Shrinkage parameter for modified linear discriminant analysis

Abdallah Mkhadri

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

Rapports de Recherche

1992
25ème anniversaire

N° 1793

Programme 5
Traitement du Signal,
Automatique et Productique

SHRINKAGE PARAMETER FOR MODIFIED LINEAR DISCRIMINANT ANALYSIS

Abdallah MKHADRI

Novembre 1992

RR-1793

Shrinkage parameter for modified linear discriminant analysis

Paramètre de rétrécissement pour la discrimination linéaire régularisée

ABDALLAH MKHADRI
INRIA[1] et Université de Cadi Ayyad[2]

Abstract

Linear discriminant analysis is considered, in the small-sample, high-dimensional setting. Alternatives, shrinkage estimators, to the usual pooled sample estimate of the covariance matrix are discussed. These estimators are characterized by a shrinkage parameter γ taking its values in (0,1). First, we show that the variance of the modified linear discriminant functions is less than those of the classical linear discriminant function. Morever, we propose two alternative simple procedures, for choosing the shrinkage parameter, which are related the discrimination problem. Our procedures are based-one on the cross-validated misclassification risk and one on the cross-validated generalized discriminant function as defined in Rayens & Greene (1991). The optimal value of the shrinkage parameter is computed explicitly. The efficacy of these methods is examined through some simulation studies.

Key-words: *discriminant analysis, shrinkage estimates, misclassification risk, cross-validation.*

Résumé

Cet article traite de la discrimination linéaire pour de petits échantillons. Des estimateurs de rétrécissement de la matrice variance et de son inverse sont considérés à la place des estimateurs classiques du maximum de vraisemblance. Ces estimateurs dépendent d'un paramètre de rétrécissement qui prend ses valeurs dans l'intervalle [0,1]. Tout d'abord, on montre que la variance des fonctions discriminantes dérivées de ces estimateurs est inférieure à celle de la fonction discriminante linéaire classique. De plus, on propose deux méthodes de sélection du paramètre de rétrécissement qui sont liées au problème de discrimination. Ces méthodes sont fondées sur la minimisation, par validation croisée, du taux d'erreur ou de la distance de Mahalanobis généralisée. Les valeurs optimales sont obtenues d'une manière explicite. L'efficacité de ces méthodes est analysée sur des données simulées.

Mots-clés : *analyse discriminante, estimateurs de rétrécissement, taux d'erreur, validation croisée.*

[1]B. P. 105, 78153 Le Chesnay, France
[2]Dépt. de Maths., Fac. des Sciences Semlalia, B. P.: S 15, Marrakech, Morocco

1 Introduction

Let $\mathbf{x}$ be a column vector observation from one of the i p-variate normal population, $\pi_i \sim \mathcal{N}(\mu_i,\Sigma)$, where μ_i represents the mean vector for the ith class and Σ is the covariance matrix of the all populations, and $i = 1,..., g$. It is desired to classify $\mathbf{x}$ into one of the g populations. Equal costs of misclassification and equal prior probabilities of membership are assumed. The probability of misclassifying a particular observation $\mathbf{x}$ is minimized by assigning $\mathbf{x}$ to the class i such that

$$\delta_i f_i(\mathbf{x}) = \max_{1 \le k \le g} \delta_k f_k(\mathbf{x}) \tag{1}$$

where δ_i denotes the prior probability that $\mathbf{x}$ belongs to the ith class and f_i is the c.d.f of the ith class. Since the populations have p-variate normal distributions, the optimal rule (1) stipulates that $\mathbf{x}$ should be assigned to the class i that minimizes

$$D_i(\mathbf{x}) = (\mathbf{x} - \mu_i)^t \Sigma^{-1} (\mathbf{x} - \mu_i) + \log \left| \Sigma \right|. \tag{2}$$

The μ_i and Σ in (2) are rarely known, and usually must be estimated from a training set consisting of samples drawn from each of the populations. The so-called Linear Discriminant Fisher (LDF) rule estimates μ_i and Σ by the mean vectors $\bar{\mathbf{x}}_i$ and the pooled sample covariance matrix S. This method has been shown to frequently behave poorly in high dimensions relative to other methods (Van Ness 1980). This is because the method uses sample estimates of the means and covariance matrix which are of the poor quality in high dimensions. In the context of estimating a covariance matrix Σ, several authors (cf. Efron & Morris 1976, Haff 1979, 1980 and Dey & Srinivasan 1985, 1986 and 1991) have shown that Stein-like biased estimators which shrink the eigenvalues of the sample covariance matrix dominate the sample covariance matrix under a variety of natural loss functions. It seems reasonable that if these unbiased estimates were replaced by estimates which are more stable in high dimensions, then the resulting linear algorithm should be an improvement as demonstrated by DiPollo (1976, 1977, 1979), Campbell (1980) and Peck & Van Ness (1982). Peck & Van Ness have substituted shrinkage estimates $(1 - t(u))aS^{-1} + (bt(u)/trS)I$ for the sample estimates S^{-1}, where the function t is nondecreasing, $0 \le t(u) \le 1$, I is the identity matrix, a and b are positive constants. They found that the resulting rules perform well than LDF rule. Similarly, DiPollo (1976) and Campbell (1980) have proposed a related modification of LDF in which ridge-like estimates of the form $(S + \gamma I)$ are subsituted for S. This reduces the ratio of the largest and smallest eigenvalues of S and thus has an effect similar to shrinking the eigenvalues of S towards equality.

