Afrika Statistika

ISSN 2316-090X

Afrika Statistika
Vol. 11(1), 2016, pages 943–953.
DOI: http://dx.doi.org/10.16929/as/2016.943.85

On misclassification probabilities of linear and quadratic classifiers

Olusola Samuel Makinde

Department of Statistics, Federal University of Technology, Akure, Nigeria

Received March 17, 2016; Accepted May 20, 2016

Abstract. We study the theoretical misclassification probability of linear and quadratic classifiers and examine the performance of these classifiers under distributional variations in theory and using simulation. We derive expression for Bayes errors for some competing distributions from the same family under location shift.

Résumé. Nous étudions les probabilités théoriques de malclassification de méthodes de classification linéaires et quadratiques. Ensuite, nous examinons les performance de ces méthodes de classifications selon différentes distributions, sur le plan théorique en les étayant avec des études de simulations. Nous exprimons l'erreur de Bayes pour des distributions de même famille en compétition selon le changement du paramètre de centralisation.

Key words: Misclassification probability; linear classifier; quadratic classifier; Bayes risk.
AMS 2010 Mathematics Subject Classification : 62H30; 60E05.

1. Introduction

Classification is aimed at getting maximum information about separability or distinction among classes or populations and then assigns each observation to one of these populations on the basis of a vector of measurements or features. It has many important applications in different fields, such as disease diagnosis in medical sciences, risk identification in finance, admission of prospective students into university based on a battery of tests among others. Anderson (1984) described classification problem as the problem of statistical decision making. A good classification procedure is the one that classifies observations from unknown populations correctly. Suppose competing populations have well defined distributions which are characterised by some location and scale parameters. Classification of observations to any of these populations can be viewed from this characterisation in terms of shift in location

Corresponding author Olusola Samuel Makinde : osmakinde@futa.edu.ng

and scale of each of the population distributions. Competing populations may have either location shift, scale shift or both (location-scale shift). Consider populations $\pi_j, j = 1, 2, \dots, J$ from multivariate distributions, F_j having probability density functions f_j with prior probabilities p_j. Bayes rule, proposed in Welch (1939), is to assign each observation $\mathbf{x}$ to population π_j, whose posterior probability $P(\pi_j|\mathbf{x})$ is the highest. It assigns $\mathbf{x}$ to population π_1, in a two class problem, if

$$\frac{f_1(\mathbf{x})p_1}{f_2(\mathbf{x})p_2} > 1$$

and to π_2 otherwise. Wald (1944) argued that if each population has a cost, $C(i|j)$ associated with misclassifying $\mathbf{x}$ whose true population is π_j into π_i, then assign observations to the class or population that has the highest expected cost of misclassification. Welch (1939) showed that for any two normally distributed populations, the ratio of log likelihood functions of the two populations is the theoretical basis for building discriminant function that best classifies new individuals to any of the two populations given that the prior probabilities of the populations are known.

For two competing populations whose principal difference is in location, Fisher (1936) described the separation between these two populations to be ratio of variance between the populations to variance within the populations. This postulation leads to discriminant analysis, called Fisher's discriminant analysis. Suppose there are two populations from the same family of multivariate distributions to which observations can be classified. If these populations are normally distributed and have the same covariance matrix, the discriminant analysis is referred to as linear discriminant analysis (LDA). Similarly, if these populations are normally distributed but have different covariance matrices, the optimal rule is nonlinear and referred to as quadratic discriminant analysis (QDA). Welch (1939) and Wald (1944) showed that linear discriminant function has optimal properties for two group classification if the populations are multivariate normally distributed. Krzanowski (1977) reviewed the performance of Fisher's linear discriminant function when underlying assumptions are violated. Many classification methods, both parametric and non-parametric, have been compared with LDA and QDA under normality and non-normality which include Ghosh and Chaudhuri (2005), Kim et al. (2011) and Li et al. (2012) among others.

