Journal de la Société Française de Statistique
Vol. 155 No. 2 (2014)

Comparing Model Selection and Regularization Approaches to Variable Selection in Model-Based Clustering

Titre: Comparaison des approches de régularisation et de sélection d'un modèle de mélange pour la sélection de variables en classification non supervisée

Gilles Celeux[1], Marie-Laure Martin-Magniette[2], Cathy Maugis-Rabusseau[3] and Adrian E. Raftery[4]

Abstract: We compare two major approaches to variable selection in clustering: model selection and regularization. Based on previous results, we select the method of Maugis et al. (2009b), which modified the method of Raftery and Dean (2006), as a current state of the art model selection method. We select the method of Witten and Tibshirani (2010) as a current state of the art regularization method. We compared the methods by simulation in terms of their accuracy in both classification and variable selection. In the first simulation experiment all the variables were conditionally independent given cluster membership. We found that variable selection (of either kind) yielded substantial gains in classification accuracy when the clusters were well separated, but few gains when the clusters were close together. We found that the two variable selection methods had comparable classification accuracy, but that the model selection approach had substantially better accuracy in selecting variables. In our second simulation experiment, there were correlations among the variables given the cluster memberships. We found that the model selection approach was substantially more accurate in terms of both classification and variable selection than the regularization approach, and that both gave more accurate classifications than K-means without variable selection. But the model selection approach is not available in a very high dimension context.

Résumé : Nous considérons deux approches importantes pour la sélection de variables en classification non supervisée : la sélection par modèle et la régularisation. Parmi les procédures existantes de sélection de variables par sélection de modèles, nous choisissons la méthode de Maugis et al. (2009b), généralisation de celle de Raftery et Dean (2006). Pour les méthodes fondées sur la régularisation, nous nous intéressons à la méthode de Witten and Tibshirani (2010). Nous comparons les performances de classification et de sélection de variables de ces deux procédures sur des données simulées. Nous montrons que la sélection de variables permet d'améliorer la classification quand les classes sont bien séparées. Les deux procédures de sélection de variables étudiées donnent des classifications analogues dans le premier exemple, mais l'approche par sélection de modèles a de meilleures performances pour la sélection de variables. Dans le second exemple, les variables sont corrélées. Nous montrons que l'approche par sélection de modèles améliore globalement la classification et la sélection de variables par rapport à la régularisation, et les deux procédures donnent de meilleurs résultats que l'algorithme des K-means (sans sélection de variables) pour la classification. Mais, il convient

[1] Inria Saclay - Île-de-France, Orsay, France.
E-mail: gilles.celeux@inria.fr
[2] Unité de Recherche en Génomique Végétale UMR INRA 1165 - UEVE ERL CNRS 8196, Evry, France.
UMR AgroParisTech/INRA MIA 518, Paris, France.
E-mail: marie_laure.martin@agroparistech.fr
[3] Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, Toulouse, France.
E-mail: cathy.maugis@insa-toulouse.fr
[4] Department of Statistics, University of Washington, Seattle, Washington, USA, and School of Mathematical Sciences, University College Dublin, Ireland.
E-mail: raftery@u.washington.edu

 ISSN: 2102-6238

de noter que la sélection par modèles est inopérante pour les très grandes dimensions. Enfin, ce travail de comparaison est également mené sur des données réelles.

Keywords: Model-based clustering, Model selection, Regularization approach, Variable selection
Mots-clés : Classification non supervisée, Mélanges gaussiens, Régularisation, Sélection de modèles, Sélection de variables
AMS 2000 subject classifications: 62H30

1. Introduction

Over the past 20 years, model-based clustering (Wolfe, 1970; McLachlan and Basford, 1988; Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 2002) has come to rival the heuristic clustering methods such as single link, complete link and K-means that dominated previously. In the past decade it has been realized that the performance of model-based clustering can be degraded if irrelevant or noise variables are present. As a result, there has been considerable interest in variable selection for model-based clustering.

Two of the most used general approaches have been model selection and regularization. Model selection approaches to the problem were pioneered by Law et al. (2004) and Tadesse et al. (2005), who partitioned the set of candidate variables into two sets, one set relevant to the clustering and the other irrelevant. They assumed that the irrelevant variables were statistically independent of the relevant ones.

Raftery and Dean (2006) — hereafter RD — realized that irrelevant variables are often correlated with relevant ones, and developed a model selection method that takes account of this, and a greedy search algorithm to implement it. Their method assumes that each irrelevant variable depends on all the relevant variables according to a linear regression model.

Maugis et al. (2009a) pointed out that the Raftery-Dean method implies a very non-parsimonious model for all the variables jointly, explaining the method's lacklustre performance in some comparative studies (Steinley and Brusco, 2008; Witten and Tibshirani, 2010). They proposed modifying it by selecting the predictor variables in the linear regression part of the model. Maugis et al. (2009b) went further and allowed explicitly for an irrelevant variable to be independent of all the relevant variables. Although at first sight these may not seem like major changes to the method, they can actually make a big difference to the results and have led to greatly improved performance. The resulting method provides both more parsimonious and realistic models. Following Celeux et al. (2011), we refer to it here as the RD-MCM method.

A different approach, via regularization, was proposed by Pan and Shen (2007); see also Xie et al. (2008); Wang and Zhu (2008); Zhou et al. (2009); Guo et al. (2010). Another regularization approach, called sparse K-means (SparseKmeans), was proposed by Witten and Tibshirani (2010) — hereafter WT; this can be viewed as a simpler version of the Clustering on Subsets of Objects (COSA) approach of Friedman and Meulman (2004). WT found in a simulation study that their method outperformed both COSA and the regularization method of Pan and Shen (2007).

WT also found that their method outperformed the RD method. However, this appears to have been due to the nonparsimonious model underlying the RD method, and the WT method gave similar results to the RD-MCM method under the simulation setup of WT (Celeux et al., 2011).

