UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 01-66
Statistics and Econometrics Series 14
November 2001

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

DIMENSION REDUCTION TRANSFORMATIONS IN DISCRIMINANT ANALYSIS

Santiago Velilla and Adolfo Hernández*

Abstract

Dimension reduction transformations in discriminant analysis are introduced. Their properties, as well as sufficient conditions for their characterization, are studied. Special attention is given to the continuous case, of particular importance in applications. An effective data based dimension reduction algorithm is proposed and its behavior illustrated in a classification problem where the class conditional probability distributions are multivariate normal with different covariance matrices.

Keywords: Bayes error, consistent sample discriminant rule, effective dimension reduction, probability of misclassification.

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

Dimension Reduction Transformations in Discriminant Analysis

Santiago Velilla and Adolfo Hernández*

Abstract

Dimension reduction transformations in discriminant analysis are introduced. Their properties, as well as sufficient conditions for their characterization, are studied. Special attention is given to the continuous case, of particular importance in applications. An effective data based dimension reduction algorithm is proposed and its behavior illustrated in a classification problem where the class conditional probability distributions are multivariate normal with different covariance matrices.

AMS 1991 subject classifications: 62H30

Key words and phrases: Bayes error, consistent sample discriminant rule, effective dimension reduction, probability of misclassification.

*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).

1. INTRODUCTION: DISCRIMINANT RULES AND TRANSFORMATIONS

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 $\mathbf{x} = (x_1, ..., x_p)'$. Although the specific form of the assignment rule that gives the optimal solution to this problem is well known (see e.g. Anderson, 1984 chap. 6), its structure depends typically on unknown parameters that must be estimated from an appropriate database. However, as explained for example in McLachlan (1992 chap. 12), the practical performance of a sample discriminant rule tends to deteriorate when the number of dimensions p increases. This phenomenon motivates then, when p is large, the construction of dimension reduction methods for optimal classification using a lower number of coordinates. The aim of this paper is to propose a general framework for dimension reduction in discriminant analysis by introducing the class of dimension reduction transformations. This section establishes notation and presents some preliminary results.

Consider the pair $(\mathbf{x}, \mathbf{g})$, where $\mathbf{g}$ is the discrete random variable, often called class index or group label, that describes the unknown true class membership of the individual corresponding to the feature vector $\mathbf{x} = (x_1, ..., x_p)'$. The class index can be conveniently represented as taking values $\mathbf{g} = i$ with class prior probabilities $\pi_i = P[\mathbf{g} = i] > 0$, $i = 1, ..., k$. The joint distribution of $(\mathbf{x}, \mathbf{g})$ can be obtained as the product $P[\mathbf{x} \in C; \mathbf{g} = i] = P[\mathbf{g} = i] P[\mathbf{x} \in C | \mathbf{g} = i]$, for each $C \in \mathcal{B}^p$ and $i = 1, ..., k$, where $\mathcal{B}^p$ is the $\sigma-$field of Borel sets in $\mathbb{R}^p$. On the other hand, if μ is the marginal distribution of $\mathbf{x}$, by standard properties of conditional probability (see e.g., Billingsley, 1995 chap. 6), the joint of $(\mathbf{x}, \mathbf{g})$ can be alternatively expressed as a function of μ and the class posterior probabilities $\pi_i(\mathbf{x}) = P[\mathbf{g} = i | \mathbf{x}]$ that satisfy,

for each $C \in \mathcal{B}^p$ and $i = 1, ..., k$, the identity

$$P\left[\mathbf{x} \in C; \mathbf{g} = i\right] = \int_C P[\mathbf{g} = i \mid \mathbf{x}]\mu\left(d\mathbf{x}\right) = \int_C \pi_i(\mathbf{x})\mu\left(d\mathbf{x}\right) \ . \tag{1}$$

Once the probabilistic structure of a given classification problem has been formulated in terms of the elements that determine the joint distribution of $(\mathbf{x}, \mathbf{g})$, the space $\mathbb{R}^p$ is partitioned into a collection of Borel sets $R_1, ..., R_k$, and the individual corresponding to $\mathbf{x}$ assigned to the *ith* group whenever $\mathbf{x} \in R_i$. This generates a discriminant rule as a mapping $r : \mathbb{R}^p \to \{1, ..., k\}$ from $\mathbb{R}^p$, the sample space of $\mathbf{x}$, onto $\{1, ..., k\}$, the sample space of $\mathbf{g}$, such that $r(\mathbf{x}) = i$ for $\mathbf{x} \in R_i$ or, equivalently, such that $r(\mathbf{x}) = \sum_{i=1}^k iI_{R_i}(\mathbf{x})$, where $I_A(.)$ is the indicator function of the subset $A \in \mathcal{B}^p$. For fixed $(\mathbf{x}, \mathbf{g})$, there is an error when $r(\mathbf{x}) \neq \mathbf{g}$. From (1), the probability of error or misclassification of rule $r(\mathbf{x})$ is

$$\begin{aligned} L[r(\mathbf{x})] &= P[r(\mathbf{x}) \neq \mathbf{g}] = 1 - P[r(\mathbf{x}) = \mathbf{g}] = 1 - \sum_{i=1}^k P[r(\mathbf{x}) = i; \mathbf{g} = i] = \\ &= 1 - \sum_{i=1}^k P\left[\mathbf{x} \in R_i; \mathbf{g} = i\right] = 1 - \sum_{i=1}^k \int_{R_i} \pi_i(\mathbf{x})\mu\left(d\mathbf{x}\right) \ . \end{aligned} \tag{2}$$

A natural criterion for optimal classification is to select those rules that minimize this probability of error. Any rule $r^*(\mathbf{x})$ that minimizes the functional $L[r(\mathbf{x})]$ is called a Bayes rule and the corresponding minimum probability of misclassification $L^* = L[r^*(\mathbf{x})]$ is the Bayes error. The following auxiliary result establishes existence and uniqueness of Bayes rules.

Proposition 1

i) The probability of misclassification is minimized by any rule $r^*(\mathbf{x}) = \sum_{i=1}^k iI_{R_i^*}(\mathbf{x})$, *where the subsets* R_1^*, ..., R_k^* *form a measurable partition of* $\mathbb{R}^p$ *such that*

$$R_i^* \subseteq \{\mathbf{x} \in \mathbb{R}^p : \pi_i(\mathbf{x}) = \max_j \pi_j(\mathbf{x})\}, \quad i = 1, ..., k \ . \tag{3}$$

ii) Under condition (C1): $P[\pi_i(\mathbf{x}) = \pi_j(\mathbf{x})] = 0$, $i \neq j$, *if* $s^*(\mathbf{x})$ *is any other rule such that* $L[s^*(\mathbf{x})] = L^*$, *then* $P[r^*(\mathbf{x}) = s^*(\mathbf{x})] = 1$.

Proof. See appendix 6.1. ■

If the binary relation between rules is defined as: $s \sim r$ if, and only if, $P[r(\mathbf{x}) = s(\mathbf{x})] = 1$, is easy to see that $' \sim '$ is an equivalence relation. By (2), if rules $r(\mathbf{x})$ and $s(\mathbf{x})$ are equivalent, $L[r(\mathbf{x})] = L[s(\mathbf{x})]$. Part *i)* of proposition 1 above assures that the equivalence class generated by any rule $r^*(\mathbf{x})$ given by subsets R_i^* satisfying condition (3), is a class of optimal rules. Under condition *(C1)* of part *ii)*, the equivalence class of optimal rules is unique. A representative in this class could be the rule associated to taking $R_i^* = \{\mathbf{x} \in \mathbb{R}^p : i$ is the smallest integer such that $\pi_i(\mathbf{x}) = \max_j \pi_j(\mathbf{x})\}$, $i = 1, ..., k$. However, to simplify notation, it is convenient to write $R_i^* = \{\mathbf{x} \in \mathbb{R}^p : \pi_i(\mathbf{x}) = \max_j \pi_j(\mathbf{x})\}$. The intuitive meaning of *(C1)* is that with probability one, once the vector $\mathbf{x}$ is observed, there is a perfect ordering $\pi_{i_1}(\mathbf{x}) > \pi_{i_2}(\mathbf{x}) > ... > \pi_{i_k}(\mathbf{x})$ of the posterior class probabilities. By (3), the natural assignment to the class with the largest posterior probability is also optimal. Other results of existence and uniqueness of Bayes rules are available in the literature but the format of Proposition 1 is convenient for the purposes of this paper.

To analyze the effect of transforming the feature vector $\mathbf{x}$ on a given classification problem, consider an invertible Borel measurable transformation $t : \mathbb{R}^p \to \mathbb{R}^p$ and put $\mathbf{y} = t(\mathbf{x})$. Given the posterior class probabilities $q_i(\mathbf{y}) = P[\mathbf{g} = i \mid \mathbf{y}]$, by proposition 1 an optimal rule in the transformed space $\mathbf{y} = t(\mathbf{x})$ is $s^*(\mathbf{y}) = \sum_{i=1}^k i I_{S_i^*}(\mathbf{y})$ where, using the convention of the previous paragraph, $S_i^* = \{\mathbf{y} \in \mathbb{R}^p : q_i(\mathbf{y}) = \max_j q_j(\mathbf{y})\}$, $i = 1, ..., k$. Under condition *(C2)*: $P[q_i(\mathbf{y}) = q_j(\mathbf{y})] = 0$, $i \neq j$, this rule is also *unique*. Given a discriminant rule $r(\mathbf{x}) = \sum_{i=1}^k i I_{R_i}(\mathbf{x})$ in the original space $\mathbf{x}$, the pair (r, t) induces in the new space $\mathbf{y} = t(\mathbf{x})$ the rule

$$r_t(\mathbf{y}) = r[t^{-1}(\mathbf{y})] = \sum_{i=1}^k i I_{R_i}[t^{-1}(\mathbf{y})] = \sum_{i=1}^k i I_{t(R_i)}(\mathbf{y}) \ , \tag{4}$$

where $\mathbf{x} = t^{-1}(\mathbf{y})$ is the inverse transformation of $\mathbf{y} = t(\mathbf{x})$ and, for $i = 1, ..., k$, $t(R_i) = \{\mathbf{y} = t(\mathbf{x}) \in \mathbb{R}^p : \mathbf{x} \in R_i\}$. Since $P[\mathbf{y} = t(\mathbf{x}) \in t(R_i); \mathbf{g} = i] =$

$P[\mathbf{x} \in R_i; \mathbf{g} = i]$, using (2) one has

$$L[r_t(\mathbf{y})] = 1 - \sum_{i=1}^{k} P[\mathbf{y} = t(\mathbf{x}) \in t(R_i); \mathbf{g} = i] = 1 - \sum_{i=1}^{k} P[\mathbf{x} \in R_i; \mathbf{g} = i] = L[r(\mathbf{x})] \ , \tag{5}$$

so the probabilities of misclassification of rules $r_t(\mathbf{y})$ and $r(\mathbf{x})$ are the same. In a dual fashion, given a rule $s(\mathbf{y}) = \sum_{i=1}^{k} iI_{S_i}(\mathbf{y})$ in the space $\mathbf{y}$, construction (4) can be applied to the pair (s, t^{-1}) to obtain the induced discriminant rule $s_{t^{-1}}(\mathbf{x}) = s[t(\mathbf{x})] = \sum_{i=1}^{k} iI_{S_i}[t(\mathbf{x})] = \sum_{i=1}^{k} iI_{t^{-1}(S_i)}(\mathbf{x})$ where, for $i = 1, ..., k$, $t^{-1}(S_i) = \{\mathbf{x} = t^{-1}(\mathbf{y}) \in \mathbb{R}^p : \mathbf{y} \in S_i\}$. The result below follows.

