UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 02-35
Statistics and Econometrics Series 06
August 2002

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

ON THE CONSISTENCY AND ROBUSTNESS PROPERTIES OF LINEAR DISCRIMINANT ANALYSIS

Santiago Velilla and Adolfo Hernández*

Abstract

Strong consistency of linear discriminant analysis is established under wide assumptions on the class conditional densities. Robustness to the presence of a mild degree of class dispersion heterogeneity is also analyzed. Results obtained may help to explain analytically the frequent good behavior in applications of linear discrimination techniques.

Keywords: Bayes error; consistent sample discriminant rule; inverse location regression models; plug-in discriminant rules.

* Velilla, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, 28903-Getafe, Madrid, Spain, e-mail: savece@est-econ.uc3m.es; Hernández, Departamento de Análisis Económico, Universidad Autónoma de Madrid, 28049-Cantoblanco, Madrid, Spain, e-mail: adolfo.hernandez@uam.es. Research partially supported by CICYT Grant BEC 2000-0167 (Spain).

On the Consistency and Robustness Properties of Linear Discriminant Analysis

Santiago Velilla and Adolfo Hernández*

Abstract

Strong consistency of linear discriminant analysis is established under wide assumptions on the class conditional densities. Robustness to the presence of a mild degree of class dispersion heterogeneity is also analyzed. Results obtained may help to explain analytically the frequent good behavior in applications of linear discrimination techniques.

AMS 2000 subject classification: 62H30, 62H99.

Key words and phrases: Bayes error, consistent sample discriminant rule, inverse location regression models, plug-in discriminant rules.

1. INTRODUCTION

Consider a discriminant problem where the goal is to assign an individual to one of a finite number of classes or groups $g_1, ..., g_k$ on the basis of p observed features $\mathbf{x} = (x_1, ..., x_p)'$. To do this, the space $\mathbb{R}^p$ is partitioned into subsets R_1, ..., R_k such

*Velilla, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, 28903-Getafe, Madrid, Spain. Hernández, Departamento de Análisis Económico, Universidad Autónoma de Madrid, 28049-Cantoblanco, Madrid, Spain. Research partially supported by CICYT Grant BEC 2000-0167 (Spain).

that, for $i = 1, ..., k$, the individual is classified in group g_i when $\mathbf{x}$ belongs to R_i. This procedure generates a discriminant rule as a mapping $r : \mathbb{R}^p \to \{1, ..., k\}$ that takes the value $r(\mathbf{x}) = i$ whenever the individual is assigned to the ith group, and that can be therefore written as $r(\mathbf{x}) = \sum_{i=1}^{k} iI_{R_i}(\mathbf{x})$, where $I_{R_i}(x)$ is the indicator function of the subset R_i. Let $\mathbf{g}$ be the discrete random variable, often called class index or group label, that represents the true membership of the individual under study. In agreement with the previous notation, the group label takes values $\mathbf{g} = i$ with class prior probabilities $\pi_i = P[\mathbf{g} = i] > 0$, $i = 1, ..., k$. Throughout this paper it is assumed that the class conditional distributions $\mathbf{x} \mid \mathbf{g} = i$ are absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^p$, that is, there exist density functions $f_i(\mathbf{x})$ such that $P[\mathbf{x} \in A| \ \mathbf{g} = i] = \int_A f_i(\mathbf{x})d\mathbf{x}$, $i = 1, ..., k$. Given $(\mathbf{x},\mathbf{g})$, rule $r(\mathbf{x}) = \sum_{i=1}^{k} iI_{R_i}(\mathbf{x})$ is in error when $r(\mathbf{x}) \neq \mathbf{g}$ and its probability of misclassification $L[r(\mathbf{x})] = P[r(\mathbf{x}) \neq \mathbf{g}] = 1 - P[r(\mathbf{x}) = \mathbf{g}] = 1 - \sum_{i=1}^{k} P[\mathbf{x} \in R_i \,; \mathbf{g} = i]$ is

$$L[r(\mathbf{x})] = 1 - \sum_{i=1}^{k} P[\mathbf{g} = i]P[\mathbf{x} \in R_i \mid \mathbf{g} = i] = 1 - \sum_{i=1}^{k} \pi_i \int_{R_i} f_i(\mathbf{x})d\mathbf{x} \ . \qquad (1)$$

The rule $r^*(\mathbf{x}) = \sum_{i=1}^{k} iI_{R_i^*}(\mathbf{x})$ that minimizes the functional $L[r(\mathbf{x})]$, or Bayes rule, is given by the partition $R_i^* = \{\mathbf{x} : \pi_i f_i(\mathbf{x}) = \max_{1 \leq j \leq k} \pi_j f_j(\mathbf{x})\}$, $i = 1, ..., k$ (see e.g. Seber 1984, chap. 6) and, according to (1), its probability of misclassification is the corresponding optimal or Bayes error

$$L^* = L[r^*(\mathbf{x})] = 1 - \sum_{i=1}^{k} \pi_i \int_{R_i^*} f_i(\mathbf{x})d\mathbf{x} \ . \qquad (2)$$

In general both π_i and $f_i(\mathbf{x})$ are unknown, so rules used in practice are sample based rules of the form $\widehat{r}_n(\mathbf{x}) = \sum_{i=1}^{k} iI_{\widehat{R}_{i,n}}(\mathbf{x})$, where the subsets $\widehat{R}_{i,n}$ depend on a data set $\mathbf{D}_n = \{(\mathbf{x}_j,\mathbf{g}_j) : j = 1, ..., n\}$ formed by i.i.d. observations from the pair $(\mathbf{x},\mathbf{g})$, obtained sampling from individuals previously classified. The appropriate measure of error of a sample rule $\widehat{r}_n(\mathbf{x})$ is its conditional probability of misclassification $L_n = P[\widehat{r}_n(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$. If the pair $(\mathbf{x}, \mathbf{g})$ is assumed to be independent of the data in

$\mathbf{D}_n$, using (1)

$$L_n = 1 - \sum_{i=1}^{k} \pi_i \int_{\widehat{R}_{i,n}} f_i(\mathbf{x}) d\mathbf{x} \tag{3}$$

is a random variable that satisfies $0 \leq L^* \leq L_n \leq 1$. Following Devroye, Györfi and Lugosi (1996, chap. 6), the sequence of rules $\{\widehat{r}_n(\mathbf{x})\}$ is weakly or strongly consistent when, as n goes to infinity, L_n converges in probability or almost everywhere (*a.e.*) to the optimum L^*.

A very common technique for constructing sample rules is the so called *linear discriminant analysis* (*LDA*) as described for example in chapter 4 of the recent book by Hastie, Tibshirani and Friedman (2001). The aim of this paper is to explore some of the asymptotic properties of the conditional probability of misclassification of *LDA*. Results obtained may help to explain the frequent correct behavior of *LDA* in applications, either with real or simulated data. Section 2 establishes notation and presents some of the issues involved in *LDA* classification procedures. Sections 3 and 4 give results on strong consistency and section 5 studies robustness to heteroscedasticity. Section 6 gives some final comments and section 7 collects proofs of some auxiliary results.

2. BACKGROUND AND MOTIVATION

Write the given database in the form $\mathbf{D}_n = \{\mathbf{x}_{ij} : i = 1, ..., k \ , \ j = 1, ..., n_i\}$, where n_i is the number of observations in class g_i. Compute the class centroids $\overline{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$, $i = 1, ..., k$, and obtain the overall sample mean vector $\overline{\mathbf{x}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \mathbf{x}_{ij}/n = \sum_{i=1}^{k}(n_i/n)\overline{\mathbf{x}}_i$ as a weighted average of the $\overline{\mathbf{x}}_i$. Given a feature vector $\mathbf{x} = (x_1, ..., x_p)'$, define its standardized version as

$$\mathbf{y} = \widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}) \ , \tag{4}$$

where

$$\widehat{\mathbf{\Sigma}}_p = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)' \ , \tag{5}$$

is a *pooled* estimator of the assumed common dispersion matrix in each group. Notice that the standardized data $\mathbf{y}_{ij} = \widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x}_{ij} - \overline{\mathbf{x}})$, $i = 1, ..., k$, $j = 1, ..., n_i$, have class centroids $\overline{\mathbf{y}}_i = \widehat{\mathbf{\Sigma}}_p^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})$, $i = 1, ..., k$, overall sample mean $\overline{\mathbf{y}} = \mathbf{0}$, and pooled dispersion estimator $\widehat{\mathbf{\Sigma}}_{p,\mathbf{y}} = \mathbf{I}_p$. *LDA* assigns $\mathbf{x} = (x_1, ..., x_p)'$ to g_i when

$$(\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i) = \min_{1 \le j \le k} (\mathbf{x} - \overline{\mathbf{x}}_j)'\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_j) = \min_{1 \le j \le k} \|\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_j)\|^2$$
$$= \min_{1 \le j \le k} \|\mathbf{y} - \overline{\mathbf{y}}_j\|^2 = \|\mathbf{y} - \overline{\mathbf{y}}_i\|^2 \ , \tag{6}$$

where $\|.\|$ is the usual euclidean norm. In the first line of (6), the feature vector is assigned to the class whose centroid is closest in the sense of the Mahalanobis distance generated by the matrix of (5). In the second line, the metric is the euclidean distance between the standardized feature vector of (4) and the corresponding standardized class centroids $\overline{\mathbf{y}}_i$. As a result of appendix 7.1, if all the class conditional distributions $\mathbf{x} \mid \mathbf{g} = i$ are absolutely continuous, the matrix $\widehat{\mathbf{\Sigma}}_p$ is positive definite (p.d.) with probability one for all $n \ge p + k$, so, for practical purposes, both its inverse $\widehat{\mathbf{\Sigma}}_p^{-1}$ and the square root $\widehat{\mathbf{\Sigma}}_p^{-1/2}$ considered above are well defined. Criterion (6) does not depend on the quadratic terms $\mathbf{x}'\widehat{\mathbf{\Sigma}}_p^{-1}\mathbf{x}$ or $\mathbf{y}'\mathbf{y}$ and produces then, either in the $\mathbf{x}$ or $\mathbf{y}$ spaces, linear boundaries of separation between classes.

On the other hand, suppose that after projecting onto a direction $\mathbf{a} \in \mathbb{R}^p$, $\|\mathbf{a}\| = 1$, *separation* between the projected standardized class centroids $\mathbf{a}'\overline{\mathbf{y}}_i = \mathbf{a}'\widehat{\mathbf{\Sigma}}_p^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})$, $i = 1, ..., k$, is calibrated by the weighted sum of squares

$$\sum_{i=1}^{k} \frac{n_i}{n}(\mathbf{a}'\overline{\mathbf{y}}_i)^2 = \mathbf{a}'(\sum_{i=1}^{k} \frac{n_i}{n}\overline{\mathbf{y}}_i\overline{\mathbf{y}}_i')\mathbf{a} = \mathbf{a}'\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2}\mathbf{a} \ , \tag{7}$$

where

$$\widehat{\mathbf{B}} = \sum_{i=1}^{k} \frac{n_i}{n}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})' \ , \tag{8}$$

is the $p \times p$ sample *between groups dispersion matrix*. As seen in appendix 7.1, if the class conditional distributions $\mathbf{x} \mid \mathbf{g} = i$ are absolutely continuous $r(\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2}) = r(\widehat{\mathbf{B}}) = q = \min(k-1,p)$, so the spectral representation of the matrix of the quadratic form in (7) is

$$\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2} = \widehat{\mathbf{C}}\widehat{\mathbf{D}}\widehat{\mathbf{C}}' \ , \tag{9}$$

