Regularized Gaussian Discriminant Analysis through Eigenvalue Decomposition

Halima Bensmail
Université Paris 6

Gilles Celeux
INRIA Rhône-Alpes

Abstract

Friedman (1989) has proposed a regularization technique (RDA) of discriminant analysis in the Gaussian framework. RDA makes use of two regularization parameters to design an intermediate classification rule between linear and quadratic discriminant analysis. In this paper, we propose an alternative approach to design classification rules which have also a median position between linear and quadratic discriminant analysis. Our approach is based on the reparametrization of the covariance matrix Σ_k of a group G_k in terms of its eigenvalue decomposition, $\Sigma_k = \lambda_k D_k A_k D_k'$ where λ_k specifies the volume of G_k, A_k its shape, and D_k its orientation. Variations on constraints concerning λ_k, A_k and D_k lead to 14 discrimination models of interest. For each model, we derived the maximum likelihood parameter estimates and our approach consists in selecting the model among the 14 possible models by minimizing the sample-based estimate of future misclassification risk by cross-validation. Numerical experiments show favorable behavior of this approach as compared to RDA.

Keywords: *Gaussian Classification, Regularization, Eigenvalue Decomposition, Maximum Likelihood.*

1 Introduction

The basic problem in discriminant analysis is to assign an unknown subject to one of K groups $G_1, \ldots, G_K$ on the basis of a multivariate observation $\mathbf{x} = (x_1, ..., x_d)'$, d denoting the number of variables. The assignment function is generally designed to minimize the expected overall error rate and consists in assigning a measurement vector $\mathbf{x}$ to the group G_k such that

$$k = \arg \max_{1 \leq j \leq K} \pi_j f_j(\mathbf{x}), \tag{1.1}$$

π_k denoting the a priori probability of belonging to group G_k and $f_k(\mathbf{x})$ denoting the group conditional density of $\mathbf{x}, (1 \leq k \leq K)$. Discriminant analysis models differ essentially by their assumptions on the group conditional densities $f_k(\mathbf{x})$, $(k = 1, \ldots, K)$. The most often applied model, the linear discriminant analysis (LDA) assumed that the group conditional distributions are d-variate normal distributions with mean vectors μ_k and identical variance matrix Σ. When the variance matrices Σ_k are not assumed to be equal, the model is called quadratic discriminant analysis (QDA). The parameters μ_k and Σ_k are usually unknown and must be estimated from a training set consisting in $(\mathbf{x}_i, z_i), i = 1, \ldots, n$, where $\mathbf{x}_i$ is the vector-valued measurement and z_i is the group of subject i. The parameters are generally chosen to maximize the likelihood of the training sample. Its lead to the plug-in estimates

$$\hat{\mu}_k = \bar{\mathbf{x}}_k = \frac{\sum_{i/z_i=k} \mathbf{x}_i}{n_k}, \ k = 1, \ldots, K, \tag{1.2}$$

where $n_k = \sum_{i=1}^{n} \mathbf{I}\{z_i = k\}$. And, for LDA,

$$\hat{\Sigma} = S = \frac{\sum_{k=1}^{K} \sum_{i/z_i=k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'}{n}, \tag{1.3}$$

or, for QDA,

$$\hat{\Sigma_k} = S_k = \frac{\sum_{i/z_i=1} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'}{n_k} \ (k = 1, \ldots, K). \tag{1.4}$$

Regularization became an important subject of investigation in discriminant analysis since in many cases the size n of the training data set is small in regard to the number d of variables (see McLachlan 1992), and standard methods such as QDA or even LDA can have a disappointing behavior in such cases. Generally, regularization techniques for discriminant analysis make use of real valued regularization parameters. For instance, one of the most employed regularization technique, the Regularized Discriminant Analysis (RDA) of Friedman (1989) specify the value of a complexity parameter and of a shrinkage parameter to design an intermediate classification rule between linear and quadratic discriminant analysis. RDA performs well but do not provide easy interpretable classification rules.

