On the number of groups in clustering

Aurélie Fischer

LSTA, Université Pierre et Marie Curie – Paris VI
Boîte 158, Couloir 15-16, 2^e étage
4 place Jussieu, 75252 Paris Cedex 05, France

Abstract

Clustering is the problem of partitioning data into a finite number k of homogeneous and separate groups, called clusters. A good choice of k is essential for building meaningful clusters. In the present paper, this task is addressed from the point of view of model selection via penalization. We design an appropriate penalty shape and derive an associated oracle-type inequality. The method is illustrated on both simulated and real-life data sets.

Index terms: k-means clustering, number of clusters, model selection, oracle inequality, penalty calibration, slope heuristics.

2010 Mathematics Subject Classification: 62H30.

1 Introduction

Clustering is the problem of dividing data into a finite number of relevant classes, so that items in the same group are as similar as possible, and items in different groups are as dissimilar as possible (Duda, Hart and Stork [13]). This unsupervised learning technique has been widely used for statistical data analysis in a variety of areas. For an integer $k \geq 1$, the so-called k-means clustering method consists in partitioning a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ taking values in $(\mathbb{R}^d, \|\cdot\|)$ in k groups by minimizing the empirical distortion

$$W(\mu_n, \mathbf{c}) = \frac{1}{n} \sum_{i=1}^{n} \min_{j=1,\ldots,k} \|\mathbf{X}_i - c_j\|^2$$

over all possible codebooks $\mathbf{c} = (c_1, \ldots, c_k) \in (\mathbb{R}^d)^k$ (see, e.g., Linder [26], Graf and Luschgy [16]).

In this context, an essential problem is how to select the right number k of clusters. Indeed, if in some situations the choice of k may be motivated by the applications, it is in general unknown.

A presentation of various procedures for choosing k can be found in Milligan and Cooper [30] and Hardy [17], while Gordon compares in [15] the performances of the best five rules exposed in [30]. These methods can be divided in two main types, global or local. Global procedures consist in performing clustering for different values of k and then retaining the value minimizing or maximizing some function of k. In local procedures, it must be decided at each step whether a cluster should be partitioned (or two groups merged into a single one).

Calinski and Harabasz [11] propose to choose the value of k which maximizes an index based on the quotient

$$\frac{B(k)/(k-1)}{W(k)/(n-k)}.$$

Here, the empirical distortion is denoted by $W(k)$ to highlight the dependency in k, whereas $B(k) = \sum_{j=1}^{k} \|c_j - \bar{c}\|^2$, where $\bar{c}$ is the mean of the data, is the between sum of squares. The method by Krzanowski and Lai [23] consists in maximizing $W(k)k^{2/d}$, or more precisely the related quantity

$$\Big|\frac{\mathrm{DIFF}(k)}{\mathrm{DIFF}(k+1)}\Big|,$$

where

$$\mathrm{DIFF}(k) = W(k-1)(k-1)^{2/d} - W(k)k^{2/d},$$

whereas in Hartigan's rule [18], a new cluster is added while the quantity

$$H(k) = \frac{W(k)}{W(k-1)}(n-k-1)$$

is less than a certain threshold. The *Silhouette* statistic of Kaufman and Rousseeuw [20] takes the form

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average distance between $\mathbf{X}_i$ and the observations belonging to the same cluster as $\mathbf{X}_i$, and $b(i)$ is the average distance between $\mathbf{X}_i$ and the observations in the nearest cluster (i.e., the cluster minimizing $b(i)$). An observation $\mathbf{X}_i$ is well clustered when $s(i)$ is large, and Kaufman and Rousseeuw [20] suggest to retain the value of k which maximizes the average of $s(i)$ for $i = 1, \ldots, n$. The *Gap Statistic* of Tibshirani, Walther and Hastie [39] compares the evolution of the logarithm of the distortion for the considered clustering problem with the function obtained for uniformly distributed observations. Kim, Park and Park [22] develop an index which allows to select $\hat{k}$ by combining two functions of opposite monotonicity presenting a jump around the optimal value of k, whereas Sugar and James [38] propose to apply to the empirical distortion a transformation $w \mapsto w^{-p}, p > 0$. Other methods rely on the stability of partitions. In this case, the number of groups is chosen thanks to a criterion which has been determined using the clustering results obtained when considering several subsamples of the data set (see, e.g., Levine and Domany [24] and Ben-Hur, Elisseeff and Guyon [8]). The relation between the number k and cluster stability has been investigated from a theoretical point of view in Shamir and Tishby [35, 36], Ben-David, Luxburg, and Pál [7], Ben-David, Pál, and Simon [5] and Ben-David and Luxburg [6].

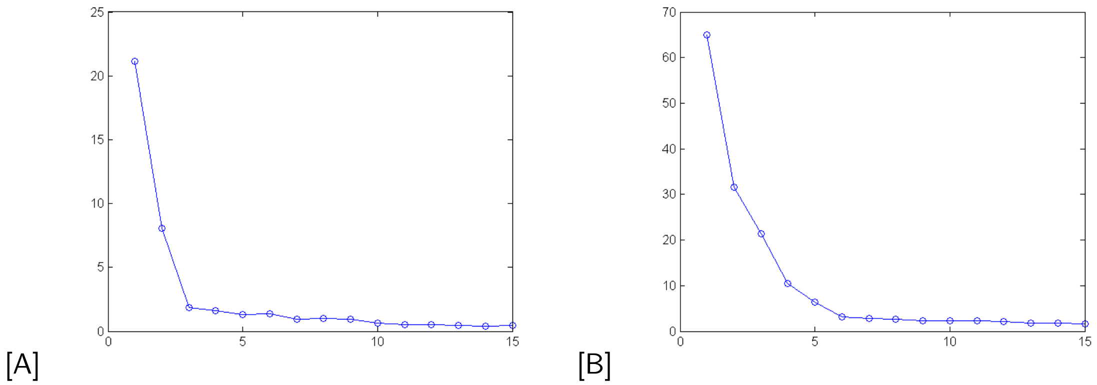


Figure 1: Graph of the empirical distortion as a function of k for two examples: [A] 3 clusters and [B] 6 clusters.

