Scientific Research Publishing

Open Journal of Statistics, 2022, 12, 443-455
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Robust Classification through a Nonparametric Kernel Discriminant Analysis

Macdonald G. Obudho, George O. Orwa, Romanus O. Otieno, Festus A. Were

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: mobudho@knbs.or.ke, gorwa@buc.ac.ke, rodhiambo@must.ac.ke, Weref87@gmail.com

How to cite this paper: Obudho, M.G., Orwa, G.O., Otieno, R.O. and Were, F.A. (2022) Robust Classification through a Nonparametric Kernel Discriminant Analysis. *Open Journal of Statistics*, **12**, 443-455.
https://doi.org/10.4236/ojs.2022.124028

Received: June 7, 2022
Accepted: August 8, 2022
Published: August 11, 2022

Abstract

The problem of classification in situations where the assumption of normality in the data is violated, and there are non-linear clustered structures in the dataset is addressed. A robust nonparametric kernel discriminant classification function, which is able to address this challenge, has been developed and the misclassification rates computed for various bandwidth matrices. A comparison with existing parametric classification functions such as the linear discriminant and quadratic discriminant is conducted to evaluate the performance of this classification function using simulated datasets. The results presented in this paper show good performance in terms of misclassification rates for the kernel discriminant classifier when the correct bandwidth is selected as compared to other identified existing classifiers. In this regard, the study recommends the use of the proposed kernel discriminant classification rule when one wishes to classify units into one of several categories or population groups where parametric classifiers might not be applicable.

Keywords

Discriminant Analysis, Kernel Discriminant, Nonparametric

1. Introduction

Application of discriminant analysis has gained interest in various fields of social science, economics, education, finance and engineering. For instance, in routine banking or commercial finance, an officer or analyst may wish to classify loan applicants as low or high credit risks on the basis of the elements of certain accounting statements [1]. According to [2], the problem of discriminant analysis is one of assigning an unknown observation to a group with a low error rate. The function or functions used for the assignment may be identical to those used in the multivariate analysis of variance. Related to this [3], defined discriminant

analysis and classification as multivariate techniques concerned with separating distinct sets of objects or observations, and with allocating new objects (observations) to previously defined groups.

Therefore, the problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several categories or population groups on the basis of these measurements [4]. In all these problems it is assumed that there are two populations, say P_1 and P_2, one representing the population of individuals fit, and the other the population of individuals unfit for the purpose under consideration. The problem is that of classifying an individual into one of the populations P_1 and P_2 on the basis of test scores based on some statistical data from past experience.

To minimize the failures of the parametric techniques, this paper presents a Robust Nonparametric Kernel Discriminant function that is a better choice whenever a non-linear classification model is needed. This is because non-parametric estimators are more robust and are useful especially when there exists auxiliary information on finite population parameters which is often used to increase precision of estimators of the parameters [5].

2. Discriminant Functions and Classification

Suppose we have a set of v populations or groups that correspond to density functions $f_1, f_2, \cdots, f_v$. The intention is to assign all points x from the sample space to one of these groups or densities. We compare the weighted heights of the density functions to obtain the Bayes discriminant rule

$$x \text{ is allocated to group } j_0 \text{ if } j_0 = \arg\max_{j\in 1,\cdots,v} \pi_j f_j(x) \tag{1}$$

where π_j is the prior probability of drawing from density f_j. Enumerating for all x from the sample space, a partition $P = \{P_1, P_2, \cdots, P_v\}$ of the sample spaceis produced using

$$x \in P_j \text{ if } x \text{ is allocated to group } j$$

The discriminant rule defined in Equation (1), contains the unknown density functions and the (possibly) unknown prior probabilities. When data is collected, this abstract rule can be modified into a practical one.

The training data $X_j = \{X_{j1}, X_{j2}, \cdots, X_{jn_j}\}$, is collected which is drawn from f_j, for $j = 1, 2, \cdots, v$. (The sample sizes n_j are known and non-random).

A priori there is a class structure in the population since it's known which data points are drawn from which density function. From these training data, a practical discriminant rule and subsequent partition can be developed.

Using this discriminant rule/partition, the test data $Y_1, Y_2, \cdots, Y_m$, drawn from

$$f = T = \sum_{i\in s} y_i + \sum_{j=i}^{v} \pi_j f_j(x)$$

can be classified.

In this case, it's not clear which populations generated which data points.

