Variable Selection for Clustering with Gaussian Mixture Models

Cathy Maugis, Gilles Celeux, Marie-Laure Martin-Magniette

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Variable Selection for Clustering with Gaussian Mixture Models

Cathy Maugis — Gilles Celeux — Marie-Laure Martin-Magniette

N° 6211

Juin 2007

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--6211--FR+ENG

Variable Selection for Clustering with Gaussian Mixture Models

Cathy Maugis* , Gilles Celeux † , Marie-Laure Martin-Magniette ‡

Thème COG — Systèmes cognitifs
Projets SELECT

Rapport de recherche n° 6211 — Juin 2007 — 35 pages

Abstract: 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 general model generalizing the model of Raftery and Dean (2006) is proposed to specify the role of each variable. This model does not need any prior assumptions about the link between the selected and discarded variables. Models are compared with BIC. Variables role is obtained through an algorithm embedding two backward stepwise variable selection algorithms for clustering and linear regression. The consistency of the resulting criterion is proved under regularity conditions. Numerical experiments on simulated datasets and a genomics application highlight the interest of the proposed variable selection procedure.

Key-words: Variable Selection, Model-based Clustering, Linear Regression, Bayes factor, BIC

* Université Paris-Sud, Projet SELECT
† INRIA Futurs, Projet SELECT, Université Paris-Sud
‡ UMR ENGREF/ INA-PG/ INRA Mathématique et Informatique Appliquées

Unité de recherche INRIA Futurs
Parc Club Orsay Université, ZAC des Vignes,
4, rue Jacques Monod, 91893 ORSAY Cedex (France)
Téléphone : +33 1 72 92 59 00 —Télécopie : +33 1 60 19 66 08

Sélection de variables pour la classification non supervisée par mélanges gaussiens

Résumé : Cet article s'intéresse à la sélection de variables en classification non supervisée. Le problème est ramené à un problème de sélection de modèles de mélanges de lois de probabilités. Un modèle global, généralisant celui de Raftery et Dean (2006) est proposé pour spécifier le rôle de chaque variable. Ce modèle ne nécessite aucune hypothèse a priori sur le lien entre les variables sélectionnées et les variables écartées pour la classification. Les modèles sont comparés avec BIC. Le statut des variables est obtenu grâce à un algorithme imbriquant deux algorithmes de sélection descendante avec remise en cause de variables pour la classification et pour la régression linéaire. La consistence du critère de sélection est démontrée sous des conditions de régularités. Des exemples numériques sur données simulées et sur une application génomique mettent en évidence l'intérêt de notre procédure de sélection de variables.

Mots-clés : Sélection de variables, Classification, Mélanges gaussiens, régression linéaire, Facteur de Bayes, BIC

1 Introduction

The goal of clustering methods is to discover structures (clusters) among n individuals described by Q variables. Many clustering methods exist and roughly fall into two categories. The first ones are based on similarity and/or dissimilarity distances. They gather hierarchical clusterings, which creates "trees" and also methods like K-means algorithm which classify data through a certain number of clusters fixed a priori. The second category is model-based methods which consist of using certain models for clusters and attempting to optimize the fit between the data and the model. In practice, each cluster is represented by a parametric distribution, like a Gaussian distribution and the entire data set is therefore modelled by a mixture of these distributions. One advantage of model-based clustering is to provide a rigourous framework to assess the number of mixture components and the role of the variables in a 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 only a subset of the available variables and a lot of variables can be useless or even harmful to detect a reasonable clustering structure. It is thus important to select the relevant variables from the cluster analysis view point. It is a recent research topic in contrast to variable selection in regression and classification models (Kohavi and John, 1997, Guyon and Elisseeff, 1990 and Miller, 1990) and this new interest for the variable selection in clustering comes from the increasingly frequent use of these methods on high-dimensional datasets.

Three types of approach dealing with variable selection in clustering have been proposed. The first one includes clustering methods with weighted variables (see for instance Friedman and Meulman 2003) and dimension reduction methods. For this later, 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 by Bouveyron *et al.* (2007) to take into account simultaneously the dimension reduction and the data clustering. The specificity of this first type of methods is an implicit variable selection in contrast to the two last approaches selecting explicitly relevant variables. The so-called "filter" approach selects the variables with data analysis tools before the clustering analysis (see for instance, Dash *et al.* 2002, Jouve and Nicoloyannis 2005). The influence of the independent selection step on the clustering result is the main weakness of these methods. In contrast, the so-called "wrapper" approach combines the variable selection and the clustering. For distance-based methods, one can cite Fowlkes *et al.* (1988) for a forward selection approach for 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 exists also wrapper methods in the model-based clustering setting. When the number of variables is greater than the number of individuals, Tadesse *et al.* (2005) proposed a fully Bayesian method using a reversible jump algorithm to choose simultaneously the number of mixture components and to select variables. Kim *et al.* (2006) use a similar approach by formulating clustering in terms of an infinite mixture of distributions via Dirichlet process mixtures. They assumed that there is no correlation

between the sets of the relevant variables for the clustering and the irrelevant ones. However this assumption is often unrealistic. Others approaches that rely on model selection methods exist. In this paper, we focus on Gaussian mixture model clustering. Law *et al.* (2004) proposed to estimate a set of real-valued quantity for each variable which allows to evaluate the importance of the variable in the clustering process. It is done using the *minimum message length* model selection criterion. Raftery and Dean (2006) recasted the problem of comparing two nested subsets of variables as a model comparison problem and address it using Bayes factors. An interesting aspect of their formulation model is that it does not require that irrelevant variables be independent of the clustering variables. They avoid thus the unrealistic independence assumption between the relevant and irrelevant variables to the clustering, considered for instance in Kim *et al.* (2006) or in Law *et al.* (2004). In their model, the subset of irrelevant variables, independent of the clustering, depends of the relevant variables through a linear regression equation. However, they do not allow the irrelevant variables to be independent of the clustering variables. It means that independent variables are enforced to enter as dependent variables in the regression linear equation. Their introduction in the regression involves additional parameters but does not lead to a significant increase of the loglikelihood. Thus, models including those variables in the clustering variables could be wrongly preferred to models considering them as dependent variables in the regression equation, when the models are compared with Bayes factor or penalized likelihood criteria as AIC or BIC.

In this paper, we propose an improvement of the method proposed by Raftery and Dean (2006) by considering another type of relation between the irrelevant variables and the clustering variables. We consider that some of irrelevant variables could be independent of the clustering variables. Also, the algorithm we make use is a backward stepwise variable selection algorithm rather than a forward stepwise selection algorithm as in Raftery and Dean (2006), because starting the search with all variables allows the model to take variable interactions into account. Finally, we consider a more general situation where the variables are partitioned into homogeneous blocks which cannot be splitted. This new variable roles modelling allows to improve the clustering and its interpretation.

The paper is organized as follows. Gaussian mixture models for clustering are reviewed in Section 2. Our variable selection approach is presented and compared with the Raftery and Dean approach in Section 3. The greedy search algorithm we propose is presented in Section 4. The consistency of the variable selection criterion is proved in Section 5. Section 6 is devoted to the presentation of numerical experiments on both simulated and real datasets. A discussion on the overall method and related approaches is presented in Section 7. The paper is completed with three technical appendices.

2 Multivariate Gaussian models and clustering

Model-based clustering (MBC) consists of assuming that the data come from a source with several subpopulations. Each subpopulation is modelled 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 component parameterized density is usually a multidimensional Gaussian density. Observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, with $\mathbf{y}_i$ in $\mathbb{R}^Q$, are assumed to be a sample from a probability distribution with density

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

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 μ_k and variance matrix Σ_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 $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) = ((\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_n, \mathbf{z}_n))$ where the missing data are $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ are binary vectors such that $z_{ik} = 1$ iff $\mathbf{y}_i$ arises from group k. The $\mathbf{z}$'s define a partition of the observed data $\mathbf{y}$, which is an ideal clustering of the data associated to the mixture model.

The mixture component variance matrix can be decomposed as in Banfield and Raftery (1993) and Celeux and Govaert (1995), into the following form

$$\Sigma_k = L_k D_k' A_k D_k$$

where $L_k = |\Sigma_k|^{1/Q}$ controls the volume of the k^{th} cluster, D_k is the matrix of eigenvectors of Σ_k which defines the cluster orientation and A_k is the diagonal matrix of normalized eigenvalues of Σ_k which controls the shape of that cluster. According to the constraints which are 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 the set of 28 models, described in Appendix A, and $K_{\max}$ is the maximum number of clusters which has to be 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

$$(\tilde{K}, \tilde{m}) = \underset{(K,m)\in\mathcal{T}}{\operatorname{argmax}} \; P[(K, m)|\mathbf{y}]$$

is to be chosen. By Bayes theorem

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

and, supposing a non informative uniform prior distribution $P[(K,m)]$ on the models, it leads to $P[(K,m)|\mathbf{y}] \propto f(\mathbf{y}|K,m)$. Thus

$$(\tilde{K},\tilde{m}) = \underset{(K,m)\in\mathcal{T}}{\operatorname{argmax}}\ f(\mathbf{y}|K,m)$$

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

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

α being the vector parameter of the model (K,m) and $\pi(\alpha|K,m)$ its prior distribution (Kass and Raftery 1995). Since this integrated likelihood is typically difficult to calculate, an asymptotic approximation of $2\ln[f(\mathbf{y}|K,m)]$ is generally used. This approximation is the Bayesian Information Criterion (BIC) defined by

$$\mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}|K,m) = 2\ln[f(\mathbf{y}|K,m,\hat{\alpha})] - \lambda_{(K,m)}\ln(n) \tag{2}$$

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

$$(\hat{K},\hat{m}) = \underset{(K,m)\in\mathcal{T}}{\operatorname{argmax}}\ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}|K,m).$$

For deriving $(\hat{K},\hat{m})$, the maximum likelihood estimate (mle) $\hat{\alpha}$ is computed using generally the EM algorithm (Dempster, Laird and Rubin, 1977). And this estimate yields the clustering Maximum a Posteriori (MAP) rule $\hat{\mathbf{z}} = (\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_n) = \mathrm{MAP}(\hat{\alpha})$ defined by

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

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 entering 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. The idea of the Raftery and Dean approach is to divide the set of variables into the subset of relevant clustering variables and the subset of irrelevant variables, independent of the clustering, but depending of the relevant variables through a linear regression equation. This is an interesting aspect of their approach, avoiding the unrealistic independence assumption between the relevant and irrelevant variables to the clustering, considered for instance in Law *et al.* (2004). And,

as they stressed, the independence assumption would often lead to wrongly declare a variable as relevant to the clustering because it is related to the clustering variables, but not necessarily to the clustering itself. However, Raftery and Dean (2006) do not allow the irrelevant variables to be independent of the clustering variables. It means that independent variables are forced to enter as dependent variables in the regression linear equation. Their introduction in the regression involves additional parameters in the model without leading to a significant increase of its loglikelihood. Thus, models including those variables in the clustering variables could be wrongly preferred to models considering them as dependent variables in the regression equation, when the models are compared with Bayes factor or penalized likelihood criteria as AIC or BIC. For this very reason, we opt for a more realistic model where an irrelevant clustering variable can be supposed to be dependent or independent of the clustering variables. Moreover, we consider a more general framework where the Q variables are partitioned into T blocks. That is there exists a function Ψ 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 6.3. 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.

