Modern Applied Science
www.ccsenet.org/journal.html
Vol. 3, No. 1
January 2009

# Strategies for Non-Parametric Smoothing of the Location Model in Mixed-Variable Discriminant Analysis

Nor Idayu Mahat (Corresponding author)
College of Arts and Sciences, Building of Faculty of Quantitative Sciences
Universiti Utara Malaysia
06010 Sintok, Kedah, Malaysia
Tel: 6-04-9284-098 E-mail: noridayu@uum.edu.my
W.J. Krzanowski
School of Engineering, Computer Science and Mathematics
University of Exeter
North Park Road, EX4 4QE Exeter, UK
Tel: 44-1392-725-279 E-mail: W.J.Krzanowski@ex.ac.uk
A. Hernandez
Escuela Universitaria de Estudios Empresariales, Universidad Complutense
Avda Filipinas 3, 28003 Madrid, Spain
Tel: 34-91-394-6746 E-mail: a.hernandez@emp.ucm.es

**Abstract**

The non-parametric smoothing of the location model proposed by Asparoukhov and Krzanowski (2000) for allocating objects with mixtures of variables into two groups is studied. The strategy for selecting the smoothing parameter through the maximisation of the pseudo-likelihood function is reviewed. Problems with previous methods are highlighted, and two alternative strategies are proposed. Some investigations into other possible smoothing procedures for estimating cell probabilities are discussed. A leave-one-out method is proposed for constructing the allocation rule and evaluating its performance by estimating the true error rate. Results of a numerical study on simulated data highlight the feasibility of the proposed allocation rule as well as its advantages over previous methods, and an example using real data is presented.

**Keywords:** Brier score, Error rate, Leave-one-out process, Location model, Pseudo-likelihood

## 1. Introduction

Various methods for constructing allocation rules in discriminant analysis with mixtures of variables have been proposed and discussed by researchers. Broadly speaking there are three possible strategies:

(i) transform the variables so they are all of the same type, and then apply an allocation rule appropriate to this type;

(ii) apply separate allocation rules to each type, and then combine the results for an overall classification;

(iii) develop a model that pays regard to the separate types, and then derive an allocation rule from this model.

Strategy (i) entails possible loss of information (Krzanowski, 1993; Hand, 1997); strategy (ii) has had limited study (Wernecke, 1992; Xu et al., 1992), but strategy (iii) has received much more wide-spread attention. Some methods available so far are non-parametric kernel and nearest neighbour approaches (reviewed by Silverman and Jones (1989)), semi-parametric methods such as logistic discriminant analysis (Anderson, 1972) and fully parametric methods based on the location model (Chang and Afifi, 1974; Krzanowski, 1975).

The use of the location model in discriminant analysis has been discussed by many researchers (Chang and Afifi, 1974; Krzanowski, 1975, 1980; Daudin, 1986; Knoke, 1982; Vlachonikolis and Marriot, 1982; Titterington et al., 1981; Kiang, 2003). In a recent development, Asparoukhov and Krzanowski (2000) have shown how non-parametric smoothing can be used to estimate the classifier's parameters. This approach is particularly useful for situations with sparse data, where traditional maximum likelihood methods run into problems.

The aim of this paper is to carry out further investigation on non-parametric smoothing of the location model. Attention is focused on classifying objects with continuous and binary variables to one of two groups, but expanding the idea to more than two groups and more general categorical variables can be executed without difficulty. Existing methodology is summarized in the second section, problems with previous methods are highlighted and some new ideas are presented in the third section, a Monte Carlo study to investigate the methods is described in the fourth section and results are presented in the fifth section. We also compare the proposed methods with standard maximum likelihood ones in situations where the latter are possible. An example of real data is then presented, and brief conclusions are given in the final section.

**2. Non-parametric smoothing of the location model**

Suppose there are two groups, $\pi_1$ and $\pi_2$, both of which consist of objects with binary and continuous variables. We denote the vector of binary variables as $\mathbf{x}$ and the vector of continuous variables as $\mathbf{y}$, and let the former have $q$ components, $\mathbf{x}^T = \{x_1, x_2, ..., x_q\}$ while the latter has $p$ components, $\mathbf{y}^T = \{y_1, y_2, ..., y_p\}$. Hence, we may present the vector of variables observed on each object in both groups as $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$.

Let each binary variable take a value of 0 or 1. Then we can express the $q$ binary variables as a multinomial $\mathbf{s} = \{s_1, ..., s_m\}$, where $m = 2^q$, and each distinct pattern of $\mathbf{x}$ defines a multinomial cell uniquely, with $\mathbf{x}$ falling in cell $s = 1 + \sum_{b=1}^{q} x_b 2^{b-1}$. We denote the probability of obtaining an object in cell $s$ of $\pi_i$ ($i$ = 1,2) by $p_{is}$. Next, we assume the vector of continuous variables to have a multivariate normal distribution with mean $\boldsymbol{\mu}_{is}$ in cell $s$ of $\pi_i$ and a homogeneous dispersion matrix across cells and populations, $\boldsymbol{\Sigma}$, thus $\mathbf{Y}_{is} \sim \mathbf{N}(\boldsymbol{\mu}_{is}, \boldsymbol{\Sigma})$. The joint probability of observing an object in group $\pi_i$ with associated $\mathbf{x}$ and $\mathbf{y}$ is

$$f(\mathbf{x}, \mathbf{y} \mid \pi_i) = \frac{p_{is}}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_{is})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{is}) \right]. \tag{1}$$

The application of this joint probability to the problem of allocating a future object with mixed variables to one of two groups was first studied by Chang and Afifi (1974) and generalised by Krzanowski (1975). By assuming that the costs due to misallocating future objects in both groups are equal and that the covariance structures in both groups are homogeneous, we allocate a future object with the vector of observed variables $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ to $\pi_1$ if its $\mathbf{x}$ falls in cell $s$ of the multinomial variable and

$$(\boldsymbol{\mu}_{1s} - \boldsymbol{\mu}_{2s})^T \boldsymbol{\Sigma}^{-1} \left\{ \mathbf{y} - \frac{1}{2} (\boldsymbol{\mu}_{1s} + \boldsymbol{\mu}_{2s}) \right\} \geq \log\left( \frac{p_{2s}}{p_{1s}} \right), \tag{2}$$

otherwise to $\pi_2$.

