FUNAI Journal of Science and Technology
3 (2), 2017, 149-158

ON ESTIMATION OF COVARIANCE MATRIX WITH MIXTURES OF CONTINUOUS AND CATEGORICAL VARIABLES

Oyeyemi, G. M.[1] & Mbaeyi, G. C.[2]

[1]*Department of Statistics, University of Ilorin, P. M. B. 1515, Ilorin, Kwara State.*
[2]*Department of Mathematics/Computer Science/Statistics/Informatics, Federal University Ndufu Alike Ikwo, Nigeria.*

(Received 20 December 2016; Revised 22 August 2017; Accepted:25 August, 2017)

Abstract

A method known as the Location Model for discriminant analysis when discriminating variables are mixtures of continuous and categorical variables, which, unlike the conventional Linear Discriminant procedure does not assign arbitrary scores to each state of the categorical variable was studied. An alternative to estimating the covariance matrix for evaluating the discriminant function was suggested since the Location Model still makes such distortion of treating categorical variable as if they are continuous when estimating the covariance matrix. We compare the performance of the two procedures using the accuracy rate produced by each method under various conditions. Our suggested method performed better than the location model over all cases considered.

Keywords: *continuous, categorical, covariance matrix, accuracy rate, discriminant, groups*.

.

1. Introduction

Consider the discriminant function below

$$W = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)'\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{y} - \frac{1}{2}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2)\right\}$$

$\overline{\boldsymbol{x}}_1$ is the vector of means for group I, $\overline{\boldsymbol{x}}_2$ is the vector of means for group II, $\boldsymbol{y}$ is the vector of a new observation and $\boldsymbol{\Sigma}$ is the common variance-covariance matrix. Assuming $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and x_{i1} represents Gender (coded as, say, Male=1 and Female=0 or as the analyst chooses). Can we have any reasonable numerical quantity as "mean of gender"?

In a situation where available data set for a discriminant analysis is a mixture of both the continuous and categorical variable type, attention is most times never given to the nature of the data. Many areas of biometry involve the collection of multivariate data in which observed variates ranges from

Correspondence: E-mail: george.chinanu@funai.edu.ng

continuous variables to categorical variables that are unordered. Most conventional and common procedure ignore the discrete nature of the categorical data and proceed with analysis as if all the data are continuous. When available data set is a mixture of continuous and categorical variables, the simplest and conventional procedure, the Linear Discriminant Function (LDF), is to assign an arbitrary numerical score to the possible states of the categorical variables and proceed as if all variables are continuous. If such scoring is not possible, Cochran and Hopkins (1961) suggested that one can perhaps categorize the continuous variables. Consequently, a categorical variable technique can be applied on all the data (Glick, 1973). Alternatively, one can separate the two types of variables and perform two separate analyses, using the appropriate technique for each of the data type (Krzanowski, 1980). Worthy to note is the fact that, most of the procedures that are silent about the discrete nature of the categorical variables are likely to always give better classification accuracy and they are easy to implement. Hence, the aim of this work is develop a procedure based on the Location model for estimating the variance-covariance matrix when data is a mixture of continuous and categorical variables. Some distinct procedures to the treatment of mixtures of continuous and categorical variables in discriminant analysis are evident in literature. These procedures have suggested ways of possibly combining the different types of variables optimally. Fisher (1936) suggested the use of the linear discriminant function by assigning arbitrary score to each possible state of the discrete variable Another procedure (hereafter referred to as Location Model) which have sufficiently combined the two variable types has been proposed by Okin and Tate (1961) and developed by Krzanowski (1975, 1980, and 1982) to handle the problem of discrimination with mixtures of continuous and categorical variables. The location model developed by Krzanowski (1975) is a parametric likelihood ratio based discriminant function derived from a probabilistic model for mixed categorical and continuous variables. Liang *et al* (2008), Lee *et al* (2008), Huberty *et al* (2010) and De-Leoan and Willianson (2011) have also the works of Krzanowski. Chang and Afifi (1974) derived a method for one dichotomous and more than one continuous variable. Aitchson and Aitken (1976) suggested a method based on the kernel density estimation. Anderson (1972, 1975) suggested the use of logistic discrimination. Vlachonikolis (1990) suggested using predictive rule instead of estimative rule for such situation. Ognain and Krzanowski (2001) compared some procedure for Binary variable discriminant including the LDA. Feldesman

(2002) suggested using Classification Tree (Breiman, et el, 1984; Venebles and Ripley, 1997, 1999) in place of the Linear Discriminant Analysis (LDA) since most users of LDA often do not consider the statistical limitations and assumptions (including data transformation) required for its usage. Many researchers have studied the Location model and have been making some modifications though not strictly for classification. Daudin and Bar-Hen (1999) applied the generalized Mahalanobis distance between populations based on the Kullback-Leibler divergence on the location model. Leing (2005) proposed a regularized classifier for the location model when there is a heteroscedastic dispersion across all location. Oyeyemi *et al* (2016) compared the performance of the Location model with the Fisher's linear discriminant analysis using only real data when continuous variables are more than binary variables.

