Slope heuristics for variable selection and clustering via Gaussian mixtures

Cathy Maugis, Bertrand Michel

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Slope heuristics for variable selection and clustering via Gaussian mixtures

Cathy Maugis — Bertrand Michel

N° 6550

Juin 2008

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--6550--FR+ENG

Slope heuristics for variable selection and clustering via Gaussian mixtures

Cathy Maugis* , Bertrand Michel †

Thème COG — Systèmes cognitifs
Projets SELECT

Rapport de recherche n° 6550 — Juin 2008 — 32 pages

Abstract: Specific Gaussian mixtures are considered to solve simultaneously variable selection and clustering problems. A penalized likelihood criterion is proposed in Maugis and Michel (2008) to choose the number of mixture components and the relevant variable subset. This criterion is depending on unknown constants to be approximated in practical situations. A "slope heuristics" method is proposed and experimented to deal with this practical problem in this context. Numerical experiments on simulated datasets, a curve clustering example and a genomics application highlight the interest of the proposed heuristics.

Key-words: Model-based clustering, Variable selection, Penalized likelihood criterion, Slope heuristics, Curve clustering.

* INRIA Futurs, Projet SELECT, Université Paris-Sud 11
† INRIA Futurs, Projet SELECT, Université Paris-Sud 11

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

Heuristique de pente pour la sélection de variables et la classification non supervisée via des mélanges gaussiens

Résumé : Des mélanges gaussiens de formes spécifiques sont considérés pour résoudre un problème de sélection de variables en classification non supervisée. Un critère de vraisemblance pénalisée est proposé dans Maugis and Michel (2008) pour sélectionner le nombre de composantes du mélange et le sous-ensemble des variables significatives pour la classification. Ce critère dépend de constantes multiplicatives inconnues qui doivent être évaluées en pratique. Une méthode heuristique dite "de la pente" est proposée et expérimentée pour résoudre ce problème. Des exemples numériques sur données simulées, un exemple de classification de courbe et une application génomique mettent en évidence l'intérêt de cette procédure.

Mots-clés : Classification, Mélanges gaussiens, Sélection de variables, Critère pénalisé, Heuristique de pente, Classification de courbes.

1 Introduction

Model-based clustering methods consist of modelling each cluster with a parametric distribution and considering the mixture of these distributions to describe the whole dataset. They provide a rigorous framework to assess the number of mixture components and to take into account the variable roles.

Currently, cluster analysis is more and more concerned with large datasets where observations are described by many variables. This large number of predictor variables could be beneficial to data clustering. Nevertheless, the useful information for clustering can be contained into a variable subset and some of the variables can be useless or even harmful to choose a reasonable clustering structure. Several authors suggest variable selection methods for Gaussian mixture clustering which is the most widely used mixture model for clustering multivariate continuous datasets. These methods are "wrapper" methods since they are included into the clustering process. Law et al. (2004) have introduced the feature saliency concept. Regardless of cluster membership, relevant variables are assumed to be independent of the irrelevant variables which are supposed to have the same distribution. Raftery and Dean (2006) recast variable selection for clustering into a global model selection problem. Irrelevant variables are explained by all the relevant clustering variables according to a linear regression. The comparison between two nested variable subsets is performed using Bayes factor. A variation of this method is proposed in Maugis et al. (2007) where irrelevant variables can only depend on a relevant clustering variable subset and variables can have different sizes (block variables). Since all these methods are based on a variable selection procedure included into the clustering process, they do not impose specific constraints on Gaussian mixture forms. On the contrary, Bouveyron et al. (2007) consider a suitable Gaussian mixture family to take into account that data live in low-dimensional subspaces hidden in the original space. However, since this dimension reduction is based on principal components, it is difficult to deduce from this approach an interpretation of the variable roles.

In this paper, a new variable selection method for clustering is proposed. It recasts variable selection and clustering problems into a model selection problem in a density estimation framework. Suppose that we observe a sample from an unknown probability distribution with density s. A specific collection of models is defined: Each model $\mathcal{S}_{(K,\mathbf{v})}$ corresponds to a particular clustering situation where the cluster number is K and $\mathbf{v}$ is the relevant clustering variable subset. A density t belonging to $\mathcal{S}_{(K,\mathbf{v})}$ is decomposed into a Gaussian mixture density with K components on the relevant clustering variable subset $\mathbf{v}$ and a multidimensional Gaussian density on the other variables. Definitions of models $\mathcal{S}_{(K,\mathbf{v})}$ are precised in Section 2. The problem can be formulated as the choice of a model among the collection since this choice automatically leads to a data clustering and a variable selection. Thus, a data-driven criterion is needed to select the "best" model among the model collection. This article is the companion of Maugis and Michel (2008) where a penalized likelihood criterion is proposed. The results obtained in this previous paper allow to specify the general shapes of the criterion penalties but additional work is necessary to use these penalties in prac-

tice. The aim of this paper is to describe the practical use of these theoretical results. The so-called "slope heuristics" method is applied to calibrate penalties on the data.

First, our methodology is applied to a curve clustering problem. Curve clustering deals with the problem of identifying homogeneous groups in a set of functional data. This situation occurs in many areas of sciences, for instance in genetics, neuroscience, economics and engineering. Many methods for making curve clustering are based on different versions of the k-means algorithm. A widely used technique consists of finding a convenient projection of the functional data into a finite dimensional subspace, and next of applying a k-means procedure on the finite dimensional data obtained. In this context, B-spline bases are currently used, see for instance Abraham et al. (2003) and García-Escudero and Gordaliza (2005). An other approach proposed by Tarpey and Kinateder (2003) is to adapt the k-means algorithm for functional spaces. With a different point of view, Ma et al. (2006) use mixture models on B-splines coefficients, as in the works of James and Sugar (2003) for sparsely sampled functional data. In most of the cited works, each curve is described with a coefficient vector and in practice, the number of these coefficients can be of the same order as the curve number. This high dimensional context makes our method desirable to solve curve clustering problems. We illustrate the application of our method for curve clustering on the study of an oil production curve sample.

Next, we show that our method can be applied to the transcriptome data clustering in a particular context. During these last years, biologists are interested in determining biological functions of genes. In this aim, clustering methods such as hierarchical clustering or k-means algorithm are commonly applied to find clusters of co-expressed genes (see for instance Sharan et al., 2002; Jiang et al., 2004, and references therein). Since the experiment number increases in available transcriptome datasets, variable selection is more and more considered in order to improve the clustering and its interpretation to help biologists. In the model-based clustering context, Maugis et al. (2007) apply their variable selection method for a transcriptome dataset analysis for instance.

The paper is organized as follows: Section 2 presents the collections of Gaussian mixture model used in this paper. In Section 3, the general framework of model selection for density estimation based on Kullback-Leibler contrast is first recalled. Next, we present a penalized criterion to select a model into the considered collections of Gaussian mixture models, and we recall the results obtained on this criterion in Maugis and Michel (2008). Section 4 is devoted to the description of slope heuristics and its practical use in our specific context. Simulations and applications for a curve clustering problem and for a genomics study are presented in Section 5.

2 Gaussian mixture models

2.1 Multivariate Gaussian models and clustering

Centered 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 unknown density s. This target distribution is proposed to be

estimated by a finite mixture model in a clustering purpose. Model-based clustering 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 the data are multivariate continuous observations, the component parameterized density is usually a multidimensional Gaussian density. The general form of a Gaussian mixture model with K components is

$$\sum_{k=1}^{K} p_k \Phi(\cdot|\eta_k, \Lambda_k)$$

where the p_k's are the mixing proportions ($\forall k = 1, \ldots, K, 0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$) and $\Phi(\cdot|\eta_k, \Lambda_k)$ denotes the Q-dimensional Gaussian density with mean η_k and variance matrix Λ_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 $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ such that $z_{ik} = 1$ iff $\mathbf{y}_i$ arises from component k. The z's define an ideal clustering of the data $\mathbf{y}$ associated to the mixture model. After an estimation of the parameter vector thanks to the EM algorithm (Dempster et al., 1977), a data clustering is deduced using the maximum a posteriori principle (MAP rule):

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

2.2 Definitions of the considered Gaussian mixture models

Currently, statistics deals with problems where individuals are explained by many variables. In principle, the more information we have about each individual, the better a clustering method is expected to perform. Nevertheless, some variables can be useless or even harmful to obtain a good clustering of data. Thus, it is important to take into account the variable role in the clustering process. In this goal, we propose to consider Gaussian mixtures with a specific form, based on the following idea. On irrelevant variables, individuals have an homogenous behavior around the null mean (centered data) allowing not to distinguish different subpopulations. Hence data are assumed to have a spherical Gaussian joint law with null mean vector on these variables. On the contrary, the different component mean vectors are free on relevant clustering variables. Variance matrices restricted on relevant variables are either taken completely free or are chosen in a specified set of definite positive matrices depending on the considered situation. This idea is now formalized.

Let $\mathcal{V}$ be the set of the nonempty subsets of $\{1, \ldots, Q\}$. A Gaussian mixture family is characterized by its number of mixture components $K \in \mathbb{N}^\star$ and its relevant variable index subset $\mathbf{v} \in \mathcal{V}$ whose cardinal is denoted α. In the sequel, the set of index couples $(K, \mathbf{v})$ is $\mathcal{M} = \mathbb{N}^\star \times \mathcal{V}$. Consider the decomposition of a vector $x \in \mathbb{R}^Q$ into its restriction on relevant variables $x_{[\mathbf{v}]} = (x_{j_1}, \ldots, x_{j_\alpha})'$ and its restriction on irrelevant variables $x_{[\mathbf{v}^c]} = (x_{l_1}, \ldots, x_{l_{Q-\alpha}})'$ where $\mathbf{v} = \{j_1, \ldots, j_\alpha\}$ and $\mathbf{v}^c = \{l_1 \ldots, l_{Q-\alpha}\} = \{1, \ldots, Q\} \backslash \mathbf{v}$. On

relevant clustering variables, a Gaussian mixture f is chosen among the following family

$$\mathcal{L}_{(K,\alpha)} = \left\{ w \in \mathbb{R}^\alpha \mapsto \sum_{k=1}^{K} p_k \Phi(w|\mu_k, \Sigma_k);\ \begin{array}{l} \forall k,\ \mu_k \in [-a,a]^\alpha, (\Sigma_1, \ldots, \Sigma_K) \in \mathcal{D}^+_{(K,\alpha)} \\ 0 < p_k < 1, \sum_{k=1}^K p_k = 1 \end{array} \right\}$$

where $0 < a$ and $\mathcal{D}^+_{(K,\alpha)}$ denotes a family of K-uples of $\alpha \times \alpha$ symmetric definite positive matrices which is related to the Gaussian mixture shape specified hereafter. On irrelevant variables, a spherical Gaussian density belonging to the following family is considered

$$\mathcal{G}_{(\alpha)} = \left\{ u \in \mathbb{R}^{Q-\alpha} \mapsto \Phi(u|0, \omega^2 I_{Q-\alpha});\ \omega^2 \in [\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}] \right\}.$$

Thus, the family of Gaussian mixture associated to $(K, \mathbf{v}) \in \mathcal{M}$ is defined by

$$\mathcal{S}_{(K,\mathbf{v})} = \left\{ x \in \mathbb{R}^Q \mapsto f(x_{[\mathbf{v}]})\, g(x_{[\mathbf{v}^c]});\ f \in \mathcal{L}_{(K,\alpha)},\ g \in \mathcal{G}_{(\alpha)} \right\}. \tag{1}$$

The dimension of the model $\mathcal{S}_{(K,\mathbf{v})}$ is denoted $D(K, \mathbf{v})$ and corresponds to the free parameter number of Gaussian mixtures in this model. It only depends on the number of components K, the Gaussian mixture shape and the number of relevant variables α.