In this paper, we propose an alternative approach to design regularized classification rules in the Gaussian framework. Following Banfiled and Raftery (1993) and Flury *et al.* (1994), our approach is based on the reparametrization of the covariance matrix Σ_k of a group G_k in terms of its eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{1.5}$$

where $\lambda_k = |\Sigma_k|^{1/d}, D_k$ is the matrix of eigenvectors of Σ_k and A_k is 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 group G_k, 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 Gaussian discriminant

models. Variations on assumptions on the parameters λ_k, D_k and A_k $(1 \leq k \leq K)$ lead to 8 general models of interest. For instance, we can assume different volumes and keep the shapes and orientations equal by requiring that $A_k = A$ (A unknown) and $D_k = D$ (D unknown) for $k = 1, \ldots, K$. We denote this model $[\lambda_k DAD']$. With this convention, writing (for instance) $[\lambda D_k A D_k']$ means that we consider a discriminant model with equal volumes, equal shapes and different orientations.
Moreover 2 other families of situation are of interest. The first one consists in assuming that the variance matrices Σ_k are diagonal matrices. In the considered parametrization, it means that the orientation matrices D_k are permutation matrices. Since, in such a case, it does not seem that variations on the shape matrices are of any interest, we write $\Sigma_k = \lambda_k B_k$ where B_k is a diagonal matrix with $|B_k| = 1$. This particular parametrization gives rise to 4 models ($[\lambda B], [\lambda_k B], [\lambda B_k]$ and $[\lambda_k B_k]$). The second family of models consists in shrinking discriminant models by assuming spherical shapes, namely $A_k = I$, I denoting the identity matrix. In such a case, two parsimonious models are in competition: $[\lambda I]$ and $[\lambda_k I]$. Finally, we get 14 different discriminant models.

The method, that we propose and that we called EDRDA (Eigenvalue Decomposition Regularized Discriminant Analysis), consists in selecting the m.l. estimated model among the 14 above mentioned models which minimizes the sample-based estimate of future misclassification risk by cross-validation.

Remark 1: The main motivation of EDRDA is to provide a regularized classification rule easily interpreted, since it can be analyzed from the volumes, the shapes and the orientations of the groups.

Remark 2: Our selection procedure (the cross-validated error rate) has been proved to provide good performances for selecting models in discriminant analysis (e.g. Friedman 1989).

Remark 3: EDRDS generalizes the approach of Flury *et al.* (1994) which analyzed the performance of models $[\lambda DAD']$, $[\lambda_k DAD']$, $[\lambda_k D A_k D']$, and $[\lambda_k D_k A_k D_k']$ and suggested to choose among different models with the cross-validated error rate.

Remark 4: AS for RDA (see Section 2), it often happens that several models provide the same cross-validated error rate. In such cases, we investigated, in the following numerical experiments, two strategies: the first one consists in selecting the most parsimonious model (*parsimonious* strategy) and the second one in selecting the most complex model (*complex* strategy). This point is discussed further in the comments of Section 4 and in Section 5.

In Section 2, we sketch RDA since this method can be regarded as a reference method of regularization in discriminant analysis. In Section 3, for each of the 14 above mentioned models from which the EDRDA classification rule is designed, we give the formulas for maximum likelihood (m.l.) estimation. In Section 4, we compare RDA

and EDRDA on the basis of Monte Carlo simulations and a short discussion section ends this paper.

2 Regularized Discriminant Analysis

The regularized discriminant analysis of Friedman (1989) makes use of a *complexity* parameter α and of a *shrinkage* parameter γ in the following way. RDA replaces the plug-in estimator (1.4) of Σ_k with

$$\hat{\Sigma}_k(\alpha, \gamma) = (1 - \gamma)\hat{\Sigma}_k(\alpha) + \frac{\gamma}{d}\mathrm{tr}[\hat{\Sigma}_k(\alpha)]I, \tag{2.6}$$

where

$$\hat{\Sigma}_k(\alpha) = \frac{(1 - \alpha)S_k + \alpha S}{(1 - \alpha)n_k + \alpha n} \tag{2.7}$$

.
Thus α ($0 \leq \alpha \leq 1$) controls the amount the S_k are shrunk towards S, while γ ($0 \leq \gamma \leq 1$) controls the shrinkage of the eigenvalues towards equality as $\mathrm{tr}[\hat{\Sigma}_k(\alpha)]/p$ is equal to the average of the eigenvalues of $\hat{\Sigma}_k(\alpha)$.