2. Materials and methods

Fisher (1936) first introduced linear discriminant analysis for two classes and the idea was to transform the multivariate observations **x** to univariate observations y, such that the y's derived from the two classes were separated as much as possible. Suppose that we have a set of *m p*-dimensional samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m$ (where $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$) belonging to two different groups, G_1 *and* G_2. If $G_i (i = 1,2)$ has p-dimensional normal distribution with mean vector $\boldsymbol{\mu}_i (i = 1,2)$ and common covariance matrix $\sum$ which are known, the Linear Discriminant Function (LDF) is constructed as (Anderson, 2003)

$$W = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \left\{ \boldsymbol{y} - \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right\} \quad (1)$$

$\boldsymbol{y}$ is the vector of a future observation, $\boldsymbol{\mu}_i$ (i=1,2) is the vector of group means and $\boldsymbol{\Sigma}$ is the covariance matrix. Assuming equal probability of group membership, $\boldsymbol{y}$ may be assigned to G_1 *if*

$$W = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \left\{ \boldsymbol{y} - \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right\} \geq log \left(\frac{p_2}{p_1} \right) \quad (2)$$

Otherwise, we assign $\boldsymbol{y}$ to G_2. (Anderson, 1958)

2.1 The Location Model (LM)

This is a likelihood ratio method proposed by Krzanowski (1975). Given a vector of observations, $\boldsymbol{w} = (\boldsymbol{y}, \boldsymbol{x})$, $\boldsymbol{y}$ is a *p*-component vector of continuous variables and $\boldsymbol{x}$ is a *q*-component vector of binary variables. The q binary variables may be expressed as a multinomial outcomes $z = (z_1, z_2, \cdots, z_k)$ where $k = 2^q$. Each distinct pattern of $\boldsymbol{x}$ uniquely defines a multinomial cell, that is, $\boldsymbol{x} = (x_1, x_2, \cdots, x_k)$ falls in cell *m* of the multinomial cells. This method assumes that $\boldsymbol{y}$ has a multivariate normal distribution with mean $\mu_{(im)}$ in cell *m* (*m*=1,2,…,k) of population (or group) G_i(*i*=1,2) and common dispersion matrix $\sum$ in

all the cells. The probability of an observation falling in cell *m* of population G_i is p_{im}. For the purpose of allocating future observations, estimates of the parameters as given by Krzanowski (1975) are as follows:

Estimate of the mean for cell *m* of population G_i is given as

$$\hat{\boldsymbol{\mu}}_{im} = \overline{\boldsymbol{y}}_{im} = \left(\frac{1}{n_{im}}\right)\sum_{j=1}^{n_{im}} \boldsymbol{y}_{jim} \qquad (3)$$

The probability of an observation falling in cell *m* of population G_i is given as

$$\hat{p}_{im} = \frac{n_{im}}{n_i} \qquad (4)$$

Let

$$\boldsymbol{y} = (y_{i1}, y_{i2}, \cdots, y_{ip}),\ \ \boldsymbol{v} = (1, x_{i1}, x_{i2}, \cdots, x_{iq})$$

$$\boldsymbol{C}_i = \sum_{i=1}^{n_i} \boldsymbol{y}_{ij}\boldsymbol{v}'_{ij}\ ,\ \ and\ \ \boldsymbol{A}_i = \sum_{i=1}^{n_i} \boldsymbol{v}_{ij}\boldsymbol{v}'_{ij}$$

Then

$$\widehat{\boldsymbol{B}}_i = \boldsymbol{C}_i\boldsymbol{A}_i^{-1} \qquad (5)$$

(5) is as defined by Anderson (1958) in Krzanowski (1975).