Using the empirical distortion in order to choose k is a natural idea. Unfortunately, it is always decreasing in k (see Figure 1). Thus, minimizing the empirical distortion would lead to choose k as large as possible, which does not present much interest (think, for example, of the situation where $k = n$, and each single observation builds a cluster). Moreover, Hastie, Tibshirani and Friedman [19] observe that it is not enough to evaluate the criterion on an independent test data set, since numerous centers would anyway densely fill the whole data space so that each observation is very close to one of them. Consequently, it is not possible to select k by cross validation like in supervised learning.

In this note, we propose a data-driven approach for evaluating k, which is based on the empirical distortion and relies on the model selection scheme introduced by Birgé and Massart [9] and Barron, Birgé, Massart [3]. The main advantage is that the performance of this method of choice of k can be assessed through an oracle-type inequality.

The rest of the note is organized as follows. In Section 2, we obtain a penalty shape and show an inequality ensuring a control of the risk of the estimator obtained by minimization of the corresponding penalized criterion. Section 3 presents some simulations and real data experiments illustrating the practical implementation of the proposed approach. For the sake of clarity, the proof of the main result is postponed to Section 4.

2 The choice of k

Throughout the paper, we let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random vectors, with the same distribution as a generic random vector $\mathbf{X}$ taking its values in $\mathbb{R}^d$ endowed with its standard Euclidean norm $\|\cdot\|$. We assume that $\mathbf{X}$ satisfies the peak power constraint, that is, for some $R > 0$,

$$\mathbb{P}\left\{\|\mathbf{X}\| \leq R\right\} = 1. \tag{1}$$

For every k, the minimization of the empirical distortion

$$W(\mu_n, \mathbf{c}) = \frac{1}{n}\sum_{i=1}^{n} \min_{c\in\mathbf{c}} \|\mathbf{X}_i - c\|^2$$

yields some codebook with k components. Our mission is to select the best one over all possible values of k.

To this aim, for all k, $1 \leq k \leq n$, let S_k denote the (countable) set of all $(c_1, \ldots, c_k) \in \mathcal{Q}^k$, where $\mathcal{Q}$ is some grid over $\mathbb{R}^d$. Observe that restricting the search of the centers to $\mathcal{Q}$ is a mild assumption, since, in practice, an algorithm can only provide centers belonging to such a grid. For every k, let $\hat{\mathbf{c}}_k$ be a minimizer of the criterion $W(\mu_n, \mathbf{c})$ over S_k, i.e., $\hat{\mathbf{c}}_k \in \arg\min_{\mathbf{c}\in S_k} W(\mu_n, \mathbf{c})$. To determine the best codebook in $\{\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_n\}$, we will search for a $\hat{k}$ minimizing a criterion of the type

$$\operatorname{crit}(k) = W(\mu_n, \hat{\mathbf{c}}_k) + \operatorname{pen}(k),$$

where $\operatorname{pen} : \{1, \ldots, n\} \to \mathbb{R}^+$ is a penalty function, whose role is to avoid the choice of a too large k.

In order to design a penalty, we follow the route of non-asymptotic model selection (Birgé and Massart [9] and Barron, Birgé, Massart [3]). We will make use of Theorem 8.1 in Massart [28], as well as of the following upper bound established by Linder [25]:

$$\mathbb{E}\left[\sup_{\mathbf{c}\in S_k}(W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] \leq aR^2\sqrt{\frac{kd}{n}},$$

where a is an absolute constant.

Theorem 2.1. *Consider a family of nonnegative weights $\{x_k\}_{1\leq k\leq n}$ such that*

$$\sum_{k=1}^{n} e^{-x_k} = \Sigma.$$

If for every $1 \leq k \leq n$,

$$\operatorname{pen}(k) \geq R^2\left[a\sqrt{\frac{kd}{n}} + 4\sqrt{\frac{x_k}{2n}}\right], \tag{2}$$

then, letting $\tilde{\mathbf{c}} = \hat{\mathbf{c}}_{\hat{k}}$ denote a minimizer of the penalized criterion

$$\operatorname{crit}(k) = W(\mu_n, \hat{\mathbf{c}}_k) + \operatorname{pen}(k),$$

we have

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq \inf_{1\leq k\leq n}\left(W_k^\star(\mu) + \operatorname{pen}(k)\right) + R^2\Sigma\sqrt{\frac{2\pi}{n}}, \tag{3}$$

where $W_k^\star(\mu) = \inf_{\mathbf{c}\in S_k} W(\mu, \mathbf{c})$.

Theorem 2.1 suggests a penalty function $\operatorname{pen}(k)$ tending to 0 at the rate $1/\sqrt{n}$ and provides an upper bound for the expectation of the distortion at $\tilde{\mathbf{c}}$. Inequality (3) ensures that if the penalty function is large enough, the expectation of the distortion at $\tilde{\mathbf{c}}$ remains relatively low, at least close to the smallest value of the distortion over k, up to a vanishing term.

The term of the order $\sqrt{k/n}$ in the penalty reflects the complexity of the models since a model is more complex when k is larger. Let us point out that the proof of Theorem 3 in Linder [25] reveals that we can set $a = 96$. However, this value of the constant a, which is an upper bound, is of limited interest in practice, all the more so since the expression (2) involves the radius R. In fact, Theorem 2.1 gives the shape of the penalty rather than an exact penalty function, but this penalty may be calibrated using the so-called slope heuristics, as discussed below in Section 3.

Then, consider the weights $\{x_k\}_{1\leq k\leq n}$. The larger they are, the smaller Σ is. Nevertheless, they should not be too large since they also appear in the penalty. In the Gaussian linear model selection framework, where each model S_m, $m \in \mathcal{M}$, has dimension D_m, a possible choice when there is no redundancy in the models dimension consists in taking x_m proportional to D_m (see Massart [28], Section 4.2.1). By analogy, the weights may here be taken proportional to k. To summarize, the penalty is a constant times $\sqrt{k/n}$ to a first approximation.