Overall, it seems that, among variable selection methods for model-based clustering, the RD-MCM method is currently one of the best performing model selection methods, and the WT

 ISSN: 2102-6238

method is one of the best performing regularization methods. In this paper, we compare these two methods in a range of simulated and real data settings.

In Section 2 we summarize the two methods we are comparing. In Section 3 we give results for a range of simulation setups previously proposed, and in Section 4 we give results for a waveform dataset and for data from a genetics experiment. In Section 5 we discuss limitations, caveats, other approaches and possible extensions of our results.

2. Variable Selection Methods for Model-Based Clustering

2.1. Model Selection Methods

Let n observations $\mathbf{y} = (y_1, \ldots, y_n)$ to be clustered be described by p continuous variables ($y_i \in \mathbb{R}^p$). In the model-based clustering framework, the multivariate continuous data $\mathbf{y}$ are assumed to come from several subpopulations (clusters) modeled with a multivariate Gaussian density. The observations are assumed to arise from a finite Gaussian mixture with K components and a model m, namely

$$f(y_i|K,m,\alpha) = \sum_{k=1}^{K} \pi_k \phi(y_i|\mu_k, \Sigma_{k(m)}),$$

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 $\sum_{k=1}^{K} \pi_k = 1$), $\phi(.|\mu_k, \Sigma_k)$ is the p-dimensional Gaussian density function with mean μ_k and variance matrix Σ_k, and $\alpha = (\boldsymbol{\pi}, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ is the parameter vector.

This framework yields a range of possible models m, each corresponding to different assumptions on the forms of the covariance matrices, arising from a modified spectral decomposition. 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, the mixture parameters are estimated via maximum likelihood using the EM algorithm, and both the model structure and the number of components K are chosen using the BIC or other penalized likelihood criteria (Fraley and Raftery, 2002). Software to implement this methodology includes the MCLUST R package (`http://www.stat.washington.edu/mclust/`), and the MIXMOD software (`http://www.mixmod.org`). The latter implements 28 Gaussian mixture models, most of which are also available in MCLUST. Here we view each mixture component as corresponding to one cluster, and so the term cluster is used hereafter.

The RD-MCM method, as described by Maugis et al. (2009b), involves three possible roles for the variables: the relevant clustering variables (S), the redundant variables (U) and the independent variables (W). Moreover, the redundant variables U are explained by a subset R of the relevant variables S, while the variables W are assumed to be independent of the relevant variables. Thus the data density is assumed to be decomposed into three parts as follows:

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

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

 ISSN: 2102-6238

The RD-MCM method recasts the variable selection problem for model-based clustering as a model selection problem. This model selection problem is solved using a model selection criterion, decomposed into the sum of the three values of the BIC criterion associated with the Gaussian mixture, the linear regression and the independent Gaussian density respectively. The method is implemented by using two backward stepwise algorithms for variable selection, one each for the clustering and the linear regression. A backward algorithm allows one 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 numerical reasons if the number of variables is large. The method is implemented in the *SelvarClustIndep* software. [1]

The RD-MCM method generalizes several previous model selection methods. The procedure of Law et al. (2004), where irrelevant variables are assumed to be independent of all the relevant variables, corresponds to $W = S^c, R = \emptyset, U = \emptyset$. The RD method (Raftery and Dean, 2006) assumes that the irrelevant variables are regressed on the whole relevant variable set ($W = \emptyset$, $U = S^c$ and $R = S$). The generalization of Maugis et al. (2009a) enriches this model by allowing the irrelevant variables to be explained by only a subset of the relevant variables $R \subset S$ $(W = \emptyset, U = S^c)$; this method is implemented in the *SelvarClust* software [2].

2.2. *Regularization Methods*

A review of sparse clustering techniques can be found in WT. Most of these methods embed sparse clustering in the model-based clustering framework. Notable exceptions are the COSA approach of Friedman and Meulman (2004) and the WT approach, which can be thought of as a simpler version of the approach of Friedman and Meulman (2004).

WT propose a sparse clustering procedure, called the *Sparse K-means* algorithm. This procedure is based on a variable weighting in the K-means algorithm. Let $\mathscr{P} = (\mathscr{P}_1, \ldots, \mathscr{P}_K)$ denote a clustering of observations and n_k the number of observations in cluster $\mathscr{P}_k$. The *sparse K-means* algorithm maximizes a weighted between-cluster sum of squares

$$\max_{\mathscr{P},\mathbf{w}} \sum_{j=1}^{p} w_j \left(\frac{1}{n} \sum_{i=1}^{n}\sum_{i'=1}^{n} (y_{ij} - y_{i'j})^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in \mathscr{P}_k} (y_{ij} - y_{i'j})^2 \right),$$

where $\mathbf{w} = (w_1, \ldots, w_p)$ is the weight vector such that for all j, $w_j \geq 0$, $\|\mathbf{w}\|^2 \leq 1$ and $\|\mathbf{w}\|_1 \leq t$ where t is a tuning parameter. This parameter is chosen by a permutation approach using the gap statistic of Tibshirani et al. (2001). Their method is implemented in the R package SPARCL.

3. Simulation Experiments

We now give comparative results for two simulation experiments with setups based on simulation experiments in the related literature. We compare three methods: K-means with no variable selection, SparseKmeans, and the RD-MCM method.

[1] *SelvarClustIndep* is available at http://www.math.univ-toulouse.fr/~maugis/
[2] *SelvarClust* is available at http://www.math.univ-toulouse.fr/~maugis/

 ISSN: 2102-6238

3.1. *Simulation Experiment 1: Conditionally Independent Variables*

This simulated example is inspired by WT's Simulation 2. It concerns the situation where the relevant variables are conditionally independent given the cluster memberships, and the irrelevant variables are independent both of the relevant variables and of one another.