The usual approach is to estimate these density functions and substitute into the discriminant rule. Parametric approaches that are well-known and widely used are linear and quadratic discriminant techniques. However, these suffer from the restrictive assumption of normality. With non-parametric discriminant analysis, this assumption can be relaxed and thus be able to tackle more complex cases. The study will focus on kernel methods for discriminant analysis. The monographs [6] [7] and [8] (Chapter 7) contain summaries of kernel discriminant analysis while [9] contains more detailed and lengthy expositions on this subject.

2.1. Parametric Discriminant Analysis

The two parametric methods that are reviewed in more detail here are the linear and quadratic discriminant analysis, being the most commonly used. Their ease of computation is a result from the underlying normality assumption, which does not necessarily hold for most datasets.

2.1.1. Linear Discriminants

Assume that the densities f_j are normal with different mean vectors μ_j and with common variance matrix Σ. The key assumption is that $f_j \sim N\left(\mu_j, \Sigma\right)$. The discriminant rule, Equation (1), reduces to (after taking logarithms of f_j)

$$x \text{ is allocated to group } j_0 \text{ if } j_0 = \arg\max_{\{j\in 1,\cdots,v\}} \log\left(\pi_j\right) - \frac{1}{2}\left(x-\mu_j\right)^{\mathrm{T}} \Sigma^{-1}\left(x-\mu_j\right) \quad (2)$$

From this equation, it can be observed that the resulting partition is obtained by intersections of ellipsoids with different centres and with the same orientation. This yields partition boundaries that are hyperplanes. Figure 1 is obtained using the sample mean $\overline{X}_j$ as estimate of μ_j and $S = \left(n-v\right)^{-1}\sum_{i\in s} y_i + \sum_{j=i}^{v} n_j S_j$ for Σ where S_j is the sample variance, forthe case of Linear Discriminant Rule.

2.1.2. Quadratic Discriminant Function

For Quadratic Discriminants, the densities are assumed to be normal with different

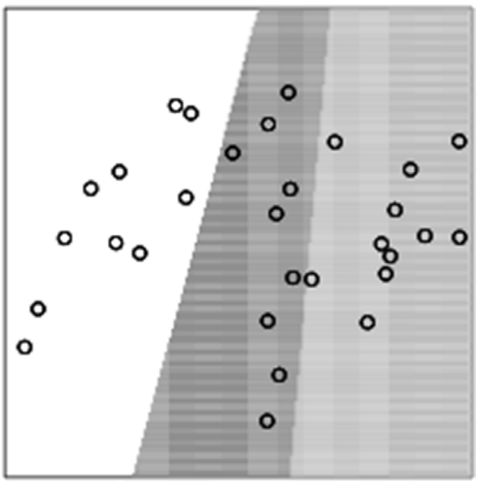

Figure 1. Partition from linear discriminant analysis.

means μ_j and different variances Σ_j. The assumption of common variance of linear discriminant analysis is relaxed. That is, $f_j \sim N\left(\mu_j, \Sigma\right)$. The discriminant rule, Equation (1), reduces to (after taking logarithms of f_j)

$$\begin{aligned} &x \text{ is allocated to group } j_0 \text{ if } j_0 \\ &= \arg \max_{\{j\in 1,\cdots,v\}} \log\left(\pi_j\right) - \frac{1}{2}\log\left|\Sigma_j\right| - \frac{1}{2}\left(x-\mu_j\right)^{\mathrm{T}} \Sigma^{-1}\left(x-\mu_j\right) \end{aligned} \tag{3}$$

This discriminant rule yields a partition defined by intersections of ellipsoids with differing centres and orientations. The boundaries are thus piecewise paraboloidal curves, as is illustrated in Figure 2, obtained by replacing the means and variances with their sample statistics.

To effectively use the parametric discriminant rules, one has to replace the unknown parameters with their usual sample estimates.

2.2. Proposed Kernel Discriminant Function

The parametric methods can be generalized to a non-parametric one in a straightforward way. Instead of assuming a normal (or any other parametric) form for the densities, we simply estimate the densities nonparametrically. In this study, the kernel density estimators constructed from the training data is used.