The parameters in (2.6) and (2.7) are chosen to minimize jointly the cross-validated error rate. Friedman proceeds in the following way. A grid of candidate (α, γ)-pair is first selected on the unit square. The cross-validation is then employed to obtain a (nearly) unbiased estimate of the overall error rate for the discriminant rule associated with each (α, γ)-pair on the grid. Then, RDA chooses the point (α, γ) with the smallest estimated error rate.
A characteristic of the RDA approach, pointed out by Rayens and Greene (1991), is that the optimal value of the cross-validated error rate rarely occurs at a single point of the grid, but for a large range of values of (α, γ). The RDA procedure resolve ties by selecting first the points with the largest value of α (parsimonious principle), and then the point with the largest value of γ (shrinking principle).
RDA provides a fairly rich class of regularization alternatives. Holding γ fixed at 0 and varying α produces models between QDA and LDA. While, holding α fixed at 0 and increasing γ attempts to unbias the sample-based eigenvalues estimates. Holding α fixed at 1 and increasing γ gives rise to the ridge-regression analog for LDA. The reported experiments in Friedman (1989) showed that RDA performs well in many circumstances as compared with LDA and QDA.

However, the resulting classification rule can have no clear interpretation especially when both parameters are far from the boundaries of $[0, 1] \times [0, 1]$.

3 Maximum likelihood estimation of the models

Table 1 summarizes some features of the 14 models considered by EDRDA. In this table, the first column specifies the model. The second column gives the number of

Table 1: Some characteristics of the 14 models. We have $\alpha = Kd + K - 1$ and $\beta = \frac{d(d+1)}{2}$; CF means that the m.l. estimates are closed form, IP means that the m.l. estimation needs an iterative procedure.

model	number of parameters	m.l. est.
$[\lambda DAD']$	$\alpha + \beta$	CF
$[\lambda_k DAD']$	$\alpha + \beta + K - 1$	IP
$[\lambda DA_k D']$	$\alpha + \beta + (K-1)(d-1)$	IP
$[\lambda_k DA_k D']$	$\alpha + \beta + (K-1)d$	IP
$[\lambda D_k AD'_k]$	$\alpha + K\beta - (K-1)d$	CF
$[\lambda_k D_k AD'_k]$	$\alpha + K\beta - (K-1)(d-1)$	IP
$[\lambda D_k A_k D'_k]$	$\alpha + K\beta - (K-1)$	CF
$[\lambda_k D_k A_k D'_k]$	$\alpha + K\beta$	CF
$[\lambda B]$	$\alpha + d$	CF
$[\lambda_k B]$	$\alpha + d + K - 1$	IP
$[\lambda B_k]$	$\alpha + Kd - K + 1$	CF
$[\lambda_k B_k]$	$\alpha + Kd$	CF
$[\lambda I]$	$\alpha + 1$	CF
$[\lambda_k I]$	$\alpha + K$	CF

parameters to be estimated. The third column indicates if the m.l. estimates can be achieved with closed form formulas (CF) or if there is the need to make use of an iterative procedure (IP).
For each model, the m.l. estimation of the group mean vectors (μ_k, $k = 1, \ldots, K$) is

$$\hat{\mu}_k = \bar{\mathbf{x}}_k = \frac{\sum_{i/z_i=k} \mathbf{x}_i}{n_k} \tag{3.8}$$

where $n_k = \#G_k$ in the learning sample.

The m.l. estimation of the variance matrices of the groups depends on the model at hand. In some cases, it leads to closed form formulas but most of the time there is a need to use an iterative procedure to derive m.l. estimates. And, in some circumstances, especially for models assuming different shape group variance matrices, designing these algorithms need some effort. In this section, we do not provide details on the m.l. calculations, since those details appear in a paper of Celeux and Govaert (1994) where the same models were considered in a cluster analysis context. In the following, we only give the formulas of m.l. estimators of the variance matrices for the 14 models. First, we need to define some matrices: The within group scattering matrix W

$$W = \sum_{k=1}^{K} \sum_{i/z_i=k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)' \tag{3.9}$$

and, the scattering matrix W_k of group G_k $(1 \leq k \leq K)$

$$W_k = \sum_{i/z_i=1} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'. \ (k = 1, \ldots, K). \tag{3.10}$$

Model $[\lambda DAD']$. This is the classical linear discriminant analysis model. This model is obtained with $\alpha = 1$ and $\gamma = 0$ in the RDA scheme. The common variance matrix Σ is estimated by

$$\hat{\Sigma} = \frac{W}{n}.$$

Model $[\lambda_k DAD']$. In this situation, it is convenient to write $\Sigma_k = \lambda_k C$ with $C = DAD'$. This model has been considered and called the *proportional covariance matrices* model by Flury (1988). The estimation of the λ_k's and C need an iterative procedure.

