).

### 3.6 Parsimonious models

All models for the random vector  $\mathbf{X}$  described in Sections 3.1 and 3.4 are based on Gaussian mixture models whose components are assumed to have unconstrained covariance matrices. Models with a reduced number of variance-covariance parameters can be obtained by resorting to the approach of Banfield and Raftery (1993) and Celeux and Govaert (1995). Namely, the covariance matrix  $\boldsymbol{\Sigma}_{k_g}^{(g)}$  is decomposed as follows:  $\boldsymbol{\Sigma}_{k_g}^{(g)} = \lambda_{k_g}^{(g)} \mathbf{D}_{k_g}^{(g)} \mathbf{A}_{k_g}^{(g)} \mathbf{D}_{k_g}^{(g)}$ , where  $\lambda_{k_g}^{(g)} = |\boldsymbol{\Sigma}_{k_g}^{(g)}|^{1/L_g}$ ,  $\mathbf{D}_{k_g}^{(g)}$  is the matrix of eigenvectors of  $\boldsymbol{\Sigma}_{k_g}^{(g)}$  and  $\mathbf{A}_{k_g}^{(g)}$  is the diagonal matrix containing the eigenvalues of  $\boldsymbol{\Sigma}_{k_g}^{(g)}$  (normalized in such a way that  $|\mathbf{A}_{k_g}^{(g)}| = 1$ ). In this parameterisation, the volume, shape and orientation of the  $k$ -th component in the mixture model (34) are determined by the parameters  $\lambda_{k_g}^{(g)}$ ,  $\mathbf{A}_{k_g}^{(g)}$  and  $\mathbf{D}_{k_g}^{(g)}$ , respectively. Thus, by constraining one or more of these parameters to be the same for all components, for  $g = 1, \dots, G$ , parsimonious and interpretable models can be obtained. Some details about these constraints are reported in Table A of the Supplementary Material. Let the class obtained by including also such parsimonious models be denoted as  $\mathcal{M}_{pars}^{(G)}$ , and a model of this widened class be  $M_{pars}^{(G)} = (S_1, \dots, S_G, U, I, K_1, \dots, K_G, P_1, \dots, P_G, P_U, P_I)$ , where  $P_g$  ( $g = 1, \dots, G$ ) denotes the parameterisation of the component-covariance matrices of the mixture model (34),  $P_U$  and  $P_I$  denote the form (spherical, diagonal, unconstrained) of the covariance matrices  $\boldsymbol{\Sigma}_U$  and  $\boldsymbol{\Sigma}_I$ , respectively, in model (36). The parsimonious Gaussian mixture models (and their ML estimators) resulting from imposing (up to) fourteen different constraints on such parameters are illustrated in Celeux and Govaert (1995). These models can be estimated using, for example, the R packages `mclust` and `mixture`

(Browne *et al.* 2015). Details on the estimation of parsimonious Gaussian clusterwise linear regression models can be found in Dang and McNicholas (2015). Models from the class  $\mathcal{M}_{pars}^{(G)}$  are estimated and compared in Sections 4.1, 4.2 and E (see the Supplementary Material).

### 3.7 Model selection

The selection of an appropriate model in a model class for a given dataset can be performed through the same methods usually employed to select the number of components or the parameterizations of the component-covariance matrices in model-based cluster analysis (see, e.g., McLachlan and Peel 2000). A widely employed information-based criterion is the *BIC*:

$$BIC_M = 2l(\hat{\boldsymbol{\theta}}_M) - npar_M \ln(n), \quad (38)$$

where  $l(\hat{\boldsymbol{\theta}}_M)$  and  $npar_M$  denote the maximum value of the log-likelihood and the number of estimated parameters in model  $M$ , respectively. Note that, for models  $M^{(G)}$  with a log-likelihood equal to the one defined in equation (37),  $BIC_{M^{(G)}}$  can be obtained by summing the *BIC* values associated with the  $G + 2$  parts of  $l(\hat{\boldsymbol{\theta}}_{M^{(G)}})$ .

In a Bayesian framework, *BIC* represents an asymptotic approximation of the log-posterior probability of a model under specific conditions (Kass and Raftery 1995). Keribin (2000) provides conditions that guarantee the consistency of the *BIC* in estimating the number of components for mixture models. The criteria for performing variable selection in Gaussian model-based cluster analysis proposed by Maugis *et al.* (2009a) and Maugis *et al.* (2009b), based on the *BIC*, are proved to be consistent under regularity conditions. A similar result is proved in Galimberti and Soffritti (2013) for selecting the partition of the variables in a parsimonious approach to model-based cluster analysis. From an applied point of view, good performances of the *BIC* as a model selection criterion for Gaussian mixture models are reported in several papers, such as Biernacki and Govaert (1999) and Fraley and Raftery (2002). All the experimental results illustrated in Section 4 and in the Supplementary Material are obtained by using this criterion.

In order to find the optimal model in a model class for a given dataset all possible models have to be fitted and compared. An approach based on an exhaustive search is clearly feasible only when the number of observed variables is small. An example is illustrated in Section D of the Supplementary Material. When this number is moderate or high an approach based on an

exhaustive search becomes computationally expensive and time-consuming; thus, non-exhaustive strategies are needed. Solutions may be represented, for example, by stepwise techniques and genetic algorithms. In this paper a greedy search algorithm and two genetic algorithms are developed. They are suitably devised to perform model selection in classes of models with two unknown independent clusterings. A general description of the two genetic algorithms is given in Section 3.8. For a detailed description of all algorithms see Sections A and B of the Supplementary Material.

### 3.8 Exploring model classes using genetic algorithms

Genetic algorithms constitute stochastic optimisation techniques that exploit principles and operators of the biological evolution of a species for solving complex problems with a vast number of possible solutions (Goldberg 1989). These algorithms are widely used in many fields of statistics (see, e.g., Chatterjee *et al.* 1996). Applications in subset selection problems can be found, for example, in Bozdogan (2004) and Scrucca (2016).

In general, a genetic algorithm starts from the examination of the chromosomes (ordered sequences of genes) that compose an initial population. Each of these chromosomes is randomly generated; it is assigned a value summarising its fitness. Then, an iterative evolution process is performed, based on three main genetic operators (selection, crossover, mutation), with the goal of generating novel populations composed of chromosomes characterised by improved fitness values. The selection operator consists in a weighted random sampling from the initial population with weights that are generally proportional to the chromosomes' fitness. The chromosomes selected in this way reproduce and their offspring will compose a novel generation. Such a generation is obtained after crossover and mutation. Namely, crossover is a random process of genome recombination that applies to pairs of chromosomes; mutation is a random alteration of a gene in a chromosome. The chromosomes of the resulting novel generation are assigned their fitness and the evolution process repeats. Usually, the algorithm stops when a maximum number of populations has been generated.

As far as the exploration of classes of models illustrated in this paper is concerned, a first genetic algorithm is developed for finding the model  $\hat{M}$  such that

$$\hat{M} = \operatorname{argmax}_{M \in \tilde{\mathcal{M}}^{(2)}} BIC_M, \quad (39)$$

where  $\tilde{\mathcal{M}}^{(2)}$  is the subclass of  $\mathcal{M}^{(2)}$  composed of models  $M^{(2)} = (S_1, S_2, U, I, K_1, K_2)$  with  $I = \emptyset$ .

In this algorithm each model  $M$  is represented as a chromosome whose fitness is given by  $BIC_M$ . The evolution process is composed of two parts:

- a) information extraction for the specification of model (1);
- b) information extraction for the specification of models (2) and (7).

In part *a*) the examined chromosomes have a binary gene for each variable in  $\mathbf{X}$  (with 1/0 denoting a variable selected/not selected for  $\mathbf{X}^{S_1}$ , respectively); they also have two genes associated with the number of components to be used in models (1) and (2). The possible values of these two latter genes are the integer numbers between two and a maximum value chosen by the researcher ( $K_{1max}$  for model (1) and  $K_{2max}$  for model (2)). Thus, chromosomes of length  $L + 2$  are examined in the first part of the algorithm. Let  $M_a$  be the best model detected at the end of part *a*) and let  $\hat{S}_1, \hat{K}_1$  be the solution for  $S_1, K_1$  obtained from  $M_a$ .