Usually, parameters $\boldsymbol{\mu}_{is}, \boldsymbol{\Sigma}$ and $p_{is}$ are unknown and have to be estimated from initial samples, known as training sets. We focus here on estimating these parameters using non-parametric smoothing methods. The mean of the $j$th continuous variable $y_j$ for cell $s$ of group $\pi_i$ is estimated through

$$\hat{\mu}_{isj} = \left\{ \sum_{k=1}^{m} n_{ik} w_{ij}(s,k) \right\}^{-1} \sum_{k=1}^{m} \left\{ w_{ij}(s,k) \sum_{r=1}^{n_{ik}} y_{rikj} \right\} \tag{3}$$

under conditions $0 \leq w_{ij}(s,k) \leq 1$ and $\sum_{k=1}^{m} n_{ik} w_{ij}(s,k) > 0$ where $s$, $k$ = 1, . . . , $m$; $i$ = 1, 2 and $j$ = 1, . . . , $p$. In this form, $n_{ik}$ is the number of objects of $\pi_i$ that fall in cell $k$, $y_{rijk}$ is the $j$th continuous variable value of the $r$th object falling in cell $k$ of $\pi_i$ and $w_{ij}(s,k)$ is the weight with respect to variable $j$ and cell $s$ of all objects of $\pi_i$ that fall in cell $k$.

One may consider any suitable function for the weights, $w_{ij}(s,k)$, but we prefer the exponential function due to its simple form of

$$w_{ij}(s,k) = \lambda_{ij}^{d(s,k)}; \quad 0 \le \lambda_{ij} \le 1 \tag{4}$$

where $d(s,k) \in \{0,1,...,q\}$. Here, *d(s, k)* is the dissimilarity coefficient between the *s*th cell and the *k*th cell of binary vectors, measured using distance function $d(\mathbf{x}_s, \mathbf{x}_k) = (\mathbf{x}_s - \mathbf{x}_k)^T (\mathbf{x}_s - \mathbf{x}_k)$. All cells that have equal dissimilarity with respect to cell *s* will thus have equal weight in the estimation of cell means.

In any practical application, the degree of smoothing represented by $\lambda_{ij}$ needs to be determined. One possible way of doing this is to select the set of smoothing parameters for continuous variables, **Λ**, as the values in the interval [0,1] that maximise the leave-one-out pseudo-likelihood function (Asparoukhov and Krzanowski, 2000)

$$\mathrm{PL}_{loo}(\mathbf{\Lambda} \mid D) = \prod_{r=1}^{n} p(\mathbf{y}_r \mid D - \mathbf{z}_r, \mathbf{\Lambda}) \tag{5}$$

where $p(\mathbf{y}_r \mid D - \mathbf{z}_r, \mathbf{\Lambda})$ is the probability density of $\mathbf{y}_r$ if object *r* falls in cell *s* of $\pi_i$ and $D - \mathbf{z}_r$ is the training set of $\pi_1$ and $\pi_2$ objects with object *r* excluded.

Next, the smoothed cell means (3) are used in the estimation of the smoothed covariance matrix

$$\mathbf{V} = \frac{1}{n_1 + n_2 - g_1 - g_2} \sum_{i=1}^{2} \sum_{s=1}^{m} \sum_{r=1}^{n_{is}} (\mathbf{y}_{ris} - \hat{\boldsymbol{\mu}}_{is})(\mathbf{y}_{ris} - \hat{\boldsymbol{\mu}}_{is})^T \tag{6}$$

where $n_i$ is the number of objects of $\pi_i$ and $g_i$ is the number of non-empty cells in the training sets from $\pi_i$. Finally, Asparoukhov and Krzanowski (2000) employed the adaptive weighted near-neighbour estimators as originally proposed by Hall (1981), for estimating the cell probabilities. These estimators have the form

$$\hat{p}_{is} = n_i^{-1} \sum_{j=0}^{L} w_{ij} N_{ij}(s,k); \quad 0 \le L \le q-1 \tag{7}$$

where *s*, *k* = 1, . . . , *m*; *i* = 1, 2 and $N_{ij}(s,k)$ is the number of training objects from $\pi_i$ that fall in cell *k* and are *j* binary variables distant from the cell *s*, $d(\mathbf{x}_s, \mathbf{x}_k) = j$. The weights, $w_{ij}$, are chosen to minimise mean squared error:

$$\Delta_i(w_{i0}, w_{i1}, ..., w_{iL}) = \sum_{s=1}^{m} E(\hat{p}_{is} - p_{is})^2 . \tag{8}$$

The proposed strategies for estimating the parameters are satisfactory, but some potential problems may arise in certain situations. The first problem concerns the choice of smoothing parameters by maximising the leave-one-out pseudo-likelihood function (5). This function may be satisfactory for data from normal populations, but less satisfactory if this assumption is not appropriate. Secondly, the estimation of cell probabilities using the adaptive weighted nearest neighbour method determines the amount of smoothing by computing the weights automatically. It is an easy technique, but Hall (1981) warned that sometimes these weights take negative or zero values. These phenomena usually happen when the probabilities are small and the sample is not sufficiently large. These problems are addressed in the rest of this paper, and some alternative methods are investigated.

## 3. Some modifications

### *3.1 Parameter estimations*

The sample-based version of allocation rule (2) that assumes equal cost of misclassifying objects in both groups and homogeneous covariance matrix in the two groups, will allocate a future object, $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ to $\pi_1$ if it satisfies

$$(\hat{\boldsymbol{\mu}}_{1s} - \hat{\boldsymbol{\mu}}_{2s})^T \mathbf{V}^{-1} \left\{ \mathbf{y} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_{1s} + \hat{\boldsymbol{\mu}}_{2s}) \right\} \ge \log\left( \frac{\hat{p}_{2s}}{\hat{p}_{1s}} \right) \tag{9}$$

and otherwise to $\pi_2$, where $\hat{\boldsymbol{\mu}}_{1s}, \hat{\boldsymbol{\mu}}_{2s}, \mathbf{V}, \hat{p}_{1s}$ and $\hat{p}_{2s}$ are the smoothed parameter estimates obtained from the training sets.