- As the matrix C is kept fixed, the λ_k's are solution of the equations $(1 \leq k \leq K)$

$$\lambda_k = \frac{\mathrm{tr}(W_k C^{-1})}{dn_k}.$$

- As the volumes λ_k's are kept fixed, the matrix C maximizing the likelihood is

$$C = \frac{\sum_{k=1}^{K} \frac{1}{\lambda_k} W_k}{|\sum_{k=1}^{K} \frac{1}{\lambda_k} W_k|^{\frac{1}{d}}}.$$

Model $[\lambda DA_k D']$. In this situation and in the next one, there is no interest to assume that the terms of the diagonal matrices A_k are in decreasing order. Thus for the models $[\lambda DA_k D']$ and $[\lambda_k DA_k D']$ we do not assume that the diagonal terms of A_k are in decreasing order. The m.l. estimates of λ, D and $(A_k, k = 1, \ldots . K)$ are derived using an iterative method, that we describe hereunder, and by a direct calculation of λ,

$$\lambda = \frac{\sum_{k=1}^{K} \mathrm{tr}(DA_k^{-1} D' W_k)}{nd}.$$

- For fixed D, compute

$$A_k = \frac{\mathrm{diag}(D' W_k D)}{|\mathrm{diag}(D' W_k D)|^{\frac{1}{d}}}$$

  where $\mathrm{diag}(M)$ denotes the diagonal matrix which has the same diagonal as the matrix M.

- For fixed $A_1, \ldots, A_K$, D is obtained using an adaptation of an algorithm of Flury and Gautschi (1986):
  Starting from an initial solution $D = (\mathbf{d}_1, \ldots, \mathbf{d}_d)$, for any couple $(\ell, m)(\ell \neq m) \in \{1, \ldots, d\}$, the couple $(\mathbf{d}_\ell, \mathbf{d}_m)$ is replaced with (δ_ℓ, δ_m) where δ_ℓ and δ_m are orthonormal vectors, linear combination of $\mathbf{d}_\ell$ and $\mathbf{d}_m$, such that

$$\delta_\ell = (\mathbf{d}_\ell, \mathbf{d}_m)\mathbf{q}_1 \text{ and } \delta_m = (\mathbf{d}_\ell, \mathbf{d}_m)\mathbf{q}_2$$

where $\mathbf{q}_1$ and $\mathbf{q}_2$ are two orthonormal vectors of $\mathbf{R}^2$ with $\mathbf{q}_1$ is the eigenvector associated to the smallest eigenvalue of the matrix $\sum_{k=1}^{K}(\frac{1}{a_k^\ell} - \frac{1}{a_k^m})Z_k$ with $Z_k = (\mathbf{d}_\ell, \mathbf{d}_m)'W_k(\mathbf{d}_\ell, \mathbf{d}_m)$.
This algorithm is repeated until it produces no increase of the likelihood.

Model $[\lambda_k D A_k D']$. In this situation, it is convenient to write $\Sigma_k = DA_kD'$ where $|A_k| = |\Sigma_k|$. This model has been considered and called the *common principal components* model by Flury (1984). The algorithm for deriving the m.l. estimates of $D, A_1, \ldots, A_K)$ is similar to the previous one:

- For fixed D, we get
$$A_k = \frac{1}{n_k}\mathrm{diag}(D'W_kD).$$
- For fixed $A_1, \ldots, A_K$, D is obtained using the same procedure as described for model $[\lambda D A_k D']$.

Model $[\lambda D_k A D_k']$. Considering for $k = 1, \ldots, K$ the eigenvalue decomposition $W_k = L_k\Omega_k L_k^{'}$ of the symmetric definite positive matrix W_k with the eigenvalues in the diagonal matrix Ω_k in decreasing order, we get

$$D_k = L_k \ k = 1, \ldots, K,$$

$$A = \frac{\sum_{k=1}^{K}\Omega_k}{|\sum_{k=1}^{K}\Omega_k|^{\frac{1}{d}}}$$

and

$$\lambda = \frac{|\sum_{k=1}^{K}\Omega_k|^{\frac{1}{d}}}{n}.$$

Model $[\lambda_k D_k A D_k']$. We use again the eigenvalue decomposition $W_k = L_k\Omega_k \mathrm{L}_k'$. Parameters λ_k, D_k and A are solutions of the equations, to be solved iteratively,

$$\lambda_k = \frac{\mathrm{tr}(W_k D_k A^{-1} D_k^{'})}{dn_k} \ (1 \leq k \leq K)$$

$$D_k = L_k \ (1 \leq k \leq K)$$

and

$$A = \frac{\sum_{k=1}^{K}\frac{1}{\lambda_k}\Omega_k}{|\sum_{k=1}^{K}\frac{1}{\lambda_k}\Omega_k|^{\frac{1}{d}}}.$$