In part *b*) a solution for  $S_2, K_2$  and  $U$  is searched by keeping fixed the solution for  $S_1, K_1$  obtained from the model  $M_a$ . Specifically, the chromosomes are composed of  $L - \hat{L}_1 + 1$  genes, where  $\hat{L}$  is the length of  $\mathbf{X}^{\hat{S}_1}$ , with a binary gene for each variable in  $\mathbf{X} \setminus \mathbf{X}^{\hat{S}_1}$  (where 1 and 0 denote a variable selected and not selected for  $\mathbf{X}^{S_2}$ , respectively), and an additional gene (with positive integer values) associated with the number of components for model (2). The model  $M_b$  obtained at the end of this second part provides  $\hat{S}_2, \hat{K}_2, \hat{U}$ , where  $\hat{U} = \mathcal{I} \setminus \hat{S}_1 \setminus \hat{S}_2$ . Thus,  $\hat{U}$  will not be empty when  $\hat{L}_1 + \hat{L}_2 < L$ , with  $\hat{L}_2$  denoting the length of  $\mathbf{X}^{\hat{S}_2}$ . The final solution of the algorithm is  $\hat{M} = (\hat{S}_1, \hat{S}_2, \hat{U}, \hat{K}_1, \hat{K}_2)$ . All models examined and estimated with this genetic algorithm have unconstrained covariance matrices. The results of an evaluation of its effectiveness, based on simulated datasets generated from models with unconstrained covariance matrices, are provided in Section 5 and Section C of the Supplementary Material.

A second genetic algorithm is developed for carrying out the model search in the subclass  $\tilde{\mathcal{M}}_{pars}^{(2)}$  of  $\mathcal{M}_{pars}^{(2)}$  that only comprises models  $M_{pars}^{(2)} = (S_1, S_2, U, I, K_1, K_2, P_1, P_2)$  with  $I = \emptyset$ . The search is still decomposed in two parts that are similar to those described above; the main difference is in the structure of the chromosomes. Namely, in part *a*) the examined chromosomes have two additional genes (with positive integer values up to 14) for distinguishing the parsimonious parameterisations to be used in models (1) and (2). In part *b*) an additional gene is necessary to denote the parsimonious parameterisation in model (2); another gene (with up to three possible values) is added for indicating the form of the covariance matrix  $\Sigma_U$  in model (7).

Sections 4.1, 4.2 and E in the Supplementary Material show the results obtained from analyses carried out using this second algorithm.

These genetic algorithms have been implemented in R by exploiting the package GA (Scrucca 2013). Each execution requires the specification of the following tuning parameters:  $K_{1max}, K_{2max}, N_1$  and  $N_2$  (dimension of the examined populations in parts *a*) and *b*), respectively),  $d_{1max}$  and  $d_{2max}$  (maximum number of generations to be examined in the two parts of the algorithm). The specific values of these tuning parameters employed in the analyses are detailed in the following Sections. In both algorithms linear-rank selection and single point crossover operators are used. The probability of crossover between pairs of chromosomes is set equal to 0.8 in all analyses. As far as the mutation is concerned, this genetic transformation is randomly carried out in a parent chromosome with a probability of 0.1.

#### 4 Experimental results from analyses of real datasets

Each real dataset considered in this Section is analysed with six different algorithms: `mclust`, `clustvarsel`, `SelvarClust`, `SelvarClustIndep`, the greedy search algorithm illustrated in Section A of the Supplementary Material and the second genetic algorithm (see Section 3.8). The main goal is to provide examples in which the proposed approach results to be useful to discover something that cannot be found by the other methods. In this perspective, all the results summarised in this Section are obtained using implementations of the greedy search algorithm and the genetic algorithm that have not carried out with the goal of being efficient from a computational point of view. Thus, the reported CPU times are merely illustrative and can be greatly reduced using more efficient implementations.

##### 4.1 Further results from the analysis of the crabs dataset

The three models illustrated in Section 2 for the crabs dataset can be seen as a special case of the models described in Sections 3.1 and 3.2. They are listed in Table 7 together with an additional model  $M_4$ .

The mixtures in models  $M_1$  and  $M_2$  have four ellipsoidal components with the same volume and shape. Model  $M_4$  is obtained from an exploration of the model class  $\tilde{\mathcal{M}}_{pars}^{(2)}$  through the greedy search algorithm. The total CPU time required by this algorithm is three hours and 58 minutes. The model selected from this algorithm

Table 7: Maximised log-likelihood and  $BIC$  value of four models fitted to the crabs dataset.

Models	$\mathbf{X}^{S_1}$	$\mathbf{X}^{S_2}$	$\mathbf{X}^U$	$K_1$	$K_2$	$l(\hat{\boldsymbol{\theta}}_M)$	$npar_M$	$BIC_M$
$M_1$	FL, RW, CW, BD, CL	$\emptyset$	$\emptyset$	4	-	-1241.0	68	-2842.3
$M_2$	FL, RW, CW, BD	-	CL	4	-	-1265.3	53	-2811.2
$M_3$	RW, CL	FL, CW, BD	$\emptyset$	2	2	-1345.4	23	-2812.7
$M_4$	FL, RW, CW, BD	CL	$\emptyset$	4	2	-1265.2	55	-2821.8

for the joint p.d.f. of FL, RW, CW, BD coincides with the one obtained using `clustvargsel`, `SelvarClust` and `SelvarClustIndep`. As far as the conditional p.d.f. of CL is concerned, a mixture of two linear regression models with the same variances is selected. Model  $M_3$  is selected through the second genetic algorithm. Namely, eight independent executions of this algorithm are performed, one for each combination of the following values for the tuning parameters:  $N_1 = 200, 300, 400, 500$ ,  $d_{1max} = 40, 50$ . The remaining tuning parameters are set as follows:  $N_1 = N_2$ ,  $d_{1max} = d_{2max}$  and  $K_{1max} = K_{2max} = 5$ . A further execution is carried out by using values of  $N_1$  and  $d_{1max}$  chosen in a way that the total CPU time required by the genetic algorithm is similar to the one of the greedy search algorithm. Namely, an execution with  $N_1 = 65$  and  $d_{1max} = 40$  has required three hours and 31 minutes. This setting makes it possible to compare the effectiveness of these two algorithms based on a similar CPU time (although their implementations are not optimised from a computational point of view). In all these nine executions the selected model is  $M_3$ . In this model the parsimonious mixture used to model the marginal p.d.f. of (RW, CL) has two ellipsoidal components with the same volume and shape; as far as the mixture model for the conditional p.d.f. of (FL, CW, BD) given (RW, CL) is concerned, its components are spherical with the same volume. According to the  $BIC$  values of these four models,  $M_2$  and  $M_3$  perform better than  $M_1$  and  $M_4$ .

In model  $M_3$  the measurements FL, CW and BD are assumed to linearly depend on both RW and CL. Since this assumption could be restrictive, models are also estimated in which a different set of regressors is allowed for each of the three regression equations in the multivariate linear regression model of frontal lobe size, carapace width and body depth (see equation (18)). Namely, given the splitting of the five measurements obtained from the genetic algorithm, a regressors selection is carried out through an exhaustive search. According to the  $BIC$ , the best solution obtained after examining these further models is a mixture of two seemingly unrelated linear regression models in which frontal lobe size and carapace width are both regressed on carapace length and rear width, while body depth only depends

Table 8: Classification of the crabs according to their colour and sex and the segmentation based on the joint examination of the two independent cluster structures detected by model  $M_5$ .

Cluster I	Cluster II	Colour and sex			
		BF	BM	OF	OM
1	1	50	7	0	0
1	2	0	0	50	3
2	1	0	43	0	0
2	2	0	0	0	47
aRi		0.873			

on carapace length. The clustering of the crabs resulting from such a model coincides with the classification of crabs based on their colour (aRi = 1). The joint model for the five morphological measurements, denoted as  $M_5$ , is given by the product of this seemingly unrelated linear regression model and the Gaussian mixture model for CL and RW described above. It has a  $BIC$  value (-2808.3) which is higher than the one of any other examined model. Model  $M_5$  also reaches the best performance in recovering the classification of the crabs based on colour and sex (aRi = 0.873) (see Table 8).

Figure 3 provides some graphical displays about model  $M_5$ . In particular, Figure 3(a) shows the joint bivariate scatterplot of RW and CL where the first cluster structure is represented. In this figure, each point is labelled according to its estimated cluster membership. The superimposed ellipses are centred on the estimated component mean vectors; their volume, shape and orientation depend on the parameterisation of the component-covariance matrices. Figures 3(b), 3(c) and 3(d) refer to the second cluster structure. They show the three two-dimensional scatterplots of the variables FL, CW and BD, after removing the estimated effects of RW and CL. Labels correspond to the clusters of the second cluster structure. The centres of the superimposed ellipses are obtained from the estimated values of  $\boldsymbol{\gamma}_{k_2}^{(2)}$ , ( $k_2 = 1, 2$ ). Since the selected model for the conditional p.d.f. of (FL, CW, BD) given (RW, CL) has spherical components with equal volume, the superimposed ellipses are circles with the same radius.

