BIOMETRICS 65, 701–709
September 2009

DOI: 10.1111/j.1541-0420.2008.01160.x

# Variable Selection for Clustering with Gaussian Mixture Models

**Cathy Maugis,[1,*] Gilles Celeux,[2] and Marie-Laure Martin-Magniette[3,4]**

[1]Department of Mathematics, University Paris-Sud 11, Orsay, France
[2]Inria Saclay—Ile de France, France
[3]UMR AgroParisTech/INRA, MIA 518, Paris, France
[4]URGV UMR INRA 1165, CNRS 8114, UEVE, Evry, France
[*]*email:* Cathy.Maugis@math.u-psud.fr

SUMMARY. This article is concerned with variable selection for cluster analysis. The problem is regarded as a model selection problem in the model-based cluster analysis context. A model generalizing the model of Raftery and Dean (2006, *Journal of the American Statistical Association* **101,** 168–178) is proposed to specify the role of each variable. This model does not need any prior assumptions about the linear link between the selected and discarded variables. Models are compared with Bayesian information criterion. Variable role is obtained through an algorithm embedding two backward stepwise algorithms for variable selection for clustering and linear regression. The model identifiability is established and the consistency of the resulting criterion is proved under regularity conditions. Numerical experiments on simulated datasets and a genomic application highlight the interest of the procedure.

KEY WORDS: Bayes' factor; BIC; Linear regression; Model-based clustering; Variable selection.

## 1. Introduction

The goal of clustering methods is to discover structures (clusters) among individuals described by several variables. Many clustering methods exist and roughly fall into two categories. The first one is based on similarity or dissimilarity distances. It gathers hierarchical clusterings, which build trees and also methods like $K$-means algorithm, which classify data through a number of clusters fixed a priori. The second category is model-based methods that consist of using a model for clusters and optimizing the fit between the data and the model. In practice, each cluster is represented by a parametric distribution, like a Gaussian one and the entire dataset is modeled by a mixture of these distributions. An advantage of model-based clustering is to provide a rigorous framework to assess the number of clusters and the role of each variable in the clustering process.

In principle, the more information we have about each individual, the better a clustering method is expected to perform. However, the structure of interest may often be contained in a subset of the available variables and a lot of variables may be useless or even harmful to detect a reasonable clustering structure. It is thus important to select the relevant variables from the cluster analysis viewpoint. It is a recent research topic in contrast to variable selection in regression and classification models (Miller, 1990; Kohavi and John, 1997; Guyon and Elisseeff, 2003). This new interest for variable selection in clustering comes from the increasingly frequent use of these methods on high-dimensional datasets, such as transcriptome datasets. It is usually considered that coexpressed genes are often implicated in the same biological function and consequently are potential candidates to be coregulated genes (see, for instance, Sharan, Elkon, and Shamir [2002] or Jiang, Tang, and Zhang [2004], and references therein). Because the number of transcriptome experiments always increases, an experiment selection in the clustering procedure is desirable to reveal important biological phenomena.

Three types of approaches dealing with variable selection in clustering have been proposed. The first one includes clustering methods with weighted variables (see, for instance, Friedman and Meulman, 2004) and dimension reduction methods. For this latter, McLachlan, Bean, and Peel (2002) use a mixture of factor analyzers to reduce the extremely high dimensionality of a gene expression problem. A suitable Gaussian mixture family is considered in Bouveyron, Girard, and Schmid (2007) to take into account the dimension reduction and the data clustering simultaneously. In contrast to this first method type, the two last approaches select explicitly relevant variables. The so-called "filter" approaches select the variables before a clustering analysis (see, for instance, Dash et al., 2002; Jouve and Nicoloyannis, 2005). Their main weakness is the influence of independent selection step on the clustering results. In contrast, the so-called "wrapper" approaches combine variable selection and clustering. For distance-based methods, one can cite Fowlkes, Gnanadesikan, and Kettenring (1988) for a forward selection approach with complete linkage hierarchical clustering, Devaney and Ram (1997) who propose a stepwise algorithm where the quality of the feature subsets is measured with the COBWEB algorithm, or the method of Brusco and Cradit (2001) based on the adjusted Rand index for $K$-means clustering. There exist also wrapper methods in the model-based clustering setting. When the number of variables is greater than the number of individuals, Tadesse, Sha, and Vannucci (2005) propose a fully Bayesian method using a reversible

jump algorithm to simultaneously choose the number of mixture components and select variables. Kim, Tadesse, and Vannucci (2006) use a similar approach by formulating clustering in terms of Dirichlet process mixtures. In Gaussian mixture model clustering, Law, Figueiredo, and Jain (2004) propose to evaluate the importance of the variables in the clustering process via "feature saliencies" and use the *minimum message length* criterion. Raftery and Dean (2006) recast the problem of comparing two nested variable subsets as a model comparison problem and address it using Bayes' factor. An interesting aspect of their model formulation is that irrelevant variables are not required to be independent of the clustering variables. They thus avoid the unrealistic independence assumption between the relevant and irrelevant variables for the clustering, considered in Tadesse et al. (2005), Kim et al. (2006), and Law et al. (2004). In their model, the whole irrelevant variable subset depends on the whole relevant variables through a linear regression equation. However, some relevant variables are not necessarily required to explain all irrelevant variables in the linear regression and their introduction involves additional parameters without a significant increase of the log likelihood.

In this article, we improve their method by considering another type of relation between the irrelevant variables for clustering and the relevant ones. We consider that the irrelevant variables can be independent of some relevant variables. This modeling allows us to improve the clustering and its interpretation. Our variable selection implementation is based on a backward stepwise algorithm. Moreover, we look at a more general situation where the variables are partitioned into blocks which cannot be split.

The article is organized as follows: Gaussian mixture models for clustering are reviewed in Section 2. Our variable selection approach is presented in Section 3. The associated search algorithm is described in Section 4. The model identifiability and the consistency of the variable selection criterion are proved in Section 5. Simulated experiments are presented to validate the method and to compare it with Raftery and Dean's approach in Section 6. An example of transcriptome data clustering is addressed in Section 7. Finally, a discussion on the overall method is given in Section 8.

