UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

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

Rapports de Recherche

N° 1481

Programme 5
Traitement du Signal,
Automatique et Productique

DISCRETE REGULARIZED DISCRIMINANT ANALYSIS

Gilles CELEUX
Abdallah MKHADRI

Juillet 1991

RR-1481

DISCRETE REGULARIZED DISCRIMINANT ANALYSIS

ANALYSE DISCRIMINANTE QUALITATIVE REGULARISEE

Gilles Celeux and Abdallah Mkhadri

INRIA Rocquencourt, 78150 Le Chesnay, France

Abstract: A method of regularized discriminant analysis for discrete data, denoted DRDA hereafter, is proposed. This method is related to Regularized Discriminant Analysis of Friedman (1989) conceived in a Gaussian framework for continuous data. Here, we are concerned with discrete data and consider the classification problem using the multinomial distribution. DRDA has been conceived in the small sample, high-dimensional setting. This method has a median position between multinomial discrimination, the first order independence model and kernel discrimination. DRDA is characterized by two parameters, the values of which are calculated by minimizing a sample-based estimate of future misclassification risk by cross-validation. The first parameter is a *complexity* parameter which provides class conditional probabilities as convex combination of those derived from the full multinomial model and the first order independence model. The second parameter is a *smoothing* parameter associated to the discrete kernel of Aitchitson and Aitken (1976). The optimal complexity parameter is calculated first, then, holding fixed this parameter, the optimal smoothing parameter is determined. The efficiency of the approach is examined with other classical methods through applications to data.

Keywords: Discrimination, Multinomial Model, Nonparametric Density Estimation, Cross-validation, Regularization.

Résumé : nous proposons une méthode de régularisation pour la discrimination sur variables qualitatives. Cette méthode s'inspire de la discrimination régularisée de Friedman (1989) conçue dans un cadre gaussien. Notre méthode, définie dans le cadre multinomial, concerne les petits échantillons. Elle utilise deux paramètres de régularisation qui sont déterminés par minimisation du risque d'erreur évalué par validation croisée. Le premier paramètre est un paramètre de *complexité* qui conduit à une règle de décision intermédiaire entre celles fournies par le modèle multinomial complet et le modèle d'indépendance d'ordre un. Le deuxième paramètre est un paramètre de *lissage* qui est associé à l'estimation par la méthode des noyaux, introduite par Aitchison et Aitken (1976), des densités par groupe. Nous déterminons tout d'abord le paramètre optimal de complexité, puis, celui-ci étant fixé, le paramètre optimal de lissage. Nous comparons à partir de simulations de Monte-Carlo et sur des données réelles les performances de notre méthode vis-à-vis de celles des méthodes classiques. Ces expériences illustrent ses qualités.

Mots-clés : discrimination, modèle multinomial, estimation non paramétrique de densités, validation croisée, régularisation.

1. Introduction

In multivariate discriminant analysis, each object is assumed to come from one of K exclusive groups E_1, ..., E_K and is associated with a p-dimensional vector $\mathbf{x} = (x_1, ..., x_p)$. The purpose is to construct a classification rule using a training sample $\mathbf{x}_1$, ..., $\mathbf{x}_n$ of multivariate observations for which membership in a specific one of the K groups is known *a priori*. The Bayes procedure minimizes the expected value of the probability of misclassification. It leads to classify an observation $\mathbf{x}$ into E_k if

$$\delta_k p(\mathbf{x}|E_k) \geq \delta_g p(\mathbf{x}|E_g) \quad \text{for } g = 1,..., K, g \neq k, \tag{1}$$

where δ_g is the prior probability that an observation belongs to the gth group and $p(.|E_g)$ is the conditional probability function for the gth group. Usually, the conditional probabilities $p(.|E_g)$ are unknown and are estimated on the basis of the training sample.

In this paper, we are concerned with discrete discriminant analysis in the small sample, high-dimensional setting. Most of the discrete classification methods can appear to perform poorly because discrete data tables are often very sparse. For instance, if we consider ten binary variables, they are $2^p = 1024$ possible binary vectors (states), and we encounter a troublesome problem of sparseness in which some of the states or multinomial cells may have no data in the training sets (for one or several groups).

We propose a method of regularized discriminant analysis for discrete data, denoted DRDA hereafter. This method is related to Regularized Discriminant Analysis (RDA) of Friedman (1989) conceived in a Gaussian framework for continuous data. For the sake of simplicity, we focus on binary data. In section 2, we sketch the discrete classification methods derived from the multinomial distribution : the full multinomial model, the first order independence model and kernel discrimination. In section 3, we summarize the Friedman's regularized discriminant analysis. In section 4, we first parallel Gaussian discrimination and multinomial discrimination, then we present the regularized model. As for RDA, DRDA is characterized by two parameters and leads to a method which has an intermediate position between three multinomial discrimination models. But, contrarily to RDA, it is possible to explicitly compute the parameters which minimize the cross-validated misclassification risk. The efficiency of DRDA is examined through simulations studies and application to data in section 5. A concluding section gives some additional remarks.

2. Discrete discriminant analysis

For discrete data, the most natural model is to assume that conditional probabilities $p(\mathbf{x}|E_k)$, where $\mathbf{x} \in \{0,1\}^p$ and $k = 1, ..., K$, are multinomial probabilities. In this case, the conditional probabilities are estimated by the observed frequencies

$$M_k(\mathbf{x}) = \frac{N_{0k}(\mathbf{x})}{n_k}, \text{ for } k = 1, ..., K, \tag{2}$$

with $n_k = \# E_k$ and $N_{0k}(\mathbf{x})$ is the number of observations of the training sample belonging to group E_k for which the state $\mathbf{x}$ occurs. Goldstein and Dillon (1978) call this model the Full Multinomial Model (FMM). This method involves 2^p -1 parameters in each group. Hence, even for moderate p, not all of parameters are identifiable and many conditional probabilities are estimated to be 0, if n is not very large.

They are two ways to deal with this high dimensional problem. The first one consists in reducing the number of parameters needed to be estimated. The loglinear model (Cox 1972, Krzanowski 1975), the Bahadur model (Bahadur 1961), the logistic discrimination (Anderson 1972) are some examples of this kind of model which reduce the complexity. Here, we concentrate attention on the first order independence model (FOIM) (see Goldstein & Dillon 1978) that we describe now. It is assumed that for given group E_k ($k = 1, ..., K$), the binary variables are independent. Hence, the estimated conditional probabilities are

$$I_k(\mathbf{x}) = \prod_{j=1}^{p} \frac{N_{0k}^j(\mathbf{x})}{n_k} \tag{3}$$

where $N_{0k}^j(\mathbf{x}) = \#\{\mathbf{y} \in E_k | y^j = x^j\}$, $j = 1, ..., p$.

