Journal of Multivariate Analysis 102 (2011) 1374–1387

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Variable selection in model-based discriminant analysis

C. Maugis [a,*], G. Celeux [b], M.-L. Martin-Magniette [c,d]

[a] *Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse, France*
[b] *Inria Saclay Île-de-France, France*
[c] *UMR AgroParisTech/INRA MIA 518, Paris, France*
[d] *URGV UMR INRA 1165, UEVE, ERL CNRS 8196, Evry, France*

ARTICLE INFO

Article history:
Received 19 May 2010
Available online 15 May 2011

AMS subject classification:
62H30

Keywords:
Discriminant, redundant or independent variables
Variable selection
Gaussian classification models
Linear regression
BIC

ABSTRACT

A general methodology for selecting predictors for Gaussian generative classification models is presented. The problem is regarded as a model selection problem. Three different roles for each possible predictor are considered: a variable can be a relevant classification predictor or not, and the irrelevant classification variables can be linearly dependent on a part of the relevant predictors or independent variables. This variable selection model was inspired by a previous work on variable selection in model-based clustering. A BIC-like model selection criterion is proposed. It is optimized through two embedded forward stepwise variable selection algorithms for classification and linear regression. The model identifiability and the consistency of the variable selection criterion are proved. Numerical experiments on simulated and real data sets illustrate the interest of this variable selection methodology. In particular, it is shown that this well ground variable selection model can be of great interest to improve the classification performance of the quadratic discriminant analysis in a high dimension context.

1. Introduction

The task of supervised classification is to build a classifier which enables us to assign an object described by predictors to one of the known classes. Such classifiers are built from a training set of objects for which the predictor measurements and the class labels are known. A lot of different methods are available; see for instance, the recent books [9,4] on statistical learning. Those methods differ in the way they approach the problem. Generative models, as Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), estimate the class-conditional densities. The predictive models (for instance logistic regression, classification trees and the k-nearest-neighbor classifier) directly estimate the posterior class probabilities. Non-probabilistic methods, as Neural networks and Kernel methods such as Support Vectors, aim at finding the decision function which characterizes the classifier.

Generative models are less parsimonious than predictive and non-probabilistic methods. Those last methods are generally preferred to generative models when the number of predictors is large in regard to the number of objects in the training set. However, generative models have some advantages since they allow us to determine the marginal density of the data. As noted in [9], LDA and QDA are widely used and perform well on an amazingly large and diverse set of classification problems. Moreover, LDA is regarded as a reference method by many practitioners, and an advantage of LDA over QDA is that it is a more parsimonious method.

[*] Corresponding address: Cathy Maugis, INSA de Toulouse, 135 avenue de Rangueil, 31077 Toulouse Cedex 4, France.
E-mail addresses: cathy.maugis@insa-toulouse.fr (C. Maugis), Gilles.Celeux@inria.fr (G. Celeux), mlmartin@agroparistech.fr (M.-L. Martin-Magniette).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.05.004

Much efforts have been paid in variable selection for classification; see the reviews of Guyon and Ellisseeff [8] and Mary-Huard et al. [12]. In this paper, we concentrate our attention on variable selection for Gaussian generative models. There exist quite efficient methods to select predictors in the LDA context. Efficient stepwise variable selection procedures are available in most statistical softwares (see [16] [Section 12.3.3]). On the contrary, there is less available material for QDA [22], and as far as we know, no variable selection procedure for QDA is available in standard statistical softwares. However, in the past few years, there is a renewal of interest in this topic. Zhang and Wang [23] proposed a variable selection procedure for QDA based on a BIC criterion and Murphy et al. [17] have adapted the variable selection procedure of Raftery and Dean [18] to the supervised classification context.

The purpose of this paper is to extend the general variable selection modeling proposed in [14], conceived for model-based clustering to the Gaussian classification models. This modeling is the result of successive improvements of variable selection modeling in model-based clustering [18,13,14]. Acting in such a way, we dramatically strengthen the appeal of nonlinear Gaussian classifiers, proposed by Bensmail and Celeux [2] which are up to now limited by the large number of parameters to be estimated. The models and variable selection algorithms proposed not only lead to interpret the roles of variables in a clear way, but they also lead to much increase the discriminative efficiency of methods such as QDA.

The paper is organized as follows. In Section 2, Gaussian models of classification are recalled. Our variable selection approach is presented in Section 3. It makes use of a model which states a clear distinction between useful, redundant and noisy variables for the classification task in the Gaussian framework. It leads to a BIC-like criterion to be optimized (Section 4). It is proved in Section 5 that our approach leads to identifiable classification models and that our variable selection criterion is consistent under mild assumptions. In Section 6, a variable selection algorithm using two forward stepwise algorithms is described to determine the roles of the predictors. Applications on simulated and real data sets are presented in Section 7. A short discussion section ends the paper and the proofs of the theorems of Section 5 are postponed to Appendices B and C.

2. Gaussian classification models

Training data for discriminant analysis are composed by n vectors

$$(\underline{\mathbf{x}}, \underline{z}) = \left\{(\mathbf{x}_1, z_n), \dots, (\mathbf{x}_n, z_n); \mathbf{x}_i \in \mathbb{R}^Q, z_i \in \{1, \dots, K\}\right\},$$

where $\mathbf{x}_i$ is the Q-dimensional predictor and z_i is the class label of the ith subject. We assume that the prior probability of the class G_k is $P(z = k) = p_k$ with $p_k > 0$ for any k, $1 \leq k \leq K$ and $\sum_{k=1}^{K} p_k = 1$. The class-conditional density of class G_k is modeled with a Q-dimensional Gaussian density: $\mathbf{x}_i \mid z_i = k \sim \mathcal{N}_Q(\mu_k, \Sigma_k)$ where $\mu_k \in \mathbb{R}^Q$ is the mean vector and Σ_k is the $Q \times Q$ variance matrix. The aim of discriminant analysis is to design a classifier from the training sample, allowing us to estimate the label of any new observation $\mathbf{x} \in \mathbb{R}^Q$.

Gaussian generative models differ essentially in their assumptions about the variance matrices. The most commonly applied method, called linear discriminant analysis (LDA), assumes that the variance matrices of the different class are equal. When the variance matrices are totally free, the method is called quadratic discriminant analysis (QDA). Bensmail and Celeux [2] generalize the LDA and QDA methods in the Eigenvalue Decomposition Discriminant Analysis (EDDA). As in [1,5], EDDA is based on the eigenvalue decomposition of the variance matrices

$$\forall k \in \{1, \dots, K\}, \quad \Sigma_k = L_k D_k A_k D_k'$$

where $L_k = |\Sigma_k|^{\frac{1}{Q}}$, D_k is the Σ_k's eigenvector matrix and A_k is the diagonal matrix of the normalized eigenvalues of Σ_k. Those elements respectively control the volume, the orientation and the shape of the density contour of class G_k. According to constraints required on the three elements of the eigenvalue decomposition, a collection $\mathcal{M}$ of 14 more or less parsimonious and easily interpreted models is available (see Table A.4 in Appendix A). Those 14 models are available in the MIXMOD software [3] and, for most of them, in the MCLUST software [6]. The LDA and QDA are implemented in several softwares as well as R in the library MASS.

The model selection in the EDDA context consists of choosing the best form of the variance matrices. The best model is usually selected by minimizing the cross-validated classification error rate [2]. Another possible selection criterion is the Bayesian information criterion [19] which is an asymptotic approximation of the integrated loglikelihood. The model which maximizes the Bayesian information criterion (BIC) is selected. In this paper a selection by the BIC is considered. Notice that BIC focuses on the model fit rather than the minimization of the misclassification rate and is much cheaper to compute.

Once a model of the collection is selected, a new observation $\mathbf{x}_0$ is assigned to the group for which its a posteriori probability is maximum. It is the *Maximum A Posteriori* rule and it is equivalent to find the class $k^\star$ such that

$$k^\star = \underset{1 \leq k \leq K}{\operatorname{argmax}}\, p_k \Phi(\mathbf{x}_0 | \mu_k, \Sigma_{k(m)}),$$

where $\Phi(.|\mu_k, \Sigma_{k(m)})$ denotes the Gaussian density with mean vector μ_k and variance matrix $\Sigma_{k(m)}$ fulfilling the form $m \in \mathcal{M}$.

3. The variable selection model collection

Each of the Q available variables brings information (its own ability to separate the classes), and noise (its sampling variance). Thus it is important to select the variables bringing more discriminant information than noise. In practice there are three kinds of variables: The discriminant variables useful for the classification task, the redundant variables linked to the discriminant variables, and the noisy variables which bring no information for the classification task. Thus, variable selection is an important part of discriminant analysis to get a reliable and parsimonious classifier. Considering the classification problem in the model-based discriminant analysis context allows us to recast variable selection into a model selection problem and to adapt the variable selection model for model-based clustering of Maugis et al. [14] in the supervised classification context.