We keep both the smoothed estimators $\hat{\boldsymbol{\mu}}_{is}$ and $\mathbf{V}$ as given in equations (3) and (6) respectively. However, we apply a restriction by having a single smoothing parameter, λ, across all continuous variables and groups. Therefore, a new

exponential function is

$$w(s,k) = \lambda^{d(s,k)}; \quad 0 < \lambda < 1 \tag{10}$$

and we use this quantity for estimating the cell means (3). The reason for imposing a single smoothing parameter is to have less complexity on the designed model and greater ease in handling the process for selecting the smoothing parameter.

We overcome the problem of obtaining negative or zero cell probabilities when using the adaptive weighted nearest neighbour (7) by imposing the following restrictions.

(i) All weights, $(w_{i0},...,w_{iL})$ should be between 0 and 1, so that any weight that is less than or equal to 0 is replaced by 0.00001 while any weight that is greater than or equal to 1 is replaced by 0.99999.

(ii) If restriction (i) fails to avoid obtaining a zero probability, then any cell probability that has zero value is replaced by 0.00001.

In addition, we consider two alternative non-parametric smoothing methods for estimating $p_{is}$. These methods are;

3.1.1 The kernel method (Aitchison and Aitken, 1976) that estimates the probability of observing cell *s* of $\pi_i$ as

$$\hat{p}_{is} = n_i^{-1}\lambda^q \sum_{j=0}^{q} N_{ij}(s,k)\left\{\frac{1-\lambda}{\lambda}\right\}^j; \quad (1/2 \le \lambda \le 1) \tag{11}$$

where $N_{ij}(s,k)$ has the same interpretation as the adaptive weighted near-neighbour estimator which was discussed earlier and $\lambda$ is a smoothing parameter.

3.1.2 The modified maximum likelihood estimator in the weighted form:

$$\hat{p}_{is} = \frac{\sum_{k=1}^{m} w(s,k)n_{is}}{n_i}. \tag{12}$$

In order to have the simplest allocation rule, we suggest that the weights $w(s,k)$ follow the exponential function (10), standardised to probabilities in each group by

$$\hat{p}_{is(std)} = \hat{p}_{is} / \sum_{s=1}^{m} \hat{p}_{is}. \tag{13}$$

We will term this the exponential method.

*3.2 Selecting optimised smoothing parameters*

Both the kernel and the exponential methods (11) and (13), as well as the estimation of cell means (3), need an identified value of the smoothing parameter, $\lambda$. Therefore, we investigate three different strategies for obtaining its value. The $\lambda$ that satisfies a given criterion in each strategy is termed $\lambda_{opt}$.

The first strategy is to select $\lambda_{opt}$ through the maximisation of the leave-one-out pseudo-likelihood function. This was successfully applied by Asparoukhov and Krzanowski (2000) and we are interested in studying its feasibility. The second and the third strategies choose $\lambda_{opt}$ as the value that provides the best allocation performance as measured by the true error rate and the Brier score respectively. The choice of $\lambda_{opt}$ based on the true error rate was used by Raudys and Jain (1991), who selected $\lambda_{opt}$ for the Kernel window classier as the value that provides the smallest classification errors. The error rate is computed by taking the proportion of the number of objects misclassified by the rule to the total number of objects in the sample.

The error rate takes a discrete value, so an alternative criterion that takes values from a continuum is considered. This criterion is known as the Brier score. Suppose $\delta(\pi_i \mid \mathbf{x}_r, \mathbf{y}_r)$ denotes the true group of object $r$ in the training set, being equal to 1 if $r$ is from $\pi_i$ and 0 otherwise, and $f(\pi_i \mid \mathbf{x}_r, \mathbf{y}_r)$ is the probability that object $r$ with measurements $\mathbf{x}_r$ and $\mathbf{y}_r$ belongs to $\pi_i$. Then for the two groups case, the Brier score is defined as (Hand, 1997)

$$\frac{1}{n}\sum_{r=1}^{n}\sum_{i=1}^{2}\{\delta(\pi_i \mid \mathbf{x}_r, \mathbf{y}_r) - f(\pi_i \mid \mathbf{x}_r, \mathbf{y}_r)\}^2 . \tag{14}$$

To have a good estimate of the Brier score, an appropriate method that gives the best estimate of $f(\pi_i \mid \mathbf{x}_r, \mathbf{y}_r)$ has to be used. If the group densities $f(\mathbf{x}_r, \mathbf{y}_r \mid \pi_i)$ are known to follow the location model, we may obtain $f(\pi_i \mid \mathbf{x}_r, \mathbf{y}_r)$ easily through *Bayes theorem*:

$$f(\pi_i \mid \mathbf{x}_r, \mathbf{y}_r) = [p_i f(\mathbf{x}_r, \mathbf{y}_r \mid \pi_i)]\left\{\sum_{i=1}^{2} p_i f(\mathbf{x}_r, \mathbf{y}_r \mid \pi_i)\right\}^{-1} \tag{15}$$

where $p_i$ is the prior probability of obtaining an object from $\pi_i$ and

$$f(\mathbf{x}_r, \mathbf{y}_r \mid \pi_i) = \frac{p_{is}}{(2\pi)^{p/2} \mid \mathbf{\Sigma} \mid^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y}_r - \boldsymbol{\mu}_{is})^T \mathbf{\Sigma}^{-1}(\mathbf{y}_r - \boldsymbol{\mu}_{is})\right]. \tag{16}$$

We estimate both (15) and (16) by replacing all the parameters ($\boldsymbol{\mu}_{is}, \mathbf{\Sigma}$ and $p_{is}$) with the values of the corresponding smoothed estimators ($\hat{\boldsymbol{\mu}}_{is}, \mathbf{V}$ and $\hat{p}_{is}$) obtained from the training set, choosing prior probabilities appropriate to the substantive application. Like error rate, we choose the $\lambda_{opt}$ that provides the smallest value of Brier score. The computation of both error rate and Brier score are performed through the leave-one-out process, so that biased estimation of these criteria is avoided and comparison among these three strategies for selecting $\lambda_{opt}$ is done without bias.