## 2. Multivariate Gaussian Models and Clustering

Model-based clustering consists of assuming that data come from several subpopulations modeled separately and the overall population is a mixture of these subpopulations. The resulting model is a finite mixture model. When data are multivariate continuous observations, the parameterized component density is usually a multidimensional Gaussian density. We consider $n$ individuals $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$ described by $Q$ variables ($\mathbf{y}_i'$ in $\mathbb{R}^Q$). Observations are assumed to be a sample from a probability distribution with density

$$f(\mathbf{y}_i \mid K, \alpha) = \sum_{k=1}^{K} p_k \phi(\mathbf{y}_i \mid \mu_k, \Sigma_k),$$

where the $p_k$'s are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} p_k = 1$), and $\phi(. \mid \mu_k, \Sigma_k)$ denotes the $Q$-dimensional Gaussian density with mean $\mu_k$ and variance matrix $\Sigma_k$. The vector parameter is denoted $\alpha = (p_1, \ldots, p_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$. The mixture model is an incomplete data structure model: The complete data are $((\mathbf{y}_1, \mathbf{z}_1)', \ldots, (\mathbf{y}_n, \mathbf{z}_n)')'$, where the missing data are $\mathbf{z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')'$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ being binary vectors such that $z_{ik} = 1$ iff $\mathbf{y}_i$ arises from the $k$th subpopulation. The $\mathbf{z}$'s define an ideal clustering of the data $\mathbf{y}$, associated to the mixture model.

As in Banfield and Raftery (1993) and Celeux and Govaert (1995), the mixture component variance matrix can be decomposed into $\Sigma_k = L_k D_k' A_k D_k$, where $L_k = |\Sigma_k|^{1/Q}$, $D_k$ is the $\Sigma_k$'s eigenvector matrix and $A_k$ is the diagonal matrix of normalized eigenvalues of $\Sigma_k$. They control, respectively, the volume, the orientation, and the shape of the $k$th cluster. According to constraints required on the different elements of this decomposition, a collection of parsimonious and interpretable models is available. Moreover, the proportions can be assumed to be equal or free. Finally, the considered model family is

$$\mathcal{T} = \{(K, m) \in \{2, \ldots, K_{\max}\} \times \mathcal{M}\}$$

where $\mathcal{M}$ is a collection of 28 models, described in Web Table 1, and $K_{\max}$ is the maximum number of clusters specified by the user. Those 28 models are available in the MIXMOD software (Biernacki et al., 2006) and, for most of them, in the MCLUST software (Fraley and Raftery, 2003).

In this inferential framework, it is possible to choose one of the models $(K, m) \in \mathcal{T}$, by using model selection methods or criteria (see McLachlan and Peel, 2000). In a Bayesian perspective, the model maximizing the posterior probability $P[(K, m) \mid \mathbf{y}]$ is to be chosen. By Bayes' theorem

$$P[(K, m) \mid \mathbf{y}] = \frac{f(\mathbf{y} \mid K, m) P[(K, m)]}{f(\mathbf{y})},$$

and supposing a noninformative uniform prior distribution $P[(K, m)]$ on the models, it leads to $P[(K, m) \mid \mathbf{y}] \propto f(\mathbf{y} \mid K, m)$. Thus the chosen model satisfies

$$(\tilde{K}, \tilde{m}) = \arg\max_{(K,m) \in \mathcal{T}} f(\mathbf{y} \mid K, m),$$

where the integrated likelihood $f(\mathbf{y} \mid K, m)$ is defined by

$$f(\mathbf{y} \mid K, m) = \int f(\mathbf{y} \mid K, m, \alpha) \pi(\alpha \mid K, m)\, d\alpha,$$

$\pi(\alpha \mid K, m)$ being the prior distribution of the vector parameter $\alpha$ of the $(K, m)$ model (Kass and Raftery, 1995). Because this integrated likelihood is typically difficult to calculate, an asymptotic approximation of $2 \ln \{f(\mathbf{y} \mid K, m)\}$ is generally used. This approximation is the Bayesian information criterion (BIC) defined by

$$\mathrm{BIC}_{\mathrm{clust}}(\mathbf{y} \mid K, m) = 2 \ln\{f(\mathbf{y} \mid K, m, \hat{\alpha})\} - \lambda_{(K,m)} \ln(n), \quad (1)$$

where $\lambda_{(K,m)}$ is the number of free parameters for the $(K, m)$ model and $f(\mathbf{y} \mid K, m, \hat{\alpha})$ is the maximum likelihood under this model (Schwarz, 1978). In this perspective, the selected model is

$$(\hat{K}, \hat{m}) = \arg\max_{(K,m) \in \mathcal{T}} \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y} \mid K, m).$$

For deriving $(\hat{K}, \hat{m})$, the maximum likelihood estimate (mle) $\hat{\alpha}$ is computed using generally the expectation–maximization algorithm (Dempster, Laird, and Rubin, 1977). Finally, the clustering is performed using the maximum a posteriori rule defined by

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \hat{p}_k \phi(\mathbf{y}_i \mid \hat{\mu}_k, \hat{\Sigma}_k) > \hat{p}_l \phi(\mathbf{y}_i \mid \hat{\mu}_l, \hat{\Sigma}_l), \forall l \neq k, \\ 0 & \text{otherwise.} \end{cases}$$

Here all the $Q$ variables are supposed to enter in the mixture models. When there are numerous variables, it can be sensible to choose which variables are actually required in the mixture models. This can be regarded as a model selection problem as well.

## 3. Selecting Variables

The approach we propose for selecting relevant variables for clustering is related to the Raftery and Dean (2006) approach that is sketched first. Their idea is to divide the variable set into a subset of relevant clustering variables and its complementary, which does not provide information for the clustering but which depends on the relevant variables through a linear regression. This is an interesting aspect because they avoid the unrealistic and usual independence assumption between the relevant and irrelevant variables for the clustering. As they stressed, the independence assumption would often lead to wrongly declaring a variable as relevant for the clustering because this variable is related to some relevant variables, but not necessarily to the clustering itself. Although relevant variables are not all required to explain the whole irrelevant variable subset, Raftery and Dean (2006) force them to enter in the regression model. This involves additional parameters in the model without necessarily leading to a significant increase of its log likelihood. One consequence is that models assigning some irrelevant variables as relevant could be wrongly preferred when model comparison is performed with Bayes' factor or penalized likelihood criteria. In this article, we opt for a more realistic model where irrelevant variables are explained by a subset of relevant variables. Moreover, we consider a more general framework where the $Q$ variables are partitioned into $T$ blocks. That is, there exists a function $\Psi$ such that each variable $j \in \{1, \ldots, Q\}$ belongs to a unique variable block $\Psi(j) \in \{1, \ldots, T\}$. This common situation appears for instance in the genomic application considered in Section 7. Obviously in the standard situation where each block reduces to a single variable, we have $T = Q$ and all the following formula can be straightforwardly particularized to this simple case. Throughout the article, $\mathbf{y}^j = (y_1^j, \ldots, y_n^j)'$ and for a subset of variable blocks $A$, $\mathbf{y}^A$ denotes the set $\{\mathbf{y}^j \in \mathbb{R}^n; \Psi(j) \in A\}$ and $\text{card}(A) = \text{card}\{j; \Psi(j) \in A\}$.