where $\widehat{\mathbf{C}} = (\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2, ..., \widehat{\boldsymbol{\gamma}}_p)$ is a $p \times p$ orthogonal matrix of eigenvectors and $\widehat{\mathbf{D}} = diag(\widehat{\lambda}_1, ..., \widehat{\lambda}_p)$ is a $p \times p$ matrix of nonnegative eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq ... \geq \widehat{\lambda}_q > 0 = \widehat{\lambda}_{q+1} = ... = \widehat{\lambda}_p$. The eigenvectors $\widehat{\boldsymbol{\gamma}}_j$ can be obtained sequentially as orthogonal directions that, as measured by criterion (7), maximize separation between projected standardized centroids. The eigenvalue $\widehat{\lambda}_j = \widehat{\boldsymbol{\gamma}}_j' \widehat{\boldsymbol{\Sigma}}_p^{-1/2} \widehat{\mathbf{B}} \widehat{\boldsymbol{\Sigma}}_p^{-1/2} \widehat{\boldsymbol{\gamma}}_j$ is the strength of separation obtained in the jth direction. Notice that only q directions are needed for reaching the total separation index $\sum_{j=1}^{q} \widehat{\lambda}_j$. Put $\widehat{\mathbf{g}}_j = \boldsymbol{\Sigma}_p^{-1/2} \widehat{\boldsymbol{\gamma}}_j$, $j = 1, ..., q$. If $\widehat{\mathbf{W}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)'/n = (n-k)\widehat{\boldsymbol{\Sigma}}_p/n$ is the $p \times p$ sample *within groups dispersion matrix*, the pairs $(n\widehat{\lambda}_j/(n-k), \widehat{\mathbf{g}}_j)$ are the eigenvalues and eigenvectors of $\widehat{\mathbf{W}}^{-1}\widehat{\mathbf{B}}$, where the eigenvectors are normalized by conditions $\widehat{\mathbf{g}}_j' \widehat{\boldsymbol{\Sigma}}_p \widehat{\mathbf{g}}_k = \delta_{jk} = 1$ for $j = k$ and $\delta_{jk} = 0$ for $j \neq k$. The $\widehat{\mathbf{g}}_j$, usually known as *discriminant directions*, can be also obtained as solutions of the Fisher-Rao discriminant criterion

$$\max_{\mathbf{g} \in \mathbb{R}^p} \frac{\mathbf{g}' \widehat{\mathbf{B}} \mathbf{g}}{\mathbf{g}' \widehat{\mathbf{W}} \mathbf{g}} , \tag{10}$$

and, therefore, maximize the ratio of the between to the within variability. In particular, the first discriminant direction $\widehat{\mathbf{g}}_1$ generates the so called Fisher's *linear discriminant function* (*LDF*) $\widehat{\mathbf{g}}_1'(\mathbf{x} - \overline{\mathbf{x}})$.

Select now an integer $1 \leq r \leq q = \min(k-1, p)$, and partition the matrix $\widehat{\mathbf{C}}$ of (9) in the form $\widehat{\mathbf{C}} = (\widehat{\mathbf{C}}_1(r) \mid \widehat{\mathbf{C}}_2(r))$, where $\widehat{\mathbf{C}}_1(r) = (\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2, ..., \widehat{\boldsymbol{\gamma}}_r)$ is of $p \times r$ and $\widehat{\mathbf{C}}_2(r) = (\widehat{\boldsymbol{\gamma}}_{r+1}, ..., \widehat{\boldsymbol{\gamma}}_p)$ of $p \times (p-r)$. Since $\widehat{\mathbf{C}}$ is orthogonal, the distances considered in (6) can be decomposed additively in the form

$$\begin{aligned} (\mathbf{x} - \overline{\mathbf{x}}_i)' \widehat{\boldsymbol{\Sigma}}_p^{-1} (\mathbf{x} - \overline{\mathbf{x}}_i) &= \|\widehat{\boldsymbol{\Sigma}}_p^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_i)\|^2 = \|\mathbf{y} - \overline{\mathbf{y}}_i\|^2 = \|\widehat{\mathbf{C}}'(\mathbf{y} - \overline{\mathbf{y}}_i)\|^2 \\ &= \|\widehat{\mathbf{C}}_1'(r)(\mathbf{y} - \overline{\mathbf{y}}_i)\|^2 + \|\widehat{\mathbf{C}}_2'(r)(\mathbf{y} - \overline{\mathbf{y}}_i)\|^2 . \end{aligned} \tag{11}$$

Generalizing (7), separation of standardized centroids after projecting onto the subspace generated by the columns of $\widehat{\mathbf{C}}_1(r)$ can be quantified by the weighted sum

$$\sum_{i=1}^{k} \frac{n_i}{n} \|\widehat{\mathbf{C}}_1'(r) \overline{\mathbf{y}}_i\|^2 = \sum_{i=1}^{k} \frac{n_i}{n} tr[\widehat{\mathbf{C}}_1'(r) \mathbf{y}_i \overline{\mathbf{y}}_i' \widehat{\mathbf{C}}_1(r)]$$

$$= tr[\widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{C}}_1(r)] = \sum_{j=1}^{r}\widehat{\lambda}_j \ . \tag{12}$$

The sum in (12) is an aggregate additive measure of the degree of separation obtained after projecting onto each one of the directions in $\widehat{\mathbf{C}}_1(r)$. Similarly, separation after projecting onto the column space of $\widehat{\mathbf{C}}_2(r)$ can be measured by the number $tr[\widehat{\mathbf{C}}_2'(r)\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{C}}_2(r)] = \sum_{j=r+1}^{q}\widehat{\lambda}_j$. Let $\widehat{p}_j = \widehat{\lambda}_j/\sum_{j=1}^{q}\widehat{\lambda}_j$ be the relative proportion of separation provided by direction $\widehat{\boldsymbol{\gamma}}_j$, $j = 1, ..., q$. When the cumulative relative proportion $\widehat{q}_r = \sum_{j=1}^{r}\widehat{p}_j = \sum_{j=1}^{r}\widehat{\lambda}_j/\sum_{j=1}^{q}\widehat{\lambda}_j$ is "close" to one, the second summand in (11) could be ignored for classification purposes. This leads to a *reduced rank linear discriminant analysis* ($RLDA$) criterion that assigns $\mathbf{x}$ to g_i when

$$\|\widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}_i)\|^2 = \|\widehat{\mathbf{C}}_1'(r)(\mathbf{y}-\overline{\mathbf{y}}_i)\|^2$$

$$= \min_{1\le j\le k}\|\widehat{\mathbf{C}}_1'(r)(\mathbf{y}-\overline{\mathbf{y}}_j)\|^2 = \min_{1\le j\le k}\|\widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}_j)\|^2 \ . \tag{13}$$

Criterion above can be expressed in terms of the *canonical* or *discriminant coordinates* $\widehat{\mathbf{y}}_r = \widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}) = \widehat{\mathbf{C}}_1'(r)\mathbf{y}$, that allow writing (13) as $\|\widehat{\mathbf{y}}_r - \widehat{\mathbf{m}}_i\|^2 = \min_{1\le j\le k}\|\widehat{\mathbf{y}}_r - \widehat{\mathbf{m}}_i\|^2$, where $\widehat{\mathbf{m}}_i = \widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}) = \widehat{\mathbf{C}}_1'(r)\overline{\mathbf{y}}_i$ are the canonical coordinates of centroid $\overline{\mathbf{x}}_i$, $i = 1, ..., k$.

LDA and $RLDA$ were developed by Fisher (1936) and Rao (1948) under no particular assumption for the class conditional densities $f_i(\mathbf{x})$, $i = 1, ..., k$. The goal was to construct a classification procedure after a search for the subspace spanned by the directions that, as measured by a criterion of the form (10), maximize separation between class centroids. A traditional justification for LDA is that (6) is a sample plug-in version of the optimal procedure obtained when the class prior probabilities are identical and the class conditional densities are multivariate normal with the same dispersion matrix. Notice that if $\pi_i = 1/k$ and $f_i(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$ for $i = 1, ..., k$, where the $\boldsymbol{\mu}_i$ are $p \times 1$ vectors and $\mathbf{\Sigma}$ is a $p \times p$ p.d. matrix, the subset R_i^* of the associated Bayes rule $r^*(\mathbf{x}) = \sum_{i=1}^{k} i I_{R_i^*}(\mathbf{x})$ is formed by all points $\mathbf{x} \in \mathbb{R}^p$ such that

$$(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) = \min_{1\le j\le k}(\mathbf{x}-\boldsymbol{\mu}_j)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) \ . \tag{14}$$

Criterion (6) is obtained after replacing in (14) $\mathbf{\Sigma}$ and $\boldsymbol{\mu}_i$ by, respectively, $\widehat{\mathbf{\Sigma}}_p$ and $\overline{\mathbf{x}}_i$. However, and as remarked recently by Hastie et al. (2001, sec. 4.3), it is well-known that LDA is fairly robust against deviations from the standard gaussian assumptions and, as indicated by the estimated behavior of its conditional probability of error, performs well in a diverse set of classification tasks even as compared with more sophisticated procedures. This is well illustrated, for example, by Michie, Spiegelhalter and Taylor (1994) in the STATLOG project. McLachlan (1992, sec. 5.6.1) reports conclusions from simulation studies. Broadly speaking, for sample sizes n large enough rule (6) seems to work well when the class conditional densities $f_i(\mathbf{x})$ are *symmetric* but not necessarily gaussian. LDA tolerates also some mild degree of class dispersion heterogeneity. On the other hand, as suggested by Johnson and Wichern (1998, p. 697), not much is known about the behavior of $RLDA$ in practice. According to Hastie, Tibshirani and Buja (1994), when $q = \min(k - 1,p)$ is relatively large as compared to p and for some $r \ll q = \min(k - 1,p)$ the cumulative relative proportion of separation among centroids $\widehat{q}_r = \sum_{j=1}^{r} \widehat{\lambda}_j / \sum_{j=1}^{q} \widehat{\lambda}_j$ is close to one, $RLDA$ eliminates *spurious* directions with no relevant information for separation-classification purposes and can be then preferable to LDA. Nevertheless, and following Flury (1997, sec. 7.3), the choice in rule (13) of the number of canonical coordinates r to be used in practice remains as a relatively undetermined question. As seen next, describing the asymptotic behavior of the conditional probability of misclassification of both LDA and $RLDA$ can provide some analytical answers for the issues presented in this paragraph.

3. STRONG CONSISTENCY

Suppose that, for $i = 1, ..., k$, the ith class conditional distribution can be represented as

$$\mathbf{x} \mid \mathbf{g} = i \stackrel{D}{=} \boldsymbol{\mu}_i + \mathbf{\Sigma}^{1/2}\mathbf{u} \ , \tag{15}$$

where $\boldsymbol{\mu}_i$ is a constant $p \times 1$ vector, $\boldsymbol{\Sigma}^{1/2}$ is the square root of a $p \times p$ p.d. matrix $\boldsymbol{\Sigma}$, and $\mathbf{u}$ is a $p \times 1$ random vector independent of the class index $\mathbf{g}$. According to Cook and Yin (2001, p. 158), when (15) holds the feature vector satisfies an *inverse location regression* model. In what follows, it is assumed that $\mathbf{u}$ has an spherical density $g(\mathbf{u}'\mathbf{u})$, where $g(.)$ is a function from [0,∞) to [0,∞). Under this assumption, the ith class conditional distribution $\mathbf{x} \mid \mathbf{g} = i$ has an elliptically symmetric density

$$f_i(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} g[(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)] \ . \tag{16}$$

If also $g(.)$ is such that $\int_0^{+\infty} t^{p/2} g(t)dt < +\infty$, then $E(\mathbf{x} \mid \mathbf{g} = i) = \boldsymbol{\mu}_i$ and $Var(\mathbf{x} \mid \mathbf{g} = i) = a\boldsymbol{\Sigma}$, where $a > 0$ is a positive constant independent of the specific value $\mathbf{g} = i$ (Muirhead, 1982 p. 34). Therefore, the marginal mean vector and dispersion matrix of the feature vector $\mathbf{x}$ are $\boldsymbol{\mu} = E(\mathbf{x}) = \sum_{i=1}^k \pi_i E(\mathbf{x} \mid \mathbf{g} = i) = \sum_{i=1}^k \pi_i \boldsymbol{\mu}_i$ and

$$\boldsymbol{\Gamma} = Var(\mathbf{x}) = Var[E(\mathbf{x} \mid \mathbf{g})] + E[Var(\mathbf{x} \mid \mathbf{g})] = \mathbf{B} + \mathbf{W} \ ,$$

where $\mathbf{B} = Var[E(\mathbf{x} \mid \mathbf{g})] = \sum_{i=1}^k \pi_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'$ and $\mathbf{W} = \sum_{i=1}^k \pi_i Var(\mathbf{x} \mid \mathbf{g} = i) = a\boldsymbol{\Sigma}$ are, respectively, the populational between and within dispersion matrices. This section presents limit results under the setup (15)-(16) for the conditional probability of misclassification of both LDA and $RLDA$ rules.