In this paper, four collections of Gaussian mixtures are considered but the same notation $\mathcal{S}_{(K,\mathbf{v})}$ is used for the four model collections to make easier the reading of this article. The Gaussian mixture notation for those collections is taken from Biernacki et al. (2006).

- For the $[L_k B_k]$ collection, the variance matrices are assumed to be diagonal and free. Thus, the variance matrices have the following form

  $$\Sigma_k = \mathrm{diag}(\sigma^2_{k1}, \ldots, \sigma^2_{k\alpha})$$

  where the eigenvalues σ^2_{kj} are assumed to be in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The associated dimension of model $\mathcal{S}_{(K,\mathbf{v})}$ is equal to $D(K, \mathbf{v}) = K(2\alpha + 1)$.

- For the $[L_k C_k]$ collection, the variance matrices are assumed to be totally free. Thus, the variance matrices are $\alpha \times \alpha$ positive definite matrices whose eigenvalues are assumed to belong to the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The associated model dimension is $D(K, \mathbf{v}) = K[1 + \alpha + \frac{\alpha(\alpha+1)}{2}]$.

- For the $[L B_k]$ collection, the variance matrices are assumed to be diagonal and to have the same volume i.e. $\forall k \neq k', |\Sigma_k|^{\frac{1}{\alpha}} = |\Sigma_{k'}|^{\frac{1}{\alpha}}$. The variance matrices are decomposed into $\Sigma_k = \beta B_k$ where the common volume β belongs to $[\beta_{\mathrm{m}}, \beta_{\mathrm{M}}]$ and B_k is a diagonal matrix with a determinant 1 and with diagonal coefficients in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. Here, the model dimension is equal to $D(K, \mathbf{v}) = 2K\alpha + 1$.

- For the $[LC]$ collection, the variance matrices are all equal to a free positive definite matrix Σ whose eigenvalues are assumed to be in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The model dimension is $D(K, \mathbf{v}) = K(1 + \alpha) + \frac{\alpha(\alpha+1)}{2}$.

In this paper, for each of the four possible model collections, the variables can be assumed to be ordered or not ordered. If variables are ordered, the relevant variable subset is $\mathbf{v} = \{1, \ldots, \alpha\}$ and can be assimilated to its cardinal α. Moreover, note that a density of $\mathcal{S}_{(K,\mathbf{v})}$ can be written as a global Gaussian mixture with mean vectors $\eta_k = (\mu_k, 0, \ldots, 0)$ and block-diagonal variance matrices Λ_k with diagonal-blocks Σ_k and $\omega^2 I_{Q-\alpha}$. Consequently, a data clustering can be deduced from the MAP rule given in Section 2.1.

These Gaussian mixture families allow to recast clustering and variable selection problems in a global model selection problem. A criterion is now required to select the best model according to the dataset.

3 A new penalized likelihood criterion

3.1 Model selection principle

Density estimation deals with the problem of estimating the unknown distribution of a sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$. In many cases, it is not obvious to choose a model of adequate dimension. For instance, a model with few parameters tends to be efficiently estimated whereas it could be far from the true distribution. On the opposite situation, a more complex model easily fits data but estimates have larger variances. The aim of model selection is to construct data-driven criterion to select a model of proper dimension among a given list. A general theory on this topic, with a non asymptotic approach is proposed in the works of Birgé and Massart (see Massart, 2007, for an overview). This model selection principle is now described in our setting.

Let $\mathcal{S}$ be the set of all densities with respect to the Lebesgue measure on $\mathbb{R}^Q$. The contrast $\gamma(t, \cdot) = -\ln\{t(\cdot)\}$ is considered, leading to the maximum likelihood criterion. The corresponding loss function is the Kullback-Leibler information. It is defined for two densities s and t in $\mathcal{S}$ by

$$\mathrm{KL}(s,t) = \int \ln\left\{\frac{s(x)}{t(x)}\right\} s(x)\,\mathrm{d}x$$

if sdx is absolutely continuous with respect to tdx and $+\infty$ otherwise. Noticing that s is a minimizer of the Kullback-Leibler information over $\mathcal{S}$, s is also a minimizer over $\mathcal{S}$ of the expectation of the empirical contrast, defined by

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \ln\left\{t(\mathbf{y}_i)\right\}.$$

A countable collection of models $\{\mathcal{S}_{(K,\mathbf{v})}\}_{(K,\mathbf{v})\in\mathcal{M}}$ is considered and let $\hat{s}_{(K,\mathbf{v})}$ be a minimizer of the empirical contrast γ_n over the model $\mathcal{S}_{(K,\mathbf{v})}$. Substituting the empirical criterion γ_n to its expectation and minimizing γ_n on $\mathcal{S}_{(K,\mathbf{v})}$ are expected to lead to a sensible estimator of s, at least if s belongs (or is close enough) to model $\mathcal{S}_{(K,\mathbf{v})}$. The model we want to select is the one presenting the smallest risk

$$(K^\star, \mathbf{v}^\star) = \operatorname*{argmin}_{(K,\mathbf{v})\in\mathcal{M}} \mathbb{E}[\mathrm{KL}(s, \hat{s}_{(K,\mathbf{v})})].$$

However, the function $\hat{s}_{(K^\star,\mathbf{v}^\star)}$, called oracle, is unknown since it depends on the true density s. Thus, the aim is to find a data-driven criterion to select an estimator such that its risk is as close as possible as the oracle risk. The model selection via penalization procedure consists of considering some penalized criterion

$$\text{crit}(K,\mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \text{pen}(K,\mathbf{v}) \tag{2}$$

where pen is a penalty function pen : $(K,\mathbf{v}) \in \mathcal{M} \mapsto \text{pen}(K,\mathbf{v}) \in \mathbb{R}_+$. Then the selected model $(\hat{K},\hat{\mathbf{v}})$ is a minimizer of the penalized criterion (2) and the associated selected estimator is $\hat{s}_{(\hat{K},\hat{\mathbf{v}})}$. The final purpose of a non asymptotic approach is to obtain a penalty function providing an oracle inequality. This oracle inequality allows to compare the risk of the penalized maximum likelihood estimator (MLE) $\hat{s}_{(\hat{K},\hat{\mathbf{v}})}$ with the benchmark $\inf_{(K,\mathbf{v})\in\mathcal{M}} \mathbb{E}[\text{KL}(s,\hat{s}_{(K,\mathbf{v})})]$, for a fixed number n of observations.

3.2 Theoretical results

From a theoretical point of view, the problem of defining a convenient penalized likelihood criterion for our specific Gaussian mixture model collections has been treated in Maugis and Michel (2008). This work has been made possible thanks to the use of a general model selection theorem for MLE, proposed by Massart (2007). The application of this theorem requires the control of bracketing entropy of Gaussian mixture families. It allows to obtain penalty function forms and associated oracle inequalities.

In the sequel, the norm $\|\sqrt{f}-\sqrt{g}\|_2$ between two nonnegative functions f and g of $\mathbb{L}_1$ is denoted $d_H(f,g)$. We note that if f and g be two densities with respect to the Lebesgue measure on $\mathbb{R}^Q$, $d_H^2(f,g)$ is the squared Hellinger distance between f and g

$$d_H^2(f,g) = \int_{\mathbb{R}^Q} \left(\sqrt{f(x)}-\sqrt{g(x)}\right)^2 \mathrm{d}x.$$

In the following, $d_H(f,g)$ is improperly called Hellinger distance even if f and g are not density functions. The following theorem summarizes the theoretical results established in Maugis and Michel (2008). These results are valid for the four Gaussian mixture models at hand. They are proved for the $[L_kB_k]$ shape and the $[L_kC_k]$ shape in Maugis and Michel (2008) and additional proofs for the two other shapes $[LB_k]$ and $[LC]$ are available in the Appendices.

Theorem 1 (Maugis and Michel (2008))**.** *For the four Gaussian mixture collections,*

1. *If the variables are ordered, there exists two absolute constants κ and C such that, if*

$$\text{pen}(K,\mathbf{v}) \geq \kappa\frac{D(K,\mathbf{v})}{n}\left[2\,A + \ln\left(\frac{1}{1\wedge\frac{D(K,\mathbf{v})}{n}A}\right)+1\right]$$

then the model $(\hat{K}, \hat{\mathbf{v}})$ *minimizing* $\mathrm{crit}(K, \mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})$ *on* $\mathcal{M}$ *exists and*

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{(\hat{K},\hat{\mathbf{v}})})\right] \leq C \left\{ \inf_{(K,\mathbf{v})\in\mathcal{M}} [\mathrm{KL}(s, \mathcal{S}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})] + \frac{1}{n} \right\}. \tag{3}$$

2. *If the variables are not ordered, there exists two absolute constants* κ *and* C *such that, if*

$$\mathrm{pen}(K, \mathbf{v}) \geq \kappa \frac{D(K, \mathbf{v})}{n} \left\{ 2\,A + \ln\left[\frac{1}{1 \wedge \frac{D(K,\mathbf{v})}{n} A} \right] + \frac{1}{2} \ln\left[\frac{8 \exp(1) Q}{(D(K, \mathbf{v}) - 1) \wedge (2Q - 1)} \right] \right\}$$

then the model $(\hat{K}, \hat{\mathbf{v}})$ *minimizing* $\mathrm{crit}(K, \mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})$ *on* $\mathcal{M}$ *exists and*

$$\mathbb{E}\left[d_H^2(s, \hat{s}_{(\hat{K},\hat{\mathbf{v}})})\right] \leq C \left\{ \inf_{(K,\mathbf{v})\in\mathcal{M}} [\mathrm{KL}(s, \mathcal{S}_{(K,\mathbf{v})}) + \mathrm{pen}(K, \mathbf{v})] + \frac{2}{n} \right\}. \tag{4}$$

In the four cases, A *is a function of parameters* λ_m, λ_M, a, Q *and also* β_m *and* β_M *for the* $[LB_k]$ *shape such that* $A = O(\sqrt{\ln Q})$ *as* Q *goes to infinity.*

The penalty functions take into account the model complexity through $D(K, \mathbf{v})$ and the richness of model family. Actually the number of models with the same dimension is larger in the non-ordered variable case, the associated penalty functions have an additional logarithm term, depending on the dimension.

The other logarithm term, common to both cases, is probably not necessary to define efficient penalties. As explain in Maugis and Michel (2008), the reason for that is certainly that the general model selection theorem for MLE is stated in a local version whereas we only manage to apply the global version in our framework. Logarithm terms are not detected in practice as shown in Section 5.1 and thus only the preponderant term in $\frac{D(K,\mathbf{v})}{n}$ is retained in the penalty form.

Contrary to classical criteria for which Q is fixed and n tends to infinity, our result allows to study cases for which Q increases with n. For specific clustering problems where the number of variables Q is of the order of n or even larger than n, the oracle inequality is still significant.

The point we want to stress here is that Theorem 1 gives the general form of penalty functions but is not totally explicit since the results depend on absolute unknown constants and mixture parameters are not bounded in practice. Consequently a method is to be applied to calibrate the penalty function for a practical use of these results.

4 Slope heuristics

The aim of this paper is to show how the theoretical results of Section 3.2 can be applied in practice. Since the lower bounds on penalty functions in Theorem 1 are defined up to an unknown multiplicative constant, this theorem does not provide directly a usable model

selection criterion. Last years, some efforts have been paid to overcome such a difficulty. Birgé and Massart (2006) propose a practical method based on a mixture of theoretical and heuristic ideas for defining efficient penalty functions from the data. This heuristics is only proved in Birgé and Massart (2006) in the framework of Gaussian regression with a homoscedastic fixed design and more recently generalized by Arlot and Massart (2008) in the heteroscedastic random-design case. Nevertheless applications of this method are developed in many other frameworks: For instance, in multiple change points detection by Lebarbier (2005), in genomics applications by Villers (2007) and in Gaussian Markov random fields by Verzelen (2008). This section first describes the main ideas of this heuristics, called the "*slope heuristics*", and next details its practical use in our framework.

