

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

On discrimination procedure with mixtures of continuous and categorical variables

Gafar Matanmi Oyeyemi, George Chinanu Mbaeyi, Saheed Ishola Salawu & Bernard Olagboyega Muse

To cite this article: Gafar Matanmi Oyeyemi, George Chinanu Mbaeyi, Saheed Ishola Salawu & Bernard Olagboyega Muse (2016): On discrimination procedure with mixtures of continuous and categorical variables, Journal of Applied Statistics

To link to this article: http://dx.doi.org/10.1080/02664763.2015.1125859

Published online: 22 Feb 2016.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=cjas20

Download by: [GEORGE FIGURES] **Date:** 23 February 2016, At: 19:32

JOURNAL OF APPLIED STATISTICS, 2016
http://dx.doi.org/10.1080/02664763.2015.1125859

Taylor & Francis
Taylor & Francis Group

On discrimination procedure with mixtures of continuous and categorical variables

Gafar Matanmi Oyeyemi[a], George Chinanu Mbaeyi[b], Saheed Ishola Salawu[a] and Bernard Olagboyega Muse[c]

[a]Department of Statistics, University of Ilorin, Ilorin, Kwara State, Nigeria; [b]Department of Statistics, Federal University, Abakaliki, Ebonyi State, Nigeria; [c]Department of Maths & Statistics, Rufus Giwa Polytechnic, Owo, Nigeria

ABSTRACT

A discrimination procedure, based on the location model is described and suggested for use in situation where the discriminating variables are mixtures of continuous and binary variables. Some procedures that have been previously employed, in a similar situation, like Fisher's linear discriminant function and the logistic regression were compared with this method using error rate (ER). Optimal ERs for these procedures are reported using real and simulated data for the case of varying sample size and number of continuous and binary variables and were used as a measure for assessing the performance of the various procedures. The suggested procedure performed considerably better in the cases considered and never did produce a result that is poor when compared with other procedures. Hence, the suggested procedure might be considered for such situations.

ARTICLE HISTORY
Received 5 September 2013
Accepted 25 November 2015

KEYWORDS
Location model; error rate; discriminant

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

1. Introduction

Multivariate analysis consists of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. Most bodies of data involve observations associated with various facets of a particular background, environment or experiment, thus in a general sense, data are always multivariate in character [9]. Many areas of biometry involve the collection of multivariate data in which the variates measured range from continuous to unordered categorical. Examples can be found in soil science (where such continuous laboratory measurements as soil pH may be mixed with such categorical attributes as soil color or texture) and in medicine (where continuous laboratory measurements may be included with such categorical variables as site of tumor and presence/absence of a certain symptom for each patient). If the categorical variables are ordered, the simplest and conventional procedure, the linear discriminant function (LDF), is to assign an arbitrary numerical score to each possible state of the variable and proceed as if all variables are continuous Truett *et al.* [22]. Consequently, many workers faced with data comprising wholly or partly of categorical (including binary) variables are tempted to

CONTACT G. M. Oyeyemi gmoyeyemi@yahoo.com

© 2016 Taylor & Francis

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

ignore its discrete nature and proceed with continuous variable techniques. This implies that a certain amount of misrepresentation and distortion is likely to be introduced by treating categorical variables as if they are continuous. Various procedures have been proposed as alternatives to classical discriminant analysis [13]. The location model is a discriminant function derived from a probabilistic model for mixed categorical (including binary) and continuous variables.

The aim of this paper is to suggest a discriminating procedure for classification in a two group discrimination problem with mixed continuous and binary variables, and to investigate the performance of the suggested model (hereafter referred to as location model) by comparing it with the conventional Fisher's Linear Discriminant Function (FLDF) and the logistic discrimination using their various error rates (ERs) under varying sample sizes and varying number of continuous and binary variables.

A review of few related literature will be continued in this section. The methodology will be presented in Section 2 wherein the model shall be introduced together with its parameter estimation and allocation rule. In Section 3, some practical examples using both real and simulated data will be applied to the suggested procedure and the other available procedures. Result of Section 3 will be discussed in Section 4 and then a conclusion to end the paper.

