American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

© Global Society of Scientific Research and Researchers

http://asrjetsjournal.org/

Testing the Optimality of Two Different Non-Parametric Discriminant Methods

Evelyn N. Okeke[a]*, Uchenna J. Okeke[b]

[a,b]Depatment of Mathematics & Statistics, Federal University, Wukari, Taraba State, Nigeria

[a]Email: evelyn70ng@yahoo.com

[b]Email: uche70ng@yahoo.com

Abstract

This paper aims at comparing the concept of data depth to classification and classification by projection pursuit using method of linear discriminant function. These two methods allow the extension of univariate concepts to the field of multivariate analysis. In particular they open the possibility of non-parametric methods to be used in multivariate data analysis. In this study, six simulated and one real life data sets were studied and, we observed that projection pursuit method is more optimal in classifying objects into their original groups.

Keywords: Variance-covariance matrix; Data depth; Spatial or L_1 depth; Linear Discriminant analysis; Probability of Misclassification (PMC).

1. Introduction

Disriminant analysis is one of the most popular method of classification, which aimed at classifying new object whose true population is not yet known into one of the known populations whose characteristics were known and have been used to set up a classification function. The classification function is what is used in assigning new object into a population in which it belongs. When the populations are normally distributed parametric discriminant functions like Fisher linear and quadratic functions, Bayes discriminant functions and others can correctly classify an object into its true population. When the populations distribution are far from normal, or when the data are ill conditioned, parametric discriminant analysis may produce misleading results. For this reason it become necessary to look for a more robust method that can work well under so many unusual situations and this is what gave rise to non-parametric discriminant analysis methods.

----

* Corresponding author.

Though we have some semi-parametric methods like M-estimators, S-estimators, MCD estimators, MWCD estimators and MVE estimators etc that works well when the populations are not normally distributed but these still have their own impediments. Most of them are not optimal in classification in singular and near singular conditions of variance-covariance matrices of the populations and also when the data are ill conditioned.

To solve singularity problem and other problems that are associated with sparse and multicollinear data, projection pursuit and data depth approaches came up as remedies. These methods aimed at reducing a high dimensional data set to low dimension so that the low dimensional data statistical tools can be applied. In this article we compared the performance of data depth method and projection pursuit method through several simulations and by applying them to a real data set. Their performance is assessed by comparing the probability of misclassification of the methods.

This paper is organized into four Sections. Section one contains the introduction. Section two contains the materials and methods. Section three shows some illustrations and the results of the study. And section four contains the summary and conclusion of the study.

2. Materials and methods

2.1. Data depth

Data depth is a modern nonparametric tool for the analysis of multivariate data. This method helps in reducing high dimensional data to low dimension where low dimension data statistical analysis can be applied. The concept of data depth is very important because it leads to a natural center-outward ordering of sample points in multivariate data sets. The notion of data depth was proposed by [1] as a graphical tool for visualizing bivariate data sets and has since then been extended to the multivariate case [2]. The depth of a point relative to a given data set measures how deep that point lies in the data cloud. The data depth concept provides center-outward ordering of points in any dimension and leads to a new nonparametric multivariate statistical analysis in which no distributional assumptions are needed.

Most depth functions are robust and affine invariant making them well suited for the study of real life high dimensional data sets that may contain outliers.

Recently there exist different types of data depth each with its different functional form. Some different types of data depth include

- ***L_1 or Spatial depth***

Let X be a p-dimensional random vector having distributional form F_x. Then, the multivariate spatial or L$_1$ depth of $x \in R^p$ relative to F_x is defined as

$$D_1(x;F_x) = 1 - \left\| E_{F_x}(x - X) / \|x - X\| \right\| \qquad (1)$$

Where $X \,\square F_x$, and $\|\cdot\|$ is the Euclidean norm.[3]

The spatial depth function above is use in finding the depth of observation(s) in a data cloud of numerous p-variate observations. This has attractive robustness and computational properties, and serves as a basis for useful non-parametric multivariate descriptive measures.

