ESAIM: PS 15 (2011) 320–339
DOI: 10.1051/ps/2010002

ESAIM: Probability and Statistics
www.esaim-ps.org

DATA-DRIVEN PENALTY CALIBRATION: A CASE STUDY FOR GAUSSIAN MIXTURE MODEL SELECTION

Cathy Maugis[1] and Bertrand Michel[2]

Abstract. In the companion paper [C. Maugis and B. Michel, A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: P&S* **15** (2011) 41–68], a penalized likelihood criterion is proposed to select a Gaussian mixture model among a specific model collection. This criterion depends on unknown constants which have to be calibrated in practical situations. A "slope heuristics" method is described and experimented to deal with this practical problem. In a model-based clustering context, the specific form of the considered Gaussian mixtures allows us to detect the noisy variables in order to improve the data clustering and its interpretation. The behavior of our data-driven criterion is highlighted on simulated datasets, a curve clustering example and a genomics application.

Mathematics Subject Classification. 62G07, 62H30.

Received September 19, 2008. Revised July 17, 2009 and January 22, 2010.

Introduction

The principle of selecting a model by penalizing loglikelihood or least squares criteria has emerged during the seventies. Akaike[2] proposed the AIC criterion (Akaike's information criterion) and Schwarz [37] suggested the BIC (Bayesian Information Criterion). A non asymptotic approach for model selection via penalization has emerged, mainly with works of Birgé and Massart [13] and Barron *et al.* [8] (an overview is available in [30]). The aim of this approach is to determine data-driven penalized criteria and associated oracle inequalities. Nevertheless, in many situations the penalties involved are only known up to a multiplicative constant. In such a situation, Birgé and Massart [14] propose a so-called "slope heuristics" method to estimate this constant. This heuristics consists of assuming that twice the minimal penalty is almost the optimal penalty. Theoretically, this rule of thumb is justified by Birgé and Massart [14] in the framework of Gaussian regression in a homoscedastic fixed design and generalized by Arlot and Massart [5] in the heteroscedastic random design case for histograms. The slope heuristics has been the subject of several practical studies. For example, it has been successfully applied for multiple change point detection [25], clustering [9], estimation of oil reserves [26] and genomics [40]. In the companion paper [33], a non asymptotic penalized criterion is proposed to solve a problem of Gaussian mixture model selection. Since the penalty of this criterion depends on an unknown multiplicative constant, the

Keywords and phrases. Slope heuristics, Penalized likelihood criterion, Model-based clustering, noisy variable detection.

[1] Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, 135 avenue de Rangueil, 31077 Toulouse Cedex 4, France. cathy.maugis@insa-toulouse.fr

[2] Laboratoire de Statistique Théorique et Appliquée, Université Paris 6, 175 rue du Chevaleret, 75013 Paris, France. bertrand.michel@upmc.fr

Article published by EDP Sciences

aim of this paper is to carry out the slope heuristics method for calibrating it, without using the "dimension jump" method as in the previous cited works.

Cluster analysis is more and more concerned with large datasets where observations are described by many variables. This large number of variables could be beneficial but in many situations, the presence of noisy variables can be harmful to detect a reasonable clustering structure. Thus, the variables are partitioned into two categories. The subset $\mathbf{v}^c$ contains the noisy variables, said irrelevant in the sequel. The distribution of such a noisy variable is homogeneous and centered around its mean, allowing not to distinguish a possible clustering of the data. The complementary set $\mathbf{v}$ is composed of the clustering variables. In the companion paper [33], we recast our variable partition problem for Gaussian mixture clustering into a model selection problem. For this, a specific collection of Gaussian mixture models characterized by the number of mixture components K and the clustering variable subset $\mathbf{v}$ is considered. A general model selection theorem for maximum likelihood estimation proposed by Massart [30] is applied to construct the penalized criterion. This construction requires the control of the bracketing entropy of Gaussian mixture families. Our theoretical results show that the penalty has to be chosen proportional to the dimension. In spite of the fact that this heuristics is not proved in our Gaussian mixture clustering context where the models are described by two entities K and $\mathbf{v}$, we show that it can be successfully used to calibrate the penalty function. This new data-driven penalized criterion is an alternative to the asymptotic criteria AIC [2], BIC [37] and ICL [11] generally used in such a framework.

The model selection method based on our calibrated penalized criterion behaves well on real datasets. First, our methodology is applied to an oil production curve clustering problem. Curve clustering deals with the problem of identifying homogeneous groups in a functional dataset. This situation occurs in many areas of sciences, for instance in genetics, neuroscience, economics and engineering. Many methods of 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 [1] and [20]. Another approach proposed by Tarpey and Kinateder [39] is to adapt the k-means algorithm for functional spaces. With a different point of view, Ma *et al.* [28] use mixture models on B-splines coefficients, as in the works of James and Sugar [22] for sparsely sampled functional data. In most of cited works, each curve is described with a coefficient vector. In practice, the number of these coefficients can be of the same order as the number of curves in the sample, a non asymptotic model selection approach is thus relevant in such a situation.

Next, our method is applied to cluster a transcriptome dataset. After sequencing the genome of some species, biologists are now interested in determining biological gene functions. In this aim, clustering methods such as hierarchical clustering , k-means algorithm or model-based clustering are commonly applied to find clusters of co-expressed genes (see for instance [23,38] and references therein). But when a gene subset is studied, all the genes can be non differentially expressed in some experiments. Thus the detection of such experiments by our procedure leads to improve the clustering and its interpretation.

The paper is organized as follows: Section 1 presents the collections of Gaussian mixture model used in this paper. In Section 2, 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 Gaussian mixture model context, and we recall the results obtained with this criterion in [33]. Section 3 is devoted to the description of the slope heuristics and its practical use in this specific context. Simulations and applications to curve clustering problem and a genomics clustering are presented in Section 4 to highlight the interest of this method. Finally, a discussion ends the paper in Section 5.

1. Gaussian mixture models

1.1. 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 s is estimated by a finite mixture model in a clustering purpose although s is not assumed to be a Gaussian mixture density itself. Model-based clustering consists of assuming that the data come from a source with several subpopulations, modelled separately and the resulting model is a finite mixture model. When the data are multivariate continuous observations, the parameterized component density is usually a multidimensional Gaussian density. A Gaussian mixture density with K components is

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

where the p_k's are the mixing proportions ($\forall k = 1, \dots, K,\ 0 < p_k < 1$ and $\sum_{k=1}^{K} p_k = 1$) and $\Phi(.|\eta_k, \Lambda_k)$ denotes the Q-dimensional Gaussian density with mean η_k and variance matrix Λ_k. The parameter vector is $(p_1, \dots, p_K, \eta_1, \dots, \eta_K, \Lambda_1, \dots, \Lambda_K)$. The mixture model is an incomplete data structure model: the complete data are $((\mathbf{y}_1, \mathbf{z}_1), \dots, (\mathbf{y}_n, \mathbf{z}_n))$ where the missing data are $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ such that $z_{ik} = 1$ if and only if $\mathbf{y}_i$ arises from the component k. The vector $\mathbf{z}$ defines an ideal clustering of the data $\mathbf{y}$ associated to the mixture model. Denoting $(\hat{p}_1, \dots, \hat{p}_K, \hat{\eta}_1, \dots, \hat{\eta}_K, \hat{\Lambda}_1, \dots, \hat{\Lambda}_K)$ the maximum likelihood estimate of the vector parameter, derived for instance from the EM algorithm [18], a data clustering is deduced from the maximum a posteriori principle:

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

1.2. Definitions of the considered Gaussian mixture models

Gaussian mixtures with a specific form are considered in order to detect the noisy variables in the clustering process. Irrelevant variables are assumed to have a homogeneous behavior around a null mean. Hence the data density is modelled with a spherical Gaussian joint distribution with zero mean vector. On the contrary, the different component mean vectors are free on clustering variables. Moreover, the variance matrices restricted on clustering variables are either taken completely free or are chosen in a specified set of positive definite matrices.

