ADAC (2007) 1:105–122
DOI 10.1007/s11634-007-0009-9

REGULAR ARTICLE

Variable selection in discriminant analysis based on the location model for mixed variables

Nor Idayu Mahat · Wojtek Janusz Krzanowski · Adolfo Hernandez

Received: 19 October 2006 / Revised: 30 May 2007 / Accepted: 14 June 2007 /
Published online: 17 July 2007

Abstract Non-parametric smoothing of the location model is a potential basis for discriminating between groups of objects using mixtures of continuous and categorical variables simultaneously. However, it may lead to unreliable estimates of parameters when too many variables are involved. This paper proposes a method for performing variable selection on the basis of distance between groups as measured by smoothed Kullback–Leibler divergence. Searching strategies using forward, backward and stepwise selections are outlined, and corresponding stopping rules derived from asymptotic distributional results are proposed. Results from a Monte Carlo study demonstrate the feasibility of the method. Examples on real data show that the method is generally competitive with, and sometimes is better than, other existing classification methods.

Keywords Brier score · Cross-validation · Discriminant analysis · Error rate · Kullback-Leibler divergence · Location model · Non-parametric smoothing procedures · Variable selection

Mathematics Subject Classification (2000) 62H30

N. I. Mahat (✉)
Faculty of Quantitative Sciences, Universiti Utara Malaysia,
06010 Sintok, Kedah, Malaysia
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
e-mail: W.J.Krzanowski@ex.ac.uk

A. Hernandez
Escuela Universitaria de Estudios Empresariales, Universidad Complutense,
Avda Filipinas 3, 28003 Madrid, Spain
e-mail: a.hernandez@emp.ucm.es

1 Introduction

In this paper we consider the problem of classifying an individual into one of two classes (populations) π_1, π_2, on the basis of a data vector $\mathbf{z}$ that contains both continuous and categorical variables. The location model was introduced by Olkin and Tate (1961) to handle such vectors, and this model was later used by Chang and Afifi (1974) and Krzanowski (1975) to form a suitable classification rule in the context of discriminant analysis.

To conduct discriminant analysis based on the location model, cells of a multinomial table must be generated from the categorical values in each group. Then, the conditional distributions of the continuous variables given values of the categorical variables are estimated from the data and an object is allocated to a group on the basis of these distributions. The number of cells grows exponentially with the number of categorical variables. Therefore in practice, unless the size of samples is large, some cells may be empty. This prevents the use of maximum likelihood estimation and limits the feasibility of the linear model approach to cases with few categorical variables. Asparoukhov and Krzanowski (2000) proposed non-parametric smoothed estimation that rectifies these weaknesses and Mahat (2006) conducted further investigations on different smoothing procedures. However, this method may obtain inaccurate estimated parameters when too many variables have been used.

This problem can be overcome by performing variable selection where a subset of variables is selected from all the observed variables, and then using the subset to construct a classification rule. Using fewer variables may be beneficial to avoid estimation problems, to improve classification performance (McLachlan 1992, p. 389), to save some costs of computation and to facilitate the interpretation of the constructed rule.

Some available criteria that have been used in previous studies for distinguishing between useful and poor variables are rule performance criteria (Krusińska 1987; Snapinn and Knoke 1989; Ganeshanandam and Krzanowski 1989), group separation criteria (McKay and Campbell 1982; Krzanowski 1983; Daudin and Bar-Hen 1999), model goodness-of-fit criteria such as AIC and BIC (Daudin 1986) and other criteria including R^2, Hotelling's T^2 and Wilk's Λ (see Rencher 1993). Choice among these criteria depends on the initial aims of the classification rule, but rule performance and group separation criteria are generally popular. In the context of the location model, Bar-Hen and Daudin (1995) derived the distance between groups for mixed variables using Kullback–Leibler divergence, and obtained a test for the null hypothesis of equality of two groups. Subsequently, Daudin and Bar-Hen (1999) showed the capability of this measure to identify some variables that give significant separation of two groups. However, the study was limited to one continuous and one binary variable, and there may be problems with the use of linear model approach for parameter estimation purposes.

This paper proposes a strategy for performing variable selection in discriminant analysis using the location model. The strategy is also based on the Kullback–Leibler divergence, but it works with multivariate continuous and binary variables. The non-parametric smoothing procedures in Mahat (2006) are used for estimating the parameters as they are feasible with large numbers of binary variables. Backgrounds

of the location model for discriminant analysis and for measuring the separation of two groups are given in Sect. 2. The proposals for performing variable selection are described in Sect. 3. These proposals are evaluated on simulated data sets, the details about generating the data being given in Sect. 4 and results on them in Sect. 5. Section 6 reports the application to three real data sets, which show that the proposed method is competitive, and sometimes is better than other existing classification methods. Finally, some discussion and conclusions are presented in Sect. 7.

2 The location model

We first review the use of the location model for discriminant analysis and for measuring the differences between two groups.

2.1 Classification rule and parameter estimation

Suppose that n objects have come from two groups, n_1 being from π_1 and n_2 from π_2 with $n = n_1 + n_2$. A vector $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ is observed on each object, where $\mathbf{x}^T = (x_1, \ldots, x_q)$ is a vector of q binary variables each having values 0 or 1 and $\mathbf{y}^T = (y_1, \ldots, y_p)$ is a vector of p continuous variables. More general categorical variables are included in this formulation, as a nominal variable with c categories can be represented by $c - 1$ dummy binary variables, and $\mathbf{x}$ is identified with cell number $1 + \sum_{b=1}^{q} x_b 2^{b-1}$ of an m-cell multinomial where $m = 2^q$. Suppose that $p_{is} = P(\mathbf{x} \in s|\pi_i)$ is the probability of observing an object of π_i in cell s for $i = 1, 2$ and $s = 1, \ldots, m$. We assume that the p continuous variables have a multivariate normal distribution in each cell, with mean $\boldsymbol{\mu}_{is}$ in cell s and class π_i and a homogeneous covariance matrix, $\boldsymbol{\Sigma}$, across all cells and classes so that $\mathbf{y}_{is} \sim \mathbf{N}(\boldsymbol{\mu}_{is}, \boldsymbol{\Sigma})$. This assumption is similar to the one routinely made in multivariate analysis of variance (MANOVA). It is possible to relax it (see, e.g., Krzanowski 1994), but this will be unnecessary in most practical applications.

