Working Paper 01-39
Statistics and Econometrics Series 25
October 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 in Nonparametric Discriminant Analysis

Adolfo Hernández and Santiago Velilla*

Abstract

A dimension reduction method in kernel discriminant analysis is presented, based on the concept of dimension reduction subspace. Examples of application are discussed.

Keywords: curse of dimensionality, dimension reduction subspaces, kernel discriminant rule, projection pursuit, separation of populations.

* Hernández, Departamento de Análisis Económico, Universidad Autónoma de Madrid, 28049-Cantoblanco, Madrid, Spain. Velilla, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, 28903-Getafe, Madrid, Spain. Research partially supported by $CICYT$ Grant $BEC2000-0167$ (Spain). The authors are grateful to F. J. Prieto for helpful computing advice.

Dimension Reduction in Nonparametric Discriminant Analysis

Adolfo Hernández and Santiago Velilla*

Abstract

A dimension reduction method in kernel discriminant analysis is presented, based on the concept of dimension reduction subspace. Examples of application are discussed.

Keywords and phrases: curse of dimensionality, dimension reduction subspaces, kernel discriminant rule, projection pursuit, separation of populations.

1. INTRODUCTION

Consider a classification 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 $\mathbf{x} = (x_1, ..., x_p)'$. If the possible distributions of $\mathbf{x}$ are assumed to be continuous, the optimal or Bayes rule, that is the rule that minimizes the probability of misclassification, assigns $\mathbf{x}$ to group g_i when

$$\pi_i f_i(\mathbf{x}) = \max_{1 \leq j \leq k} \pi_j f_j(\mathbf{x}) \,, \tag{1}$$

*Hernández, Departamento de Análisis Económico, Universidad Autónoma de Madrid, 28049-Cantoblanco, Madrid, Spain. Velilla, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, 28903-Getafe, Madrid, Spain. Research partially supported by $CICYT$ Grant $BEC2000 - 0167$ (Spain). The authors are grateful to F. J. Prieto for helpful computing advice.

where, for $i = 1, ..., k$, $f_i(\mathbf{x})$ is the ith class conditional density and, if $\mathbf{g}$ is the random label that denotes the true class membership of the individual under study, $\pi_i = P[\mathbf{g} = i] > 0$ is the ith class prior probability (see e.g. Seber 1984, sec. 6.9). Given that the pairs $(\pi_i, f_i(\mathbf{x}))$ are typically unknown, rule (1) is often implemented in practice once the unknown quantities involved in its construction have been replaced by adequate estimators computed from

$$\mathbf{D}_n = \{(\mathbf{x}_j, \mathbf{g}_j) : \ j = 1, ..., n\}, \tag{2}$$

a database of size n formed by i.i.d. observations from the pair $(\mathbf{x},\mathbf{g})$ obtained from individuals previously classified. For example, the prior probabilities π_i may be estimated by the proportions $\widehat{\pi}_i = n_i/n$, where n_i represents the number of observations $(\mathbf{x}_j,\mathbf{g}_j)$ in $\mathbf{D}_n$ such that $\mathbf{g}_j = i$. If, on the other hand, the class conditional densities are not supposed to follow any particular model, $f_i(\mathbf{x})$ may be estimated by a nonparametric kernel density estimator of the form

$$\widehat{f}_i(\mathbf{x}) = \frac{1}{n_i h_i^p} \sum_{j=1}^{n} K[\frac{1}{h_i}(\mathbf{x} - \mathbf{x}_j)] I_{(i)}(\mathbf{g}_j) \ , \tag{3}$$

see e.g. Scott (1992, chap. 6), where $K(.)$ is a suitable kernel function, h_i is an smoothing parameter and $I_{(i)}(\mathbf{g}_j)$ is an indicator function that takes the value 1 when $\mathbf{g}_j = i$ and 0 otherwise. After replacing in (1) the pairs $(\pi_i, f_i(\mathbf{x}))$ by the pairs $(\widehat{\pi}_i, \widehat{f}_i(\mathbf{x}))$, the sample based plug-in rule that assigns $\mathbf{x}$ to group g_i when

$$(n_i/n)\widehat{f}_i(\mathbf{x}) = \max_{1 \le j \le k} (n_j/n)\widehat{f}_j(\mathbf{x}) \ , \tag{4}$$

is the so called kernel discriminant rule and a classification procedure based on (4) is commonly denoted as kernel discriminant analysis (KDA).

Despite its natural construction and well established theoretical properties (see e.g. Devroye, Györfi and Lugosi 1996, chap. 10), the performance of the KDA rule in applications deteriorates as the dimension p of the feature vector $\mathbf{x}$ increases (Hand

1997, chap. 5). This phenomenon, usually referred to as the *"curse of dimensionality"*, motivates the need of constructing efficient dimension reduction methods. The aim of this paper is to develop a dimension reduction procedure that, after projecting the vector of features onto a lower dimensional subspace, allows to perform KDA in a nearly optimal fashion. Section 2 establishes notation and contains some background and motivation. Section 3 presents the theoretical foundations of the procedure. Section 4 studies related practical implementation issues and section 5 develops some applications on real or simulated data. Section 6 gives some final comments.

2. BACKGROUND AND MOTIVATION

In the continuous feature vector case, the joint distribution of the pair ($\mathbf{x}$,$\mathbf{g}$) is characterized by the prior probabilities $\pi_i = P[\mathbf{g} = i]$ and class conditional densities $f_i(\mathbf{x})$, or, alternatively, by the marginal density of $\mathbf{x}$, $f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) + \ ... \ + \pi_k f_k(\mathbf{x})$, combined with the posterior class probabilities

$$\pi_i(\mathbf{x}) = P[\mathbf{g} = i \,|\mathbf{x}] = \frac{\pi_i f_i(\mathbf{x})}{f(\mathbf{x})} \ , \quad i = 1, ..., k \ . \tag{5}$$

When the densities $f_i(\mathbf{x})$ have finite moments of order two, the $p \times p$ matrices $\mathbf{\Sigma}_i = Var(\mathbf{x} \,|\mathbf{g} = i)$ are positive definite for $i = 1, ..., k$, so $\mathbf{x}$ can be conveniently standardized in the form $\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = E(\mathbf{x})$ is the expected value with respect to the marginal density $f(\mathbf{x})$, and

$$\mathbf{\Sigma} = E[Var(\mathbf{x} \,|\mathbf{g})] = \sum_{i=1}^{k} \pi_i \mathbf{\Sigma}_i \, , \tag{6}$$

is the $p \times p$ *within groups* dispersion matrix. If the linear transformation

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} = \mathbf{A}' \mathbf{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu}) \ , \tag{7}$$

is now considered, where $\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_p)$ is a non singular $p \times p$ matrix of generic column $\mathbf{a}_j$, $j = 1, ..., p$, the joint distribution of $(\mathbf{y},\mathbf{g})$ is characterized by the priors π_i and class conditional densities $f_{\mathbf{y},i}(\mathbf{y})$, $i = 1, ..., k$, or, in other form, by the marginal density $f_{\mathbf{y}}(\mathbf{y}) = \pi_1 f_{\mathbf{y},1}(\mathbf{y}) + ... + \pi_k f_{\mathbf{y},k}(\mathbf{y})$ and posterior probability functions

$$q_i(\mathbf{y}) = P[\mathbf{g} = i \,|\mathbf{y}] = \frac{\pi_i f_{\mathbf{y},i}(\mathbf{y})}{f_{\mathbf{y}}(\mathbf{y})} \;, \quad i = 1, ..., k \;. \tag{8}$$

Using standard arguments of change of variable, the optimal or Bayes error corresponding to rule (1) can be written in the form

$$L^* = 1 - \sum_{i=1}^{k} P[\mathbf{x} \in R_i \,;\, \mathbf{g} = i] =$$

$$= 1 - \sum_{i=1}^{k} \pi_i \int_{R_i} f_i(\mathbf{x}) d\mathbf{x} = 1 - \sum_{i=1}^{k} \pi_i \int_{S_i} f_{\mathbf{y},i}(\mathbf{y}) d\mathbf{y} \,, \tag{9}$$

where, for $i = 1, ..., k$, $R_i = \{\mathbf{x} \in \mathbb{R}^p : \pi_i f_i(\mathbf{x}) = \max_{1 \le j \le k} \pi_j f_j(\mathbf{x})\}$ and $S_i = \{\mathbf{y} \in \mathbb{R}^p : \pi_i f_{\mathbf{y},i}(\mathbf{y}) = \max_{1 \le j \le k} \pi_j f_{\mathbf{y},j}(\mathbf{y})\}$. The Bayes error is then the same in both the $\mathbf{x}$ and $\mathbf{y}$ spaces and, as a natural idea, transformation (7) could be designed to achieve the optimum error L^* using only the first s coordinates of the transformed feature vector $\mathbf{y} = (y_1, ..., y_s, y_{s+1}, ..., y_p)'$ where, hopefully, $1 \le s \ll p$.

To that end, let $f_{\mathbf{y},i}(y_1, ..., y_s)$ be the marginal density of $(y_1, ..., y_s)$ under $f_{\mathbf{y},i}(\mathbf{y})$, and write

$$f_{\mathbf{y},i}(\mathbf{y}) = f_{\mathbf{y},i}(y_1, ..., y_s) f_{\mathbf{y},i}(y_{s+1}, ..., y_p \,|y_1, ..., y_s) \;, \tag{10}$$

where $f_{\mathbf{y},i}(y_{s+1}, ..., y_p \,|y_1, ..., y_s)$ is the conditional density of $(y_{s+1}, ..., y_p)$ given $(y_1, ..., y_s)$. Similarly, write

$$f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{y}}(y_1, ..., y_s) f_{\mathbf{y}}(y_{s+1}, ..., y_p \,|y_1, ..., y_s) \;, \tag{11}$$

where, respectively, $f_{\mathbf{y}}(y_1, ..., y_s)$ and $f_{\mathbf{y}}(y_{s+1}, ..., y_p \,|y_1, ..., y_s)$ are, in the probability distribution in $\mathbb{R}^p$ defined by $f_{\mathbf{y}}(\mathbf{y})$, the marginal density of $(y_1, ..., y_s)$ and the

conditional density of $(y_{s+1},\ ...,\ y_p)$ given $(y_1,\ ...,\ y_s)$. If, for some $1 \leq s < p$, the condition below holds:

(C1) For $i = 1,\ ...,\ k$,

$$f_{\mathbf{y},i}(y_{s+1},\ ...,\ y_p\,|y_1,\ ...,\ y_s) = f_{\mathbf{y}}(y_{s+1},\ ...,\ y_p\,|y_1,\ ...,\ y_s)\ , \tag{12}$$

one has, using (10),

$$S_i = \{\mathbf{y} \in \mathbb{R}^p : \pi_i f_{\mathbf{y},i}(\mathbf{y}) = \max_{1\leq j\leq k} \pi_j f_{\mathbf{y},j}(\mathbf{y})\} = U_i \times \mathbb{R}^{p-s}\ , \tag{13}$$

where $U_i = \{(y_1, ..., y_s) \in \mathbb{R}^s : \pi_i f_{\mathbf{y},i}(y_1, ..., y_s) = \max_{1\leq j\leq k} \pi_j f_{\mathbf{y},j}(y_1, ..., y_s)\}$. The Bayes error in equation (9) can be then reexpressed as

$$\begin{aligned} L^* &= 1 - \sum_{i=1}^{k} \pi_i \int_{S_i} f_{\mathbf{y},i}(\mathbf{y})d\mathbf{y} = 1 - \sum_{i=1}^{k} \pi_i \int_{U_i\times\mathbb{R}^{p-s}} f_{\mathbf{y},i}(\mathbf{y})d\mathbf{y} = \\ &= 1 - \sum_{i=1}^{k} \pi_i \int_{U_i} f_{\mathbf{y},i}(y_1, ..., y_s)dy_1\ ...\ dy_s\ , \end{aligned} \tag{14}$$

and, in conclusion, if, for $1 \leq s \leq p$,

$$\mathbf{A}_s = (\mathbf{a}_1, ..., \mathbf{a}_s) \tag{15}$$

is the $p \times s$ matrix formed by the first s columns of the matrix $\mathbf{A}$ in (7), assigning $(y_1,\ ...,\ y_s)' = \mathbf{A}_s'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ to group g_i when

$$\pi_i f_{\mathbf{y},i}(y_1, ..., y_s) = \max_{1\leq j\leq k} \pi_j f_{\mathbf{y},j}(y_1, ..., y_s)\,, \tag{16}$$

is a classification procedure that, as desired, achieves the Bayes error L^* using only the first s coordinates of the transformed feature vector $\mathbf{y} = \mathbf{A}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$.

Using now standard properties of conditional probability, for $i = 1, ..., k$ the identity $\pi_i(\mathbf{x}) = q_i[\mathbf{A}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})]$ holds. Combining this expression with (8), (10) and (11) above, condition *(C1)* can be seen to be equivalent to the alternative condition:

(C2) For $i = 1, ..., k$, there exists some function $h_i(.)$ such that

$$\pi_i(\mathbf{x}) = P[\mathbf{g} = i \,|\mathbf{x}] = q_i[\mathbf{A}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})] = h_i[\mathbf{A}_s^{'}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})]\,. \tag{17}$$

In other words, and following Cook (1998, chap. 6), the conditional density functions $f_{\mathbf{y},i}(y_{s+1}, ..., y_p \,|y_1, \, ..., \, y_s)$ are identical across groups if, and only if, $Col(\mathbf{A}_s)$, the column space spanned by the columns of the $p \times s$ matrix $\mathbf{A}_s$ in (15), is a *dimension reduction subspace* for $\mathbf{g} \left| \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \right.$, the conditional distribution of the group label $\mathbf{g}$ given the standardized feature vector $\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. Condition (17), and therefore condition (12), can be seen to be equivalent to the statement that $\mathbf{g}$ and $(y_{s+1}, \, ..., \, y_p)$ are conditionally independent once the first s coordinates $(y_1, ..., y_s)$ of the transformed feature vector $\mathbf{y} = \mathbf{A}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ have been determined. That is, if $Col(\mathbf{A}_s)$ is a dimension reduction subspace, the original feature vector $\mathbf{x}$ can be, for classification purposes, replaced by the projected coordinates $(y_1, \, ..., \, y_s)' = \mathbf{A}_s^{'}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ without loss of discriminatory information.

As a consequence of the above, the problem of optimal dimension reduction in nonparametric discriminant analysis by means of a linear transformation of the form (7), can be solved by finding a dimension reduction subspace relative to the conditional distribution $\mathbf{g} \left| \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \right.$ with the smallest possible dimension. According to Cook (1998, chap. 6), assuming that the support of the marginal density $f(\mathbf{x})$ is a convex subset of $\mathbb{R}^p$, this subspace, termed the *central dimension reduction subspace* or simply the *central subspace*, exists, is unique and coincides with the intersection of all dimension reduction subspaces. If by convention the central subspace is an $r-$dimensional subspace of the form $L_0 = Col(\mathbf{A}_{r,0})$, where $\mathbf{A}_{r,0}$ is a $p \times r$ matrix of rank r, the "canonical" coordinates

$$\begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix} = \mathbf{A}_{r,0}^{'}\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \;, \tag{18}$$

are, in a sense, the smallest possible amount of information needed for optimal classification by means of the following reduced optimal rule: assign $\mathbf{x}$ to g_i when

$$\pi_i f_{\mathbf{y},i}(y_1, ..., y_r) = \max_{1 \le j \le k} \pi_j f_{\mathbf{y},j}(y_1, ..., y_r) \ , \tag{19}$$

where the $f_{\mathbf{y},i}(y_1, ..., y_r)$ are the class conditional densities of coordinates (18).

3. DETECTION OF DIMENSION REDUCTION SUBSPACES

Paradigm (18)-(19) motivates the need of determining the dimension and specific structure of the central subspace $L_0 = C(\mathbf{A}_{r,0})$. This is done in this section by introducing two numeric functionals of the pairs $(\pi_i, f_i(\mathbf{x}))$ that can be used to characterize when, for some $1 \le s \le p$, $Col(\mathbf{A}_s)$ is a dimension reduction subspace. For every invertible $s \times s$ matrix $\mathbf{B}$ the identity $Col(\mathbf{A}_s) = Col(\mathbf{A}_s\mathbf{B})$ holds, so if $\mathbf{A}_s$ spans a dimension reduction subspace the same is true for $\mathbf{A}_s\mathbf{B}$. The matrix $\mathbf{A}_s$ can then be taken as being suborthogonal for all $1 \le s \le p$ or, in other words, it is enough to consider linear transformations of the form (7) where the matrix $\mathbf{A}$ is orthogonal. In what follows, emphasis is in the functionals as tools for subspace detection, leaving a detailed discussion of their properties for appendix A.

3.1 Dimension reduction functionals

Given two densities $g(\mathbf{x})$ and $h(\mathbf{x})$ in $\mathbb{R}^p$, the quantity

$$I(g, h) = \int_{\mathbb{R}^p} \log[\frac{g(\mathbf{x})}{h(\mathbf{x})}] g(\mathbf{x}) d\mathbf{x} \ , \tag{20}$$

is the well-known relative entropy between $g(\mathbf{x})$ and $h(\mathbf{x})$ (see e.g. Huber 1985, secs. 11 and 12). $I(g, h)$ is always a non negative number, possibly infinite, and $I(g, h) = 0$ only when $g(\mathbf{x})$ and $h(\mathbf{x})$ coincide. With this notation, if $f(\mathbf{x}) = \pi_1 f_1(\mathbf{x}) +$ ... $+\pi_k f_k(\mathbf{x})$, the functional

$$H = \sum_{i=1}^{k} \pi_i I(f, f_i) \, , \tag{21}$$

will be called the *global entropy* of the discriminant problem defined by the pairs $(\pi_i, f_i(\mathbf{x}))$, $i = 1, ..., k$. Assuming $I(f, f_i)$ finite for all i, the magnitude H of (21) is such that $0 \leq H < +\infty$. Taking into account the well established notion of the relative entropy (20) as a measure of discrepancy, the index H can be interpreted as an aggregate *measure of separation* between the conditional class densities $f_i(\mathbf{x})$ and the marginal density $f(\mathbf{x})$. Also, if for a fixed $p \times p$ orthogonal matrix $\mathbf{A}$ the linear transformation $\mathbf{y} = \mathbf{A}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ of (7) is considered, by change of variable arguments the global entropy is the same in the classification problems defined by either the pairs $(\pi_i, f_i(\mathbf{x}))$ or the pairs $(\pi_i, f_{\mathbf{y},i}(\mathbf{y}))$. That is, separation among densities remains constant under non singular linear transformations. As it can be seen in appendix A, for every $1 \leq s \leq p$ the additive decomposition below holds:

$$H = H_s(\mathbf{A}_s) + J_s(\mathbf{A}) \ , \tag{22}$$

where, if $f_{\mathbf{y},i}(y_1, ..., y_s)$ and $f_{\mathbf{y}}(y_1, ..., y_s)$ are as in (10) and (11), the index

$$H_s(\mathbf{A}_s) = \sum_{i=1}^{k} \pi_i I[f_{\mathbf{y}}(y_1, ..., y_s), f_{\mathbf{y},i}(y_1, ..., y_s)] \ , \tag{23}$$

is the global entropy in the projected coordinates $(y_1, ..., y_s)' = \mathbf{A}_s'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, and

$$J_s(\mathbf{A}) = \sum_{i=1}^{k} \pi_i E\{I[f_{\mathbf{y}}(y_{s+1}, .., y_p \,|y_1, ..., y_s), f_{\mathbf{y},i}(y_{s+1}, .., y_p \,|y_1, ..., y_s)]\} \, , \tag{24}$$

where in (24) expectation is taken with respect to the marginal density $f_{\mathbf{y}}(y_1, ..., y_s)$. Index $J_s(\mathbf{A})$ is always nonnegative or equivalently using (22), $0 \leq H_s(\mathbf{A}_s) \leq H$. In other words, as measured respectively by indexes $H_s(\mathbf{A}_s)$ in (23) and H in (21), the degree of separation among densities after projecting $\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ onto $Col(\mathbf{A}_s)$, is always smaller than the degree of separation among the original densities $f_i(\mathbf{x})$.

