

RESEARCH

Open Access



# Asymptotic performance of the quadratic discriminant function to skewed training samples

Atinuke Adebanji<sup>1†</sup>, Michael Asamoah-Boaheng<sup>2\*†</sup> and Olivia Osei-Tutu<sup>3†</sup>

\*Correspondence:

asboaheng@yahoo.com

<sup>†</sup>All the three authors contributed equally in conducting this research.

<sup>2</sup>Institute of Research, Innovation and Development (IRID), Kumasi Polytechnic, Box 854, Kumasi, Ghana  
Full list of author information is available at the end of the article

## Abstract

This study investigates the asymptotic performance of the quadratic discriminant function (QDF) under skewed training samples. The main objective of this study is to evaluate the performance of the QDF under skewed distribution considering different sample size ratios, varying the group centroid separators and the number of variables. Three populations ( $\pi_i, i = 1, 2, 3$ ) with increasing group centroid separator function were considered. A multivariate normal distributed data was simulated with MatLab R2009a. There was an increase in the average error rates of the sample size ratios 1:2:2 and 1:2:3 as the total sample size increased asymptotically in the skewed distribution when the centroid separator increased from 1 to 3. The QDF under the skewed distribution performed better for the sample size ratio 1:1:1 as compared to the other sampling ratios and under centroid separator ( $\delta = 5$ ).

**Keywords:** Group centroid separator, Lognormal distribution, Error rates, Coefficient of Variation

## Background

Discriminant analysis is used in situations where the clusters are known a priori. The main aim of discriminant analysis is to classify an observation, or several observations, into these known groups (Härdle and Simar 2007). The problem of multiple group discrimination under normality and non-normality for a long time has posed a challenge to researchers and several attempts have been made at deriving parsimonious rules that address this hurdle (Asamoah-Boaheng et al. 2014). This study evaluates the asymptotic performance of a three group quadratic discriminant function (QDF) under non-normal distribution with varying degrees of sample sizes, varying variable selections and under increasing group centroid separators.

Lachenbruch et al. (1977) studied the performance of the QDF under non-normality. They generated random samples from non-normal distributions and the samples were transformed into components by using Johnson's system of transformation. Among their findings, they found that, the overall sample standard deviation, the between sample variability of the individual error rates of the function (QDF) under normal or non-normal distributions was quite large. In the computation of the overall sample standard

deviation, the between sample variability of the individual error rates in the QDF on normal or non-normal distributions was quite large and for that instability of QDF is pronounced. Also the actual error rates were considerably larger than the optimal rates in the case of zero mean difference (this is a very difficult problem in assignment). The QDF for non-normal samples generally did not do substantially worse than when the QDF was derived under normal samples which were obtained after transformation. Lachenbruch et al. (1977) compared the re-substitution method and the leave-one-out method. The re-substitution method had an unacceptably high bias. The leave-one-out method was far superior in respect of generally having a far lesser bias.

Hosseini and Armacost (1992) presented a study on two group discriminant problem with equal group mean vectors with several methods and mathematical formulations. For comparative purposes, both Fishers linear discriminant function (FLDF) and that of QDF were used. Both methods performed better in the case of multivariate non-normal distributions than compared to that of the one generated from a multivariate normal distribution. All the various discriminatory methods performed better generally when the covariance matrices for the two populations were assumed to be unequal. Also, less favourable performance was observed for FLDF as well as QDF with presence of outliers than when there was absence of outliers/noise. Lachenbruch and Goldstein (1979) considered the effects of initial misclassification on the QDF. In his simulation, a population of two with equal priori probabilities, mean of 0 and 2 and number of variables, 2, 4, 8 and a fraction  $\alpha_i$  of the  $n_i$ , which are actually from the other population, were considered. He then suggested that if initial misclassification is suspected, all sample points should be carefully checked and reassigned if needed. Krzanowski and Hand (1977) considered an assessment of error rate estimators paying special attention to the leave-one-out method. The estimator was investigated in a simulation study, both in absolute terms and in comparison with a popular bootstrap estimator. Motivated by this, extension of leave-one-out, the leave-two-out was looked at considering the variance. As expected, the leave-two-out method yields a slight variance reduction relative to the leave-one-out method, but was not enough to make it a good competitor.