The classification rule based on the location model can be derived easily using these population parameters (Krzanowski 1975). We allocate a future object $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ to π_1 if its $\mathbf{x}$ falls in cell s and its $\mathbf{y}$ satisfies

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

otherwise to π_2. In this allocation rule, the constant a depends on costs due to misallocation and prior probabilities for the two groups. We will assume equal costs and equal priors in both groups, hence $log(a) = 0$.

The sample-based classification rule is obtained by replacing all the population parameters in (1) with their estimates computed from the $n = n_1 + n_2$ sample observations. By using non-parametric smoothing procedures to obtain the estimates, each cell mean $\boldsymbol{\mu}_{is}$ is fitted by a weighted average of all continuous variables from the data

in the relevant group π_i. The vector of means of the p continuous variables $\mathbf{y}$ for cell s of π_i is

$$\hat{\boldsymbol{\mu}}_{is} = \left\{ \sum_{k=1}^{m} n_{ik} w(s,k) \right\}^{-1} \sum_{k=1}^{m} \left\{ w(s,k) \sum_{r=1}^{n_{ik}} \mathbf{y}_{rik} \right\} \tag{2}$$

subject to

$$0 \leq w(s,k) \leq 1; \quad \text{and} \quad \left\{ \sum_{k=1}^{m} n_{ik} w(s,k) \right\} > 0$$

where $s, k = 1, \ldots, m$ and $i = 1, 2$. Here, n_{ik} is the number of objects of π_i that fall in cell k ($\sum_{k=1}^{m} n_{ik} = n_i$), $\mathbf{y}_{rik}$ is the vector of the continuous variables of the rth object falling in cell k of π_i and $w(s,k)$ is a weight with respect to cell s of objects that fall in cell k.

In this paper the weights are chosen in the form $w(s,k) = \lambda^{d(s,k)}$. They incorporate a smoothing parameter, λ ($0 < \lambda < 1$), that is the same for all p continuous variables, cells and groups to avoid having too many parameters to be estimated. In this definition, $d(s,k)$ is the dissimilarity coefficient between the sth cell and the kth cell of the binary vectors, given by the number of binary variables whose values differ between the two cells. Thus if we let $\mathbf{x}_s$ denote the vector of binary variable values defining the sth cell, then we can formally write $d(s,k) = d(\mathbf{x}_s, \mathbf{x}_k) = (\mathbf{x}_s - \mathbf{x}_k)^T (\mathbf{x}_s - \mathbf{x}_k)$. All cells that have equal dissimilarity with respect to cell s will have equal weight in the estimation of cell means, and $w(s,k)$ decreases as $d(s,k)$ increases for a given value of λ.

The obtained vectors of estimated cell means, $\hat{\boldsymbol{\mu}}_{1s}$ and $\hat{\boldsymbol{\mu}}_{2s}$, are then used to compute a smoothed pooled covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{(n_1 + n_2 - g_1 - g_2)} \sum_{i=1}^{2} \sum_{s=1}^{m} \sum_{r=1}^{n_{is}} \left(\mathbf{y}_{ris} - \hat{\boldsymbol{\mu}}_{is} \right) \left(\mathbf{y}_{ris} - \hat{\boldsymbol{\mu}}_{is} \right)^T \tag{3}$$

where n_{is} is the number of objects that fall in cell s of π_i, $\mathbf{y}_{ris}$ is the vector of continuous variables of the rth object in cell s of π_i and g_i is the number of non-empty cells in the data from π_i.

Finally, estimates $\hat{p}_{is}$ for the cell probabilities p_{is} can be obtained using exponential smoothing

$$\hat{p}_{is} = \frac{\sum_{k=1}^{m} w(s,k) n_{is}}{\sum_{s=1}^{m} \sum_{k=1}^{m} w(s,k) n_{is}}. \tag{4}$$

This is an easy method and suitable if both groups can be smoothed by a single parameter (Mahat 2006), otherwise the following smoothing methods can also be considered.

- Kernel smoothing (Aitchison and Aitken 1976):

$$\hat{p}_{is} = n_i^{-1}\lambda^q \sum_{b=0}^{q} N_{ib}(s,k)\left\{\frac{1-\lambda}{\lambda}\right\}^b; \quad \frac{1}{2} \le \lambda \le 1. \tag{5}$$

  Here, $N_{ib}(s,k)$ is the number of objects that fall in cell k of π_i whose binary vector $\mathbf{x}$ is b binary variables distant from the cell s ($d(s,k) = b$), and λ is a smoothing parameter for both π_1 and π_2.
- Nearest neighbour smoothing (Hall 1981):

$$\hat{p}_{is} = n_i^{-1} \sum_{b=0}^{L} w_{ib} N_{ib}(s,k); \quad 0 \le L \le q-1 \tag{6}$$

  where weights, w_{ib}, are chosen to minimise the mean squared error

$$\Delta_i(w_{i0}, w_{i1}, \ldots, w_{iL}) = \sum_{k=1}^{m} E\left(\hat{p}_{ik} - p_{ik}\right)^2,$$

  the expectation being with respect to repeated sampling from a multinomial distribution.

Considering different methods for obtaining $\hat{p}_{is}$ in (4)–(6), three classification rules can be constructed. These rules give similar results in general (Mahat 2006), but for simplicity, this paper focuses on constructing a rule using exponential smoothing for estimating parameters of continuous and binary variables, unless stated otherwise.

The smoothing parameter, λ, must be obtained before parameters can be estimated. It can be estimated in many ways, but we suggest choosing the value λ that gives good performance of a classification rule. One characteristic of good performance is minimum error rate, but error rate takes discrete values and its function has a non-smooth curve with several local minima. An alternative measure that has a continuous function is Brier score (see Hand 1997, p. 101), so this is our preferred measure (to be minimized w.r.t. λ):

$$\frac{1}{n}\sum_{r=1}^{n}\sum_{i=1}^{2}\left\{\delta\left(\pi_i|g_r, \mathbf{x}_r, \mathbf{y}_r\right) - f\left(\pi_i|\mathbf{x}_r, \mathbf{y}_r\right)\right\}^2. \tag{7}$$