The model is as follows. Let $\mathcal{V}$ be the collection of nonempty subsets of $\{1, \dots, Q\}$. A Gaussian mixture family is characterized by the number of components $K \in \mathbb{N}^*$ and the 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 $\mathcal{M} = \mathbb{N}^* \times \mathcal{V}$. Consider the decomposition of a vector $x \in \mathbb{R}^Q$ into its restriction on relevant variables $x_{[\mathbf{v}]} = (x_{j_1}, \dots, x_{j_\alpha})'$ and its restriction on irrelevant variables $x_{[\mathbf{v}^c]} = (x_{l_1}, \dots, x_{l_{Q-\alpha}})'$ where $\mathbf{v} = \{j_1, \dots, j_\alpha\}$ and $\mathbf{v}^c = \{l_1, \dots, l_{Q-\alpha}\} = \{1, \dots, Q\} \backslash \mathbf{v}$. On clustering variables, a Gaussian mixture f is chosen among the following mixture family

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

where $a \in \mathbb{R}^*_+$ and $\mathcal{D}^+_{(K,\alpha)}$ denotes a family of K-tuples of $\alpha \times \alpha$ symmetric positive definite matrices which is related to the Gaussian mixture form specified hereafter. On irrelevant variables, the data density is modelled by a spherical Gaussian density g belonging to the following family

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

where $0 < \lambda_{\mathrm{m}} < \lambda_{\mathrm{M}}$. Finally, the family of Gaussian mixture associated to $(K, \mathbf{v}) \in \mathcal{M}$ is defined by

$$\mathcal{S}_{(K,\mathbf{v})} = \left\{ x \in \mathbb{R}^Q \mapsto f(x_{[\mathbf{v}]})\, g(x_{[\mathbf{v}^c]});\ f \in \mathcal{L}_{(K,\alpha)},\ g \in \mathcal{G}_{(\alpha)} \right\}. \tag{1.1}$$

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 K of components, the Gaussian mixture form and the number of clustering variables α. Note that a density of $\mathcal{S}_{(K,\mathbf{v})}$ can be written as a 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 1.1.

In this paper, four forms of Gaussian mixtures based on the eigenvalue decomposition of the variance matrices as in [7,17] are considered. The same notation $\mathcal{S}_{(K,\mathbf{v})}$ is used for the four model forms for brevity. The Gaussian mixture notation for those forms is taken from [12].

- 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
$$\Sigma_k = \mathrm{diag}(\sigma_{k1}^2, \ldots, \sigma_{k\alpha}^2)$$
where the eigenvalues σ_{kj}^2 belong to the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The 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 with eigenvalues in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. The model dimension is $D(K, \mathbf{v}) = K[1 + \alpha + \frac{\alpha(\alpha+1)}{2}]$.
- For the $[LB_k]$ 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 determinant 1 and with diagonal coefficients in the interval $[\lambda_{\mathrm{m}}, \lambda_{\mathrm{M}}]$. 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 $D(K, \mathbf{v}) = K(1 + \alpha) + \frac{\alpha(\alpha+1)}{2}$.

These Gaussian mixture models allow to recast clustering problem and detection of noisy variables into a model selection problem.

2. A new penalized likelihood criterion

2.1. Model selection principle

Density estimation deals with the problem of estimating the distribution of a sample $\mathbf{y}$. In many cases, it is not easy to choose a model of adequate dimension. For instance, a model with few parameters tends to be efficiently estimated but can be far from the true distribution. In the opposite situation, a more complex model easily fits data but estimates could have too large variances. The aim of model selection is to construct a data-driven penalized criterion to select a proper dimension model among a model collection. A general theory on this topic, using a non asymptotic view point is proposed by Birgé and Massart (see [30] for an overview). In the density estimation framework, their model selection principle is as follows.

Let $\mathcal{S}$ be the set of all densities with respect to the Lebesgue measure on $\mathbb{R}^Q$. The contrast $\gamma(t, \cdot) = -\ln\{t(\cdot)\}$ leading to the maximum likelihood criterion is considered. 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 $s\mathrm{d}x$ is absolutely continuous with respect to $t\mathrm{d}x$ and $+\infty$ otherwise. The density s being the unique minimizer of the Kullback-Leibler information on $\mathcal{S}$, s is also a minimizer over $\mathcal{S}$ of the expectation of the empirical contrast, defined by
$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n} \ln\{t(\mathbf{y}_i)\}.$$
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 $\mathcal{S}_{(K,\mathbf{v})}$ is expected to lead to a sensible estimator of s, at least if s is close enough to model $\mathcal{S}_{(K,\mathbf{v})}$. The ideal model is minimizing the expected risk

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

However, the oracle function $\hat{s}_{(K^\star,\mathbf{v}^\star)}$ 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 to the oracle risk. A penalized model selection procedure consists of considering a penalized criterion

$$\mathrm{crit}(K,\mathbf{v}) = \gamma_n(\hat{s}_{(K,\mathbf{v})}) + \mathrm{pen}(K,\mathbf{v}) \tag{2.1}$$

where pen is a penalty function pen : $(K,\mathbf{v}) \in \mathcal{M} \mapsto \mathrm{pen}(K,\mathbf{v}) \in \mathbb{R}_+$. Then the selected model $(\hat{K},\hat{\mathbf{v}})$ is a minimizer of the penalized criterion (2.1). The purpose of a non asymptotic approach is to obtain a penalty function providing an oracle inequality. Such an oracle inequality would allow to compare the risk of the penalized maximum likelihood estimator (MLE) $\hat{s}_{(\hat{K},\hat{\mathbf{v}})}$ with the benchmark $\inf_{(K,\mathbf{v})\in\mathcal{M}} \mathbb{E}[\mathrm{KL}(s,\hat{s}_{(K,\mathbf{v})})]$, for a fixed number n of observations.

2.2. **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 [33]. This work has been made possible thanks to the use of a general model selection theorem for MLE, proposed by Massart [30]. 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 are two densities with respect to the Lebesgue measure on $\mathbb{R}^Q$, the squared Hellinger distance between f and g is

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

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 [33] for the four Gaussian mixture models we consider.

Theorem 2.1 (Maugis and Michel [33])**.** *For the four Gaussian mixture collections,*

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

$$\mathrm{pen}(K,\mathbf{v}) \geq \kappa \frac{D(K,\mathbf{v})}{n}\left[2\,A + \ln\left(\frac{1}{1 \wedge \frac{D(K,\mathbf{v})}{n} A}\right) + 1\right] \tag{2.2}$$

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*

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

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

$$\mathrm{pen}(K,\mathbf{v}) \geq \kappa \frac{D(K,\mathbf{v})}{n}\left\{2\,A + \ln\left[\frac{1}{1 \wedge \frac{D(K,\mathbf{v})}{n} A}\right] + \frac{1}{2}\ln\left[\frac{8\exp(1)Q}{(D(K,\mathbf{v})-1)\wedge(2Q-1)}\right]\right\}, \tag{2.3}$$

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*

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

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

The penalty functions take the model complexity into account through $D(K,\mathbf{v})$ as well as the richness of model family. Indeed in the non-ordered variable case, the number of models with the same dimension is larger, and 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 explained in [33], the reason for that is certainly that the general model selection theorem for MLE is stated in a local version whereas we are only able to apply the global version in our framework. Logarithm terms are not detected in practice as shown in Section 4.1 and thus only the preponderant term in $\frac{D(K,\mathbf{v})}{n}$ is retained in the penalty form.

Contrary to classical model selection 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 sensible.