Model $[\lambda D_k A_k D_k']$. In this situation, it is convenient to write $\Sigma_k = \lambda C_k$ where $C_k = D_k A_k D_k'$. We get

$$C_k = \frac{W_k}{|W_k|^{\frac{1}{d}}} \ (1 \leq k \leq K),$$

and

$$\lambda = \frac{\sum_{k=1}^{K}|W_k|^{\frac{1}{d}}}{n}.$$

Model $[\lambda_k D_k A_k D_k']$. This is the most general situation corresponding to ordinary quadratic discriminant analysis. This model is obtained with $\alpha = 0$ and $\gamma = 0$ in the RDA scheme. The m.l. estimates of variance matrices Σ_k are

$$\hat{\Sigma}_k = \frac{1}{n_k} W_k.$$

We now present the m.l. estimates for models with diagonal variance matrices. For this more parsimonious family of models, the eigenvectors of Σ_k $(1 \leq k \leq K)$ are the vectors generating the basis associated to the d variables $(D_k = J_k)$. If the J_k are equal, the variable are independent. If the J_k are different, the variables are independent conditionally to the groups to be classified.

Model $[\lambda B]$. In this situation, we get

$$B = \frac{\text{diag}(W)}{|\text{diag}(W)|^{\frac{1}{d}}}$$

and

$$\lambda = \frac{|\text{diag}(W)|^{\frac{1}{d}}}{n}.$$

Model $[\lambda_k B]$. In this situation, the m.l. estimates are derived from the following iterative procedure:

- As the matrix B is kept fixed, the λ_k's are

$$\lambda_k = \frac{\text{tr}(W_k B^{-1})}{d n_k} \ (1 \leq k \leq K).$$

- As the volumes λ_k's are kept fixed, the matrix B is

$$B = \frac{\text{diag}\left(\sum_{k=1}^{K} \frac{1}{\lambda_k} W_k\right)}{|\text{diag}\left(\sum_{k=1}^{K} \frac{1}{\lambda_k} W_k\right)|^{\frac{1}{d}}}.$$

Model $[\lambda B_k]$. In this situation, we get

$$B_k = \frac{\text{diag}(W_k)}{|\text{diag}(W_k)|^{\frac{1}{d}}} \ (1 \leq k \leq K)$$

and

$$\lambda = \frac{\sum_{k=1}^{K} |\text{diag}(W_k)|^{\frac{1}{d}}}{n}.$$

Model $[\lambda_k B_k]$. In this situation, we get

$$B_k = \frac{\text{diag}(W_k)}{|\text{diag}(W_k)|^{\frac{1}{d}}} \ (1 \leq k \leq K)$$

and

$$\lambda_k = \frac{|\mathrm{diag}(W_k)|^{\frac{1}{d}}}{n_k} \ (1 \leq k \leq K).$$

We consider now models for which the variance matrices are spherical. Two situations have to be considered: $\Sigma_k = \lambda I$ and $\Sigma_k = \lambda_k I$, I denoting the $(d \times d)$ identity matrix. We derive the m.l. estimations of the volumes of the groups for these models.

Model $[\lambda I]$. This model is obtained with $\alpha = 1$ and $\gamma = 1$ in the RDA scheme. It has been called the *nearest-means classifier* by Friedman (1989). In this situation, we get

$$\lambda = \frac{\mathrm{tr}(W)}{nd}.$$

Model $[\lambda_k I]$. In this situation, we get

$$\lambda_k = \frac{\mathrm{tr}(W_k)}{dn_k}.$$

4 Numerical experiments

We now present Monte Carlo simulations to compare RDA and EDRDA. We essentially used the same simulation scheme as Friedman (1989). We called D1-D5 the simulated data structures for dimensions $d = 6$ and $d = 20$ and sample size $n = 40$. For each data structure D1-D5, we randomly generated 100 replications and we ran RDA and EDRDA. The data structures are respectively corresponding to Tables 2-6 of Friedman's paper. Roughly speaking, D1 provides spherical groups with different volumes and means; D2 and D3 provide ellipsoidal groups with same shapes and orientations, with poorly separated means for D2 and well separated means for D3; D4 and D5 provide unequal ellipsoidal groups with equal means for D4 and different means for D5. More precisely the simulated distribution parameters were the following.