*3.3 Assessment of smoothed allocation rule*

There are nine possible allocation rules altogether, obtained from the combination of the three different methods for smoothing cell probabilities and the three different strategies to choose $\lambda_{opt}$. In each of these rules the exponential smoothing method is used for estimating the cell means and the covariance matrix. We compare these rules by measuring their performance using misclassification of future objects as the criterion. For this purpose, the leave-one-out error rate is preferred because it removes the bias in the apparent error rate and easily represents a classifier's inaccuracy by giving the percentage of objects it misclassifies.

The methods for selecting $\lambda_{opt}$ and evaluating rule performance both need the leave-one-out process. We conduct these leave-one-out processes in nested fashion; the *inner* leave-one-out is used to select the optimised smoothing parameter and the *outer* leave-one-out is used to measure rule performance. The details of the proposed design are outlined as follows:

1. Omit each object *r* in turn from the sample ($r = 1,..,n$).

    1.1 Determine $\lambda_{opt}$ from the sample without object *r*. Choose $\lambda_{opt}$ that

        1.1.1 maximises the leave-one-out log-likelihood function, or

        1.1.2 minimises the leave-one-out error rate, or

        1.1.3 minimises the leave-one-out Brier score.

of the (*n* - 1) remaining object.

    1.2 Compute $\hat{\boldsymbol{\mu}}_{is}, \mathbf{V}$ and $\hat{p}_{is}$ using the obtained value of $\lambda_{opt}$ and the sample without object *r*.

    1.3 Construct an allocation rule using the estimates obtained in (1.2).

    1.4 Predict the group of object *r*. If correct, then $\text{error}_\text{r} = 0$, otherwise $\text{error}_\text{r} = 1$.

2. Compute the leave-one-out error rate, $\sum_{r=1}^{n} \text{error}_r / n$.

For option (1.1.2) in step 1.1, we evaluate the allocation rule performance over a set of values of $\lambda$ in (0, 1) and choose the value which provides the smallest leave-one-out error rate as $\lambda_{opt}$. However, we obtain $\lambda_{opt}$ for both options (1.1.1) and (1.1.3) using built-in functions ('*optimize*' and '*nlminb*') as provided in S-Plus. It is worth

emphasising that the optimised value obtained from any of these strategies is not necessarily the global optimum.

Since there are nine rules, we distinguish them by labelling each rule with the smoothing methods used to smooth the cell probabilities, thus we have (i) Nearest neighbour$_{LL}$, Nearest neighbour$_{ER}$ and Nearest neighbour$_{BS}$, (ii) Kernel$_{LL}$, Kernel$_{ER}$ and Kernel$_{BS}$ and (iii) Exponential$_{LL}$, Exponential$_{ER}$ and Exponential$_{BS}$. The subscripts refer to the strategies (log-likelihood, error rate and Brier score) for obtaining the optimised smoothing parameter, $\lambda_{opt}$.

**4. Monte Carlo simulation**

The rules discussed in the previous section were evaluated through simulations. We generated data having a mixture of continuous and binary variables by following Everitt and Merette (1990). Each object in group $\pi_i$ is first assumed to have $p$+$q$ continuous random variables, $y_{i1},...,y_{ip},y_{i(p+1)},...,y_{i(p+q)}$, taken from a multivariate normal distribution with mean $\boldsymbol{\mu}_i$ and homogeneous dispersion matrix, $\boldsymbol{\Sigma}$. The first $p$ variables $y_{i1},...,y_{ip}$ are treated as observed variables, and the remaining $y_{i(p+1)},...,y_{i(p+q)}$ as unobserved variables. The binary variables arise by discretising the latter continuous variables. This is done by converting $y_{i(p+1)},...,y_{i(p+q)}$ to $x_{i1},...,x_{iq}$ using

$$x_{ij} = \begin{cases} 0 & \text{if} \quad -\infty \le y_{i(p+j)} < \delta_{ij} \\ 1 & \text{if} \quad \delta_{ij} \le y_{i(p+j)} < \infty \end{cases} \tag{17}$$

where $\delta_{ij}$ are thresholds, $i$ = 1,2 and $j$ = 1,…,$q$. These thresholds are given by

$$\delta_{ij} = \Phi^{-1}(p_{ij}) \times \sigma_{y_{i(p+j)}} + \mu_{y_{i(p+j)}}$$

where $p_{ij}$ is the target proportion of objects of $\pi_i$ having $x_{ij} = 0$, $\mu_{y_{i(p+j)}}$ and $\sigma_{y_{i(p+j)}}$ are the mean and standard deviation of observed variable $y_{p+j}$ in $\pi_i$ respectively, and $\Phi^{-1}(.)$ is the inverse cumulative standard normal integral. Assuming that the mean of $\mathbf{y}$ is $\mathbf{0}$ and all variances are unity, then $\delta_{ij}$ is equal to the standard normal ordinate corresponding to the target proportion.

The dispersion matrix, $\boldsymbol{\Sigma}$, unites the dispersion matrix of the continuous variables, $\boldsymbol{\Sigma}_p$, the dispersion matrix of the binary variables, $\boldsymbol{\Sigma}_q$, and the covariance between these two variables, $\boldsymbol{\Sigma}_{pq}$, such that

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_p & \boldsymbol{\Sigma}_{pq} \\ \boldsymbol{\Sigma}_{pq}^T & \boldsymbol{\Sigma}_q \end{pmatrix}$$

The set of parameters for generating artificial data sets is

$n_i$ = sample size of $\pi_i$

$p$ = number of continuous variables

$q$ = number of binary variables

$\boldsymbol{\mu}_i$ = ( $\mu_{i1},...,\mu_{ip},\mu_{i(p+1)},...,\mu_{i(p+q)}$ )

the vector of means in $\pi_i$ where $\mu_{i(p+1)} = ... = \mu_{i(p+q)} = 0$

$\boldsymbol{\Sigma}$ = dispersion matrix

$i$ = 1, 2

We set the following values

$n_i = 25,50 \text{ and } 100 \qquad p = 3 \qquad q = 2$

$\boldsymbol{\mu}_1 = (1,0,0,0,0) \quad \boldsymbol{\mu}_2 = (1.5,1.5,1.5,0,0)$

For the dispersion matrix, $\boldsymbol{\Sigma}$, we defined two different matrices as follows:

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 & 0.2 & 0 \\ 0 & 1 & 0 & 0.5 & 0.5 \\ 0 & 0 & 1 & 0 & 0.4 \\ 0.2 & 0.5 & 0 & 1 & 0.5 \\ 0 & 0.5 & 0.4 & 0.5 & 1 \end{pmatrix}, \quad \mathbf{\Sigma}_2 = \begin{pmatrix} 16 & 0 & 0 & 0.2 & 0 \\ 0 & 15 & 0 & 0.5 & 0.5 \\ 0 & 0 & 12 & 0 & 0.4 \\ 0.2 & 0.5 & 0 & 1 & 0.5 \\ 0 & 0.5 & 0.4 & 0.5 & 1 \end{pmatrix}$$

The differences between these two matrices is in their settings of $\mathbf{\Sigma}_p$ where the variances of the continuous variables are very small (unity) in the first matrix, whilst they are very dispersed in the second matrix. Finally, the largest proportions of the binary variable values are given in Table 1.

SET 1 in Table 1 gives large separation of cells, so we will have more objects in one category than in the other category for the same binary variable. Therefore, we expect to obtain many empty cells with this set especially when the size of sample is very small. SET 2 and SET 3 give moderate and small separation of cells respectively. In SET 3, full cells are expected. There are 18 combinations of (i) size of groups, $n_i = 25, 50 \text{ or } 100$, (ii) dispersion matrices, $\mathbf{\Sigma}_1$ or $\mathbf{\Sigma}_2$ and (iii) proportions in category of binary variables (see Table 1). Since the simulation process is extensive, we only had one artificial data set for each combination and we restricted our investigation to have equal sample size in both groups for these 18 data sets.

## 5. Results for simulated data

### *5.1 Selecting optimised smoothing parameters*

Figure 1 shows the variability of $\lambda_{opt}$ across one data set for the three allocation rules when they are chosen to minimise the leave-one-out Brier score. These examples are taken from data set 4, one of the 18 data sets generated in the previous section. The patterns of other data sets and the patterns that correspond to the other two optimisation criteria were similar.

Figure 1(a) illustrates the variability of $\lambda_{opt}$ for smoothing cell means using the nearest neighbour rule while Figure 1(c) illustrates the variability of $\lambda_{opt}$ for smoothing cell means and cell probabilities using the exponential rule. In both plots, the *Y*-axis represents the value of $\lambda_{opt}$ and the *X*-axis represents the object *r*. Each point in these plots indicates the smoothing parameter value for the allocation rule constructed from the sample on omitting object *r*. The kernel rule applies two different smoothing parameters, one each for smoothing cell means and cell probabilities. So, we plot the corresponding $\lambda_{opt}$ in the *Y*-axis and *X*-axis (Figure 1(b)). Each point in this plot represents the sample on omitting object *r*. All these plots show a clear scatter of points especially in Figure 1(c). They describe some differences in the amount of smoothing from one sample of size ($n - 1$) to other samples of the same size. This variability occurs because different samples have slightly different distributions when a different object *r* is omitted.

We have labelled several points that are further from the majority by their omitted objects (Figure 1(a) and Figure 1(c)) and circled numbers (Figure 1(b)), to highlight the possibility of having potential outliers. This is quite interesting to highlight because we can see in each plot that different methods identify different potential outliers.

### *5.2 Performance of the smoothed allocation rules for normal populations*

The performances of the smoothed allocation rules are depicted in Figure 2. Each plot displays the behaviour of the performance of each smoothed allocation rule, estimated through the leave-one-out error rate (*Y*-axis) and its relation to the distance between two groups measured through the Kullback-Leibler divergence (*X*-axis). The three lines in each plot represent the three different strategies for selecting $\lambda_{opt}$. As clearly seen, all plots show the same decreasing pattern when the distance between two groups increases. It is quite difficult to spot whether any single rule always gives the smallest estimated error rate when comparing all Figures 2(a), 2(b) and 2(c), so Table 2 shows the frequencies with which each allocation rule was best.

In this table, the best allocation rule refers to the one that gives the lowest error rate among the three for each of the 18 artificial data sets. The scoring system is as follows; we assign 1 point to a single best rule, 1/2 point to both rules if there are two winners, and 1/3 for all rules if all of them are the lowest. Among the smoothing methods, the nearest neighbour rule is clearly best when its $\lambda_{opt}$ is chosen through the minimisation of the leave-one-out Brier score and the maximisation of the leave-one-out log-likelihood function, while the kernel rule is the winner when its $\lambda_{opt}$ is chosen through the minimisation of the leave-one-out error rate. However, in terms of this criterion all three methods perform very similarly.

*5.3 Investigation with non-normal populations*

The rules where $\lambda_{opt}$ is obtained through the maximisation of the normal log-likelihood function should show some deterioration when the data are in fact non-normal. In such a case, their performance should differ from those of the other rules because the other two criteria (i.e. optimising error rate and Brier score) do not rely on distributional assumptions.

To investigate the performance of the proposed rules under non-normal conditions, we chose randomly five out of the 18 generated data sets and transformed all continuous variables in these subsets using the inverse of Johnson's systems (Lachenbruch et al., 1973) so that they no longer had a multivariate normal distribution. The chosen transformations were: log-normal (data sets 1 and 8), logit (data sets 2 and 14) and hyperbolic sine normal function (data set 17). Details on these transformation can be found in Lachenbruch et al. (1973). To avoid complexity, the same transformation function was applied to the two groups. The same investigations as performed for normal populations were repeated, and results are shown in Table 3.

Our results do not show the expected effect, possibly because of the estimation of parameters, in which they have been smoothed for density estimation purposes. However, it is strongly advised that extra precautions need to be taken for dealing with this situation, and choosing smoothing parameter through inaccuracy measures is advised.