In our modeling, the variables have three possible roles: relevant, redundant or independent for the discriminant analysis. The nonempty set of relevant predictors is denoted as S and the independent variable subset is denoted as W. The redundant variables, whose the subset is denoted as U, are explained by a variable subset R of S according to a linear regression while the variables in W are assumed to be independent of all the relevant variables. Note that if U is empty, R is empty too and otherwise R is assumed to be not empty. Thus denoting $\mathcal{F}$ the family of variable index subsets of $\{1, \dots, Q\}$, the variable partition set can be described as follows:

$$\mathcal{V} = \left\{ (S, R, U, W) \in \mathcal{F}^4; \begin{array}{l} S \cup U \cup W = \{1, \dots, Q\} \\ S \cap U = \emptyset, S \cap W = \emptyset, U \cap W = \emptyset \\ S \neq \emptyset, R \subseteq S \\ R = \emptyset \text{ if } U = \emptyset \text{ and } R \neq \emptyset \text{ otherwise} \end{array} \right\}.$$

Throughout this paper, a quadruplet (S, R, U, W) of $\mathcal{V}$ is denoted as $\mathbf{V} = (S, R, U, W)$.

The law of the training sample is modeled by, $\forall (\mathbf{x}, z) \in \mathbb{R}^Q \times \{1, \dots, K\}$,

$$\begin{cases} f(\mathbf{x}|z = k, m, r, l, \mathbf{V}) = \Phi(\mathbf{x}^S|\mu_k, \Sigma_{k(m)})\, \Phi(\mathbf{x}^U|a + \mathbf{x}^R\beta, \Omega_{(r)})\, \Phi(\mathbf{x}^W|\gamma, \tau_{(l)}) \\ (\mathbb{I}_{z=1}, \dots, \mathbb{I}_{z=K}) \sim \text{Multinomial}(1; p_1, \dots, p_K) \end{cases}$$

where

- on the discriminant variable subset S, the variance matrices $\Sigma_{1(m)}, \dots, \Sigma_{K(m)}$ fulfill the constraints of the form $m \in \mathcal{M}$ (see Section 2);
- on the redundant variable subset U, the density $\Phi(\mathbf{x}^U|a + \mathbf{x}^R\beta, \Omega_{(r)})$ corresponds to the linear regression density of $\mathbf{x}^U$ on $\mathbf{x}^R$, where the vector a is the intercept vector, β is the regression coefficient matrix and $\Omega_{(r)}$ is the variance matrix; this last matrix is assumed to have a spherical ($[LI]$), diagonal ($[LB]$) or a general ($[LC]$) form, and this form is specified by $r \in \mathcal{T}_{reg} = \{[LI], [LB], [LC]\}$;
- on the independent variable subset W, the marginal density is assumed to be a Gaussian density with mean γ and variance matrix $\tau_{(l)}$ which can be spherical or diagonal and is specified by $l \in \mathcal{T}_{indep} = \{[LI], [LB]\}$.

Finally the model collection is

$$\mathcal{N} = \left\{(m, r, l, \mathbf{V});\, m \in \mathcal{M}, r \in \mathcal{T}_{\text{reg}}, l \in \mathcal{T}_{\text{indep}}, \mathbf{V} \in \mathcal{V}\right\} \tag{1}$$

and the likelihood $f(\underline{\mathbf{x}}, \underline{z}|m, r, l, \mathbf{V}, \theta)$ of model $(m, r, l, \mathbf{V})$ is given by

$$\prod_{i=1}^{n} \prod_{k=1}^{K} \left[p_k \Phi(\mathbf{x}_i^S|\mu_k, \Sigma_{k(m)})\, \Phi(\mathbf{x}_i^U|a + \mathbf{x}_i^R\beta, \Omega_{(r)})\, \Phi(\mathbf{x}_i^W|\gamma, \tau_{(l)})\right]^{\mathbb{I}_{z_i=k}}$$

where the parameter vector $\theta = (\alpha_{(m)}, a, \beta, \Omega_{(r)}, \gamma, \tau_{(l)})$ with

$$\alpha_{(m)} = (p_1, \dots, p_K, \mu_1, \dots, \mu_K, \Sigma_{1(m)}, \dots, \Sigma_{K(m)})$$

belongs to a parameter vector set $\Upsilon_{(m,r,l,\mathbf{V})}$.

4. Model selection criterion

The model collection $\mathcal{N}$ allows us to recast the variable selection problem for Gaussian discriminant analysis into a model selection problem. Ideally, we search the model maximizing the integrated loglikelihood

$$(\tilde{m}, \tilde{r}, \tilde{l}, \tilde{\mathbf{V}}) = \underset{(m,r,l,\mathbf{V}) \in \mathcal{N}}{\operatorname{argmax}} \ln[f(\underline{\mathbf{x}}, \underline{z}|m, r, l, \mathbf{V})]$$

where

$$f(\underline{\mathbf{x}}, \underline{z}|m, r, l, \mathbf{V}) = \int f(\underline{\mathbf{x}}, \underline{z}|m, r, l, \mathbf{V}, \theta)\Pi(\theta|m, r, l, \mathbf{V})\mathrm{d}\theta,$$

Π being the prior distribution of the vector parameter. Since this integrated loglikelihood is difficult to evaluate, it could be approximated by the BIC criterion [19]. Then the selected model satisfies

$$(\hat{m}, \hat{r}, \hat{l}, \hat{\mathbf{V}}) = \underset{(m,r,l,\mathbf{V})\in\mathcal{N}}{\operatorname{argmax}} \operatorname{crit}(m, r, l, \mathbf{V}) \tag{2}$$

where the model selection criterion is defined by

$$\operatorname{crit}(m, r, l, \mathbf{V}) = \mathrm{BIC}_{\mathrm{da}}(\underline{\mathbf{x}}^S, \underline{z}|m) + \mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^U|r, \underline{\mathbf{x}}^R) + \mathrm{BIC}_{\mathrm{indep}}(\underline{\mathbf{x}}^W|l), \tag{3}$$

where

- the BIC criterion for the Gaussian discriminant analysis on the relevant variable subset S is given by

$$\mathrm{BIC}_{\mathrm{da}}(\underline{\mathbf{x}}^S, \underline{z}|m) = 2\sum_{i=1}^{n} \ln\left[\sum_{k=1}^{K} \hat{p}_k \Phi(\mathbf{x}_i^S|\hat{\mu}_k, \hat{\Sigma}_{k(m)})\mathbb{I}_{z_i=k}\right] - \lambda_{(m,S)}\ln(n)$$

  where $\hat{\alpha}_{(m)}$ is the maximum likelihood estimator and $\lambda_{(m,S)}$ is the number of free parameters for the model m on the variable subset S.
- the BIC criterion for the linear regression of the variable subset U on R is defined by

$$\mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^U|r, \underline{\mathbf{x}}^R) = 2\sum_{i=1}^{n} \ln[\Phi(\mathbf{x}_i^U|\hat{a} + \mathbf{x}_i^R\hat{\beta}, \hat{\Omega}_{(r)})] - \nu_{(r,U,R)}\ln(n) \tag{4}$$

  where $\hat{a}$, $\hat{\beta}$ and $\hat{\Omega}_{(r)}$ are the maximum likelihood estimators and $\nu_{(r,U,R)}$ is the number of free parameters of the linear regression.
- the BIC criterion associated to the Gaussian density on the variable subset W is given by

$$\mathrm{BIC}_{\mathrm{indep}}(\underline{\mathbf{x}}^W|l) = 2\sum_{i=1}^{n} \ln[\Phi(\mathbf{x}_i^W|\hat{\gamma}, \hat{\tau}_{(l)})] - \rho_{(l,W)}\ln(n).$$

  The parameters $\hat{\gamma}$ and $\hat{\tau}_{(l)}$ denote the maximum likelihood estimators and $\rho_{(l,W)}$ is the number of free parameters of the Gaussian density.
- the maximum likelihood estimator is denoted as $\hat{\theta} = (\hat{\alpha}_{(m)}, \hat{a}, \hat{\beta}, \hat{\Omega}_{(r)}, \hat{\gamma}, \hat{\tau}_{(l)})$ and the overall number of free parameters is $\Xi_{(m,r,l,\mathbf{V})} = \lambda_{(m,S)} + \nu_{(r,U,R)} + \rho_{(l,W)}$.

5. Theoretical properties

The theoretical properties established in [14] in the model-based clustering framework can be adapted to the Gaussian discriminant analysis context. First, necessary and sufficient conditions are given to ensure the identifiability of the model collection. Second, a consistency theorem of the model selection criterion is stated.

