International Journal of Mathematical Analysis and Applications
2014; 1(2): 27-30
Published online May 30, 2014 (http://www.aascit.org/journal/ijmaa)

Keywords
Classification,
Mean Probability,
Unbiased,
Robust,
Admissible

Received: April 14, 2014
Revised: May 04, 2014
Accepted: May 05, 2014

Supervised learning techniques based on Fisher and Filter linear classification procedures for two groups problem

Friday Zinzendoff Okwonu

Department of Mathematics and Computer Science, Delta State University, Abraka, Nigeria

Email address
fzokwonu_delsu@yahoo.com

Citation
Friday Zinzendoff Okwonu. Supervised Learning Techniques Based on Fisher and Filter Linear Classification Procedures for Two Groups Problem. *International Journal of Mathematical Analysis and Applications.* Vol. 1, No. 2, 2014, pp. 27-30.

Abstract
The conventional Fisher linear discriminant analysis was proposed to investigate separation between two groups of object. This procedure performs optimally if the data set for the two groups is normally distributed and the variance covariance matrices are homoscedastic. When these assumptions are violated, the Fisher's technique underperforms. To remedy this deficiency, a supervised learning technique based on the Filter linear classification technique is proposed. The comparative classification performance of these techniques is investigated via Monte Carlo Simulation using data set generated from contaminated normal model. The classification results based on the mean of the optimal probability of correct classification indicate that both procedures are unbiased. Although, the analyses revealed that the Filter technique is robust and admissible over the Fisher's method.

1. Introduction

The Fisher linear discriminant analysis (FLDA) [1] was introduced when it was applied to study the Iris data set for two groups. The coefficient of the Fisher's technique can be computed if the sample size is greater than the sample size dimension. This procedure assist in gaining information regarding the separation between the two groups with regards to the within group centroid and the contribution of the profile variables[2, 3]. It is a dimension reduction technique and belongs to the class of supervised learning technique[4]. The basic assumptions of the FLCA are homoscedasticity of the covariance matrices and normality of the data set. From now on, the term *classification* is used in place of *discrimination* since the focus of this paper is classification. The FLCA technique performs optimally if the above assumptions are satisfied. Conventionally, the FLCA procedure was proposed for two groups. However, it has been generalized to more than two groups. During the mid 1960's researchers opined that separation and estimation be included as part of the objective of the FLCA [5]. It is the suggestion of this author to state that without separation the coefficient of FLCA is infeasible.

The coefficient of the FLCA is computed based on the difference between the within group mean vectors and pooled covariance matrix. These sample statistics are the building blocks of most classical multivariate techniques including the FLCA but are sensitive to influential observations [6-14]. The sample mean vectors and covariance matrices computed based on data set generated from a multivariate

normal distribution enhances the performance of the FLCA maximally [15, 16]. On the other hand, if the data set are not drawn from a multivariate normal distribution, the sample statistics computed are influenced by influential observations hence when these sample statistics are applied to develop the FLCA, the misclassification rate for the FLCA tends to increase maximally[17]. Due to the aforementioned, this paper focused on supervised learning technique to filter the influential observations and hence enhancing robust performance when the data set are not normally distributed. The sample statistics of this technique is computed based on the filtered sample observations derived by comparing the values of the Mahalanobis distance with a fixed constant. This method is called the filter linear classification rule (MYROB) and it is a dimension reduction technique that encompasses filtered sample data[18]. This technique like the Fisher's approach can be applied to numerous fields of study. Both the FLCA and MYROB techniques are supervised learning techniques.

This paper is organized as follows. The conventional Fisher linear classification analysis is described in Section two. The filter linear classification rule (MYROB) is contained in Section three. Simulation and conclusion is contain in Sections four and five respectively.

