Variable selection in model-based clustering and discriminant analysis with a regularization approach

Gilles Celeux, Cathy Maugis-Rabusseau, Mohammed Sedki

Advances in Data Analysis and Classification manuscript No.
(will be inserted by the editor)

Variable selection in model-based clustering and discriminant analysis with a regularization approach

Gilles Celeux · Cathy Maugis-Rabusseau · Mohammed Sedki

Received: date / Accepted: date

Gilles Celeux
Inria and Université Paris-Sud
Dept. de mathématiques
Btiment 425
91405 Orsay Cedex, France
E-mail: gilles.celeux@inria.fr

Cathy Maugis-Rabusseau
Institut de Mathématiques de Toulouse; UMR 5219
Université de Toulouse
INSA de Toulouse
135 avenue de Rangueil
31077 Toulouse, Cedex 4, France
E-mail: cathy.maugis@insa-toulouse.fr

Mohammed Sedki
INSERM U1181 B2PHI, Institut Pasteur and UVSQ
Bâtiment. 15/16 Inserm, Hôpital Paul Brousse
16 avenue Paul Vaillant Couturier
94807 Villejuif Cedex, France
E-mail: mohammed.sedki@u-psud.fr

Abstract Several methods for variable selection have been proposed in model-based clustering and classification. These make use of backward or forward procedures to define the roles of the variables. Unfortunately, such stepwise procedures are slow and the resulting algorithms inefficient when analyzing large data sets with many variables. In this paper, we propose an alternative regularization approach for variable selection in model-based clustering and classification. In our approach the variables are first ranked using a lasso-like procedure in order to avoid slow stepwise algorithms. Thus, the variable selection methodology of Maugis et al (2009b) can be efficiently applied to high-dimensional data sets.

Keywords Variable Selection · Lasso · Gaussian Mixture · Clustering · Classification

1 Introduction

In data mining and statistical learning, available datasets are larger and larger. As a result, there is more and more interest in variable selection procedures for clustering and classification tasks. See, among others, the contributions of Law et al (2004); Tadesse et al (2005); Raftery and Dean (2006); Fraiman et al (2008); Maugis et al (2009a,b); Nia and Davison (2012); Kim et al (2012); Lee and Li (2012). There are also several other recent proposals for variable selection through regularization, as for instance Sun et al (2012); Bouveyron and Brunet (2014); Galimberti et al (2009).

After a series of papers on variable selection in model-based clustering (Law et al, 2004; Tadesse et al, 2005; Raftery and Dean, 2006; Maugis et al, 2009a), Maugis et al (2009b) proposed a general model for selecting variables for clustering with Gaussian mixtures. This model, called SRUW, distinguishes between relevant variables (S) and irrelevant variables (S^c) for clustering. In addition, the irrelevant variables are divided into two categories. A part of the irrelevant variables (U) may be dependent on a subset R of the relevant variables, and another part (W) are independent of the other variables. In Maugis et al (2009b), a procedure using embedded stepwise variable selection algorithms is used to identify the SRUW sets. It compares two models at each step in order to determine which variable should be excluded or included in the sets S, R, U and W. However, these stepwise procedures, implemented in *SelvarClustIndep*[1], do not work so well as soon as the number of variables is a few dozen or more. The SRUW model was also modified to work for Gaussian model-based classification in Maugis et al (2011). In this supervised framework, the identification of the sets S, R, U and W is simpler since the stepwise procedures are not performed inside an EM algorithm, but the stepwise algorithms still encounter combinatorial problems. Note that Raftery and Dean's method was adapted to the semi-supervised setting in Murphy et al (2010).

[1] *SelvarClustIndep* is implemented in C++ and is available at `http://www.math.univ-toulouse.fr/~maugis/`

In parallel, Pan and Shen (2007) were inspired by the success of lasso regression to develop a method of variable selection in model-based clustering using ℓ_1 regularization of the likelihood. This approach was successively extended in Xie et al (2008); Wang and Zhu (2008), and following this, Zhou et al (2009) proposed a regularized Gaussian mixture model with unconstrained covariance matrices.

The general methodology proposed in Raftery and Dean (2006), and improved in Maugis et al (2009a) and Maugis et al (2009b) in the Gaussian model-based clustering framework, has been proven to lead to parsimonious and realistic models by allowing different possible roles for each variable. It has been proven to be efficient in many situations (Celeux et al, 2014). However, it is slow as soon as the data set has several dozen variables. In order to overcome this, we propose in the present paper a variant of this methodology in which the variables are first ranked using an ℓ_1 penalty placed on the Gaussian mixture components' mean vectors and precision matrices.

This is made feasible by exploiting the abundant literature on lasso penalization in Gaussian graphical models (see Friedman et al, 2007; Meinshausen and Bühlmann, 2006). Using the resulting ranking of the variables can avoid combinatorial problems in stepwise variable selection algorithms. Also, it is hoped that using this lasso-like ranking of the variables instead of stepwise algorithms does not deteriorate identification of the sets S, R, U and W.

The article is organized as follows. In Section 2, the SRUW model is reviewed in the Gaussian model-based clustering context, and its simple extension to the Gaussian model-based classification context is sketched out. The variable selection procedure using lasso-like penalization is presented in Section 3, focusing on model-based clustering. Section 4 is devoted to the comparison of *SelvarClustIndep* and *SelvarMix*[2] procedures on several simulated and real datasets, as well as with other variable selection procedures. A short discussion section ends the article.

2 The SRUW model

2.1 Model-based clustering

Let n observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ each be made up of p continuous variables ($\mathbf{y}_i \in \mathbb{R}^p$). In the model-based clustering framework, the multivariate continuous data $\mathbf{y}$ are assumed to come from several subpopulations (clusters), each modeled by a multivariate Gaussian density. The observations are assumed to arise from a finite Gaussian mixture with K components and a mixture form m (explained below):

$$f\big(\mathbf{y}_i \mid K, m, \alpha\big) = \sum_{k=1}^{K} \pi_k \phi\big(\mathbf{y}_i \mid \mu_k, \Sigma_{k(m)}\big),$$

[2] *SelvarMix* R package is available at `https://CRAN.R-project.org/package=SelvarMix`

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ is the mixing proportion vector ($\pi_k \in (0,1)$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$), $\phi\big(\cdot \mid \mu_k, \Sigma_k\big)$ is the p-dimensional Gaussian density function with mean μ_k and covariance matrix Σ_k, and $\alpha = \big(\boldsymbol{\pi}, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K\big)$ the parameter vector. Several Gaussian mixture forms m are available, each corresponding to different assumptions on the forms of the covariance matrices, arising from modified spectral decompositions. These include whether the volume, shape and orientation of each mixture component vary between components or are constant across clusters (Banfield and Raftery, 1993; Celeux and Govaert, 1995).

Typically, mixture parameters are estimated via maximum likelihood using the EM algorithm (Dempster et al, 1977), and both the number of components K and the mixture form m are chosen using the Bayesian Information Criterion (BIC) (Schwarz, 1978) or other penalized likelihood criteria as the Integrated Completed Likelihood (ICL) criterion (Biernacki et al, 2000) in the clustering context. R packages which implement this methodology include the *mclust* (Scrucca et al, 2016), and *Rmixmod* (Lebret et al, 2015) packages.