$$D1 \begin{cases} \mu_1' = (0, 0, 0, \ldots, 0) \\ \mu_2' = (0, 3, 0, \ldots, 0) \\ \mu_3' = (0, 0, 4, \ldots, 0) \\ \Sigma_1 = I \\ \Sigma_2 = 2I \\ \Sigma_3 = 3I \end{cases}$$

It can be remarked that data set D1 can be related to models $[\lambda_k I]$, $[\lambda_k B]$, and $[\lambda_k DAD']$.

The variances matrices for data sets D2 and D3 are identical. They are the same for each group and diagonal, with general diagonal term

$$a_j = [9(j-1)/(d-1) + 1]^2, \quad 1 \leq j \leq d.$$

For D2, the group mean vectors are

$$\begin{aligned}\mu_1' &= (0,\ldots,0)\\ \mu_{2j} &= 2.5\sqrt{a_j/d}\frac{d-j}{d/2-1}, 1\le j\le d\\ \mu_{3j} &= (-1)^j\mu_{2j},\quad 1\le j\le d.\end{aligned}$$

For D3, the group mean vectors are

$$\begin{aligned}\mu_1' &= (0,\ldots,0)\\ \mu_{2j} &= 2.5\sqrt{a_j/d}\frac{j-1}{d/2-1}, 1\le j\le d\\ \mu_{3j} &= (-1)^j\mu_{2j},\quad 1\le j\le d.\end{aligned}$$

Data set D2 and D3 are related to models $[\lambda B]$ and $[\lambda DAD^t]$.

The variance matrices for data sets D4 and D5 are identical. They are diagonal but different for each group. For the group G_1, the general diagonal term is

$$a_{1j} = [9(j-1)/(d-1)+1]^2,\quad 1\le j\le d.$$

For group G_2, it is

$$a_{2j} = [9(d-j)/(d-1)+1]^2,\quad 1\le j\le d.$$

And, for group G_3, it is

$$a_{3j} = [9[j-(d-1)/2]/(d-1)]^2,\quad 1\le j\le d.$$

The group mean vectors are equal for D4 and are for D5

$$\begin{aligned}\mu_1 &= (0,\ldots,0)\\ \mu_{2j} &= 14/\sqrt{d},\quad j=1,\ldots,d\\ \mu_{3j} &= (-1)^j\mu_{2j}\quad j=1,\ldots,d.\end{aligned}$$

Data sets D4 and D5 are related to models $[\lambda_k B_k]$ and $[\lambda_k DA_k D']$.

As Friedman, for each simulated data set, we used an additional test sample of size 100 to obtain an estimate of the compared classification rules. The experiments results are summarized in Tables 2-8. Tables 2 and 3 gives the means error rates and (into parentheses) the standard deviations of the error rates. Table 4 displays the mean values of the complexity (α) and the shrinkage (γ) parameters of RDA with their respective standard deviations into parentheses. Tables 5-8 give the frequencies of the selected model by EDRDA among the 14 models in competition for the two strategies (parsimonious and complex) and the two dimension ($d=6$ and $d=20$). The main points arising from these experiments are the following.

Table 2: Mean error rates on the test sample for RDA and EDRDA, using the parsimonious strategy (par) or the complex strategy (com) for the dimension $d = 6$.

$d = 6$	error rate(**RDA**)	error rate(**EDRDA**):par	error rate(**EDRDA**):com
D1	0.15(0.043)	0.04(0.04)	0.06(0.07)
D2	0.13(0.03)	0.18(0.09)	0.15(0.08)
D3	0.07(0.031)	0.08(0.08)	0.07(0.06)
D4	0.16(0.05)	0.13(0.07)	0.11(0.05)
D5	0.07(0.035)	0.05(0.04)	0.04(0.02)

Table 3: Mean error rates on the test sample for RDA and EDRDA, using the parsimonious strategy (par) or the complex strategy (com) for the dimension $d = 20$.