5.1. Identifiability

In order to ensure the model identifiability, some natural conditions are required to distinguish the discriminant density part to the regression and the independent Gaussian density parts. For instance, if s is a nonempty subset strictly included into the relevant variable subset S and $\bar{s}$ is its complement in S then the identifiability cannot be ensured if the regression density of $\bar{s}$ on s can be regrouped with the regression density of U on R. Despite the fact that Conditions (C1)–(C3) of Theorem 1 look rather technical, they are quite natural and Theorem 1 is saying that our variable selection model is identifiable in all situations of interest.

The following additional notation is introduced to state the model identifiability theorem. Recall that $\Phi(.|\mu_k, \Sigma_k)$ denotes the Gaussian density with mean μ_k and variance matrix Σ_k. The parameters can be decomposed into $\mu_k = (\mu_{ks}, \mu_{k\bar{s}})$ and Σ_k into submatrices $\Sigma_{k,ss}$, $\Sigma_{k,s\bar{s}}$ and $\Sigma_{k,\bar{s}\bar{s}}$, where s is a nonempty subset of S and $\bar{s}$ its complement in S. Moreover, conditional parameters are defined by $\mu_{k,\bar{s}|s} = \mu_{k\bar{s}} - \mu_{ks}\Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$, $\Sigma_{k,\bar{s}|s} = \Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$ and $\Sigma_{k,\bar{s}\bar{s}|s} = \Sigma_{k,\bar{s}\bar{s}} - \Sigma_{k,\bar{s}s}\Sigma_{k,ss}^{-1}\Sigma_{k,s\bar{s}}$. For two subsets s and t, the following restrictions of a $I \times J$ matrix Λ are considered: $\Lambda_{st} = (\Lambda_{ij})_{i\in s, j\in t}$, $\Lambda_{.t} = (\Lambda_{ij})_{1\leq i\leq I, j\in t}$ and $\Lambda_{s.} = (\Lambda_{ij})_{i\in s, 1\leq j\leq J}$.

Theorem 1. *Let $\Theta_{(m,r,l,\mathbf{V})}$ be a subset of the parameter set $\Upsilon_{(m,r,l,\mathbf{V})}$ such that elements $\theta = (\alpha, a, \beta, \Omega, \gamma, \tau)$*

(C1): *contain couples (μ_k, Σ_k) fulfilling $\forall s \subsetneq S$, $\exists(k, k')$, $1 \leq k < k' \leq K$*

$$\mu_{k,\bar{s}|s} \neq \mu_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}|s} \neq \Sigma_{k',\bar{s}|s} \text{ or } \Sigma_{k,\bar{s}\bar{s}|s} \neq \Sigma_{k',\bar{s}\bar{s}|s},$$

where $\bar{s}$ denotes the complement in S of any nonempty subset s of S.

(C2): *if $U \neq \emptyset$,*
- $*$ *for all variables j of R, there exists a variable u of U such that the restriction β_{uj} of the regression coefficient matrix β associated with j and u is not equal to zero.*
- $*$ *for all variables u of U, there exists a variable j of R such that $\beta_{uj} \neq 0$.*

(C3): *Parameters Ω and τ strictly respect the forms r and l respectively: They are both diagonal matrices with at least two different eigenvalues if $r = [LB]$ and $l = [LB]$ and Ω has at least a non-zero entry outside the main diagonal if $r = [LC]$.*

Let $(m, r, l, \mathbf{V})$ and $(m^\star, r^\star, l^\star, \mathbf{V}^\star)$ be two models. If there exist $\theta \in \Theta_{(m,r,l,\mathbf{V})}$ and $\theta^\star \in \Theta_{(m^\star,r^\star,l^\star,\mathbf{V}^\star)}$ such that

$$f(.|m, r, l, \mathbf{V}, \theta) = f(.|m^\star, r^\star, l^\star, \mathbf{V}^\star, \theta^\star)$$

then $(m, r, l, \mathbf{V}) = (m^\star, r^\star, l^\star, \mathbf{V}^\star)$ and $\theta = \theta^\star$.

The complete proof of Theorem 1 is postponed to Appendix B.

5.2. Consistency

A consistency property of our criterion can be checked. In this section, it is proved that the probability of selecting the true model by maximizing Criterion (3) approaches 1 as $n \to \infty$. Denoting h the density function of the sample $(\underline{\mathbf{x}}, \underline{z})$, the two following vectors are considered

$$\begin{aligned}\theta^\star_{(m,r,l,\mathbf{V})} &= \underset{\theta_{(m,r,l,\mathbf{V})}\in\Theta_{(m,r,l,\mathbf{V})}}{\operatorname{argmin}} \mathrm{KL}[h, f(.|m, r, l, \mathbf{V}, \theta)]\\ &= \underset{\theta_{(m,r,l,\mathbf{V})}\in\Theta_{(m,r,l,\mathbf{V})}}{\operatorname{argmax}} \mathbb{E}\{\ln f(X, Z|m, r, l, \mathbf{V}, \theta)\},\end{aligned}$$

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

$$\hat{\theta}_{(m,r,l,\mathbf{V})} = \underset{\theta_{(m,r,l,\mathbf{V})}\in\Theta_{(m,r,l,\mathbf{V})}}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \ln\{f(\mathbf{x}_i, z_i|m, r, l, \mathbf{V}, \theta)\}.$$

Recall that $\Theta_{(m,r,l,\mathbf{V})}$'s are the subsets defined in Theorem 1 for ensuring the model identifiability.

The following assumption is considered:

(H1) The density h is assumed to be one of the densities in competition. By identifiability, there exists a unique model $(m_0, r_0, l_0, \mathbf{V}_0)$ and an associated parameter $\theta^\star$ such that $h = f(.|m_0, r_0, l_0, \mathbf{V}_0, \theta^\star)$.

Moreover, an additional technical assumption is considered:

(H2) For all models $(m, r, l, \mathbf{V}) \in \mathcal{N}$, the vectors $\theta^\star$ and $\hat{\theta}$ are supposed to belong to a compact subspace $\Theta'_{(m,r,l,\mathbf{V})}$ in the intersection between $\Theta_{(m,r,l,\mathbf{V})}$ and

$$\begin{pmatrix}\mathcal{P}_{K-1}(\rho) \times \mathcal{B}(\eta, \mathrm{card}(S))^K \times \mathcal{D}^K_{\mathrm{card}(S)} \times \mathcal{B}(\eta, \mathrm{card}(U))\\ \times \mathcal{B}(\eta, \mathrm{card}(R), \mathrm{card}(U)) \times \mathcal{D}_{\mathrm{card}(U)} \times \mathcal{B}(\eta, \mathrm{card}(W)) \times \mathcal{D}_{\mathrm{card}(W)}\end{pmatrix}$$

where
- $\mathcal{P}_{K-1}(\rho) = \left\{(p_1, \ldots, p_K) \in [\rho, 1]^K; \sum_{k=1}^K p_k = 1\right\}$ where $\rho > 0$,
- $\mathcal{B}(\eta, r)$ is the closed ball in $\mathbb{R}^r$ of radius η centered at zero for the l^2-norm defined by $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^r x_i^2}$, $\forall \mathbf{x} \in \mathbb{R}^r$,
- $\mathcal{B}(\eta, r, q)$ is the closed ball in $\mathcal{M}_{r\times q}(\mathbb{R})$ of radius η centered at zero for the matricial norm $|||.|||$ defined by $\forall A \in \mathcal{M}_{r\times q}(\mathbb{R}),\ |||A||| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{x}A\|$,
- $\mathcal{D}_r$ is the set of the $r \times r$ positive definite matrices with eigenvalues in $[s_{\mathrm{m}}, s_{\mathrm{M}}]$ with $0 < s_{\mathrm{m}} < s_{\mathrm{M}}$.

Theorem 2. *Under assumptions (H1) and (H2), the model $(\hat{m}, \hat{r}, \hat{l}, \hat{\mathbf{V}})$ maximizing Criterion (3) is such that*

$$P((\hat{m}, \hat{r}, \hat{l}, \hat{\mathbf{V}}) = (m_0, r_0, l_0, \mathbf{V}_0)) \underset{n\to\infty}{\to} 1.$$

The proof of this theorem is given in Appendix C.

6. The variable selection procedure

Theorem 2 is reassuring about the theoretical behavior of the model selection Criterion (3). Unfortunately, the number of models given by (1) being huge, an exhaustive search for the model maximizing Criterion (3) is impossible. Thus we design a procedure, embedding forward stepwise algorithms, to determine the best variable roles and the best variance matrix forms.