The point we want to stress here is that Theorem 2.1 gives the general form of penalty functions but it does not provide explicit penalties since (2.2) and (2.3) depend on absolute unknown constants and mixture parameters are not bounded in practice. Consequently a method has to be proposed to calibrate the penalty function for a practical use of these results.

3. Slope heuristics

Since the lower bounds on penalty functions in (2.2) and (2.3) are defined up to an unknown multiplicative constant, this theorem does not provide directly an usable model selection criterion. Recently, some efforts have been paid to overcome such a difficulty. Birgé and Massart [14] propose a practical method based on a mixture of theoretical and heuristic ideas for defining efficient penalty functions from the data. This heuristics is proved in [14] in the framework of Gaussian regression with a homoscedastic fixed design and has been generalized by Arlot and Massart [5] in the heteroscedastic random-design case. Nevertheless applications of this method are developed in other frameworks: For instance, in multiple change point detection by Lebarbier [25] and in genomics applications by Villers [40]. This section first describes the main ideas of this heuristics, the so-called "*slope heuristics*", and next details its practical use in our framework.

3.1. 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. A minimizer of $\mathrm{KL}(s,\cdot)$ on $\mathcal{S}_D$ is denoted

$$s_D = \underset{t\in\mathcal{S}_D}{\mathrm{argmin}}\ \mathrm{KL}(s,t)$$

and a minimizer of $\gamma_n(\cdot)$ on $\mathcal{S}_D$ is denoted

$$\hat{s}_D = \underset{t\in\mathcal{S}_D}{\mathrm{argmin}}\ \gamma_n(t).$$

As for criteria due to Mallows [29] and Akaike [2,3], Birgé and Massart's criterion is based on an unbiased risk estimation. The ideal model is minimizing the risk $\mathbb{E}\,[\mathrm{KL}(s, \hat{s}_D)]$. A solution to estimate this ideal model is to find a penalty function, called *optimal penalty* such that the empirical risk is as close as possible to $\inf_{D\in\mathcal{D}} \mathbb{E}\left[\mathrm{KL}(s,\hat{s}_D)\right]$. To express the risk of each estimator $\hat{s}_D$, the following decomposition is considered

$$\begin{aligned}\mathrm{KL}(s,\hat{s}_D) &= \int \ln\left[\frac{s(x)}{s_D(x)}\right] s(x)\mathrm{d}x + \int \ln\left[\frac{s_D(x)}{\hat{s}_D(x)}\right] s(x)\mathrm{d}x \\ &= b_D + V_D \end{aligned} \tag{3.1}$$

where $b_D := \mathrm{KL}(s,s_D)$ is a bias term and $V_D := \int \ln(s_D/\hat{s}_D)s$ is a variance term. Note that the bias b_D decreases whereas the variance term V_D tends to increase when the dimension D increases. Taking the expectation of (3.1) leads to

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

Among the model collection $\mathcal{D}$, the selected model $\hat{D}$ is the one minimizing a criterion of the form

$$D \mapsto \gamma_n(\hat{s}_D) + \mathrm{pen}(D). \tag{3.2}$$

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 is minimizing

$$\begin{aligned}\gamma_n(\hat{s}_D) - \gamma_n(s) + \mathrm{pen}(D) &= \gamma_n(\hat{s}_D) - \gamma_n(s_D) + \gamma_n(s_D) - \gamma_n(s) + \mathrm{pen}(D) \\ &= \hat{b}_D - \widehat{V}_D + \mathrm{pen}(D). \end{aligned} \tag{3.3}$$

Introducing the term of interest (3.1) into (3.3) leads to

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

Because of the law of large numbers, it is reasonable to assume that $\hat{b}_D - b_D \approx 0$. Furthermore, concentration arguments (see [30] p. 9) allow us to suppose that $\mathrm{KL}(s,\hat{s}_D)$ is close to its expectation. Thus

$$\gamma_n(\hat{s}_D) - \gamma_n(s) + \mathrm{pen}(D) \approx \mathbb{E}[\mathrm{KL}(s,\hat{s}_D)] - (V_D + \widehat{V}_D) + \mathrm{pen}(D). \tag{3.4}$$

In order to make (3.4) close to the risk $\mathbb{E}\left[\mathrm{KL}(s,\hat{s}_D)\right]$, the *optimal penalty* is

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

Next, the main point of this heuristics is to assume that $\widehat{V}_D \approx V_D$. An argument to justify this assumption is that the probability measure and the corresponding empirical measure play a similar role, in the expressions of V_D and $\widehat{V}_D$. 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 assumption leads to $\mathrm{pen}_{\mathrm{opt}}(D) = 2\widehat{V}_D$. On an other hand, $\widehat{V}_D$ can be written

$$\begin{aligned}\widehat{V}_D &= \gamma_n(s_D) - \gamma_n(s) + \gamma_n(s) - \gamma_n(\hat{s}_D) \\ &= \hat{b}_D + \gamma_n(s) - \gamma_n(\hat{s}_D).\end{aligned}$$

As the dimension increases, the bias term $\hat{b}_D$ becomes stable as soon as the approximation of the model cannot be appreciably improved. Thus, the behavior of $\widehat{V}_D$ according to the model dimension is derived by $-\gamma_n(\hat{s}_D)$ for large dimensions. In our framework, penalty functions could be regarded as proportional to the dimension (see remarks at the end of Sect. 2.2). Hence $\mathrm{pen}_{\mathrm{opt}}(D)$ is of the form

$$\mathrm{pen}_{\mathrm{opt}}(D) = 2\widehat{V}_D = 2C_{\mathrm{opt}}D$$

where C_{opt} is a constant to be fixed. In order to use the slope heuristics to calibrate the penalty, a required condition is to observe a linear behavior of $D \mapsto -\gamma_n(\hat{s}_D)$ for large dimensions. If this condition is fulfilled, C_{opt} can be estimated by the slope $\hat{C}$ of the linear part of $D \mapsto -\gamma_n(\hat{s}_D)$ and the final penalty is

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

3.2. 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 form have been fixed by the user. Our model selection procedure, based on the slope heuristics, is decomposed into the three following steps:

- *Estimation step*: The maximum likelihood estimator $\hat{s}_{(K,\mathbf{v})}$ is computed for each model $\mathcal{S}_{(K,\mathbf{v})}$. According to (1.1), the mixture parameters and ω^2 can be independently estimated. Thus

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

  where 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 [18] using MIXMOD software [12] and $\hat{\omega}^2 = \frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_{i[\mathbf{v}^c]}\|^2$.
- *Penalty determination step*: First, models are grouped according to their dimension in order to obtain the model collection $(\mathcal{S}_D)_{D\in\mathcal{D}}$. For each dimension $D \in \mathcal{D}$, $\hat{s}_D$ is the estimator providing the largest loglikelihood value among the estimators associated to a model of dimension D. The associated model is denoted $(K_D, \mathbf{v}_D)$ $(\hat{s}_D = \hat{s}_{(K_D,\mathbf{v}_D)})$.

  The function $D \mapsto -\gamma_n(\hat{s}_D)$ is plotted. The user has to check that this function has a linear behavior for large dimensions. If this condition is not fulfilled, the slope heuristics cannot be applied. This situation may occur when the estimation step is performed not for large enough dimensions or when the model collection is not adapted to the dataset.

  Assume that the linear behavior of $D \mapsto -\gamma_n(\hat{s}_D)$ is observed for large dimensions. The user has to choose a threshold D_0 such that the restriction of $-\gamma_n(\hat{s}_D)$ for dimensions greater than D_0 has a linear behavior. Then the slope $\hat{C}$ of this linear part is evaluated. Since possible estimation errors in the first step can damage the slope estimation, a robust regression procedure [21] is used. This procedure which makes use of an iteratively weighted least squares algorithm, is expected to give lower weights to suboptimal and spurious parameter estimates. Finally, the calibrated penalty function is $\text{pen}(D) = 2\hat{C}D$.
- *Model selection step*: The minimizer $\hat{D}$ of the criterion $D \mapsto \gamma_n(\hat{s}_D) + 2\hat{C}D$ is determined and the model $(\hat{K},\hat{\mathbf{v}}) = (K_{\hat{D}}, \mathbf{v}_{\hat{D}})$ is selected. Finally, the estimated parameter vector associated to $(\hat{K},\hat{\mathbf{v}})$ provides a data clustering using the MAP rule (see Sect. 1.1).