The function values $t(u)$, and also γ, control the choice between the extrems LDF ($t(u) = 1$ or $\gamma = 1$) and the Euclidean Distance Classifier (EDC) rules ($t(u) = 0$ or $\gamma = 0$). The sample EDC rule can outperform the sample LDF rule as shown in Raudys & Pikelis (1980) and Marco, Young & Turner (1987). Likewise, Biseay et al. (1991) have proposed a new rule based on a metric which is a linear combinaison of the Mahalanobis distance in the subspace of the first k components and the euclidean distance in its orthogonal complement. Simulation results showed the usefulness of this method.

In the two modified linear discriminant rules described above, the authors did not find how to obtain the optimal choice of the shrinkage parameter $t(u)$ or γ of the shrinkage estimators.
In this paper, we consider the intermediate method between LDF and EDC defined by shrinkage estimators of Peck & Van Ness. Here, we assume that $t(u) = \gamma \in (0,1)$. So, taking $\gamma = 1/2$ gives the estimator proposed by Efron & Morris (1976). Holding $\gamma = 1$ gives the sample LDF; while $\gamma = 0$ gives the sample EDC. Varing γ between 0 and 1 yields the rules intermediate between LDF and EDC rules.
We propose two alternative simple procedures to choose the optimal *shrinkage parameter* γ, which are related to the discrimination problem. Our procedures are based-one on the cross-validated misclassification risk and one on the cross-validated generalized discriminant function as defined in Rayens & Greene (1991). The optimal values of γ are computed explicitly without using optimization algorithms. Section 2 describes the two modified linear discriminant functions and we show that their variances are smaller than the variance of LDF function. Section 3 describes the optimal procedures of the choice of the *shrinkage parameter* γ. In Section 4, the performance of the modified linear rules are investigated through simulations studies which show the usefulness of these procedures.

2 The two-population statistical discrimination problem

For the sake of simplicity, we consider the case of two groups problem, $g = 2$. The classification rule (2) is based on the (sample) linear function $\mathcal{W}$ defined by

$$\mathcal{W}(\mathbf{x}) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t S^{-1}[\mathbf{x} - \frac{(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)}{2}] \tag{3}$$

where $nS = \sum_{j=1}^{n_1}(\mathbf{x}_j - \overline{\mathbf{x}}_1)^t(\mathbf{x}_j - \overline{\mathbf{x}}_1) + \sum_{j=1}^{n_2}(\mathbf{x}_j - \overline{\mathbf{x}}_2)^t(\mathbf{x}_j - \overline{\mathbf{x}}_2)$
and $\overline{\mathbf{x}}_i = (1/n_i)\sum_{j=1}^{n_i}\mathbf{x}_j$, with $n_i = \#\pi_i$ and $n = n_1 + n_2 - 2$, are the unbiased sample estimates. The rule (3) classifies $\mathbf{x}$ as coming from π_2 if $\mathcal{W}(\mathbf{x}) > c$ and from π_1 if $\mathcal{W}(\mathbf{x}) \leq c$, where c may be a constant, particularly 0, or a function of $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2$ and S. Hence, the optimum probability of misclassification (Popt) no longer is achieved and the probability of misclassification (PMC) depends upon the particular values of $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2$ and S obtained from the sample. It follows, then, that the probability of misclassification is not minimized.
Therefore, DiPollo (1976) and Campbell (1980) have proposed an alternative rule which incorporates the biasing feature of the Ridge Technique. This classification rule is based of the modified linear function $\overline{\mathcal{W}}$ defined by

$$\overline{\mathcal{W}}(\mathbf{x}, \gamma) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t \overline{S}^{-1}(\gamma)[\mathbf{x} - \frac{(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)}{2}] \tag{4}$$

where $\overline{S}^{-1}(\gamma) = (S + \gamma I)^{-1}$ and $\gamma \in (0,1)$.
Another alternative classification rule was proposed by Peck & Van Ness (1982) who have substituted S^{-1} in (3) by the shrinkage estimates $S^{*-1}(\gamma) = (1-\gamma)aS^{-1} +$

$(b\gamma/trS)I$, where $a = (n - p - 3)/n$, so aS^{-1} is an unbiased estimate of Σ^{-1}. Here b is a positive constant. Provided that we assume $\Sigma = \sigma^2 I$, $[b/tr(S)]I$ is a natural estimator of Σ^{-1}. We can take b equal to *(np - 2)/n* since it yielded the unbiased estimate of Σ^{-1} in this case (cf. Peck & Van Ness). But here we take it equal to p/n because it is convenable for our approximations in proposition 3. Then, the new rule is

$$\mathcal{W}^{*-1}(\mathbf{x}, \gamma) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t S^{*-1}(\gamma)[\mathbf{x} - \frac{(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)}{2}] \tag{5}$$

Here, we have considered the case where $t(u) = \gamma \in (0, 1)$ for any scalar u. These two new rules aim to reduce the variability of the two functions $\mathcal{W}^*(., \gamma)$ and $\overline{\mathcal{W}}(., \gamma)$, while introducing a little bias in the means. Implied in the concept of variance reduction is the improvement in the stability of the sample estimates. We illustrate this in the following section.