One way of evaluating the performance of a classification rule is to calculate its misclassification probabilities. One can define the *total probability of misclassification* (Δ) as

$$\Delta = p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x},$$

where

$$R_1 = \left\{\mathbf{x} \in \mathbb{R}^d : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)p_2}{c(2|1)p_1}\right\} \text{ and } R_2 = \left\{\mathbf{x} \in \mathbb{R}^d : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)p_2}{c(2|1)p_1}\right\}.$$

The classification regions R_1 and R_2 can be constructed only when the distributions F and G are fully known (Makinde and Chakraborty, 2015). This will rarely be the case, we have to work with the empirical versions of the classification regions and calculate the error rates. In this paper, we study the misclassification probabilities of linear and quadratic classifiers with emphasis to multivariate normal distributions and deduce the mathematical expression for Bayes error for some multivariate distributions under suitable conditions.

2. Classification Rules Based on Normality

We consider J populations having density function of the form

$$f_j(\mathbf{x}) = g((\mathbf{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)), \quad \mathbf{x} \in \mathbb{R}^d,$$

$j = 1, \ldots, J$, for some strictly decreasing, continuous, non-negative scalar function g, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are mean vector and covariance matrix of jth population respectively. Assuming normality, equal prior probabilities and equal cost of misclassification for the J populations, Bayes rule can be defined as

$$\text{assign } \mathbf{x} \text{ to } \pi_k \text{ if } D_k(\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \min_{1\leq j\leq J} D_j(\mathbf{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{1}$$

where $D_j(\mathbf{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\mathbf{x} - \boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) + \log|\boldsymbol{\Sigma}_j|$. This classification rule is known as quadratic discriminant analysis(QDA). It is linear if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_J$. Define R_k, the region of classification into kth population, as

$$R_k = \{\mathbf{X} : D_k(\mathbf{X}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \min_{1\leq j\leq J} D_j(\mathbf{X}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$$

Then the probability of misclassification of the optimal rule in (1) is

$$\Delta = \sum_{j=1}^{J} p_j P\left(\mathbf{x} \notin R_j | \mathbf{x} \in \pi_j\right). \tag{2}$$

A good classification method is the one that minimises Δ.
Let us consider two classes for simplicity. Suppose $\pi_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. The classification in (1) can be expressed as

$$\text{assign } \mathbf{x} \text{ to } \pi_1 \text{ if } (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 - 2\mathbf{x}) > 0 \tag{3}$$

if For $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. Observe that if $\boldsymbol{\Sigma}$ is a constant multiple of identity matrix, (3) is the generalisation of component-wise centroid classifier in (see, Hall et al., 2009, page 1598). Then, the probability of misclassification of $\mathbf{x}$ into either π_1 or π_2 is $\Delta = \Phi\left(\frac{-c_0}{2}\right)$, where $c_0^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and Φ is the cumulative distribution function of the standard normal distribution. See Johnson and Wichern (2007) for further discussion.

To illustrate the probability of misclassification, let π_1 and π_2 be two d-variate normal populations with mean vector and covariance matrix, $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ respectively. Assume that the prior probabilities of π_1 and π_2 are equal. Consider $\boldsymbol{\mu}_1^\top = \big(0, 0\big)$, $\boldsymbol{\mu}_2^\top = \big(\delta, 0\big)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2$. The total probability of misclassification associated with LDA is a function of non-centrality parameter δ and is obtained as

$$\Delta = \Phi\left(-\frac{\delta}{2}\right).$$

When covariance matrix of a population is a scalar multiple of the other. The following results hold:

Theorem 1. *Let F and G be two competing distributions with prior probabilities p_1 and p_2 respectively. Suppose $F \equiv N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $G \equiv N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Take $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \big(0, 0\big)^\top$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_2, \boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_2$ for $\sigma \neq 1$ and $p_1 = p_2 = 0.5$. Then*