4.1 Rationale for the slope heuristics

In many situations, the considered model collection contains several models with the same dimension. In order to penalize each model of dimension D in the same way, a new collection $(\mathcal{S}_D)_{D\in\mathcal{D}}$ is considered such that $\mathcal{S}_D$ is the union of all the models $\mathcal{S}_{(K,\mathbf{v})}$ having the same dimension $D(K,\mathbf{v}) = D$. It is recalled that γ and γ_n are the Kullback-Leibler contrast and its associated empirical contrast respectively (see Section 3.1). Moreover s_D and $\hat{s}_D$ denote a minimizer of $\mathrm{KL}(s,\cdot)$ and $\gamma_n(\cdot)$ on $\mathcal{S}_D$ respectively.

As for criteria due to Mallows (1973) and Akaike (1973, 1974), Birgé and Massart criterion is based on an unbiased risk estimation. The ideal model to estimate s is the one minimizing the risk $\mathbb{E}\left[\mathrm{KL}(s,\hat{s}_D)\right]$. Nevertheless, it is impossible to choose this optimal model since s is unknown. A solution is to find a penalty function, called *optimal penalty* such that the empirical risk is as close as possible to the risk.

The following classical decomposition of the risk of each estimator $\hat{s}_D$ is considered

$$\mathbb{E}\left[\mathrm{KL}(s,\hat{s}_D)\right] = b_D + \mathbb{E}(V_D)$$

where the bias term $b_D := \mathrm{KL}(s,s_D)$ and the variance term is defined by $V_D := \int \ln(s_D/\hat{s}_D)s$. Note that the bias b_D decreases whereas the variance term V_D tends to increase when the dimension D increases. Among the model collection $\mathcal{D}$, the selected model $\hat{D}$ is the one minimizing the criterion

$$D \mapsto \gamma_n(\hat{s}_D) + \mathrm{pen}(D). \tag{5}$$

Defining $\hat{b}_D := \gamma_n(s_D) - \gamma_n(s)$ and $\widehat{V}_D := \gamma_n(s_D) - \gamma_n(\hat{s}_D)$, the selected model dimension is also a minimizer of

$$\begin{aligned}\gamma_n(\hat{s}_D) - \gamma_n(s) + \mathrm{pen}(D) &= \hat{b}_D - \widehat{V}_D + \mathrm{pen}(D)\\ &= \mathrm{KL}(s,\hat{s}_D) + (\hat{b}_D - b_D) - (V_D + \widehat{V}_D) + \mathrm{pen}(D). \end{aligned}\tag{6}$$

Because of the law of large numbers, it is reasonable to assume that $\hat{b}_D - b_D \approx 0$. Furthermore, concentration arguments allow to suppose that $\mathrm{KL}(s,\hat{s}_D)$ is close to its expectation which is the risk of $\hat{s}_D$. In order to make the quantity in (6) close to the risk $\mathbb{E}\left[\mathrm{KL}(s,\hat{s}_D)\right]$,

the *optimal penalty* is defined by

$$\mathrm{pen}_{\mathrm{opt}}(D) = V_D + \widehat{V}_D.$$

Next, the main hypothesis of this heuristics is to assume that $\widehat{V}_D \approx V_D$. An argument to justify this hypothesis is that, in the expressions of V_D and $\widehat{V}_D$, the probability measure and the corresponding empirical measure play a similar role. If one permutes these measures inside the definitions of V_D and $\widehat{V}_D$, and also in the definitions of s_D and $\hat{s}_D$, then V_D is changed in $\widehat{V}_D$ and reciprocally. Finally, this hypothesis leads to $\mathrm{pen}(D) = 2\widehat{V}_D$. Turning back on the expression of $\widehat{V}_D$, it can be written

$$\widehat{V}_D = \hat{b}_D + \gamma_n(s) - \gamma_n(\hat{s}_D).$$

For large dimensions, the bias term stabilizes itself since the approximation of the model cannot be appreciably improved. Thus, the behavior of $\widehat{V}_D$ according to the model dimension is known for large dimensions via $-\gamma_n(\hat{s}_D)$. In our framework, penalty functions could be regarded as proportional to the dimension (see remarks after Theorem 1) and if the slope $\hat{C}$ of the linear part of $-\gamma_n(\hat{s}_D)$ is known, the final penalty is

$$\mathrm{pen}(D) = 2\hat{C}D.$$

4.2 Using the slope heuristics

This section details how the slope heuristics is applied to select a Gaussian mixture model among a family $\left(\mathcal{S}_{(K,\mathbf{v})}\right)_{(K,\mathbf{v})\in\mathcal{M}}$ with $\mathcal{M} := \{(K,\mathbf{v});\ 2 \leq K \leq K_{\max}, \mathbf{v} \in \mathcal{V}\}$ where the maximum number of mixture components $K_{\max}$ and the mixture shape are fixed by the user. The heuristics makes use of three steps:

1. *Estimation step*: The MLE is computed for each model $\mathcal{S}_{(K,\mathbf{v})}$. According to (1), the mixture parameters and ω^2 can be independently estimated. The estimated mixture parameters $\left(\hat{p}_1, \ldots, \hat{p}_K, \hat{\mu}_1, \ldots, \hat{\mu}_K, \widehat{\Sigma}_1, \ldots, \widehat{\Sigma}_K\right)$ are computed with the Expectation Maximization (EM) algorithm (Dempster et al., 1977) using MIXMOD software (Biernacki et al., 2006) and $\hat{\omega}^2 = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_{i[\mathbf{v}^c]}\|^2$. Thus, the MLE is

$$\hat{s}_{(K,\mathbf{v})}(x) = \sum_{k=1}^{K} \hat{p}_k \Phi\left(x_{[\mathbf{v}]}|\hat{\mu}_k, \widehat{\Sigma}_k\right) \times \Phi(x_{[\mathbf{v}^c]}|0, \hat{\omega}^2 I_{Q-\alpha}).$$

2. *Penalty determination step*: First, models are grouped according to their dimension in order to obtain the model collection $(\mathcal{S}_D)_{D\in\mathcal{D}}$. The function $D \mapsto -\gamma_n(\hat{s}_D)$ is plotted and a threshold D_0 is chosen by the user such that the function has a linear behavior for $D \geq D_0$. Using a robust regression (Huber, 1981) of $-\gamma_n(\hat{s}_D)$ on D, the slope $\hat{C}$ of the linear part is estimated. Thus the optimal penalty function is defined by

$$\mathrm{pen}_{\mathrm{opt}}(D) = 2\hat{C}D.$$

The robust regression procedure allows to mitigate the influence of possible estimation errors of the first step.

3. *Model selection step*: The minimizer $\hat{D}$ of the criterion $D \mapsto \gamma_n(\hat{s}_D) + 2\hat{C}D$ is determined. Then, among the initial model collection, we find the model $(\hat{K}, \hat{\mathbf{v}})$ fulfilling $D(\hat{K}, \hat{\mathbf{v}}) = \hat{D}$ and $\hat{s}_{(\hat{K},\hat{\mathbf{v}})} = \hat{s}_{\hat{D}}$. Finally, the parameter estimators associated to $(\hat{K}, \hat{\mathbf{v}})$ provide a data clustering using the MAP rule.

5 Applications

This section is devoted to the application of our method on simulated and real datasets. We show that our method allows us to determinate the variable role to improve the clustering. Moreover, we check that the penalized estimator mimics the oracle. The slope heuristics is compared with the classical criteria used for Gaussian mixture model selection: AIC, BIC and ICL. They are respectively defined by

$$\mathrm{crit}_{\mathrm{AIC}}(D) = \gamma_n(\hat{s}_D) + \frac{D}{n},\ \mathrm{crit}_{\mathrm{BIC}}(D) = \gamma_n(\hat{s}_D) + \frac{D\ln(n)}{2n} \text{ and } \mathrm{crit}_{\mathrm{ICL}}(D) = \mathrm{crit}_{\mathrm{BIC}} + \frac{\mathrm{ENT}}{n}$$

with the entropy term $\mathrm{ENT} = -\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln(t_{ik})$ where z is given by the MAP rule and $t_{ik} = \frac{\hat{p}_k\Phi(y_i|\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{l=1}^{K}\hat{p}_l\Phi(y_i|\hat{\mu}_l,\hat{\Sigma}_l)}$. The interested reader is respectively referred to Akaike (1973, 1974), Schwarz (1978) and Biernacki et al. (2000) for more details on these criteria.

The method is applied on simulated datasets in Section 5.1 and then on real datasets. In Section 5.2, our procedure is carried out on a curve clustering example for oil production profiles. In Section 5.3, a transcriptome dataset is studied with our method to obtain co-expressed gene clusters.

5.1 Simulated datasets

Comparison to the oracle and to other criteria. The aim of this first example is to compare the slope estimator with other penalized estimators and to study its behavior with respect to the oracle. The dataset consists of $n = 2000$ points described by $Q = 32$ variables. The data are simulated according to a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_k, \Sigma_k)$ where

$$\mu_1 = (3, 2, 1, 0.7, 0.3, 0.2, 0.1, 0.07, 0.05, 0.025),\ \mu_2 = 0_{10},\ \mu_3 = -\mu_1,$$
$$\mu_4 = (3, -2, 1, -0.7, 0.3, -0.2, 0.1, -0.07, -0.05, -0.025),$$

and

$$\Sigma_1 = \Sigma_3 = \Sigma_4 = I_{10} \text{ and } \Sigma_2 = \mathrm{diag}(2, 1.9, 1.8, \ldots, 1.1).$$

The vector 0_p denotes the null vector of length p. Twenty two independent variables, sampled from a $\mathcal{N}(0, 1)$, are appended. Consequently, the true density belongs to the model $\mathcal{S}_{(K_0,\mathbf{v}_0)}$

where $K_0 = 4$ and $\mathbf{v}_0 = \{1, \ldots, 10\}$ ($\alpha_0 = 10$) and the variables are ordered. Note that the discriminant power of the relevant variables decreases with respect to the variable index. In other words the four subpopulations of the mixtures are progressively gathered together into a unique Gaussian distribution, as shown in Figure 1.

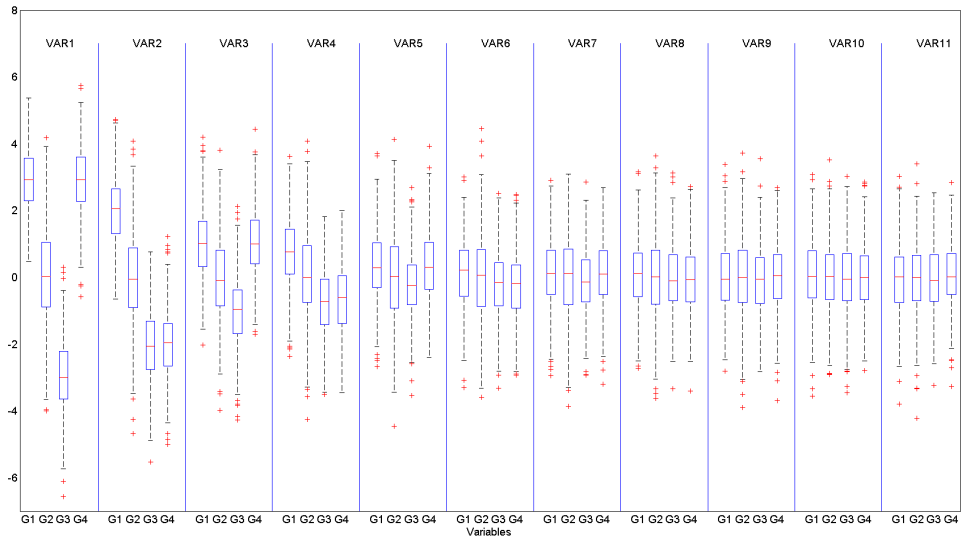

Figure 1: Boxplots of the first eleven variables (VAR1,...,VAR11) on the four mixture components (G1,G2,G3,G4).