2.1 Aspect of variance reduction

For a given $\overline{\mathbf{x}}_1$, $\overline{\mathbf{x}}_2$ and S, $\mathcal{W}(\mathbf{x})$ is normally distributed with mean

$$E[\mathcal{W}(\mathbf{x})|\pi_1] = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t S^{-1}\mu_1 - \frac{1}{2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t S^{-1}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)$$

and variance

$$var[\mathcal{W}(\mathbf{x})|\pi_1] = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t S^{-1}\Sigma S^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2).$$

Suppose now that the rule (4) or (5) is used, then it can be shown (Anderson 1984) that $\tilde{\mathcal{W}}(\mathbf{x}, \gamma)$ (equal to $\mathcal{W}^*(., \gamma)$ or equal to $\overline{\mathcal{W}}(., \gamma)$) is normally distributed with mean

$$E[\tilde{\mathcal{W}}(\mathbf{x}, \gamma)|\pi_1] = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t \tilde{S}^{-1}(\gamma)\mu_1 - \frac{1}{2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t \tilde{S}^{-1}(\gamma)(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)$$

and variance

$$var[\tilde{\mathcal{W}}(\mathbf{x}, \gamma)|\pi_1] = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t \tilde{S}^{-1}(\gamma)\Sigma\tilde{S}^{-1}(\gamma)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2).$$

where $\tilde{S}^{-1}(\gamma) = S^{*-1}(\gamma)$ or $\overline{S}^{-1}(\gamma)$.
We have the following proposition, which gives the relationship between the mean and the variance of the biased rule ((4) or (5)) and the unbiased LDF rule.

Proposition 1: The relationship between the mean and the variance of $\tilde{\mathcal{W}}(\mathbf{x}, \gamma)$ can be expressed by

$$E[\tilde{\mathcal{W}}(\mathbf{x}, \gamma)|\pi_1] = E[\mathcal{W}(\mathbf{x})|\pi_1] + B$$

and

$$var[\tilde{\mathcal{W}}(\mathbf{x}, \gamma)|\pi_1] \leq var[\mathcal{W}(\mathbf{x})|\pi_1] \tag{6}$$

where B is some function of n, γ, $\overline{\mathbf{x}}_1$, $\overline{\mathbf{x}}_2$, and S.

The proof is given in appendix I.

So, when biasing is introduced, the variance reduction is also accompanied by a shift in the location of $E[\tilde{W}(\mathbf{x},\gamma)|\pi_1]$.
In the section 4, it is shown that the variance reduction overcomes the location shift consistently. The net result is an improved classification rule which decreases the PMC of the commonly used rule.

2.2 Probabilities of Misclassification

The total probability of misclassification of the sample-based LDF rule (PMC) is conditional upon $\overline{\mathbf{x}}_1$, $\overline{\mathbf{x}}_2$, and S and is given by (cf. DiPollo 1976)

$$PMC = \frac{1}{2}[1 - \Phi(z_1) - \Phi(z_2)] \tag{7}$$

where

$$z_i = \frac{(1/2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t S^{-1}(\overline{\mathbf{x}}_2 + \overline{\mathbf{x}}_1) - (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t S^{-1}\mu_i}{[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t S^{-1}\Sigma S^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_1)]^{1/2}}$$

and Φ is the cumulative function of the normal distribution. Likewise, the total probability of misclassification of the biased rule (PMC*) is

$$PMC^* = \frac{1}{2}[1 - \Phi(z_1^*) - \Phi(z_2^*)] \tag{8}$$

where

$$z_i^* = \frac{(1/2)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t \tilde{S}^{-1}(\gamma)(\overline{\mathbf{x}}_2 + \overline{\mathbf{x}}_1) - (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t \tilde{S}^{-1}(\gamma)\mu_i}{[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^t \tilde{S}^{-1}(\gamma)\Sigma\tilde{S}^{-1}(\gamma)(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)]^{1/2}}$$

with i = 1, 2. For $\gamma = 0$, PMC = PMC*. DiPollo (1976) attempted to determine the optimum choice for γ, a recurring problem with ridge technique. He found the analytical solution to this problem intractable, and he used a simulation study to choose the optimal value for γ. We can do the same experiments with the rule (5), but, these experiments will not indicate how to choose the optimal shrinkage parameter γ. Therefore, in the following sections, we propose two alternative methods for the choice of the optimal shrinkage parameter for the rule (5). These two methods are based on the minimization of loss functions which are related to the discrimination problem; morever, the optimal values for γ are determined analytically and without need of optimization algorithms.

3 Methods of the choice of the shrinkage parameter

The rule (5) depends on the shrinkage parameter γ. The simple method for choosing γ is to select a grid of values in $(0,1)$ and take, for the optimal value, the value which gives the smallest value of an estimator of the PMC (eg. Leaving-one out estimator, Lachenbruch 1975). But, this loss function is discontinous in γ and we

are not certain to reach the optimal value. Morever, while this method has the advantage of selecting the shrinkage parameter on the basis of an estimate of the actual misclassification rate, it can partially ignore information from a substantial portion of the data in selecting γ (cf. Rayens & Greene 1991). An alternative is to use a method which estimates γ automatically from the training sample. We have chosen two methods, to select automatically γ from the training set, which take into account the discrimination problem. To reduce the variability in data, we used the cross-validation method in the selection model.
The first method is based on the minimization of the cross-validated misclassification risk. It was used in the discrete case by Celeux & Mkhadri (1992). The second one, is proposed by Rayens & Greene (1991), for estimating the regularization parameters of regularized discriminant analysis, and is based on the minimization of cross-validated generalized discriminant function. They needed an optimization algorithm to compute the regularization parameters. Here, we show by approximation that the optimal shrinkage parameter γ can be computed explicitly.