*5.4 Comparison of maximum likelihood and smoothed classification rules*

Smoothing is essential whenever some of the multinomial cells are either empty or have very few observations. In the former case it is impossible to construct the maximum likelihood classification rule in such cells, while in the latter case the rule will be very poorly estimated. However, it is also of interest to enquire how the smoothed rules will compare with maximum likelihood ones when all cells have sufficient numbers of objects in them. If the smoothed rules show good performance in this comparison, they can be applied very generally rather than only in cases where cell membership is sparse.

To investigate this question, we extracted from the foregoing data sets those that had observations in all the multinomial cells; there were 7 such sets among the normal ones and 3 among the non-normal ones. For each data set, the error rate and the Brier score were found for the maximum likelihood (ML) rule and for each of the three smoothed rules.

Results are shown in Table 4, where we give the number of sets in which the ML rule performed better than all the smoothed rule ("ML best"), the number in which all the smoothed rules performed better than ML ("ML worst"), and the number in which either one or two smoothed rules performed better than ML ("ML intermediate"). Noting that ML was beaten by at least one smoothed rule on both assessment criteria in nine of the ten data sets, these results therefore suggest that the smoothed rules constitute good classifiers whether the multinomial data are sparse or not.

## 6. Practical example

An example of using the proposed rules for a real problem is provided by data set 2 reported by Krzanowski (1975). These data concern patients suffering from jaundice. These patients comprise two groups: $\pi_1$ denotes patients who required medical treatment and $\pi_2$ denotes patients who required surgical treatment. There are 30 patients in $\pi_1$ and 63 patients in $\pi_2$. Four continuous and three binary variables are available for distinguishing these two groups. Since the costs of misclassifying patients to wrong groups are unavailable, we assume equal costs and prior probabilities.

The performance of the proposed allocation rules is shown in Table 5. Since there are many cells without patients, no allocation rule can be derived through the maximum likelihood estimation. Among smoothed allocation rules, Nearest neighbour$_{LL}$ is best. In fact, the numbers of misclassifications recorded by nearest neighbour rules are smaller than those for kernel and exponential rules. These findings are consistent with the ones in Table 2, where the nearest neighbour rules are best, followed by the kernel rules and finally the exponential rules.

## 7. Discussion

The results that we have obtained through simulation study and numerical example show that the choice of the optimised smoothing parameter depends on both the criterion used for the optimisation and the smoothing method. In respect of criterion, we recommend selecting $\lambda_{opt}$ through minimisation of the leave-one-out Brier score due to its continuous nature and since it does not make any distributional assumptions. The continuous nature makes it more amenable to standard function optimisation routines, and avoids the problem of having many local optima with discrete error rate. Lack of distributional assumptions circumvents potential objections to the use of log-likelihood ratio.

As regards smoothing method, our findings show that the nearest neighbour rule is the best and is followed by the kernel rule. This situation may relate to the lower restriction of the smoothing applied by them. Despite always being the best, however, nearest neighbour needs to be handled carefully because it is sensitive to the occurrence of cells without objects. On the other hand, the kernel method is heavy in computational time relative to its competitors. Alternatively, one may choose the exponential method if using a single smoothing parameter is preferred. It is possible to use this method because its results are not much poorer than the others.

As a final comment, the use of smoothing methods in the location model should be based on the problem in hand, rather than being a mere tool of the estimation method.

**References**

Aitchison, J. & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63, 413-420.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59, 19-35.

Asparoukhov, O. & Krzanowski, W. J. (2000). Non-parametric smoothing of the location model in mixed variable discrimination. *Statistics and Computing*, 10, 289-297.

Chang, P. C. & Afifi, A. A (1974). Classification based on dichotomous and continuous variables. *Journal of the American Statistical Association*, 69, 336-339.

Daudin, J. J. (1986). Selection of variables in mixed-variable discriminant analysis. *Biometrics*, 42, 473-481.

Everitt, B. S. & Merette, C. (1990). The clustering of mixed-mode data: A comparison of possible approaches. *Journal of Applied Statistics*, 17, 283-297.

Hall, P. (1981). Optimal near neighbour estimator for use in discriminant analysis. *Biometrika*, 68, 572-575.

Hand, D. J. (1997). Construction and assessment of classification rules. Chichester: John Wiley & Son.

Kiang, M. Y. (2003). A comparative assessment of classification methods. *Decision Support Systems*, 35, 441-454.

Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, 38, 191-200.

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

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 36, 493-499.

Krzanowski, W. J. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, 10, 25-49.

Lachenbruch, P. A., Sneeringer, C. & Revo, L. T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics*, 1, 39-56.

Raudys, S. J. & A. K. Jain (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Systems, Man and Cybernetics*, 13, 252-264.

Silverman, B. W. & M. C. Jones (1989). E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review*, 57, 233-247.

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F.

& Gelpke, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society*, 144, 145-175.

Vlachonikolis, I. G. & Marriot, F. H. C. (1982). Discriminant analysis with mixed binary and continuous data. *Applied Statistics*, 31, 23-31.

Wernecke, K. D. (1992). A coupling procedure for the discrimination of mixed data. *Biometrics*, 48, 497-506.

Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22, 418-435.

