INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Choosing Models in Model-based Clustering and Discriminant Analysis

Christophe Biernacki, Gérard Govaert

No 3509

Octobre 1998

THÈME 4

Rapport de recherche

ISSN 0249-6399

Choosing Models in Model-based Clustering and Discriminant Analysis

Christophe Biernacki, Gérard Govaert

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet IS2

Rapport de recherche n° 3509 — Octobre 1998 — 22 pages

Abstract: Using an eigenvalue decomposition of variance matrices, Celeux and Govaert (1993) obtained numerous and powerful models for Gaussian model-based clustering and discriminant analysis. Through Monte Carlo simulations, we compare the performances of many classical criteria to select these models: information criteria as AIC, the Bayesian criterion BIC, classification criteria as NEC and cross-validation. In the clustering context, information criteria and BIC outperform the classification criteria. In the discriminant analysis context, cross-validation shows good performance but information criteria and BIC give satisfactory results as well with, by far, less time-computing.

Key-words: Gaussian mixture models, eigenvalue decomposition, cross-validation, information, Bayesian and classification criteria

(Résumé : tsvp)

Collaboration avec le laboratoire HEUDIASYC UMR CNRS 6599, Université de Technologie de Compiègne, BP 529, 60205 Compiègne Cedex, France.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN (France)
Téléphone : 04 76 61 52 00 - International: +33 4 76 61 52 00
Télécopie : 04 76 61 52 52 - International: +33 4 76 61 52 52

Choix de modèles en classification automatique et en discrimination

Résumé : Par le moyen d'une décomposition spectrale des matrices de variance, Celeux et Govaert (1993) ont obtenus de nombreux modèles très utiles pour la classification automatique ou la discrimination lorsque ces méthodes reposent sur des modèles de mélange gaussiens. De nombreuses simulations de Monte-Carlo nous permettent de comparer la performance de plusieurs critères pour choisir ces modèles : des critères d'information comme AIC, le critère bayésien BIC, des critères de classification comme NEC et le critère de validation croisée. En classification automatique, les critères d'information et BIC donnent de meilleurs résultats que les critères de classification. En discrimination, le critère de validation croisée a un bon comportement mais les critères d'information et BIC donnent aussi de bons résultats avec beaucoup moins de calculs.

Mots-clé : Modèles de mélange gaussiens, décomposition spectrale, validation croisée, critères d'information, critères de classification

1 Introduction

Finite multivariate Gaussian mixture distributions lead to commonly used models for multivariate data analysis and statistical pattern recognition (see for instance McLachlan 1992 and Ripley 1996). Recently several authors have exploited the eigenvalue decomposition of the group variance matrices in Gaussian mixtures to propose numerous and powerful models for clustering (Banfield and Raftery 1993, Celeux and Govaert 1995, Bensmail, Celeux, Raftery and Robert 1997) and discriminant analysis Flury, Schmid and Narayanan 1993, Bensmail and Celeux 1996). This parametrization of the mixture components provides a general and flexible framework to give raise to efficient, although somewhat unusual, clustering criteria and classification rules. It consists in writing the variance matrix Σ_k in terms of its eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{1.1}$$

where $\lambda_k = |\Sigma_k|^{1/d}, d$ denoting the number of variables, D_k the matrix of eigenvectors of Σ_k and A_k a diagonal matrix, such that $|A_k| = 1$, with the normalized eigenvalues of Σ_k on the diagonal in a decreasing order.

The parameter λ_k determines the volume of the kth group, D_k its orientation and A_k its shape. By allowing some but not all of these quantities to vary between groups, we obtain parsimonious and easily interpreted models which are appropriate to describe various clustering or classification situations. For instance Celeux and Govaert (1995) and Bensmail and Celeux (1996) considered 14 different models related to different assumptions on the group variance matrices. Eight of these models are obtained by assuming equal or different volumes, shapes or orientations ($[\lambda DAD']$, $[\lambda_k DAD']$, $[\lambda DA_kD']$, $[\lambda_k DA_kD']$, $[\lambda D_kAD_k']$, $[\lambda_k D_kAD_k']$, $[\lambda D_kA_kD_k']$, and $[\lambda_k D_kA_kD_k']$.) We use the following convention: writing, for instance, $[\lambda DA_kD']$ means that we consider a mixture model with equal volumes, equal orientations, and different shapes. Four models assume diagonal variance matrices, we denoted them by $[\lambda B]$, $[\lambda_k B]$, $[\lambda B_k]$, $[\lambda_k B_k]$ with $|B| = 1$ or $|B_k| = 1$, and two models assume spherical shapes $[\lambda I]$, $[\lambda_k I]$, where I denotes the identity matrix.

In this framework, selecting a relevant and parsimonious model is a difficult task of crucial importance. In this paper, we review different approaches of model selection in cluster analysis and in discriminant analysis. For both situations, we report Monte Carlo numerical experiments to illustrate the performance of the considered approaches.

In the cluster analysis context, we propose to compare the performance of information criteria, criteria derived from approximations of the integrated likelihood and classification criteria for choosing the model producing the lowest empirical error rate.

In the discriminant analysis context, we compare the performance of the cross-validation procedure to information and Bayesian criteria for selecting the model producing the lowest error rate in a small sample setting.

2 Choosing a mixture model in cluster analysis

In the multivariate Gaussian mixture model, data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in $\mathbb{R}^d$ are assumed to be a sample from a probability distribution with density

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} p_k \phi(\boldsymbol{x}, \boldsymbol{a}_k) \tag{2.1}$$

where the p_k's are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \ldots, K$ and $\sum_k p_k = 1$) and $\phi(\boldsymbol{x}, \boldsymbol{a}_k)$ denotes the d-dimensional Gaussian density with mean $\boldsymbol{\mu}_k$ and variance matrix Σ_k with $\boldsymbol{a}_k = (\boldsymbol{\mu}_k, \Sigma_k)$. In what follows, the number of clusters is assumed to be known and the variance matrices Σ_k are supposed to be modelized according to one of the 14 models described in the introduction. The maximized log likelihood of $((p_1, \boldsymbol{a}_1), \ldots, (p_K, \boldsymbol{a}_K))$ for the sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is denoted

$$L(M) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \hat{p}_k \phi(\boldsymbol{x}_i, \hat{\boldsymbol{a}}_k) \right], \tag{2.2}$$

with $\hat{p}_k$ and $\hat{\boldsymbol{a}}_k$ denoting the maximum likelihood estimates of the corresponding parameters. In this formula, M is one of 28 models: namely, the 14 models previously defined with equal or different mixing proportions.