And the estimate of the common covariance matrix for all cells in both populations is

$$\boldsymbol{\Sigma} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\boldsymbol{y}_i\boldsymbol{y}_i' - \widehat{\boldsymbol{B}}_i\boldsymbol{A}_i\widehat{\boldsymbol{B}}_i')}{(n_1+n_2-2k)} \qquad (6)$$

The allocation rule is to allocate to G_1 if

$$(\hat{\boldsymbol{\mu}}_{1m} - \hat{\boldsymbol{\mu}}_{2m})'\boldsymbol{\Sigma}^{-1}\left\{\boldsymbol{y} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_{1m} + \hat{\boldsymbol{\mu}}_{2m})\right\} \geq log\left(\frac{\hat{p}_{2m}}{\hat{p}_{1m}}\right) \qquad (7)$$

otherwise allocate to group G_2

A close look at (6) clearly shows that the expression $\widehat{\boldsymbol{B}}_i\boldsymbol{A}_i\widehat{\boldsymbol{B}}_i'$ can be traced as being computed using $\boldsymbol{v} = (1, x_{i1}, x_{i2}, \cdots, x_{iq})$ which is a vector of arbitrary scores assigned to each state of the categorical variables, the same distortion done by the conventional LDA procedure in assuming that and treating all variables, including the categorical as been continuous. The probabilities of misclassification $P(1/2)$ and $P(2/1)$ for G_1 and G_2 can be obtained from (7), conditional on the observation falling in multinomial cell m_i and noting that the overall probability of misclassification from $G_i(i = 1,2)$ is the sum of the probabilities of misclassification for each multinomial cell of $G_i(i = 1,2)$ weighted by the probability of its occurrence, we have

$$P(2/1) = \sum_{m=1}^{k} p_{1m}\Phi\left\{\frac{log\left(\frac{p_{2m}}{p_{1m}}\right) - \frac{1}{2}D_m^2}{D_m}\right\} \qquad (8)$$

and

$$P(1/2) = \sum_{m=1}^{k} p_{2m}\Phi\left\{\frac{log\left(\frac{p_{1m}}{p_{2m}}\right) - \frac{1}{2}D_m^2}{D_m}\right\} \qquad (9)$$

Where $\Phi(.)$ is the cummulative density function of standard normal distribution and $D_m^2 = (\boldsymbol{\mu}_{1m} - \boldsymbol{\mu}_{2m})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{1m} - \boldsymbol{\mu}_{2m})$ is the mahalanobis squared distance between G_1 and $G_2$with respect to the multinomial cell m_i. Estimates of $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}$ and the probabilities of misclassification are obtained, as most times, these parameters are unknown.

2.2.The Modified Location Model (MLM)

Suppose that we have available a set of n_1 individuals known to have come from a first

population (or group) G_1, and a set of n_2 individuals also known to have come from a second population (or group), G_2. We wish to set up a discriminating rule for, G_1 and G_2 based on a vector $\boldsymbol{x}$ of *q*-binary variables and a vector $\boldsymbol{y}$ of *p*-continuous variables for an observation vector $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{y})$. As already existing in literature, the binary variables can form a contingency classification by treating the binary variables as multinomial with $k = 2^q$ classes (Krzanowski, 1975). Thus, an incidence table with *m*-cells can arise from such construction for each of the population. The probability of any given observation falling into one of the *m* (*m=1,2,..,k*) cells could be estimated using the proportion of the observations falling into each cell of the incidence table. Let the estimate of the probability that an observation falls in cell m of population G_i be denoted by p_{im} (i=1,2 *m*=1, 2,…,*k*). We assume that the conditional distribution of $\boldsymbol{y}$, given each unique pattern of $\boldsymbol{x}$ in cell *m* of the multinomial is $N(\mu_{im}, \boldsymbol{\Sigma})$ for population G_i.

In classifying an observation, $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{y})$ to one of two populations, G_1 & G_2, we may generalize by assuming that $\boldsymbol{y}$ has a multivariate normal distribution with mean μ_{im} in cell *m* and population $G_i (i = 1,2;\ m = 1,2, \dots, k)$ and a common variance-covariance matrix $\boldsymbol{\Sigma}$ in all cells of both populations.

$$(\boldsymbol{y} / \boldsymbol{x}) \sim N_p(\mu_{im}, \boldsymbol{\Sigma}) \quad (10)$$

For estimating the parameters needed for allocation, we have

$$\hat{\mu}_{im} = \bar{\boldsymbol{y}}_{im} = \left(\frac{1}{n_{im}}\right) \sum_{j=1}^{n_{im}} \boldsymbol{y}_{jim} \quad (11)$$

$$\hat{p}_{im} = \frac{n_{im}}{n_i} \quad (12)$$

$$\boldsymbol{\Sigma} = \frac{\sum_{i=1}^{2} \sum_{m=1}^{k} \sum_{j=1}^{n_{im}} (\boldsymbol{y}_{jim} - \bar{\boldsymbol{y}}_{im})' (\boldsymbol{y}_{jim} - \bar{\boldsymbol{y}}_{im})}{n_1 + n_2 - 2k} \quad (13)$$

While retaining the allocation rule given in (7), the modification made on the location model is given by (13) as compared with (6). The modification (13) stems from the fact that we can possibly compute a separate *m*-covariance matrices corresponding to each cell of the multinomial using the continuous values in that cell, and then pool these covariance matrices to represent the common covariance matrix for that data set (Sorum, 1973). In this way, the inclusion and usage of the categorical variable by the LM when computing the covariance has been avoided.

