SUPERVISE DIFFERENCE LINEAR CLASSIFICATION TECHNIQUE FOR TWO GROUP'S PROBLEM

F. Z. Okwonu
Department of Mathematics and Computer Science, Delta State University, Abraka, Delta State, Nigeria. E-mail: fzokwonu_delsu@yahoo.com

ABSTRACT
Classification techniques can be categorized as parametric and nonparametric. The former strictly depends on assumptions, and is a supervised linear classification procedure whereas the later does not rely on any assumptions and is an unsupervised linear classification technique. This paper considers supervised linear classification techniques in which the course of dimensionality is negated. In this regard, the sample size is greater than the sample dimension. We investigated the performance of the Fisher, and the difference linear classification techniques based on the classification performance and the acceptance or rejection of the null or alternative hypothesis in which the mean of the optimal probability of correct classification is used as the hypothesized mean, and the computed mean probability is derived from the mean probability of correct classification based on 1000 replications. The robustness of these techniques is determined by the sensitivity of these methods to contaminated data set. Relying on the performance analysis, we further investigated the acceptance or rejection of the null or alternate hypothesis based on the proportion of contamination using the Hoteling test statistics. The comparative analysis indicates that the difference linear classification rule outperformed the conventional Fisher's technique. The analysis revealed that using 95% level of significant, the null hypothesis is rejected.

Key words: Mean probability, sample size, hypothesized mean, robust.

INTRODUCTION
Classification is a statistical tool that is applied to model, and also to predict the characteristics of interest. This procedure helps to predict precisely the characteristics of interest to the exact groups. It is a decision based technique that is widely applied to scientific fields. Generally, classification techniques based on supervised learning is developed using assignment rule, which implies that an observation maybe correctly assigned to the correct groups or otherwise.

The model based linear classification rules suffer the curse of dimensionality in which the dimension of the data set is greater than the sample size. Detailing this aspect, the covariance matrices tend to be singular as such the coefficient cannot be computed. High dimensional data set are often envisaged in application area such as micro arrays, genomics and mass spectrometry (Bouveyron, 2013).

In specific term, virtually not all the predictor variables are useful or contribute meaningfully to the required objective(s). Hence technique such as the principal component analysis and factor analysis are often applied to reduce the data dimension. In other to perform the conventional classification task using data set with high dimension, dimension reduction technique is often applied as initial step before the classical classification coefficient can be obtained. Thus, dimension reduction techniques imply that information loss is inevitable. It may be on the contrary that the reduced dimension or deleted predictor variable may contribute meaningfully to decision making. Several other classification techniques such as nearest mean classifier; variable selection and subspace classification procedure have been proposed to handle high dimensional data set.

Classification technique can be considered as classical or robust. Developing linear classification model based on parametric procedure strictly relies on the model assumptions, say normality of the data set and homoscedasticity of the covariance matrices (Rencher, 2002). In this respect, the nature of the training data set used in building the model is of paramount importance.

The fundamental objective of the classification techniques being supervised or unsupervised is to correctly predict group membership, with the sole aim of minimizing misclassification error. But with the assumptions violation, the classical techniques often misclassify group membership maximally. Though, the linear classification approach based on the Fisher's technique strictly depends on the above assumptions. On the contrary, the quadratic classification approach only depends on normality of the data set to predict accurately (Bouveyron, 2013, Johnson and Wichern 2007).

In 2014, Okwonu and Othman (2014) investigated the effect of unequal variance covariance matrices and mixture of contaminated normal data set for low dimensional data set. In that paper, as the contamination proportion increases, the rate of misclassification increases for the Fisher's technique for small and medium sample sizes. Though for large sample size and relying on the central limit theorem, the rate of misclassification is reduced. Thus, the study also revealed that linear classification technique can be used to predict accurately if the normality assumption is violated (Okwonu and Othman, 2013).

This paper investigates the classification performance, and the rejection/acceptance of the null/alternate hypothesis based on the percentage of contaminated normal data set applied to predict group membership using the Fisher's and the difference linear classification procedures. The difference linear classification technique was proposed as comparison to the conventional Fisher's approach. The concept of applying hypothesis to determine the acceptance or rejection of classification performance for these techniques are obviously novel but though not new concept in the field of statistics, but considered as infusion of hypothesis testing procedures to determine accuracy of group membership.