Five scenarios are considered with $p = 25$ variables in Scenarios 1-4 and $p = 100$ variables in Scenario 5. The first five variables are distributed according to a mixture of three equiprobable spherical Gaussian distributions $\mathcal{N}(\mu_k, I_5)$ with $\mu_1 = -\mu_2 = (\mu, \ldots, \mu) \in \mathbb{R}^5$ and $\mu_3 = 0_5$. Twenty (respectively ninety-five) noisy standard centered Gaussian variables are appended in Scenarios 1-4 (respectively in Scenario 5).

The number of observations is $n = 30$ in Scenarios 1 and 2, and $n = 300$ in Scenarios 3, 4 and 5. In Scenarios 1 and 3, $\mu = 0.6$, while $\mu = 1.7$ in Scenarios 2, 4 and 5. Note that the second scenario is the one considered by WT. Since the SparseKmeans method requires the user to know the number of clusters, the true number $K = 3$ is assumed known for all three methods. Moreover, the true mixture shape is fixed for the RD-MCM method. Twenty-five datasets were simulated for each scenario.

We evaluate the three methods according to three criteria. Classification performance is evaluated by the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985), reported in percent. The ARI is 100 for a perfect classification and 0 for a random one; larger ARI is better. Performance in selecting the right variables is evaluated using the Variable Selection Error Rate (VSER), defined as the average number of errors in selecting (or not selecting) variables, as a percentage of the total number of variables considered. SparseKmeans is defined as selecting a variable if the corresponding weight is positive. Smaller values of VSER are better. Finally, we also report the average number of variables selected (#VarSel). The true number of variables in this experiment is 5, so the closer #VarSel is to 5 the better.

The results are shown in Table 1. For classification, both variable selection methods greatly outperformed K-means without variable selection for Scenarios 2 and 4 (with $\mu = 1.7$), while for Scenarios 1 and 3 (with $\mu = 0.6$) there was not a big difference. The two variable selection methods performed similarly for classification across the first four scenarios. In Scenario 5, the three methods had similar behaviour.

For variable selection, however, there were clear distinctions. The RD-MCM method had a better variable selection error rate than SparseKmeans in the five scenarios, with a substantial difference for the larger sample size (Scenario 5). The number of variables selected by RD-MCM was closer to the true number (5) for all five scenarios. SparseKmeans selected substantially more variables in Scenarios 1-3 and all the variables in Scenarios 4-5.

Figure 1 displays the distribution of the variable roles with RD-MCM and the variations of the weights given by SparseKmeans for Scenarios 1-4. In Scenario 1, neither method accurately distinguished between clustering and non-clustering variables, which is not surprising as both the sample size and the between-cluster separation were small. In Scenario 2 the RD-MCM method identified the five clustering variables correctly in most cases, while the weights from SparseKmeans varied considerably. In Scenarios 3 and 4, both methods distinguished clearly between clustering and non-clustering methods, with RD-MCM selecting them and not the non-clustering variables, and SparseKmeans giving much higher weights on average to the clustering than the non-clustering variables. SparseKmeans gave positive, although small, weights

 ISSN: 2102-6238

TABLE 1. *Simulation Experiment 1 Results. ARI is the Adjusted Rand Index expressed in percent (the higher the better), VSER is the variable selection error rate in percent (the lower the better). #VarSel is the average number of variables selected (correct number = 5). Standard errors are shown in parentheses. The method performing best in each scenario under each criterion is shown in bold.*

Scenario	n	μ	Method	ARI	VSER	#VarSel
1	30	0.6	K-means	**11 (9)**	80 (—)	25.0 (—)
25 variables			SparseKmeans	8 (7)	45 (22)	14.4 (6.3)
			RD-MCM	8 (7)	**36 (11)**	**8.1 (1.9)**
2	30	1.7	K-means	44 (10)	80 (—)	25.0 (—)
25 variables			SparseKmeans	**81 (17)**	**13 (16)**	8.2 (4.0)
			RD-MCM	71 (24)	17 (12)	**6.8 (1.4)**
3	300	0.6	K-means	20 (3)	80 (—)	25.0 (—)
25 variables			SparseKmeans	14 (4)	76 (11)	24.0 (2.7)
			RD-MCM	**23 (3)**	**10 (7)**	**7.0 (1.8)**
4	300	1.7	K-means	64 (14)	80 (—)	25.0 (—)
25 variables			SparseKmeans	**89 (3)**	80 (0)	25.0 (0)
			RD-MCM	88 (3)	**2 (3)**	**5.6 (0.9)**
5	300	1.7	K-means	86 (3)	95 (—)	100.0 (—)
100 variables			SparseKmeans	**88 (4)**	95 (0)	100.0 (0)
			RD-MCM	84 (18)	**4 (2)**	**8.4 (2.3)**

to non-clustering variables in many cases.

Figure 2 shows the results for Scenario 5, with 100 variables, of which five were relevant. RD-MCM selected the correct variables in most cases, and also correctly identified the non-clustering variables as independent (rather than just redundant). SparseKmeans gave much higher weights to the clustering variables than to the non-clustering ones on average, although again the weights for the non-clustering variables were always positive.

3.2. *Simulation Experiment 2: Correlated Variables*

We now consider situations where the variables are correlated conditionally on the cluster memberships. Two of the seven simulated situations considered in (Maugis et al., 2009b, Sect. 6.1) are now considered. The $n = 2000$ observations are described by $p = 14$ variables. The first two variables are distributed according to a 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 $\mu_4 = (4,2)$. In the first situation (Maugis et al., 2009b, Scenario 3), the mixing proportion vector is $\boldsymbol{\pi} = (0.2, 0.3, 0.3, 0.2)$. The last twelve variables are simulated as follows:

$$\begin{cases} y_i^3 = 3y_i^1 + \varepsilon_i \text{ with } \varepsilon_i \sim \mathcal{N}(0, 0.5), \\ y_i^{\{4-14\}} \sim \mathcal{N}((0, 0.4, 0.8, \ldots, 3.6, 4), I_{11}). \end{cases}$$