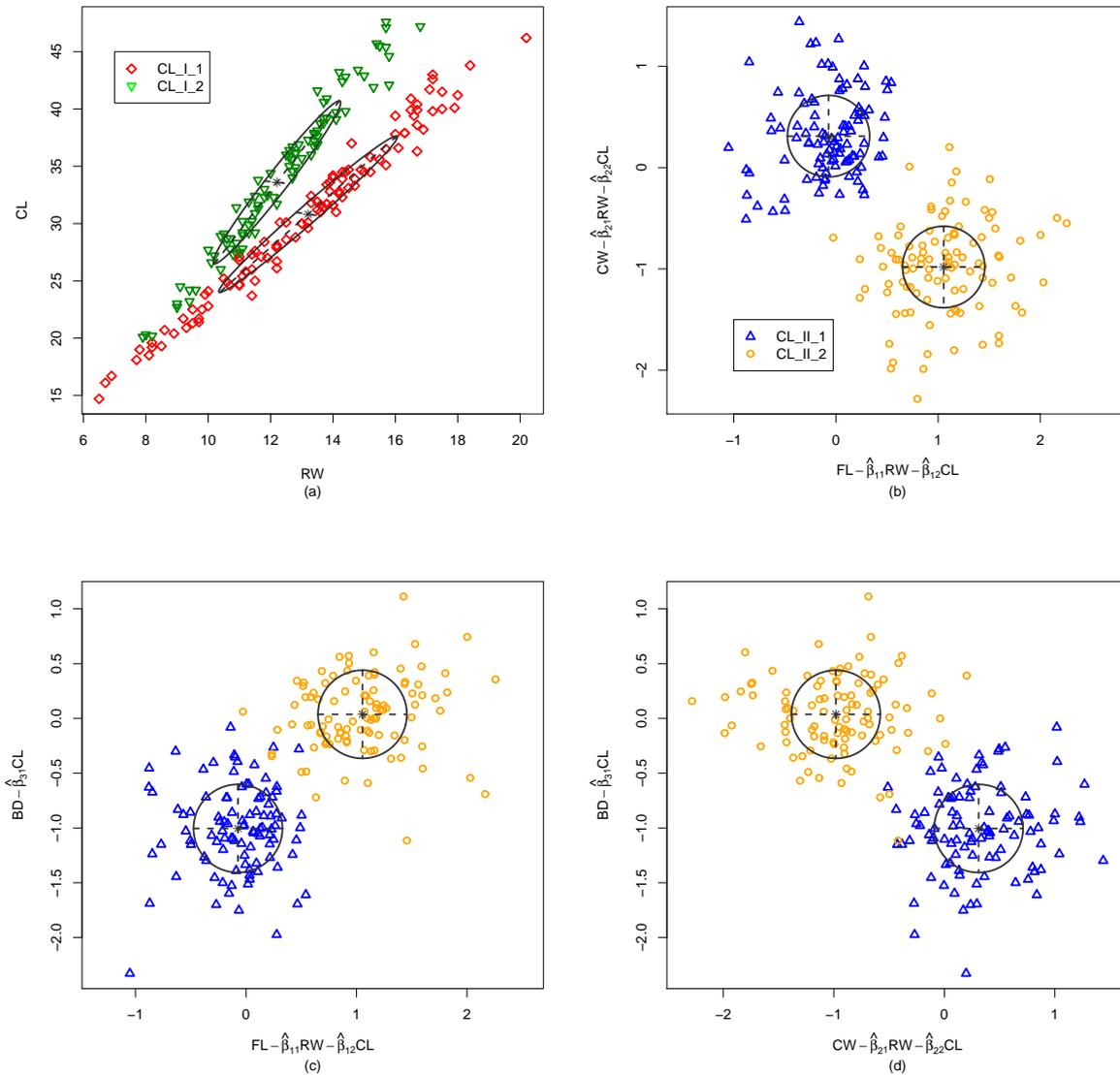


Fig. 3: Scatterplots for the crabs dataset. Points are labelled according to their estimated cluster memberships. Panel (a) refers to the first cluster structure and is based on the observed values of variables RW and CL. Panels (b), (c) and (d) refer to the second cluster structure and are based on the values of FL, CW and BD after removing the estimated effects of RW and CL.

#### 4.2 Analysis of the quality of life in Italy

Since 1990 the Italian financial newspaper *Il Sole 24 Ore* ([www.ilsole24ore.com](http://www.ilsole24ore.com)) has carried out yearly a research about the quality of life in the Italian provinces. The dataset considered in this Section refers to the year 2002 and contains the following five indicators for 103 provinces: the deposits per inhabitant ( $X_1$ ), the value added at market prices per inhabitant ( $X_2$ ), the per capita disposable income ( $X_3$ ), the number of retired people per 1000 working people ( $X_4$ ) and the unem-

ployment rate ( $X_5$ ). Variables in the dataset are generally characterised by moderate/high pairwise linear dependencies (see Table 9).

The models selected by the six algorithms for clustering the Italian provinces based on the illustrated indicators are summarised in Table 10. Model  $M_1$  results from an analysis performed through *mclust* with a maximum number of components equal to six; it is a mixture of five Gaussian ellipsoidal components with the same volume, orientation and shape. Model  $M_2$  is obtained using the variable selection methods imple-

Table 9: Pearson correlation matrix in the dataset concerning the quality of life in Italy.

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$
$X_1$	1.000	0.869	0.772	-0.503	-0.681
$X_2$	0.869	1.000	0.895	-0.564	-0.845
$X_3$	0.772	0.895	1.000	-0.318	-0.789
$X_4$	-0.503	-0.564	-0.318	1.000	0.491
$X_5$	-0.681	-0.845	-0.789	0.491	1.000

mented by `clustvarel` (setting the maximum number of components for the p.d.f. of the informative variables equal to six). According to these methods, the relevant indicators for clustering the Italian provinces are the deposits per inhabitant and the unemployment rate. The best Gaussian mixture model fitted to p.d.f. of these variables is a mixture of four Gaussian ellipsoidal components with the same volume and orientation. According to the *BIC* values, model  $M_2$  performs better than  $M_1$ . Three independent executions of `SelvarClust` lead to model  $M_3$ ; the same model is also obtained after three independent executions of `SelvarClustIndep`. According to these methods, a third indicator (value added at market prices per inhabitant) is also relevant for clustering the Italian provinces. The best model for the joint p.d.f. of  $(X_1, X_2, X_5)$  detected by `SelvarClust` and `SelvarClustIndep` is a mixture of four Gaussian ellipsoidal components with the same volume and orientation. As far as the Gaussian linear regression model for the two uninformative indicators (the per capita disposable income and the number of retired people per 1000 working people) is concerned, only the value added at market prices per inhabitant is used as a regressor; the covariance matrix is unconstrained. According to the *BIC* values, model  $M_3$  is better than  $M_1$  and  $M_2$ .

Models  $M_4$  and  $M_5$  are obtained using the greedy search algorithm and the second genetic algorithm, respectively. The total CPU time of the analysis based on the greedy search algorithm is 4 hours and 34 minutes. According to model  $M_4$  the first cluster structure that characterises the examined dataset is defined in the joint marginal distribution of the following indicators: value added at market prices per inhabitant, per capita disposable income and unemployment rate. The model for the p.d.f. of these indicators selected by the greedy search algorithm is a mixture of two Gaussian ellipsoidal components with the same volume. The second cluster structure is detected in the conditional p.d.f. of the deposits per inhabitant given the value added at market prices per inhabitant, the per capita disposable income and the unemployment rate. The selected model for this p.d.f. is a mixture of two Gaussian ellipsoidal components with equal volume, orientation and shape.

For the conditional p.d.f. of the number of retired people per 1000 working people given the other indicators a Gaussian linear regression model is selected; thus, in model  $M_4$  this latter indicator composes  $\mathbf{X}^U$  and represents an uninformative variable. However, according to the *BIC* values, model  $M_4$  is worse than  $M_2$  and  $M_3$ .