The models in competition will be compared with their integrated likelihoods which can be decomposed into two multiplicative parts. The first part is the integrated likelihood of the Gaussian mixture model (K, m) on the relevant clustering variables. The second part is the integrated likelihood of the regression of the irrelevant variables on a subset of the clustering variables. It is the main difference with the Raftery and Dean model: Here, the clustering variables do not necessarily enter in the regression equation explaining the irrelevant variables. Let $\mathcal{F}$ be the family of variable block index subset, $S \in \mathcal{F}$ the set of clustering variable block indexes, and S^c its complementary in $\mathcal{F}$.

In order to distinguish the role of each clustering block of variables in a regression, the set of clustering variable blocks entering the regression equation of irrelevant variables S^c will be denoted J. The information is summarized into a couple (S, J) and the set $\mathcal{V} = \{(S, J); (S, J) \in \mathcal{F}^2,\ J \subseteq S\}$ is defined. The division of the variable block roles is illustrated in Figure 1. Finally, the considered model set is defined by

$$\mathcal{N} = \{(K, m, S, J); (K, m) \in \mathcal{T};\ (S, J) \in \mathcal{V}\}.$$

Throughout the paper, for an union of variable blocks A, $\mathbf{y}^A$ denotes the set $\{\mathbf{y}^j \in \mathbb{R}^n / \Psi(j) \in A\}$ and $\text{card}(A) = \text{card}\{j; \Psi(j) \in A\}$. For each model (K, m, S, J), the associated integrated likelihood has the form

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

where $f_{\text{clust}}(\mathbf{y}^S|K, m) = \int f_{\text{clust}}(\mathbf{y}^S|K, m, \alpha)\pi(\alpha|K, m, S)d\alpha$ is the mixture integrated likelihood and $f_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J) = \int f_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J, B, \Omega)\pi(B, \Omega|S, J)dBd\Omega$ the multidimensional regression integrated likelihood, B denoting the vector of regression coefficients and Ω the

variance matrix of the regression model (see Appendix B).

The model to be chosen is supposed to maximize this integrated likelihood

$$\begin{aligned}(\tilde{K},\tilde{m},\tilde{S},\tilde{J}) &= \underset{(K,m,S,J)\in\mathcal{N}}{\operatorname{argmax}} f_{\text{clust}}(\mathbf{y}^S|K,m) f_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J) \\ &= \underset{(K,m,S,J)\in\mathcal{N}}{\operatorname{argmax}} 2\ln\left[f_{\text{clust}}(\mathbf{y}^S|K,m)\right] + 2\ln\left[f_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J)\right].\end{aligned}$$

In practice, the integrated likelihoods are approximated using the BIC approximation as in (2), and the chosen model is

$$(\hat{K},\hat{m},\hat{S},\hat{J}) = \underset{(K,m,S,J)\in\mathcal{N}}{\operatorname{argmax}} \{\text{BIC}_{\text{clust}}(\mathbf{y}^S|K,m) + \text{BIC}_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J)\} \tag{3}$$

where

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

and

$$\text{BIC}_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J) = 2\ln[f_{\text{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J,\hat{\beta},\hat{\Omega})] - \nu_{(S,J)}\ln(n),$$

$\lambda^S_{(K,m)}$ being the number of free parameters of the (K,m) model with $\text{card}(S)$ variables, $(\hat{\beta},\hat{\Omega})$ the maximum likelihood estimate of the regression parameters, and $\nu_{(S,J)} = (\text{card}(J)+1)\text{card}(S^c) + \frac{\text{card}(S^c)(\text{card}(S^c)+1)}{2}$. We refer to formula (18) in Appendix B where the computation of the BIC criterion for multidimensional multivariate regression is detailed.

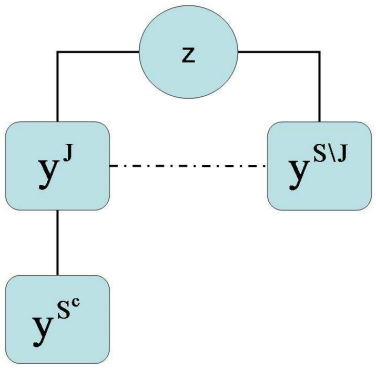

Figure 1: Graphical representation of the variables repartition into three groups.

4 The variable selection procedure

The number of models in $\mathcal{N}$ is $28(K_{\max}-1)\sum_{t=1}^{T}\binom{T}{t}\sum_{l=1}^{t}\binom{t}{l}$, where $K_{\max}$ is the maximum number of clusters and an exhaustive research of the optimal model is impossible in most situations. The algorithm we propose is a two-nested-step algorithm.

- (i) For all (K, m), we search

$$(\hat{S}(K,m), \hat{J}(K,m)) = \underset{(S,J)\in\mathcal{V}}{\operatorname{argmax}} \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S|K,m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J) \right\}$$

  by a backward stepwise procedure detailed hereafter.

- (ii) We determine

$$(\hat{K}, \hat{m}) = \underset{(K,m)\in\mathcal{T}}{\operatorname{argmax}} \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^{\hat{S}(K,m)}|K,m) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{\hat{S}^c(K,m)}|\mathbf{y}^{\hat{J}(K,m)}) \right\}.$$

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

In a backward stepwise selection, all the variables are selected at the beginning and, at each step, a block of variables is excluded or included of the significant variable set. We opt for a backward stepwise selection algorithm rather than a forward stepwise selection algorithm as in Raftery and Dean (2006), because starting the search with all variables included allows the model to take variable block interactions into account.

4.1 The models in competition

The variable set $\{1, \ldots, T\}$ is divided at each step into three subgroups: S the set of already selected clustering variable blocks, C the candidate block being considered for inclusion into or exclusion from the set of clustering variables and R the remaining variables. The decision of exclusion (resp. inclusion) of variable block C from (resp. in) the set of clustering variables is made by the comparison of the following two models:

1. $M_1(K, m)$:

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

   Model M_1 specifies that given $\mathbf{y}^S$, $\mathbf{y}^C$ is explained by a subset $\mathbf{y}^{S[C]}$ of $\mathbf{y}^S$ and gives no additional information for the clustering.

2. $M_2(K,m)$:

$$\begin{aligned} f_2(\mathbf{y}|K,m) &= f_2(\mathbf{y}^R,\mathbf{y}^C,\mathbf{y}^S|K,m) \\ &= \sum_{\mathbf{z}} f_2(\mathbf{y}^R,\mathbf{y}^C,\mathbf{y}^S|\mathbf{z},K,m) f_2(\mathbf{z}|K,m) \\ &= f_2(\mathbf{y}^R|\mathbf{y}^C,\mathbf{y}^S) \sum_{\mathbf{z}} f_2(\mathbf{y}^C,\mathbf{y}^S|\mathbf{z},K,m) f_2(\mathbf{z}|K,m) \\ &= f_2(\mathbf{y}^R|\mathbf{y}^C,\mathbf{y}^S) f_{\text{clust}}(\mathbf{y}^C,\mathbf{y}^S|K,m). \end{aligned}$$

In model M_2, after the observation of $\mathbf{y}^S$, $\mathbf{y}^C$ provides additional information for the clustering.

The two models are compared with the Bayes factor, $B_{12}(K,m)$ for $M_1(K,M)$ against $M_2(K,M)$:

$$B_{12}(K,m) = \frac{f_1(\mathbf{y}|K,m)}{f_2(\mathbf{y}|K,m)}. \tag{4}$$

Because the conditional distribution of $(\mathbf{y}^R|\mathbf{y}^C,\mathbf{y}^S)$ is unaffected by the distribution of $(\mathbf{y}^C,\mathbf{y}^S)$, $f_1(\mathbf{y}^R|\mathbf{y}^C,\mathbf{y}^S) = f_2(\mathbf{y}^R|\mathbf{y}^C,\mathbf{y}^S)$ and the Bayes factor can be written

$$B_{12}(K,m) = \frac{f_{\text{reg}}(\mathbf{y}^C|\mathbf{y}^{S[C]}) f_{\text{clust}}(\mathbf{y}^S|K,m)}{f_{\text{clust}}(\mathbf{y}^C,\mathbf{y}^S|K,m)}.$$

Since the integrated likelihoods are difficult to evaluate, $-2\ln[B_{12}(K,m)]$ is approximated with

$$\text{BIC}_{\text{diff}}(\mathbf{y}^C|K,m) = \text{BIC}_{\text{clust}}(\mathbf{y}^S,\mathbf{y}^C|K,m) - \left[\text{BIC}_{\text{reg}}(\mathbf{y}^C|\mathbf{y}^{S[C]}) + \text{BIC}_{\text{clust}}(\mathbf{y}^S|K,m)\right]. \tag{5}$$

4.2 The backward stepwise selection algorithm

Initialisation Let (K,m) fixed, $S = \{1,\dots,T\}$, $i_E = \emptyset$ and $i_I = \emptyset$.

This algorithm is making use of an exclusion and an inclusion procedures now described. The decision of excluding or including a multidimensional variable block from the set of clustering variables is based on the comparison of the two models with the BIC approximation of the Bayes factor.

Exclusion step In this step, the proposed multidimensional variable for removal from the set of currently selected clustering variables is chosen to be the variable block from this set which gives the smallest value of BIC_{diff} defined in (5). It is as follows:

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

$$i_E = \underset{i\in S}{\text{argmin}}\, \text{BIC}_{\text{diff}}(\mathbf{y}^i|K,m).$$

- If $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^{i_E}|K,m) \leq 0$,
  - $S = S - i_E$
  - if $i_E = i_I$ stop
  - otherwise go to the inclusion step;
- otherwise
  - if $i_I = \emptyset$ stop
  - go to the inclusion step.

Inclusion step In this step, the proposed new multidimensional clustering variable is chosen to be the variable block from this set which gives the greatest difference between value for $\mathrm{BIC}_{\mathrm{diff}}$. It is as follows:

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

$$i_I = \underset{i \in S^c}{\operatorname{argmax}} \mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^i|K,m).$$

- If $\mathrm{BIC}_{\mathrm{diff}}(\mathbf{y}^{i_I}|K,m) > 0$,
  - if $i_I = i_E$ stop
  - otherwise $S = S \cup i_I$ 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 the exclusion and the inclusion steps.