In fact, suppose we have a classification rule based on $p+q$ measurements and g_r is the group that object r with vectors of measurement $\mathbf{x}_r$ and $\mathbf{y}_r$ originally comes from (either from group 1 or group 2). Then, $\delta\left(\pi_i|g_r, \mathbf{x}_r, \mathbf{y}_r\right)$ is the indicator function characterizing the true group of object r in the training set, so it takes value 1 if $\pi_i = g_r$ and 0 otherwise, while $f\left(\pi_i|\mathbf{x}_r, \mathbf{y}_r\right)$ is the posterior probability of object r belonging

to the class π_i given the observed values $\mathbf{x}_r$ and $\mathbf{y}_r$. $f\left(\pi_i|\mathbf{x}_r, \mathbf{y}_r\right)$ is given by Bayes' formula

$$f(\pi_i|\mathbf{x}_r, \mathbf{y}_r) = \frac{p_i f(\mathbf{x}_r, \mathbf{y}_r|\pi_i)}{\sum_{i=1}^{2} p_i f(\mathbf{x}_r, \mathbf{y}_r|\pi_i)} \qquad (i = 1, 2) \tag{8}$$

where p_i is the prior probability of obtaining an object from π_i and

$$f\left(\mathbf{x}_r, \mathbf{y}_r|\pi_i\right) = \frac{p_{is}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} exp\Big[-\frac{1}{2}(\mathbf{y}_r - \boldsymbol{\mu}_{is})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}_r - \boldsymbol{\mu}_{is})\Big]. \tag{9}$$

The prior probabilities p_1, p_2 may be known from previous experience, or may be estimated by n_1/n and n_2/n if the training data have arisen from mixture sampling of the two classes, or may be set equal to each other (at 0.5) if neither of these conditions holds. We used the last option in the applications described below. All the other parameters can be estimated through leave-one-out cross-validation in order to obtain unbiased estimates: omit object r from the sample; use the remaining objects to obtain smoothed estimators ($\hat{\boldsymbol{\mu}}_{is}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{p}_{is}$); insert these estimators into (9), then into (8) to obtain $\hat{f}\left(\pi_i|\mathbf{x}_r, \mathbf{y}_r\right)$; repeat for all objects r in the sample.

Finally, error rate is estimated using leave-one-out for measuring performance of the constructed rule. To avoid having biased estimates and assessment, the whole analysis is performed using a double leave-one-out arranged in a nested fashion: omit each object r from the samples in turn for $r = 1, \ldots, n = n_1 + n_2$; obtain a value of λ that minimises the leave-one-out Brier score from the sample of size $n - 1$ without object r; compute the smoothed estimators $\hat{\boldsymbol{\mu}}_{is,-r}$, $\hat{\boldsymbol{\Sigma}}_{-r}$ and $\hat{p}_{is,-r}$ using the obtained value of λ and the sample without object r; use the estimators to construct the classification rule; predict the group of the omitted object r; if the prediction is correct, then $error_r = 0$ otherwise $error_r = 1$; repeat for all objects r in the sample; compute the leave-one-out error rate using $\sum_{r=1}^{n} error_r/n$.

2.2 Distance between groups

Bar-Hen and Daudin (1995) derived the Kullback-Leibler divergence that measures separation of two groups when variables are mixed as

$$\Delta_J = \Delta_{J_1} + \Delta_{J_2} \tag{10}$$

where

$$\Delta_{J_1} = \sum_{s=1}^{m}(p_{1s} - p_{2s})log\Big[\frac{p_{1s}}{p_{2s}}\Big] \tag{11}$$

and

$$\Delta_{J_2} = \frac{1}{2}\sum_{s=1}^{m}(p_{1s} + p_{2s})(\boldsymbol{\mu}_{1s} - \boldsymbol{\mu}_{2s})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{1s} - \boldsymbol{\mu}_{2s}). \tag{12}$$

They obtained the sample-based divergence, D_J, by replacing all the parameters in (11) and (12) with the respective estimators using the linear model, and derived the asymptotic distribution of D_J under the null hypothesis that both groups are equal, $H_0 : \Delta_J = 0$ as

$$X_J = \frac{n_1 n_2}{(n_1 + n_2)} D_J \sim \chi^2(\theta_1 + \theta_2) \quad \text{when } n_1 \to +\infty, \quad n_2 \to +\infty \quad \text{and} \quad \frac{n_1}{n_2} \to u. \tag{13}$$

Here D_J is the estimate of Δ_J; θ_1 and θ_2 are the degrees of freedom in relation to binary and continuous variables separating the two groups (see also Sect. 3.1); n_1 and n_2 are the size of the sample from π_1 and π_2; and $0 < u < \infty$.

The null hypothesis is rejected if $X_J > \chi^2(\theta_1 + \theta_2, 1 - \alpha)$ for type I error α. As mentioned in Sect. 1, this measure is capable of identifying variables that give significant separation of two groups. We therefore propose using this measure for variable selection with multivariate situations.

3 The procedure

The overall procedure can be summarised in three steps: (i) select useful variables that give maximum separation of two groups from $p + q$ variables, (ii) construct a classification rule using the selected variables and finally, (iii) evaluate the constructed rule. The variable selection process requires a criterion for identifying useful variables, and a search process for selecting useful variables that optimises the chosen criterion.

3.1 Criterion for selecting useful variables

We use the Kullback–Leibler divergence in (10) to identify useful variables. The non-parametric estimators in Sect. 2.1 are employed to obtain a sample-based Kullback–Leibler divergence, D_J, by replacing the parameters in functions (11)–(12) with the relevant smoothed estimators, i.e. $\hat{\boldsymbol{\mu}}_{is}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{p}_{is}$ so that more variables can be considered.

Initially, Bar-Hen and Daudin (1995) employed linear model approaches for estimating the parameters. These parameters were divided into some that contributed to the separation between both groups and some that did not. The former were used to determine the degrees of freedom under the null hypothesis $H_0 : \Delta_J = 0$, where θ_1 is the number of parameters in estimating cell probabilities using a log-linear model and θ_2 is the number of parameters in estimating the mean vector of the continuous variables using a linear model of analysis of variance.

This strategy however may not be suitable with non-parametric smoothing, so we propose new values of θ_1 and θ_2 that represent the numbers of estimated parameters in the location model using smoothing methods. If there are g groups and q categorical

variables, with k_i categories in the ith such variable, then

$$\theta_1 = g\left\{\left(\prod_{i=1}^{q} k_i\right) - 1\right\} \quad \text{and} \quad \theta_2 = p(g-1)\left\{\left(\prod_{i=1}^{q} k_i\right) - 1\right\}. \tag{14}$$

Thus, if all the categorical variables are binary, then $\theta_1 = g(2^q - 1)$ and $\theta_2 = p(g-1)(2^q - 1)$. We interpret θ_1 as the number of estimated cell probabilities and θ_2 as the number of elements of the estimated $\hat{\boldsymbol{\mu}}_{1s} - \hat{\boldsymbol{\mu}}_{2s}$.