For a penalty of the form

$$\mathrm{pen}(k) = a'R^2\sqrt{\frac{kd}{n}},$$

observe that Theorem 2.1 allows to derive a rate of convergence for $\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right]$. Indeed, an extension of the Pierce Lemma [34] due to Luschgy and Pagès [27] ensures that, for a suitable grid $\mathcal{Q}$, $W_k^\star(\mu)$ is not larger than

$$\frac{A(d)R^2}{k^{2/d}},$$

where $A(d)$ is a positive constant depending only on the dimension d. Thus, inequality (3) can be rewritten as

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq R^2\left(\inf_{1\leq k\leq n}\left[\frac{A(d)}{k^{2/d}} + a'\sqrt{\frac{kd}{n}}\right] + \Sigma\sqrt{\frac{2\pi}{n}}\right).$$

Optimizing the two terms in the brackets leads to k of the order $n^{\frac{d}{d+4}}$ and shows that the expected distortion at the selected codebook $\tilde{\mathbf{c}}$ vanishes at the rate $\mathcal{O}(n^{-\frac{2}{d+4}})$, as expressed in the next corollary.

Corollary 2.1. *If*

$$\mathrm{pen}(k) = a'R^2\sqrt{\frac{kd}{n}}$$

and $\tilde{\mathbf{c}} = \hat{\mathbf{c}}_{\hat{k}}$ *is obtained by minimization of the penalized criterion*

$$\mathrm{crit}(k) = W(\mu_n, \hat{\mathbf{c}}_k) + \mathrm{pen}(k),$$

then

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \leq C(d, R)n^{-\frac{2}{d+4}},$$

where $C(d, R)$ *is a constant which depends only on* d *and* R.

Remark 2.1. Let $\rho \geq 0$ (typically, $\rho = n^{-2}$). If for all k, $\hat{\mathbf{c}}_k$ is an approximate minimizer of the empirical risk $W(\mu_n, \mathbf{c})$, in the sense that for all $\mathbf{c} \in S_k$,

$$W(\mu_n, \hat{\mathbf{c}}_k) \leq W(\mu_n, \mathbf{c}) + \rho,$$

then Theorem 2.1 remains true provided one adds ρ in the right-hand term of the inequality.

3 Experimental results

In the present section, we propose to illustrate on some simulated and real data examples the choice of the number k of clusters using the penalized criterion suggested by Theorem 2.1. As mentioned earlier, we can assume to a first approximation that the penalty shape is $c\sqrt{k/n}$, where c is a constant which has to be determined in practice. To this end, we will use the slope heuristics, introduced by Birgé and Massart [10] and further developed by Arlot and Massart [2]. This method precisely allows calibrating a penalty known up to a multiplicative constant. An essential condition for applying the method is that the empirical contrast must be a decreasing function of the complexity of the models and the penalty shape an increasing function. This condition is clearly satisfied in our clustering framework. Two techniques based on the slope heuristics may be employed in order to calibrate a penalty. The dimension jump method consists in identifying an abrupt jump in the models complexity, whereas the other possibility is to observe that the empirical contrast is proportional to the penalty shape for complex models and compute the slope of this line. Both methods have been implemented in MATLAB by Baudry, Maugis and Michel [4] as an interface called CAPUSHE (CAlibrating Penalty Using Slope HEuristics).

From a computational point of view, the k-means algorithm used in all examples is initialized by taking as the unique center for $k = 1$ the mean of all observations. Then, at the kth step, a new center chosen uniformly at random among the observations is added to the $k - 1$ centers resulting from the previous step—the algorithm is therefore random. This procedure is repeated 50 times and the set of centers yielding the lower distortion is kept. Let us point out that there is an abundant literature about the interesting problem of the initialization of the k-means algorithm and that the strategy could be replaced by a more robust one (see, e.g., Pena, Lozano and Larranaga [32], Su and Dy [37], Khan and Ahmad [21], Perim, Wandekokem and Varejão [33], Al-Shboul and Myaeng [1]).

3.1 Simulated data

3.1.1 Some different numbers of groups and dimensions

In a first series of simulations, we tried to recover the number of clusters for 5 types of samples, which are different by dimension d and underlying number k of groups. In these examples, the right number of clusters is found in general. We observe that the direct slope estimation and the dimension jump method perform approximately similarly.

G1 A single group. We begin with a situation where it is not relevant to cluster the data and consider 200 points distributed uniformly in the unit hypercube in dimension 10.

G2 3 groups in dimension 2. The observations were sampled following a normal bivariate distribution with variance the identity matrix and build 3 groups, centered at $(0, 0), (0, 6)$ and $(5, -3)$ respectively, each containing 30 observations (see Figure 2). An example of CAPUSHE output for this data set is visible in Figure 3.

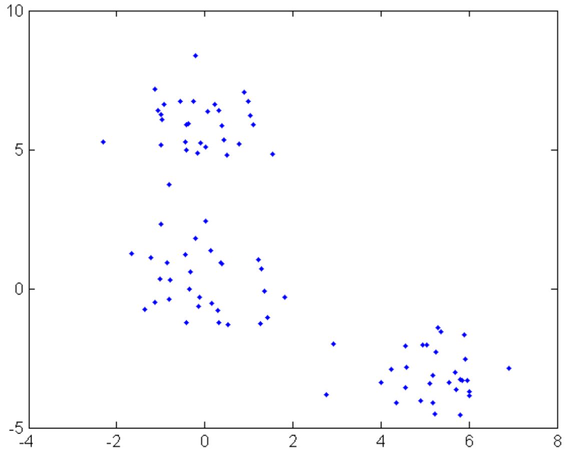


Figure 2: Groups of 30 observations following a normal distribution centered at $(0,0),(0,6)$ and $(5,-3)$.