5 Consistency of our criterion

In this section, it is proved that the probability of selecting the true couple of variables (S_0, J_0) by maximizing criterion (3) approaches 1 as $n \to \infty$ when the sampling distribution is one of the mixture models in competition, and the true model m_0 and number of mixture components K_0 are known. The density function of the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is denoted h, and

$$\begin{aligned} \theta^{\star}_{(K,m,S,J)} &= \underset{\theta_{(K,m,S,J)} \in \Theta_{(K,m,S,J)}}{\operatorname{argmin}} \mathrm{KL}[h, f(.|\theta_{(K,m,S,J)})] \\ &= \underset{\theta_{(K,m,S,J)} \in \Theta_{(K,m,S,J)}}{\operatorname{argmax}} \mathbb{E}_X[\ln f(X|\theta_{(K,m,S,J)})], \end{aligned}$$

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

$$\hat{\theta}_{(K,m,S,J)} = \underset{\theta_{(K,m,S,J)} \in \Theta_{(K,m,S,J)}}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln[f(\mathbf{y}_i|\theta_{(K,m,S,J)})].$$

The following assumption is considered:

(H1) There exists a unique (K_0, m_0, S_0, J_0) such that $h = f(.|\theta^\star_{(K_0,\, m_0,\, S_0,\, J_0)})$ 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 of models (K_0, m_0) is omitted in the following. Moreover, an additional technical assumption is considered:

(H2) The vectors $\theta^\star_{(S,J)}$ and $\hat{\theta}_{(S,J)}$ are supposed to belong to a compact subspace $\Theta'_{(S,J)}$ of $\Theta_{(S,J)}$ defined by

$$\Theta'_{(S,J)} = \mathcal{P} \times \mathcal{B}(\eta, \mathrm{card}(S))^{K_0} \times \mathcal{D}^{K_0}_{\mathrm{card}(S)} \times \mathcal{B}(\rho, \mathrm{card}(S^c), 1 + \mathrm{card}(J)) \times \mathcal{D}_{\mathrm{card}(S^c)}$$

with

- $\mathcal{P} = \left\{(p_1, \ldots, p_K) \in [0,1]^K; \sum_{k=1}^{K} p_k = 1\right\}$ denotes the set of possible proportions,
- $\mathcal{B}(\eta, r) = \{\mathbf{x} \in \mathbb{R}^r,\ \|\mathbf{x}\| \leq \eta\}$ where $\forall \mathbf{x} \in \mathbb{R}^r,\ \|\mathbf{x}\| = \sqrt{\sum_{i=1}^{r} x_i^2}$,
- $\mathcal{B}(\rho, q, r) = \{A \in \mathcal{M}_{q\times r}(\mathbb{R}),\ |||A||| \leq \rho\}$ where the norm $|||.|||$ is defined by

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

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

Theorem 1.
Under assumptions (H1), (H2), the couple of variable sets $(\hat{S}, \hat{J})$ maximizing the criterion (3) with fixed (K_0, m_0) is such that $P((\hat{S}, \hat{J}) = (S_0, J_0)) \underset{n\to\infty}{\to} 1$.

Proof. We have $(\hat{S}, \hat{J}) = \underset{(S,J)\in\mathcal{V}}{\operatorname{argmax}}\ \mathbf{BIC}(S, J)$ with

$$\begin{aligned}
\mathbf{BIC}(S,J) &= \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^S) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J) \\
&= 2\ln[f_{\mathrm{clust}}(\mathbf{y}^S|\hat{\alpha})] - \lambda^S \ln(n) + 2\ln[f_{\mathrm{reg}}(\mathbf{y}^{S^c}|\mathbf{y}^J, \hat{B}, \hat{\Omega})] - \nu_{(S,J)} \ln(n) \\
&= 2\sum_{i=1}^{n} \ln[f(\mathbf{y}_i|\hat{\theta}_{(\mathrm{S,J})})] - \Xi_{(S,J)} \ln(n),
\end{aligned}$$

where $\Xi_{(S,J)} = \lambda^S + \nu_{(S,J)}$ is the number of model parameters for variable set (S, J). Thus

$$\begin{aligned} P((\hat{S}, \hat{J}) = (S_0, J_0)) &= P(\mathbf{BIC}(S_0, J_0) \geq \mathbf{BIC}(S, J), \forall (S, J) \in \mathcal{V}) \\ &= P(\mathbf{BIC}(S_0, J_0) - \mathbf{BIC}(S, J) \geq 0, \forall (S, J) \in \mathcal{V}). \end{aligned} \quad (6)$$

Denoting $\gamma_{(S,J)} = \Xi_{(S,J)} - \Xi_{(S_0,J_0)}$ and $\Delta\mathbf{BIC}(S, J) = \mathbf{BIC}(S_0, J_0) - \mathbf{BIC}(S, J)$, $\Delta\mathbf{BIC}(S, J)$ can be written

$$\Delta\mathbf{BIC}(S, J) = 2n \left\{ \frac{1}{n} \sum_{i=1}^{n} \ln \left[\frac{f(\mathbf{y}_i | \hat{\theta}_{(S_0,J_0)})}{h(\mathbf{y}_i)} \right] - \frac{1}{n} \sum_{i=1}^{n} \ln \left[\frac{f(\mathbf{y}_i | \hat{\theta}_{(S,J)})}{h(\mathbf{y}_i)} \right] \right\} + \gamma_{(S,J)} \ln(n). \quad (7)$$

Under regularity conditions and from (H1), $\Delta\mathbf{BIC}(S, J)$ converges to $-\mathrm{KL}[h, f(.|\theta^\star_{(S,J)})]$ when n tends to infinity . If $\mathrm{KL}[h, f(.|\theta^\star_{(S,J)})] \neq 0$ then the term into braces in (7) dominates and tends to infinity with n. Otherwise, by the unicity assumption in (H1), we have $S = S_0$ and $J_0 \subset J$. And, the term into braces in (7) is the loglikelihood ratio statistic of two nested models which tends to a chi-squared distribution. Thus, the term $\gamma_{(S,J)} \ln(n)$ dominates and tends to infinity with n. It leads to consider that $\mathcal{V}$ can be decomposed as follows

$$\mathcal{V} = \{(S_0, J_0)\} \cup \mathcal{V}_1 \cup \mathcal{V}_2$$

where

$\mathcal{V}_1 = \{(S, J) \in \mathcal{V};\ \mathrm{KL}[h, f(.|\theta^\star_{(S,J)})] \neq 0\}$ and

$\mathcal{V}_2 = \{(S_0, J) \in \mathcal{V}; J_0 \subset J\}$.

From (6), the theorem is demonstrated if it is proved that

$$\forall (S, J) \in \mathcal{V}_1 \cup \mathcal{V}_2,\ P\left(\Delta\mathbf{BIC}(S, J) < 0\right) \underset{n\to\infty}{\to} 0.$$

Case $(S, J) \in \mathcal{V}_1$**.** Denoting $\mathbb{M}_n(S, J) = \frac{1}{n} \sum_{i=1}^{n} \ln \left[\frac{f(\mathbf{y}_i | \hat{\theta}_{(S,J)})}{h(\mathbf{y}_i)} \right]$ and $M(S, J) = -\mathrm{KL}[h, f(.|\theta^\star_{(S,J)})]$, from (7) we have

$$\begin{aligned} & P(\Delta\mathbf{BIC}(S, J) < 0) = P(2n[\mathbb{M}_n(S_0, J_0) - \mathbb{M}_n(S, J)] + \gamma_{(S,J)} \ln(n) < 0) \\ = \ & P\left(\mathbb{M}_n(S_0, J_0) - M(S_0, J_0) + M(S_0, J_0) - M(S, J) + M(S, J) - \mathbb{M}_n(S, J) + \frac{\gamma_{(S,J)} \ln(n)}{2n} < 0 \right). \end{aligned}$$

Thus, for all $\epsilon > 0$, according to Lemma 5 in Appendix D,

$$\begin{aligned} P(\Delta\mathbf{BIC}(S, J) < 0) \leq \ & P(M(S_0, J_0) - \mathbb{M}_n(S_0, J_0) > \epsilon) + P(\mathbb{M}_n(S, J) - M(S, J) > \epsilon) \\ + \ & P\left(M(S_0, J_0) - M(S, J) + \frac{\gamma_{(S,J)} \ln(n)}{2n} < 2\epsilon \right). \end{aligned} \quad (8)$$

From Proposition 1, stated hereafter, $\forall (S, J), \mathbb{M}_n(S, J) \underset{n\to\infty}{\overset{P}{\to}} M(S, J)$. Thus,

$$\forall \epsilon > 0,\ P(\mathbb{M}_n(S, J) - M(S, J) > \epsilon) \leq P(|\mathbb{M}_n(S, J) - M(S, J)| > \epsilon) \underset{n\to\infty}{\to} 0.$$

Now,

$$P\left(M(S_0,J_0)-M(S,J)+\frac{\gamma_{(S,J)}\ln(n)}{2n}<2\epsilon\right)\leq P\left(M(S_0,J_0)-M(S,J)-2\epsilon<\left|\frac{\gamma_{(S,J)}\ln(n)}{2n}\right|\right).$$

But $\frac{\gamma_{(S,J)}\ln(n)}{2n}\underset{n\to\infty}{\to}0$ and $M(S_0,J_0)-M(S,J)>0$ because $(S,J)\in\mathcal{V}_1$. Taking $\epsilon=\frac{M(S_0,J_0)-M(S,J)}{4}>0$, we get

$$P\left(M(S_0,J_0)-M(S,J)+\frac{\gamma_{(S,J)}\ln(n)}{2n}<2\epsilon\right)\leq P\left(\frac{M(S_0,J_0)-M(S,J)}{2}<\left|\frac{\gamma_{(S,J)}\ln(n)}{2n}\right|\right)\underset{n\to\infty}{\to}0.$$

Finally, $P(\Delta\mathbf{BIC}(S,J)<0)\underset{n\to\infty}{\to}0$.