Kernel density estimation [6] [10], is a popular method for nonparametric density estimation, and it has one well known application in kernel discriminant analysis (KDA) [11]. In a J class classification problem, if we have a training sample $S = \left\{\left(x_i, c_i\right); x_i \in \mathbb{R}^d, C_i \in \left(1,2,\cdots,J\right), i = 1,2,\cdots,n\right\}$ of n observations, the kernel estimate for the density function $f_j\left(j = 1,2,\cdots,J\right)$ can be expressed as

$$\hat{f}_{jb}\left(x\right) = \frac{1}{nb^d} \sum_{i:c_i=j} K\left\{\frac{1}{b}\left(x - x_i\right)\right\} \tag{4}$$

where n_j is the number of observations from the *jth* class $\sum n_j = n$ K is a d-dimensional density function symmetric around 0, and b is the associated

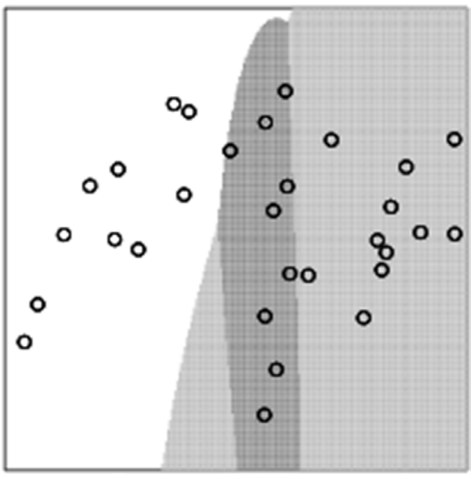

Figure 2. Partition from quadratic discriminant analysis.

smoothing parameter known as the bandwidth. These kernel density estimates are used to construct the kernel discriminant rule (*KDR*) given by

$$KDR: \text{is allocated to group } j_0 \text{ if } j_0 = \arg\max_{j\in\{1,\cdots,v\}} \hat{\pi}_j \hat{f}_j\left(x, B_j\right) \tag{5}$$

where $\hat{f}_j\left(x, B_j\right)$ is the kernel density estimate corresponding to the *jth* group and where π_j is the prior probability of the *jth* group. If these priors are not known, one usually estimates them using training sample proportions $\hat{\pi}_j = \frac{n_j}{n}, \left(j = 1, 2, \cdots, J\right)$ of different groups. Many choices for the kernel function K are available in the literature [6] [10]. Equation (5) forms our proposed classification rule.

To illustrate its implementation, the resulting partition is in **Figure 3** where the plug-in bandwidth selectors for B_j has been used.

Now that the kernel density estimators for discriminant analysis are being used, selection of appropriate bandwidths is crucial. On one hand, one can attempt to find optimal bandwidths for optimal individual kernel density estimates. On the other hand, optimal bandwidths which directly optimise the misclassification rate (*MR*), as [11] attempt for the two can be found.

2.3. Misclassification Rate (*MR*)

This rate is the proportion of points that are assigned to an incorrect group based on a discriminant rule. Then we have

$$\begin{aligned} 1 - MR &= P\left(Y \text{ is classified correctly}\right) \\ &= E_Y\left[1\left\{Y \text{ is classified correctly}\right\}\right] \\ &= E_X\left[E_Y\left[1\left\{Y \text{ is classified correctly}\right\}\right] \mid X_1, X_2, \cdots, X_v\right] \\ &= 1 - \frac{TP + TN}{TP + FP + TN + FN} \end{aligned} \tag{6}$$

where E_Y is expectation with respect to Y or $\sum_{j=1}^{v} \pi_j f_j$, and E_X is expectation with respect to $X_1, X_2, \cdots, X_v$ or $\pi_1 f_1, \pi_2 f_2, \pi_v f_v$.

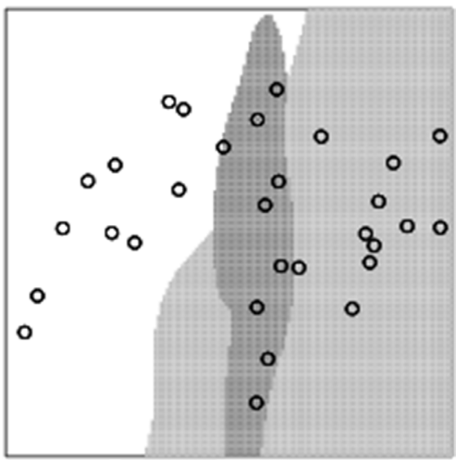

Figure 3. Partition from kernel discriminant analysis.

- True positive (*TP*): Observation is predicted positive and is actually positive.
- False positive (*FP*): Observation is predicted positive and is actually negative.
- True negative (*TN*): Observation is predicted negative and is actually negative.
- False negative (*FN*): Observation is predicted negative and is actually positive.

[9] recommends the former approach for three reasons. First, accurate estimates of the individual density functions are useful in their own right; second, accurate density estimates can be used in other, more complex discriminant problems which look at measures other than the misclassification rate; and third, direct optimisation with respect to a misclassification rate poses many difficult mathematical obstacles.