The model collection associated to the $[L_k B_k]$ Gaussian mixture shape is considered and variables are assumed to be ordered. After the estimation step, the function $D \mapsto -\gamma_n(\hat{s}_D)$ is plotted on the top of Figure 2. For $D \geq D_0 = 300$, we observe that the function $D \mapsto -\gamma_n(\hat{s}_D)$ has a linear behavior as expected (see Section 4.1). The residuals of the linear regression are plotted on the bottom of Figure 2. This defends the use of penalties proportional to the dimension since no trend can be observed in the residuals. The estimation of $\hat{C}$ leads to the penalty choice in criterion (5) and the selected model according to this penalized criterion is $\hat{K}_{\text{slope}} = 4$ and $\hat{\alpha}_{\text{slope}} = 7$.

This procedure is repeated 1000 times with new simulated dataset each time. The distribution used for simulations is the same as before except that the last ten variables have been removed in order to reduce the computation times. Consequently, the "true model" still corresponds to $(K_0, \alpha_0) = (4, 10)$ but the total number of variables is now $Q = 22$. These simulations allow to compare the behavior of the slope estimator with the oracle and with the behaviors of three other estimators given by AIC, BIC and ICL criteria. Since the true density is known, a Monte Carlo procedure gives the following oracle model estimation $K_{\text{oracle}} = 4$ and $\alpha_{\text{oracle}} = 9$. Note that even if the true density belongs to the density collection, the oracle model is not equal to the corresponding true model. The

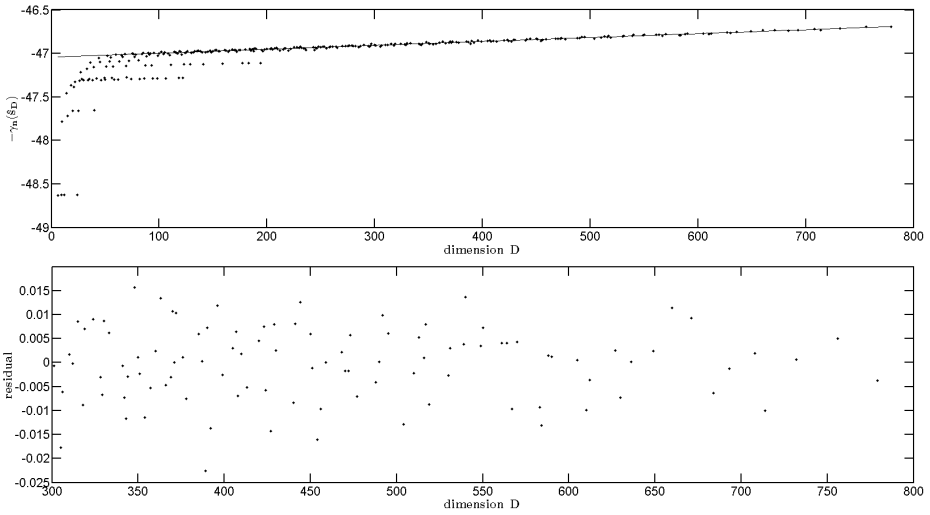

Figure 2: On the top graph, the function $D \mapsto -\gamma_n(\hat{s}_D)$ is plotted. The linear regression is made for $D \geq D_0 = 300$. The associated residuals are drawn on the bottom graph.

results are summarized in Table 5.1. The two criteria BIC and ICL select a model with 4 components and most of the times with 6 relevant variables. It is shown in Keribin (2000) that a model selection procedure using BIC is consistent to find the number of components of a Gaussian mixture when the component densities are bounded. But as far as we know, there no consistency result for such a variable selection and clustering problem. The results of Table 5.1 show that the model selected by BIC is not the true model. In this context, even if BIC tries to find the true model, this could be only done for unrealistic large samples. The behavior of the ICL method is not surprising since the aim of this criterion is to provide a mixture model leading to a sensible partitioning of the data. The AIC method selects too many components and too many relevant variables since it underpenalizes models in the mixture context. From a clustering point of view, BIC, ICL and the slope method have similar performances. The interest of this first simulated example is to illustrate the different behaviors of criteria. As expected, the slope method selects a model close to the oracle model.

criterion	$\hat{K}$	$\hat{\alpha}$								
		5	6	7	8	9	10	11	12	≥ 13
ICL	4	22	792	184	2					
BIC	4	29	859	111	1					
AIC	≤ 5			2	21	26	11	3	2	
	6			7	42	62	26	9	2	3
	≥ 7			57	155	237	170	93	41	41
Slope Heuristics	4		43	417	456	58	1			
	5		8	13	13	4				

Table 1: For each criterion, number of times that a model (K, α) is selected among the 1000 simulations.

Waveform dataset The waveform dataset, available at the UCI repository Blake and Merz (1999), is composed of three groups based on a random convex combination of two of three wave functions sampled at the integers from 1 to 21, with noise added. A detailed description is available in Breiman et al. (1984). The dataset consists of 5000 observations described by 40 variables. The last nineteen are noisy variables, sampled from a $\mathcal{N}(0, 1)$ density. By construction, the first and the twenty-first variables have also the same distribution $\mathcal{N}(0, 1)$. Consequently they are both irrelevant for the clustering and thus there is 19 variables which are potentially relevant for clustering.

First, the data have been centered. Contrary to the previous example, the variables are not ordered. Ideally, the model collection should be based on all the possible relevant variable subsets $\mathbf{v}$. Nevertheless, the selection among this model family is impossible because of the large cardinal of this collection and the resulting computation times. To get round this problem, the variables are ordered by decreasing order of their variances. With this ordering, the last twenty-one variables are the variables sampled from the $\mathcal{N}(0, 1)$ density.

The model selection is proceeded with the two mixture shapes $[L_k B_k]$ and $[L_k C_k]$. The plots of $D \mapsto -\gamma_n(\hat{s}_D)$ for the estimation of $\hat{C}$ are given on the top of Figure 4 for these two collections of models. To compare the two model collections, both corresponding fittings of the dimension model surfaces on maximum loglikelihood surfaces are presented on the bottom of Figure 4. As expected, we observe that the fitting is dramatically better for the model collection associated to the $[L_k C_k]$ collection. Indeed the relevant variables are dependent by construction. The $[L_k B_k]$ model collection is not rich enough for this problem. This $[L_k C_k]$ collection leads to select a model with $\hat{K} = 3$ clusters and $\hat{\alpha} = 19$ relevant variables. Despite the simulated data do not follow a Gaussian mixture, the procedure provides a stable and sensible solution. As to other criteria, they all select $\hat{\alpha} = 19$ with respectively $\hat{K} = 2$, 3 and 10 for ICL, BIC and AIC.

It has been said before that a clustering based on an ill-chosen collection of significant variables can lead to a large clustering error rate. As a matter of fact, Table 2 shows that the three true clusters are found with an error rate of 14.3%. Figure 3 plots the error rate in function of the number α of significant variables. Each curve corresponds to a fixed number

of components in the mixture. Choosing a model with more significant variables deteriorates the clustering performance.

	cluster 1	cluster 2	cluster 3	total
group 1	1331	185	176	1692
group 2	95	99	1459	1653
group 3	65	1494	96	1655
total	1491	1778	1731	5000

Table 2: Contingency table for the clustering obtained with the slope heuristics.

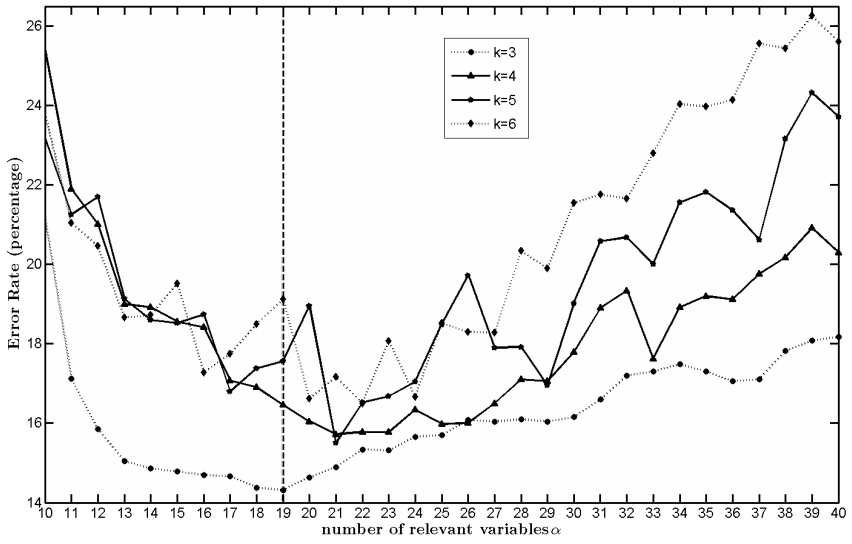

Figure 3: Evolution of the clustering error rate in function of the chosen number α of significant variables. Each curve corresponds to a fixed number K of mixture components.

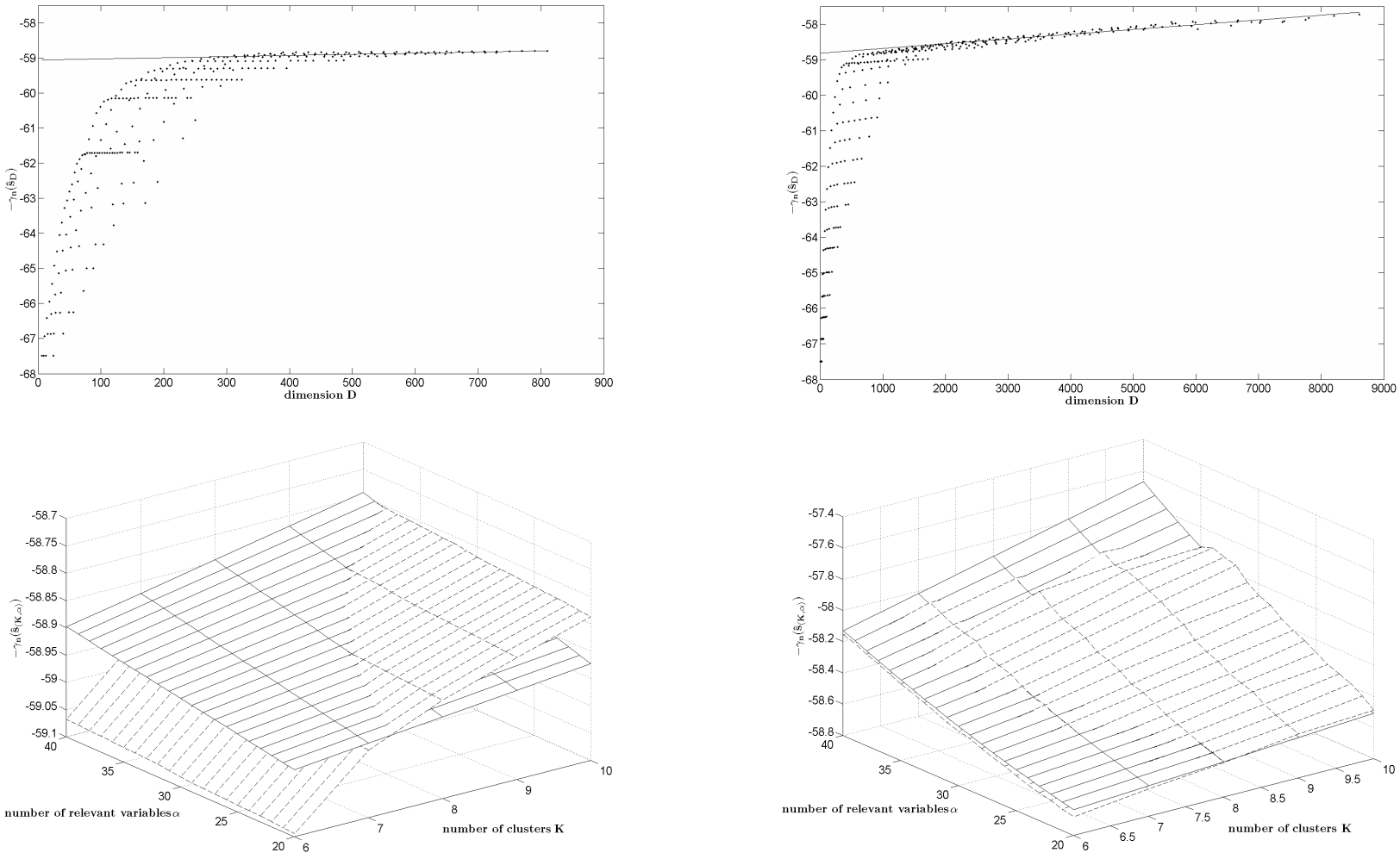

Figure 4: On the top, graphical representation of $D \mapsto -\gamma_n(\hat{s}_D)$ leading the estimation of $\hat{C}$ for the $[L_k B_k]$ collection (on the left) and the $[L_k C_k]$ collection (on the right). On the bottom, fitting of the dimension surface on $(K, \alpha) \in [|6, 10|] \times [|20, 40|] \mapsto -\gamma_n(\hat{s}_{(K,\alpha)})$ for the $[L_k B_k]$ collection (on the left) and the $[L_k C_k]$ collection (on the right).