Case $(S,J)\in\mathcal{V}_2$. In this case, because of the unicity assumption in (H1), we have $S=S_0$ and $J_0\subset J\subseteq S_0$.

$$\begin{aligned}P(\Delta\mathbf{BIC}(S,J)<0) &= P\left(2\sum_{i=1}^{n}\ln\left[\frac{f(\mathbf{y}_i|\hat{\theta}_{(S_0,J_0)})}{f(\mathbf{y}_i|\hat{\theta}_{(S_0,J)})}\right]+\gamma_{(S_0,J)}\ln(n)<0\right)\\ &= P\left(2\ln\left[\frac{f_{\text{reg}}(\mathbf{y}^{S_0^c}|\mathbf{y}^{J_0},\hat{B},\hat{\Omega})}{f_{\text{reg}}(\mathbf{y}^{S_0^c}|\mathbf{y}^{J},\hat{B},\hat{\Omega})}\right]+\left[\nu_{(S_0,J)}-\nu_{(S_0,J_0)}\right]\ln(n)<0\right).\end{aligned}$$

Denoting $\mathcal{A}_n=\ln\left[\frac{f_{\text{reg}}(\mathbf{y}^{S_0^c}|\mathbf{y}^{J_0},\hat{B},\hat{\Omega})}{f_{\text{reg}}(\mathbf{y}^{S_0^c}|\mathbf{y}^{J},\hat{B},\hat{\Omega})}\right]$ the loglikelihood ratio between the regression models, since $J_0\subset J$, we have,

$$\mathcal{A}_n\underset{n\to\infty}{\overset{\mathcal{L}}{\to}}-\chi^2(\nu_{(S_0,J)}-\nu_{(S_0,J_0)}),$$

and the test statistic $\mathcal{A}_n$ is uniformly tight

$$\forall\xi>0,\exists M>0,\ \forall n,\ P(|\mathcal{A}_n|>M)<\xi. \tag{9}$$

From Lemma 5 in Appendix D,

$$\begin{aligned}P(\Delta\mathbf{BIC}(S,J)<0) &= P\left(\mathcal{A}_n+\left[\frac{\nu_{(S_0,J)}-\nu_{(S_0,J_0)}}{2}\right]\ln(n)<0\right)\\ &\leq P(-\mathcal{A}_n>M)+P\left(\left[\nu_{(S_0,J)}-\nu_{(S_0,J_0)}\right]\ln(n)<2M\right)\\ &\leq P(|\mathcal{A}_n|>M)+P\left(\left[\nu_{(S_0,J)}-\nu_{(S_0,J_0)}\right]\ln(n)<2M\right).\end{aligned}$$

Since $(S,J)\in\mathcal{V}_2$, $\nu_{(S_0,J)}-\nu_{(S_0,J_0)}>0$, $\left[\nu_{(S_0,J)}-\nu_{(S_0,J_0)}\right]\ln(n)\underset{n\to\infty}{\to}+\infty$. Therefore

$$P\left(\left[\nu_{(S_0,J)}-\nu_{(S_0,J_0)}\right]\ln(n)<2M\right)\underset{n\to\infty}{\to}0,$$

which is equivalent to

$$\forall \xi > 0,\ \exists N_0,\ \forall n > N_0, P\left(\left[\nu_{(S_0,J)} - \nu_{(S_0,J_0)}\right]\ln(n) < 2M\right) < \xi. \tag{10}$$

Finally, by (9) and (10) we obtain

$$\forall \xi > 0,\ \exists M > 0,\ \exists N_0 \in \mathbb{N}^\star,\ \forall n > N_0,\ P(|\mathcal{A}_n| > M) + P\left(\left[\nu_{(S_0,J)} - \nu_{(S_0,J_0)}\right]\ln(n) < 2M\right) < 2\xi.$$

Hence

$$\forall (S,J) \in \mathcal{V}_2,\ P(\Delta\mathbf{BIC}(S,J) < 0) \underset{n\to\infty}{\to} 0.$$

□

The following proposition will imply that $\forall (S,J),\ \mathbb{M}_n(S,J) \underset{n\to\infty}{\overset{P}{\to}} M(S,J)$.

Proposition 1.
Under assumption (H1), (H2), $\forall (S,J) \in \mathcal{V},\ \frac{1}{n}\sum_{i=1}^{n} \ln\left[\frac{h(\mathbf{y}_i)}{f(\mathbf{y}_i|\hat{\theta}_{(S,J)})}\right] \underset{n\to\infty}{\overset{P}{\to}} KL[h, f(.|\theta^\star_{(S,J)})]$.

Proof.
For making easier the reading of this proof, the notation Card(S) is replaced by $\sharp S$. Let $(S,J) \in \mathcal{V}$. By the law of large numbers, if $\mathbb{E}[|\ln(h(X))|] < \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left[h(\mathbf{y}_i)\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln(h(X))]. \tag{11}$$

And, if the Proposition 2 can be applied with the family

$$\mathcal{G}_{(\mathrm{S,J})} := \{\ln[f(.|\theta)];\ \theta \in \Theta'_{(\mathrm{S,J})}\}$$

thus

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left[f(\mathbf{y}_i|\hat{\theta}_{(\mathrm{S,J})})\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln f(X|\theta^\star_{(\mathrm{S,J})})]. \tag{12}$$

Then (11) and (12) give the result. Thus we have to prove that (H2) allows to verify the hypotheses of the Proposition 2 and $\mathbb{E}_X[|\ln h(X)|] < \infty$.
Firstly, according to (H2), $\Theta'_{(S,J)}$ is a compact metric space. Moreover, for all $\mathbf{x}$ in $\mathbb{R}^Q$, $\theta_{(S,J)} \in \Theta'_{(S,J)} \mapsto \ln[f(\mathbf{x}|\theta_{(S,J)})]$ is continuous. Let us verify now that there is an envelope function G of $\mathcal{G}_{(S,J)}$ being h-integrable. Recalling that

$$\ln[f(\mathbf{x}|\theta_{(S,J)})] = \ln[f_{\mathrm{clust}}(\mathbf{x}^S|\alpha)] + \ln[f_{\mathrm{reg}}(\mathbf{x}^{S^c}|\mathbf{x}^J, B, \Omega)],$$

these two terms are bounded separately.

Study of the first term:

Due to $\|\mathbf{x}^S-\mu_k\|^2_{\Sigma_k^{-1}}\geq 0$, $|\Sigma_k|^{-\frac{1}{2}}\leq a^{-\frac{\sharp S}{2}}$ according to Lemma 3 and $\sum_{k=1}^{K}p_k=1$, the upper bound of this first term is given by

$$\begin{aligned}\ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)] &= \ln\left[\sum_{k=1}^{K}p_k|2\pi\Sigma_k|^{-\frac{1}{2}}\exp\left(-\frac{\|\mathbf{x}^S-\mu_k\|^2_{\Sigma_k^{-1}}}{2}\right)\right]\\ &\leq \ln\left[\sum_{k=1}^{K}p_k(2\pi a)^{-\frac{\sharp S}{2}}\right]\\ &\leq -\frac{\sharp S}{2}\ln\left[2\pi a\right]\end{aligned}$$

where $\|\mathbf{x}^S-\mu_k\|^2_{\Sigma_k^{-1}}=(\mathbf{x}^S-\mu_k)'\Sigma_k^{-1}(\mathbf{x}^S-\mu_k)$.

For obtaining a lower bound, the concavity of the logarithm function is used thus

$$\begin{aligned}\ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)] &\geq \sum_{k=1}^{K}p_k\ln\left[|2\pi\Sigma_k|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\|\mathbf{x}^S-\mu_k\|^2_{\Sigma_k^{-1}}\right)\right]\\ &= -\frac{\sharp S}{2}\ln[2\pi]-\frac{1}{2}\sum_{k=1}^{K}p_k\left\{\ln\left[|\Sigma_k|\right]+\left[\|\mathbf{x}^S-\mu_k\|^2_{\Sigma_k^{-1}}\right]\right\}\end{aligned}$$

since $\forall k,\ |\Sigma_k|\leq b^{\sharp S}$ according to Lemma 3 and

$$\begin{aligned}\|\mathbf{x}^S-\mu_k\|^2_{\Sigma_k^{-1}} &\leq \frac{\|\mathbf{x}^S-\mu_k\|^2}{a}\\ &\leq \frac{2(\|\mathbf{x}^S\|^2+\|\mu_k\|^2)}{a}\\ &\leq \frac{2(\|\mathbf{x}^S\|^2+\eta^2)}{a}\end{aligned}$$

because $\mu_k\in\mathcal{B}(\eta,\sharp S)$. Thus,

$$\begin{aligned}\ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)] &\geq -\frac{\sharp S}{2}\ln[2\pi]-\frac{1}{2}\sum_{k=1}^{K}p_k\left\{\ln[b^{\sharp S}]+\frac{2}{a}(\|\mathbf{x}\|^2+\eta^2)\right\}\\ &= -\frac{\sharp S}{2}\ln[2\pi b]-\frac{\|\mathbf{x}\|^2+\eta^2}{a}.\end{aligned}$$

Finally the first term is bounded by

$$-\frac{\sharp S}{2}\ln[2\pi b]-\frac{\|\mathbf{x}\|^2+\eta^2}{a}\leq\ln[f_{\text{clust}}(\mathbf{x}^S|\alpha)]\leq-\frac{\sharp S}{2}\ln\left[2\pi a\right]. \tag{13}$$

Study of the second term:

Let $\mathbf{w}^J$ denote $\begin{pmatrix} 1 \\ \mathbf{x}^J \end{pmatrix} \in \mathbb{R}^{1+\sharp J}$.

$$\begin{aligned} \ln\left[f_{\text{reg}}(\mathbf{x}^{S^c}|\mathbf{x}^J, B, \Omega)\right] &= \ln\left[|2\pi\Omega|^{-1/2}\exp\left(-\frac{1}{2}\|\mathbf{x}^{S^c} - B\mathbf{w}^J\|^2_{\Omega^{-1}}\right)\right] \\ &= -\frac{\sharp S^c}{2}\ln[2\pi] - \frac{1}{2}\ln[|\Omega|] - \frac{1}{2}\|\mathbf{x}^{S^c} - B\mathbf{w}^J\|^2_{\Omega^{-1}}. \end{aligned}$$

Using Lemma 3, the following upper bound is found

$$\ln\left[f_{\text{reg}}(\mathbf{x}^{S^c}|\mathbf{x}^J, B, \Omega)\right] \leq -\frac{\sharp S^c}{2}\ln[2\pi a].$$

According to Lemma 3, $|\Omega| \leq b^{\sharp S^c}$ and $\|\mathbf{x}^{S^c} - B\mathbf{w}^J\|^2_{\Omega^{-1}} \leq a^{-1}\|\mathbf{x}^{S^c} - B\mathbf{w}^J\|^2$. In addition,

$$\begin{aligned} \|\mathbf{x}^{S^c} - B\mathbf{w}^J\|^2 &\leq 2(\|\mathbf{x}^{S^c}\|^2 + \|B\mathbf{w}^J\|^2) \\ &\leq 2(\|\mathbf{x}^{S^c}\|^2 + |||B|||^2\|\mathbf{w}^J\|^2) \\ &\leq 2(\|\mathbf{x}^{S^c}\|^2 + \rho^2[1 + \|\mathbf{x}^J\|^2]) \end{aligned}$$

because $\|\mathbf{w}^J\|^2 = 1 + \|\mathbf{x}^J\|^2$ and $B \in \mathcal{B}(\rho, \sharp S^c, 1 + \sharp J)$. Moreover, $\|\mathbf{x}^{S^c}\|^2 \leq \|\mathbf{x}\|^2$ and $\|\mathbf{x}^J\|^2 \leq \|\mathbf{x}\|^2$ hence

$$\|\mathbf{x}^{S^c} - B\mathbf{w}^J\|^2 \leq 2([1 + \rho^2]\|\mathbf{x}\|^2 + \rho^2).$$