Eight independent executions of the genetic algorithm are carried out by combining the values 200, 300, 400 and 500 of  $N_1$  with the values 40 and 50 of  $d_{1max}$ . As far as the remaining tuning parameters are concerned, they are set as follows:  $N_1 = N_2$ ,  $d_{1max} = d_{2max}$  and  $K_{1max} = K_{2max} = 5$ . In all these executions the selected model is  $M_5$ . The CPU times required by the executions with  $N_1 = 200$ ,  $d_{1max} = 40$  and  $N_1 = 200$ ,  $d_{1max} = 50$  are almost the same (4 hours and 37 minutes, 4 hours and 35 minutes, respectively) as the CPU time of executing the greedy search algorithm. According to model  $M_5$ , the information about the first cluster structure is given by the joint marginal distribution of three indicators: the deposits per inhabitant, the number of retired people per 1000 working people and the unemployment rate. The selected model for the p.d.f. of these indicators is a mixture of four Gaussian ellipsoidal components with the same volume and orientation. Panels (a)-(c) of Figure 4 show the clustering of the Italian provinces associated with this model. Cluster `CL_I_1` is composed of 64 provinces that are all located in the northern and central parts of Italy. This cluster contains Italian provinces that have high mean values of deposits and low mean values of the unemployment rate. Cluster `CL_I_4` only contains three provinces (Rome, Milan and Trieste), whose distinctive feature with respect to the previous cluster is a higher value of deposits. Cluster `CL_I_2` is composed of 23 provinces that are characterised by low mean values of deposits and high mean values of both the number of retired people per 1000 working people and the unemployment rate. The 13 provinces assigned to cluster `CL_I_3` are similar to the ones of cluster `CL_I_2`, but have higher values of both the number of retired people per 1000 working people and the unemployment rate. Provinces belonging to clusters `CL_I_2` and `CL_I_3` are all located in the southern part of Italy. Thus, the first cluster structure mainly reflects geographical differences. The second clustering of the Italian provinces is detected in the conditional p.d.f. of the per capita disposable income given the amount of deposits per inhabitant, the number of retired people per 1000 working people and the unemployment rate. The selected model for this p.d.f. is a mixture of two Gaussian ellipsoidal components with equal volume, orientation and shape. Panel (d) of Figure 4 shows that the p.d.f. of the disposable income after removing the estimated

Table 10: Models for the dataset concerning the quality of life in Italy.

Models	$\mathbf{X}^{S_1}$	$\mathbf{X}^{S_2}$	$\mathbf{X}^U$	$K_1$	$K_2$	$BIC_M$
$M_1$	$X_1, X_2, X_3, X_4, X_5$	$\emptyset$	$\emptyset$	5	-	-7289.8
$M_2$	$X_1, X_5$	-	$X_2, X_3, X_4$	4	-	-7260.8
$M_3$	$X_1, X_2, X_5$	-	$X_3, X_4$	4	-	-7258.6
$M_4$	$X_2, X_3, X_5$	$X_1$	$X_4$	2	2	-7274.9
$M_5$	$X_1, X_4, X_5$	$X_3$	$X_2$	4	2	-7255.4

Table 11: Cluster structures in the dataset concerning the quality of life in Italy (detected by model  $M_5$ ) and their mutual association.

Cluster II	Cluster I			
	1	2	3	4
1	40	8	10	1
2	24	15	3	2
aRi	0.037			

effects of the three indicators  $X_1$ ,  $X_4$ , and  $X_5$  is bimodal. The two clusters detected by the genetic algorithm in the conditional p.d.f. of  $X_3$  given  $(X_1, X_4, X_5)$  are composed of 59 and 44 provinces. The estimated mean value of the per capita disposable income after removing the estimated effects of the three indicators  $X_1$ ,  $X_4$ , and  $X_5$  within cluster  $CL\_II\_1$ , given by  $\hat{\gamma}_1^{(2)}$ , is equal to 9487.64; it is higher than that in the other cluster ( $\hat{\gamma}_2^{(2)} = 7354.80$ ). This second clustering of Italian provinces is not associated with the first one (see Table 11); thus, the second cluster structure detected in the dataset is reasonably affected by latent factors that are independent of geographical differences among provinces. Finally, a Gaussian linear regression model is selected for the value added at market prices per inhabitant. Thus, this latter indicator composes  $\mathbf{X}^U$  and represents an uninformative variable. No improvement of the linear regression mixture model with two components selected for  $X_3$  is obtained after performing a regressor selection through an exhaustive search. The same result also holds for the linear regression model selected for  $X_2$ . Thus,  $X_3$  linearly depends on  $X_1$ ,  $X_4$  and  $X_5$ ; similarly,  $X_2$  linearly depends on  $X_1$ ,  $X_3$ ,  $X_4$  and  $X_5$ . In these linear regression models both the value added per inhabitant and the per capita disposable income are positively affected by the deposits per inhabitant and negatively affected by the unemployment rate; the effect of the number of retired people per 1000 working people is positive on the disposable income and negative on the value added; finally, there is a positive effect of the disposable income on the deposits per inhabitant. According to the  $BIC$  values of the five models summarised in Table 10,  $M_5$  reaches the best performance for this dataset.

The two cluster structures obtained from model  $M_5$  are compared with the ones discovered by models  $M_1$ ,  $M_2$  and  $M_3$ . In particular, the results concerning the comparison with the first cluster structure associated with  $M_5$  (see Table 12) show that using methods `mclust`, `clustvarsel`, `SelvarClust` and `SelvarClustIndep` lead to detect clusters that are quite similar to the ones that compose the first cluster structure. Thus, the three partitions obtained through `mclust`, `clustvarsel`, `SelvarClust` and `SelvarClustIndep` capture the geographical differences among the provinces. However, they miss the second source of clustering associated with latent factors that are independent of geographical differences.

## 5 Experimental results from analyses of simulated datasets

The performance of the first genetic algorithm is evaluated through two Monte Carlo experiments where artificial datasets are generated from a given known model. The main goal is to evaluate the effectiveness of this algorithm (with the  $BIC$  as a fitness measure) in selecting the model the datasets come from. This Section reports the results obtained in the first experiment. As far as the second Monte Carlo study is concerned see Section C of the Supplementary Material.

In the first experiment the artificial datasets are generated in the Euclidean space  $\mathbb{R}^8$  using model (8), where  $\mathbf{X}^{S_1} = (X_1, X_2, X_3)$ ,  $K_1 = 2$ ,  $\mathbf{X}^{S_2} = (X_4, X_5, X_6)$ ,  $K_2 = 2$ , and  $\mathbf{X}^U = (X_7, X_8)$ .

Specifically, the parameters of the marginal p.d.f. of  $\mathbf{X}^{S_1}$  are:  $\pi_1^{(1)} = 0.5$ ,

$$\boldsymbol{\mu}_1^{(1)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1^{(1)} = \begin{pmatrix} 1 & -0.6 & -0.3 \\ -0.6 & 1 & -0.4 \\ -0.3 & -0.4 & 1 \end{pmatrix},$$

$$\boldsymbol{\mu}_2^{(1)} = \begin{pmatrix} 5 \\ -5 \\ 5 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2^{(1)} = \begin{pmatrix} 1 & 0.6 & 0.3 \\ 0.6 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{pmatrix}.$$