Lemma 2 *Given an invertible and measurable transformation $\mathbf{y} = t(\mathbf{x})$, the optimal probabilities of misclassification or Bayes errors are the same in both the original and transformed spaces. Moreover, the rules induced by Bayes rules in a given space, either $\mathbf{x}$ or $\mathbf{y}$, are also Bayes rules in the corresponding transformed space.*

Proof. If $r^*(\mathbf{x})$ and $s^*(\mathbf{y})$ are Bayes rules in the spaces $\mathbf{x}$ and $\mathbf{y} = t(\mathbf{x})$ respectively, by (5) one has $L^* = L[r^*(\mathbf{x})] = L[r_t^*(\mathbf{y})] \geq L[s^*(\mathbf{y})]$ and $L[s^*(\mathbf{y})] = L[s_{t^{-1}}^*(\mathbf{x})] \geq L[r^*(\mathbf{x})] = L^*$. As a conclusion, $L[r^*(\mathbf{x})] = L[s^*(\mathbf{y})] = L^* = L[r_t^*(\mathbf{y})] = L[s_{t^{-1}}^*(\mathbf{x})]$. ■

2. DIMENSION REDUCTION TRANSFORMATIONS

Let $\mathbf{y} = t(\mathbf{x}) = (y_1, ..., y_p)'$ be an invertible measurable transformation and consider, for $r \leq p$, the $r-$dimensional random vector $\mathbf{y}_r = (y_1, ..., y_r)'$. This generates the partition $\mathbf{y} = (\mathbf{y}_r', \mathbf{y}_{(r)}')'$, where $\mathbf{y}_{(r)} = (y_{r+1}, ..., y_p)'$ is the $(p-r) \times 1$ vector formed by the coordinates not in $\mathbf{y}_r$. This notation can be easily adapted to the case in which $\mathbf{y}_r$ is formed by any subset of r components from $\mathbf{y} = (y_1, ..., y_p)'$.

By proposition 1, if the posterior class probabilities $\eta_i(\mathbf{y}_r) = P[\mathbf{g} = i \mid \mathbf{y}_r]$ satisfy condition *(C3):* $P[\eta_i(\mathbf{y}_r) = \eta_j(\mathbf{y}_r)] = 0$, $i \neq j$, the *unique* Bayes rule for classification

into g_1, ..., g_k with the information provided by $\mathbf{y}_r = (y_1, ..., y_r)'$, is

$$d^*(\mathbf{y}_r) = \sum_{i=1}^{k} i I_{U_i^*}(\mathbf{y}_r) \ , \tag{6}$$

where, under the usual convention, $U_i^* = \{\mathbf{y}_r \in \mathbb{R}^r : \eta_i(\mathbf{y}_r) = \max_j \eta_j(\mathbf{y}_r)\} \subseteq \mathbb{R}^r$, $i = 1, ..., k$. The following result proves that the Bayes error L^* is a lower bound for the probability of misclassification $L[d^*(\mathbf{y}_r)]$.

Proposition 3 *If $r^*(\mathbf{x})$ and $s^*(\mathbf{y})$ are Bayes rules in the spaces $\mathbf{x}$ and $\mathbf{y}$ respectively, the discriminant rule $d^*(\mathbf{y}_r)$ of (6) satisfies the inequality*

$$L[d^*(\mathbf{y}_r)] \geq L[s^*(\mathbf{y})] = L[r^*(\mathbf{x})] = L^* \ . \tag{7}$$

Proof. Let $\mu_{\mathbf{y}}$ and $\mu_{\mathbf{y}_r}$ be the probability distributions of $\mathbf{y} = t(\mathbf{x}) = (y_1, ..., y_p)'$ and $\mathbf{y}_r = (y_1, ..., y_r)'$ respectively. Using the subsets U_i^* construct, in the space $\mathbf{y} = (\mathbf{y}_r', \mathbf{y}_{(r)}')' \in \mathbb{R}^p$, the discriminant rule $u^*(\mathbf{y}) = \sum_{i=1}^{k} i I_{U_i^* \times \mathbb{R}^{p-r}}(\mathbf{y})$. Taking into account that, for each $C \in \mathcal{B}^r$,

$$\begin{aligned} \int_C \eta_i(\mathbf{y}_r) \mu_{\mathbf{y}_r}(d\mathbf{y}_r) &= P[\mathbf{y}_r \in C; \mathbf{g} = i] \\ &= P[\mathbf{y} \in C \times \mathbb{R}^{p-r}; \mathbf{g} = i] = \int_{C \times \mathbb{R}^{p-r}} q_i(\mathbf{y}) \mu_{\mathbf{y}}(d\mathbf{y}) \ , \end{aligned} \tag{8}$$

one has, by (2) and (8),

$$\begin{aligned} L[d^*(\mathbf{y}_r)] &= 1 - \sum_{i=1}^{k} \int_{U_i^*} \eta_i(\mathbf{y}_r) \mu_{\mathbf{y}_r}(d\mathbf{y}_r) \\ &= 1 - \sum_{i=1}^{k} \int_{U_i^* \times \mathbb{R}^{p-r}} q_i(\mathbf{y}) \mu_{\mathbf{y}}(d\mathbf{y}) \\ &= L[u^*(\mathbf{y})] \geq L[s^*(\mathbf{y})] = L[r^*(\mathbf{x})] = L^*. \end{aligned}$$

■

Inequality (7) will be, in general, strict. When equality holds, there is a dimension reduction in the classification problem from p to r dimensions.

Definition 4 *In the notation of proposition 3, the invertible measurable transformation* $\mathbf{y} = t(\mathbf{x})$ *is said to be a dimension reduction transformation (d.r.t.) when, for some* $r < p$,

$$L[d^*(\mathbf{y}_r)] = L[s^*(\mathbf{y})] = L[r^*(\mathbf{x})] = L^* \, .$$

The following result gives a sufficient condition for $\mathbf{y} = t(\mathbf{x})$ to be a d.r.t.

Theorem 5 *Consider an invertible measurable transformation* $\mathbf{y} = t(\mathbf{x})$ *and any optimal rule* $s^*(\mathbf{y}) = \sum_{i=1}^{k} i I_{S_i^*}(\mathbf{y})$ *in the space* $\mathbf{y}$. *Transformation* $\mathbf{y} = t(\mathbf{x})$ *is a d.r.t. if the subsets* S_1^*, ..., S_k^* *do not depend on the coordinates* $\mathbf{y}_{(r)} = (y_{r+1}, ..., y_p)'$, *i.e., there exists a measurable partition* $T_1, ..., T_k$ *of* $\mathbb{R}^r$ *such that* $\sum_{i=1}^{k} P\left[\mathbf{y} \in S_i^* \, \triangle \, (T_i \times \mathbb{R}^{p-r})\right] = 0$, *where* $\triangle$ *is the operator symmetric difference of subsets.*

Proof. Consider the rules $v(\mathbf{y}_r) = \sum_{i=1}^{k} i I_{T_i}(\mathbf{y}_r)$ and $T(\mathbf{y}) = \sum_{i=1}^{k} i I_{T_i \times \mathbb{R}^{p-r}}(\mathbf{y})$ in the spaces $\mathbf{y}_r = (y_1, ..., y_r)'$ and $\mathbf{y} = (\mathbf{y}_r', \mathbf{y}_{(r)}')'$ respectively. By assumption, $P[s^*(\mathbf{y}) \neq T(\mathbf{y})] \leq \sum_{i=1}^{k} P\left[\mathbf{y} \in S_i^* \, \triangle \, (T_i \times \mathbb{R}^{p-r})\right] = 0$ so, using (2), $T(\mathbf{y})$ has the same probability of misclassification than rule $s^*(\mathbf{y})$. Also, by (8), $L[v(\mathbf{y}_r)] = L[T(\mathbf{y})]$ and, as a consequence, $L^* \leq L[d^*(\mathbf{y}_r)] \leq L[v(\mathbf{y}_r)] = L[T(\mathbf{y})] = L[s^*(\mathbf{y})] = L^*$. This leads to $L[d^*(\mathbf{y}_r)] = L^*$. ■

An alternative sufficient condition is also of interest.

Theorem 6 *The invertible measurable transformation* $\mathbf{y} = t(\mathbf{x})$ *is a d.r.t. if the class posterior probabilities* $q_i(\mathbf{y}) = P[\mathbf{g} = i \mid \mathbf{y}]$ *depend only on* $\mathbf{y}_r = (y_1, ..., y_r)'$, *that is, if for* $i = 1$, ..., k *there exist functions* $h_i(\mathbf{y}_r)$ *such that*

$$q_i(\mathbf{y}) = h_i(\mathbf{y}_r), \quad a.e. \ (\mu_{\mathbf{y}}) \ . \tag{9}$$

Proof. The first step is to verify that, under (9), $q_i(\mathbf{y}) = \eta_i(\mathbf{y}_r)$, *a.e.* $(\mu_{\mathbf{y}})$ for $i = 1$, .., k. By construction of $\eta_i(\mathbf{y}_r) = P[\mathbf{g} = i \mid \mathbf{y}_r]$ one has, for all $C_1 \in \mathcal{B}^r$,

$$P[\mathbf{y}_r \in C_1; \mathbf{g} = i] = \int_{C_1} \eta_i(\mathbf{y}_r) \mu_{\mathbf{y}_r}(d\mathbf{y}_r) \ . \tag{10}$$

On the other hand, using $\mathbf{y} = (\mathbf{y}_r^{'},\mathbf{y}_{(r)}^{'})'$, the left hand side of (10) coincides by assumption with

$$\begin{aligned} P\left[\mathbf{y} \in C_1 \times \mathbb{R}^{p-r}; \mathbf{g} = i\right] &= \int_{C_1 \times \mathbb{R}^{p-r}} q_i(\mathbf{y}) \mu_{\mathbf{y}}(d\mathbf{y}) = \int_{\mathbb{R}^p} h_i(\mathbf{y}_r) I_{C_1 \times \mathbb{R}^{p-r}}(\mathbf{y})\, \mu_{\mathbf{y}}(d\mathbf{y}) \\ &= E[h_i(\mathbf{y}_r) I_{C_1 \times \mathbb{R}^{p-r}}(\mathbf{y})] = E[h_i(\mathbf{y}_r) I_{C_1}(\mathbf{y}_r)] \\ &= \int_{C_1} h_i(\mathbf{y}_r) \mu_{\mathbf{y}_r}(d\mathbf{y}_r)\ . \end{aligned} \tag{11}$$