Deriving definite degrees of freedom mathematically is crucial due to the complexity of the model. Investigation using simulated data sets gave evidence on the adequacy of the new degrees of freedom (14) over those proposed by Bar-Hen and Daudin (1995), at least when all categorical variables are binary. The investigation was as follows. For each pair from a range of values of p and q, two groups with 150 objects each were randomly generated in such a way that $\Delta_J = 0$ in each case. This was ensured by generating the continuous variables in each group from a single normal distribution with mean $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ and covariance matrix, $\boldsymbol{\Sigma}$, and setting $p_{is} = \frac{1}{m}$ for all groups i and cells s of the categorical variables. 100 objects from each group were used to obtain the optimum smoothing parameter that minimises the leave-one-out Brier score, λ. The obtained λ and the remaining 50 objects from each group were used to compute D_J. Finally, D_J was compared to the χ^2 distribution with ν degrees of freedom at a fixed value of type I error, α. These processes were repeated 1,000 times and using the property of binomial distribution, we expect to obtain 950 ± 14 and 900 ± 16 non-significant cases if the null hypothesis is true at $\alpha = 5\%$ and $\alpha = 10\%$, respectively.

Three settings were used to determine the appropriate degrees of freedom, $\nu = \theta_1 + \theta_2$: (i) the one proposed by Bar-Hen and Daudin (1995), ν_L; (ii) the one given in (14), ν_S; and (iii) a value that was searched manually through a grid of integers, ν_G. ν_G returns the value in which the non-significant cases fall in the acceptance range, either 950 ± 14 or 900 ± 16 depending on the size of α. A suitable grid of $a \leq \nu_G \leq b$ was used, where a and b are a minimum and a maximum value chosen by trial and error, e.g. for a case of two continuous and two binary variables, a grid of interval $10 \leq \nu_G \leq 15$ may be used. To decide whether either ν_L or ν_S is appropriate, both were compared to ν_G and the one that gave values in closest agreement with ν_G was chosen.

Figure 1 shows the location of ν_L and ν_S relative to ν_G at $\alpha = 5\%$ and $\alpha = 10\%$. In this figure, the X-axis represents the simulated data sets: each set has different numbers of continuous and binary variables, with the number of binary variables increasing as we move from left to the right, so that there are more cells and more parameters to be estimated. Meanwhile, the Y-axis gives the information on the degrees of freedom. The figure shows that the curve of ν_S is much closer to the curve of ν_G compared to ν_L either at $\alpha = 5\%$ or $\alpha = 10\%$, showing that ν_S is more appropriate than ν_L. The latter deteriorates considerably when the number of binary variables increases, thus should be avoided for the smoothed location model.

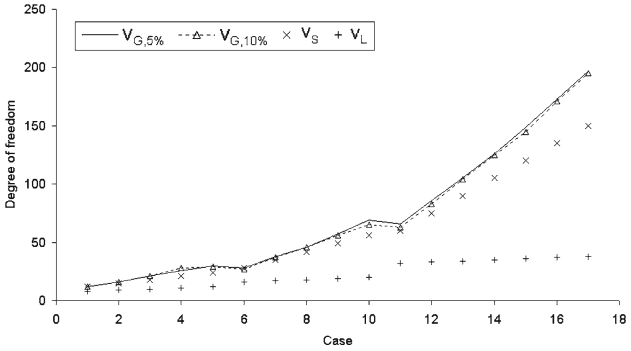

Fig. 1 Agreement between v_L, v_S and v_G when Λ_{opt} minimises the leave-one-out Brier score

3.2 Searching process

Variables that maximise D_J can be found using forward, backward and stepwise selections, based on the test for no additional information (Rao 1973, pp. 554-555). To have an automatic searching process, the inclusion or the elimination of variables must be stopped when there is no meaningful change in the value of D_J. Let C be the set of selected variables. A summary of the stopping rule in each selection process is as follows, where we assume that all categorical variables are binary.

(a) Forward selection.
Find the best of the remaining variables, whose inclusion into C yields the highest value of D_J in each step. If the chosen variable does not increase the group separation, then the values of D_J with and without that variable are estimates of the same population quantity Δ. Thus, the difference between D_{J_j} and $D_{J_{j-1}}$ for $j = 2, \ldots, p+q$ is an estimate of $\Delta = 0$ and hence, by the results above, should approximately follow the chi-squared distribution with modified degrees of freedom $\nu = v_j - v_{j-1}$

$$X_0^f = \frac{n_1 n_2}{n_1 + n_2}(D_{J_j} - D_{J_{j-1}}) \sim \chi^2(\nu = v_j - v_{j-1}). \tag{15}$$

Here, n_i is the size of the training set for π_i; D_{J_j} and $D_{J_{j-1}}$ are the estimated Kullback-Leibler divergence at steps j and $j-1$ respectively; and v_j and v_{j-1} are the values of the degrees of freedom for the chi-squared distribution of D_{J_j} and $D_{J_{j-1}}$ respectively. This test statistic (15) is compared to $\chi^2(\nu = 2^q - 1, 1-\alpha)$ if a continuous variable is selected at step j or $\chi^2(\nu = 2^q(2+p), 1-\alpha)$ if a binary variable is selected at step j, where p and q are the number of continuous and binary variables at step $j-1$ and α is the type I error. So, the tested variable is selected if $X_0^f > \chi^2(\nu, 1-\alpha)$, otherwise the process is stopped.

(b) Backward elimination.
Find the worst variable, whose omission from C yields the highest value of D_J. Then, compare the test statistic

$$X_0^b = \frac{n_1 n_2}{n_1 + n_2}(D_{J_{j-1}} - D_{J_j}) \tag{16}$$

with $\chi^2(\nu = 2^q - 1, 1 - \alpha)$ if a continuous variable is selected at step j or $\chi^2(\nu = 2^q + p2^{q-1}, 1 - \alpha)$ if a binary variable is selected at step j. p and q are the number of continuous and binary variables at step $j - 1$, and the selected variable is deleted if $X_0^b < \chi^2(\nu, 1 - \alpha)$, otherwise the process is stopped. This strategy has a potential to delete all variables and this will go against the construction of the classification rule. Therefore, the backward elimination is also stopped if there is only one continuous and one binary variable left in C.

