Statistics and Computing

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

## Supplementary material

**Giuliano Galimberti · Annamaria Manisi · Gabriele Soffritti**

### A Discovering two cluster structures through a greedy search

For the description of the greedy search algorithm employed in the analyses reported in Sections 4 and E it is convenient to use the following notation. A model for the marginal distribution of a variable sub-vector $\mathbf{X}^a$ obtained from equation (1) with $K_a$ mixture components and the parameterisation $P_a$ of the component-covariance matrices is denoted as $M = (\mathbf{X}^a, K_a, P_a)$; the $BIC_M$ of such a model is $BIC(\mathbf{X}^a, K_a, P_a)$. Furthermore, let $\mathcal{M}_1(\mathbf{X}^a)$ be the class of models for $\mathbf{X}^a$ defined using equation (1) with $K_1 \in \{2, \ldots, K_{1max}\}$ and by considering all the possible parameterisations for the component-covariance matrices. In an analogous way, a model for the conditional distribution of a variable sub-vector $\mathbf{X}^b$ given $\mathbf{X}^a$, defined according to equations (2)-(6) with $K_b$ mixture components and the parsimonious parameterisation $P_b$, is denoted as $M = (\mathbf{X}^b|\mathbf{X}^a, K_b, P_b)$; $BIC(\mathbf{X}^b|\mathbf{X}^a, K_b, P_b)$ is the $BIC$ value of such a model and $\mathcal{M}_2(\mathbf{X}^b|\mathbf{X}^a)$ is the class of the models for the conditional distribution of $\mathbf{X}^b$ given $\mathbf{X}^a$ obtained from equations (2)-(6) by admitting that $K_2 \in \{2, \ldots, K_{2max}\}$ and by considering all the parameterisations for the component-covariance matrices.

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

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

G. Soffritti (corresponding author)
Department of Statistical Sciences, University of Bologna
via delle Belle Arti 41 - 40126 Bologna, Italy
E-mail: gabriele.soffritti@unibo.it

The greedy search algorithm is composed of three parts. In the first part the selection of the variables that provide information on the first cluster structure (i.e., $\mathbf{X}^{S_1}$) is carried out; the second part identifies the relevant variables for the second cluster structure ($\mathbf{X}^{S_2}$); in the last part the uninformative variables ($\mathbf{X}^U$) are defined. The stages that characterise the first two parts are similar to the ones of the algorithm described in Raftery and Dean (2006) for performing variable selection. The main difference with the first part is given by the use of model (6), that makes it possible to discover a second cluster structure in the conditional p.d.f. of $\mathbf{X}^{S_2}$ given $\mathbf{X}^{S_1}$. As far as the second part is concerned, the most important difference with the algorithm of Raftery and Dean (2006) is that the analyses are carried out by conditioning all the examined models on the variables selected in the first part.

Specifically, the main stages of the greedy search algorithm are described in the following. A sequence of adding and removing steps characterises both the first and the second part.

*Part 1 Selection of the relevant variables for the first cluster structure*

*(1a)* (*First adding step*) In order to select the first relevant variable for the first cluster structure the evidence of clustering in the univariate marginal distributions of the $L$ variables has to be evaluated. This task is carried out through the following measure:

$$\Delta BIC(X_l) = BIC_{clust_1}(X_l) - BIC_{notclust}(X_l), \quad l \in \mathcal{I}, \tag{A}$$

where $\mathcal{I} = \{1, \ldots, L\}$ is the variable index set,

$$BIC_{clust_1}(X_l) = \max_{\mathcal{M}_1(X_l)} BIC(X_l, K, P)$$

and $BIC_{notclust}(X_l)$ is the $BIC$ of the univariate Gaussian model for the variable $X_l$. Thus, the first variable selected as informative for the first cluster structure is the one that has the highest value of the measure in equation (A):

$$\hat{X}_1^{(1)} = \underset{l \in \mathcal{I}}{\operatorname{argmax}}\, \Delta BIC(X_l).$$

Let $\hat{S}_1$ be the set composed of the integer number associated with $\hat{X}_1^{(1)}$.

*(1b)* (*Second adding step*) In order to choose the second relevant variable for the first cluster structure $L-1$ bivariate distributions have to be examined. Namely, the evidence of clustering in the joint bivariate distributions of $(X_l, \mathbf{X}^{\hat{S}_1})$, $l \in \mathcal{I} \setminus \hat{S}_1$, is evaluated through the following measure:

$$\begin{aligned}&\Delta BIC(X_l \cup \mathbf{X}^{\hat{S}_1}) = BIC_{clust_1}(X_l \cup \mathbf{X}^{\hat{S}_1})\\ &-BIC_{clust_2}(X_l|\mathbf{X}^{\hat{S}_1})\\ &-BIC_{clust_1}(\mathbf{X}^{\hat{S}_1}), \quad l \in \mathcal{I} \setminus \hat{S}_1, \end{aligned} \tag{B}$$

where

$$BIC_{clust_1}(X_l \cup \mathbf{X}^{\hat{S}_1}) = \max_{\mathcal{M}_1(X_l \cup \mathbf{X}^{\hat{S}_1})} BIC(X_l \cup \mathbf{X}^{\hat{S}_1}, K, P),$$

and

$$BIC_{clust_2}(X_l|\mathbf{X}^{\hat{S}_1}) = \max_{\mathcal{M}_2(X_l|\mathbf{X}^{\hat{S}_1})} BIC(X_l|\mathbf{X}^{\hat{S}_1}, K, P). \tag{C}$$

Then, the second variable selected as informative for the first cluster structure is the one that has the highest value of the measure in equation (B):

$$\hat{X}_2^{(1)} = \underset{l \in \mathcal{I} \setminus \hat{S}_1}{\operatorname{argmax}}\, \Delta BIC(X_l \cup \mathbf{X}^{\hat{S}_1}), \tag{D}$$

and $\hat{S}_1$ becomes the set composed of the variable indices associated with $\hat{X}_1^{(1)}$ and $\hat{X}_2^{(1)}$.

*(1c)* (*General adding step*) The next variable selected for the first cluster structure is the one that shows the strongest *positive* evidence of clustering in the joint multivariate distributions of $(X_l, \mathbf{X}^{\hat{S}_1})$, $l \in \mathcal{I} \setminus \hat{S}_1$, according to the measure $\Delta BIC(X_l \cup \mathbf{X}^{\hat{S}_1})$ defined in equation (B). Namely:

$$\hat{X}_q^{(1)} = \underset{l \in \mathcal{I} \setminus \hat{S}_1}{\operatorname{argmax}}\, \Delta BIC(X_l \cup \mathbf{X}^{\hat{S}_1});$$

furthermore, if $\Delta BIC(\hat{X}_q^{(1)} \cup \mathbf{X}^{\hat{S}_1}) > 0$, then $\hat{X}_q^{(1)}$ is selected as a relevant variable for the first cluster structure and the variable index associated with $\hat{X}_q^{(1)}$ is added to $\hat{S}_1$.

*(1d)* (*General removing step*) The next variable to be removed from the current vector $\mathbf{X}^{\hat{S}_1}$ of the relevant variables for the first cluster structure is the one that shows the weakest *negative* evidence of clustering in the joint p.d.f. of $\mathbf{X}^{\hat{S}_1}$. In order to identify this variable, the following measure is computed:

$$\begin{aligned}&\Delta BIC(\mathbf{X}^{\hat{S}_1 \setminus l}) = BIC_{clust_1}(\mathbf{X}^{\hat{S}_1})\\ &-BIC_{clust_2}(X_l|\mathbf{X}^{\hat{S}_1 \setminus l})\\ &-BIC_{clust_1}(\mathbf{X}^{\hat{S}_1 \setminus l}), \quad l \in \hat{S}_1. \end{aligned} \tag{E}$$

The variable that shows the weakest evidence of clustering is identified as the one that registers the lowest value of the measure in equation (E):

$$\hat{X}_r^{(1)} = \underset{l \in \hat{S}_1}{\operatorname{argmin}}\, \Delta BIC(\mathbf{X}^{\hat{S}_1 \setminus l}).$$

If $\Delta BIC(\mathbf{X}^{\hat{S}_1 \setminus s}) < 0$, where $s$ is the variable index associated with $\hat{X}_r^{(1)}$, then $\hat{X}_r^{(1)}$ is removed from the current vector of the relevant variables for the first cluster structure and $s$ is removed from $\hat{S}_1$.