Comparing (10) and (11), $\int_{C_1} h_i(\mathbf{y}_r)\mu_{\mathbf{y}_r}(d\mathbf{y}_r) = \int_{C_1} \eta_i(\mathbf{y}_r)\mu_{\mathbf{y}_r}(d\mathbf{y}_r)$ for all $C_1 \in \mathcal{B}^r$ and this implies $h_i(\mathbf{y}_r) = \eta_i(\mathbf{y}_r)$, *a.e.* $(\mu_{\mathbf{y}_r})$. Since $h_i(\mathbf{y}_r)$ and $\eta_i(\mathbf{y}_r)$ depend only on $\mathbf{y}_r = (y_1, ..., y_r)'$ this also implies $h_i(\mathbf{y}_r) = \eta_i(\mathbf{y}_r)$, *a.e.* $(\mu_{\mathbf{y}})$. As a conclusion, $q_i(\mathbf{y}) = \eta_i(\mathbf{y}_r)$, *a.e.*$(\mu_{\mathbf{y}})$ for all $i = 1, ..., k$, and the subsets $S_i^* = \{\mathbf{y} \in \mathbb{R}^p : q_i(\mathbf{y}) = \max_j q_j(\mathbf{y})\} \subseteq \mathbb{R}^p$ and $U_i^* = \{\mathbf{y}_r \in \mathbb{R}^r : \eta_i(\mathbf{y}_r) = \max_j \eta_j(\mathbf{y}_r)\} \subseteq \mathbb{R}^r$ are such that $\sum_{i=1}^k P\left[\mathbf{y} \in S_i^* \vartriangle (U_i^* \times \mathbb{R}^{p-r})\right] = 0$ so, by theorem 5, $\mathbf{y} = t(\mathbf{x})$ is d.r.t. ■

Condition of theorem 6 is stronger than condition of theorem 5 as it will be illustrated by example in subsection 3.2 below. The next result gives an equivalent formulation for sufficient condition (9).

Theorem 7 *Condition (9) holds if, and only if, the class label* $\mathbf{g}$ *and the random vector* $\mathbf{y}_{(r)} = (y_{r+1}, ..., y_p)'$ *are conditionally independent given* $\mathbf{y}_r = (y_1, ..., y_r)'$*, that is, if for all* $C_2 \in \mathcal{B}^{p-r}$ *and* $i = 1, ..., k$

$$P[\mathbf{y}_{(r)} \in C_2; \mathbf{g} = i \mid \mathbf{y}_r] = P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{y}_r] P[\mathbf{g} = i \mid \mathbf{y}_r], \quad a.e.\ (\mu_{\mathbf{y}_r})\ . \tag{12}$$

Proof. Since $\mathbf{y} = (\mathbf{y}_r^{'},\mathbf{y}_{(r)}^{'})'$, if $q_i(\mathbf{y}) = \eta_i(\mathbf{y}_r)$, *a.e.* $(\mu_{\mathbf{y}})$ one has, for all $C_1 \in \mathcal{B}^r$, $C_2 \in \mathcal{B}^{p-r}$ and $i = 1, ..., k$,

$$\begin{aligned} P\left[\mathbf{y} \in C_1 \times C_2; \mathbf{g} = i\right] &= \int_{C_1 \times C_2} q_i(\mathbf{y}) \mu_{\mathbf{y}}(d\mathbf{y}) = E[q_i(\mathbf{y}) I_{C_1 \times C_2}(\mathbf{y})] \\ &= E[\eta_i(\mathbf{y}_r) I_{C_1 \times C_2}(\mathbf{y})] = E[\eta_i(\mathbf{y}_r) I_{C_1}(\mathbf{y}_r) I_{C_2}(\mathbf{y}_{(r)})] \\ &= E\{\eta_i(\mathbf{y}_r) I_{C_1}(\mathbf{y}_r) E[I_{C_2}(\mathbf{y}_{(r)}) \mid \mathbf{y}_r]\} \end{aligned}$$

$$
\begin{aligned}
&= E\{\eta_i(\mathbf{y}_r) I_{C_1}(\mathbf{y}_r) P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{y}_r]\} \\
&= \int_{C_1} \eta_i(\mathbf{y}_r) P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{y}_r] \mu_{\mathbf{y}_r}(d\mathbf{y}_r) \ . \qquad (13)
\end{aligned}
$$

On the other hand,

$$
P[\mathbf{y} \in C_1 \times C_2; \mathbf{g} = i] = \int_{C_1} P[\mathbf{y}_{(r)} \in C_2; \mathbf{g} = i \mid \mathbf{y}_r] \mu_{\mathbf{y}_r}(d\mathbf{y}_r) \ , \qquad (14)
$$

and (12) follows after comparing (13) and (14). Conversely, under the assumption of conditional independence between $\mathbf{g}$ and $\mathbf{y}_{(r)} = (y_{r+1}, ..., y_p)'$ given $\mathbf{y}_r = (y_1, ..., y_r)'$, one has, using similar arguments as above,

$$
\begin{aligned}
P[\mathbf{y} \in C_1 \times C_2; \mathbf{g} = i] &= \int_{C_1} P[\mathbf{y}_{(r)} \in C_2; \mathbf{g} = i \mid \mathbf{y}_r] \mu_{\mathbf{y}_r}(d\mathbf{y}_r) \\
&= \int_{C_1} \eta_i(\mathbf{y}_r) P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{y}_r] \mu_{\mathbf{y}_r}(d\mathbf{y}_r) \\
&= E\{\eta_i(\mathbf{y}_r) I_{C_1}(\mathbf{y}_r) P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{y}_r]\} \\
&= E\{\eta_i(\mathbf{y}_r) I_{C_1}(\mathbf{y}_r) E[I_{C_2}(\mathbf{y}_{(r)}) \mid \mathbf{y}_r]\} \\
&= E[\eta_i(\mathbf{y}_r) I_{C_1}(\mathbf{y}_r) I_{C_2}(\mathbf{y}_{(r)})] = E[\eta_i(\mathbf{y}_r) I_{C_1 \times C_2}(\mathbf{y})] \\
&= \int_{C_1 \times C_2} \eta_i(\mathbf{y}_r) \mu_{\mathbf{y}}(d\mathbf{y}) \ . \qquad (15)
\end{aligned}
$$

Also, by construction of $q_i(\mathbf{y})$,

$$
P[\mathbf{y} \in C_1 \times C_2; \mathbf{g} = i] = \int_{C_1 \times C_2} q_i(\mathbf{y}) \mu_{\mathbf{y}}(d\mathbf{y}) \ . \qquad (16)
$$

Comparing now (15) and (16), and using the extension theorem for finite measures, one has $P[\mathbf{y} \in C; \mathbf{g} = i] = \int_C \eta_i(\mathbf{y}_r) \mu_{\mathbf{y}}(d\mathbf{y}) = \int_C q_i(\mathbf{y}) \mu_{\mathbf{y}}(d\mathbf{y})$, for all $C \in \mathcal{B}^p$ and $i = 1, ..., k$. Hence $q_i(\mathbf{y}) = \eta_i(\mathbf{y}_r)$, *a.e.* $(\mu_{\mathbf{y}})$, $i = 1, ..., k$. ■

Condition (12) relative to conditional independence between the class label $\mathbf{g}$ and $\mathbf{y}_{(r)} = (y_{r+1}, ..., y_p)'$ once $\mathbf{y}_r = (y_1, ..., y_r)'$ is given, formalizes an intuitive aspect of dimension reduction transformations: if the r components $\mathbf{y}_r = (y_1, ..., y_r)'$ of $\mathbf{y}$ are known, the remaining $p - r$ components $\mathbf{y}_{(r)} = (y_{r+1}, ..., y_p)'$ do not carry relevant information on the particular class membership of the individual under study.

3. CONTINUOUS CASE

This section considers specialization of the previous results to the case in which the class conditional probability distributions $\mathbf{x} \mid \mathbf{g} = i$ are absolutely continuous with respect to Lebesgue measure in $\mathbb{R}^p$, that is, when for $i = 1, ..., k$ there exist density functions $f_i(\mathbf{x})$ such that, for all $C \in \mathcal{B}^p$, $P[\mathbf{x} \in C \mid \mathbf{g} = i] = \int_C f_i(\mathbf{x})d\mathbf{x}$. Let $f(\mathbf{x}) = \sum_{i=1}^k \pi_i f_i(\mathbf{x})$ be the marginal density of $\mathbf{x}$. In the continuous case, a *regular* version of the class posterior probabilities $\pi_i(\mathbf{x})$ is obtained by defining

$$\pi_i(\mathbf{x}) = P[\mathbf{g} = i \mid \mathbf{x}] = \frac{\pi_i f_i(\mathbf{x})}{f(\mathbf{x})} , \tag{17}$$

if $f(\mathbf{x}) > 0$ and, for example, $\pi_i(\mathbf{x}) = \pi_i$ if $f(\mathbf{x}) = 0$. The joint probability distribution of the pair $(\mathbf{x}, \mathbf{g})$ is then given by

$$P[\mathbf{x} \in C; \mathbf{g} = i] = P[\mathbf{g} = i] P[\mathbf{x} \in C \mid \mathbf{g} = i] = \int_C \pi_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \pi_i \int_C f_i(\mathbf{x}) d\mathbf{x} . \tag{18}$$

Using (18), the probability of misclassification of a rule $r(\mathbf{x}) = \sum_{i=1}^k i I_{R_i}(\mathbf{x})$ is

$$L[r(\mathbf{x})] = 1 - \sum_{i=1}^k P[\mathbf{x} \in R_i; \mathbf{g} = i] = 1 - \sum_{i=1}^k \pi_i \int_{R_i} f_i(\mathbf{x}) d\mathbf{x} . \tag{19}$$

Adapting adequately the proof of proposition 1 in section 1, a Bayes rule is determined by a measurable partition $R_1^*, ..., R_k^*$ where, by the usual convention, $R_i^* = \{\mathbf{x} \in \mathbb{R}^p : \pi_i f_i(\mathbf{x}) = \max_j \pi_j f_j(\mathbf{x})\}$, $i = 1, ..., k$. Under condition *(C4)*: $P[\pi_i f_i(\mathbf{x}) = \pi_j f_j(\mathbf{x})] = 0$, $i \neq j$, rule $r^*(\mathbf{x}) = \sum_{i=1}^k i I_{R_i^*}(\mathbf{x})$ is *unique*.