In the second situation, an equiprobable mixture is considered on the first two variables and the last twelve variables are simulated as follows:

$$\begin{cases} y_i^{\{3-11\}} = (0, 0, 0.4, 0.8, \ldots, 2) + y_i^{\{1,2\}} b + \varepsilon_i \text{ with } \varepsilon_i \sim \mathcal{N}(0_9, \Omega), \\ y_i^{\{12-14\}} \sim \mathcal{N}((3.2, 3.6, 4), I_3), \end{cases} \tag{1}$$

 ISSN: 2102-6238

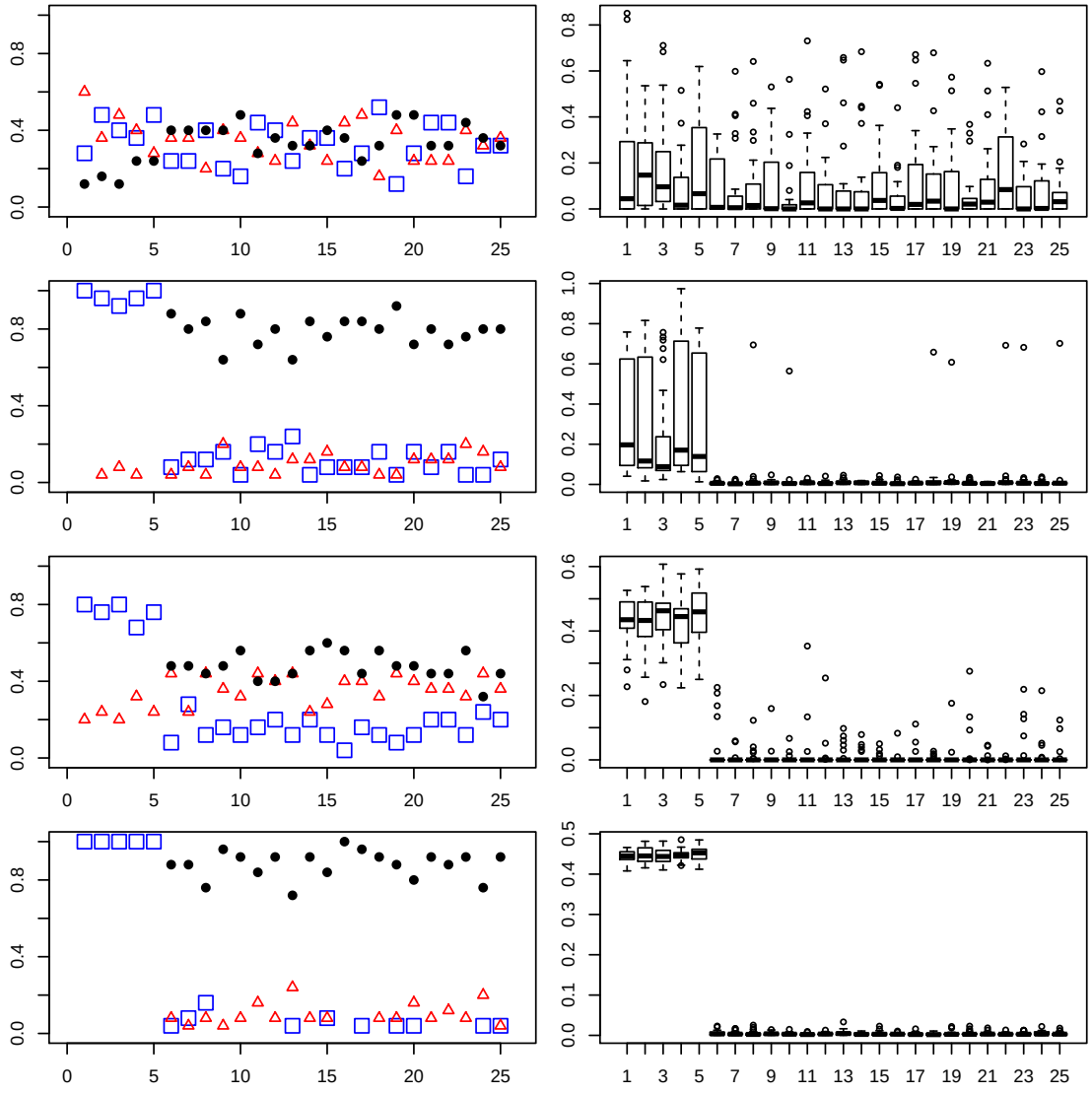

FIGURE 1. *Simulation Experiment 1: On the left, the proportion of times each variable was declared relevant (square), redundant (triangle) or independent (circle) by RD-MCM in the first four scenarios (Scenario 1 at top to Scenario 4 at bottom). Zero values are not shown. On the right, boxplots of the weights given by SparseKmeans for each variable in the four scenarios.*