*(1e)* Steps *(1c)* and *(1d)* are iterated until two consecutive steps do not lead to any change in the vector of the selected relevant variables for the first cluster structure.

*Part 2 Selection of the relevant variables for the second cluster structure*

*(2a)* (*First adding step*) In order to select the first relevant variable for the second cluster structure the evidence of clustering in the conditional univariate p.d.f. of $X_l$ given $\mathbf{X}^{\hat{S}_1}$ has to be evaluated for $l \in \mathcal{I} \setminus \hat{S}_1$. This task is carried out through the following measure:

$$\begin{aligned}&\Delta BIC(X_l|\mathbf{X}^{\hat{S}_1}) = BIC_{clust_2}(X_l|\mathbf{X}^{\hat{S}_1})\\ &-BIC_{notclust}(X_l|\mathbf{X}^{\hat{S}_1}), \quad l \in \mathcal{I} - \hat{S}_1, \end{aligned} \tag{F}$$

where $BIC_{clust_2}(X_l|\mathbf{X}^{\hat{S}_1})$ is defined in equation (C) and $BIC_{notclust}(X_l|\mathbf{X}^{\hat{S}_1})$ is the $BIC$ of the univariate Gaussian linear regression model with $X_l$ as the response and $\mathbf{X}^{\hat{S}_1}$ as predictors. Thus, the first variable selected as informative for the second cluster structure is the one that has the highest value of the measure in equation (F):

$$\hat{X}_1^{(2)} = \underset{l \in \mathcal{I} \setminus \hat{S}_1}{\operatorname{argmax}}\, \Delta BIC(X_l|\mathbf{X}^{\hat{S}_1}).$$

Then, let $\hat{S}_2$ be the set composed of the variable index associated with $\hat{X}_1^{(2)}$.

*(2b)* (*Second adding step*) The second relevant variable for the second cluster structure is selected as the one that, together with $\hat{X}_1^{(2)}$, shows the strongest evidence of clustering in their conditional bivariate p.d.f. given $\mathbf{X}^{\hat{S}_1}$, according to the following measure:

$$\begin{aligned}&\Delta BIC(X_l \cup \mathbf{X}^{\hat{S}_2}|\mathbf{X}^{\hat{S}_1}) = BIC_{clust_2}(X_l \cup \mathbf{X}^{\hat{S}_2}|\mathbf{X}^{\hat{S}_1})\\ &-BIC_{notclust}(X_l|\mathbf{X}^{\hat{S}_1 \cup \hat{S}_2})\\ &-BIC_{clust_2}(\mathbf{X}^{\hat{S}_2}|\mathbf{X}^{\hat{S}_1}), \quad l \in \mathcal{I} \setminus \hat{S}_1 \setminus \hat{S}_2. \qquad \text{(G)}\end{aligned}$$

Thus,

$$\hat{X}_2^{(2)} = \underset{l \in \mathcal{I} \setminus \hat{S}_1 \setminus \hat{S}_2}{\operatorname{argmax}} \Delta BIC(X_l \cup \mathbf{X}^{\hat{S}_2}|\mathbf{X}^{\hat{S}_1}).$$

Then, the variable index associated with $\hat{X}_2^{(2)}$ is added to $\hat{S}_2$.

*(2c)* (*General adding step*) The next relevant variable for the second cluster structure is the one that shows the strongest *positive* evidence of clustering in the joint conditional p.d.f.'s of $(X_l, \mathbf{X}^{\hat{S}_2})$ given $\mathbf{X}^{\hat{S}_1}$, $l \in \mathcal{I} \setminus \hat{S}_1 \setminus \hat{S}_2$, according to the measure defined in equation (G). Namely:

$$\hat{X}_q^{(2)} = \underset{l \in \mathcal{I} \setminus \hat{S}_1 \setminus \hat{S}_2}{\operatorname{argmax}} \Delta BIC(X_l \cup \mathbf{X}^{\hat{S}_2}|\mathbf{X}^{\hat{S}_1});$$

if $\Delta BIC(\hat{X}_q^{(2)} \cup \mathbf{X}^{\hat{S}_2}|\mathbf{X}^{\hat{S}_1}) > 0$, then $\hat{X}_q^{(2)}$ is selected as a relevant variable for the second cluster structure and its variable index is added to $\hat{S}_2$.

*(2d)* (*General removing step*) The next variable to be removed from the current vector $\mathbf{X}^{\hat{S}_2}$ of the relevant variables for the second cluster structure is the one that shows the weakest *negative* evidence of clustering in the joint conditional p.d.f. of $\mathbf{X}^{\hat{S}_2}$ given $\mathbf{X}^{\hat{S}_1}$. In order to identify this variable the following measure is computed:

$$\begin{aligned}&\Delta BIC(\mathbf{X}^{\hat{S}_2 \setminus l}|\mathbf{X}^{\hat{S}_1}) = BIC_{clust_2}(\mathbf{X}^{\hat{S}_2}|\mathbf{X}^{\hat{S}_1})\\ &-BIC_{notclust}(X_l|\mathbf{X}^{\hat{S}_1 \cup \hat{S}_2 \setminus l})\\ &-BIC_{clust_2}(\mathbf{X}^{\hat{S}_2 \setminus l}|\mathbf{X}^{\hat{S}_1}), \quad l \in \hat{S}_2, \qquad \text{(H)}\end{aligned}$$

where $BIC_{notclust}(X_l|\mathbf{X}^{\hat{S}_1 \cup \hat{S}_2 \setminus l})$ is the $BIC$ of the univariate Gaussian linear regression model with $X_l$ as the response and $\mathbf{X}^{\hat{S}_1 \cup \hat{S}_2 \setminus l}$ as predictors. The variable showing the weakest evidence of clustering is the one with the lowest value of the measure in equation (H):

$$\hat{X}_r^{(2)} = \underset{l \in \hat{S}_2}{\operatorname{argmin}}\, \Delta BIC(\mathbf{X}^{\hat{S}_2 \setminus l}|\mathbf{X}^{\hat{S}_1}).$$

If $\Delta BIC(\mathbf{X}^{\hat{S}_2 \setminus s}|\mathbf{X}^{\hat{S}_1}) < 0$, where $s$ is the variable index associated with $\hat{X}_r^{(2)}$, then $\hat{X}_r^{(2)}$ is removed from the vector of the relevant variables for the second cluster structure and $s$ is removed from $\hat{S}_2$.

*(2e)* Steps *(2c)* and *(2d)* are iterated until two consecutive steps do not lead to any change in the vector of the selected variables for the second cluster structure.

*Part 3 Definition of the uninformative variables* $\mathbf{X}^{\hat{U}} = \mathbf{X} \setminus \mathbf{X}^{\hat{S}_1} \cup \mathbf{X}^{\hat{S}_2}$.

At steps *(1a)*-*(1b)* and *(2a)*-*(2b)* the largest $\Delta BIC(.)$ value is not required to be positive because it may happen that the evidence of clustering in univariate and bivariate analyses can be weak, but it can become stronger when a multivariate analysis is performed (see Raftery and Dean (2006)). These steps simply allow to initialise the search for the vectors of relevant variables for the first and the second cluster structure, respectively. Removal steps *(1d)* and *(2d)* are introduced in order to overcome the well-known drawback of forward selection procedures as well as possible wrong initializations of $\mathbf{X}^{S_1}$ and $\mathbf{X}^{S_2}$. It is worth noting that the algorithm may stop after step *(1e)*. This happens when the evidence of one cluster structure in the observed data is strongest than the one in favour of two cluster structures. Thus, this algorithm can also be employed as a diagnostic tool to check the model assumptions (namely, the presence of two cluster structures in the dataset).