Consider now a measurable and invertible transformation $\mathbf{y} = t(\mathbf{x}) = (t_1(\mathbf{x}), ..., t_p(\mathbf{x}))' = (y_1, ..., y_p)'$ and assume for the rest of this section that the inverse transformation $\mathbf{x} = t^{-1}(\mathbf{y}) = (t_1^{-1}(\mathbf{y}), ..., t_p^{-1}(\mathbf{y}))' = (x_1, ..., x_p)'$ is continuously differentiable. By the well-known change of variable formula (see, e.g. Billingsley, 1995 Chap. 4), for $i = 1, ..., k$ the class conditional distribution $\mathbf{y} \mid \mathbf{g} = i$ has a density

$$f_{\mathbf{y},i}(\mathbf{y}) = f_i[t^{-1}(\mathbf{y})] \mid \det[\partial t^{-1}(\mathbf{y})/\partial \mathbf{y}] \mid , \tag{20}$$

where $\partial t^{-1}(\mathbf{y})/\partial \mathbf{y} = (\partial t_i^{-1}(\mathbf{y})/\partial y_j : i,j = 1,...,p)$ is the $p \times p$ Jacobian matrix of $\mathbf{x} = t^{-1}(\mathbf{y})$. If $f_{\mathbf{y}}(\mathbf{y}) = \sum_{i=1}^{k} \pi_i f_{\mathbf{y},i}(\mathbf{y})$ is the marginal density of the transformed feature vector $\mathbf{y} = t(\mathbf{x})$, the class posterior probabilities in the space $\mathbf{y}$ are

$$q_i(\mathbf{y}) = P[\mathbf{g} = i \mid \mathbf{y}] = \frac{\pi_i f_{\mathbf{y},i}(\mathbf{y})}{f_{\mathbf{y}}(\mathbf{y})} \ , \tag{21}$$

for $f_{\mathbf{y}}(\mathbf{y}) > 0$. Also, if the discriminant problem in the space $\mathbf{y}$ is restricted to the first r variables $\mathbf{y}_r = (y_1,...,y_r)'$, the class posterior probabilities are

$$\eta_i(\mathbf{y}_r) = P[\mathbf{g} = i \mid \mathbf{y}_r] = \frac{\pi_i f_{\mathbf{y},i}(\mathbf{y}_r)}{f_{\mathbf{y}}(\mathbf{y}_r)} \ , \tag{22}$$

for $i = 1, ..., k$, where, in (22), $f_{\mathbf{y},i}(\mathbf{y}_r)$ is the marginal of $\mathbf{y}_r = (y_1,...,y_r)'$ relative to the density $f_{\mathbf{y},i}(\mathbf{y})$ of (20), and $f_{\mathbf{y}}(\mathbf{y}_r) = \sum_{i=1}^{k} \pi_i f_{\mathbf{y},i}(\mathbf{y}_r) > 0$ is the corresponding marginal of $\mathbf{y}_r = (y_1,...,y_r)'$ relative to $f_{\mathbf{y}}(\mathbf{y}) = \sum_{i=1}^{k} \pi_i f_{\mathbf{y},i}(\mathbf{y})$.

3.1 A sufficient condition in terms of conditional densities

Theorems 6 and 7 in section 2 established that $\mathbf{y} = t(\mathbf{x})$ is a d.r.t. if identity $q_i(\mathbf{y}) = \eta_i(\mathbf{y}_r)$ holds *a.e.* $(\mu_{\mathbf{y}})$ for $i = 1, ..., k$ or, equivalently, if $\mathbf{g}$ and $\mathbf{y}_{(r)} = (y_{r+1},...,y_p)'$ are conditionally independent given $\mathbf{y}_r = (y_1, ..., y_r)'$. In the continuous case, this sufficient condition can be formulated in terms of the conditional densities

$$f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) = \frac{f_{\mathbf{y},i}(\mathbf{y})}{f_{\mathbf{y},i}(\mathbf{y}_r)} \ , \tag{23}$$

and

$$f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) = \frac{f_{\mathbf{y}}(\mathbf{y})}{f_{\mathbf{y}}(\mathbf{y}_r)} \ , \tag{24}$$

that are well defined for $\mathbf{y}_r = (y_1, ..., y_r)' \in A_i = \{\mathbf{y}_r \in \mathbb{R}^r : f_{\mathbf{y},i}(\mathbf{y}_r) > 0\} \subseteq A = \{\mathbf{y}_r \in \mathbb{R}^r : f_{\mathbf{y}}(\mathbf{y}_r) > 0\}$.

Theorem 8 *In the continuous case, $q_i(\mathbf{y}) = \eta_i(\mathbf{y}_r)$ a.e. $(\mu_{\mathbf{y}})$ for all $i = 1, ..., k$ if, and only if, there exists a subset $B \in \mathcal{B}^r$, with $B \subseteq A$ and $P[\mathbf{y}_r \in B] = 1$, such that,*

for all $i = 1, ..., k$ *and* $\mathbf{y}_r \in A_i \cap B$, *the condition below holds*

$$f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) = f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r), \quad a.e.\ (m_{p-r}) \ , \tag{25}$$

where m_{p-r} *is Lebesgue measure on the* σ*-field* $\mathcal{B}^{p-r}$ *of Borel sets in* $\mathbb{R}^{p-r}$.

Proof. For each $C_2 \in \mathcal{B}^{p-r}$ and $i = 1, ..., k$, a version of the conditional probability $P[\mathbf{y}_{(r)} \in C_2; \mathbf{g} = i \mid \mathbf{y}_r]$ is given by the product

$$P[\mathbf{y}_{(r)} \in C_2; \mathbf{g} = i \mid \mathbf{y}_r] = \eta_i(\mathbf{y}_r) P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{g} = i; \mathbf{y}_r] \ , \tag{26}$$

where $\eta_i(\mathbf{y}_r) = \pi_i f_{\mathbf{y},i}(\mathbf{y}_r)/f_{\mathbf{y}}(\mathbf{y}_r)$ is as in (22) and $P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{g} = i; \mathbf{y}_r]$ is the function defined for $\mathbf{y}_r \in A_i$ as

$$P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{g} = i; \mathbf{y}_r] = \int_{C_2} f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} \ , \tag{27}$$

where, in (27), $d\mathbf{y}_{(r)}$ represents integration with respect to the measure m_{p-r}. To verify this statement, recall that $\mathbf{y} = (\mathbf{y}_r^{'}, \mathbf{y}_{(r)}^{'})'$ and notice that, with definitions (22), (23) and (27), by Fubini's theorem one has, for all $C_1 \in \mathcal{B}^r$,

$$\int_{C_1} \eta_i(\mathbf{y}_r) P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{g} = i; \mathbf{y}_r] f_{\mathbf{y}}(\mathbf{y}_r) d\mathbf{y}_r =$$

$$\begin{aligned}
&= \int_{C_1 \cap A_i} \frac{\pi_i f_{\mathbf{y},i}(\mathbf{y}_r)}{f_{\mathbf{y}}(\mathbf{y}_r)} \left[\int_{C_2} f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} \right] f_{\mathbf{y}}(\mathbf{y}_r) d\mathbf{y}_r \\
&= \pi_i \int_{(C_1 \cap A_i) \times C_2} f_{\mathbf{y},i}(\mathbf{y}) d\mathbf{y} = P[\mathbf{g} = i] P[\mathbf{y}_r \in C_1 \cap A_i; \mathbf{y}_{(r)} \in C_2 \mid \mathbf{g} = i] \\
&= P[\mathbf{y}_r \in C_1 \cap A_i; \mathbf{y}_{(r)} \in C_2; \mathbf{g} = i] = P[\mathbf{y}_r \in C_1; \mathbf{y}_{(r)} \in C_2; \mathbf{g} = i] \, ,
\end{aligned}$$

where the last identity above follows from definition of A_i and inequality $P[\mathbf{y}_r \in C_1 \cap A_i^c; \mathbf{y}_{(r)} \in C_2; \mathbf{g} = i] \leq P[\mathbf{y}_r \in A_i^c; \mathbf{g} = i] = P[\mathbf{g} = i] P[\mathbf{y}_r \in A_i^c \mid \mathbf{g} = i] = 0$.

Suppose now that $q_i(\mathbf{y}) = \eta_i(\mathbf{y}_r)$ $a.e.(\mu_{\mathbf{y}})$ for all $i = 1, ..., k$. According to theorem 7, this is equivalent to conditional independence between $\mathbf{g}$ and $\mathbf{y}_{(r)} = (y_{r+1}, ..., y_p)'$

once the information in $\mathbf{y}_r = (y_1, ..., y_r)'$ is given. In other words, for each $C_2 \in \mathcal{B}^{p-r}$ and $i = 1, ..., k$,

$$P[\mathbf{y}_{(r)} \in C_2; \mathbf{g} = i \mid \mathbf{y}_r] = \eta_i(\mathbf{y}_r) P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{y}_r], \quad a.e.\ (\mu_{\mathbf{y}_r}) \ , \tag{28}$$

where, in the continuous case,

$$P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{y}_r] = \int_{C_2} f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} \ . \tag{29}$$

Comparing (26)-(27) with (28)-(29) it turns out that, for each $C_2 \in \mathcal{B}^{p-r}$ and $i = 1$, ..., k, there exists a Borel set $B(C_2,i) \subseteq \mathbb{R}^r$ that depends on C_2 and i, such that $P[\mathbf{y}_r \in B(C_2,i)] = 0$ and, if $\mathbf{y}_r \in [B(C_2,i)]^c$,

$$\eta_i(\mathbf{y}_r) \int_{C_2} f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} = \eta_i(\mathbf{y}_r) \int_{C_2} f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} \ . \tag{30}$$

Define now, for real numbers x_{r+1}, ..., x_p, the infinite rectangle $C_2(x_{r+1}, ..., x_p) = (-\infty, x_{r+1}] \times ... \times (-\infty, x_p]$ and put $D = \underset{s_{r+1}, ..., s_p \in \mathbb{Q}\,; i \leq 1 \leq k}{\cup} B[C_2(s_{r+1}, ..., s_p), i] \cup A^c$, where $\mathbb{Q}$ is the set of rational numbers. Since the union that defines D is countable,

$$P[\mathbf{y}_r \in D] \leq \sum_{s_{r+1}, ..., s_p \in \mathbb{Q}; 1 \leq i \leq k} P\{\mathbf{y}_r \in B[C_2(s_{r+1}, ..., s_p), i]\} + P[\mathbf{y}_r \in A^c] = 0 \ ,$$

so taking $B = D^c \subseteq A$, one has $P[\mathbf{y}_r \in B] = 1$. Also, if $i = 1$, ..., k and $\mathbf{y}_r \in A_i \cap B$, cancelling $\eta_i(\mathbf{y}_r) = \pi_i f_{\mathbf{y},i}(\mathbf{y}_r)/f_{\mathbf{y}}(\mathbf{y}_r) > 0$ in both sides of (30) one has, for all rational numbers s_{r+1}, ..., $s_p \in \mathbb{Q}$,

$$\int_{C_2(s_{r+1}, ..., s_p)} f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} = \int_{C_2(s_{r+1}, ..., s_p)} f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} \, . \tag{31}$$

Therefore, for each x_{r+1}, ..., $x_p \in \mathbb{R}$, taking sequences $s_j \to x_j$ and passing to the limit in (31),

$$\int_{C_2(x_{r+1}, ..., x_p)} f_{\mathbf{y},i}(\mathbf{y}_{(r)} \,|\mathbf{y}_r) d\mathbf{y}_{(r)} = \lim_{\substack{s_j \to x_j \\ r+1 \leq j \leq p}} \int_{C_2(s_{r+1}, ..., s_p)} f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)}$$

$$= \lim_{\substack{s_j \to x_j \\ r+1 \le j \le p}} \int_{C_2(s_{r+1},...,s_p)} f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} = \int_{C_2(x_{r+1},...,x_p)} f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} \,. \qquad (32)$$

