

Research Article

Open Access

On Rank-Rank Classification Method for High Dimensional Data

Emmanuel Idowu Olamide* and Femi Afolayan Aluko

Department of Statistics, Federal University of Technology, P.M.B 704, Akure, Nigeria

ABSTRACT

Advances in technology have made available several high-dimensional data, with dimension greater than sample size. In multivariate analysis, data depth was defined and well discussed in literature. Recently, the notion of rank-rank (RR) plot was introduced for classification in low dimension based on the notions of data depth. However, the performance of RR classification method needs to be investigated in high dimension especially when dimension is much greater than sample size. One of the primary motivations for using spatial depth is its computational tractability, particularly when the dimension is high. However, spatial depth function does not account for correlation among variables. It has been proven that correlation values between variables in high dimensional data are very strong, therefore there is need for modification of spatial depth function which is implemented for three different known classification methods. The performance of spatial depth function and modified depth function were compared in depth-depth classification method, depth-oriented classification method and rank-rank classification method. The performance of the classifiers was evaluated through the estimates of mean proportion of correct classification using simulation studies and analysis of real data. The result in this study showed that rank-rank classification method outperformed other methods for high dimensional data.

*Corresponding author

Emmanuel Idowu Olamide, Department of Statistics, Federal University of Technology, P.M.B 704, Akure, Nigeria.

Received: September 25, 2026; **Accepted:** October 13, 2026; **Published:** October 19, 2026

Keywords: Rank-Rank Classification, Spatial Depth, Modified Depth, Depth-Oriented Ranking, Depth-Depth Classification, High Dimension

Introduction

Classification is one of the widely used practical tools in statistics due to the online information available these days. A good statistical classification aims at assigning some basic measurement on an unclassified observation to one of several groups or populations on the basis of vector of measurement on characteristics correctly. Anderson (1984) described classification problem as the problem of statistical decision making [1].

Advances in modern technology have made several high dimensional data available online these days. Dimension of each of such datasets is greater than the sample size. High dimensional data classification problems are available nowadays due to the technological innovations everywhere. It has some statistical challenges and makes many traditional classification algorithms methods impractical to use.

Data depth has been defined as alternative to some parametric ways in analyzing many multivariate data. Data depth is a quantitative measure which expresses how central (or deep) a point is with respect to its data cloud or distribution [2].

(2008) proposed maximum depth classifier based on modified projection depth. The classifier works well only when samples are multivariate normally distributed.

Makinde and Adewumi (2017) defined classification rule as a set of steps for assigning an unclassified observation to one of several

groups or populations on the basis of some measurement [3].

Makinde (2019) studied the notions of functional depth and proposed a classification method based on distribution functions of data depth for functional data. The performance of the classification method was examined by using simulations and real datasets, the results were compared with the results from existing methods [4].

Statistical depth functions are proving to be a promising tool for ordering multi-dimensional observations. The use of statistical depth functions in nonparametric inference for multivariate data have greatly increased. In spite of that, structural properties for functional data and methods of analyzing them may resemble those for conventional multivariate data such as linear and nonlinear regression models, principal components analysis, among others. The main idea of depth functions is to provide from the "deepest" point a "center-outward" ordering of multi-dimensional observations. Though linear discriminant analysis (LDA) is robust against distributional assumption, high dimensionality makes LDA formidable due to singularity of the covariance estimate of such data and hence makes its applicability difficult. In low dimension, many classification methods have been proposed ranging from parametric method to nonparametric method. Their performances vary from one method to another depending on their properties. Using nonparametric approach, the notions of the data depth have been applied in many applications.

Vencalek and Makinde (2021) developed a rank – rank classification method for low dimensional data. The method competes well with others. However, a thorough investigation is needed to fully understand its performance when sample size is less than dimension (high dimension) of the data irrespective of

sample size. They are versatile and more attractive in handling different data structure. Their robustness against outliers and extreme values is one of its advantages. Nevertheless, some non-parametric classifiers are more flexible and free from distributional assumption but not suitable for classification of high dimensional data (Scott, 2015). The spatial depth function is computationally simple, making it suitable for classification when dimension is greater than sample size. Similarly, modified band depth is well applied in functional statistics. Their application for analyzing high dimensional data and as depth function for determining ranks in rank – rank classification method is highly significant and novel [5].

Methods

Depth Function

A depth function is a non – negative real – valued function which allows the ordering of points in d – dimensional space ($\mathbb{R}^d$) in relation to probability measure P defined on the space. The concept of depth function serves as a technique for extending various attributes of a univariate distribution to multivariate distribution. It provides effective ways through which relevant information from multivariate data is extracted. The performance of depth functions is centered on its probability of correct classification, computational simplicity, theoretical mathematical formations, and hypothetical properties.

The dataset used for this research is a multivariate dataset (say, d -dimensional) and the median vector is the point in $\mathbb{R}^d$ for which a statistical depth functional is maximized. Based on the sample median for multivariate data for describing the center of dataset, quantitative and graphical methods were used for analyzing multivariate distributional characteristics such as location, scale, bias, skewness and kurtosis, as well as for comparing inferential methods. The depth distribution classifiers are nonparametric classification methodologies based on the use of data depth and are Bayes – optimal rule under certain conditions.

In low dimension, especially when dimension of data is 2 or 3, all these depth functions are computationally simple. That is, they are easy to compute. However, when dimension is greater than 3, many of them are practically difficult to compute. However, where sample size is less than dimension, only spatial depth function is computationally feasible.

Spatial Depth Function

Vardi and Zhang (2000) and Serfling (2002) introduced and studied a popular notion of data depth known as spatial depth. The spatial depth function of an observation $x \in \mathbb{R}^d$ with regards to a distribution function, F , defined as:

$$SPD(x, F) = 1 - \left\| E \left(\frac{x - x}{\|x - x\|} \right) \right\|$$

where, E is the mathematical expectation and F is the distribution of X [6].

Spatial depth possesses the following attributes; its highest depth is attained at the spatial median; it is invariant under rotation and orthogonal transformation of the data. One of the primary motivations for using spatial depth is its computational tractability, particularly when the dimension is high. However, spatial depth function does not account for correlation among variables. It has been proven that correlation values between variables in high dimension may be very strong [7].