6.1. The models in competition

At a fixed step of the algorithm, the variable set $\{1, \ldots, Q\}$ is divided into the subset of selected discriminant variables S, the subset U of redundant variables which are linked to some discriminant variables, the subset W of independent irrelevant variables and j the candidate variable for inclusion into or exclusion from the discriminant variable subset. Under the model (m, r, l), the integrated likelihood can be decomposed as

$$\begin{aligned} f(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{\mathbf{x}}^U, \underline{\mathbf{x}}^W, \underline{z}|m, r, l) &= f(\underline{\mathbf{x}}^U, \underline{\mathbf{x}}^W|\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z}, m, r, l) f(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z}|m, r, l) \\ &= f_{\text{indep}}(\underline{\mathbf{x}}^W|l) f_{\text{reg}}(\underline{\mathbf{x}}^U|r, \underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j) f(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z} \mid m, r, l) \end{aligned}$$

where $f_{\text{indep}}(\underline{\mathbf{x}}^W|l)$ is the integrated likelihood on the independent irrelevant variable subset W and $f_{\text{reg}}(\underline{\mathbf{x}}^U|r, \underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j)$ corresponds to the integrated likelihood on the subset U regressed on variable subset S and the candidate variable j. The expression of the integrated likelihood restricted on $S \cup \{j\}$ depends on the three situations which can occur for the candidate variable j:

- *First situation*: Given $\underline{\mathbf{x}}^S$, $\underline{\mathbf{x}}^j$ provides additional information for the discriminant analysis thus

  $$f(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z}|m, r, l) = f_{\text{da}}(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z}|m)$$

  corresponds to the integrated likelihood for the discriminant analysis on variable subset $S \cup \{j\}$.
- *Second situation:* Given $\underline{\mathbf{x}}^S$, $\underline{\mathbf{x}}^j$ does not provide additional information for the discriminant analysis but has a linear link with a nonempty subset denoted $R[j]$ of S containing the relevant variables for the regression of $\underline{\mathbf{x}}^j$ on $\underline{\mathbf{x}}^S$:

  $$f(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z}|m, r, l) = f_{\text{da}}(\underline{\mathbf{x}}^S, \underline{z}|m) f_{\text{reg}}(\underline{\mathbf{x}}^j|[LI], \underline{\mathbf{x}}^{R[j]}).$$

  The second term in the right-hand side corresponds to the integrated likelihood of the regression of $\underline{\mathbf{x}}^j$ on $\underline{\mathbf{x}}^{R[j]}$. Since j is a single variable, the variance matrix is spherical ($[LI]$).
- *Third situation:* Given $\underline{\mathbf{x}}^S$, $\underline{\mathbf{x}}^j$ does not provide additional information for the discriminant analysis and is independent of all the variables of S:

  $$f(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z}|m, r, l) = f_{\text{da}}(\underline{\mathbf{x}}^S|m) f_{\text{indep}}(\underline{\mathbf{x}}^j|[LI]).$$

  The second term in the right-hand side corresponds to the integrated likelihood of the independent Gaussian density on the variable j with a spherical variance matrix since j is a single variable.

In order to compare those three situations in an efficient way, we remark that $f_{\text{indep}}(\underline{\mathbf{x}}^j|[LI])$ can be written $f_{\text{reg}}(\underline{\mathbf{x}}^j|[LI], \underline{\mathbf{x}}^{\emptyset})$. Thus instead of considering the nonempty subset $R[j]$ we consider a new explicative variable subset denoted $\tilde{R}[j]$ and defined by $\tilde{R}[j] = \emptyset$ if j follows the third situation and $\tilde{R}[j] = R[j]$ if j follows the second situation. This allows us to recast the comparison of the three situations into the comparison of two situations with the Bayes factor

$$\frac{f_{\text{da}}(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z}|m)}{f_{\text{da}}(\underline{\mathbf{x}}^S, \underline{z}|m) f_{\text{reg}}(\underline{\mathbf{x}}^j|[LI], \underline{\mathbf{x}}^{\tilde{R}[j]})}.$$

This Bayes factor being difficult to evaluate, it is approximated by

$$\text{BIC}_{\text{diff}}(j|m) = \text{BIC}_{\text{da}}(\underline{\mathbf{x}}^S, \underline{\mathbf{x}}^j, \underline{z}|m) - \left\{\text{BIC}_{\text{da}}(\underline{\mathbf{x}}^S, \underline{z}|m) + \text{BIC}_{\text{reg}}(\underline{\mathbf{x}}^j|[LI], \underline{\mathbf{x}}^{\tilde{R}[j]})\right\}. \tag{5}$$

6.2. The general steps of the algorithm

First, this algorithm consists of separating variables into relevant and irrelevant variables for the discriminant analysis via a forward stepwise algorithm described in Section 6.3. Second, the irrelevant variables are partitioned into redundant variables, if regressors are chosen inside the relevant variables, and independent variables otherwise. It remains then to determine the set of regressors from the relevant variables for the multidimensional regression of the redundant variables and the general variance structures.

This algorithm, detailed in Fig. 1, is based on two embedded forward stepwise algorithms. In what follows, FSAC refers to the forward stepwise algorithm for classification described in Section 6.3 and FSAR refers to the forward stepwise algorithm for regression given in Appendix D. These forward stepwise algorithms allow to study data sets where the individual number n is smaller than the variable number Q. Nevertheless, for studying a data set where $Q \leq n$, a backward procedure (starting the search with all variables) could be preferred because it takes variable interactions into account.

```
Step 1: For each mixture form $m$:
    Step A: FSAC (described in Section 6.3)
        * Initialisation: $S(m) = \emptyset$, $S^c(m) = \{1, \dots, Q\}$
        * Preliminary inclusion step: selection of the first relevant predictor ($\sharp S(m) = 1$)
        * Alternate inclusion step and exclusion step until the algorithm stops.
          Note that these steps use FSAR.
        $\Rightarrow$ Variables are divided into relevant predictors $\hat{S}(m)$ and irrelevant predictors $\hat{S}^c(m)$.
    Step B: $\hat{S}^c(m)$ is divided into redundant variables $\hat{U}(m)$ and independent variables $\hat{W}(m)$ using FSAR:
        $\forall j \in \hat{S}^c(m)$, the variable subset $\tilde{R}[j]$ ($\subseteq \hat{S}(m)$) allowing to explain $j$ by a linear regression is
        determined with FSAR. If $\tilde{R}[j] = \emptyset$, $j \in \hat{W}(m)$ and otherwise, $j \in \hat{U}(m)$.
    Step C: For each form $r$:
        * Determination of the variable subset $\hat{R}(m,r)$ ($\subseteq \hat{S}(m)$) explaining $\hat{U}(m)$ using FSAR.
        * For each form $l$:
            ▷ Parameter vector estimation $\hat{\theta}$
            ▷ Calculation of the criterion value
              $\widetilde{\mathrm{crit}}(m,r,l) = \mathrm{crit}(m,r,l,\hat{S}(m),\hat{R}(m,r),\hat{U}(m),\hat{W}(m))$.
Step 2: Selection of $(\hat{m},\hat{r},\hat{l})$ fulfilling $(\hat{m},\hat{r},\hat{l}) = \underset{(m,r,l)\in\mathcal{M}\times\mathcal{T}_{\mathrm{reg}}\times\mathcal{T}_{\mathrm{indep}}}{\mathrm{argmax}} \widetilde{\mathrm{crit}}(m,r,l)$
    and selection of the complete model: $\left(\hat{m},\hat{r},\hat{l},\hat{S}(\hat{m}),\hat{R}(\hat{m},\hat{r}),\hat{U}(\hat{m}),\hat{W}(\hat{m})\right)$
```

Fig. 1. Description of the general steps of the algorithm. The abbreviation FSAC (resp. FSAR) refers to the forward stepwise algorithm for classification (resp. for regression) described in Section 6.3 (resp. Appendix D).

6.3. *The forward stepwise algorithm for classification (FSAC)*

Initialization Let m fixed, $S(m) = \emptyset$, $S^c(m) = \{1, \dots, Q\}$, $j_I = \emptyset$ and $j_E = \emptyset$. The algorithm is making use of an inclusion and an exclusion steps now described. The decision of including (resp. excluding) a variable in (resp. from) the discriminant variable subset is based on (5). Starting from a preliminary inclusion step, the forward variable selection algorithm consists of alternating inclusion and exclusion steps. It returns the discriminant variable subset $\hat{S}(m)$ and the irrelevant variable subset $\hat{S}^c(m)$. These different steps are now described.

Preliminary inclusion step This step consists of selecting the first discriminant variable. For all j in $S^c(m)$, compute

$$\mathrm{BIC}_{\mathrm{diff}}(j|m) = \mathrm{BIC}_{\mathrm{da}}(\underline{\mathbf{x}}^j, \underline{z}|m) - \mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^j|[LI], \underline{\mathbf{x}}^{\emptyset})$$

and determine

$$j_I = \underset{j\in S^c(m)}{\mathrm{argmax}}\, \mathrm{BIC}_{\mathrm{diff}}(j|m).$$

Then $S(m) = \{j_I\}$, $S^c(m) = S^c(m) \setminus \{j_I\}$ and go to the inclusion step.