Table 1. Proportions in categories of binary variables.

| | Group | Proportions in category (binary variables, $x_{ij}$ ) | | | |
|---|---|---|---|---|---|
| | | $P(x_{i1}=0)$ | $P(x_{i1}=1)$ | $P(x_{i2}=0)$ | $P(x_{i2}=1)$ |
| SET 1 | 1 | 0.07 | 0.93 | 0.90 | 0.10 |
| | 2 | 0.85 | 0.15 | 0.05 | 0.95 |
| SET 2 | 1 | 0.75 | 0.25 | 0.40 | 0.60 |
| | 2 | 0.30 | 0.70 | 0.15 | 0.85 |
| SET 3 | 1 | 0.50 | 0.50 | 0.65 | 0.35 |
| | 2 | 0.40 | 0.60 | 0.55 | 0.45 |

Table 2. Frequency of being a winner among smoothing methods based on leave-one-out criteria.

| Leave-one-out optimisation criteria | Smoothing method | | |
|---|---|---|---|
| | Nearest neighbour | Kernel | Exponential |
| Error rate | 6.00 | 7.00 | 5.00 |
| Brier score | 10.33 | 4.33 | 3.83 |
| Log-likelihood | 8.17 | 4.67 | 5.17 |

Table 3. Performance of the rules for non-normal populations.

| Data set | Optimisation method for selecting $\lambda_{opt}$ | | |
|---|---|---|---|
| | Error rate | Brier score | Normal log-likelihood |
| Allocation rule: Nearest neighbour | | | |
| 1 | 0.06 | 0.06 | 0.06 |
| 2 | 0.06 | 0.05 | 0.05 |
| 8 | 0.23 | 0.20 | 0.21 |
| 14 | 0.31 | 0.30 | 0.31 |
| 17 | 0.46 | 0.47 | 0.46 |
| Allocation rule: Kernel | | | |
| 1 | 0.04 | 0.06 | 0.06 |
| 2 | 0.04 | 0.05 | 0.05 |
| 8 | 0.21 | 0.19 | 0.21 |
| 14 | 0.29 | 0.30 | 0.30 |
| 17 | 0.47 | 0.47 | 0.47 |

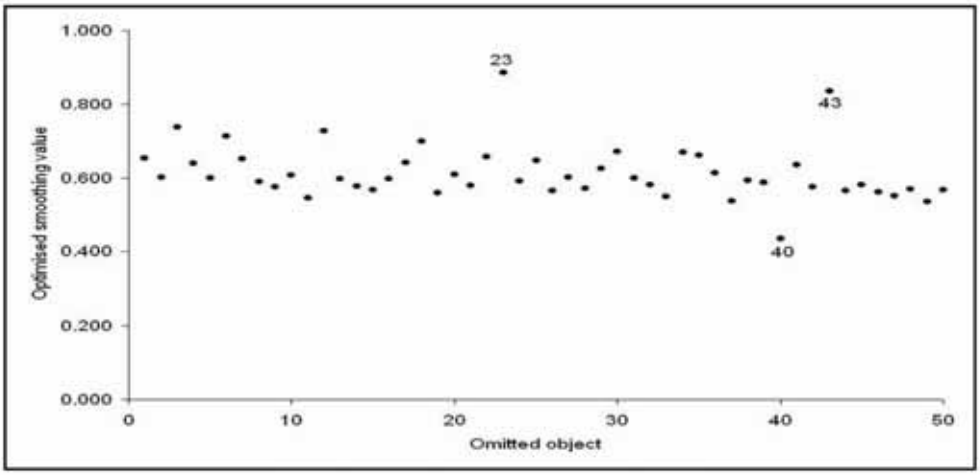
| Allocation rule: Exponential | | | |
|---|---|---|---|
| 1 | 0.06 | 0.06 | 0.06 |
| 2 | 0.04 | 0.05 | 0.05 |
| 8 | 0.22 | 0.20 | 0.22 |
| 14 | 0.32 | 0.31 | 0.33 |
| 17 | 0.49 | 0.48 | 0.45 |

Table 4. Comparison of the performance of the ML rule with those of the smoothed rules.

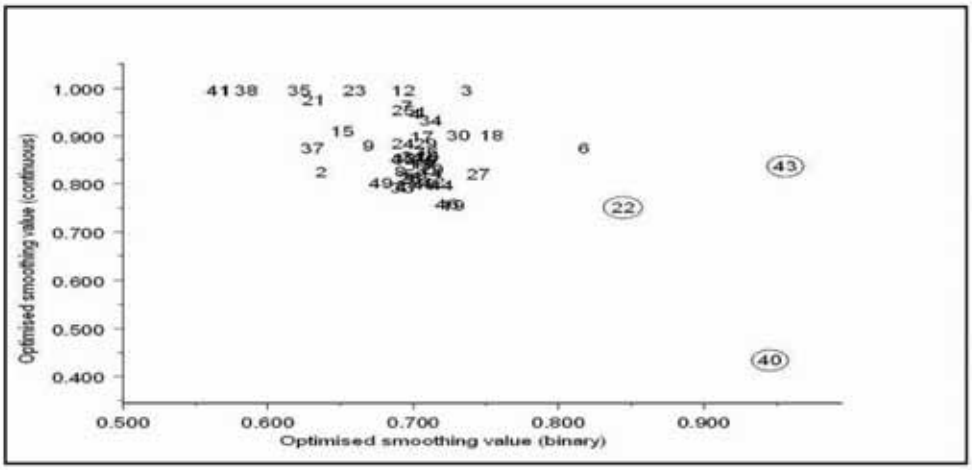
| Criterion | Performance | Normal sets | Non-normal sets |
|---|---|---|---|
| Error rate | ML best | 0 | 1 |
| | ML intermediate | 4 | 2 |
| | ML worst | 3 | 0 |
| Brier score | ML best | 0 | 0 |
| | ML intermediate | 2 | 1 |
| | ML worst | 5 | 2 |

Table 5. Location model assessment for jaundice patients.