The rest of this paper is organized as follows. Section Two describes the Fisher linear classification analysis. Section Three contains the proposed difference linear classification technique. Simulation is contained in Section Four.
Finally, Section Five offers the concluding remarks.

Fisher linear classification analysis (FLCA)

The importance of Fisher linear classification rule (Fisher, 1936) strictly depends on the accuracy of prediction if the assumptions are satisfied, easy for computation and interpretation. This technique is obtained by deriving a set of variable otherwise called the linear classification coefficient which is applied to predict group membership accurately.

The linear classification score is computed by post multiplying the coefficient with the sample observations. This technique is similar in a way to the regression equation in which the predictor variables are post multiplied by the coefficient to yield the linear classification score. The depen-dent variable is generally the classification score whereas the predictor variable is the independent variable. In general, the study assumed that the cutoff point defines a boundary which strictly determines the allocation of an observation to the respective groups.

Based on this concept, the classification score is compared with the cutoff point. To be specific, if the classification score is greater than the cutoff point, this implies that the observation is assigned to group one otherwise if the classification score is less than the cutoff point, the observation is assigned to group two, respectively. This analysis strictly defines the Fisher linear classification analysis which is based on two fundamental assumptions. Generally, the Fisher's technique is a dimension reduction procedure (Rencher, 2002).

Difference linear classification rule (DLCR)

This procedure relies on the selection of a simple random sample for the two groups. The data set for each group is chosen one after the other. The order of the data set is randomly reassigned to the respective groups. The unique mean vector is computed by obtaining the difference between the sample observations for the respective groups, for example, $d_j = g_1 - g_2$. Where $g_i, i = 1, 2,$ denotes the respective groups. The difference d_j is summed in order to compute the unique mean $\bar{d}$. The sample covariance matrix for this technique is computed based on the difference between the sample observation and $\bar{d}$, for example,

$$S_i = \frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{d})(x_{ij}-\bar{d})'}{n_i - 1}, i = 1,2,\ j = 1,2,3,.....,n_i. \quad (1)$$

Where n_i is the sample size for each group and x_{ij} denote the sample observations for each group. The computation of the coefficient requires that the common covariance matrices are combined. Relying on this, the coefficient hdf and the classification score kvb are stated as follows;

$$hdf = \bar{d}S_c^{-1},$$
$$kvb = hdfx' + \bar{x}/8,$$
$$dgp = \bar{d}(8hdf')^{-1},$$

where S_c is the combined common covariance matrix, and dgp denote the cutoff point. The assignment of an observation to group one depends on $kvb \geq dgp$, otherwise the observation is assigned to the second group if the following equation is satisfied, $kvb < dgp$.

Simulation

This section reveals the performance of the Fisher and the difference linear classification techniques via Monte Carlo simulation. The data set is generated based on the contaminated normal model in which by convention, large proportion of the data set come from the normal distribution while the other fraction is generated from the contaminated normal model. The contaminated normal data set is generated with different mean vectors and large variances respectively.

The simulation is designed such that the data set is divided into two groups, say training (60%) and validation (40%) samples. This implies that the training data is different from the validation data, respectively. This process is devoid of upward biased. The comparative classification performance of these techniques is based on different sample sizes; say small, medium, large and comparable dimensions. The performance of these techniques is measured based on the mean probability of correct classification compared to the mean of the optimal probability of correct classification computed from the uncontaminated data set.

The mean of the optimal probability of correct classification is used as the performance benchmark. The result reported is based on 1000 replications. Figureure.1 showed the performance of these techniques based on small sample size; $n_1 = n_2 = 20$. The analysis revealed that DLCR outperformed the FLCA method. The analysis indicates that DLCR has reduced misclassification error rate compared to the FLCA.