Inclusion step For all j in $S^c(m)$, use the forward stepwise regression algorithm (see Appendix D) to determine the subset $\tilde{R}[j]$ for the regression of $\underline{\mathbf{x}}^j$ on $\underline{\mathbf{x}}^{S(m)}$. And, compute

$$\mathrm{BIC}_{\mathrm{diff}}(j|m) = \mathrm{BIC}_{\mathrm{da}}(\underline{\mathbf{x}}^{S(m)}, \underline{\mathbf{x}}^j, \underline{z}|m) - \left\{\mathrm{BIC}_{\mathrm{da}}(\underline{\mathbf{x}}^S, \underline{z}|m) + \mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^j|[LI], \underline{\mathbf{x}}^{\tilde{R}[j]})\right\}.$$

Then, compute

$$j_I = \underset{j\in S^c(m)}{\mathrm{argmax}}\, \mathrm{BIC}_{\mathrm{diff}}(j \mid m).$$

- If $\mathrm{BIC}_{\mathrm{diff}}(j_I \mid m) > 0$, $S(m) = S(m) \cup \{j_I\}$, $S^c(m) = S^c(m) \setminus \{j_I\}$ and, if $j_I \neq j_E$, go to the exclusion step and stop otherwise.
- Otherwise, $j_I = \emptyset$. If $j_E \neq \emptyset$, go to the exclusion step and stop otherwise.

Exclusion step For all j in $S(m)$, use the forward stepwise regression algorithm (see Appendix D) to determine the subset $\tilde{R}[j]$ for the regression of $\underline{\mathbf{x}}^j$ on $\underline{\mathbf{x}}^{S(m)\setminus j}$. And, compute

$$\mathrm{BIC}_{\mathrm{diff}}(j|m) = \mathrm{BIC}_{\mathrm{da}}(\underline{\mathbf{x}}^{S(m)}, \underline{z}|m) - \left\{\mathrm{BIC}_{\mathrm{da}}(\underline{\mathbf{x}}^{S(m)\setminus\{j\}}, \underline{z}|m) + \mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^j|[LI], \underline{\mathbf{x}}^{\tilde{R}[j]})\right\}.$$

Then, compute

$$j_E = \underset{j\in S(m)}{\mathrm{argmin}}\, \mathrm{BIC}_{\mathrm{diff}}(j|m).$$

Table 1
Averaged classification error rate (±standard deviation) for LDA, QDA and EDDA methods, with and without variable selection for the simulated data sets.

With variable selection			Without variable selection		
LDA	QDA	EDDA	LDA	QDA	EDDA
4.94 (±0.13)	4.19 (±0.06)	4.18 (±0.06)	5.30 (±0.18)	6.23 (±0.38)	5.29 (±0.18)

- If $\text{BIC}_{\text{diff}}(j_I|m) < 0, S^c(m) = S^c(m) \cup \{j_E\}, S(m) = S(m) \setminus \{j_E\}$. If $j_E \neq j_I$, go to the inclusion step and stop otherwise.
- Otherwise, $j_E = \emptyset$. If $j_I \neq \emptyset$, go to the inclusion step and stop otherwise.

The way to include the first predictor in the subset $S(m)$ is standard, other initializations are possible. In practice, we have observed that the initialization step has no impact on the predictor partition. It is due to the use of stepwise algorithms which allow one to reevaluate the predictor role at each iteration.

7. Applications

We present numerical experiments to assess our variable selection procedure. First, the interest of variable selection for nonlinear discriminant analysis models such as QDA is highlighted on simulated data. Then, two applications on real data sets are presented. The application on the Landsat Satellite data set allows us to illustrate the interest of precising the role of the variables in an explicative perspective and again the great interest of our variable selection procedure to improve the classification performances of QDA. The second application concerns the Leukemia data of Golub et al. [7], a classical genomics example where the number of variables is greater than the number of observations.

7.1. Simulated example

This simulated example consists of considering samples described by $Q = 16$ variables. The prior probabilities of the four classes are assumed to be $p_1 = 0.15$, $p_2 = 0.3$, $p_3 = 0.2$ and $p_4 = 0.35$. On the three discriminant variables, data are distributed from $\mathbf{x}_i^{\{1-3\}}|z_i = k \sim \Phi(.|\mu_k, \Sigma_k)$ with mean vectors $\mu_1 = (1.5, -1.5, 1.5)$, $\mu_2 = (-1.5, 1.5, 1.5)$, $\mu_3 = (1.5, -1.5, -1.5)$, $\mu_4 = (-1.5, 1.5, -1.5)$, and variance matrices $\Sigma_k = \left(\rho_k^{|i-j|}\right)_{1\leq i,j\leq 3}$ with $\rho_1 = 0.85$, $\rho_2 = 0.1$, $\rho_3 = 0.65$ and $\rho_4 = 0.5$. Four redundant variables simulated from

$$\mathbf{x}_i^{\{4-7\}} \sim \mathcal{N}\left(\mathbf{x}_i^{\{1,3\}} \begin{pmatrix} 1 & 0 & -1 & 2 \\ 0 & -2 & 2 & 1 \end{pmatrix}; I_4\right)$$

and nine independent variables are appended, sampled from $\mathbf{x}_i^{\{8-16\}} \sim \mathcal{N}(\gamma, \tau)$ with

$$\gamma = (-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2)$$

and

$$\tau = \text{diag}(0.5, 0.75, 1, 1.25, 1.5, 1.25, 1, 0.75, 0.5).$$

A total of 100 simulation replications are considered where the training sample is composed of $n = 500$ data points and the same test sample with 50 000 points is used. The LDA, QDA and EDDA methods with and without variable selection are compared according to the averaged classification error rates.

Results summarized in Table 1 show that the variable selection procedure allows to improve the classification performance of LDA, QDA and EDDA. In particular, QDA becomes superior to LDA with variable selection. In all replications, only the first three variables are declared discriminant and the redundant variables are detected. When the true variable partition is not selected, it is due to one independent variable (21, 21 and 22 times for EDDA, QDA and LDA respectively) or two independent variables (4, 4 and 5 times for EDDA, QDA and LDA) which are declared redundant.

7.2. The Landsat Satellite Data

The Landsat Satellite Data, available at the UCI Machine Learning Repository (see http://www.ics.uci.edu/~mlearn/) is considered. This data set consists of the multi-spectral values of pixels in a tiny sub-area of a satellite image. Each line is a vector of length $Q = 36$, composed of the pixel values in four spectral bands (two in the visible region and two in the near infrared) of each of the 9 pixels in the 3 × 3 neighborhood. These data points are split into six classes. The original data set has already been divided into a training set with 4435 samples and a testing set with 2000 samples. The same experiment conditions than in [23] are considered: 1000 samples (randomly selected from the training data) are used to estimate and select the model, and this experiment is randomly replicated 100 times. Only QDA and LDA are considered in this study.

According to Table 2, QDA and LDA perform the same without variable selection, while QDA outperforms LDA with variable selection. In all replications, our variable selection procedure selects the QDA model ($\hat{m} = [L_k C_k]$), and all the

Table 2
Averaged classification error rate (±standard deviation) for LDA and QDA methods, without and with variable selection for the Landsat Satellite Data.

With variable selection		Without variable selection	
LDA	QDA	LDA	QDA
21 (±0.53)	16.21 (±0.68)	18.05 (±0.48)	17.90 (±0.57)

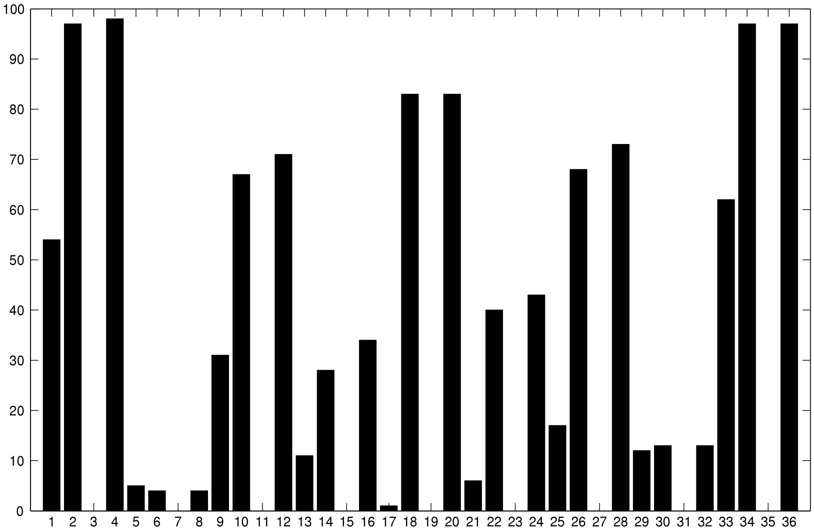

Fig. 2. Number of times each variable is declared discriminant among 100 replications for the Landsat Satellite Data.

irrelevant classification variables are redundant ($\hat{W} = \emptyset$) and regressed on all discriminant variables ($\hat{R} = \hat{S}$) with a general covariance matrix structure ($\hat{r} = [LC]$).