Let $\mathcal{F}$ be the family of variable block index subset, $S \in \mathcal{F}$ the set of relevant clustering variable block indexes and, $S^c$ its complement in $\{1, \ldots, T\}$, denoting the irrelevant variables. In order to distinguish the role of each clustering variable block in the regression, those entering in the regression equation of the irrelevant variables constitute the subset $R$. The information is thus summarized into a couple $(S, R)$ belonging to $\mathcal{V} = \{(S, R); (S, R) \in \mathcal{F}^2,\ S \neq \emptyset,\ R \subseteq S\}$. This division of the variable block roles is illustrated in Figure 1. Finally, the considered model set is defined by $\mathcal{N} = \{(K, m, S, R); (K, m) \in \mathcal{T}; (S, R) \in \mathcal{V}\}$.

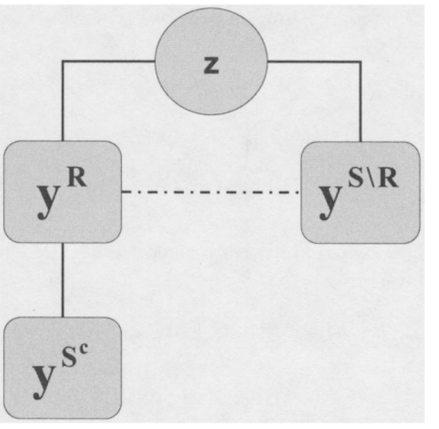


**Figure 1.** Graphical representation of the variable repartition in the model.

The models in competition are compared with their integrated likelihoods decomposed into two multiplicative parts

$$f(\mathbf{y} \mid K, m, S, R) = f_{\text{clust}}(\mathbf{y}^S \mid K, m) f_{\text{reg}}(\mathbf{y}^{S^c} \mid \mathbf{y}^R).$$

The function $f_{\text{clust}}(\mathbf{y}^S \mid K, m) = \int f_{\text{clust}}(\mathbf{y}^S \mid K, m, \alpha) \pi(\alpha \mid K, m, S)\, d\alpha$ is the integrated likelihood of the $(K, m)$ mixture model on the relevant clustering variables. The function $f_{\text{reg}}(\mathbf{y}^{S^c} \mid \mathbf{y}^R) = \int f_{\text{reg}}(\mathbf{y}^{S^c} \mid \mathbf{y}^R, B, \Omega)\, \pi(B, \Omega \mid S^c, R)\, dB\, d\Omega$ is the integrated likelihood of multidimensional regression of the irrelevant variables on a subset of the clustering variables, $B$ denoting the vector of regression coefficients and $\Omega$ the variance matrix (see Web Appendix C). In practice, the integrated likelihoods are approximated using the BIC approximation as in equation (1), and the chosen model is

$$(\hat{K}, \hat{m}, \hat{S}, \hat{R}) = \underset{(K,m,S,R) \in \mathcal{N}}{\arg\max} \left\{ \text{BIC}_{\text{clust}}(\mathbf{y}^S \mid K, m) + \text{BIC}_{\text{reg}}(\mathbf{y}^{S^c} \mid \mathbf{y}^R) \right\}, \quad (2)$$

where

$$\text{BIC}_{\text{clust}}(\mathbf{y}^S \mid K, m) = 2 \ln \left\{ f_{\text{clust}}(\mathbf{y}^S \mid K, m, \hat{\alpha}) \right\} - \lambda^S_{(K,m)} \ln(n),$$

and

$$\text{BIC}_{\text{reg}}(\mathbf{y}^{S^c} \mid \mathbf{y}^R) = 2 \ln \left\{ f_{\text{reg}}(\mathbf{y}^{S^c} \mid \mathbf{y}^R, \hat{B}, \hat{\Omega}) \right\} - \nu_{(S,R)} \ln(n), \quad (3)$$

$\lambda^S_{(K,m)}$ being the number of free parameters of the $(K, m)$ mixture model with $\text{card}(S)$ variables, $(\hat{B}, \hat{\Omega})$ the maximum likelihood estimate of the regression parameters, and $\nu_{(S,R)} = \{\text{card}(R) + 1\} \text{card}(S^c) + \frac{\text{card}(S^c)\{\text{card}(S^c)+1\}}{2}$. The computation of BIC for multidimensional multivariate regression is detailed in Web Appendix C.

## 4. The Variable Selection Procedure

The number of models in $\mathcal{N}$ is $28(K_{\max} - 1) \sum_{t=1}^{T} \binom{T}{t} \sum_{l=0}^{t} \binom{t}{l}$, where $K_{\max}$ is the maximum number of clusters. Thus an exhaustive search of the optimal model is impossible in most

situations. The algorithm we propose is a two-nested-step algorithm.

(1) For all $(K, m)$, we search

$$(\hat{S}(K,m), \hat{R}(K,m)) = \arg\max_{(S,R)\in\mathcal{V}} \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S \mid K, m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c} \mid \mathbf{y}^R) \right\}$$

by a backward stepwise procedure detailed hereafter.

(2) We determine

$$(\hat{K}, \hat{m}) = \arg\max_{(K,m)\in\mathcal{T}} \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^{\hat{S}(K,m)} \mid K, m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{\hat{S}^c(K,m)} \mid \mathbf{y}^{\hat{R}(K,m)}) \right\}.$$

Finally, the selected model is $(\hat{K}, \hat{m}, \hat{S}(\hat{K}, \hat{m}), \hat{R}(\hat{K}, \hat{m}))$.

We opt for a backward stepwise selection algorithm. It means that all the variables are selected at the beginning and at each step, a variable block is excluded or included.