2. Fisher Linear Classification Analysis (FLCA)

The sample mean vectors and covariance matrices are computed from the training samples. These sample statistics are applied to learn the Fisher linear classification rule. Based on the information provided by the Fisher linear classification rule via the training or validation sample, the objective is to classify an observation as belonging to one of the two groups accurately. The Fisher linear classification analysis [1] for two groups problem is defined mathematically as follows,

$$c = \mathbf{q}^T \mathbf{x}, \tag{1}$$

where $\mathbf{q}$ denote the Fisher linear coefficient, $\mathbf{x}$ is the sample observation and c denote the Fisher's classification sore, a scalar. The following equation in comparison with the classification score allows an observation to be assigned to the correct group, say,

$$\overline{c} = \frac{\sum_{i=1}^{2} \overline{\mathbf{x}}_i}{2} \mathbf{q}^T. \tag{2}$$

Where $\overline{c}$ denote the midpoint and $\overline{\mathbf{x}}_i$ is the within group mean vectors. The computation of the Fisher linear coefficient is possible if the group means are unequal. This condition is vital to enable separation, classification and discrimination feasible. The Fisher's coefficient maximizes the "between" group variability relative to the "within" group variability [4].

The comparison between the classification score and the midpoint defines the linear classification rule. The Fisher linear classification rule is obtained by comparing the classification score with the classification midpoint. The allocation rule is based on Equations (1-2). An observation is assign to group one if the classification score is greater than or equal to the midpoint otherwise the observation is assign to group two if the classification score is less than the midpoint.

3. Filter Linear Classification Rule (MYROB)

This technique involves computing the classical estimates, filtering the data set, then computing the filtered mean vectors and covariance matrices using the weighted sample data set[18]. The filtered sample means and covariance matrices are applied to develop the filter linear classification method. This procedure is based on the following steps:

Step 1 requires the computation of the Mahalanobis distance and then comparing the values with a fixed constant. This process yields the weighted values w for the respective groups.

Step 2 the sample observations is transformed by pre-multiplying the weighted values by the sample observations. This process yields the filtered sample observations, say;

$$d_{ij} = w_i \mathbf{x}_{ij}, i = 1, 2, j = 1, 2, .., n_i. \tag{3}$$

Step 3 the filtered sample data set is used to compute the sample mean vectors, covariance matrices and pooled covariance matrix defined as follows,

$$\overline{\mathbf{x}}_i = \frac{\sum_{j=1}^{n_i} d_{ij}}{n_i}, (i = 1, 2), \tag{4}$$

where n_i is the sample size of the filtered sample observations and

$$\mathbf{S}_i = \frac{\sum_{i=1}^{n_i} (d_{ij} - \overline{\mathbf{x}}_i)(d_{ij} - \overline{\mathbf{x}}_i)^T}{(n_i - 1)},$$

$$\mathbf{S}_{pooled} = \frac{\sum_{i=1}^{2} (n_i - 1)\mathbf{S}_i}{\sum_{i=1}^{2} n_i - 2}.$$

Step 1 to Step 3 is applied to formulate the filter linear classification rule. The linear classification score for this approach is described mathematically as,

$$g = \frac{hd}{\mathbf{S}_{pooled}}\mathbf{x} = fdu'\mathbf{x}, hd = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2. \quad (5)$$

The midpoint is described mathematical as

$$\bar{g} = \frac{4hdk}{8} fdu', hdk = \bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2. \quad (6)$$

The classification rule for the MYROB[19] is formulated as follows;

$$g \geq \bar{g} \quad (7)$$

The implication of Equation (7) is that an observation in group one is correctly assigned otherwise is assigned to group two if the following equation is satisfied $g < \bar{g}$.

4. Simulation

The simulation is designed to investigate the effect of contamination on these supervised learning linear classification techniques. The Monte Carlo simulation is conducted for three sample sizes (small, medium and large). The objective of this section is to investigate the comparative performance of these techniques when the assumptions of the FLCA are violated. The mean of the optimal probability of correct classification is used as the performance benchmark to determine robustness as against the mean probability of correct classification obtained from each technique. The mean probability of correct classification is plotted against the proportion of contamination. The proportion of contamination is the percentage of contaminated normal data set introduced to contaminate the normal data set. The idea is to investigate the effect the fraction of contaminated normal data has on the performance of these techniques.

