Probability Base Classification Technique: A Preliminary Study for Two Groups

Friday Zinzendoff Okwonu [1,2*] Abdul Rahman Othman [2]

1. Department of Mathematics and Computer Science, Delta State University, P.M.B.1, Abraka, Nigeria
2. Center for Mathematical Sciences, School of Distance Education, Universiti Sains Malaysia, 11800, Penang, Malaysia

* E-mail:fzokwonu_delsu@yahoo.com

Abstract
The conventional Fisher linear classification technique to perform classification for two groups problem is strictly developed based on the within group sample mean vectors and within group sample variance covariance matrices. A comparable classification procedure that incorporate the within group probabilities is considered. The conventional procedure based on the Fisher's technique assumed equality of the within group probabilities as such the computational procedure negate the within groups probabilities to solve classification problems. The new approach is a modification of the coefficient of the Fisher's technique by applying the within group probability for the respective groups to solve classification problems.The classification performance of these techniques is investigated based on generated contaminated normal data set using homoscedastic and heteroscedastic variance covariance matrices for various sample sizes and dimensions. The comparative performance of these procedures are investigated by comparing the mean probabilities of correct classification based on the contaminated date set with the mean of the optimal probability computed from the uncontaminated data set. The comparative classification performance revealed that both techniques perform comparable. Though, the Monte Carlo simulation indicate that as the proportion of contamination increases, the probability base approach perform better for homoscedastic covariance matrices, on the other hand, the Fisher's technique outperformed the probability base procedure for heteroscedastic covariance matrices. The comparative analysis indicate that the probability base approach performed comparable with the conventional procedure. The implication of this procedure indicate that classification problems can be solved by incorporating the respective within group probabilities to develop the classification model.

Keywords: Classification, Homoscedastic and Heteroscedastic Covariance Matrices, Mean Probability

1. Introduction

Conventionally, the linear classification problem for two groups is accomplished using the Fisher Linear Classification Analysis (FLCA). This procedure strictly depends on the within group sample mean vectors and the within group sample variance covariance matrices. The Fisher's technique is based on the assumption of multivariate normal data set and the variance covariance matrices are homoscedastic. The sample mean vectors and sample covariance matrices are unstable because these parameters are susceptible or easily influenced by influential observations (Maronna *et al.* 2006; Munoz-Pichardo *et al.* 2011). Sajobi *et al.* (2012) proposed to robustify the sample mean vectors and the covariance matrices by replacing the maximum likelihood estimates by the maximum likelihood estimators computed based on coordinate wise trimming. Hubert *et al.* (2010) proposed permutation invariant technique called deterministic algorithm for the minimum covariance determinant procedure. This procedure uses permutation/deterministic method rather than the random subset to robustify the sample mean and covariance matrix. Bouveyron & Brunet (2012) proposed robust and flexible Fisher linear discriminant analysis based on probabilistic concept that "relax" the equal covariance assumption. This technique, basically does not incorporate the within group probabilities in computing the classification coefficient.

This paper consider the modification of the Fisher's technique by introducing the within group probabilities to the separation parameter w. The new procedure solve classification problems for two groups by incorporating the information the within group probabilities provides and to obtain maximum correct classification rate. This procedure adheres strictly to the homoscedastic assumption of the covariance matrices. The performance of these methods is investigated for contaminated normal data set, equal and unequal variance covariance matrices.

The methodology section contains the Fisher linear classification analysis followed by the probability base classification technique. Simulation results are contained in results section followed by discussion and conclusions, respectively.

2. Method

The method section consists of the Fisher linear classification analysis and the Probability base classification technique. Both procedures are applied to perform classification for two groups problem.