$d = 20$	error rate(**RDA**)	error rate(**EDRDA**):par	error rate(**EDRDA**):com
D1	0.10(0.05)	0.05(0.02)	0.06(0.03)
D2	0.27(0.07)	0.14(0.03)	0.18(0.03)
D3	0.14(0.04)	0.05(0.05)	0.09(0.06)
D4	0.12(0.05)	0.20(0.06)	0.25(0.04)
D5	0.06(0.04)	0.03(0.05)	0.06(0.03)

Table 4: Mean values of the complexity parameter α and of the shrinking parameter γ for RDA.

	$d = 6$	$d = 20$
D1	$\alpha = 0.098(0.092)$ $\gamma = 0.81(0.030)$	$\alpha = 0.04(0.071)$ $\gamma = 0.90(0.14)$
D2	$\alpha = 0.80(0.23)$ $\gamma = 0.02(0.06)$	$\alpha = 0.71(0.23)$ $\gamma = 0.18(0.30)$
D3	$\alpha = 0.90(0.34)$ $\gamma = 0.69(0.36)$	$\alpha = 0.75(0.35)$ $\gamma = 0.70(0.30)$
D4	$\alpha = 0.028(0.05)$ $\gamma = 0.19(0.18)$	$\alpha = 0.07(0.07)$ $\gamma = 0.41(0.18)$
D5	$\alpha = 0.08(0.11)$ $\gamma = 0.23(0.19)$	$\alpha = 0.098(0.09)$ $\gamma = 0.56(0.17)$

Table 5: Frequencies of the selected model for EDRDA using the parsimonious strategy ($d = 6$).

Model	**D1**	**D2**	**D3**	**D4**	**D5**
$[\lambda DAD^t]$	3	9	4	0	0
$[\lambda_k DAD^t]$	65	2	1	0	0
$[\lambda DA_k D^t]$	0	1	2	0	0
$[\lambda_k DA_k D^t]$	2	4	2	39	9
$[\lambda D_k AD_k^t]$	7	0	0	0	0
$[\lambda_k D_k AD_k^t]$	4	0	0	6	5
$[\lambda D_k A_k D_k^t]$	2	0	0	1	3
$[\lambda_k D_k A_k D_k^t]$	0	0	0	0	0
$[\lambda I]$	12	0	43	0	0
$[\lambda_k I]$	2	0	11	0	0
$[\lambda B]$	0	63	27	0	0
$[\lambda_k B]$	2	9	7	1	1
$[\lambda B_k]$	1	12	3	49	59
$[\lambda_k B_k]$	0	0	0	4	23

Table 6: Frequencies of the selected model for EDRDA using the complex strategy ($d = 6$).

Model	**D1**	**D2**	**D3**	**D4**	**D5**
$[\lambda DAD^t]$	4	26	16	0	0
$[\lambda_k DAD^t]$	55	3	12	0	0
$[\lambda DA_k D^t]$	0	1	1	1	0
$[\lambda_k DA_k D^t]$	9	35	13	55	38
$[\lambda D_k AD_k^t]$	14	0	0	0	0
$[\lambda_k D_k AD_k^t]$	8	0	0	7	13
$[\lambda D_k A_k D_k^t]$	7	1	0	4	19
$[\lambda_k D_k A_k D_k^t]$	0	0	1	0	16
$[\lambda I]$	0	0	3	0	0
$[\lambda_k I]$	1	0	14	0	0
$[\lambda B]$	0	10	6	0	0
$[\lambda_k B]$	2	13	23	1	0
$[\lambda B_k]$	0	11	11	30	7
$[\lambda_k B_k]$	0	0	0	2	7

Table 7: Frequencies of the selected model for EDRDA using the parsimonious strategy ($d = 20$).

Model	**D1**	**D2**	**D3**	**D4**	**D5**
$[\lambda DAD^t]$	2	6	2	0	0
$[\lambda_k DAD^t]$	4	0	2	0	0
$[\lambda DA_k D^t]$	0	0	0	6	2
$[\lambda_k DA_k D^t]$	0	0	0	0	0
$[\lambda D_k AD_k^t]$	0	0	0	0	0
$[\lambda_k D_k AD_k^t]$	0	0	0	0	0
$[\lambda D_k A_k D_k^t]$	0	0	0	0	0
$[\lambda D_k A_k D_k^t]$	0	0	0	0	0
$[\lambda I]$	44	0	32	0	0
$[\lambda_k I]$	36	0	0	0	0
$[\lambda B]$	8	74	44	0	0
$[\lambda_k B]$	6	6	12	0	2
$[\lambda B_k]$	0	8	8	94	96
$[\lambda_k B_k]$	0	6	0	0	0

Table 8: Frequencies of the selected model for EDRDA using the complex strategy ($d = 20$).