Depth-Oriented Ranking

Makinde (2019) defined distribution function of depth function as a centre–outward ranking in ordering multivariate objects with the use of data depth. He explores the notions of functional depth and proposed a classification method based on distribution functions of data depth for functional data. The performance of the classification method was examined by using simulations and real datasets, the results were compared with the results from existing methods. The depth – oriented ranking is said to assign a test of observation to the class or population with respect to the maximal depth classification rules [4].

The depth-oriented rank, defined as:

$$R_1(y) = P(D(X, F_1) \geq D(y, F_1))$$

Where, $D(X, F_1)$ is the depth function of X with respect to F_1 and $D(y, F_1)$ is the depth function of y with respect to F_1 , which is necessary for ordering multivariate observations, especially in high dimension will be determined.

Rank-Rank Classifiers

The RR-classifiers are based on relative ranks with respect to empirical distributions gained from the available training set. It has the following form:

Consider two distributions F_1 and F_2 on $\mathbb{R}^d$;

assign x to F_1 if $R_1(x) \geq \eta(R_2(x))$ and to F_2 otherwise,

where, $\eta(\cdot)$ is certain function, e.g., polynomial of a given order (other functions can be employed too). Here, polynomial of the

form $\eta(r) = \sum_{i=1}^{k_0} b_i r^i$, where $\mathbf{b} = (b_1, b_2, \dots, b_{k_0}) \in \mathbb{R}^{k_0}$

is a vector of coefficients of the polynomial, is considered. To determine optimal choice of k_0 , cross validation of error rate is suggested. The performance of the proposed Rank – Rank classification method will be further investigated using simulated datasets.

Depth-Depth Classification

A DD-plot is a simple graph on a plane regardless of the size of the data dimension. DD-plot denotes a plot of depth values of those sample points with respect to the underlying distribution of the given multivariate data. Given two random samples $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^d$ with distribution F and

G respectively, depth – depth classifier are stated as

If $D_G(x) > r(D_F(x)) \rightarrow \text{assign } x \text{ to } G$

If $D_G(x) < r(D_F(x)) \rightarrow \text{assign } x \text{ to } F$

where $D_F(x) = (f(x))$ and $D_G(x) = (f(g))$ for some strictly increasing function.

Suppose $X = \{X_1, \dots, X_n\}$ and $Y = \{Y_1, \dots, Y_m\}$ are two random samples from F and G respectively with distributions specified in $\mathbb{R}^d$. If $F = G$, the DD-plot will concentrate along the 45 degree line. Also, if the distribution of F and G are different, there will be a clear deviation from the 45 degree line in the DD-plot. The best separating curve in the DD-plot is calculated by the data underlying probabilistic geometry and is controlled strictly by the data. In a two-dimensional DD-plot, the classification result can be quickly identified. It is much simpler than monitoring the result of the classification in the initial sample space which is prohibitively challenging if the samples are huge.

Results and Discussion

Simulation Studies

Simulation studies in high dimensions were presented to evaluate the performance of the classifiers based on their mean proportion of correct classification. Each simulation experiment was replicated 1000 times. Each simulation study consists of 100 random observations in each training set and 100 random observations in each test set. The five underlisted simulation procedures are considered for depth – depth classification method, support vector machine and rank – rank classification method. The performances of the classifiers are evaluated based on their mean proportion of correct classification of the classifiers.

Simulation 1: Suppose i^{th} observation is in j^{th} group, then

$$X_i \sim N(\mu_j, \Sigma), \text{ where } \mu_j = \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \\ \vdots \\ \mu_{j500} \end{pmatrix} \text{ with } \mu_{1k}=1 \text{ for } 1 \leq k \leq 100$$

and $\mu_{1k}=2$ for $101 \leq k \leq 500$ and $\mu_{2k}=2$ for $1 \leq k \leq 500$, $j=1,2$. The covariance matrix Σ has elements $\rho^{|k-k'|}$. That is,

$$\Sigma = \begin{pmatrix} 1 & \rho & \dots & \rho^{499} \\ \rho & 1 & \dots & \rho^{498} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{499} & \rho^{498} & \dots & 1 \end{pmatrix}$$

with $\rho=0.7$.

The competing groups differ only in their mean vectors. This covariance matrices of the two groups is not identity matrix. Also, the two competing population are elliptically symmetric.

Simulation 2: Suppose i^{th} observation is in j^{th} group, then

$$X_{ji} \sim N(\mu_j, I), \text{ where } \mu_j = \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \\ \vdots \\ \mu_{j500} \end{pmatrix} \text{ with } \mu_{1k}=1 \text{ for } 1 \leq k \leq 100 \text{ and}$$

$\mu_{1k}=2$ for $101 \leq k \leq 500$ while $\mu_{2k}=2$ for $1 \leq k \leq 500$ and group label $j=1,2$. The term I denotes 500×500 identity matrix.

Covariance matrices of the competing classes are identity matrices. The competing populations are spherically symmetric.

Simulation 3: Suppose A_1 and A_2 are populations with normal mixture distributions F_1 and F_2 , defined as

$$F_1 = \begin{cases} N(\mu_1^a, \Sigma), & \text{with probability } \rho \\ N(\mu_1^b, \Sigma), & \text{with probability } 1 - \rho \end{cases}$$

and

$$F_2 = \begin{cases} N(\mu_2^a, \Sigma), & \text{with probability } \rho \\ N(\mu_2^b, \Sigma), & \text{with probability } 1 - \rho \end{cases}$$

respectively, where $\rho \in (0,1)$ is the mixing proportion,

$\mu_{1k}^a = 0$ for $1 \leq k \leq 500$, $\mu_{1k}^b = 0.5$ if $1 \leq k \leq 100$ and

$\mu_{1k}^b = 0$ if $101 \leq k \leq 500$, $\mu_{2k}^a = 0.5$ if $1 \leq k \leq 100$ and $\mu_{2k}^a = 0$

if $101 \leq k \leq 500$, $\mu_{2k}^b = 0.7$ if $1 \leq k \leq 100$ and $\mu_{2k}^b = 0$

if $101 \leq k \leq 500$, the covariance matrix Σ is defined as

$$\Sigma = \begin{pmatrix} \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho \\ \rho & \rho & \rho & \rho & \rho \end{pmatrix}$$

and

$$V = \begin{pmatrix} 1 & 0.5 & \dots & 0.5^{98} & 0.5^{99} \\ 0.5 & 1 & \dots & 0.5^{97} & 0.5^{98} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.5^{99} & 0.5^{98} & \dots & 0.5 & 1 \end{pmatrix}$$

This involves mixture distributions. The mixing proportion p in simulation 3 is taken to be 0.8.