To derive a second dimension reduction functional, the true membership of a given individual is modelled by a random vector $\mathbf{G}$ that takes when $\mathbf{g} = i$ the value $\mathbf{G} = \mathbf{e}_i$,

where $\mathbf{e}_i$ is the *ith* canonical vector of $\mathbb{R}^k$. Let

$$G_p(\mathbf{x}) = E(\mathbf{G}\,|x_1,\, ...,\, x_p) =$$

$$= (\pi_1(\mathbf{x}),\, ...,\, \pi_k(\mathbf{x}))' = (\pi_1 f_1(\mathbf{x})/f(\mathbf{x}),\, ...,\, \pi_k f_k(\mathbf{x})/f(\mathbf{x}))'\ , \tag{25}$$

be the best mean square prediction of $\mathbf{G}$ in terms of the feature vector $\mathbf{x} = (x_1,\, ...,\, x_p)'$. The functional

$$C = tr\, Var[G_p(\mathbf{x})] = \sum_{i=1}^{k} var[\pi_i(\mathbf{x})] =$$

$$= \sum_{i=1}^{k}\{E[\pi_i^2(\mathbf{x})] - E^2[\pi_i(\mathbf{x})]\} = \sum_{i=1}^{k}\int_{\mathbb{R}^p}\left(\frac{\pi_i f_i(\mathbf{x})}{f(\mathbf{x})}\right)^2 f(\mathbf{x})d\mathbf{x} - \sum_{i=1}^{k}\pi_i^2\ , \tag{26}$$

will be called the *total prediction capacity* for the new "group label" $\mathbf{G}$ of the feature vector $\mathbf{x}$. The properties of C in (26) are similar to the ones of H in (21). For example, by change of variable arguments, the transformed feature vector $\mathbf{y} = \mathbf{A}'\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ of (7) possesses the same prediction capacity for $\mathbf{G}$ than the original vector $\mathbf{x}$. Moreover, as established in appendix A, for all integers $1 \leq s \leq p$ the additive decomposition below, similar to decomposition (22), holds:

$$C = C_s(\mathbf{A}_s) + D_s(\mathbf{A})\ , \tag{27}$$

where, if $G_s(\mathbf{y}) = E(\mathbf{G}\,|y_1,\, ...,\, y_s)$ is the total prediction of $\mathbf{G}$ offered by the projected coordinates $(y_1,\, ...,\, y_s)' = \mathbf{A}_s'\mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$,

$$C_s(\mathbf{A}_s) = trVar[G_s(\mathbf{y})] =$$

$$= \sum_{i=1}^{k}\int_{\mathbb{R}^s}\left(\frac{\pi_i f_{\mathbf{y},i}(y_1,\, ...,\, y_s)}{f_{\mathbf{y}}(y_1,\, ...,\, y_s)}\right)^2 f_{\mathbf{y}}(y_1,\, ...,\, y_s) - \sum_{i=1}^{k}\pi_i^2\ , \tag{28}$$

is the prediction capacity in $(y_1,\, ...,\, y_s)'$ and

$$D_s(\mathbf{A}) = trVar[G_p(\mathbf{y}) - G_s(\mathbf{y})]\ . \tag{29}$$

Definition (26) above depends on the variances of the ratios $\pi_i(\mathbf{x}) = \pi_i f_i(\mathbf{x})/f(\mathbf{x})$, $i = 1, ..., k$, so index C can be again interpreted as an aggregate measure of separation between the densities $f_i(\mathbf{x})$ and $f(\mathbf{x})$. $D_s(\mathbf{A})$ in (29) is always nonnegative so, according to (27), $0 \leq C_s(\mathbf{A}_s) \leq C$ and, as with the relative entropy, separation among densities decreases after projecting onto $Col(\mathbf{A}_s)$.

For every pair ($\mathbf{A}_s$,$\mathbf{F}$), where $\mathbf{A}_s$ is suborthogonal of $p \times s$ and $\mathbf{F}$ orthogonal of $s \times s$, by arguments of change of variable one has the identities

$$H_s(\mathbf{A}_s) = H_s(\mathbf{A}_s\mathbf{F}) \quad , \quad C_s(\mathbf{A}_s) = C_s(\mathbf{A}_s\mathbf{F}) \ . \tag{30}$$

Since $Col(\mathbf{A}_s) = Col(\mathbf{A}_s\mathbf{F})$, both $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ take then values that are specific to the subspace $Col(\mathbf{A}_s)$ and not to the particular form of the columns $\mathbf{a}_j$, $j = 1, ..., s$. (30) defines an invariance property that seems naturally adapted to the task of detecting dimension reduction subspaces.

3.2 Finding the central subspace

The usefulness of $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ for detecting dimension reduction subspaces is extracted from the following condition, taken from appendix A:

(C3) Given the orthogonal $p \times p$ matrix $\mathbf{A}$ of transformation (7), for a given integer value $1 \leq s \leq p$, $Col(\mathbf{A}_s)$ is a dimension reduction subspace if, and only if, $J_s(\mathbf{A}) = D_s(\mathbf{A}) = 0$, or, according to representations (22) and (27), if, and only if,

$$H_s(\mathbf{A}_s) = H \quad \text{or} \quad C_s(\mathbf{A}_s) = C \ . \tag{31}$$

Phrased differently, and recalling the interpretation given in subsection 3.1 of $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ as separation indexes, $Col(\mathbf{A}_s)$ is a dimension reduction subspace according to (31) if, and only if, the degree of separation among densities after projecting onto $Col(\mathbf{A}_s)$ is the same than in the original formulation of the problem. Condition *(C3)* can be exploited to develop a search criterion for dimension reduction

subspaces of a predetermined dimension $1 \leq s \leq p$. Under adequate regularity conditions given in appendix A, once the location vector $\boldsymbol{\mu}$ and within dispersion matrix $\boldsymbol{\Sigma}$ have been fixed in transformation (7), $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ are continuous functions of argument $\mathbf{A}_s = (\mathbf{a}_1, ..., \mathbf{a}_s)$ in the *Stiefel manifold* $V_{s,p}$ of orders s and p (Muirhead 1982, p. 67), defined as the space of all suborthogonal $p \times s$ matrices $\mathbf{A}_s$ such that $\mathbf{A}_s'\mathbf{A}_s = \mathbf{I}_s$, where $\mathbf{I}_s$ is the identity of order s. The criterion below follows:

(C4) For a given integer $1 \leq s \leq p$ there exists a dimension reduction subspace L_s of dimension s, if, and only if,

$$H_s = \max_{\mathbf{A}_s \in V_{s,p}} H_s(\mathbf{A}_s) = H \quad \text{or} \quad C_s = \max_{\mathbf{A}_s \in V_{s,p}} C_s(\mathbf{A}_s) = C \ . \tag{32}$$

Moreover, $L_s = Col(\mathbf{A}_{s,0})$, where $\mathbf{A}_{s,0}$ is any optimizer of either $H_s(\mathbf{A}_s)$ or $C_s(\mathbf{A}_s)$. The "if" part follows from *(C.3)* and continuity of $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ in the compact set $V_{s,p}$. The "only if" part from (31) and $H_s(\mathbf{A}_s) \leq H$ or $C_s(\mathbf{A}_s) \leq C$.

Criteria (32) above transform then the task of detecting dimension reduction subspaces of dimension s by making the conditional densities $f_{\mathbf{y},i}(y_{s+1}, .., y_p \,|y_1, ..., y_s)$ identical for $i = 1, ..., k$, into the more accessible task of detection by separating, in as much as possible according to the indexes H_s and C_s of (32), the marginals $f_{\mathbf{y},i}(y_1, ..., y_s)$ of the projected coordinates $(y_1, ..., y_s)' = \mathbf{A}_s'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. Maximization criteria (32) lead to the following characterization of $L_0 = Col(\mathbf{A}_{r,0})$:

(C5) The dimension r of L_0 is the first integer $1 \leq r \leq p$ such that

$$H_r = H_{r+1} = \ ... \ = H_p = H \quad \text{or} \quad C_r = C_{r+1} = \ ... \ = C_p = C \ . \tag{33}$$

Moreover, $L_0 = Col(\mathbf{A}_{r,0})$, where $\mathbf{A}_{r,0}$ is given, up to an orthogonal rotation, by

$$\mathbf{A}_{r,0} = \arg \max_{\mathbf{A}_r \in V_{r,p}} H_r(\mathbf{A}_r) = \arg \max_{\mathbf{A}_r \in V_{r,p}} C_r(\mathbf{A}_r) \ . \tag{34}$$

According to *(C5)* L_0 is then the "smallest" subspace for reaching, after projecting onto L_0, the maximum possible degree of separation among densities.

4. EFFECTIVE DIMENSION REDUCTION

Criteria (33)-(34) suggest that the dimension and specific shape of the central subspace $L_0 = Col(\mathbf{A}_{r,0})$ could be determined after conducting sequentially for $s = 1$, 2, ... , maximization processes over $V_{s,p}$ of the functionals $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ until observing stability of the corresponding optima. These processes are not directly feasible since, according to their definitions (23) and (28), $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ depend both on the unknown prior probabilities π_i and on the unknown class conditional densities $f_{\mathbf{y},i}(y_1, .., y_s)$ of $(y_1, ..., y_s)' = \mathbf{A}_s'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. If a database $\mathbf{D}_n$ of individuals previously classified is available, a possible solution is to replace $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ by sample based objective functions $\widehat{H}_s(\mathbf{A}_s)$ and $\widehat{C}_s(\mathbf{A}_s)$, constructed in such way that their sequential optimization for $s = 1, 2, ...,$ is "informative" on the optimization of their theoretical counterparts. In what follows, it is convenient to write the database, rather than in the original notation of expression (2) in section 1, as $\mathbf{D}_n = \{\mathbf{x}_{ij} : i = 1, ..., k\ ,\ j = 1, ..., n_i\}$, where $\mathbf{x}_{ij}$ is the jth individual in class g_i and, for $i = 1, ..., k$, n_i is the total number of individuals in group g_i.

4.1 Objective functions

Proceeding in order, the prior probabilities π_i are estimated by the proportions $\widehat{\pi}_i = n_i/n$. If, on the other hand, the matrix of directions $\mathbf{A}_s$, the location vector $\boldsymbol{\mu}$, and the within group covariance $\boldsymbol{\Sigma} = \sum_{i=1}^{k} \pi_i \boldsymbol{\Sigma}_i$ were known, $f_{\mathbf{y},i}(y_1, .., y_s)$ could be approximated by a nonparametric density "estimator" computed from the "data" $\mathbf{A}_s'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_{ij} - \boldsymbol{\mu})$, $j = 1, ..., n_i$. Since the goal is to construct sample based objective functions depending solely on $\mathbf{A}_s$, $\boldsymbol{\mu}$ is estimated by the sample mean $\overline{\mathbf{x}} = \sum_{i=1}^{k}(n_i/n)\overline{\mathbf{x}}_i$, where $\overline{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$ is the ith class centroid, and $\boldsymbol{\Sigma}$ by the weigthed covariance matrix $\widehat{\boldsymbol{\Sigma}} = \sum_{i=1}^{k}(n_i/n)\widehat{\boldsymbol{\Sigma}}_i$, where $\widehat{\boldsymbol{\Sigma}}_i = (1/n_i)\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)'$.

For $i = 1, \ldots\ k$, $f_{\mathbf{y},i}(y_1, .., y_s)$ may be then approximated by the function

$$\widehat{f}_{\mathbf{y},i}(y_1, ..., y_s;\mathbf{A}_s) = \frac{1}{n_i h_i^s} \sum_{j=1}^{n_i} K_i\{\frac{1}{h_i}[\begin{pmatrix} y_1 \\ \vdots \\ y_s \end{pmatrix} - \mathbf{A}_s' \widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}_{ij} - \overline{\mathbf{x}})]\} \ , \qquad (35)$$

where

$$K_i(\mathbf{z}) = (2\pi)^{-(s/2)} \left|\widehat{\mathbf{V}}_i\right|^{-1/2} \exp(-\mathbf{z}'\widehat{\mathbf{V}}_i^{-1}\mathbf{z}/2) \ , \quad \mathbf{z} \in \mathbb{R}^s \, , \qquad (36)$$

is an $s-$variate gaussian kernel with $s \times s$ dispersion matrix

$$\widehat{\mathbf{V}}_i = \mathbf{A}_s' \widehat{\mathbf{Q}}_i \mathbf{A}_s \, , \qquad (37)$$

where $\widehat{\mathbf{Q}}_i = \widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{\Sigma}}_i\widehat{\mathbf{\Sigma}}^{-1/2}$, and h_i is an smoothing parameter. Once that for $i = 1$, ..., k each $f_{\mathbf{y},i}(y_1, ..., y_s)$ has been approximated by the corresponding function in (35), the density $f_{\mathbf{y}}(y_1, ..., y_s) = \sum_{i=1}^k \pi_i f_{\mathbf{y},i}(y_1, ..., y_s)$ is approximated by the mixture $\widehat{f}_{\mathbf{y}}(y_1, ..., y_s;\mathbf{A}_s) = \sum_{i=1}^k (n_i/n)\widehat{f}_{\mathbf{y},i}(y_1, ..., y_s;\mathbf{A}_s)$.