Various criteria to be minimized have been proposed to measure a model's suitability by balancing model fit and model complexity. In this paper, we only mention those criteria that we have experimented: information criteria (AIC, AIC3, ICOMP), a Bayesian information criterion (BIC) and classification criteria (entropy, NEC, fuzzy classification likelihood, classification likelihood). Other approaches to the problem of assessing mixture models, including their resampling approach, are cited in McLachlan and Peel (1996).

2.1 Information criteria

The *Akaike information criterion* (Akaike 1974) takes the form

$$\mathrm{AIC}(M) = -2L(M) + 2\nu(M), \tag{2.3}$$

where $\nu(M)$ is the number of free parameters in the mixture model M. Because the regularity conditions on which the AIC criterion relies do no hold when the likelihood ratio λ is designed to contrast two hypotheses on the number of components (see for instance Aitkin and Rubin 1985), Bozdogan (1987) proposed to use the approximation to the null distribution of $-2\log(\lambda)$ given by Wolfe (1971). It leads to a modified AIC criterion

$$\mathrm{AIC3}(M) = -2L(M) + 3\nu(M). \tag{2.4}$$

In this paper, only the Gaussian model has to be selected since the number of clusters is fixed, and, so, AIC3 has no justification in our context. For choosing parsimonious models, Bozdogan (1990) proposed also an *informational complexity criterion*,

$$\mathrm{ICOMP}(M) = -2L(M) + \frac{\nu(M)}{2} \log \frac{\mathrm{tr} F^{-1}}{\nu(M)} - \frac{1}{2} \log |F^{-1}|, \tag{2.5}$$

where F is the Fisher information matrix of the model. Thus, when measuring the complexity of a model with ICOMP, there is a need to approximate F. The calculation of F depends on the parametrization of the model and can be difficult. In our experiments, following Cutler and Windham (1993), we approximate F with its empirical mean given by

$$\hat{F} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x_i})) \bigg|_{\hat{\boldsymbol{\theta}}} \cdot \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{x_i})) \bigg|_{\hat{\boldsymbol{\theta}}}' \right),$$

where $\boldsymbol{\theta}$ is the vector of parameters $((p_1, \boldsymbol{a}_1), \ldots, (p_K, \boldsymbol{a}_K))$ of M.

2.2 A Bayesian information criterion

In a fully Bayesian inference for Gaussian mixture models, a simple way to determine the appropriate model is to calculate the integrated likelihood (Kass and Raftery 1995). Integrated likelihood of the data $\boldsymbol{d} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ given the model M is

$$\Pr(\boldsymbol{d}|M) = \int \Pr(\boldsymbol{d}|M, \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}|M) \mathrm{d}\boldsymbol{\theta}, \tag{2.6}$$

where $\Pr(\boldsymbol{\theta}|M)$ is the prior density of $\boldsymbol{\theta}$. A classical way to approximate the integrated likelihood consists in using the *Bayesian information criterion* (Schwarz 1978). Noting $\hat{\boldsymbol{\theta}}$ the maximum likelihood estimate of $\boldsymbol{\theta}$, this approximation is

$$\log \Pr(\boldsymbol{d}|M) = \log \Pr(\boldsymbol{d}|M, \hat{\boldsymbol{\theta}}) - \frac{\nu(M)}{2} \log n + O(1). \tag{2.7}$$

Thus the Bayesian information criterion (BIC) is given by

$$\mathrm{BIC}(M) = -2L(M) + \nu(M) \log n. \tag{2.8}$$

2.3 Classification criteria

The classification criteria we propose in this section measure the ability of a mixture model to provide well-separated clusters. They are derived from a relation emphasizing the differences between the likelihood and the “fuzzy” classification likelihood of the mixture (Hathaway 1986) or, in the same manner, between the likelihood and the classification likelihood of the mixture (Biernacki and Govaert 1997). Let

$$t_{ik} = \frac{\hat{p}_k \phi(\boldsymbol{x}_i, \hat{\boldsymbol{a}}_k)}{\sum_{j=1}^{K} \hat{p}_j \phi(\boldsymbol{x}_i, \hat{\boldsymbol{a}}_j)}$$

be the estimated conditional probability that $\boldsymbol{x}_i$ arises from the kth mixture component ($1 \leq i \leq n$ and $1 \leq k \leq K$). Direct calculations show that

$$\mathrm{C}(M) = L(M) - \mathrm{E}(M), \tag{2.9}$$

with the fuzzy classification likelihood

$$\mathrm{C}(M) = \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \log[\hat{p}_k \phi(\boldsymbol{x}_i, \hat{\boldsymbol{a}}_k)],$$

and the entropy term

$$\mathrm{E}(M) = -\sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} \log t_{ik} \geq 0.$$

$\mathrm{C}(M)$ is related to the fuzzy classification matrix $\boldsymbol{t} = \{t_{ik}\}$. If the mixture components are well-separated, the classification matrix $\boldsymbol{t}$ tends to define a partition of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $\mathrm{E}(M) \approx 0$. But if the mixture components are not well-separated, $\mathrm{E}(M)$ has a large value. Thus, $\mathrm{E}(M)$ can be regarded as a measure of the ability of the K-component mixture model to provide a relevant partition of the data $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Relation (2.9) shows that the classification likelihood term $\mathrm{C}(M)$ can be regarded as a compromise between the fit of the data to the mixture model, measured with the log likelihood $L(M)$, and the ability of the mixture model to provide a classification in well-separated clusters, measured with the entropy term $\mathrm{E}(M)$.

As a consequence, the entropy of the classification matrix $\boldsymbol{t}$ gives raise to several classification criteria (see Celeux and Soromenho 1996, Biernacki 1997) which are $\mathrm{E}(M)$, its normalized version

$$\mathrm{NEC}(M) = \mathrm{E}(M)/[L(M) - L_1(M)],$$

where $L_1(M)$ denotes the maximized likelihood for a single Gaussian distribution and $\mathrm{C}(M)$ the fuzzy classification likelihood.

A relation between the likelihood and the classical classification likelihood exists in the same manner (Biernacki and Govaert 1997). Direct calculations show that

$$\mathrm{CLM}(M) = L(M) - \mathrm{EC}(M), \tag{2.10}$$

with the classification likelihood

$$\mathrm{CLM}(M) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \log[\hat{p}_k \phi(\boldsymbol{x}_i, \hat{\boldsymbol{a}}_k)],$$

and a kind of entropy term

$$\mathrm{EC}(M) = -\sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} \log t_{ik} \geq 0,$$

where $z_{ik} = 1$ if $\arg\max_\ell t_{i\ell} = k$ and 0 otherwise. The behaviour of EC(M) is analogous to the behaviour of E(M) since EC measures the cluster overlapping and EC(M) ≈ 0 if clusters are well-separated. Relation (2.10) shows that CLM, as C, makes a compromise between the fit of the data and the ability of the mixture model to provide a classification. So, CLM(M) and EC(M) are two other classification criteria of interest (Biernacki 1997, Biernacki and Govaert 1997).