Simulation 4: Suppose i^{th} observation is in group 1, then

$$X_i \sim N(\mu_j, I), \text{ where } \mu_j = \begin{pmatrix} \mu_{j1} \\ \mu_{j2} \\ \vdots \\ \mu_{j500} \end{pmatrix} \text{ with}$$

$\mu_j=1$ for $1 \leq j \leq 100$ and $\mu_j=2$ for $101 \leq j \leq 500$. Suppose i^{th} observation is in group 2, then $Y_i \sim N(\mu_k, 3I)$, where $\mu_k=2$ for $1 \leq k \leq 500$. The term I denotes 500×500 identity matrix.

Simulation 5: Suppose each experiment consists of measurements on independent features such that if observation i belong to group G_1 , a random variable $X_k \sim \text{Exp}(1)$ for $1 \leq k \leq 100$, $X_k \sim \text{Exp}(2)$ for $101 \leq k \leq 300$ and $X_k \sim \text{Exp}(1.5)$ for $301 \leq k \leq 500$. Also, if observation i belong to group G_2 , a random variable $Y_k \sim \text{Exp}(2)$ for $1 \leq k \leq 500$.

Simulations 4 and 5 involve mean shift with independent features for normally and uniformly distributed samples respectively.

The figures below present plot of competing groups in depth-depth and rank-rank spaces using simulations 1 to 5.

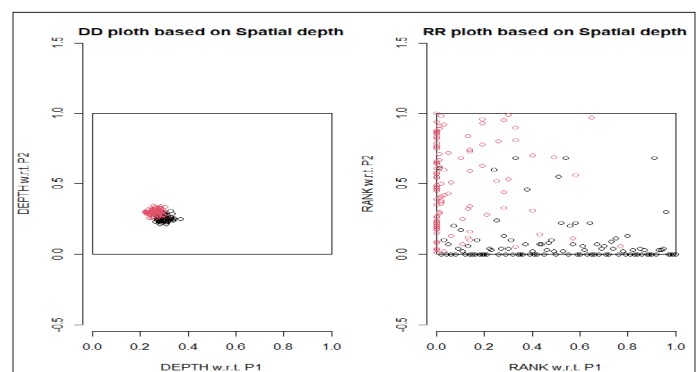


Figure 1: Plot of Competing Groups in Depth – Depth and Rank – Rank Spaces for Simulation 1

Observations from the two populations in simulation 1 are presented in DD space and RR space. In DD space, population 1 and 2 are clustered together while in RR space, population 1 and 2 are scattered along the vertical and horizontal axis together.

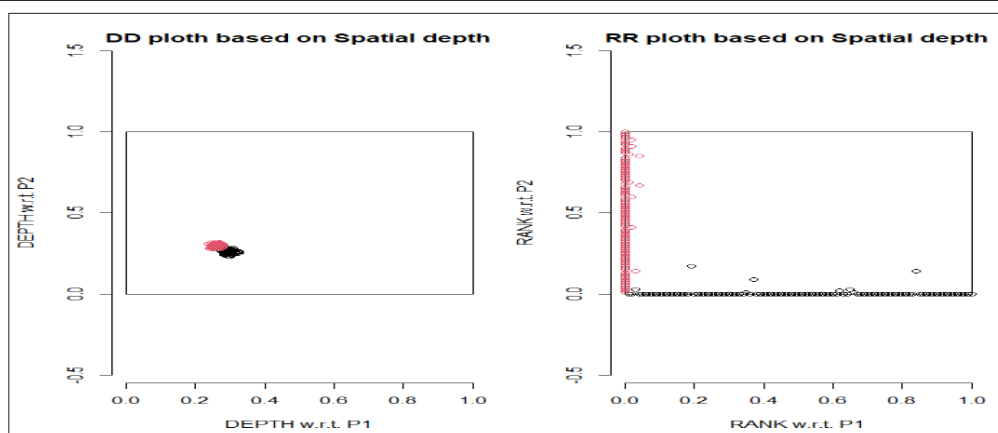


Figure 2: Plot of Competing Groups in Depth – Depth and Rank - Rank Spaces for Simulation 2

In simulation 2, classification of the population can be seen using DD space and RR space. The DD space is clustered together in population 1 and 2 while in RR space population 1 are horizontally displaced and population 2 are vertically displaced.

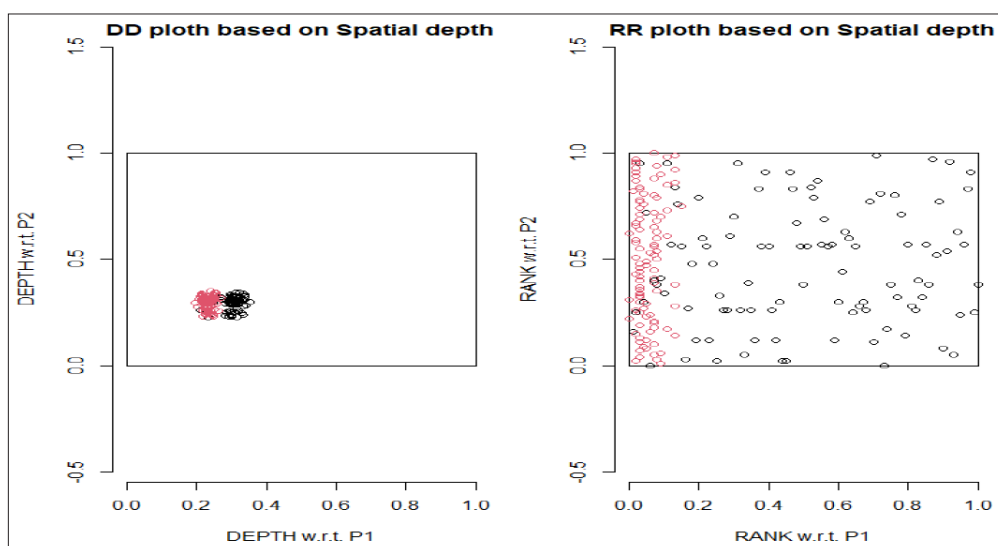


Figure 3: Plot of Competing Groups in Depth – Depth and Rank - Rank Spaces for Simulation 3