## B Two genetic algorithms for discovering two cluster structures

This section provides some details about the actual implementation of the genetic algorithms introduced in Section 3.8.

### B.1 Genetic algorithm 1

Recall that the generic model in the class $\tilde{\mathcal{M}}^{(2)}$ is denoted by $(S_1, S_2, U, K_1, K_2)$. The first genetic algorithm is organized into two parts:

*a*) selection of $S_1$ and $K_1$ (information extraction for the first cluster structure); for all models examined in this part $U = \emptyset$;

*b*) selection of $S_2$, $K_2$ and $U$ (information extraction for the second cluster structure and the uninformative variables), given the solution obtained in part *a*).

Part *a*) is structured as follows:

*a*.1) *Generation of the initial population*
$N_1$ chromosomes are randomly generated according to the genetic coding scheme described in Section 3.8. In particular:

- genes in position from one to $L$ are generated according to $L$ independent draws from a Bernoulli random variable with probability 0.5;
- gene in position $L+1$ is randomly selected from the set of the integer values $\{2, \ldots, K_{1max}\}$;
- gene in position $L+2$ is randomly selected from the set of the integer values $\{2, \ldots, K_{2max}\}$.

For example, if $L = 6$, the chromosome $(0, 1, 1, 0, 0, 1, 3, 2)$ corresponds to the model with $S_1 = \{2, 3, 6\}$, $S_2 = \{1, 4, 5\}$, $K_1 = 3$ and $K_2 = 2$.
,

*a*.2) *Fitness evaluation*
For each chromosome, the $BIC$ value of the corresponding model is used as a fitness measure. By default, the fitness measure for chromosomes corresponding to models with $S_1 = \emptyset$ (all genes in positions from 1 to $L$ are equal to zero) or $S_2 = \emptyset$ (all genes in positions from 1 to $L$ are equal to one) is considered a missing value and set to `NA`. Furthermore, if the maximum likelihood estimation for a model fails because of numerical issues (for example, due to singularities in some matrices that must be inverted in the EM algorithms), the fitness measure of the corresponding chromosome is also set to `NA`.
A preliminary check is performed on each chromosome, in order to establish whether it corresponds to a model that has been already fitted. If this holds true, the corresponding model is not refitted, since its $BIC$ is already available (this allows to save computational time).

*a*.3) *Generation of a new population*

*a*.3.i) *Selection*: after excluding chromosomes with `NA` fitness values, parent chromosomes for the new population are selected using a linear-rank method (Scrucca 2013). In particular, $\lfloor N_1/2 \rfloor$ pairs of parent chromosomes are randomly generated, where $\lfloor N_1/2 \rfloor$ denotes the largest integer lower than $N_1/2$.

*a*.3.ii) *Crossover*: given a pair of parent chromosomes, single point crossover is performed. In particular, the probability of crossover is set to 0.8. Pairs of parent chromosomes that are not subject to crossover are directly inserted in the new population. For pairs of parent chromosomes that are subject to crossover, a crossover point is selected at random among the values of the set $\{1, \ldots, L+2\}$.

*a*.3.iii) *Mutation*: the probability of a mutation is set to 0.1 for each chromosome obtained after crossover. The gene on which the mutation occurs is selected at random. The actual mutation depends on the position of the mutating gene:

- if the mutating gene is in position from 1 to $L$, the gene is changed from 0 to 1 or from 1 to 0;
- if the mutating gene is in position $L+1$, the mutation depends on the value of the gene. Denoting this value by $\tilde{k}$, the possible mutations are:
  - if $\tilde{k} = 2$, it is increased by one;
  - if $\tilde{k} = K_{1max}$, it is decreased by one;
  - otherwise, $\tilde{k}$ can be either increased or decreased by one. The direction of the mutation is randomly selected, with a probability of an increase equal to 0.5;
- if the mutating gene is in position $L+2$, the mutation scheme is similar to the one for gene in position $L+1$ (after replacing $K_{1max}$ with $K_{2max}$).

Steps *a*.2) and *a*.3) are iteratively repeated $d_{1max} - 1$ times, so that a total of $d_{1max}$ populations are examined. Among all examined chromosomes, the one with the largest fitness measure is selected. The optimal values $\hat{S}_1$ and $\hat{K}_1$ are derived from the corresponding model. Furthermore, let $\hat{L}_1$ denote the number of elements in $\hat{S}_1$ (the number of variables for the first cluster structure). These optimal values are considered as fixed in the second part of the genetic algorithm.

Part *b*) is structured as follows:

*b*.1) *Generation of the initial population*
$N_2$ chromosomes are randomly generated according to the genetic coding scheme described in Section 3.8. In particular:

- genes in position from 1 to $L - \hat{L}_1$ are generated according to $L - \hat{L}_1$ independent draws from a Bernoulli random variable with probability 0.5;
- gene in position $L - \hat{L}_1 + 1$ is randomly selected from the set of the integer values $\{2, \ldots, K_{2max}\}$;

For example, assuming that $L = 6$, $\hat{L}_1 = 2$, $\hat{S}_1 = \{2, 3\}$, and $\hat{K}_1 = 3$, the chromosome $(0, 0, 0, 1, 2)$ corresponds to the model with $S_1 = \{2, 3\}$, $S_2 = \{6\}$, $U = \{1, 4, 5\}$, $K_1 = 3$ and $K_2 = 2$.

*b*.2) *Fitness evaluation*
For each chromosome, the $BIC$ value for the corresponding model is used as a fitness measure. By default, the fitness measure for chromosomes corresponding to models with $S_2 = \emptyset$ (all genes in positions from 1 to $L - \hat{L}_1$ are equal to zero) is set to `NA`.

The fitness measure of a chromosome is also set to `NA` whenever the maximum likelihood estimation of the corresponding model fails because of numerical issues.
Furthermore, if a chromosome corresponds to a model already fitted, that model is not refitted.
Note that, given the particular structure of the log-likelihood functions of the proposed models (see Section 3.5), the ML estimates of $\boldsymbol{\theta}_1$ do not vary among the chromosomes examined in this part of the algorithm, and can be obtained from the results of part *a*). Thus, they must not be recomputed.

*b*.3) *Generation of a new population*

*b*.3.i) *Selection*: the selection process is similar to the one described in step *a*.3.i). The only difference is in the number of pairs of parent chromosomes, that changes to $\lfloor N_2/2 \rfloor$.

*b*.3.ii) *Crossover*: the crossover process is similar to the one described in step *a*.3.ii). The only difference is in the selection of the crossover point, that is randomly chosen among the values of the set $\left\{1, \ldots, L - \hat{L}_1 + 1\right\}$

*b*.3.iii) *Mutation*: the mutation process is similar to the one described in step *a*.3.iii). The only difference is that there is only one gene (in position $L - \hat{L}_1 + 1$) that can be increased or decreased by one.

Steps *b*.2) and *b*.3) are iteratively repeated $d_{2max} - 1$ times, so that a total of $d_{2max}$ populations are examined. Among all examined chromosomes, the one with the largest fitness measure is selected. The optimal values $\hat{S}_2$ and $\hat{K}_2$ and $\hat{U}$ are obtained from the corresponding model. These optimal values, along with the ones obtained with the execution of part *a*), define the optimal model $(\hat{S}_1, \hat{S}_2, \hat{U}, \hat{K}_1, \hat{K}_2)$.

## B.2 Genetic algorithm 2

This second algorithm extends the one just described by considering also the parsimonious covariance structures introduced in Section 3.6. Details about these parameterisations are provided in Table A, along with the genetic coding schemes that are used in the algorithm. Note that the fourteen parameterisations described is Table A are meaningfully defined when $L_g > 1$ ($g = 1, 2$). Whenever $L_g = 1$, the corresponding component-covariance matrices are scalar and some parameterisations lead to the same model. In particular, the fourteen parameterisations can be collapsed into only two different situations: constant component-variances or varying component-variances (see column 8 in Table A).