3.1 The cross-validated misclassification risk

This method can be used only for the shrinkage estimator defined by equation (5). For the sake of simplicity, calculations will be detailed for two groups case ($g = 2$) and for the general case, calculations are similar to those of Celeux & Mkhadri (1992). In effect, equation (5) can be written as

$$\mathcal{W}^*(\mathbf{x}, \gamma) = (1-\gamma)a\mathcal{W}(\mathbf{x}) + \gamma(b/trS)\mathcal{E}(\mathbf{x}), \tag{9}$$

where $\mathcal{W}(.)$ is the LDF function (3), defined for group i, and $\mathcal{E}(.)$ is the EDC function (3) in which S is replaced by I.

Proposition 2: the optimal shrinkage parameter is either 0, 1 or takes

$$\frac{a\mathcal{W}^{(j)}(\mathbf{x}_j)}{a\mathcal{W}^{(j)}(\mathbf{x}_j) + (b/trS_{/j})\mathcal{E}^{(j)}(\mathbf{x}_j)} \tag{10}$$

for each $\mathbf{x}_j$, $1 \leq j \leq n$.

Proof: it is immediate to show that the cross-validated classification rule for any $\mathbf{x}_j$ ($1 \leq j \leq n$) is: $\mathbf{x}_j$ is assigned to group π_1 if and only if $\mathcal{W}^{*(j)}(\mathbf{x}_j, \gamma) \leq 0$, where

$$\mathcal{W}^{*(j)}(\mathbf{x}_j, \gamma) = (1-\gamma)a\mathcal{W}^{(j)}(\mathbf{x}_j) + \gamma(b/trS)\mathcal{E}^{(j)}(\mathbf{x}_j) \tag{11}$$

with $\mathcal{W}^{(j)}(\mathbf{x}_j)$ (resp. $\mathcal{E}^{(j)}(\mathbf{x}_j)$) denotes the estimated discriminant function of LDF (resp. EDC) when $\mathbf{x}_j$ is removing from the training sample.
Thus, different values of γ give different assignments for $\mathbf{x}_j$ if and only if there exists γ_0 in (0,1) such that $\mathcal{W}^{*(j)}(\mathbf{x}_j, \gamma_0) = 0$. It then follows that γ_0 has the form (11).
If $\mathcal{W}^{*(j)}(\mathbf{x}_j, \gamma)$ has a constant sign on $(0, 1)$, the assignment of $\mathbf{x}_j$ to one of the two groups does not depend on γ. In such case, $\mathbf{x}_j$ would be assigned following the LDF

rule ($\gamma = 0$) or following the EDC rule ($\gamma = 1$). Δ

Remark: In practical situations, the number of sample points $\mathbf{x}_j$ $(1 \leq j \leq n)$ for which the linear equation $\mathcal{W}^{*(j)}(\mathbf{x}_j, \gamma) = 0$ has a solution γ in (0,1) is very small. This number represents the number of observations for which both methods (LDF and EDC) provide different assignments.

From proposition 2, it is possible to find easily and explicitly the optimal solution for the shrinkage parameter γ. SLDF1 denotes this procedure of selection of γ defined by (11).
This procedure can be generalized in the same lines as in Celeux & Mkhadri (1992).

3.2 Cross-validation based on generalized distance

All of the discriminant rules considered in this article assign an unknown $\mathbf{x}$ to the class i which minimizes the generalized distance $\mathcal{D}_i(\mathbf{x})$ defined in (2), with Σ and μ_i estimated from the training set. The distinction between the rules lies solely in their use of different estimators of the Σ. So, as Rayens & Greene (1991) have suggested, this fundamental role of the distances suggests that misclassification risk may be reduced by estimating the Σ with a goal of minimizing the resulting generalized distance between the training set observations and the mean vectors of the classes to which they belong. It turns out that discriminant rules of this form can be devolopped by estimating the density functions within each class.
So, for choosing their complexity parameter m say, Rayens & Greene suggested to minimize the generalized distance

$$\hat{\mathcal{D}}(m) = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\hat{\mathcal{D}}_{/j}(\mathbf{x}_{ji}, m)$$

with

$$\hat{\mathcal{D}}_{/j}(\mathbf{x}_{ji}, m) = (\mathbf{x}_{ji} - \bar{\mathbf{x}}_{i/j})^t(\hat{\Sigma}_{/j}(m))^{-1}(\mathbf{x}_{ji} - \bar{\mathbf{x}}_{i/j}) + \log\left|\ \hat{\Sigma}_{/j}(m)\ \right|$$

where the notation $/j$ represents the corresponding quantity with $\mathbf{x}_{ji}$ removed. Hence, in our problem $\hat{\Sigma}_{/j}^{-1}(\gamma) = {S_{/j}}^{*-1}(\gamma)$, and we minimize

$$\hat{\mathcal{D}}^*(\gamma) = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\hat{\mathcal{D}}^*_{/j}(\mathbf{x}_{ji}, \gamma) \tag{12}$$

where $\hat{\mathcal{D}}^*_{/j}(\mathbf{x}_{ji}, \gamma)$ can be written as

$$\hat{\mathcal{D}}^*_{/j}(\mathbf{x}_{ji}, \gamma) = (1-\gamma)a\|\mathbf{x}_{ji} - \bar{\mathbf{x}}_{i/j}\|^2_{(S_{/j})^{-1}} + (b\gamma/tr(S_{/j}))\|\mathbf{x}_{ji} - \bar{\mathbf{x}}_{i/j}\|^2$$
$$- \log\left|\ (1-\gamma)aS_{/j}^{-1} + (b\gamma/tr(S_{/j}))I\ \right|$$

Note that $\widehat{\mathcal{D}}^*_{/j}(\mathbf{x}_{ji}, \gamma)$ is the (sample) generalized distance between $\mathbf{x}_{ji}$ and the ith group mean. Thus, selecting γ to minimize $\widehat{\mathcal{D}}^*(\gamma)$ amounts to minimizing a measure of the total of generalized distances between the training set observations and their respective group means. We have the following proposition which gives us the analytical optimal value of γ.