3 Choosing a discriminant analysis model

In the discriminant analysis context, the partition and the number of classes K are known. A natural way to choose a model is to select the model that minimizes the sample based estimate of future misclassification risk by cross-validation. This is done in Bensmail and Celeux (1996) to choose a model among the 14 models mentioned in the introduction.

It is worth noting that in many circumstances several models provide exactly the same cross-validated misclassification rate. In such cases, several strategies are possible and in the present paper we investigated two strategies. The first one consists in selecting the most parsimonious model (i.e. the model for which the number of parameters is the smallest); we denoted this strategy by CV$-$. The second one consists in selecting the most complex model for which the number of parameters to be estimated is the greatest. This strategy is denoted by CV+.

However, it can happen that, for some specific values of K and d, different models have the same number of parameters. In such cases, we complete the CV$-$ strategy in the following manner: at first, a spherical model is preferred to a diagonal model which is preferred to a non diagonal model; secondly, a model with different volumes is preferred to a model with different shapes which is preferred to a model with different orientations. For the strategy CV+, we proceed exactly in the opposite way.

But, it appears that cross-validation procedures are painfully slow, even if it is generally possible to reduce the calculations when computing the cross-validated classification rules (see Appendix A for details). Then, it would be of interest to use one of the criteria presented in Section 2. Thus, we experimented with the criteria AIC, AIC3, BIC in comparison with the cross-validation criteria CV$-$ and CV+.

4 Numerical experiments

4.1 Choosing a clustering model

We assessed the practical ability of the criteria L (the log likelihood), AIC, AIC3, BIC, ICOMP, NEC, EC, E, CLM, and C to choose a model when the number of clusters K is known. We simulated two-component bivariate Gaussian mixtures with different variance matrices. The variance matrices were determined according to the 14 models based on their eigenvalue decomposition as described in Table 1.

Moreover, two kinds of mixing proportions have been choosen (Table 2). The centres of the two components were $\boldsymbol{\mu}_1 = (0,0)'$ and $\boldsymbol{\mu}_2 = (t,0)'$. Defining the optimal misclassification rate as the misclassification rate obtained with the true parameter $\boldsymbol{\theta}$, we choose the value t to get three degrees of overlapping: a small one corresponding to an optimal misclassification rate of 5%, a medium one corresponding to an optimal misclassification rate of 15% and a large one corresponding to an optimal misclassification rate of 30%. Note that for some quite different variance matrices, we were unable to reach the large misclassification rate, and finally we get 73 different mixtures models.

For each of those 73 models, we simulated 30 times a sample of size $n = 40$ and 30 times a sample of larger size $n = 200$. Then the EM algorihm (Dempster *et al.* 1977) is started with the true underlying centres for each model in turn. Figures 1 and 2 summarize the results. They give, for each criterion, the histogram of the ratio of the minimum empirical misclassification rate obtained with one of the 28 models over this misclassification rate obtained with the model chosen by the criterion. This ratio takes value between zero and one, and its ideal value is one. Moreover, Tables 3 and 4 give the mean number of parameters in the mixture model selected by the criteria which has to be compared with the mean number 8.14 of the true underlying model. Note that the mean number of parameters of the models selected by L is little less than the expected number of parameters of the most complex model (11 parameters) because a suboptimal solution may be found by the EM algorithm since this one is started only once with the true centres. In these tables, we also mentioned the mean number of parameters obtained when using the empirical misclassification rate (column mis. r.) as a criterion of selection.

In the small sample size case ($n = 40$), each criterion gives similar low performance (less than 6.5) with an especially bad performance for the classification criteria NEC, EC and E (ratio less than 6.0). Moreover, it is very amazing that the likelihood L criterion leads to the best performance since it theoretically selects only the most complex model.

All criteria improve their performance with the larger sample size $n = 200$. The mean number of parameters for the selected model increases for information criteria and the BIC criterion whereas it decreases for the classification criteria. It appears that the best results are obtained with the criteria AIC3 and AIC and that the information criteria and BIC outperformed the classification criteria. It is somewhat amazing that AIC3 gives the best performance since, in the context where K is known, it has no theoretical justification.

4.2 Choosing a discriminant analysis model

In a discriminant analysis context we simulated the same models than in section 4.1. The only differences are that we only consider equal proportion models and that for each situation, we consider two small sample sizes $n = 10$ and $n = 50$ instead of $n = 40$ and $n = 200$. As before, each situation (37 models with $n = 10$ and 37 models with $n = 50$) was simulated 30 times. The compared criteria were L, AIC, AIC3, BIC, CV$-$ and CV$+$. Figures 3 and 4 give, for each criterion, the histogram of the ratio of the minimum misclassification rate obtained with one of the 14 models over this misclassification rate obtained with the model

volumes	
equal $[\lambda]$	$\lambda_1 = 1$, $\lambda_2 = 1$
different $[\lambda_k]$	$\lambda_1 = 1$, $\lambda_2 = 3$

	shapes and orientations						
	$[I]$	$[B]$	$[B_k]$	$[C]$	$[C_k]$	$[DA_kD']$	$[D_kAD'_k]$
α_1	1	2	2	2	2	2	2
α_2	1	2	4	2	4	4	2
δ_1	-	90	90	45	45	45	45
δ_2	-	90	0	45	-45	45	-45

Table 1: Variance matrices related to the 14 simulated models. α_k is the first diagonal term of the shape matrix A_k and δ_k is the angle in degrees of the rotation matrix D_k.

proportions	
equal $[p]$	$p_1 = 0.5$, $p_2 = 0.5$
different $[p_k]$	$p_1 = 0.3$, $p_2 = 0.7$

Table 2: Proportions related to the simulated models.