Recall that the generic model in the class $\tilde{\mathcal{M}}^{(2)}_{pars}$ is denoted by $(S_1, S_2, U, K_1, K_2, P_1, P_2, P_U)$. The second genetic algorithm is organized into two parts:

$\bar{a}$) selection of $S_1$, $K_1$ and $P_1$; for all models examined in this part $U = \emptyset$;

$\bar{b}$) selection of $S_2$, $K_2$, $P_2$, $U$ and $P_U$, given the solution obtained in part $\bar{a}$).

Part $\bar{a}$) is structured as follows:

$\bar{a}$.1) *Generation of the initial population*

$N_1$ chromosomes are randomly generated according to the genetic coding scheme described in Section 3.8. Chromosomes in this part have $L+4$ genes: the first $L+2$ genes have the same meaning of genes described in step *a*.1). The two additional genes are used to encode the parameterisations $P_1$ and $P_2$. Thus, for each chromosome:

- genes in position from 1 to $L + 2$ are randomly generated following the rules given in step *a*.1);
- gene in position $L + 3$ is randomly selected from the set of the integer values $\{1, \ldots, 14\}$;
- gene in position $L + 4$ is randomly selected from the set of the integer values $\{1, \ldots, 14\}$ .

For example, assuming that $L = 6$, the chromosome $(0, 1, 1, 0, 0, 1, 3, 2, 4, 12)$ correspond to the model with $S_1 = \{2, 3, 6\}$, $S_2 = \{1, 4, 5\}$, $K_1 = 3$, $K_2 = 2$, $P_1 =$ `VII` and $P_2 =$ `VVE`.
Before proceeding to the following steps, each chromosome is examined, in order to check whether it assigns only one variable to $S_1$ or to $S_2$. Consider, for example, the following chromosome: $(0, 0, 0, 0, 0, 1, 3, 2, 4, 12)$. The corresponding model has $S_1 = \{6\}$, $K_1 = 3$ and $P_1 =$ `VII`, thus implying a univariate Gaussian mixture for the marginal distribution of $X_6$, with varying component variances. Note that also the chromosome $(0, 0, 0, 0, 0, 1, 3, 2, 13, 12)$ leads to the same univariate Gaussian mixture for the marginal distribution of $X_6$. Thus, the two chromosomes encode the same model, although they are different. In order to avoid such inconsistencies, whenever in a chromosome only one of the first $L$ genes is equal to one (only one variable is assigned to $S_1$), gene in position $L + 3$ is modified according to the genetic coding scheme reported in the last column of Table A. A similar modification is performed on gene in position $L+4$ whenever only one of the first $L$ genes is equal to zero (only one variable is assigned to $S_2$).

$\bar{a}$.2) *Fitness evaluation*

This step is similar to step *a*.2).
The only difference is in the estimation of the component covariance matrices, that is carried out un-

**Table A** Parsimonious parameterisations for the component covariance matrices ($g = 1, 2$)

| Acronym | Model | Distribution | Volume | Shape | Orientation | Gene coding scheme | |
|---|---|---|---|---|---|---|---|
| | | | | | | $L_g > 1$ | $L_g = 1$ |
| EEE | $\lambda^{(g)}\mathbf{D}^{(g)}\mathbf{A}^{(g)}\mathbf{D}'^{(g)}$ | Ellipsoidal | Equal | Equal | Equal | 1 | 1 |
| VVV | $\lambda_{k_g}^{(g)}\mathbf{D}_{k_g}^{(g)}\mathbf{A}_{k_g}^{(g)}\mathbf{D}_{k_g}'^{(g)}$ | Ellipsoidal | Variable | Variable | Variable | 2 | 2 |
| EII | $\lambda^{(g)}\mathbf{I}_{L_g}$ | Spherical | Equal | Equal | – | 3 | 1 |
| VII | $\lambda_{k_g}^{(g)}\mathbf{I}_{L_g}$ | Spherical | Variable | Equal | – | 4 | 2 |
| EEI | $\lambda^{(g)}\mathbf{A}^{(g)}$ | Diagonal | Equal | Equal | – | 5 | 1 |
| VEI | $\lambda_{k_g}^{(g)}\mathbf{A}$ | Diagonal | Variable | Equal | – | 6 | 2 |
| EVI | $\lambda^{(g)}\mathbf{A}_{k_g}^{(g)}$ | Diagonal | Equal | Variable | – | 7 | 1 |
| VVI | $\lambda_{k_g}^{(g)}\mathbf{A}_{k_g}^{(g)}$ | Diagonal | Variable | Variable | – | 8 | 2 |
| EEV | $\lambda^{(g)}\mathbf{D}_{k_g}^{(g)}\mathbf{A}^{(g)}\mathbf{D}_{k_g}'^{(g)}$ | Ellipsoidal | Equal | Equal | Variable | 9 | 1 |
| VEV | $\lambda_{k_g}^{(g)}\mathbf{D}_{k_g}^{(g)}\mathbf{A}^{(g)}\mathbf{D}_{k_g}'^{(g)}$ | Ellipsoidal | Variable | Equal | Variable | 10 | 2 |
| EVE | $\lambda^{(g)}\mathbf{D}^{(g)}\mathbf{A}_{k_g}^{(g)}\mathbf{D}'^{(g)}$ | Ellipsoidal | Equal | Variable | Equal | 11 | 1 |
| VVE | $\lambda_{k_g}^{(g)}\mathbf{D}^{(g)}\mathbf{A}^{(g)}\mathbf{D}_{k_g}'^{(g)}$ | Ellipsoidal | Variable | Variable | Equal | 12 | 2 |
| VEE | $\lambda_{k_g}^{(g)}\mathbf{D}^{(g)}\mathbf{A}^{(g)}\mathbf{D}'^{(g)}$ | Ellipsoidal | Variable | Equal | Equal | 13 | 2 |
| EVV | $\lambda_{k_g}^{(g)}\mathbf{D}^{(g)}\mathbf{A}_{k_g}^{(g)}\mathbf{D}_{k_g}'^{(g)}$ | Ellipsoidal | Equal | Variable | Variable | 14 | 1 |

der the restrictions introduced by the parsimonious parameterisations.

$\bar{a}$.3) *Generation of a new population*

$\bar{a}$.3.i) *Selection*: see step *a*.3.i).

$\bar{a}$.3.ii) *Crossover*: this step is similar to step *a*.3.ii). Before proceeding to the following steps, new chromosomes obtained after crossover are examined and modified as described in step $\bar{a}$.1), in order to avoid the previously illustrated inconsistencies.

$\bar{a}$.3.iii) *Mutation*: this step is similar to step *a*.3.iii). The gene on which the mutation occurs is selected at random. The actual mutation depends on the position of the mutating gene:

- if the mutating gene is in position from 1 to $L + 2$, the same rules described in step *a*.3.iii) apply;
- if the mutating gene is in position $L + 3$, the mutation depends on the values of genes from 1 to $L$. In particular:
  - if only one of the first $L$ genes is equal to one (only one variable in $S_1$), the mutating gene is changed from 1 to 2 or from 2 to 1;
  - otherwise, the new value for the mutating gene is randomly selected among the set $\{1, \ldots, 14\} \setminus \tilde{k}$, where $\tilde{k}$ is the current value for that gene;
- if the mutating gene is in position $L + 4$, the mutation scheme is similar to the one for gene in position $L + 3$.

Before proceeding to the following steps, new chromosomes obtained after mutations in genes from 1 to $L$ are examined and modified in order to avoid the same inconsistencies described in step $\bar{a}$.1).

Steps $\bar{a}$.2) and $\bar{a}$.3) are iteratively repeated $d_{1max}-1$ times, so that a total of $d_{1max}$ populations are examined. Among all examined chromosomes, the one with the largest fitness measure is selected. The optimal values $\hat{S}_1$, $\hat{K}_1$ and $\hat{P}_1$ are derived from the corresponding model. Furthermore, let $\hat{L}_1$ denote the number of elements in $\hat{S}_1$ (the number of variables for the first cluster structure). These optimal values are considered as fixed in the second part of the algorithm.