The most widely used method of discrimination, introduced by Fisher [8] assumes that $\mathbf{x}$ is distributed according to the multivariate normal distribution, the covariance matrix being the same whether $\mathbf{x}$ arises from Π_1 or Π_2 but the vector of means being different [17]. In the discriminant analysis, if categorical variables are ordered, the simplest procedure is to assign an arbitrary numerical score to each possible state of the variable and proceed as if all variables are continuous [15]. If such scoring is not possible, then one can perhaps categorize the continuous variable and use a categorical variable technique on all the data Cochran and Hopkin [5]. Alternatively, one can separate the two types of variables and perform two separate analyses using the appropriate technique for each analysis [16]. Another common method of discrimination is that described by Warner *et al.* [23], which has found considerable favor among statisticians engaged on diagnostic applications, assumes that each variable may take only two values, corresponding to a symptom recorded simply 'Present' or 'Absent', and that the separate variables are independent. Although both methods of discrimination have been used successfully in many problems, they are not always suitable [15]. Olkin and Tate [20] proposed the location model which is a probability model for the dichotomous and continuous variables. Aitchson and Aitken [1] suggested a method based on the kernel approach to estimating densities [21]. Krzanowski [14] derived a model of likelihood ratio based on the location model. Anderson [3] advocates the use of logistic discrimination in which the probability of group membership is assumed to be a logistic function of the observed variates. James and Sandra [12] found logistic regression with maximum likelihood estimates outperforming the classical linear discriminant analysis though not by a large amount. Thus, Halperin *et al.* [11] puts it this way, 'use of the maximum likelihood method would be preferable, whenever practical, in situations where the normality assumptions are violated, especially when many of the independent variables are qualitative'. Cochran and Hopkins [5] identified some problems relating to the discriminant analysis involving mixtures of categorical and continuous variables irrespective of the approach used in the analysis. Some of the problems he identified include: the effect of the initial sample sizes on the performances of the proposed classification rule,

the relative discriminating power of qualitative and continuous variates and use of classification experience for improvement of the rule. It is in the light of these identified problems that this paper would compare the performance of few available techniques alongside the suggested techniques under various sample sizes and number of discriminating variables.

2. Methodology

It is assumed that we have r explanatory variables $x_1, x_2, \ldots, x_n$ for a two-group discriminant analysis problem. Let Π_1 denote the first group and Π_2 the second group, assuming that Π_1 and Π_2 are exhaustive groups of the population of interest. Let the prior probability for Π_1 be the ratio of occurrence of Π_1 among the population, let this be denoted by p_1 implying that $p_2 = 1 - p_1$ for the other group Π_2. We also assume that our approach for discrimination is based on a sample containing both the population groups. Our objective is to predict through the use of an allocation rule.

2.1. The model

Description of the location model method for mixtures of binary and continuous variables is done as described by Krzanowski [14]. Suppose that we have an available set of n_1 individuals known to have come from the first group, Π_1, and a set of n_2 individuals known to have come from the second group,Π_2, and we wish to set up a discriminant rule between Π_1 and Π_2 on the basis of a vector of binary variables denoted by $\mathbf{x}$ and a vector of continuous variables denoted by $\mathbf{y}$. The vector of the binary variables has q components and that of the continuous variables has p components each of variables observed on each individual. The q binary variables can be treated as a multinomial $\boldsymbol{Z}' = (z_1, z_2, \ldots, z_k)$ with $k = 2^q$ states. Thus, each distinct pattern of $\mathbf{x}$ defines a multinomial cell uniquely, with $\boldsymbol{x}' = (x_1, x_2, \ldots, x_q)$ falling in cell m. Generally, we would assume that the probability of obtaining an observation in cell m is p_{im} for population Π_i $(i = 1, 2)$. For the problem of classifying an observation $\boldsymbol{w}' = (\boldsymbol{x}', \boldsymbol{y}')$ to one of two populations,Π_1 and Π_2, the model may be generalized by assuming that $\boldsymbol{y}$ has a multivariate normal distribution with mean $\boldsymbol{\mu}^{(m)}$ in cell m $(m = 1, 2, \ldots, k)$ and common dispersion matrix $\boldsymbol{\Sigma}$ in all cells of both population groups, i.e.

