International Journal of Applied Mathematics and Statistics,
Int. J. Appl. Math. Stat.; Vol. 42; Issue No. 12; Year 2013,
ISSN 0973-1377 (Print), ISSN 0973-7545 (Online)

Optimal Ratio of Continuous to Categorical Variables for the Two Group Location Model

Philemon Baah[1], **Atinuke Adebanji**[1] **and Romain Glèlè Kakaï**[2]

[1]Department of Mathematics,
Kwame Nkrumah University of Science and Technology,
Kumasi, Ghana
brophily@gmail.com; tinuadebanji@gmail.com

[2]Faculty of Agronomic Sciences, University of Abomey-Calavi,
01 BP 526, Cotonou, Bénin
gleleromain@yahoo.fr

ABSTRACT

We investigated the effect of different combinations of (p) *continuous to* (q) *categorical variables and increasing group centroid separation function* $(\delta = 1, 2, 3)$ *on the performance of the Location model for two groups* $(\Pi_i, i = 1, 2)$*. The number of predictor variables were 4 and 8 with 1:3, 1:1 and 3:1 being the predetermined ratios for* $p : q$*. We generated* $N(\mu_1, \mathbf{I})$ *of sizes 40, 80 and 120 with MatLab R2007b for* p *variables within* 2^q *binary cells in* Π_1*. The size of* Π_2 *was determined using sample ratios 1:1, 1:2, 1:3 and 1:4 for* $n_1 : n_2$ *within* 2^q *cells. Group1 has mean* $\mu_1^{(1)} = \mathbf{0}$ *in the first cell (for* p *continuous variables) and* $\mu_2^{(1)} = \delta$*, subsequent cells,* $\mu_i^{(m+1)} = \mu_i^{(m)} + \mathbf{1}$ *. Error rates reduced more rapidly for increase in* δ *than asymptotically. The optimal* $p : q$ *was 3:1 and the model deteriorated at 1:3 with larger variability. The 8 variable model performed better than the 4 variable model for large sample sizes of* $p : q = 1 : 1$ *and outperformed it for all sample sizes of* $p : q = 3 : 1$*. Results showed that to use the Location model for classification problems with equal (or more) categorical to continuous variables, it should be compensated with increased distance function and sample sizes.*

Keywords: Location model, classification, categorical to continuous variables, contingency table, leave-one-out method
2000 Mathematics Subject Classification: 62H30, 62H17

1 Introduction

Traditionally, discriminant analysis is used for differentiating groups (categorical dependent variables) which are known *a priori* while the independent variables are quantitative and normally distributed. When the independent variables used in discriminant analysis constitute both qualitative (discrete) and quantitative (continuous), a familiar technique is the application of *the location model*, which was first proposed by Olkin and Tate (1961). The model assumes that

the conditional distribution of the continuous variables given the discrete variables are multivariate normally distributed with constant covariance matrix across all locations determined by the discrete variables (McLachlan, 1992). Chang and Afifi (1974) extended the concept of the location model to two-group situations deriving a Bayes classification procedure for classifying an observation consisting of both dichotomous and continuous variables. A generalization of their results has been considered by Krzanowski (1975), in which optimum and estimated allocation rules were derived for mixed binary and continuous variables using likelihood ratio. He later looked at the location model for mixtures of all types of variables (Krzanowski, 1980), and when there exists more than two differentiating groups for more general discrete and continuous mixtures (Krzanowski, 1986).

In many multivariate situations the statistician is presented with data on a very large number of independent variables, and the question arises whether they are all necessary and if not which can be discarded. In discriminant analysis the problem is to choose a subset of the available variables without seriously impairing the discriminating power of the set. Murray (1977) in a simulated study presented three different procedures used in selecting subsets of available variables. Selection of variables for mixtures of continuous and discrete variables, with reference to the location model, has also been looked at in literature (e.g. Krzanowski, 1983; Gutiérrez, Merbouha, Gutiérrez-Sánchez and Nafidi, 2008; Hamid, 2010).