Part $\bar{b}$) is structured as follows:

$\bar{b}$.1) *Generation of the initial population*

$N_2$ chromosomes are randomly generated according to the genetic coding scheme described in Section 3.8. Chromosomes in this part of the second genetic algorithm have $L - \hat{L}_1 + 3$ genes: the first $L - \hat{L}_1 + 1$ genes have the same meaning of genes described in step *b*.1). The two additional genes are used to encode the parameterisations $P_2$ and $P_U$. Thus, for each chromosome:

- genes in position from 1 to $L - \hat{L}_1 + 1$ are randomly generated following the rules given in step *b*.1);
- gene in position $L - \hat{L}_1 + 2$ is randomly selected from the set of the integer values $\{1, \ldots, 14\}$;
- gene in position $L - \hat{L}_1 + 2$ is randomly selected from the set of the integer values $\{1, 2, 3\}$. These

three values correspond to isotropic, diagonal or unconstrained covariance matrices, respectively.
For example, assuming that $L = 6$, $\hat{L}_1 = 2$, $\hat{S}_1 = \{2, 3\}$, $\hat{K}_1 = 3$ and $\hat{P}_1 =$`VEE`, the chromosome $(0, 0, 0, 1, 2, 1, 2)$ corresponds to the model with $S_1 = \{2, 3\}$, $S_2 = \{6\}$, $U = \{1, 4, 5\}$, $K_1 = 3$, $K_2 = 2$, $\hat{P}_1 =$`VEE`, $\hat{P}_2 =$`EEE` and diagonal $\boldsymbol{\Sigma}_U$.

Before proceeding to the following steps, each chromosome is examined, in order to avoid the same inconsistencies already described. In particular, whenever in a chromosome only one of the first $L$ genes is equal to one (only one variable is assigned to $S_2$), gene in position $L - \hat{L}_1 + 2$ is modified according to the genetic coding scheme reported in the last column of Table A. A similar modification is performed on gene in position $L - \hat{L}_1 + 3$ whenever only one of the first $L$ genes is equal to zero (only one variable is assigned to $U$). In this latter situation, gene in position $L - \hat{L}_1 + 3$ is set equal to 1.

$\bar{b}$.2) *Fitness evaluation*
For each chromosome, the $BIC$ value for the corresponding model parameter is used as a fitness measure. By default, the fitness measure for chromosomes corresponding to models with $S_2 = \emptyset$ (all genes in positions from 1 to $L - \hat{L}_1$ are equal to zero) is set to `NA`. Furthermore, if the maximum likelihood estimation for a model fails because of numerical issues, the fitness measure of the corresponding chromosome is also set to `NA`.
Similarly to part $b$) of the first genetic algorithm, given the particular structure of the log-likelihood functions of the proposed models (see Section 3.5), the ML estimates of $\boldsymbol{\theta}_1$ do not vary among chromosomes, and can be obtained from the results of part $\bar{a}$). Thus, they must not be recomputed.

$\bar{b}$.3) *Generation of a new population*
$\bar{b}$.3.i) *Selection*: see step $b$.3.i).

$\bar{b}$.3.ii) *Crossover*: this step is similar to step $b$.3.ii). Before proceeding to the following steps, new chromosomes obtained after crossover are examined and modified as described in step $\bar{b}$.1).

$\bar{b}$.3.iii) *Mutation*: this step is similar to step $b$.3.iii). The gene on which the mutation occurs is selected at random. The actual mutation depends on the position of the mutating gene:

- if the mutating gene is in position from 1 to $L - \hat{L}_1 + 1$, the same rules described in step $b$.3.iii) apply;
- if the mutating gene is in position $L - \hat{L}_1 + 2$, the mutation depends on the values of genes from 1 to $L - \hat{L}_1$. In particular:
  - if only one of the first $L - \hat{L}_1$ genes is equal to one (only one variable in $S_2$), the mutating gene is changed from 1 to 2 or from 2 to 1;
  - otherwise, the new value for the mutating gene is randomly selected among the 13 values of the set $\{1, \ldots, 14\}$ remaining after excluding the current value for that gene;
- if the mutating gene is in position $L - \hat{L}_1 + 3$, the mutation depends on the values of genes from 1 to $L - \hat{L}_1$. In particular:
  - if only one of the first $L - \hat{L}_1$ genes is equal to zero (only one variable in $U$), no mutation is allowed;
  - otherwise, the new value for the mutating ene is randomly selected among the two values of the set $\{1, 2, 3\}$ remaining after excluding the current value for that gene.

Again, before proceeding to the following steps, new chromosomes obtained after mutations in genes from 1 to $L - \hat{L}_1$ are examined and modified as described in step $\bar{b}$.1), so that to avoid the above illustrated inconsistencies.

Steps $\bar{b}$.2) and $\bar{b}$.3) are iteratively repeated $d_{2max} - 1$ times, so that a total of $d_{2max}$ populations are examined. Among all examined chromosomes, the one with the largest fitness measure is selected. The optimal values $\hat{S}_2$, $\hat{K}_2$, $\hat{P}_2$, $\hat{U}$ and $\hat{P}_U$ are derived from the corresponding model. These values, along with the ones obtained with the execution of part $\bar{a}$), define the optimal model $(\hat{S}_1, \hat{S}_2, \hat{U}, \hat{K}_1, \hat{K}_2, \hat{P}_1, \hat{P}_2, \hat{P}_U)$.

## C Second Monte Carlo study

In order to extend the performance evaluation of the first genetic algorithm (with the $BIC$ as a fitness measure), a second Monte Carlo experiment is performed. This second experiment considers a more challenging situation than the one examined in the first study. It is obtained by increasing the number of variables and by reducing the separation among clusters in both cluster structures. In particular, artificial datasets are generated in the Euclidean space $\mathbb{R}^{12}$ 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 $\mathbf{X}^U = (X_7, X_8, X_9, X_{10}, X_{11}, X_{12})$. Thus, in comparison with the first Monte Carlo experiment,

the increase in the total number of variables is only due to the presence of a greater number of uninformative variables.

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

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

$$\boldsymbol{\mu}_2^{(1)} = \begin{pmatrix} 3 \\ -3 \\ 3 \end{pmatrix}, \ \boldsymbol{\Sigma}_2^{(1)} = \begin{pmatrix} 1 & 0.6 & 0.3 \\ 0.6 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{pmatrix}.$$

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

$$\boldsymbol{\gamma}_1^{(2)} = \begin{pmatrix} -2 \\ -1 \\ 3.5 \end{pmatrix}, \ \boldsymbol{\Sigma}_1^{(2)} = \begin{pmatrix} 1 & 0.5 & 0.6 \\ 0.5 & 1 & 0.4 \\ 0.6 & 0.4 & 1 \end{pmatrix},$$

$$\boldsymbol{\gamma}_2^{(2)} = \begin{pmatrix} 0 \\ 1 \\ 1.5 \end{pmatrix}, \ \boldsymbol{\Sigma}_2^{(2)} = \begin{pmatrix} 1 & -0.5 & -0.6 \\ -0.5 & 1 & -0.4 \\ -0.6 & -0.4 & 1 \end{pmatrix},$$

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

These parameters coincide with the ones employed in the first Monte Carlo study except for $\boldsymbol{\mu}_2^{(1)}$ and $\boldsymbol{\gamma}_2^{(2)}$, whose values are modified such that the separation between clusters in every univariate variable subspace is reduced.