In order to study the asymptotic error rates of linear, quadratic and logistic rules, Kakai and Pelz (2010) conducted a Monte Carlo study in two, three and five-group discriminant analysis. The simulation study took into account the overlap of the populations ( $e = 0.05$ ,  $e = 0.1$ ,  $e = 1.5$ ), their common distribution (normal, chi-square with 4, 8 and 12 df) and their heteroscedasticity degree,  $\Gamma$ , measured by the value of the power function,  $1 - \beta$  of the homoscedasticity test related to  $\Gamma$  ( $1 - \beta = 0.05$ ,  $1 - \beta = 0.4$ ,  $1 - \beta = 0.6$ ,  $1 - \beta = 0.8$ ). They found that the three rules gave similar error rates for normal homoscedastic populations. For non-normal populations, quadratic rule still gave lowest relative error except for two-group where logistic was the best. The quadratic and logistic rules were more influenced by the number of groups irrespective of their lowest relative error. Also linear and quadratic were more influenced by non-normality. The study deviates from Lachenbruch et al. (1977) by focusing on three populations, unequal sample sizes and log-normal distribution for the skewness. Croux (2004) studied the influence of observations on the misclassification probability in quadratic discriminant analysis. They also studied the effect of observations in the training sample on the performance of the associated classification rule. MacFarland (2001) investigated

into the exact misclassification probabilities for plug-in normal quadratic functions; the case of equal mean. A stochastic representations for the exact distributions of the “plug-in” quadratic discriminant functions was derived for classifying a newly obtained observation.

As evident in the above literatures, several researchers have done extensive work on the performance of various discriminant and classification functions under skewed or non normal distributions. However, not much attention has been focused on studying and evaluating the performance of these classifiers using three populations under skewed distribution considering different sampling ratios, under different centroid separators and under varying variable selections. This study therefore seeks to investigate the performance of a single classifier (i.e the QDF) under skewed distribution considering different variable selections, varying sampling ratios and varying centroid separators considering three groups/populations.

**Methods**

**The quadratic classifier ( $\Sigma_1 \neq \Sigma_2$ )**

Suppose that the joint densities of  $X' = [X_1, X_2, \dots, X_p]$  for population  $\Pi_1$  and  $\Pi_2$  are given by

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right] \tag{1}$$

When the multivariate normal densities have different covariance structures, the terms in the density ratio involving  $|\Sigma_i|^{1/2}$  do not cancel as they do when we have equal covariance matrices and also the quadratic forms in the exponents of  $f_i(\mathbf{x})$  do not combine. Therefore substituting multivariate normal densities with different covariance matrices into Eq. (1) and after taking the natural logarithms and simplifying, the likelihood of the density ratios gives the quadratic function (assuming equal misclassification cost). Allocate  $\mathbf{x}$  to  $\Pi_1$  if

$$-\frac{1}{2} \mathbf{x}' (\Sigma_1^{-1} - \Sigma_2^{-1}) \mathbf{x} + (\mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1}) \mathbf{x} - k \geq \ln \left[ \left( \frac{p_2}{p_1} \right) \right],$$

where

$$k = \frac{1}{2} \ln \left( \frac{|\Sigma_1|}{|\Sigma_2|} \right) + \frac{1}{2} (\mu_1' \Sigma_1^{-1} \mu_1 - \mu_2' \Sigma_2^{-1} \mu_2) \tag{2}$$

otherwise,  $\mathbf{x} \in \Pi_2$ . Considering the Mahalanobis distance, the function is sometimes written as

$$f(\mathbf{x}) = D_1^2(\mathbf{x}) - D_2^2(\mathbf{x}) + \ln \left[ \frac{|\Sigma_1|}{|\Sigma_2|} \right] - 2 \ln \left( \frac{p_1}{p_2} \right) \tag{3}$$

The quantity  $D_i^2(\mathbf{x}) = (\mathbf{x} - \mu_i)' \Sigma_i^{-1} (\mathbf{x} - \mu_i)$  is the Mahalanobis square distance.

When  $\Sigma_1 = \Sigma_2$  the function reduces to the linear classifier rule.

This function is easily extended to the three group classification where two cut off points are required for assigning observations to the three groups (Johnson and Wichern 2007).

**Simulation design**

We evaluated the performance of QDF in case of skewed training samples following non normal distribution. In the simulation procedure, multivariate normally correlated random data was generated for three populations with their mean vector  $\mu_1 = (0, \dots, 0)$ ,  $\mu_2 = (0, \dots, \delta)$  and  $\mu_3 = (0, \dots, 2\delta)$  respectively using MatLab R2009a.

The covariance matrices,  $\Sigma_i (i = 1, 2, 3)$ , where  $k \neq l, \sigma_{kl} = 0.7$  for all groups except the diagonal entries given as  $\sigma_k^2 = i$ , for  $i = 1, 2, 3$  were obtained. Three different groups or populations which are normally correlated data were generated. Since the researchers were interested in evaluating the performance of the QDF under skewed uncorrelated data, the data was transformed from correlated normal to skewed data. In transforming the data, skewed data was generated by taking an exponents of the normally correlated/ log normal data.