 ISSN: 2102-6238

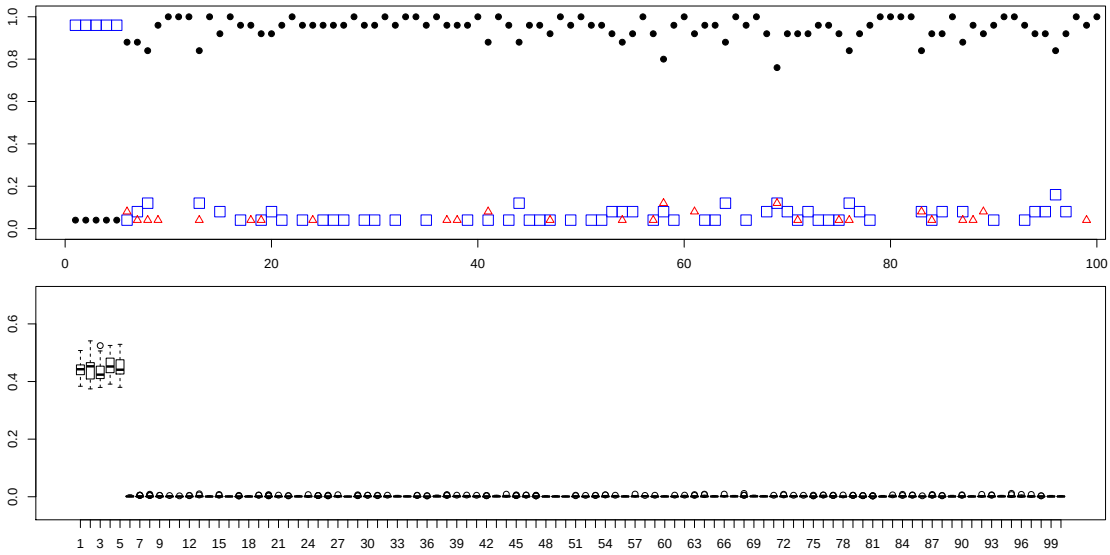

FIGURE 2. *Simulation Experiment 1: On the top, proportion of times each variable was declared relevant (square), redundant (triangle) or independent (circle) by RD-MCM in Scenario 5. Zero values are not shown. On the bottom, boxplots of the weights given by SparseKmeans for each variable in Scenario 5.*

where the regression coefficients are

$$b = ((0.5,1)', (2,0)', (0,3)', (-1,2)', (2,-4)', (0.5,0)', (4,0.5)', (3,0)', (2,1)').$$

In (1), $\Omega = \text{diag}(I_3, 0.5I_2, \Omega_1, \Omega_2)$ is the block diagonal regression variance matrix with $\Omega_1 = \text{Rot}(\pi/3)'\text{diag}(1,3)\text{Rot}(\pi/3)$ and $\Omega_2 = \text{Rot}(\pi/6)'\text{diag}(2,6)\text{Rot}(\pi/6)$, where $\text{Rot}(\theta)$ denotes the plane rotation matrix with angle θ.

The third situation is analogous to the second situation with many more noisy variables:

$$\begin{cases} y_i^{\{3-11\}} = (0,0,0.4,0.8,\ldots,2) + y_i^{\{1,2\}} b + \varepsilon_i \text{ with } \varepsilon_i \sim \mathcal{N}(0_9, \Omega) \\ y_i^{\{12-41\}} \sim \mathcal{N}(0_{30}, I_{30}); y_i^{\{42-71\}} \sim \mathcal{N}(2_{30}, I_{30}); y_i^{\{72-101\}} \sim \mathcal{N}(4_{30}, I_{30}) \end{cases}$$

The correlations of the variables in the three situations are shown in Figure 3. The true number of clusters, $K = 4$, is assumed known for all the procedures. Each situation has been replicated 50 times. The results are shown in Table 2.

For both classification and variable selection, RD-MCM substantially outperformed both K-means without variable selection, and SparseKmeans. This may be because the simulation experiment involves correlation between the variables beyond the clustering. Neither K-means nor SparseKmeans takes account of this explicitly, while RD-MCM does.

Figures 4 and 5 display the proportion of times each variable was declared relevant, redundant or independent by RD-MCM, and the variations of the weights given by SparseKmeans. In most cases, RD-MCM correctly selected the clustering variables and not the non-clustering variables in all three scenarios. SparseKmeans tended to give high weights to redundant variables, and in Scenarios 2 and 3 it gave low weights on average to the first two variables, which are the relevant ones.

 ISSN: 2102-6238

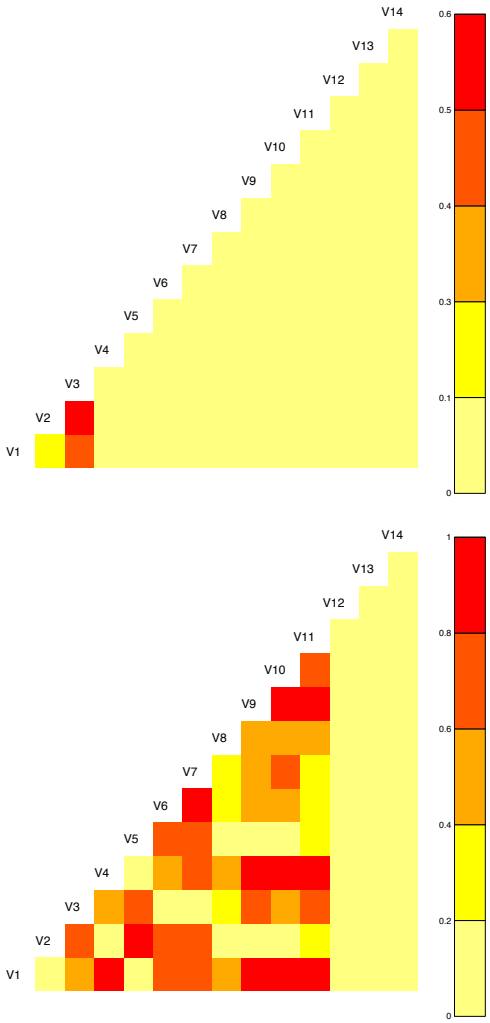

FIGURE 3. *Simulation Experiment 2: Correlation matrix for all 14 variables in one simulation of Scenario 1 (top) and Scenarios 2 and 3 (bottom).*

4. Real Data Examples

4.1. Waveform data

This dataset consists of $n = 5,000$ points with $p = 40$ variables. The first 21 variables are relevant to the classification problem, and are based on a random convex combination of two of three waveforms sampled at the integers $\{1,\ldots,21\}$ with noise added. Nineteen noisy standard centered Gaussian variables are appended. A detailed description of the waveform dataset is given by (Breiman et al., 1984, pp.43–49).