Simulation 3 shows that populations 1 and 2 are clustered together when using DD space while they are scattered together in populations 1 and 2 when using RR space.

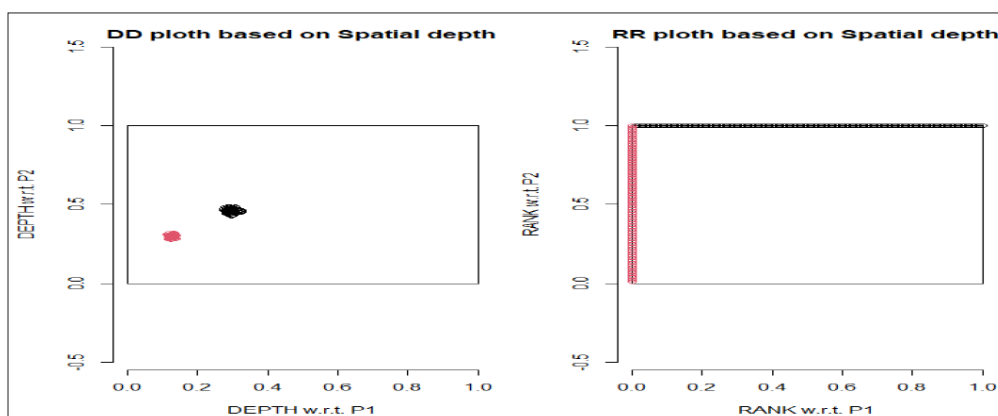


Figure 4: Plot of Competing Groups in Depth – Depth and Rank - Rank Spaces for Simulation 4

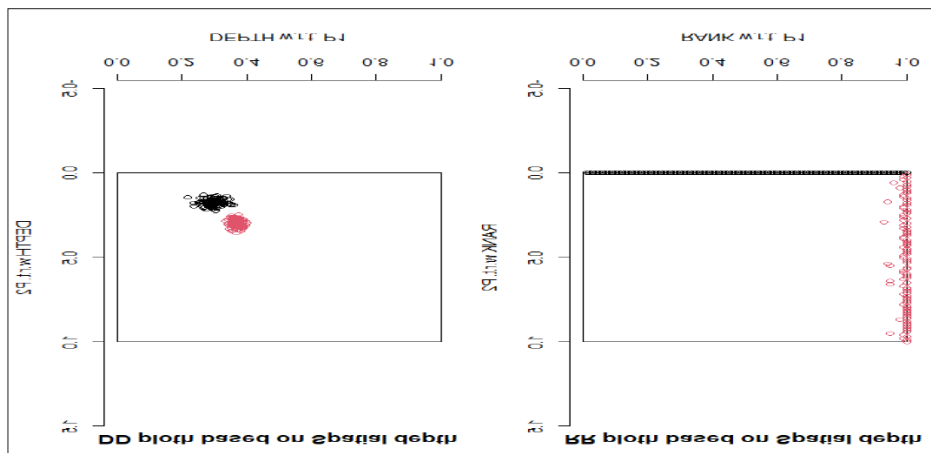


Figure 5: Plot of Competing Groups in Depth – Depth and Rank - Rank Spaces for Simulation 5

Observations in simulation 4 and 5 have the same features, that is, populations 1 and 2 can efficiently displace data and the classification exercise can be seen through the plot using DD space while data in population 1 are horizontally displaced and are vertically displaced in population 2 using RR space.

Descriptive Statistics

Tables 1 - 5 below present descriptive statistics showing the minimum, 1st quartile, median, 3rd quartile, maximum, mean and standard deviation of the proportion of correct classification for some competing classifiers for simulations procedures 1-5.

Table 1: Descriptive Statistics of the Proportion of Correct Classification for Some Competing Classifiers for Simulation 1

Descriptive Statistics	RR-SPD Lin	RR-SPD Quad	DD-SPD Lin	DD-SPD Quad	RR-MBD Lin	RR-MBD Quad	DD-MBD Lin	DD-MBD Quad	SVM
Minimum	0.6700	0.6111	0.9950	0.5000	0.8525	0.7692	0.8400	0.8750	0.9350
1st Quartile	0.7400	0.6891	0.9950	0.5000	0.9025	0.8750	0.9000	0.9150	0.9700
Median	0.7600	0.7105	0.9950	0.5000	0.9150	0.8944	0.9100	0.9250	0.9750
3rd Quartile	0.7800	0.7296	1.0000	0.5000	0.9275	0.9116	0.9250	0.9400	0.9850
Maximum	0.8425	0.8053	1.0000	1.0000	0.9650	0.9568	0.9600	0.9750	1.0000
Mean	0.7598	0.7102	0.9974	0.5180	0.9149	0.8927	0.9108	0.9274	0.9761
SD	0.0289	0.0299	0.0025	0.0932	0.0193	0.0283	0.0200	0.0165	0.0106

In simulation 1, depth – depth spatial depth linear (DD – SPD Lin) has the highest mean proportion of correct classification, followed by support vector machine (SVM) and depth – depth modified band depth quadratic (DD – MBD Quad). The performances of RR – MBD Lin, DD – MBD Lin, RR – MBD Quad, RR – SPD Lin and rank – rank spatial depth quadratic RR – SPD Quad are fairly better than depth – depth spatial depth quadratic (DD – SPD Quad) which has the least mean proportion of correct classification.