Model	**D1**	**D2**	**D3**	**D4**	**D5**
$[\lambda DAD^t]$	0	4	2	0	0
$[\lambda_k DAD^t]$	18	6	2	0	0
$[\lambda DA_k D^t]$	0	4	4	82	20
$[\lambda_k DA_k D^t]$	14	18	8	6	26
$[\lambda D_k AD_k^t]$	2	0	0	0	0
$[\lambda_k D_k AD_k^t]$	0	0	0	0	0
$[\lambda D_k A_k D_k^t]$	2	0	0	0	0
$[\lambda_k D_k A_k D_k^t]$	2	2	0	0	0
$[\lambda I]$	0	0	4	0	0
$[\lambda_k I]$	14	0	12	0	0
$[\lambda B]$	18	30	10	0	0
$[\lambda_k B]$	24	34	30	0	0
$[\lambda B_k]$	0	0	28	2	0
$[\lambda_k B_k]$	6	2	0	10	54

- In most cases EDRDA outperforms RDA (see Tables 1 and 2). The only case where RDA do significatively better is D4 with $d = 20$. In this case, where group means are equal, it seems that the shriking parameter γ of RDA plays an important role as it appears from Table 4. Indeed, RDA could outperform EDRDA when shrinking is an important factor, since EDRDA do not propose many shrinking models. But quite generally, EDRDA performs favorably as compared with RDA.

- Not surprisingly, the complex strategy (resp. the parsimonious strategy) of EDRDA tends to provide smaller error rates for $d = 6$ (resp. $d = 20$). But, for $d = 6$ the advantage of the complex strategy is not so marked, and, on the contrary, the advantage of the parsimonious strategy can be important for $d = 20$.

- From Tables 5 and 7, it appears that the parsimonious strategy of EDRDA selected reasonable models among the 14 possible models. However, and not surprisingly, this strategy has a tendency to select too simple models especially when the groups are well separated. In such cases, the criterion of selecting the model, namely the cross-validated error rate, can indicate that simpler models provide a quit performing classification rule (as for data set D5 with $d = 6$, where the model $[\lambda B]$ is preferred to the model $[\lambda_k D]$) .

- From Tables 6 and 8, it appears that the complex strategy can select reasonable models (see for instance the selected model for D5 with $d = 20$), but it can give also some disconcerting choices, as the model $[\lambda_k D A_k D']$ for the data set D2 with $d = 6$ or the models $[\lambda_k B]$ and $[\lambda B_k]$ for the data set D3 with $d = 20$.

- As a consequence, the parsimonious strategy can be preferred to the complex strategy: It gives often better error rates and moreover, it provides more realistic or reliable models in most cases.

5 Discussion

We have proposed a regularization approach, EDRDA, for Gaussian discriminant analysis based on the eigenvalue decomposition of the group variance matrices. One of the main interest of this approach is to provide a clear classification rule. The reported numerical experiments show that EDRDA can be expected to perform as least as well as RDA by producing a more user friendly classification rule. Moreover, in our opinion, the usefulness of EDRDA is not reduced to a small sample size setup and can provide quite performing classification rules where LDA and QDA give poor error rates.

We have proposed two strategies (a parsimonious one and a complex one) to solve the problem of tied models in a context of Monte Carlo numerical experiments. And, from those experiments, it appears that the parsimonious strategy can be preferred. But, we think that a better solution when models give close cross-validated error rates is to suggest to the user to choose one of the models in competition from its

own point of view: It is one of the interest of EDRDA to allow users to select a reasonable and good performing model from simple geometrical interpretations.

References

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

Celeux, G. and Govaert, G. (1994). Gaussian parsimonious clustering models. *Pattern Recognition*, to appear.

Flury, B. W. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, **79**, 892-897.

Flury B. (1988). *Common principal components and Related multivariate models.* New York: John Wiley.

Flury, B. W., Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal of Scientific Statistics Computation*, **7**, 169-184.

Flury, B. W., Schmid, M. J. and Narayanan, A. (1993) Error rates in quadratic discrimination with constraints on the covariance matrices. *Journal of Classification*, to appear.

Friedman, J. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association* **84**, 165-175.

Rayens, W. S. and Greene, T. (1991). Covariance pooling and stabilization for classification. *Computational Statistics & Data Analysis*, **11**, 17-42.

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