- ***Mahalanobis depth***

Let μ_{F_x} be a vector that measures the location of X in a continuous and affine equivariant, and $\sum_x$

covariance matrix of F$_x$ which depend on the distribution of F$_x$. Based on the estimates of location and scatter (μ_{F_x} an $\sum_x$) a simple depth statistic is constructed, the Mahalanobis depth is given by

$$MD(X, F_x) = [1 + (x - \mu_{F_x})' \Sigma_x^{-1} (x - \mu_{F_x})]^{-1} \tag{2}$$

The sample version of MD is

$$MD(x_1, \dots, x_n;\ F_x) = [1 + (x - \bar{x})' S_x^{-1} (x - \bar{x})]^{-1} \tag{3}$$

where $\bar{x}$ is the mean vector and S_x^{-1} is the empirical covariance matrix [4]

- ***L$_2$- Depth***

L$_2$-depth D_2 is one of the depth function where the outlyingness of a point, and hence its depth, can be measured by a distance from a properly chosen center of the distribution. It is based on the mean outlyingness of a point, as measured by the L$_2$ distance,

$$D_2(y|X) = (1 + E\|y - X\|)^{-1} \tag{4}$$

where y is a data point whose depth you want to find and X is the entire data one is working on. For an empirical distribution on points $x_{i,}$ where $i = 1, \dots, n$ we have that

$$D_2(y|X) = (1 + 1/\text{n}(\Sigma\|y - x_i\|))^{-1} \tag{5}$$

L$_2$-depth vanishes at infinity and is maximum at spatial median of X, which is the point that minimizes $E\|y - X\|$. If the distribution is centrally symmetric, the centre is spatial median; hence the maximum is at the centre. This depth converges in the probability distribution: For a uniformly integrable and weakly convergent sequence $P_n \rightarrow P$ it holds that $lim_n D(y|P_n) = D(y|P)$

L$_2$-depth is invariant against rigid Euclidean motions, but not affine invariant. Then an affine invariant L$_2$-depth is given by

$$D_2(y|X) = (1 + E\|y - X\|_{S_X})^{-1} \qquad (6)$$

S_X is a positive definite $d \times d$ matrix that depends continuously (in weak convergence) on the distribution and measures the dispersion of X in an affine equivariant way, that is,

$S_{XA+b} = AS_XA$ holds for any matrix A of full rank and any b.

A simple choice for S_X is the covariance matrix of X [5].

- Regression depth

The regression depth (r depth) of a fit β relative to a data set Z_n is the smallest number of observations that need to be removed to make β a nonfit. Equivalently, r depth (β, Z_n) is the smallest number of residual that need to change sign to make β a nonfit.

The regression depth measures the quality of any candidate fit. Fits with higher regression depth fit the data better than do fits with lower regression depth. Hence, the regression depth ranks all possible fit from worst (r depth = 0) to best (maximal depth). This leads to the deepest regression estimator.

The deepest regression estimator $DR(Z_n)$ according to [6] is the fit β with maximal regression depth relative to the data set, i.e.,

$$DR(Z_n) = argmax_\beta \; r\ depth\ (\beta, Z_n) \qquad (7)$$

In the univariate case it is easy to see that the deepest regression of a data set is its median. Hence, deepest regression generalized the univariate median to linear regression. The maximum value of regression depth is achieved when all observation lie on a line.

- ***Location depth***

In 1975, [1] introduced the concept of location depth (called halfspace depth and Turkey depth). In the bivariate case, the location depth of a point u relative to a bi-dimensional data set S_n is defined as the smallest number of data points lying in one of the sides of a line passing through u. This definition can be extended to higher dimension.

Given a set of n points, P, the location depth of a point u is the minimum number of points contained in any half-plane passing through u. This notion of depth is not restricted to points in R^2 −the location depth of a point can be defined in any dimension. Note that in R, a number’s location depth can be computed solely by its rank. The point in R^d with the highest location depth is called the Tukey median.

This estimator is called deepest location estimator and is defined as follows:

$$DL(Zn) = argmax_{\beta} HD(\beta, Zn) \qquad (8)$$

If there is more than one observation with maximal depth, then the deepest location will be the mean of those observations. Note that in the univariate case this estimator is equivalent to the median, and in the multivariate case the deepest location can be seen as a multivariate median.

The half-space depth function is given by

$$\text{HD (X;F}_x\text{)= inf}_H\text{\{P(H): H is a close half-space in } R^p, X \in H \qquad (9)$$

It turns out that T_x(c) is the half-space depth of c in one dimension with respect to the population F_x, that is, T_x (c) =HD (c; F_x). Half-space depth is sometimes referred to as Tukey depth .