QDF was then performed in each case and the leave-one-out method was used to estimate the proportion of observations misclassified. Factors considered in this study were:

1. Mean vector separator which is set at  $\delta$  from 1 to 5 where  $\delta$  is determined by the difference between the mean vectors.
2. Sample sizes which are also specified. Here 14 values of  $n_1$  set at 30, 60, 100, 150, 200, 250, 300, 400, 500, 600, 700, 800, 1000, 2000 and the sample size of  $n_2$  and  $n_3$  are determined by the sample ratios at 1:1:1, 1:2:2 and 1:2:3 and these ratios also determined the prior probabilities to be considered.
3. The number of variables were set at 4, 6 and 8 following (Murray 1977).
4. The size of population 1 ( $n_1$ ) was fixed throughout the study and the sizes of populations 2 and 3,  $n_2$  and  $n_3$  respectively are determined by the sample size ratio under consideration.

**Evaluating the performance of the QDF**

Let  $r$  denote the classification rule obtained on individuals belonging to  $p$ -variate populations with mixture density  $\mathbf{F}$ . The error rate can be defined as the overall probability of misclassification associated with the classification rule. The probability  $e_{jk}(r, \mathbf{F})$  that  $r$  allocates a random observation vector  $\mathbf{X}$  to  $G_j$  while it belongs to  $G_k$  and is computed as follows [McLachlan (1992) as cited in Kakai and Pelz (2010)].

$$e_{jk}(r, \mathbf{F}) = P((\mathbf{X}, \mathbf{F}) = P(r(\mathbf{X}, \mathbf{F}) = j | \mathbf{X} \in G_k), \quad (j, k = 1, \dots, g - 1; \quad j \neq k).$$

The overall error rate  $e(r, \mathbf{F})$  associated with  $r$  is computed as shown below.

$$e(r, \mathbf{F}) = \sum_{k=1}^g P_k \sum_{j(\neq k)=1}^{g-1} e_{jk}(r, \mathbf{F})$$

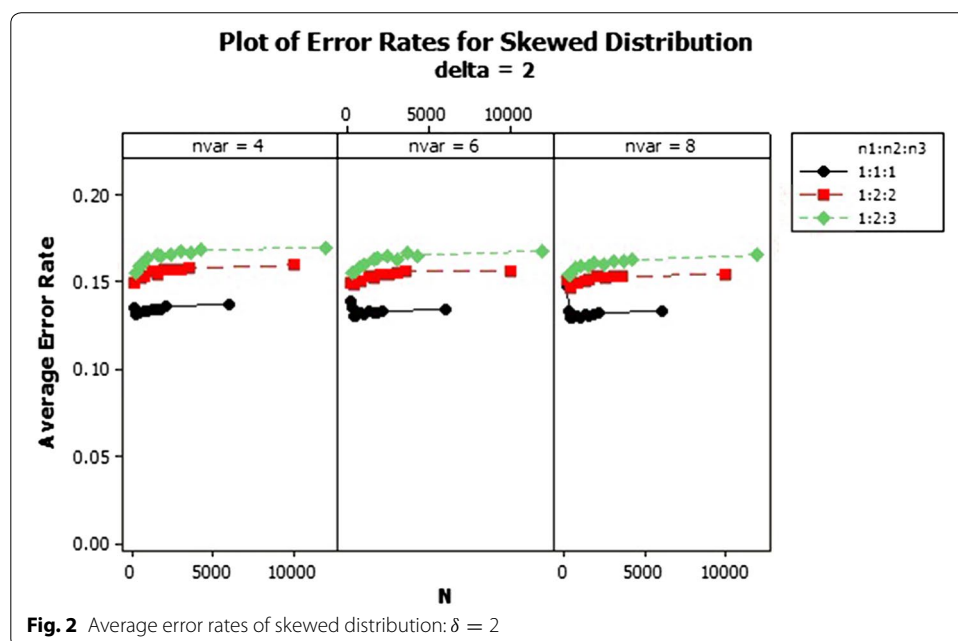
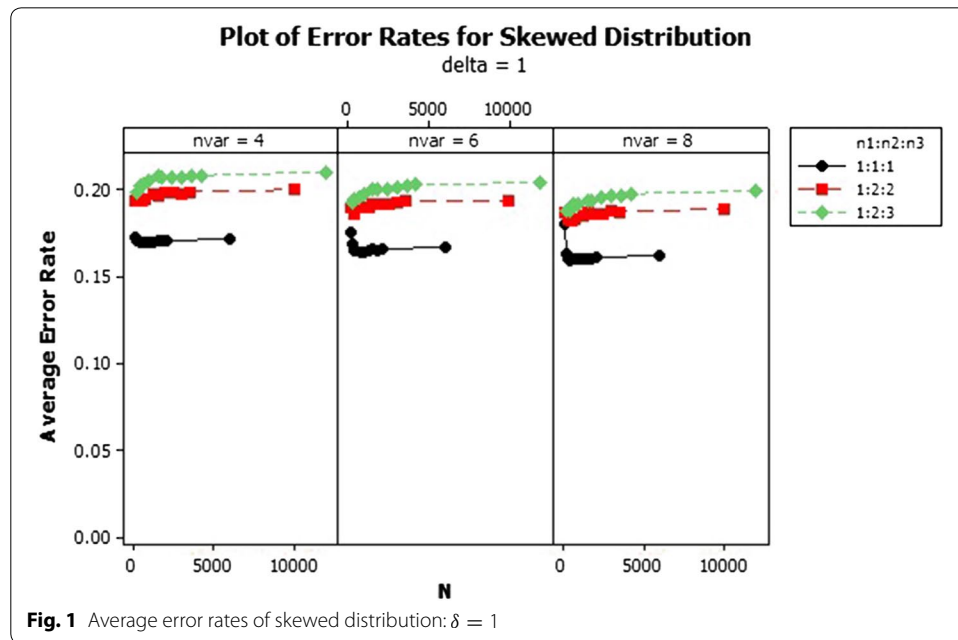
where  $p_k (k = 1, \dots, g)$  is the group prior probability of  $G_k$