Whilst we will not use the misclassification rate to select bandwidths, we will still use it as our performance measure of a discriminant rule. So we need to estimate it. The most appropriate estimate depends on whether we have test data or not. If we do, as is the usual case for simulated data, then a simple estimate is obtained by counting the number of Y_j that are assigned to an incorrect group, divided by the total number of data points *m*. On the other hand, if we do not have test data, as is the usual case for real data, then we use the cross validation estimate of *MR*, as recommended by [6] and [9]. This involves leaving out each X_{ji}, constructing a corresponding leave-one-out density estimate and subsequent discriminant rule. We then compare the label assigned to X_{ji} based on the leave-one-out discriminant rule to its correct group label. These counts are then summed and divided by *n*.

2.4. Algorithm for Proposed Kernel Discriminant Analysis

The algorithm for the proposed kernel discriminant analysis is given below. The algorithms for linear and quadratic discriminant analysis are similar except that any kernel methods are replaced by the appropriate parametric methods. We put these algorithms into practice with both simulated and real data.

1) For each training sample $X_j = \left\{X_{j1}, X_{j2}, \cdots, X_{jn_j}\right\}$, $j = 1, 2, \cdots, v$, compute a kernel density estimate

$$\hat{f}\left(x; B_j\right) = n_j^{-1} \sum_{i=1}^{n_j} K_{B_j}\left(x - X_{ji}\right) \tag{7}$$

We can use any sensible bandwidth selector B_j.

2) If prior probabilities are available then use these. Otherwise estimate them using the training sample proportions $\hat{\pi}_j = n_j / n$.

3a) Allocate test data points $Y_1, Y_2, \cdots, Y_m$ according to KDR/Equation (5) or

3b) Allocate all points *x* from the sample space according to KDR/Equation (5).

4a) If we have test data then the estimate of the misclassification rate is

$$\hat{MR} = 1 - m^{-1} \sum_{k=1}^{v} 1\left\{Y_K \text{ is classified correctly using } KDR\right\} \tag{8}$$

4b) If we do not have test data the cross validation estimate of the misclassifi-

cation rate is

$$\hat{MR}_{CV} = 1 - n^{-1} \sum_{j=1}^{v} \sum_{i=1}^{n_j} 1\left\{ X_{ji} \text{ is classified correctly using } KDR_{ji} \right\} \tag{9}$$

where KDR_{ji} is similar to KDR except that $\hat{f}_j\left(.; B_j\right)$ and $\hat{\pi}_j$ are replaced by their leave one out estimates obtained by removing X_{ji} that is $\hat{\pi}_{ji} = \left(n_j - 1\right)/n$ and

$$\hat{f}_{j,-i}\left(x; B_j\right) = \left(n_j - 1\right)^{-1} \sum_{i'=1, i' \neq 1}^{n_j} K_{B_{j,-i}}\left(x - X_{ji'}\right) \tag{10}$$

That is, we repeat step 3 to classify all X_{ji} using KDR_{ji}.

2.5. Bandwidth Selection

The bandwidth of a kernel is a free parameter which exhibits a strong influence on the resulting estimate. Kernel smoothing requires the choice of a bandwidth parameter. This choice is critical, as under- or over-smoothing can substantially reduce precision. To illustrate its effect, we take a simulated random sample from a random sample of 100 points from a standard normal distribution as shown in Figure 4. The grey curve is the true density (a normal density with mean 0 and variance 1). In comparison, the red curve is under-smoothed since it contains too many spurious data artifacts arising from using a bandwidth b = 0.05, which is too small. The green curve is over-smoothed since using the bandwidth b = 2 obscures much of the underlying structure. The black curve with a bandwidth of b = 0.337 is considered to be optimally smoothed since its density estimate is close to the true density.

The most common optimality criterion used to select this parameter is the expected $L2$ risk function, also termed the mean integrated squared error:

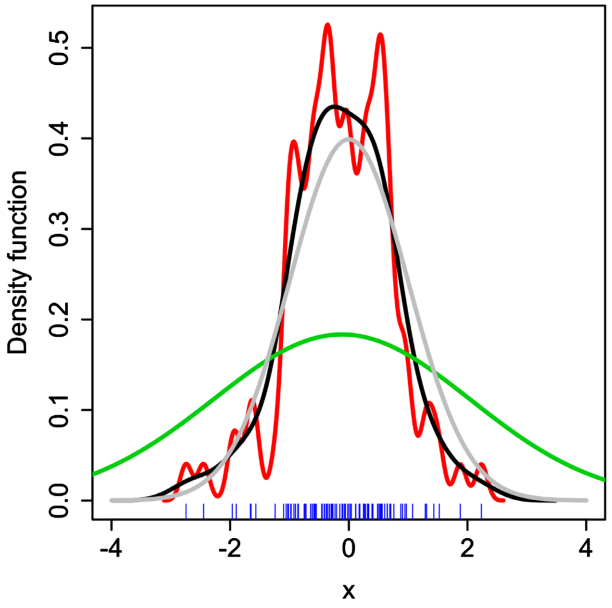

Figure 4. Kernel Density Estimate (KDE) with different bandwidths of a random sample.

$$MISE(b) = E\left[\int \left(f_b(x) - f(x)\right)^2 \mathrm{d}x\right]$$

Under weak assumptions on f and K, $MISE(b) = AMISE(b) + o\left(\frac{1}{nb} + b^4\right)$ where o is the little o notation. The $AMISE$ is the Asymptotic MISE which consists of the two leading terms

$$AMISE(b) = \frac{R(K)}{nb} + \frac{1}{4} m_2(K)^2 b^4 R(f'')$$

where $R(g) = \int g(x)^2$ a function g

$$m_2(K) = \int x^2 K(x) \mathrm{d}x$$

To be able to prove the theoretical results, the following assumptions are made; and f'' is the second derivative of f. The minimum of this $AMISE$ is the solution to this differential equation

$$\frac{\mathrm{d}}{\mathrm{d}b} AMISE(b) = -\frac{R(K)}{nb^2} + m_2(K)^2 b^3 R(f'') = 0$$

or

$$b_{AMISE} = \frac{R(K)^{\frac{1}{5}}}{m_2(K)^{\frac{2}{5}} R(f'')^{\frac{1}{5}} n^{\frac{1}{5}}}$$

Neither the $AMISE$ nor the b_{AMISE} formulas can be used directly since they involve the unknown density function f or its second derivative f'', so a variety of automatic, data-based methods have been developed for selecting the bandwidth. Many review studies have been carried out to compare their efficacies, b with the general consensus that the plug-in selectors and cross validation selectors are the most useful over a wide range of data sets.

Substituting any bandwidth b which has the same asymptotic order $n^{\frac{-1}{5}}$ as b_{AMISE} into the $AMISE$ gives that $AMISE(b) = O\left(n^{\frac{-4}{5}}\right)$, where O is the big O notation. It can be shown that, under weak assumptions, there cannot exist a non-parametric estimator that converges at a faster rate than the kernel estimator. Note that the $n^{\frac{-4}{5}}$ rate is slower than the typical $n-1$ convergence rate of parametric methods.

If the bandwidth is not held fixed, but is varied depending upon the location of either the estimate (balloon estimator) or the samples (pointwise estimator), this produces a particularly powerful method termed adaptive or variable bandwidth kernel density estimation.

Further, the bandwidths vary with the kernel function chosen. An optimal bandwidth of one kernel function cannot be regarded in the same way for another function. Because of this, many researchers have been carrying out studies aimed at determining techniques of obtaining bandwidths that minimize MSE or $AMSE$ functions that can be used with the different kernel functions.

Two methods, the "plug-in" and "cross validation" are the common ways in which this problem can be tackled. The plug-in method simply involves the replacement of the unknown functions in the expression of interest. The AMISE optimal bandwidth equation (13) depends on the unknown roughness R_1. A simple choice is a normal scale estimate. if $f = \phi_\sigma$

$$R_1 = \int_{-\infty}^{\infty} \left(\phi_\sigma^{(1)}(y)\right)^2 = \frac{1}{\sigma^3}\int_{-\infty}^{\infty} y^2 \phi(y)^2 = \frac{1}{\sigma^3 4\sqrt{\pi}}$$

Thus, a reference bandwidth is

$$\hat{b}_0 = \hat{\sigma}\left(4\sqrt{\sigma\psi}\right)^{\frac{1}{3}} n^{-\frac{1}{3}} \tag{11}$$

where $\hat{\sigma}$ is the sample standard deviation. In particular, for the normal kernel $K = \phi$ then $\hat{b}_1 = 1.59\hat{\sigma}n^{-\frac{1}{3}}$. The reference bandwidth, however, may work poorly for distributions which are far from the normal. As shown by Jones (1990) if $bn^{\frac{1}{2}} \to \infty$ as $n \to \infty$ then

$$AMISE(b) = \int_{-\infty}^{\infty} E\left(\hat{F}_b(y) - F(y)\right)^2 \mathrm{d}y = \frac{V}{n} - \frac{b\psi}{n} + \frac{b^4 R_1}{4} + O\left(b^4\right) \tag{12}$$

where, $V = \int_{-\infty}^{\infty} F(y)\left(1-F(y)\right)^2 \mathrm{d}y$, $\psi = 2\int_{-\infty}^{\infty} xK(x)k(x)\mathrm{d}x > 0$ is a constant which depends only on the kernel. For example, if $k(x) = \phi(x)$ then $\psi = \frac{1}{\sqrt{\pi}}$.