2.1 Fisher Linear Classification Analysis (FLCA)

It is observed that the two groups linear classification technique based on Fisher's technique assumed that the within group probability and misclassification cost are equal, as such its classification rule negate the probabilities for each group, that is:

$$\xi - \overline{\xi} \geq \ln\left(\frac{\aleph_{c2}(1/2)}{\aleph_{c1}(2/1)}\left(\frac{p_2}{p_1}\right)\right) = 0 \tag{1}$$

$$\xi\text{-}\overline{\xi} < \ln\left(\frac{\aleph_{c2}(1/2)}{\aleph_{c1}(2/1)}\left(\frac{p_2}{p_1}\right)\right) = 0 \tag{2}$$

As observed in the literature, the Fisher's technique performs optimally if the data set is drawn from the multivariate normal distribution and if the variance covariance matrices are equal. When the classification coefficient is inconsistent, the misclassification rate tends to increase. The within group mean vector, variance covariance matrices and the pooled common covariance matrix are defined as follows:

$$\overline{x}_i = \sum_{j=1}^{N_i} x_{ij} / N_i, i = 1,2 \tag{3}$$

$$S_i = \sum_{j=1}^{N_i} (x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)' / (N_i - 1) \tag{4}$$

$$S_{pooled} = \frac{\sum_{i=1}^{2} (N_i - 1) S_i}{\sum_{i=1}^{2} N_i - 2} \tag{5}$$

Equations (3-5) are applied to develop Equations (1-2). Based on the equality assumptions in Equations (1-2), the Fisher's procedure reduces to:

$$\xi \geq \overline{\xi} \tag{6}$$

$$\xi < \overline{\xi} \tag{7}$$

where, $\xi = (\overline{x}_1 - \overline{x}_2) S_{pooled}^{-1} x = q'x$ is the classification score and $\overline{\xi} = ((\overline{x}_1 + \overline{x}_2)/2)q'$ is the cutoff point. Equations (6-7) defines the Fisher's classification rule. Equation (6) implies that an observation in group one is allocated correctly to group one otherwise the observation is assigned to group two if Equation (7) is satisfied, respectively.

2.2 Probability Base Classification Technique (PCT)

This section describe classification procedure that includes the within group probabilities to develop the classification coefficient. Based on Equation (3), the within group mean vectors difference for the two groups is obtained, say, $d = \overline{x}_1 - \overline{x}_2$ and the sum of the within group mean vectors is given as $\widehat{d} = \overline{x}_1 + \overline{x}_2$, respectively. To formulate the coefficient of the new procedure, the following are obtained:

$$\begin{aligned} &\partial = | d |, \\ &\tilde{d} = 1 + \sqrt{\partial}, \\ &\beta = d^2 / \tilde{d}, \\ &\varepsilon = 1 - \left|\beta^2\right|. \end{aligned} \tag{8}$$

Based on the definitions in Equation (8), the following is obtained:

$$w = e^{\beta} + e^{\beta^2/\varepsilon} + p_i \tag{9}$$

where, $P_i = N_i/N$ is the within group probabilities, N_i is the sample size for each group, N is the total sample size for the two groups and $p = \sum_{i=1}^{2} p_i$, is the total probability. The classification model is given as:

$$z = \left(\frac{w}{S_{\text{pooled}}^{-1}}\right)' x = u'x, \qquad u = \frac{w}{S_{\text{pooled}}^{-1}}. \tag{10}$$

The classification cutoff point is given as follows:

$$\bar{z} = \frac{\hat{d}}{2}u' \tag{11}$$

The classification rule is defined as:

$$z < \bar{z} \tag{12}$$

in this regard, an observation is assigned to group one if Equation (12) is satisfied otherwise the observation is classified to group two if the following equation hold:

$$z \geq \bar{z} \tag{13}$$

3. Result

The Monte Carlo simulation is designed to investigate the comparative classification performance of the above techniques for unequal and equal variance covariance matrices based on contaminated normal data set. The contamination normal model used in this study for the respective groups is given as:

$$(1-\varepsilon)N_{d_p}(0,1) + \varepsilon N_{d_p}(\mu, \sigma^2 I_{d_p}) \tag{14}$$