Then a lower bound of $\ln[f_{\text{reg}}(\mathbf{x}^{S^c}|\mathbf{x}^J, B, \Omega)]$ is

$$\ln\left[f_{\text{reg}}(\mathbf{x}^{S^c}|\mathbf{x}^J, B, \Omega)\right] \geq -\frac{\sharp S^c}{2}\ln[2\pi b] - \frac{\rho^2}{a} - \frac{1+\rho^2}{a}\|\mathbf{x}\|^2.$$

Finally the second term is bounded by

$$-\frac{\sharp S^c}{2}\ln[2\pi b] - \frac{\rho^2}{a} - \frac{1+\rho^2}{a}\|\mathbf{x}\|^2 \leq \ln\left[f_{\text{reg}}(\mathbf{x}^{S^c}|\mathbf{x}^J, B, \Omega)\right] \leq -\frac{\sharp S^c}{2}\ln[2\pi a]. \quad (14)$$

Using (13), (14) and $\sharp S + \sharp S^c = Q$, each function of the family $\mathcal{G}_{(S,J)}$ is bounded by

$$-\frac{Q}{2}\ln[2\pi b] - \frac{\rho^2+\eta^2}{a} - \frac{2+\rho^2}{a}\|\mathbf{x}\|^2 \leq \ln\left[f(\mathbf{x}|\theta_{(S,J)})\right] \leq -\frac{Q}{2}\ln\left[2\pi a\right].$$

Thus, for all $\theta_{(\text{S,J})} \in \Theta'_{(\text{S,J})}$ and all $\mathbf{x} \in \mathbb{R}^Q$,

$$|\ln[f(\mathbf{x}|\theta_{(S,J)})]| \leq C_1(a, b, Q, \eta, \rho) + C_2(\rho, a)\|\mathbf{x}\|^2$$

defining the envelope function G, where $C_1(a, b, Q, \eta, \rho)$ and $C_2(\rho, a)$ are two positive constantes.

To verify that G is h-integrable, we have to show that $\int \|\mathbf{x}\|^2 h(\mathbf{x})dx < \infty$.

$$\begin{aligned}
\int \|\mathbf{x}\|^2 h(\mathbf{x})d\mathbf{x} &= \int \|\mathbf{x}\|^2 f(\mathbf{x}|\theta^\star_{(S_0,\,J_0)})dx \\
&= \int (\|\mathbf{x}^{S_0}\|^2 + \|\mathbf{x}^{S_0^c}\|^2) f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|\mathbf{x}^{J_0}, B^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&\leq \int \|\mathbf{x}^{S_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int 2(\|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2 + \|B^\star \mathbf{w}^{J_0}\|^2) f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|\mathbf{x}^{J_0}, B^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&= \int \|\mathbf{x}^{S_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} + 2\int \|B^\star \mathbf{w}^{J_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&+ \int 2\|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|\mathbf{x}^{J_0}, B^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&= A_1 + A_2 + A_3. \qquad (15)
\end{aligned}$$

The behavior of the three integrals A_1, A_2 and A_3 is studied separately.

$$\begin{aligned}
A_1 &= \int \|\mathbf{x}^{S_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&= \sum_{k=1}^{K} p_k \int \|\mathbf{x}^{S_0}\|^2 \phi(\mathbf{x}^{S_0}|\mu_k, \Sigma_k) d\mathbf{x}^{S_0} \\
&\leq \sum_{k=1}^{K} p_k [2\|\mu_k\|^2 + 2\,\text{tr}(\Sigma_k)]
\end{aligned}$$

according to Lemma 4. Thus, from Lemma 3 and since $\sum_{k=1}^{K} p_k = 1$,

$$A_1 \leq 2\eta^2 + 2b\sharp S_0.$$

$$\begin{aligned}
A_2 &= \int \|B^\star \mathbf{w}^{J_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&\leq \int \rho^2 (1 + \|\mathrm{x}^{S_0}\|^2) f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \\
&\leq \rho^2 \int f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} + \rho^2 A_1 \\
&\leq \rho^2 + \rho^2 [2\eta^2 + 2b\sharp S_0].
\end{aligned}$$

$$
\begin{aligned}
A_3 &= \int \|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2 f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) f_{\text{reg}}(\mathbf{x}^{S_0^c}|\mathbf{x}^{J_0}, B^\star, \Omega^\star) d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&= \int f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) \int \|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2 |2\pi\Omega^\star|^{-1/2} \exp\left[-\frac{1}{2}\|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2_{\Omega^\star}\right] d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0} \\
&\leq \int f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) \int \|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2 (2\pi b)^{-\sharp S_0^c/2} \exp\left[-\frac{1}{2b}\|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2\right] d\mathbf{x}^{S_0^c} d\mathbf{x}^{S_0}
\end{aligned}
$$

because $|\Omega^\star|^{-1/2} \leq b^{-\sharp S_0^c/2}$ and $\|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2_{\Omega^\star} \geq b^{-1}\|\mathbf{x}^{S_0^c} - B^\star \mathbf{w}^{J_0}\|^2$ according to Lemma 3. Thus, from Lemma 4,

$$
\begin{aligned}
A_3 &\leq \int f_{\text{clust}}(\mathbf{x}^{S_0}|\alpha^\star) d\mathbf{x}^{S_0} \times \int \|y\|^2 \phi(y|0, bI_{\sharp S_0^c}) dy \\
&= b\sharp S_0^c.
\end{aligned}
$$

So turning back (15), $\int \|\mathbf{x}\|^2 h(\mathbf{x}) d\mathbf{x} \leq 2(b\sharp S_0^c + \rho^2) + (1 + 2\rho^2)(2\eta^2 + 2b\sharp S_0)$ and finally G is h-integrable. Since $\ln(h) \in \mathcal{G}_{(S_0,\, J_0)}$, it implies that $\mathbb{E}[|\ln h(X)|] \leq \mathbb{E}[G(X)] < \infty$ and the law of large numbers can be applied to obtain (11). □

6 Numerical experiments

This section is devoted to the presentation of numerical experiments to illustrate the behavior of our methodology. First in Section 6.1, datasets presented in Raftery and Dean (2006) are considered to compare the behavior of both approaches. In Section 6.2, results of numerical experiments on simulated datasets are reported. Those numerical experiments aim to highlight some important features of the variable selection model we consider. Finally an application of the methodology for clustering a genomic dataset is described in Section 6.3.

6.1 Real datasets examples

Results on two real datasets, considered in Raftery and Dean (2006), where the correct number of clusters is supposed to be known are summarized.

Iris Data This well-known dataset consists of 150 samples of equally distributed Iris species described with four measurements (Anderson, 1935). We do not detail the results of our variables selection procedure since they do not differ from those of the Raftery and Dean procedure. Mixture model $[pL_kC_k]$ with $K = 3$ clusters with three clustering variables (all but sepal length which is explained with by the three clustering variables) has been selected with our procedure.

Leptograpsus crabs Data This dataset (Campbell and Mahon, 1974) consists of 200 subjects, divided into 100 of species orange and 100 of species blue, each with 50 males and 50 females. Each crab is described with five measurements: width of frontal lip (FL),

rear width (RW), length along the mid-line of the carapace (CL), maximum width of the carapace (CW) and body depth (BD) in mm. When no variable selection is done and the number of groups is not fixed, we obtain the clustering model $[pLC]$ with seven clusters by maximizing $\text{BIC}_{\text{clust}}$. The corresponding error rate with the four class species is 42%, as can be seen from the confusion matrix in Table 1. But, if the number of clusters is

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7
class 1	18	0	0	**32**	0	0	0
class 2	18	0	0	0	**32**	0	0
class 3	0	0	0	0	0	**28**	22
class 4	0	**24**	21	0	0	0	5

Table 1: Confusion table between the clusters and the class species with all variables and $K = 7$ for the Crabs dataset.

fixed to four and no variable selection is done, the error rate comes down to 7% with model $[pLD'_kAD_k]$ as seen from the confusion matrix in Table 2. Now, using our variable selection

	cluster 1	cluster 2	cluster 3	cluster 4
class 1	**42**	8	0	0
class 2	0	**49**	1	0
class 3	0	0	0	**50**
class 4	0	0	**45**	5

Table 2: Confusion table between the clusters and the class species with all variables and $K = 4$ for the Crabs dataset.

procedure with a number of clusters varying from two to ten, a four cluster solution for model $[pLD'_kAD_k]$ with clustering variables FL, RW, CW and BD is selected and variable CL is explained by all the clustering variables. Thus the error rate of the selected clustering is 7% as in Raftery and Dean (2006). Here again there is little difference between both approaches. Note that the variable selection procedures do not improve the error rate. And their gain, as for Iris dataset, concerns essentially the interpretation of the variable roles. The point to be remarked here is that assuming equal proportions leads to a much more interesting clustering model for this dataset.

	cluster 1	cluster 2	cluster 3	cluster 4
class 1	0	9	**41**	0
class 2	0	**49**	1	0
class 3	**49**	0	1	0
class 4	3	0	0	**47**

Table 3: Confusion table between the clusters and the class species with our variable selection procedure and $K = 4$ for the Crabs dataset.

6.2 Simulation data

First example This is the first simulated example in Law *et al.* (2004). The dataset consists of 800 points from a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_i, I),\ i \in \{1,2,3,4\}$ where

$$\mu_1 = \begin{pmatrix} 0 \\ 3 \end{pmatrix} \quad \mu_2 = \begin{pmatrix} 1 \\ 9 \end{pmatrix} \quad \mu_3 = \begin{pmatrix} 6 \\ 4 \end{pmatrix} \quad \mu_4 = \begin{pmatrix} 7 \\ 10 \end{pmatrix}.$$

Eight noisy variables, sampled from a $\mathcal{N}(0,1)$ density, have been appended to these data . The true model is $(K_0 = 4, m_0 = [pLI], S_0 = \{1,2\}, J_0 = \emptyset)$. As in Raftery and Dean (2006), if a variable selection procedure is not introduced in the regression, all the variables are declared significant for the clustering by our algorithm. On the contrary, with the variable selection procedure in the regression, our algorithm finds the true model with an error rate of $0,25\%$. By comparison, MIXMOD, using all the variables, chooses the model (K_0, m_0) and provides the same error rate. But the model incorporating a variable selection procedure is more parsimonious, and its interpretation of the variable roles is more significant.

Second example Each dataset consists of 800 data points from a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_i, \Sigma_i)$ with

$$\mu_1 = \begin{pmatrix} -a \\ -a \end{pmatrix} \quad \mu_2 = \begin{pmatrix} -a \\ a \end{pmatrix} \quad \mu_3 = \begin{pmatrix} a \\ -a \end{pmatrix} \quad \mu_4 = \begin{pmatrix} a \\ a \end{pmatrix}$$

where $a \in \{2,3,5\}$ and

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 3 & 0 \\ 0 & 0.5 \end{pmatrix} \quad \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \Sigma_4 = \Sigma_1.$$

Five noisy variables, sampled from a $\mathcal{N}(0,1)$ density, have been appended to these data. The true model is $(K_0 = 4, m_0 = [pL_kB_k], S_0 = \{1,2\}, J_0 = \emptyset)$. The results summarized in Table 4, show that our procedure selects the true role of the variables, the true number of clusters and a model m in the diagonal family, in the three considered cases. It can be remarked that our procedure outperforms MIXMOD, which does not proceed to a variable selection, when the clusters are poorly separated $(a = 2)$. Actually, variable selection in clustering can be expected the more useful in such cases.