Finally, the parameters of the conditional p.d.f. of $\mathbf{X}^U$ given $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2})$ are:

$$\boldsymbol{\alpha}_0 = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{pmatrix}, \ \mathbf{A}_1 = \begin{pmatrix} 2 & 2 & 2 \\ -2 & -2 & -2 \\ 2 & 2 & 2 \\ -2 & -2 & -2 \\ 2 & 2 & 2 \\ -2 & -2 & -2 \end{pmatrix},$$

$$\mathbf{A}_2 = -\mathbf{A}_1, \ \boldsymbol{\Sigma}_U = 2.025 \cdot \mathbf{I}_6 + 0.225 \cdot \mathbf{1}_6\mathbf{1}_6',$$

where $\mathbf{I}_6$ is the identity matrix of order 6 and $\mathbf{1}_6$ is a unit-column vector of length 6. Thus, the detection of the two cluster structures is made more complex than in the first experiment because of a reduction in the evidence of clustering both in the p.d.f. of $\mathbf{X}^{S_1}$ and in the conditional p.d.f of $\mathbf{X}^{S_2}$ given $\mathbf{X}^{S_1}$. The complexity of the task is also increased by the fact that now the number of uninformative variables equals the total number of informative ones.

One hundred samples of $n = 400$ observations each are generated. The genetic algorithm is executed three times on each sample, by changing the values of the tuning parameter $N_1$ that controls the information extraction for the specification of model (1). Namely, the examined values are 120, 240 and 360. The other tuning parameters are kept constant throughout the experiment; they are set as follows: $K_{1max} = K_{2max} = 3$, $d_{1max} = 30$, $N_2 = 80$ and $d_{2max} = 20$.

The obtained results are summarized in Table B. Similarly to the first experiment, the percentage of samples for which $\mathbf{X}^{S_1}$, $\mathbf{X}^{S_2}$ and $\mathbf{X}^U$ are correctly identified tends to increase as the value of $N_1$ increases. However, since the increase in the number of variables (from 8 to 12) implies an enlargement of the model space, larger values of $N_1$ are needed in order to obtain a satisfactory performance with the genetic algorithm. It is interesting to note that, despite the errors in identifying the correct variable partition, there is a good agreement between true and estimated cluster structures also in this second experiment (see the mean values and standard deviations of the adjusted Rand index reported in Table B).

## D An example of an exhaustive search

The dataset used in this example is described and analysed in Ingrassia *et al.* (2014) and is available in the `R` package `flexCWM` (Mazza *et al.* 2015). This dataset reports $L = 3$ measurements for $n = 270$ students (151 females, 119 males) attending a statistics course: weight (WEIGHT, in kilograms) and height (HEIGHT, in centimeters) of the student and height of the student's father (HEIGHT.F, in centimeters).

As for the examples in Sections 2, 4.1 and E, this dataset is analysed using an unsupervised approach and ignoring the information about students' gender. In particular, the optimal model for this dataset is searched by comparing different kinds of models for the joint distribution of WEIGHT, HEIGHT and HEIGHT.F.

The first group of models is obtained from equation (8), by setting $K_1 = 1, 2, 3$ and $K_2 = 2, 3$. For each examined value of $K_1$ and $K_2$, also the parsimonious models illustrated in Section 3.6 are estimated. The Gaussian mixture models for clustering and regression described in Section 3.2 are included in the search. This leads to 888 distinct models: 768 are characterised by the presence of two cluster structures ($K_1 \geq 2$ and $K_2 \geq 2$), the remaining 120 models consider a single cluster structure defined on the conditional distribution of $\mathbf{X}^{S_2}$ given $\mathbf{X}^{S_1}$ ($K_1 = 1$ and $K_2 \geq 2$). Furthermore, 96 distinct models with only one cluster structure and at least one uninformative variable are considered. These latter models correspond to the solutions explored in the approach of Raftery and Dean (2006)

**Table B** Summary of results obtained in the second Monte Carlo Study.

| | $N_1$ | | |
|---|---|---|---|
| | 120 | 240 | 360 |
| Correct classification of all the variables | 50 | 70 | 78 |
| aRi for the first cluster structure: | | | |
| mean | 0.986 | 0.986 | 0.988 |
| s.d. | 0.012 | 0.012 | 0.010 |
| aRi for the second structure: | | | |
| mean | 0.840 | 0.843 | 0.843 |
| s.d. | 0.042 | 0.043 | 0.040 |

to perform variable selection in model-based clustering. They are obtained by setting $\mathbf{X}^{S_2} = \emptyset$, $K_1 = 2, 3$ and by examining all the possible parameterisations for the component covariance matrices. Gaussian mixture models with only one cluster structure on the joint distribution of the three observed variables (i.e., without uninformative variables) are also examined. These models are derived by considering $\mathbf{X}^{S_2} = \mathbf{X}^{U} = \emptyset$ and $K_1 = 2, 3$. For each of these values of $K_1$, all the 14 parsimonious parameterisations have been estimated, thus leading to 28 models. Finally, also the Gaussian model for the joint distribution of the observed variable is included in the analysis ($K_1 = 1$, $\mathbf{X}^{S_2} = \mathbf{X}^{U} = \emptyset$). The total number of examined model is 1013.

According to the $BIC$, if one restricts the attention to models with $\mathbf{X}^{S_2} = \emptyset$, the best model for the joint distribution of the three measurements is obtained with the following splitting of the variable vector: $\mathbf{X}^{S_1}$ = (HEIGHT, HEIGHT.F), $\mathbf{X}^{U}$ = WEIGHT. The same splitting is obtained using `clustvarsel`. The BIC value of this model is $-5347.7$. A mixture of two Gaussian components with equal covariance matrices is selected for modelling the joint distribution of students' and fathers' height.

By allowing $\mathbf{X}^{S_2} \neq \emptyset$ (such as in equation (8)), the model with the largest $BIC$ value for the joint distribution of the three measurements has the following splitting: $\mathbf{X}^{S_1}$ = HEIGHT.F, $\mathbf{X}^{S_2}$ = HEIGHT, $\mathbf{X}^{U}$ = WEIGHT. The BIC value of this model is $-5342.7$. As far as the model for the marginal distribution of the fathers' height is concerned, a Gaussian distribution is selected. Thus, this marginal distribution does not provide any evidence of clustering of the students (note that this information cannot be directly recovered from the results obtained when setting $\mathbf{X}^{S_2} = \emptyset$). The $BIC$ value of this marginal model is $-1739.1$. A mixture of two Gaussian regression models with the same regression coefficients and unconstrained variances is used to model the linear dependence of the students' height on the height of their fathers (see Figure A). The $BIC$ value of this conditional model is $-1850.6$. This model allows to detect a latent clustering of the students that is strongly associated with their grouping based on gender (see Table C). As already noted, HEIGHT.F does not provide information about this cluster structure. This is consistent with the fact that the gender of a student is expected to be independent from his/her father's height. Finally, a Gaussian linear regression model is employed for the conditional distribution of the students' weight in which both heights are used as predictors ($BIC = -1753.0$). According to this latter result, conditionally on HEIGHT.F and HEIGHT, the students' weight does not provide any information about the clustering of the students.

**Table C** Classification of the students according to their gender and the cluster membership estimated by the Gaussian mixture of regression models selected for the conditional distribution of HEIGHT|HEIGHT.F.

| | Cluster | |
|---|---|---|
| Gender | 1 | 2 |
| F | 151 | 0 |
| M | 7 | 112 |
| aRi | 0.899 | |

The results obtained using models defined according to equation (8) are consistent with the ones described in Ingrassia *et al.* (2014). It is worth noting that Ingrassia *et al.* (2014) analyse this dataset using an approach that differs from the one proposed in this paper. Namely, Ingrassia *et al.* (2014) focus on two distinct bivariate analyses (one for the joint distribution of HEIGHT.F and HEIGHT, the other for the joint distribution of WEIGHT and HEIGHT) instead of considering the joint distribution of the three measurements. Furthermore, in each of these two bivariate analysis, they follow a regression approach, by specifying in advance which variable plays the role of dependent variable and using the other variable as a regressor. These *a priori* distinctions in the roles of the variables are not required when using the approach proposed in this paper.