$0 \le DF \le 1$where DF in any dept function.

X_1 is more central to (or deeper) in F_X than X_2 in F_x if DF (X_1; F_x) > DF ($X_{2;}$ F_x). This is true for any depth function (DF). Let *f* be the class of distributions on the Borel sets of R^P: a statistical depth function is a bounded, nonnegative mapping D: R^P x $f \to R.$

- ***Oja depth***

Another measure of depth was proposed by [7], as follows. Given a set of points P in the plane, the Oja depth of a point $u \in R^2$ is the sum of the area of triangles formed by u and all pairs of points in P, i.e.,

$$ODepth\ (u) = \Sigma_{ij} triangle\ area(u, p_{i,} p_j) \qquad (10)$$

The same measure can also be extended to p-dimensions by considering p-simplices formed by u and every other configuration of p points of P.

The point (not necessarily unique) with minimum Oja depth is called the Oja median. It is known that the points with minimum Oja median form a convex set. Furthermore, it is known that Oja median occurs as an intersection of lines formed by pairs of points of P.

Take any point $q \in R^p$. Let $p_i, p_j \in P$ be two points in the input pointset. The Oja depth of q is simply the sum of the areas of the triangles formed by q and all pairs of points in P. Given p_i and p_j, let l_{ij} be the line passing through p_i and p_j. The area of the triangle formed by q, p_i, p_j is $\frac{1}{2}$. d($p_i p_j$).d($p_j p_i$). Essentially, the area of the triangle can be thought of as the weighted distance of q from l_{ij} , $w_{ij.}.d(q, l_{ij})$ where

$$w_{ij} = d(p_i, p_j)/2 \qquad (11)$$

So the area of triangle formed by points r, p_i, p_j is then simply $w_{ij.}.d(r, l_{ij})$. And the Ojo depth of a point q becomes $\sum_i \sum_{j\neq i} w_{ij.} . d(q, l_{ij})$

Given a point $p_i = (x_i, x_j)$, look at the cone

$$z = w_{ij}. \sqrt{(x - x_i)^2 (y - y_i)^2} \qquad (12)$$

where $w_{ij.}$ is some constant. It is easy to verify that the height of the cone at point q is $w_{ij.}. d(q, p_i)$. Therefore take the line, l_{ij}, through the two points p_i and p_j, and let $w_{ij.}$ be as defined above. Then if we place the cone defined above at every point on the line, we essentially get a wedge defined by the line. It follows that the height of this wedge at point q is $w_{ij.}. d(q, l_{ij})$ which is exactly the area of the triangle q, p_i, p_j.

- ***Simplical depth***

Oja depth for a point measured the sum of the areas of $\binom{n}{2}$ triangle in the plane. We can also define a

measure which simply counts the number of triangles containing the required point. This leads to simplicial depth of a points set, defined by [8]. Let T be the set of all triangles formed by vertices of P- each triangle requires three vertices, and therefore T has $\binom{n}{3}$ triangles. Given a point $u \in R^p$, the simplicial depth of u, denoted S Depth(u) is the number of triangles of T that contains u.

The simplical median is the point with the highest simplical depth. [8] showed that the simplical median is invariant to affine transformations.

- ***Convex Hull Peeling Depth*** **[9]**

Convex hull peeling depth of a point $x \in S_n \subset R^p$ relative to a p-dimensional data set S_n is simply the level of the convex layer to which x belongs to .

A convex layer is defined as follows: Construct the smallest convex hull and enclose all data points. The points on the perimeter are designated the first convex layer which is removed. The convex hull of the remaining points is constructed; these points on the perimeter are the second convex layer. The process is repeated, and a sequence of the nested convex layers is formed. The higher a point belongs to, the deeper the point is within the data cloud. The disadvantages of convex hull peeling depth are: it is not a robust measure and it's impossible to associate it a theoretical distribution.

2.2. Projection Pursuit Tables

Projection pursuit is a thorough process of searching all the projection directions to find the most "interesting" projection. This procedure derives its name from the fact that it interprets high dimensional data through well-chosen lower-dimensional projections. The "pursuit" part of the name refers to optimization with respect to the projection direction. The method is computationally intensive and gets complicated further as the dimension increases, but this technique is gaining popularity with increased attention given to savvy computer programming techniques and improvement in computer technology.