$$(\boldsymbol{y}/z_m = 1, z_j = 0, j \neq m = 1, 2, \ldots, k) \sim N(\boldsymbol{\mu}_i^{(m)}, \boldsymbol{\Sigma}) \tag{1}$$

in $\Pi_i (i = 1, 2)$. Also, the probability of obtaining an observation in cell m is p_{im} for population $\Pi_i (i = 1, 2)$. While the conditional dispersion matrix of $\mathbf{y}$ given $\mathbf{x}$ are the same in Π_1 and Π_2, the dispersion matrices of $\mathbf{y}$ are not.

2.2. Estimation of the parameters of the location model

For the estimation of the parameters in the optimum allocation rule in Equation (1), the frequencies of occurrence of each possible pattern of $\mathbf{x}$ for each initial sample may be written as a q-way contingency table containing $k = 2^q$ cells. Let n_{1m} and n_{2m} denote the number of observations falling in cell m of the tables from Π_1 and Π_2, respectively, and let $\boldsymbol{y}_{ji}^{(m)}$ denote the vector of continuous variables associated with the jth observation in cell

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

m of the sample from Π_i, if

$$\bar{\mathbf{y}}_i^{(m)} = \left(\frac{1}{n_{im}}\right) \sum_{j=1}^{n_{im}} \mathbf{y}_{ji}^{(m)},$$

then the maximum likelihood estimates of the population parameters P_{im}, $\boldsymbol{\mu}_i^{(m)}$ and $\boldsymbol{\Sigma}$ are given by

$$\hat{p}_{im} = \frac{n_{im}}{n_i},$$

$$\hat{\boldsymbol{\mu}}_i^{(m)} = \bar{\mathbf{y}}_i^{(m)},$$

$$\boldsymbol{\Sigma} = \frac{1}{(n_1 + n_2 - 2k)} \sum_{i=1}^{2} \sum_{m=1}^{k} \sum_{j=1}^{n_{im}} (\mathbf{y}_{ji}^{(m)} - \bar{\mathbf{y}}_i^{(m)})(\mathbf{y}_{ji}^{(m)} - \bar{\mathbf{y}}_i^{(m)})'$$

$$(i = 1, 2;\ m = 1, 2, \ldots, k).$$

In practice, however, unless n_1 and n_2 are large relative to k, it is very possible or probable that some n_{im} will be either small or zero. As such, the corresponding parameter estimates will at best be poor, and at worst unobtainable. Hence, for the allocation rule to be of more practical use, it is therefore necessary to seek an approximation which will yield parameter estimates for all possible cells. Such an approximation may be obtained by imposing some further structure on the distribution of the variables. Using the information contained in the binary variables, the initial samples may be written in the form of contingency tables. The log linear model is one of the most frequently used for the analysis of such arising contingency table [14]. The maximum likelihood estimation of the parameters in this model can be effected by means of a simple iterative procedure called iterative scaling [7]. An algorithm readily available for carrying out this procedure is called the Haberman's iterative algorithm [10]. Fitting the full model would lead to recovery of the observed cell frequencies. The simplest structure which can be imposed on the mean vectors and which is compatible with the location model is a linear additive model, with components which may be interpreted as main effects of each binary variable and interactions between the binary variables of all order up to q. Thus, to estimate $\boldsymbol{\mu}$, the mean continuous variable vector in Π_i may be written as

$$\boldsymbol{\mu}_i^{(m)} = \boldsymbol{\nu}_i + \sum_{j=1}^{q} \boldsymbol{\alpha}_{i,j} x_j^{(m)} + \sum \sum_{j<k} \boldsymbol{\beta}_{i,jk} x_k^{(m)} + \cdots + \gamma_{i,jk\ldots q} x_j^{(m)} x_k^{(m)}, \ldots, x_q^{(m)}. \quad (2)$$