Proposition 3: The optimal value of the shrinkage parameter which minimizes $\widehat{\mathcal{D}}^*(\gamma)$ is

$$\gamma_{op} = \frac{aD_1 - bD_2 + \sum_i \sum_j \sum_{v=1}^{p} (\beta_j^v - 1)}{\sum_i \sum_j \sum_{v=1}^{p} (\beta_j^v - 1)^2} \tag{13}$$

where $D_1 = \sum_i \sum_j \|\mathbf{x}_{ji} - \overline{\mathbf{x}}_{i/j}\|^2_{(S_{/j})^{-1}}$, $D_2 = \sum_i \sum_j \frac{\|\mathbf{x}_{ji} - \overline{\mathbf{x}}_{i/j}\|^2}{trS - \mathbf{r}_j{}^t \mathbf{r}_j}$, $\beta_j^v = \frac{b(\lambda_v - \mathbf{r}_j{}^t \mathbf{r}_j)}{a(trS - \mathbf{r}_j{}^t \mathbf{r}_j)}$, $\mathbf{r}_j = \frac{(\mathbf{x}_j - \overline{\mathbf{x}})}{\sqrt{n-1}}$ and λ_v is the vth eigenvalue of the pooled covariance matrix S.

Proof: Following Friedman (1989) (pp. 13, eq. 22a), it is trivial to show that $S_{/v} = S - \mathbf{r}_v \mathbf{r}_v^t$ where $\mathbf{r}_v = [n_i/(n-1)]^{1/2}(\mathbf{x}_{vi} - \overline{\mathbf{x}}_i)$, here we assume the observation v belongs the class i. Likewise, it is easy to see that $trS_{/v} = trS - \mathbf{r}_v^t \mathbf{r}_v$, since $\mathbf{r}_v \mathbf{r}_v^t$ is rank 1 with non-zero eigenvalue $\mathbf{r}_v^t \mathbf{r}_v$. Actually, we have

$$\left| (S_{/v})^{-1}(\gamma) \right| = \left| (1-\gamma)a(S_{/v})^{-1} + [b\gamma/(trS - \mathbf{r}_v^t \mathbf{r}_v)]I \right|$$

$(S_{/v})^{-1}$ can be calculated for all n_v in groupe i by using its spectral decomposition. That is

$$(S_{/v})^{-1} = \sum_{j=1}^{p} \frac{e_j e_j{}^t}{\lambda_j - \mathbf{r}_v^t \mathbf{r}_v},$$

where $(\lambda_j)_{j=1}^p$ are eigenvalues of the pooled covariance matrix S, and $(e_j)_{j=1}^p$ are the corresponding eigenvectors. So, we have

$$\left| (S_{/v})^{-1}(\gamma) \right| = \prod_{j=1}^{p} [\frac{(1-\gamma)a}{\lambda_j - \mathbf{r}_v^t \mathbf{r}_v} + \frac{\gamma b}{trS - \mathbf{r}_v^t \mathbf{r}_v}].$$

Hence, we obtain

$$\log \left| (S_{/v})^{-1}(\gamma) \right| = \sum_j \log[a/(\lambda_j - \mathbf{r}_v^t \mathbf{r}_v)] + \sum_j \log[1 + \gamma(\beta_v^j - 1)],$$

where $\beta_v^j = \frac{b\lambda_j(1 - \mathbf{r}_v^t \mathbf{r}_v/\lambda_j)}{atrS(1 - \mathbf{r}_v^t \mathbf{r}_v/trS)}$. We have chosen a equal to $(n$ -p - $3)/n$ (we assume that n $\geq p + 4$), and we have chosed b equal to p/n. Note that b/a is less or equal to 1, and since λ_j is less than trS, we conclued that β_v^j is less or equal to 1. Hence, using Taylor series in power of γ, we have

$$\log \left| (S_{/v})^{-1}(\gamma) \right| = \sum_j \log[a/(\lambda_j - \mathbf{r}_v^t \mathbf{r}_v)] + \sum_j \gamma(\beta_v^j - 1)$$

$$+ \sum_j [\frac{\gamma^2(\beta_v^j - 1)^2}{2} + O(\gamma^3(\beta_v^j - 1)^3)]$$

Thus, approximatively, the derivative of $\hat{D}(\gamma)$ can be written as

$$\frac{\delta \hat{D}(\gamma)}{\delta \gamma} = -\sum_{i=1}^{K}\sum_{v=1}^{n}\sum_{j=1}^{p}[(\beta_v^j - 1) + \gamma(\beta_v^j - 1)^2]$$

$$-\sum_{i=1}^{K}\sum_{v=1}^{n} a\|\mathbf{x}_{vi} - \overline{\mathbf{x}}_{i/v}\|^2_{(S_{/v})^{-1}}$$

$$+\sum_{i=1}^{K}\sum_{v=1}^{n}\frac{b}{trS - \mathbf{r}^t\mathbf{r}}\|\mathbf{x}_{vi} - \overline{\mathbf{x}}_{i/v}\|^2$$

Thus, the result of equation (14) is direct from this equation which is equal to 0. Δ

Remark: the admissible solutions are $0 \leq \gamma_{op} \leq 1$, so we take $\gamma_{op} = 1$ if (13) is greater than unity and $\gamma_{op} = 0$ if (13) is less than 0. However, we have never met these cases in our applications in section 4.