1. *for* $\sigma^2 > 1$
$$\Delta = \frac{1}{2}\left[1 - F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right) + F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right)\right]$$
where $F_2(.)$ *denotes distribution function of central Chi-square distribution with 2 degrees of freedom.*
2. *for* $0 < \sigma^2 < 1$
$$\Delta = \frac{1}{2}\left[1 + F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right) - F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right)\right]$$

Proof of Theorem 1: For $\mathbf{x} \in \mathbb{R}^d$, if $\mathbf{x} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathbf{x}^\top\mathbf{x} \sim \chi^2_d$ and if $\mathbf{x} \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $\mathbf{x}^\top\mathbf{x} \sim \sigma^2\chi^2_d$. $f_1(\mathbf{x})/f_2(\mathbf{x}) \geq 1$ implies $(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \leq \log_e|\boldsymbol{\Sigma}_2| - \log_e|\boldsymbol{\Sigma}_1|$. This gives $\left(\frac{\sigma^2-1}{\sigma^2}\right)\mathbf{x}^\top\mathbf{x} \leq 2\log_e(\sigma^2)$. For $\sigma^2 > 0$, we consider two cases. These are $\sigma^2 > 1$ and $\sigma^2 < 1$.

1. When $\sigma^2 > 1$, the region of classification is
$$R_1 : \mathbf{x}^\top\mathbf{x} \leq \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \text{ and } R_2 : \mathbf{x}^\top\mathbf{x} > \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2.$$
Define $p_1P(2|1)$ as probability that $\mathbf{x}$ comes from population π_1 but eventually falls in the region of classification into population π_2 and $p_2P(1|2)$ as probability that $\mathbf{x}$ comes from population π_2 but eventually falls in the region of classification into population π_1. Then
$$P(2|1) = P\left(\mathbf{x}^\top\mathbf{x} > \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \middle| \mathbf{x}^\top\mathbf{x} \sim \chi^2_2\right) = 1 - F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right),$$
$$P(1|2) = P\left(\mathbf{x}^\top\mathbf{x} \leq \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \middle| \mathbf{x}^\top\mathbf{x} \sim \sigma^2\chi^2_2\right) = F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right)$$
and Δ, probability of misclassification is
$$\Delta = \frac{1}{2}\left[1 - F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right) + F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right)\right].$$
2. When $\sigma^2 < 1$, the region of classification is
$$R_1 : \mathbf{x}^\top\mathbf{x} \geq \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \text{ and } R_2 : \mathbf{x}^\top\mathbf{x} < \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2.$$
$$P(2|1) = P\left(\mathbf{x}^\top\mathbf{x} < \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \middle| \mathbf{x}^\top\mathbf{x} \sim \chi^2_2\right) = F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right)$$
$$P(1|2) = P\left(\mathbf{x}^\top\mathbf{x} \geq \frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2 \middle| \mathbf{x}^\top\mathbf{x} \sim \sigma^2\chi^2_2\right) = 1 - F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right).$$
The probability of misclassification is
$$\Delta = \frac{1}{2}\left[1 + F_2\left(\frac{2\sigma^2}{\sigma^2-1}\log_e \sigma^2\right) - F_2\left(\frac{2}{\sigma^2-1}\log_e \sigma^2\right)\right].$$

□

The results in Theorem 1 are compared with empirical results based on simulation. The procedure for the simulation follows from Section 3. The mean vectors and covariance matrices of competing distributions are taken to be $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \left(0, 0\right)^\top$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_2, \boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_2$ for $\sigma \neq 1$. The numerical results are presented in Figure 1(b).