It follows that the number of parameters to be estimated for each group is reduced from 2^p -1 to p. Intensive experiments show that FOIM performs well for small or moderate sample size (cf. Titterington *et al.* 1981).

The second way consists in smoothing the observed frequencies. The nearest neighbour models (Hills 1967, Hall 1981b), the procedures using orthogonal polynomials (Martin & Bradley 1972, Ott & Kronmal 1976) are examples of this kind of models. Our approach is to concentrate our attention on kernel discriminant analysis (KER) (Aitchison & Aitken 1976) which turns out to provide good results for sparse data (Hand 1982, 1983). This model estimates the group conditional probability in the following way

$$p(\mathbf{x}|E_k,\lambda_k) = \frac{1}{n_k} \sum_{\mathbf{x}_i \in E_k} \lambda_k^{p-d(\mathbf{x}_i,\mathbf{x})}(1-\lambda_k)^{d(\mathbf{x}_i,\mathbf{x})}, \quad (4)$$

where $d(\mathbf{x}_i,\mathbf{x}) = \sum_{j=1}^{p} | x^j - x_i^j |$ and $\lambda_k \in [1/2,1]$.

Setting $\gamma_k = \frac{1-\lambda_k}{\lambda_k} \in [0,1]$, we get

$$p(\mathbf{x}|E_k,\gamma_k) = \frac{1}{n_k(\gamma_k + 1)^p} \sum_{\mathbf{x}_i \in E_k} \gamma_k^{d(\mathbf{x}_i,\mathbf{x})}. \quad (5)$$

In the following we will use this particular parametrization.

3. Regularized Discriminant Analysis

Friedman (1989) proposed a regularization scheme for continuous discriminant analysis. RDA has been conceived in the Gaussian framework and is devoted to the construction of efficient classification rules from small training samples. RDA provides alternative regularization using two parameters α et γ in [0,1]. The first one, α, is a *complexity* parameter which provides an intermediate classification rule between linear and quadratic discriminant analysis. The second one, γ, is a shrinkage parameter for covariance matrix estimates. More precisely, α is devoted to choose a regularized estimate of the covariance matrix of group E_k (k = 1, ..., K) in the following way

$$\hat{\Gamma}_k(\alpha) = \frac{(1-\alpha)\, S_k + \alpha\, S}{(1-\alpha)\, n_k + \alpha\, n}, \quad (6)$$

where $n_k = \#\, E_k$, $n = \sum_{k=1}^{K} n_k$ is the size of the training sample $\mathbf{x}_1,...,\mathbf{x}_n$ in $\mathbf{R}^p$, and

$$S_k = \frac{1}{n} \sum_{\mathbf{x}_i \in E_k} (\mathbf{x}_i - \mathbf{g}_k)\,(\mathbf{x}_i - \mathbf{g}_k)' \text{ with } \mathbf{g}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in E_k} \mathbf{x}_i \text{ and } S = \sum_{k=1}^{K} S_k, \quad (7)$$

and where $\mathbf{x}'$ denotes the transpose of $\mathbf{x}$.

The second parameter, γ, is used to regularize the sample group covariance estimates beyond that provided by equation (6) through

$$\hat{\Gamma}_k(\alpha, \gamma) = (1-\gamma)\ \hat{\Gamma}_k(\alpha) + \frac{\gamma}{p} \mathrm{tr}\,[(\hat{\Gamma}_k(\alpha)\,]\,\mathbf{I}, \quad (8)$$

with $\hat{\Gamma}_k(\alpha)$ given by equation (6) and **I** is the pxp identity matrix.

The RDA classification rule is then

$$d_{\hat{k}}(\mathbf{x}) = \min_{1 \le k \le K} d_k(\mathbf{x})$$

with

$$d_k(\mathbf{x}) = (\mathbf{x} - \mathbf{g}_k)'[\hat{\Gamma}_k(\alpha, \gamma)]^{-1}(\mathbf{x} - \mathbf{g}_k) + \ln |\hat{\Gamma}_k(\alpha, \gamma)| - 2 \ln \delta_k . \qquad (9)$$

Friedman's approach is to choose α and γ that jointly minimize the cross-validated estimates of future misclassification risk. Its strategy is to choose a grid of points on the α, γ plane ($0 \le \alpha \le 1$, $0 \le \gamma \le 1$), evaluate the cross-validated estimate of future misclassification risk at each prescribed point on the grid, and then choose the point with the smallest estimated risk as its estimate for the optimal regularization parameter values $\hat{\alpha}$ and $\hat{\gamma}$. Typically the size of the optimization grid is taken to be from 25 to 50 points. Friedman presented numerical experiments which indicated that RDA performs very well in many circumstances.

4. Discrete Regularized Discriminant Analysis

The underlying ideas of DRDA arose from striking analogies between continuous classification models and discrete classification models. We first state those informal ressemblances before describing DRDA in a detailed way.

- In the continuous case, the most often classification rules are based on the normal distributions. Whereas in the discrete case, they are based on the multinomial distributions.
- In the continuous case, linear discriminant analysis (LDA) provides, by assuming equality of group conditional covariance matrix, a considerable degree of regularization by substantially reducing the number of parameters to be estimated. For this reason, it leads often to superior performance, especially in small-sample setting. In the discrete case, reducing dramatically the number of parameters is achieved with the first order independence model. Here again, FOIM performs well in many situations (cf. Titterington *et al.* 1981).
- In the same way, we can parallel quadratic discrimination which make use of unrestricted Gaussian distributions and the full multinomial model. Both of them involve the estimation of many parameters and for this very reason appear to perform poorly in small-sample settings.

- The shrinkage parameter γ, introduced by Friedman is aimed at smoothing the group conditional covariance matrix estimates to attenuate the influence of eigenvectors associated with the smallest eigenvalues. In the same manner, kernel discrete discrimination aims to smooth the observed frequencies to avoid all the problems caused by empty cells.