This model require that majority of the data set come from the normal distribution while the rest come from the contaminated distribution (Cont. Dist.). In each case, the data set is randomly reshuffled and divided into two categories; say training set (60%) and validation set (40%). To determine the performance of each procedure, the mean of the optimal probability (Opt.) is used as the performance benchmark. The comparative analyses are based on the comparison between the mean of the optimal probability computed from the uncontaminated normal data set and the mean probabilities of correct classification obtain from each technique. In the respective figures, the straight line is the performance benchmark. Figure 1 and Figure 2 show that the Fisher's technique performed better than the probability based approach for increasing proportion of contamination for the unequal variance covariance matrices. Figure 3 revealed that the probability base approach performed better than the Fisher's technique for the equal variance covariance matrices and performed comparable in Figure 4. The following results in Tables 1 and 2 reveal the performance of these technqiues for heteroscedastic matrices while Table 3 and 4 show the performance of both techniques for homoscedastic matrices. The best procedure appears in bold. The analysis reveals that the FLCA and the PCT techniques are comparable in all cases investigated.

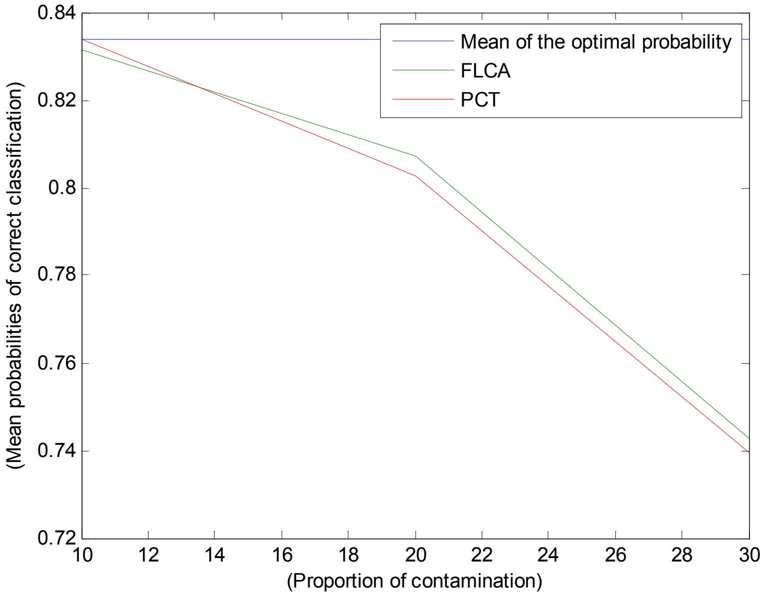

Figure 1.Effect of contamination on the mean probability of correct classification

Table 1. Mean probability of correct classification and standard deviation (In Bracket), Optimal = 0.8340

Con. Dist.	N_i	d_p	ε	FLCA	PCT	OPT-FLCA	OPT-PCT
$\varepsilon \mathrm{N}_2(3,10)$	30	2	10	0.8314 (0.0055)	**0.8338** (0.0120)	0.0026	0.0002
$\varepsilon \mathrm{N}_2(3,10)$	30	2	20	**0.8072** (0.0065)	0.8026 (0.0140)	0.0268	0.0314
$\varepsilon \mathrm{N}_2(3,10)$	30	2	30	**0.7425** (0.0100)	0.7393 (0.0096)	0.0915	0.0947

FLCA: Fisher linear classification analysis
PCT: Probability base classification technique
OPT-FLCA: Difference between the mean of the optimal probability and the mean probability of FLCA
OPT-PCT: Difference between the mean of the optimal probability and the mean probability of PCT