3.1 Strong consistency of LDA

For identical class prior probabilities $\pi_i = 1/k$, the ith subset of the Bayes rule $r^*(\mathbf{x}) = \sum_{i=1}^k i I_{R_i^*}(\mathbf{x})$ is determined by condition $f_i(\mathbf{x}) = \max_{1 \le j \le k} f_j(\mathbf{x})$ so, if $f_i(\mathbf{x})$ is as in (16), R_i^* is formed by all the points $\mathbf{x}$ such that

$$|\boldsymbol{\Sigma}|^{-1/2} g[(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)] = \max_{1 \le j \le k} |\boldsymbol{\Sigma}|^{-1/2} g[(\mathbf{x} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)] \ . \tag{17}$$

Moreover, if the function $g(.)$ is positive and strictly decreasing, using $\mathbf{W} = a\boldsymbol{\Sigma}$, $a > 0$, (17) is equivalent to $(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{W}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) = \min_{1 \le j \le k} (\mathbf{x} - \boldsymbol{\mu}_j)'\mathbf{W}^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)$. Replacing $\boldsymbol{\mu}_i$ and $\mathbf{W}$ by respectively estimators $\overline{\mathbf{x}}_i$ and $\widehat{\boldsymbol{\Sigma}}_p$, the corresponding sample version of criterion (17) is then $(\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\boldsymbol{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i) = \min_{1 \le j \le k} (\mathbf{x} - \overline{\mathbf{x}}_j)'\widehat{\boldsymbol{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_j)$, exactly as in the first line of (6) in section 2.

Theorem 1 *If the class prior probabilities* $\pi_i = P[\mathbf{g} = i]$ *are identical and the feature vector* $\mathbf{x}$ *follows an inverse location regression model (15) with class conditional densities (16), where* $g(.)$ *is a continuous and strictly decreasing function such that* $\int_0^{+\infty} t^{p/2} g(t)dt < +\infty$ *and* $g(t) > 0$ *for all* $t \geq 0$*, then the LDA rule is strongly consistent.*

Proof. Put $\widehat{l}_n(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{L}_{i,n}}(\mathbf{x})$ for the *LDA* rule, where $\widehat{L}_{i,n}$ is the subset of $\mathbb{R}^p$ formed by the points $\mathbf{x}$ that satisfy condition $(\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i) = \min_{1 \leq j \leq k} (\mathbf{x} - \overline{\mathbf{x}}_j)'\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_j)$. According to section 1, the goal is to proof that the conditional probability of error $L_n = P[\widehat{l}_n(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$ converges *a.e.* as $n \to \infty$ to $L^* = P[r^*(\mathbf{x}) \neq \mathbf{g}]$, the optimum probability of error of the Bayes rule $r^*(\mathbf{x}) = \sum_{i=1}^{k} i I_{R_i^*}(\mathbf{x})$. To do this, notice that $\widehat{L}_{i,n}$ can be reexpressed as

$$\widehat{L}_{i,n} = \{\mathbf{x} : \widehat{f}_{i,n}(\mathbf{x}) = \max_{1 \leq j \leq k} \widehat{f}_{j,n}(\mathbf{x})\}, \tag{18}$$

where, for $i = 1, ..., k$,

$$\widehat{f}_{i,n}(\mathbf{x}) = \mid \widehat{\mathbf{\Sigma}} \mid^{-1/2} g[(\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i)], \tag{19}$$

and $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_p/a$. Since $\overline{\mathbf{x}}_i$ is an estimator of $\boldsymbol{\mu}_i$ and $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_p/a$ of $\mathbf{W}/a = \mathbf{\Sigma}$, $\widehat{f}_{i,n}(\mathbf{x})$ in (19) is an estimator of $f_i(\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2} g[(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)]$ in (16) so, by (18), the *LDA* rule can be seen as a plug-in version of the Bayes rule $r^*(\mathbf{x}) = \sum_{i=1}^{k} i I_{R_i^*}(\mathbf{x})$ given by subsets $R_i^* = \{\mathbf{x} : f_i(\mathbf{x}) = \max_{1 \leq j \leq k} f_j(\mathbf{x})\}$. By theorem 1 in Devroye and Györfi (1985, p. 254), the difference $L_n - L^*$ can be bounded in the form

$$0 \leq L_n - L^* \leq \frac{1}{k} \sum_{i=1}^{k} \int_{\mathbb{R}^p} \mid f_i(\mathbf{x}) - \widehat{f}_{i,n}(\mathbf{x}) \mid d\mathbf{x} \, . \tag{20}$$

For each fixed $1 \leq i \leq k$, the sequence of random functions $\{\widehat{f}_{i,n}(\mathbf{x})\}$ is, with probability one, a sequence of densities such that

$$0 \leq \int_{\mathbb{R}^p} \mid f_i(\mathbf{x}) - \widehat{f}_{i,n}(\mathbf{x}) \mid d\mathbf{x} = 2 \int_{\mathbb{R}^p} [f_i(\mathbf{x}) - \widehat{f}_{i,n}(\mathbf{x})]_+ \, d\mathbf{x} \, , \tag{21}$$

where $[f_i(\mathbf{x}) - \widehat{f}_{i,n}(\mathbf{x})]_+$ is the positive part of the difference $f_i(\mathbf{x}) - \widehat{f}_{i,n}(\mathbf{x})$. By the results of appendix 7.2, as $n \to \infty$, $\overline{\mathbf{x}}_i \to E(\mathbf{x} \mid \mathbf{g} = i) = \boldsymbol{\mu}_i$ and $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_p/a \to \mathbf{W}/a =$

$\mathbf{\Sigma}$, *a.e.*, and thus, since the function $g(.)$ in (19) is continuous, $[f_i(\mathbf{x}) - \widehat{f}_{i,n}(\mathbf{x})]_+$ converges *a.e.* to zero for all $\mathbf{x} \in \mathbb{R}^p$. On the other hand, $0 \leq [f_i(\mathbf{x}) - \widehat{f}_{i,n}(\mathbf{x})]_+ \leq f_i(\mathbf{x})$ so by lemma 3.1.3 in Glick (1974) (see also Prakasa Rao 1983, p. 191) $\int_{\mathbb{R}^p}[f_i(\mathbf{x}) - \widehat{f}_{i,n}(\mathbf{x})]_+ d\mathbf{x}$ converges to zero *a.e.* for all $i = 1, ..., k$, and by (20) and (21) this leads to $L_n \to L^*$ *a.e.* ■

Under the assumptions of theorem 1, the conditional probability of error $L_n = P[\widehat{l}_n(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$ is asymptotically close to the optimum L^*. Phrased differently, LDA should have a good behavior as long as the sample size n is large enough, the prior class probabilities π_i are identical and the class conditional distributions $\mathbf{x} \mid \mathbf{g} = i$ are described by an inverse location regression as (15), where the "error" $\mathbf{u}$ has an adequate spherically symmetric density $g(\mathbf{u}'\mathbf{u})$. This is a flexible model that includes a variety of distributions, among others: *i)* the multivariate normal, taking $g(t) = (2\pi)^{-p/2}\exp(-t/2)$; *ii)* mixtures of normals with the same dispersion shape, taking $g(t) = (2\pi)^{-p/2}[(1-\varepsilon)\exp(-t/2) + \varepsilon\sigma^{-p}\exp(-t/2\sigma^2)]$, where $0 < \varepsilon < 1$ and $\sigma > 0$; and *iii)* the multivariate Student's t_k distribution with $k > 2$ degrees of freedom, taking $g(t) = c(k,p)[1 + (t/k)]^{-(k+p)/2}$, where $c(k,p)$ is a constant depending only on k and p. Theorem 1 establishes then a robustness property of the LDA rule indicating that its good performance does not depend on specific gaussian assumptions for the class conditional densities $f_i(\mathbf{x})$, but on the existence instead of a wider homoscedastic inverse location regression model as (15)-(16) for the class conditional distributions $\mathbf{x} \mid \mathbf{g} = i$. Finally, when the class prior probabilities are not all identical, theorem 1 might not be true in general. For arbitrary class priors π_i, a weaker result can be however obtained. Specifically, if $f_i(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$, the modified LDA type criterion that assigns $\mathbf{x}$ to g_i when $(\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i) - 2\log(n_i/n) = \min_{1 \leq j \leq k}[(\mathbf{x} - \overline{\mathbf{x}}_j)'\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}}_j) - 2\log(n_j/n)]$, is strongly consistent. This can be verified combining the arguments in the proof above with the convergences $n_i/n \to \pi_i$ *a.e.* as $n \to \infty$ for $i = 1, ..., k$.

3.2 Asymptotic properties of *RLDA*

Using the notation of section 2, write $\widehat{l}_{r,n}(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{L}_{ri,n}}(\mathbf{x})$ for the sample *RLDA* rule based on r coordinates where, for $i = 1, ..., k$, $\widehat{L}_{ri,n}$ is the subset of $\mathbb{R}^p$ formed by the $\mathbf{x}$ that satisfy condition (13), namely $\|\widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_i)\|^2 = \min_{1\leq j\leq k} \|\widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_j)\|^2$, where $\widehat{\mathbf{C}}_1(r) = (\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2, ..., \widehat{\boldsymbol{\gamma}}_r)$ is the $p \times r$ suborthogonal matrix formed by the first r eigenvectors of $\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2}$. This section analyzes, under the same assumptions than in theorem 1, the asymptotic behavior of the conditional probability of misclassification $L_n(r) = P[\widehat{l}_{r,n}(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$ as a function of the number of coordinates $1 \leq r \leq q = \min(k-1, p)$ used in (13).

As a first step, let $r_0 = r(\mathbf{B})$ be the rank of the populational between variation matrix $\mathbf{B} = \sum_{i=1}^{k}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'/k$, where $\boldsymbol{\mu} = \sum_{i=1}^{k} \boldsymbol{\mu}_i/k$, and consider the spectral representation $\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2} = \mathbf{CDC}'$, where $\mathbf{C} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p)$ is a $p \times p$ orthogonal matrix of normalized eigenvectors $\boldsymbol{\gamma}_j$ and $\mathbf{D} = diag(\lambda_1, ..., \lambda_{r_0}, \lambda_{r_0+1}, ..., \lambda_p)$ is a $p \times p$ diagonal matrix of eigenvalues $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_{r_0} > 0 = \lambda_{r_0+1} = ... = \lambda_p$. For an adequate value of r, partition $\mathbf{C} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_r \,|\boldsymbol{\gamma}_{r+1}, ..., \boldsymbol{\gamma}_p) = (\mathbf{C}_1(r) \,|\mathbf{C}_2(r))$ into matrices $\mathbf{C}_1(r) = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_r)$ of $p \times r$ and $\mathbf{C}_2(r) = (\boldsymbol{\gamma}_{r+1}, ..., \boldsymbol{\gamma}_p)$ of $p \times (p-r)$. For $1 \leq r \leq r_0$, the intention is to proof convergence of $L_n(r) = P[\widehat{l}_{r,n}(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$ to $L_r = L[l_r(\mathbf{x})] = P[l_r(\mathbf{x}) \neq \mathbf{g}]$, the probability of error of the populational *RLDA* rule based on r coordinates $l_r(\mathbf{x}) = \sum_{i=1}^{k} i I_{L_{r,i}}(\mathbf{x})$, given by subsets $L_{r,i} = \{\mathbf{x} : \|\mathbf{C}_1'(r)'\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)\|^2 = \min_{1\leq j\leq k} \|\mathbf{C}_1'(r)'\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_j)\|^2\}$. To do this, define for $i = 1, ..., k$ the random functions

$$\widehat{f}_{i,n}(r; \mathbf{x}) = |\widehat{\mathbf{V}}(r)|^{1/2} g[\widehat{Q}_i(r; \mathbf{x})] \;, \tag{22}$$

where $\widehat{\mathbf{V}}(r) = \widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{C}}_1(r)\widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}^{-1/2} + \mathbf{\Sigma}^{-1/2}\mathbf{C}_2(r)\mathbf{C}_2'(r)\mathbf{\Sigma}^{-1/2}$ is a $p \times p$ matrix, $\widehat{Q}_i(r; \mathbf{x}) = (\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{C}}_1(r)\widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_i) + (\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1/2}\mathbf{C}_2(r)\mathbf{C}_2'(r)\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, and $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}_p/a$ is as in the proof of theorem 1. Since the second summand in $\widehat{Q}_i(r; \mathbf{x})$ does not depend on i and the function $g(.)$ is strictly decreasing, rule

$\widehat{l}_{r,n}(\mathbf{x}) = \sum_{i=1}^{k} iI_{\widehat{L}_{ri,n}}(\mathbf{x})$ is equivalent to the *pseudo plug-in* classification criterion

$$\widehat{f}_{i,n}(r;\mathbf{x}) = \max_{1\leq j\leq k} \widehat{f}_{j,n}(r;\mathbf{x}) \ . \tag{23}$$

The asymptotic behavior of $L_n(r) = P[\widehat{l}_{r,n}(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$ depends then on the limit properties of the functions $\widehat{f}_{i,n}(r;\mathbf{x})$. These are summarized next in the following auxiliary result.