From Figure. 1 we observed that both techniques did not attain the mean of the optimal

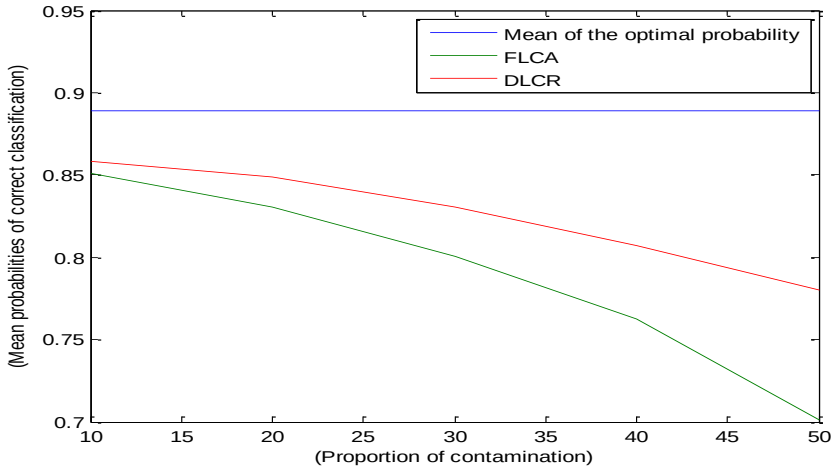

Figure. 1. Effect of contamination on classification performance.

probability; however, the difference method (DLCR) is robust over the FLCA. Figure. 2 revealed that DLCR approach outperformed the conventional Fisher's technique for medium

sample size; $n_1 = n_2 = 30$. In Figure. 3, for large sample size; $n_1 = n_2 = 50$, both techniques performed comparable. From the earlier mentioned analyses the study observed that as the proportion of contamination increases, the rate of misclassification increases. Based on the earlier mentioned performance analysis, the study investigates if the computed mean μ_c is equal to the hypothesized mean μ_h or otherwise for the different techniques investigated. The

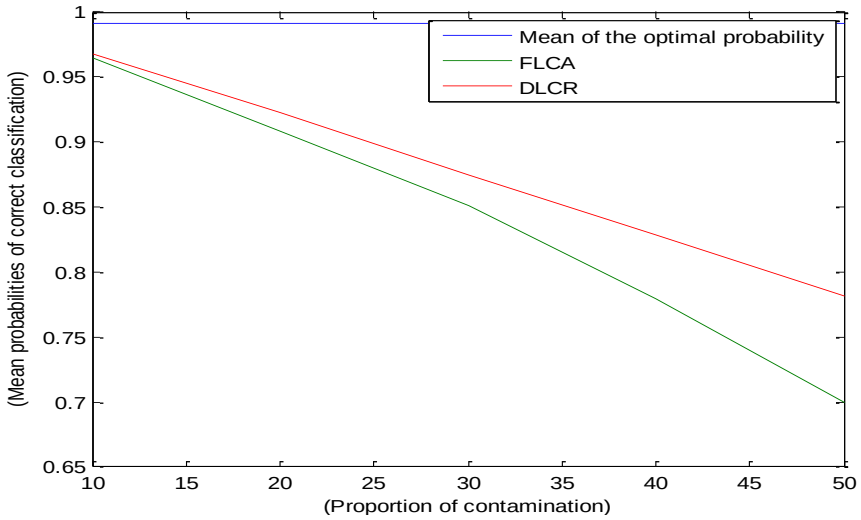

Figure. 2. Effect of contamination on classification performance.