G3 4 groups in dimension 3. Next, we used 4 groups of observations following a normal distribution in dimension 3 with variance the identity matrix. These groups, depicted in Figure 4, are centered at $(0,0,0),(3,5,-1),(-5,0,0)$, and $(6,6,6)$.

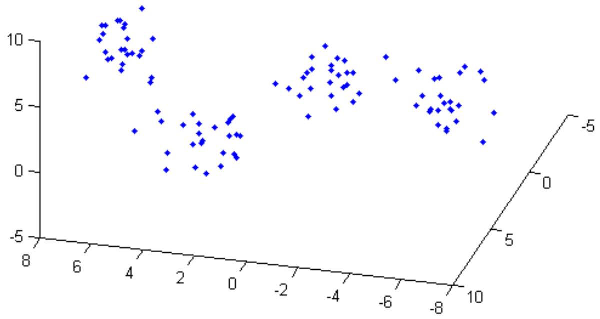
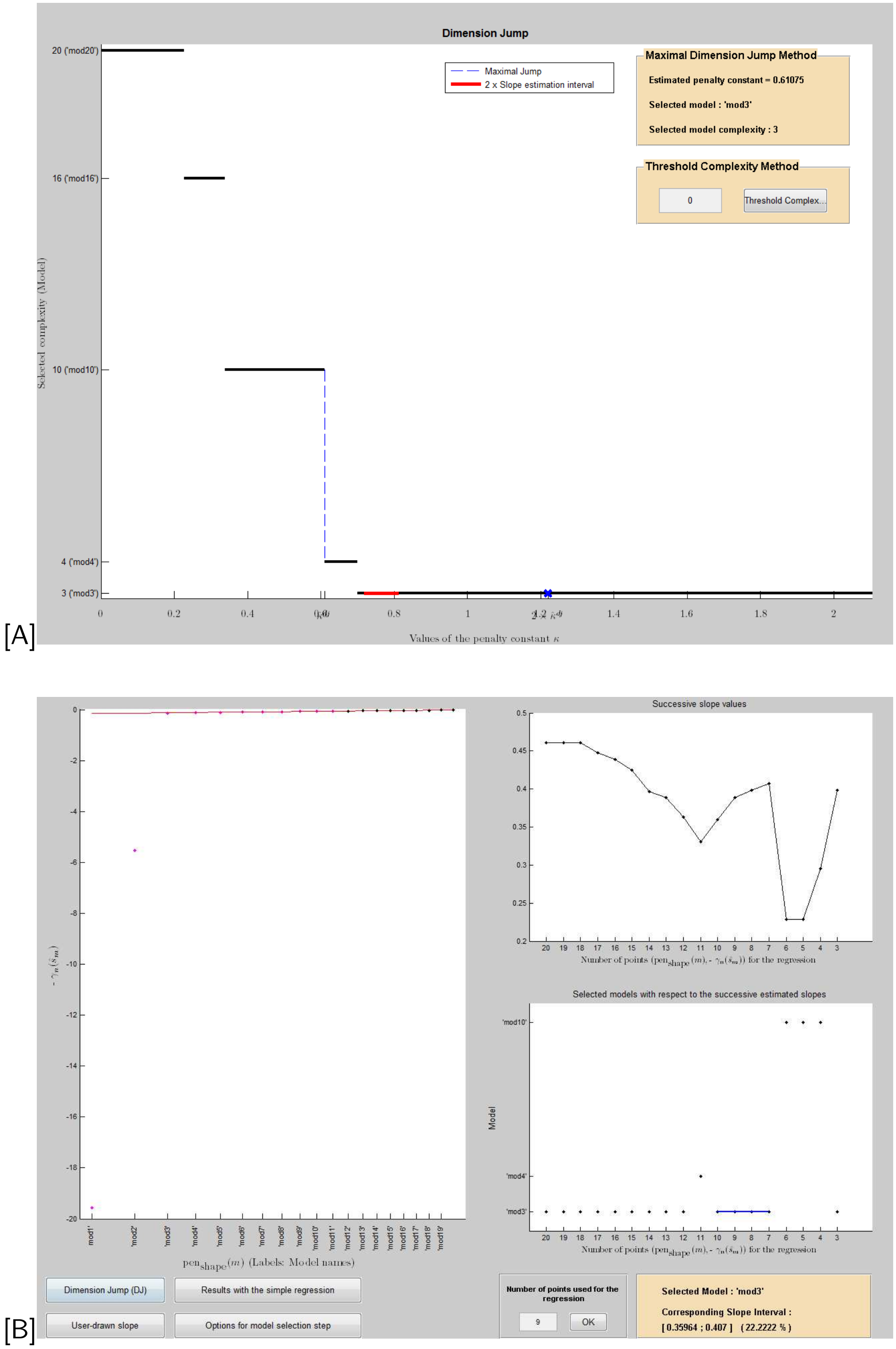


Figure 4: Groups of 25 observations following a normal distribution centered at $(0,0,0),(3,5,-1),(-5,0,0)$ and $(6,6,6)$.

G4 5 groups in dimension 4. This data set is made of 5 normal groups in dimension 4. The 5 groups are centered at $(0,0,0,0),(3,5,-1,0),(-5,0,0,0)$, $(1,1,6,-2)$ and $(1,-3,-2,5)$ respectively.

G5 4 groups in dimension 10. Finally, we have simulated, still using the normal distribution, 4 groups of data in dimension 10. For each of them, the 10 components of the mean vector were chosen uniformly at random between 0 and 10.

Table 1 shows for the 5 simulated data sets the number $\hat{k}$ of clusters obtained with the slope estimation method by averaging over 20 repeated trials.



Maximal Dimension Jump Method

Estimated penalty constant = 0.61075

Selected model : 'mod3'

Selected model complexity : 3

Threshold Complexity Method

0

Number of points used for the regression

9

Selected Model : 'mod3'

Corresponding Slope Interval :
[0.35964 ; 0.407] (22.2222 %)

Figure 3: Output of CAPUSHE (n=90, d=2, k=3). [A] Dimension jump: Selected $\hat{k}$ versus penalty constant values. [B] Slope estimation: **Left**: Graph of the criterion $-W(\mu_n, \hat{\mathbf{c}}_k)$ as a function of $\sqrt{k/n}$. **Upper right**: Successive estimated slope values versus the number of points used for the slope estimation. **Bottom right**: Selected values of $\hat{k}$ versus the number of points used for the slope estimation.