Proposition 1 *If* $\lambda_r > \lambda_{r+1}$,

$$\widehat{f}_{i,n}(r;\mathbf{x}) \to f_i(r;\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2}\, g[Q_i(r;\mathbf{x})] \ , \ a.e. \ , \tag{24}$$

as $n \to \infty$ *for all* $\mathbf{x}$, *where* $Q_i(r;\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1/2}\mathbf{C}_1(r)\mathbf{C}_1'(r)'\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i) + (\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1/2}\mathbf{C}_2(r)\mathbf{C}_2'(r)\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$. *Moreover, with probability one,* $\widehat{f}_{i,n}(r;\mathbf{x})$ *is a density function for* n *large enough.*

Proof. From appendix 7.2, $\widehat{\mathbf{\Sigma}}_p \to \mathbf{W} = a\mathbf{\Sigma}$ and $\widehat{\mathbf{B}} \to \mathbf{B}$ so $\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2} \to \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$. By lemma 2.1 in Tyler (1981, p.726), the orthogonal projection operator $\widehat{\mathbf{C}}_1(r)\widehat{\mathbf{C}}_1'(r)$ converges then *a.e.* to the orthogonal projection operator defined by the first r eigenvectors of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$. From identity $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} = \mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}/a$ this operator is $\mathbf{C}_1(r)\mathbf{C}_1'(r)$, where $\mathbf{C}_1(r) = (\boldsymbol{\gamma}_1, \ ..., \ \boldsymbol{\gamma}_r)$ is as defined above. As a consequence, $\widehat{\mathbf{V}}(r) \to \mathbf{\Sigma}^{-1/2}[\mathbf{C}_1(r)\mathbf{C}_1'(r) + \mathbf{C}_2(r)\mathbf{C}_2'(r)]\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1/2}\mathbf{C}\mathbf{C}'\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1}$ and $\widehat{Q}_i(r;\mathbf{x}) \to Q_i(r;\mathbf{x})$. On the other hand, consider the change of variable

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_r \\ \mathbf{u}_{(r)} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}) \\ \mathbf{C}_2'(r)\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}) \end{pmatrix} = \mathbf{A}\mathbf{x} + \mathbf{b} \ , \tag{25}$$

where $\mathbf{A} = (\widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{C}}_1(r) \left| \mathbf{\Sigma}^{-1/2}\mathbf{C}_2(r) \right.)'$ and $\mathbf{b} = -(\overline{\mathbf{x}}'\widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{C}}_1(r) \left| \boldsymbol{\mu}'\mathbf{\Sigma}^{-1/2}\mathbf{C}_2(r) \right.)'$. Since $\widehat{\mathbf{V}}(r) \to \mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}^{-1}$ is p.d., with probability one $\widehat{\mathbf{V}}(r)$ is also p.d. for n large enough so, since $\mathbf{A}'\mathbf{A} = \widehat{\mathbf{V}}(r)$, one has $r(\mathbf{A}) = r(\mathbf{A}'\mathbf{A}) = r(\widehat{\mathbf{V}}(r)) = p$ and $|\partial\mathbf{x}/\partial\mathbf{u}| = |\partial\mathbf{u}/\partial\mathbf{x}|^{-1} = |\mathbf{A}|^{-1} = |\widehat{\mathbf{V}}(r)|^{-1/2}$. By change of variable

$$\int_{\mathbb{R}^p} \widehat{f}_{i,n}(r;\mathbf{x})d\mathbf{x} = \int_{\mathbb{R}^p} g((\|\mathbf{u}_r - \widehat{\mathbf{M}}_{i,r}\|^2 + \|\mathbf{u}_{(r)}\|^2)d\mathbf{u} = \int_{\mathbb{R}^p} g(\mathbf{u}'\mathbf{u})d\mathbf{u} = 1 \ , \tag{26}$$

where $\widehat{\mathbf{M}}_{i,r} = \widehat{\mathbf{C}}_1'(r)\widehat{\mathbf{\Sigma}}^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})$, $i = 1, ..., k$. ■

The properties of the limit $f_i(r;\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2}\, g[Q_i(r;\mathbf{x})]$ in (24) are also of interest. The proof of the result below is given in appendix 7.3

Proposition 2 *$f_i(r;\mathbf{x})$ is a density function for each r. Moreover, for $r \leq s$, the probability of error $L_r = P[l_r(\mathbf{x}) \neq \mathbf{g}]$ can be obtained in terms of the family $\{f_i(s;\mathbf{x}) : 1 \leq i \leq k\}$ by means of the formula*

$$L_r = L[l_r(\mathbf{x})] = 1 - \frac{1}{k}\sum_{i=1}^{k}\int_{L_{r,i}} f_i(s;\mathbf{x})d\mathbf{x}\ . \tag{27}$$

Expression (27) leads to $L_r = 1 - \sum_{i=1}^{k}\int_{L_{r,i}} f_i(r;\mathbf{x})d\mathbf{x}/k$ so, by expression (2) in section 1 and observing that $L_{r,i} = \{\mathbf{x} : \|\mathbf{C}_1'(r)\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i)\|^2 = \min_{1\leq j\leq k}\|\mathbf{C}_1'(r)\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_j)\|^2\} = \{\mathbf{x} : f_i(r,\mathbf{x}) = \max_{1\leq j\leq k} f_j(r,\mathbf{x})\}$, $l_r(\mathbf{x}) = \sum_{i=1}^{k} i I_{L_{r,i}}(\mathbf{x})$ is the optimal rule in the discriminant problem defined by priors $\pi_i = 1/k$ and class conditional densities $f_i(r;\mathbf{x})$. (27) also implies $L_{r-1} = 1 - \sum_{i=1}^{k}\int_{L_{r-1,i}} f_i(r;\mathbf{x})d\mathbf{x}/k$. Using the partition $\mathbf{C}_1(r) = (\mathbf{C}_1(r-1)\,|\boldsymbol{\gamma}_r)$ and the identity

$$\sum_{i=1}^{k}[\boldsymbol{\gamma}_r'\mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})]^2/k =$$

$$= \sum_{i=1}^{k}\boldsymbol{\gamma}_r'\mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1/2}\boldsymbol{\gamma}_r/k = \boldsymbol{\gamma}_r'\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}\boldsymbol{\gamma}_r = \lambda_r\ , \tag{28}$$

it turns out that, if $\lambda_r > 0$, the subsets $L_{r-1,i} = \{\mathbf{x} : \|\mathbf{C}_1'(r-1)\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i)\|^2 = \min_{1\leq j\leq k}\|\mathbf{C}_1'(r-1)\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_j)\|^2\} = \{\mathbf{x} : f_i(r-1,\mathbf{x}) = \max_{1\leq j\leq k} f_j(r-1,\mathbf{x})\}$ define a different partition than the one used by rule $l_r(\mathbf{x}) = \sum_{i=1}^{k} i I_{L_{r,i}}(\mathbf{x})$ so, again by (2), $L_{r-1} = 1 - \sum_{i=1}^{k}\int_{L_{r-1,i}} f_i(r;\mathbf{x})d\mathbf{x}/k > 1 - \sum_{i=1}^{k}\int_{L_{r,i}} f_i(r;\mathbf{x})d\mathbf{x}/k = L_r$. Finally, the family $\{f_i(r_0;\mathbf{x}) : 1 \leq i \leq k\}$ coincides with the family of class conditional densities $\{f_i(\mathbf{x}) : 1 \leq i \leq k\}$. To see this, recall first that

$$(\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) = (\mathbf{x} - \boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1/2}\mathbf{C}\mathbf{C}'\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_i)$$

$$= (\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1/2}\mathbf{C}_1(r_0)\mathbf{C}_1'(r_0)\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i)+(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1/2}\mathbf{C}_2(r_0)\mathbf{C}_2'(r_0)\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i)\ . \tag{29}$$

Using repeatedly identity (28) for $r = r_0+1$, ..., p, it turns out that for all $i = 1$, ..., k $\mathbf{C}_2'(r_0)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu} = \mathbf{C}_2'(r_0)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}_i$ so, by (29), $Q_i(r_0;\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1/2}\mathbf{C}_1(r_0)\mathbf{C}_1'(r_0)$ $\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i)+(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1/2}\mathbf{C}_2(r_0)\mathbf{C}_2'(r_0)\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ and then $f_i(r_0,\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2}\, g[Q_i(r_0;\mathbf{x})] = |\boldsymbol{\Sigma}|^{-1/2}\, g[(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)] = f_i(\mathbf{x})$. This leads to $L_{r_0,i} = \{\mathbf{x} : f_i(r_0,\mathbf{x}) = \max_{1\le j\le k} f_j(r_0,\mathbf{x})\} = \{\mathbf{x} : f_i(\mathbf{x}) = \max_{1\le j\le k} f_j(\mathbf{x})\} = R_i^*$, that is, the *RLDA* rule $l_{r_0}(\mathbf{x}) = \sum_{i=1}^k iI_{L_{r_0,i}}(\mathbf{x})$ is identical to the Bayes rule $r^*(\mathbf{x}) = \sum_{i=1}^k iI_{R_i^*}(\mathbf{x})$ and thus $L_{r_0} = L[l_{r_0}(\mathbf{x})] = L[r^*(\mathbf{x})] = L^*$. The asymptotic behavior of $L_n(r) = P[\widehat{l}_{r,n}(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$ is characterized next for $1 \le r \le r_0$.

Theorem 2 *Under the assumptions of theorem 1, let $r_0 = r(\mathbf{B})$. If $1 \le r \le r_0$ and $\lambda_r > \lambda_{r+1}$, $L_n(r)$ converges a.e. as $n \to \infty$ to the probability of error $L_r = L[l_r(\mathbf{x})]$, where $L_1 > L_2 > \ldots > L_{r_0} = L^*$. In particular, $L_n(r_0)$ converges a.e. to the Bayes error $L_{r_0} = L^*$.*

Proof. Let $1 \le r \le r_0$ and consider the function $\widehat{h}_{i,n}(r;\mathbf{x}) = |\widehat{\mathbf{V}}(r)|^{1/2} g[\widehat{H}_i(r;\mathbf{x})]$, where $\widehat{H}_i(r;\mathbf{x}) = (\mathbf{x}-\overline{\mathbf{x}}_i)'\widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\mathbf{C}}_1(r)\widehat{\mathbf{C}}_1'(r)\widehat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}_i)+(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1/2}\mathbf{C}_2(r)\mathbf{C}_2'(r)$ $\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i)$. $\widehat{h}_{i,n}(r;\mathbf{x})$ has a similar structure than $\widehat{f}_{i,n}(r;\mathbf{x})$ in (22) so, by the same type arguments used in proposition 2, $\widehat{h}_{i,n}(r;\mathbf{x})$ is a density function for n large enough such that, if $\lambda_r > \lambda_{r+1}$, $\widehat{h}_{i,n}(r;\mathbf{x}) \to f_i(\mathbf{x})$ *a.e.* for all $\mathbf{x}$. Using expression (3) in section 1, the conditional probability of error $L_n(r) = P[\widehat{l}_{r,n}(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n]$ can be written as

$$L_n(r) = 1 - \frac{1}{k}\sum_{i=1}^k \int_{\widehat{L}_{r,i}} f_i(\mathbf{x})d\mathbf{x}$$

$$= 1 - \frac{1}{k}\sum_{i=1}^k \int_{\widehat{L}_{r,i}} [f_i(\mathbf{x}) - \widehat{h}_{i,n}(r;\mathbf{x})]d\mathbf{x} - \frac{1}{k}\sum_{i=1}^k \int_{\widehat{L}_{r,i}} \widehat{h}_{i,n}(r;\mathbf{x})d\mathbf{x}\ . \tag{30}$$