### 4.1 *The Models in Competition*

At each step of this algorithm, the variable set $\{1, \ldots, T\}$ is divided into three subgroups: $S$ the set of selected clustering variable blocks, $j$ the candidate variable block being considered for inclusion into or exclusion from the set of clustering variables, and $U$ the irrelevant variable set. The integrated likelihood can be thus decomposed into

$$f(\mathbf{y} \mid K, m) = f(\mathbf{y}^U \mid \mathbf{y}^j, \mathbf{y}^S) f(\mathbf{y}^j, \mathbf{y}^S \mid K, m).$$

The decision of exclusion (respectively, inclusion) of variable block $j$ from (respectively, in) the set of clustering variables is made by the comparison of the following two models:

(1) $M_1(K, m)$ specifies that given $\mathbf{y}^S$, $\mathbf{y}^j$ does not provide additional information for the clustering and is explained by a subset $\mathbf{y}^{R[j]}$ of $\mathbf{y}^S$,

$$\begin{aligned} f_1(\mathbf{y}^j, \mathbf{y}^S \mid K, m) &= \sum_{\mathbf{z}} f(\mathbf{y}^j, \mathbf{y}^S \mid \mathbf{z}, K, m) f(\mathbf{z} \mid K, m) \\ &= \sum_{\mathbf{z}} f_{\mathrm{reg}}(\mathbf{y}^j \mid \mathbf{y}^{R[j]}) f(\mathbf{y}^S \mid \mathbf{z}, K, m) f(\mathbf{z} \mid K, m) \\ &= f_{\mathrm{reg}}(\mathbf{y}^j \mid \mathbf{y}^{R[j]}) f_{\mathrm{clust}}(\mathbf{y}^S \mid K, m). \end{aligned}$$

(2) $M_2(K, m)$ specifies that given $\mathbf{y}^S$, $\mathbf{y}^j$ provides additional information for the clustering,

$$f_2(\mathbf{y}^j, \mathbf{y}^S \mid K, m) = f_{\mathrm{clust}}(\mathbf{y}^j, \mathbf{y}^S \mid K, m).$$

The two models are compared with the following Bayes' factor:

$$B_{12}(K, m) = \frac{f_{\mathrm{reg}}(\mathbf{y}^j \mid \mathbf{y}^{R[j]}) f_{\mathrm{clust}}(\mathbf{y}^S \mid K, m)}{f_{\mathrm{clust}}(\mathbf{y}^j, \mathbf{y}^S \mid K, m)}.$$

Because the integrated likelihoods are difficult to compute, $-2 \ln B_{12}(K, m)$ is approximated by

$$\begin{aligned} \mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j \mid K, m) = \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S, \mathbf{y}^j \mid K, m) - \{ &\mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S \mid K, m) \\ &+ \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^j \mid \mathbf{y}^{R[j]}) \}. \end{aligned} \quad (4)$$

If $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j \mid K, m)$ is positive, Model $M_2$ is chosen, otherwise, Model $M_1$ is chosen.

### 4.2 *The Backward Stepwise Selection Algorithm*

Let $(K, m)$ be fixed, this algorithm is making use of an exclusion and an inclusion procedure now described. The decision of excluding (respectively, including) a variable block from (respectively, into) the set of clustering variables is based on the comparison of the two models with the BIC approximation of the Bayes' factor.

*Initialization* $S = \{1, \ldots, T\}$, $j_E = \emptyset$ and $j_I = \emptyset$.

**Exclusion step** In this step, the proposed variable block for removal from the set of currently selected clustering variables is chosen to be the variable block, which gives the smallest value of $\mathrm{BIC}_{\mathrm{diff}}$ defined in (4). It is as follows:

For all $j$ in $S$, use the backward stepwise selection algorithm, described in the Appendix, to choose the subset $R[j]$ of dependent variables for the regression of $\mathbf{y}^j$ on $\mathbf{y}^{S-j}$, and compute $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j \mid K, m)$. Then, compute

$$j_E = \arg\min_{j\in S} \mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j \mid K, m).$$

- If $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^{j_E} \mid K, m) \leqslant 0, S = S - j_E$ and: Stop if $j_E = j_I$, otherwise go to the inclusion step;
- otherwise: Stop if $j_I = \emptyset$ or go to the inclusion step otherwise.

**Inclusion step** In this step, the proposed new variable block is chosen to be the variable block which gives the greatest value of $\mathrm{BIC}_{\mathrm{diff}}$. It is as follows:

For all $j$ in $S^c$, use the backward stepwise selection algorithm, described in the Appendix, to choose the subset $R[j]$ of dependent variables for the regression of $\mathbf{y}^j$ on $\mathbf{y}^S$, and compute $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j \mid K, m)$. Then, compute

$$j_I = \arg\max_{j\in S^c} \mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^j \mid K, m).$$

- If $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^{j_I} \mid K, m) > 0$, stop if $j_I = j_E$, otherwise $S = S \cup j_I$ and go to the exclusion step;
- otherwise go to the exclusion step.

Starting from the exclusion step, the backward variable selection algorithm consists of alternating exclusion and inclusion steps. It returns the relevant clustering variable subset $\hat{S}(K, m)$. Next, $\hat{R}(K, m)$ is obtained using the backward stepwise algorithm for the regression of $\mathbf{y}^{S^c}$ on $\mathbf{y}^S$ (see Appendix).

## 5. Theoretical Properties

In this section, a necessary and sufficient condition is given to ensure the model identifiability and a consistency theorem of the criterion is stated. In model $(K, m, S, R)$, the parameterized densities are denoted $f(. \mid \theta)$, where $\theta = (\alpha, B, \Omega) \in \Upsilon_{(K,m,S,R)}$ in the sequel.

***Identifiability.*** The model identifiability is based on the following remark: Let $s$ be a nonempty subset included strictly into $S$ and $\bar{s}$ be its complement in $S$, then the density $f(. \mid \theta)$

under the model $(K, m, S, R)$ can be decomposed as

$$f(x\,|\,\theta) = \sum_{k=1}^{K} p_k \phi\big(x^S \,\big|\, \mu_k, \Sigma_k\big)\phi\big(x^{S^c} \,\big|\, x^R, B, \Omega\big)$$

$$= \phi\big(x^{S^c}\,|\,x^R, B, \Omega\big) \sum_{k=1}^{K} p_k \phi\big(x^s \,\big|\, \mu_{ks}, \Sigma_{k,ss}\big)$$

$$\times \phi\big(x^{\bar{s}} \,\big|\, \mu_{k,\bar{s}|s} + x^s \Sigma_{k,\bar{s}|s}, \Sigma_{k,\bar{s}\bar{s}|s}\big)$$