The *AMISE* is minimized by setting b equal to $b_0 = \left(\frac{\psi}{R_1}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}$. The optimal AMISE is

$$AMISE(b_0) = \frac{V}{n} - \frac{3\psi^{\frac{3}{8}}}{n^{\frac{3}{4}} R_1^{\frac{1}{8}}} \tag{13}$$

3. Empirical Study

Sometimes in survey sampling, we do not usually observe all the survey information. That is, the survey variable is not observable for all the population units. Auxiliary variable X is often used to estimate the unobserved survey variables. One way of overcoming the above problem is the super population approach in which the working model relating the auxiliary variables to the response variable is assumed.

We conduct a similar comparison to the simulation studies contained in [9], examining the performance of the following discriminant analysers:

1) Linear discriminant (LD).
2) Quadratic discriminant (QD).
3) Kernel discrinant with 2-stage AMSE diagonal bandwidth matrices (KDD2).
4) Kernel discrinant with 2-stage SAMSE full bandwidth matrices (KDS2).
5) Kernel discrinant with 1-stage SCV full bandwidth matrices (KDSC).

The R code for kernel discriminant analysers is based on the bandwidth matrix selection and density functions in the *ks* library. The R code for LDA and QDA are supplied within the MASS library in the R software by the function *lda*() and *qda*() respectively.

We simulate from the following normal mixture densities for 500 trials, using training sample sizes $n=100$ and $n=1000$ and test data sample size $k=1000$. The target densities D and E are used from previous studies except that we keep track of which mixture component an observation is drawn from. Density D contains fairly distinct components discriminant analyser are expected to perform well here. Density E has three components of various shapes and sizes, therefore, is more challenging case than density D. Density K is a pair of bimodal normal mixtures, with alternating modes. Density L is a large mode separating a bimodal density with narrower modes. For these two latter densities it is expected that, the linear and quadratic discriminant analysers to perform poorly since it is difficult to distinguish the different components using only linear or quadratic cuts. Alternatively, densities K and L are viewed as being highly non-normal so the assumptions of normality for the parametric methods are invalid. Thus it is expected that the kernel methods will demonstrate their efficiency here. The formulas for these target densities are in Table 1 and their corresponding contour plots are in Figure 5.

Table 1. Formulas for target densities D, E, K & L.

Target density	Formular
D	$\pi_1=\frac{1}{2}f_1\sim N\left(\begin{bmatrix}1\\-1\end{bmatrix},\begin{bmatrix}\frac{4}{9} & \frac{14}{45}\\ \frac{14}{45} & \frac{4}{9}\end{bmatrix}\right);\ \pi_2=\frac{1}{2},\ f_2=N\left(\begin{bmatrix}-1\\1\end{bmatrix},\begin{bmatrix}\frac{4}{9} & 0\\ 0 & \frac{4}{9}\end{bmatrix}\right)$
E	$\pi_1=\frac{3}{7}f_1\sim N\left(\begin{bmatrix}-1\\0\end{bmatrix},\begin{bmatrix}\frac{9}{25} & \frac{63}{250}\\ \frac{63}{250} & \frac{49}{100}\end{bmatrix}\right);\ \pi_2=\frac{3}{7},\ f_2=N\left(\begin{bmatrix}1\\\frac{2}{\sqrt{3}}\end{bmatrix},\begin{bmatrix}\frac{9}{25} & 0\\ 0 & \frac{49}{100}\end{bmatrix}\right)$ $\pi_3=\frac{1}{7}f_3\sim N\left(\begin{bmatrix}1\\-\frac{2}{\sqrt{3}}\end{bmatrix},\begin{bmatrix}\frac{9}{25} & 0\\ 0 & \frac{49}{100}\end{bmatrix}\right)$
K	$\pi_1=\frac{1}{2}f_1\sim\frac{1}{2}N\left(\begin{bmatrix}-\frac{3}{2}\\-\frac{3}{2}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right)+\frac{1}{2}N\left(\begin{bmatrix}\frac{1}{2}\\\frac{1}{2}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right);$ $\pi_2=\frac{1}{2}f_2\sim\frac{1}{2}N\left(\begin{bmatrix}\frac{3}{2}\\\frac{3}{2}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right)+\frac{1}{2}N\left(\begin{bmatrix}-\frac{1}{2}\\-\frac{1}{2}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right);$