In using the location model for discriminatory problems, one has to limit the number of discrete variables, otherwise the number of parameters to be estimated will be excessive. Krzanowski (1983) suggested an upper limit of six binary variables if the sizes of the initial samples available from each group are not large, with corresponding reduction in number when some variables have more than two states. If the sizes of the initial samples are large, the computational effort needed to estimate error rates was found to increase disproportionately with the number of discrete variables, which becomes a problem. Krzanowski (1983) therefore proposed a backward elimination method of discrete variable selection which can be used to identify a suitable, reduced location model for discriminant applications when the number of discrete variables are too large for direct use. In another instance, some authors proposed a combination of both non-parametric smoothing and regularization to address the problem of over-parameterisation and instability of the covariance matrix in the location model (Gutiérrez et al., 2008). More recently, Hamid (2010) proposed an idea that integrates a dimensionality reduction technique via principal component analysis and a discriminant function based on the location model. The aim is to offer another technique of classification when the observed variables are mixed and too large.

Undoubtedly, in order to reduce the number of parameters estimated and to overcome the problem of instability or singularity in the location model, reducing the number of discrete variables in the location model and the number of variables in discriminant analysis as a whole has been the concern of researchers for some time now. Since the estimation of parameters in the location model is based on the number of multinomial cells created by the discrete variables

and the continuous variables as well, we looked at the problem in which the number of binary variables is a scalar multiple of the continuous variables. The aim was to find a continuous-binary variable ratio combination which will give minimum error rates of misclassification in the location model for the two group case.

2 The Location Model and Estimation of its Error Rates

Let $\mathbf{v}$ denote a random vector of observations made on any individual which is a mixture of q binary variables $\mathbf{x}$ and p continuous variables $\mathbf{y}$. The contingency table formed from $\mathbf{x}$ has $s = 2^q$ locations or cells; and denote these locations by $z_1, z_2, \ldots, z_s$. Then the location model (LM) as proposed by Olkin and Tate (1961) assumes that the conditional distribution of $\mathbf{y}$ given that $\mathbf{x}$ falls in location z_m is $N_p(\mu^{(m)}, \mathbf{\Sigma})$ and the marginal distribution of the locations is given by $P(z = z_m) = p_m,$ with $\sum_{m=1}^{s} p_m = 1$.

From the normality assumption of the model, the conditional probability density of $\mathbf{y}$, given that the binary variables locate the individual in cell m, is

$$\frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{y} - \mu_i^{(m)})'\mathbf{\Sigma}^{-1}(\mathbf{y} - \mu_i^{(m)})\}$$

in $\Pi_i,\ (i = 1, 2)$. Thus the joint probability density of obtaining the individual cell m and observing the continuous variable values $\mathbf{y}$ is

$$\frac{p_{im}}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{y} - \mu_i^{(m)})'\mathbf{\Sigma}^{-1}(\mathbf{y} - \mu_i^{(m)})\}$$

in $\Pi_i,\ (i = 1, 2)$. Hence we deduce that the Bayes allocation rule for an observation $\mathbf{v}' = (\mathbf{y}', \mathbf{x}')$ is: allocate to Π_1 if

$$(\mu_1^{(m)} - \mu_2^{(m)})'\mathbf{\Sigma}^{-1}\{\mathbf{y} - \frac{1}{2}(\mu_1^{(m)} + \mu_2^{(m)})\} \geq \ln(p_{2m}/p_{1m}); \tag{2.1}$$

and otherwise to Π_2. (Krzanowski, 1975)

In practice, the parameter values are unknown and the usual procedure is to replace the optimum classification rule (2.1) by a sample-based allocation rule from given training sample. The classical approach of using a sample-based allocation rule is to use the allocation rule (2.1) but to replace the parameters by their unbiased maximum likelihood estimates. Let n_{1m} and n_{2m} denote the number of observations falling in cell m of the multinomial cells from Π_1 and Π_2, and let $\mathbf{y}_{ji}^{(m)}$ denote the vector of continuous variables associated with the jth observation in cell m of the sample from Π_i. Then the unbiased maximum likelihood estimates are:

$$\hat{p}_{im} = \frac{n_{im}}{n_i};\ \hat{\mu}_i^{(m)} = \bar{\mathbf{y}}_i^{(m)} = \frac{1}{n_{im}} \sum_{j=1}^{n_{im}} \mathbf{y}_{ji}^{(m)}$$

and

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

(Krzanowski, 1975).
Once an allocation rule has been derived, it is important to have a reliable method for estimating the error rates that it gives rise to, in order to assess its performance relative to other classification rules since parameters are rarely known (McLachlan, 1992). A non-parametric method which is known to work better than most other estimative procedures is the leave-one-out method of Lachenbruch and Mickey (1968). This method is also adopted in this paper because of its relatively good performance compared with other error rate estimators (Glèlè Kakaï and Palm, 2004; Kakaï and Palm, 2009) and also because it is available in most of the statistical software. The error rate from each group is estimated as the proportion of observations misclassified from each sample.