Identity (32) shows that, when $\mathbf{y}_r \in A_i \cap B$, the two probability distributions $\int_{C_2} f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)}$ and $\int_{C_2} f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)}$ are identical for all $C_2 \in \mathcal{B}^{p-r}$and, therefore, $f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) = f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r)$, *a.e.* (m_{p-r}). Conversely, if (25) holds, one has

$$\eta_i(\mathbf{y}_r) \int_{C_2} f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} = \eta_i(\mathbf{y}_r) \int_{C_2} f_{\mathbf{y}}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) d\mathbf{y}_{(r)} \qquad (33)$$

for all $C_2 \in \mathcal{B}^{p-r}$, $i = 1, ..., k$ and $\mathbf{y}_r \in B = (B \cap A_i) \cup (B \cap A_i^c) \subseteq A$. Therefore, by (26)-(27) and (28)-(29), for each $C_2 \in \mathcal{B}^{p-r}$ and $i = 1, ..., k$

$$P[\mathbf{y}_{(r)} \in C_2; \mathbf{g} = i \mid \mathbf{y}_r] = \eta_i(\mathbf{y}_r) P[\mathbf{y}_{(r)} \in C_2 \mid \mathbf{y}_r], \quad a.e.\ (\mu_{\mathbf{y}_r}) \,.$$

By theorem 7 above, this is equivalent to $q_i(\mathbf{y}) = \eta_i(\mathbf{y}_r)$ *a.e.* $(\mu_{\mathbf{y}})$ for $i = 1, ..., k$. ■

3.2 Example

Suppose $\pi_i = 1/k$ for $i = 1, ..., k$, and assume also conditional class densities $f_i(\mathbf{x})$ elliptically symmetric with density

$$f_i(\mathbf{x}) = |\mathbf{\Sigma}|^{-1/2} g[(\mathbf{x} - \boldsymbol{\mu}_i)' \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)] \,, \qquad (34)$$

where $\boldsymbol{\mu}_i$ is a $p \times 1$ vector, $\mathbf{\Sigma}$ is a $p \times p$ positive definite (p.d.) matrix, and $g : [0,\infty) \to [0,\infty)$ is an strictly decreasing and continuous function such that $\int_0^{+\infty} t^{p/2} g(t) dt < +\infty$. Under (34), $E(\mathbf{x} \mid \mathbf{g} = i) = \boldsymbol{\mu}_i$ and, therefore, the $p \times p$ *between* dispersion matrix is

$$\mathbf{B} = Var[E(\mathbf{x} \mid \mathbf{g})] = \frac{1}{k} \sum_{i=1}^{k} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' \,, \qquad (35)$$

where $\boldsymbol{\mu} = E(\mathbf{x}) = E[E(\mathbf{x} \mid \mathbf{g})] = \sum_{i=1}^{k} \boldsymbol{\mu}_i / k$ is the marginal mean of $\mathbf{x}$. Notice that, since $\mathbf{\Sigma}$ is p.d., the square root $\mathbf{\Sigma}^{-1/2}$ of $\mathbf{\Sigma}^{-1}$ is well defined. Let $r = rank(\mathbf{B})$ and consider the spectral decomposition

$$\mathbf{\Sigma}^{-1/2} \mathbf{B} \mathbf{\Sigma}^{-1/2} = \mathbf{C} \mathbf{D} \mathbf{C}' \,, \qquad (36)$$

where $\mathbf{C} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p)$ is a $p \times p$ orthogonal matrix of eigenvectors and $\mathbf{D} = diag(\lambda_1, ..., \lambda_r, 0, ..., 0)$ is a $p \times p$ diagonal matrix of eigenvalues such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$. If the linear transformation

$$\mathbf{y} = \mathbf{C}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \tag{37}$$

is considered, the class conditional distribution $\mathbf{y} \mid \mathbf{g} = i$ has, for $i = 1, ..., k$, a density

$$f_{\mathbf{y},i}(\mathbf{y}) = g(\| \mathbf{y} - \mathbf{M}_i \|^2) \ , \tag{38}$$

where $\|.\|$ is the usual euclidean norm and the $\mathbf{M}_i = \mathbf{C}'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}) = E(\mathbf{y} \,|\mathbf{g} = i)$ are $p \times 1$ vectors such that

$$\frac{1}{k}\sum_{i=1}^{k} \mathbf{M}_i = E[E(\mathbf{y} \,|\mathbf{g})] = E(\mathbf{y}) = \mathbf{0} \ . \tag{39}$$

Using (36), (37) and (39), the between dispersion matrix $\mathbf{B}_\mathbf{y} = Var[E(\mathbf{y} \,|\mathbf{g})]$ is

$$\begin{aligned} \mathbf{B}_\mathbf{y} &= Var[E(\mathbf{y} \,|\mathbf{g})] = \frac{1}{k}\sum_{i=1}^{k} \mathbf{M}_i\mathbf{M}_i' \\ &= Var[\mathbf{C}'\boldsymbol{\Sigma}^{-1/2}E(\mathbf{x} \,|\mathbf{g})] = \mathbf{C}'\boldsymbol{\Sigma}^{-1/2}\mathbf{B}\boldsymbol{\Sigma}^{-1/2}\mathbf{C} = \mathbf{D} \ , \end{aligned} \tag{40}$$

so, since $\mathbf{D} = diag(\lambda_1, ..., \lambda_r, 0, ..., 0)$, from (40) all vectors $\mathbf{M}_i$ are of the form $\mathbf{M}_i = (\mathbf{m}_i', 0, \overset{p-r)}{...}, 0)'$ where $\mathbf{m}_i$ is of $r \times 1$. Writing $\mathbf{y} = (\mathbf{y}_r', \mathbf{y}_{(r)}')'$, the identity

$$\| \mathbf{y} - \mathbf{M}_i \|^2 = \| \mathbf{y}_r - \mathbf{m}_i \|^2 + \| \mathbf{y}_{(r)} \|^2 \ , \tag{41}$$

holds for all $i = 1, ..., k$.

Given the Bayes rule $s^*(\mathbf{y}) = \sum_{i=1}^{k} i I_{S_i^*}(\mathbf{y})$ in the $\mathbf{y} = \mathbf{C}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ space, the subset S_i^* is formed by all transformed feature vectors such that $f_{\mathbf{y},i}(\mathbf{y}) = \max_{1 \leq j \leq k} f_{\mathbf{y},j}(\mathbf{y})$. Since the function $g(.)$ is strictly decreasing, from (38) and (41) maximizing $f_{\mathbf{y},i}(\mathbf{y}) = g(\| \mathbf{y} - \mathbf{M}_i \|^2)$ in i is equivalent to minimizing expression $\| \mathbf{y}_r - \mathbf{m}_i \|^2$ across groups, operation that clearly does not depend on the coordinates $\mathbf{y}_{(r)} = (y_{r+1}, ..., y_p)'$. By the sufficient condition of theorem 5, the linear transformation of (37) is a d.r.t.

If $g(t) = (2\pi)^{-p/2}\exp(-t/2)$, that is, when the conditional class densities $f_i(\mathbf{x})$ are multivariate normal $N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$, the sufficient condition of theorem 8 also holds. To see this notice that, since $\mathbf{y} \mid \mathbf{g} = i \sim N_p(\mathbf{M}_i,\mathbf{I}_p)$ and $\mathbf{M}_i = (\mathbf{m}_i', \mathbf{0}')'$, $f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r) \sim N_{p-r}(\mathbf{0};\mathbf{I}_{p-r})$ for $i = 1, ..., k$, and all the conditional densities $f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r)$ are then identical. However, condition of theorem 8 does not hold in general for an arbitrary function $g(.)$. As it can be seen for example in Johnson (1987, p. 109), if $f_{\mathbf{y},i}(\mathbf{y})$ is as in (38), the conditional density $f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r)$ corresponds to an elliptically symmetric distribution with mean $\mathbf{0}$ and dispersion matrix of the form $w(\| \mathbf{y}_r - \mathbf{m}_i \|^2)I_{p-r}$, where $w(.)$ is some nonnegative real function. As remarked by Muirhead (1982, p. 36), by results in Kelker (1970) $w(\| \mathbf{y}_r - \mathbf{m}_i \|^2)$ is constant if, and only if, $\mathbf{y} \mid \mathbf{g} = i$ is $N_p(\mathbf{M}_i,\mathbf{I}_p)$. As a consequence, unless $\mathbf{y} \mid \mathbf{g} = i$ is normal, densities $f_{\mathbf{y},i}(\mathbf{y}_{(r)} \mid \mathbf{y}_r)$ cannot be identical because the conditional covariance matrix $w(\| \mathbf{y}_r - \mathbf{m}_i \|^2)I_{p-r}$ depends on the group index i through vector $\mathbf{m}_i$.

4. AN EFFECTIVE DIMENSION REDUCTION ALGORITHM

Suppose that, after application of some of the conditions presented, it has been determined that transformation $\mathbf{y} = t(\mathbf{x}) = (t_1(\mathbf{x}), ..., t_p(\mathbf{x}))' = (y_1, ..., y_p)'$ is a d.r.t. from the original value p to the number $r < p$ of coordinates in $\mathbf{y}_r = (y_1, ..., y_r)'$. Typically, this transformation will depend on some of the unknown elements that determine the joint probability distribution of the pair $(\mathbf{x}, \mathbf{g})$. On the other hand, the posterior class probabilities $\eta_i(\mathbf{y}_r) = P[\mathbf{g} = i \mid \mathbf{y}_r]$ are unknown as well, so the subsets U_i^* of the optimal rule $d^*(\mathbf{y}_r) = \sum_{i=1}^k iI_{U_i^*}(\mathbf{y}_r)$ of definition 4 in section 2 are not feasible. This type of problems motivate the need of considering data based effective dimension reduction procedures.