Remark about estimated oracle model: when simulated datasets (known density s) are studied, the model $(\hat{K},\hat{\mathbf{v}})$ selected with our penalized criterion can be compared to the oracle model

$$(K^\star, \mathbf{v}^\star) = \underset{(K,\mathbf{v})\in\mathcal{M}}{\operatorname{argmin}}\ \mathbb{E}\left[-\int \ln\{\hat{s}_{(K,\mathbf{v})}(x)\} s(x)\mathrm{d}x\right] \tag{3.5}$$

with a Monte Carlo procedure. A first sample $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$, with a large size n, is simulated from the density s in order to approximate the integral of the right-hand side of (3.5) by

$$f_{\mathbf{x}}(\hat{s}_{(K,\mathbf{v})}) = -\frac{1}{n}\sum_{i=1}^{n} \ln\{\hat{s}_{(K,\mathbf{v})}(\mathbf{x}_i)\}.$$

Then a collection of M independent samples from s is considered. For the jth sample, the maximum likelihood estimators $(\hat{s}^{(j)}_{(K,\mathbf{v})})_{(K,\mathbf{v})\in\mathcal{M}}$ are computed. Thus the estimated oracle model is defined by

$$(\hat{K}_{\text{oracle}}, \hat{\mathbf{v}}_{\text{oracle}}) = \underset{(K,\mathbf{v})\in\mathcal{M}}{\operatorname{argmin}} \frac{1}{M}\sum_{j=1}^{M} f_{\mathbf{x}}(\hat{s}^{(j)}_{(K,\mathbf{v})}).$$

4. Applications

This section is devoted to the application of our method on simulated and real datasets. The method is applied on simulated datasets in Sections 4.1 and 4.2. In Section 4.3, our procedure is carried out on a curve clustering example for oil production profiles. In Section 4.4, a transcriptome dataset is studied with our method to obtain coexpressed gene clusters.

For each example we check the conditions ensuring the slope heuristics application and our data-driven criterion is compared with the classical criteria used for Gaussian mixture model selection: AIC, BIC and ICL. They are respectively defined by

$$\begin{aligned}\text{crit}_{\text{AIC}}(D) &= \gamma_n(\hat{s}_D) + \frac{D}{n},\\ \text{crit}_{\text{BIC}}(D) &= \gamma_n(\hat{s}_D) + \frac{D\ln(n)}{2n}\end{aligned}$$

and

$$\text{crit}_{\text{ICL}}(D) = \text{crit}_{\text{BIC}} + \frac{\text{ENT}}{n}$$

with the entropy term $\text{ENT} = -\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\ln(t_{ik})$ where z is given by the MAP rule and

$$t_{ik} = \frac{\hat{p}_k\Phi(\mathbf{y}_i|\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{l=1}^{K}\hat{p}_l\Phi(\mathbf{y}_i|\hat{\mu}_l,\hat{\Sigma}_l)}\cdot$$

The reader is respectively referred to [2], [37] and [11] for more details on these criteria.

4.1. Assessment of the slope heuristics

The aim of this first example is threefold: (i) checking the validity of the linear penalty shape assumption, (ii) comparing the slope estimator with other penalized estimators, (iii) analyzing its behavior with respect to the oracle. The dataset consists of $n = 2000$ points described by $Q = 22$ variables. The data are simulated according to a mixture of four equiprobable Gaussian distributions $\mathcal{N}(\mu_k,\Sigma_k)$ where

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

and

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

The vector 0_p denotes the null vector of length p. Twelve 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 clustering 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.

The model collection associated to the $[L_kB_k]$ Gaussian mixture form 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

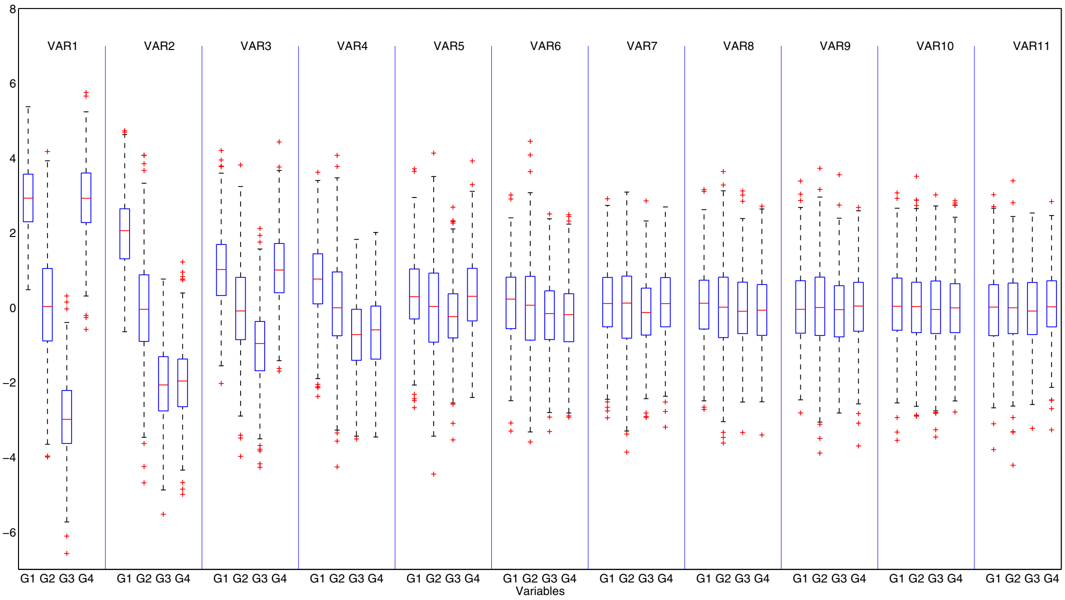

Figure 1. Boxplots of the first eleven variables (VAR1,...,VAR11) on the four mixture components (G1,G2,G3,G4).

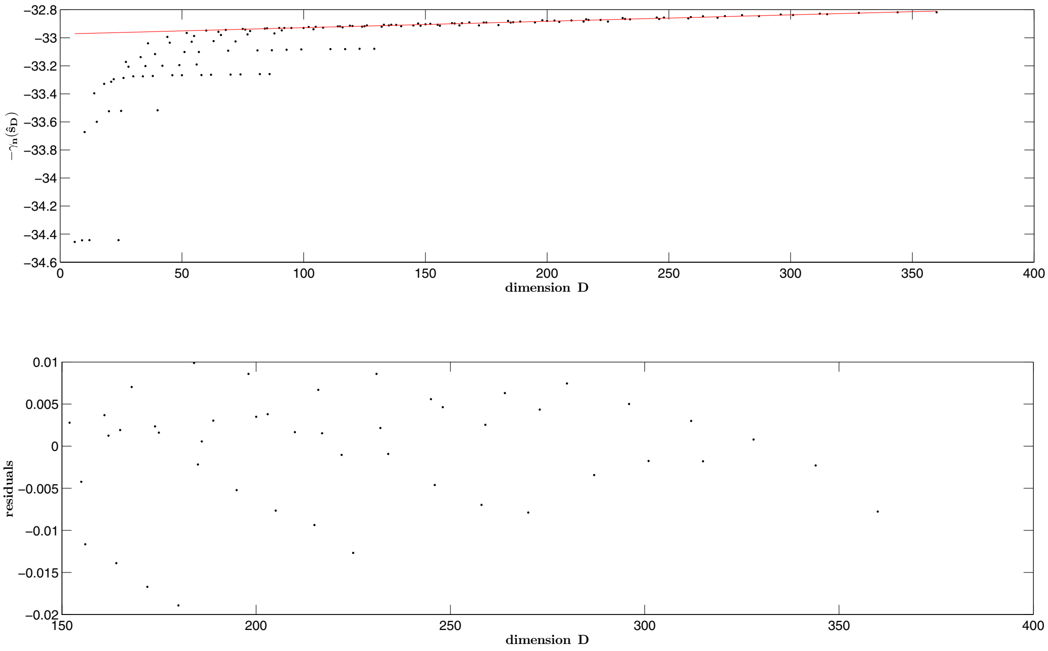

Figure 2. On the top graph, the function $D \mapsto -\gamma_n(\hat{s}_D)$ is plotted. The linear regression is performed for $D \geq D_0 = 150$. The associated residuals are drawn on the bottom graph.