3 Simulation Design and Efficiency Criteria

Data were simulated for two groups, Π_1 and Π_2 with MatLab R2007b. We generated $N(\mu_1, \mathbf{I})$ where $\mathbf{I}$ (an identity matrix) denotes the common covariance matrix across the groups and cells, of sizes 40, 80 and 120 for p variables within 2^q binary cells in Π_1. The size of Π_2 is determined using sample ratios 1:1, 1:2, 1:3 and 1:4 for $n_1 : n_2$ within 2^q cells. Group1 has mean $\mu_1^{(1)} = \mathbf{0}$ in the first cell (for p continuous variables) and $\mu_2^{(1)} = \delta$ (where $\delta = 1, 2, 3$ is the group centroid separator), subsequent cells, $\mu_i^{(m+1)} = \mu_i^{(m)} + \mathbf{1}$, in this case we have restricted the mean of each continuous variable to be a positive integer. The number of predictor variables are 4 and 8 with 1:3, 1:1 and 3:1 being the predetermined ratios for $p : q$. The following was also considered during the data simulations: we let $\hat{p}_{im} = \hat{p}_{ij}$ for $m \neq j^{th}$ locations, which implies the probability of locating an observation in, say, the mth cell is $\hat{p}_m = 1/2^q$, a constant. That is, the total number of observations across the locations are the same with $n_{im} = n_{ij}$ for $m \neq j$. Then $\hat{p}_{im} = \hat{p}_i$, the estimated prior probability of $\Pi_i,\ (i = 1, 2)$. In order to overcome the problem of singularity of the covariance matrices within locations, we ensured that $n_{1m} + n_{2m} - 2 \geq p$ for the mth cell (Johnson and Wichern, 2007, pp. 591).

For each combination of factors considered, thirty samples of different sizes given above were generated and the leave-one-out error rate estimator was used in estimating the error rates in each case . The average error rates of classification together with their standard deviations and coefficients of variation (variations in short) were computed for the thirty replications.

4 Results

4.1 Presentation of the Results

Results of the misclassification rates obtained from the simulations is presented in Table 1 and also pictorially using comparative box plots in Figures 1 to 3, for all factor combinations. The Table is displayed as follows. The first column is the total number of mixed continuous and binary predictor variables $p + q$; the second column of the table is the total sample size $n_1 + n_2$ used in the simulations predetermined by the sample ratios $n_1 : n_2 = 1 : 1, 1 : 2, 1 : 3, 1 : 4$; the next three major columns are the misclassification rates obtained for the centroid

separators $\delta = 1, 2, 3$. The results for the different δ's is made up of three subcolumns each. The subcolumns are the misclassification rates for the continuous-binary ratio combinations $p : q = 1 : 3, 1 : 1, 3 : 1$. For each $p + q = 4$ and 8 number of predictor variables, results are displayed for the different sample sizes of group1 ($n_1 = 40, 80, 120$). In order to avoid singularity of covariance matrices across locations, simulations were not carried out for continuous-binary ratio $1 : 3$ for $p + q = 8$ and for $n_1 : n_2 = 1 : 1$ when $n_1 = 40$ for the same $p + q$.

4.2 Error Rate of Misclassification of the Location Model According to the Factors Considered

Table 1: Mean error rates of misclassification of the location model for the different factor combinations