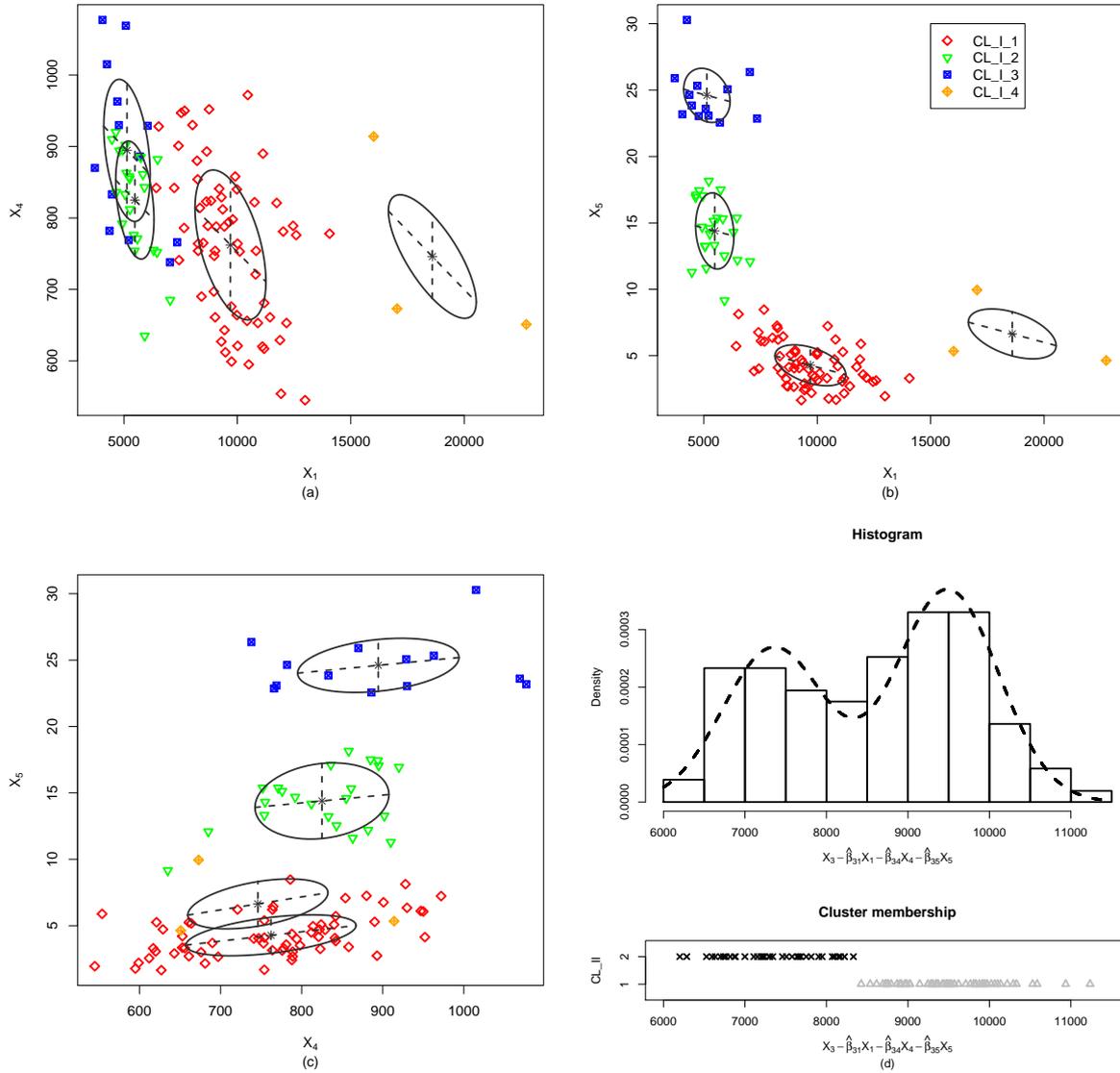


Fig. 4: Plots for the dataset concerning the quality of life in Italy. Panels (a), (b) and (c) show the bivariate scatterplots for the three indicators ( $X_1$ ,  $X_4$ ,  $X_5$ ) that provide information about the first cluster structure associated with model  $M_5$  ( $CL I$ ). Points are labelled according to the provinces' estimated cluster memberships. Panel (d) refers to the second cluster structure ( $CL II$ ) associated with  $M_5$ . It contains the density histogram of  $X_3$  after removing the estimated effects of  $X_1$ ,  $X_4$ , and  $X_5$ ; the dotted line denotes the corresponding estimated p.d.f.; the estimated cluster memberships of the provinces for the second structure are also illustrated.

Table 12: Comparison between the first cluster structure associated with model  $M_5$  (cluster I) and the clusterings of the Italian provinces estimated by models  $M_1$ ,  $M_2$  and  $M_3$ .

Cluster I	Clusters from $M_1$					Clusters from $M_2$				Clusters from $M_3$			
	1	2	3	4	5	1	2	3	4	1	2	3	4
1	30	34	0	0	0	64	0	0	0	64	0	0	0
2	0	0	0	23	0	0	0	23	0	0	0	13	10
3	0	0	0	0	13	0	0	0	13	0	0	0	13
4	0	0	3	0	0	0	3	0	0	0	3	0	0
aRi	0.590					1.000				0.900			

Table 13: Distribution of 100 simulated datasets according to the partition of  $\mathbf{X}$  obtained with the genetic algorithm in the executions with  $n = 200$ .

	$N_1$											
	80			120			160			200		
	30	40	50	30	40	50	30	40	50	30	40	50
Correct classification of all variables	62	62	61	76	79	80	91	88	91	91	91	90
Correct recovery of $\mathbf{X}^{S_1}$ only	1	2	2	4	3	3	3	4	3	3	3	4
$\hat{S}_1 = (1, 3, 7, 8)$ , $\hat{S}_2 = (2, 4, 5, 6)$	25	24	26	14	14	13	5	8	5	6	6	5
Other wrong partitions	12	12	11	6	4	4	1	0	1	0	0	1

Table 14: Distribution of 100 simulated datasets according to the partition of  $\mathbf{X}$  obtained with the genetic algorithm in the executions with  $n = 400$ .

	$N_1$											
	80			120			160			200		
	30	40	50	30	40	50	30	40	50	30	40	50
Correct classification of all variables	75	75	75	87	87	87	92	92	92	95	95	95
Correct recovery of $\mathbf{X}^{S_1}$ only	2	2	2	1	1	1	2	2	2	2	2	2
$\hat{S}_1 = (1, 3, 7, 8)$ , $\hat{S}_2 = (2, 4, 5, 6)$	17	17	17	9	9	9	5	5	5	3	3	3
Other wrong partitions	6	6	6	3	3	3	1	1	1	0	0	0

The parameters of the conditional p.d.f. of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$  are:  $\pi_1^{(2)} = 0.5$ ,

$$\gamma_1^{(2)} = \begin{pmatrix} -2 \\ -1 \\ 3.5 \end{pmatrix}, \quad \Sigma_1^{(2)} = \begin{pmatrix} 1 & 0.5 & 0.6 \\ 0.5 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{pmatrix},$$

$$\gamma_2^{(2)} = \begin{pmatrix} 4 \\ 5 \\ -2.5 \end{pmatrix}, \quad \Sigma_2^{(2)} = \begin{pmatrix} 1 & -0.5 & -0.6 \\ -0.5 & 1 & -0.4 \\ -0.6 & -0.4 & 1 \end{pmatrix},$$

$$\mathbf{B}_{21} = \begin{pmatrix} 1.5 & 2 & 1.5 \\ 1.5 & -2.5 & -2 \\ 1.5 & 2 & -2.5 \end{pmatrix}.$$

Finally, the parameters of the conditional p.d.f. of  $\mathbf{X}^U$  given  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2})$  are:  $\alpha_0 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ ,  $\mathbf{A}_1 = \begin{pmatrix} 2 & 2 & 2 \\ -2 & -2 & -2 \end{pmatrix}$ ,  $\mathbf{A}_2 = -\mathbf{A}_1$  and  $\Sigma_U = \begin{pmatrix} 2.25 & 0 \\ 0 & 1 \end{pmatrix}$ .

One hundred samples of  $n = 200$  observations each are generated and analysed using the first genetic algorithm. Since the algorithm's performance may depend on how large is the exploration of the model space, the algorithm is executed by changing the values of the tuning parameters  $N_1$  and  $d_{1max}$  that control the information extraction for the specification of model (1). Namely, the examined values are 80, 120, 160 and 200 for  $N_1$ ; 30, 40 and 50 for  $d_{1max}$ . The other tuning parameters are kept constant throughout the experiment; they are set as follows:  $K_{1max} = K_{2max} = 3$ ,  $N_2 = 80$  and  $d_{2max} = 20$ . Greater values of  $N_2$  and  $d_{2max}$  are not examined because some preliminary analyses have highlighted that increasing them has a little impact on

the results. This is mainly due to the fact that in part *b*) of the algorithm the exploration of the model space is carried out conditionally on the results obtained in part *a*). For each sample, twelve executions of the algorithm are performed. The same analysis is carried out with the sample size  $n = 400$ .

The effectiveness of the genetic algorithm is evaluated with respect to the ability to recover the correct variable partition. The percentage of datasets for which  $\mathbf{X}^{S_1}$ ,  $\mathbf{X}^{S_2}$  and  $\mathbf{X}^U$  are successfully identified is generally high, especially with  $n = 400$  (see Tables 13 and 14, first row). The most common error is represented by a partition in which both uninformative variables are allocated to the variable sub-vector that defines the first cluster structure, and  $X_2$  is wrongly inserted in the variable sub-vector that defines the second cluster structure.