Continued

L

$$\pi_1 = \frac{1}{3} f_1 \sim \frac{1}{2} N\left(\begin{bmatrix} -\frac{3}{2} \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{3}{10} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{10} \end{bmatrix}\right) + \frac{1}{2} N\left(\begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{3}{10} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{10} \end{bmatrix}\right);$$

$$\pi_2 = \frac{2}{3} f_2 \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{4}{5} & \frac{2}{5} \\ \frac{2}{5} & 1 \end{bmatrix}\right)$$

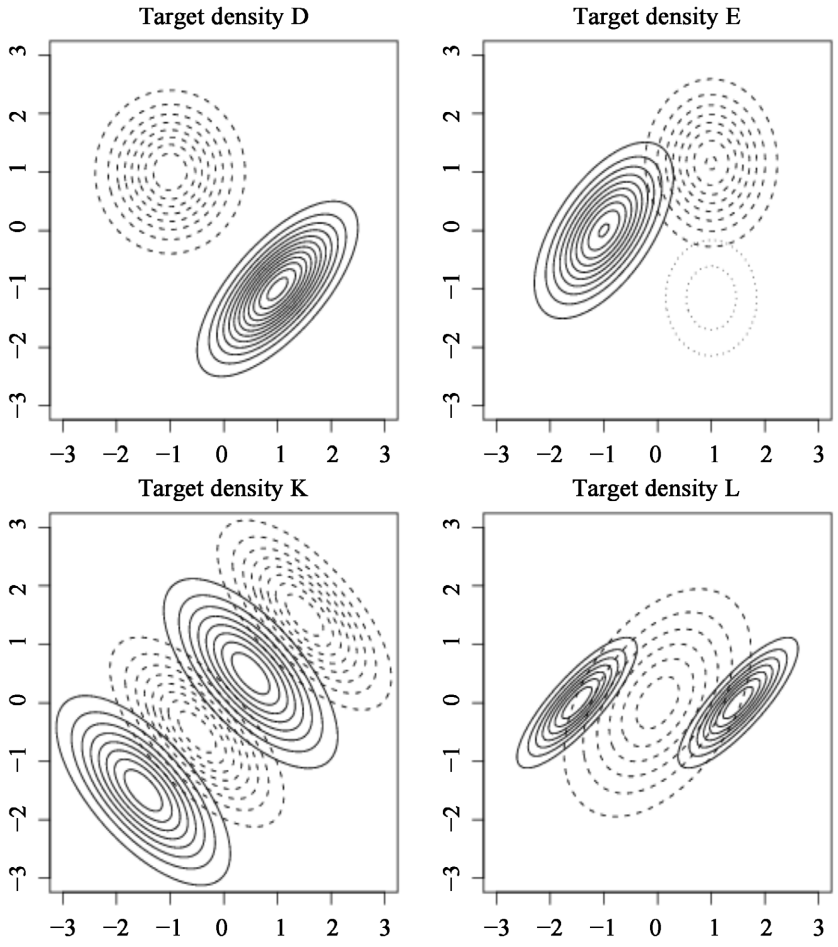

Figure 5. Contour plots for the target densities D, E, K, L for discriminant analysers: solid contours- $\pi_1 f_1$, dashed lines- $\pi_2 f_2$ and dotted lines- $\pi_3 f_3$.

Misclassification Rates Using Simulated Data

The average and standard deviation of misclassification rates are in **Table 2**. From this table, for density D and E, the LD performed poorly compared to QD and the kernel discriminant analysers. For density K, our expectations are confirmed: KDD2, KDS2, KDSC all outperform the linear and quadratic counterparts. For density L, the advantage of the kernel methods over the linear method is maintained while it is reduced compared to the quadratic method. The increased

Table 2. Misclassification rates for discriminant analysers.