TABLE 1. Risk estimation of four models in the collection around the oracle model ($K = 4, \alpha = 9$).

α	7	8	**9**	10
$\mathbb{E}[\ln(\hat{s}_{(4,\alpha)})]$	–32.9443	–32.9399	**–32.9382**	–32.9383

TABLE 2. For each criterion, number of times a model (K, α) is selected among the 1000 simulations.

criterion	$\hat{K}$	$\hat{\alpha}$								
		5	6	7	8	9	10	11	12	≥13
ICL	4	17	**771**	209	3					
BIC	4	26	**838**	136						
AIC	≤4			2	5	8			1	
	5			2	24	21	13	5	2	2
	6			9	56	66	29	12	7	3
	7			15	67	108	64	22	14	10
	8			23	101	**126**	99	32	29	23
Slope Heuristics	4		56	**438**	**438**	63				
	5			3	1	1				

$D \geq D_0 = 150$, we observe that the function $D \mapsto -\gamma_n(\hat{s}_D)$ has a linear behavior as expected (see Sect. 3.1). The residuals of the linear regression are plotted on the bottom of Figure 2. This supports the use of penalties proportional to the dimension since no trend can be observed in the residuals. The estimation of $\hat{C}$ leads to choose the penalty of criterion (3.2) and the selected model is $\hat{K}_{\text{slope}} = 4$ and $\hat{\alpha}_{\text{slope}} = 7$.

In order to compare slope heuristics, the oracle, AIC, BIC and ICL criteria, 1000 datasets have been simulated as described before. Since the true density is known, a Monte Carlo procedure (see Sect. 3.2) gives the following oracle model estimation $\hat{K}_{\text{oracle}} = 4$ and $\hat{\alpha}_{\text{oracle}} = 9$ and other models show a risk close to the oracle risk (see Tab. 1). Note that even if the true density belongs to the density collection, the oracle model is not equal to the corresponding true model. The results for the 1000 simulations are summarized in Table 2.

The AIC criterion selects too many components and relevant variables since it underpenalizes models in the mixture context. Its averaged classification error rate is 8.11% in this context. The two criteria BIC and ICL select a model with four components and most of the times with six relevant variables. It is shown in [24] 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 is no consistency result for such a noisy variable detection for mixtures. The model selected by BIC is not the true model. In this context, even if BIC tries to find the true model, the asymptotics could be not achieved. 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. From a clustering point of view, BIC and ICL lead to an averaged classification error rate of 7.72%. As expected, the slope method selects one of the best models around the oracle model (see Tab. 1). An interesting fact is that the error rate of the oracle model ($K = 4, \alpha = 9$) is a bit lower (7.62%) than those of AIC, ICL and BIC criteria. This is also the case for the true model ($K = 4, \alpha = 10$) which error rate is 7.63%. By selecting a model close to the oracle one, the slope estimator shows a better averaged classification error rate (7.65%) than AIC, ICL and BIC criteria.

The interest of this first simulated example is to illustrate the behavior of the different criteria. For this particular simulation, it is clear that BIC and ICL criteria tend to provide a clustering and an associated sparse set of clustering variables whereas the slope estimator tends to provide a clustering based on the detection of the noisy variables.

TABLE 3. Contingency table for the clustering obtained with the slope heuristics.

	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

4.2. Waveform dataset

The waveform dataset, available at the UCI repository [15] and described in detail in [16], 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 added noise. The dataset consists of 5000 observations described by 40 variables. The last nineteen variables are noisy, sampled from a $\mathcal{N}(0,1)$ density. By construction, Variables 1 and 21 have the same distribution $\mathcal{N}(0,1)$. Consequently they are both irrelevant for the clustering and thus there are 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 selection should be based on all the possible relevant variable subsets $\mathbf{v}$. Nevertheless, it is impossible because of the large cardinal of the model collection. 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 with a $\mathcal{N}(0,1)$ density.

The model selection is performed with the two mixture forms $[L_kB_k]$ and $[L_kC_k]$. The plots of $D \mapsto -\gamma_n(\hat{s}_D)$ for the estimation of $\hat{C}$ are given on the top of Figure 3 for these two model collections. The fit of the dimension model surface on maximum loglikelihood surface is presented for each model collection at the bottom in Figure 3. As expected, we observe that the fitting is dramatically better for the model collection associated to the $[L_kC_k]$ collection. Indeed the relevant variables are dependent by construction and the $[L_kB_k]$ model collection is too simple for this problem. This $[L_kC_k]$ 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.

Table 3 gives the contingency table for the waveform data clustering obtained with our procedure. The three clusters are well related with the three true groups, with an error rate of 14.3%. Figure 4 plots the error rate in function of the number α of relevant variables, each curve is associated to a fixed number of components in the mixture. Choosing a model with an ill-chosen subset of relevant variables deteriorates the clustering performance. Note that the classification error rate is lower for the slope heuristics than for AIC (19.32%) and ICL (44.3%).

4.3. 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 [6] 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. 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 behavior with productions of three fields of different size. In order to compare the production profile forms, we consider production profiles normalized by the amount of reserves contained in each field. The aim of the study is to validate that a clustering of normalized production profiles is consistent with the values of the reserves variable.