Data set	**G1**	**G2**	**G3**	**G4**	**G5**
Number of groups $\hat{k}$	1.05	3.2	4.1	5	4.05

Table 1: Number of clusters given by the algorithm based on the slope heuristics (average over 20 repeated trials).

3.1.2 More or less separate groups

In this subsection, two different configurations corresponding to 4 groups in dimension 3 are studied (see Figure 5). In the first example (Figure 5 [A]), the clusters are not as well separated as in the second (Figure 5 [B]). We compared the results of the algorithm based on the slope method and the *Gap Statistic* of Tibshirani et al. [39] in both situations.

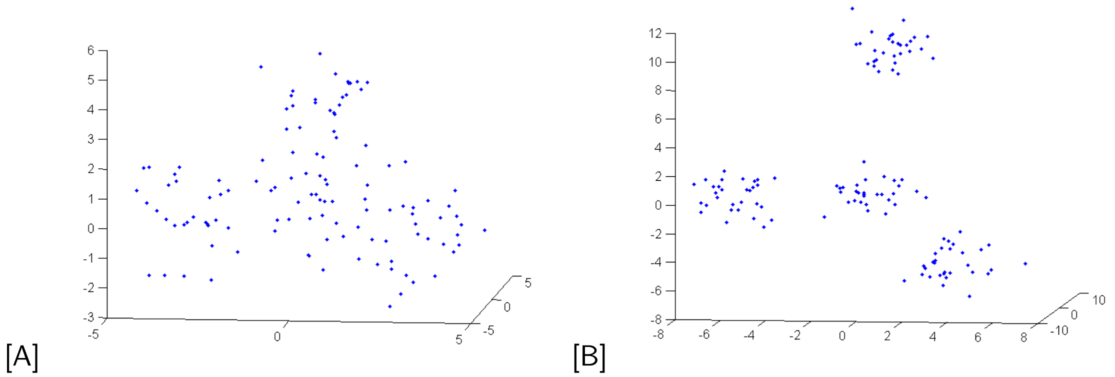


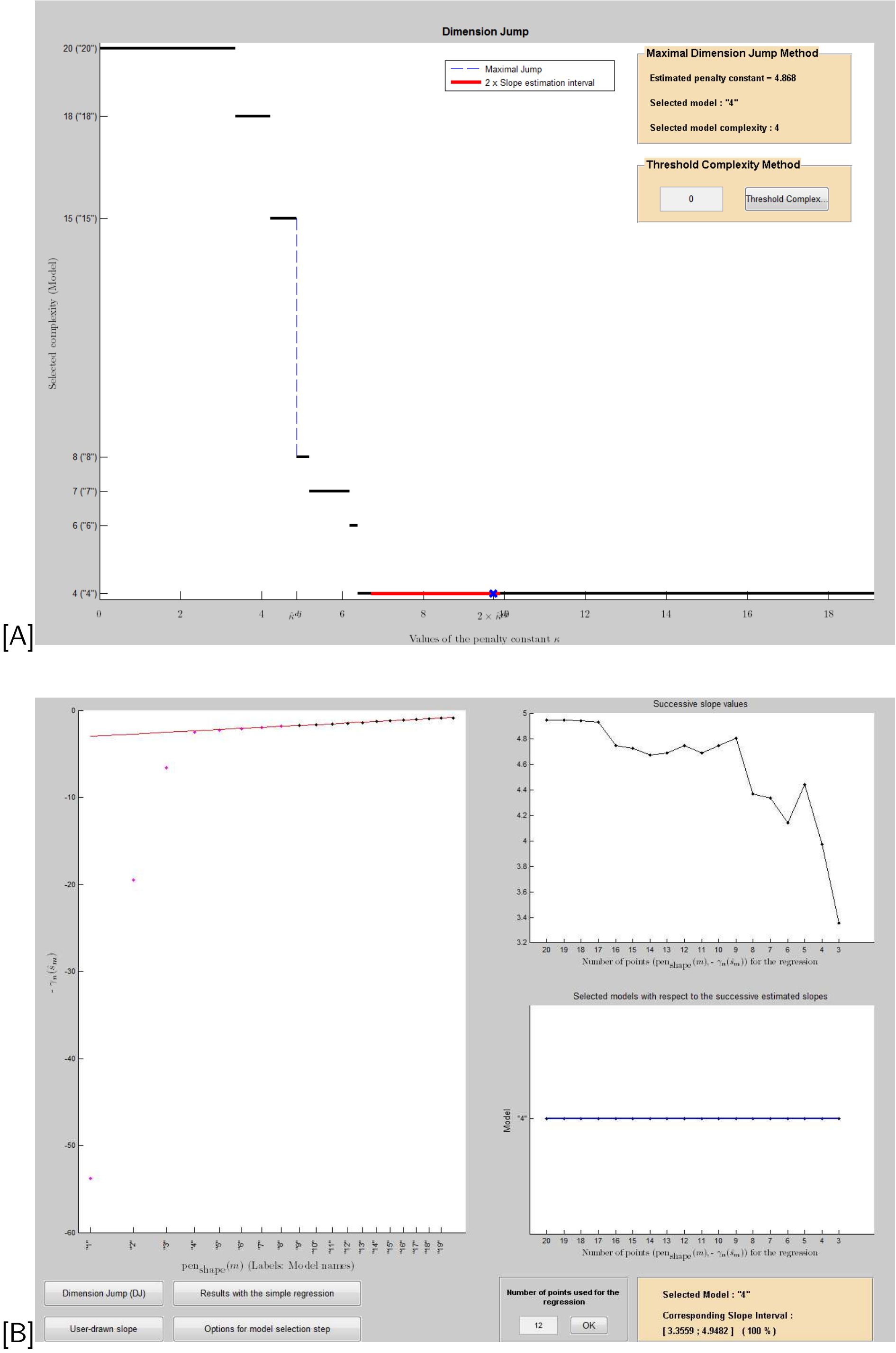
Figure 5: More or less separate groups: 4 groups of 30 observations following a normal distribution. [A] Groups centered at $(0,0,0),(0,2,3),(3,0,-1)$ and $(-3,-1,0)$. [B] Groups centered at $(0,0,0),(0,6,10),(3,0,-5)$ and $(-6,-3,0)$.