It is noteworthy that QDA and LDA select the same variables in the 100 replications, with an average selection of 12 discriminant variables as in [23]. It is worth mentioning some variable selection tendencies (see Fig. 2). First, variables tend to be selected by couple: for instance Variables 34 and 36, Variables 18 and 20, and Variables 2 and 4 are both declared discriminant in the same replications. Second, we can note that Variables 3, 7, 11, 15, 19, 23, 27, 31 and 35, corresponding to one measure in the near infrared for each pixel of the 3 × 3 neighborhood are never declared discriminant. Third, the variables corresponding to Pixels 1, 3, 5, 7 and 9 in the 3 × 3 neighborhood are more often declared discriminant than the one of the other pixels, certainly because these pixels have more neighbor pixels in common.

7.3. Leukemia data set

These data come from a study of gene expression in two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) published by Golub et al. [7]. Gene expression levels were measured for 47 ALL tumor samples and 25 AML tumor samples, using Affymetrix high-density oligonucleotide arrays containing 6817 human genes. After the pre-processing steps, as image analysis, standardization and some gene filtering, $Q = 3571$ genes are conserved. The interest of this data set is the large number of genes describing the samples, and the importance to detect the genes whose expression pattern separate the two types of leukemias. This data set is known as a benchmark data set and numerous results are available [7,20,10,11,21]. It is considered from two points of view in this section. First, the relevance and stability of our variable selection procedure are measured using a leave-one-out procedure for LDA and QDA models. Second, the interest of variable selection to improve the prediction accuracy is assessed using the training and test samples used by Golub et al. [7].

In the leave-one-out procedure for LDA and QDA, the averaged classification error rate is equal to zero and so as good as the error rates of different methods already applied (see [11]). For LDA, 3529 genes are never declared discriminant for the classification by our variable selection procedure and 44 of them are always declared independent. Among the 42 genes declared at least once discriminant, seven genes (Macmarcks, CD33 antigen, CST3, DF D, CCND3, GLUTATHIONE S-TRANSFERASE and RABAPTIN-5 protein) are already known to be discriminant and three (PLZF, Adrenal-Specific Protein Pg2 and, PRSS1) are known to be implicated in cancers. The first two genes declared the most time relevant are DF D (67 times) and GLUTATHIONE S-TRANSFERASE (53 times); the other ones are more declared redundant than relevant. We also compare our results with the list of [20] which contains 100 genes reported as discriminant by at least one discriminant method. A lack of precision in the gene names leads to consider 88 genes among these 100 genes. According to our method,

Table 3
Variable selection and misclassification error rate. The first four lines indicate the number of variables in S, R, U and W sets. The last line gives the number of misclassified test observations according to the two leukemia types ALL and AML.

Models	LDA	QDA	$[L_kC]$
Card($\hat{S}$)	8	8	3
Card($\hat{R}$)	2	2	3
Card($\hat{U}$)	3058	2848	1912
Card($\hat{W}$)	505	715	1656
Misc. test obs. (ALL, AML)	(2, 4)	(0, 0)	(0, 0)

83 genes are mainly declared redundant and six of them (Macmarcks, CD33 antigen, CST3, DF D (adipsin), CCND3 and GLUTATHIONE S-TRANSFERASE) at least once relevant. The five remaining genes are declared at least once independent (W) and never discriminant for the classification. Among these five genes, three are very often in W (more than 52 times) but are identified in the Su list by only the t-test procedure. They may be false candidates. The two others (Lamp2 and GLYCYLPEPTIDE TETRADECANOYLTRANSFERASE) are declared redundant 67 and 60 times and otherwise independent. Their status is thus more redundant than independent and it is coherent with their presence in the Su list. We do not report in detail the results in QDA which are quite analogous. But it is worthwhile to remark that the number of discriminant variables selected at least one time is 122 for QDA instead of 32 for LDA. It indicates a greater stability of the variable selection procedure for the simpler model LDA.

We then analyze the Leukemia data set using 38 (27 are ALL and 11 are AML) samples in the training and 34 (20 are ALL and 14 are AML) samples in the test. Results of our method are given in Table 3 and are compared with the performance of other methods given in [21]. Despite the variable selection, LDA performs poorly, on the contrary the quadratic methods (QDA and $[L_kC]$) are greatly improved by variable selection leading to zero misclassified test observation. Moreover, this is achieved with a small number of discriminant variables especially for the parsimonious model $[L_kC]$ (see Appendix A).

8. Discussion

We have proposed a variable selection methodology for a large family of Gaussian generative models in discriminant analysis. Regarding the problem as a model selection problem, we have proposed a BIC-like criterion to distinguish between discriminant, redundant and noisy variables. We proved the identifiability of our model collection and the consistency of the proposed BIC-like criterion. A procedure embedding two forward stepwise variable selection algorithms for classification and regression has been defined. Numerical experiments highlight the potentially great interest of our variable selection procedure to improve the classification performances of nonlinear Gaussian classification models. Actually those models involve many parameters when the number of variables is large with respect to the training sample size. But our variable selection procedure allows us to overcome the dimensionality problem leading to powerful classifiers with a nice interpretation of variable roles. Those results confirm the promising performances obtained by Murphy et al. [17] and Zhang and Wang [23] with less general methods. Our opinion is that our methodology is able to make the nonlinear generative classification methods such as quadratic discriminant analysis much more efficient in high dimensional contexts and competitive with gold standard classifiers such as LDA, logistic regression, k-nearest neighbor classifier or support vector classifiers in many situations.

Appendix A. The different model forms

Table A.4 gives the list of the 14 different model forms available in the MIXMOD software.

Appendix B. Proof of the model identifiability

Theorem 1 concerning the model identifiability can be proved quickly as below using the SRUW model identifiability established in [14] in the model-based clustering context. It is also possible to prove this theorem in the discriminant analysis context: the complete proof is available in [15].

Proof. Let $(m, r, l, \mathbf{V})$ and $(m^\star, r^\star, l^\star, \mathbf{V}^\star)$ be two models. Let $\theta \in \Theta_{(m,r,l,\mathbf{V})}$ and $\theta^\star \in \Theta_{(m^\star,r^\star,l^\star,\mathbf{V}^\star)}$ two parameters vectors such that $\forall \mathbf{x} \in \mathbb{R}^Q,\ \forall z \in \{1,\dots,K\}$,

$$f(\mathbf{x}, z|m, r, l, \mathbf{V}, \theta) = f(\mathbf{x}, z|m^\star, r^\star, l^\star, \mathbf{V}^\star, \theta^\star).$$

It is equivalent to the following system: $\forall \mathbf{x} \in \mathbb{R}^Q,\ \forall k \in \{1,\dots,K\}$,

$$\begin{aligned} &p_k\Phi(\mathbf{x}^S|\mu_k, \Sigma_{k(m)})\,\Phi(\mathbf{x}^U|a + \mathbf{x}^R\beta, \Omega_{(r)})\,\Phi(\mathbf{x}^W|\gamma, \tau_{(l)}) \\ &\quad = p_k^\star\Phi(\mathbf{x}^{S^\star}|\mu_k^\star, \Sigma_{k(m^\star)}^\star)\,\Phi(\mathbf{x}^{U^\star}|a^\star + \mathbf{x}^{R^\star}\beta^\star, \Omega_{(r^\star)}^\star)\,\Phi(\mathbf{x}^{W^\star}|\gamma^\star, \tau_{(l^\star)}^\star). \end{aligned} \tag{B.1}$$

Table A.4
List of model forms available in MIXMOD.

Family	Model	Volume	Orientation	Shape
Spherical	$[LI]$	Equal	NA	Equal
	$[L_kI]$	Variable	NA	Equal
Diagonal	$[LB]$	Equal	Coordinate axes	Equal
	$[L_kB]$	Variable	Coordinate axes	Equal
	$[LB_k]$	Equal	Coordinate axes	Variable
	$[L_kB_k]$	Variable	Coordinate axes	Variable
General	$[LC]$	Equal	Equal	Equal
	$[L_kC]$	Variable	Equal	Equal
	$[LDA_kD]$	Equal	Equal	Variable
	$[L_kDA_kD]$	Variable	Equal	Variable
	$[LD_kAD_k]$	Equal	Variable	Equal
	$[L_kD_kAD_k]$	Variable	Variable	Equal
	$[LC_k]$	Equal	Variable	Variable
	$[L_kC_k]$	Variable	Variable	Variable