Third example The same mixture of four equiprobable Gaussian distributions described in the previous example is considered. An additional variable is defined by $\mathbf{y}_i^3 = 3\mathbf{y}_i^1 + \epsilon_i$, ϵ_i being a Gaussian noise $\mathcal{N}(0, 0.5)$ independent of $\mathbf{y}_i^1$, for $i = 1, \ldots, n$. And, five noisy variables, sampled from a $\mathcal{N}(0,1)$ density, have been appended. The true model is $(K_0 = 4, m_0 = [pL_kB_k], S_0 = \{1,2\}, J_0 = \{1\})$. The results are summarized in Table 5. Our variable selection procedure improves the clustering (error rates are much smaller than with MIXMOD) and chooses the right variable roles.

	our algorithm					MIXMOD		
a	$\hat{K}$	$\hat{m}$	$\hat{S}$	$\hat{J}$	error rate	$\hat{K}$	$\hat{m}$	error rate
5	4	$[pL_kB_k]$	{1,2}	$\emptyset$	0%	4	$[pLB_k]$	0%
3	4	$[pLB_k]$	{1,2}	$\emptyset$	1%	4	$[pLB_k]$	0,875%
2	4	$[pLB_k]$	{1,2}	$\emptyset$	6,875%	4	$[pLI]$	11,25%

Table 4: Clustering results with our algorithm and MIXMOD for the second simulated example.

	our algorithm					MIXMOD		
a	$\hat{K}$	$\hat{m}$	$\hat{S}$	$\hat{J}$	error rate	$\hat{K}$	$\hat{m}$	error rate
5	4	$[pL_kB_k]$	{1,2}	{1}	0%	4	$[pL_kC_k]$	0%
3	4	$[pL_kB_k]$	{1,2}	{1}	0,875%	5	$[p_kLC]$	11,875%
2	4	$[pLB_k]$	{1,2}	{1}	9,25%	5	$[pLC]$	22,875%

Table 5: Results with our algorithm and MIXMOD for the third simulated example.

Remark about identifiability of the proposed model. Apparently, in this example, the role of vectors $\mathbf{y}^3$ and $\mathbf{y}^1$ can be exchanged since, we have $\mathbf{y}_i^1 = \mathbf{y}_i^3/3 - \epsilon_i$. But, $\mathbf{y}_i^1$ and ϵ_i are independent while $\mathbf{y}_i^3$ and ϵ_i are not independent. And the conditional distribution of $\mathbf{y}^1$ knowing $\mathbf{y}^3$ is not a Gaussian distribution in general. For this very reason, models $(S = \{1, 2\}, J = \{1\})$ and $(S = \{3, 2\}, J = \{3\})$ are not equivalent.

Waveform dataset The waveform dataset, available at the UCI repository (Blake *et al.*, 1999), is composed of three groups based on a random convex combination of two of three waveforms sampled at integers with noise added. A detailed description is available in Breiman *et al.* (1984). The dataset, studied in this paper, is extracted at random from this waveform dataset and consists of 900 observations divided into 300 of each group and described by 40 variables where the nineteen last are noisy variables, sampled from a $\mathcal{N}(0, 1)$ density. Among all models and with a cluster number varying between two and ten, MIXMOD selects a Gaussian mixture $[p_kLI]$ with six clusters. This clustering reveals the group construction according to the three wave functions. When our variable selection procedure is applied with $K \in \{3, 6\}$ and with spherical and diagonal models. The selected model is

$$(\hat{K} = 6, \hat{m} = [pLI], \hat{S} = (3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 18, 19), \hat{J} = (8, 11, 15)).$$

All of the noisy variables and seven "real" variables are declared irrelevant. This resulting variable partition gives additional information on the clustering. Moreover the final clustering is coherent with the construction of the sample: three clusters correspond to the three wave functions and the three others to convex combinations of two wave functions, as illustrated in Figure 2.

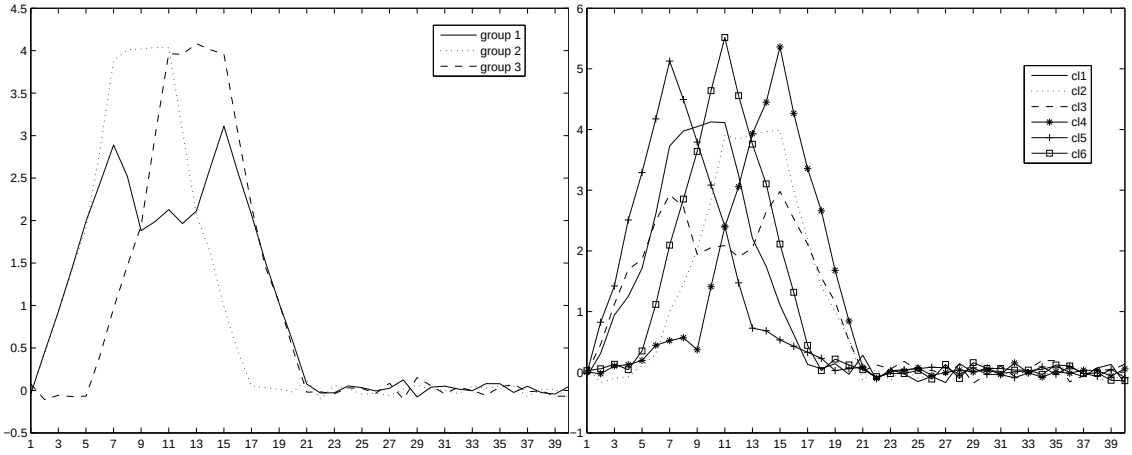

Figure 2: On the left, average profiles of the three real groups and on the right, average profiles of the six clusters found with our variable selection procedure.

6.3 Application to transcriptome data

For few years, the number of transcriptome data has been increased so that it is now feasible to investigate them with a global point of view. Studying transcriptomes under different physiological conditions or different stress situations allows us either to better understand how an organism works or to obtain more information on a subset of genes of interest. For that purpose clustering methods such as hierarchical clustering or K-means algorithm are commonly applied to find clusters of co-expressed genes (see for instance Sharan, Elkon, and Shamir, 2002, or Jiang, Tang and Zhang, 2004, and references therein). Indeed it is usually considered that co-expressed genes are often implicated in a same biological function and consequently are potential candidates to be co-regulated genes. In contrast to clustering methods based on distance, our method which belongs to model-based clustering methods, allows to take data variability into account, and select relevant variables which could be informative from a biological point of view.

To assess the behavior of our method on such type of data, it is applied on transcriptome data of the model plant *Arabidopsis thaliana*. Considered data are extracted from the database CATdb developed by Brunaud *et al.* (2007). An advantage of this database is that all data are produced with a microarray CATMA (Crowe *et al.* 2003) on the same platform with the same protocol. Moreover in CATdb statistical analyses required to remove the technical biases (normalization) and to determine the genes significantly differentially expressed (differential analysis) between two conditions are identical for all transcriptome experiments. The reader is referred to Lurin *et al.* (2004) for a description of such an analysis. From a statistical point of view, it means that the technical variability of the dataset is controllable and it is possible to consider that the level of technical variability

is homogeneous among the experiments. We focus on 1020 genes of *Arabidopsis thaliana* declared differentially expressed at least once in the time course of the hypotocotyl growth switch (Project 6 in Table 6). The aim of the biologists is to study them to characterize more precisely their biological functions and to determine their role in the hypotocotyl growth. The behavior of these 1020 genes on seven transcriptome projects have been studied. Each project is composed of a set of experiments dedicated to a specific biological question (see Table 6). Finally $Q = 27$ experiments partitioned into $T = 7$ block variables are available. Each gene is described with a vector $\mathbf{y}_g \in \mathbb{R}^{27}$, the component y_{gj} corresponding to the test statistic calculated in the experiment j for the differential analysis.

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 cell division to differentiation
4	3	transcriptome profiling from a protoplast culture
5	3	transcriptome response to nematode infection
6	3	transcriptome time course of the hypotocotyl growth switch
7	2	transcriptome of the hypotocotyl growth switch to isoxaben treatment

Table 6: 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.

Gaussian mixture model including our variable selection procedure has been performed with a maximal number of mixture components fixed to $K_{\max} = 20$ and with equal volume (see Table A). The selected model is the model $[p_k LC]$ with $\hat{K} = 17$ clusters. The relevant block variables are Projects 1, 3, 4, 6 and 7 and the four last clustering block variables enter in the regression model. It is worthwhile to remark that Project 6, used to define the gene subset under study, has been declared relevant for the clustering, as Project 7, relating also to the hypotocotyl growth switch. Among the 17 clusters with different size (see Table 7), some already known gene subsets are recovered and some clusters interested 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 well-studied genes (cluster 13).

cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
nb genes	7	120	25	11	42	424	14	29	43	94	11	149	19	4	6	2	20

Table 7: Size of the 17 estimated clusters.

In a comparison purpose, the data have been modelled with a Gaussian mixture including all the variables and with equal volume. The selected model is the mixture model $[p_k LC]$ with $\hat{K} = 16$ clusters. It shows some genes sets, common to the two clusterings (see Table 8) but the clusters are less homogeneous than with variable selection.