For the least separate groups, which are centered at $(0,0,0)$, $(0,2,3)$, $(3,0,-1)$ and $(-3,-1,0)$, the method selecting $\hat{k}$ by means of the slope heuristics yields $\hat{k}=4$ a little more than half of the time. The other values given by the algorithm are 3, 5, 6 and 7. The value 3 is obtained rarely, whereas the *Gap Statistic* founds $\hat{k}=3$ almost every time. For 10 realizations of such groups, Table 2 shows, for both methods, the average value of $\hat{k}$ over 20 trials. The fact that the *Gap Statistic* does not perform very well here suggests that these clusters are two close from each other to estimate $\hat{k}$ accurately.

Slope method	4.25	5.2	5.2	4.6	4.6	4.5	4.55	4.55	4.6	4.15
Gap Statistic	3	4	3	3	3	3.1	3	3	3	3.2

Table 2: Number of clusters given by the algorithm based on the slope method and the *Gap Statistic*, for 10 realizations of Gaussian groups centered at $(0,0,0)$, $(0,2,3)$, $(3,0,-1)$ and $(-3,-1,0)$ (average over 20 repeated trials).

However, when the groups are well separated, the algorithm relying on the slope method seems to perform well, and the results are very similar to those obtained with the *Gap Statistic*: both methods almost always recover the expected result. For the well separate normal clusters, centered at $(0,0,0),(0,6,10),(3,0,-5)$ and $(-6,-3,0)$, visible in Figure 5 [B], we obtain for example the CAPUSHE outputs shown in Figure 6.



[A]
Dimension Jump
20 ("20")
18 ("18")
15 ("15")
8 ("8")
7 ("7")
6 ("6")
4 ("4")
Selected complexity (Model)
0
2
4
$\hat{\kappa}^{dj}$
6
8
$2 \times \hat{\kappa}^{dj}$
10
12
14
16
18
Values of the penalty constant κ
Maximal Jump
2 x Slope estimation interval
Maximal Dimension Jump Method
Threshold Complexity Method
0
Threshold Complex...
[B]
0
-10
-20
-30
-40
-50
-60
$-\gamma_n(\hat{s}_m)$
"1"
"2"
"3"
"4"
"5"
"6"
"7"
"8"
"9"
"10"
"11"
"12"
"13"
"14"
"15"
"16"
"17"
"18"
"19"
$\text{pen}_{\text{shape}}(m)$ (Labels: Model names)
Successive slope values
5
4.8
4.6
4.4
4.2
4
3.8
3.6
3.4
3.2
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3
Number of points $(\text{pen}_{\text{shape}}(m), -\gamma_n(\hat{s}_m))$ for the regression
Selected models with respect to the successive estimated slopes
Model
"4"
20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3
Number of points $(\text{pen}_{\text{shape}}(m), -\gamma_n(\hat{s}_m))$ for the regression
Dimension Jump (DJ)
Results with the simple regression
User-drawn slope
Options for model selection step
Number of points used for the regression
12
OK

Estimated penalty constant = 4.868

Selected model : "4"

Selected model complexity : 4

Selected Model : "4"

Corresponding Slope Interval :
[3.3559 ; 4.9482] (100 %)

Figure 6: Output of CAPUSHE. [A] Dimension jump. [B] Slope estimation.

3.2 Real-life data

3.2.1 Zoo

The data, available from the UCI Machine Learning Repository [14], contains information about different animal species (such as wolf, herring, chicken, crab...). For each animal, 16 features have been registered: hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic, catsize (Booleans), number of legs (integer in $\{0, 2, 4, 5, 6, 8\}$). We consider a sample made of 92 data items building 5 underlining groups: mammal, fish, invertebrate, bird, insect. Most of the time, the output of the algorithm is $\hat{k} = 5$, the other values obtained being 4 and 6. The average number $\hat{k}$ over 20 trials was 5.05.

3.2.2 Dyslexia

Here, the data arises from a study about dyslexia, carried out at the Laboratoire de Sciences Cognitives et Psycholinguistique located in Paris in the Département d'Etudes Cognitives (DEC) of Ecole Normale Supérieure. In order to better understand this disability affecting a person's fluency or accuracy in reading, speaking, and spelling, several hypotheses have to be tested by comparing the performance of dyslexic and non-dyslexic adults (`http://www.ehess.fr/lscp/persons/ramus/fr/phonodysfr.html`). People aged from 18 to 31 took part to experiments based on RAN tests (Rapid Automatized Naming, see, e.g., Denkla and Rudel [12]), which consist in naming rapidly digits, colors and objects, and on listening to a list of words and "non-words". Here, "non-word" means syllables put together that do not build an existing word. For instance, "distu", "malani" and "sonper" are non-words. For each of the 57 considered people, we have results such as response time, number of errors and response accuracy rate.

On this problem, the most often selected value of $\hat{k}$ was 4, and the average over 20 runs 3.9. The algorithm did not often choose $\hat{k} = 2$, which would correspond to the two classes, dyslexic people and control group. Nevertheless, the result can be explained by some "false positive" and "false negative" results in the study. Indeed, a few dyslexic individuals answered quite accurately and rapidly compared to the other dyslexic people, whereas some people not affected by dyslexia were slower and made more errors than expected.