(c) Stepwise selection.
In each step, stepwise selection performs the forward selection first, thus the stopping rule outlined in (a) is employed. Then, the current set of selected variables is tested using the backward elimination stopping rules. However, the test statistic is

$$X_0^s = \frac{n_1 n_2}{n_1 + n_2}(D_{J_j}^f - D_{J_j}^b), \tag{17}$$

where $D_{J_j}^f$ is the estimated distance at step j obtained from the forward selection sequence and $D_{J_j}^b$ is the estimated distance at step j obtained from the backward elimination sequence. X_0^s is compared to the chi-squared distributions as given in (b), either $\chi^2(\nu = 2^q - 1, 1 - \alpha)$ or $\chi^2(\nu = 2^q + p2^{q-1}, 1 - \alpha)$, depending on which type of variable has been selected. Here, p and q are the number of continuous and binary variables at step j. The searching is terminated when both tests of forward and backward sequences give the instruction to stop.

If some or all of the categorical variables have more than two categories then, as already mentioned in Sect. 2.1, they can be replaced by an appropriate number of dummy binary variables. The above procedure can thus be carried out formally, but it is evident from equation (14) that the degrees of freedom of the various chi-squared tests need to be adjusted. This can be done, for example, as in the fitting of log-linear models to incomplete multiway tables (Fienberg 1972). Unfortunately, differing numbers of categories in different variables disrupts the automatic nature of the process because individual tailoring of the computational routine is required to ensure that correct degrees of freedom are used in each test. Thus, for simplicity, we suggest keeping the process exactly as given above. Although this means that the degrees of freedom are larger in each test than they should be (as $k_i < 2^{k_i - 1}$ when $k_i > 2$), this suggestion has the merit that only those variables that are very clearly significant will be retained, and the selection will therefore be as parsimonious as possible. Of course if some relaxation is desired, then significance levels can always be set less stringently in each of the tests.

Once the variable selection process has been completed, the classification rule can be constructed and then evaluated. To avoid having a biased evaluation on discriminant analysis with variable selection, triple leave-one-out could be used. Unless the size of samples is small, this strategy demands excessive computing so we use independent training and test sets: first, objects in both groups are divided into separate training and test sets. All $p+q$ variables and training set are used to find a set of useful variables using the outlined forward, backward or stepwise selection. Second, the obtained set of variables and training set are used to construct a classification rule. All the estimation processes in both steps are conducted in leave-one-out fashion. Finally, the constructed rule is evaluated on the test set by counting the proportion of misclassified objects.

4 Monte Carlo study

The proposed strategies of variable selection were evaluated on simulated data sets. These data sets consist of p continuous and q binary variables generated using the finite mixture model following Everitt and Merette (1990). This model involves two groups, with n_1 and n_2 objects respectively, and starts from $p+q$ continuous random variables, $y_{i1}, \ldots, y_{ip}, y_{i(p+1)}, \ldots, y_{i(p+q)}$, that are generated from a multivariate normal distribution with mean $\boldsymbol{\mu}_i$ and a homogeneous covariance matrix, $\boldsymbol{\Sigma}$ of size $(p+q) \times (p+q)$. Then the mixture of continuous and binary variables is produced by keeping $y_{i1}, \ldots, y_{ip}$ as generated, but discretising $y_{i(p+1)}, \ldots, y_{i(p+q)}$ using

$$x_{ij} = \begin{cases} 0 & \text{if} \quad -\infty \leqslant y_{i(p+j)} < \delta_{ij} \\ 1 & \text{if} \quad \delta_{ij} \leqslant y_{i(p+j)} < \infty \end{cases} \tag{18}$$

where δ_{ij} are thresholds, $i = 1, 2$ and $j = 1, \ldots, q$. The threshold for each $y_{i(p+1)}, \ldots, y_{i(p+q)}$, was determined such that $\delta_{ij} = \Phi^{-1}(p_{ij}) \times \sigma_{y_{i(p+j)}} + \mu_{y_{i(p+j)}}$ where p_{ij} is the target proportion of objects of π_i having $x_{ij} = 0$, $\mu_{y_{i(p+j)}}$ and $\sigma_{y_{i(p+j)}}$ are the mean and standard deviation of y_{p+j} in π_i respectively, and $\Phi^{-1}(.)$ is the inverse cumulative standard normal integral. Assuming the mean of each $y_{i(p+1)}, \ldots, y_{i(p+q)}$ is zero and all variances are unity, then δ_{ij} is equal to the standard normal ordinate corresponding to the target proportion.

In the finite mixture model, a homogeneous covariance matrix is

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

where $\boldsymbol{\Sigma}_p, \boldsymbol{\Sigma}_q$ and $\boldsymbol{\Sigma}_{pq}$ are the covariance matrices for the relevant types of variables. Diagonal elements of $\boldsymbol{\Sigma}_q$ were set as unity and low correlations among binary variables and low correlations between binary and continuous variables were assumed. Correlations among continuous variables were determined by the parametrisation procedure of Costanza and Afifi (1979). A factor υ $(0 < \upsilon < 1)$ must be specified to determine the eigenvalues λ_k of $\boldsymbol{\Sigma}_p$ through

$$\lambda_k = a\upsilon^{k-1} + 0.1 \quad \text{for} \quad k = 1, \ldots, p, \tag{19}$$

where

$$a = \begin{cases} 0.9p(1-\upsilon)/(1-\upsilon^p) & 0 < \upsilon < 1 \\ 0.9 & \upsilon = 1. \end{cases}$$

υ represents the degree of independence among continuous variables: the variables are highly dependent when υ is close to 0 and they are mutually independent when $\upsilon = 1$. Define $\mathbf{\Sigma}_p = \mathbf{L}\mathbf{\Lambda}\mathbf{L}^T$, where $\mathbf{L}$ is the matrix of eigenvectors of $\mathbf{\Sigma}_p$, $\mathbf{L}^T$ is the transpose matrix of $\mathbf{L}$ and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues λ_k. Once λ_k has been specified, then $\mathbf{\Lambda}$ can be obtained. A random orthonormal matrix $\mathbf{L}$ was generated using the Gram-Schmidt procedure.

The artificial data sets were generated in such a way that for each type of variables some of these variables contributed to the separation between groups and some did not. This is necessary because the selection strategies (Sect. 3.2) must have at least one continuous and one binary variable for construction of the classification rule. The values for the necessary characteristics were chosen as follows:

- Number of continuous variables: (i) $p = 3$; (ii) $p = 4$; (iii) $p = 8$, while number of binary variables is $q = 4$ for all cases.
- Covariance structure of continuous variables, $\mathbf{\Sigma}_p$, ranging from dependent to independent cases: (i) $\upsilon = 0.25$; (ii) $\upsilon = 0.45$; (iii) $\upsilon = 0.85$.
- Size of training sets: (i) $n_1 = n_2 = 25$; (ii) $n_1 = n_2 = 100$; (iii) $n_1 = 25, n_2 = 100$.
- Test sets for π_1 and π_2 fixed at 50 objects each.