Table 2: Descriptive Statistics of the Proportion of Correct Classification for Some Competing Classifiers for Simulation 2

Descriptive Statistics	RR-SPD Lin	RR-SPD Quad	DD-SPD Lin	DD-	RR-MBD Lin	RR-MBD Quad	DD-MBD Lin	DD-MBD Quad	SVM
Minimum	0.9700	0.7895	0.9950	0.5000	0.9725	0.8462	0.9750	0.9800	1.0000
1st Quartile	0.9900	0.9706	0.9950	0.5000	0.9900	0.9697	0.9900	0.9950	1.0000
Median	0.9950	1.0000	0.9950	0.5000	0.9950	1.0000	0.9950	0.9950	1.0000
3rd Quartile	1.0000	1.0000	0.9950	0.5000	0.9975	1.0000	0.9950	1.0000	1.0000
Maximum	1.0000	1.0000	0.9950	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Mean	0.9943	0.9836	0.9950	0.5280	0.9936	0.9831	0.9927	0.9964	1.0000
SD	0.0051	0.0251	0.0000	0.1000	0.0056	0.0247	0.0059	0.0040	0.0000

In simulation 2, support vector machine (SVM) has the highest mean proportion of correct classification and its performance is slightly better than depth – depth modified band depth quadratic (DD – MBD Quad) followed by depth – depth spatial depth linear (DD – SPD Lin) while RR – SPD Lin, RR – MBD Lin, DD – MBD Lin, RR – SPD Quad and rank – rank modified band depth quadratic are far better than depth – depth spatial depth quadratic which has the least mean proportion of correct classification.

Table 3: Descriptive Statistics of the Proportion of Correct Classification for Some Competing Classifiers for Simulation 3

Descriptive Statistics	RR-SPD Lin	RR-SPD Quad	DD-SPD Lin	DD-	RR-MBD Lin	RR-MBD Quad	DD-MBD Lin	DD-MBD Quad	SVM
Minimum	0.6725	0.6158	0.9950	0.5000	0.6050	0.6471	0.6200	0.6600	0.0800
1st Quartile	0.7394	0.6902	0.9950	0.5000	0.7150	0.6902	0.6550	0.7000	0.2400
Median	0.7600	0.7121	1.0000	0.5000	0.7350	0.7047	0.6700	0.7150	0.4450
3rd Quartile	0.7800	0.7315	1.0000	0.5000	0.7525	0.7196	0.6850	0.7300	0.7413
Maximum	0.8325	0.8131	1.0000	1.0000	0.8225	0.7684	0.7600	0.7800	0.9300
Mean	0.7592	0.7114	0.9976	0.5120	0.7332	0.7050	0.6711	0.7160	0.4836
SD	0.0288	0.0306	0.0025	0.0766	0.0317	0.0223	0.0236	0.0200	0.2539

The depth – depth spatial depth linear (DD – SPD Lin) achieves the highest mean proportion of correct classification in simulation 3 followed by rank – rank spatial depth linear (RR – SPD Lin) and rank – rank modified band depth linear (RR – MBD Lin). The performances of DD – MBD Quad, RR – SPD Quad, RR – MBD Quad, DD – MBD Lin and depth – depth spatial depth quadratic are far better than that of support vector machine (SVM) in simulation 3.

Table 4: Descriptive Statistics of the Proportion of Correct Classification for Some Competing Classifiers for Simulation 4

Descriptive Statistics	RR-SPD Lin	DD-SPD Lin	DD-SPD Quad	RR-MBD Lin	DD-MBD Lin	DD-MBD Quad	SVM
Minimum	0.5025	0.9950	0.5000	0.8950	0.5000	0.9950	1.0000
1st Quartile	0.5025	0.9950	0.5000	0.9375	0.5000	1.0000	1.0000
Median	0.5025	0.9950	0.5000	0.9450	0.5000	1.0000	1.0000
3rd Quartile	0.5025	0.9950	1.0000	0.9550	0.5000	1.0000	1.0000
Maximum	0.5025	0.9950	1.0000	0.9900	0.5000	1.0000	1.0000
Mean	0.5025	0.9950	0.7310	0.9450	0.5000	0.9999	1.0000
SD	0.0000	0.0000	0.2495	0.0137	0.0000	0.0007	0.0000

In simulation 4, support vector machine (SVM) and depth – depth modified band depth quadratic (DD – MBD Quad) perform competitively, followed by depth – depth spatial depth linear (DD – SPD Lin) while RR – MBD Lin, DD – SPD Quad and rank – rank spatial depth linear are far better than depth – depth modified depth linear (DD-MD Lin) which has the least mean proportion of correct classification. RR – SPD Quad and RR – MBD Quad are insignificant.

Table 5: Descriptive Statistics of the Proportion of Correct Classification for Some Competing Classifiers for Simulation 5

Descriptive Statistics	RR-SPD Lin	RR-SPD Quad	DD-SPD Lin	DD-SPD Quad	RR-MBD Lin	RR-MBD Quad	DD-MBD Lin	DD-MBD Quad	SVM
Minimum	0.5025	0.9700	0.9950	0.5000	0.7525	0.7640	0.5000	0.7800	0.9900
1st Quartile	0.5025	0.9900	0.9950	0.5000	0.7975	0.8105	0.5000	0.8250	1.0000
Median	0.5075	0.9900	0.9950	0.5000	0.8150	0.8265	0.5000	0.8350	1.0000
3rd Quartile	0.5100	0.9900	0.9950	0.5000	0.8325	0.8449	0.5000	0.8550	1.0000
Maximum	0.5275	1.0000	0.9950	1.0000	0.8850	0.8967	0.5150	0.9000	1.0000
Mean	0.5075	0.9898	0.9950	0.6170	0.8157	0.8275	0.5011	0.8389	0.9990
SD	0.0048	0.0038	-	0.2119	0.0257	0.0254	0.0023	0.0222	0.0022

In simulation 5, support vector machine (SVM) has the highest mean proportion of correct classification, followed by depth – depth spatial depth linear (DD – SPD Lin) and rank – rank spatial depth quadratic (RR – SPD Quad). The performances of DD – MBD Quad, RR – MBD Quad, RR – MBD Lin, DD – SPD Quad and rank – rank spatial depth linear (RR – SPD Lin) are fairly better than depth – depth modified depth linear (DD – MBD Lin) which has the least mean proportion of correct classification.