5.2 Curve clustering

An oil field production profile is the curve of oil production versus time. In the following, the term reserves (or ultimate reserves) denotes the amount of oil that is produced during the exploitation of an oil field. The reader is referred to Babusiaux et al. (2007) for more details about the exploration and the production of oil. It is well known by the oil industry that production profiles of large fields have a different shape than production profiles of little fields. Indeed, little fields tend to produce their reserves in a short time and early pass the production peak. On the contrary, large fields slowly produce their reserves and their production presents a plateau during several years at the top level. Figure 5 illustrates this behaviour with productions of three fields of different size. In order to compare the production profile shapes, we consider production profiles normalized by the amount of reserves contained in each field. The study's aim is to validate that a clustering of normalized production profiles is consistent with the values of the reserves variable.

The database is composed of several hundred of oil production profiles corresponding to hydrocarbon layers in the North Sea[1]. The data used in the procedure are obtained from the original curves as follow. First, each production profile is normalized by the reserves of the corresponding field[2]. Ideally, it is desirable to proceed to the clustering on complete production profile. This is impossible since most of the fields are still in production nowadays. Figure 5 suggests that the beginning of the production curve is sufficient to distinguish different shapes in the curve family. Thus, we only consider the subsample composed of 180 fields which have started their production more than 64 months ago. Next, a discrete wavelet transform (DWT) is proceeded on each of these normalized curves. This decomposition has the advantage of giving information on each curve at different resolution levels. This transformation has already been used in curve classification (see for instance Berlinet et al., 2008). The reader is referred to Percival and Walden (2000) for details on the DWT. Let $\mathbf{W}_i$ be the wavelet coefficient vector of the i^{th} curve. Since the length of each curve is 64, the dimension of $\mathbf{W}_i$ is also 64. The vector $\mathbf{W}_i$ is defined by

$$\mathbf{W}_i = (\mathbf{V}'_{i6}, \mathbf{W}'_{i6}, \dots, \mathbf{W}'_{i1})'$$

where $\mathbf{W}_{ij}$ is a vector of length $64/2^j$ which is composed of all the wavelet coefficients corresponding to the scale j. The coefficient $\mathbf{V}_{i6}$ is equal to the mean of the curve i divided by $\sqrt{64}$. The hierarchical structure of the DWT suggests a natural order of the wavelet coefficient variables according their resolution. Indeed, $\mathbf{V}_{i6}$ and $\mathbf{W}_{i6}$ give informations about the general shape of the curve i whereas $\mathbf{W}_{i1}$ and $\mathbf{W}_{i2}$ give informations about details on it. We do not use the coefficients in $\mathbf{W}_{i1}$ and $\mathbf{W}_{i2}$ since they correspond to the finer resolution. We will see that the remaining coefficients are sufficient to propose a sensible clustering. Moreover, all the wavelet coefficient variables are centered and scaled to unit variance to

[1]Data is available on the website of the Norwegian Petroleum Directorate : `www.npd.no/engelsk/cwi/pbl/en/index.htm`, and the website of the English Department of Trade and Industry (DTI): `www.og.dti.gov.uk/fields/fields_index.htm`.

[2]The DTI does not provide estimations of reserves of their fields, consequently we use the IHS database `http://energy.ihs.com` for the english fields of the sample

make easier the fitting of the multidimensional Gaussian distribution $\mathcal{N}(0, \omega^2 I_{Q-\alpha})$ on the coefficient vectors which are not used for the clustering. These new coefficients are denoted $\widetilde{\mathbf{V}}_{i6}$ and $\widetilde{\mathbf{W}}_{ij}$ where $j \in \{3, \ldots, 6\}$. The procedure is performed on the sample $\mathbf{y}$ where $\mathbf{y}_i = \left(\widetilde{\mathbf{V}}'_{i6}, \widetilde{\mathbf{W}}'_{i6}, \ldots, \widetilde{\mathbf{W}}'_{i3}\right)'$ for an ordered model collection $[LB_k]$. This mixture collection allows to avoid estimation problems when the variances are too small.

Figure 6 clearly shows the expected linear behavior of $D \mapsto -\gamma_n(\hat{s}_D)$ in large dimensions. The selected model minimizing the penalized criterion deduced from the slope heuristics has $\hat{K} = 3$ components and $\hat{\alpha} = 20$ clustering variables. Finally, the MAP rule gives a clustering of the curves. The clusters contain 31, 140 and 9 curves respectively. Figure 7 displays the mean cluster of the normalized production profiles in each cluster. Boxplot of the logarithm of the reserves variable for each cluster are displayed in Figure 8. The second cluster mainly corresponds to the large fields whereas the first and the third clusters contain fields of medium size and small size respectively. The shape of normalized production profiles can be explained by the reserves variable. The reader is referred to Michel (2008) for more details.

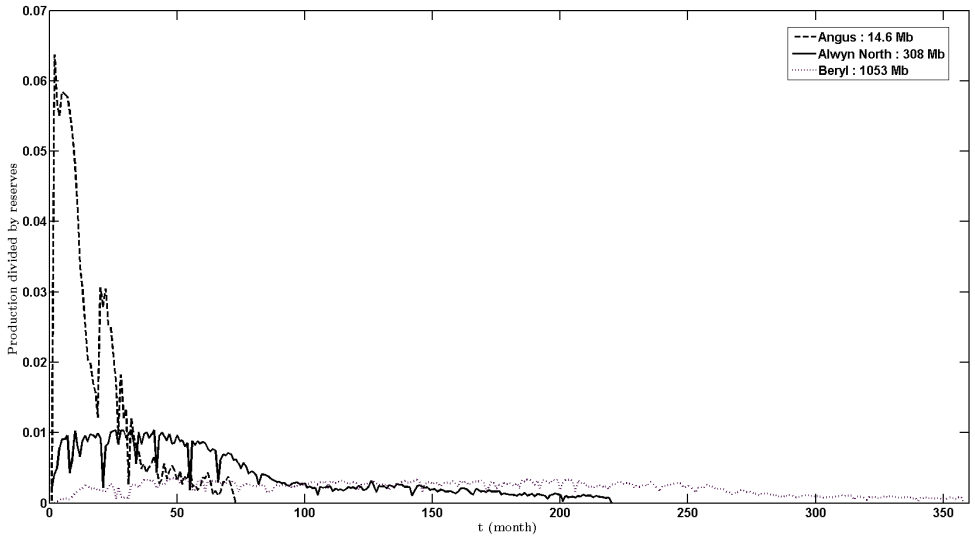

Figure 5: Oil production profiles normalized by reserves of three fields located in the North Sea.

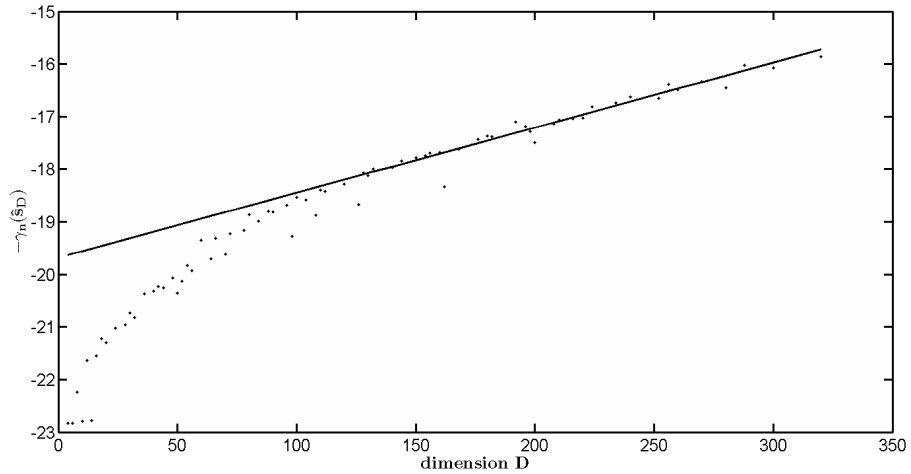

Figure 6: Slope method applied to the $[LB_k]$ collection for the wavelet coefficient curve data.

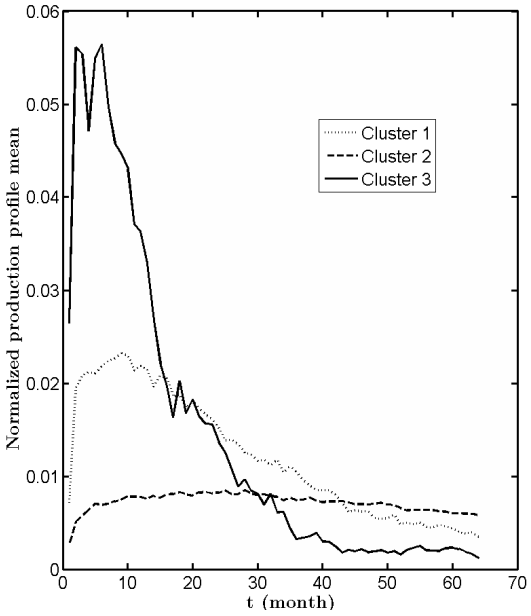

Figure 7: Normalized production profile means for each cluster.

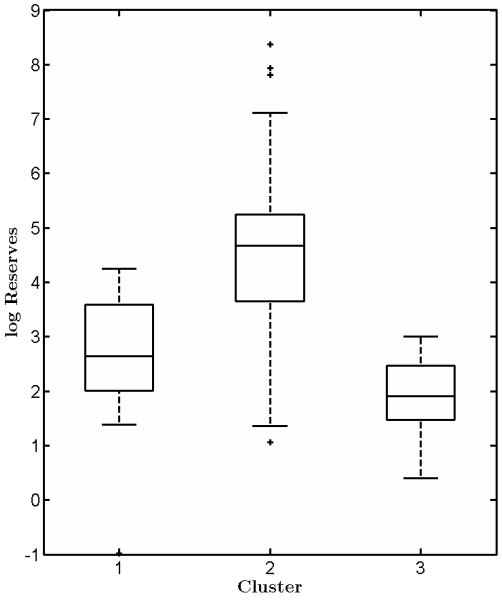

Figure 8: Boxplots of the logarithm of the reserves variable for each cluster.

5.3 Analysis of a transcriptome dataset

Currently, biologists are interested in gene functional analysis. It is usually considered that co-expressed genes are often implicated in the same biological function and consequently they are potential candidate to be co-regulated genes. Thus biologists try to extract groups of co-expressed genes according to transcriptome datasets in order to characterize more precisely their biological functions. Moreover an experiment selection for the clustering is desirable to improve the clustering and its interpretation with a biological point of view.