2.2 Variable selection in model-based clustering

The SRUW model, as described in Maugis et al (2009b), involves three possible roles for variables: relevant clustering variables (S), redundant variables (U), and independent variables (W). Moreover, redundant variables U are dependent on a subset R of the relevant variables S, while the variables W are assumed to be independent of them. Therefore, the data density is assumed to break down into three parts, as follows:

$$f(\mathbf{y}_i \mid K, m, r, \ell, \mathbf{V}, \theta) = \sum_{k=1}^{K} \pi_k \phi\left(\mathbf{y}_i^S \mid \mu_k, \Sigma_{k(m)}\right) \times \phi\left(\mathbf{y}_i^U \mid a + \mathbf{y}_i^R b, \Omega_{(r)}\right) \times \phi\left(\mathbf{y}_i^W \mid \gamma, \tau_{(\ell)}\right),$$

where $\theta = (\alpha, a, b, \Omega, \gamma, \tau)$ is the full parameter vector and $\mathbf{V} = (S, R, U, W)$. The form of the regression covariance matrix Ω is denoted by r; it can be spherical, diagonal or general. The form of the covariance matrix τ of the independent variables W is denoted by ℓ, and can be spherical or diagonal.

The SRUW model recasts the variable selection problem for model-based clustering as a model selection problem, where the model collection is indexed by $(K, m, r, \ell, S, R, U, W)$. This model selection problem is solved maximizing the following BIC-type criterion:

$$\begin{aligned}&\mathrm{crit}\big(K, m, r, \ell, \mathbf{V}\big) = \\ &\mathrm{BIC}_{\mathrm{clust}}\left(\mathbf{y}^S \mid K, m\right) + \mathrm{BIC}_{\mathrm{reg}}\left(\mathbf{y}^U \mid r, \mathbf{y}^R\right) + \mathrm{BIC}_{\mathrm{indep}}\left(\mathbf{y}^W \mid \ell\right), \qquad (1)\end{aligned}$$

where $\mathrm{BIC}_{\mathrm{clust}}$ represents the BIC criterion of the Gaussian mixture model with the variables S, $\mathrm{BIC}_{\mathrm{reg}}$ the BIC criterion of the regression model of the

variables U on the variables R, and $\mathrm{BIC}_{\mathrm{indep}}$ the BIC criterion of the Gaussian model with the variables W.

Since the SRUW model collection is large, two embedded backward or forward stepwise algorithms for variable selection, one for the clustering and one for the linear regression, are used to solve this model selection problem. A backward algorithm allows us to start with all variables, in order to take variable dependencies into account. A forward procedure, starting with an empty clustering variable set or a small variable subset, could be preferred for computational reasons when the number of variables is large. The method is implemented in the *SelvarClustIndep* software, and a simplified version is implemented in the *clustvarsel*[3] R package. However, in a high-dimensional setting, even the variable selection method with the two forward stepwise algorithms becomes slow, and alternative methods are required.

2.3 Variable selection in classification

In the supervised classification framework, the labels of the training dataset are known and the variable selection problem is analogous to but simpler than in the clustering framework. The training data set is

$$(\mathbf{y}, \mathbf{z}) = \{(\mathbf{y}_1, z_1), \ldots, (\mathbf{y}_n, z_n);\ \mathbf{y}_i \in \mathbb{R}^p, z_i \in \{1, \ldots, K\}\},$$

where $\mathbf{y}_i$, $i = 1, \ldots, n$ are p-dimensional i.i.d. predictors and z_i, $i = 1, \ldots, n$ are the corresponding class labels. The number of classes K is known. The subset S is now the discriminant variable subset. Under a model $(m, r, \ell, \mathbf{V})$ with $\mathbf{V} = (S, R, U, W)$, the distribution of the training sample is modeled by

$$\begin{cases} f(\mathbf{y}_i | z_i = k, m, r, \ell, \mathbf{V}) = \phi(\mathbf{y}_i^S | \mu_k, \Sigma_{k(m)})\, \phi(\mathbf{y}_i^U | a + \mathbf{y}^R \beta, \Omega_{(r)})\, \phi(\mathbf{y}_i^W | \gamma, \tau_{(\ell)}), \\ \\ \quad (\mathbb{1}_{z_i=1}, \ldots, \mathbb{1}_{z_i=K}) \sim \text{Multinomial}(1; \pi_1, \ldots, \pi_K). \end{cases}$$

According to the assumed form of the covariance matrices involving their eigenvalue decomposition, a collection of more or less parsimonious mixture forms (m) is available, as in the clustering context.

Considering the same SRUW model, the variable selection problem is solved by using the model selection criterion:

$$\mathrm{crit}(m, r, \ell, \mathbf{V}) = \mathrm{BIC}_{\mathrm{clas}}(\mathbf{y}^S, \mathbf{z} | m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^U | r, \mathbf{y}^R) + \mathrm{BIC}_{\mathrm{indep}}(\mathbf{y}^W | \ell),$$

where $\mathrm{BIC}_{\mathrm{clas}}$ denotes the BIC criterion for the Gaussian classification on the discriminant variable subset S. Maximizing this criterion is an easier task in this supervised context, since there is no need to use the EM algorithm to derive the parameter estimates of the Gaussian classification model, unlike in the model-based clustering situation. See Maugis et al (2011) for further details. However, the variable selection procedure with two embedded stepwise procedures remain computationally costly and alternative procedures are still required.

[3] *clustvarsel* R package is available at `https://CRAN.R-project.org/package=clustvarsel`

3 Variable selection through regularization

In order to avoid the costly computational requirements of stepwise algorithms, we now propose an alternative variable selection procedure with two steps. First, the variables are ranked using a lasso-like procedure, and second, the variables' roles are determined using criterion (1) on these ranked variables. The procedure is detailed here for the clustering framework; the simplification to classification is presented in Section 3.3. This variable selection procedure is implemented in the *SelvarMix* R package.

3.1 Variable ranking by regularization

In the first step, the variables are ranked through the lasso-like procedure of Zhou et al (2009). First, data are centered and scaled ($\bar{\mathbf{y}} = (\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_n)'$). For any $K \in \mathbb{N}^\star$, the criterion to be maximized is

$$\sum_{i=1}^{n} \ln \Big[\sum_{k=1}^{K} \pi_k \phi\big(\bar{\mathbf{y}}_i \mid \mu_k, \Sigma_k\big) \Big] - \lambda \sum_{k=1}^{K} \|\mu_k\|_1 - \rho \sum_{k=1}^{K} \big\|\Sigma_k^{-1}\big\|_1 , \tag{2}$$

where

$$\|\mu_k\|_1 = \sum_{j=1}^{p} \big|\mu_{kj}\big|, \big\|\Sigma_k^{-1}\big\|_1 = \sum_{\substack{j'=1 \\ j' \neq j}}^{p} \sum_{j=1}^{p} \big|(\Sigma_k^{-1})_{jj'}\big| ,$$

and λ and ρ are two non-negative regularization parameters defined on two grids of values $\mathcal{G}_\lambda$ and $\mathcal{G}_\rho$ respectively. The estimated mixture parameters for fixed tuning parameters λ and ρ,

$$\widehat{\alpha}(\lambda, \rho) = \big(\widehat{\boldsymbol{\pi}}(\lambda, \rho), \widehat{\mu}_1(\lambda, \rho), \ldots, \widehat{\mu}_K(\lambda, \rho), \widehat{\Sigma}_1(\lambda, \rho), \ldots, \widehat{\Sigma}_K(\lambda, \rho)\big),$$

are computed with the EM algorithm of Zhou et al (2009). In particular, the glasso algorithm (Friedman et al, 2007) involving a coordinate descent procedure for the lasso is used to estimate the sparse precision matrices $\Sigma_k^{-1}, k = 1, \ldots, K$. This procedure is summarized in Appendix A.1.