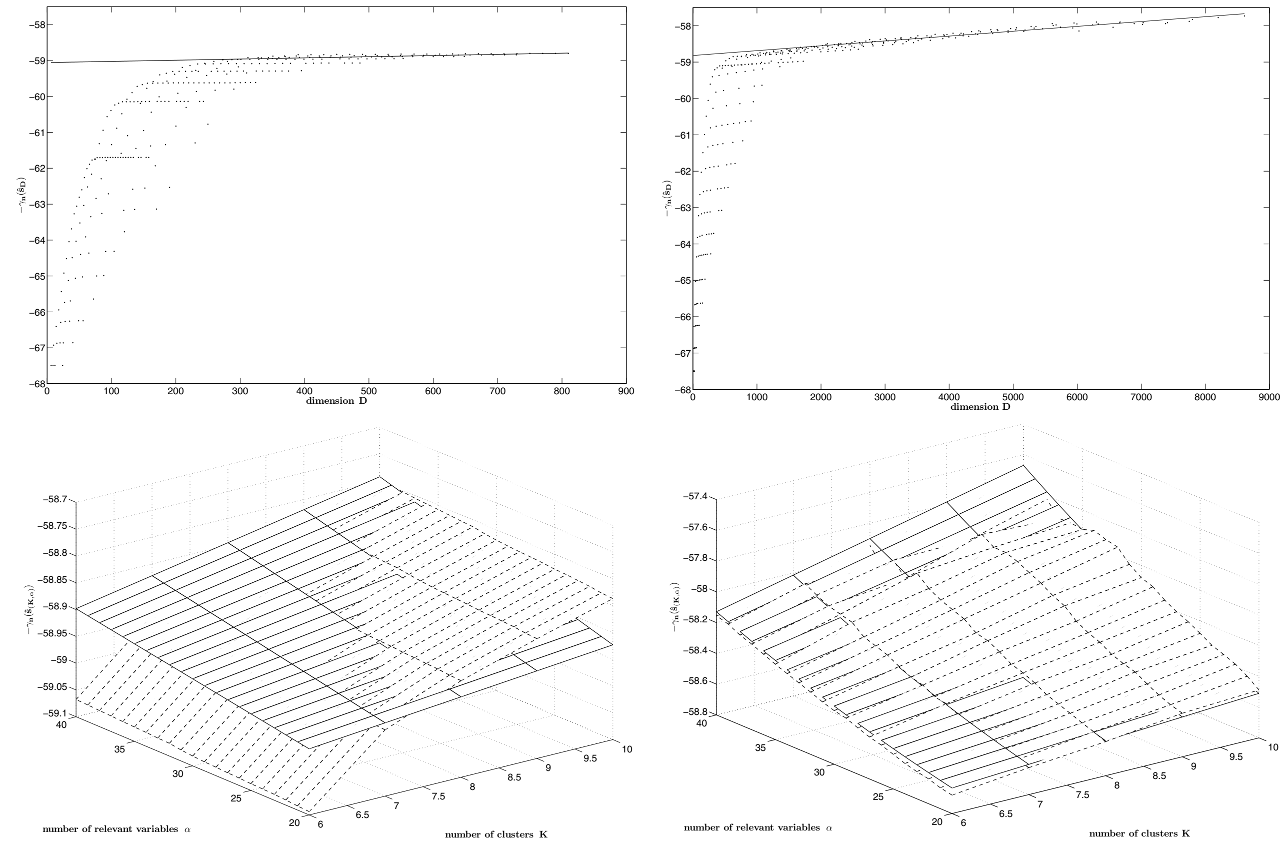

FIGURE 3. On the top, graphical representation of $D \mapsto -\gamma_n(\hat{s}_D)$ leading to the estimation of $\hat{C}$ for the $[L_kB_k]$ form (on the left) and the $[L_kC_k]$ form (on the right). On the bottom, fitting of the dimension surface on $(K, \alpha) \in [|6, 10|] \times [|20, 40|] \mapsto -\gamma_n(\hat{s}_{(K,\alpha)})$ for the $[L_kB_k]$ form (on the left) and the $[L_kC_k]$ form (on the right).

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 follows. 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. 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 [10]). The reader is referred to [35] for details on the DWT. Let $\mathbf{W}_i$ be the wavelet coefficient vector of the ith 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

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

[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

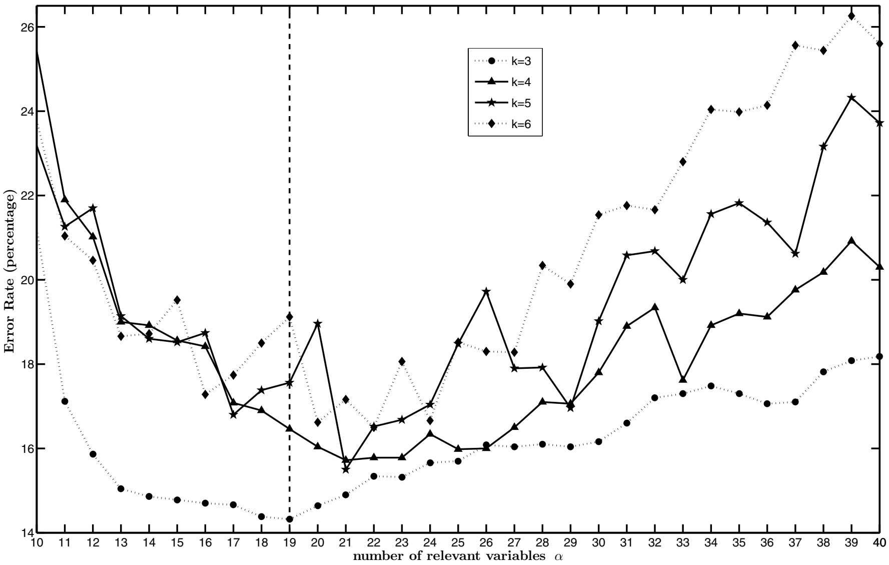

FIGURE 4. Evolution of the clustering error rate in function of the chosen number α of relevant variables. Each curve corresponds to a fixed number K of mixture components.

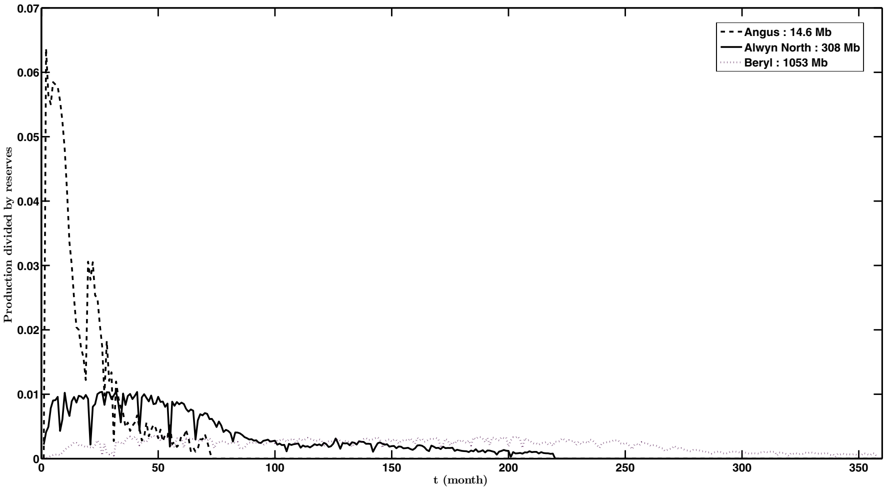

FIGURE 5. Oil production profiles normalized by reserves of three fields located in the North Sea.

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 to 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 finest 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 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

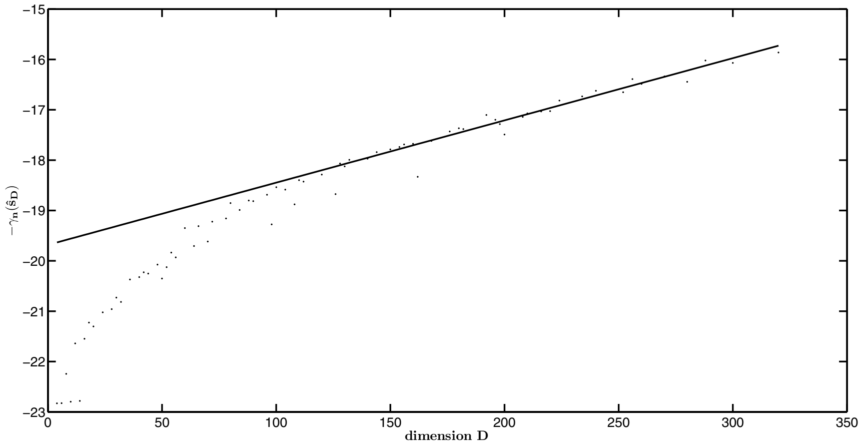

FIGURE 6. Slope method applied to the $[LB_k]$ collection for the wavelet coefficient curve data.