Thus, the computation of the optimal value is simple whenever we have calculated the spectral decomposition of S. The advantage of this procedure over Rayens & Greene's procedure is that we did not need to use any optimization algorithm. SLDF2 denotes the procedure described in this Section and defined by (13).

4 The simulation study

In the following Section, the performance of these procedures (LDF, SLDF1 and SLDF2) is examined through simulation studies.

4.1 Sampling experiments

As in Peck & Van Ness (1982) and DiPollo (1976), we limited the data to the two underlying normal populations $\mathcal{N}(\mathbf{0}, \Sigma)$ and $\mathcal{N}(\mu, \Sigma)$, where $\mathbf{0}$ is the p-dimensional vector, μ is an arbitrary p-dimensional vector and Σ is a p-dimensional nonsingular matrix. Two different forms for the mean vector μ are selected, namely, $(m^*, 0, \ldots, 0)^t$ and $(m, m, \ldots, m)^t$ (denoted Mod1 and Mod2 respectively). The values of m^* and m were chosen so that the Mahalanobis distance, $\Delta = \mu^t\Sigma^{-1}\mu$, would be the same for either case. The mean vectors were chosen using tree values of Δ (0.5, 1.5 and 2.5). The covariance matrix Σ is chosen to have the form of an intraclass covariance. That is, Σ is of the form $\Sigma = (1-\rho)I + \rho J$, where $-1/(p-1) \leq \rho \leq 1$, I is a pxp identity matrix, and J is a pxp matrix of $1's$. Two different covariance matrices were obtained by choosing two different values of the parameter ρ, namely, 0.2 and 0.4. For each of the different combinaisons of μ and Σ, we have compared LDF rule with the two modified linear discriminant rules, SLDF1 and SLDF2, resulting from the two procedures of selection of the shrinkage parameter γ in Section 3, for one dimensions size $p = 12$. For each triple (p, μ, Σ), we have performed 100 replications of the following experiment:

Table 1 : Misclassification risk and shrinkage parameter values for population structure Mod1

		$\rho = .2$		$\rho = .4$	
		test	$\overline{\gamma}$	test	$\overline{\gamma}$
$\delta = 2.5$	LDF	.25 (.10)	.0	.35 (.08)	.0
	SLDF2	.24 (.10)	.10 (.05)	.35 (.07)	.05 (.04)
	SLDF1	.17 (.10)	.75 (.21)	.35 (.07)	.07 (.15)
$\delta = 1.5$	LDF	.42 (.08)	.0	.42 (08)	.0
	SLDF2	.40 (.07)	.05 (.00)	.41 (.08)	.06 (.05)
	SLDF1	.39 (.07)	.10 (.20)	.41 (.07)	.08 (.21)
$\delta = .5$	LDF	.44 (.06)	.0	.46 (.06)	.0
	SLDF2	.44 (.06)	.05 (.02)	.46 (.06)	.05 (.04)
	SLDF1	.44 (.06)	.11 (.20)	.44 (.06)	.08 (.20)

Table 2 : Misclassification risk and shrinkage parameter values for population structure Mod2

		$\rho = .2$		$\rho = .4$	
		test	$\overline{\gamma}$	test	$\overline{\gamma}$
$\delta = 2.5$	LDF	.23 (.08)	.0	.16 (.07)	.0
	SLDF2	.19 (.07)	.06 (.01)	.11 (.07)	.06 (.04)
	SLDF1	.18 (.09)	.17 (.30)	.11 (.08)	.14 (.23)
$\delta = 1.5$	LDF	.30 (.07)	.0	.23 (07)	.0
	SLDF2	.26 (.07)	.05 (.00)	.18 (.07)	.06 (.04)
	SLDF1	.24 (.09)	.20 (.30)	.18 (.09)	.14 (.24)
$\delta = .5$	LDF	.40 (.07)	.0	.36 (.07)	.0
	SLDF2	.38 (.07)	.05 (.00)	.32 (.07)	.06 (.03)
	SLDF1	.37 (.08)	.12 (.21)	.32 (.09)	.13 (.23)

1) For each population, we generate ten p-dimensional training data vectors and 50 p-dimensional test vectors according to the distribution being considered.
2) We use the training data to calculate the linear and the two shrinkage rules (see Section 3) and classify the test data according to each.
3) For each replication, we compute all relevant statistics.
Tables 1 and 2 summarize the results for each form of the mean vector (Mod1 and Mod2). They give, for each parameter value of ρ, the average test misclassification risk (column test) over 100 replications for each of the three classification rules: LDF, SLDF1 and SLDF2. Also shown, is the mean of the selected shrinkage parameter γ for the modified linear discriminant rules over 100 replications. The quantities in parentheses are the standard deviations of the respective quantities.