It is worth noting that this lasso-like criterion does not take into account the division of the variables induced by the SRUW model. Strictly speaking, it only distinguishes two possible roles: a variable is declared related to, or independent of, the clustering. A variable j is declared independent of the clustering if for all $k = 1, \ldots, K$, $\widehat{\mu}_{kj}(\lambda, \rho) = 0$. The variance matrices are not considered in this definition. In fact, their role is secondary in clustering, and taking them into account would lead to serious computational difficulties.

By varying the regularization parameters (λ, ρ) in $\mathcal{G}_\lambda \times \mathcal{G}_\rho$, a "clustering" score can be defined for each variable $j \in \{1, \ldots, p\}$ and for fixed K by

$$\mathcal{O}_K(j) = \sum_{(\lambda,\rho) \in \mathcal{G}_\lambda \times \mathcal{G}_\rho} \mathcal{B}_{(K,\rho,\lambda)}(j),$$

where

$$\mathcal{B}_{(K,\rho,\lambda)}(j) = \begin{cases} 0 & \text{if } \widehat{\mu}_{1j}(\lambda,\rho) = \ldots = \widehat{\mu}_{Kj}(\lambda,\rho) = 0 \\ 1 & \text{otherwise.} \end{cases}$$

The larger the value of $\mathcal{O}_K(j)$, the more related to the clustering the variable j is expected to be. Variables are thus ranked by their decreasing values of $\mathcal{O}_K(j)$. This variable ranking is denoted $\mathcal{I}_K = (j_1, \ldots, j_p)$, with $\mathcal{O}_K(j_1) > \mathcal{O}_K(j_2) > \ldots > \mathcal{O}_K(j_p)$.

Remarks:

- The variance matrices are not taken into account for the variable ranking task. This limitation has been introduced to stop the procedure becoming overly complex. More precisely, the goal is to reduce computing time. In the model-based clustering context, this limitation can be thought of as reasonable since users are essentially interested in clusters with different means. To limit the impact of the variances, the data are initially centered and scaled.
- Concerning the sensitivity of the choice of the regularization grids, the default values did not show any pathological behavior in our experiments. However, we recommend that users try several regularization grids to ensure robust variable selection.

3.2 Determination of the variables' roles

The relevant clustering variable set: $S_{(K,m)}$, is first determined. The variable set is scanned according to the order in $\mathcal{I}_K$. One variable is added to $S_{(K,m)}$ if

$$\begin{aligned} \text{BIC}_{\text{diff}}(j_v) = \text{BIC}_{\text{clust}}\Big(\mathbf{y}^{S_{(K,m)}}, \mathbf{y}^{j_v} \mid K, m\Big) \\ -\text{BIC}_{\text{clust}}\Big(\mathbf{y}^{S_{(K,m)}} \mid K, m\Big) - \text{BIC}_{\text{reg}}\Big(\mathbf{y}^{j_v} \mid \mathbf{y}^{R[j_v]}\Big) \end{aligned} \quad (3)$$

is positive, $R[j_v]$ being the variables of $S_{(K,m)}$ required to linearly characterize $\mathbf{y}^{j_v}$. This subset $R[j_v]$ is determined with the standard backward stepwise algorithm for variable selection in linear regression. The scanning of $\mathcal{I}_K$ is stopped as soon as c successive variables have a non-positive BIC_{diff}, c being a fixed positive integer. Once the relevant variable set $S_{(K,m)}$ is determined, the independent variable set W is determined as follows. Scanning the variable set according to the reverse order of I_K, a variable j_v is added to $W_{(K,m)}$ if the subset $R[j_v]$ of $S_{(K,m)}$ (derived from the backward stepwise algorithm) is empty. The algorithm stops as soon as c successive variables are not declared independent. The redundant variables are thus declared to be $U_{(K,m)} = \{1, \ldots, p\} \setminus \{S_{(K,m)} \cup W_{(K,m)}\}$, and the subset $R_{(K,m,r)}$ of $S_{(K,m)}$ required to linearly explain $\mathbf{y}^{U_{(K,m)}}$ is derived from the backward stepwise algorithm, for each covariance shape r. The ideal position of the variable sets S, U and W in the variable ranking is shown in Figure 1. Lastly, the model

$(\hat{K}, \hat{m}, \hat{r}, \hat{\ell})$ maximizing the criterion $\text{crit}\big(K, m, r, \ell, \mathbf{V}_{(K,m,r)}\big)$ defined in Equation (1) with $\mathbf{V}_{(K,m,r)} = (S_{(K,m)}, R_{(K,m,r)}, U_{(K,m)}, W_{(K,m)})$ is selected.

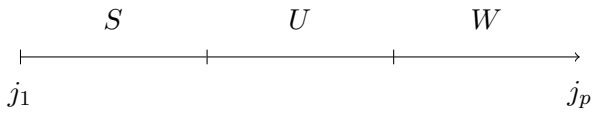

Fig. 1 The ideal position of the sets S, U and W in the ranking of the variables.

Some comments are in order:

- It is possible to use a lasso procedure instead of the stepwise variable selection algorithm in the linear regression step. However, this is not expected to be highly beneficial since stepwise variable selection in linear regression is not overly expensive, and moreover, the number of variables in the set S is not expected to be great.
- There is no guarantee that the variable order defined in 3.1 would be in accordance with the ideal ranking of the variables displayed in Figure 1. In particular when the variables are highly correlated, lasso-like procedures could be expected to lead to confusion between the sets S and U. This is the reason why we wait c (> 1) steps before deciding on the variables' roles; we give the procedure a chance to capture more variables in S and in W.

3.3 Variable selection through regularization for classification

In the classification context, K is fixed and the regularization criterion to be maximized is

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}_{\{z_i=k\}} \ln\Big[\pi_k \phi\big(\bar{\mathbf{y}}_i \mid \mu_k, \Sigma_k\big)\Big] - \lambda \sum_{k=1}^{K} \|\mu_k\|_1 - \rho \sum_{k=1}^{K} \big\|\Sigma_k^{-1}\big\|_1, \qquad (4)$$

with the same notation as in Section 3.1. Assuming that the training data set has been obtained according to the mixture sampling scheme, the proportions $\pi_1, \ldots, \pi_K$ are estimated by

$$\widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{z_i=k\}},\ k = 1, \ldots, K.$$

The maximization of criterion (4) is done according to the procedure described in Appendix A.2. The only difference with the clustering context is that the labels z_i are known, and no EM algorithm is required. Thus, a ranking $\mathcal{I}_K$ of the variables is obtained and the procedure described in Section 3.2 for model-based clustering is easily adapted to the supervised classification context, where (3) is replaced by

$$\begin{aligned}&\mathrm{BIC}_{\mathrm{diff}}(j_v) = \\ &\mathrm{BIC}_{\mathrm{clas}}\left(\mathbf{y}^{S_{(m)}}, \mathbf{y}^{j_v}, \mathbf{z} \mid m\right) - \mathrm{BIC}_{\mathrm{clas}}\left(\mathbf{y}^{S_{(m)}}, \mathbf{z} \mid m\right) - \mathrm{BIC}_{\mathrm{reg}}\left(\mathbf{y}^{j_v} \mid \mathbf{y}^{R[j_v]}\right).\end{aligned}$$

4 Simulations

This section is devoted to comparing our procedure, implemented in the *SelvarMix* R package, with the forward/backward stepwise procedures of Maugis et al (2009b, 2011) in both model-based clustering and classification settings.