## E Results from the analysis of the AIS dataset

The AIS dataset is described in Cook and Weisberg (1994) and is available in the `R` package `sn` (Azzalini

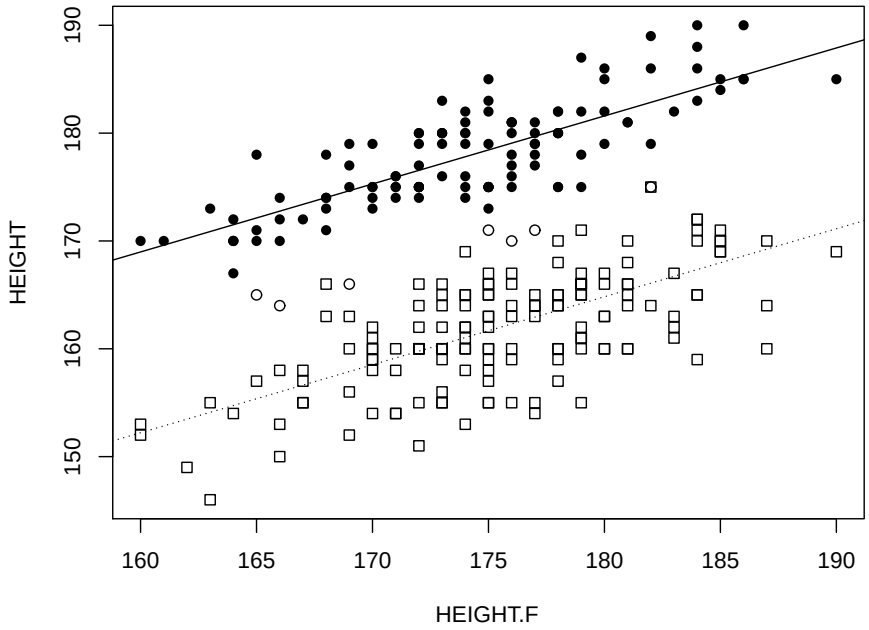


**Fig. A** Scatterplot of the students' and fathers' heights and fitted regression lines using a mixture of two Gaussian linear regression models with the same regression coefficients. Male and female students are represented using circles and squares, respectively. White and black colours are used to distinguish between the two clusters detected by the mixture model.

2014). It contains information concerning $n = 202$ athletes (102 males and 100 females) at the Australian Institute of Sport. The analysis described in this Section focuses on $L = 9$ variables: red cell count (RCC), white cell count (WCC), hematocrit (Hc), hemoglobin (Hg), plasma ferritin concentration (Fe), body mass index (BMI), sum of skin folds (SSF), body fat percentage (Bfat) and lean body mass (LBM). This dataset is analysed using `mclust`, `clustvarsel`, `SelvarClust`, `SelvarClustIndep`, the greedy search algorithm (see Section A) and the second genetic algorithm (Section B.2).

The best Gaussian mixture model resulting from the analysis performed through `mclust` (with a maximum number of components equal to five) is a mixture of three Gaussian ellipsoidal components with the same orientation. This result is obtained with the `mclust` option for the initialisation of the EM algorithm that transforms the variables using a singular value decomposition. The value of $BIC$ for such a mixture model is $-9028.2$. The clustering obtained from this model reproduces quite well the two classes of athletes based on their gender (see Table D, left part).

The three examined variable selection methods lead to different decisions about the variables that provide relevant information on the clustering of the athletes (see Table E). Using `clustvarsel` (with a maximum number of components for the p.d.f. of the informative variables equal to five), only the biometrical variables are selected. The best Gaussian mixture model fitted to the p.d.f. of these variables is a mixture of three Gaussian ellipsoidal components with the same shape. The $BIC$ value of the resulting joint model for the nine variables is $-9008.1$. Thus, according to the $BIC$, this joint model is better than the best model detected without variable selection. However, the partition of the athletes resulting from the best mixture model for the biometrical variables shows a slightly lower agreement with the partition based on gender (see Table D).

In addition to the four biometrical variables, softwares `SelvarClust` and `SelvarClustIndep` also select two blood composition variables: one is plasma ferritin concentration and the other is hemoglobin or hematocrit. Namely, the best joint model obtained after three independent executions of `SelvarClust` is given by the product of a Gaussian mixture model with three equally-oriented components for the joint marginal distribution of BMI, SSF, Bfat, LBM, Fe and Hg, and a Gaussian linear regression model for the conditional distribution of the remaining variables in which only Hg is used as a regressor and the covariance matrix is unconstrained. The $BIC$ value of the joint model obtained in this way is $-8935.3$. Using `SelvarClustIndep` the best model is

**Table D** Classification of the athletes according to their gender and the cluster membership estimated by the models selected using `mclust`, `clustvarsel`, `SelvarClust` and `SelvarClustIndep`.

| | `mclust` | | | `clustvarsel` | | | `SelvarClust` | | | `SelvarClustIndep` | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cluster | | | Cluster | | | Cluster | | | Cluster | | |
| Gender | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| F | 97 | 2 | 1 | 39 | 61 | 0 | 2 | 1 | 97 | 2 | 97 | 1 |
| M | 2 | 40 | 60 | 13 | 1 | 88 | 25 | 75 | 2 | 23 | 2 | 77 |
| aRi | 0.682 | | | 0.586 | | | 0.735 | | | 0.745 | | |

composed of a mixture of three Gaussian components with the same orientation for the joint marginal distribution of BMI, SSF, Bfat, LBM, Fe and Hc, and a Gaussian linear regression model for the conditional distribution of the remaining variables in which the selected regressors are Hc, BMI and Bfat and the covariance matrix is diagonal. None of the uninformative variables results to be independent of all the informative ones. Overall, this joint model registers a $BIC$ of $-8934.5$. As far as the recovery of the classification based on gender is concerned, the partitions of the athletes resulting from the mixture models for the variables selected by `SelvarClust` and `SelvarClustIndep` reach a very similar performance, that is better than the ones obtained using both `mclust` and `clustvarsel` (see Table D).

**Table E** Variables selected by the packages `clustvarsel`, `SelvarClust` and `SelvarClustIndep` from the AIS dataset.

| Package | Selected variables |
|---|---|
| `clustvarsel` | BMI, SSF, Bfat, LBM |
| `SelvarClust` | BMI, SSF, Bfat, LBM, Fe, Hg |
| `SelvarClustIndep` | BMI, SSF, Bfat, LBM, Fe, Hc |

The splitting of the nine measurements obtained using the greedy search algorithm (with $K_{1max} = K_{2max} = 4$) is $\mathbf{X}^{\hat{S}_1}$ = (BMI, SSF, Bfat, LBM), $\mathbf{X}^{\hat{S}_2}$ = (RCC, Fe), $\mathbf{X}^{\hat{U}}$=(WCC, Hc, Hg). The CPU time requires by this algorithm is 9 hours and 35 minutes. The model selected for the joint p.d.f. of the variable sub-vector (BMI, SSF, Bfat, LBM) coincides with the one detected by `clustvarsel`. Thus, the first cluster structure discovered through the greedy search coincides with the clustering of the athletes obtained from the variable selection methods implemented in `clustvarsel` (see Table D). As far as the conditional p.d.f. of the variable sub-vector (RCC, Fe) given (BMI, SSF, Bfat, LBM) is concerned, a mixture of three Gaussian ellipsoidal components with the same shape and orientation is selected. The second cluster structure detected from this conditional model is not associated with the athletes' gender (see Table F). Finally, a Gaussian linear regression model with unconstrained covariance matrix is selected for the conditional distribution of (WCC, Hc, Hg) given all the other measurements. The $BIC$ value of the joint model for the nine variables is $-8950.22$. Thus, this model is better than the model selected by `clustvarsel` but worse than the models selected through `SelvarClust` and `SelvarClustIndep`.