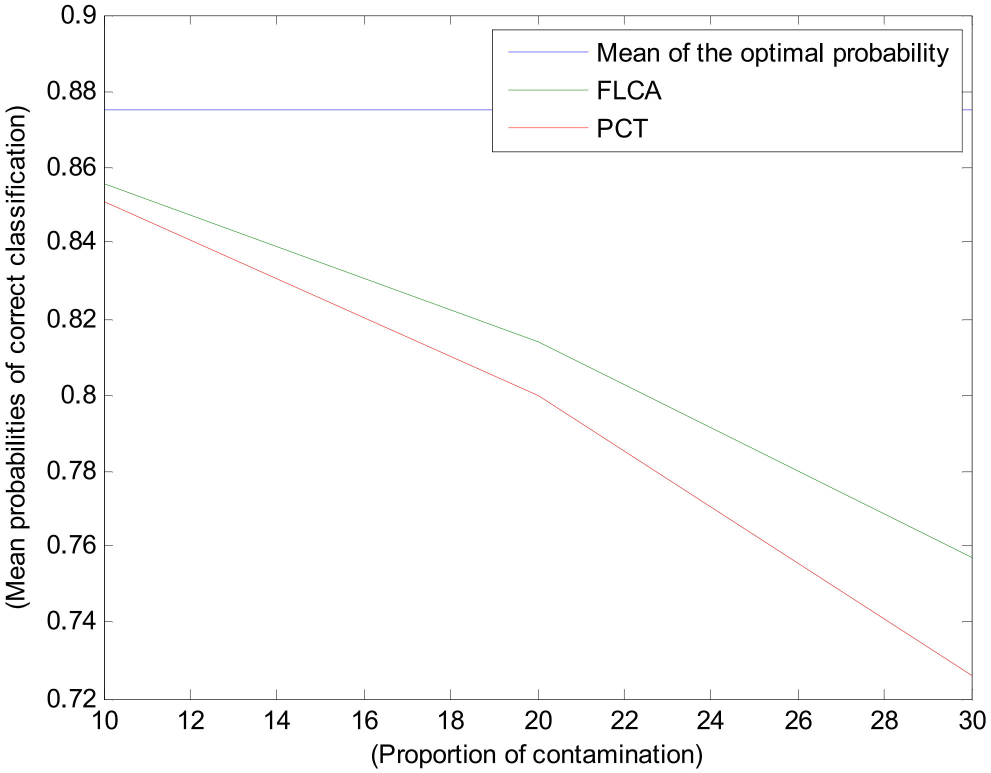

Figure 2.Effect of contamination on the mean probability of correct classification

Table 2. Mean probability of correct classification and standard deviation (In Bracket), Optimal = 0.8749

Con. Dist.	N_i	d_p	ε	FLCA	PCT	OPT-FLCA	OPT-PCT
$\varepsilon \mathrm{N}_3(4.25)$	60	3	10	**0.8553** (0.0068)	0.8506 (0.0033)	0.0196	0.0244
$\varepsilon \mathrm{N}_3(4.25)$	60	3	20	**0.8141** (0.0084)	0.7997 (0.0030)	0.0608	0.0752
$\varepsilon \mathrm{N}_3(4.25)$	60	3	30	**0.7570** (0.0096)	0.7261 (0.0070)	0.1179	0.1488

FLCA: Fisher linear classification analysis
PCT: Probability base classification technique
OPT-FLCA: Difference between the mean of the optimal probability and the mean probability of FLCA
OPT-PCT: Difference between the mean of the optimal probability and the mean probability of PCT