We are now in position to describe DRDA. For the sake of simplicity, we only consider binary variables. Straightforward extensions to general categorical variables will be discussed briefly in the concluding section.

4.1 Regularization scheme

Recall that $\mathbf{x}_1, ..., \mathbf{x}_n$ denotes the training sample in $\{0,1\}^p$. Let $P_k(\mathbf{x}|\alpha, \gamma_k)$ be the DRDA estimates of the group conditional probability for any $\mathbf{x}$ in $\{0,1\}^p$, with α and γ_k ($0 \leq \alpha \leq 1$, $0 \leq \gamma_k \leq 1$ for $k = 1,..., K$) denoting the regularization parameters. The regularization equations take the form

$$P_k(\mathbf{x}|\alpha,\gamma_k) = (1- \alpha)\, P_M\,(\mathbf{x}|k,\gamma_k) + \alpha\, P_I\,(\mathbf{x}|k,\gamma_k) \tag{10a}$$

with

$$P_M\,(\mathbf{x}|k,\gamma_k) = \frac{1}{n_k(1+\gamma_k)^p} \sum_{\mathbf{x}_i \in E_k} \gamma_k^{d(\mathbf{x},\mathbf{x}_i)} \tag{10b}$$

where

$$d(\mathbf{x},\mathbf{x}_i) = \sum_{j=1}^{p} |x^j - x_i^j|,$$

and

$$P_I\,(\mathbf{x}|k,\gamma_k) = \frac{1}{\{n_k(1+\gamma_k)\}^p} \prod_{j=1}^{p} \sum_{\mathbf{x}_i \in E_k} \gamma_k^{|x^j - x_i^j|}. \tag{10c}$$

The full multinomial model corresponds to the case $\alpha = 0$ and $\gamma_k = 0$ for $1 \leq k \leq K$. The first order independence model corresponds to the case $\alpha = 1$ and $\gamma_k = 0$ for $1 \leq k \leq K$. Holding the γ_k's fixed at 0 and varying α produces models between FMM and FOIM. Holding α fixed at 0 and varying the γ_k's leads to the kernel discriminant analysis (KER) of Aitchison and Aitken.

The problem is to choose good values for α and the γ_k's. Contrarily to Friedman, it is possible to select the regularization parameters in a nearly optimal fashion. Holding the γ_k's fixed, we can find in closed form the *complexity* parameter α which minimizes the cross-validated misclassification risk; holding α fixed, we can find the γ_k's which

minimize the cross-validated misclassification risk. We opted for the following strategy which is saving a substantial amount of computation: we first choose the optimal *complexity* parameter; and then, holding this parameter fixed, we select the optimal *smoothing parameters*. This approach seems natural since the *complexity parameter* is the most important and is expected to provide a small error rate in most cases.

4.2 The optimal complexity parameter

In this section, the γ_k's are fixed. For simplicity we assume that $\gamma_k = 0$ for all k. The overall probability of misclassification risk of DRDA can be written (cf. Tutz 1986)

$$T(\alpha) = \sum_{k=1}^{K} \delta_k \sum_{\mathbf{x}} P_k(\mathbf{x}) [\, 1 - \mathbb{1}_k\{ \delta_1 P_1(\mathbf{x}|\alpha, 0),..., \delta_K P_K(\mathbf{x}|\alpha, 0)\}], \qquad (11)$$

where

$$\mathbb{1}_k\{z_1,...,z_K\} = \begin{cases} 1 & \text{if } z_k > z_j \text{ for } j \neq k \\ 1/r & \text{if } z_k = z_i \text{ for } i \in \{i_1,...,i_r\},\ z_k > z_j \text{ for } i \notin \{i_1,...,i_r\} \\ 0 & \text{otherwise.} \end{cases}$$

and where $P_k(\mathbf{x})$ denotes the unknown group-conditional probabilities for $\mathbf{x} \in \{0,1\}^p$ and $1 \leq k \leq K$. The problem is to find α minimizing $T(\alpha)$. Since $T(\alpha)$ is unknown, the optimal α has to be estimated from data. It is well known that the method of resubstitution provides misleading biased error rate, especially for small samples (cf. Glick 1972). An alternative approach would be to consider a parametric estimation of $T(\alpha)$ which consists of substituting $P_k(\mathbf{x})$ by $P_k(\mathbf{x}|\alpha, 0)$ $(1 \leq k \leq K)$ in equation (11). The drawback of this method is that the resulting estimate of $T(\alpha)$ depends highly of the involved parametric model (Hand 1986). More effecient methods of estimating the misclassification risk are based on resampling techniques such as cross-validation and bootstrapping (cf. Efron 1983). Following Friedman (1989) we choose the cross-validation technique for its computational advantages. We search the *complexity* parameter α which minimizes the cross-validated probability of misclassification $T^*(\alpha)$

$$T^*(\alpha) = \sum_{k=1}^{K} \delta_k \sum_{\mathbf{x}_i \in E_k} \frac{1}{n_k} [1 - \mathbb{1}_k\{\delta_1 P_1(\mathbf{x}_i|\alpha, 0),...,\delta_k P_{k\setminus\mathbf{x}_i}(\mathbf{x}_i|\alpha, 0),...,\delta_K P_K(\mathbf{x}_i|\alpha, 0)\}] \qquad (12)$$

where $P_{k\setminus\mathbf{x}_i}(\mathbf{x}_i|\alpha, 0)$ denotes the estimation of $P_k(\mathbf{x}_i|\alpha, 0)$ using $E_k - \{\mathbf{x}_i\}$; it can be written

$$P_{k\backslash \mathbf{x}_i}(\mathbf{x}_i|\alpha, 0) = (1 - \alpha)\, M_k^{(i)}(\mathbf{x}_i) + \alpha\, I_k^{(i)}(\mathbf{x}_i), \tag{13}$$

where

$$M_k^{(i)}(\mathbf{x}_i) = \begin{cases} \dfrac{n_k M_k(\mathbf{x}_i) - 1}{n_k - 1} & \text{if} \quad \mathbf{x}_i \in E_k \\ M_k(\mathbf{x}_i) & \text{if} \quad \mathbf{x}_i \notin E_k \end{cases}$$

and

$$I_k^{(i)}(\mathbf{x}_i) = \begin{cases} \dfrac{1}{(n_k - 1)^p} \prod_{j=1}^{p} (N_{0k}^j(\mathbf{x}_i) - 1) & \text{if} \quad \mathbf{x}_i \in E_k \\ I_k(\mathbf{x}_i) & \text{otherwise} \end{cases}$$

with $N_{0k}^j(\mathbf{x}_i) = \#\{\mathbf{x}_h \text{ in } E_k \text{ such that } | x_h^j - x_i^j | = 0,\ h = 1, \ldots, n_k\}$, $j = 1, \ldots, p$; $k = 1, \ldots, K$.