Here we study a transcriptome dataset of *Arabidopsis thaliana* extracted from the database CATdb (Gagnot et al., 2008). To build this database, an identical statistical analysis for all transcriptome experiments has been performed to remove the technical biases (normalization) and to determine the gene significantly differentially expressed (differential analysis) between two conditions. In this differential analysis, we test if a gene is non-differentially expressed or not in the experiment j. For this test, a test statistic corresponding to the normalized differential expression and a p-value adjusted by the Bonferroni method are determined. Then a gene is declared to be differentially expressed when its Bonferroni p-value is lower than 0.05. The reader is referred to Lurin (2004) for a description of such an analysis.

We focus on 305 genes of *Arabidopsis thaliana* studied on ten experiments which correspond to mutant conditions or different stress situations. These genes are declared differentially expressed in the two last experiments and non-differentially expressed in five experiments. Table 3 gives the number of differentially expressed genes per experiment. Each gene is described with a vector $\mathbf{y}_i \in \mathbb{R}^{10}$, the component y_{ij} corresponding to the test statistic calculated in the experiment j for the differential analysis.

experiment number	1	2	3	4	5	6	7	8	9	10
number of differentially expressed genes	0	0	207	0	219	118	0	0	305	305

Table 3: Number of differentially expressed genes per experiment.

Since there is not a natural way to order the variables, our procedure for non-ordered variables is performed with the $[LC]$ mixture collection. The maximum number of components is fixed to $K_{\max} = 40$. After the estimation step, we notice that the function $D \mapsto -\gamma_n(\hat{s}_D)$ has a linear behavior for $D \geq 220$ (see Figure 9), thus the slope method can be applied. The procedure selects a clustering with $\hat{K} = 8$ clusters based on the seven variables $\hat{\mathbf{v}} = \{3, 5, 6, 7, 8, 9, 10\}$. The eight clusters have different size (see Table 4) and the clustering shows some interesting similar behaviors of expression profiles (see Figure 10). A similar clustering can be found if all the variables are considered ($\alpha = 10$ fixed) but with the variable selection, the interpretation of the clustering is made clearer.

First, we note that the two benchmark experiments (9 and 10) where all genes are differentially expressed are selected. Moreover, the three variables which are not selected for the clustering are three variables where all genes are non-differentially expressed. The average behavior of genes per cluster is the same in the irrelevant experiments 1, 2 and 4 since it is concentrated around zero. On the contrary, genes of the cluster 2 have a particular

behavior in experiments 7 and 8. Their expression difference decreases between the two experiments (7 and 8) whereas the genes of the other clusters have the same expression in these two experiments (see Figure 10). This remark may explain why the two experiments 7 and 8 where all genes are non-differentially expressed are selected for the clustering while experiments 1, 2 and 4 are not. This clustering can help biologists to find gene biological functions. For instance, 12 genes for which biologists know any biological function are clustered with other genes for which biologists have some information.

cluster	1	2	3	4	5	6	7	8
number of genes	60	39	47	12	82	51	9	5

Table 4: Number of genes per cluster.

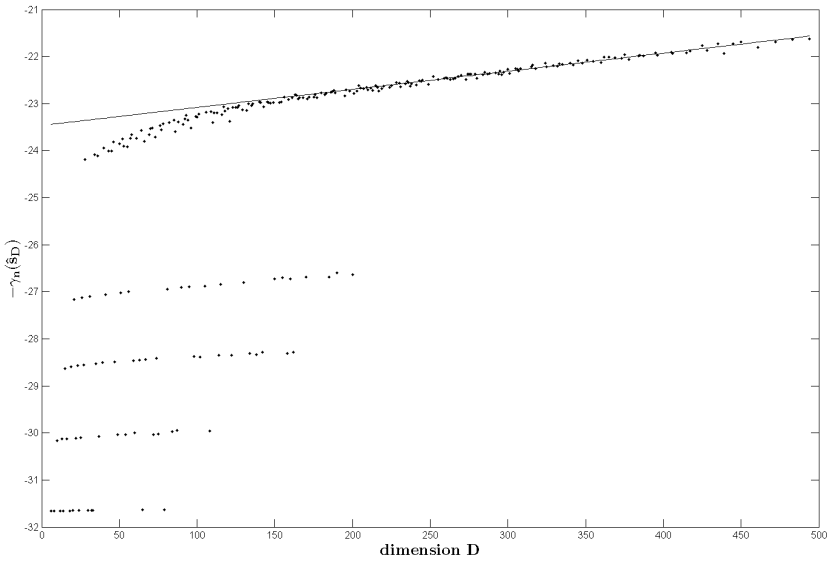

Figure 9: Penalty determination on the linear behavior of the function $D \mapsto -\gamma_n(\hat{s}_D)$.

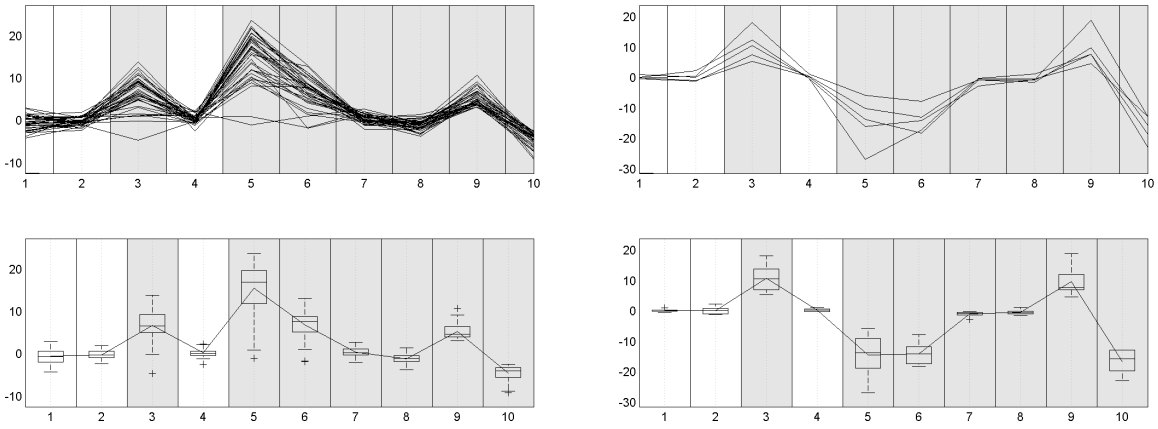

Figure 10: Graphical representation of genes profiles in clusters 2 (on the left) and 8 (on the right). Relevant experiments are colored in grey.

6 Discussion

In this paper, a methodology has been proposed to take into account the variable role in a clustering process in a model-based cluster analysis setting. The interest of our approach is to recast these two problems into a model selection problem where model collections are indexed by two quantities (K and $\mathbf{v}$). The practical use of the penalized likelihood criterion, proposed in Maugis and Michel (2008), is based on a slope heuristics method allowing to calibrate the multiplicative constant in the penalty term. The slope $\hat{C}$ is estimated on the restriction of $D \mapsto -\gamma_n(\hat{s}_D)$ to $D \geq D_0$, namely where this function has a linear behaviour. A robust regression is used allowing to mitigate the influence of possible estimation errors. The estimated slope $\hat{C}$ which depends on the user choice threshold D_0 can be confirmed by plotting the estimated slope in function of different values of the threshold. In a neighbourhood of $\hat{C}$, the slope has to be stable. The behaviour of this slope heuristics method has been studied on simulated and real datasets. It has been compared with standard criteria used in Gaussian mixture clustering context. BIC, ICL, AIC and our criterion have different goals thus it is natural to observe different behaviours in the examples. In particular, our criterion tends to select the oracle model.

Our method can be efficiently applied when the variables could be ordered as illustrated in the curve clustering study in Section 5.2. When the variables cannot be ordered in a natural way, our method is difficult to use when the number of variables is too large. An exhaustive research of the best model becomes untractable. A possible way to circumvent this problem is to find a convenient strategy allowing to run the estimation step only for a model subset with a reasonable size.

In all cases, the user has to check that the linear behaviour of $D \mapsto -\gamma_n(\hat{s}_D)$ is observed and also that the dimension model surface fits the maximum likelihood surface on high dimensions (see Figure 4). If it is not the case, several explanations can be given. First, the model dimension can be too low, namely the maximum number of components $K_{\max}$ has to be increased. The problem can be also related to the model family choice. Roughly speaking, the family model leads to a stabilization of the bias in large dimension only if the family model efficiently approaches the true density. Otherwise, the collection of models has to be changed in order to obtain a better fitting between the two surfaces.

Our theoretical and practical results could be extended for most of the twenty-eight Gaussian mixture shapes proposed by Celeux and Govaert (1995). In particular, without variable selection ($\alpha = Q$), our method allows us to select the number of clusters which is the fundamental problem in model-based clustering. It could be envisaged to adapt our works to select also the Gaussian mixture type.

Acknowledgements

The authors are grateful to Gilles Celeux (INRIA), Pascal Massart (Université Paris-Sud 11) and Marie-Laure Martin-Magniette (INRA) for helpful discussions and valuable comments.

Appendices

In this paper, the two model collections $[LB_k]$ and $[LC]$ used for the real dataset examples in Section 5 are not entering the theoretical study of Maugis and Michel (2008). Nevertheless it can be proved that the penalty functions have the same form as in Theorem 1. We sketch the proof in this appendix. According to the work of Maugis and Michel (2008), it is only required to determine an upper bound of the bracketing entropy of the collection $\mathcal{S}_{(K,\mathbf{v})}$ for the Hellinger distance. To determine this upper bound, we cannot use Proposition 1 of Appendix A in Maugis and Michel (2008) which allows to recast the problem as the study of the bracketing entropy of the associated mixture density family. Indeed, in these two collections, the variance matrices of mixtures have the same volume or are identical. Here we use the following result which can be proved along the lines of the proof of Theorem 2 in Genovese and Wasserman (2000).

The notion of bracketing number, denoted $\mathcal{N}_{[.]}$, for a family of nonnegative integrable functions $\mathcal{F}_{(K,\alpha)} = \{f = (f_1, \ldots, f_K); \forall k,\ f_k : \mathbb{R}^\alpha \to \mathbb{R}\}$ is now specified. Consider two functions $L(x) = (l_1(x), \ldots, l_K(x))$ and $U(x) = (u_1(x), \ldots, u_K(x))$ from $\mathbb{R}^\alpha$ to $\mathbb{R}^K$ such that $l_k \leq u_k$ for all $k \in \{1, \ldots, K\}$. The set $[L, U]$ is composed of all the functions $f = (f_1, \ldots, f_K)$ from $\mathbb{R}^\alpha$ to $\mathbb{R}^K$ such that $l_k \leq f_k \leq u_k$ for all k. This set $[L, U]$ is called an ε-bracket for the Hellinger distance if $d_H(l_k, u_k) \leq \varepsilon$ for all $k \in \{1, \ldots, K\}$. Then the smallest number of such ε-brackets $[L, U]$ requiring to cover $\mathcal{F}_{(K,\alpha)}$ is called the bracketing number.

If $\mathcal{F}_{(K,\alpha)}$ denotes the set of K-uples of Gaussian densities used in the mixtures of $\mathcal{L}_{(K,\alpha)}$, Proposition 1 allows to give an upper bound of the bracketing number of $\mathcal{L}_{(K,\alpha)}$ from an upper bound of the bracketing entropy of the $\mathcal{F}_{(K,\alpha)}$ for the Hellinger distance.