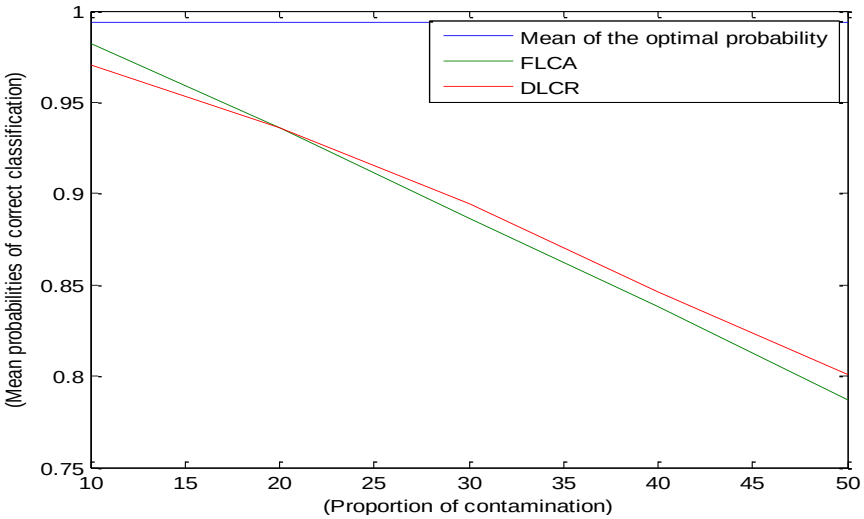

Figure. 3. Effect of contamination on classification performance.

test hypothesis is based on

$$H_0 : \mu_c = \mu_h$$
$$H_1 : \mu_c \neq \mu_h$$

At the α level of significant, the null hypothesis H_0 is rejected in favor of the alternate hypothesis H_1 if

$$T^2 = n(\mu_c - \mu_h)'\mathbf{S}^{-1}(\mu_c - \mu_h) > \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha),$$

and the alternate hypothesis is accepted if $T^2 \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$. The T^2 (Hoteling T^2) is distributed as $\frac{(n-1)p}{(n-p)} F_{p,n-p}$, where $F_{p,n-p}$ denotes a random variable with an F distribution with

p and $n-p$ degree of freedom at 95% level of significant.

The hypothesis testing revealed that the null hypothesis is rejected for the proportion of contamination considered. This implies that the computed mean differs from the hypothesized mean; hence the alternative hypothesis is accepted. The implication of this is that the mean vector of the difference linear classification technique is closer to the hypothesized mean vector than the mean vector of the Fisher's approach. Figure. 5 below revealed the rejection of the null hypothesis for the two procedures considered.

The hypothesis testing based on these techniques indicates that as the proportion of contamination increases, both techniques revealed the rejection of the null hypothesis. In general, as the proportion of contamination increases the mean difference between the hypothesized mean vector and the computed mean vectors differs. In practical terms, supposed that the null hypothesis was accepted, classification based on these techniques would have been infeasible since the coefficient would be impractical to compute. Figure. 1 through Figure. 3 corresponds to Figure. 4 through Figure. 6, respectively.

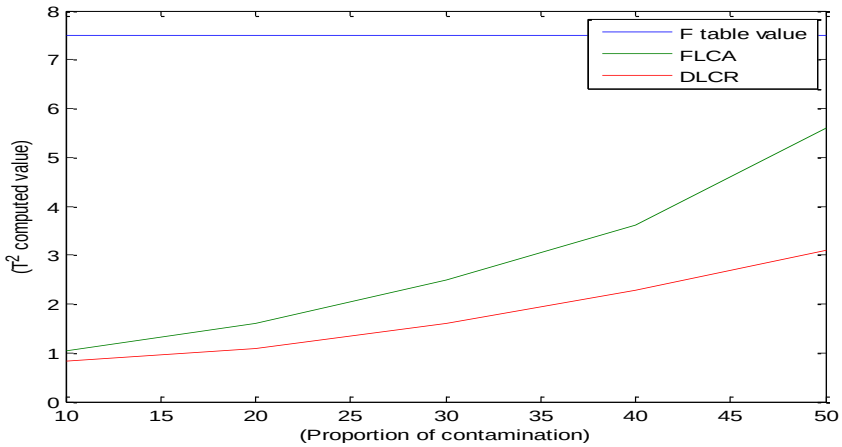

Figure. 4. Hypothesis testing based on hoteling T^2 for small sample size.