Construction (35) is similar to the usual structure of a nonparametric kernel density estimator, as studied for example in Scott (1992, chap. 6). Other alternative kernel functions could be considered, but a gaussian kernel seems to be convenient for computing work. $\widehat{\mathbf{V}}_i$ in (37) is the dispersion matrix of the "data" $\mathbf{A}_s'\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}_{ij} - \overline{\mathbf{x}})$, $j = 1, ..., n_i$, so, according to Silverman (1986, pp. 77 and 78), a kernel of the form (36) justifies using, as in (35), a single smoothing parameter instead of one for each of the s dimensions in $(y_1, .., y_s)'$. In principle, h_i should depend on the scale of the "data" $\mathbf{A}_s'\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}_{ij} - \overline{\mathbf{x}})$, $j = 1, ..., n_i$, and therefore on the specific form of the matrix of directions $\mathbf{A}_s$. However, and again to simplify numerical matters, h_i is selected in an automatic form using a proposal in Silverman (1986, p. 87), specifically

$$h_i = \left(\frac{4}{n_i(s+2)}\right)^{1/(s+4)} \, . \qquad (38)$$

Replacing now in definitions (23) of $H_s(\mathbf{A}_s)$ and (28) of $C_s(\mathbf{A}_s)$, unknown elements by estimations or approximations and, at the same time, expectations by averages

over the "data" $\widehat{\mathbf{z}}_{lj} = \mathbf{A}_s' \widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x}_{lj} - \overline{\mathbf{x}})$, $l = 1, ..., k$, $j = 1, ..., n_l$, the following objective functions are obtained after some routine algebra. On one hand,

$$\widehat{H}_s(\mathbf{A}_s) =$$

$$= \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) [\frac{1}{n} \sum_{l=1}^{k} \sum_{j=1}^{n_l} - \log \left(\frac{(n_i/n) \widehat{f}_{\mathbf{y},i}(\widehat{\mathbf{z}}_{lj}; \mathbf{A}_s)}{\widehat{f}_{\mathbf{y}}(\widehat{\mathbf{z}}_{lj}; \mathbf{A}_s)} \right)] - \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) [- \log \left(\frac{n_i}{n}\right)] \ , \quad (39)$$

is the *sample global entropy* or *entropy criterion* and, on the other,

$$\widehat{C}_s(\mathbf{A}_s) = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 [\frac{1}{n} \sum_{l=1}^{k} \sum_{j=1}^{n_l} \left(\frac{\widehat{f}_{\mathbf{y},i}(\widehat{\mathbf{z}}_{lj}; \mathbf{A}_s)}{\widehat{f}_{\mathbf{y}}(\widehat{\mathbf{z}}_{ij}; \mathbf{A}_s)} \right)^2] - \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 \ , \quad (40)$$

is the *sample total prediction capacity* or *trace criterion*. The second summands in (39) and (40) can be ignored for maximization purposes. From (35), (36) and (37), the two objective functions above depend on $\mathbf{A}_s$ only through $\left|\mathbf{A}_s' \widehat{\mathbf{Q}}_i \mathbf{A}_s\right|$ and $\mathbf{A}_s(\mathbf{A}_s' \widehat{\mathbf{Q}}_i \mathbf{A}_s)^{-1} \mathbf{A}_s'$. Therefore $\widehat{H}_s(\mathbf{A}_s)$ and $\widehat{C}_s(\mathbf{A}_s)$ satisfy, for every orthogonal matrix $\mathbf{F}$ of $s \times s$ and $\mathbf{A}_s$ in $V_{s,p}$, an orbit invariance property of the form

$$\widehat{H}_s(\mathbf{A}_s) = \widehat{H}_s(\mathbf{A}_s \mathbf{F}) \quad , \quad \widehat{C}_s(\mathbf{A}_s) = \widehat{C}_s(\mathbf{A}_s \mathbf{F}) \ , \quad (41)$$

that is similar to property (30) satisfied by $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$.

As established in appendix A, under adequate regularity conditions on the posterior probability functions $\eta_i(y_1, ..., y_s) = P[\mathbf{g} = i \,|y_1, ..., y_s]$, $i = 1, ..., k$, for every $1 \leq s \leq p$, $\widehat{H}_s(\mathbf{A}_s)$ and $\widehat{C}_s(\mathbf{A}_s)$ are such that

$$\widehat{H}_s = \max_{\mathbf{A}_s \in V_{s,p}} \widehat{H}_s(\mathbf{A}_s) \cong H_s \quad , \quad \widehat{C}_s = \max_{\mathbf{A}_s \in V_{s,p}} \widehat{C}_s(\mathbf{A}_s) \cong C_s \ , \quad (42)$$

where, as in (32), $H_s = \max_{\mathbf{A}_s \in V_{s,p}} H_s(\mathbf{A}_s)$ and $C_s = \max_{\mathbf{A}_s \in V_{s,p}} C_s(\mathbf{A}_s)$. Approximations in (42) are in some asymptotic sense explained in the appendix. Sequential optimization of the objective functions (39) and (40) is then, as intended, "equivalent" to sequential optimization of their theoretical versions (23) and (28).

4.2 Steps in dimension reduction

A two step procedure can be now introduced for effective dimension reduction in KDA using the information obtained from sequential optimization of (39) and (40):

Step 1: *Estimation of the central subspace* $L_0 = Col(\mathbf{A}_{r,0})$. The dimension of L_0 is declared as the value of the first integer r such that stability of the optima $\widehat{H}_s$ and $\widehat{C}_s$ is observed for $s \geq r$. From approximations (42), this stability can be interpreted as empirical evidence of an structure of the form (33), that is either $H_s = H$ or $C_s = C$ for $s \geq r$. As seen in next section, this procedure, of essentially exploratory nature, has a good behaviour in applications. Once a decision on the value r of $\dim(L_0)$ has been adopted, $\mathbf{A}_{r,0}$ is estimated by any optimizer $\widehat{\mathbf{A}}_{r,0}$ of $\widehat{H}_r(\mathbf{A}_r)$ or $\widehat{C}_r(\mathbf{A}_r)$;

Step 2. Construction of a reduced KDA *rule* ($RKDA$). After estimation of $L_0 = Col(\mathbf{A}_{r,0})$, the next natural stage is to construct a sample based version of the reduced optimal rule defined by paradigm (18)-(19) in section 2. To do that, coordinates (18) are estimated by the sample canonical coordinates

$$\begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_r \end{pmatrix} = \widehat{\mathbf{A}}'_{r,0}\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}}) \ , \tag{43}$$

and the prior probabilities π_i by the proportions $\widehat{\pi}_i = n_i/n$. Also, for $i = 1, ..., k$, the unknown $f_{\mathbf{y},i}(y_1, ..., y_r)$ is approximated by the corresponding function in (35). Replacing in $\widehat{f}_{\mathbf{y},i}(y_1, ..., y_r; \mathbf{A}_{r,0})$ coordinates $(y_1, ..., y_r)'$ by coordinates $(\widehat{y}_1, ..., \widehat{y}_r)'$ of (43) and matrix $\mathbf{A}_{r,0}$ by optimizer $\widehat{\mathbf{A}}_{r,0}$ leads finally to the following sample rule: assign $\mathbf{x}$ to group g_i when

$$(n_i/n)\widehat{f}_{\mathbf{y},i}(\widehat{y}_1, ..., \widehat{y}_r; \widehat{\mathbf{A}}_{r,0}) = \max_{1 \leq j \leq k}(n_j/n)\widehat{f}_{\mathbf{y},j}(\widehat{y}_1, ..., \widehat{y}_r; \widehat{\mathbf{A}}_{r,0}) \ . \tag{44}$$

4.3 Additional considerations

For a given integer $1 \leq s \leq p$, put $\widehat{\mathbf{A}}_{s,0}$ for any optimizer of the objective functions $\widehat{H}_s(\mathbf{A}_s)$ or $\widehat{C}_s(\mathbf{A}_s)$. If in (35) $\mathbf{A}_s$ is replaced by $\widehat{\mathbf{A}}_{s,0}$, the function $\widehat{f}_{\mathbf{y},i}(y_1, ..., y_s; \widehat{\mathbf{A}}_{s,0})$ is a first estimate of $f_{\mathbf{y},i}(y_1, ..., y_s)$, the ith class conditional density of the projected coordinates $(y_1, ..., y_s)' = \mathbf{A}_s' \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. The maximization problems defined in (42), $\widehat{H}_s = \max_{\mathbf{A}_s \in V_{s,p}} \widehat{H}_s(\mathbf{A}_s)$ and $\widehat{C}_s = \max_{\mathbf{A}_s \in V_{s,p}} \widehat{C}_s(\mathbf{A}_s)$, can be then interpreted as search procedures to separate estimators $\widehat{f}_{\mathbf{y},i}(y_1, ..., y_s; \widehat{\mathbf{A}}_{s,0})$ in as much as possible after projecting onto $Col(\widehat{\mathbf{A}}_{s,0})$. This is in the same spirit than $s-$ dimensional projection pursuit methods, as studied in Huber (1985, sec. 7).

Once the dimension $r = \dim(L_0)$ has been determined, rule $RKDA$ (43)-(44) is the result of a three stage process: *i) separate* estimated densities using sequential application of (42) to get $\widehat{\mathbf{A}}_{r,0}$; *ii) project* onto $Col(\widehat{\mathbf{A}}_{r,0})$ to determine coordinates $(\widehat{y}_1, ..., \widehat{y}_r)' = \widehat{\mathbf{A}}_{r,0}' \widehat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}})$ in (43), and *iii) classify* according to the sample based criterion (44). This is a nonparametric extension of the classical ideas of linear discriminant analysis (LDA) where, as described for example in Johnson and Wichern (1998, chap. 11), assuming that the underlying class conditional densities are approximately multivariate normal with the same dispersion matrix, classification (stage *iii))* is performed after projection (stage *ii))* onto an adequately chosen subspace in which the class centroids $\overline{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i$, $i = 1, ..., k$, reach their maximum possible separation (stage *i)).*

According to Gill, Murray and Wright (1981, chap. 6), given the nonlinear nature of both objective functions and restrictions $\mathbf{A}_s'\mathbf{A}_s = \mathbf{I}_s$, a sequential quadratic programming (SQP) algorithm is appropriate to solve the numerical problems (42). In experience of the authors, the optima of $\widehat{H}_s(\mathbf{A}_s)$ and $\widehat{C}_s(\mathbf{A}_s)$ offer exploratory evidence of stability after an integer r typically smaller than the original dimension p. In the spirit of circumventing the curse of dimensionality in KDA, the range of integers

s in which both the entropy and trace criteria need to be effectively evaluated can be then restricted to some subset $1 \leq s \leq s_0$, where $r < s_0 \ll p$.