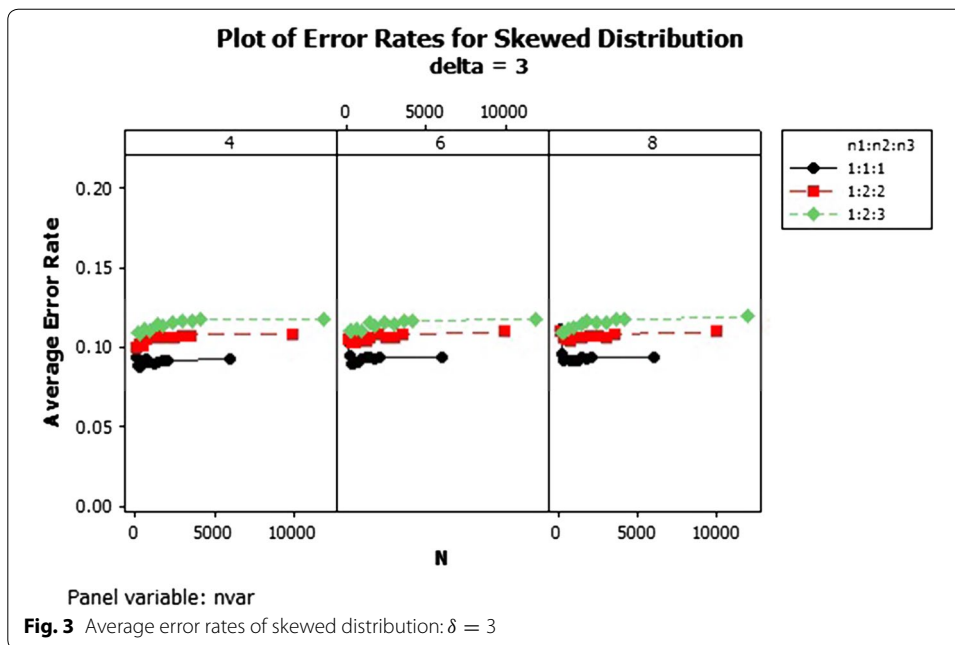
**Results and discussion**

This sections presents the outcome and discussion of the simulation results of the asymptotic performance of the QDF under skewed training samples.

**Performance of QDF under varying sampling ratios**

From the results, there was an increase in the average error rates of the sample size ratios 1:2:2 and 1:2:3 as the total sample size increased asymptotically in the skewed distribution for  $\delta = 1-3$  as shown in Figs. 1, 2 and 3. In Fig. 1 for  $\delta = 1$  the lowest error rates were reported for equal sample size ratios (1:1:1). The error rates reduced marginally across the number of variables. Improvement in the performance was achieved with increased Mahalanobis distance and not asymptotically. The patterns of the error rates did not change significantly beyond  $\delta = 3$  as shown in Fig. 3. The average error rates for