**Table F** Second cluster structure detected by the greedy search algorithm and its association with the classification of the athletes based on gender.

| | Structure 2 | | |
|---|---|---|---|
| | Cluster | | |
| Gender | 1 | 2 | 3 |
| F | 90 | 8 | 2 |
| M | 77 | 21 | 4 |
| aRi | 0.014 | | |

**Table G** Cluster structures detected by the genetic algorithm and their association with the classification of the athletes based on gender.

| | Structure 1 | | | Structure 2 | | |
|---|---|---|---|---|---|---|
| | Cluster | | | Cluster | | |
| Gender | 1 | 2 | 3 | 1 | 2 | 3 |
| F | 98 | 1 | 1 | 50 | 49 | 1 |
| M | 1 | 74 | 27 | 54 | 34 | 14 |
| aRi | 0.754 | | | 0.015 | | |

Nine independent executions of the second genetic algorithm are performed, one for each combination of the following values for the tuning parameters: $N_1 = 300, 500, 700$, $d_{1max} = 30, 50, 70$. The remaining tuning parameters are set as follows: $N_1 = N_2$, $d_{1max} = d_{2max}$ and $K_{1max} = K_{2max} = 4$. A tenth execution is carried out with $N_1 = 100$ and $d_{1max} = 30$. Using this setting the CPU time of the analysis is 8 hours and 39 minutes. The model selected by the algorithm in this latter execution coincides with the best overall model, that is selected in other five executions. According to this model, a first cluster structure is defined in the sub-vector $\mathbf{X}^{\hat{S}_1}$ = (BMI, SSF, Bfat, LBM, Hg). The best model for the p.d.f. of this sub-vector is a mixture of three Gaussian ellipsoidal components with the same orientation. The recovery of males and females classes obtained using the segmentation of the athletes based on this model is slightly improved over the previous models (see Table G, left part). The second cluster structure is found in the conditional distribution of the sub-vector $\mathbf{X}^{\hat{S}_2}$ = (Hc, Fe) given $\mathbf{X}^{\hat{S}_1}$, resulting from

a mixture of three Gaussian components with diagonal covariance matrices having the same volume. The partition of the athletes obtained from this second mixture model is not associated with the athletes' gender (see Table G, right part). Since in model (8) the latent variables $Z_1$ and $Z_2$ are assumed to be independent, this latter result is not surprising. Thus, the second structure is reasonably associated with other (unobserved) factors independent of the gender. Finally, red and white cell counts compose $\mathbf{X}^{\hat{U}}$ and, thus, result to be uninformative variables. Their conditional distribution is modelled using a Gaussian linear regression model with a diagonal regression covariance matrix. The $BIC$ value of the joint model for the nine variables is $-8933.2$.

An improvement of the linear regression mixture model with three components selected for (Hc, Fe) is obtained after performing a regressors selection through an exhaustive search, given the splitting of the variables detected by the genetic algorithm. This task is carried out by allowing each dependent variable to have its own specific set of regressors (see equation (18)). Furthermore, all fourteen parameterisations are estimated for each examined model. According to the $BIC$, the best solution is obtained using a model for the linear dependence of (Hc, Fe) on (BMI, SSF, Bfat, LBM, Hg) in which haematocrit is regressed on hemoglobin, sum of skin folds and body fat percentage, while the selected predictors for plasma ferritin concentration are hemoglobin, body mass index, body fat percentage and lean body mass. The component-covariance matrices of this model are unconstrained. In a similar way, the best Gaussian linear regression model for (RCC, WCC) obtained after performing regressors selection is the one that has haematocrit as a predictor for both dependent variables and sum of skin folds only for WCC; furthermore, the covariance matrix in this model is diagonal. The joint model for the nine variables obtained in this way has a $BIC$ value of $-8856.9$; thus, it provides a description of the relevant information contained in the AIS dataset which is better than the previously illustrated models.

## F 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, $\boldsymbol{\theta}_M$ is composed of four non empty subvectors. 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 $\boldsymbol{\theta}_M$ and $\boldsymbol{\theta}^*_{M^*}$ be such that

$$f(\mathbf{x};\boldsymbol{\theta}_M) = f(\mathbf{x};\boldsymbol{\theta}^*_{M^*}) \ \forall \ \mathbf{x} \in \mathbb{R}^L. \tag{I}$$

In the following it is shown that the equality (I) implies that $M = M^*$ and $\boldsymbol{\theta}_M = \boldsymbol{\theta}^*_{M^*}$. 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 $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^*$; the second part proves that $K_2 = K_2^*$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^*$; finally, the last two parts demonstrate that $\boldsymbol{\theta}_U = \boldsymbol{\theta}_U^*$ and $\boldsymbol{\theta}_I = \boldsymbol{\theta}_I^*$, respectively.

According to equations (14) and (1), integrating each side of the equality (I) with respect to $\mathbf{X}^{S_2}$, $\mathbf{X}^U$ and $\mathbf{X}^I$ yields $f\left(\mathbf{x}^{S_1};\boldsymbol{\theta}_1\right) = f\left(\mathbf{x}^{S_1};\boldsymbol{\theta}_1^*\right) \ \forall \ \mathbf{x}^{S_1} \in \mathbb{R}^{L_1}$, that is:

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

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

Given the constraints *(I1)* on $\boldsymbol{\theta}_1$, the class of distributions that contains $f\left(\mathbf{x}^{S_1};\boldsymbol{\theta}_1\right)$ and $f\left(\mathbf{x}^{S_1};\boldsymbol{\theta}_1^*\right)$ is identifiable. Thus, $K_1 = K_1^*$ and $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^*$ (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

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

Let

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

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 (I) with respect to $\mathbf{X}^U$ and $\mathbf{X}^I$ and then conditioning on any $\mathbf{x}^{S_1} \in C^{(1)}$ leads to $f\left(\mathbf{x}^{S_2}|\mathbf{x}^{S_1};\boldsymbol{\theta}_2\right) = f\left(\mathbf{x}^{S_2}|\mathbf{x}^{S_1};\boldsymbol{\theta}_2^*\right) \ \forall \, \mathbf{x}^{S_2} \in \mathbb{R}^{L_2}$, that is:

$$\sum_{k_2=1}^{K_2} \pi_{k_2}^{(2)} \phi_{L_2}\left(\mathbf{x}^{S_2};\boldsymbol{\gamma}_{k_2}^{(2)} + \mathbf{B}_{21}\mathbf{x}^{S_1},\boldsymbol{\Sigma}_{k_2}^{(2)}\right) =$$

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

Given the constraints *(I2)* on $\boldsymbol{\theta}_2$, for each $\mathbf{x}^{S_1} \in C^{(1)}$ the class of distributions that contains $f\left(\mathbf{x}^{S_2}|\mathbf{x}^{S_1};\boldsymbol{\theta}_2\right)$ and $f\left(\mathbf{x}^{S_2}|\mathbf{x}^{S_1};\boldsymbol{\theta}_2^*\right)$ is identifiable. Thus, $K_2 = K_2^*$ and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^*$ 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

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

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 $K_1K_2$ 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 (I) with respect to $\mathbf{X}^I$ and then conditioning on any $(\mathbf{X}^{S_1}, \mathbf{X}^{S_2}) \in C^{(2)}$ yields $f\left(\mathbf{x}^U|\mathbf{x}^{S_1}, \mathbf{x}^{S_2};\boldsymbol{\theta}_U\right) = f\left(\mathbf{x}^U|\mathbf{x}^{S_1}, \mathbf{x}^{S_2};\boldsymbol{\theta}_U^*\right)$ $\forall\ \mathbf{x}^U \in \mathbb{R}^{L_U}$, that is:

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

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

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

## References


Azzalini, A.: The `R` package `sn`: the skew-normal and skew-*t* distributions (version 1.1-2). `URL` `http://azzalini.stat.unipd.it/SN` (2014)

Cook, R.D., Weisberg, S.: An Introduction to Regression Graphics. Wiley, New York (1994)

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

Ingrassia, S., Minotti, S.C., Punzo, A.: Model-based clustering via linear cluster-weighted models. Comput. Stat. Data Anal. 71, 159–182 (2014)

Mazza, A., Punzo, A, Ingrassia, S.: `flexCWM`: Flexible Cluster-Weighted Modeling. `R` package version 1.5 (2015)

Raftery, A.E., Dean, N.: Variable selection for model-based cluster analysis. J. Am. Stat. Assoc. 101, 168–178 (2006)

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