To compare the RD-MCM method with SparseKmeans on an equal footing, we assume that the

 ISSN: 2102-6238

TABLE 2. *Simulation Experiment 2 Results. ARI is the Adjusted Rand Index in percent (the higher the better), VSER is the variable selection error rate in percent (the lower the better). #VarSel is the average number of variables selected (correct number = 2). Standard errors are into parentheses. The method performing best in each scenario under each criterion is shown in bold.*

Scenario	Method	ARI	VSER	#VarSel
1	K-means	52 (1)	86 (—)	14.0 (—)
	SparseKmeans	47 (2)	86 (0)	14.0 (0)
	RD-MCM	**57 (4)**	**0 (0)**	**2 (0)**
2	K-means	57 (2)	86 (—)	14.0 (—)
	SparseKmeans	31 (3)	85 (5)	13.8 (1)
	RD-MCM	**60 (2)**	**1 (3)**	**2.0 (0.2)**
3	K-means	56 (2)	86 (—)	101.0 (—)
	SparseKmeans ($w > 0$)	34 (6)	80 (36)	82.9 (36.5)
	RD-MCM	**57 (7)**	**1 (2)**	**2.02 (0.14)**

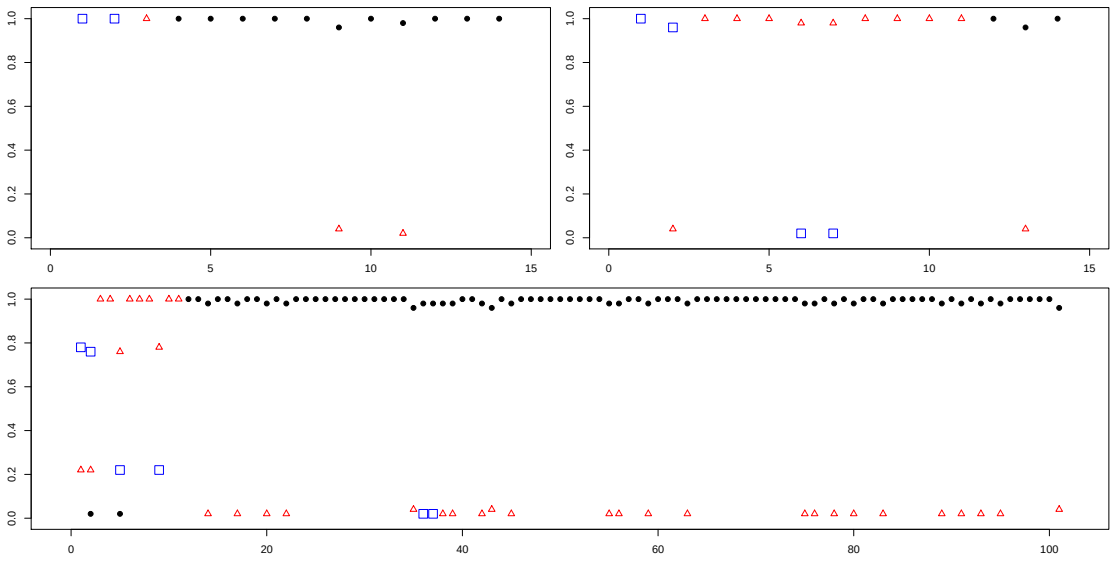

FIGURE 4. *Simulation Experiment 2: Proportion of times each variable was declared relevant (blue square), redundant (red triangle) or independent (black circle) by RD-MCM in Scenario 1 (top left), Scenario 2 (top right) and Scenario 3 (bottom). Zero values are not shown. In each scenario, only the first two variables were relevant.*

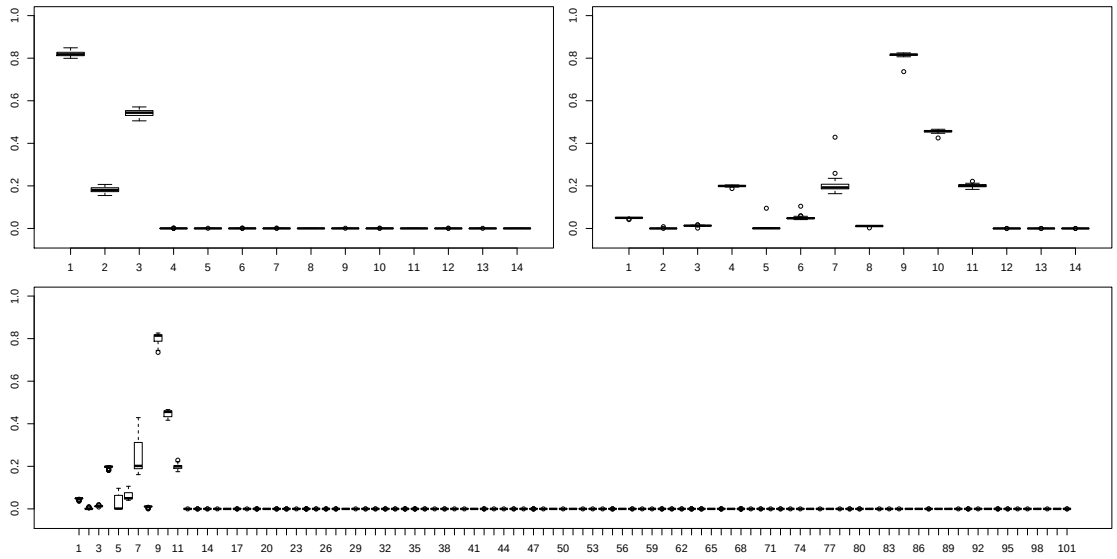

FIGURE 5. *Simulation Experiment 2: Boxplots of the weights* **w** *given by SparseKmeans for each variable in Scenario 1 (top left), Scenario 2 (top right) and Scenario 3 (bottom) respectively. In each scenario, only the first two variables were relevant.*

correct number of clusters, $K = 3$, is known. The RD-MCM method selects clustering variables $\hat{S} = \{4-7, 9-18\}$, redundant variables $\hat{U} = \{2,3,8,19,20,38\}$, and independent variables $\hat{W} = \{1, 21-37, 39, 40\}$. For the SparseKmeans, all variables are selected because all weights are positive. The Adjusted Rand Index between the RD-MCM clustering and the three waves is 0.25, while for SparseKmeans it is 0.23. Thus RD-MCM performs slightly better in terms of classification, while using fewer variables..