4.1 Model-based clustering

4.1.1 Comparison on simulated data

We consider one of the seven simulated data sets studied in Maugis et al (2009b, Section 6.1). The data consist of $n = 2000$ observations of $p = 14$ variables. For the first two variables ($S = \{1, 2\}$), data are distributed from an equiprobable mixture of four Gaussian distributions $\mathcal{N}(\mu_k, I_2)$ with $\mu_1 = (0, 0), \mu_2 = (4, 0), \mu_3 = (0, 2)$ and $\mu_4 = (4, 2)$. For the nine redundant variables ($U = \{3, \ldots, 11\}$), data are simulated as follows: for $i = 1, \ldots, n$,

$$\mathbf{y}_i^{\{3-11\}} = (0, 0, 0.4, 0.8, \ldots, 2) + \mathbf{y}_i^{\{1,2\}} b + \varepsilon_i,$$

where the regression coefficients are

$$b = \left((0.5, 1)', (2, 0)', (0, 3)', (-1, 2)', (2, -4)', (0.5, 0)', (4, 0.5)', (3, 0)', (2, 1)'\right)$$

and ε_i are i.i.d. $\mathcal{N}(0_9, \Omega)$. The regression covariance matrix Ω is block diagonal

$$\Omega = \mathrm{diag}\left(I_3, 0.5I_2, \Omega_1, \Omega_2\right)$$

with $\Omega_1 = \mathrm{Rot}\left(\frac{\pi}{3}\right)'\mathrm{diag}(1, 3)\mathrm{Rot}\left(\frac{\pi}{3}\right)$ and $\Omega_2 = \mathrm{Rot}\left(\frac{\pi}{6}\right)'\mathrm{diag}(2, 6)\mathrm{Rot}\left(\frac{\pi}{6}\right)$, where $\mathrm{Rot}(\theta)$ is a plane rotation matrix with angle θ. The last three independent variables are standard Gaussian random variables $\mathbf{y}_i^{\{12-14\}} \sim \mathcal{N}(0_3, I_3)$.

In the first scenario, the CPU time and the variable selection of *SelvarMix*, *SelvarClustIndep* and *clustvarsel* (Scrucca and Raftery, 2014) were compared. The calculations were carried out on a machine with 80 Intel Xeon 2.4 GHz processors. The comparison is based on 100 replicates of the simulated dataset, $K = 4$ is fixed, and only spherical mixture models were considered. The variable ranking procedure (see Section 3.2) of the *SelvarClustLasso* function of *SelvarMix* is parallelized, whereas the combinatorial nature of the forward version of *SelvarClustIndep* made it difficult to do so. The *clustvarsel* function was used with the forward direction, and in the headlong approach, defined as follows. This approach, proposed in Murphy et al (2010), assesses

the univariate clustering/classification performance of each variable (based on both mean and variance information) at the first stage of the forwards stepwise algorithms. Thus, it gives a potential ranking of variables since it is used to order the variables for subsequent steps of the algorithm. As a result, a significant improvement of the run time with *SelvarMix* was obtained: it took 47.0($\pm$3.2) seconds, whereas *SelvarClustIndep* needed 450.64($\pm$104.0) and *clustvarsel* 124.0($\pm$25.97) seconds. Figure 2 displays the distribution of the variable roles with *SelvarClustIndep* (top) and *SelvarMix* (bottom). Globally, the true variable roles are recovered well. Surprisingly, *SelvarMix* detects the relevant variables better than *SelvarClustIndep*, which sometimes selects variables 5 and 9 instead of the first two variables. *clustvarsel* always declared the first two variables as relevant, except once where variable 12 was also declared relevant.

In the second scenario, the first 50 previous simulated datasets were considered (but now with the number of clusters K varying between 2 and 6), as well as the 28 Gaussian mixture forms m. The true cluster number was always correctly selected by *SelvarClustIndep* and *SelvarMix*. Variable selection was similar with both procedures (see Figure 3): the true variable partition was selected 46 (resp. 48) times by *SelvarClustIndep* (resp. *SelvarMix*). The clustering performance was preserved with *SelvarMix*, since the average adjusted rand index (ARI) was 0.6($\pm$0.017) with *SelvarMix* and 0.6($\pm$0.015) with *SelvarClustIndep*.

We also applied Zhou et al (2009)'s lasso-like procedure, the sparse Fisher-EM algorithm (*sfem*) of Bouveyron and Brunet (2014), and *clustvarsel* to this second scenario. For Zhou et al (2009)'s procedure, the penalization parameters, including the means $\widehat{\mu}_1, \ldots, \widehat{\mu}_K$, were estimated according to the penalized likelihood criterion (2), whereas the number of clusters K was selected using BIC. A variable j was declared relevant if there was at least one cluster k where $|\widehat{\mu}_{kj}| > 10^{-1}$. For the 50 datasets, this procedure failed to select the true number of clusters ($K = 4$) and the true set of relevant variables ($S = \{1, 2\}$). It always selected $K = 6$, both relevant and redundant variables were declared relevant, and the ARI was lower (0.45 ± 0.025). The *sfem*'s procedure always selected $K = 6$ clusters, with a lower ARI than *SelvarMix* (0.445 ± 0.021), and did not detect the first two relevant clustering variables since 12.7 ± 0.76 variables were used to create a five-dimensional discriminative subspace. Lastly, *clustvarsel* selected $K = 2$ (22 times) and $K = 4$ (28 times) clusters, 2.24 ± 0.48 variables were declared relevant for the clustering, and the ARI was 0.53 ± 0.07.

In the third scenario, we considered 50 datasets consisting of $n = 400$ observations made up of 100 variables. On the first eleven variables, data were distributed as previously. Then, 89 standard Gaussian variables were appended. As previously, the number of clusters varied from 2 to 6, and the 28 Gaussian mixtures forms m were considered. The selection of the number of clusters by *SelvarMix* was less efficient: the true cluster number $K = 4$ was selected 23 times, the model with $K = 3$ was selected 10 times, and the models with $K = 2, 5$ and 6 were respectively selected $6, 7$ and 4 times. Compared