Let

$$\mathbf{D}_n = \{(\mathbf{x}_j, \mathbf{g}_j) : j = 1, ..., n\} \tag{42}$$

be a set of i.i.d. observations from $(\mathbf{x}, \mathbf{g})$ that can be interpreted as a database of individuals previously classified. Consider, for $i = 1, ..., k$, an estimator $\widehat{\eta}_i(\mathbf{y}_r)$ of $\eta_i(\mathbf{y}_r)$ computed from $\mathbf{D}_n$ and put $\widehat{\mathbf{y}}_r = (\widehat{y}_1, ..., \widehat{y}_r)' = (\widehat{t}_1(\mathbf{x}), ..., \widehat{t}_r(\mathbf{x}))'$ where, for $j = 1, ..., r$, $\widehat{y}_j = \widehat{t}_j(\mathbf{x})$ is an estimator of $t_j(\mathbf{x})$. In applications, it is natural to replace $d^*(\mathbf{y}_r) = \sum_{i=1}^k i I_{U_i^*}(\mathbf{y}_r)$ by the sample rule $\widehat{d}_n^*(\mathbf{x}) = \sum_{i=1}^k i I_{\widehat{U}_i^*}(\mathbf{x})$ where, for $i = 1, ..., k$, the subsets

$$\widehat{U}_i^* = \{\mathbf{x} \in \mathbb{R}^p : \widehat{\eta}_i(\widehat{\mathbf{y}}_r) = \max_j \eta_j(\widehat{\mathbf{y}}_r)\} \tag{43}$$

are plug-in versions of the subsets U_i^*. Optimality of $d^*(\mathbf{y}_r)$ in the $\mathbf{y}_r = (y_1, .., y_r)'$ space can be replaced by *consistency* of rule $\widehat{d}_n^*(\mathbf{x})$, that is, by convergence of the conditional probability of error

$$L_n = L[\widehat{d}_n^*(\mathbf{x})] = P[\widehat{d}_n^*(\mathbf{x}) \neq \mathbf{g} \,|D_n] = 1 - \sum_{i=1}^k \int_{\widehat{U}_i^*} \pi_i(\mathbf{x}) \mu(d\mathbf{x}) \ , \tag{44}$$

where the pair $(\mathbf{x}, \mathbf{g})$ is independent of the database $\mathbf{D}_n$, to the Bayes error L^*, either weakly or in probability or strongly or with probability one (see e.g., Devroye, Györfi and Lugosi, 1996 chap. 6). These ideas can be summarized in a three step *effective dimension reduction algorithm*: *i)* determine the theoretical expression of the d.r.t. $\mathbf{y} = t(\mathbf{x})$; *ii)* choose estimators $\widehat{\eta}_i(\mathbf{y}_r)$ and sample coordinates $\widehat{y}_j = \widehat{t}_j(\mathbf{x})$; and *iii)* form rule $\widehat{d}_n^*(\mathbf{x}) = \sum_{i=1}^k i I_{\widehat{U}_i^*}(\mathbf{x})$ and study its consistency properties. As an illustration, this algorithm is applied to perform data based dimension reduction in a classification problem with heteroscedastic normal class conditional densities.

4.1 Heteroscedastic normal models

Suppose that, for $i = 1, ..., k$, the conditional class densities $f_i(\mathbf{x})$ are $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ where the $\boldsymbol{\mu}_i$ are $p \times 1$ vectors and the $\boldsymbol{\Sigma}_i$ $p \times p$ p.d. matrices. Given the class prior probabilities $\pi_i > 0$, the marginal mean of $\mathbf{x}$ is $\boldsymbol{\mu} = E(\mathbf{x}) = E[E(\mathbf{x} \,|\mathbf{g})] = \sum_{i=1}^k \pi_i \boldsymbol{\mu}_i$

and the marginal dispersion matrix is

$$Var(\mathbf{x}) = Var[E(\mathbf{x}\,|\mathbf{g})] + E[Var(\mathbf{x}\,|\mathbf{g})] \; = \mathbf{B} + \mathbf{\Sigma} \; , \tag{45}$$

where the between dispersion matrix is of the form

$$\mathbf{B} = Var[E(\mathbf{x}\,|\mathbf{g})] = \sum_{i=1}^{k} \pi_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' \; , \tag{46}$$

and the within dispersion matrix is

$$\mathbf{\Sigma} = E[Var(\mathbf{x}\,|\mathbf{g})] = \sum_{i=1}^{k} \pi_i \mathbf{\Sigma}_i \; . \tag{47}$$

Taking into account that if all $\mathbf{\Sigma}_i$ are p.d. $\mathbf{\Sigma}$ is also p.d., Schott (1993), based on previous ideas of Decell, Odell and Coberly (1981), proposes taking

$$\mathcal{D} = \sum_{i=1}^{k} \pi_i \mathbf{\Sigma}^{-1/2} (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1/2} + \sum_{i=1}^{k} \pi_i [\mathbf{\Sigma}^{-1/2} (\mathbf{\Sigma}_i - \mathbf{\Sigma}) \mathbf{\Sigma}^{-1/2}]^2 \; , \tag{48}$$

as the dimension matrix of the discriminant problem. Notice that the matrix above reflects differences in both the conditional means and dispersion matrices of the standardized feature vector $\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. If $s = rank(\mathcal{D})$, the spectral representation of the matrix of (48) is

$$\mathcal{D} = \mathbf{U} \mathbf{\Delta} \mathbf{U}' \; , \tag{49}$$

where $\mathbf{U} = (\mathbf{u}_1, ..., \mathbf{u}_p)$ is an orthogonal $p \times p$ matrix of normalized eigenvectors and $\mathbf{\Delta} = diag(\delta_1, ..., \delta_s, 0, ..., 0)$ is a diagonal matrix of eigenvalues $\delta_1 \geq ... \geq \delta_s > 0$.

4.2 Application of the algorithm

The effective dimension reduction algorithm is now applied in an stepwise fashion.

- Step *i)* Consider the linear transformation

$$\mathbf{y} = \mathbf{U}' \mathbf{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu}) \; . \tag{50}$$

To verify that (50) is a d.r.t., write $\mathbf{U} = (\mathbf{U}_1 \,|\mathbf{U}_2)$, where $\mathbf{U}_1 = (\mathbf{u}_1, ..., \mathbf{u}_s)$ is of $p \times s$ and $\mathbf{U}_2 = (\mathbf{u}_{s+1}, ..., \mathbf{u}_p)$ of $p \times (p-s)$. From (48) and (49), one has

$$\begin{aligned} \mathbf{U}'\mathcal{D}\mathbf{U} &= \sum_{i=1}^{k} \pi_i \mathbf{U}'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1/2}\mathbf{U} \\ &\quad + \sum_{i=1}^{k} \pi_i [\mathbf{U}'(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_p)\mathbf{U}][\mathbf{U}'(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_p)\mathbf{U}]' \\ &= \sum_{i=1}^{k} \pi_i \mathbf{a}_i \mathbf{a}_i' + \sum_{i=1}^{k} \pi_i \begin{pmatrix} \mathbf{E}_i\mathbf{E}_i' & \mathbf{E}_i\mathbf{F}_i' \\ \mathbf{F}_i\mathbf{E}_i' & \mathbf{F}_i\mathbf{F}_i' \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Delta}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} , \end{aligned} \tag{51}$$

where $\boldsymbol{\Delta}_s = diag(\delta_1, ..., \delta_s)$ and, for $i = 1, ..., k$, the $\mathbf{a}_i = \mathbf{U}'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})$ are $p \times 1$ vectors, the $\mathbf{E}_i = \mathbf{U}_1'(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_p)\mathbf{U}$ $s \times p$ matrices and the $\mathbf{F}_i = \mathbf{U}_2'(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1/2} - \mathbf{I}_p)\mathbf{U}$ $(p-s) \times p$ matrices. Using (51), it can be seen that, for $i = 1, ..., k$, $\mathbf{a}_i = (\mathbf{m}_i', \mathbf{0}')'$, where $\mathbf{m}_i = \mathbf{U}_1'\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu})$ is of $s \times 1$, and $\mathbf{F}_i = \mathbf{0}$. As a conclusion, under (50), the conditional class distributions are

$$\mathbf{y} \mid \mathbf{g} = i \sim N_p[\begin{pmatrix} \mathbf{m}_i \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{Q}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-s} \end{pmatrix}] , \tag{52}$$

where $\mathbf{Q}_i = \mathbf{U}_1'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1/2}\mathbf{U}_1$ is an $s \times s$ p.d. matrix. From (52) the conditional densities $f_{\mathbf{y},i}(\mathbf{y}_{(s)} \mid \mathbf{y}_s)$ are $N_{p-s}(\mathbf{0}, \mathbf{I}_{p-s})$ and, hence, they are all identical. By the sufficient condition of theorem 8, the linear transformation (50) is a d.r.t. from p to $s = rank(\mathcal{D})$ coordinates. For further use, it is useful to retain the identity

$$Var(\mathbf{y} \,|\mathbf{g} = i) = \mathbf{U}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}^{-1/2}\mathbf{U} = \begin{pmatrix} \mathbf{Q}_i & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-s} \end{pmatrix} ; \tag{53}$$

• Step *ii)* For the estimation phase of the algorithm, write the database of (42) in the more standard notation $\mathbf{D}_n = \{\mathbf{x}_{ij} : i = 1, ..., k, \; j = 1, ..., n_i\}$ where, for $i = 1, ..., k$, n_i is the number of cases in $\mathbf{D}_n$ that belong to class g_i. Consider also the

sample version of the dimension matrix of (48), namely

$$\widehat{\mathcal{D}} = \sum_{i=1}^{k} \widehat{\pi}_i \widehat{\mathbf{\Sigma}}^{-1/2} \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}\right) \left(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}\right)' \widehat{\mathbf{\Sigma}}^{-1/2} + \sum_{i=1}^{k} \widehat{\pi}_i [\widehat{\mathbf{\Sigma}}^{-1/2} (\widehat{\mathbf{\Sigma}}_i - \widehat{\mathbf{\Sigma}}) \widehat{\mathbf{\Sigma}}^{-1/2}]^2 \ , \qquad (54)$$

where $\widehat{\pi}_i = n_i/n$, $\overline{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$, $\overline{\mathbf{x}} = \sum_{i=1}^{k} (n_i/n) \overline{\mathbf{x}}_i$, $\widehat{\mathbf{\Sigma}}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)'/n_i$ and $\widehat{\mathbf{\Sigma}} = \sum_{i=1}^{k} (n_i/n) \widehat{\mathbf{\Sigma}}_i$. Notice that $\widehat{\mathcal{D}}$ is constructed replacing the unknown elements in $\mathcal{D}$ by their natural estimators computed from the database $\mathbf{D}_n$. Once an specific value for $s = rank(\mathcal{D})$ has been accepted, compute the spectral representation

$$\widehat{\mathcal{D}} = \widehat{\mathbf{U}} \widehat{\mathbf{\Delta}} \widehat{\mathbf{U}}' \ , \qquad (55)$$

where $\widehat{\mathbf{U}} = (\widehat{\mathbf{U}}_1 \mid \widehat{\mathbf{U}}_2)$ is a $p \times p$ matrix of eigenvectors, being $\widehat{\mathbf{U}}_1$ of $p \times s$ and $\widehat{\mathbf{U}}_2$ of $p \times (p-s)$, and $\widehat{\mathbf{\Delta}} = diag(\widehat{\delta}_1, ..., \widehat{\delta}_p)$ is a $p \times p$ diagonal matrix of nonnegative eigenvalues. From (52), the marginal density $f_{\mathbf{y},i}(\mathbf{y}_s)$ is $N_s(\mathbf{m}_i, \mathbf{Q}_i)$, so a natural estimator of the posterior class probability $\eta_i(\mathbf{y}_s) = P[\mathbf{g} = i \mid \mathbf{y}_s] = \pi_i f_{\mathbf{y},i}(\mathbf{y}_s) / \sum_{j=1}^{k} \pi_j f_{\mathbf{y},j}(\mathbf{y}_s)$ is

$$\widehat{\eta}_i(\mathbf{y}_s) = \frac{\widehat{\pi}_i \widehat{f}_{\mathbf{y},i}(\mathbf{y}_s)}{\sum_{j=1}^{k} \widehat{\pi}_j \widehat{f}_{\mathbf{y},j}(\mathbf{y}_s)} \ , \qquad (56)$$

where $\widehat{f}_{\mathbf{y},i}(\mathbf{y}_s)$ is the estimator of $f_{\mathbf{y},i}(\mathbf{y}_s)$ given by a $N_s(\widehat{\mathbf{m}}_i, \widehat{\mathbf{Q}}_i)$, where $\widehat{\mathbf{m}}_i = \widehat{\mathbf{U}}_1' \widehat{\mathbf{\Sigma}}^{-1/2}$ $(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})$ and $\widehat{\mathbf{Q}}_i = \widehat{\mathbf{U}}_1' \widehat{\mathbf{\Sigma}}^{-1/2} \widehat{\mathbf{\Sigma}}_i \widehat{\mathbf{\Sigma}}^{-1/2} \widehat{\mathbf{U}}_1$. Estimator (56) is complemented with the sample coordinates

$$\widehat{\mathbf{y}}_s = \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_s \end{pmatrix} = \widehat{\mathbf{U}}_1' \widehat{\mathbf{\Sigma}}^{-1/2} (\mathbf{x} - \overline{\mathbf{x}}) \ . \qquad (57)$$