"Projection pursuit is the most powerful technique that can lift a one-dimensional technique to higher dimension." [10]. This implies that a projection pursuit technique can be used to reduce dimension to one and then any one-dimensional statistical technique can be applied. Every traditional projection pursuit methodology mainly differs in the choice of projection index. There are certain demands of a good projection index [11]: Robust to deviations, Approximately affine invariant, Consistent, Simple enough to permit quick computation even for large data sets, and Others.

In addition to lowering the dimension, projection pursuit also allows us to overcome problems associated with sparsity of data in high dimensions [12] which is also termed "curse of dimensionality" [13]. With increasing dimension, the need for more and more data to meet the requirement of sufficient data increases like a curse. Many techniques fail to perform well under the conditions of sparse data. There are also situations where the number of variables (P) is much higher than the amount of data or number of observations (n). References [14,12] gave strong heuristic arguments indicating that a projection is less interesting the more it is normal.

Any method that effectively helps to reduce the dimension from high to low can be treated as a form of projection pursuit. There exist different forms of projection pursuit: principal component analysis, linear discriminant analysis, and projection pursuit regression etc. In this paper we shall only concentrate on projection pursuit by linear discriminant function.

- ***Linear discriminant analysis***

Discriminant analysis determines some optimal combination of variables called the discriminant function so that the first function provides the most overall discrimination between groups; the second provides the second most overall discrimination, and so on. The functions will be orthogonal, that is their combinations to the discrimination between groups will not overlap. The first function picks up the most variation; the second function picks up the greatest part of the unexplained variation, etc. The maximum number of discriminant functions will be equal to the degree of freedom or the number of variables in the analysis, which is smaller. The discriminant function is what is used in classifying objects in the groups in which they had the highest classification score.

In this paper the first discriminant function will be used to reduce the p-dimensional data to one dimensional data. This reduced form of the data will then be fixed into the point group transvariation which is then used in classify new subjects with unknown group into one of the existing groups.

2.3. Allocation Based on Distance

Given two independent trainng samples $X_1,\ldots, X_m$ and $Y_1,\ldots,Y_n$ from populations Π_x and Π_y , respectively, defined on R^P ($P\geq1$), a new observation $Z=z$ is classified in Π_x if

$$\left|\hat{u}_{opt}\, z - m_x\left(\hat{u}_{opt}\right)\right| < \left|\hat{u}_{opt}\, z - m_y\left(\hat{u}_{opt}\right)\right|$$

Otherwise classify it in Π_y. Here $m_x\left(\hat{u}_{opt}\right)$ and $m_y\left(\hat{u}_{opt}\right)$ are centres of the two projected groups. One may take either the mean or the median as a measure of centres. We will consider mean in this study. Hereafter the classifier obtained using this allocation method will be referred to as a Transvariation – Distance (TD) classifier. In this article this classifier was used in classification after the data have been transformed [15].

2.4. Data presentation

To compare the discriminant procedures discussed in section one, two different data sets are studied in this article: simulated and real life data sets.

- ***Simulation Data***

The discriminant procedures by data depth and projection pursuit (by linear discriminant function) are evaluated using six simulated data sets. The procedures are then evaluated on data sets generated from a variety of specifications with different dimensions P = 2,3,4,5,6, and 7; the same number of groups $g = 2$; and different size of samples $n.$ In all the cases the class distributions are binomial, but the generated data sets differ in size and probability of success of the groups. The various specifications of the data sets are presented below:

Table 1: Data specifications and their optimal probability of misclassification (pmc)

S/N	Sample Size	No. of variables	No. of trials		Probability of success		P(MC)
			Group X	Group Y	Group X	Group Y	
1	120	5	25	40	0.5, …,0.5	0.7,…,0.7	0.3300
2	100	4	50	80	0.6,…,0.6	0.3,…,0.3	0.5883
3	80	2	40	50	0.5,…,0.5	0.5,…,0.5	0.5000
4	50	7	30	60	0.8,…,0.8	0.6,…,0.6	0.698
5	40	3	20	30	0.4,…,0.4	0.6,…,0.6	0.352
6	10	6	25	30	0.3,…,0.3	0.6,…,0.6	0.365

- ***Real data***

The real life data we used are obtained from Ph.D seminar paper presented at the Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria by [16], sourced from Nigeria Institute for Oil Palm Research with emphasis on the characteristics and yield of two different progenies of palm tree. The characteristics considered for classification are leaf count in the nursery, height in nursery, leaf count in field, height in field, canopy spread in meters, sex ratio (%), and yield in 4 years. The number of sample size studied is forty.