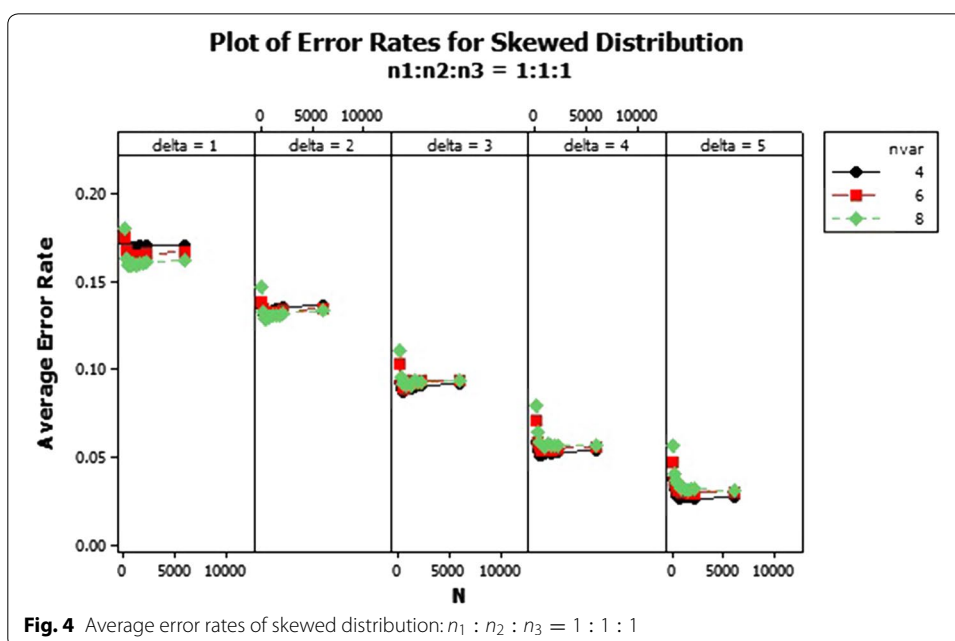


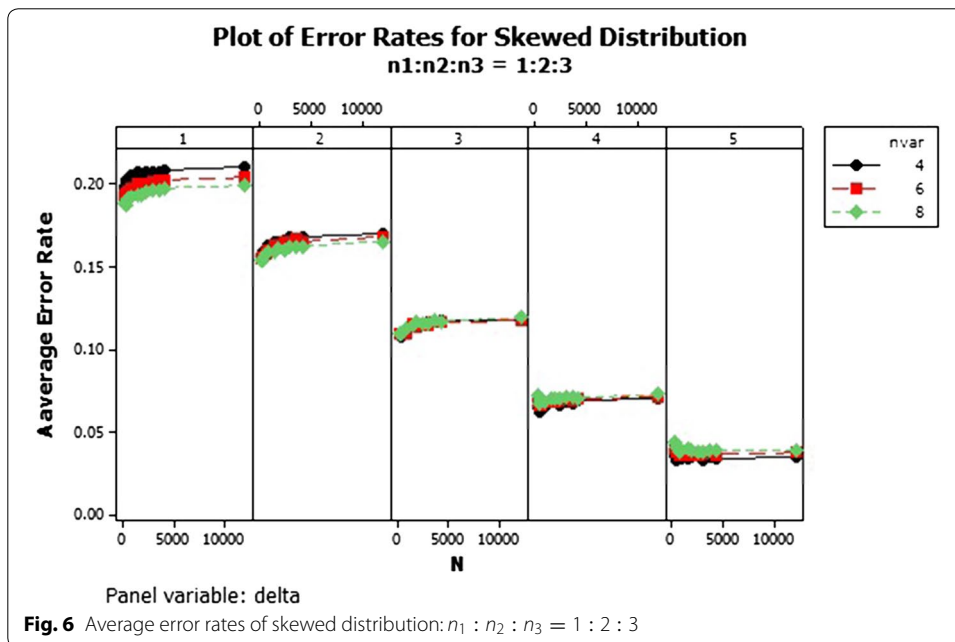
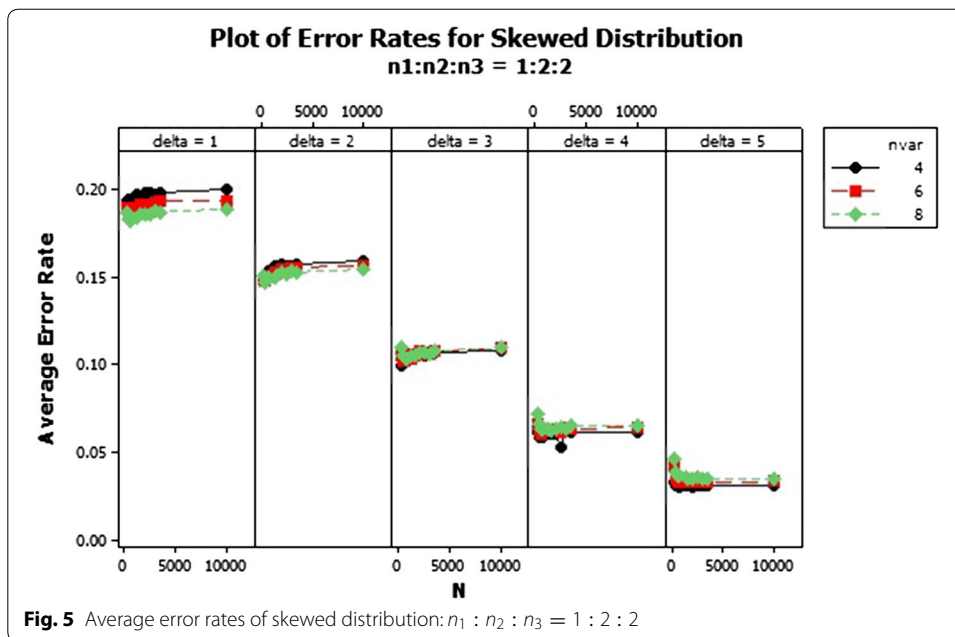