As far as the choice of the tuning parameters is concerned, the beneficial effect of increasing  $N_1$  is quite clear. Larger values of  $N_1$  allow a wider exploration of the model space. This reduces the chance of selecting the wrong model. It is interesting to note that the percentage of datasets for which  $\mathbf{X}^{S_1}$ ,  $\mathbf{X}^{S_2}$  and  $\mathbf{X}^U$  are successfully identified increases nonlinearly with  $N_1$ . Similar conclusions about the effect of  $N_1$  can be drawn from the results of the second Monte Carlo study (see Section C of the Supplementary Material). The choice of  $d_{1max}$  seems to be less crucial. When the sample size is 200 the differences in the performance obtained for different values of  $d_{1max}$  are negligible, for any given value of  $N_1$ . Using a value greater than 30 for the maximum number of examined generations, in association

with any examined value of  $N_1$ , is completely useless with  $n = 400$ .

The effectiveness is also evaluated with respect to the ability of the genetic algorithm to recover the two latent cluster structures. This task is carried out by computing the adjusted Rand index between the true cluster structures and the structures estimated by the algorithm. Although in some datasets the true variable partition is not correctly detected, the first cluster structure identified by the genetic algorithm perfectly coincides with the first true one in all samples ( $aRi = 1$ ). Also the agreement between the second true cluster structure and the second estimated one is very high: the mean (over 100 datasets) of the  $aRi$  is greater than 0.996 in all executions of the algorithm for both sample sizes. These results are due to the fact that there is a clear-cut separation between clusters in both  $\mathbf{X}^{S_1}$  and  $\mathbf{X}^{S_2}$ ; furthermore, in all datasets in which the genetic algorithm selects wrong variable partitions, two variables in both  $\mathbf{X}^{\hat{S}_1}$  and  $\mathbf{X}^{\hat{S}_2}$  are always correctly selected.

## 6 Conclusions

The approach for modelling the role of variables in cluster analysis proposed in this paper relies on a model that makes it possible to compare the results of different (supervised and unsupervised) analyses carried out on a given dataset. Namely, this model allows to perform variable selection in model-based clustering according to the methods proposed by Raftery and Dean (2006), Maugis *et al.* (2009a) and Maugis *et al.* (2009b). It also allows to carry out model selection in multivariate linear regression analysis and seemingly unrelated linear regression analysis by assuming either a Gaussian model or a Gaussian mixture model for the distribution of the errors (Soffritti and Galimberti 2011; Galimberti *et al.* 2015). Furthermore, using the proposed model enables the detection of the presence of multiple cluster structures from possibly correlated variable sub-vectors.

As already remarked in Section 3.3, the identifiability conditions provided in this paper do not guarantee general identifiability. Thus, some caution is needed when interpreting the results obtained from fitting models from the proposed class to a given dataset. From a practical point of view, in order to reveal whether a fitted model suffers from identifiability problems, bootstrap methods could be employed, as suggested by (Grün and Leisch 2004).

The process of selecting a model in the proposed approach may be complex, particularly when  $G > 2$  cluster structures are assumed. This may prevent the methodology illustrated in this paper from being used

with high-dimensional datasets. In this paper two different selection strategies are examined, in the special case  $G = 2$ : one performs a greedy search and the other one is based on genetic algorithms. Resorting to these strategies can partly mitigate the drawback just mentioned. From the results obtained on two real datasets (see Sections 2, 4.1 and ??) it emerges that an approach based on a genetic algorithm performs better than a greedy search technique. This result is due to the fact that genetic algorithms are able to globally explore a given model class, thus avoiding the main problems that typically characterise stepwise and greedy search strategies. Clearly, the effectiveness of a genetic algorithm greatly depends on how large is the exploration of the model class. A proper choice of the parameters that control this exploration represents the main issue of genetic algorithms. No general rule about how to perform this choice is available. For this reason it is always advisable to run a genetic algorithm different times by varying the values of the tuning parameters. This approach is exploited in all analyses illustrated in Section 4. In particular, different settings for the population size and the number of generations are considered.

From the Monte Carlo studies summarised in Section 5 and in the Supplementary Material it emerges that an important role is played by the population size  $N_1$  employed to discover the first cluster structure. In general, the choice of this tuning parameter is strictly related to the number of observed variables, as this number directly affects the dimension of the model class. In principle, a large value ensures a large exploration of the model class and, thus, should be employed. However, increasing the population size also slows down the search and increases the computational burden. Note that, according to some authors (see, e.g., Bozdogan 2004; Scrucca 2016), after some limit an increase of this tuning parameter does not lead to any additional benefit. As far as the population size  $N_2$  is concerned, its choice has a lower impact on the results. This is due to the fact that the genetic algorithms developed in this paper search for the two cluster structures in a sequential way. For this reason,  $N_2$  is set equal to or lower than  $N_1$  in all analyses. It is worth mentioning that the computational burden of the proposed genetic algorithms depends not only on the choice of the tuning parameters but also on the effort required to fit a model and to compute its *BIC*.

Another issue is represented by the choice of the model selection criterion. All results described in this paper are obtained using the *BIC*. The Monte Carlo studies show that the performance of this criterion is satisfactory when the true model is in the examined

model class. However, since this criterion tends to overestimate the number of clusters when the correct model is not in the considered model class, criteria more robust to violation of some of the mixture model assumptions should be considered and evaluated (see, e.g., Bieracki *et al.* 2000).

A proper choice of the tuning parameters in a genetic algorithm as well as other aspects concerning model selection (e.g.: a comparison with other non-exhaustive strategies; how to deal with high-dimensional datasets) represent topics for future research. Another issue to be investigated is represented by the computational efficiency of genetic algorithms: for example, resorting to parallel computing can reduce the overall computational time. Nevertheless, the experimental results illustrated in Sections 4.1 and ?? show that for some datasets the joint use of supervised and unsupervised learning methods allows to extract unknown relevant information that otherwise would be missed, thus supporting the usefulness of the approach illustrated in this paper.

## 7 Appendix. Proof of Theorem 1

The proof exploits arguments similar to the ones used by Hennig (2000). It refers to any given splitting  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}, \mathbf{X}^U, \mathbf{X}^I)$  of  $\mathbf{X}$  in which both  $\mathbf{X}^U$  and  $\mathbf{X}^I$  are not empty. Thus,  $\theta_M$  is composed of four non empty sub-vectors. This proof can be easily modified so as to deal with situations in which  $\mathbf{X}^U = \emptyset$  and/or  $\mathbf{X}^I = \emptyset$ .

Let  $\theta_M$  and  $\theta_{M^*}$  be such that

$$f(\mathbf{x}; \theta_M) = f(\mathbf{x}; \theta_{M^*}) \quad \forall \mathbf{x} \in \mathbb{R}^L. \quad (40)$$

In the following it is shown that the equality (40) implies that  $M = M^*$  and  $\theta_M = \theta_{M^*}$ . This is the only implication that needs to be proved in order to guarantee identifiability (Hennig 2000).

The proof is composed of four parts. In the first part it is shown that  $K_1 = K_1^*$  and  $\theta_1 = \theta_1^*$ ; the second part proves that  $K_2 = K_2^*$  and  $\theta_2 = \theta_2^*$ ; finally, the last two parts demonstrate that  $\theta_U = \theta_U^*$  and  $\theta_I = \theta_I^*$ , respectively.

According to equations (14) and (1), integrating each side of the equality (40) with respect to  $\mathbf{X}^{S_2}$ ,  $\mathbf{X}^U$  and  $\mathbf{X}^I$  yields  $f(\mathbf{x}^{S_1}; \theta_1) = f(\mathbf{x}^{S_1}; \theta_1^*) \quad \forall \mathbf{x}^{S_1} \in \mathbb{R}^{L_1}$ , that is:

$$\begin{aligned} \sum_{k_1=1}^{K_1} \pi_{k_1}^{(1)} \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{(1)}, \boldsymbol{\Sigma}_{k_1}^{(1)}) = \\ \sum_{k_1=1}^{K_1^*} \pi_{k_1}^{*(1)} \phi_{L_1}(\mathbf{x}^{S_1}; \boldsymbol{\mu}_{k_1}^{*(1)}, \boldsymbol{\Sigma}_{k_1}^{*(1)}) \quad \forall \mathbf{x}^{S_1} \in \mathbb{R}^{L_1}. \end{aligned}$$

Given the constraints (I1) on  $\theta_1$ , the class of distributions that contains  $f(\mathbf{x}^{S_1}; \theta_1)$  and  $f(\mathbf{x}^{S_1}; \theta_1^*)$  is identifiable. Thus,  $K_1 = K_1^*$  and  $\theta_1 = \theta_1^*$  (up to a permutation of the mixture components).