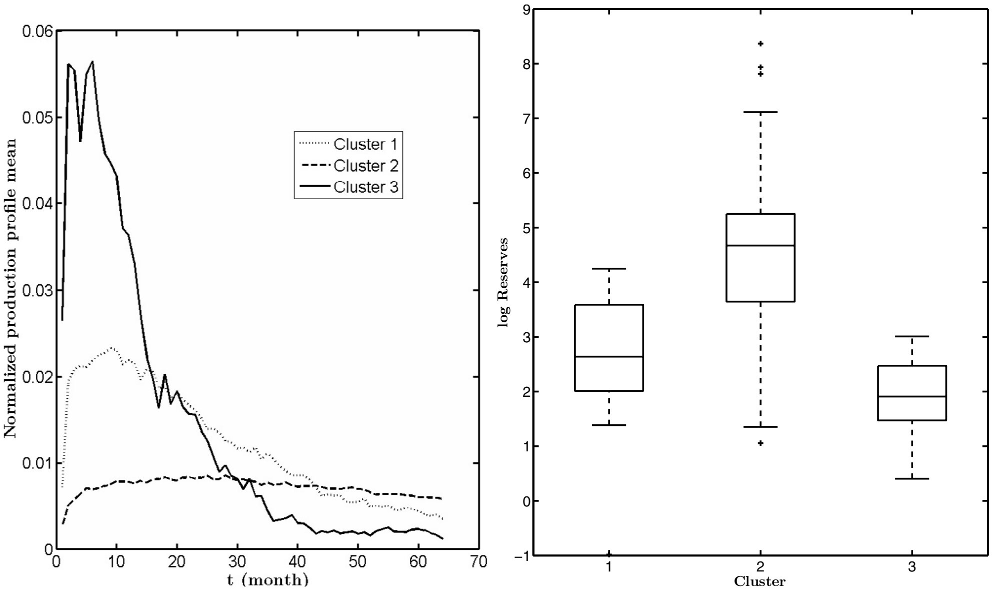

FIGURE 7. On the left: normalized production profile means for each cluster. On the right: boxplots of the logarithm of the reserves variable for each cluster.

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 avoids 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 $\hat{\alpha} = 10$ clustering variables. On this dataset, the classical methods BIC ($\hat{K} = 5$, $\hat{\alpha} = 16$) and AIC ($\hat{K} = 10$, $\hat{\alpha} = 16$) select all the variables, whereas ICL is more relevant with the slope estimator since it gives $\hat{K} = 5$ and $\hat{\alpha} = 11$.

The clusters derived from the slope heuristics have 31, 140 and 9 curves respectively. Figure 7 displays on the left the mean cluster of the normalized production profiles in each cluster. Boxplot of the logarithm of

Table 4. Number of differentially expressed genes per experiment.

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 5. Number of genes per cluster.

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

Table 6. Selected model for each criterion.

criterion	$\hat{K}$	$\hat{\alpha}$	$\hat{\mathbf{v}}$
slope heuristics	8	7	$\{3; 5–10\}$
BIC	8	7	$\{3; 5–10\}$
AIC	39	9	$\{2–10\}$
ICL	8	9	$\{2–10\}$

the reserves variable for each cluster is displayed on the right part of this figure. 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. As expected, the shape of normalized production profiles can be explained by the reserves variable. The reader is referred to [34] for more details.

4.4. Analysis of a transcriptome dataset

Currently, biologists are interested in gene functional analysis. It is usually considered that coexpressed genes are often implicated in the same biological function and consequently are potential candidate to be co-regulated genes. Thus biologists try to extract groups of coexpressed genes according to transcriptome datasets in order to characterize more precisely their biological functions. Moreover the detection of irrelevant experiments can be 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 [19]. 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, a test statistic corresponding to the normalized differential expression and a p-value adjusted by the Bonferroni method are calculated in order to test if a gene is non-differentially expressed or not in an experiment. A gene is declared to be differentially expressed when its Bonferroni p-value is lower than 0.05. The reader is referred to [19] for a description of such an analysis and [27] for an application.

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 4 provides 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.

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 Fig. 8), thus the slope heuristics 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 Tab. 5) and the clustering shows some interesting similar behaviors of expression profiles (see Fig. 9). A similar clustering can be found if all the variables are considered ($\alpha = 10$ fixed) but with the noisy variable detection, the interpretation of the clustering is made clearer. The selected models for each of the classical criteria are given in Table 6.

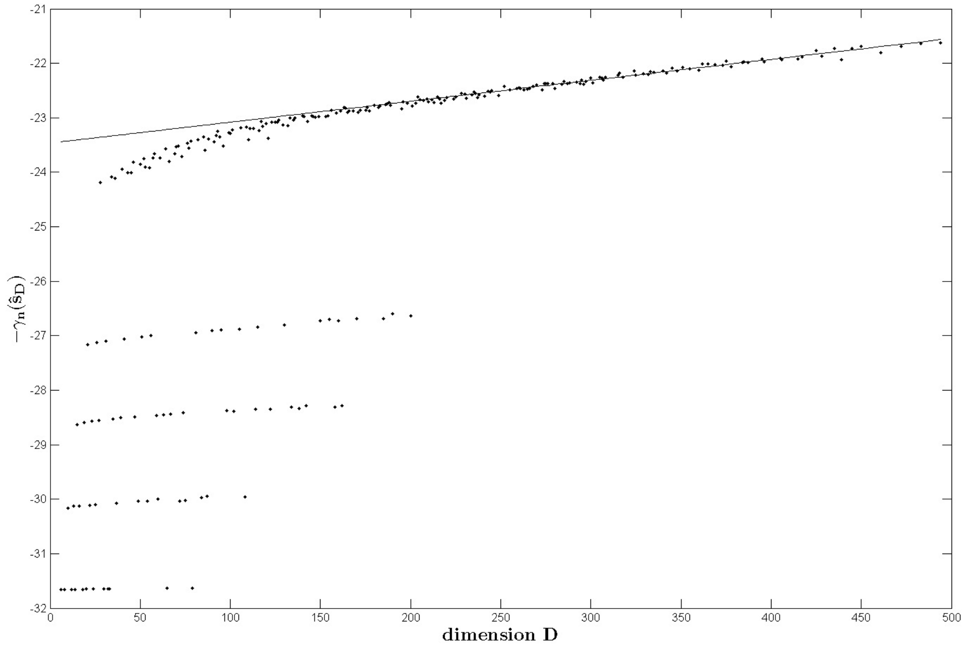

FIGURE 8. Penalty determination on the linear behavior of the function $D \mapsto -\gamma_n(\hat{s}_D)$.

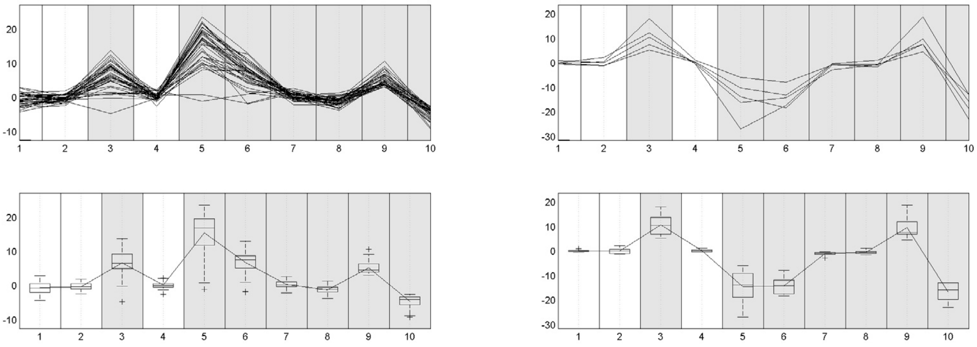

FIGURE 9. Graphical representation of gene profiles in Clusters 2 (on the left) and 8 (on the right). Relevant experiments are colored in grey.