where mixture parameters are decomposed into $\mu_k = (\mu_{ks}, \mu_{k\bar{s}})$, and $\Sigma_k$ into submatrices $\Sigma_{k,ss}, \Sigma_{k,s\bar{s}}$, and $\Sigma_{k,\bar{s}\bar{s}}$ (according to Theorem 2.5.1 of Anderson, 2003, p. 35). The conditional parameters are defined by $\mu_{k,\bar{s}|s} = \mu_{k\bar{s}} - \mu_{ks}\Sigma^{-1}_{k,ss}\Sigma_{k,s\bar{s}}, \Sigma_{k\bar{s}|s} = \Sigma^{-1}_{k,ss}\Sigma_{k,s\bar{s}}$ and $\Sigma_{k,\bar{s}\bar{s}|s} = \Sigma_{k,\bar{s}\bar{s}} - \Sigma_{k,\bar{s}s}\Sigma^{-1}_{k,ss}\Sigma_{k,s\bar{s}}$. If these parameters $\mu_{k,\bar{s}|s}, \Sigma_{k,\bar{s}|s}$, and $\Sigma_{k,\bar{s}\bar{s}|s}$ are identical for all clusters, the identifiability cannot be ensured because the regression density of $\bar{s}$ on $s$ can be factorized from the Gaussian mixture and regrouped with the regression density of $S^c$ on $R$. This remark leads to the following identifiability theorem.

THEOREM 1: *Let $\Theta_{(K,m,S,R)}$ be a subset of $\Upsilon_{(K,m,S,R)}$ whose elements $\theta$ contain distinct couples $(\mu_k, \Sigma_k)$ fulfilling*

$$\forall s \subsetneq S, \exists (k,k'), 1 \leqslant k < k' \leqslant K; \mu_{k,\bar{s}|s} \neq \mu_{k',\bar{s}|s} \quad \text{or} \quad \Sigma_{k,\bar{s}|s} \neq \Sigma_{k',\bar{s}|s} \quad \text{or} \quad \Sigma_{k,\bar{s}\bar{s}|s} \neq \Sigma_{k',\bar{s}\bar{s}|s}. \tag{5}$$

*Let $(K, m, S, R)$ and $(K^\star, m^\star, S^\star, R^\star)$ be two models. If there exist $\theta \in \Theta_{(K,m,S,R)}$ and $\theta^* \in \Theta_{(K^\star,m^\star,S^\star,R^\star)}$ such that $f(.\,|\,\theta) = f(.\,|\,\theta^\star)$ then $(K, m, S, R) = (K^\star, m^\star, S^\star, R^\star)$ and $\theta = \theta^\star$ (up to a permutation of mixture components).*

Proof steps are given below and a complete proof can be found in Web Appendix A.

*Sketch of the proof.* First, each density $f(.\,|\,\theta)$ is written as a Gaussian mixture allowing us to use the identifiability of Gaussian mixture models. Thus, $K = K^\star$ and the parameters of the Gaussian mixture are equal, up to a permutation of mixture components. Second, it is proved by contradiction that $S \cap S^\star \neq \emptyset$ because the $(\mu_k, \Sigma_k)$'s are not identical. And, it is deduced from (5) that the only possible case is $S = S^\star$ leading to $m = m^\star, R = R^\star$, and, $\theta = \theta^\star$.

***Consistency of our criterion.*** In this paragraph, it is proved that the probability of selecting the true couple $(S_0, R_0)$ by maximizing criterion (2) approaches 1 as $n \to \infty$ when the sampling distribution is one of the densities in competition and the true mixture model $(K_0, m_0)$ is known. Denoting $h$ the density function of the sample $\mathbf{y}$

$$\theta^\star_{(K,m,S,R)} = \underset{\theta_{(K,m,S,R)} \in \Theta_{(K,m,S,R)}}{\arg\min} \mathrm{KL}[h, f(.\,|\,\theta_{(K,m,S,R)})]$$

$$= \underset{\theta_{(K,m,S,R)} \in \Theta_{(K,m,S,R)}}{\arg\max} \mathbb{E}_X\{\ln f(X\,|\,\theta_{(K,m,S,R)})\},$$

where $\mathrm{KL}[h, f] = \int \ln\left\{\frac{h(x)}{f(x)}\right\} h(x)dx$ is the Kullback-Leibler divergence between the densities $h$ and $f$, and

$$\hat{\theta}_{(K,m,S,R)} = \underset{\theta_{(K,m,S,R)} \in \Theta_{(K,m,S,R)}}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \ln\{f(\mathbf{y}_i\,|\,\theta_{(K,m,S,R)})\}.$$

The following assumption is considered:

(H1) There exists a unique $(K_0, m_0, S_0, R_0)$ such that $h = f(.\,|\,\theta^\star_{(K_0,m_0,S_0,R_0)})$ for some parameter value $\theta^\star$, and the couple $(K_0, m_0)$ is supposed to be known.

To simplify the notation, all the dependencies over this couple $(K_0, m_0)$ is omitted in the following. Moreover, an additional technical assumption is considered:

(H2) The vectors $\theta^\star_{(S,R)}$ and $\hat{\theta}_{(S,R)}$ are supposed to belong to a compact subspace $\Theta'_{(S,R)}$ of the following subset (included into $\Theta_{(S,R)}$)

$$\big(\mathcal{P} \times \mathcal{B}(\eta, \mathrm{card}(S))^{K_0} \times \mathcal{D}^{K_0}_{\mathrm{card}(S)} \times \mathcal{B}(\rho, \mathrm{card}(S^c), 1 + \mathrm{card}(R)) \times \mathcal{D}_{\mathrm{card}(S^c)}\big) \cap \Theta_{(S,R)}$$

with

(i) $\mathcal{P} = \{(p_1, \ldots, p_K) \in [0,1]^K; \sum_{k=1}^K p_k = 1\}$ denotes the set of possible proportions,

(ii) $\mathcal{B}(\eta, r) = \{\mathbf{x} \in \mathbb{R}^r, \ \|\mathbf{x}\| \leqslant \eta\}$ where $\forall \mathbf{x} \in \mathbb{R}^r, \ \|\mathbf{x}\| = \sqrt{\sum_{i=1}^r x_i^2}$,

(iii) $\mathcal{B}(\rho, q, r) = \{A \in \mathcal{M}_{q\times r}(\mathbb{R}), \ |||A||| \leqslant \rho\}$ where the norm $|||.|||$ is defined by

$$\forall A \in \mathcal{M}_{q\times r}(\mathbb{R}), \ |||A||| = \sup_{\substack{y \in \mathbb{R}^q \\ \|y\|=1}} \|Ay\|$$

(iv) $\mathcal{D}_r$ is the set of the $r \times r$ positive definite matrices with eigenvalues in $[a, b]$ with $0 < a < b$.