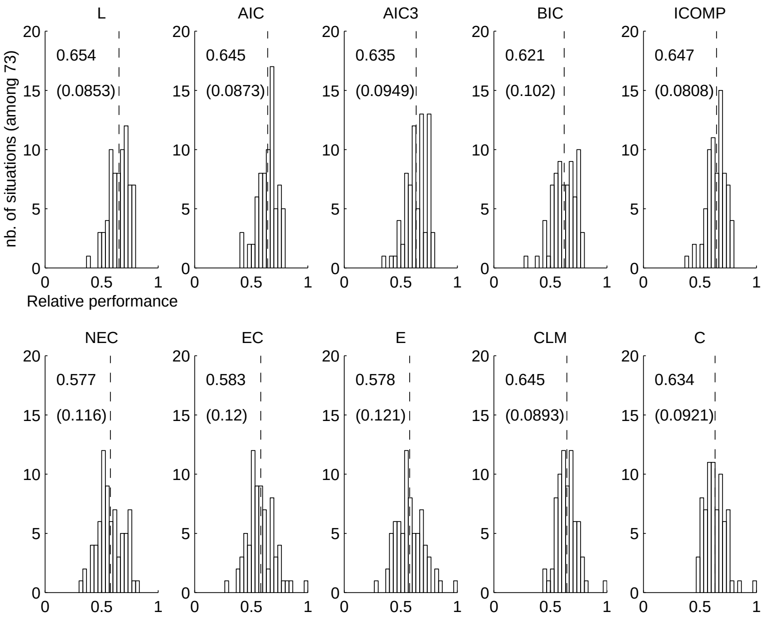

Figure 1: For each criterion, histogram, mean value, and, in parentheses, standard deviation of the ratio of the minimum misclassification rate over the misclassification rate of the model chosen by the criterion with the sample size $n = 40$.

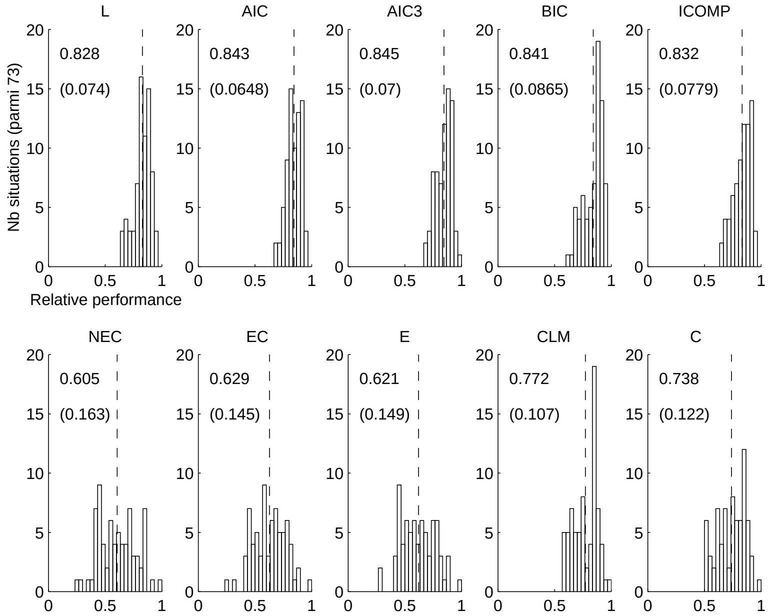

Figure 2: For each criterion, histogram, mean value, and, in parentheses, standard deviation of the ratio of the minimum misclassification rate over the misclassification rate of the model chosen by the criterion with the sample size $n = 200$.

mis. r.	L	AIC	AIC3	BIC	ICOMP	NEC	EC	E	CLM	C
7.90	10.59	8.16	7.56	7.19	8.65	8.34	8.59	8.61	10.03	9.88

Table 3: For each criterion, the mean number of parameters for the selected model with the sample size $n = 40$. The mean number of parameters for the true underlying model is 8.14.

mis. r.	L	AIC	AIC3	BIC	ICOMP	NEC	EC	E	CLM	C
8.60	10.72	8.40	8.04	7.67	9.08	7.95	8.46	8.50	9.67	9.47

Table 4: For each criterion, the mean number of parameters for the selected model with the sample size $n = 200$. The mean number of parameters for the true underlying model is 8.14.

chosen by the criterion. This ratio takes value between zero and one, and its ideal value is one. Tables 5 and 6 summarize the results in the same way as Tables 3 and 4.

Those simulation experiments show that for the very small sample size $n = 10$ the cross-validation criteria can be preferred to information criteria. Moreover, the parsimonious cross-validation strategy CV$-$ outperforms the strategy CV$+$.

For $n = 50$, information criteria now outperform the two cross-validation strategies. The best results are obtained with AIC3 and the same remark concerning the lack of justification of this criterion is in order... Maybe the BIC criterion can be preferred since its performance is just behind AIC3 and its justification holds in this context. Moreover, all the criteria, except L, prefer a more complex model in comparison to the smaller sample size $n = 10$. It is interesting also to note that the strategy CV$+$, which selects quite complex models, give best performance that the more parsimonious criterion CV$-$.

We simulated also 30 samples of size $n = 60$ from two quite overlapping clusters in high dimension ($d = 10$):

$$p_1 = p_2 = 0.5,\ \boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},\ \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},\ \Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 4 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 6 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 7 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 8 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 10 \end{pmatrix}.$$

Ratio of the minimum misclassification rate over this misclassification rate obtained with the selected model and also mean number of parameters of this model are diplayed in Table 7 for previous criteria. The good behaviour of CV$-$ is strengthened by this numerical experiment and the information criteria and the BIC criteria confirm a good performance as well.

We considered also a model with a non-Gaussian class in dimension two. The first class is Gaussian with mixing proportion 0.5, center $(2, 0)'$ and variance matrix identity whereas the second class is a mixture of two Gaussian distributions. Parameters of this two component mixture are: equal mixing proportions, same center at $(0, 0)'$, variance matrices equal to $\mathrm{diag}(0.25, 4)$ and $\mathrm{diag}(4, 0.25)$. Figure 5 displays isodensity curve of the three Gaussian distributions and the optimal classification boundary as well. Overlapping of the Gaussian and the non-Gaussian class is moderate.

As before, 30 samples of size $n = 10$ and 30 samples of size $n = 50$ are generated and all the criteria are computed on each sample for the 14 models. Results are displayed in Tables 8 and 9 in the same way as Table 7.

The CV$-$ criterion outperforms all other criteria for both sample sizes. Contrary to previous simulations, mean number of parameters for the selected models for information criteria and BIC decreases with the sample size.