Theorem 2. *Suppose the conditions of Theorem 1 hold and take* $\boldsymbol{\mu}_1 = \begin{pmatrix}0\\0\end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix}\delta\\0\end{pmatrix}$ *and* $\boldsymbol{\Sigma}_1 = \mathbf{I}_2, \boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_2$, *then*

$$\Delta = \begin{cases} p_1 P(\chi^2_{f_1} > \frac{k}{c_1}) + p_2 P(\chi^2_{f_2} < \frac{k}{c_2}), & \textit{for } \sigma^2 > 1 \\ p_1 P(\chi^2_{f_1} < -\frac{k}{c_1}) + p_2 P(\chi^2_{f_2} > -\frac{k}{c_2}), & \textit{for } \sigma^2 < 1 \end{cases}$$

where

$$k = \log_e \sigma^2 + \frac{1}{4}\frac{\delta^2(\sigma^2+1)}{\sigma^2-1}, \quad c_i = \frac{\overline{\sigma}_i^2}{\overline{\mu}_i}, \quad f_i = \frac{\overline{\mu}_i^2}{c_i}, \quad i = 1, 2,$$

$$\boldsymbol{\Sigma} = \mathbf{I}_2 + \sigma^2\mathbf{I}_2, \quad \mathbf{A} = \mathbf{I}_2, \quad \boldsymbol{\upsilon} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

$$U^2 = \boldsymbol{\upsilon}^\top\boldsymbol{\Sigma}\boldsymbol{\upsilon} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \delta^2,$$

$$\overline{\mu}_1 = \frac{1}{2\sigma^2}\left\{\frac{\frac{1}{2}[1+\sigma^2]\delta^2}{|\sigma^2-1|} + 2|\sigma^2-1|\right\}, \quad \overline{\mu}_2 = \frac{1}{2}\left\{\frac{\frac{1}{2}\sigma^2[1+\sigma^2]\delta^2}{|\sigma^2-1|} + 2|\sigma^2-1|\right\},$$

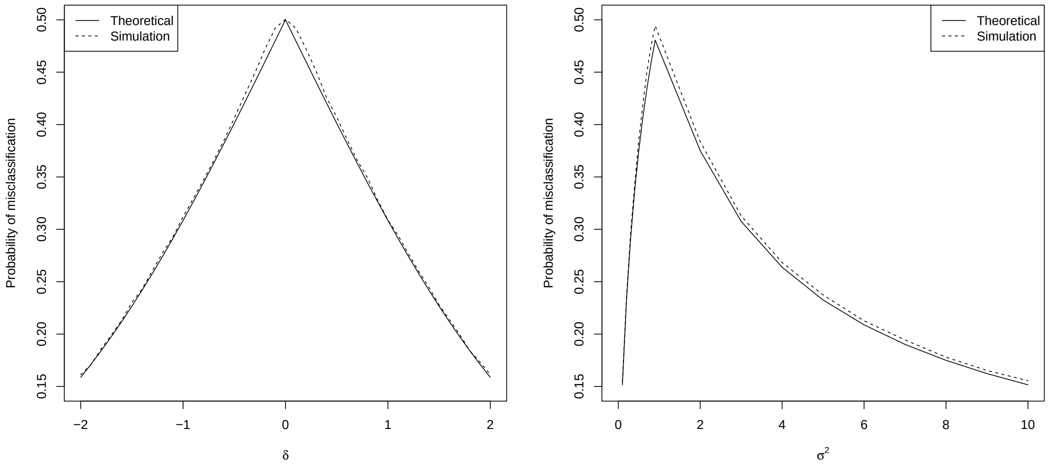
$$\overline{\sigma}_1^2 = \frac{1}{\sigma^2}\left\{\frac{1}{2}[1+\sigma^2]\delta^2 + (\sigma^2-1)^2\right\}, \quad \overline{\sigma}_2^2 = \frac{1}{2}\sigma^2[1+\sigma^2]\delta^2 + (\sigma^2-1)^2.$$

The proof of Theorem 2 follows from Gilbert (1969).

3. Numerical Examples

As illustration of actual error rates of LDA and QDA, we present a small simulation study. Let us consider the two populations π_1 and π_2 to be bivariate spherically symmetric with centre of symmetries $\boldsymbol{\mu}_1 = (0, 0)^\top$ and $\boldsymbol{\mu}_2 = (\delta, 0)^\top$, respectively. The sample sizes for $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from π_1 and $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ from π_2 are taken to be $n = m = 100$. We simulate a new random sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$ from π_1 and $\mathbf{Z}_{m+1}, \ldots, \mathbf{Z}_{2m}$ from π_2 with $m = 100$ and estimate the actual error rates by the proportion of misclassification in $\mathbf{Z}_1, \ldots, \mathbf{Z}_{2m}$. The simulation size is 1000.