Summing the K previous equations, we obtain that

$$\begin{aligned}&\left\{\sum_{k=1}^{K} p_k\Phi(\mathbf{x}^S|\mu_k,\Sigma_{k(m)})\right\}\Phi(\mathbf{x}^U|a+\mathbf{x}^R\beta,\Omega_{(r)})\,\Phi(\mathbf{x}^W|\gamma,\tau_{(l)})\\ &\quad=\left\{\sum_{k=1}^{K} p_k^\star\Phi\left(\mathbf{x}^{S^\star}|\mu_k^\star,\Sigma_{k(m^\star)}^\star\right)\right\}\Phi(\mathbf{x}^{U^\star}|a^\star+\mathbf{x}^{R^\star}\beta^\star,\Omega_{(r^\star)}^\star)\,\Phi(\mathbf{x}^{W^\star}|\gamma^\star,\tau_{(l^\star)}^\star).\end{aligned}$$

Next, using the identifiability result established in [14] in the clustering framework with a fix number of components K, we obtain that $\mathbf{V}=\mathbf{V}^\star$, $m=m^\star$, $r=r^\star$, $l=l^\star$, $a=a^\star$, $\beta=\beta^\star$, $\Omega_{(r)}=\Omega_{(r^\star)}^\star$ and the parameters $p_k, p_k^\star, \mu_k, \mu_k^\star$ and $\Sigma_{k(m)}^\star, \Sigma_{k(m^\star)}^\star$ are equal up to a permutation of Gaussian mixture components. But this permutation is the identity according to (B.1) thus $p_k=p_k^\star$, $\mu_k=\mu_k^\star$ and $\Sigma_{k(m)}=\Sigma_{k(m^\star)}^\star$ for all $k\in\{1,\ldots,K\}$. □

Appendix C. Proof of the criterion consistency theorem

This appendix is devoted to the proof of Theorem 2 given the criterion consistency.

Proof. According to the expressions (2) and (3), the selected model satisfies

$$(\hat{m},\hat{r},\hat{l},\hat{\mathbf{V}})=\underset{(m,r,l,\mathbf{V})\in\mathcal{N}}{\operatorname{argmax}}\ \operatorname{crit}(m,r,l,\mathbf{V})$$

with $\operatorname{crit}(m,r,l,\mathbf{V})=2\sum_{i=1}^n\ln[f(\mathbf{x}_i,z_i|m,r,l,\mathbf{V},\hat{\theta})]-\Xi_{(m,r,l,\mathbf{V})}\ln(n)$. Thus

$$\begin{aligned}P((\hat{m},\hat{r},\hat{l},\hat{\mathbf{V}})&=(m_0,r_0,l_0,\mathbf{V}_0))\\ &=P(\operatorname{crit}(m_0,r_0,l_0,\mathbf{V}_0)-\operatorname{crit}(m,r,l,\mathbf{V})\geq 0,\quad\forall\,(m,r,l,\mathbf{V})\in\mathcal{N}).\end{aligned}\tag{C.1}$$

Denoting $\Delta\operatorname{crit}(m,r,l,\mathbf{V})=\operatorname{crit}(m_0,r_0,l_0,\mathbf{V}_0)-\operatorname{crit}(m,r,l,\mathbf{V})$, we get

$$\begin{aligned}\Delta\operatorname{crit}(m,r,l,\mathbf{V})&=2n\left[\frac{1}{n}\sum_{i=1}^n\ln\left\{\frac{f(\mathbf{x}_i,z_i|m_0,r_0,l_0,\mathbf{V}_0,\hat{\theta})}{h(\mathbf{x}_i,z_i)}\right\}-\frac{1}{n}\sum_{i=1}^n\ln\left\{\frac{f(\mathbf{x}_i,z_i|m,r,l,\mathbf{V},\hat{\theta})}{h(\mathbf{x}_i,z_i)}\right\}\right]\\ &\quad+\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n).\end{aligned}\tag{C.2}$$

Let $\mathcal{N}_1=\{(m,r,l,\mathbf{V})\in\mathcal{N};\ \operatorname{KL}[h,f(.|m,r,l,\mathbf{V},\theta^\star)]\neq 0\}$. We have that $\mathcal{N}_1=\mathcal{N}\setminus\{(m_0,r_0,l_0,\mathbf{V}_0)\}$ since if $\operatorname{KL}[h,f(.|m,r,l,\mathbf{V},\theta^\star)]=0$ then the true density function $h=f(.|m_0,r_0,l_0,\mathbf{V}_0,\theta^\star)=f(.|m,r,l,\mathbf{V},\theta^\star)$ and according to the model identifiability, $(m_0,r_0,l_0,\mathbf{V}_0)=(m,r,l,\mathbf{V})$. Thus from (C.1), the theorem is established if it is proved that

$$\forall\,(m,r,l,\mathbf{V})\in\mathcal{N}_1,\quad P\left(\Delta\operatorname{crit}(m,r,l,\mathbf{V})<0\right)\underset{n\to\infty}{\to}0.\tag{C.3}$$

Let $(m,r,l,\mathbf{V})\in\mathcal{N}_1$. Denoting $\mathbb{M}_n(m,r,l,\mathbf{V})=\frac{1}{n}\sum_{i=1}^n\ln\left\{\frac{f(\mathbf{x}_i,z_i|m,r,l,\mathbf{V},\hat{\theta})}{h(\mathbf{x}_i,z_i)}\right\}$ and $M(m,r,l,\mathbf{V})=-\operatorname{KL}[h,f(.|m,r,l,\mathbf{V},\theta^\star)]$, from (C.2) we have

$$\begin{aligned}P(\Delta\text{crit}(m,r,l,\mathbf{V})<0) &= P\left(2n\{\mathbb{M}_n(m_0,r_0,l_0,\mathbf{V}_0)-\mathbb{M}_n(m,r,l,\mathbf{V})\}+\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n)<0\right)\\ &= P\Bigg(\mathbb{M}_n(m_0,r_0,l_0,\mathbf{V}_0)-M(m_0,r_0,l_0,\mathbf{V}_0)+M(m,r,l,\mathbf{V})-\mathbb{M}_n(m,r,l,\mathbf{V})\\ &\quad + M(m_0,r_0,l_0,\mathbf{V}_0)-M(m,r,l,\mathbf{V})+\frac{\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n)}{2n}<0\Bigg).\end{aligned}$$

Thus, using the property that for two real random variables A and B and for all $u\in\mathbb{R}$,

$$P(A+B\le 0)\le P(A\le u)+P(-B>u)$$

we get that for all $\epsilon>0$,

$$\begin{aligned}P(\Delta\text{crit}(m,r,l,\mathbf{V})<0) &\le P(M(m_0,r_0,l_0,\mathbf{V}_0)-\mathbb{M}_n(m_0,r_0,l_0,\mathbf{V}_0)>\epsilon)+P(\mathbb{M}_n(m,r,l,\mathbf{V})-M(m,r,l,\mathbf{V})>\epsilon)\\ &\quad +P\left(M(m_0,r_0,l_0,\mathbf{V}_0)-M(m,r,l,\mathbf{V})+\frac{\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n)}{2n}<2\epsilon\right).\end{aligned}$$

From Lemma 3, stated hereafter, $\forall(m,r,l,\mathbf{V})\in\mathcal{N}$, $\mathbb{M}_n(m,r,l,\mathbf{V})\underset{n\to\infty}{\overset{P}{\to}}M(m,r,l,\mathbf{V})$. Thus, for all $\epsilon>0$,

$$P(\mathbb{M}_n(m,r,l,\mathbf{V})-M(m,r,l,\mathbf{V})>\epsilon)\le P(|\mathbb{M}_n(m,r,l,\mathbf{V})-M(m,r,l,\mathbf{V})|>\epsilon)\underset{n\to\infty}{\to}0.$$

For the third term, note

$$\begin{aligned}&P\left(M(m_0,r_0,l_0,\mathbf{V}_0)-M(m,r,l,\mathbf{V})+\frac{\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n)}{2n}<2\epsilon\right)\\ &\quad\le P\left(M(m_0,r_0,l_0,\mathbf{V}_0)-M(m,r,l,\mathbf{V})-2\epsilon<\left|\frac{\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n)}{2n}\right|\right).\end{aligned}$$

Since $M(m_0,r_0,l_0,\mathbf{V}_0)-M(m,r,l,\mathbf{V})>0$ because $(m,r,l,\mathbf{V})\in\mathcal{N}_1$ and

$$\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n)/2n\underset{n\to\infty}{\to}0,$$

taking $\epsilon=\{M(m_0,r_0,l_0,\mathbf{V}_0)-M(m,r,l,\mathbf{V})\}/4>0$, we get

$$\begin{aligned}&P\left(M(m_0,r_0,l_0,\mathbf{V}_0)-M(m,r,l,\mathbf{V})+\frac{\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n)}{2n}<2\epsilon\right)\\ &\quad\le P\left(\frac{M(m_0,r_0,l_0,\mathbf{V}_0)-M(m,r,l,\mathbf{V})}{2}<\left|\frac{\left[\Xi_{(m,r,l,\mathbf{V})}-\Xi_{(m_0,r_0,l_0,\mathbf{V}_0)}\right]\ln(n)}{2n}\right|\right)\underset{n\to\infty}{\to}0.\end{aligned}$$