For the sake of simplicity, calculations will be detailed for two groups (K = 2) and the general case will be discussed afterwards.

Proposition 1: The optimal *complexity* parameter is either 0, either 1 or takes the form

$$\frac{\delta_1 M_1^{(i)}(\mathbf{x}_i) - \delta_2 M_2^{(i)}(\mathbf{x}_i)}{\delta_1 \{M_1^{(i)}(\mathbf{x}_i) - I_1^{(i)}(\mathbf{x}_i)\} - \delta_2 \{M_2^{(i)}(\mathbf{x}_i) - I_2^{(i)}(\mathbf{x}_i)\}}, \tag{14}$$

with $\mathbf{x}_i \in \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

Proof: Holding α fixed, it is immediate that the cross-validated classification rule for any $\mathbf{x}_i$ ($1 \le i \le n$) is: $\mathbf{x}_i$ is assigned to E_1 if and only if $C(\mathbf{x}_i,\alpha) \ge 0$, where

$$C(\mathbf{x}_i,\alpha) = (1 - \alpha)\, [\, \delta_1 M_1^{(i)}(\mathbf{x}_i) - \delta_2 M_2^{(i)}(\mathbf{x}_i)] + \alpha\, [\, \delta_1 I_1^{(i)}(\mathbf{x}_i) - \delta_2 I_2^{(i)}(\mathbf{x}_i)\,]. \tag{15}$$

Thus, different values of α give different assignments for $\mathbf{x}_i$ if and only if there exists α_0 in (0,1) such that $C(\mathbf{x}_i,\alpha_0) = 0$. It then follows that α_0 has the form (14).

If $C(\mathbf{x}_i,\alpha)$ has a constant sign on [0,1], the assignment of $\mathbf{x}_i$ to one of the two groups does not depend on α. In such a case, $\mathbf{x}_i$ would be assigned following the FMM ($\alpha = 0$) or following the FOIM ($\alpha = 1$).

Remark : In pratical situations, the number of sample points $\mathbf{x}_i$ ($1 \le i \le n$) for which the linear equation $C(\mathbf{x}_i,\alpha) = 0$ has a solution α in (0,1) is very small. This number represents the number of states for which both models (FMM and FOIM) provide different assignments.

From Proposition 1 it is possible to find easily and explicitly the optimal solution for the *complexity* parameter α.

General case: The problem can be treated in the same manner when $K > 2$. The only difference is that, for each sample point $\mathbf{x}_i$ $(1 \leq i \leq n)$, there may be several optimal α's as possible optimal solutions. Figure 1 gives an illustration of this situation with $K = 4$ groups.

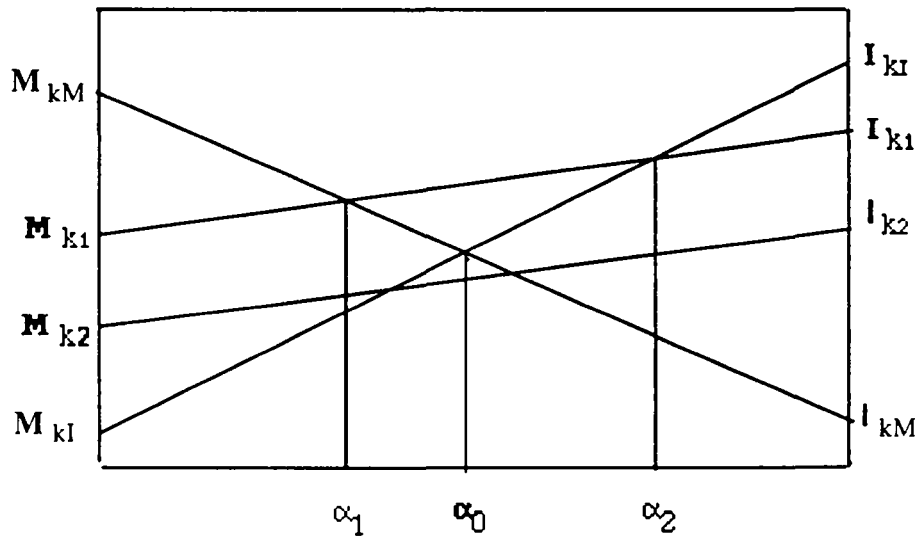

Figure1: *Variation of the estimated conditional probabilities as a function of the parameter α on a 4 group case*

This figure shows, for one fixed $\mathbf{x}_i$, the $P_{k\backslash \mathbf{x}_i}(\mathbf{x}_i|\alpha, 0)$ variation $(1 \leq k < K)$ on α. k_M (resp. k_I) denotes the assignment group of $\mathbf{x}_i$ for $\alpha = 0$ (resp. $\alpha = 1$). In this case, two critical values α_1 and α_2 of the *complexity* parameter can be found in (0,1). They represent values for which the classification of $\mathbf{x}_i$ changes and they have to be considered as possible optimal values.

Thus, each $\mathbf{x}_i$ generates at most $(K - 1)$ such critical α–values different from 0 and 1. All the critical α–values can be found explicitly with the following procedure.