THEOREM 2: *Under assumptions (H1), (H2), the couple of variable sets $(\hat{S}, \hat{R})$ maximizing Criterion (2) with fixed $(K_0, m_0)$ is such that $P((\hat{S}, \hat{R}) = (S_0, R_0)) \to_{n\to\infty} 1$.*

The proof is given in Web Appendix B.

## 6. Method Validation

Through the presentation of simulated experiments, we highlight the difference with Raftery and Dean's method and discuss the robustness of our methodology. Other examples are presented in Web Appendix D and in Maugis, Celeux, and Martin-Magniette (2007).

*Comparison with Raftery and Dean's method.* The dataset consists of 2000 data points from a mixture of four Gaussian distributions $\mathcal{N}(\mu_k, I_2)$ with $\mu_1 = (-2, -2)$, $\mu_2 = (-2, 2)$, $\mu_3 = -\mu_2$, $\mu_4 = -\mu_1$ and with a proportion vector $\mathbf{p} = (0.3, 0.2, 0.3, 0.2)$. Eight irrelevant variables are appended, simulated according to $\boldsymbol{y}_i^{\{3,\ldots,10\}} = \boldsymbol{y}_i^{\{1,2\}}\beta + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \Omega)$. Different scenarios for the eight variables are proposed, ranging from all are independent of the relevant variables to all depend on the relevant variables (see Table 1).

The two compared algorithms choose the true number of components $\hat{K} = 4$ in all scenarios. Raftery and Dean's procedure selects all variables when the irrelevant ones are simulated from $\mathcal{N}(0,1)$ but selects the true relevant variables when the noise is lower (scenarios 1 and 2). For scenarios 3–5 where the number of independent variables is larger than the regressed variables, their procedure selects the independent variables. Nevertheless their method has a good behavior when the number of regressed variables is larger (scenarios 6 and 7). Our procedure selects the true variable partition in all

**Table 1**

*Selected models with Raftery and Dean's algorithm and ours in seven scenarios, where* $\beta_1 = ((0.5, 0)', (0, 1)', (2, 0)', (0, 3)')$, $\beta_2 = (2, 0.5)', (0.5, 1)', \beta_3 = (2, 0)', (0, 3)', 0_p$ *is the* $2 \times p$ *zero matrix. In all cases, both methods select a mixture with* $\hat{K} = 4$ *components.*

| | Raftery & Dean | | Our algorithm | | |
|---|---|---|---|---|---|
| Scenario | $\hat{m}$ | $\hat{S}$ | $\hat{m}$ | $\hat{S}$ | $\hat{R}$ |
| 1: $\beta = 0_8$, $\Omega = I_8$ | $[p_k LI]$ | $\{1 - 10\}$ | $[p_k LI]$ | $\{1, 2\}$ | $\emptyset$ |
| 2: $\beta = 0_8$, $\Omega = 0.5I_8$ | $[p_k LI]$ | $\{1, 2\}$ | $[p_k LI]$ | $\{1, 2\}$ | $\emptyset$ |
| 3: $\beta = ((2, 0)', 0_7)$, $\Omega = I_8$ | $[p_k LI]$ | $\{1, 2, 4 - 10\}$ | $[p_k LI]$ | $\{1, 2\}$ | $\{1\}$ |
| 4: $\beta = ((0.5, 0)', (0, 1)', 0_6)$, $\Omega = I_8$ | $[p_k LI]$ | $\{1, 2, 5 - 10\}$ | $[p_k LI]$ | $\{1, 2\}$ | $\{1, 2\}$ |
| 5: $\beta = (\beta_1, 0_4)$, $\Omega = \text{diag}(I_2, 0.5I_2, I_4)$ | $[p_k LI]$ | $\{1, 2, 7 - 10\}$ | $[p_k LI]$ | $\{1, 2\}$ | $\{1, 2\}$ |
| 6: $\beta = (\beta_1, \beta_2, 0_2)$, $\Omega = \text{diag}(I_2, 0.5I_4, I_2)$ | $[p_k LI]$ | $\{1, 2\}$ | $[p_k LC]$ | $\{1, 2\}$ | $\{1, 2\}$ |
| 7: $\beta = (\beta_1, \beta_2, \beta_3)$, $\Omega = \text{diag}(I_2, 0.5I_4, I_2)$ | $[p_k LI]$ | $\{1, 2\}$ | $[p_k LC]$ | $\{1, 2\}$ | $\{1, 2\}$ |

scenarios but in the two last, it selects a too complex mixture form. This result seems to be related to the use of two nested backward stepwise algorithms.

*Waveform dataset.* This experiment allows us to assess the method behavior on a non-Gaussian mixture dataset. It is extracted at random from the waveform dataset, available at the UCI repository (Blake, Keogh, and Merz, 1999). It consists of 900 observations divided into three equiprobable groups, based on a random convex combination of two of three waveforms sampled at integers $\{1, \ldots, 21\}$ with noise added. Nineteen noisy standard centered Gaussian variables are appended. A detailed description of the waveform dataset is available in Breiman et al. (1984). We compare our method with the one of Raftery and Dean for $K \in \{3, 6\}$ with spherical, diagonal models and models of the form $[p\_L\_C\_]$ (see Web Table 1). The selected model with our method is $(\hat{K} = 6, \hat{m} = [pLI], \hat{S} = \{3, 4, 6 - 15, 18, 19\}, \hat{R} = \{8, 11, 15\})$. All noisy variables and variables {1, 2, 5, 16, 17, 20, 21} are declared irrelevant. The final clustering is coherent with the construction of the sample: Three clusters correspond to the three wave functions and the three others are convex combinations of two wave functions (see Figure 2). Raftery and Dean's method selects a mixture ($K$, $m$) = (3, $[pL_kB]$) and declares that all variables are relevant except variables 5 and 16. Their resulting clustering reveals the three wave functions but not their convex combinations. But when the complete waveform dataset (5000 individuals) is analyzed, there are no sensitive differences between the two variable selection methods. It highlights the fact that our more realistic model is able to detect more easily the variable roles for smaller datasets.