Proposition 1. *Let* $\varepsilon \in (0, 1]$,

$$\mathcal{N}_{[.]}(\varepsilon, \mathcal{L}_{(K,\alpha)}, d_H) \leq \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{T}_{K-1}, d_H\right) \mathcal{N}_{[.]}\left(\frac{\varepsilon}{3}, \mathcal{F}_{(K,\alpha)}, d_H\right)$$

where $\mathcal{T}_{K-1} = \{(p_1, \ldots, p_K) \in [0, 1]^K;\ \sum_{k=1}^K p_k = 1\}$ *is the K-1 dimensional simplex.*

A Bracketing entropy of the model collection $[LB_k]$

For $[LB_k]$ mixture shape, the variance matrices of a mixture have the same volume and are all diagonal. These variance matrices can be decomposed into $\Sigma_k = \beta B_k$ where $\beta = |\Sigma_k|^{\frac{1}{\alpha}}$ is the common volume and B_k is a diagonal matrix with determinant 1. Thus, the family $\mathcal{F}_{(K,\alpha)}$ is

$$\mathcal{F}_{(K,\alpha)} = \left\{ (\Phi(.|\mu_1, \beta B_1), \ldots, \Phi(.|\mu_K, \beta B_K)) \ ; \ \begin{array}{l} \forall k \in \{1, \ldots, K\},\ \mu_k \in [-a, a]^\alpha \\ B_k \in \Delta^1_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}), \beta \in [\beta_{\mathrm{m}}, \beta_{\mathrm{M}}] \end{array} \right\}$$

where $\Delta^1_{(\alpha)}(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ is the set of $\alpha \times \alpha$ diagonal matrices with determinant 1 and which eigenvalues are in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$ where $0 < \lambda_{\mathrm{m}} < \lambda_{\mathrm{M}}$. According to Proposition 1, the

aim is to build a bracket family for $\mathcal{F}_{(K,\alpha)}$ and to determine its cardinal in order to obtain the following result:

Proposition 2. *For all* $\varepsilon \in (0,1]$, $\mathcal{N}_{[.]}\left(\varepsilon, \mathcal{F}_{(K,\alpha)}, d_H\right) \leq \left(A_2 \frac{\alpha}{\varepsilon}\right)^{K(2\alpha-1)+1}$ *where the constant* A_2 *only depends on* λ_m, λ_M, β_m, β_M *and* a *and* $K(2\alpha-1)+1$ *is the dimension of* $\mathcal{F}_{(K,\alpha)}$.

Proof. The proof for the unidimensional case ($\alpha = 1$) is already available in Genovese and Wasserman (2000). Let $\varepsilon \in (0,1]$ and assume $K \geq 2$ and $\alpha \geq 2$ fixed. Let $\delta = \varepsilon/(3\alpha)$. For $j \in \{2, \ldots, r\}$, we define

$$b_j^2 = (1+\delta)^{1-\frac{j}{2}}\ \lambda_{\mathrm{M}}$$

where $r = \left\lceil 2\, \frac{\ln\left\{\frac{\lambda_{\mathrm{M}}}{\lambda_{\mathrm{m}}}(1+\delta)\right\}}{\ln(1+\delta)} \right\rceil$ in order to have $b_r^2 \leq \lambda_{\mathrm{m}} \leq b_2^2 = \lambda_{\mathrm{M}}$. $\lceil h \rceil$ denotes the smallest integer greater than or equal to h. For $z \in \{0, \ldots, r'\}$ we consider

$$\beta_z = (1+\delta)^{-z}\beta_{\mathrm{M}}$$

where $r' = \left\lceil \frac{\ln\left\{\frac{\beta_{\mathrm{M}}}{\beta_{\mathrm{m}}}\right\}}{\ln(1+\delta)} \right\rceil$ in order to have $\beta_{r'} \leq \beta_{\mathrm{m}} \leq \beta_0 = \beta_{\mathrm{M}}$.

For a vector $J = (j(1), \ldots, j(\alpha-1)) \in \{2, \ldots, r\}^{\alpha-1}$, the diagonal matrices B_J^l and B_J^u are defined by

$$B_J^l = \mathrm{diag}\left(b_{j(1)+1}^2, \ldots, b_{j(\alpha-1)+1}^2, \lambda_{\mathrm{M}}^{1-\alpha}(1+\delta)^{\frac{S_J}{2}-(\alpha-1)}\right)$$

and

$$B_J^u = \mathrm{diag}\left(b_{j(1)}^2, \ldots, b_{j(\alpha-1)}^2, \lambda_{\mathrm{M}}^{1-\alpha}(1+\delta)^{\frac{S_J}{2}-\frac{\alpha-1}{2}}\right)$$

with $S_J = \sum_{q=1}^{\alpha-1} j(q)$. The q^{th} diagonal coefficients of these matrices are denoted $B_{J,q}^l$ and $B_{J,q}^u$ respectively.

First, a function $\Phi(.|\mu, \beta B)$ such that $\beta \in [\beta_m, \beta_{\mathrm{M}}]$, $\mu \in [-a,a]^\alpha$ and $B \in \Delta_{(\alpha)}^1(\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}})$ is considered. Let $z \in \{0, \ldots, r'\}$ be the unique integer of $\{0, \ldots, r'\}$ such that $\beta_{z+1} < \beta \leq \beta_z$ and let J be the unique vector of $\{2, \ldots, r\}^{\alpha-1}$ such that $\forall q \in \{1, \ldots, \alpha-1\}$, $B_{J,q}^l \leq B_{qq} \leq B_{J,q}^u$. Hence for all $q \in \{1, \ldots, \alpha\}$,

$$\beta_{z+1}\, B_{J,q}^l \leq \beta\, \Sigma_{qq} \leq \beta_z\, B_{J,q}^u.$$

For a couple (z, J), we also consider a regular lattice of mean vector $\nu^{(z,J)} = \left(\nu_1^{(z,J)}, \ldots, \nu_\alpha^{(z,J)}\right) \in [-a,a]^\alpha$ such that for all $q \in \{1, \ldots, \alpha-1\}$,

$$\nu_q^{(z,J)} = (1+\delta)^{-\frac{j(q)+1}{4}-\frac{z}{2}} \sqrt{\lambda_{\mathrm{M}}\ \beta_{\mathrm{M}}\ c_1}\, \delta s_q,$$

with $s_q \in \{-N_q, \ldots, N_q\}$ where $N_q = \left\lfloor \frac{a(1+\delta)^{\frac{j(q)+1}{4}+\frac{z}{2}}}{\sqrt{\lambda_{\mathrm{M}}\ \beta_{\mathrm{M}}\ c_1}\, \delta} \right\rfloor$,

$$\nu_\alpha^{(z,J)} = (1+\delta)^{\frac{S_J}{4}-\frac{\alpha+z}{2}} \sqrt{\lambda_{\mathrm{M}}^{1-\alpha}\, \beta_{\mathrm{M}}\ c_1}\, \delta s_\alpha,$$

with $s_\alpha \in \{-N_\alpha, \ldots, N_\alpha\}$ where $N_\alpha = \left\lfloor \frac{a(1+\delta)^{\frac{\alpha+z}{2}-\frac{S_J}{4}}}{\sqrt{\lambda_{\mathrm{M}}^{1-\alpha}\,\beta_{\mathrm{M}}\,c_1\delta}} \right\rfloor$ and $c_1 := 1 - 2^{-\frac{1}{4}}$. For a given couple (z, J), this insures that for all vectors $\mu \in [-a, a]^\alpha$, there exists a vector $\nu^{(z,J)}$ of this lattice such that

$$\frac{(1+\delta)^z}{\beta_{\mathrm{M}}\,\lambda_{\mathrm{M}}} \left\{ \sum_{q=1}^{\alpha-1} \left(\nu^{(z,J)} - \mu\right)_q^2 (1+\delta)^{\frac{j(q)+1}{2}} + \left(\nu^{(z,J)} - \mu\right)_\alpha^2 (1+\delta)^{-\frac{S_J}{2}+\alpha} \lambda_{\mathrm{M}}^\alpha \right\} \leq c_1\,\alpha\,\delta^2. \quad (7)$$

For a couple (z, J) and a vector $\nu^{(z,J)}$ defined as before, the two following associated functions are considered

$$\begin{cases} l(x) = (1+\delta)^{-2\alpha}\,\Phi\left(x|\nu^{(z,J)}, (1+\delta)^{-\frac{1}{4}}\,\beta_{z+1}\,B_J^l\right) \\ u(x) = (1+\delta)^{2\alpha}\,\Phi\left(x|\nu^{(z,J)}, (1+\delta)\,\beta_z\,B_J^u\right). \end{cases} \quad (8)$$

Second, we check that for all $x \in \mathbb{R}^Q$, $l(x) \leq \Phi(x|\mu, \beta B) \leq u(x)$. According to Proposition 6 of Appendix C in Maugis and Michel (2008) which allows to upper bound the ratio of two Gaussian densities, we get

$$\begin{aligned} \frac{\Phi(x)}{u(x)} &\leq (1+\delta)^{-2\alpha} \sqrt{\frac{|(1+\delta)\beta_z B_J^u|}{|\beta B|}}\, \exp\left[\frac{1}{2}(\nu^{(z,J)} - \mu)' \left\{(1+\delta)\beta_z B_J^u - \beta B\right\}^{-1} (\nu^{(z,J)} - \mu)\right] \\ &\leq (1+\delta)^{-\frac{3\alpha+1}{4}}\, \exp\left\{\frac{1}{2\delta}(\nu^{(z,J)} - \mu)'(\beta_z B_J^u)^{-1}(\nu^{(z,J)} - \mu)\right\} \end{aligned}$$

and

$$\begin{aligned} \frac{l(x)}{\Phi(x)} &\leq (1+\delta)^{-2\alpha} \sqrt{\frac{|\beta B|}{|(1+\delta)^{-\frac{1}{4}}\beta_{z+1}B_J^l|}}\, \exp\left[\frac{1}{2}(\nu^{(z,J)} - \mu)' \left\{\beta B - (1+\delta)^{-\frac{1}{4}}\beta_{z+1}B_J^l\right\}^{-1} (\nu^{(z,J)} - \mu)\right] \\ &\leq (1+\delta)^{-(\frac{9\alpha}{8}+\frac{1}{4})}\, \exp\left\{\frac{1}{2\delta(1-2^{-\frac{1}{4}})}(\nu^{(z,J)} - \mu)'(\beta_{z+1}B_J^l)^{-1}(\nu^{(z,J)} - \mu)\right\} \end{aligned}$$

using the concavity of the function $\delta \mapsto 1 - (1+\delta)^{-\frac{1}{4}}$. The following inequalities

$$(\nu^{(z,J)} - \mu)'\,(\beta_z B_J^u)^{-1}\,(\nu^{(z,J)} - \mu) \leq \frac{\delta^2}{4}(3\alpha + 1) \quad (9)$$

and

$$(\nu^{(z,J)} - \mu)'(\beta_{z+1}B_J^l)^{-1}(\nu^{(z,J)} - \mu) \leq \delta^2(1 - 2^{-\frac{1}{4}})\frac{9\alpha + 2}{8} \quad (10)$$

are then sufficient to have $l \leq \Phi \leq u$. We can check that condition (7) implies the two inequalities (9) and (10).