Finally, $P(\Delta\text{crit}(m,r,l,\mathbf{V})<0)\underset{n\to\infty}{\to}0$. □

Lemma 3. *Under assumptions (H1) and (H2),*

$$\forall\,(m,r,l,\mathbf{V})\in\mathcal{N},\quad \frac{1}{n}\sum_{i=1}^{n}\ln\left[\frac{h(\mathbf{x}_i,z_i)}{f(\mathbf{x}_i,z_i|m,r,l,\mathbf{V},\hat{\theta})}\right]\underset{n\to\infty}{\overset{P}{\to}}KL[h,f(.|m,r,l,\mathbf{V},\theta^{\star})].$$

Proof. A complete proof can be found in [15]. It essentially applies the following Proposition 4 to the family $\mathcal{F}_{(m,r,l,\mathbf{V})}:=\{\ln[f(.|m,r,l,\mathbf{V},\theta)];\theta\in\Theta'_{(m,r,l,\mathbf{V})}\}$. We point out that Assumption (H2) is required to ensure the boundedness inequalities:

$$\begin{aligned}\ln(\rho)-\frac{\text{Card}(S)}{2}\ln[2\pi s_{\text{M}}]-\frac{\|\mathbf{x}\|^2+\eta^2}{s_{\text{m}}} &\le \ln\left[\sum_{k=1}^{K}p_k\Phi(\mathbf{x}^S|\mu_k,\Sigma_{k(m)})\mathbb{I}_{z=k}\right]\\ &\le -\frac{\text{Card}(S)}{2}\ln\left[2\pi s_{\text{m}}\right],\\ -\frac{\text{Card}(U)}{2}\ln[2\pi s_{\text{M}}]-\frac{\eta^2}{s_{\text{m}}}-\frac{1+\eta^2}{s_{\text{m}}}\|\mathbf{x}\|^2 &\le \ln\left[\Phi(\mathbf{x}^U|a+\mathbf{x}^R\beta,\Omega_{(r)})\right]\\ &\le -\frac{\text{Card}(U)}{2}\ln[2\pi s_{\text{m}}],\end{aligned}$$

and

$$-\frac{\mathrm{Card}(W)}{2}\ln[2\pi s_{\mathrm{M}}] - \frac{(\|\mathbf{x}\|^2+\eta^2)}{s_{\mathrm{m}}} \leq \ln\left[\Phi(\mathbf{x}^W|\gamma,\tau_{(l)})\right] \leq -\frac{\mathrm{Card}(W)}{2}\ln[2\pi s_{\mathrm{m}}]$$

from which it is easily verified that there exists a h-integrable envelop function of $\mathcal{F}_{(m,r,l,\mathbf{V})}$. □

Proposition 4. *Assume that*

1. $(Y_1,\ldots,Y_n)$ *is a n-sample with unknown density h.*
2. Θ *is a compact metric space.*
3. $\theta \in \Theta \mapsto \ln[f(\mathbf{y}|\theta)]$ *is continuous for every* $\mathbf{y} \in \mathbb{R}^Q$.
4. *F is an envelope function of* $\mathcal{F} := \{\ln[f(.|\theta)];\, \theta \in \Theta\}$ *which is h-integrable.*
5. $\theta^\star = \underset{\theta\in\Theta}{argmin}\, KL[h, f(.|\theta)]$
6. $\hat{\theta} = \underset{\theta\in\Theta}{argmax} \sum_{i=1}^n f(Y_i|\theta)$.

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

This proposition is proved in [15].

Appendix D. The forward variable selection in regression (FSAR)

The following algorithm allows us to determine the subset $R[u]$ of variables among S required to explain $\mathbf{x}^u$ with a linear regression, u being a set of redundant variables. The model comparison is performed with criterion $\mathrm{BIC}_{\mathrm{reg}}$ defined in (4). The algorithm is making use of the inclusion and exclusion steps now described.

Initialization $R[u] = \emptyset$, $j_E = \emptyset$ and $j_I = \emptyset$.

Inclusion step For all j in $S \setminus R[u]$, compute

$$\mathrm{B}_{\mathrm{diffreg}}(j) = \mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^u|r, \underline{\mathbf{x}}^{R[u]\cup j}) - \mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^u|r, \underline{\mathbf{x}}^{R[u]}).$$

Then, compute $j_I = \underset{j\in S-R[u]}{\mathrm{argmax}}\, \mathrm{B}_{\mathrm{diffreg}}(j)$.

- If $\mathrm{B}_{\mathrm{diffreg}}(j_I) > 0$,
  - if $j_I = j_E$, stop
  - otherwise, $R[u] = R[u] \cup j_I$ and go to the exclusion step.
- Otherwise, $j_I = \emptyset$. If $j_E \neq \emptyset$, go to the exclusion step and stop otherwise.

Exclusion step For all j in $R[u]$, compute

$$\mathrm{B}_{\mathrm{diffreg}}(j) = \mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^u|r, \underline{\mathbf{x}}^{R[u]}) - \mathrm{BIC}_{\mathrm{reg}}(\underline{\mathbf{x}}^u|r, \underline{\mathbf{x}}^{R[u]-j}).$$

Then, compute $j_E = \underset{j\in R[u]}{\mathrm{argmin}}\, \mathrm{BIC}_{\mathrm{diffreg}}(j)$.

- If $\mathrm{B}_{\mathrm{diffreg}}(j_E) \leq 0$, set $R[u] = R[u] - j_E$ and go to the inclusion step if $j_E \neq j_I$ or stop otherwise.
- otherwise, $j_E = \emptyset$ and go to the inclusion step.

Starting from the inclusion step, the forward variable selection algorithm consists of alternating the inclusion and exclusion steps.

References

[1] J.D. Banfield, A.E. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics 49 (1993) 803–821.
[2] H. Bensmail, G. Celeux, Regularized Gaussian discriminant analysis through eignenvalue decomposition, Journal of the American Statistical Association 91 (1996) 1743–1748.
[3] C. Biernacki, G. Celeux, G. Govaert, F. Langrognet, Model-based cluster and discriminant analysis with the MIXMOD software, Computational Statistics and Data Analysis 51 (2006) 587–600.
[4] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[5] G. Celeux, G. Govaert, Gaussian parsimonious clustering models, Pattern Recognition 28 (1995) 781–793.
[6] C. Fraley, A.E. Raftery, Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST, Journal of Classification 20 (2003) 263–286.
[7] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[8] I. Guyon, A. Ellisseeff, An introduction to variable and feature selection, Journal of Machine Learning research 3 (2003) 1157–1182.
[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Second Ed., Springer, New York, 2009.
[10] B. Krishnapuram, L. Carin, A. Hartemink, Gene expression analysis: Joint feature selection and classifier design, in: Kernel Methods in Computational Biology, MIT Press, Cambridge, MA, 2004.
[11] T. Mary-Huard, S. Robin, Tailored aggregation for classification, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 2098–2105.

[12] T. Mary-Huard, S. Robin, J.J. Daudin, A penalized criterion for variable selection in classification, Journal of Multivariate Analysis 98 (2007) 695–705.

[13] C. Maugis, G. Celeux, M.L. Martin-Magniette, Variable Selection for Clustering with Gaussian mixture models, Biometrics 65 (2009) 701–709.

[14] C. Maugis, G. Celeux, M.L. Martin-Magniette, variable selection in model-based clustering: A general variable role modeling, Computational Statistics and Data Analysis 53 (2009) 3872–3882.

[15] C. Maugis, G. Celeux, M.L. Martin-Magniette, Variable selection in model-based discriminant analysis, Technical Report RR-7290, INRIA, 2010.

[16] G. McLachlan, Discriminant Analysis and Statistical Pattern Analysis, Wiley-Interscience, New York, 1992.

[17] B.T. Murphy, A.E. Raftery, N. Dean, Variable selection and updating in model-based discriminant analysis for high-dimensional data with food authenticity applications, Annals of Applied Statistics 4 (2010) 396–421.

[18] A.E. Raftery, N. Dean, variable selection for model-based clustering, Journal of the American Statistical Association 101 (2006) 168–178.

[19] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (1978) 461–464.

[20] Y. Su, T. Murali, V. Pavlovic, M. Schaffer, S. Kasif, Rank Gene: Identification of diagnostic genes based on expression data, Bioinformatics 19 (2003) 1578–1579.

[21] A.J. Yang, S. Xin-Yuan, Bayesian variable selection for disease classification using gene expression data, Bioinformatics 26 (2010) 215–222.

[22] D.M. Young, P.L. Odell, Feature-subset selection for statistical classification problems involving unequal covariance matrices, Communication in Statistics–Theory and Methods 15 (1986) 137–157.

[23] Q. Zhang, H. Wang, A bic criterion for gaussian mixture model selection with application in discriminant analysis, Technical Report, Guanghua School of Management, Peking University, 2008.