## 7. Analysis of Transcriptome Data

As explained in the Introduction, variable selection is desirable in cluster analysis of transcriptome data. We illustrate it by studying a transcriptome dataset of *Arabidopsis thaliana*, extracted from the database CATdb (Gagnot et al., 2008). In this database, data are organized by project, each project being composed of a set of experiments dedicated to a specific biological question. We focus on 1020 genes of *Arabidopsis thaliana* declared differentially expressed at least once in a time course of the hypocotyl growth switch (project 6 in

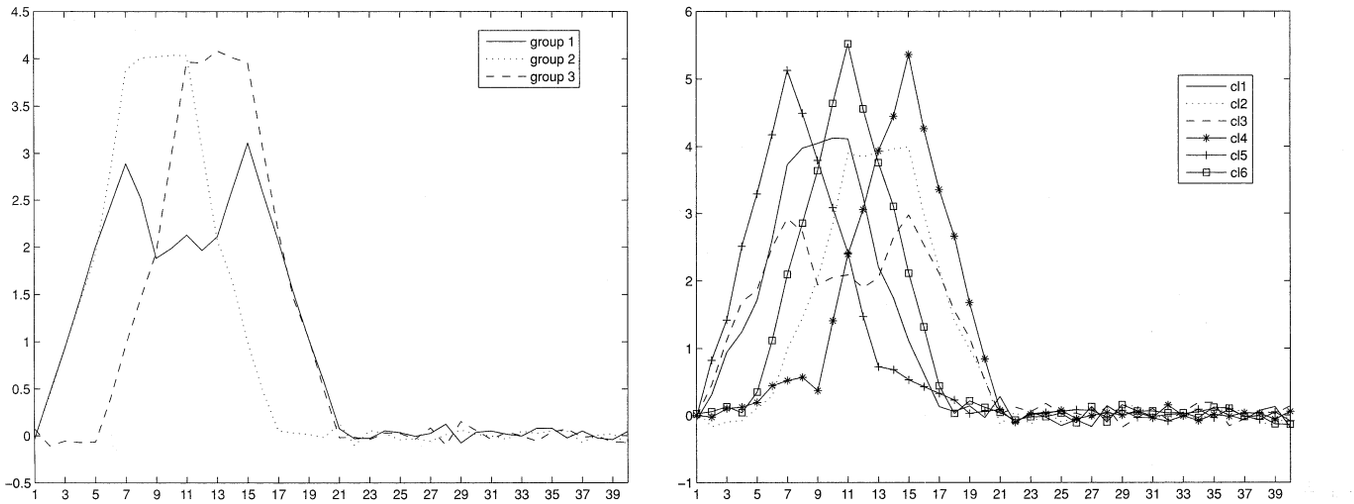


**Figure 2.** On the left, average profiles of the three real groups and on the right, average profiles of the six clusters found with our variable selection procedure.

**Table 2**

*Description of the transcriptome projects used to define the seven block variables. The number of experiments and the aim of each project are given in columns 2 and 3.*

| Project | Exper. num. | Aim of the project |
|---|---|---|
| 1 | 8 | Transcriptome of the circadian cycle |
| 2 | 4 | Transcriptome response to iron signaling |
| 3 | 4 | Transcriptome profiling from a protoplast culture |
| 4 | 3 | Transcriptome profiling from cell division to differentiation |
| 5 | 3 | Transcriptome response to nematode infection |
| 6 | 3 | Transcriptome time course of the hypocotyl growth switch |
| 7 | 2 | Transcriptome of the hypocotyl growth switch to isoxaben treatment |

Table 2) and study their behavior in other projects (see Table 2): We study $n = 1020$ genes described by $Q = 27$ experiments partitioned into $T = 7$ block variables. Gene $i$ is described with a vector $\mathbf{y}_i' \in \mathbb{R}^{27}$, the component $y_i^j$ corresponding to the test statistic calculated in the experiment $j$ for the differential analysis (see Gagnot et al., 2008, for details on the normalization and the differential analysis steps).

In this example, Gaussian mixtures with equal volumes and a maximal number of mixture components fixed to $K_{\max} = 20$ (see Web Table 1) are considered. Using all variables, the selected model is $[p_k LC]$ with 16 clusters. When our variable selection procedure is performed, the selected mixture model is $[p_k LC]$ with $\hat{K} = 17$ clusters. The relevant block variables are projects 1, 3, 4, 6, and 7 and the four last ones enter in the regression model. Some gene sets are common to the two clusterings (see Web Figure 1). Nevertheless with our variable selection procedure, clusters seem to be more homogeneous than without variable selection. Among the 17 clusters with different sizes (see Web Table 2), some already known gene subsets are recovered and some clusters interesting from a biological point of view are highlighted (see for example Figure 3). The result of our procedure allows biologists to formulate new assumptions. As an example, they will take a biological interest in 15 genes clustered with four recently discovered genes (cluster 13).

Concerning the variable selection, project 6, used to define the gene subset, has been declared relevant for the clustering, as project 7, also related to the hypocotyl growth switch. Regression parameters are given in Web Figure 2. The two irrelevant projects are related to stress conditions: project 5 investigates root-knot nematode infection, known to induce the redifferentiation of root cells and formation of giant cells (Jammes et al., 2005). These two phenomena are, respectively, studied in project 3 and projects 6, 7. Project 2 investigates iron stress of cells. It is mainly explains by projects 3 and 6 but no biological interpretation is yet available. Finally we explore whether the irrelevant projects offer a different gene clustering and we find six clusters poorly related to the 17 clusters previously discussed.