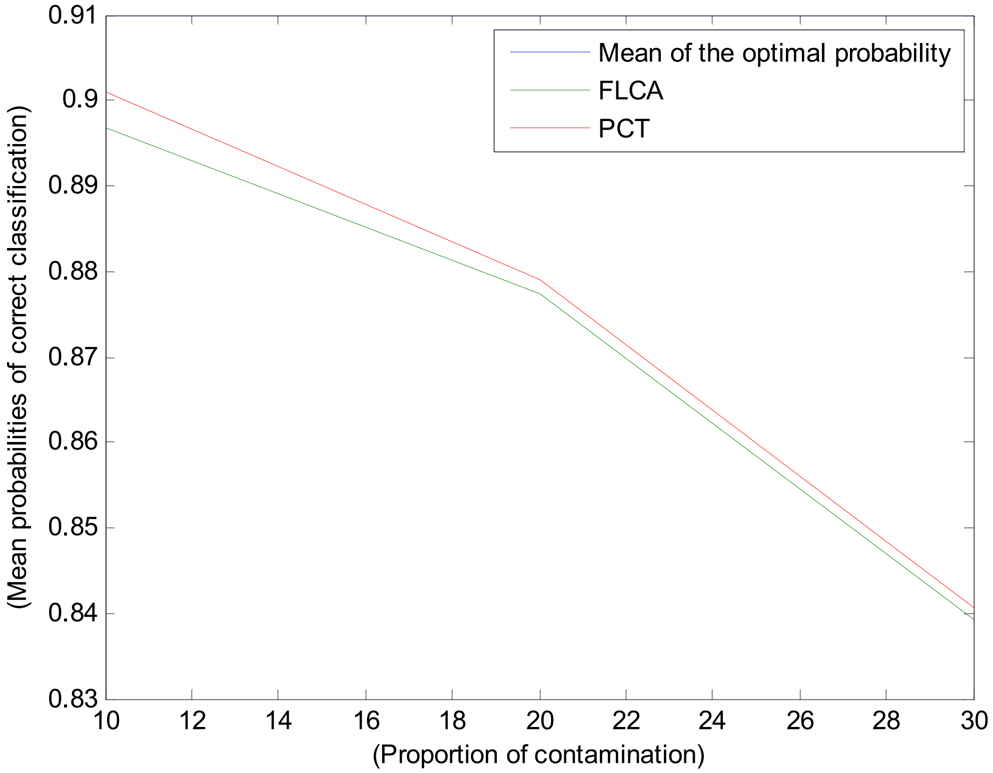

Figure 3 .Effect of contamination on the mean probability of correct classification

Table 3. Mean probability of correct classification and standard deviation (In Bracket), Optimal = 0.9099

Con. Dist.	N_i	d_p	ε	FLCA	PCT	OPT-FLCA	OPT-PCT
$\varepsilon N_3(2,9)$	30	3	10	0.8967 (0.0106)	**0.9009** (0.0063)	0.0132	0.009
$\varepsilon N_3(2,9)$	30	3	20	0.8774 (0.0100)	**0.8791** (0.0128)	0.0325	0.0308
$\varepsilon N_3(2,9)$	30	3	30	0.8392 (0.0146)	**0.8406** (0.0128)	0.0707	0.0694

FLCA: Fisher linear classification analysis
PCT: Probability base classification technique
OPT-FLCA: Difference between the mean of the optimal probability and the mean probability of FLCA
OPT-PCT: Difference between the mean of the optimal probability and the mean probability of PCT