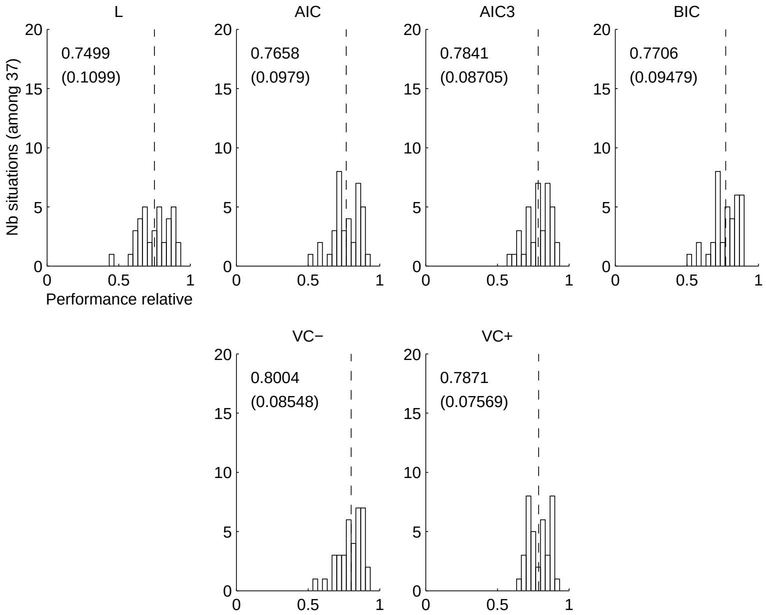

Figure 3: For each criterion, histogram, mean value, and, in parentheses, standard deviation of the ratio of the minimum misclassification rate over the misclassification rate of the model chosen by the criterion with the sample size $n = 10$.

mis. r.	L	AIC	AIC3	BIC	CV−	CV+
7.33	10.00	7.49	6.91	7.32	6.10	8.14

Table 5: For each criterion, the mean number of parameters for the selected model with sample size $n = 10$. The mean number of parameters for the true model is 7.54.

mis. r.	L	AIC	AIC3	BIC	CV−	CV+
8.11	10.00	7.84	7.59	7.45	7.06	8.84

Table 6: For each criterion, the mean number of parameters for the selected model with sample size $n = 50$. The mean number of parameters for the true model is 7.54.

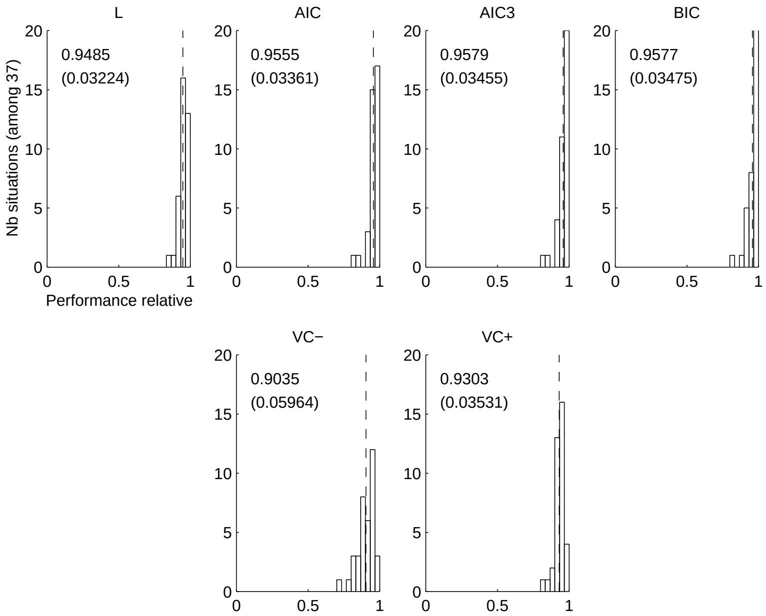

Figure 4: For each criterion, histogram, mean value, and, in parentheses, standard deviation of the ratio of the minimum misclassification rate over the misclassification rate of the model chosen by the criterion with the sample size $n = 50$.

	mis. r.	L	AIC	AIC3	BIC	CV−	CV+
perf.	1.000	0.972	0.985	0.994	0.996	0.997	0.972
nb. par.	87.00	130.00	80.70	76.63	75.40	75.00	129.93

Table 7: Ratio of the minimum misclassification rate over this misclassification rate obtained with the selected model (perf.) and mean number of parameters for this model (nb. par.) with samples in dimension $d = 10$.

	mis. r.	L	AIC	AIC3	BIC	CV−	CV+
perf.	1.000	0.685	0.723	0.758	0.751	0.832	0.799
nb. par.	7.13	10.00	7.03	6.50	6.70	5.96	7.73

Table 8: Ratio of the minimum misclassification rate over this misclassification rate obtained with the selected model (perf.) and mean number of parameters for this model (nb. par.) with non-Gaussian class samples of size $n = 10$.

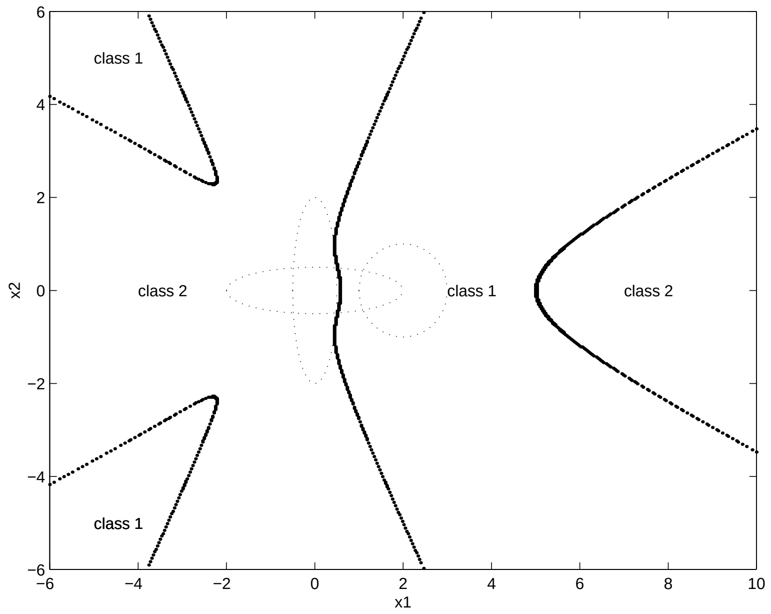

Figure 5: Isodensity curves and optimal classification boundary in the case of the non-Gaussian class.

	mis. r.	L	AIC	AIC3	BIC	CV−	CV+
perf.	1.000	0.882	0.881	0.882	0.888	0.918	0.899
nb. par.	6.96	10.00	6.76	6.23	6.10	6.43	7.76

Table 9: Ratio of the minimum misclassification rate over this misclassification rate obtained with the selected model (perf.) and mean number of parameters for this model (nb. par.) with non-Gaussian class samples of size $n = 50$.

5 Discussion

We compared many criteria in the ability to choose a Gaussian model in both the clustering and the discriminant analysis context. It emerges the following remarks from this study.