Note that the RD-MCM method can select not only the clustering variables, but also the number of clusters, if this is unknown. This is unlike SparseKmeans, which requires the number of clusters to be assumed or known in advance. If we do not assume the number of clusters known here, the RD-MCM method selects $K = 6$ clusters with clustering variables $\hat{S} = \{4-18\}$, redundant variables $\hat{U} = \{2,3,19,20,38\}$ that are explained by the predictor variables $\hat{R} = \{5-7, 9-12, 14, 15, 17\}$, and independent variables $\hat{W} = \{1, 21-37, 39, 40\}$.

We compare these results with the SparseKmeans method where $K = 6$ is fixed. In this case also, SparseKmeans selects all the variables. The ARI between the RD-MCM six-cluster solution and the three waves is 0.257, which is actually slightly better than with the three-cluster solution; and is 0.27 for SparseKmeans, also better than when the number of clusters is taken to be three. This suggests that in some cases there may be some advantage to selecting the number of clusters based on the data, even when the true number of clusters is known a priori.

4.2. Transcriptome data

We turn to a transcriptome dataset of *Arabidopsis thaliana*, extracted from the database CATdb (Gagnot et al., 2008) which has been considered by Maugis et al. (2009c) for clustering. This dataset consists of $n = 4616$ genes described by $p = 33$ biotic stress experiments (partitioned

 ISSN: 2102-6238

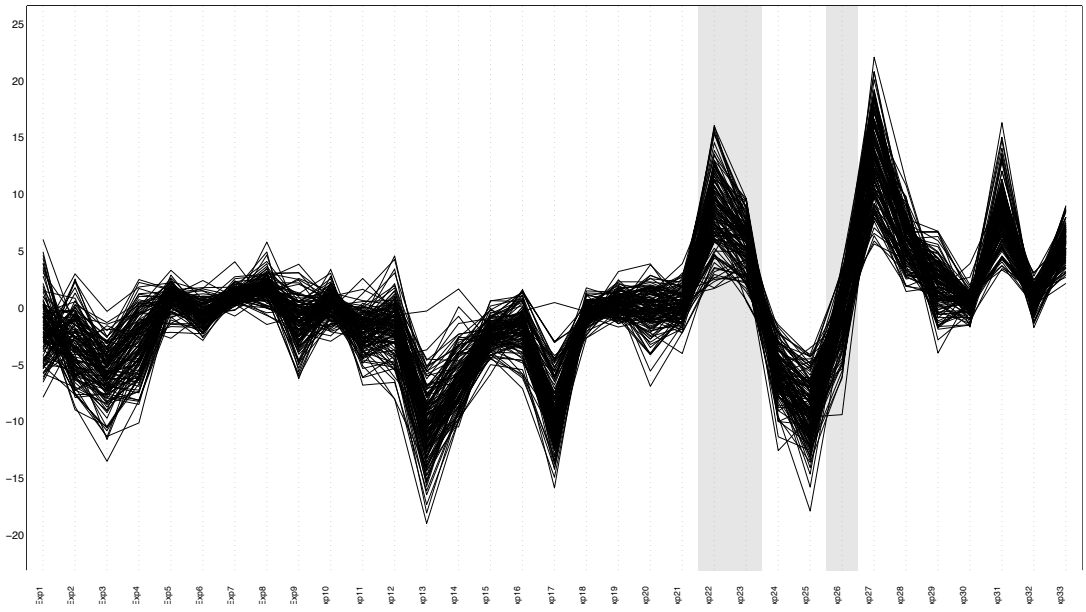

FIGURE 6. *Transcriptome Data: Gene Profiles in Cluster 19*

into 9 biological projects). Each gene is described by a vector $y_i \in \mathbb{R}^{33}$, where y_{ij} is the test statistic from the experiment j for the differential analysis. For details of the normalization and the differential analysis steps, see Gagnot et al. (2008).

We first analyzed these data using the RD-MCM method. Based on previous analyses of transcriptome datasets (Maugis et al., 2009a), we allowed the number of mixture components to vary between 10 and 30, and we considered the mixture forms EEE (identical variance matrices) and VEE (different volumes and the same orientation and shape for the variance matrices) in the notation of Fraley and Raftery (2002). The method selected the VEE model with $\hat{K} = 26$ components. Experiments 22, 23 and 26, which come from the same biological project, were identified as redundant and the other 30 experiments were declared relevant for clustering. The three redundant experiments were explained by the relevant variables $\hat{R} = \{1,5,8,13,24,25,27,28-31,33\}$. They were mainly explained by experiments 24, 25 and 27, which come from the same biological project. The sizes of the 26 clusters varied widely, from 45 to 602 genes per cluster.

As an example, Figure 6 shows the expression profiles of the genes in cluster 19, which are quite homogeneous and clearly coexpressed. This corresponds to other biological information, because most of the genes in cluster 19 are in the cell nucleus. In addition, the biological function of 101 of the 129 genes in cluster 19 is linked to transcription. Of the remaining 28 genes, the role of eight is unknown according to the functional annotation. Thus this cluster analysis can potentially shed light on the biological functions of these eight genes for which this is not currently known (Maugis et al., 2009c).

The results from the SparseKmeans and K-means procedures were quite different from one another. Repeating the SparseKmeans procedure from random initial positions with $K = 26$ led to an average ARI of $0.135(SD\ 0.002)$ compared with the RD-MCM clustering. For instance, the 129 genes of cluster 19 of the RD-MCM solution shown in Figure 6 were divided between cluster

 ISSN: 2102-6238

8 (100 genes among 193), cluster 21 (3 genes among 358), cluster 22 (1 gene among 224) and cluster 26 (25 genes among 320) in the "best" solution provided by SparseKmeans. It was also surprising to see that the ARI between SparseKmeans and K-means was rather low at 0.349. On the other hand, the ARI between the RD-MCM partition and the partition obtained with the VEE model with $\hat{K} = 26$ components without selecting the variables was higher (0.578), suggesting more stable results.