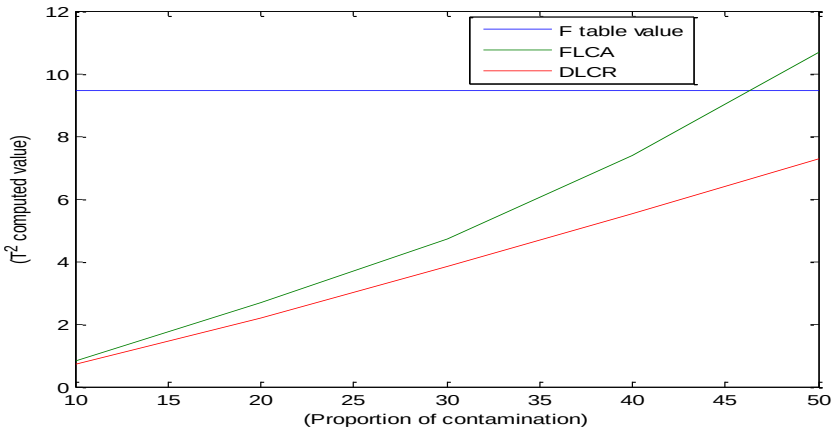

Figure. 5. Hypothesis testing based on Hotelling T^2 for medium sample size.

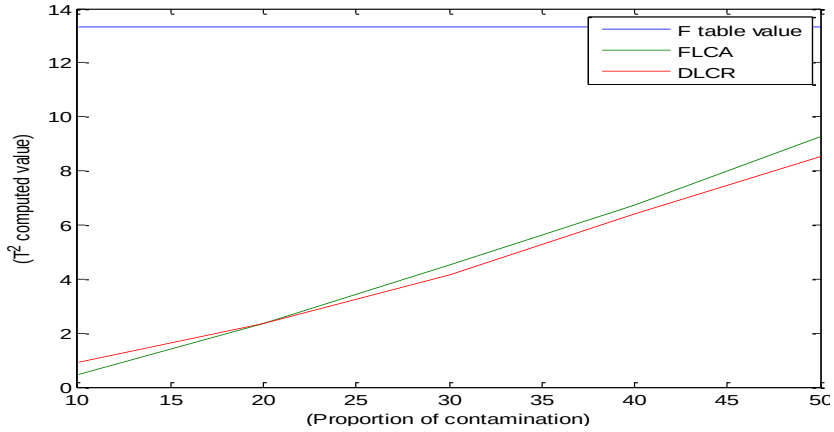

Figure. 6. Hypothesis testing based on Hotelling T^2 for large sample size.

Conclusion

Conventionally, the Fisher linear classification analysis is applied to study the separation and classification between observations or objects. For several decades, different classification techniques have been proposed with respect to robustness in which the influential observations are modeled or deleted. This paper focused on the comparative performance between the Fisher's approach and the proposed difference linear classification rule. The difference classification rule relies on the difference between the sample observations for the two groups to compute its coefficient. The comparative classification performance revealed that the difference linear classification rule performed better than the Fisher's approach as the proportion of contamination increases. In each case, both techniques are unable to attain the performance benchmark. The Monte Carlo simulations revealed that the mean probabilities of correct classification for the difference linear classification rule is closer to the performance benchmark for all cases considered. The hypothesis testing revealed the rejection of the null hypothesis in favor of the alternate hypothesis. The hypothesis testing revealed that as the proportion of contamination increases, the mean difference between the hypothesized mean and the computed mean differ at 95% level of significance, which implies acceptance of the alternate hypothesis.

Conflict of interests

The author has not declared any conflict of interests.

REFERENCES

Bouveyron, C., (2013). Probabilistic model-based discriminant analysis and clustering methods in chemometrics. *Journal of Chemometrics In* press, Pp. 1-21.

Fisher, R. A., (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Pp.179 - 188.

Okwonu, F. Z. and Othman, A. R., (2013). Heteroscedastic variance covariance matrices for unbiased two groups linear classification methods. *Applied Mathematical Sciences*, Pp.6855-6865.

Okwonu, F. Z. and Othman, A. R. (2014). Effect of heteroscedastic variance covariance matrices on two groups linear classification techniques. *Journal of Mathematics and System Science*, Pp.133-138.

Rencher, A. C., (2002). *A methods of multivariate analysis*. A John Wiley & Sons, Inc.

Johnson, R. A. and Wichern, D. W., (2007). *Applied multivariate statistical analysis*. Pearson Prentice Hall, Upper Saddle River.