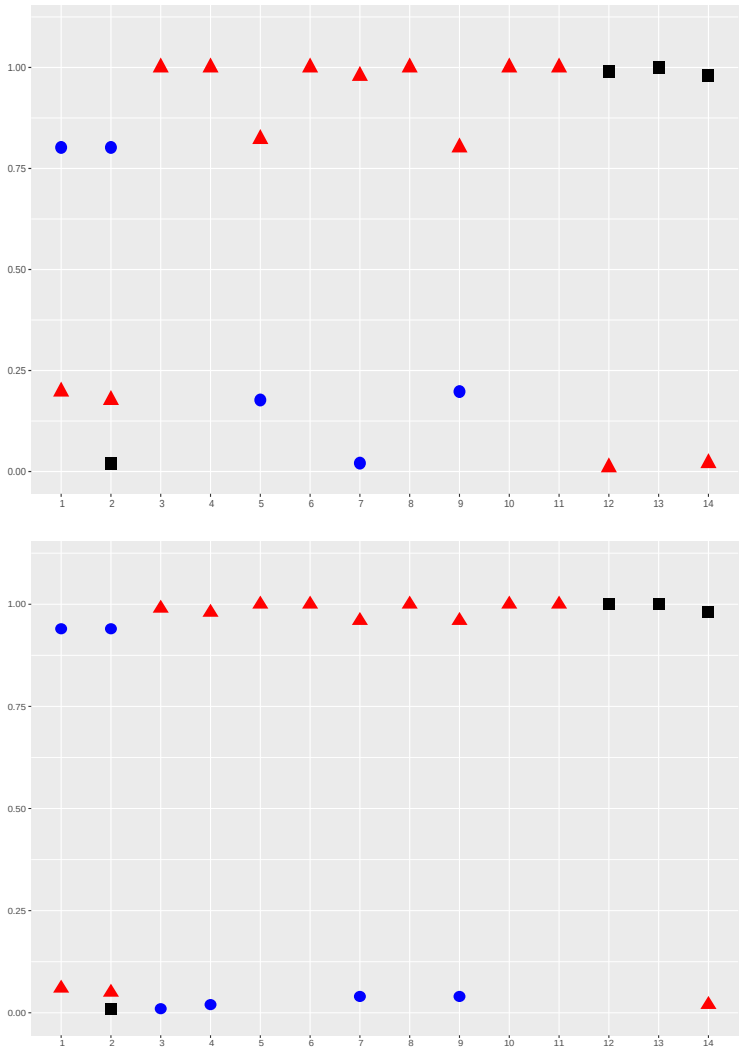

Fig. 2 Proportion of times each variable was declared relevant (square), redundant (triangle) or independent (circle) by *SelvarClustIndep* (top) and *SelvarMix* (bottom) in the first scenario. Zero values are not shown.

to the previous low-dimensional scenario ($p = 14$), variable selection using *SelvarMix* deteriorated slightly (see Figure 4). The true relevant variable set $S = \{1, 2\}$ was selected 23 times. Occasionally, *SelvarMix* declared one of the redundant variables as relevant, and one of the independent variables was declared relevant seven times. Moreover, the clustering performance deteriorated slightly: the ARI was $0.49(\pm 0.1)$. The *sfem* procedure always selected $K = 6$ clusters, the ARI was lower (0.43 ± 0.036) and several noise variables were declared relevant since 62.3 ± 6.82 variables were used to create a five-dimensional

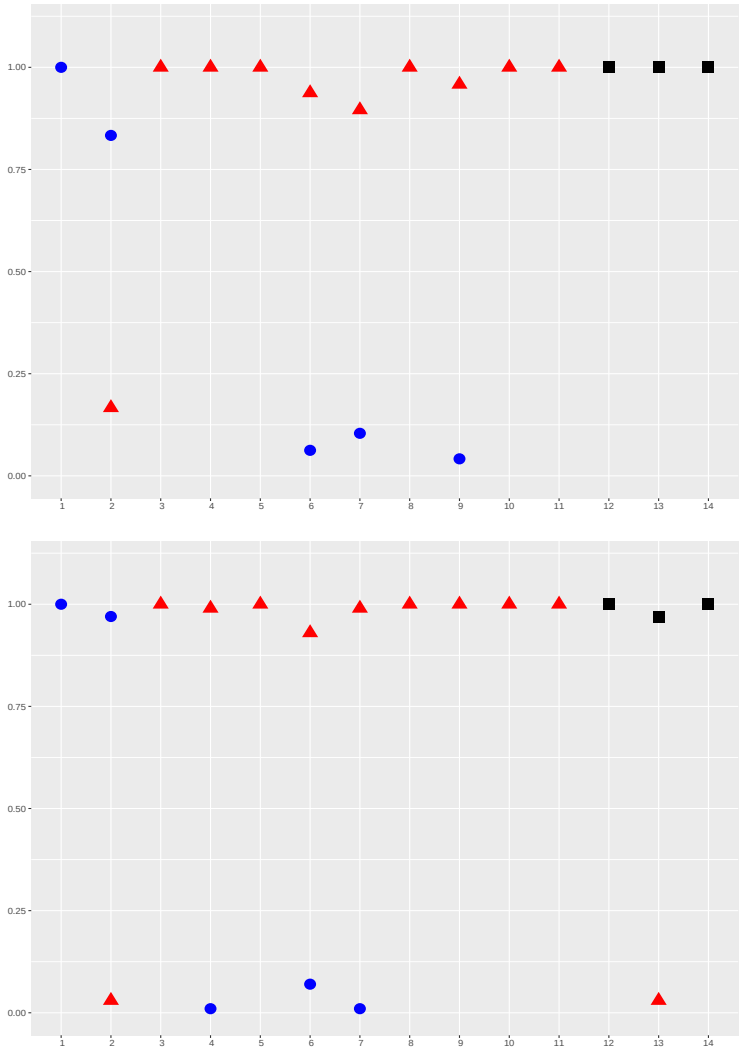

Fig. 3 Proportion of times each variable was declared relevant (square), redundant (triangle) or independent (circle) by *SelvarClustIndep* (top) and *SelvarMix* (bottom) in the second scenario. Zero values are not shown.

discriminative subspace. The *clustvarsel* function with the forward direction and the headlong approach failed to run more than half of the time. When it worked, *clustvarsel* selected 25 relevant variables on average, and always chose $K = 2$.

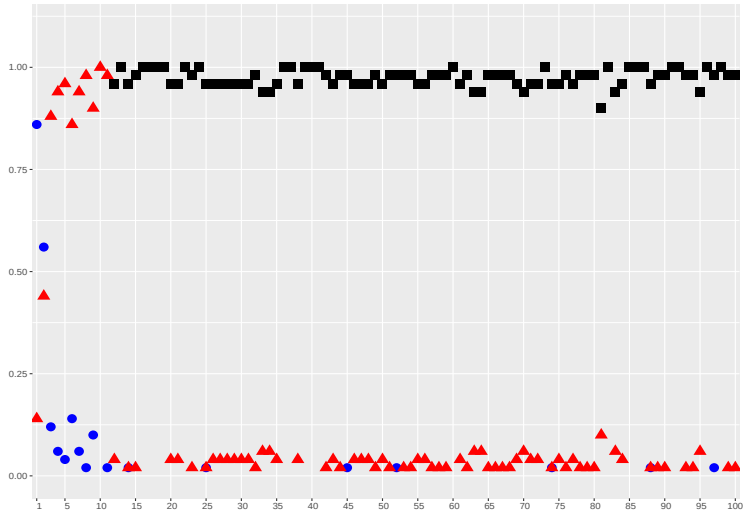

Fig. 4 Proportion of times each variable was declared discriminant (square), redundant (triangle) or independent (circle) by *SelvarMix* in the clustering setting, for the datasets with $p = 100$ variables. Zero values are not shown.