$\delta = 5$  were the lowest as compared to the other  $\delta$ s and they decreased as the total sample size increased.

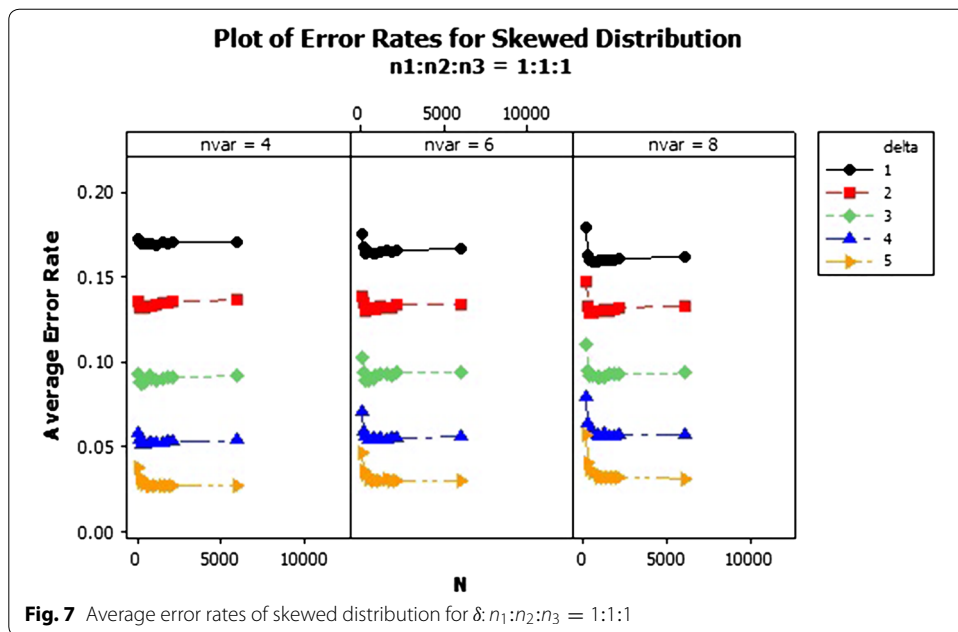
**Effects of number of variables on the performance of the QDF**

The QDF performs differently with increasing number of variables. For sample size ratio 1:1:1, the average error rates of the variables reduced and curved upward as the total sample size increased for all  $\delta$ s, as shown in Fig. 4. The average error rates of sample ratios 1:2:2 and 1:2:3 were different as shown in Figs. 5 and 6. Also from Figs. 5 and 6 the average