5. APPLICATIONS

The dimension reduction algorithm of the previous section is now applied to three classification problems with continuous data $\mathbf{x} = (x_1, ..., x_p)'$ and moderately large dimension p. In all the examples that follow, optimization of the objective functions $\widehat{H}_s(\mathbf{A}_s)$ and $\widehat{C}_s(\mathbf{A}_s)$ is performed using the SQP algorithm $NPSOL$, released by the System Optimization Laboratory of Stanford University. For details, see Gill, Murray, Saunders and Wright (1986). Analysis of the values of the optima $\widehat{H}_s$ and $\widehat{C}_s$ is complemented, for each integer $s = 1, 2, ...$, with an adequate estimation of L_n, the conditional probability of error of the sample rule that assigns $(\widehat{y}_1, ..., \widehat{y}_s)' = \widehat{\mathbf{A}}'_{s,0}\widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{x} - \overline{\mathbf{x}})$ to the ith group when

$$(n_i/n)\widehat{f}_{\mathbf{y},i}(\widehat{y}_1, ..., \widehat{y}_s; \widehat{\mathbf{A}}_{s,0}) = \max_{1 \leq j \leq k}(n_j/n)\widehat{f}_{\mathbf{y},j}(\widehat{y}_1, ..., \widehat{y}_s; \widehat{\mathbf{A}}_{s,0}) \ , \tag{45}$$

where $\widehat{f}_{\mathbf{y},i}(\widehat{y}_1, ..., \widehat{y}_s; \widehat{\mathbf{A}}_{s,0})$ is constructed in an obvious manner from the function $\widehat{f}_{\mathbf{y},i}(y_1, ..., y_s; \mathbf{A}_{s,0})$ of (35). Clearly, when $s = r = \dim(L_0)$, rule (45) above reduces to the $RKDA$ rule (43)-(44). Recall that, as for example in Devroye et al. (1996, chap. 1), L_n is defined as the random variable

$$L_n = 1 - \sum_{i=1}^{k} P[\mathbf{x} \in \widehat{U}_i \, ; \, \mathbf{g} = i \, | \mathbf{D}_n] = 1 - \sum_{i=1}^{k} \pi_i \int_{\widehat{U}_i} f_i(\mathbf{x}) d\mathbf{x} \ , \tag{46}$$

where the pair ($\mathbf{x}$,$\mathbf{g}$) is independent from the database $\mathbf{D}_n$ and $\widehat{U}_i = \{\mathbf{x} \in \mathbb{R}^p : (n_i/n)\widehat{f}_{\mathbf{y},i}(\widehat{y}_1, ..., \widehat{y}_s; \widehat{\mathbf{A}}_{s,0}) = \max_{1 \leq j \leq k}(n_j/n)\widehat{f}_{\mathbf{y},j}(\widehat{y}_1, ..., \widehat{y}_s; \widehat{\mathbf{A}}_{s,0})\}$.

5.1 Wisconsin breast cancer data

The problem in this example, as presented in Lim, Lo and Shih (2000), is to decide whether a tissue sample of nine measurements $\mathbf{x} = (x_1, ..., x_9)'$ obtained from a

patient's breast, is either g_1 : *benign* or g_2 : *malignant*. Using information from Merz and Murphy (1996), a database $\mathbf{D}_n$ of size $n = 683$ is available with $n_1 = 444$ observations in group g_1 and $n_2 = 239$ observations in group g_2. For completeness, in each of the subtables of table 1 the first column contains the values of the optima $\widehat{H}_s$ and $\widehat{C}_s$ for the full range of integers $1 \leq s \leq 9$. The second column gives for each s the value of $L_n^{(R)}$, the usual apparent or resubstitution error rate estimator of the conditional error L_n in (46), as defined, for example, in Devroye et al. (1996, chap. 23). The third column contains the estimations of L_n given by the ten-fold cross validation estimator $\widehat{L}_{n,10}$ of Breiman, Friedman, Olsen and Stone (1984, p. 85). The fourth column reports standard errors of $\widehat{L}_{n,10}$.

Table 1

According to the results in table 1, the optima of the trace criterion seem to be stable starting from $s = 1$, while the optima of the entropy criterion are clearly stable from $s = 2$. An explanation for this apparent contradiction is given by figures 1 and 2 below, that display the functions $\widehat{f}_{\mathbf{y},i}(y_1, ..., y_s;\widehat{\mathbf{A}}_{s,0})$, $i = 1, 2$, for respectively $s = 1$ and $s = 2$. According to figure 1, the one dimensional functions $\widehat{f}_{\mathbf{y},i}(y_1;\widehat{\mathbf{A}}_{1,0})$ are quite separated for both the entropy and trace criterions but still with a common part, due mainly to the pronounced skewness to the right of $\widehat{f}_{\mathbf{y},1}(y_1;\widehat{\mathbf{A}}_{1,0})$. As seen in figure 2, the functions $\widehat{f}_{\mathbf{y},1}(y_1,y_2;\widehat{\mathbf{A}}_{2,0})$ and $\widehat{f}_{\mathbf{y},2}(y_1,y_2;\widehat{\mathbf{A}}_{2,0})$ are almost completely separated. The minor extra degree of separation in the (y_1,y_2) plane forces only a slight change in the optima of the trace criterion from $\widehat{C}_1 = .9609$ to $\widehat{C}_2 = .9740$, and a relatively larger change in the optima of the entropy criterion from $\widehat{H}_1 = 3.8217$ to $\widehat{H}_2 = 5.1562$. This should be expected since, in the format (39) and (40), both objective functions depend on the ratios $0 < (n_i/n)\widehat{f}_{\mathbf{y},i}(\widehat{\mathbf{z}}_{ij};\mathbf{A}_s)/\widehat{f}_{\mathbf{y}}(\widehat{\mathbf{z}}_{lj};\mathbf{A}_s) < 1$, the trace criterion measuring separation in a quadratic scale and the entropy criterion in the $-\log$ scale. As a conclusion of this analysis, the dimension of the central subspace

is declared as $r = 2$, the second dimension offering only a marginal contribution for density separation purposes.

Figure 1

Figure 2

Turning now to error estimation, the values of $L_n^{(R)}$ in table 1 tend to cero as s increases. This confirms the well-known optimistically biased character of $L_n^{(R)}$ as an estimator of L_n. The message from the more reliable estimator $\widehat{L}_{n,10}$ in the third column is different. For both criteria, the estimated error rates for $s = 1$ and $s = r = 2$ are substantially smaller than the values of $\widehat{L}_{n,10}$ for the rules of the form (45) that include spurious directions $s = 3, 4, \ldots$. In particular observe, for $s = p = 9$, the poor behaviour of the KDA rule. This is an empirical justification of dimension reduction methods as tools for a more efficient use of the available sample information.

Notice that, in this example, the sample rule (45) for $s = 1$ is slightly better than the $RKDA$ rule that corresponds to taking $s = r = 2$ in (45). This is not surprising given the marginal separation of densities provided by the second direction. Finally, it is interesting to analyze the relative performances of these two rules as compared with other classification methods applied to this data set. The two rules derived from the trace criterion for $s = 1$ and $s = r = 2$ improve over the "best" classification method out of the thirty three studied in the comparative study of Lim et al. (2000), a neural net with estimated error $\widehat{L}_{n,10} = .0278$. The entropy rule (45) for $s = 1$ is immediately after this neural net. However, all these comparisons are, rather than among the true unknown conditional errors L_n, among estimators $\widehat{L}_{n,10}$ of relatively imprecise sampling distributions, as indicated by the large standard errors in the fourth columns of table 1, and should be therefore taken with some degree of caution.

5.2 Wave form data

This is an artificial classification problem with $k = 3$ groups and $p = 21$ variables, introduced by Breiman et al. (1984, p. 49). The probabilistic structure of $\mathbf{x}$ is:

$$g_1 \quad : \quad x_i = uh_1(i) + (1-u)h_2(i) + \varepsilon_i \; ; \tag{47}$$

$$g_2 \quad : \quad x_i = uh_1(i) + (1-u)h_3(i) + \varepsilon_i \; ; \tag{48}$$

$$g_3 \quad : \quad x_i = uh_2(i) + (1-u)h_3(i) + \varepsilon_i \; , \tag{49}$$

where, for $i = 1, ..., 21$, $u \sim U(0,1)$ and $\varepsilon_1, ..., \varepsilon_{21} \sim N(0,1)$ independently of u. In equations (47), (48) and (49), the $h_i(.)$ are the shifted wave form functions $h_1(i) = \max(6 - |i-11|,0)$, $h_2(i) = h_1(i-4)$ and $h_3(i) = h_1(i+4)$. The priors are all set to $\pi_i = 1/3$. This relatively complex example has been used as a test case of classification methods for several authors, among others Michie, Spiegelhalter and Taylor (1994), and Hastie and Tibshirani (1996). Breiman et al. (1984, p. 84) estimate the Bayes error as $L^* = .1400$. As in Lim et al. (2000), simulated "training" and "testing" databases $\mathbf{D}_n$ and $\mathbf{T}_m$ of sizes $n = 600$ and $m = 3000$ are available.

Table 2

Table 2 displays, for the range $1 \leq s \leq 7$, the results after sequential optimization of $\widehat{H}_s(\mathbf{A}_s)$ and $\widehat{C}_s(\mathbf{A}_s)$. Clearly, for both the trace and entropy criterions, the decision on the dimension of the central subspace L_0 is $r = 2$. This is confirmed by figure 3 that shows a poor degree of separation between the one dimensional functions $\widehat{f}_{\mathbf{y},i}(y_1;\widehat{\mathbf{A}}_{1,0})$, $i = 1, 2, 3$, improved in figure 4 for the $\widehat{f}_{\mathbf{y},i}(y_1,y_2;\widehat{\mathbf{A}}_{2,0})$, $i = 1, 2, 3$.

Figure 3

Figure 4

As in the previous example, the resubstitution error estimates $L_n^{(R)}$ do not offer reliable information. The adequate error estimator in this example is the so called

error-counting $\widehat{L}_{n,m}$ (Devroye et al. 1996, chap. 8), defined as the relative frequency of errors on the testing database $\mathbf{T}_m$ of a rule of the form (45) constructed using the training database $\mathbf{D}_n$. According to the values of $\widehat{L}_{n,m}$ in the third columns of table 2, the best rule of the form (45) is the $RKDA$ rule corresponding to $s = r = 2$. Again, the error rates deteriorate with the inclusion of spurious directions $s = 3, 4, \ldots$.