3.2.3 Tasmanian Abalone

Abalone, also called ear-shell or sea ear, is a kind of sea snail. This marine gastropod mollusk in the family Haliotidae and the genus Haliotis presents several shell layers ("rings"), which can be used to learn its age, an important task to study the biology and ecology of a species. In fact, the number of "rings" plus 1.5 gives the age in years. More precisely, to determine the age of abalone, the shell is cut through the cone, stained, and the biologist counts the number of rings through a microscope. To avoid this time-consuming task, it can be of interest to predict the age of abalone from other physical measurements, which are easier to obtain. Here, we used a data set originating from the Marine Resources Division of the Marine Research Laboratories, Taroona, Department of Primary Industry and Fisheries, Tasmania (Nash, Sellers, Talbot, Cawthorn, and Ford [31]) and available from the UCI Machine Learning Repository [14]. The data contains information relative to 4177 abalones, labeled "female", "male" or "infant". For each of them, seven representative features have been measured: length (longest shell measurement), diameter (perpendicular to length), height, whole weight, shucked weight, viscera weight (after bleeding), shell weight (after being dried).

Considering a subsample of 1303 female abalones, with number of rings ranging from 5 to 23, we intend to recover the number of age groups from the physical measurements. The algorithm yields a number $\hat{k}$ of groups between 16 and 22 and the average value of $\hat{k}$ over 20 trials is 18.

4 Proof of Theorem 2.1

Theorem 2.1 is an adaptation of Theorem 8.1 in Massart [28]. For the proof, we will need the following lemma, which is a consequence of McDiarmid's inequality [29] (see Massart [28, Theorem 5.3]).

Lemma 4.1. *If $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent random variables and $\mathcal{G}$ is a finite or countable class of real-valued functions such that $a \leq g \leq b$ for all $g \in \mathcal{G}$, then if $Z = \sup_{g \in \mathcal{G}} \sum_{i=1}^n (g(\mathbf{X}_i) - \mathbb{E}[g(\mathbf{X}_i)])$, we have, for every $\varepsilon \geq 0$,*

$$\mathbb{P}\left\{Z - \mathbb{E}[Z] \geq \varepsilon\right\} \leq \exp\left(-\frac{2\varepsilon^2}{n(b-a)^2}\right).$$

Proof of the theorem. Observe that, by the definition of $\tilde{\mathbf{c}}$,

$$W(\mu_n, \tilde{\mathbf{c}}) + \mathrm{pen}(\hat{k}) \leq W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k)$$

for all $k, 1 \leq k \leq n$ and $\mathbf{c}_k \in S_k$. Thus,

$$W(\mu_n, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) \leq \mathrm{pen}(k) - \mathrm{pen}(\hat{k}),$$

which leads to

$$W(\mu, \tilde{\mathbf{c}}) \leq W(\mu_n, \mathbf{c}_k) + W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \tilde{\mathbf{c}}) + \mathrm{pen}(k) - \mathrm{pen}(\hat{k}). \tag{4}$$

Consider nonnegative weights $\{x_k\}_{1 \leq k \leq n}$ such that

$$\sum_{k=1}^n e^{-x_k} = \Sigma,$$

and let $z > 0$. Applying Lemma 4.1, we obtain, for all $k', 1 \leq k' \leq n$ and all $\varepsilon \geq 0$,

$$\mathbb{P}\left\{\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \geq \mathbb{E}\left[\sup_{\mathbf{c} \in S'_k} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] + \varepsilon\right\} \leq \exp\left(-\frac{n\varepsilon^2}{8R^4}\right).$$

It follows that for every $k', 1 \leq k' \leq n$,

$$\mathbb{P}\left\{\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \geq \mathbb{E}\left[\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right] + 4R^2\sqrt{\frac{x_{k'} + z}{2n}}\right\} \leq e^{-x_{k'} - z}.$$

Setting $E_{k'} = \mathbb{E}\left[\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c}))\right]$, we have, for all $k', 1 \leq k' \leq n$,

$$\sup_{\mathbf{c} \in S_{k'}} (W(\mu, \mathbf{c}) - W(\mu_n, \mathbf{c})) \leq E_{k'} + 4R^2\sqrt{\frac{x_{k'} + z}{2n}},$$

except on a set with probability not larger than Σe^{-z}. By inequality (4), we thus get

$$\begin{aligned} W(\mu, \tilde{\mathbf{c}}) &\le W(\mu_n, \mathbf{c}_k) + E_{\hat{k}} + 4R^2\sqrt{\frac{x_{\hat{k}} + z}{2n}} - \mathrm{pen}(\hat{k}) + \mathrm{pen}(k) \\ &\le W(\mu_n, \mathbf{c}_k) + E_{\hat{k}} + 4R^2\sqrt{\frac{x_{\hat{k}}}{2n}} - \mathrm{pen}(\hat{k}) + \mathrm{pen}(k) + 4R^2\sqrt{\frac{z}{2n}}, \end{aligned}$$

except on a set of probability not larger than Σe^{-z}. Next, according to Linder [25, Theorem 3], there exists a constant $a > 0$ such that

$$E_{k'} \le aR^2\sqrt{\frac{k'd}{n}}.$$

Thus, if for all $k', 1 \le k' \le n$,

$$\mathrm{pen}(k') \ge R^2\left[a\sqrt{\frac{k'd}{n}} + 4\sqrt{\frac{x_{k'}}{2n}}\right],$$

then

$$W(\mu, \tilde{\mathbf{c}}) \le W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k) + 4R^2\sqrt{\frac{z}{2n}},$$

except on a set of probability not larger than Σe^{-z}. This may be rewritten

$$\mathbb{P}\left\{(4R^2)^{-1}\sqrt{2n}[W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k)] \ge \sqrt{z}\right\} \le \Sigma e^{-z},$$

or, setting $z = u^2$,

$$\mathbb{P}\left\{(4R^2)^{-1}\sqrt{2n}[W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k)] \ge u\right\} \le \Sigma e^{-u^2}.$$