Considering now the change of variable (25), $\widehat{h}_{i,n}(r;\mathbf{x})$ transforms into $g(\|\mathbf{u}_r - \widehat{\mathbf{M}}_{i,r}\|^2 + \|\mathbf{u}_{(r)} - \mathbf{M}_{i2,r}\|^2)$, where $\widehat{\mathbf{M}}_{i,r} = \widehat{\mathbf{C}}_1'(r)\widehat{\boldsymbol{\Sigma}}^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})$ and $\mathbf{M}_{i2,r} = \mathbf{C}_2'(r)\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})$. Also, $\widehat{f}_{i,n}(r;\mathbf{x})$ transforms into $g(\|\mathbf{u}_r - \widehat{\mathbf{M}}_{i,r}\|^2 + \|\mathbf{u}_{(r)}\|^2)$ and the subset $\widehat{L}_{r,i} =$

$\{\mathbf{x} : \widehat{f}_{i,n}(r;\mathbf{x}) = \max_{1\le j\le k} \widehat{f}_{j,n}(r;\mathbf{x})\}$ into $\widehat{L}_{r,i}(\mathbf{u}_r) \times \mathbb{R}^{p-r}$, where $\widehat{L}_{r,i}(\mathbf{u}_r) = \{\mathbf{u}_r : \|\mathbf{u}_r - \widehat{\mathbf{M}}_{i,r}\|^2 = \min_{1\le j\le k} \|\mathbf{u}_r - \widehat{\mathbf{M}}_{j,r}\|^2\}$. By Fubini's theorem, the third term in (30) is

$$\frac{1}{k}\sum_{i=1}^{k}\int_{\widehat{L}_{r,i}} \widehat{h}_{i,n}(r;\mathbf{x})d\mathbf{x} = \frac{1}{k}\sum_{i=1}^{k}\int_{\widehat{L}_{r,i}(\mathbf{u}_r)\times\mathbb{R}^{p-r}} g(\|\mathbf{u}_r - \widehat{\mathbf{M}}_{i,r}\|^2 + \|\mathbf{u}_{(r)} - \mathbf{M}_{i2,r}\|^2)d\mathbf{u} =$$

$$= \frac{1}{k}\sum_{i=1}^{k}\int_{\widehat{L}_{r,i}(\mathbf{u}_r)} [\int_{\mathbb{R}^{p-r}} g(\|\mathbf{u}_r - \widehat{\mathbf{M}}_{i,r}\|^2 + \|\mathbf{u}_{(r)} - \mathbf{M}_{i2,r}\|^2)d\mathbf{u}_{(r)}]d\mathbf{u}_r =$$

$$= \frac{1}{k}\sum_{i=1}^{k}\int_{\widehat{L}_{r,i}(\mathbf{u}_r)\times\mathbb{R}^{p-r}} g(\|\mathbf{u}_r - \widehat{\mathbf{M}}_{i,r}\|^2 + \|\mathbf{u}_{(r)}\|^2)d\mathbf{u} = \frac{1}{k}\sum_{i=1}^{k}\int_{\widehat{L}_{r,i}} \widehat{f}_{i,n}(r;\mathbf{x})d\mathbf{x}\ , \quad (31)$$

so combining identity $L_r = 1 - \sum_{i=1}^{k}\int_{L_{r,i}} f_i(r;\mathbf{x})d\mathbf{x}/k$ with (30) and (31) it turns out that

$$L_n(r) - L_r = \frac{1}{k}\sum_{i=1}^{k}[\int_{L_{r,i}} f_i(r;\mathbf{x})d\mathbf{x} - \int_{\widehat{L}_{r,i}} \widehat{f}_{i,n}(r;\mathbf{x})d\mathbf{x}]$$

$$-\frac{1}{k}\sum_{i=1}^{k}\int_{\widehat{L}_{r,i}} [f_i(\mathbf{x}) - \widehat{h}_{i,n}(r;\mathbf{x})]d\mathbf{x}\ . \quad (32)$$

To proceed in (32), notice that since $L_{r,i} = \{\mathbf{x} : f_i(r,\mathbf{x}) = \max_{1\le j\le k} f_j(r,\mathbf{x})\}$ and $\widehat{L}_{ri,n} = \{\mathbf{x} : \widehat{f}_{i,n}(r;\mathbf{x}) = \max_{1\le j\le k} \widehat{f}_{j,n}(r;\mathbf{x})\}$, the following inequalities hold. On one hand,

$$\sum_{i=1}^{k}\int_{L_{r,i}} f_i(r;\mathbf{x})d\mathbf{x} = \sum_{i=1}^{k}\sum_{j=1}^{k}\int_{L_{r,i}\cap\widehat{L}_{r,j}} f_i(r;\mathbf{x})d\mathbf{x} = \sum_{j=1}^{k}\sum_{i=1}^{k}\int_{L_{r,i}\cap\widehat{L}_{r,j}} f_i(r;\mathbf{x})d\mathbf{x}$$

$$\ge \sum_{j=1}^{k}\sum_{i=1}^{k}\int_{L_{r,i}\cap\widehat{L}_{r,j}} f_j(r;\mathbf{x})d\mathbf{x} = \sum_{j=1}^{k}\int_{\widehat{L}_{r,j}} f_j(r;\mathbf{x})d\mathbf{x} \quad (33)$$

and, similarly, $\sum_{i=1}^{k}\int_{\widehat{L}_{r,i}} \widehat{f}_i(r;\mathbf{x})d\mathbf{x} \ge \sum_{i=1}^{k}\int_{L_{r,i}} \widehat{f}_i(r;\mathbf{x})d\mathbf{x}$. Therefore,

$$-\sum_{i=1}^{k}\int_{\mathbb{R}^p} \left|f_i(r;\mathbf{x}) - \widehat{f}_{i,n}(r;\mathbf{x})\right| d\mathbf{x} \le \sum_{i=1}^{k}\int_{\widehat{L}_{r,i}} [f_i(r;\mathbf{x}) - \widehat{f}_{i,n}(r;\mathbf{x})]d\mathbf{x}$$

$$\le \sum_{i=1}^{k}\int_{L_{r,i}} f_i(r;\mathbf{x})d\mathbf{x} - \sum_{i=1}^{k}\int_{\widehat{L}_{r,i}} \widehat{f}_{i,n}(r;\mathbf{x})d\mathbf{x}$$

$$\leq \sum_{i=1}^{k} \int_{L_{r,i}} [f_i(r;\mathbf{x}) - \widehat{f}_{i,n}(r;\mathbf{x})] d\mathbf{x} \leq \sum_{i=1}^{k} \int_{\mathbb{R}^p} \left| f_i(r;\mathbf{x}) - \widehat{f}_{i,n}(r;\mathbf{x}) \right| d\mathbf{x} \ ,$$

so, by (32), the difference $L_n(r) - L_r$ is such that

$$|L_n(r) - L_r| \leq \frac{1}{k} \sum_{i=1}^{k} \int_{\mathbb{R}^p} \left| f_i(r;\mathbf{x}) - \widehat{f}_{i,n}(r;\mathbf{x}) \right| d\mathbf{x} + \frac{1}{k} \sum_{i=1}^{k} \int_{\mathbb{R}^p} \left| f_i(\mathbf{x}) - \widehat{h}_{i,n}(r;\mathbf{x}) \right| d\mathbf{x}.$$

$$= \frac{2}{k} \sum_{i=1}^{k} \int_{\mathbb{R}^p} [f_i(r;\mathbf{x}) - \widehat{f}_{i,n}(r;\mathbf{x})]_+ d\mathbf{x} + \frac{2}{k} \sum_{i=1}^{k} \int_{\mathbb{R}^p} [f_i(\mathbf{x}) - \widehat{h}_{i,n}(r;\mathbf{x})]_+ d\mathbf{x} \ . \qquad (34)$$

If $\lambda_r > \lambda_{r+1}$, $\widehat{f}_{i,n}(r;\mathbf{x}) \to f_i(r;\mathbf{x})$ and $\widehat{h}_{i,n}(r;\mathbf{x}) \to f_i(\mathbf{x})$, *a.e.*, for all $\mathbf{x}$, so using in (34) lemma 3.1.3 in Glick (1974) as in the proof of theorem 1, $|L_n(r) - L_r|$ is bounded above by a quantity that converges to zero and then $L_n(r) \to L_r$, *a.e.*. In particular, $\lambda_{r_0} > 0 = \lambda_{r_0+1}$ so $L_n(r_0) \to L^*$, *a.e.*. The ordering $L_1 > L_2 > ... > L_{r_0} = L^*$ is a direct consequence of the eigenvalue structure $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{r_0} > 0 = \lambda_{r_0+1} = ...$ $= \lambda_p$ of $\mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}$ and the inequality $L_{r-1} > L_r$, valid for $\lambda_r > 0$. ■

The rank $r_0 = r(\mathbf{B})$ is easily seen to be less or equal than $q = \min(k-1,p)$. If $r_0 < q = \min(k-1,p)$, two possibilities exist when the number of directions used in the *RLDA* rule $\widehat{l}_{r,n}(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{L}_{ri,n}}(\mathbf{x})$ is $r > r_0$:

i) If $r_0 < r < q = \min(k-1,p)$, $\widehat{l}_{r,n}(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{L}_{ri,n}}(\mathbf{x})$ is equivalent to the *pseudo plug-in* criterion (23). However, in this case, $\lambda_r/a = 0$ is a multiple eigenvalue of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} = \mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}/a$ so, by lemma 2.1 in Tyler (1981, p.726), $\widehat{\mathbf{C}}_1(r)\widehat{\mathbf{C}}_1'(r)$ cannot be guaranteed to converge to $\mathbf{C}_1(r)\mathbf{C}_1'(r)$. Therefore $\widehat{f}_{i,n}(r;\mathbf{x})$ does not necessarily converge to $f_i(r;\mathbf{x})$ and the argument of theorem 2 does not apply. The asymptotic behavior of $L_n(r) > L^*$ remains then undetermined;

ii) If $r = q = \min(k-1,p)$, $\widehat{l}_{r,n}(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{L}_{ri,n}}(\mathbf{x})$ is equivalent to the *LDA* rule of (6). To verify this, recall that with probability one $r(\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2}) = q$ so, proceeding similarly as (12) in section 2, $\sum_{i=1}^{k} n_i \|\widehat{\mathbf{C}}_2'(q)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})\|^2/n = tr[\widehat{\mathbf{C}}_2'(q)\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{C}}_2(q)] = \sum_{j=q+1}^{p} \widehat{\lambda}_j = 0$, and therefore $\widehat{\mathbf{C}}_2'(q)\widehat{\mathbf{\Sigma}}_p^{-1/2}\overline{\mathbf{x}}_i = \widehat{\mathbf{C}}_2'(q)$ $\widehat{\mathbf{\Sigma}}_p^{-1/2}\overline{\mathbf{x}}$, *a.e.*, for $i = 1, ..., k$. Minimizing the quantities $\|\widehat{\mathbf{C}}_1'(q)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}_i)\|^2$

considered in (13) is then equivalent to minimizing

$$\|\widehat{\mathbf{C}}_1'(q)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}_i)\|^2 + \|\widehat{\mathbf{C}}_2'(q)\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}_i)\|^2 = \|\widehat{\mathbf{C}}'\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}_i)\|^2$$
$$= \|\widehat{\mathbf{\Sigma}}_p^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}_i)\|^2 = (\mathbf{x}-\overline{\mathbf{x}}_i)'\widehat{\mathbf{\Sigma}}_p^{-1}(\mathbf{x}-\overline{\mathbf{x}}_i) \ ,$$

exactly as in criterion (6). By theorem 1, $L_n(q)$ converges *a.e.* to the Bayes error L^*.

Summarizing the results of this section, $L_n(r)$ is consistent only for $r = r_0$ and $r = q = \min(k-1,p)$. The impact of the "non consistency" is worse for $1 \leq r < r_0$ than for $r_0 < r < q$. In the former case, *RLDA* ignores directions that are relevant for classification, while in the latter *RLDA* considers directions with no effective separative power. In particular, when $r_0 = r(\mathbf{B}) > 1$ classifying using the *LDF* function $\widehat{\mathbf{g}}_1'(\mathbf{x}-\overline{\mathbf{x}})$, which is just *RLDA* for $r = 1$, might have a poor behavior in applications. In conclusion, for $r = 1, 2, ..., q$, a plot of the conditional probability of error $L_n(r)$ versus r can be conjectured to have a marked decreasing pattern for $1 \leq r < r_0$. After reaching its "minimum" at $r = r_0$ the plot should have, as a result of the inclusion of *spurious* directions, an increasing erratic pattern for $r_0 < r \leq q$ with a trend to stability as r approaches towards q, due to the consistency of *LDA*. This is in agreement with the empirical behavior of the plot of the estimated error rates $\widehat{L}_n(r)$ versus r in some well studied classification problems with a large number of groups, as for example the *vowel data set*, studied thoroughly in Hastie et al. (2001, sec. 4.3), in which $k = 11$, $p = 10$ and $q = 10$. Of particular interest is figure 4.10 in page 96 of this book.