Boxplots

The results of the simulations are summarized in form of boxplots.

Figures 6-10 below present a boxplot for the mean proportion of correct classification obtained from simulations 1 to 5.

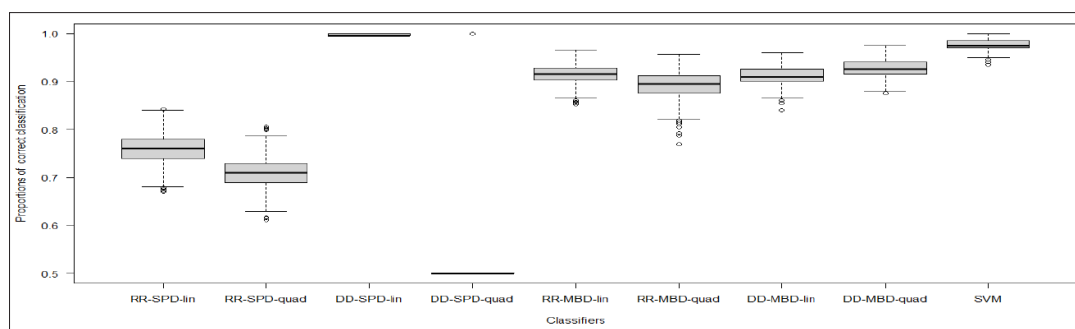


Figure 6: Boxplot for the Proportions of Correct Classification Obtained from Simulation 1

In simulation 1, DD-SPD-Lin outperforms all other methods as the mean proportion of correct classification approaches 1. Also, the SVM performs better than the remaining methods. However, DD-SPD-quad has the least mean proportion of correct classification.

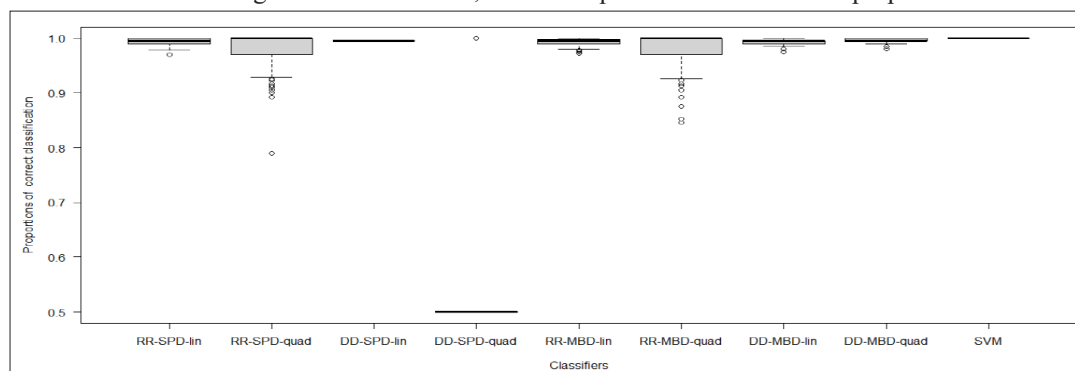


Figure 7: Boxplot for the Proportions of Correct Classification Obtained from Simulation 2

In simulation 2, DD-SPD-quad has the least mean proportion of correct classification while all other methods performed fairly well with their mean proportion of correct classification as it approaches 1.

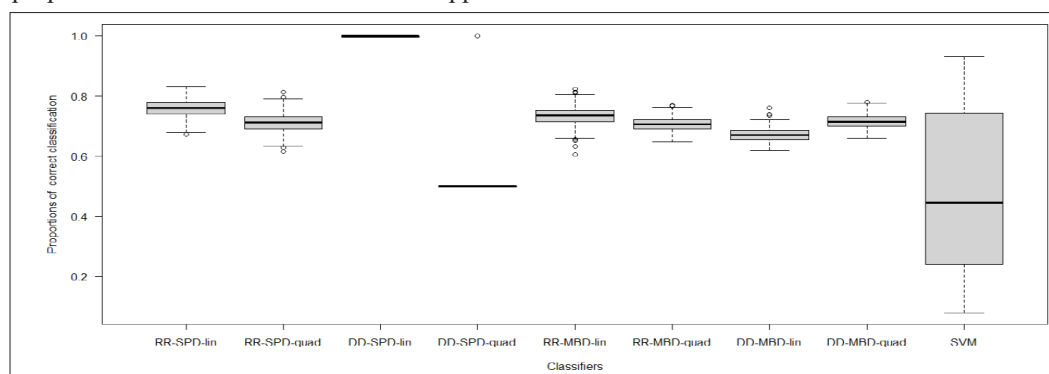


Figure 8: Boxplot for the Proportions of Correct Classification Obtained from Simulation 3

Observation in simulation 3 shows that, the mean proportion of correct classification using linear function in DD-SPD outperforms all other methods as the mean proportion of correct classification approaches 1. However, SVM do not perform competitively in this example as the mean proportion is small and it varies from 0.05 -0.8

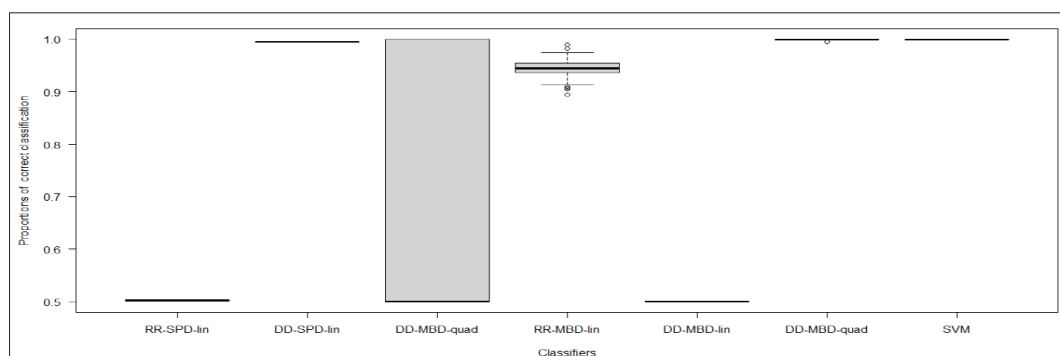


Figure 9: Boxplot for the Proportions of Correct Classification Obtained from Simulation 4

In simulation 4, there is a perfect mean proportion of correct classification using modified band depth (DD-MBD-quad), SVM and DD-SPD-Lin as they achieve highest performances. However, RR-SPD-Lin and DD-MBD-Lin have the least mean proportion of correct classification.