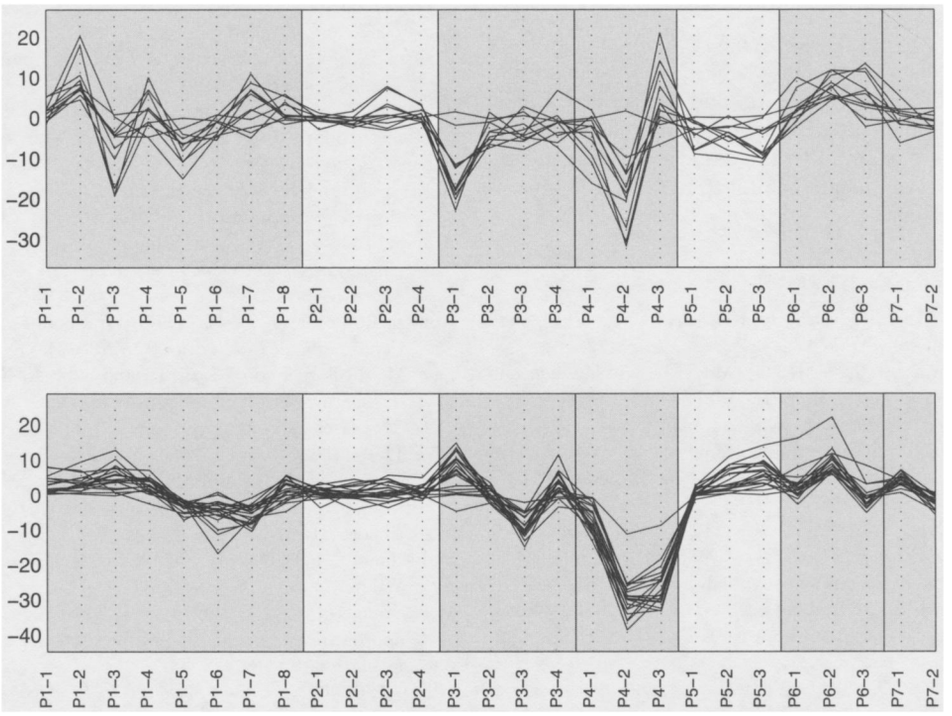


**Figure 3.** Graphical representation of gene profiles in cluster 4 on the top and cluster 17 on the bottom. Relevant clustering projects are colored in gray and, on the $x$-axis, $Pi - j$ denotes experiment $j$ in project $i$.

## 8. Discussion

We have presented a general variable selection methodology for cluster analysis when the individual number is greater than the variable number. Following Raftery and Dean (2006), this methodology considers the problem in the model-based cluster analysis context. In our approach, variables could be partitioned into blocks and the role of the clustering variables with respect to the other variables is more versatile and is expected to be more realistic. Compared to Raftery and Dean (2006), our more general definition of variable role could avoid over-penalizing models with independent variables as illustrated in the numerical experiments. On the theoretical side, we have established the model identifiability and proved the consistency of our variable selection criterion under reasonable assumptions.

One of the interests of our model is to allow for a better and, sometimes subtle, interpretation of the variable role. Thus, this method can be regarded as promising in the field of microarray gene expression dataset analysis where the behavior of several thousand genes are described by an increasing number of experiments. Finally, we want to stress that the defined procedure can work with alternative models linking the clustering and remaining variables, provided that a BIC-like criterion analogous with our $BIC_{reg}$ criterion can be computed.

## 9. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2, 3, 5, 6, and 7 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

## Acknowledgements

The authors thank Sylvie Huet (INRA) and Adrian Raftery (University of Washington) for helpful discussions, and Sandra Pelletier, Sébastien Aubourg, and Jean-Pierre Renou (URGV) for their implication in the analysis of the transcriptome data. They are grateful to the associate editor and reviewers for valuable comments helpful to improve the presentation of this article.

## References

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd edition. New York: Wiley.

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

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis* **51,** 587–600.

Blake, C., Keogh, E., and Merz, C. (1999). *UCI Repository of Machine Learning Algorithms Databases.* Available at `http://www.ics.uci.edu/~mlearnMLRepository.html`

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

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

Brusco, M. J. and Cradit, J. D. (2001). A variable selection heuristic for $k$-means clustering. *Psychometrika* **66,** 249–270.

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

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002). Feature selection for clustering—a filter solution. *Proceedings of the Second IEEE International Conference on Data Mining*, 115–122. Washington, DC: IEEE Computer Society.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Devaney, M. and Ram, A. (1997). Efficient feature selection in conceptual clustering. *Machine Learning: Proceedings of the Fourteenth International Conference*, 92–97. San Francisco: Morgan Kaufman.

Fowlkes, E. B., Gnanadesikan, R., and Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification* **5,** 205–228.

Fraley, C. and Raftery, A. E. (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *Journal of Classification* **20,** 263–286.

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.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research* **3,** 1157–1182.

Jammes, F., Lecomte, P., Almeida-Engler, J., Bitton, F., Martin-Magniette, M.-L., Renou, J.-P., Abad, P., and Favery, B. (2005). Genome-wide expression profiling of the host response to root-knot nematode infection in *Arabidopsis. The Plant Journal* **44,** 447–458.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering* **16,** 1370–1386.

Jouve, P.-E. and Nicoloyannis, N. (2005). A filter feature selection method for clustering. *Proceedings of International Symposium on Methodologies for Intelligent Systems*, 583–593. Berlin: Springer.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773–795.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93,** 877–893.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence* **97,** 273–324.

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.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2007). *Variable selection for clustering with Gaussian mixture models.* Technical Report RR-6211, Inria, France.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

McLachlan, G., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18,** 413–422.

Miller, A. J. (1990). *Subset Selection in Regression.* London: Chapman and Hall.

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

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

Sharan, R., Elkon, R., and Shamir, R. (2002). Cluster analysis and its applications to gene expression data. In *Ernst Schering Workshop on Bioinformatics and Genome Analysis.* Berlin: Springer Verlag.

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.

*Received June* 2007. *Revised June* 2008.
*Accepted June* 2008.

## Appendix

### *The Backward Variable Selection in Regression*

The following algorithm allows us to determine the subset $R[\ell]$ of variables among $S$ required to explain a variable $\mathbf{y}^\ell$ with a linear regression. The model comparison is performed with criterion $\mathrm{BIC}_{\mathrm{reg}}$ defined in equation (3). The algorithm is making use of exclusion and inclusion steps now described.

*Initialization* $R[\ell] = S$, $j_E = \emptyset$ and $j_I = \emptyset$.

**Exclusion step** For all $j$ in $R[\ell]$, compute $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j) = \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^\ell \,|\, \mathbf{y}^{R[\ell]}) - \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^\ell \,|\, \mathbf{y}^{R[\ell]-j})$. Then, compute $j_E = \arg\min_{j \in R[\ell]} \mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j)$.

- If $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^{j_E}) \leqslant 0$, set $R[\ell] = R[\ell] - j_E$ and go to the inclusion step if $j_E \neq j_I$ or stop otherwise.
- otherwise go to the inclusion step if $j_I \neq \emptyset$ or stop otherwise.

**Inclusion step** For all $j$ in $S - R[\ell]$, compute $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j) = \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^\ell \,|\, \mathbf{y}^{R[\ell] \cup j}) - \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^\ell \,|\, \mathbf{y}^{R[\ell]})$. Then, compute $j_I = \arg\max_{j \in S - R[\ell]} \mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j)$.

- If $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^{j_I}) > 0$, $R[\ell] = R[\ell] \cup j_I$ and go to the exclusion step if $j_I \neq j_E$ or stop otherwise.
- otherwise go to the exclusion step.

Starting from the exclusion step, the backward variable selection algorithm consists of alternating the exclusion and the inclusion steps.