And the conditional mean vector in cell m, $\boldsymbol{\mu}_i{}^{(m)}$ is obtained by inserting the values of the binary variables in cell m into the right hand side of Equation (2). In practice, an approximation may be obtained by treating all high-order terms of Equation (2) as residual/error and estimating the remaining parameters. The estimates of the parameters can be obtained using the multivariate multiple regression [2]. For consistency with the estimation of the multinomial probabilities, it is proposed that the order of model fitted is the same in both cases. The model is said to be of second order if only main effects and the first-order interactions have been retained in both the log linear model for the incidence tables and the

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

linear model for the multivariate regression. If only main effects are included throughout, then only single x_j terms will appear in the multivariate regression and the model is said to be of first order. In this case, we write

$$\mathbf{v}' = (1, x_1, x_2, \ldots, x_q) \tag{3}$$

and

$$\boldsymbol{\mu}_i = (\boldsymbol{\nu}_i, \boldsymbol{\alpha}_{1,i}, \boldsymbol{\alpha}_{2,i}, \ldots, \boldsymbol{\alpha}_{q,i}), \tag{4}$$

where $\boldsymbol{\alpha}$ is the coefficient of the main effect of each of the binary variable x. Thus, maximum likelihood estimates of the unknown parameters are given by

$$\mathbf{B}_{i=}\mathbf{C}_i\mathbf{A}_i^{-1} \tag{5}$$

and

$$(n_1 + n_2)\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{2}\left(\sum_{j=1}^{n_i} \boldsymbol{y}_{ji}\boldsymbol{y}'_{ji} - \hat{\boldsymbol{B}}_i\boldsymbol{A}_i\hat{\boldsymbol{B}}'_i\right),$$

where

$$\mathbf{C}_i = \sum_{j=1}^{n_i} \mathbf{y}_{ij}\mathbf{v}'_{ij},$$

$$\mathbf{A}_i = \sum_{j=1}^{n_i} \mathbf{v}_{ij}\mathbf{v}'_{ij}.$$

Thus, estimates of all parameters are readily available for the restricted model, and the appropriate allocation rule is constructed by replacing all parameters by their estimates.

2.3. Optimum allocation rule of the location model

It is assumed that all population parameters are known, in which case the optimum allocation rule can be derived readily from the general theory of classification. Let $\overline{X}_1, \overline{X}_2$ and $\mathbf{S}$ denote the sample means and sample covariance matrix, respectively, based on two samples n_1 and n_2 from Π_1 and Π_2, respectively. Anderson [2] classification statistic is given by

$$W = \left\{\mathbf{X} - \frac{1}{2}(\bar{X}_1 - \bar{X}_2)\right\}' \mathbf{S}^{-1}(\bar{X}_1 - \bar{X}_2). \tag{6}$$

If $P_i(\mathbf{w})$ is the probability density of $\mathbf{w}$ in Π_i $(i = 1, 2)$, it can be shown that the optimal rule (assuming equal costs and equal a priori probabilities) is to assign $\mathbf{w}$ to Π_1 if $P1(\mathbf{w})/P2(\mathbf{w}) > 1$, and otherwise to Π_2. Now $P_i(\mathbf{w}) = P_i(\mathbf{x},\mathbf{y}) = P_i(\mathbf{x})\cdot P_i(\mathbf{y}/\mathbf{x})$ $(i = 1, 2)$. Conditioning first on $\mathbf{x}$, we suppose that the observation falls in cell m. Thus, $P_i(\mathbf{x},\mathbf{y}) = P_{im}\cdot P_i(\mathbf{y}/z_m)$. Hence, using Equation (1), we deduce that the allocation rule for observation $(\mathbf{x}', \boldsymbol{y}')$ is as follows:

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

Allocate to Π_1 if

$$(\boldsymbol{\mu}_1^{(m)} - \boldsymbol{\mu}_2^{(m)})' \boldsymbol{\Sigma}^{-1} \left\{ \mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_1^{(m)} + \boldsymbol{\mu}_2^{(m)}) \right\} \geq \log \left(\frac{p_{2m}}{p_{1m}} \right). \tag{7}$$

and otherwise to Π_2 with cutoff points determined in each case by the discrete component of the model. The probabilities of misclassification $p(2/1)$ and $p(1/2)$ from Π_1 and Π_2, respectively, associated with location model may be derived from Equation (7) by writing

$$\boldsymbol{D}_{\boldsymbol{m}}^2 = (\boldsymbol{\mu}_1^{(m)} - \boldsymbol{\mu}_2^{(m)})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1^{(m)} - \boldsymbol{\mu}_2^{(m)}) \tag{8}$$

for the Mahalanobis squared distance between Π_1 and Π_2 conditional on the observation falling in multinomial cell m and noting that the overall probability of misclassification from Π_i is the sum of the probabilities of misclassification for each multinomial cell of Π_i weighted by the probability of its occurrence. Thus

$$p\left(\frac{2}{1}\right) = \sum_{m=1}^{k} p_{1m} \emptyset \left\{ \frac{(\log(p_{2m}/p_{1m}) - (1/2)D_m^2)}{D_m} \right\}, \tag{9}$$

$$p\left(\frac{1}{2}\right) = \sum_{m=1}^{k} p_{2m} \emptyset \left\{ \frac{(\log(p_{1m}/p_{2m}) - (1/2)D_m^2)}{D_m} \right\}. \tag{10}$$

The utility of an allocation rule can be assessed by the probabilities of misclassification, or ER that it gives rise to. When parameters are known in the location model, the ERs are obtained by adding Equations (9) and (10). When the parameters are unknown, various other types of ER estimation can be considered. This is very essential so as to have a measure of its utility and to be able to assess its performance relative to other allocation rules. A method that is achievable and amenable to the location model is that proposed by Lachembruch and Mickey [18] in which each unit of the initial samples is classified in turn, using the remaining $n_1 + n_2 - 1$ units to obtain the allocation rule. The ER from each population is then estimated by the proportion of units/observation misclassified in each sample using this method. The overall ER is the sum of ERs from each population. A great deal of work would be required, even for small samples, if the whole iterative scaling and multivariate regression had to be repeated whenever a sample point is omitted from the calculation of the ER. However, a simplification can be achieved by the matrix inversion approach suggested by Bartlett [4]. Lachembruch's method for estimating the ERs using the allocation rule derived from the location model is computationally feasible. However, a drawback encountered in practice is that resulting in zero cell estimates whenever a unit is omitted. In such circumstances, the original estimates of p_{1m} and p_{2m} from the sample have been taken. This will introduce some bias, but the circumstances have been sufficiently rare in practice for this bias to be negligible. Alternatively, rather than formulating another rule, it is easiest to consider such an individual as contributing .5 to the overall ER [15].

3. Application

3.1. Data used

Real and simulated data were used. The real data are surveys on hypertensive and diabetic patients carried out at The Federal Medical Centre, Bida, Niger State, Nigeria. Data-set I

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

comprises 32 individuals (19 hypertensive and 13 non-hypertensive) with 4 continuous and 3 binary variables. Data-set II comprises 80 individuals (36 diabetic and 44 non-diabetic) with 5 continuous and 4 binary variables. Data-set III comprises 153 individuals (86 hypertensive and 67 non-hypertensive) with 3 continuous and 2 binary variables. These three different data-sets would represent different situations under which comparison shall be made and from them we shall obtain parameters for simulation. The simulated data comprise samples of size $n = 30$, 50 and 100 representing small, medium and large sample sizes, respectively, in each of the situations been considered. A careful setup was used to generate multivariate continuous normal and uniformly distributed binary data using the R statistical package for various sample sizes and corresponding analysis carried out accordingly. The setup was such that a cross validation analysis (using methods amenable for each model) was done to assess the ER produced by each model. The ER for each technique is obtained by counting the number of observations misclassified in each group by each technique in each sample and this number is divided by the corresponding sample size.