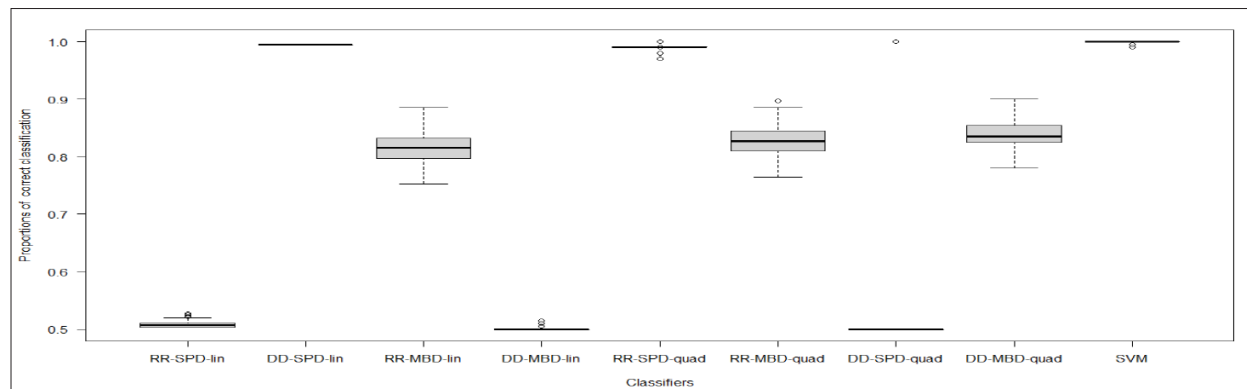


Figure 10: Boxplot For the Proportions of Correct Classification Obtained from Simulation 5

The performances of DD-SPD-Lin, RR-SPD-Quad and SVM are equivalent while RR-MBD-Lin, RR-SPD-Quad and DD-MBD-Quad perform better than other versions of spatial depth function in simulation 5 while RR-SPD-Lin, DD-MBD-Lin and DD-SPD-Quad have the least mean proportion of correct classification.

Real Data to Validate the Simulation Results

We analyse three benchmark data sets to illustrate the performances of the classification methods. These datasets are Arcene data, Gisette data and Madelon data.

Arcene data consist of two classes, positive and negative (+1 and -1) examples, with training sample sizes 44 and 56, and validation sample size 44 and 56, respectively, each with 10,000 features. A random training sample of size 30 and a random validation sample of size 30 are selected from each of the two classes with the 10,000 features. Arcene data are available at <https://archive.ics.uci.edu/ml/machine-learning-databases/arcene/ARCENE/>. To select feature with high classification power, t statistic of each feature was computed. Eight hundred features with highest absolute values of t-test were selected for the classification exercise.

We consider the Gisette data which available at <https://archive.ics.uci.edu/ml/machine-learning-databases/gisette/GISETTE/gisette-train.data>, and it arose from the study of LeCun et al. (1998). The data consist of two classes, each with size 300 and 50,001 features. We removed zero and near-zero-valued features in the datasets based on the percentage uniqueness as implemented in R package caret. The minimum percentage uniqueness of 20% was set. It is observed that only 4639 out of 50,001 features achieve percentage uniqueness more than 0.2. A random training and test sample of size 100 each are selected from each of the classes.

Furthermore, considering the Madelon data, this is artificial dataset and consists of two groups with 500 continuous input variables. The groups are positive and negative (+1 and -1), a random training sample of size 200 is selected from each group. The experiment was replicated 500 times. In each replication, p-value is estimated and the power is determined as the proportion of p-values are less than level of significance.

Table 6: Descriptive Statistics of the Proportions of Correct Classification Obtained from Arcene Data

Descriptive Statistics	RR-SPD Lin	RR-SPD Quad	DD-SPD Lin	DD-SPD Quad	RR-MBD Lin	DD-MBD Lin	SVM
Minimum	0.6250	0.6571	0.6000	0.7125	1	1	0.7000
1st Quartile	0.6625	0.6957	0.6625	0.7375	1	1	0.7375
Median	0.6813	0.7121	0.6750	0.7375	1	1	0.7500
3rd Quartile	0.7000	0.7286	0.7000	0.7375	1	1	0.7750
Maximum	0.7625	0.8136	0.7750	0.7375	1	1	0.8250
Mean	0.6830	0.7130	0.6816	0.7375	1.0000	1.0000	0.7568
SD	0.0241	0.0245	0.0295	0.0011	-	-	0.0237

In arcene data using the above analyses, RR-MBD Lin and DD-MBD Lin are both equal and have the highest mean proportion of correct classification, followed by SVM and DD-SPD Quad. The performances of RR-SPD Quad and RR-SPD Lin are better than DD-SPD Lin which has the least mean proportion of correct classification.