4.2 Discussion of results

From Table1, it can be seen that, for the data set Mod1, SLDF1 outperforms LDF in the case of small correlation ($\rho = .2$). This superiority decreases when the parameter of separation δ decreases. On the other hand, the performances of SLDF1 and SLDF2 are similar, with a slight advantage to SLDF1 when δ is large or moderate ($\delta = 2.5$ or 1.5). While for the large correlation parameter value ($\rho = .4$), the performances of the three procedures are more comparable with a slight advantage of SLDF1 when δ increases. However, note that the shrinkage parameter values of SLDF1 are always greater than those of SLDF2 which are close to 0. Thus, SLDF2 has the characteristic to choose small values for the shrinkage parameter, while SLDF1 overcomed this limitation and selected the shrinkage parameters which are close to optimal values.
For the data set Mod2, Table2 shows that SLDF1 and SLDF2 outperform well LDF in all situations. Here, SLDF1 performed slightly better (1-2 percent) than SLDF2 for the small correlation case. While, for the large correlation case, both methods give the same results on the test sample. On the other hand, the characteristic result of the shrinkage parameter values is similar to those of the data set Mod1.
The conclusion of these experiments is that the modified linear rules perfomed better than LDF rule. SLDF1 and SLDF2 differed only in the data set Mod1 for the small values of correlation and separation parameters.

5 Conclusion

We have presented the shrinkage linear discriminant rules in the Gaussian framework. The numerical experiments showed that good performances can be expected from these modified linear rules. The performance of SLDF1 is slightly better than the one of SLDF2. Morever, SLDF1 has the property of selecting the shrinkage parameters which are close to optimal values. On the other hand, SLDF2 has the characteristic to choose small values for the shrinkage parameter. Despite these restrictions, we think that SLDF rules shoud be quite beneficial for Gaussian discriminant analysis in setting for which sample sizes are small. In the case of g populations, $g \geq 3$, SLDF1 will be time comsuming. Because for each $\mathbf{x}_i$ ($1 \leq i \leq n$), there are, generally, at most two possible optimal γ values (cf. Celeux & Mkhadri 1992). In the later case, we recommande to use SLDF2 which is less time consuming than SLDF1.
Morever, the proposed procedures are related the discrimination problem and compute explicitly the shrinkage parameter without using any optimisation algorithm.

APPENDIX I

The proof of Proposition 1: The distribution of $\mathcal{W}$ is invariant with respect to the transformation $\mathbf{x}^* = A\mathbf{x} + d$, $\mathbf{x}_j^{*(1)} = A\mathbf{x}_j^{(1)} + d$, and $\mathbf{x}_j^{*(2)} = A\mathbf{x}_j^{(2)} + d$, where A is nonsingular, $j = 1,..., n$. We choose A and d to transform Σ to I, $\mu_1 - \mu_2$ to

$\delta = (\Delta, 0, ..., 0)$, where $\Delta = [(\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)]^{1/2}$, and μ_1 to $\mathbf{0}$ (see Anderson 1973). Let $\mathbf{Y}$, $\mathbf{Z}$ and V be defined by

$\mathbf{d} = \overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(2)} = \delta + \frac{1}{n^{1/2}}\mathbf{Y}$, $\overline{\mathbf{x}}^{(1)} = \frac{1}{n^{1/2}}\mathbf{Z}$ and $S = I + \frac{1}{n^{1/2}}V$.

Then, the joint distribution of $(\mathbf{Y}, \mathbf{Z})^t$ is normal with vector mean $(0, 0)^t$ and variance

$$\begin{pmatrix} (n/n_1 + n/n_2)I & (n/n_1)I \\ (n/n_1)I & (n/n_1)I \end{pmatrix}.$$

We have

$$E[W(\mathbf{x})|\pi_1] = -\mathbf{d}^t S^{-1} \mathbf{d}^*$$

where $\mathbf{d}^* = (\overline{\mathbf{x}}^{(1)} + \overline{\mathbf{x}}^{(2)})/2$. The mean becomes, after some algebraic manipulations, by using Taylor series in power of V (cf. Anderson 1973)

$$E[W(\mathbf{x})|\pi_1] = -d^t d^* + \frac{1}{n^{1/2}}\mathbf{d}^t V \mathbf{d}^* - \frac{1}{n}\mathbf{d}^t V^2 \mathbf{d}^* + O(\frac{1}{n^{3/2}})$$

Again, by using Taylor series in power of V we have

$$E[\overline{W}(\mathbf{x}, \gamma)|\pi_1] = -\frac{1}{1+\gamma} d^t d^* + \frac{1}{n^{1/2}(1+\gamma)^2}\mathbf{d}^t V \mathbf{d}^* - \frac{1}{n(1+\gamma)^3}\mathbf{d}^t V^2 \mathbf{d}^* + O(n^{-3/2}(1+\gamma)^{-1})$$

Now, using Taylor series in power of γ, we get

$$E[\overline{W}(\mathbf{x}, \gamma)|\pi_1] = E[W(\mathbf{x})|\pi_1] + \gamma \mathbf{d}^t \mathbf{d}^* + \frac{2\gamma}{n^{1/2}}\mathbf{d}^t V \mathbf{d}^* - \frac{3\gamma}{n}\mathbf{d}^t V^2 \mathbf{d}^* + O(\gamma^2 n^{-3/2})$$

Hence, we obtain

$$E[\overline{W}(\mathbf{x}, \gamma)|\pi_1] = E[W(\mathbf{x})|\pi_1] - \gamma\Delta^2 + O(\gamma n^{-1/2})$$

Similarly, the variance of $W(., \gamma)$ becomes, after some algebraic calculations,

$$var[W(\mathbf{x})|\pi_1] = \mathbf{d}^t \mathbf{d} - \frac{1}{n^{1/2}}\mathbf{d}^t V \mathbf{d} + \frac{2}{n}\mathbf{d}^t V^2 \mathbf{d} + O(n^{-3/2})$$