- If $k_M(\mathbf{x}_i) = k_I(\mathbf{x}_i)$, then $\mathbf{x}_i$ is assigned to $k_M(\mathbf{x}_i)$ for any α.
- Otherwise compute α_0 in (0,1) such that

$$(1 - \alpha_0)\delta_{k_M} M^{(i)}_{k_M}(\mathbf{x}_i) + \alpha_0 \delta_{k_M} I^{(i)}_{k_M}(\mathbf{x}_i) = (1 - \alpha_0)\delta_{k_I} M^{(i)}_{k_I}(\mathbf{x}_i) + \alpha_0 \delta_{k_I} I^{(i)}_{k_I}(\mathbf{x}_i). \qquad (16a)$$

If, for any k, we have

$$(1 - \alpha_0)\delta_k M_k^{(i)}(\mathbf{x}_i) + \alpha_0 \delta_k I_k^{(i)}(\mathbf{x}_i) \leq (1 - \alpha_0)\delta_{k_M} M_{k_M}^{(i)}(\mathbf{x}_i) + \alpha_0 \delta_{k_M} I_{k_M}^{(i)}(\mathbf{x}_i), \qquad (16b)$$

then α_0 is the only possible optimal value, associated to $\mathbf{x}_i$, to be considered,

Else, let k_0 ($\neq k_M$ and $\neq k_I$) be the value which maximizes

$$(1 - \alpha_0)\delta_{k_0} M_{k_0}^{(i)}(\mathbf{x}_i) + \alpha_0 \delta_{k_0} I_{k_0}^{(i)}(\mathbf{x}_i).$$

Compute $\alpha_1 \in (0, \alpha_0)$ and $\alpha_2 \in (\alpha_0, 1)$ such that

$$(1 - \alpha_1)\delta_{k_M} M_{k_M}^{(i)}(\mathbf{x}_i) + \alpha_1 \delta_{k_M} I_{k_M}^{(i)}(\mathbf{x}_i) = (1 - \alpha_1)\delta_{k_0} M_{k_0}^{(i)}(\mathbf{x}_i) + \alpha_1 \delta_{k_0} I_{k_0}^{(i)}(\mathbf{x}_i) \qquad (16c)$$

and

$$(1 - \alpha_2)\delta_{k_0} M_{k_0}^{(i)}(\mathbf{x}_i) + \alpha_2 \delta_{k_0} I_{k_0}^{(i)}(\mathbf{x}_i) = (1 - \alpha_2)\delta_{k_I} M_{k_I}^{(i)}(\mathbf{x}_i) + \alpha_2 \delta_{k_I} I_{k_I}^{(i)}(\mathbf{x}_i). \qquad (16d)$$

Then proceed in the same way, used for α_0, to check if α_1 and α_2 are possible optimal values. And so on ...

For instance, for the situation described in Figure 1, α_1 and α_2 are candidates for optimality and the procedure stops.

Remark: Obviously, when $K > 2$ the algorithm to find the optimal *complexity* parameter is more complicated. However, in practical situations, for each $\mathbf{x}_i$ they are only one or two possible optimal α–values and the procedure remains competitive.

4.3. The optimal smoothing parameters

From the paper of Aitchison & Aitken (1976), many authors have considered the kernel density estimation for categorical variables and have proposed methods for estimating the *smoothing* parameter γ_k ($1 \leq k \leq K$) involved in (5) (Bowman 1980, Titterington 1980, Hall (1981a), Bowman *et al.* 1984, Brown & Rundell 1985). Some authors have considered the problem in the classification setting and proposed methods to estimate the *smoothing* parameter γ_k ($1 \leq k \leq K$) using loss criteria that refer directly to the discrimination problem (Tutz 1986, 1989, Hall & Wand 1988). Tutz (1986) proposed to choose $(\gamma_1, \ldots, \gamma_K)$ minimizing the leaving-one-out error rate stated in equation (12). For two groups E_1 and E_2, Hall & Wand (1988) considered the minimization of the distance between $\delta_1 p(\mathbf{x}|E_1) - \delta_2 p(\mathbf{x}|E_2)$ and $\delta_1 p(\mathbf{x}|E_1, \gamma_1) - \delta_2 p(\mathbf{x}|E_2, \gamma_2)$. Tutz (1989) proposed to choose $\gamma = (\gamma_1, \ldots, \gamma_K)$ minimizing other loss

functions than the leaving-one-out error rate, using cross-validation. These loss functions are continuous in γ. Acting in such a way, Tutz aimed to avoid the numerical problems arising when using the leaving-one-out error rate which is a jump function of γ.

Since we are concerned with discriminant analysis, it is natural to choose γ minimizing a discriminant loss function. Now, since we are mainly concerned with small or very small sample sizes, it is not realistic to introduce too many *regularization* parameters. Thus, we introduce only one *smoothing* parameter ($\gamma = \gamma_1 = ... = \gamma_K$), as Friedman has done for the *shrinkage* parameter. Another argument can be used for choosing an unique *smoothing* parameter γ. Numerical experiments reported in Hand (1982 p. 162, 1983) and Tutz (1986) showed that, in many situations, the smallest classification risk occurs with an unique γ. One of the continuous discriminant functions considered by Tutz (1989) could be preferred since they induce non numerical difficulties. Nevertheless, Tutz's approach needs optimization algorithms. On an other hand, Hand (1982, section 4.4) has noticed that the choice of the smoothing method is not very critical, and that the computationally less demanding methods should be used.

For these reasons, we adopt for determining γ a simple method which does not use optimization algorithms. This method uses the classical misclassification risk as loss function and is described now. First, we suppose that the optimal *complexity* parameter α^* has been found out, as described in the previous section, and is held fixed.

Let $P_M^{(i)}(\mathbf{x}_i|k,\gamma)$ (resp. $P_I^{(i)}(\mathbf{x}_i|k,\gamma)$) denote $P_M(\mathbf{x}_i|E_k-\{\mathbf{x}_i\},\gamma)$ (resp. $P_I(\mathbf{x}_i|E_k\{\mathbf{x}_i\},\gamma)$) the estimated probabilities by cross-validation ($1 \le i \le n$) under the FMM (resp. FOIM). Using Taylor's series to expand $P_M^{(i)}(.|k,\gamma)$ in power of γ, we obtain

$$P_M^{(i)}(\mathbf{x}_i|k,\gamma) = (n_k^{(i)})^{-1}\{N_{0k}^{(i)}(\mathbf{x}_i) + \gamma[N_{1k}(\mathbf{x}_i) - pN_{0k}^{(i)}(\mathbf{x}_i)]\} + O((n_k^{(i)})^{-1}\gamma^2) \qquad (17)$$

where

$$N_{0k}^{(i)}(\mathbf{x}_i) = \begin{cases} N_{0k}(\mathbf{x}_i) - 1 & \text{if } \mathbf{x}_i \in E_k \\ N_{0k}(\mathbf{x}_i) & \text{otherwise} \end{cases} \quad \text{et} \quad n_k^{(i)} = \begin{cases} n_k - 1 & \text{if } \mathbf{x}_i \in E_k \\ n_k & \text{otherwise} \end{cases} . \qquad (18)$$

Thus, $P_M^{(i)}(\mathbf{x}_i|k,\gamma)$ can be approximated in the following way, since the second order term in γ can be neglected.