nvar.[a]	S/Size[b]	$\delta = 1$			$\delta = 2$			$\delta = 3$		
		Var. Ratio[c]			Var. Ratio			Var. Ratio		
		1:3	1:1	3:1	1:3	1:1	3:1	1:3	1:1	3:1
4	$n_1 = 40$									
	80	0.1754	0.1394	0.1029	0.0867	0.0490	0.0256	0.0346	0.0125	0.0031
	120	0.1515	0.1211	0.0931	0.0761	0.0368	0.0201	0.0335	0.0100	0.0026
	160	0.1191	0.0991	0.0766	0.0651	0.0373	0.0200	0.0295	0.0101	0.0019
	200	0.0974	0.0795	0.0689	0.0593	0.0299	0.0181	0.0302	0.0073	0.0019
	$n_1 = 80$									
	160	0.1537	0.1215	0.1030	0.0776	0.0457	0.0250	0.0333	0.0098	0.0034
	240	0.1410	0.1118	0.0917	0.0746	0.0399	0.0203	0.0313	0.0081	0.0023
	320	0.1145	0.0928	0.0788	0.0667	0.0337	0.0171	0.0278	0.0081	0.0028
	400	0.0944	0.0822	0.0684	0.0577	0.0303	0.0163	0.0263	0.0073	0.0020
	$n_1 = 120$									
	240	0.1587	0.1265	0.0982	0.0810	0.0407	0.0227	0.0372	0.0099	0.0019
	360	0.1357	0.1110	0.0908	0.0736	0.0390	0.0208	0.0314	0.0082	0.0016
	480	0.1121	0.0967	0.0793	0.0653	0.0347	0.0175	0.0290	0.0070	0.0017
	600	0.0954	0.0799	0.0664	0.0586	0.0301	0.0153	0.0258	0.0072	0.0016
8	$n_1 = 40$									
	120	-	0.1650	0.0736	-	0.0714	0.0068	-	0.0321	0.0004
	160	-	0.1157	0.0623	-	0.0389	0.0058	-	0.0090	0.0001
	200	-	0.1005	0.0558	-	0.0316	0.0052	-	0.0059	0.0000
	$n_1 = 80$									
	160	-	0.1355	0.0670	-	0.0483	0.0059	-	0.0143	0.0005
	240	-	0.1044	0.0610	-	0.0266	0.0039	-	0.0041	0.0001
	320	-	0.0911	0.0511	-	0.0209	0.0047	-	0.0023	0.0001
	400	-	0.0744	0.0463	-	0.0166	0.0039	-	0.0015	0.0000
	$n_1 = 120$									
	240	-	0.1104	0.0641	-	0.0246	0.0058	-	0.0044	0.0000
	360	-	0.0933	0.0595	-	0.0197	0.0043	-	0.0019	0.0000
	480	-	0.0767	0.0498	-	0.0148	0.0039	-	0.0014	0.0001
	600	-	-	-	-	0.0135	0.0033	-	0.0015	0.0000

[a] Total Number of Variables
[b] Total Sample Size
[c] Continuous-Binary Variable Ratio

The table above shows that as δ increases from 1 to 3, the mean error rates decrease for all variables and continuous-binary variable ratio combinations for each total sample size. The following pattern is also evident in general. As the total sample size increases, the error rates decrease appreciably for all factor combinations, the error rates were found to be smaller for the 8 variable models than the 4. There was also decrease in error rates we move the continuous-binary variable ratios from $1 : 3$ through $3 : 1$.

The graphical display of results are presented as pairs of figures. In each pair, the graph for the mean error rates is on the left with that of the variations on the right. Each pair represents a matrix of comparative boxplots for the different centroid separators (δ's) considered. Also, each graph (either for mean error rates or variations) comprises of a 1×3 matrix plots, each

column showing results for the different sizes of group1, n_1 predefined. For each n_1, the plots show results for the number of variables considered ($nvar = 4, 8$) and the continuous-binary variable ratios ($var\ ratio = 1:1, 1:3, 3:1$).

4.3 Stability of the Performance of the Location Model According to the Factors Considered

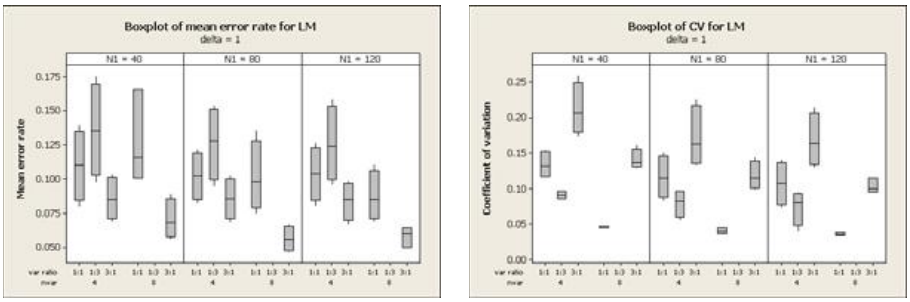

Figure 1: Box plots of mean error rates of classification and coefficients of variation for $\delta = 1$