Figure 1 presents the comparison between results from theory and simulation based on information above for misclassification probabilities of linear and quadratic classifiers under location shift and scale shift, as described in Section 2. It is clearly seen that the sample estimate of probability of misclassification is a good approximation for the population version of it. As expected, the error rate is nearly 0.5 when $\delta = 0$ and it decreases as δ goes away from 0 and the separation between the populations increases for location shift case. Consider the nonlinear case where $\boldsymbol{\mu}_1 = (0, 0)^T$ and $\boldsymbol{\mu}_2 = (\delta, 0)^T$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_2$ and $\boldsymbol{\Sigma}_2 = \sigma^2\mathbf{I}_2$. Figure 2 shows that error rate is affected by the difference in location and scale parameters of competing populations. The error rate is smaller with σ^{-2} than σ^2, given that $\sigma^2 > 1$. It can be ascertained from figure 2 that the error rate for σ^2 and σ^{-2} at the medians of symmetric distributions are the same. Also, misclassification rate decreases as σ^2 goes farther from 1.



(a) LDA

(b) QDA

Fig. 1: Misclassification probability: Theoretical versus Simulation.

Theoretically, LDA and QDA can not used for discriminating between distributions whose first and second moments do not exist. Furthermore, the presence of an outlying training sample point will affect the performance of LDA and QDA. Hence, both linear and quadratic classifiers are not robust against outliers and extreme values. However, Hubert and Van Driessen (2004) proposed replacing the estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ in Equation(1) by reweighted minimum covariance determinant (MCD) estimator of multivariate location and scatter based on FAST-MCD algorithm of Rousseeuw and Van Driessen (1999).

4. Theoretical Bayes Risk - Location Shift

We want to derive misclassification probability associated with Bayes rule for some competing distributions with location shift in a two-class problem. The distributions are multivariate t distribution with k degree of freedom and multivariate Laplace distribution. For multivariate normal distributions, the probability of misclassification associated with Bayes rule is discussed in Section 2 above.

Multivariate t distributions

Let $\mathbf{Z} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ and $U \sim \chi^2_k$ be independent, where k is the degree of freedom of Chi-squared distribution. Define $\mathbf{X} = \left(\mathbf{Z}\sqrt{\frac{k}{U}}\right) + \boldsymbol{\mu}$. The distribution of $\mathbf{X}$ is multivariate t distribution with k degree of freedom, denoted by $t(k, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The probability density function

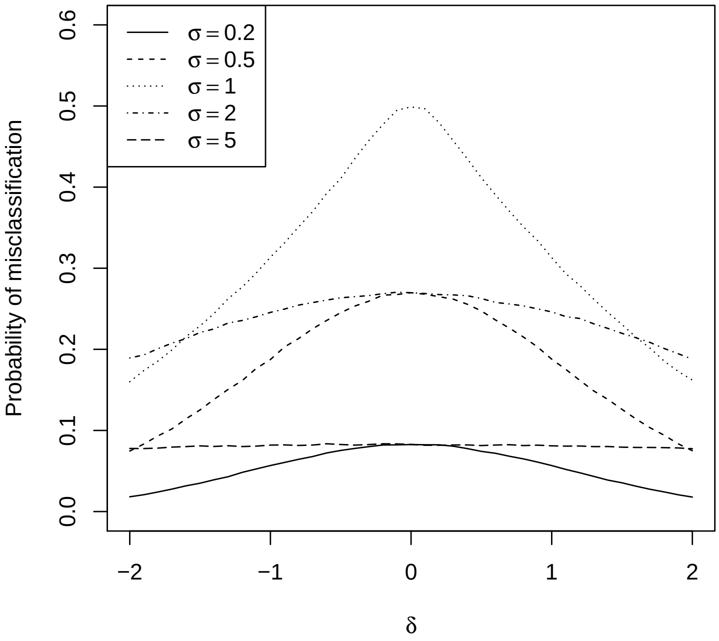


Fig. 2: Plot of error rate for location-scale shift problem using QDA.