As with respect the relative performance of the $RKDA$ rule in this example as compared with other classification methods, the reduced rules corresponding to, respectively, the entropy and the trace criterions occupy the third and fourth positions in the ranking of methods reported by Lim et al. (2000). The "best" method is a neural net, with estimated error $\widehat{L}_{n,m} = .151$. This shows that, despite its rather sophisticated non linear simulation mechanism (47)-(48)-(49), the information needed in this example for nearly optimal classification is essentially captured by a projection onto a linear subspace of dimension $r = 2$. See also the comments in Hastie and Tibshirani (1996, sec. 4) on the intrinsic "bidimensional structure" of this classification problem.

5.3 Image segmentation data

This is a classification problem analyzed in the STATLOG project reported by Michie et al. (1994). The samples are from a database of outdoor images that are hand-segmented to create a classification in $k = 7$ classes that can be either $g_1 : brickface$, $g_2 : sky$, $g_3 : foliage$, $g_4 : cement$, $g_5 : window$, $g_6 : path$ or $g_7 : grass$. The feature vector $\mathbf{x}$ contains $p = 18$ continuous variables. This is a rather numerically complex example and, for simplicity, only the results of application of the trace criterion for the range of integers $1 \leq s \leq 5$ are reported in table 3 below. Classification errors are estimated using either the resubstitution estimator $L_n^{(R)}$ or the ten-fold cross-validation estimator $\widehat{L}_{n,10}$.

Table 3

Given the limited range of dimensions analyzed in this example, it is somewhat premature to get a definite conclusion on the dimension on the central subspace L_0, although $s = 3$ and $s = 4$ appear as strong candidates. Taking again as a reference the study in Lim et al. (2000), the reduced rules (45) for these two integers have a good behaviour. The best rule is a *kth* nearest neighbor plug-in rule with $\widehat{L}_{n,10} = .0221$. The rules for $s = 3$ and $s = 4$ occupy, respectively, the third and fourth positions. Figures 5 and 6 display, for $s = 1$ and $s = 2$, the estimated densities $\widehat{f}_{\mathbf{y},i}(y_1, ..., y_s; \widehat{\mathbf{A}}_{s,0})$, $i = 1, ..., 7$. Notice the improvement for differentiating among densities of figure 6 over figure 5, particularly for the case of groups g_3 and g_7.

Figure 5

Figure 6

5.4 Conclusions

In the three examples above, the $RKDA$ rule (43)-(44) is, in applications of non-parametric discriminant analysis, a valuable analytical tool for dimension reduction. The methodology has the advantage of *visualization* since the process of sequential separation of densities can be monitored graphically using, for $s = 1, 2, ...$, plots of the functions $\widehat{f}_{\mathbf{y},i}(y_1, ..., y_s; \widehat{\mathbf{A}}_{s,0})$, $i = 1, ..., k$. These are natural nonparametric extensions of the plots of canonical coordinates in standard LDA, that are used to calibrate separation among projected class centroids. See, e.g. Flury (1997, chap. 7).

As for comparison of criteria, the trace criterion (40) is by construction numerically more tractable than the entropy criterion (39), due to the unbounded character of the logarithm in regions of low density. The trace criterion is then more manageable, although the entropy criterion can offer, as in the Wisconsin breast cancer data, additional information on the dimension of the central subspace.

Finally, a natural question to be asked is whether a computationally simpler *one*

direction at a time optimization strategy, would be perhaps preferable to joint optimization as suggested in (42). For example, if $\widehat{\mathbf{a}}_1 = \arg\max_{\|\mathbf{a}\|=1} \widehat{C}_1(\mathbf{a})$, one has

$$\max_{\|\mathbf{b}\|=1,\ \widehat{\mathbf{a}}_1'\mathbf{b}=0} \widehat{C}_2(\widehat{\mathbf{a}}_1,\mathbf{b}) \leq \max_{(\mathbf{a},\ \mathbf{b})\in V_{2,p}} \widehat{C}_2(\mathbf{a},\mathbf{b}) = \widehat{C}_2 \ , \tag{50}$$

and therefore a one at a time approach defines, in general, a suboptimal search procedure that can lead to misleading conclusions on the value of $\widehat{C}_2$. Joint optimization, not trivial but made feasible by the proper use of an adequate SQP algorithm, is then more recommendable. See also Huber (1985, sec. 7) on the advantages and disadvantages of stepwise versus multidimensional projection pursuit procedures.

6. FINAL COMMENTS

Dimension reduction to avoid the curse of dimensionality in KDA is an standard applied problem. See the review in McLachlan (1992, sec. 12.4.5, p. 405) of earlier work in this area. Recently, Zhu (2001) and Hastie and Zhu (2001) have presented a method for feature extraction in nonparametric discriminant analysis. Their approach is inspired on the ideas of projection pursuit density estimation, developed by Friedman, Stuetzle and Schroeder (1984). This paper offers an alternative approach for dimension reduction in KDA, exploiting the concept of central subspace in the context of a classification problem defined by the pairs $(\pi_i,f_i(\mathbf{x}))$, $i = 1, ..., k$. Dimension reduction in discriminant analysis by subspaces was already used in Flury, Boukai and Flury (1997), who introduced the notion of *discrimination subspace model* for separating two normal populations. For related work and applications of the concept of dimension reduction subspace in classification, see Cook and Yin (2001) and Hastie and Zhu (2001).

APPENDIX A: DIMENSION REDUCTION FUNCTIONALS

For (22), integrate in the decomposition below

$$\log\left[\frac{f_{\mathbf{y}}(\mathbf{y})}{f_{\mathbf{y},i}(\mathbf{y})}\right] f_y(\mathbf{y}) = \log\left[\frac{f_{\mathbf{y}}(y_1,\, ...,\, y_s)}{f_{\mathbf{y},i}(y_1,\, ...,\, y_s)}\right] f_y(\mathbf{y}) +$$

$$+ \log\left[\frac{f_{\mathbf{y}}(y_{s+1}, ..., y_p\,|y_1,\, ...,\, y_s)}{f_{\mathbf{y},i}(y_{s+1}, ..., y_p\,|y_1,\, ...,\, y_s)}\right] f_{\mathbf{y}}(y_1,\, ...,\, y_s) f_{\mathbf{y}}(y_{s+1}, ..., y_p\,|y_1,\, ...,\, y_s)\,. \qquad (51)$$

(27) follows from uncorrelation of the $k \times 1$ random vectors $G_p(\mathbf{y})$ and $G_p(\mathbf{y}) - G_s(\mathbf{y})$, as it can be checked using the law of iterated expectations as, for example, in Billingsley (1995, chap. 6). Using properties of the relative entropy, $J_s(\mathbf{A}) = 0$ if, and only if, all conditionals $f_{\mathbf{y},i}(y_{s+1}, ..., y_p\,|y_1,\, ...,\, y_s)$ are identical. If $D_s(\mathbf{A}) = 0$, this is because $G_p(\mathbf{y}) - G_s(\mathbf{y}) = \mathbf{0}$ or, in other words, because, for $i = 1,\, ...,\, k$, $q_i(\mathbf{y}) = P[\mathbf{g} = i\,|\mathbf{y}] = \eta_i(y_1, ..., y_s) = P[\mathbf{g} = i\,|y_1,\, ...,\, y_s]$. From (8), (10) and (11) this leads again to $f_{\mathbf{y},i}(y_{s+1}, ..., y_p\,|y_1,\, ...,\, y_s) = f_{\mathbf{y}}(y_{s+1}, ..., y_p\,|y_1,\, ...,\, y_s)$, for $i = 1,\, ...,\, k$.

Continuity of $H_s(\mathbf{A}_s)$ and $C_s(\mathbf{A}_s)$ in $\mathbf{A}_s \in V_{s,p}$ can be derived from their expressions as functions of $\eta_i(y_1, ..., y_s) = P[\mathbf{g} = i\,|y_1,\, ...,\, y_s]$. After some algebra, one has

$$H_s(\mathbf{A}_s) = \sum_{i=1}^{k} \pi_i E\{-\log \eta_i[\mathbf{A}_s' \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})]\} - \sum_{i=1}^{k} \pi_i[-\log(\pi_i)]\,. \qquad (52)$$

By application of the bounded convergence theorem, $H_s(\mathbf{A}_s)$ is continuous if the $\eta_i(y_1, ..., y_s)$ are uniformly continuous and bounded in the form $0 < c < \eta_i(y_1, ..., y_s) \leq 1$ for some $c > 0$. The case of $C_s(\mathbf{A}_s)$ can be treated similarly.

APPENDIX B: OBJECTIVE FUNCTIONS

For reasons of conciseness, only the case of the entropy criterion is treated. Approximate $\widehat{H}_s(\mathbf{A}_s)$ by the *pseudo objective function*

$$\widehat{H}_s^{ps}(\mathbf{A}_s) =$$

$$= \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) [\frac{1}{n} \sum_{l=1}^{k} \sum_{j=1}^{n_l} - \log \left(\frac{(n_i/n) f_{\mathbf{y},i}(\mathbf{z}_{lj})}{f_{\mathbf{y}}(\mathbf{z}_{lj})} \right)] - \sum_{i=1}^{k} \left(\frac{n_i}{n}\right) [- \log \left(\frac{n_i}{n}\right)] \ , \qquad (53)$$

where $\mathbf{z}_{lj} = \mathbf{A}_s' \boldsymbol{\Sigma}^{-1/2} (\mathbf{x}_{lj} - \boldsymbol{\mu})$, $l = 1, ..., k$, $j = 1, ..., n_l$. The goal is to establish

$$\max_{\mathbf{A}_s \in V_{s,p}} \left| \widehat{H}_s^{ps}(\mathbf{A}_s) - H_s(\mathbf{A}_s) \right| \to 0 \ , \quad a.s. \ , \qquad (54)$$

as $n \to \infty$. From (52) and (53) it suffices to obtain, for $i = 1, ..., k$, the convergence

$$\max_{\mathbf{A}_s \in V_{s,p}} \left| \frac{1}{n} \sum_{l=1}^{k} \sum_{j=1}^{n_l} \{- \log[\eta_i(\mathbf{z}_{lj})]\} - E\{- \log[\eta_i(y_1, ..., y_s)]\} \right| \to 0 \ , \quad a.s. . \qquad (55)$$

To do this, consider, for fixed $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the function $h(\mathbf{A}_s \, , \mathbf{x}) = - \log\{\eta_i[\mathbf{A}_s' \boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu})]\}$ that has two arguments: $\mathbf{A}_s$ in the compact set $V_{s,p}$ and $\mathbf{x}$ in $\mathbb{R}^p$. Using $h(\mathbf{A}_s \, , \mathbf{x})$, (55) can be reexpressed in the form

$$\max_{\mathbf{A}_s \in V_{s,p}} \left| \frac{1}{n} \sum_{l=1}^{k} \sum_{j=1}^{n_l} h(\mathbf{A}_s \, , \mathbf{x}_{lj}) - E[h(\mathbf{A}_s \, , \mathbf{x})] \right| \to 0 \ , \quad a.s. . \qquad (56)$$

Under the regularity condition of appendix A on $\eta_i(y_1, ..., y_s)$, (56) follows from the uniform strong law of the large numbers in Rubin (1956, sec. 1). Recall that (54) leads to $\widehat{H}_s = \max_{\mathbf{A}_s \in V_{s,p}} \widehat{H}_s(\mathbf{A}_s) \cong \max_{\mathbf{A}_s \in V_{s,p}} \widehat{H}_s^{ps}(\mathbf{A}_s) \cong \max_{\mathbf{A}_s \in V_{s,p}} H_s(\mathbf{A}_s) = H_s$.