4. CHOOSING THE NUMBER OF DIRECTIONS IN *RLDA*

As mentioned in section 2, an important issue in *RLDA* is the choice of the number r of canonical coordinates to use in practice. The analysis after theorem 2 suggests that, in general, choosing $r = r_0 = r(\mathbf{B})$ can be recommended. As an illustration, in a two group problem with equal class priors $\pi_i = 1/2$, one has $k = 2$ and

$$\mathbf{B} = \frac{1}{2}\sum_{i=1}^{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' = \frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \ ,$$

so, if $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, $r_0 = r(\mathbf{B}) = 1 = q = \min(k-1,p) = r(\widehat{\mathbf{B}})$, independently of the dimension p of the feature vector $\mathbf{x} = (x_1, ..., x_p)'$. According to the previous section, *LDA*, *RLDA* and classifying using the values of the *LDF* function $\widehat{\mathbf{g}}_1'(\mathbf{x} - \overline{\mathbf{x}}) \sim (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\widehat{\boldsymbol{\Sigma}}_p^{-1}(\mathbf{x} - \overline{\mathbf{x}})$ are equivalent for a two group problem, and by theorems 1 and/or 2 consistent under an inverse location regression model with elliptical class densities (16). For problems with a moderate to large number of groups $k > 2$, $r_0 = r(\mathbf{B}) \leq q = \min(k-1,p) = r(\widehat{\mathbf{B}})$ is in general an unknown constant, and its true value should be assessed by some formal testing method. McLachlan (1992, p. 187) reviews inference techniques for $r_0 = r(\mathbf{B})$.

A classical alternative is to proceed by trial and error since, as mentioned by Hastie et al. (1994, p. 1256), in practice it is often enough to consider a low number $r \leq 3$ of canonical coordinates even in problems with a large number of groups k. This section explores the properties of a classification procedure based in selecting the number of directions as a function of the data $\mathbf{D}_n$ by means of the criterion

$$\widehat{r} = \widehat{r}(\mathbf{D}_n) = \text{first integer } 1 \leq r \leq q \text{ such that } \widehat{q}_r \geq C \ , \tag{35}$$

where, as introduced in section 2, $\widehat{q}_r = \sum_{j=1}^{r} \widehat{\lambda}_j / \sum_{j=1}^{q} \widehat{\lambda}_j$ is the cumulative relative proportion of separation among centroids, and C is a fixed positive constant close to one. This is in the original spirit of *RLDA* as motivated in expressions (11), (12) and (13) of section 1. In fact, for an adequate choice of C the consistency of a *feasible RLDA* rule of the form

$$\widehat{l}_{\widehat{r},n}(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{L}_{\widehat{r}i,n}}(\mathbf{x}) \tag{36}$$

can be established under the assumptions of theorem 1. To do this, recall the eigenvalue structure of $\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2}$ $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_{r_0} > 0 = \lambda_{r_0+1} = ... = \lambda_p$, define the populational cumulative separation proportions $q_r = \sum_{j=1}^{r} \lambda_j / \sum_{j=1}^{r_0} \lambda_j$, $r = 1$, ..., r_0, and put $q_0 = 0$.

Theorem 3 *Under the assumptions of theorem 1, the feasible RLDA rule of (35)-(36) is strongly consistent for all values of C such that $q_{r_0-1} < C < q_{r_0} = 1$.*

Proof. Consider a sequence $\mathbf{D}_\infty = \{(\mathbf{x}_k,\mathbf{g}_k) : k \geq 1\}$ of independent observations with the same distribution than the pair $(\mathbf{x},\mathbf{g})$. If $(\mathbf{x},\mathbf{g})$ and $\mathbf{D}_\infty$ are independent and $I_{\{0\}}(.)$ is the indicator function of the singleton $\{0\} \subset \mathbb{R}$, using standard properties of conditional expectation, the conditional probability of error $L_n = 1 - P[\widehat{l}_{\widehat{r},n}(\mathbf{x}) = \mathbf{g} \mid \mathbf{D}_n]$ can be represented as

$$L_n = 1 - E[I_{\{0\}}(\widehat{l}_{\widehat{r},n}(\mathbf{x}) - \mathbf{g}) \mid \mathbf{D}_n] = 1 - E[I_{\{0\}}(\widehat{l}_{\widehat{r},n}(\mathbf{x}) - \mathbf{g}) \mid \mathbf{D}_n,\{(\mathbf{x}_k, \mathbf{g}_k) : k > n\}]$$

$$= 1 - E[I_{\{0\}}(\widehat{l}_{\widehat{r},n}(\mathbf{x}) - \mathbf{g}) \mid \mathbf{D}_\infty] \ . \tag{37}$$

Similarly as in (37), the conditional probability of error of the $RLDA$ rule based on r_0 coordinates is $L_n(r_0) = P[\widehat{l}_{r_0}(\mathbf{x}) \neq \mathbf{g} \mid \mathbf{D}_n] = 1 - E[I_{\{0\}}(\widehat{l}_{r_0}(\mathbf{x}) - \mathbf{g}) \mid \mathbf{D}_\infty]$. By theorem 2 $L_n(r_0) \to L^*$, so to get convergence of L_n to L^* is then enough to establish $L_n - L_n(r_0) \to 0$, *a.e.*, as $n \to \infty$. If $I_{A_n}(.)$ is the indicator function of the subset $A_n = \{\mathbf{D}_n : \widehat{r} = \widehat{r}(\mathbf{D}_n) = r_0 = r(\mathbf{B})\}$, the feasible rule $\widehat{l}_{\widehat{r},n}(\mathbf{x})$ can be decomposed additively in the form

$$\widehat{l}_{\widehat{r},n}(\mathbf{x}) = \widehat{l}_{\widehat{r},n}(\mathbf{x})I_{A_n}(\mathbf{D}_n) + \widehat{l}_{\widehat{r},n}(\mathbf{x})I_{A_n^c}(\mathbf{D}_n)$$

$$= \widehat{l}_{r_0}(\mathbf{x})I_{A_n}(\mathbf{D}_n) + \widehat{l}_{\widehat{r},n}(\mathbf{x})I_{A_n^c}(\mathbf{D}_n) = \widehat{l}_{r_0}(\mathbf{x}) + Z_n \ , \tag{38}$$

where $Z_n = Z_n(\mathbf{x},\mathbf{D}_n) = [\widehat{l}_{\widehat{r},n}(\mathbf{x}) - \widehat{l}_{r_0}(\mathbf{x})]I_{A_n^c}(\mathbf{D}_n)$. Putting things together, the difference $L_n - L_n(r_0)$ can be written as

$$L_n - L_n(r_0) = E(W_n \mid \mathbf{D}_\infty) \ , \tag{39}$$

where, from (37) and (38), $W_n = I_{\{0\}}(\widehat{l}_{\widehat{r},n}(\mathbf{x}) - \mathbf{g}) - I_{\{0\}}(\widehat{l}_{r_0}(\mathbf{x}) - \mathbf{g}) = I_{\{0\}}([\widehat{l}_{r_0}(\mathbf{x}) - \mathbf{g}] + Z_n) - I_{\{0\}}(\widehat{l}_{r_0}(\mathbf{x}) - \mathbf{g})$. Observe that $|W_n| \leq 1$ so, by (39) and the dominated convergence theorem for conditional expectations (see e.g. Shiryayev 1984, p. 216), to get $L_n - L_n(r_0) \to 0$ *a.e.* is enough to verify that, as $n \to \infty$, $W_n \to 0$, *a.e.*.

Fix $\varepsilon > 0$. From the definition of $Z_n = [\widehat{l}_{\widehat{r},n}(\mathbf{x}) - \widehat{l}_{r_0}(\mathbf{x})]I_{A_n^c}(\mathbf{D}_n)$ given after (38), and using the structure of W_n,

$$P[\sup_{m \geq n} |W_m| \geq \varepsilon] \leq P[\bigcup_{m=n}^{\infty} \{Z_m \neq 0\}] \leq P[\bigcup_{m=n}^{\infty} A_m^c] \ , \tag{40}$$

so the task is then to check that the upper bound of (40) converges to zero as $n \to \infty$. Since $\widehat{\mathbf{\Sigma}}_p^{-1/2}\widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}_p^{-1/2} \to \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} = \mathbf{\Sigma}^{-1/2}\mathbf{B}\mathbf{\Sigma}^{-1/2}/a$, from lemma 2.1 in Tyler (1981, p.726), $\widehat{\lambda}_j \to \lambda_j/a > 0$ for $j = 1, ..., r_0$ and $\widehat{\lambda}_j \to 0$ for $j = r_0 + 1, ..., p$, so $\widehat{q}_r = \sum_{j=1}^{r} \widehat{\lambda}_j / \sum_{j=1}^{q} \widehat{\lambda}_j \to \sum_{j=1}^{r} \lambda_j / \sum_{j=1}^{r_0} \lambda_j = q_r$ for $r = 1, ..., r_0$, where all the convergences are in an *a.e.* sense. In what follows, the notational convention $\widehat{q}_0 = q_0 = 0$ is used. Since $A_n = \{\mathbf{D}_n : \widehat{r} = \widehat{r}(\mathbf{D}_n) = r_0 = r(\mathbf{B})\} = \bigcap_{r=0}^{r_0-1}\{\mathbf{D}_n : \widehat{q}_r < C\} \cap \{\mathbf{D}_n : \widehat{q}_{r_0} \geq C\}$, the inequality below holds for any $0 < a < \min\{C - q_{r_0-1}, q_{r_0} - C\} = \min\{C - q_{r_0-1}, 1 - C\}$,

$$P[\sup_{m \geq n} \max_{0 \leq r \leq r_0} |\widehat{q}_r - q_r| \leq a] \leq P[\bigcap_{m=n}^{\infty} A_m] \ . \tag{41}$$

Since $\max_{0 \leq r \leq r_0} |\widehat{q}_r - q_r| \to 0$, *a.e.*, the left hand side of inequality (41) converges to 1. By (40) $P[\sup_{m \geq n} |W_m| \geq \varepsilon] \leq P[\bigcup_{m=n}^{\infty} A_m^c] = 1 - P[\bigcap_{m=n}^{\infty} A_m] \to 0$ for all $\varepsilon > 0$, and then $W_n \to 0$, *a.e.* . ■

As a consequence of the proof above, one has

$$P[\sup_{m \geq n} |\widehat{r}(\mathbf{D}_m) - r_0| \geq \varepsilon] \leq P[\bigcup_{m=n}^{\infty} A_m^c] = 1 - P[\bigcap_{m=n}^{\infty} A_m] \to 0 \ ,$$

so $\widehat{r} \to r_0 = r(\mathbf{B})$, *a.e.* That is, the construction of the feasible rule (35)-(36) replaces in the theoretical *RLDA* rule $\widehat{l}_{r_0}(\mathbf{x})$ the unknown quantity r_0 by the strongly consistent estimator $\widehat{r}$ of (35). In a way, theorem 3 justifies then asymptotically the usual exploratory practice in *RLDA* of considering a number of directions r such that $\widehat{q}_r \geq C$, where C is a constant close enough to one and such that condition $q_{r_0-1} < C < 1$ holds. For example, in the vowel data example mentioned at the end of subsection 3.2, an analysis of the quantities $\widehat{q}_j$ leads to $\widehat{q}_1 = .5617$ and $\widehat{q}_2 = .9135$ so for $C = .90$ a choice of $\widehat{r} = 2$ seems appropriate for this data set. Notice that $\widehat{L}_n(2) = .4913$ is the minimum value in the plot of estimated error rates $\widehat{L}_n(r)$ based on test data, as displayed in figure 4.10 in Hastie et al. (2001, p. 96).