4.1.2 Comparison on real datasets

In this section, we compare *SelvarMix*, *SelvarClustIndep*, *sfem* and *clustvarsel* on the two following moderately high-dimensional datasets:

- **waveform**: This dataset consists of $n = 5,000$ points with $p = 40$ variables. The first 21 variables are relevant to the clustering problem. Nineteen standard centered Gaussian variables representing noise are appended. For the four procedures, the number of clusters varies from 3 to 6, *sfem* is used with its set of 12 models, and 20 Gaussian mixture models are considered for the three other methods.
- **transcriptome**: This dataset was extracted from the CATdb database (Gagnot et al (2008)) and was examined by Maugis et al (2009a) for clustering. This dataset consists of $n = 4,616$ genes of *Arabidopsis thaliana* characterized in terms of $p = 33$ biotic stress experiments. For the four methods, the number of clusters varies from 10 to 30, *sfem* is used with its set of 12 models, and two Gaussian mixture models are considered for the three other methods.

A detailed description of these two datasets is available in Maugis et al (2009b).

Results are shown in Table 1. The CPU time of *SelvarMix* is dramatically lower than that of *SelvarClustIndep* and *clustvarsel* for both datasets. Note that the comparison with *sfem* was difficult because the collection of models in *sfem* is not comparable with that of the other three methods. The number of relevant variables chosen by the four procedures was similar. Nevertheless,

sfem declared 12 noise variables as relevant for the waveform dataset. The SRUW modeling gives more information about the variables' roles.

Dataset	Software	$\widehat{K}$	$\widehat{m}$	Card(S)	Card(R)	Card(U)	Card(W)	time
waveform	SelvarClustIndep	6	pkLC	16	8	3	21	$>$ 24 hr
	SelvarMix	6	pLkC	15	11	8	18	1.29 min
	sfem	6	AkBk	20	-	-	-	57 min
	clustvarsel	6	pkLC	15	-	-	-	2 hr 2 min
transcriptome	SelvarClustIndep	24	pkLkC	30	16	3	0	$>$ 24 hr
	SelvarMix	27	pkLkC	29	15	4	0	1 hr 49 min
	sfem	24	DkBk	32	-	-	-	15 hr
	clustvarsel	26	pkLkC	28	-	-	-	$>$ 24 hr

Table 1 Results and CPU time for the *SelvarClustIndep*, *SelvarMix*, *sfem* and *clustvarsel* methods on both datasets.

4.2 Supervised classification

Here, we consider the simulated example presented in Maugis et al (2011), in which samples were made up of $p = 16$ variables. The proportions of the four classes were $\boldsymbol{\pi} = (0.15, 0.3, 0.2, 0.35)$. For the three discriminant variables, data were distributed according to

$$\mathbf{y}_i^{\{1-3\}} \mid z_i = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

with means $\mu_1 = (1.5, -1.5, 1.5)$, $\mu_2 = (-1.5, 1.5, 1.5)$, $\mu_3 = (1.5, -1.5, -1.5)$ and $\mu_4 = (-1.5, 1.5, -1.5)$. The covariance matrices were $\Sigma_k = \left(\rho_k^{|i-j|}\right)$ with $\rho_1 = 0.85, \rho_2 = 0.1, \rho_3 = 0.65$ and $\rho_4 = 0.5$. There were four redundant variables, simulated according to $\mathbf{y}_i^{\{4-7\}} \sim \mathcal{N}\left(\mathbf{y}_i^{\{1,3\}}\beta, I_4\right)$, with

$$\beta = \begin{pmatrix} 1 & 0 & -1 & 2 \\ 0 & -2 & 2 & 1 \end{pmatrix}.$$

Nine independent variables were appended, sampled from $\mathbf{y}_i^{\{8-16\}} \sim \mathcal{N}(\gamma, \tau)$, with

$$\gamma = (-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2)$$

and the diagonal covariance matrix

$$\tau = \mathrm{diag}(0.5, 0.75, 1, 1.25, 1.5, 1.25, 1, 0.75, 0.5).$$

To begin, 100 repeated simulation were performed, where the training sample was composed of $n = 500$ observations and the same test sample with 50 000 points was used. The fourteen forms m were considered. The *SelvarMix* method, and the version of *SelvarClustIndep* devoted to variable selection in supervised classification (still called *SelvarClustIndep* in the following), were

compared. Figure 5 shows the variable selection obtained by these two methods. We see that *SelvarMix* occasionally declared variables 6 and 7 as relevant alongside the first three relevant variables, and had a tendency to declare redundant certain independent variables, more often so than *SelvarClustIndep*. In terms of prediction, both procedures gave similar results: the misclassification error rate was 4.5%(±0.19) for *SelvarMix* and 4.18%(±0.06) for *SelvarClustIndep*.

Second, we considered 100 training samples composed of $n = 500$ observations of $p = 100$ variables, and 84 standard Gaussian variables were appended to the previous training samples. The test sample was similarly modified. As expected, *SelvarMix* allowed us to rapidly analyse such data sets involving a large number of variables. Prediction performance did not decrease: the misclassification error rate was 4.34%(±0.18), and variable selection remained similar, as shown in Figure 6.

5 Discussion

The SRUW model is a powerful model for defining the roles of variables in the Gaussian model-based clustering and classification contexts. However, the practical use of the model is hampered by the stepwise selection algorithms used in its previous versions. Indeed, not only sub-optimal, these algorithms are also highly computationally costly.

The regularization approach we have proposed allows us to avoid such stepwise procedures by designing an unmodifiable order in which the variables in S, U and W are chosen. Simulations performed with the R package *SelvarMix* implementing this approach encouraging results. *SelvarMix* is much faster than *SelvarClustIndep* while providing analogous (and sometimes better) performance than the reference *SelvarClustIndep* program.

In practice, the number c of steps, defined at the end of Subsection 3.2, for which the regularization algorithm provides the same selection is not sensitive. In our numerical experiments, a default value $c = 3$ seems reasonable. However the influence of c should be further investigated in order to try to propose heuristics to select it as a simple function of the total number of variables, in such a way that stable selections are obtained.

A Procedures to maximize penalized empirical contrasts

A.1 The model-based clustering case

The EM algorithm for maximizing criterion (2) is as follows (Zhou et al, 2009). The penalized complete loglikelihood of the centered data set $\bar{\mathbf{y}} = \left(\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_n\right)'$ is given by

$$L_{\mathrm{c},(\lambda,\rho)}(\bar{\mathbf{y}}, \mathbf{z}, \alpha) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\ln(\pi_k) + \ln\phi(\bar{\mathbf{y}}_i \mid \mu_k, \Sigma_k)\right] - \lambda\sum_{k=1}^{K}\|\mu_k\|_1 - \rho\sum_{k=1}^{K}\|\Theta_k\|_1\,, \quad (5)$$