In the cluster analysis context, information criteria as AIC3 and also BIC criterion show reasonable performances to choose a good mixture model when the number of groups K is known and when the sample size is moderate. Nevertheless, the BIC criterion has to be preferred to AIC3 since it is more justified in this context. But, both these criteria are known to overestimate the number of clusters when the model is fixed (Biernacki and Govaert, 1997). On the contrary, although often having a good behaviour to detect the number of clusters, classification criteria C, CLM and NEC give poor results to select one of the considered model. Such results are somewhat disappointing to recommand one criterion in order to select both the number of clusters and the Gaussian model.

In the discriminant analysis context, results are much more encouraging. The cross-validation criterion has been shown to be a good way to choose a model especially in a small sample setting. But simple information criteria as AIC3 or the simple approximation of the integrated likelihood BIC can be regarded as advantageous alternatives to it in a large or even moderate sample size setting. We have to remind that, in the case of large sample size, the cross-validation criterion is a very time-consuming criterion. Moreover, as in the clustering context, the BIC criterion has to be preferred to the AIC3 criterion from a theoretical point of view.

Acknowlegments We are indebted to Van Mô Dang for its contribution in the cluster analysis studies.

A Reducing calculations for cross-validation

With a sample $((\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_n, z_n))$, the label z_i being equal to k if $\boldsymbol{x}_i$ belongs to the kth cluster, the cross-validation criterion of a model M is given by

$$\mathrm{CV}(M) = \frac{1}{n} \sum_{i=1}^{n} c(r^{(i)}(\boldsymbol{x}_i), z_i),$$

where $\delta(a,b)$ is the 0-1 cost and $r^{(i)}$ is the "plug-in" discriminant rule obtained from the whole sample without the element $(\boldsymbol{x}_i, z_i)$. Calculation of these n discriminant rules is time-consuming and we establish, for some models M, updating formulae of $r^{(i)}$ from r (the rule obtained from the whole sample).

The decision rule r is entirely determined with the mixing proportions $\hat{p}_k$, the centers $\hat{\boldsymbol{\mu}}_k$, the determinants $|\hat{\Sigma}_k|$ and the inverse matrices $\hat{\Sigma}_k^{-1}$. For eight models ($[\lambda_k I]$, $[\lambda_k B_k]$, $[\lambda_k D_k A_k D_k']$, $[\lambda I]$, $[\lambda B]$, $[\lambda B_k]$, $[\lambda DAD']$ and $[\lambda D_k A_k D_k']$), updated determinants and inverse of variance matrices can be directly computed from some updated terms: the sample size, the kth cluster size n_k, the within cluster scattering matrix $W_k = \sum_{i=1}^{n} \sum_{z_i=k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'$ of each cluster, their determinant, inverse and trace and same characteristics for the scattering matrix $W = \sum_{k=1}^{K} W_k$. The six remaining models use an iterative procedure to compute the variance matrix and, so, their is no way to simplify calculation although the iterative procedure needs also some of the previous updated terms.

Tables 10 and 11 give the updating formula for the terms to be calculated. Note that a slightly more general situation is taken into account: a point $\boldsymbol{y}$ is *added* ($\epsilon = +1$) to or is *removed* ($\epsilon = -1$) from the cluster t ($t = 1, \ldots, K$). The formulae are proved in Section B of Appendix.

Term to be updated	Updating formula $k = t$	Updating formula $k \neq t$
n_k	$n_t + \epsilon$	unchanged
equal $\hat{p}_k$	unchanged	unchanged
different $\hat{p}_k$	$(n_t + \epsilon)/(n + \epsilon)$	$n_k/(n + \epsilon)$
$\hat{\boldsymbol{\mu}}_k$	$(n_t\hat{\boldsymbol{\mu}}_t + \epsilon\boldsymbol{y})/(n_t + \epsilon)$	unchanged
W_k	$W_t + \epsilon\boldsymbol{h}_t\boldsymbol{h}_t'$	unchanged
$\lvert W_k \rvert$	$\lvert W_t \rvert(1 + \epsilon\boldsymbol{h}_t'W_t^{-1}\boldsymbol{h}_t)$	unchanged
W_k^{-1}	$W_t^{-1} - \{\epsilon(W_t^{-1}\boldsymbol{h}_t)(W_t^{-1}\boldsymbol{h}_t)'\}/(1 + \epsilon\boldsymbol{h}_t'W_t^{-1}\boldsymbol{h}_t)$	unchanged
$\mathrm{tr}(W_k)$	$\mathrm{tr}(W_t) + \epsilon\boldsymbol{h}_t'\boldsymbol{h}_t$	unchanged

Table 10: Updating formulae of sizes, mixing proportions, centres, scattering matrices, their determinant, inverse and trace. We noted $\boldsymbol{h}_t = \omega_t(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_t)$ and $\omega_t^2 = \frac{n_t}{n_t+\epsilon}$.

Term to be updated	Updating formula
n	$n + \epsilon$
W	$W + \epsilon\boldsymbol{h}_t\boldsymbol{h}_t'$
$\lvert W \rvert$	$\lvert W \rvert(1 + \epsilon\boldsymbol{h}_t'W^{-1}\boldsymbol{h}_t)$
W^{-1}	$W^{-1} - \{\epsilon(W^{-1}\boldsymbol{h}_t)(W^{-1}\boldsymbol{h}_t)'\}/(1 + \epsilon\boldsymbol{h}_t'W^{-1}\boldsymbol{h}_t)$
$\mathrm{tr}(W)$	$\mathrm{tr}(W) + \epsilon\boldsymbol{h}_t'\boldsymbol{h}_t$

Table 11: Updating formulae of the sample size, the within cluster scattering matrix, its determinant, inverse and trace. We noted $\boldsymbol{h}_t = \omega_t(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_t)$ and $\omega_t^2 = \frac{n_t}{n_t+\epsilon}$.

B Proof of updating formulae

The aim of this section is to detail how updating formulae displayed in Table 10 have been obtained. Results displayed in Table 11 are simply deduced from these formulae.

Let E be a set of n points $\boldsymbol{x}$ of $\mathbb{R}^d$ with center $\boldsymbol{g} = \sum_{x \in E} \boldsymbol{x}/n$ and within scattering matrix $Q = \sum_{x \in E} (\boldsymbol{x} - \boldsymbol{g})(\boldsymbol{x} - \boldsymbol{g})'$. Adding (respectively removing) a point $\boldsymbol{y} \in R^d$ to (respectively from) the set E gives a new set E^* of sample size $n^* = n + \epsilon$ with $\epsilon = +1$ (respectively $\epsilon = -1$). Noting g^* the center and Q^* the within scattering matrix of E^*, relations between g^*, Q^*, Q^{*-1}, $|Q^*|$, $\mathrm{tr}(Q^*)$ and g, Q, Q^{-1}, $|Q|$, $\mathrm{tr}(Q)$ are now proved.