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 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 Fig. 9). 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. To validate this explanation, t-test between experiments 7 and 8 for the eight clusters have been performed at level 0.05. Only the test for Cluster 2 is significant (p-value $< 5.10^{-4}$). This clustering can help biologists to find gene biological functions. For instance, 12 genes for which biologists do not know biological function are clustered with 27 other genes in Cluster 2. According to biological knowledge, the 27 genes have the same subcellular localisation (in plastid) and are involved in the photosynthesis of *Arabidopsis thaliana*. Thus the function of the 12 unknown genes is certainly related to the photosynthesis. This hypothesis remains to be confirmed by biologists experimentally.

5. Discussion

This paper illustrates how the slope heuristics can be applied to calibrate a penalty function whose form is known in a model-based clustering problem. More precisely, we are interested in the detection of noisy variables for improving the Gaussian mixture clustering and its interpretation. The problem is recast into a Gaussian mixture model selection problem where model collections are indexed by the number K of clusters and the clustering variable subset $\mathbf{v}$. The slope heuristics allows us to calibrate the multiplicative constant in the penalty term of the theoretical criterion whose form is justified in the companion paper [33]. The behavior of this slope heuristic method is studied on simulated and real datasets. It has been compared with the standard asymptotic criteria BIC, ICL and AIC, currently used in this Gaussian mixture clustering context. Note that without noisy variable detection ($\alpha = Q$), our method allows one to select the number of clusters which is the fundamental problem in model-based clustering. Furthermore, our theoretical and practical results could be extended for most of the twenty-eight Gaussian mixture forms proposed by Celeux and Govaert [17]. Thus our work could be adapted for selecting also the Gaussian mixture form.

In the previous applications of the slope heuristics, the penalty function was calibrated according to the "dimension jump" method. It consists of determining the minimal penalty and using the rule of thumb of the slope heuristics, saying that the optimal penalty is twice the minimal penalty. This minimal penalty pen_{min} is such that the selected model has a too large dimension if the penalty is lower than pen_{min} and has a reasonable dimension if the penalty is greater than pen_{min}. Thus the key point of this method consists of determining this transition corresponding to the maximal dimension jump. This calibration method was applied for instance in [5,25,40]. In this paper, we do not use this method to calibrate the penalty since the maximal dimension jump does not often clearly appear, especially for real data (see [4] p. 94).

Our heuristic method is based on the linear behavior of $D \mapsto -\gamma_n(\hat{s}_D)$. Thus in any case, the user should check that this linear behavior in large dimensions is observed. The absence of a linear behavior can have different reasons: (i) the model dimension can be too low, namely the maximum number of components $K_{\max}$ has to be increased, (ii) the problem can be 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. When this linear behavior is observed for dimensions larger than a threshold D_0, the slope of the restriction of $D \mapsto -\gamma_n(\hat{s}_D)$ to $D \geq D_0$ is estimated. A robust regression is used to attenuate the influence of possible estimation errors. Ideally, we would like to propose a totally data-driven method which determines automatically the threshold D_0 but this automatic choice is not an easy task. An alternative solution would be to base the method on the stability of the selected model in function of the threshold value. In a work in progress, this strategy is adopted for a graphical user interface devoted to the general application of the slope heuristics, without restriction to Gaussian mixture clustering context.

Recent works [31,32,36] on variable selection in model-based clustering have been proposed to progressively improve the variable role modelling. In particular, the models involved in these methods allow the irrelevant clustering variables to be linked to some relevant clustering variables through a linear regression. With a variable selection point of view, our hypotheses about the correlation between variables are less realistic than the ones considered in the three cited papers. One possible challenge would be to extend our procedure for such variable role modellings. Unfortunately, the first step consisting of adapting the theoretical results given in [33] in order to find the penalty shape is a difficult task since the bracketing entropy of the multivariate regression should be controlled. Next, the number of models involved in such variable role modellings being very large, an exhaustive search of the best model is impossible. A convenient strategy allowing us to run the estimation step only for a model subcollection is necessary. A natural idea is to use a backward or a forward stepwise algorithm for the variable selection as in [31,32,36]. Nevertheless, the slope heuristics method being not totally automatic, the BIC criterion used in the three associated variable selection algorithms cannot be directly replaced by our penalized criterion for the moment. Efforts should be concentrated on this question in future work.

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.

References

[1] C. Abraham, P.A. Cornillon, E. Matzner-Løber and N. Molinari. Unsupervised curve clustering using B-splines. *Scand. J. Stat. Th. Appl.* **30** (2003) 581–595.

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

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

[4] S. Arlot, *Rééchantillonnage et sélection de modèles*, Ph.D. thesis, Université Paris-Sud XI (2007).

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

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

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

[8] A. Barron, L. Birgé and P. Massart, Risk bounds for model selection via penalization. *Prob. Th. Rel. Fields* **113** (1999) 301–413.

[9] J.-P. Baudry, Clustering through model selection criteria. *Poster session at One Day Statistical Workshop in Lisieux.* http://www.math.u-psud.fr/~baudry, June (2007).

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

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

[12] C. Biernacki, G. Celeux, G. Govaert and F. Langrognet, Model-based cluster and discriminant analysis with the MIXMOD software. *Comp. Stat. Data Anal.* **51** (2006) 587–600.

[13] L. Birgé and P. Massart, Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** (2001) 203–268.

[14] L. Birgé and P. Massart, Minimal penalties for Gaussian model selection. *Prob. Th. Rel. Fields* **138** (2006) 33–73.

[15] K.-E. Blake and C. Merz, *Uci repository of machine learning databases* (1999). http://mlearn.ics.uci.edu/MLSummary.html.

[16] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and regression trees.* Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA (1984).

[17] G. Celeux and G. Govaert, Gaussian parsimonious clustering models. *Patt. Recog.* **28** (1995) 781–793.

[18] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Methodol.* **39** (1977) 1–38, With discussion.

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

[20] L.A. García-Escudero and A. Gordaliza, A proposal for robust curve clustering. *J. Class.* **22** (2005) 185–201.

[21] P.J. Huber, *Robust Statistics.* Wiley (1981).

[22] G.M. James and C.A. Sugar, Clustering for sparsely sampled functional data. *J. Am. Stat. Assoc.* **98** (2003) 397–408.

[23] D. Jiang, C. Tang and A. Zhang, Cluster analysis for gene expression data: A survey. *IEEE Trans. Knowl. Data Eng.* **16** (2004) 1370–1386.

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

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

[26] V. Lepez, *Potentiel de réserves d'un bassin pétrolier: modélisation et estimation*, Ph.D. thesis, Université Paris Sud (2002).

[27] C. Lurin, C. Andréas, S. Aubourg, M. Bellaoui, F. Bitton, C. Bruyère, M. Caboche, J. Debast, C. Gualberto, B. Hoffmann, M. Lecharny, A. Le Ret, M.-L. Martin-Magniette, H. Mireau, N. Peeters, J.-P. Renou, B. Szurek, L. Taconnat and I. Small, Genome-wide analysis of arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16** (2004) 2089–103.

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

[29] C.L. Mallows, Some comments on *Cp. Technometrics* **37** (1973) 362–372.

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

[31] C. Maugis, G. Celeux and M.-L. Martin-Magniette, Variable selection for clustering with Gaussian mixture models. *Biometrics* **65** (2009) 701–709.

[32] C. Maugis, G. Celeux and M.-L. Martin-Magniette, Variable selection in model-based clustering: A general variable role modeling. *Comput. Stat. Data Anal.* **53** (2009) 3872–3882.

[33] C. Maugis and B. Michel, A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: P&S* **15** (2011) 41–68.

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

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

[36] A.E. Raftery and N. Dean, Variable selection for model-based clustering. *J. Am. Stat. Assoc.* **101** (2006) 168–178.

[37] G. Schwarz, Estimating the dimension of a model. *Ann. Stat.* **6** (1978) 461–464.

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

[39] T. Tarpey and K.K.J. Kinateder, Clustering functional data. *J. Class.* **20** (2003) 93–114.

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