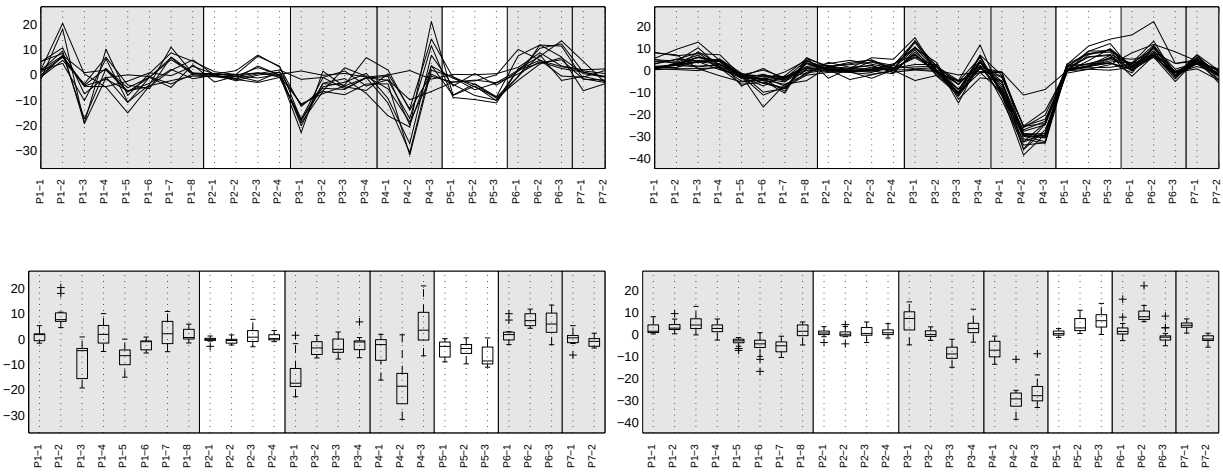

Figure 3: Graphical representation of genes profiles in Clusters 4 (on the left) and 17 (on the right). Relevant projects are colored in grey and on the x-axis $Pi - j$ denotes experiment j in project i.

			clustering without variable selection															
			cl1	cl2	cl3	cl4	cl5	cl6	cl7	cl8	cl9	cl10	cl11	cl12	cl13	cl14	cl15	cl16
			21	3	27	18	144	10	35	46	15	8	397	5	97	13	40	141
clustering with variable selection	cl1	7		1		6												
	cl2	120					107			1		6			1	5		
	cl3	25			22			1					2					
	cl4	11		2		6				2			1					
	cl5	42				1	2	5	2	29			3					
	cl6	424			2	3	2	3	26	1			373		11	1	2	
	cl7	14								12						1	1	
	cl8	29					11		3				11				1	3
	cl9	43	1				6	1					1				33	1
	cl10	94				1	9						1		83			
	cl11	11							2				4	1		2	2	
	cl12	149					5		1			2			2	2		137
	cl13	19					1			1	15					2		
	cl14	4			2	1			1									
	cl15	6					1						1	4				
	cl16	2	1		1													
	cl17	20	19														1	

Table 8: Comparison of the two clusterings, with and without variable selection.

7 Discussion

We have presented a general variable selection methodology for cluster analysis. Following Raftery and Dean (2006), this methodology considers the problem in the model-based cluster analysis context. In our approach, the role of the clustering variables with respect to the other variables is more versatile and can be expected to be more realistic especially for large dimension problems. As shown in the numerical experiments, this more general definition of variable role could avoid to overpenalize models with independent variables and our approach takes into account a common situation where the variables are partitioned into blocks. One of the interests of our model is to allow for a better and, sometimes subtle, interpretation of the variable role. For instance, in genomic applications, it could help biologists to improve functional annotation.

On the theoretical side , we have proved the consistency of our variable selection criterion under reasonable assumptions. Moreover, the identifiability of the proposed model parameter has been discussed with a simple example. Strictly speaking, there is no identifiability problem. But, in practice, it can happen that the clustering and the regression roles of two variables are permuted. However, since the regression equations involved in the resulting model are easy to be interpreted, ambiguous situations can be controlled by users.

Variable selection procedures need to use greedy algorithms which are expensive especially for large dimension data sets. Thus, in practical situations, it can be advantageous to avoid to select the model, the number of clusters and to define the variables status in the same exercise. We advocate to define the model and the number of clusters first, using all the variables as clustering variables, and to choose the variables status, in a second step, with a fixed model and number of clusters. Such a strategy is expected to provide a reliable and informative selection of the clustering variables in a reasonable amount of time. Finally, we want to stress that the procedure defined can work with alternative models linking the clustering and the remaining variables, provided that a BIC-like criterion analogous with our $\mathrm{BIC}_{\mathrm{reg}}$ criterion can be computed.

Acknowledgments

The authors thank Sylvie Huet (INRA) and Adrian Raftery (University of Washington) for helpful discussions, and Sebastien Aubourg (URGV), Jean-Pierre Renou (URGV) and Sandra Pelletier (IJPB-INRA) for their implication in the analysis of the transcriptome data.

Appendices

A The different Gaussian mixture model forms

This is the list of the 28 different Gaussian mixture model forms available in the MIXMOD software. Here $a = KQ$, $c = a + (K-1)$ and $b = \frac{Q(Q+1)}{2}$:

Family	Model	Proportion	Volume	Orientation	Shape	number of free parameters
Spherical	$[pLI]$	equal	equal	equal	NA	$a+1$
	$[pL_kI]$	equal	variable	equal	NA	$a+K$
Diagonal	$[pLB]$	equal	equal	coordinate axes	equal	$a+Q$
	$[pL_kB]$	equal	variable	coordinate axes	equal	$a+Q-1+K$
	$[pLB_k]$	equal	equal	coordinate axes	variable	$a+KQ-K+1$
	$[pL_kB_k]$	equal	variable	coordinate axes	variable	$a+KQ$
General	$[pLC]$	equal	equal	equal	equal	$a+b$
	$[pL_kC]$	equal	variable	equal	equal	$a+b+K-1$
	$[pLD'A_kD]$	equal	equal	equal	variable	$a+b-(K-1)(Q-1)$
	$[pL_kD'A_kD]$	equal	variable	equal	variable	$a+b-(K-1)Q$
	$[pLD'_kAD_k]$	equal	equal	variable	equal	$a+Kb-(K-1)Q$
	$[pL_kD'_kAD_k]$	equal	variable	variable	equal	$a+Kb-(K-1)(Q-1)$
	$[pLC_k]$	equal	equal	variable	variable	$a+Kb-(K-1)$
	$[pL_kC_k]$	equal	variable	variable	variable	$a+Kb$
Spherical	$[p_kLI]$	variable	equal	equal	NA	$c+1$
	$[p_kL_kI]$	variable	variable	equal	NA	$c+K$
Diagonal	$[p_kLB]$	variable	equal	coordinate axes	equal	$c+Q$
	$[p_kL_kB]$	variable	variable	coordinate axes	equal	$c+Q+K-1$
	$[p_kLB_k]$	variable	equal	coordinate axes	variable	$c+KQK+1$
	$[p_kL_kB_k]$	variable	variable	coordinate axes	variable	$c+KQ$
General	$[p_kLC]$	variable	equal	equal	equal	$c+b$
	$[p_kL_kC]$	variable	variable	equal	equal	$c+b+K-1$
	$[p_kLD'A_kD]$	variable	equal	equal	variable	$c+b-(K-1)(Q-1)$
	$[p_kL_kD'A_kD]$	variable	variable	equal	variable	$c+b-(K-1)Q$
	$[p_kLD'_kAD_k]$	variable	equal	variable	equal	$c+Kb-(K-1)Q$
	$[p_kL_kD'_kAD_k]$	variable	variable	variable	equal	$c+Kb-(K-1)(Q-1)$
	$[p_kLC_k]$	variable	variable	variable	variable	$c+Kb-(K-1)$
	$[p_kL_kC_k]$	variable	variable	variable	variable	$c+Kb$

Table 9: List of model forms available in MIXMOD

B Multidimensional Multivariate Regression

Let H be a $n \times V$ observed matrix of V response variables on each of n individuals. Let M be a $n \times A$ known matrix which represents a matrix of A observed variables on each of the

n individuals. The regression model is defined by

$$H_i = (H_i^1, \ldots, H_i^V) = (a_k + \sum_{j=1}^{A} \beta_j^k M_j^i, k = 1, \ldots, V) + E_i \text{ where } E_i \sim \mathcal{N}_V(0, \Omega).$$

It can be written

$$H = XB + E,$$

with

$$H = \begin{pmatrix} H_1^1 & \ldots & H_1^V \\ \vdots & & \vdots \\ H_n^1 & \ldots & H_n^V \end{pmatrix}, X = \left(\begin{array}{c|ccc} 1 & M_1^1 & \ldots & M_1^A \\ \vdots & \vdots & & \vdots \\ 1 & M_n^1 & \ldots & M_n^A \end{array}\right) \text{ and } B = \left(\begin{array}{ccc} a_1 & \ldots & a_V \\ \hline \beta_1^1 & \ldots & \beta_1^V \\ \vdots & & \vdots \\ \beta_A^1 & \ldots & \beta_A^V \end{array}\right)$$

The following theorem is proved in Mardia, Kent and Bibby (1979, Theorems 6.2.1) or Anderson (2003).

Theorem 2. *If $(X'X)^{-1}$ exists then defining $P = I - X(X'X)^{-1}X'$, the ml estimates of B and Ω are*

$$\hat{B} = (X'X)^{-1}X'H \text{ and } \hat{\Omega} = \frac{1}{n}H'PH. \tag{16}$$

BIC criterion for multidimensional multivariate regression is now derived. The model likelihood for data H is

$$f(H|M, B, \Omega) = |2\pi\Omega|^{-n/2} \exp\left[-\frac{1}{2}\operatorname{tr}[(H - XB)\Omega^{-1}(H - XB)']\right]. \tag{17}$$

The integrated likelihood is defined by

$$f(H|M) = \int f(H|M, B, \Omega)\pi(B, \Omega)d(B, \Omega)$$

where π is the prior distribution of the parameters. It can be written

$$f(H|M) = \int e^{nL_n(B,\Omega)}d(B, \Omega)$$

where

$$nL_n(B, \Omega) = -\frac{n}{2}\ln[|2\pi\Omega|] - \frac{1}{2}\operatorname{tr}\left[(H - XB)\Omega^{-1}(H - XB)'\right] + \ln[\pi(B, \Omega)].$$

Using Laplace approximation along a line detailed in Burnham and Anderson (2002) or in Lebarbier and Mary-Huard (2006), we get

$$f(H|M) = e^{nL_n(B^\star,\Omega^\star)}\left(\frac{2\pi}{n}\right)^{\nu/2} |-L_n''(B^\star, \Omega^\star)|^{-1/2}[1 + \mathcal{O}(n^{-1/2})],$$

where $(B^\star, \Omega^\star) = \text{argmax}\, L_n((B, \Omega)$ and $\nu = (A+1)V + \frac{V(V+1)}{2}$ is the number of free parameters in the regression model. Replacing $(B^\star, \Omega^\star)$ by $(\hat{B}, \hat{\Omega}) = \underset{B,\Omega}{\text{argmax}}\ f(H|M, B, \Omega)$ defined by (16) and $|-L_n''(B^\star, \Omega^\star)|$ by the Fisher information $I_{(\hat{B},\hat{\Omega})}$ which must be bounded, we get

$$2\ln[f(H|M)] = 2nL_n(\hat{B}, \hat{\Omega}) + \nu \ln\left(\frac{2\pi}{n}\right) - \ln[I_{(\hat{B},\hat{\Omega})}] + \mathcal{O}(n^{-1/2}).$$

From

$$2nL_n(\hat{B}, \hat{\Omega}) = -n\ln(|2\pi\hat{\Omega}|) - \text{tr}\left[(H - X\hat{B})\hat{\Omega}^{-1}(H - X\hat{B})'\right] + 2\ln(\pi(\hat{B}, \hat{\Omega}))$$

and remarking that

$$\text{tr}\left[(H - X\hat{B})\hat{\Omega}^{-1}(H - X\hat{B})'\right] = nV,$$

we get

$$\begin{aligned} 2\ln[f(H|M)] &= -n\ln[|2\pi\hat{\Omega}|] - nV + \nu\ln\left(\frac{2\pi}{n}\right) - \ln(I_{(\hat{B},\hat{\Omega})}) + 2\ln[\pi(\hat{B}, \hat{\Omega})] + \mathcal{O}(n^{-1/2}) \\ &\approx -n\ln[|2\pi\hat{\Omega}|] - nV - \nu\ln(n). \end{aligned}$$

We conclude that the BIC criterion for multivariate regression is

$$\text{BIC}_{\text{reg}}(H|M) = -n\ln[|2\pi\hat{\Omega}|] - nV - \nu\ln(n) \tag{18}$$

and in the simple regression context, ($V = 1$ and $\Omega = \sigma^2 > 0$), it becomes

$$\text{BIC}_{\text{reg}}(H|M) = -n\ln[2\pi\hat{\sigma}^2] - n - (A+2)\ln(n).$$

C The backward variable selection in regression

The procedure now described is comparing the models in competition with criterion BIC_{reg} defined in (18). Let $\mathbf{y}^c$ a variable to be explained with a linear regression model on a set S of dependent variables.