B.1 Updating the center

Proposition B.1 *The center $\boldsymbol{g}^*$ of E^* is given by*

$$\boldsymbol{g}^* = \frac{n\boldsymbol{g} + \epsilon \boldsymbol{y}}{n + \epsilon}.$$

Proof The definition of the center directly gives:

$$\boldsymbol{g}^* = \frac{\sum_{x \in E^*} \boldsymbol{x}}{n + \epsilon} = \frac{\sum_{x \in E} \boldsymbol{x} + \epsilon \boldsymbol{y}}{n + \epsilon} = \frac{n\boldsymbol{g} + \epsilon \boldsymbol{y}}{n + \epsilon}.$$

B.2 Updating the within scattering matrix

Proposition B.2 *The within scattering matrix Q^* of E^* is given by*

$$Q^* = Q + \epsilon \boldsymbol{h}\boldsymbol{h}'$$

with $\boldsymbol{h} = \omega(\boldsymbol{y} - \boldsymbol{g})$ and $\omega^2 = \frac{n}{n+\epsilon}$.

Proof We have

$$\begin{aligned}
Q^* &= \sum_{x \in E^*} (\boldsymbol{x} - \boldsymbol{g}^*)(\boldsymbol{x} - \boldsymbol{g}^*)' \\
&= \sum_{x \in E} (\boldsymbol{x} - \boldsymbol{g} + \boldsymbol{g} - \boldsymbol{g}^*)(\boldsymbol{x} - \boldsymbol{g} + \boldsymbol{g} - \boldsymbol{g}^*)' + \epsilon(\boldsymbol{y} - \boldsymbol{g}^*)(\boldsymbol{y} - \boldsymbol{g}^*)' \\
&= \underbrace{\sum_{x \in E} (\boldsymbol{x} - \boldsymbol{g})(\boldsymbol{x} - \boldsymbol{g})'}_{Q} + \underbrace{\sum_{x \in E} (\boldsymbol{x} - \boldsymbol{g})(\boldsymbol{g} - \boldsymbol{g}^*)'}_{0} + \underbrace{\sum_{x \in E} (\boldsymbol{g} - \boldsymbol{g}^*)(\boldsymbol{x} - \boldsymbol{g})'}_{0} \\
&\quad + \sum_{x \in E} (\boldsymbol{g} - \boldsymbol{g}^*)(\boldsymbol{g} - \boldsymbol{g}^*)' + \epsilon(\boldsymbol{y} - \boldsymbol{g}^*)(\boldsymbol{y} - \boldsymbol{g}^*)' \\
&= Q + n(\boldsymbol{g} - \boldsymbol{g}^*)(\boldsymbol{g} - \boldsymbol{g}^*)' + \epsilon(\boldsymbol{y} - \boldsymbol{g}^*)(\boldsymbol{y} - \boldsymbol{g}^*)'.
\end{aligned}$$

The updating formula of the center gives also

$$\begin{cases} \boldsymbol{g} - \boldsymbol{g}^* = -\frac{\epsilon}{n+\epsilon}(\boldsymbol{y} - \boldsymbol{g}) \\ \boldsymbol{y} - \boldsymbol{g}^* = \frac{n}{n+\epsilon}(\boldsymbol{y} - \boldsymbol{g}). \end{cases}$$

Consequently

$$\begin{aligned} Q^* &= Q + n\frac{\epsilon^2}{(n+\epsilon)^2}(\boldsymbol{y}-\boldsymbol{g})(\boldsymbol{y}-\boldsymbol{g})' + \epsilon\frac{n^2}{(n+\epsilon)^2}(\boldsymbol{y}-\boldsymbol{g})(\boldsymbol{y}-\boldsymbol{g})' \\ &= Q + \epsilon\frac{n}{n+\epsilon}(\boldsymbol{y}-\boldsymbol{g})(\boldsymbol{y}-\boldsymbol{g})'. \end{aligned}$$

B.3 Updating inverse of the within scattering matrix

Proposition B.3

$$Q^{*-1} = Q^{-1} - \epsilon\frac{(Q^{-1}\boldsymbol{h})(Q^{-1}\boldsymbol{h})'}{1+\epsilon\boldsymbol{h}'Q^{-1}\boldsymbol{h}}.$$

Proof It suffices to prove that $Q^*Q^{*-1} = I$:

$$\begin{aligned} Q^*Q^{*-1} &= (Q + \epsilon\boldsymbol{h}\boldsymbol{h}')\left(Q^{-1} - \epsilon\frac{(Q^{-1}\boldsymbol{h})(Q^{-1}\boldsymbol{h})'}{1+\epsilon\boldsymbol{h}'Q^{-1}\boldsymbol{h}}\right) \\ &= I + \epsilon\frac{\overbrace{\boldsymbol{h}\boldsymbol{h}'Q^{-1} - \boldsymbol{h}\boldsymbol{h}'Q^{-1}}^{0} + \epsilon\overbrace{\boldsymbol{h}\boldsymbol{h}'Q^{-1}\boldsymbol{h}'Q^{-1}\boldsymbol{h} - \boldsymbol{h}\boldsymbol{h}'Q^{-1}\boldsymbol{h}\boldsymbol{h}'Q^{-1}}^{0 \text{ because } \boldsymbol{h}'Q^{-1}\boldsymbol{h}\in\mathbb{R}}}{1+\epsilon\boldsymbol{h}'Q^{-1}\boldsymbol{h}}. \end{aligned}$$

B.4 Updating determinant of the within scattering matrix

Lemma B.1 *For all $\boldsymbol{h} \in \mathbb{R}^d$, we have*

$$|I + \epsilon\boldsymbol{h}\boldsymbol{h}'| = 1 + \epsilon\boldsymbol{h}'\boldsymbol{h}.$$