The outlined strategy was used to generate both groups separately, $\pi_1 \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$ and $\pi_2 \sim \mathbf{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$. The values of means and the target proportion of binary variables are summarised in Tables 1 and 2, respectively. Thus, the selection methods were supposed to choose the continuous variables listed in the last column of Table 1, and based on Table 2, X_1, X_3 and X_4, were also supposed to be identified. We call these variables the "useful" ones.

Table 1 Expectations of the continuous variables

Number of variables	π_1	π_2	Useful variables
p =3	(0,0,0)	(0,0,1)	Y_3
p =4	(0,0,0,0)	(0,0,1,3)	Y_3, Y_4
p =8	(0,0,0,0,0,0,0,0)	(0,0,1,3,1,1,2,1.5)	$Y_3, Y_4, Y_5, Y_6, Y_7, Y_8$

Table 2 Target proportion $P(X_j = 0)$ for the binary variables

Group	X_1	X_2	X_3	X_4
1	0.30	0.50	0.50	0.90
2	0.80	0.40	0.15	0.50

5 Monte Carlo results

Table 3 summarises results of the simulation study in the following format, (p_1,q_1)/ (p_2,q_2);z. Here p_1 and q_1 are the numbers of continuous and binary variables chosen by the method; p_2 and q_2 are the numbers of "useful" continuous and binary variables among those in (p_1,q_1); and z is the estimated error rate.

Consideration of (p_1,q_1) shows that backward elimination retains more variables compared to the other strategies in general. Therefore, considering (p_2,q_2), backward elimination is able to identify more useful variables in many cases and in the case $p = 3, q = 4$ and $n_1 = n_2 = 25$ (bold), it has identified all the useful variables correctly. Forward and stepwise selections agree with each other most of the time and this might be influenced by the nature of identifying a new useful variable at every searching step. Both methods potentially give different results if the number of variables is large, for example in the case $p = 8, q = 4$ and $n_1 = n_2 = 100$.

In real applications, useful variables are unknown, thus performance of the rule (e.g. error rate) is often used as a criterion of good variables. Table 3 shows that there is no clear winner among the selection strategies based on error rate. Two interesting situations from this table are worth highlighting.

(i) The first situation ($p = 4; q = 4; n_1 = n_2 = 25$) shows the biggest difference in error rates between methods. Backward elimination has much lower error rate (0.06) than both forward and stepwise selections (0.23). This is due to the differences in the final chosen variables. Details from printed results showed that backward elimination retained x_1, y_1 and y_4 while the other two strategies selected x_4, y_1, y_2 and y_4. Two factors may explain this situation. Firstly, there were different sets of binary variables. Secondly, both forward and stepwise selections selected the useless y_2.

(ii) The second situation is when all strategies obtain zero value for the error rate (see case $p = 8, q = 4$). This means that the associated classification rules perfectly allocate test objects to their groups. These results occur because both groups are very disperse. In this case, all proposed search strategies are equivalent.

In Table 3, the 'Full' method means that all $p + q$ variables are used to construct the smoothed location model. In 52% of cases the error rate of the rules with variable selection is equal to or lower than the error rate of the full model (marked '[a]').

The obtained results have shown some interesting features about selecting variables on the basis of smoothed Kullback–Leibler divergence. Initially, this criterion was chosen instead of error rate due to lack of asymptotic distributions of the latter, which creates difficulties to perform automatic variable selection. Despite criticisms that have previously been made by several researchers (Habbema and Hermans 1977; Raudys and Jain 1991), there is evidence that this criterion chooses useful variables.

Both assessment based on the selecting of useful variables and error rate show that none of the selection strategies is always the winner, so the choice of strategy depends on the problem in hand. However, backward elimination is not a good strategy for a large number of binary variables because it includes all the variables at the beginning of the process and so requires estimation of many parameters. In such case, forward and stepwise selections are preferable.

Table 3 Numbers of selected variables and values of error rate for variable selection strategies

Number of		Group size		Selection	Covariance structure		
p	q	π_1	π_2	method	Independent	Moderate	Dependent
				Full	0.20	0.19	0.13
				Backward	**(3,3)/(1,3)**; 0.20[a]	(3,2)/(1,2); 0.22	(2,1)/(1,1); 0.25
		25	25	F & S	(2,1)/(1,1); 0.31	(1,1)/(0,1); 0.30	(2,1)/(1,1); 0.25
3	4			Full	0.20	0.49	0.21
				Backward	(1,2)/(1,2); 0.23	(1,2)/(0,1); 0.43[a]	(3,1/(1,1); 0.23
		100	100	F & S	(1,2)/(1,2); 0.23	(1,1)/(1,1); 0.47	(3,1)/(1,1); 0.22
				Full	0.25	0.13	0.09
				Backward	(2,3)/(1,2); 0.26	(3,2)/(1,2); 0.18	(3,2)/(1,2); 0.18
		25	100	F & S	(1,2)/(1,2); 0.23[a]	(1,1)/(1,1); 0.25	(1,1)/(1,1); 0.24
				Full	0.11	0.09	0.04
				Backward	(2,1)/(1,1); 0.06[a]	(4,2)/(2,2); **0.09**	(3,2)/(1,2); 0.06
		25	25	F & S	(3,1)/(1,1); 0.23	(3,2)/(2,2); 0.08[a]	(3,1)/(2,1); 0.02[a]
4	4			Full	0.15	0.08	0.01
				Backward	(4,2)/(2,2); 0.15[a]	(4,1)/(2,1); 0.09	(3,2)/(1,2); 0.02
		100	100	F & S	(4,2)/(2,2); 0.15[a]	(3,1)/(1,1); 0.09	(3,2)/(1,2); 0.02
				Full	0.12	0.16	0.06
				Backward	(2,2)/(2,2); 0.17	(3,3)/(1,3); 0.15[a]	(4,2)/(2,2); 0.05
		25	100	F & S	(2,2)/(2,2); 0.17	(2,1)/(2,1); 0.18	(4,1)/(2,1); 0.04[a]
				Full	0.13	0.00	0.01
				Backward	(5,1)/(5,1); 0.14	(8,2)/(6,2); 0.00[a]	(7,1)/(6,1); 0.00[a]
		25	25	F & S	(2,2)/(2,1); 0.19	(8,2)/(6,2); 0.00[a]	(5,2)/(5,1); 0.01
8	4			Full	0.18	0.00	0.00
				Backward	(4,3)/(3,3); 0.15	(8,3)/(6,2); 0.00[a]	(7,3)/(5,3); 0.00[a]
		100	100	Forward	(3,2)/(3,2); 0.15	(8,2)/(6,2); 0.00[a]	(7,3)/(5,3); 0.00[a]
				Stepwise	(3,1)/(3,1); 0.13[a]	(7,2)/(5,2); 0.00[a]	(4,1)/(4,1); 0.00[a]
				Full	0.10	0.02	0.01
				Backward	(6,2)/(4,1); 0.11	(7,1)/(5,1); 0.04	(6,2)/(5,2); 0.03
		25	100	F & S	(1,2)/(1,2); 0.20	(7,1)/(5,1); 0.03	(6,2)/(5,2); 0.03