Initialisation $S[c] = S$, $j_E = \emptyset$ and $j_I = \emptyset$.

The algorithm is making use of an exclusion and an inclusion steps now described.

Exclusion step For all j in $S[c]$, compute

$$\text{B}_{\text{diffreg}}(\mathbf{y}^j) = \text{BIC}_{\text{reg}}(\mathbf{y}^c|\mathbf{y}^{S[c]}) - \text{BIC}_{\text{reg}}(\mathbf{y}^c|\mathbf{y}^{S[c]-j}).$$

Then, compute

$$j_E = \underset{j \in S[c]}{\text{argmin}}\, \text{BIC}_{\text{diffreg}}(\mathbf{y}^j).$$

- If $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^{j_E}) \leq 0$,
  - $S[c] = S[c] - j_E$
  - if $j_E = j_I$ stop,
  - otherwise go to the inclusion step;
- otherwise
  - if $j_I = \emptyset$ stop,
  - otherwise go to the inclusion step.

Inclusion step For all j in $S - S[c]$, compute

$$\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j) = \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^c|\mathbf{y}^{S[c]\cup j}) - \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^c|\mathbf{y}^{S[c]}).$$

Then, compute

$$j_I = \underset{j\in S-S[c]}{\mathrm{argmax}}\, \mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^j).$$

- If $\mathrm{B}_{\mathrm{diffreg}}(\mathbf{y}^{j_I}) > 0$,
  - if $j_I = j_E$ stop
  - otherwise $S[c] = S[c] \cup j_I$ 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 the exclusion and the inclusion steps.

D Technical results related to the consistency proof

Proposition 2.
Assume that

1. $(X_1, \ldots, X_n)$ *is a n-sample with unknown density* h.
2. Θ *is a compact metric space.*
3. $\theta \in \Theta \mapsto \ln[f(\boldsymbol{x}|\theta)]$ *is continuous for every* $\boldsymbol{x} \in \mathbb{R}^Q$.
4. G *is an envelope function of* $\mathcal{G} := \{\ln[f(.|\theta)];\ \theta \in \Theta\}$ *which is* h*-integrable.*
5. $\theta^\star = \underset{\theta\in\Theta}{\mathrm{argmax}}\, KL[h, f(.|\theta)]$
6. $\hat{\theta} = \underset{\theta\in\Theta}{\mathrm{argmax}} \sum_{i=1}^n f(X_i|\theta)$.

Then $\frac{1}{n}\sum_{i=1}^{n}\ln\left[f(X_i|\hat{\theta})\right] \underset{n\to\infty}{\overset{P}{\to}} \mathbb{E}_X[\ln f(X|\theta^\star)]$.

Proof.

$$\begin{aligned}\left|\mathbb{E}_X[\ln f(X|\theta^\star)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\hat{\theta})]\right| &\leq \left|\mathbb{E}_X[\ln f(X|\theta^\star)] - \mathbb{E}_X[\ln f(X|\hat{\theta})]\right| \\ &\quad + \sup_{\theta\in\Theta}\left|\mathbb{E}_X[\ln f(X|\theta)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta)]\right|.\end{aligned}$$

According to the definition of $\theta^\star$, $\mathbb{E}_X[\ln(f(X|\theta^\star))] - \mathbb{E}_X[\ln(f(X|\hat{\theta}_n))] \geq 0$, thus

$$\begin{aligned}\left|\mathbb{E}_X[\ln f(X|\theta^\star)] - \mathbb{E}_X[\ln f(X|\hat{\theta})]\right| &= \mathbb{E}_X[\ln f(X|\theta^\star)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta^\star)] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta^\star)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\hat{\theta})] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\hat{\theta})] - \mathbb{E}_X[\ln f(X|\hat{\theta})] \\ &\leq 2\sup_{\theta\in\Theta}\left|\mathbb{E}_X[\ln f(X|\theta)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta)]\right|.\end{aligned}$$

According to Example 19.8 in Van der Vaart (1998), the bracketing numbers of $\mathcal{G}$ are finite under the assumptions. Hence, using Theorem 19.4 in Van der Vaart (1998), $\mathcal{G}$ is P-Glivenko-Cantelli. Thus $\sup_{\theta\in\Theta}\left|\mathbb{E}_X[\ln f(X|\theta)] - \frac{1}{n}\sum_{i=1}^{n}\ln[f(X_i|\theta)]\right| \underset{n\to\infty}{\overset{P}{\to}} 0$, which concludes the proof. □

Lemma 3. *Let* $\Sigma \in \mathcal{D}_r$ *where* $\mathcal{D}_r$ *is defined in (H2). Then*

1. $a^r \leq |\Sigma| \leq b^r$ *and* $\mathrm{tr}(\Sigma) \leq br$
2. $\forall \boldsymbol{x} \in \mathbb{R}^r, b^{-1}\|\boldsymbol{x}\|^2 \leq \|\boldsymbol{x}\|^2_{\Sigma^{-1}} \leq a^{-1}\|\boldsymbol{x}\|^2$

Proof. The proof is based on the eigenvalue decomposition of the variance matrix Σ and the bounded constraint on the eigenvalues because $\Sigma \in \mathcal{D}_r$. □

Lemma 4.
Let $\phi(.|\mu,\Sigma)$ *be the density of the multivariate Gaussian distribution* $\mathcal{N}_r(\mu,\Sigma)$. *Then*

1. $\int \|y\|^2\phi(y|0,\Sigma)dy = \mathrm{tr}(\Sigma)$
2. $\int \|y\|^2\phi(y|\mu,\Sigma)dy \leq 2\left[\|\mu\|^2 + \mathrm{tr}(\Sigma)\right]$

Proof. The first result is a classical property of multivariate Gaussian densities. The second result is deduced from the first one using the triangle inequality. □

Lemma 5.
Let A and B be two real random variables,

$$\forall \epsilon \in \mathbb{R},\ P(A+B \leq 0) \leq P(A \leq \epsilon) + P(-B > \epsilon).$$

References

[1] Anderson, E. (1935), The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, **59**, 2-5.

[2] Anderson, T.W. (2003), *Introduction to Multivariate Statistical Analysis*. Third edition. New York, John Wiley & sons.

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

[4] 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.

[5] Blake, C., Keogh, E. and Merz, C. (1999), UCI repository of Machine Learning databases.

[6] Bouveyron, C., Girard, S. and Schmid, C. (2007), High-Dimensional Data Clustering. *Computational Statistics and Data Analysis*, to appear.

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

[8] Brunaud *et al.* (2007) cat*db: a Complete Arabidopsis Transcriptome database*, http://urgv.evry.inra.fr/CATdb.

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

[10] Burnham, K.P. and Anderson, D.R. (2002), *Model selection and multimodel inference*. New York, Springer-Verlag.

[11] Campbell, N.A. and Mahon, R.J. (1974), A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. *Australian Journal of Zoology*, **22**, 417-425.

[12] Celeux, G. and Govaert, G. (1995), Gaussian parsimonious clustering models. *Pattern Recognition*, **5**, 781-793.

[13] Crowe, M.L. *et al.* (2003), CATMA: a Complete Arabidopsis GST database. *Nucleic Acids Res*, **31**, 156-158.

[14] Dash, M., Choi, K., Scheuermann, P. and Liu, H. (2002) Feature Selection for Clustering - A Filter Solution. *Second IEEE International Conference on Data Mining (ICDM'02)*, pp. 115.

[15] 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.

[16] Devaney, M. and Ram, A. (1997), Efficient feature selection in conceptual clustering. *Machine Learning: Proceedings of the Fourteenth International Conference, Nashville, TN*, 92-97.

[17] Eisen, M., Spellman, P.L., Brown, P.O. and Botstain, D. (1998), Cluster Analysis and Display of Genome-wide Expression Patterns. *Proc. Nat. Acad. Sci. USA*, **95**, 14863-14868.

[18] Fisher, R.A. (1936), The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

[19] Fowlkes, E.B., Gnanadesikan, R. and Kettering, J.R. (1988), Variable selection in clustering. *Journal of classification*, **5**, 205-228.

[20] 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.

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

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

[23] 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.

[24] Jouve, P.E. and Nicoloyannis, N. (2005), A Filter Feature Selection Method for Clustering. *Proceedings ISMIS 2005*, pp. 583-593 .

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

[26] Kim, S., Tadesse, M.G. and Vannucci, M. (2006), Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93**, 321-344.

[27] Kohavi, R. and John, G.H. (1997), Wrapper for feature subset selection. *Artificial Intelligence*, **97**, 273-324.

[28] Law, M.H., Jain, A.K. and Figueiredo, M.A.T. (2004), Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence - PAMI*, **26**, 1154-1166.

[29] Lebarbier, E. and Mary-Huard, T. (2006), Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la SFdS.*, **147**, 39-58.

[30] Lurin, C. *et al.* (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell*, **16**(8), 2089-2103.

[31] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979), *Multivariate analysis*. Academic press inc.(London) LTD.

[32] Maugis, C., Celeux, G. and Martin-Magniette, M.-L. (2007), Variable Selection for Clustering with Gaussian Mixture Models. *Technical Report INRIA*.

[33] McLachlan, G.J. and Peel, D. (2000), *Finite Mixture Model*. New York: Wiley.

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

[35] Miller, A.J. (1990), *Subset Selection in Regression.* Chapman and Hall.

[36] Raftery, A.E. and Dean, N. (2006), Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, **101**, 168-178.

[37] Schwarz, G. (1978), Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464.

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

[39] 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.

[40] Van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge University Press.

Unité de recherche INRIA Futurs
Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399