3. Illustration and Result of the Study

In projection pursuit approach we started by finding the linear discriminant function of each original data set. The resulted first discriminant function was used to sweep the p-dimensional data space R^p to one dimension R. With reduced data space, Transvariation distance classifier that is univariate statistical tool was then used to cross validate the training samples.

For data depth method because our simulated data do not have different scale of measurement and of its computational ease, spatial or L_1 depth is used in this work to find the depth of each data point. With the data depth of entire data calculated objects were correctly classify into the population they belong to. For the real life data, $\sum_x^{-\frac{1}{2}}(x - X)$ were used in place of $x - X$ in (8) to make L_1 depth classifier affine invariant.

From the results of the analyses we obtained the probability of miss classified data in Table 2 below.

Table 2: Estimated probability of misclassification according to sample size

Sample size (validated data)	60	50	40	25	20	5	Life Data
PMC for data depth	0.0000	0.0123	0.0167	0.025	0.0500	0.0000	0.2500
PMC for PP	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

4. Summary and Conclusion

The performances of the procedures were evaluated by the misclassification probabilities obtained using apparent error rate of the validated data sets. Based on our observations during iteration and our findings after the analysis (see Table 2), we conclude that projection pursuit (by linear discriminant function) has highest predictive power over the method of data depth we considered.

Acknowledgements

Our sincere gratitude goes to all the authors whose works are cited in this paper.

References

[1] J. Tukey. (1975) Mathematics and picturing data. In R. James, ed., Proceedings of the 1994 International congress of mathematicians. Vancouver vol 2, 523-531.

[2] D. L.Donoho and M. Gasko. (1992). “Breakdown properties of multivariate location estimates based on halfspace depth and projected outlyingness.” Annals of Statistics,20(4), pp. 1803-1827.

[3] P. Chaudhuri. (1996). “On a geometric notion of quantiles for multivariate data.” Journal of American Statistical Association 91, pp. 862-872.

[4] R. Y. Liu. (1992). Data depth and multivariate rank tests. In L_1-statistical analysis and related methods (Neuch $\hat{a}$ tel, 1992), pages 279-294. North-Holland, Amsterdam.

[5] Y. Zuo and R. Serfling. (2000). “General notions of statistical depth function.” Annals of Statistics 28, 461-482.

[6] P.J. Rousseew and M. Hubert. (1990). “Regression depth (with discussion).” Journal of the American Statistical Association. 94, pp. 388-433.

[7] H. Oja. (1983). “Discriptive statistics for multivariate distributions.” Statistics and Probability Letters. 1, pp. 327-332.

[8] R. Y. Liu. (1990). “On a notion of data depth based on random simplices.” Annals of Statistics. 18(1):405- 414. Analysis, 20:669-687.

[9] V. Barnet. (1976). “The ordering of multivariate data.” Journal of the Royal Statistical Society ser. A 139, pp. 318- 355

[10] Z. Y. Chen. “Robust linear discriminant procedures using projection pursuit methods.” PhD Dissertation, University of Michigan, USA, 1989.

[11] C. Posse. (1995). “Projection pursuit exploratory data analysis.”Computational Statistics and Data lib.dr.iastate.edu/cgi/viewcontent.cgi?article=2440&content=rtd

[12] P. J. Huber. (1985). “Projection pursuit.”.Annals of Statistics, 13920,pp. 435-525.With discussion. projecteuclid.org/Euclid.aos/1176349530

[13] M. Goldstein. (1987). [a review of multivariate analysis]. Comment. Statistical Science, 2(4):418-420.

[14] M. C. Jones. (1983). The projection pursuit algorithm for exploratory data Analysis, PhD Thesis, University of Bath.

[15] C. Gini. (1916). II concetto di transvariazione e le sue prime applicazioni. Giornale degli economisti Rivista di statistica.

[16] D.D. Ekezie. “A biometric study of oil palm (Elaeis guneensis Jacq) nursery characteristics and yield by the method of multivariate analysis.” Ph.D seminar paper, Nnamdi Azikiwe University Awka, Nigeria, Oct. 2010.