5. ROBUSTNESS TO HETEROSCEDASTICITY

Consider the following generalization of the setup (15), namely the model

$$\mathbf{x} \mid \mathbf{g} = i \stackrel{D}{=} \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{1/2}\mathbf{u} \ , \tag{42}$$

where, for $i = 1, ..., k$, $\boldsymbol{\mu}_i$ is a $p \times 1$ vector, $\boldsymbol{\Sigma}_i$ is a $p \times p$ p.d. matrix and $\mathbf{u}$ is a $p \times 1$ random vector independent of the class label $\mathbf{g}$. When (42) holds, $\mathbf{x}$ is said to follow an *inverse location-scale regression* model (Cook and Yin 2001, p. 160). Under the assumption that $\mathbf{u}$ has the radial density $g(\mathbf{u}'\mathbf{u})$, the class conditional densities are now

$$p_i(\mathbf{x}) = |\boldsymbol{\Sigma}_i|^{-1/2}\, g[(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)] \ . \tag{43}$$

If moments of order two exist and the class priors are identical, the populational within variation $p \times p$ matrix is $\mathbf{W} = E[Var(\mathbf{x} \mid \mathbf{g})] = \sum_{i=1}^{k} Var(\mathbf{x} \mid \mathbf{g} = i)/k = a\sum_{i=1}^{k} \boldsymbol{\Sigma}_i/k$, where $a > 0$ is the same constant than in section 3.

Assuming a *mild* degree of heterogeneity among the $\boldsymbol{\Sigma}_i$, LDA can be seen to possess some *approximately* optimal properties under the setup (42)-(43). Let $L_n = P[\widehat{l}_n(\mathbf{x}) = \mathbf{g} \mid \mathbf{D}_n]$ be the conditional probability of error of the LDA rule $\widehat{l}_n(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{L}_{i,n}}(\mathbf{x})$ where, recalling the notation of section 3, $\widehat{L}_{i,n} = \{\mathbf{x} : \widehat{f}_{i,n}(\mathbf{x}) = \max_{1\le j\le k} \widehat{f}_{j,n}(\mathbf{x})\}$, $\widehat{f}_{i,n}(\mathbf{x}) = \mid \widehat{\boldsymbol{\Sigma}} \mid^{-1/2} g[(\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i)]$, and $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_p/a$. Define also for $i = 1, ..., k$ the density $h_i(\mathbf{x}) = |\mathbf{W}/a|^{-1/2}\, g[(\mathbf{x} - \boldsymbol{\mu}_i)'(\mathbf{W}/a)^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)]$, where $\mathbf{W}/a = \sum_{i=1}^{k} \boldsymbol{\Sigma}_i/k$ is as above. Let $L^* = P[r^*(\mathbf{x}) \neq \mathbf{g}]$ be the probability of error of the Bayes rule $r^*(\mathbf{x}) = \sum_{i=1}^{k} i I_{R_i^*}(\mathbf{x})$ given by subsets $R_i^* = \{\mathbf{x} : p_i(\mathbf{x}) = \max_{1\le j\le k} p_j(\mathbf{x})\}$. By the same argument than in the proof of theorem 1, the inequality below holds

$$0 \le L_n - L^* \le \frac{1}{k}\sum_{i=1}^{k} \int\limits_{\mathbb{R}^p} \left| \widehat{f}_{i,n}(\mathbf{x}) - p_i(\mathbf{x}) \right| d\mathbf{x}$$

$$\le \frac{1}{k}\sum_{i=1}^{k} \int\limits_{\mathbb{R}^p} \left| \widehat{f}_{i,n}(\mathbf{x}) - h_i(\mathbf{x}) \right| d\mathbf{x} + \frac{1}{k}\sum_{i=1}^{k} \int\limits_{\mathbb{R}^p} |p_i(\mathbf{x}) - h_i(\mathbf{x})|\, d\mathbf{x} \ . \tag{44}$$

By the convergences of appendix 7.2, $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_p/a \to \mathbf{W}/a$ and $\overline{\mathbf{x}}_i \to \boldsymbol{\mu}_i$ so $\widehat{f}_{i,n}(\mathbf{x}) \to h_i(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$. Proceeding as in theorem 1, the random first summand in

the upper bound (44) converges to zero *a.e.* as $n \to \infty$. On the other hand, if the location vectors $\boldsymbol{\mu}_i$ are kept fixed while *all* the $\boldsymbol{\Sigma}_i$ *tend* to a common $p \times p$ p.d. matrix $\boldsymbol{\Sigma}$, then $\mathbf{W}/a = \sum_{i=1}^{k} \boldsymbol{\Sigma}_i/k \to \boldsymbol{\Sigma}$ and therefore, for all $\mathbf{x} \in \mathbb{R}^p$, $p_i(\mathbf{x})$ and $h_i(\mathbf{x})$ converge to $u_i(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} g[(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)]$. Since $p_i(\mathbf{x})$, $h_i(\mathbf{x})$ and $u_i(\mathbf{x})$ are densities, it turns out that

$$\frac{1}{k}\sum_{i=1}^{k}\int_{\mathbb{R}^p}|p_i(\mathbf{x})-h_i(\mathbf{x})|\,d\mathbf{x} \le \frac{1}{k}\sum_{i=1}^{k}\int_{\mathbb{R}^p}|p_i(\mathbf{x})-u_i(\mathbf{x})|\,d\mathbf{x} + \frac{1}{k}\sum_{i=1}^{k}\int_{\mathbb{R}^p}|u_i(\mathbf{x})-h_i(\mathbf{x})|\,d\mathbf{x}$$

$$= \frac{2}{k}\sum_{i=1}^{k}\int_{\mathbb{R}^p}[u_i(\mathbf{x})-p_i(\mathbf{x})]_+ d\mathbf{x} + \frac{2}{k}\sum_{i=1}^{k}\int_{\mathbb{R}^p}[u_i(\mathbf{x})-h_i(\mathbf{x})]_+ d\mathbf{x}\ , \qquad (45)$$

so, since $0 \le [u_i(\mathbf{x})-p_i(\mathbf{x})]_+ \le u_i(\mathbf{x})$ and $[u_i(\mathbf{x})-h_i(\mathbf{x})]_+ \le u_i(\mathbf{x})$, by the dominated convergence theorem the right hand side of (45) and thus the second summand of (44), converge to zero when $\boldsymbol{\Sigma}_i \to \boldsymbol{\Sigma}$, $i = 1, ..., k$. In conclusion, when all the $\boldsymbol{\Sigma}_i \cong \boldsymbol{\Sigma}$, the difference $L_n - L^*$ is bounded above by the sum of two terms close to zero, and L_n should be then close to the optimum L^* for a sample size n large enough. This argument might serve as an analytical explanation for the robustness of *LDA* to some small degree of class dispersion heterogeneity (see e.g. Seber 1984, p. 299).

Suppose finally that the $\boldsymbol{\Sigma}_i$ are markedly *different* to each other but such that in the log scale their determinants are *similar,* that is, such that $\log|\boldsymbol{\Sigma}_i| \cong c$, $i = 1, ..., k$, where c is some fixed constant. For equal class priors $\pi_i = 1/k$ and class conditional densities (43), the *ith* subset of the Bayes rule $r^*(\mathbf{x}) = \sum_{i=1}^{k} i I_{R_i^*}(\mathbf{x})$ is determined by condition $p_i(\mathbf{x}) = \max_{1\le j\le k} p_j(\mathbf{x})$, that is,

$$|\boldsymbol{\Sigma}_i|^{-1/2} g[(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)] = \max_{1\le j\le k} |\boldsymbol{\Sigma}_j|^{-1/2} g[(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)] \qquad (46)$$

or, taking logs, by $-1/2\log|\boldsymbol{\Sigma}_i| + \log g[(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)] = \max_{1\le j\le k} -1/2\log|\boldsymbol{\Sigma}_j| + \log g[(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)]$. But, since $g(.)$ is strictly decreasing and all the $\log|\boldsymbol{\Sigma}_i| \cong c$, (46) is *approximately* equivalent to condition

$$\log|\boldsymbol{\Sigma}_i| + (\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) = \min_{1\le j\le k} \log|\boldsymbol{\Sigma}_j| + (\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\ . \qquad (47)$$

Replacing $\boldsymbol{\Sigma}_i$ by $\widehat{\boldsymbol{\Sigma}}_i = \sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)'/n_i$ and $\boldsymbol{\mu}_i$ by $\overline{\mathbf{x}}_i$, the sample version of (47) is the familiar *quadratic discriminant analysis* (QDA) rule $\widehat{q}_n(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{Q}_{i,n}}(\mathbf{x})$, where $\widehat{Q}_{i,n}$ is formed by all the $\mathbf{x}$ such that

$$\log \left|\widehat{\boldsymbol{\Sigma}}_i\right| + (\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i) = \min_{1 \le j \le k} \log \left|\widehat{\boldsymbol{\Sigma}}_j\right| + (\mathbf{x} - \overline{\mathbf{x}}_j)'\widehat{\boldsymbol{\Sigma}}_j^{-1}(\mathbf{x} - \overline{\mathbf{x}}_j) \ . \quad (48)$$

Assuming $\log \left|\widehat{\boldsymbol{\Sigma}}_i\right| \cong c$ for $i = 1, ..., k$, and using again the strict monotonicity of $g(.)$, (48) is approximately equivalent to the plug-in criterion $\widehat{p}_{i,n}(\mathbf{x}) = \max_{1 \le j \le k} \widehat{p}_{j,n}(\mathbf{x})$, where $\widehat{p}_{i,n}(\mathbf{x}) = \left|\widehat{\boldsymbol{\Sigma}}_i\right|^{-1/2} g[(\mathbf{x} - \overline{\mathbf{x}}_i)'\widehat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x} - \overline{\mathbf{x}}_i)]$. Let $\widehat{r}_n(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{R}_{i,n}}(\mathbf{x})$ be the sample rule determined by the partition $\widehat{R}_{i,n} = \{\mathbf{x} : \widehat{p}_{i,n}(\mathbf{x}) = \max_{1 \le j \le k} \widehat{p}_{j,n}(\mathbf{x})\}$, $i = 1, ..., k$. Since $\widehat{p}_{i,n}(\mathbf{x})$ is an estimator of $p_i(\mathbf{x}) = |\boldsymbol{\Sigma}_i|^{-1/2} g[(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)]$, arguing as in the proof of theorem 1, it turns out that under the setup (42)-(43) the conditional probability of error $P[\widehat{r}_n(\mathbf{x}) = \mathbf{g} \mid \mathbf{D}_n] \to L^*$, *a.e.*. That is, since $\widehat{q}_n(\mathbf{x}) \cong \widehat{r}_n(\mathbf{x})$, the conditional probability of misclassification of the QDA rule $P[\widehat{q}_n(\mathbf{x}) = \mathbf{g} \mid \mathbf{D}_n] \cong P[\widehat{r}_n(\mathbf{x}) = \mathbf{g} \mid \mathbf{D}_n] \cong L^*$ should be close to the Bayes error for sample sizes n large enough.

As a summary of the results of this section, suppose equal class prior probabilities $\pi_i = 1/k$ and class conditional densities (43). When all the $\boldsymbol{\Sigma}_i \cong \boldsymbol{\Sigma}$, the LDA rule $\widehat{l}_n(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{L}_{i,n}}(\mathbf{x})$ should have a good behavior for n large enough. If, on the contrary, the $\boldsymbol{\Sigma}_i$ are different but $\log |\boldsymbol{\Sigma}_i| \cong c$, the conditional probability of error of the QDA rule $\widehat{q}_n(\mathbf{x}) = \sum_{i=1}^{k} i I_{\widehat{Q}_{i,n}}(\mathbf{x})$ should be expected to be close to the Bayes error. Notice that the condition $\log |\boldsymbol{\Sigma}_i| \cong c$, $i = 1, ..., k$, is quite flexible, since even in the case when the determinants $|\boldsymbol{\Sigma}_i|$ are large and different, they will tend to be more similar in the log scale.

6. FINAL COMMENTS

Hastie et al. (2001, p. 89), taking as a reference the results reported in the STATLOG project by Michie et al. (1994), comment on the good track record of LDA and QDA in a diverse set of applications. According to these authors, the reason for this property does not seem to lie in the approximate gaussianity of the class

conditional densities but on the fact that the data can only support simple linear or quadratic separation boundaries. Robustness of LDA and QDA has received recent attention in Cook and Yin (2001) who study the connection of LDA and QDA with, respectively, the dimension reduction methods *Sliced Inverse Regression* (SIR) of Li (1991) and *Sliced Average Variance Estimation* ($SAVE$) of Cook and Weisberg (1991). Hastie and Zhu (2001) provide additional insights on the relationships $LDA-SIR$ and $QDA-SAVE$.