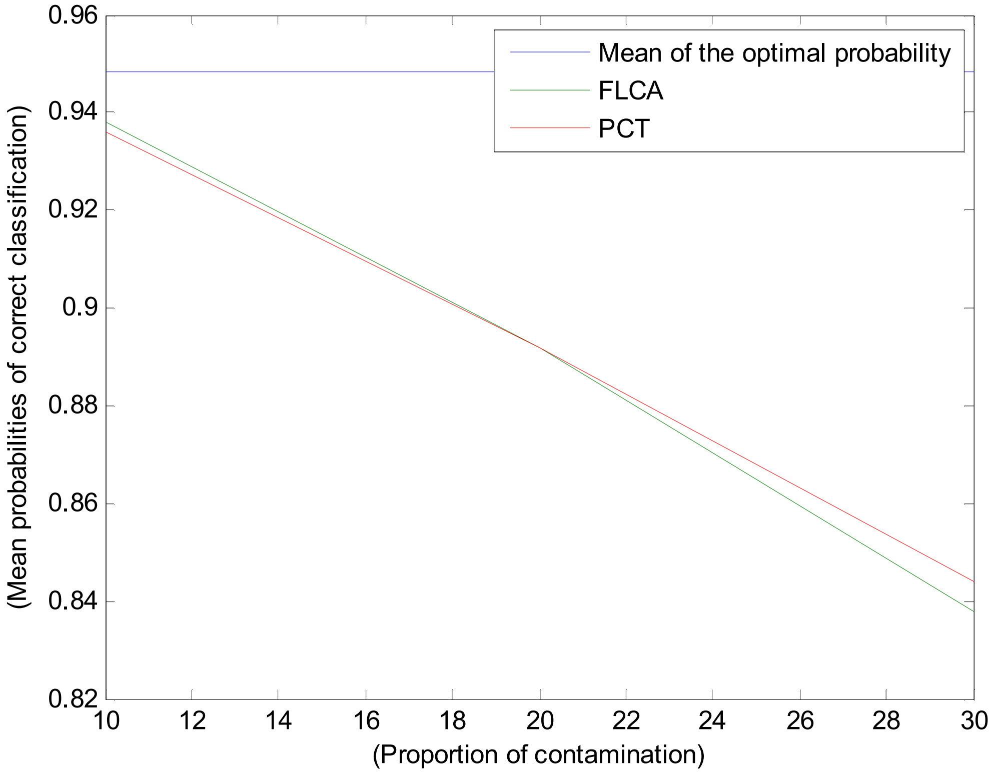

Figure 4 .Effect of contamination on the mean probability of correct classification

Table 4. Mean probability of correct classification and standard deviation (In Bracket), Optimal = 0.9484

Con. Dist.	N_i	d_p	ε	FLCA	PCT	OPT-FLCA	OPT-PCT
$\varepsilon \mathrm{N}_5(4,16)$	100	5	10	**0.9383** (0.0085)	0.9362 (0.0074)	0.0101	0.0122
$\varepsilon \mathrm{N}_5(4,16)$	100	5	20	0.8917 (0.0072)	**0.8920** (0.0012)	0.0567	0.0564
$\varepsilon \mathrm{N}_5(4,16)$	100	5	30	0.8379 (0.0042)	**0.8438** (0.0024)	0.1105	0.1046

FLCA: Fisher linear classification analysis
PCT: Probability base classification technique
OPT-FLCA: Difference between the mean of the optimal probability and the mean probability of FLCA
OPT-PCT: Difference between the mean of the optimal probability and the mean probability of PCT

3.1 Discussion

The conventional technique to solve classification problem based on the Fisher's technique does not incorporate the within group probabilities to develop the Fisher's classification coefficient, see Equations (1-2). A comparable classification technique that incorporate the within group probabilities to formulate the classification coefficient was proposed. The classification performance of these techniques was investigated by violating the homoscedastic and multivariate normality assumptions. The Monte Carlo simulations performed are based on the following controlled variables; the mean vector shift, variance shift, sample size and dimension, proportion of contamination. The comparative classification performace based on the figures and tables revealed that these techniques performed comparable. These techniques ultilize all the information glean from the data set. The probability base approach provide more information to the end user than the conventional technique.

4. Conclusion

A comparable classification technique based on probability concept for two groups problem was compared with the conventional Fisher linear classification procedure. The new technique based on the within group probabilities is suitable to perform classification for two groups problem where the probability of the respective groups are given. The comparative analyses revealed that both techniques performed comparable.

Acknowledgement

This research work was funded through the short term grant of the Universiti Sains Malaysia, Penang, Malaysia.

References

Bouveyron, C. & Brunet, C. (2012), " Probabilistic Fisher Discriminant analysis: A robust and Flexible Alternative to Fisher Discriminant Analysis", *Neurocomputing* **90**,12-22.

Hubert, M., Rousseeuw, P. J. & Verdonck, T. (2010), "A Deterministic Algorithm for the *MCD. Citeseerx.ist.psu.edu/viewdoc/summary?*, 1-26.

Maronna, R., Martin, R. D. & Yohai, V. J. (2006), "Robust Statistics: Theory and Methods", *John Wiley, New York.*

Munoz-Pichardo, J. M., Enguix-Gonzalez, A., Munoz -Garcia, J. & Moreno-Rebollo, J. L. (2011)," Influence Analysis on Discriminant Coordinates", *Communications in Statistics-Simulation and Computation,* **40**(60), 793-807.

Sajobi, T. T., Lix, L. M., Dansu, B. M., Laverty, W. & Li, L. (2012), "Robust Descriptive Discriminant Analysis for Repeated Measures Data", *Computational Statistics and Data Anal.ysis,* **56**(9), 2782-2794.