$$P_M^{(i)}(\mathbf{x}_i|k,\gamma) \approx (1-\gamma)M_k^{(i)}(\mathbf{x}_i) + \gamma V_k^{(i)}(\mathbf{x}_i) \qquad (19)$$

where

$$V_k^{(i)}(\mathbf{x}_i) = \frac{N_{1k}(\mathbf{x}_i) - (p-1)N_{0k}^{(i)}(\mathbf{x}_i)}{n_k^{(i)}}. \tag{20}$$

In the same manner, we obtain by using a Taylor expansion of $P_I^{(i)}(\mathbf{x}_i|k,\gamma)$ about γ

$$P_I^{(i)}(\mathbf{x}_i|k,\gamma) \approx (1-\gamma)I_k^{(i)}(\mathbf{x}_i) + \gamma Q_k^{(i)}(\mathbf{x}_i) \tag{21}$$

where

$$Q_k^{(i)}(\mathbf{x}_i) = B_k^{(i)}(\mathbf{x}_i) - (p-1)I_k^{(i)}(\mathbf{x}_i) \tag{22}$$

with

$$B_k^{(i)}(\mathbf{x}_i) = (n_k^{(i)})^{-p} \sum_{h=1}^{p} N_{1k}^{(i)h}(\mathbf{x}_i) \prod_{j \neq h} N_{0k}^{(i)j}(\mathbf{x}_i), \tag{23}$$

$$N_{1k}^{(i)j}(\mathbf{x}_i) = n_k - N_{0k}^{(i)j}(\mathbf{x}_i) \text{ and } N_{0k}^{(i)j}(\mathbf{x}_i) = \begin{cases} N_{0k}^{j}(\mathbf{x}_i) - 1 & \text{if } \mathbf{x}_i \in E_k \\ N_{0k}^{j}(\mathbf{x}_i) & \text{otherwise} \end{cases}. \tag{24}$$

From Equations (10), (19) and (21) $P_k^{(i)}(\mathbf{x}_i|\alpha^*,\gamma)$ is approximately an affine function in γ, and hence the optimal γ^* can be derived using the same line as for the optimal α^*. More precisely, in the two-group case, we have for any $\mathbf{x}_i$

$\mathbf{x}_i$ is assigned to E_1 (using cross-validation) if

$$C(\mathbf{x}_i,\alpha^*,\gamma) \geq 0, \tag{25}$$

where

$$C(\mathbf{x}_i,\alpha^*,\gamma) = (1-\alpha^*)[(1-\gamma)\delta_1 M_1^{(i)}(\mathbf{x}_i) + \gamma\delta_1 V_1^{(i)}(\mathbf{x}_i) - (1-\gamma)\delta_2 M_2^{(i)}(\mathbf{x}_i) - \gamma\delta_2 V_2^{(i)}(\mathbf{x}_i)]$$
$$+ \alpha^*[(1-\gamma)\delta_1 I_1^{(i)}(\mathbf{x}_i) + \gamma\delta_1 Q_1^{(i)}(\mathbf{x}_i) - (1-\gamma)\delta_2 I_2^{(i)}(\mathbf{x}_i) - \gamma\delta_2 Q_2^{(i)}(\mathbf{x}_i)]. \tag{26}$$

It turns out that $C(\mathbf{x}_i,\alpha^*,\gamma)$ is an affine function in γ and we have the following proposition.

Proposition 2: Holding the *complexity* parameter α^* fixed, the optimal *smoothing* parameter is either 0, either 1 or is one of the solutions of the equations $C(\mathbf{x}_i,\alpha^*,\gamma) = 0$ for i = 1,..., n.

Now, the general case ($K > 2$) can be tackled in the same manner as in § 4.2 for the *complexity* parameter α. Here again, it is worth noting that, for each $\mathbf{x}_i$ ($1 \leq i \leq n$), there are, generally, at most two possible optimal γ values.

5. Numerical experiments

We investigated the performance of DRDA compared with FOIM, KER (where the unique smoothing parameter has been chosen along the line described § 4.3) and linear discriminant analysis (LDA) on both real and simulated binary data. LDA is not really related with DRDA, but since this method gives often good results even in discrete case (see Titterington *et al.* 1980), it is worthwile to compare it with specific methods of discrete discriminant analysis.

5.1 Simulated data

We have performed Monte-Carlo sampling experiments implemented from the Bahadur model as discussed in Dillon and Goldstein (1978). This representation expresses the group-conditional probabilities in the following manner

$$p(\mathbf{x}|E_g) = \prod_{j=1}^{p} (\theta_{gj})^{x^j} (1 - \theta_{gj})^{1-x^j} \{1 + \sum_{k \neq j} \rho_g(jk) Z_{gj} Z_{gk}\},$$

where for g = 1, 2 and j = 1, ..., p X_{gj} is denoting a Bernoulli variable with parameter θ_{gj}, such that

$$\theta_{gj} = E(X_{gj})$$

$$Z_{gj} = \frac{X_{gj} - \theta_{gj}}{[\theta_{gj} (1 - \theta_{gj})]^{1/2}}$$

and

$$\rho_g(jk) = E(Z_{gj} Z_{gk}).$$

We selected 3 population structures with $p = 6$ variables and $g = 2$ groups that we describe below. For each population structure, we considered 3 sample sizes $n = 100$, 50, 20. The prior probability of each group was taken to be equal using the diagnostic paradigm so that $n_1 = n_2 = n/2$. Each experiment consisted of 100 replications of each

case. The 3 different population structures were generated from the following sets of parameters.

For each structure, the θ_{gj}'s were

E_1: 0.6, 0.4, 0.6, 0.5, 0.5, 0.6

E_2: 0.5, 0.3, 0.5, 0.4, 0.4, 0.5

The first population structure, denoted IND, was generated according to the first order independence model. It means that $\rho_g(jj) = 1$ and $\rho_g(jk) = 0$ if $j \neq k$ for $g = 1, 2, j = 1,..., 6, k = 1,..., 6$.

The second one was generated with

$$\rho_1(jj) = 1 \text{ and } \rho_1(jk) = 0.2 \text{ if } j \neq k \text{ for } j = 1,..., 6, k = 1,..., 6$$

$$\rho_2(jj) = 1 \text{ and } \rho_2(jk) = 0.4 \text{ if } j \neq k \text{ for } j = 1,..., 6, k = 1,..., 6.$$

it is denoted DIFF.