This paper offers an alternative analytical explanation for the good performance in applications of LDA and QDA. The explanation is based on the description of the asymptotic behavior of the corresponding probabilities of misclassification under a wide set of assumptions, among others, the existence of second order inverse location regression models for the class conditional densities with an error modelled by a radially symmetric density. Resorting to asymptotics can be justified by the typical use in practice of moderate to large data sets. Section 3.1 gives results relative to the behavior of LDA in the homoscedastic case while section 5 offers some arguments of approximate consistency of LDA and QDA in the heteroscedastic case. Combining all these results together, it turns out that LDA and QDA are bound to behave properly in a large collection of situations. In addition, the results obtained in section 3.2 can offer, as developed in section 4, some guidelines for choosing in practice the number of directions in $RLDA$.

7. APPENDIX

7.1 Results on ranks

Lemma 1 *Let* $\mathbf{X}_1$, ..., $\mathbf{X}_k$ *be independent data matrices such that* $\mathbf{X}_i$ *is a matrix of* $n_i \times p$ *whose rows are i.i.d. random vectors with density* $f_i(.)$. *With probability one: i)* $\widehat{\mathbf{\Sigma}}_p$ *is p.d. if* $n = \sum_{i=1}^{k} n_i \geq p + k$ *and ii) the rank of* $\widehat{\mathbf{B}}$ *is* $\min(k-1, p)$.

Proof. For reasons of conciseness, the proof of this result is only sketched. To see

i), write $(n-k)\widehat{\mathbf{\Sigma}}_p = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\overline{\mathbf{x}}_i)(\mathbf{x}_{ij}-\overline{\mathbf{x}}_i)' = \sum_{i=1}^{k}\mathbf{X}_i'\mathbf{P}_i\mathbf{X}_i = \mathbf{X}'\mathbf{P}\mathbf{X}$, where the $\mathbf{P}_i = \mathbf{I}_{n_i} - \mathbf{1}_{n_i}\mathbf{1}_{n_i}'/n_i$ are orthogonal projection matrices of $n_i \times n_i$, $\mathbf{1}_{n_i} = (1, \overset{(n_i)}{...}, 1)'$ is the vector of ones of order n_i, $\mathbf{X}' = (\mathbf{X}_1' \mid ... \mid \mathbf{X}_k')$ is the combined data matrix of $p \times n$ and $\mathbf{P} = diag(\mathbf{P}_1, ..., \mathbf{P}_k)$ is a block diagonal matrix of $n \times n$. It is easy to see that the rank of $\mathbf{P}$ is $n-k$ so if $n-k \geq p$ the matrix $\mathbf{X}'\mathbf{P}\mathbf{X}$ is p.d. with probability one by theorem 2.3 in Eaton and Perlman (1973, p. 711). For part *ii)*, notice that $n\widehat{\mathbf{B}} = \sum_{i=1}^{k} n_i(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})' = \mathbf{Y}'\mathbf{A}\mathbf{Y}$, where $\mathbf{Y}' = (\overline{\mathbf{x}}_1, ..., \overline{\mathbf{x}}_k)$, $\mathbf{A} = \mathbf{C}'diag(n_1, ..., n_k)\mathbf{C}$ and $\mathbf{C} = \mathbf{I}_k - n^{-1}\mathbf{1}_k(n_1, ..., n_k)$. Also, $r(\mathbf{A}) = r(\mathbf{C}) = k-1$. Under the assumptions of the lemma, the sample means $\overline{\mathbf{x}}_i$, $i = 1, ..., k$, are independent and absolutely continuous random vectors with a joint density in $\mathbb{R}^{pk}$. Therefore, if $k-1 \geq p$, by theorem 2.3 in Eaton and Perlman (1973) with probability one $r(\widehat{\mathbf{B}}) = r(n\widehat{\mathbf{B}}) = r(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = p = \min(k-1,p)$. If $k-1 < p$, the rank of $\widehat{\mathbf{B}}$ is as the rank of the matrix of $(k-1) \times p$ $\mathbf{Y}_{(k)}$, where $\mathbf{Y}_{(k)}' = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}} \mid ... \mid \overline{\mathbf{x}}_{k-1} - \overline{\mathbf{x}})$. By relating $\mathbf{Y}_{(k)}$ to the rows of $\mathbf{Y}$, it can be seen that the rows of $\mathbf{Y}_{(k)}$ have a density and then $r(\widehat{\mathbf{B}}) = r(\mathbf{Y}_{(k)}) = k-1 = \min(k-1,p)$. ■

In problems in which the class conditional densities exist, lemma 1 shows that $P[r(\widehat{\mathbf{B}}) = \min(k-1,p) \mid \mathbf{G} = G] = 1$ as long as the class labels $\mathbf{G} = (\mathbf{g}_1, ..., \mathbf{g}_n)$ take a value $\mathbf{G} = G \in \{1, ..., k\}^n$ such that the sample sizes $n_i = \sum_{j=1}^{n} I_i(\mathbf{g}_j) \geq 1$, $i = 1, ..., k$, where $I_i(.)$ is the indicator function of the singleton $\{i\}$. Since $P(n_i = 0) = (1-\pi_i)^n \to 0$, the samples $\mathbf{D}_n$ in which some $n_i = 0$ form a set with probability tending to zero as $n \to \infty$. Notice that this result for the rank of $\widehat{\mathbf{B}}$ holds independently of the value of the rank $r_0 = r(\mathbf{B}) \leq \min(k-1,p) = r(\widehat{\mathbf{B}})$ of the populational matrix $\mathbf{B}$. In fact, in applications in which k is relatively large as compared to p it may well occur that $r_0 = r(\mathbf{B}) \ll r(\widehat{\mathbf{B}})$.

7.2 Auxiliary convergences

All the auxiliary convergences used in the paper are a consequence of the law of the large numbers for *i.i.d.* random variables with finite first order moments. For

example, $n_i/n = \sum_{j=1}^{n} I_i(\mathbf{g}_j)/n \rightarrow E[I_i(\mathbf{g})] = P(\mathbf{g} = i) = \pi_i$ *a.e.* as $n \rightarrow \infty$. Also, $\overline{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i = (\sum_{j=1}^{n} \mathbf{x}_j I_i(\mathbf{g}_j)/n)/(\sum_{j=1}^{n} I_i(\mathbf{g}_j)/n) \rightarrow E[\mathbf{x}I_i(\mathbf{g})]/\pi_i = \sum_{j=1}^{k} P(\mathbf{g} = j)E[\mathbf{x}I_i(\mathbf{g}) \mid \mathbf{g} = j]/\pi_i = P(\mathbf{g} = i)E[\mathbf{x} \mid \mathbf{g} = i]/\pi_i = \boldsymbol{\mu}_i$. Convergences as $n \rightarrow \infty$ of $\widehat{\mathbf{B}}$ to $\mathbf{B}$ and of $\widehat{\boldsymbol{\Sigma}}_p$ to $\mathbf{W} = \sum_{i=1}^{k} \pi_i Var(\mathbf{x} \mid \mathbf{g} = i)$, can be treated similarly.

7.3 Proof of proposition 2

Consider the change of variable transformation below,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_r \\ \mathbf{y}_{(r)} \end{pmatrix} = \mathbf{C}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) = \begin{pmatrix} \mathbf{C}_1'(r)\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \\ \mathbf{C}_2'(r)\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \end{pmatrix} . \tag{49}$$

Under (49) the class conditional mean vector $\boldsymbol{\mu}_i$ transforms into $\mathbf{M}_i = \mathbf{C}'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}) = (\mathbf{M}_{i1,r}', \mathbf{M}_{i2,r}')'$, where $\mathbf{M}_{i1,r} = \mathbf{C}_1'(r)\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})$ and $\mathbf{M}_{i2,r} = \mathbf{C}_2'(r)\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})$. Also, the *ith* class conditional density $f_i(\mathbf{x}) = |\boldsymbol{\Sigma}|^{-1/2} g[(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)]$ transforms into $g(\|\mathbf{y} - \mathbf{M}_i\|^2)$, and $f_i(s; \mathbf{x})$ into $g(\|\mathbf{y}_s - \mathbf{M}_{i1,s}\|^2 + \|\mathbf{y}_{(s)}\|^2)$. Finally, the subset $L_{r,i}$ transforms into $L_{r,i}(\mathbf{y}_r) \times \mathbb{R}^{p-r}$, where $L_{r,i}(\mathbf{y}_r) = \{\mathbf{y}_r : \|\mathbf{y}_r - \mathbf{M}_{i1,r}\|^2 = \min_{1 \leq j \leq k} \|\mathbf{y}_r - \mathbf{M}_{j1,r}\|^2\}$. If $s \geq r$, using Fubini's theorem and equation (1) in section 1, the probability of misclassification of $l_r(\mathbf{x}) = \sum_{i=1}^{k} i I_{L_{r,i}}(\mathbf{x})$ is

$$\begin{aligned}
L_r &= L[l_r(\mathbf{x})] = 1 - \frac{1}{k}\sum_{i=1}^{k} \int_{L_{r,i}} f_i(\mathbf{x})d\mathbf{x} = 1 - \frac{1}{k}\sum_{i=1}^{k} \int_{L_{r,i}(\mathbf{y}_r) \times \mathbb{R}^{p-r}} g(\|\mathbf{y} - \mathbf{M}_i\|^2)d\mathbf{y} \\
&= 1 - \frac{1}{k}\sum_{i=1}^{k} \int_{L_{r,i}(\mathbf{y}_r) \times \mathbb{R}^{s-r}} [\int_{\mathbb{R}^{p-s}} g(\|\mathbf{y}_s - \mathbf{M}_{i1,s}\|^2 + \|\mathbf{y}_{(s)} - \mathbf{M}_{i2,s}\|^2)d\mathbf{y}_{(s)}]d\mathbf{y}_s \\
&= 1 - \frac{1}{k}\sum_{i=1}^{k} \int_{L_{r,i}(\mathbf{y}_r) \times \mathbb{R}^{s-r}} [\int_{\mathbb{R}^{p-s}} g(\|\mathbf{y}_s - \mathbf{M}_{i1,s}\|^2 + \|\mathbf{y}_{(s)}\|^2)d\mathbf{y}_{(s)}]d\mathbf{y}_s \\
&= 1 - \frac{1}{k}\sum_{i=1}^{k} \int_{L_{r,i}} f_i(s;\mathbf{x})d\mathbf{x} \ ,
\end{aligned}$$

which is just (27).

REFERENCES

[1] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression for Dimension Reduction" by Li (1991). *Journal of the American Statistical Association*, **86**, 328-332.

[2] Cook, R. D. and Yin, X. (2001). Dimension Reduction and Visualization in Discriminant Analysis (with discussion). *Australian and New Zealand Journal of Statistics*, **43**(2), 147-199.

[3] Devroye, L. and Györfi, L. (1985). *Nonparametric density estimation.* New York: John Wiley.

[4] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition.* New York: Springer Verlag.

[5] Eaton, M. L. and Perlman, M. D. (1973). The Nonsingularity of Generalized Sample Covariance Matrices. *The Annals of Statistics*, **1**, 710-717.

[6] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

[7] Flury, B. (1997). *A First Course in Multivariate Analysis.* New York: John Wiley.

[8] Glick, N. (1974). Consistency conditions for probability estimators and integrals of density estimators. *Utilitas Mathematica*, **6**, 61-74.

[9] Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible Discriminant Analysis by Optimal Scoring. *Journal of the American Statistical Association*, **89**, 1255-1270.

[10] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York; Springer Verlag.

[11] Hastie, T. and Zhu, M. (2001). Discussion of "Dimension Reduction and Visualization in Discriminant Analysis", by Cook and Yin. *Australian and New Zealand Journal of Statistics*, **43**(2), 179-185.

[12] Johnson, R. A. and Wichern, D. W. (1998). *Applied Statistical Multivariate Analysis*, 4th Edn, Upper Saddle River NJ: Prentice Hall.

[13] Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-342.

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

[15] Michie, D., Spiegelhalter, D. and Taylor, C. (eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood Series in Artificial Intelligence, Ellis Horwood.

[16] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: John Wiley.

[17] Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*. New York: Academic Press.

[18] Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification (with discussion). *Journal of the Royal Statistical Society, Series B*, **10**, 159 - 203.

[19] Seber, G.A.F. (1984). *Multivariate Observations*. New York: John Wiley.

[20] Shiryayev, A. N. (1984). *Probability*. New York: Springer Verlag.

[21] Tyler, D. E. (1981). Asymptotic Inference for Eigenvectors. *The Annals of Statistics*, **9**, 725-736.