See text for explanation
Lines F & S refer to forward and stepwise selections - they produce the same results
[a] Improved performance on selection

6 Examples and comparisons

The proposed variable selection strategies were also investigated on three real data sets. As well as evaluating their performance, it was of interest to compare them with other classification rules. There is potentially a very large set of such candidate rules, ranging from the simple "Naive Bayes" approach (Bickel and Levina 2004) through to complex neural networks and support vector machines (Webb 2002), but we decided to focus specifically on those commonly used rules that are routinely available in standard package software. Working with S-Plus 6.1, the full set of rules for

comparison was therefore: smoothed location models (LM) with exponential, kernel and nearest neighbour smoothing to obtain cell probabilities; linear discriminant functions (LDF); quadratic discriminant functions (QDF); logistic discrimination (logistic); regression based models with forward, backward and stepwise selections; and tree classifiers. As a full tree may be unnecessarily large, a tree with error rate pruning was also considered: a full tree was first constructed and then it was pruned to t nodes such that the pruned tree of size t gives the lowest error rate (Venables and Ripley 1994, p. 345).

The obtained data sets contain various types of variables, but the location model is suitable with continuous and binary variables. Therefore, other types of variables were transformed to follow either continuous or binary type. We treated ordinal variables as continuous variables and nominal variables were dichotomised into new binary variables: if a nominal variable has k states, then it was replaced by $k-1$ binary variables, $b_1, b_2, \ldots, b_{k-1}$. These binary variables all take zero value except b_j, which takes value one, when the corresponding nominal variable is observed in its jth state for $j = 1, \ldots, k-1$, and they are all zero when the nominal variable is in state k. For all three data sets we used the automatic (parsimonious) selection method suggested in Sect. 3.2, the degrees of freedom of chi-squared tests being those derived from the full set of binary variables.

The first data set concerns the influences of psychosocial behaviour among breast cancer patients, conducted at King's College Hospital, London. It comprises 78 patients who had a benign tumour (π_1) and 59 patients who had a malignant tumour (π_2). 15 measurements were taken for each patient: two continuous variables, six ordinal variables having 11 states each, four nominal variables having three states each and three binary variables. By treating the ordinal as continuous and dichotomising the nominal into binary, then these data consist of eight continuous and 11 binary variables. The second data set is the subset of the first data set that was analysed by Krzanowski (1975, 1980), where it was known as data Set 5. This data set contains a reduced number of variables with one continuous variable, six ordinal, two nominal and two binary variables. After transforming both the ordinal and nominal variables, this subset has seven continuous and six binary variables. This data set was also considered in this study so that some comparisons between the full and the reduced sets could be made.

Finally, the third data set is a subset of the original "Cleveland Heart" data of the StatLog project. It concerns the presence and absence of heart disease from various medical tests carried out on patients in the Cleveland Clinic Foundation. These patients came from one of two groups: 150 patients without heart disease (π_1) and 120 patients with heart disease (π_2), and there were seven continuous, three nominal and three binary variables. Transforming the nominal variables then leads to seven continuous and nine binary variables.

Comparisons were made in terms of performance of the rules and size of set of selected variables. The values of the 10-fold cross-validation error rate for each data set are presented in Table 4. As can be seen, logistic discrimination is the best rule in the full breast cancer data and heart data set, while exponential LM is the best rule in the reduced breast cancer data set. On the other hand, QDF has the worst performance in the full breast cancer data and the heart data, and the full tree is the worst in the reduced breast cancer data.

Table 4 Classification performance based on 10-fold cross-validation error rate, with lowest error rates shown in bold for each data set

Method	Selection direction	Data set		
		Breast cancer		Heart
		Full	Reduced	
LDF	Include all variables	0.2920	0.3066	0.1667
QDF	Include all variables	0.4453	0.3212	0.2778
Logistic	Include all variables	**0.2847**	0.2847	**0.1519**
Full tree	Auto-selection	0.3139	0.3650	0.2519
Pruned tree	lowest error rate	0.3139	0.3577	0.2519
Regression	Forward selection	0.3139	0.2847	0.1815
	Backward elimination	0.2920	0.2847	0.1889
	Stepwise selection	0.2920	0.2847	0.1889
Smoothed location model:				
Exponential	Forward selection	0.3139	**0.2628**	0.2037
	Stepwise selection	0.3139	**0.2628**	0.2037
N. neighbour	Forward selection	0.3066	0.2920	0.1667
	Stepwise selection	0.3066	0.2920	0.1667
Kernel	Forward selection	0.3066	–	0.1852
	Stepwise selection	0.3066	–	0.1852

We did not perform smoothed LM with backward elimination as it is inappropriate due to the large number of variables to begin with. We encountered a problem of matrix singularity in the eighth fold of the reduced breast cancer data when evaluating the kernel LM. Thus, this rule could not obtain the 10-fold cross-validation error rate. The results for the full and reduced breast cancer data are different, with LDF and trees being the worst while other rules show better performance in the reduced case. Such results may be due to the different use of variables.

The results in Table 4 also give evidence that sometimes, using a reduced size of variables is better than using them all. For example, in the full breast cancer data, LDF and regression with backward and stepwise selections are the second best rules, and in the heart data, LDF and nearest neighbour LM have the same error rate. In such cases, we usually favour having fewer variables for reasons of simplicity (see Hoadley 2001) and to avoid the effect of multicollinearity problems. We therefore compared the average number of selected variables over the 10-fold among classification rules with variable selection. These values are given in Table 5.