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

4. Discussion of result

From Table 1, analysis using the real data, in data-set 1, the location model performed better than both the FLDF and logistic regression with the FLDF providing the poorest result. In set 2, the location model also performed better than the other two methods with a larger sample size and number of both continuous and binary variables than we had in set 1. In set 3, with an increased sample size and lesser number of continuous and binary variables, the logistic regression method performed better while the location model had an equal performance with the FLDF.

From Table 2, the analysis of the simulated data, a general look at Table 2 will reveal a better performance of the location model on comparison with the FLDF. The logistic regression method performed better than the location model in few cases, the two methods had comparably equal performance in a large sample of set 2, however, the location model performed better than the logistic regression and FLDF in the medium and large samples of set 1 and set 2. The logistic regression also performed better than the FLDF in all the cases considered. The analysis using the location model was difficult and therefore was not performed in a small sample of set 2, reason being that most of the cells had no observations falling in them due to the number of binary variables involved in the data-set. The number of $k = 2^q$ cells that requires estimating probabilities is more than the sample size observed and available. The performance of the location model on the average increased positively with relative increase in sample size when compared with the FLDF except for set 1 where

Table 1. Number of misclassification and the ERs obtained in the three data-sets using the real data.

		Location model			FLDF			Logistic regression		
Data sets	Sample sizes N	Π_1	Π_2	ER	Π_1	Π_2	ER	Π_1	Π_2	ER
1	32	2	3	.156	7	6	.406	3	4	.213
2	80	7	7	.175	15	13	.35	12	8	.25
3	154	21	16	.243	22	15	.213	13	21	.222

Table 2. Number of misclassification and the ERs obtained in the three data-sets from the simulation study.

Data sets	Sample sizes *N*	Location model			FLDF			Logistic regression		
		Π_1	Π_2	ER	Π_1	Π_2	ER	Π_1	Π_2	ER
1	30	9	3	.400	7	3	.333	4	2	.200
	50	9	7	.322	12	13	.500	9	9	.360
	100	23	13	.360	29	19	.480	15	30	.450
2	30	**	**	**	6	4	.333	1	1	.067
	50	3	7	.200	9	12	.420	7	8	.300
	100	12	25	.370	21	32	.530	23	14	.370
3	30	4	6	.333	6	8	.467	4	4	.267
	50	11	7	.360	13	12	.500	11	7	.360
	100	22	17	.390	30	20	.500	3	32	.350

Note: **Analysis was not possible.

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

its performance reduced. The FLDF had a performance that appears to be approximately the same with and/or without increase in sample size.

Therefore, we can deduce that the performance of any allocation rule varies from sample to sample. This result agrees with Cochran and Hopkins [5]. Hence, varying sample sizes could be said to have effect on the performance of any allocation rule. Also, the location model, though was not consistently better in all the cases considered yet has not yielded any poor result, it rather performed alongside with the supposed 'classical and best' methods. The location model provides a method that is considerably better than the distortive technique of the FLDF. The performance of the logistic regression could possibly be attributed to the fact that the logistic discrimination analysis procedure ensures that variables are treated the way they are and it does not make any assumption about the distribution of the observed data, if it does, perhaps, the performance might not be this better. This result agrees with Ognian and Krzanowski [19] and Efron [6]. The varying number of variables has not significantly reduced the performance of the location model owing to a limit to the possible number of binary variables it can handle and cope with.

5. Conclusion

The location model so far has performed considerable and satisfactory for situations where we have mixtures of binary and continuous variables. The method that has been suggested and applied here will cope with mixtures of all types of variables in the discriminant analysis without altering the natural form of the data-set by either making transformations or assigning scores and codes to the observed data. It will do so without making any unrealistic assumptions implicit whenever a categorical variable is converted into a numerical one by arbitrary scoring. Effort was not made in this work to compare this result with other proposed methods of dealing with categorical (including binary) and continuous variables in the discriminant analysis; however, it has addressed this issue using some common and widely used methods as a benchmark. Thus, the suggested method might be considered as an alternative when the analysis involves mixtures of continuous and binary variables with lesser number of binary variables and a larger sample size. The logistic discrimination would have been considered in any case but due to its weakness in terms of making assumption about the distribution of the data, it might not be a better option. The FLDF

though has often been used conventionally, might not be considered for use in this situation especially with large sample size but since we interested in the distortions caused by using the FLDF to treat binary variables as if they are continuous, it might not also be considered. Our proposed method is computationally feasible and produces a reasonable advantage in terms of time and effort required.