of $\mathbf{x}$ is

$$f(\mathbf{x}) = (k\pi)^{-\frac{d}{2}} \frac{\Gamma(\frac{k+d}{2})}{\Gamma(\frac{k}{2})} |\mathbf{\Sigma}|^{-\frac{1}{2}} \{1 + \frac{1}{k}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\}^{-(\frac{k+d}{2})}. \quad (4)$$

Suppose π_1 has distribution $t(k, \boldsymbol{\mu}_1, \mathbf{\Sigma})$ with probability density function $f_1(\mathbf{x})$ and π_2 has distribution $t(k, \boldsymbol{\mu}_2, \mathbf{\Sigma})$ with probability density function $f_2(\mathbf{x})$. Under the assumption of equal prior probabilities, Bayes rule is to assign $\mathbf{x}$ to π_1 if $f_1(\mathbf{x}) > f_2(\mathbf{x})$, which is equivalent to

$$\text{assign } \mathbf{x} \text{ to } \pi_1 \text{ if } (\mathbf{x} - \boldsymbol{\mu}_1)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) < (\mathbf{x} - \boldsymbol{\mu}_2)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2).$$

This holds since the competing distributions have the same degree of freedom. The expression above is equivalent to $-2\mathbf{x}^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0$ and can also be written as

$$\left(\mathbf{z}\sqrt{\frac{k}{u}} + \boldsymbol{\mu}\right)^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0,$$

where $\mathbf{x} = \mathbf{z}\sqrt{\frac{k}{u}} + \boldsymbol{\mu}$ and $\mathbf{z}$ is distributed as $N_d(\mathbf{0}, \boldsymbol{\Sigma})$. If $\mathbf{x}$ is from π_1, $\boldsymbol{\mu} = \boldsymbol{\mu}_1$. Similarly, if $\mathbf{x}$ is from π_2, $\boldsymbol{\mu} = \boldsymbol{\mu}_2$. It follows that

$$\begin{aligned} P(2|1) &= P\left[\sqrt{\frac{k}{u}}\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0\right] \\ &= P\left[\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \frac{-1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]. \end{aligned}$$

This holds because u takes values in $[0, \infty)$. For either of the population, $E\big(\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big) = \mathbf{0}$ and $\text{var}\big(\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, then

$$\begin{aligned} P(2|1) &= P\left[R < \frac{\frac{-1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\big((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big)^{1/2}}\right] \\ &= P\left[R < \frac{-1}{2}\sqrt{\frac{u}{k}}c_0\right] = \int \Phi(c_1) f_u(u)du \end{aligned}$$

where R is a standard normal random variable defined as

$$R = \frac{\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - E[\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]}{\big(\text{var}(\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))\big)^{1/2}},$$

Φ is the distribution function of the standard normal distribution, f_u is probability density function of χ^2_k, u is Chi-squared distributed random variable,

$$c_1 = \frac{-1}{2}\sqrt{\frac{u}{k}}c_0, \quad c_2 = \frac{1}{2}\sqrt{\frac{u}{k}}c_0.$$

Similarly,

$$\begin{aligned} P(1|2) &= P\left[\sqrt{\frac{k}{u}}\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0\right] \\ &= P\left[\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right] \\ &= P\left[R > \frac{\frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\big((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big)^{1/2}}\right] = 1 - P\left[R < \frac{\frac{1}{2}\sqrt{\frac{u}{k}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\big((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\big)^{1/2}}\right] \\ &= 1 - P\left[R < \frac{1}{2}\sqrt{\frac{u}{k}}c_0\right] = 1 - \int \Phi(c_2) f_u(u)du \end{aligned}$$