For each simulation, the sample size was divided into two categories, training (60%) and validation (40%), respectively. The data set was randomly generated based on the contaminated normal model. This model stipulates that large proportion of the data set is generated from the normal distribution and the other fraction is generated from the contaminated normal distribution, both proportions were added and randomly reshuffled using the uniform distribution. The mean probabilities of correct classification are based on 1000 replications. In this experiment, the proportion of contamination is set as follows: 10, 15, 20 and 25 respectively. The performances of these techniques are shown in the figures below.

Based on Figure 1, the misclassification rate for the FLCA was higher than that of the MYROB, this indicates that the MYROB technique is robust and admissible over the FLCA for the small sample size $(n_1 = n_2 = 30, p = 2)$, where p is the sample size dimension. The mean of the optimal probability for the small sample size is 0.9998. Though, the analysis revealed that as the proportion of contamination increases, the misclassification rate increases more for the FLCA than the MYROB.

The mean of the optimal probability of correct classification for the medium sample size $(n_1 = n_2 = 50, p = 3)$ is 0.9982. As shown in Figure 2, the misclassification rate for the MYROB technique is minimum compared to the FLCA technique. This also revealed that MYROB is robust and admissible over the Fisher's approach.

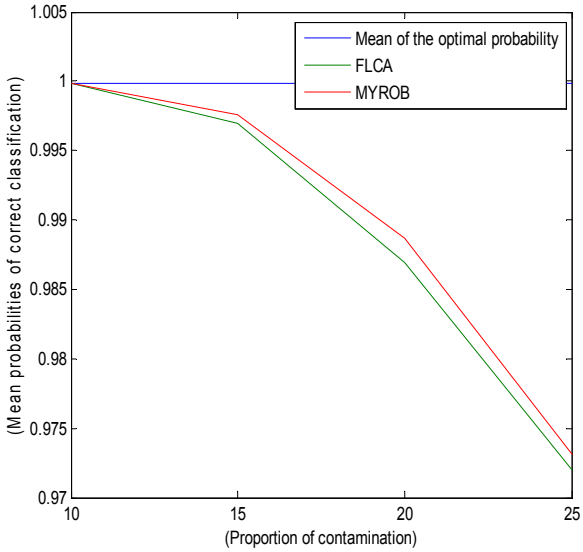

Figure 1. *Effect of proportion of contamination on mean probability of correct classification.*

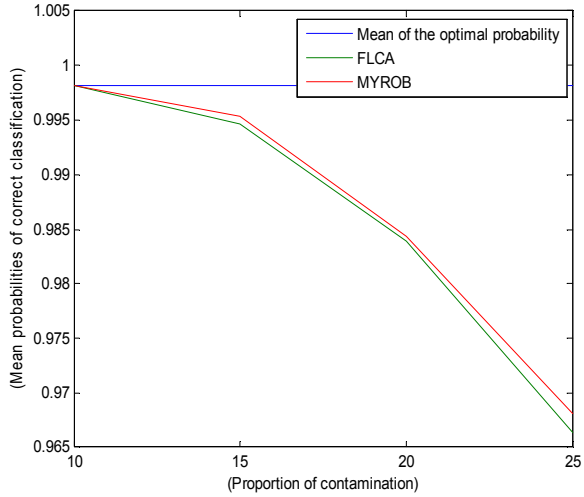

Figure 2. *Effect of proportion of contamination on mean probability of correct classification.*

The mean of the optimal probability of correct classification for large sample size $(n_1 = n_2 = 100, p = 5)$ is 0.9949. In this analysis, both techniques performed comparable up to 15% contamination but as the proportion of contamination increases, the MYROB outperformed the FLCA technique.