Table 7: Descriptive Statistics of the Proportions of Correct Classification Obtained from Gisette Data

Descriptive Statistics	RR-SPD Lin	RR-SPD Quad	DD-SPD Lin	DD-SPD Quad	RR-MBD Lin	DD-MBD Lin	SVM
Minimum	0.4950	0.5685	0.4800	0.5700	1	1	0.4950
1st Quartile	0.5700	0.6089	0.5600	0.6100	1	1	0.5700
Median	0.5900	0.6231	0.5850	0.6250	1	1	0.7500
3rd Quartile	0.7000	0.7286	0.7000	0.7375	1	1	0.7750
Maximum	0.7625	0.8136	0.7750	0.7375	1	1	0.8250
Mean	0.6830	0.7130	0.6816	0.7375	1.0000	1.0000	0.7568
SD	0.0241	0.0245	0.0295	0.0011	-	-	0.0237

In Gisette data, RR-MBD Lin and DD-MBD Lin are equal with the highest mean proportion of correct classification and their performances are much better than DD-SPD Quad, followed by RR-SPD Quad. The performance of SVM and RR-SPD Lin are better than DD-SPD Lin which has the least mean proportion of correct classification.

Table 8: Descriptive Statistics of The Proportions of Correct Classification Obtained from Madelon Data

Descriptive Statistics	RR-SPD Lin	RR-SPD Quad	DD-SPD Lin	DD-SPD Quad	RR-MBD Lin	DD-MBD Lin	SVM
Minimum	0.4788	0.5388	0.4700	0.5450	1	1	0.4925
1st Quartile	0.5350	0.5749	0.5300	0.5775	1	1	0.5500
Median	0.5481	0.5854	0.5450	0.5875	1	1	0.5650
3rd Quartile	0.5625	0.5950	0.5575	0.5975	1	1	0.5806
Maximum	0.6088	0.6288	0.6125	0.6300	1	1	0.6200
Mean	0.5482	0.5850	0.5438	0.5874	1.0000	1.0000	0.5657
SD	0.0210	0.0157	0.0212	0.0152	-	-	0.0213

Analyzing Madelon data, RR-MBD Lin and DD-MBD Lin both perform equivalently high in terms of their mean proportion of correct classification; DD-SPD Quad is better than both RR-SPD Quad and SVM, and the performance of RR-SPD Lin is better than DD-SPD Lin which has the least mean proportion of correct classification [8-21].

Conclusion

The notions of depth functions for high dimensional data have been developed to classify observations from unknown populations correctly after getting the most information between groups or populations. The performance of spatial depth function and modified depth function were compared in depth-oriented ranking classification method, rank-rank classification method and depth-depth classification method using five simulations studies. It was found that rank-rank classification method competes favorably with other methods under some conditions. Also, three real data sets were analyzed to illustrate the performances of the three depth functions. The performances of the classifiers were evaluated based on the mean proportion of correct classification.

In this study, modified band depth function was introduced and their performances were investigated for the classification of high dimensional data using three depth-based method of classification which are depth-oriented ranking, rank-rank classification and depth-depth classification methods. The mean proportion of correct classification was used to identify the optimal spatial depth functions for classification of high dimensional data based on simulation studies and real data analysis. The result in this study shows that rank-rank classification method outperforms other methods in high dimensions.

Competing Interest

The authors declare they have no conflicts of interest.

Acknowledgements

The authors acknowledge the support of the Department of Statistics, Federal University of Technology Akure, Nigeria.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Anderson TW (1984) An introduction to multivariate statistical analysis. John Wiley and Sons, Inc, New York <https://scispace.com/pdf/an-introduction-to-multivariate-statistical-analysis-second-msxr2pcton.pdf>.
- Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: Descriptive statistics, graphic and inference. The Annals of Statistics 27: 783-858.
- Makinde OS, Adewumi AD (2017) A comparison of depth functions in maximal depth classification rules. Journal of Modern Applied Statistical Methods 16: 388-405.
- Makinde OS (2019) Classification rules based on distribution functions of functional depth. Statistical Papers 60: 629-640.
- Vencalek O, Makinde OS (2021) RR-classifier: A nonparametric classification procedure in multidimensional space based on relative ranks. Asta Advances in Statistical Analysis https://ideas.repec.org/a/spr/alstar/v105y2021i4d10.1007_s10182-021-00423-7.html.
- Serfling R (2002) A depth function and a scale curve based on spatial quantiles. In: Dodge Y (Ed.), Statistical data analysis based on the L1-norm and related methods 25-38.
- Makinde OS, Fasoranbaku OA (2018) On maximum depth classifiers: Depth distribution approach. Journal of Applied Statistics 45: 1106-1117.
- Dabney AR (2005) Classification of microarrays to nearest centroids. Bioinformatics 21: 4148-4154.

9. Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179-188.
10. Ghosh AK, Chaudhuri P (2005) On maximum depth and related classifiers. *Scandinavian Journal of Statistics* 32: 327-350.
11. Hall P, Titterton DM, Xue J (2009) Median based classifiers for high dimensional data. *Journal of the American Statistical Association* 104: 1597-1608.
12. Li J, Cuesta-Albertos JA, Liu RY (2012) DD-classifier: Nonparametric classification procedure based on DD-plot. *Journal of the American Statistical Association* 107: 773-783.
13. Li J, Liu RY (2004) New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science* 19: 686-696.
14. Liu RY (1990) On a notion of data depth based on random simplices. *The Annals of Statistics* 18: 405-414.
15. Liu RY, Singh K (1993) A quality index based on multivariate data depth and multivariate rank tests. *Journal of the American Statistical Association* 88: 252-260.
16. Makinde OS (2016) On misclassification probabilities of linear and quadratic classifiers. *Afrika Statistika* 11: 943-953.
17. Makinde OS, Chakraborty B (2015) On some nonparametric classifiers based on distribution functions of multivariate ranks. *Springer Switzerland* 249-264.
18. Scott DW (2015) *Multivariate density estimation: Theory, practice and visualization*. Wiley New York <https://www.scrip.org/reference/referencespapers?referenceid=3822835>.
19. Tukey JW (1975) Mathematics and picturing data. *Proceedings of the 1975 International Congress of Mathematics* 2: 523-531.
20. Vapnik VN (1982) *Estimation of dependences based on empirical data*. Springer New York <https://www.semanticscholar.org/paper/Estimation-of-Dependences-Based-on-Empirica>.
21. Wald A (1944) On a statistical problem arising in the classification of an individual into one of two groups. *The Annals of Mathematical Statistics* 15: 145-162.