Acknowledgements

We wish to acknowledge the reviewers and the editor of this paper for their constructive comments and inputs. We also wish to thank Dr Abu of Federal Medical Centre, Bida, for supplying the data-sets used.

Disclosure statement

No potential conflict of interest was reported by the authors.

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016

References

[1] J. Aitchson and C.G.G. Aitken, *Multivariate binary discrimination by the kernel method*, Biometrika 63 (1976), pp. 413–420.
[2] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, Inc., New York, 1958.
[3] J.A. Anderson, *Separate sample logistic discrimination*, Biometrika 59 (1972), pp. 19–35.
[4] M.S. Bartlett, *An inversion matrix adjustment arising in discriminant analysis*, Ann. Math. Stat. 22 (1951), pp. 107–111.
[5] W.G. Cochran and C.E. Hopkins, *Some classification problems with multivariate qualitative data*, Biometrics 17 (1961), pp. 10–32.
[6] B. Efron, *Defining the curvature of a statistical problem (with application to second order efficiency)*, Ann. Stat. 3 (1975), pp. 1189–1242.
[7] S.E. Fienberg, *An iterative procedure for estimation in contingency tables*, Ann. Math. Stat. 41 (1970), pp. 907–917.
[8] R.A. Fisher, *The statistical utilization of multiple measurements*, Ann. Eng. Lond. 7 (1938), pp. 179–88.
[9] R. Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley & Sons, New York, 1977.
[10] S.J. Haberman, *Log linear fit for contingency tables. Algorithm AS51*, Appl. Stat. 21 (1972), pp. 218–225.
[11] M. Halperin, W.C. Blackwelder, and J.I. Verter, *Estimation of the multivariate logistic risk function: A comparison of the discriminant function and maximum likelihood approaches*, J. Chronic Dis. 24 (1971), pp. 125–158.
[12] S.P. James and W. Sandra, *Choosing between logistic regression and discriminant analysis*, J. Am. Stat. Assoc. 73 (1978), pp. 699–705.
[13] J.D. Knoke, *Discriminant analysis with discrete and continuous variables*, Biometrics 38 (1982), pp. 191–200.
[14] W.J. Krzanowski, *Discrimination and classification using both binary and continuous variables*, J. Am. Stat. Assoc. 70 (1975), pp. 782–790.
[15] W.J. Krzanowski, *Mixtures of continuous and categorical variables in discriminant analysis*, Biometrics 36 (1980), pp. 493–499.
[16] W.J. Krzanowski, *Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis testing approach*, Biometrics 38 (1982), pp. 991–1002.
[17] P.A. Lachembruch and M. Goldstein, *Discriminant analysis*, Biometrics 35 (1979), pp. 69–85.
[18] P.A. Lachembruch and M.R. Mickey, *Estimation of error rates in discriminant analysis*, Technometrics 10 (1968), pp. 1–11.

[19] K.A. Ognian and W.J. Krzanowski, *A comparison of discriminant procedures for binary variables*, Comput. Stat. Data Anal. 38 (2001), pp. 139–160.
[20] I. Olkin and R.F. Tate, *Multivariate correlation models with mixed discrete and continuous variables*, Ann. Math. Stat. 32 (1961), pp. 448–465.
[21] E. Parzen, *On estimation of a probability density function and mode*, Ann. Math. Stat. 33 (1962), pp. 1065–1076.
[22] J. Truett, J. Cornfield, and W. Kannel, *A multivariate analysis of the risk of coronary heart disease in Framingham*, J. Chronic Dis. 20 (1967), pp. 511–524.
[23] H. Warner, A. Toronto, L. Veesey, and R. Stephenson, *A mathematical approach to medical diagnosis*, J. Am. Med. Assoc. 177 (1961), pp. 177–183.

Downloaded by [GEORGE FIGURES] at 19:32 23 February 2016