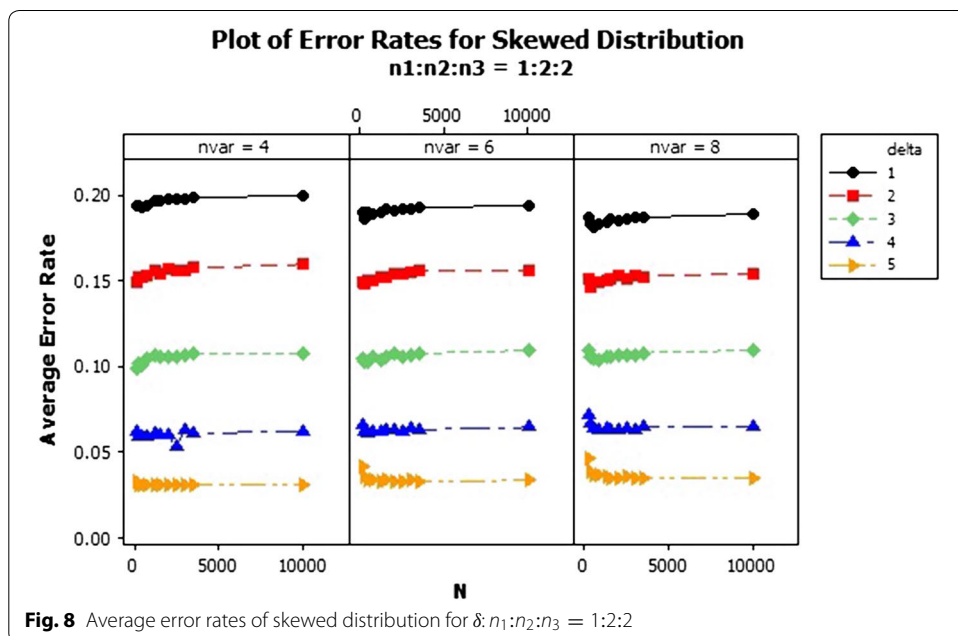


error rate of the QDF for the respective populations increased as the total sample size increased and reduced with increasing number of variable for  $\delta = 1$  and  $2$ . In  $\delta = 3$  and  $4$  of ratios  $1:2:2$  and  $1:2:3$ , as the number of variables increased the average error rate of the QDF dropped from the total sample size of  $150-300$  and increased as the sample size also increased respectively while that of  $\delta = 5$  decreased marginally. In general the average error rate increased as the number of variables increased with increasing  $\delta$ .



**Effects of group centroid separator on the performance of QDF**

The average error rate of the skewed distribution for sample size ratio 1:1:1 in Fig. 7 revealed that, as the sample size increases, the average error rates of the individual  $\delta$ s generally reduces. Also from Fig. 8, the error rates increased marginally for the individual deltas (centroids separators) as the sample sizes increases. However the performance of the QDF was quite abysmal when the centroid separator was set at  $\delta = 1$  as compared to the other deltas since it recorded the highest error rates with respect to each of the variable selections as 0.20. Also as clearly indicated in Fig. 8, the error rates of the QDF



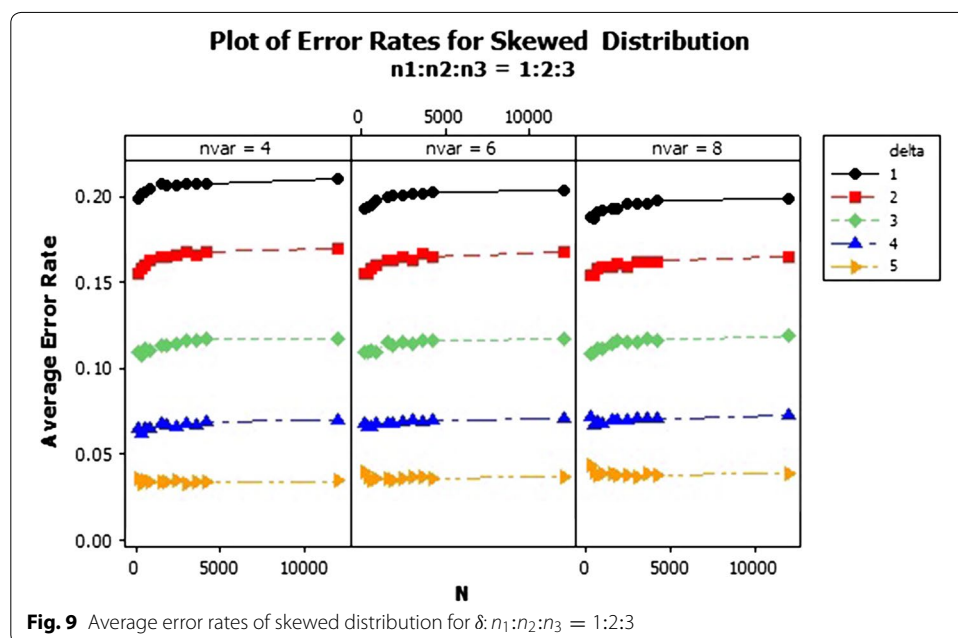


was minimised when the group centroid separator was set at  $\delta = 5$ . Hence increasing the group centroid separators minimizes the misclassification rates thereby enhancing the performance of the QDF under the sample ratio of 1:2:2. Finally the performance of the QDF was evaluated under the sampling ratio of 1:2:3 with respect to the three groups/populations,  $\pi_1, \pi_2, \pi_3$  with different selections of group centroids as shown in Fig. 9. From Fig. 9, similar results were obtained and the performance of the QDF was better under increasing group centroid separators, irrespective of the number of variables considered at a particular instance but was also dependent on the sample size selection.