Target density		Misclasifcation rate				
		KDD2	KDS2	KDSC	LD	QD
				$n = 100$, $k = 1000$		
D	mean	0.0793	0.00600	0.0050	0.0101	0.0050
	SD	0.0013	0.0013	0.0013	0.0042	0.0014
E	mean	0.00798	0.0719	0.0798	0.0703	0.0701
	SD	0.0120	0.0110	0.0071	0.0082	0.0081
K	mean	0.3008	0.2810	0.17989	0.6125	0.5998
	SD	0.0152	0.01320	0.0128	0.0396	0.0324
L	mean	0.1615	0.1517	0.1607	0.4010	0.2115
	SD	0.0156	0.0124	0.0146	0.0167	0.0189
				$n = 1000$, $k = 1000$		
D	mean	0.0048	0.0050	0.0044	0.0059	0.039
	SD	0.0015	0.0015	0.0013	0.0014	0.0016
E	mean	0.0499	0.0499	0.0489	0.0509	0.0498
	SD	0.00146	0.0014	0.0014	0.0016	0.0013
K	mean	0.0498	0.0501	0.0498	0.0498	0.0497
	SD	0.0015	0.0016	0.0015	0.0014	0.0013
L	mean	0.4998	0.4996	0.4500	0.5189	0.4984
	SD	0.0154	0.0155	0.0154	0.0121	0.0164

performance of the kernel discriminant analysers for the latter two densities is apparent for both sample sizes. Moreover, even with the increased burden of selecting an increased number of bandwidths which comprise the bandwidth matrix, the full matrix selectors overall produce smaller standard deviations.

The differences between the diagonal matrix KDD2 and the full matrix KDSC and KDS2 are more subtle than the differences between the kernel methods and the parametric methods. We can see that both full bandwidth matrix methods KDS2 and KDSC in the majority of cases considered here have lower mean misclassification rates than KDD2.

4. Conclusions and Recommendation

In this paper, a nonparametric Kernel discriminant classifier has been proposed and studied. The classification rates and classification algorithm has been developed and computed for the case of datasets through a simulation and the findings compared with those of existing classifiers such linear discriminant and quadratic discriminant classifier that have continued to be used in practice.

From the results, the following observations and conclusions have been made.

1) The kernel methods, with appropriately chosen bandwidth matrices, outperform the parametric methods; and the kernel methods with full bandwidth matrices outperform those with diagonal bandwidth matrices.

2) Both full bandwidth matrix methods KDS2 and KDSC in the majority of cases considered here have lower mean misclassification rates than KDD2.

3) The parametric discriminant classifiers perform poorly especially where the underlying structure of the model and the data do not obey the assumption of normality.

The main conclusion is therefore that the classification estimator based on the nonparametric kernel discriminant function has proved to yield results with great precision and therefore it is recommended for classification problems. The paper recommends other classification techniques which can handle the high dimensional spaces such Neural Networks to be considered in future studies so as to see if efficiency of classification can be improved.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Manly, M. and Van Waas, L. (2014) The State of Statelessness Research: A Human Rights Imperative. *Tilburg Law Review*, **19**, 3-10.

[2] Lachenbruch, P.A. (1974) Discriminant Analysis When the Initial Samples Are Misclassified II: Non-Random Misclassification Models. *Technometrics*, **16**, 419-424. https://doi.org/10.1080/00401706.1974.10489211

[3] Johnson, R.A., Wichern, D.W., *et al.* (2014) Applied Multivariate Statistical Analysis. Volume 6, Pearson, London.

[4] Flury, B. (2013) A First Course in Multivariate Statistics. Springer Science & Business Media, Dordrecht.

[5] Cochran, W.G. (2007) Sampling Techniques. John Wiley & Sons, Inc., New York.

[6] Silverman, B.W. (2018) Density Estimation for Statistics and Data Analysis. Routledge, New York. https://doi.org/10.1201/9781315140919

[7] Scott, D.W. (1991) Feasibility of Multivariate Density Estimates. *Biometrika*, **78**, 197-205. https://doi.org/10.1093/biomet/78.1.197

[8] Simonoff, J.S. (1996) Further Applications of Smoothing. In: Simonoff, J.S., Ed., *Smoothing Methods in Statistics*, Springer, New York, 252-274. https://doi.org/10.1007/978-1-4612-4026-6_7

[9] Hand, D.J. (1982) Kernel Discriminant Analysis. John Wiley & Sons, Inc., New York.

[10] Scott, D.W. (1992) Multivariate Density Estimation: Theory, Practice and Visualisation. John Willey and Sons, Inc., New York.

[11] Hall, P. and Wand, M.P. (1988) On Nonparametric Discrimination Using Density Differences. *Biometrika*, **75**, 541-547. https://doi.org/10.1093/biomet/75.3.541