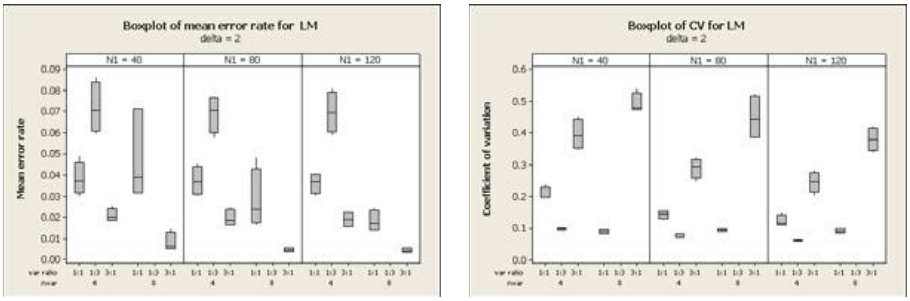

Figure 2: Box plots of mean error rates of classification and coefficients of variation for $\delta = 2$

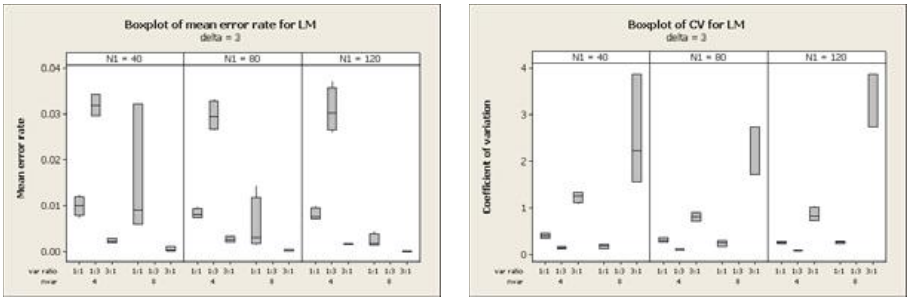

Figure 3: Box plots of mean error rates of classification and coefficients of variation for $\delta = 3$

The box plots for each n_1 shows comparative results for $p : q = 1 : 3, 1 : 1, 3 : 1$ for $nvar = 4$ and $p : q = 1 : 1, 3 : 1$ for $nvar = 8$. For the different δ's considered, variable ratio $p : q = 1 : 3$ recorded the highest error rate with that of $p : q = 3 : 1$ having the least for $nvar = 4$. For both $nvar = 4$ and 8, $p : q = 3 : 1$ had the least error rate though it recorded higher variability

compared to that of $p : q = 1 : 1$. The variations were also found to be very high as compared to their mean error rates, especially for $p : q = 3 : 1$. As δ increased from 1 to 3, the error rates were found to decrease but with increasing variations. Though the error rates for $\delta = 3$ were very low, the observed variations were very high.

4.4 Discussion

The location model is known to give much better results than the traditional linear discriminant analysis for a mixed variable case (e.g. Krzanowski, 1975; Krzanowski, 1980; Vlachonikolis and Marriott, 1982). Problem however arises when the number of qualitative variables are large, and has necessitated various variable selection procedures (e.g. Krzanowski, 1983; Krzanowski, 1995; Gutiérrez et al., 2008; Hamid, 2010; Kakaï and Palm, 2009; Kakaï, Pelz and Palm, 2010). We assessed the error rates of the location model through Monte Carlo experiments for the two groups classification model. The simulation design took into account the total number of mixed continuous-binary variables, their respective ratios, the distance between the two groups and the total sample size.

The results showed a reduction in error rates when the sample size of group1, n_1 increased from 40 through to 120 for all factor combinations. In general it can also be inferred that an increase in the total sample size resulted in decreased rates of misclassification (see Table 1). The model also recorded an improvement when the group centroid separator was increased. The rates of misclassification were minimal (Lei and Koehly, 2003). Based on our simulation studies, the error rates decreased as we increased the distance between the populations, δ from 1 to 3 but with increased variability in reported error rates. (Figures 1 to 3).