3. Results and discussion

For the purpose of applying the modification, *p*-continuous data was generated from multivariate normal $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ and *q*-binary data generated from Uniform $U(0,1)$ distributions respectively using the R-statistical package for various sample sizes.

We considered cases of, equal number of continuous and categorical variables ($p=q$), continuous variables are more than the categorical ($p> q$), and continuous variables are less than the categorical variables ($p <q$). Accuracy rate was obtained using the re-substitution method of error rate estimation. The re-substitution method uses the original (generated in this case) data to obtain the discriminant function and then use the discriminant function obtained to re-classify the original data

Table 1: Accuracy Rates for *p=4, q=3 and p=3, q=4*

	p=4, q=3		*p=3, q=4*	
Sample Size	MLM	LM	MLM	LM
20	.**	**	**	**
30	0.7500	0.5500	**	**
50	0.8100	0.5000	0.6400	0.4200
100	0.8100	0.4900	0.7600	0.5200
150	0.8433	0.5167	0.6800	0.5033
200	0.8100	0.5050	0.7575	0.4650
250	0.8320	0.5060	0.6700	0.4420
500	0.8070	0.4980	0.7560	0.527
1000	0.8295	0.5133	0.7010	0.4720
1500	0.8190	0.5133	0.7333	0.5157
2000	0.8355	0.5148	0.7390	0.5075

** analysis was not done due to zero/empty cell

The result of analysis for *p=4, q=3* and *p=3, q=4* are obtained as shown in table1 above. Three binary variables having eight multinomial cells resulted to possibilities of one or more cells having no observation, referred to as zero/empty cells. This was the case for small sample of size n=20, hence analysis was not possible for that situation. The MLM performed better than the LM. For the *p=3, q=4* case, four binary variables having sixteen cells would result to a case of empty/zero cells in which the probability of group membership for that cell cannot be estimated, for samples as small as 20 and 30, hence analysis could not be carried out. Again, the MLM had a better performance than the LM but not as in the reverse case of *p=4, q=3*.

Table 2: Accuracy Rates for *p=2, q=4 and p=4, q=2*

	p=2, q=4		*p=4, q=2*	
Sample Size	MLM	LM	MLM	LM
20	.**	**	0.7750	0.4750
30	**	**	0.7333	0.5500
50	0.6200	0.5700	0.8400	0.5500
100	0.7150	0.6300	0.8500	0.5250
150	0.6100	0.4867	0.7900	0.5300
200	0.6475	0.5675	0.8025	0.5225
250	0.7320	0.5200	0.8000	0.5240
500	0.7010	0.4950	0.7790	0.5190
1000	0.6940	0.5515	0.7950	0.5310
1500	0.7113	0.4993	0.7873	0.5270
2000	0.7023	0.5355	0.8003	0.5308

** analysis was not done due to zero/empty cell

Table 3: Accuracy Rates for *p=3, q=3 and p=4, q=4*

	p=3, q=3		*p=4, q=4*	
Sample Size	MLM	LM	MLM	LM
20	0.6500	0.6500	**	**
30	0.6000	0.4667	**	0.600
50	0.6600	0.4800	0.8000	0.5400
100	0.7300	0.4950	0.7500	0.4050
150	0.7267	0.4933	0.8267	0.5133
200	0.7250	0.4975	0.7650	0.4800
250	0.7200	0.5000	0.8680	0.4800
500	0.7170	0.5010	0.8060	0.5270
1000	0.7510	0.5190	0.7650	0.4375
1500	0.7320	0.5143	0.7650	0.5027
2000	0.7400	0.5125	0.8375	0.5400

The result of analysis for *p=2, q=4* and *p=4, q=2* are obtained as shown in Table 2 above. With the number of binary variable twice that of the continuous, the LM still had the least accuracy rate over all sample sizes. With *p=4, q=2*, the continuous variables now twice the number of the binary, the MLM produced result that far outperform that of LM. The result of analysis for *p=3, q=3* and *p=4, q=4* are obtained as shown in Table 3. With equal number of continuous and binary variables, the MLM still performed better than the LM. However, both had performance that improved as the sample sizes increases. For *p=4, q=4*, the MLM and LM could not produce result for n=20 due to possibilities of empty/zero cell resulting from the sixteen cells obtainable with four binary variables, however, the MLM still outperformed the LM.