5. Discussion

We have carried out a comparison between model selection and regularization approaches to variable selection in model-based clustering. These are two of the leading approaches. For each general approach we have selected a specific method based on results from previous results in the literature. The model selection method is the method of Maugis et al. (2009b) which modified that of Raftery and Dean (2006); we refer to it as the RD-MCM method. The regularization method is the SparseKmeans method of Witten and Tibshirani (2010). We also compared them with K-means without variable selection.

We compared the methods by simulation in terms of their accuracy in both classification and variable selection. In the first simulation experiment all the variables were conditionally independent given cluster membership. We found that variable selection (of either kind) yielded substantial gains in classification accuracy when the clusters were well separated, but few gains when the clusters were close together. We found that the two variable selection methods had comparable classification accuracy, but that the model selection approach had substantially better accuracy in selecting variables.

In our second simulation experiment, there were correlations among the variables given the cluster memberships. We found that the model selection approach was substantially more accurate in terms of both classification and variable selection than the regularization approach, and that both gave more accurate classifications than K-means without variable selection.

Another advantage of the model selection approach is that it allows one to select the number of clusters based on the data, while the regularization approach requires that it be known or specified in advance by the user. It also allows one to select among a range of models for the covariance structure. Also, we found that the results of SparseKmeans were quite sensitive to the tuning parameter. However, for computational and numerical reasons, the model selection approach is not well adapted to ill-posed high dimension cases: typically cases where the number of observations is smaller than the number of variables and where the number of variables is huge (say $p > 1000$).

There are several other recent proposals for variable selection in model-based clustering. Fraiman et al. Fraiman et al. (2008) proposed a method for variable selection *after* the clustering has been carried out that also assumes the number of clusters known. Thus it is not fully comparable with the methods considered here, which carry out clustering and variable selection simultaneously. Nia and Davison Nia and Davison (2012) proposed a fully Bayesian approach using a spike and slab prior, while Kim et al. Kim et al. (2012) proposed a Bayesian approach using Bayes factors to compare the different models. Lee and Li Lee and Li (2012) proposed an approach to variable selection for model-based clustering using ridgelines.

It is also worth mentioning the dimension reduction methods of Bouveyron et al. (2007); McNicholas and Murphy (2008); McLachlan et al. (2011); Scrucca (2010) for model-based

 ISSN: 2102-6238

clustering. Their goal is not variable selection. Also, Poon et al. Poon et al. (2013) proposed a method for facet determination in model-based clustering; this is related to but not the same as variable selection. There are also several other recent proposals for variable selection through regularisation. See for instance Sun et al. (2012); Bouveyron and Brunet (2013a); Galimberti et al. (2009). Finally, it is worth mentioning the review paper Bouveyron and Brunet (2013b).

Acknowledgements: Raftery's research was supported by NIH grants R01 GM084163, R01 HD054511 and R01 HD070936, as well as by Science Foundation Ireland E.T.S. Walton Visitor Award 11/W.1/I2079.

References

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

Bouveyron, C. and Brunet, C. (2013a). Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. *Computational Statistics*. to appear.

Bouveyron, C. and Brunet, C. (2013b). Model-based clustering of high-dimensional data : A review. *Computational Statistics and Data Analysis*. to appear.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52:502–519.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International, Belmont, California.

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

Celeux, G., Martin-Magniette, M.-L., Maugis-Rabusseau, C., and Raftery, A. E. (2011). Letter to the editor. *Journal of the American Statistical Association*, 105:383.

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

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society, Series B*, 66:815–849.

Gagnot, S., Tamby, J.-P., Martin-Magniette, M.-L., Bitton, F., Taconnat, L., Balzergue, S., Aubourg, S., Renou, J.-P., Lecharny, A., and Brunaud, V. (2008). CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Research*, 36:986–990.

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

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66:793–804.

Hubert, L. J. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Kim, S., Song, D. K. H., and DeSarbo, W. S. (2012). Model-based segmentation featuring simultaneous segment-level variable selection. *Journal of Marketing Research*, 49:725–736.

Law, M. H., Figueiredo, M. A. T., and Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1154–1166.

Lee, H. and 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., and Martin-Magniette, M.-L. (2009a). Variable selection for clustering with Gaussian mixture models. *Biometrics*, 65:701–709.

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

Maugis, C., Martin-Magniette, M.-L., Tamby, J.-P., Renou, J.-P., Lecharny, A., Aubourg, S., and Celeux, G. (2009c). Sélection de variables pour la classification par mélanges gaussiens pour prédire la fonction des gènes orphelins. *La Revue Modulad*, 40:69–80.

 ISSN: 2102-6238

McLachlan, G., Baek, J., and Rathnayake, S. I. (2011). Mixtures of factor analyzers for the analysis of high-dimensional data. In Mengersen, K., Robert, C., and Titterington, D., editors, *Mixture Estimation and Applications*, pages 171–191. New Jersey: Wiley.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

McNicholas, P. and Murphy, T. (2008). Parsimonious Gaussian Mixture Models. *Statistics and Computing*, 18:285–296.

Nia, V. P. and Davison, A. C. (2012). High-dimensional Bayesian clustering with variable selection: The R package bclust. *Journal of Statistical Software*, 47:Issue 5.

Pan, W. and Shen, X. (2007). Penalized Model-Based Clustering with Application to Variable Selection. *Journal of Machine Learning Reserach*, 8:1145–1164.

Poon, L. K. M., Zhang, N. L., and Liu, A. H. (2013). Model-based clustering of high-dimensional data: Variable selection versus facet determination. *International Journal of Approximate Reasoning*, 54:196–215.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101:168–178.

Scrucca, L. (2010). Dimension reduction for model-based clustering. *Statistics and Computing*, 20:471–484.

Steinley, D. and Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, 73:125–144.

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

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

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 63:411–423.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, 64:440–448.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of American Statistical Association*, 105:713–726.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350.

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

 ISSN: 2102-6238