• Step *iii)* Given the choices (56) and (57), it is straightforward to verify that the subset $\widehat{U}_i^*$ of (43) is formed by those points $\mathbf{x} \in \mathbb{R}^p$ such that

$$\begin{aligned} & -2 \log \widehat{\pi}_i + \log \mid \widehat{\mathbf{Q}}_i \mid + (\mathbf{x} - \overline{\mathbf{x}}_i)' \widehat{\mathbf{\Sigma}}^{-1/2} \widehat{\mathbf{U}}_1 \widehat{\mathbf{Q}}_i^{-1} \widehat{\mathbf{U}}_1' \widehat{\mathbf{\Sigma}}^{-1/2} (\mathbf{x} - \overline{\mathbf{x}}_i) \\ = \ & \min_j [-2 \log \widehat{\pi}_j + \log \mid \widehat{\mathbf{Q}}_j \mid + (\mathbf{x} - \overline{\mathbf{x}}_j)' \widehat{\mathbf{\Sigma}}^{-1/2} \widehat{\mathbf{U}}_1 \widehat{\mathbf{Q}}_j^{-1} \widehat{\mathbf{U}}_1' \widehat{\mathbf{\Sigma}}^{-1/2} (\mathbf{x} - \overline{\mathbf{x}}_j)] \ , \qquad (58) \end{aligned}$$

or, equivalently, such that

$$\widehat{\pi}_i \widehat{g}_i(\mathbf{x}) = \max_j \widehat{\pi}_j \widehat{g}_j(\mathbf{x}) \ , \tag{59}$$

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

$$\widehat{g}_i(\mathbf{x}) = (2\pi)^{-p/2} \left(\mid \widehat{\mathbf{\Sigma}} \mid\mid \widehat{\mathbf{Q}}_i \mid\right)^{-1/2} \exp[-\frac{1}{2}\widehat{W}_i(\mathbf{x})] \ , \tag{60}$$

and

$$\widehat{W}_i(\mathbf{x}) = (\mathbf{x}-\overline{\mathbf{x}}_i)'\widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{U}}_1\widehat{\mathbf{Q}}_i^{-1}\widehat{\mathbf{U}}_1'\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}-\overline{\mathbf{x}}_i) + (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1/2}\mathbf{U}_2\mathbf{U}_2'\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i) \ . \tag{61}$$

Recall that, using (51), $\mathbf{U}_2'\mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}) = 0$ for $i = 1, ..., k$, so the second summand in the quadratic form of (61) does not depend on the group index i. Equivalence between (58) and (59) is finally justified by monotonicity of the function $-2\log(.)$.

To establish strong consistency of the sample rule defined by criterion (59), notice first that the optimal rule in this context is defined by criterion

$$\pi_i f_i(\mathbf{x}) = \max_j \pi_j f_j(\mathbf{x}) \ , \tag{62}$$

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

$$f_i(\mathbf{x}) = (2\pi)^{-p/2} \left|\mathbf{\Sigma}_i\right|^{-1/2} \exp[-\frac{1}{2}W_i(\mathbf{x})] \ , \tag{63}$$

is the *ith* class conditional density of a $N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$, where using identity (53), the quadratic form of the exponent can be written in the form

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

$$= (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1/2}\mathbf{U}_1\mathbf{Q}_i^{-1}\mathbf{U}_1'\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i) + (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}^{-1/2}\mathbf{U}_2\mathbf{U}_2'\mathbf{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_i) \ . \tag{64}$$

By theorem 1 in Devroye and Györfi (1985, p. 254), the relationship between the conditional probability of error L_n of the *pseudo* plug-in rule (59) and the Bayes error L^* of rule (62) is such that

$$0 \leq L_n - L^* \leq \sum_{i=1}^{k} \int_{\mathbb{R}^p} |\widehat{\pi}_i \widehat{g}_i(\mathbf{x}) - \pi_i f_i(\mathbf{x})| \, d\mathbf{x} \; . \tag{65}$$

Therefore, to establish $L_n \rightarrow L^*$ *a.e.* it is enough to verify that, for $i = 1, ..., k$, $\int_{\mathbb{R}^p} |\widehat{\pi}_i \widehat{g}_i(\mathbf{x}) - \pi_i f_i(\mathbf{x})| \, d\mathbf{x} \rightarrow 0$**,** *a.e.* To do this, fix an integer i and put $\widehat{a}_i = \int_{\mathbb{R}^p} \widehat{g}_i(\mathbf{x}) d\mathbf{x}$. Considering the integrals of the positive and negative parts of the difference $f_i(\mathbf{x}) - \widehat{g}_i(\mathbf{x})$, the inequality below follows:

$$\begin{aligned} \int_{\mathbb{R}^p} |\widehat{\pi}_i \widehat{g}_i(\mathbf{x}) - \pi_i f_i(\mathbf{x})| \, d\mathbf{x} &\leq |\widehat{\pi}_i - \pi_i| \, \widehat{a}_i + \pi_i \int_{\mathbb{R}^p} |\widehat{g}_i(\mathbf{x}) - f_i(\mathbf{x})| \, d\mathbf{x} \\ &\leq |\widehat{\pi}_i - \pi_i| \, \widehat{a}_i + \pi_i(\widehat{a}_i - 1) + 2\pi_i \int_{\mathbb{R}^p} [f_i(\mathbf{x}) - \widehat{g}_i(\mathbf{x})]_+ d\mathbf{x} \; , \end{aligned} \tag{66}$$

so it suffices to check that all the summands of the upper bound of (66) tend to zero *a.e.* as $n \rightarrow \infty$. By the auxiliary results of appendix 6.2, as $n \rightarrow \infty$, $\widehat{\pi}_i \rightarrow \pi_i$ and $\widehat{a}_i \rightarrow 1$ *a.e.* so the first two summands tend to zero. On the other hand, as it can also be seen in the appendix, $\{\widehat{g}_i(\mathbf{x})\}$ is, for n large enough, a sequence of nonnegative integrable functions that, for all $\mathbf{x} \in \mathbb{R}^p$, converges *a.e.* to $f_i(\mathbf{x})$. Also, $0 \leq [f_i(\mathbf{x}) - \widehat{g}_i(\mathbf{x})]_+ \leq f_i(\mathbf{x})$ so the third summand converges to zero by lemma 3.1.3 in Glick (1974) (see also Prakasa Rao 1983, p. 191). Notice finally that, to facilitate matters, the previous convergence is obtained treating $s = rank(\mathcal{D})$ as a fixed known constant. Schott (1993) develops a formal test for the true value of $rank(\mathcal{D})$.

5. FINAL COMMENTS

This paper presents a proposal for dimension reduction in discriminant analysis. The problem of dimension reduction in classification problems is not trivial as the following remarks illustrate. Consider the subset

$$\mathcal{S} = \{1 \leq r \leq p : \exists \mathbf{y} = t(\mathbf{x}) \text{ with } L[d_t^*(\mathbf{y}_r)] = L[r^*(\mathbf{x})]\} \, ,$$

where, as in section 2, $d_t^*(\mathbf{y}_r)$ is the optimal rule in the space $\mathbf{y}_r = (y_1, ..., y_r)' = (t_1(\mathbf{x}), ..., t_r(\mathbf{x}))'$. This set is not empty because at least $p \in \mathcal{S}$. If $R = \min \mathcal{S}$, the dimension is optimally reduced when a transformation $\mathbf{y} = t(\mathbf{x})$ can be found such that $L[d_t^*(\mathbf{y}_R)] = L[r^*(\mathbf{x})]$. That is, R is the minimum number of coordinates needed to attain the Bayes error $L^* = L[r^*(\mathbf{x})]$. Borrowing terminology from Fukunaga (1990), R can be interpreted as the *intrinsic dimension* of the discriminant problem. To determine the pair (t, R) is generally a complex and infeasible problem and, as in the examples considered in this paper, it may be convenient to restrict attention to the class of linear transformations $\mathbf{y} = t(\mathbf{x}) = \mathbf{A}'(\mathbf{x} - \mathbf{b})$. Methods for dimension reduction in classification using linear transformations have been considered previously by several authors, among others, Decell, Odell and Coberly (1981) and McCulloch (1986). The general framework presented in this paper, based in the concept of dimension reduction transformation and the accompanying data based algorithm of section 4, can be a useful tool for dimension reduction in discriminant analysis.

6. APPENDIX

6.1 Proof of Proposition 1

From (2), the probability of misclassification can be written as

$$\begin{aligned} L[r(\mathbf{x})] &= 1 - \sum_{i=1}^{k} \int_{R_i} \pi_i(\mathbf{x}) \mu(d\mathbf{x}) \\ &= 1 - \sum_{i=1}^{k} \int_{\mathbb{R}^p} \pi_i(\mathbf{x}) I_{R_i}(\mathbf{x}) \mu(d\mathbf{x}) = 1 - \int_{\mathbb{R}^p} h_r(\mathbf{x}) \mu(d\mathbf{x}) \,, \end{aligned} \tag{67}$$

where $h_r(\mathbf{x}) = \sum_{i=1}^{k} \pi_i(\mathbf{x}) I_{R_i}(\mathbf{x})$. To see part *i)*, notice that for any rule $r^*(\mathbf{x}) = \sum_{i=1}^{k} i I_{R_i^*}(\mathbf{x})$ such that the R_i^* satisfy condition (3), if $h_{r^*}(\mathbf{x}) = \sum_{i=1}^{k} \pi_i(\mathbf{x}) I_{R_i^*}(\mathbf{x})$, for

all $\mathbf{x} \in \mathbb{R}^p$ one has the inequality

$$\begin{aligned} h_r(\mathbf{x}) &= \sum_{i=1}^{k} \pi_i(\mathbf{x}) I_{R_i}(\mathbf{x}) = \sum_{i=1}^{k} \pi_i(\mathbf{x}) [\sum_{j=1}^{k} I_{R_i \cap R_j^*}(\mathbf{x})] = \sum_{j=1}^{k} [\sum_{i=1}^{k} \pi_i(\mathbf{x}) I_{R_i \cap R_j^*}(\mathbf{x})] \\ &\leq \sum_{j=1}^{k} [\sum_{i=1}^{k} \pi_j(\mathbf{x}) I_{R_i \cap R_j^*}(\mathbf{x})] = \sum_{j=1}^{k} \pi_j(\mathbf{x}) [\sum_{i=1}^{k} I_{R_i \cap R_j^*}(\mathbf{x})] \\ &= \sum_{j=1}^{k} \pi_j(\mathbf{x}) I_{R_j^*}(\mathbf{x}) = h_{r^*}(\mathbf{x}) \ . \end{aligned} \tag{68}$$