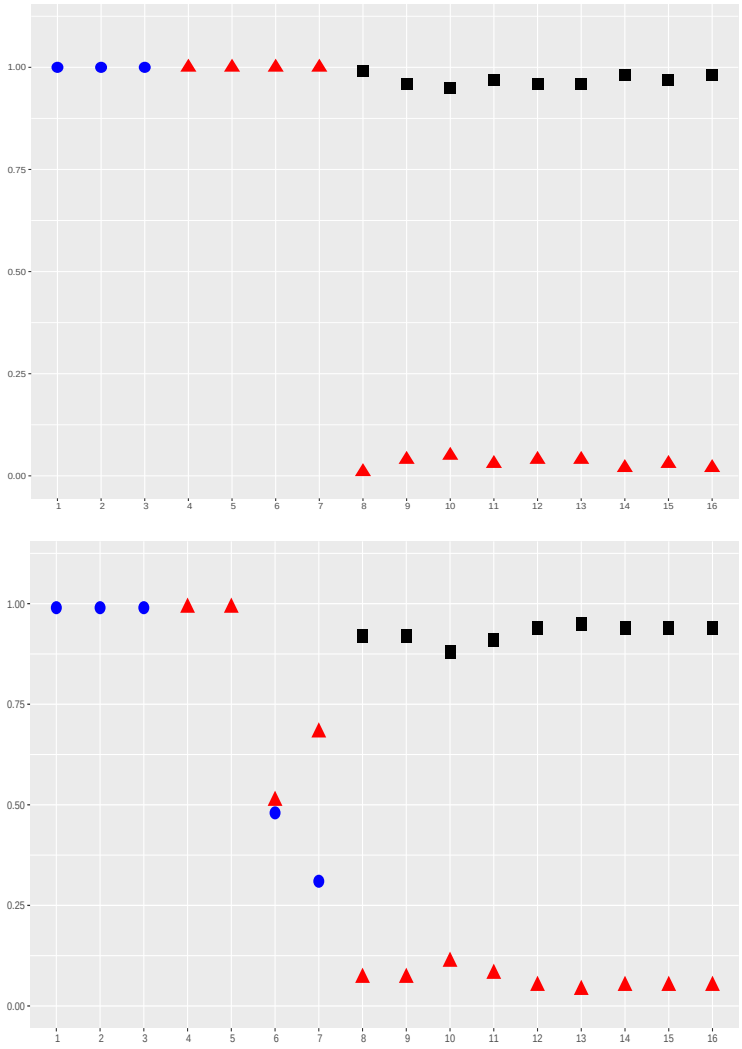

Fig. 5 Proportion of times each variable was declared discriminant (square), redundant (triangle) or independent (circle) by *SelvarClustIndep* (top) and *SelvarMix* (bottom) in the classification setting, for the data sets with $p = 16$ variables. Zero values are not shown.

where $\Theta_k = \Sigma_k^{-1}$ denotes the precision matrix of the k-th mixture component. The EM algorithm of Zhou et al (2009) maximizes at each iteration the conditional expectation of (5) given $\bar{\mathbf{y}}$ and a current parameter vector $\alpha^{(s)}$: $\mathbb{E}\Big[L_{c,(\lambda,\rho)}\big(\bar{\mathbf{y}}, \mathbf{z}, \alpha\big) \mid \bar{\mathbf{y}}, \alpha^{(s)}\Big]$. The following two steps are repeated from an initial $\alpha^{(0)}$ until convergence. At the s-th iteration of the EM algorithm:

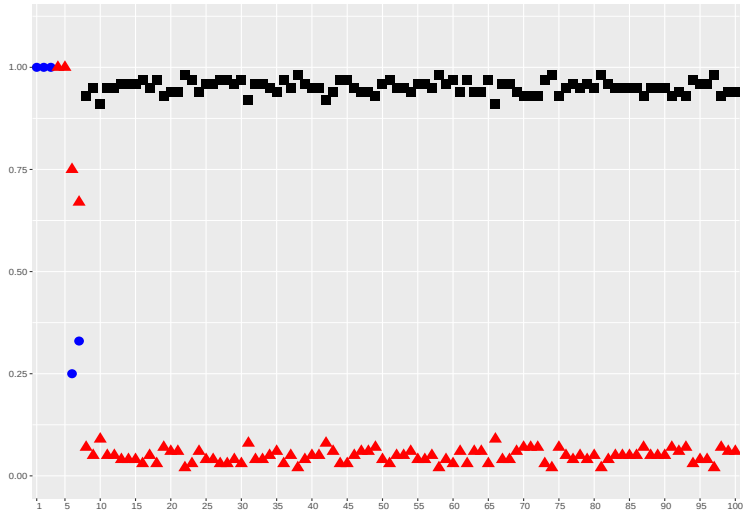

Fig. 6 Proportion of times each variable was declared discriminant (square), redundant (triangle) or independent (circle) by *SelvarMix* in the classification setting, for the data sets with $p = 100$ variables. Zero values are not shown.

- **E-step:** The conditional probabilities $t_{ik}^{(s)}$ that the i-th observation belongs to the k-th cluster are computed for $i = 1, \ldots, n$ and $k = 1, \ldots, K$,

$$t_{ik}^{(s)} = \mathbb{P}(z_{ik} = 1 \mid \bar{\mathbf{y}}, \alpha^{(s)}) = \frac{\pi_k^{(s)} \phi\left(\bar{\mathbf{y}}_i \mid \mu_k^{(s)}, \Sigma_k^{(s)}\right)}{\sum_{k'=1}^{K} \pi_{k'}^{(s)} \phi\left(\bar{\mathbf{y}}_i \mid \mu_{k'}^{(s)}, \Sigma_{k'}^{(s)}\right)}.$$

- **M-step :** This step consists of maximizing the expected complete log-likelihood derived from the E-step. It leads to the following mixture parameter updates:
  - The updated proportions are $\pi_k^{(s+1)} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}^{(s)}$ for $k = 1, \ldots, K$.
  - Compute the updated means $\mu_1^{(s+1)}, \ldots, \mu_K^{(s+1)}$ using formulas (14) et (15) of Zhou et al (2009): the j-th coordinate of $\mu_k^{(s+1)}$ is the solution of the following equations:

$$\text{if} \quad \left| \sum_{i=1}^{n} t_{ik}^{(s)} \left[\sum_{\substack{v=1 \\ v \neq j}}^{p} \left(\bar{\mathbf{y}}_{ij} - \mu_{kv}^{(s)} \right) \Theta_{k,vj}^{(s)} + \bar{\mathbf{y}}_{ij} \Theta_{k,jj}^{(s)} \right] \right| \leq \lambda, \quad \text{then} \quad \mu_{kj}^{(s+1)} = 0,$$

    otherwise:

$$\left[\sum_{i=1}^{n} t_{ik}^{(s)} \right] \mu_{kj}^{(s+1)} \Theta_{k,jj}^{(s)} + \lambda \operatorname{sign}\left(\mu_{kj}^{(s+1)} \right) = \sum_{i=1}^{n} t_{ik}^{(s)} \sum_{v=1}^{p} \bar{y}_{iv} \Theta_{k,vj}^{(s)} - \left[\sum_{i=1}^{n} t_{ik}^{(s)} \right] \left[\left(\sum_{v=1}^{p} \mu_{kv}^{(s)} \Theta_{k,vj}^{(s)} \right) - \mu_{kj}^{(s)} \Theta_{k,jj}^{(s)} \right].$$

  - For all $k = 1, \ldots, K$, the covariance matrix $\Sigma_k^{(s+1)}$ is obtained via the precision matrix $\Theta_k^{(s+1)}$. The *glasso* algorithm (available in the R package *glasso* of Friedman et al, 2011) is used to solve the following minimization problem on the set of

symmetric positive definite matrices (denoted $\Theta \succ 0$):

$$\underset{\Theta \succ 0}{\arg\min} \left\{ -\ln\det(\Theta) + \operatorname{trace}\left(S_k^{(s+1)}\Theta\right) + \rho_k^{(s+1)}\|\Theta\|_1 \right\},$$

where $\rho_k^{(s+1)} = 2\rho\left(\sum_{i=1}^n t_{ik}^{(s)}\right)^{-1}$ and

$$S_k^{(s+1)} = \frac{\sum_{i=1}^n t_{ik}^{(s)}(\bar{\mathbf{y}}_i - \mu_k^{(s+1)})(\bar{\mathbf{y}}_i - \mu_k^{(s+1)})^\top}{\sum_{i=1}^n t_{ik}^{(s)}}.$$

A.2 The classification case

The maximization of the regularized criterion (4) at $\mu_1, \ldots, \mu_K$ and $\Theta_1, \ldots, \Theta_K$ is achieved using an algorithm similar to the one presented in Section A.1 when the labels z_i are known.