Proof If $\boldsymbol{h} = (x_1, \ldots, x_d)'$, we have

$$|I + \epsilon\boldsymbol{h}\boldsymbol{h}'| = \epsilon^d \begin{vmatrix} \frac{1}{\epsilon} + x_1^2 & x_1x_2 & \ldots & x_1x_d \\ x_1x_2 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_1x_d & \ldots & \ldots & \frac{1}{\epsilon} + x_d^2 \end{vmatrix} = \epsilon^d(x_1\ldots x_d)^2 \begin{vmatrix} \frac{1}{\epsilon x_1^2} + 1 & 1 & \ldots & 1 \\ 1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 1 & \ldots & \ldots & \frac{1}{\epsilon x_d^2} + 1 \end{vmatrix}$$

$$= \epsilon^d (x_1 \dots x_d)^2 \begin{vmatrix} \frac{1}{\epsilon x_1^2}+1 & 1 & 1 & \dots & 1 \\ -\frac{1}{\epsilon x_1^2} & \frac{1}{\epsilon x_2^2} & 0 & \dots & 0 \\ -\frac{1}{\epsilon x_1^2} & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -\frac{1}{\epsilon x_1^2} & 0 & \dots & 0 & \frac{1}{\epsilon x_d^2} \end{vmatrix} = \begin{vmatrix} 1+\epsilon x_1^2 & \epsilon x_2^2 & \epsilon x_3^2 & \dots & \epsilon x_d^2 \\ -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -1 & 0 & \dots & 0 & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 1+\epsilon(x_1^2+\ldots+x_d^2) & \epsilon x_2^2 & \epsilon x_3^2 & \dots & \epsilon x_d^2 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & 1 \end{vmatrix} = 1+\epsilon(x_1^2+\ldots+x_d^2) = 1+\epsilon \boldsymbol{h}'\boldsymbol{h}.$$

Proposition B.4

$$|Q^*| = |Q|(1+\epsilon \boldsymbol{h}' Q^{-1} \boldsymbol{h}).$$

Proof The non-singular symetric matrix Q can be written $Q = BB'$ where B is a non-singular matrix. Then, we obtain

$$\begin{aligned} |Q^*| &= |Q+\epsilon \boldsymbol{h}\boldsymbol{h}'| \\ &= |BB'+\epsilon \boldsymbol{h}\boldsymbol{h}'| \\ &= |B||I+\epsilon(B^{-1}\boldsymbol{h})(\boldsymbol{h}'B'^{-1})||B'| \\ &= |B|(1+\epsilon(B^{-1}\boldsymbol{h})'(B^{-1}\boldsymbol{h}))|B'| \quad \text{(from Lemma B.1)} \\ &= |Q|(1+\epsilon \boldsymbol{h}'Q^{-1}\boldsymbol{h}). \end{aligned}$$

B.5 Updating trace of the within scattering matrix

Proposition B.5

$$tr(Q^*) = tr(Q) + \epsilon \boldsymbol{h}'\boldsymbol{h}.$$

Proof The function trace is a linear operator and, so,

$$\mathrm{tr}(Q^*) = \mathrm{tr}(Q) + \epsilon \mathrm{tr}(\boldsymbol{h}\boldsymbol{h}').$$

Since $\mathrm{tr}(\boldsymbol{h}\boldsymbol{h}') = \mathrm{tr}(\boldsymbol{h}'\boldsymbol{h})$, we have

$$\mathrm{tr}(Q^*) = \mathrm{tr}(Q) + \epsilon \mathrm{tr}(\boldsymbol{h}'\boldsymbol{h}).$$

We conclude by noting that the quantity $\boldsymbol{h}'\boldsymbol{h}$ is a scalar and, so, that $\mathrm{tr}(\boldsymbol{h}'\boldsymbol{h}) = \boldsymbol{h}'\boldsymbol{h}$.

References

Aitkin, M. and Rubin, D. B. (1985). Estimation and Hypothesis Testing in Finite Mixture Models. *Journal of the Royal Statistical Society*, Series B, **47**, 67-75.

Akaike, H. (1974). A New Look at the Statistical Identification Model. *IEEE Transactions on Automatic Control*, **19**, 716-723.

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

Bensmail H. and Celeux G. (1996). Regularized Gaussian Discriminant Analysis through Eigenvalue Decomposition. *Journal of the American Statistical Association*, **91**, 1743-1748.

Bensmail H., Celeux G., Raftery A. E. and Robert C. P. (1997). Inference in Model-based Cluster Analysis. *Statistics and Computing*, **7**, 1-10.

Biernacki C. (1997). Choix de modèles en classification. PhD. thesis, UTC Compiègne.

Biernacki C. and Govaert G. (1997). Using the Classification Likelihood to Choose the Number of Clusters. *Computing Science and Statistics*, **29**(2), 451-457.

Bozdogan H. (1987). Model Selection and Akaike Information Criterion (AIC): The General Theory and its Analytic Extensions. *Psychometrika*, **52**, 345-370.

Bozdogan, H. (1990). On the Information-Based Measure of Covariance Complexity and its Application to the Evaluation of Multivariate Linear Models. *Communications in Statistics, Theory and Methods*, **19**, 221-278.

Celeux, G. and Govaert, G. (1995). Gaussian Parsimonious Clustering Models. *Pattern Recognition*, **28**, 781-793.

Celeux G. and Soromenho G. (1996). An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model. *Journal of Classification*, **13**, 195-212.

Cutler, A., and Windham, M. P. (1993). "Information-Based Validity Functionals for Mixture Analysis," *Proceedings of the first US-Japan Conference on the Frontiers of Statistical Modeling.* (H. Bozdogan ed.), Amsterdam: Kluwer, pp. 149-170.

Dempster, A. P., Laird N. M. and Rubin D. B. (1997). Maximum Likelihood from Incomplete Data with the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Flury, B. W., Schmid, M. J. and Narayanan, A. (1993). Error Rates in Quadratic Discrimination with Constraints on the Covariance Matrices. *Journal of Classification*, **11**, 101-120.

Hathaway, R. J. (1986). Another Interpretation of the EM Algorithm for Mixture Distributions. *Statistics and Probability Letters*, **4**, 53-56.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factors and Model Uncertainty. *Journal of the American Statistical Association*, **90**, 773-795.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York, Wiley.

McLachlan, G. J. and Peel D. (1996). On a Resampling Approach to Choose the Number of Components in Normal Mixtures Models. Research Report #58, Centre for Statistics Research Report, The University of Queensland.

Raftery, A.E. (1996). Hypothesis Testing and Model Selection via Posterior Simulation. In *Practical Markov Chain Monte Carlo* (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.), London: Chapman and Hall, pp. 163-188.

Richardson S. and Green P. J. (1997). Fully Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society*, B, **59**, 731-792.

Ripley, B. D. (1996). *Neural Networks and Pattern Recognition.* New York: Cambridge University Press.

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

Wolfe, J. H. (1971). A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions. US Naval Personnel Research Activity. *Technical Bulletin STB 72-2*, San Diego, California.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

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