The third one, called CORR, was generated with

$$\rho_g(jj) = 1 \text{ and } \rho_g(jk) = 0.2 \text{ if } j \neq k \text{ for } g = 1, 2, j = 1,..., 6, k = 1,..., 6.$$

For each data set we performed the 4 methods : FOIM, KER, LDA and DRDA. An additional test data set of size n =100 was randomly generated from the same population structure and classified with the four rules derived from the training set, thereby obtaining an unbiased estimate of the misclassification risk.

Tables 1-3 summarize the results for each situation and give, for the three sample sizes, the average cross-validated misclassification risk (column CV) and the average test misclassification risk (column TEST) over 100 replications for each of the four classification rules. Also shown, are the average of the parameter γ for the KER rule over the 100 replications, and the means of the selected regularization parameters (α, γ) for the DRDA rule over the 100 replications. The quantities in parentheses are the standard deviations of the respective quantities over the 100 replications.

From Table 1, it can be seen that, for the data set IND and as would be hoped, DRDA is choosing a high degree of regularization for the complexity parameter. The first order independence model gave a slightly lower misclassification risk. Remark that this advantage decreases as the sample size decreases.

Table 1: *Misclassification risk and regularization parameter values for population structure IND.*

	n = 100		n = 50		n = 20	
	CV	TEST	CV	TEST	CV	TEST
LDA	.46 (.07)	.36 (.04)	.44 (.09)	.39 (.06)	.44 (.14)	.43 (.07)
FOIM	.39 (.05)	.36 (.04)	.40 (.07)	.39 (.07)	.40 (.12)	.41 (.05)
KER	.29 (.04)	.41 (.05)	.32 (.05)	.44 (.06)	.38 (.10)	.46 (.05)
$\tilde{\gamma}$	.08 (.08)		.08 (.08)		.07 (.08)	
DRDA	.27 (.03)	.38 (.05)	.27 (.05)	.40 (.07)	.18 (.07)	.42 (.05)
$\tilde{\alpha}$	.63 (.13)		.87 (.12)		.98 (.06)	
$\tilde{\gamma}$	.19 (.15)		.14 (.14)		.13 (.10)	

Table 2: *Misclassification risk and regularization parameter values for population structure DIFF.*

	n = 100		n = 50		n = 20	
	CV	TEST	CV	TEST	CV	TEST
LDA	.37 (.08)	.37 (.06)	.40 (.10)	.38 (.06)	.41 (.15)	.40 (.08)
FOIM	.42 (.05)	.47 (.07)	.42 (.07)	.46 (.08)	.42 (.10)	.47 (.09)
KER	.20 (.03)	.25 (.02)	.19 (.05)	.26 (.02)	.15 (.07)	.27 (.05)
$\tilde{\gamma}$	.00 (.00)		.00 (.00)		.00 (.00)	
DRDA	.20 (.03)	.25 (.02)	.19 (.05)	.26 (.02)	.15 (.07)	.28 (.06)
$\tilde{\alpha}$	.00 (.00)		.00 (.00)		.07 (.25)	
$\tilde{\gamma}$	.00 (.00)		.00 (.00)		.01 (.03)	

Table 3: *Misclassification risk and regularization parameter values for population structure CORR.*

	n = 100		n = 50		n = 20	
	CV	TEST	CV	TEST	CV	TEST
LDA	.46 (.08)	.48 (.04)	.46 (.10)	.49 (.04)	.46 (.16)	.50 (.04)
FOIM	.42 (.05)	.42 (.05)	.41 (.06)	.43 (.06)	.42 (.13)	.45 (.08)
KER	.35 (.04)	.43 (.05)	.36 (.05)	.44 (.06)	.39 (.09)	.46 (.06)
$\bar{\gamma}$	.16 (.02)		.17 (.01)		.15 (.07)	
DRDA	.32 (.04)	.42 (.05)	.31 (.04)	.43 (.07)	.24 (.07)	.44 (.07)
$\bar{\alpha}$	.63 (.15)		.80 (.18)		.95 (.10)	
$\bar{\gamma}$	.23 (.14)		.06 (.10)		.09 (.10)	

For the data set DIFF, Table 2 shows that our regularization strategy always led to the full multinomial model. Both methods differed only very slightly for the smallest sample size and outperformed dramatically FOIM and LDA. Surprisingly, there was no need for regularization even for very small sample size.

For the data set CORR, Table 3 shows that FOIM and DRDA gave rise to quite analogous classification rules and provided better misclassification risk than LDA and KER. However, note that DRDA had some marked differences with FOIM for the sample size n = 100, but gave almost the same misclassification risk. Moreover, remark that DRDA performed slightly better than FOIM for the smallest sample size. At last, LDA appeared to perform poorly. It is surprising since the correlation structure is identical in each group.

5.2 Medical data

The data set consists of 241 patients suffering from arthrose disease. The whole sample was divided into two groups. The first group contained patients for which an aggravation of disease has been discovered from a radiology examination and the second contained the other patients. For each patient, the values of 10 binary variables were available. These were: sexe, obesity, pain, subjective impression of the disease, generalisation of arthrose, presence of 3 kinds of medical care and 2 binary

characterizations of arthrose. These data have been extracted from a large data base concerning gonarthrose from the rumathologic clinic of Cochin (INSERM, clinical and biostatistic unit U88). This data base has been analysed by the biostatistic team of this INSERM unit, and an article written by Professor Dougados and his collaborators has been submitted to the Journal of Rhumathology. For our illustrative experiment, we drew at random a training sample of 141 patients and the rest constitute the test sample. Table 4 summarizes the results of the four methods for this data set. For each method, we give the cross-validated misclassification risk and the misclassification risk estimated on the test sample. The prior probabilities were taken to be equal, $\delta_k = 1/2$ ($k = 1, 2$), for each group.

Table 4: *Misclassification risk and regularization parameter values for the Medical Data.*

	LDA	FOIM	KER	DRDA
CV	48.23	43.26	41.85	38.30
TEST	42.00	43.00	49.00	40.00
$\tilde{\alpha}$				0.970
$\tilde{\gamma}$			0.100	0.278

The data set is very sparse (1024 states and only 141 observations). Thus, as it would be expected, DRDA provided large values for both regularization parameters. And, from the cross-validation and test estimates of the misclassification risk, DRDA could be preferred.