The main finding of this study is the recommended number of mixed continuous-binary variables and their respective ratios. The study could only take into account two sets of mixed variables – 4 and 8. Under the various factor combinations respectively for both the 4 and 8 mixed variables, a close look at the results shown in Table 1 shows that the 8 variable recorded the least error rates (see also Figures 1 to 3). However, there is little deviation to this conclusion drawn only for $n_1 = 40$, $p : q = 1 : 1$, and $\delta = 1$ and 2. It can also be seen from the boxplots (shown in Figures 1 to 3) that the variations recorded for the two sets of mixed variables was lower for the 8 variable cases than for the 4 $nvar$ under the respective n_1's and continuous-binary variable ratios $p : q = 1 : 1$, 3:1 and for $\delta = 2$,3. In the results for $\delta = 1$ the variations of $p : q = 3 : 1$ were lower for the 4 than the 8 variable. As far as the continuous-binary variable ratios were concerned, the results for the 4 variable model showed that the ratio $p : q = 1 : 3$ recorded the highest error rate with $p : q = 3 : 1$ having the least. Thus, as the group centroids moved further apart, the performance of the function improved especially for $p : q = 3 : 1$ and fairly large sample sizes. Though the error rates displayed larger variability, increasing the sample size improved the stability.

5 Conclusion

This study has shown that the performance of the location model improved more rapidly when the distance factor δ increased than when the sample size increased asymptotically. The $p : q = 3 : 1$ model recorded the least error rates for all sample sizes considered but there was an increase in the variability of the reported erro rates. The model reported considerably higher error rates for $p : q = 1 : 3$ although the reported rates were less volatile than those observed for the the $p : q = 3 : 1$ model. The 8 variable model performed marginally better than the 4 variable model for large sample sizes of $p : q = 1 : 1$ and outperformed it for all sample sizes of $p : q = 3 : 1$. The 8 variable model with continuous to binary variable ratio $p : q = 3 : 1$ was found to be the optimum allocation model is the . We conclude that to use the location model for classification problems with equal (or more) categorical to continuous variables, it should be compensated with increased distance function and large samples.

References

Chang, P. C. and Afifi, A. A. (1974). Classification based on dichotomous and continuous variables, *Journal of the American Statistical Association* **69**(346): 336–339.

Glèlè Kakaï, R. and Palm, R. (2004). Performance relative des règles linéaire, quadratique et logistique en analyse discriminante, *XXXVIèmes Journes de Statistique* .

Gutiérrez, R., Merbouha, A., Gutiérrez-Sánchez, R. and Nafidi, A. (2008). Non-parametric smoothing and regularization of the location model in mixed variable discrimination, *Monografías del Seminario Matemático García de Galdeano* **34**: 107116.

Hamid, H. (2010). A new approach for classifying large number of mixed variables, *World Academy of Science, Engineering and Technology* **70**: 156–161.

Johnson, R. A. and Wichern, D. W. (2007). *Applied multivariate statistical analysis*, sixth edn, Pearson Education, Inc., NJ.

Kakaï and Palm, R. (2009). Empirical comparison of error rate estimators in logistic discriminant analysis, *Journal of Statistical Computation and Simulation* **2**(79): 111–120.

Kakaï, R. L. G., Pelz, D. and Palm, R. (2010). On the efficiency of the linear classification rule in multi-group discriminant analysis, *African Journal of Mathematics and Computer Science Research* **3**(1): 19–25.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables, *Journal of the American Statistical Association* **70**(352): 782–790.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis, *Biometrics* **36**(3): 493–499.

Krzanowski, W. J. (1983). Stepwise location model choice in mixed-variable discrimination, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **32**(3): 260–266.

Krzanowski, W. J. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data, *Comp. & Maths. with Appls.* **12A**(2): 179–185.

Krzanowski, W. J. (1995). Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis, *Computational Statistics & Data Analysis* **19**: 419–431.

Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis, *Technometrics* **10**(1): 1–11.

Lei, P. and Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case, *The Journal of Experimental Education* **72**(1): 25–49.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*, John Wiley and Sons, Inc., Hoboken, New Jersey.

Murray, G. D. (1977). A cautionary note on selection of variables in discriminant analysis, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **26**(3): 246–250.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables, *The Annals of Mathematical Statistics* **32**(2): 448–465.

Vlachonikolis, I. G. and Marriott, F. H. C. (1982). Discrimination with mixed binary and continuous data, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **31**(1): 23–31.