Recalling that $\int_0^{+\infty} e^{-u^2}\mathrm{d}u = \frac{\sqrt{\pi}}{2}$ and letting $g_+ = \max(g, 0)$, we get

$$\mathbb{E}\left[(W(\mu, \tilde{\mathbf{c}}) - W(\mu_n, \mathbf{c}_k) + \mathrm{pen}(k))_+\right] \le R^2\Sigma\sqrt{\frac{2\pi}{n}}.$$

Since $\mathbb{E}[W(\mu_n, \mathbf{c}_k)] = W(\mu, \mathbf{c}_k)$,

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \le W(\mu, \mathbf{c}_k) + \mathrm{pen}(k) + R^2\Sigma\sqrt{\frac{2\pi}{n}}.$$

As this is true for every k,

$$\mathbb{E}\left[W(\mu, \tilde{\mathbf{c}})\right] \le \inf_{1\le k\le n}(W_k^\star(\mu) + \mathrm{pen}(k)) + R^2\Sigma\sqrt{\frac{2\pi}{n}},$$

where $W_k^\star(\mu) = \inf_{\mathbf{c}\in S_k} W(\mu, \mathbf{c})$. This completes the proof of the theorem.

Acknowledgement

The author wishes to thank Audrey Proust for having provided the data on dyslexia.

References

[1] B. Al-Shboul and S.-H. Myaeng. Initializing k-means using genetic algorithms. *World Academy of Science, Engineering and Technology*, 54:114–118, 2009.

[2] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

[3] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

[4] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 2011. In press. Available at `http://hal.archives-ouvertes.fr/docs/00/46/16/39/PDF/RR-7223.pdf`.

[5] S. Ben-David, D. Pál, and H. U. Simon. Stability of k-means clustering. In N. Bshouty and C. Gentile, editors, *Proceedings of the 20th Annual Conference on Learning Theory (COLT 2007)*, pages 20–34. Springer, 2007.

[6] S. Ben-David and U. von Luxburg. Relating clustering stability to properties of cluster boundaries. In R. A. Servedio and T. Zhang, editors, *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 379–390, Madison, 2008. Omnipress.

[7] S. Ben-David, U. von Luxburg, and D. Pál. A sober look on clustering stability. In G. Lugosi and H. U. Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006)*, pages 5–19, Berlin, 2006. Springer.

[8] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Proceedings of the 7th Pacific Symposium on Biocomputing*, volume 7, pages 6–17, 2002.

[9] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

[10] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138:33–73, 2007.

[11] R. B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[12] M. B. Denckla and R. G. Rudel. Rapid "automatized" naming (R.A.N.): dyslexia differentiated from other learning disabilities. *Neuropsychologia*, 14:471–479, 1976.

[13] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2000.

[14] A. Frank and A. Asuncion. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2010. `http://archive.ics.uci.edu/ml`.

[15] A. D. Gordon. *Classification*, volume 82 of *Monographs on Statistics and Applied Probability*. Chapman Hall/CRC, Boca Raton, 1999.

[16] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions.* Lecture Notes in Mathematics. Springer-Verlag, Berlin, Heidelberg, 2000.

[17] A. Hardy. On the number of clusters. *Computational Statistics and Data Analysis*, 23:83–96, 1996.

[18] J. A. Hartigan. *Clustering Algorithms.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York, 1975.

[19] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, New York, 2001.

[20] L. Kaufman and P. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, Hoboken, 1990.

[21] S. S. Khan and A. Ahmad. Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, 25:1293–1302, 2004.

[22] D. J. Kim, Y. W. Park, and D. J. Park. A novel validity index for determination of the optimal number of clusters. *IEICE Transactions on Information and System*, E84D:281–285, 2001.

[23] W. J. Krzanowski and Y. T. Lai. A criterion for determining the number of clusters in a data set. *Biometrics*, 44:23–34, 1985.

[24] E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Journal of Neural Computation*, 13:2573–2593, 2002.

[25] T. Linder. On the training distortion of vector quantizers. *IEEE Transactions on Information Theory*, 46:1617–1623, 2000.

[26] T. Linder. Learning-theoretic methods in vector quantization. In L. Györfi, editor, *Principles of Nonparametric Learning.* Springer-Verlag, Wien, 2002.

[27] H. Luschgy and G. Pagès. Functional quantization rate and mean regularity of processes with an application to Levy processes. *The Annals of Applied Probability*, 18:427–469, 2008.

[28] P. Massart. *Concentration Inequalities and Model Selection.* Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003, Lecture Notes in Mathematics. Springer, Berlin, Heidelberg, 2007.

[29] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge, 1989.

[30] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–79, 1985.

[31] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B Ford. The population biology of Abalone (Haliotis species) in Tasmania. 1, Blacklip Abalone (H. rubra) from the north coast and islands of Bass Strait. Technical Report 48, Sea Fisheries Division, 1994.

[32] J. M. Pena, J. A. Lozano, and P. Larranaga. An empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognition Letters*, 20:1027–1040, 1999.

[33] G. T. Perim, E. D. Wandekokem, and F. M. Varejão. K-means initialization methods for improving clustering by simulated annealing. In *Advances in artificial intelligence – Iberamia 2008*, volume 5290, pages 133–142. Springer-Verlag, Berlin, Heidelberg, 2008.

[34] J. N. Pierce. Asymptotic quantizing error for unbounded random variables. *IEEE Transactions on Information Theory*, 16:81–83, 1970.

[35] O. Shamir and N. Tishby. Cluster stability for finite samples. In J. C. Platt, D. Koller, Y. Singer, and S. Rowseis, editors, *Advances in Neural Information Processing Systems 20*, pages 1297–1304, Cambridge, 2008. MIT Press.

[36] O. Shamir and N. Tishby. Model selection and stability in k-means clustering. In R. A. Servedio and T. Zhang, editors, *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 367–378, Madison, 2008. Omnipress.

[37] T. Su and J. Dy. A deterministic method for initializing k-means clustering. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, volume 141, pages 784–786, 2004.

[38] C. A. Sugar and G. M. James. Finding the number of clusters in a data set: an information theoretic approach. *Journal of the American Statistical Association*, 98:750–763, 2003.

[39] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423, 2001.