6. Concluding remarks

We have presented DRDA for binary data. It can be generalized in a simple way to categorical variables with more than two categories. For example, we can use the kernel introduced by Aitchison and Aitken (1976) for unordered categorical variables.

$$p(\mathbf{x}|E_k,\lambda_k) = \frac{1}{n_k} \sum_{\mathbf{x}_i \in E_k} K(\mathbf{x}|\mathbf{x}_i, \lambda_k)$$

where

$$K(\mathbf{x}|\mathbf{x}_i, \lambda_k) = \prod_{j=1}^{p} K_j(x^j|x_i^j, \lambda_k)$$

with

$$K_j(x^j|x_i^j, \lambda_k) = \begin{cases} \lambda_k & \text{if} \quad x_i^j = x^j \\ \dfrac{1 - \lambda_k}{c_j - 1} & \text{if} \quad x_i^j \neq x^j \end{cases} \tag{27}$$

where c_j $(1 \leq j \leq p)$ is the number of unordered categories for variable j. For alternative kernel forms taking account of ordered categorical variables (see Titterington and Bowman 1985).

Concerning FOIM, we recommend using the modified formula proposed in Titterington et al. (1981)

$$I_k(\mathbf{x}) = \prod_{j=1}^{p} \frac{N_{0k}^j(\mathbf{x}) + 1/c_j}{n_k + 1}, \tag{28}$$

which can be viewed as a smoothing version of formula (3). From formulas (27) and (28), it is straightforward to see that the method of selection of regularization parameters, defined in Section 4, can be used for categorical data. However, if the *smoothing* parameter, involved in (27), depends on the variables, or if the kernel takes account of ordered variables, it is impossible to get *smoothing* parameters in a closed form. And there is the need to use an optimization algorithm (as Tutz 1989) to derive the optimal *smoothing* parameters.

The numerical experiments showed that good performances can be expected from DRDA in many situations. However, contrarily to RDA in the Gaussian framework (Friedman 1989), we did not yet exhibit situations where DRDA improved substantially both FOIM and KER. Roughly speaking, in our experiments, DRDA is related to FOIM or to KER. In fact, it is difficult to detect natural situations for which DRDA could be expected to dominate both FOIM and FMM (or KER). Moreover, we have performed some experiments, not reported here, for well-separated groups. In such situations, all the above mentioned methods perform well and the DRDA is not really useful.

Despite these restrictions, we think that DRDA should be quite beneficial for discrete discriminant analysis in setting for which sample sizes are small and the groups not well-separated. On another hand, a common method of regularization is variable subset selection. But in the discrete context, there is not really reliable variable-selection

procedures especially for sparse tables. Now, sparseness is certainly the main problem of discrete discriminant analysis in practical situations. Hence, DRDA could be expected to improve the power of discrete classification in high-dimensional setting.

References

AITCHISON J. & AITKEN C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika 63*, 413-20.

ANDERSON J. A. (1972). Separate sample logistic discrimination. *Biometrika 66*, 19-35.

BAHADUR R. R. (1961). A representation of the joint distribution of reponses to n dichotomous items. In H. Salomon, ed., *Studies in Item Analysis and Prediction*, Stanford, Calif. : Stanford University Press, 158-68.

BOWMAN A. W. (1980). A note on consistency of kernel method for the analyse of categorical data. *Biometrika 67*, 682-4.

BOWMAN A. W., HALL P. & TITTERINGTON D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika 71*, 341-51.

BROWN P. J. & RUNDELL P. W. K. (1985). Kernel estimates for categorical data. *Technometrics 27*, 293-9.

COX D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B 32*, 283-301.

DILLON W. R.& GOLDSTEIN M. (1978). On the performance of some multinomial classification rules. *Journal of American Statistical Association 73*, 305-313.

EFRON, B (1983). Estimating the error rate of a prediction rule : Improvement on cross-validation. *Journal of American Statistical Association* **78**, 316-331.

FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *Journal of American Statistical Association* **84**, 165-175.

GLICK N. (1972). Sample-based classification procedures derived from density estimates. *Journal of American Statistical Association* **67**, 116-21.

GOLDSTEIN M. & DILLON W. R. (1978). *Discrete discriminant analysis*. J. Wiley & Sons, New York.

HALL P. (1981a). On nonparametric multivariate binary discrimination. *Biometrika 68*, 287-94.

HALL P. (1981b). Optimal near neighbour estimator for use in discriminant analysis. *Biometrika 68*, 572-5.

HALL P. & WAND P. (1988). Nonparametric discrimination using density differences. *Biometrika 75*, 541-7.

HAND D. J. (1982). *Kernel discriminant analysis*. Chichester : Research Studies Press. Wiley.

HAND D. J. (1983). A comparative of two methods of discriminant anlysis applied to binary data. *Biometrics 39*, 683-94.

HAND D. J. (1986). Recent advance in error rate Estimation. *Pattern Rocognition Letters* **4**, 335-346.

HILLS M. (1967). Discrimination and allocation with discrete data. *Applied Statistics 16*, 237-250.

KRZANOWSKI W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of American Statistical Association 70*, 782-790.

MARTIN D. C. & BRADLEY R. A. (1972). Probability models, estimation, and classification for multivariate dichotomous populations. *Biometrics 28*, 203-22.

OTT J. & KRONMAL R. A. (1976). Some classification procedures for binary data using orthogonal functions. *Journal of American Statistical Association 71*, 391-99.

TITTERINGTON D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics 22*, 259-68.

TITTERINGTON D. M., MURRAY G. D., MURRAY L. S., SPIEGELHALTER D. J., SKENE A. M., HABBEMA J. D. F. & GELPKE G. J. (1981). Comparative of discrimination techniques applied to a computer data set of head injured patients. *Journal of the Royal Statistical Society* **A 144**, 145-175.

TITTERINGTON D. M & BOWMAN A. W. (1985). A comparative study of smoothing procedures for ordered categorical data. *Journal of Statistical Computation and Simulation 21*, 291-312.

TUTZ G. (1986). An alternative choice of smoothing for kernel-based density estimates in discrete discriminant analysis. *Biometrika 73*, 405-11.

TUTZ G. (1989). On cross-validation for discrete kernel estimation in discrimination. *Communication in Statistics-Theory and Methods 18* (**11**), 4145-4162.

ISSN 0249-6399