In Table 5, different classification rules select different subsets of variables. This is due to the use of different criteria for selecting variables. The regression rule selects variables based on the AIC criterion, trees select variables from the information of categorical variables and the Gini criterion and finally, the smoothed LM rules choose the variables that give the biggest separation between two groups.

The results show that smoothed LM rules use fewer variables compared to the other rules (perhaps a consequence of our "parsimonious" selection strategy). We marked the rule(s) with the lowest error rate with '[a]'; representing the best performance. Except for the full breast cancer data, other data sets give evidence that using fewer

Table 5 Average number of selected variables over 10-fold cross-validation

Classification rules	Breast cancer		Heart
	Full	Reduced	
Full tree	8.4	8.8	11.1
Pruned tree	6.9	8.7	10.8
Regression:			
Forward selection	5.1	3.5	9.2
Backward elimination	11.3[a]	4.1	11.1
Stepwise selection	11.3[a]	4.2	11.1
Smoothed location model:			
Exponential	2.4	3.5[a]	6.0
Nearest neighbour	2.3	2.8	4.6[a]
Kernel	4.3	2.5	4.6

[a] The rules with lowest error rates for each data set

variables is better: exponential LM is the best in the reduced breast cancer data and nearest neighbour LM is the best in heart data.

The similar performance of the full and pruned trees may be influenced by the strategy of pruning, where a pruned tree is constructed from a full tree. Maybe a good tree could be obtained by searching for a tree in which the combination of variables gives the lowest error rate. The difficulty of using trees is that they are best used by experts, because growing a tree not only involves mathematics but needs one to be creative to build it (Duin 1996).

7 Discussion and conclusions

In this paper, we have demonstrated variable selection for the location model in forward, backward and stepwise selections. Results from a simulation study show evidence that the proposed strategies are able to handle multivariate data and that they are also reliable for various types of variables.

Examples of real data show the potential use of the location model. It is sometimes the best, and if not, its results are not much different from the best rules. As a conclusion, the smoothed location model should be considered as another potential tool in discriminant analysis when the feature variables are mixed. Also, there are no problems in employing this rule with large number of variables.

It would very beneficial to extend this study to more than two groups. The implementation to this case is generally straightforward, but the selection strategies need to be planned carefully to avoid constructing inappropriate rules. Furthermore, the possibility of the inappropriate use of distance between groups for more than two groups (see Habbema and Hermans 1977; Raudys and Jain 1991; Aeberhard et al. 2000) have to be noted. Perhaps, the best strategy is probably to find the subset which gives the smallest error rate, but much work has to be done in order to find the 'best' subset automatically through this approach.

Acknowledgments We thank the Editor and two anonymous referees for all their comments and suggestions, which have considerably improved our presentation.

References

Aeberhard S, Vel OYD, Coomans DH (2000) New fast algorithms for error rate-based stepwise variable selection in discriminant analysis. SIAM J Sci Comput 22:1036–1052

Aitchison J, Aitken CGG (1976) Multivariate binary discrimination by the kernel method. Biometrika 63:413–420

Asparoukhov O, Krzanowski WJ (2000) Non-parametric smoothing of the location model in mixed variable discrimination. Stat Comput 10:289–297

Bar-Hen A, Daudin JJ (1995) Generalization of the Mahalanobis distance in the mixed case. J Multivar Anal 53:332–342

Bickel PJ, Levina E (2004) Some theory for Fisher's Linear Discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. Bernoulli 10:989–1010

Chang PC, Afifi AA (1974) Classification based on dichotomous and continuous variables. J Am Stat Assoc 69:336–339

Costanza MC, Afifi AA (1979) Comparison of stopping rules in forward stepwise discriminant analysis. J Am Stat Assoc 74:777–785

Daudin JJ (1986) Selection of variables in mixed-variable discriminant analysis. Biometrics 42:473–481

Daudin JJ, Bar-Hen A (1999) Selection in discriminant analysis with continuous and discrete variables. Comput Stat Data Anal 32:161–175

Duin RPW (1996) A note on comparing classifiers. Patt Recognit Lett 17:529–536

Everitt BS, Merette C (1990) The clustering of mixed-mode data: A comparison of possible approaches. J Appl Stat 17:283–297

Fienberg SE (1972) The analysis of incomplete multiway contingency tables. Biometrics 28:177–202

Ganeshanandam S, Krzanowski WJ (1989) On selecting variables and assessing their performance in linear discriminant analysis. Aust J Stat 31:433–447

Habbema JDF, Hermans J (1977) Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics 19:487–493

Hall P (1981) Optimal near neighbour estimator for use in discriminant analysis. Biometrika 68:572–575

Hand DJ (1997) Construction and assessment of classification rules. Wiley, Chichester

Hoadley B (2001) Comment on "Statistical modelling: The two cultures", by Breiman, L. Stat Sci 16: 220–224

Krusińska E (1987) A valuation of state of object based on weighted Mahalanobis distance. Patt Recognit 20:413–418

Krzanowski WJ (1975) Discrimination and classification using both binary and continuous variables. J Am Stat Assoc 70:782–790

Krzanowski WJ (1980) Mixtures of continuous and categorical variables in discriminant analysis. Biometrics 36:493–499

Krzanowski WJ (1983) Stepwise location model choice in mixed-variable discrimination. Appl Stat 32: 260–266

Krzanowski WJ (1994) Quadratic location discriminant functions for mixed categorical and continuous data. Stat Prob Lett 19:91–95

Mahat NI (2006) Some investigations in discriminant analysis with mixed variables. Ph. D. thesis, Exeter University, U.K.

McKay RJ, Campbell NA (1982) Variable selection techniques in discriminant analysis ii. allocation. British J Math Stat Psychol 35:30–41

McLachlan GJ (1992) Discriminant analysis and statistical pattern recognition. Wiley, New York

Olkin I, Tate RF (1961) Multivariate correlation models with mixed discrete and continuous variables. Ann Math Stat 32:448–465

Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York

Raudys SJ, Jain AK (1991) Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Trans Syst Man Cyber 13:252–264

Rencher AC (1993) The contribution of individual variables to Hotelling's T^2, Wilk's λ, and R^2. Biometrics 49:479–489

Snapinn SM, Knoke JD (1989) Estimation of error rates in discriminant analysis with selection of variables. Biometrics 45:289–299

Venables WN, Ripley BD (1994) Modern applied statistics with S-Plus. Springer, New York

Webb A (2002) Statistical pattern recognition, 2nd edn. Wiley, Chichester