REFERENCES

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

[2] Breiman, L., Friedman, J. H., Olsen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Bellmont, CA: Wadsworth.

[3] Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: John Wiley.

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

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

[6] Flury, B. (1997), *A First Course in Multivariate Analysis,* New York: Springer.

[7] Flury, L., Boukai, B. and Flury, B. (1997), "The Discrimination Subspace Model," *Journal of the American Statistical Association,* **92**, 758-766.

[8] Friedman, J., Stuetzle, W. and Schroeder, A. (1984), "Projection Pursuit Density Estimation," *Journal of the American Statistical Association,* **79**, 599-608.

[9] Gill, P. E., Murray, W., Saunders, M. A. and Wright, M. H. (1986), "User's guide for $NPSOL$ (version 4.0): A $FORTRAN$ package for linear programming," Technical Report SOL 86-2. Stanford University.

[10] Gill, P. E., Murray, W. and Wright, M. H. (1981), *Practical Optimization*, New York: Academic Press.

[11] Hand, D. J. (1997), *Construction and Assessment of Classification Rules*, New York: John Wiley.

[12] Hastie, T. and Tibshirani, R. (1996), "Discriminant Analysis by Gaussian Mixtures," *Journal of the Royal Statistical Society, Series B,* **58**, 155-176.

[13] Hastie, T. and Zhu, M. (2001), Discussion of Cook and Yin (2001), *Australian and New Zealand Journal of Statistics,* **43**, 179-185.

[14] Huber, P. J. (1985), "Projection pursuit (with discussion)," *The Annals of Statistics*, **13**, 435-475.

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

[16] Lim, T. S. Lo, W. Y. and Shih, Y. S. (2000), "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms," *Machine Learning*, **40**, 203-228.

[17] Merz, C. J. and Murphy, P. M. (1996), *UCI Repository of Machine Learning Databases*, University of California, Irvine, CA: Dept. of Information and Computer Science.

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

[19] Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994), *Machine Learning, Neural and Statistical Classification*, London: Chapman and Hall.

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

[21] Rubin, H. (1956), "Uniform Convergence of Random Functions with Applications to Statistics," *The Annals of Mathematical Statistics*, **27**, 200-203.

[22] Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, New York: John Wiley.

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

[24] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

[25] Zhu, M. (2001), "Feature Extraction and Dimension Reduction with Applications to Classification and the Analysis of Co-occurrence Data," PhD Dissertation, Stanford University.

FIGURES AND TABLES

	Trace Criterion				Entropy Criterion			
s	$\widehat{C}_s$	$L_n^{(R)}$	$\widehat{L}_{n,10}$	$s.\ e.$	$\widehat{H}_s$	$L_n^{(R)}$	$\widehat{L}_{n,10}$	$s.\ e.$
1	.9609	.0249	.0263	.0156	3.8217	.0278	.0306	.0219
2	.9740	.0132	.0277	.0152	5.1562	.0190	.0350	.0217
3	.9894	.0029	.0453	.0165	5.6754	.0059	.0407	.0228
4	.9978	.0044	.0468	.0215	5.8736	.0015	.0467	.0222
5	.9997	.0015	.0423	.0186	5.8850	.0000	.0467	.0250
6	.9997	.0029	.0467	.0202	5.8576	.0000	.0453	.0247
7	.9999	.0000	.0497	.0185	5.7738	.0000	.0408	.0201
8	.9995	.0015	.0423	.0228	5.5753	.0000	.0409	.0242
9	.9983	.0000	.0409	.0233	5.4777	.0000	.0409	.0233

Table 1. Wisconsin Breast Cancer Data: Sequential optimization of the trace and entropy criterions and estimation of error rates.

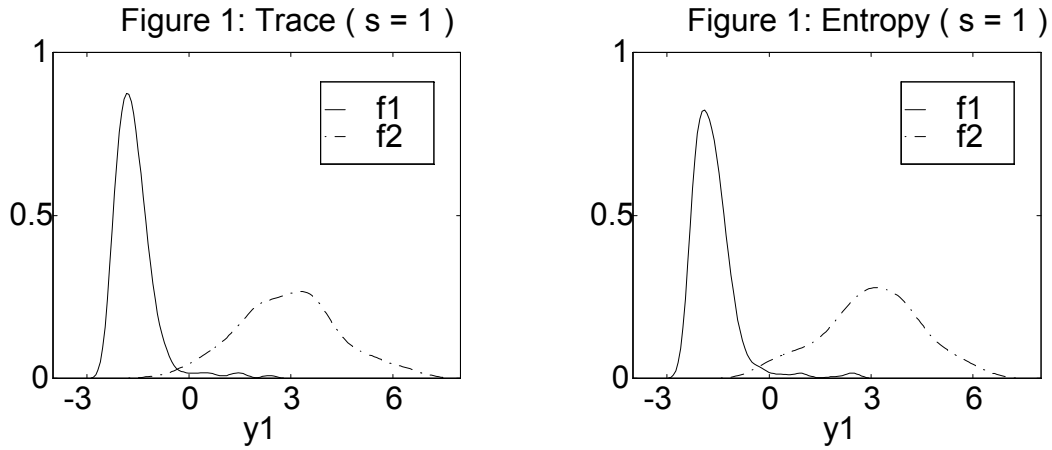

Figure 1. Wisconsin Breast Cancer Data: Density estimators for the trace and entropy criterions for $s = 1$. For $i = 1, 2$, 'fi' refers to estimator $\widehat{f}_{\mathbf{y},i}(y_1; \widehat{\mathbf{A}}_{1,0})$.

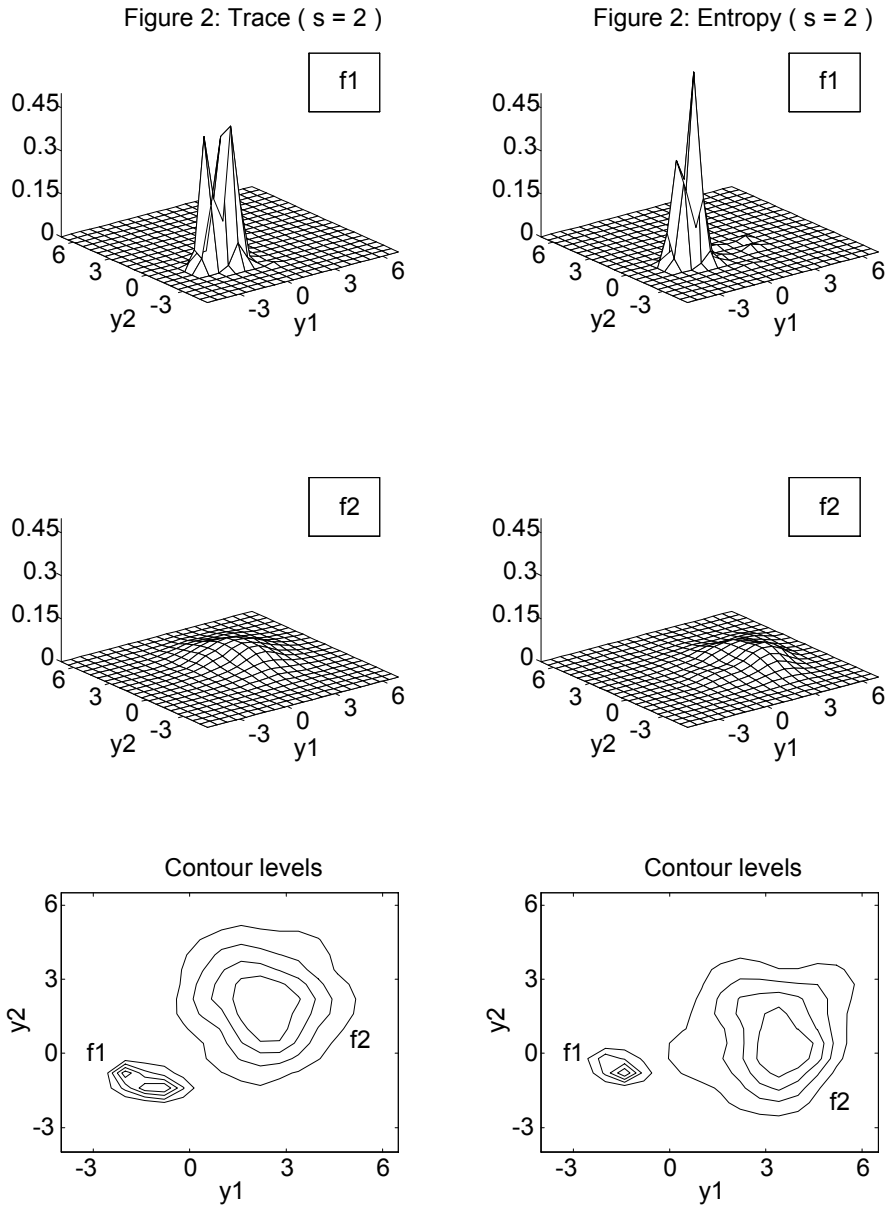

Figure 2. Wisconsin Breast Cancer Data: Density estimators for the trace and entropy criterions for $s = 2$. For $i = 1, 2$, 'fi' refers to estimator $\widehat{f}_{\mathbf{y},i}(y_1,y_2;\widehat{\mathbf{A}}_{2,0})$.

	Trace Criterion			Entropy Criterion		
s	$\widehat{C}_s$	$L_n^{(R)}$	$\widehat{L}_{n,m}$	$\widehat{H}_s$	$L_n^{(R)}$	$\widehat{L}_{n,m}$
1	.5573	.350	.398	2.900	.405	.435
2	.8357	.085	.165	9.393	.113	.162
3	.8645	.063	.174	9.006	.088	.164
4	.8977	.038	.177	8.739	.043	.174
5	.9316	.008	.186	8.471	.018	.185
6	.9645	.000	.192	8.372	.003	.192
7	.9842	.000	.209	8.247	.000	.200
⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	1.0000	.000	.231	4.063	.000	.231

Table 2. Wave Form Data: Sequential optimization of the trace and entropy criterions and estimation of error rates for the range $1 \leq s \leq 7$. Row $s = p = 21$ corresponds to the full KDA rule.