For the second part of the proof it is useful to recall from equation (6) that the expected value of  $\mathbf{X}^{S_2}$  given  $\mathbf{X}^{S_1}$  within the  $k_2$ -th component of the model (2) is

$$\boldsymbol{\mu}_{k_2}^{(2)} = \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \quad k_2 = 1, \dots, K_2.$$

Let

$$\begin{aligned} C^{(1)} &= \{\mathbf{x}^{S_1} \in \mathbb{R}^{L_1} : \forall j \in \{1, \dots, K_1\}, \forall l \in \{1, \dots, K_1^*\}, \\ &\boldsymbol{\gamma}_j^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1} = \boldsymbol{\gamma}_l^{*(2)} + \mathbf{B}_{21}^* \mathbf{x}^{S_1} \\ &\Rightarrow \boldsymbol{\gamma}_j^{(2)} = \boldsymbol{\gamma}_l^{*(2)}, \mathbf{B}_{21} = \mathbf{B}_{21}^*\}. \end{aligned}$$

The set  $C^{(1)}$  contains all the vectors  $\mathbf{x}^{S_1}$  that can be used to distinct different values of  $(\boldsymbol{\gamma}_{k_2}^{(2)}, \mathbf{B}_{21})$  by different values of  $\boldsymbol{\mu}_{k_2}^{(2)}$ . This set is the complement of a finite union of  $(L_1 - 1)$ -dimensional hyperplanes of  $\mathbb{R}^{L_1}$ . Thus,  $\mathbb{P}(\mathbb{R}^{L_1} \setminus C^{(1)}) = 0$  and  $\mathbb{P}(C^{(1)}) = 1$  according to the Gaussian mixture model defined in equation (1).

Integrating each side of the equality (40) with respect to  $\mathbf{X}^U$  and  $\mathbf{X}^I$  and then conditioning on any  $\mathbf{x}^{S_1} \in C^{(1)}$  leads to  $f(\mathbf{x}^{S_2} | \mathbf{x}^{S_1}; \theta_2) = f(\mathbf{x}^{S_2} | \mathbf{x}^{S_1}; \theta_2^*) \quad \forall \mathbf{x}^{S_2} \in \mathbb{R}^{L_2}$ , that is:

$$\begin{aligned} \sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} \phi_{L_2}(\mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21} \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2}^{(2)}) = \\ \sum_{k_2=1}^{K_2^*} \pi_{k_2}^{*(2)} \phi_{L_2}(\mathbf{x}^{S_2}; \boldsymbol{\gamma}_{k_2}^{*(2)} + \mathbf{B}_{21}^* \mathbf{x}^{S_1}, \boldsymbol{\Sigma}_{k_2}^{*(2)}) \quad \forall \mathbf{x}^{S_2} \in \mathbb{R}^{L_2}. \end{aligned}$$

Given the constraints (I2) on  $\theta_2$ , for each  $\mathbf{x}^{S_1} \in C^{(1)}$  the class of distributions that contains  $f(\mathbf{x}^{S_2} | \mathbf{x}^{S_1}; \theta_2)$  and  $f(\mathbf{x}^{S_2} | \mathbf{x}^{S_1}; \theta_2^*)$  is identifiable. Thus,  $K_2 = K_2^*$  and  $\theta_2 = \theta_2^*$  with a probability equal to one (up to a permutation of the mixture components).

According to equation (7), the conditional expected value of  $\mathbf{X}^U$  given  $\mathbf{X}^{S_1}$  and  $\mathbf{X}^{S_2}$  is  $\boldsymbol{\mu}_{U|1,2} = \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{x}^{S_1} + \mathbf{A}_2 \mathbf{x}^{S_2}$ . Let

$$\begin{aligned} C^{(2)} &= \{(\mathbf{x}^{S_1}, \mathbf{x}^{S_2}) \in \mathbb{R}^{L_1+L_2} : \\ &\boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{x}^{S_1} + \mathbf{A}_2 \mathbf{x}^{S_2} = \boldsymbol{\alpha}_0^* + \mathbf{A}_1^* \mathbf{x}^{S_1} + \mathbf{A}_2^* \mathbf{x}^{S_2} \\ &\Rightarrow \boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_0^*, \mathbf{A}_1 = \mathbf{A}_1^*, \mathbf{A}_2 = \mathbf{A}_2^*\}. \end{aligned}$$

The set  $C^{(2)}$  contains all the vectors  $(\mathbf{x}^{S_1}, \mathbf{x}^{S_2})$  that can be used to distinct different values of  $(\boldsymbol{\alpha}_0, \mathbf{A}_1, \mathbf{A}_2)$  by different values of  $\boldsymbol{\mu}_{U|1,2}$ . This set is the complement of a  $(L_1 + L_2 - 1)$ -dimensional hyperplane of  $\mathbb{R}^{L_1+L_2}$ . According to equation (10), the joint marginal distribution of  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2})$  is a Gaussian mixture model with  $K_1 K_2$  components. Given assumptions (A1) and (A2), the

component-covariance matrices of this Gaussian mixture for  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2})$  are positive definite. Thus, according to such a mixture,  $\mathbb{P}(C^{(2)}) = 1$ .

Integrating both sides of the equality (40) with respect to  $\mathbf{X}^I$  and then conditioning on any  $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}) \in C^{(2)}$  yields  $f(\mathbf{x}^U | \mathbf{x}^{S_1}, \mathbf{x}^{S_2}; \boldsymbol{\theta}_U) = f(\mathbf{x}^U | \mathbf{x}^{S_1}, \mathbf{x}^{S_2}; \boldsymbol{\theta}_U^*) \forall \mathbf{x}^U \in \mathbb{R}^{L_U}$ , that is:

$$\begin{aligned} \phi_{L_U}(\mathbf{x}^U; \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{x}^{S_1} + \mathbf{A}_2 \mathbf{x}^{S_2}, \boldsymbol{\Sigma}_U) = \\ \phi_{L_U}(\mathbf{x}^U; \boldsymbol{\alpha}_0^* + \mathbf{A}_1^* \mathbf{x}^{S_1} + \mathbf{A}_2^* \mathbf{x}^{S_2}, \boldsymbol{\Sigma}_U^*) \quad \forall \mathbf{x}^U \in \mathbb{R}^{L_U}. \end{aligned}$$

Thus,  $\boldsymbol{\theta}_U = \boldsymbol{\theta}_U^*$  with a probability equal to one.

Finally, integrating both sides of the equality (40) with respect to  $\mathbf{X}^{S_1}, \mathbf{X}^{S_2}$  and  $\mathbf{X}^U$  leads to  $\phi_{L_I}(\mathbf{x}^I; \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I) = \phi_{L_I}(\mathbf{x}^I; \boldsymbol{\mu}_I^*, \boldsymbol{\Sigma}_I^*) \forall \mathbf{x}^I \in \mathbb{R}^{L_I}$ . From this result it follows that  $\boldsymbol{\theta}_I = \boldsymbol{\theta}_I^*$ . This completes the proof.