**Conclusion**

This paper investigated the asymptotic performance of QDF on skewed training data for three populations ( $\pi_i, i = 1, 2, 3$ ) with increasing group centroid ( $\delta$ ), with chosen variables and sample size ratios. Results from the study indicates that, the QDF performed quite poorly with an increase in error rates under sample ratios 1:2:2 and 1:2:3 for  $\delta = 1 - \delta = 3$ . Other results also indicates that, the QDF performs better under an equal sample size ratio (1:1:1) resulting in a reduced misclassification rate with minimized error rates. The group centroid separators increased with decreasing group error rates and sample sizes. In other words, the QDF performed better in classifying the observations into their respective groups when the group centroid separators were increased. Also with increasing number of variables, from 4 to 8, the average error rate for evaluating the performance of the QDF dropped under  $\delta = 3, 4$  for sample ratios 1:2:2 and 1:2:3.

Generally, the study found that, there is always a pronouncement in the reduction of misclassification error rates as the group centroid separator increases as compared to an increasing sample size ratios. The results obtained from this study (skewed distribution) shows some conformity with Lachenbruch et al. (1977). Lachenbruch et al. (1977) generated random samples through simulations under non-normal distribution. Johnson’s



system of transformation was used to transform the generated random samples into components by components. After the transformation, the QDF was derived and its performance was evaluated by the estimated mean error rates, standard deviation and sample variability. From their study the QDF recorded very high and increasing error rates, standard deviation under non-normality compared with the performance of the function under normally distributed data/training samples. In other words, they discovered that the QDF under non normal samples generally performs quite poorly as compared to when their performance are evaluated under normal distribution.

#### Authors' contributions

AA worked on the literature review and background of the study as well as suggestion of the appropriate method and software used and also supervised the entire research work. MAB prepared the manuscript and discussed part of the findings and also revised the entire manuscript. OOT conducted the data analysis through simulation with the used of MATLAB software and also discussed part of the findings. All authors read and approved the final manuscript.

#### Author details

<sup>1</sup> Department of Mathematics, Kwame Nkrumah University of Science and Technology, PMB KNUST, Kumasi, Ghana.

<sup>2</sup> Institute of Research, Innovation and Development (IRID), Kumasi Polytechnic, Box 854, Kumasi, Ghana. <sup>3</sup> Department of Mathematics, Kwame Nkrumah University of Science and Technology, PMB KNUST, 24105 Kumasi, Ghana.

#### Competing interests

The authors declare that they have no competing interests.

Received: 22 February 2016 Accepted: 2 September 2016

Published online: 13 September 2016

#### References

- Asamoah-Boaheng M, Adebanji AO, Ababio N (2014) Comparing the performance of zero mean classification functions under unequal misclassification cost. *Am J Appl Math Stat* 2(6):409–415. doi:10.12691/ajams-2-6-9
- Croux C (2004) Influence of observations on the misclassification probability in quadratic discriminant analysis. *J Multivar Anal* 96:384–403
- Härdle W, Simar L (2007) *Applied multivariate statistical analysis*, 2nd edn. Springer, Berlin
- Hosseini J, Armacost R (1992) The two-group discriminant problem with equal group mean vectors: an experimental evaluation of six linear/nonlinear programming formulations. *Eur J Oper Res* 77:241–252
- Johnson RA, Wichern DW (2007) *Applied multivariate statistical analysis*. Pearson Education, Hillsdale
- Kakai GR, Pelz D (2010) Asymptotic error rate of linear quadratic and logistic rules in multi-group discriminant analysis. *Int J Appl Stat Math* 18(10):70–80
- Krzanowski WJ, Hand DJ (1977) Assessing error rate estimators: the leave-one-out reconsidered. *Aust J Stat* 39(1):35–46
- Lachenbruch PA, Clarke W, Brot B, Lin L (1977) The effect of non-normality on the quadratic discrimination function. *MEDINFO 77*, Shives/Wolf. IFEP. North Holland Publishing Co, Amsterdam, pp 101–104
- Lachenbruch P, Goldstein M (1979) Discriminant analysis. *Biometrics* 35:69–85
- MacFarland R (2001) Exact misclassification probabilities for plug-in normal quadratic discriminant functions: the equal mean case. *J Multivar Anal* 77:21–53
- McLachlan GJ (1992) *Discriminant analysis and statistical pattern recognition*. Wiley, New Jersey
- Murray GD (1977) A cautionary note on selection of variables in discriminant analysis. *Appl Stat* 26(3):246–250

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)