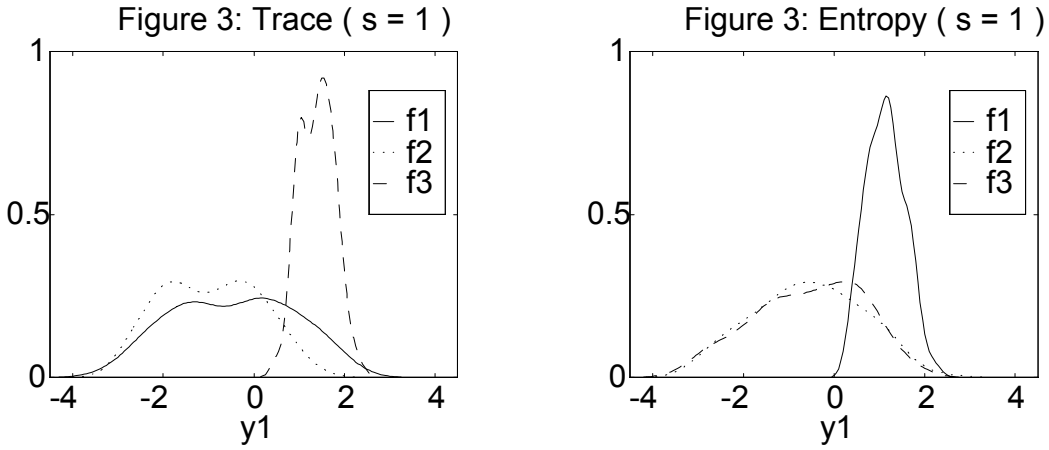

Figure 3. Wave Form Data: Density estimators for the trace and entropy criterions for $s = 1$. For $i = 1, 2, 3$, 'fi' refers to estimator $\widehat{f}_{\mathbf{y},i}(y_1; \widehat{\mathbf{A}}_{1,0})$.

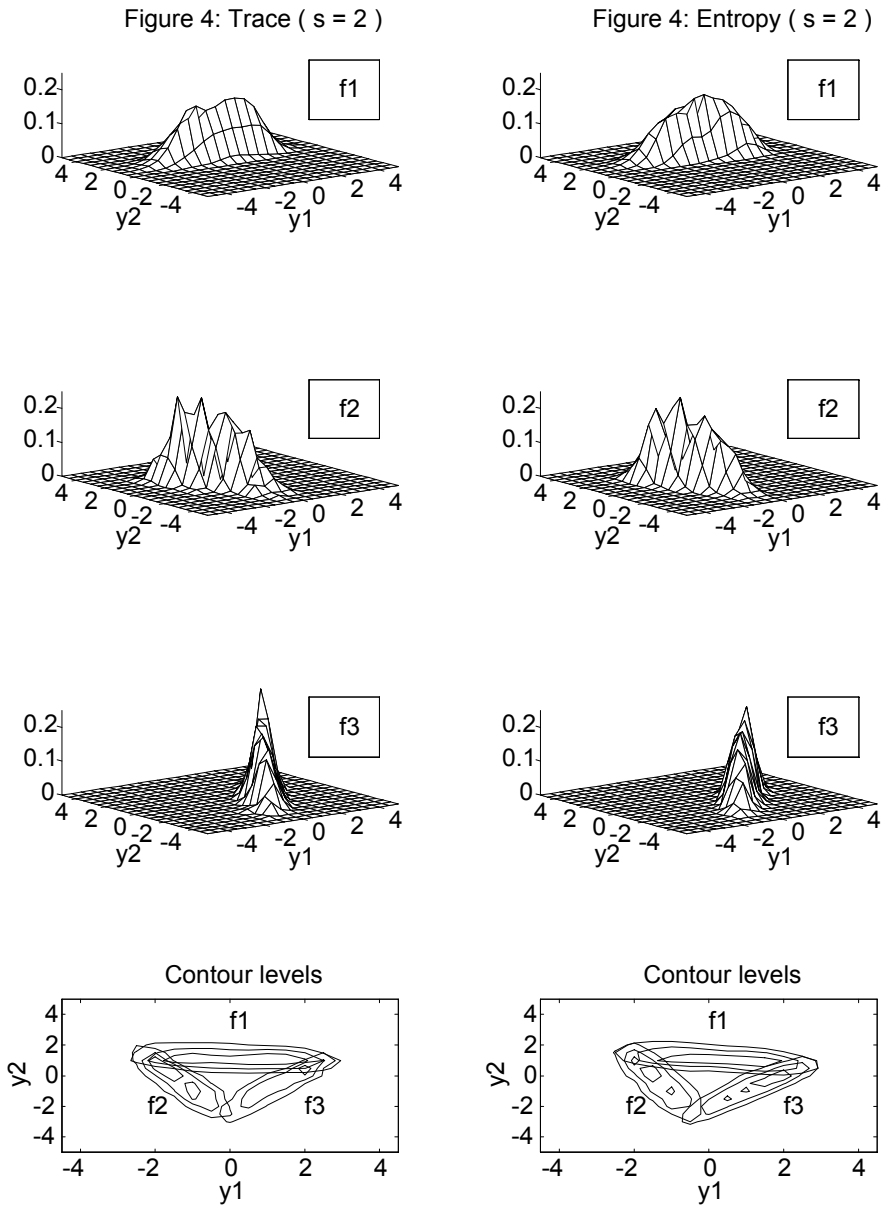

Figure 4. Wave Form Data: Density estimators for the trace and entropy criterions for $s = 2$. For $i = 1, 2, 3$, $'\mathsf{fi}'$ refers to estimator $\widehat{f}_{\mathbf{y},i}(y_1,y_2;\widehat{\mathbf{A}}_{2,0})$.

	Trace Criterion			
s	$\widehat{C}_s$	$L_n^{(R)}$	$\widehat{L}_{n,10}$	$s.e.$
1	.6846	.2113	.2030	.0211
2	.8681	.0589	.0701	.0207
3	.9538	.0234	.0320	.0118
4	.9657	.0117	.0260	.0124
5	.9722	.0108	.0420	.0298
⋮	⋮	⋮	⋮	⋮
18	.9867	.0593	.2688	.0957

Table 3. Image Segmentation Data: Sequential optimization of the trace criterion and estimation of error rates for the range $1 \leq s \leq 5$. Row $s = p = 18$ corresponds to the full KDA rule.

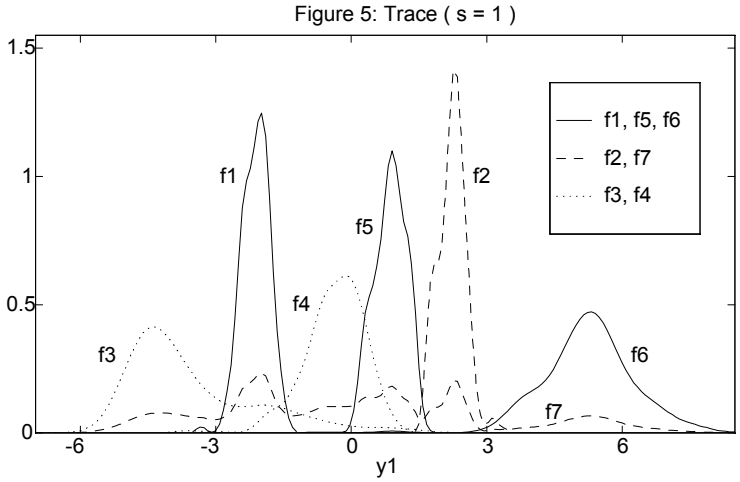

Figure 5. Image Segmentation Data: Density estimators for the trace criterion for $s = 1$. For $i = 1, \ldots, 7$, 'fi' refers to estimator $\widehat{f}_{\mathbf{y},i}(y_1; \widehat{\mathbf{A}}_{1,0})$.

Figure 6: Trace (s = 2)

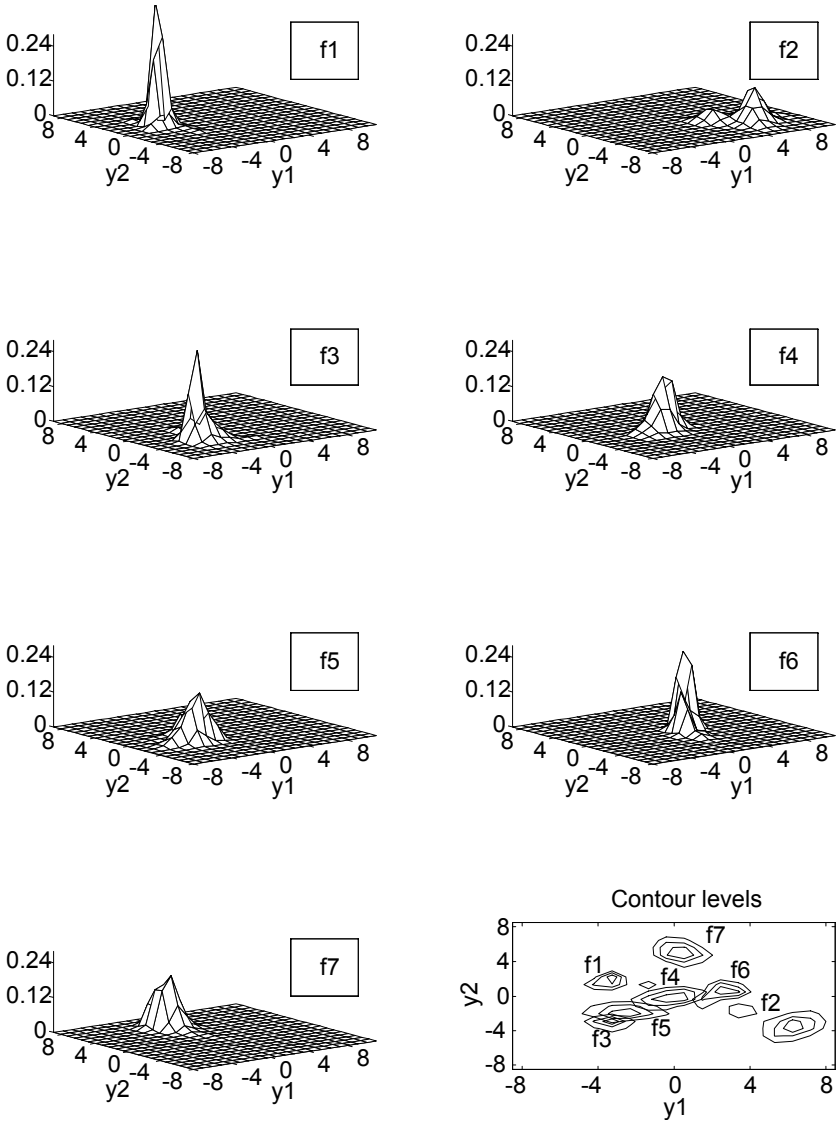

Figure 6. Image Segmentation Data: Density estimators for the trace criterion for $s = 2$. For $i = 1, \ldots, 7$, 'fi' refers to estimator $\widehat{f}_{\mathbf{y},i}(y_1,y_2;\widehat{\mathbf{A}}_{2,0})$.