## References

- Anderson, T.: An Introduction to Multivariate Statistical Analysis, Third edn. Wiley, New York (2003)
- Andrews, J.L., McNicholas, P.D.: Variable selection for clustering and classification. *J. Classif.* 31, 136–153 (2014)
- Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49, 803–17821 (1993)
- Belitskaya-Levy, I.: A generalized clustering problem, with application to DNA microarrays. *Stat. Appl. Genet. Mol. Biol.* 5, Article 2 (2006)
- Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 719–725 (2000)
- Biernacki, C., Govaert, G.: Choosing models in model-based clustering and discriminant analysis. *J. Stat. Comput. Simul.* 64, 49–71 (1999)
- Bozdogan, H.: Intelligent statistical data mining with information complexity and genetic algorithms. In: Bozdogan, H. (ed.) *Statistical Data Mining and Knowledge Discovery*. Chapman & Hall/CRC, London, 15–56 (2004)
- Browne, R.P., ElSherbiny, A., McNicholas, P.D.: *mixture*: mixture models for clustering and classification. R package version 1.4 (2015)
- Brusco, M.J., Cradit, J.D.: A variable-selection heuristic for k-means clustering. *Psychometrika* 66, 249–270 (2001)
- Campbell, N.A., Mahon, R.J.: A multivariate study of variation in two species of rock crab of the genus *Leptograpsus*. *Aust. J. Zool.* 22, 417–425 (1974)
- Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. *Pattern Recognit.* 28, 781–793 (1995)
- Celeux, G., Martin-Magniette, M.-L., Maugis, C., Raftery, A.E.: Letter to the editor. *J. Am. Stat. Assoc.* 106, 383 (2011)
- Celeux, G., Martin-Magniette, M.-L., Maugis-Rabusseau, C., Raftery, A.E.: Comparing model selection and regularization approaches to variable selection in model-based clustering. *J. Soc. Fr. Statistique* 155, 57–71 (2014)
- Chatterjee, S., Laudato, M., Lynch, L.A.: Genetic algorithms and their statistical applications: an introduction. *Comput. Stat. Data Anal.* 22, 633–651 (1996)
- Dang, X.H., Bailey, J.: A framework to uncover multiple alternative clusterings. *Mach. Learn.* 98, 7–30 (2015)
- Dang, U.J., McNicholas, P.D.: Families of parsimonious finite mixtures of regression models. In: Morlini, I., Minerva, T., Vichi, M. (eds) *Statistical Models for Data Analysis*. Springer, Berlin, 73–84 (2015)
- 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)
- De Sarbo, W.S., Cron, W.L.: A maximum likelihood methodology for clusterwise linear regression. *J. Classif.* 5, 249–282 (1988)
- Dy, J.G., Brodley, C.E.: Feature selection for unsupervised learning. *J. Mach. Learn. Res.* 5, 845–889 (2004)
- Fowlkes, E.B., Gnanadesikan, R., Kettenring, J.R.: Variable selection in clustering. *J. Classif.* 5, 205–228 (1988)
- Fraiman, R., Justel, A., Svarc, M.: Selection of variables for cluster analysis and classification rules. *J. Am. Stat. Assoc.* 103, 1294–1303 (2008)
- 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)
- Friedman, J.H., Meulman, J.J.: Clustering objects on subsets of attributes (with discussion). *J. R. Stat. Soc. Ser. B* 66, 815–849 (2004)
- Frühwirth-Schnatter, S.: *Finite Mixture and Markov Switching Models*. Springer, New York (2006)
- Galimberti, G., Montanari, A., Viroli, C.: Penalized factor mixture analysis for variable selection in clustered data. *Comput. Stat. Data Anal.* 53, 4301–4310 (2009)
- Galimberti, G., Soffritti, G.: Model-based methods to identify multiple cluster structures in a data set. *Comput. Stat. Data Anal.* 52, 520–536 (2007)
- Galimberti, G., Soffritti, G.: Using conditional independence for parsimonious model-based Gaussian clustering. *Stat. Comput.* 23, 625–638 (2013)
- Galimberti, G., Scardovi, E., Soffritti, G.: Using mixtures in seemingly unrelated linear regression models with non-normal errors. *Stat. Comput.* doi:10.1007/s11222-015-9587-0 (2015)
- Gnanadesikan, R., Kettenring, J.R., Tsao, S.L.: Weighting and selection of variables for cluster analysis. *J. Classif.* 12, 113–136 (1995)
- Goldberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading (1989)
- Gordon, A.D.: *Classification*, Second edn. Chapman & Hall, Boca Raton (1999)
- Grün, B., Leisch, F.: Bootstrapping finite mixture models. In: Antoch, J. (ed) *Compstat 2004. Proceedings in Computational Statistics*. Physica-Verlag/Springer, Heidelberg, 1115–1122 (2004)
- Guo, J., Levina, E., Michailidis, G., Zhu, J.: Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* 66, 793–804 (2010)
- Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second edn. Springer, New York (2009)
- Hennig C.: Identifiability of models for clusterwise linear regression. *J. Classif.* 17, 273–296 (2000)
- Hubert, L., Arabie, P.: Comparing partitions. *J. Classif.* 2, 193–218 (1985)
- Kass, R.E., Raftery, A.E.: Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795 (1995)

- Keribin, C.: Consistent estimation of the order of mixture models. *Sankhyā Ser. A* 62, 49–66 (2000)
- Law, M.H.C., Figueiredo, M.A.T., Jain, A.K.: Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 1154–1166 (2004)
- Liu, T.-F., Zhang, N.L., Chen, P., Liu, A.H., Poon, L.K.M., Wang, Y.: Greedy learning of latent tree models for multi-dimensional clustering. *Mach. Learn.* 98, 301–330 (2015)
- Malsiner-Walli, G., Frühwirth-Schnatter, S., Grün, B.: Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* 26, 303–324 (2016)
- Maugis, C., Celeux, G., Martin-Magniette, M.-L.: Variable selection for clustering with Gaussian mixture models. *Biometrics* 65, 701–709 (2009a)
- 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 (2009b)
- McLachlan, G.J., Peel, D.: *Finite Mixture Models*. John Wiley & Sons, 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)
- McNicholas, P.D., Murphy, T.B., McDaid, A.F., Frost, D.: Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Comput. Stat. Data Anal.* 54, 711–723 (2010)
- Melnykov, V., Maitra, R.: Finite mixture models and model-based clustering. *Stat. Surv.* 4, 80–116 (2010)
- Montanari, A., Lizzani, L.: A projection pursuit approach to variable selection. *Comput. Stat. Data Anal.* 35, 463–473 (2001)
- Pan, W., Shen, X.: Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* 8, 1145–1164 (2007)
- Poon, L.K.M., Zhang, N.L., Liu, T.-F., Liu, A.H.: Model-based clustering of high-dimensional data: variable selection versus facet determination. *Int. J. Approx. Reason.* 54, 196–215 (2013)
- Quandt, R.E., Ramsey, J.B.: Estimating mixtures of normal distributions and switching regressions. *J. Am. Stat. Assoc.* 73, 730–738 (1978)
- Raftery, A.E., Dean, N.: Variable selection for model-based cluster analysis. *J. Am. Stat. Assoc.* 101, 168–178 (2006)
- R Core Team: *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org> (2015)
- Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* 6, 461–464 (1978)
- Scrucca, L.: **GA**: a package for genetic algorithms in R. *J. Stat. Softw.* 53(4) (2013)
- Scrucca, L.: Genetic algorithms for subset selection in model-based clustering. In: Celebi, M.E., Aydin, K. (eds) *Unsupervised Learning Algorithms*. Springer, Berlin, 55–1770 (2016)
- Scrucca, L., Raftery, A.E.: **clustvarsel**: a package implementing variable selection for model-based clustering in R. Pre-print available at <http://arxiv.org/abs/1411.0606> (2014)
- Scrucca, L., Raftery, A.E.: Improved initialisation of model-based clustering using Gaussian hierarchical partitions. *Adv. Data Anal. Classif.* 9, 447–17460 (2015)
- Soffritti, G.: Identifying multiple cluster structures in a data matrix. *Commun. Stat. Simul.* 32, 1151–1177 (2003)
- Soffritti, G., Galimberti, G.: Multivariate linear regression with non-normal errors: a solution based on mixture models. *Stat. Comput.* 21, 523–536 (2011)
- Srivastava, M.S.: *Methods of Multivariate Statistics*. John Wiley & Sons, New York (2002)
- Steinley, D., Brusco, M.J.: A new variable weighting and selection procedure for k-means cluster analysis. *Multivar. Behav. Res.* 43, 77–108 (2008a)
- Steinley, D., Brusco, M.J.: Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika* 73, 125–144 (2008b)
- Tadesse, M.G., Sha, N., Vannucci, M.: Bayesian variable selection in clustering high-dimensional data. *J. Am. Stat. Assoc.* 100, 602–617 (2005)
- Venables, W.N., Ripley, B.D.: *Modern Applied Statistics with S*, Fourth edn. Springer, New York (2002)
- Viroli, C.: Dimensionally reduced model-based clustering through mixtures of factor mixture analyzers. *J. Classif.* 31, 363–388 (2010)
- Wang, S., Zhu, J.: Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* 64, 440–448 (2008)
- Witten, D.M., Tibshirani, R.: A framework for feature selection in clustering. *J. Am. Stat. Assoc.* 105, 713–726 (2010)
- Xie, B., Pan, W., Shen, X.: Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics* 64, 921–930 (2008)
- Zeng, H., Cheung, Y.-M.: A new feature selection method for Gaussian mixture clustering. *Pattern Recognit.* 42, 243–250 (2009)
- Zhou, H., Pan, W., Shen, X.: Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.* 3, 1473–1496 (2009)
- Zhu, X., Melnykov, V.: Manly transformation in finite mixture modeling. *Comput. Stat. Data Anal.* doi:10.1016/j.csda.2016.01.015 (2016)