4. Conclusion

Our objective is to consider a modification or procedure void of any distortion resulting from treating categorical data as if they are continuous, and this resulted to the Modified Location Model. The performance of these procedures was studied and analyzed using generated data over various sample sizes and combinations of continuous and binary variables. The MLM gave better accuracy than the LM over all cases considered. However, we noted that the LM produce accuracy rate that does alternates around average despite changes in sample size and variable combinations. The MLM tends to produce higher accuracy when we have more of continuous variables relative to binary variables.

References

Aitchson, J. and Aitken, C. G. G. (1976). Multivariate binary discrimination by the Kernel method. *Biometrika 63*(3): 413-420

Anderson, J. A. (1972). Separate Sample Logistic Discrimination. *Biometrika*, 59, 19-35.

Anderson, J. A. (1975). Quadratic Logistic Discrimination, *Biometrika*, **62**, 149-54

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Third Edition. John Wiley & Sons, Inc. Hoboken, New Jersey.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* New York, Chapman and Hall

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.

Cochran, W. G. and Hopkins, C. E. (1961). Some Classification Problems with Multivariate Qualitative Data. *Biometrics, 17*(1), 10-32.

Daudin, J. J. and Bar-Hen, A. (1999). Selection in Discriminant Analysis with Continuous and Discrete Variables. *Computational Statistics and Data Analysis, 32*, 161 – 175

De-Leoan, A. R., Soo, A., and Willianson, T. (2011). Classification with discrete and continuous Variables via general mixed-data models. *Journal of Applied Statistics, 5*(38), 1021 – 1032

Feldesman, M. R. (2002). Classification Tree as an Alternative to Linear Discriminant Analysis. *American Journal of Physical Anthropology, 119*, 257 – 275

Fisher, R. A. (1936). The Use of multiple measurements in taxonomic problems. *Ann. Eugen., 7*, 179-88.

Glick, N. (1973). Sample-based multinomial classification. *Biometrics, 29*, 241-256.

Huberty, C. J., Wisenbaker, J. M., Smith, J. D. and Smith, J. C. (2010). Using Categorical Variables in Discriminant Analysis. *Multivariate Behavioural Research, 21*(4), 479 – 496

Krzanowski, W. J. (1975). Discrimination and Classification Using Both Binary and Continuous Variables. *Journal of the American Statistical Association, 70*, 782-790.

Krzanowski, W. J. (1980). Mixtures of Continuous and Categorical Variables in Discriminant Analysis. *Biometrics, 36*, 493-499.

Krzanowski, W. J. (1982). Mixtures of Continuous and Categorical Variables in Discriminant Analysis; A hypothesis testing approach. *Biometrics 38*, 991-1002.

Lee, S. Y., Song, X. Y., and Lu, B. (2008) Discriminant Analysis using Mixed Continuous, Dichotomous and Ordered Categorical Variables. *Multivariate Behavioural Research, 4*(42), 631 – 645

Leing, C. Y. (2005). Regularized Classification for mixed Continuous and Categorical Variables under Cross-location heteroscedasticity. *Journal of Multivariate Analysis, 93*, 358 – 374

Liang, Z., Li, Y. and Shi, P. (2008). A note on two-dimensional linear discriminant Analysis. *Pattern Recognition Letters, 29*, 2122 – 2128.

Ognain, K. A. and Krzanowski, W. J. (2001). A Comparison of Discriminant procedure for Binary Variables. *Computational Statistics and Data Analysis, 38*, 139 – 160

Okin, I. and Tate, R. F. (1961). Multivariate Correlation Models with Mixed Discrete and Continuous variables. *Annals of Mathematics Statistics, 32*, 448-465

Oyeyemi, G. M., Mbaeyi, G. C., Salawu, S. I. and Muse, B.(2016). On discrimination procedure with mixtures of continuous and categorical variables. *Journal of Applied Statistics*.http://dx.doi.org/10.1080/02664763.2015.1125859

Sorum, M. (1973). Estimating the expected Probability of Misclassification for a Rule based on the Linear Discriminant Function: Univariate Normal Case. *Technometrics, 2*(15), 329 – 339

Venebles, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-PLUS*. Second Edition, New York, Springer

Venebles, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*. Third Edition, New York, Springer

Vlachonikolis, I. G. (1990) Predictive Discrimination and Classification with Mixed Binary and Continuous Variables. *Biometrika*, *77*(3), 657-662.