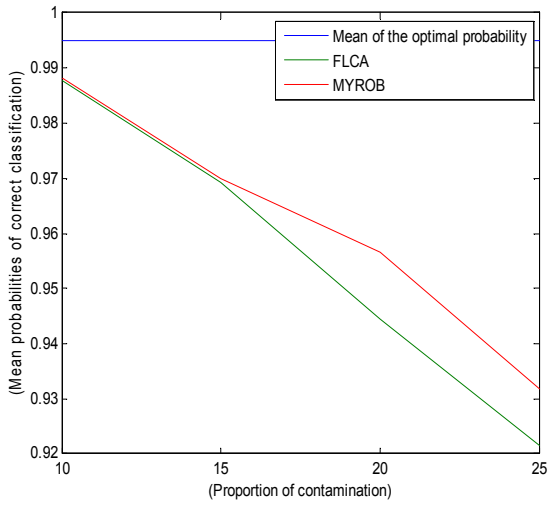

Figure 3. *Effect of proportion of contamination on mean probability of correct classification.*

5. Conclusion

The Monte Carlo simulation revealed that the MYROB procedure and the Fisher's technique are unbiased supervised linear classification techniques. The analyses showed that as the proportion of contamination increases, the Fisher's technique tends to misclassify more observations. On the other hand, the MYROB technique tends to reduce misclassification rate. The simulations also indicate that for small proportion of contamination both techniques tend to perform comparable. The analysis illustrate that the MYROB procedure performed better than the Fisher's procedure when the data set is generated from the contaminated normal model. The supervised learning technique based on the filter linear classification approach is robust over the Fisher's technique for this data set.

References

[1] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics.1936; 7:179 - 188.

[2] Kuhn M, Johnson K. Applied predictive modeling. Springer.2013.

[3] Johnson RA, Wichern DW. Applied multivariate statistical analysis. 6th ed. Pearson Prentice Hall: Upper Saddle River. 2007.

[4] Thiago GM. Reduced- rank discriminant analysis.http://tgmstat.wordpress.com/2013/12/12/reduced-rank-discriminant-analysis. 2013;1-3.

[5] Huberty CJ. Discriminant analysis. Review of Education Research. 1975; 45:543-598.

[6] Filzmoser P, Hron K. Outlier detection for compositional data using robust methods. Mathematical Geosciences. 2008; 40: 233-248.

[7] Basak I. Robust *M*-estimation in discriminant analysis. The Indian Journal of Statistics. 1998; 60:246-268.

[8] Devlin SJ, Gnanadesikan R, Kettenring JR. Robust estimation of dispersion matrices and principal components. Journal American Statistiscal Association. 1981;76:354-362.

[9] Hubert M, Rousseeuw PJ, Van Aelst S. High breakdown robust multivariate methods. Statistical Science.2008;23; 92-119.

[10] Jin J, An J. Robust discriminant analysis and its application to Identify protein coding regions of rice genes. Mathematical Biosciences. 2011; 232: 96-100.

[11] Kim SJ, Magnani A, Boyd SP. Robust Fisher discriminant analysis. Advances in Neural Information Processing System.2005; 18:659-666.

[12] Pires AM, Branco JA. Generalization of Fisher's linear discriminant. www.math.ist.utl.pt/~apires/PDF/APJB_RP96.pdf, 1996.

[13] Roelant E, Van Aelst S, Williems G. The minimum weighted covariance determinant estimator. Metrika. 2009; 70:177-204.

[14] Wu G, Chen C, Yan X. Modified minimum covariance determinant estimator and its application to outlier detection of chemical process data. Journal of Applied Statistics. 2011; 38:1007-1020.

[15] Linnet K. On the sensitivity of linear discriminant analysis to sampling variation and analytic errors. Computers and Biomedical Research. 1988; 21:158-168.

[16] Zuo Y. Robust location and scatter estimators in multivariate analysis. *WSPC/Trim Size:9in x6in for Review*.2005; 0-31.

[17] Okwonu FZ, Othman AR. Heteroscedastic variance covariance matrices for unbiased two groups linear classification methods. Applied Mathematical Sciences. 2013; 7: 6855-6865.

[18] Okwonu FZ, Othman AR. Effect of heteroscedastic variance covariance matrices on two groups linear classification techniques. Journal of Mathematics and System Science.2014;4: 133-138.

[19] Okwonu FZ, Dieng H, Othman AR, Hu OS. Classification of aede adult mosquitoes in two distinct groups based on Fisher linear discriminant analysis and FZOARO techniques. Mathematical Theory and Modelling. 2012;2:22-30.