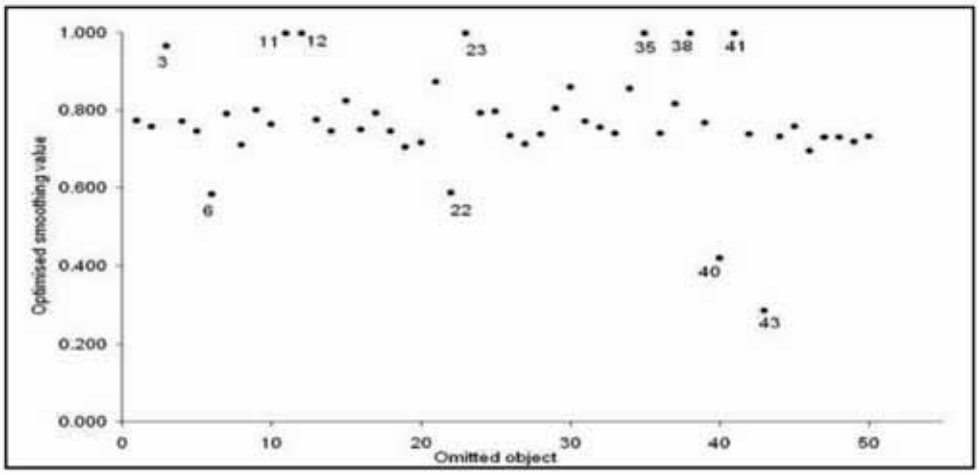
| Procedure | Number of misclassifications |
|---|---|
| Maximum Likelihood Estimation | - |
| Nearest neighbour$_{ER}$ | 25 |
| Nearest neighbour$_{BS}$ | 24 |
| Nearest neighbour$_{LL}$ | 24 |
| Kernel$_{ER}$ | 28 |
| Kernel$_{BS}$ | 27 |
| Kernel$_{LL}$ | 26 |
| Exponential$_{ER}$ | 28 |
| Exponential$_{BS}$ | 26 |
| Exponential$_{LL}$ | 29 |




(a)



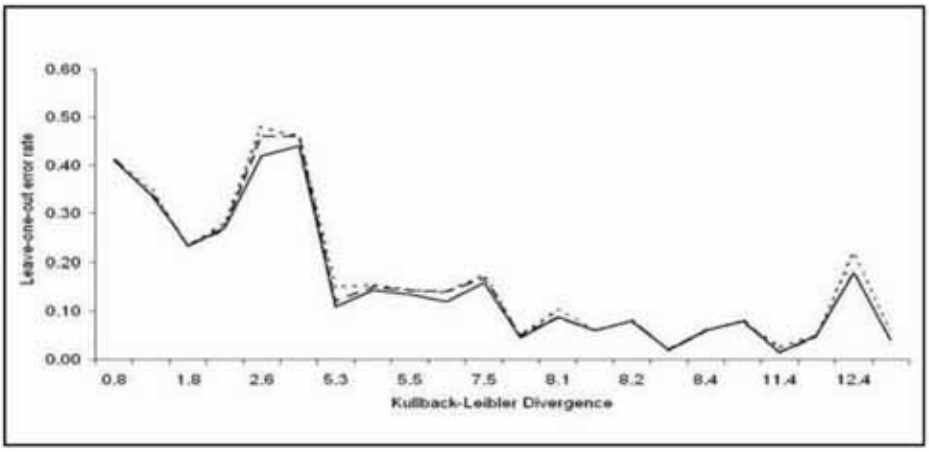

(b)



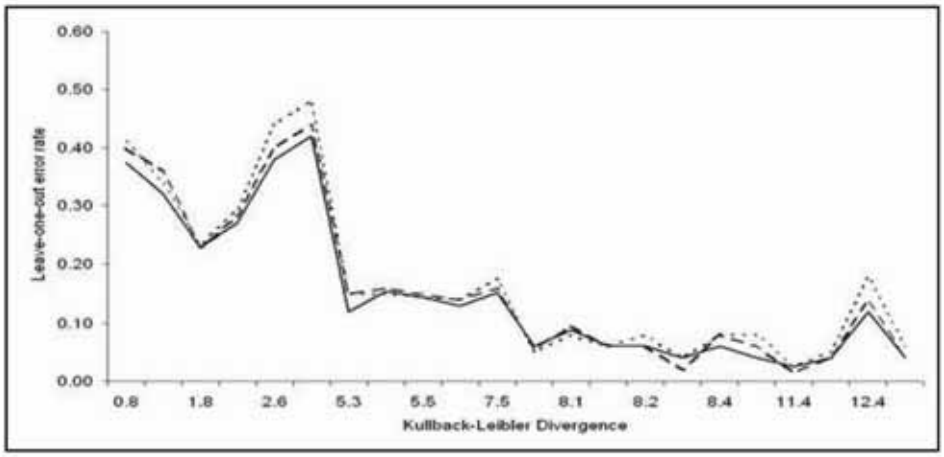

(c)

Figure 1. Example: $\lambda_{opt}$ that minimises the leave-one-out Brier score

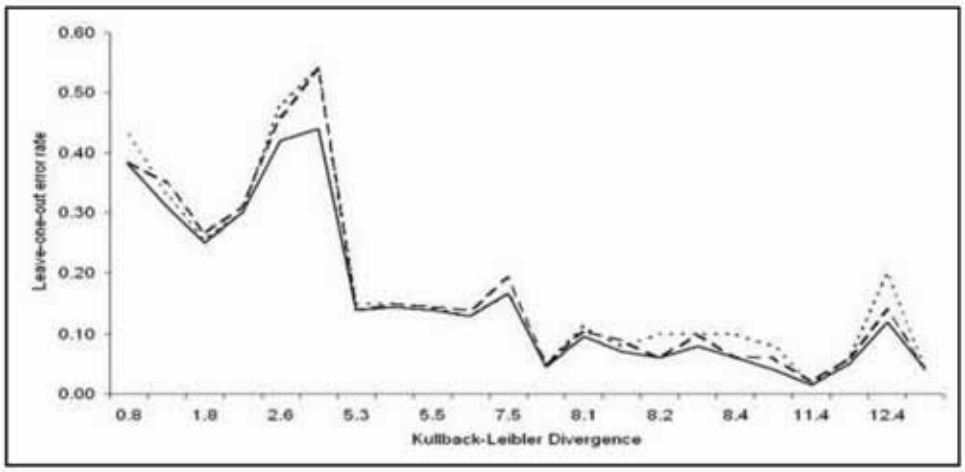
for (a) Nearest neighbour$_{BS}$, (b) Kernel$_{BS}$ and (c) Exponential$_{BS}$.




(a) Nearest neighbour smoothing of the location model.




(b) Kernel smoothing of the location model.




(c) Exponential smoothing of the location model.

Figure 2. Performance of the smoothed allocation rules measured through leave-one-out error rate. $\lambda_{oopt}$ chosen through the optimisation of the leave-one-out …. log-likehood function, ---- Brier score and ____error rate.