Third, we show that $d_H(u,l) \leq \varepsilon$. According to Proposition 7 of Appendix C in Maugis and Michel (2008), we have that the quantity $d_H^2(l,u)$ is equal to

$$
\begin{aligned}
& (1+\delta)^{-2\alpha} + (1+\delta)^{2\alpha} - 2^{\frac{\alpha}{2}+1} \left|(1+\delta)\beta_z B_J^u (1+\delta)^{-\frac{1}{4}} \beta_{z+1} B_J^l\right|^{-\frac{1}{4}} \left|\frac{(B_J^u)^{-1}}{(1+\delta)\beta_z} + \frac{(1+\delta)^{\frac{1}{4}}(B_J^l)^{-1}}{\beta_{z+1}}\right|^{-\frac{1}{2}} \\
= \;& (1+\delta)^{-2\alpha} + (1+\delta)^{2\alpha} - 2\left\{\frac{(1+\delta)^{\frac{11}{8}} + (1+\delta)^{-\frac{11}{8}}}{2}\right\}^{-\frac{\alpha-1}{2}} \left\{\frac{(1+\delta)^{\frac{2\alpha+7}{8}} + (1+\delta)^{-\frac{2\alpha+7}{8}}}{2}\right\}^{-\frac{1}{2}} \\
= \;& [2\cosh\{2\alpha\ln(1+\delta)\} - 2] + \left[2 - 2\left\{\cosh\left(\frac{11}{8}\ln(1+\delta)\right)\right\}^{-\frac{\alpha-1}{2}}\right] \\
& + 2\left\{\cosh\left(\frac{11}{8}\ln(1+\delta)\right)\right\}^{-\frac{\alpha-1}{2}} \left[1 - \left\{\cosh\left(\frac{7+2\alpha}{8}\ln(1+\delta)\right)\right\}^{-\frac{1}{2}}\right] \\
\leq \;& 4\sinh(1)\alpha^2\delta^2 + 2\frac{\alpha-1}{2}\frac{11}{8}\delta^2 + 2\frac{7+2\alpha}{8}\frac{1}{2}\delta^2 \leq 9\alpha^2\delta^2 = \varepsilon^2.
\end{aligned}
$$

Finally, we can construct an ε-bracket family (with respect to d_H) to cover $\mathcal{F}_{(K,\alpha)}$. Let $(\Phi(.|\mu_1, \beta B_1), \ldots, \Phi(.|\mu_K, \beta B_K))$ be an element of $\mathcal{F}_{(K,\alpha)}$. Let $z \in \{0, \ldots, r'\}$ and $J_1, \ldots, J_K$ in $\{2, \ldots, r\}^{\alpha-1}$ such that for all $k \in \{1, \ldots, K\}$ and all $q \in \{1, \ldots, \alpha\}$,

$$\beta_{z+1}\, B^l_{J_k,q} \leq \beta\, B_{k,qq} \leq \beta_z\, B^u_{J_k,q}.$$

For all k, there exists a vector $\nu^{(z,J_k)}$ such that condition (7) is satisfied for the mean vector μ_k. For z, J_k and $\nu^{(zJ_k)}$, the two associated functions defined by (8) are denoted u_k and l_k. Then we define $L := (l_1, \ldots, l_K)$ and $U := (u_1, \ldots, u_K)$. The set of all such brackets $[L,U]$ covers the family $\mathcal{F}_{(K,\alpha)}$ and is denoted $\mathcal{R}(\varepsilon, K, \alpha)$. An upper bound of the bracketing entropy of $\mathcal{F}_{(K,\alpha)}$ is thus determined by the computation of the cardinal of $\mathcal{R}(\varepsilon, K, \alpha)$. If $2a \geq \sqrt{\beta_{\mathrm{M}}\, c_1}(\sqrt{\lambda_{\mathrm{M}}} \vee \sqrt{\lambda_{\mathrm{M}}^{1-\alpha}})$ then

$$
\begin{aligned}
\mathcal{N}_{[.]}(\varepsilon, \mathcal{F}_{(K,\alpha)}, d_H) &\leq \text{card } \mathcal{R}(K, \varepsilon, \alpha) \\
&\leq \sum_{z=0}^{r'} \left[\sum_J \left\{\prod_{q=1}^{\alpha-1} \frac{2\,a\,(1+\delta)^{\frac{j(q)+1}{4}+\frac{z}{2}}}{\sqrt{\beta_M\,\lambda_{\mathrm{M}}\,c_1}\,\delta}\right\} \left\{\frac{2\,a\,(1+\delta)^{\frac{\alpha+z}{2}-\frac{S_J}{4}}}{\sqrt{\beta_{\mathrm{M}}\,c_1\,\lambda_{\mathrm{M}}^{1-\alpha}}\,\delta}\right\}\right]^K \\
&\leq \sum_{z=0}^{r'} \left[(r-1)^{\alpha-1}\left(\frac{2a}{\sqrt{\beta_{\mathrm{M}}\,c_1}\delta}\right)^{\alpha}(1+\delta)^{\frac{3\alpha-1}{4}+\frac{z}{2}}\right]^K \qquad (11) \\
&\leq (r-1)^{K(\alpha-1)}\left(\frac{2a}{\sqrt{\beta_{\mathrm{M}}\,c_1}\delta}\right)^{K\alpha}(1+\delta)^{K(\frac{3\alpha-1}{4}+\frac{z}{2})}(1+\delta)^{\frac{K\alpha r'}{2}}(r'+1) \\
&\leq \left(\frac{A_2\alpha}{\varepsilon}\right)^{K\alpha+K(\alpha-1)+1} \qquad (12)
\end{aligned}
$$

using $1+\delta \leq 2$ and the definition of r, r' and δ. The constant A_2 depends on parameters a, β_{m}, β_{M}, λ_{M} and λ_{m}. If $2a \leq \sqrt{\beta_{\mathrm{M}}\, c_1}(\sqrt{\lambda_{\mathrm{M}}} \vee \sqrt{{\lambda_{\mathrm{M}}}^{1-\alpha}})$, the upper bound (12) is again satisfied even if the constant A_2 is modified. □

B Bracketing entropy of the model collection $[LC]$

For $[LC]$ mixture shape, the variance matrices in mixture are all equal to the same positive definite matrix. Thus, we need to find an upper bound of the number bracketing of the following set

$$\mathcal{F}_{(K,\alpha)} = \{(\Phi(.|\mu_1,\Sigma),\ldots,\Phi(.|\mu_K,\Sigma);\ \mu_k \in [-a,a]^\alpha,\ \Sigma \in \mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}})\}$$

where $\mathcal{D}^+_{(\alpha)}(\lambda_{\mathrm{m}},\lambda_{\mathrm{M}})$ is the set of $\alpha\times\alpha$ symmetric positive definite matrices whose eigenvalues are in the interval $[\lambda_{\mathrm{m}},\lambda_{\mathrm{M}}]$.

Proposition 3. *For all $\varepsilon \in (0,1]$, $\mathcal{N}_{[.]}\left(\mathcal{F}_{(K,\alpha)}\right) \leq \left(A_3 \frac{\alpha}{\varepsilon}\right)^{K\alpha+\frac{\alpha(\alpha+1)}{2}}$ where the constant A_3 only depends on λ_m, λ_M, and a, and $K\alpha+\frac{\alpha(\alpha+1)}{2}$ is the dimension of $\mathcal{F}_{(K,\alpha)}$.*

This result is obtained by considering the following bracket family. Its construction is inspired by the bracket family used in the study of the bracketing entropy of $\mathcal{L}_{(K,\alpha)}$ for the model $[L_kC_k]$ stated in Maugis and Michel (2008).

Proposition 4. *For all $\varepsilon \in (0,1]$, let $\delta = \frac{\varepsilon}{\sqrt{3}\,\alpha}$ and $\beta = \frac{\lambda_m\,\varepsilon}{3\sqrt{3}\alpha^2}$. The following set*

$$\left\{([l_1,u_1],\ldots,[l_K,u_K]);\ \begin{array}{l} u_k(x) = (1+2\delta)^\alpha\, \Phi\left(x|\,\nu_k,(1+\delta)A\right) \\ l_k(x) = (1+2\delta)^{-\alpha}\, \Phi\left(x|\,\nu_k,(1+\delta)^{-1}A\right) \end{array}\ ;\ A \in \mathcal{R}(\beta),\ \nu_k \in \mathcal{X}(\varepsilon,a,\lambda_m,\alpha)\right\}$$

where

$$\mathcal{R}(\beta) = \left\{A = (A_{ij})_{1\leq i,j\leq\alpha};\ A_{ij} = a_{ij}\beta;\ a_{ij} = a_{ji} \in \mathbb{Z} \cap \left[-\left\lfloor\frac{\lambda_M}{\beta}\right\rfloor, \left\lfloor\frac{\lambda_M}{\beta}\right\rfloor\right]\right\}$$

and

$$\mathcal{X}(\varepsilon,a,\lambda_m,\alpha) = \left\{\nu = (\nu_1,\ldots,\nu_\alpha);\ \nu_q = \frac{\sqrt{\lambda_m}\,\varepsilon}{3\alpha}\, s_q;\ s_q \in \mathbb{Z} \cap \left[-\left\lfloor\frac{3\,a\,\alpha}{\sqrt{\lambda_m}\,\varepsilon}\right\rfloor, \left\lfloor\frac{3\,a\,\alpha}{\sqrt{\lambda_m}\,\varepsilon}\right\rfloor\right]\right\},$$

is an ε-bracket set over $\mathcal{F}_{(K,\alpha)}$.

Finally the bracketing number of $\mathcal{F}_{(K,\alpha)}$ is upper bounded by

$$\begin{aligned}
\mathcal{N}_{[.]}(\varepsilon, \mathcal{F}_{(K,\alpha)}, d_H) &\leq \operatorname{card}(\mathcal{R}(\beta)) \times \operatorname{card}(\mathcal{X}(\varepsilon, a, \lambda_{\mathrm{m}}, \alpha))^K \\
&\leq \left(\frac{2\,\lambda_{\mathrm{M}}}{\beta}\right)^{\frac{\alpha(\alpha+1)}{2}} \times \left(\frac{6a\alpha}{\sqrt{\lambda_{\mathrm{m}}}\varepsilon}\right)^{K\alpha} \\
&\leq \left(\frac{6\sqrt{3}\,\lambda_{\mathrm{M}}\,\alpha^2}{\varepsilon\,\lambda_{\mathrm{m}}}\right)^{\frac{\alpha(\alpha+1)}{2}} \times \left(\frac{6a\alpha}{\sqrt{\lambda_{\mathrm{m}}}\varepsilon}\right)^{K\alpha} \\
&\leq \left(A_3\frac{\alpha}{\varepsilon}\right)^{K\alpha+\frac{\alpha(\alpha+1)}{2}}
\end{aligned}$$

where the constant A_3 only depends on λ_{m}, λ_{M}, and a.

References

Abraham, C., Cornillon, P. A., Matzner-Løber, E., and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics. Theory and Applications*, 30:581–595.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723. System identification and time-series analysis.

Arlot, S. and Massart, P. (2008). Slope heuristics for heteroscedastic regression on a random design. Submitted to the Annals of Statistics.

Babusiaux, D., Barreau, S., and Bauquis, P.-R. (2007). *Oil and gas exploration and production, reserves, costs, contracts.* Technip, Paris.

Berlinet, A., Biau, G., and Rouvière, L. (2008). Functional classification with wavelets. Technical report. To appear in Annales de l'ISUP.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:719–725.

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

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73.

Blake, C., K.-E. and Merz, C. (1999). Uci repository of machine learning databases. `http://mlearn.ics.uci.edu/MLSummary.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*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA.

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

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

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.

García-Escudero, L. A. and Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification*, 22:185–201.

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28:1105–1127.

Huber, P. (1981). *Robust Statistics*. Wiley.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98:397–408.

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.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, 62:49–66.

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

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85:717–736.

Lurin, C. *et al.* (2004). Genome-wide analysis of arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell*, 16:2089–103.

Ma, P., Castillo-Davis, C.and Zhong, W., and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research*, 34:1261–1269.

Mallows, C. (1973). Some comments on *Cp*. *Technometrics*, 37:362–372.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

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

Maugis, C. and Michel, B. (2008). A penalized criterion for Gaussian mixture model selection. Technical Report 6549, INRIA.

Michel, B. (2008). *Modélisation de la production d'hydrocarbures dans un bassin pétrolier*. PhD thesis, Université Paris-Sud 11.

Percival, B. and Walden, A. T. (2000). *Wavelet methods for time series analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge university press, New York.

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*. Springer Verlag.

Tarpey, T. and Kinateder, K. K. J. (2003). Clustering functional data. *Journal of Classification*, 20:93–114.

Verzelen, N. (2008). Adaptive estimation of regular Gaussian Markov random fields. In preparation.

Villers, F. (2007). *Tests et sélection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, Université Paris-Sud 11.

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