The probability of misclassification associated with Bayes rule, denoted by Δ_B, is

$$\Delta_B = p_1 P(2|1) + p_2 P(1|2) = p_1 \int \Phi(c_1) f_u(u)du + p_2\left(1 - \int \Phi(c_2) f_u(u)du\right).$$

Multivariate Laplace distributions

Suppose the distribution of $\mathbf{X} \in \mathbb{R}^d$ is multivariate Laplace distribution $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean vector and covariance matrix of the distribution respectively. The probability density function of $\mathbf{x}$ is of the form

$$f(\mathbf{x}) \propto e^{-\sqrt{(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}}.$$

Without loss of generality, let $d = 2$, $r \sim \text{Gamma}(d)$, $\theta \sim \text{Uniform}(0, 2\pi)$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$. Define

$$Z_1 = r\cos\theta, \quad Z_2 = r\sin\theta, \quad \mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}.$$

Then, $\mathbf{X} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}$ has bivariate Laplace distribution $BL(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean and covariance of the distribution respectively. It follows that $\mathbf{X} - \boldsymbol{\mu} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z}$. Suppose populations π_1 and π_2 have distribution functions $BL(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $BL(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ respectively. If $\mathbf{x} \in \pi_1$, then $\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}_1)$, $\sqrt{(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)} = \sqrt{\mathbf{z}^\top \mathbf{z}} = r \sim \text{Gamma}(d)$ and $(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \neq r^2$ except $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, where $d = 2$. Similarly, if $\mathbf{x} \in \pi_2$, then $\mathbf{Z} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}_2)$, $\sqrt{(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)} = \sqrt{\mathbf{z}^\top \mathbf{z}} = r \sim \text{Gamma}(d)$ and $(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) \neq r^2$ except $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. It follows that

$$\log\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) = -\sqrt{(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)} + \sqrt{(\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)}.$$

Assuming equal prior probabilities, and for $\mathbf{x}$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2 \in \mathbb{R}^d$ and $d \geq 2$, the separating hyperplane between π_1 and π_2 can be written as

$$(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) = (\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2).$$

This is equivalent to

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

It follows that if $\mathbf{x}$ is distributed as population π_1, $\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ implies $(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and can be written as

$$\mathbf{z}^\top \mathbf{a} = -\frac{1}{2}\mathbf{a}^\top \mathbf{a},$$

where $\mathbf{a} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\mathbf{z}$ is a standard multivariate Laplace distributed random variable. Kotz et al. (2001) has shown that linear combination of standard multivariate Laplace random variables has a univariate symmetric Laplace distribution $\mathcal{L}(0, \sigma_l)$ (See Proposition 5.1.1 in pp. 232). That is, $w = \mathbf{a}^\top \mathbf{z}$ has a univariate Laplace distribution with mean 0 and variance σ_l, where $\sigma_l = \sqrt{\text{var}(\mathbf{a}^\top \mathbf{z})}$ and $\mathbf{a}$ is a vector of constant real numbers. Similarly, if $\mathbf{x}$ is distributed as population π_2, the separating hyperplane remains unchanged.

Observe that $f_1(\mathbf{x}) > f_2(\mathbf{x})$ implies $(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) < (\mathbf{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)$ and $\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) > \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$. It follows that

$$\begin{aligned}
&P(2|1) = P\big(\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) < \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)|\mathbf{x}\in\pi_1\big)\\
&= P\big(\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) < \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)\\
&= P\big(\mathbf{z}^\top\mathbf{a} < -\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big) = P\big(w < -\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\mu_2)\big)\\
&= F\Big(-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big),
\end{aligned}$$

where F is the distribution function of 1-dimensional symmetric Laplace distribution $\mathcal{L}\big(0, c_0\big)$ with $c_0^2 > 0$. Similarly,