The j-th coordinate of the mean vector μ_k is the solution of the following equations:

$$\text{if} \quad \left| \sum_{i=1}^n \mathbb{1}_{\{z_i=k\}} \left[\sum_{\substack{v=1 \\ v\neq j}}^p (\bar{\mathbf{y}}_{ij} - \mu_{kv})\,\Theta_{k,vj} + \bar{\mathbf{y}}_{ij}\Theta_{k,jj} \right] \right| \leq \lambda, \quad \text{then} \quad \mu_{kj} = 0,$$

otherwise:

$$\left[\sum_{i=1}^n \mathbb{1}_{\{z_i=k\}}\right] \mu_{kj}\Theta_{k,jj} + \lambda\ \operatorname{sign}\left(\mu_{kj}\right) = \sum_{i=1}^n \mathbb{1}_{\{z_i=k\}} \sum_{v=1}^p \bar{y}_{iv}\Theta_{k,vj} - \left[\sum_{i=1}^n \mathbb{1}_{\{z_i=k\}}\right] \left[\left(\sum_{v=1}^p \mu_{kv}\Theta_{k,vj} \right) - \mu_{kj}\Theta_{k,jj} \right].$$

To estimate the sparse precision matrices $\Theta_1, \ldots, \Theta_K$ from the data set $\mathbf{y}$ and the labels $\mathbf{z}$, we use the glasso algorithm to solve the following minimization problem on the set of symmetric positive definite matrices

$$\widehat{\Theta}_k = \underset{\Theta \succ 0}{\arg\min} \Big\{ -\ln\det\big(\Theta\big) + \operatorname{trace}\big(S_k\Theta\big) + \rho_k\|\Theta\|_1 \Big\}, \tag{6}$$

for each $k = 1, \ldots, K$. The ℓ_1 regularization parameter in (6) is given by $\rho_k = 2\rho\left(\sum_{i=1}^n \mathbb{1}_{\{z_i=k\}}\right)^{-1}$ and the empirical covariance matrix S_k by

$$S_k = \frac{\sum_{i=1}^n \mathbb{1}_{\{z_i=k\}}(\bar{\mathbf{y}}_i - \mu_k)(\bar{\mathbf{y}}_i - \mu_k)^\top}{\sum_{i=1}^n \mathbb{1}_{\{z_i=k\}}}.$$

Then, coordinate descent maximization in $(\mu_1, \ldots, \mu_K)$ and $(\Theta_1, \ldots, \Theta_K)$ is run until convergence.

Acknowledgments

One of the authors was supported by Fondation de Coopération Scientifique Campus Paris-Saclay-DIGITEO during his one year post-doc at INRIA Saclay Île-de-France. This work was partially supported by the French Agence Nationale de la Recherche (ANR), under grant MixStatSeq (ANR-13-JS01-0001-01). The authors thank the coordinating editor and two anonymous referees for their constructive remarks and comments.

References

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49(3):803–821

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(7):719–725

Bouveyron C, Brunet C (2014) Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. Computational Statistics 29:489–513

Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recognition 28(5):781 – 793

Celeux G, Maugis C, Martin-Magniette ML, Raftery AE (2014) Comparing model selection and regularization approaches to variable selection in model-based clustering. Journal of the French Statistical Society 155:57–71

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society Series B 39(1):1–38

Fraiman R, Justel A, Svarc M (2008) Selection of variables for cluster analysis and classification rules. Journal of the American Statistical Association 103:1294–1303

Friedman J, Hastie T, Tibshirani R (2007) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441

Friedman J, Hastie T, Tibshirani R (2011) glasso: Graphical lasso- estimation of Gaussian graphical models. URL `http://cran.r-project.org/web/packages/glasso/`

Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V (2008) Catdb: a public access to arabidopsis transcriptome data from the urgv-catma platform. Nucleic Acids Research 36(suppl 1):D986–D990

Galimberti G, Montanari A, Viroli C (2009) Penalized factor mixture analysis for variable selection in clustered data. Computational Statistics and Data Analysis 53:4301–4310

Kim S, Song DKH, DeSarbo WS (2012) Model-based segmentation featuring simultaneous segment-level variable selection. Journal of Marketing Research 49:725–736

Law MH, Figueiredo MAT, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9):1154–1166

Lebret R, Iovleff S, Langrognet F, Biernacki C, Celeux G, Govaert G (2015) Rmixmod: the R package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. Journal of Statistical Software 67(6):241–270

Lee H, Li J (2012) Variable selection for clustering by separability based on ridgelines. Journal of Computational and Graphical Statistics 21:315–337

Maugis C, Celeux G, Martin-Magniette M (2009a) Variable selection for clustering with Gaussian mixture models. Biometrics 65(3):701–709

Maugis C, Celeux G, Martin-Magniette ML (2009b) Variable selection in model-based clustering: A general variable role modeling. Computational Statistics and Data Analysis 53:3872–3882

Maugis C, Celeux G, Martin-Magniette ML (2011) Variable selection in model-based discriminant analysis. Journal of Multivariate Analysis 102:1374–1387

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the Lasso. The Annals of Statistics 34(3):1436–1462

Murphy TB, Dean N, Raftery AE (2010) Variable selection and updating in model-based discriminant analysis for high-dimensional data with food authenticity applications. Annals of Applied Statistics 4:396–421

Nia VP, Davison AC (2012) High-dimensional Bayesian clustering with variable selection: The R package bclust. Journal of Statistical Software 47:Issue 5

Pan W, Shen X (2007) Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research 8:1145–1164

Raftery AE, Dean N (2006) Variable Selection for Model-Based Clustering. Journal of the American Statistical Association 101(473):168–178

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464

Scrucca L, Raftery AE (2014) clustvarsel: A Package Implementing Variable Selection for Model-based Clustering in R. ArXiv e-prints 1411.0606

Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. The R Journal 8(1):289

Sun W, Wang J, Fang Y (2012) Regularized k-means clustering of high dimensional data and its asymptotic consistency. Electronic Journal of Statistics 6:148–167

Tadesse MG, Sha N, Vannucci M (2005) Bayesian variable selection in clustering high-dimensional data. Journal of the American Statistical Association 100(470):602–617

Wang S, Zhu J (2008) Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data. Biometrics 64(2):440–448

Xie B, Pan W, Shen X (2008) Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. Electronic journal of statistics 2:168–212

Zhou H, Pan W, Shen X (2009) Penalized model-based clustering with unconstrained covariance matrices. Electronic Journal of Statistics 3:1473–1496