It follows that,

$$var[\overline{W}(\mathbf{x}, \gamma)|\pi_1] = \frac{1}{(1+\gamma)^2}[\mathbf{d}^t \mathbf{d} - \frac{1}{n^{1/2}(1+\gamma)}\mathbf{d}^t V \mathbf{d} + \frac{2}{(1+\gamma)^2 n}\mathbf{d}^t V^2 \mathbf{d} + O(n^{-3/2})]$$

Using Taylor series in power of γ, it can be written by

$$var[\overline{W}(\mathbf{x}, \gamma)|\pi_1] = var[W(\mathbf{x})|\pi_1] - 2\gamma\Delta^2 + O(\gamma n^{-1/2})$$

In the same way, it is easy to show that

$$E[W^*(\mathbf{x},\gamma)|\pi_1] = aE[W(\mathbf{x})|\pi_1] + \gamma[a - b/(p + trV)]\Delta^2 + O(\gamma n^{-1/2})$$

Since a is approximatly equal to 1, then

$$E[W^*(\mathbf{x})|\pi_1] = E[W(\mathbf{x})|\pi_1] + \gamma[1 - b/(p + trV)]\Delta^2 + O(\gamma n^{-1/2})$$

The variance of W^* can be computed in the same way as

$$var[W^*(\mathbf{x},\gamma)|\pi_1] = a^2 var[W(\mathbf{x})|\pi_1] + [-2a^2 + \gamma^2 a^2]\mathbf{d}^t S^{-2}\mathbf{d} + \frac{2ab\gamma(1-\gamma)}{p + trV}\mathbf{d}^t S^{-1}\mathbf{d} + \frac{b^2\gamma^2}{p + trV}\mathbf{d}^t\mathbf{d}$$

It can be written by (since a $\approx$ 1)

$$var[W^*(\mathbf{x},\gamma)|\pi_1] = var[W(\mathbf{x})|\pi_1] + \psi(\gamma)\Delta^2 + O(\gamma^2 n^{-1/2})$$

where

$$\psi(\gamma) = -2\gamma(1-\gamma) + \frac{2b\gamma(1-\gamma)}{p + trV} + \frac{b^2\gamma^2}{p + trV}\mathbf{d}^t\mathbf{d}.$$

Note that $\psi(0) = 0$ and $\psi(1) = -1 + [b/(p + trV)]^2$ and is negative when $b \leq p$, which is true. Morever we can show easily that the function ψ is decreasing in (0,1), which completes the proof of proposition 1. Δ

BIBLIOGRAPHY

Anderson T. W. (1984). *An introduction to multivariate analysis.* New York: John Wiley.

Biscay R., Valdes P. & Pascual R. (1991). Modithied Fisher's linear discriminant function with reduction of dimensionality. *J. Statist. Comput. Simul.*, **vol. 36**, p. 1-8.

Campbell N. A. (1980). Shrunken estimator in discriminant and canonical variate analysis. *Appl. Stat.*, **29**, p. 5-14.

Celeux G. & Mkhadri A. (1992). Regularized discrete discriminant analysis. *Statis. & Computing*, vol. **2**, p. 143-151.

Dey D. K. & Srinivasan C. (1985). Estimation of a coavriance matrix under Stein's loss. *Ann. Stat.*, **13**, 1581-1591.

Dey D. K. & Srinivasan C. (1986). Trimmed minimax estimator of a covariance matrix. *Ann. Inst. Stat. Math.*, **39**, 101-108.

Dey D. K. & Srinivasan C. (1991). On estimation of discriminant coefficients. *Stat. & Prob. Letters*, **11**, 189-193.

DiPollo P. J. (1976). The application of bias to discriminant analysis. *Comm. Stat.-Theor. Meth.*, **A5**, 843-854 .

DiPollo P. J. (1977). Further applications of bias discriminant analysis. *Comm. Stat.-Theor. Meth.*, **A6**, 933-943.

DiPollo P. J. (1979). Biased discriminant analysis: Evaluation of the optimum probability of misclassification. *Comm. stat.-Theor. Meth.*, **A8**, 1447-57.

Efron B. & Morris C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Stat.*, **4**, 22-32.

Friedman J. (1989). Regularized discriminant analysis. *JASA* **84**, $n^0405, 165 - 175$.

Haff L. R. (1979). Estimation of the inverse of covariance matrix: random mixtures of the inverse Wishart matrix and the identity. *Ann. Stat*, **7**, 1264.

Haff L. R. (1980). Empirical Bayes estimation of the multivarite normal covariance matrix. *Ann. Stat.*, **8**, 586-597.

Marco V. R., Young D. M. & Turner D. N. (1987). The euclidean distance classifier : An alternative to the linear discriminant function. *Comm. Stat.-Simul.*, **16**(2), 485-505.

Peck R. & Van Ness J. (1982). The use of shrinkage estimators in linear discriminant analysis. *IEEE Trans. Patt. Anal. Mach. Intell.*, **PAMI-4**, No 5, 530-37.

Rands S. & Pikelis V. (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in Pattern Rocognition. *IEEE Trans. Patt. Anal. Mach. Intell.*, **PAMI-2**, $n^0 3, 242 - 252$.

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

Van Ness J. (1980). On the effects of dimension in discriminant analysis for unequal covariance populations. *Technometrics*, **21**, 119-127.

ISSN 0249-6399