Therefore, $\int_{\mathbb{R}^p} h_r(\mathbf{x}) \mu\,(d\mathbf{x}) \leq \int_{\mathbb{R}^p} h_{r^*}(\mathbf{x}) \mu\,(d\mathbf{x})$ and, according to (67), $L[r^*(\mathbf{x})] \leq L[r(\mathbf{x})]$. To see part *ii),* let $s^*(\mathbf{x})$ be another Bayes rule corresponding to a partition S_1^*, ..., S_k^* and consider the function $h_{s^*}(\mathbf{x}) = \sum_{i=1}^{k} \pi_i(\mathbf{x}) I_{S_i^*}(\mathbf{x})$ of representation (67) for $L[s^*(\mathbf{x})]$. From (68) $h_{r^*}(\mathbf{x}) - h_{s^*}(\mathbf{x}) \geq 0$ and, since $L[r^*(\mathbf{x})] = L^* = L[s^*(\mathbf{x})]$, using (67) one also has $\int_{\mathbb{R}^p} [h_{r^*}(\mathbf{x}) - h_{s^*}(\mathbf{x})] \mu\,(d\mathbf{x}) = 0$. This leads to $h_{r^*}(\mathbf{x}) - h_{s^*}(\mathbf{x}) = 0$, *a.e.* (μ). By *(C1)* there exists a set M in $\mathcal{B}^p$ with $\mu(M) = 1$ such that, if $\mathbf{x} \in M$, $h_{r^*}(\mathbf{x}) = h_{s^*}(\mathbf{x})$ and $\pi_i(\mathbf{x}) \neq \pi_j(\mathbf{x})$ if $i \neq j$. Therefore, for $i = 1, ..., k$, $\mathbf{x} \in R_i^* \cap M$ if, and only if, $\mathbf{x} \in S_i^* \cap M$, $i = 1, ..., k$. As a consequence, the symmetric difference $R_i^* \triangle S_i^* = [R_i^* \cap (S_i^*)^c] \cup [(R_i^*)^c \cap S_i^*] \subseteq M^c$, and hence, for $i = 1, ..., k$, $P\,[\mathbf{x} \in R_i^* \triangle S_i^*] = \mu\,(R_i^* \triangle S_i^*) \leq \mu(M^c) = 0$, that is, $P[r^*(\mathbf{x}) \neq s^*(\mathbf{x})] \leq \sum_{i=1}^{k} P\,[\mathbf{x} \in R_i^* \triangle S_i^*] = 0$. ■

6.2 Auxiliary convergences

Let $I_{(i)}(.)$ be the indicator function of class i. By the law of large numbers

$$\widehat{\pi}_i = \frac{n_i}{n} = \frac{1}{n} \sum_{j=1}^{n} I_{(i)}(\mathbf{g}_j) \to E[I_{(i)}(\mathbf{g})] = \sum_{j=1}^{k} \pi_j E[I_{(i)}(\mathbf{g}) \mid \mathbf{g} = j] = \pi_i \quad a.e. \ ,$$

as $n \to \infty$. Convergence of $\overline{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$, $\overline{\mathbf{x}} = \sum_{i=1}^{k} (n_i/n) \overline{\mathbf{x}}_i$, $\widehat{\mathbf{\Sigma}}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)'/n_i$ and $\widehat{\mathbf{\Sigma}} = \sum_{i=1}^{k} (n_i/n) \widehat{\mathbf{\Sigma}}_i$ to, respectively, $\boldsymbol{\mu}_i = E(\mathbf{x} \,|\mathbf{g} = i\,)$, $\boldsymbol{\mu} = E(\mathbf{x})$,

$\boldsymbol{\Sigma}_i = Var(\mathbf{x}\,|\mathbf{g}=i)$ and $\boldsymbol{\Sigma} = \sum_{i=1}^{k} \pi_i \boldsymbol{\Sigma}_i$ is obtained similarly. For example,

$$\begin{aligned} \overline{\mathbf{x}}_i &= \frac{\sum_{j=1}^{n_i} \mathbf{x}_{ij}}{n_i} = \frac{\sum_{j=1}^{n} \mathbf{x}_j I_{(i)}(\mathbf{g}_j)/n}{\sum_{j=1}^{n} I_{(i)}(\mathbf{g}_j)/n} \to \frac{E[\mathbf{x} I_{(i)}(\mathbf{g})]}{\pi_i} \\ &= \frac{\sum_{j=1}^{k} \pi_j E[\mathbf{x} I_{(i)}(\mathbf{g}) \mid \mathbf{g}=j]}{\pi_i} = E(\mathbf{x}\,|\mathbf{g}=i) = \boldsymbol{\mu}_i \quad a.e. \ . \end{aligned}$$

By lemma 2.1 in Tyler (1981, p.726), when $s = rank(\mathcal{D})$, $\widehat{\mathbf{U}}_1 \widehat{\mathbf{U}}_1' \to \mathbf{U}_1 \mathbf{U}_1'$ *a.e.*, and then $\mid \widehat{\mathbf{Q}}_i \mid = \mid \widehat{\mathbf{U}}_1' \widehat{\boldsymbol{\Sigma}}^{-1/2} \widehat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{\Sigma}}^{-1/2} \widehat{\mathbf{U}}_1 \mid \to \mid \mathbf{U}_1' \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}^{-1/2} \mathbf{U}_1 \mid = |\mathbf{Q}_i|$ and $\widehat{\mathbf{U}}_1 \widehat{\mathbf{Q}}_i^{-1} \widehat{\mathbf{U}}_1' \to \mathbf{U}_1 \widehat{\mathbf{Q}}_i^{-1} \mathbf{U}_1'$. Since, from identity (53), $|\mathbf{Q}_i| = |\boldsymbol{\Sigma}|^{-1} |\boldsymbol{\Sigma}_i|$, using expressions (60) and (63) it turns out that, for all $\mathbf{x} \in \mathbb{R}^p$, $\widehat{g}_i(\mathbf{x})$ converges *a.e.* to the density $f_i(\mathbf{x})$.

Finally, since both $\boldsymbol{\Sigma}$ and $\mathbf{Q}_i$ are positive definite matrices, for n large enough the following change of variable can be considered

$$\mathbf{z} = \begin{pmatrix} \widehat{\mathbf{Q}}_i^{-1/2} \widehat{\mathbf{U}}_1' \widehat{\boldsymbol{\Sigma}}^{-1/2} (\mathbf{x} - \overline{\mathbf{x}}_i) \\ \mathbf{U}_2' \boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu}_i) \end{pmatrix} = \widehat{\mathbf{A}}_i \mathbf{x} + \widehat{\mathbf{b}}_i \ ,$$

where

$$\widehat{\mathbf{A}}_i = \begin{pmatrix} \widehat{\mathbf{Q}}_i^{-1/2} \widehat{\mathbf{U}}_1' \widehat{\boldsymbol{\Sigma}}^{-1/2} \\ \mathbf{U}_2' \boldsymbol{\Sigma}^{-1/2} \end{pmatrix}, \quad \widehat{\mathbf{b}}_i = - \begin{pmatrix} \widehat{\mathbf{Q}}_i^{-1/2} \widehat{\mathbf{U}}_1' \widehat{\boldsymbol{\Sigma}}^{-1/2} \overline{\mathbf{x}}_i \\ \mathbf{U}_2' \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\mu}_i \end{pmatrix} . \tag{69}$$

Define $\widehat{\mathbf{V}}_i = \widehat{\mathbf{A}}_i' \widehat{\mathbf{A}}_i = \widehat{\boldsymbol{\Sigma}}^{-1/2} \widehat{\mathbf{U}}_1 \widehat{\mathbf{Q}}_i^{-1} \widehat{\mathbf{U}}_1' \widehat{\boldsymbol{\Sigma}}^{-1/2} + \boldsymbol{\Sigma}^{-1/2} \mathbf{U}_2 \mathbf{U}_2' \boldsymbol{\Sigma}^{-1/2}$. By (53), as $n \to \infty$, $\widehat{\mathbf{V}}_i \to \boldsymbol{\Sigma}^{-1/2} \mathbf{U}_1 \mathbf{Q}_i^{-1} \mathbf{U}_1' \boldsymbol{\Sigma}^{-1/2} + \boldsymbol{\Sigma}^{-1/2} \mathbf{U}_2 \mathbf{U}_2' \boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}_i^{-1}$ *a.e.*, so it might also be assumed that $r(\widehat{\mathbf{A}}_i) = r(\widehat{\mathbf{A}}_i' \widehat{\mathbf{A}}_i) = r(\widehat{\mathbf{V}}_i) = p$. Since $|\partial \mathbf{x} / \partial \mathbf{z}| = |\partial \mathbf{z} / \partial \mathbf{x}|^{-1} = \mid \widehat{\mathbf{V}}_i \mid^{-1/2}$,

$$\begin{aligned} \widehat{a}_i &= \int_{\mathbb{R}^p} \widehat{g}_i(\mathbf{x}) dx = (\mid \widehat{\boldsymbol{\Sigma}} \mid\mid \widehat{\mathbf{Q}}_i \mid\mid \widehat{\mathbf{V}}_i \mid)^{-1/2} (2\pi)^{-p/2} \int_{\mathbb{R}^p} \exp(- \|z\|^2 / 2) dz \\ &= (\mid \widehat{\boldsymbol{\Sigma}} \mid\mid \widehat{\mathbf{Q}}_i \mid\mid \widehat{\mathbf{V}}_i \mid)^{-1/2} \to 1 \ . \blacksquare \end{aligned}$$

REFERENCES

[1] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis,* 2nd Edn. New York: John Wiley.

[2] Billingsley, P. (1995). *Probability and Measure*, 3rd Edn. New York: John Wiley.

[3] Decell, H. P., Odell, P.L. and Coberly, W. A. (1981). Linear dimension reduction and Bayes classification. Pattern Recognition, **13**, 241-243.

[4] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation.* New York: John Wiley.

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

[6] Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition.* New York: Academic Press.

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

[8] Johnson, M. E. (1987). *Multivariate Statistical Simulation,* New York: John Wiley.

[9] Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization, *Sankhyā.* A, **32**, 419-430.

[10] McCulloch, R. E. (1986). Some remarks on allocatory and separatory linear discrimination. Journal of Statistical Planning and Inference, **14**, 323-340.

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

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

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

[14] Schott, J.R. (1993). Dimensionality reduction in quadratic discriminant analysis. *Computational Statistics and Data Analysis*, **16**, 161-174.

[15] Tyler, D.E. (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, **9**, 725-736.