$$\begin{aligned}
&P(1|2) = P\big(\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) > \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)|\mathbf{x}\in\pi_2\big)\\
&= P\big(\mathbf{x}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) > \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)\\
&= P\big(\mathbf{z}^\top\mathbf{a} > \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big) = P\big(w > \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)\\
&= 1 - F\Big(\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big)
\end{aligned}$$

where F is as defined above. The Bayes probability of misclassifying of $\mathbf{x}$ into either π_1 or π_2, denoted by Δ_B, is

$$\begin{aligned}
\Delta_B &= p_1P(2|1) + p_2P(1|2)\\
&= p_1F\Big(-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big) + p_2\Big[1 - F\Big(\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big)\Big]
\end{aligned}$$

where $p_1 + p_2 = 1$. Suppose G is a Laplace distribution function which is symmetric about c, then $G(-c) = 1 - G(c)$ for all $c \in \mathbb{R}$. Hence

$$\begin{aligned}
\Delta_B &= p_1F\Big(-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big) + p_2\Big[F\Big(-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big)\Big]\\
&= F\Big(-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big)
\end{aligned}$$

5. Concluding Remarks

In this paper, we have considered probabilities of misclassification between two populations only. However, these can be extended to more than two populations easily. The optimal performance of linear and quadratic discriminant functions are investigated and provide solutions of some theoretical examples. The theoretical probabilities of misclassification are compared with empirical error rates based on simulation, when competing populations differ in location and scale. The sample estimates of probability of misclassification associated with LDA and QDA are good approximation for their respective population versions. We derive expressions for Bayes error for multivariate Laplace distributions and multivariate t distributions with the same degree of freedom, under location shift.

References

Anderson, T.W., 1984. *An introduction to multivariate statistical analysis.* John Wiley & Sons, Inc, New York.

Chang, P.C. and Afifi, A.A., 2008. Classification based on dichotomous and continuous variables. *JASA*, **69**, 336-339.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. *Ann. Eug.*, **7**, 179-188.

Ghosh, A.K. and Chaudhuri, P., 2005. On maximum depth and related classifiers. *Scand. J. of Stat.*, **32**, 327–350.

Gilbert, E.S., 1969. The effect of unequal variance-covariance matrices on Fisher's linear discriminant function, *Biometrics*, **25(3)**, 505-515.

Hall, P., Titterington, D.M. and Xue, J., 2009. Median Based classifiers for High Dimensional Data. *Journal of the American Statistical Association*, **104**(488), 1597-1608.

Hubert, M. and Van Driessen, K., 2004. Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, **45**, 301-320.

Johnson, R.A. and Wichern, D.W., 2007. *Applied multivariate statistical analysis.* Sixth edition, Pearson Prentice Hall inc. New Jersey.

Kim, K.S., Choi, H.H., Moon, C.S. and Mun, C.W., 2011. Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Curr. Appl. Phy.*, **11**, 740-745.

Kotz, S., Kozubowski, T. and Podgorski, K., 2001. *The Laplace distribution and generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance.* Springer Science+Business Media, LLC.

Krzanowski, W.J., 1977. The performance of Fisher's linear discriminant function under non-optimal conditions, *Technometrics*, **19**(2), 191-200.

Li, J., Cuesta-Alberstos J.A. and Liu, R.Y., 2012. DD-Classifier: Nonparametric Classification Procedure Based on DD-plot, *JASA*, **107**, 737–753.

Makinde, O.S. and Chakraborty, B., 2015. On some nonparametric classifiers based on distribution functions of multivariate ranks. In Nordhausen, K and Taskinen, S.(eds): Modern Nonparametric, Robust and Multivariate Methods, Festschrift in Honour of Hannu Oja. Springer, 249-264.

Rousseeuw, P.J. and Van Driessen, K., 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, **41**, 212-223.

Wald, A., 1944. On a statistical problem arising in the classification of an individual into one of two groups. *Annals of Math. Stat.*, **15**, 145-162.

Welch, B.L., 1939. Note on discriminant functions. *Biometrika*, **31**, 218-220.