International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Multivariate Rank Discriminant Classifier of Small Sample

Evelyn Nkiruka Okeke [a*], Joseph Uchenna Okeke [b], Sidney I. Onyeagu[c]

[a,b] Department of Mathematics and Statistics, Federal University Wukari, Nigeria

[c] Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria.

[a]evelyn70ng@yahoo.com

[b]uche70ng@yahoo.com

[b]sidneyonyeagu@yahoo.ca

Abstract

This article studied discriminant analysis procedure that is based on multivariate ranking with emphasis on Spatial or L_1 depth classifier using Eviews and SPSS computer packages. The performance of the classifier is assessed using both simulated and real life data. The result of the study revealed that the classifier is optimal in classifying observations into one of the two pre-defined groups.

Keywords: Data depth; Spatial or L_1 depth; linear discriminant analysis; nonparametric discriminant analysis; Probability of misclassification (PMC)

1. Introduction

Discriminant analysis is one of the data mining techniques used to discriminate a single classification variable using multiple attributes. Discriminant analysis assigns observations to one of the pre-defined groups based on the knowledge of the multi-attributes.

* Corresponding author. Tel.: +23408181278549

E-mail address: velyn70ng@yahoo.com.

When the distribution within each group is multivariate normal, a parametric method can be used to develop a discriminant function using a generalized squared distance measure. Linear discriminant analysis (LDA) based methods (parametric methods) suffer a fundamental limitations originating from the parametric nature of covariance matrices which are based on Gaussian distribution assumption. The performance of these methods is not optimal when the actual distribution is Non-Gaussian. LDA is guaranteed to find the best direction when each class has a Gaussian density with a common nonsingular covariance matrix. Non-parametric discriminant methods are used in finding important discriminant directions without assuming that the class densities belong to any particular parametric family.

A discriminant analysis procedure uses all the variables that the training data contains and uses their correct classification information to create a discriminant rule or a classifier. Classification is done by feeding new observations into this classifier and getting the group membership to which the new observations belong. The performance of a classifier could be evaluated by estimating probabilities of misclassification (PM) of new observations in the validation data. When two or more classifiers performed equally well in terms of their probability of misclassification (PMC), classifier that is robust to deviations is more preferred.

A lot of research works have been done in the field of discriminant analysis in an effort to come up with classifiers that are robust to violations of certain assumptions. Most of the work done to make LDA robust concentrated on replacing the measures of location and scatter of LDA classifiers by their robust counterparts. However when the covariance structure is singular or close to it, the later methods may fail to be optimal. To solve singularity problem, projection pursuit approach has come up as a remedy. This method aimed at reducing a high dimensional data set to low dimension so that the statistical tool for the low dimensional data can be applied. It is observed that most of the projection pursuit methods fail in the presence of multivariate outlier. As a solution to some of these problems, discriminant procedure based on multivariate rank was proposed by [1]. This procedure works in high-dimensional spaces and aimed at reducing the dimension to one. In this article we are going to investigate the performance of this method through several simulations and by applying it to a real data set.

This paper would be organizes into Sex Sections. Section one contains the introduction. In Section two we have the description of discriminant procedures based on multivariate rank. Section three contains the simulated and real data sets. The illustration of L_1 depth classifier discussed in Section two and the result of the data are presented in Section four using one of the generated samples. The summary and conclusion of the study are in section five. References are in Section seven.

2. Procedures Based on Multivariate Ranking

In univariate setting, the statistical method that used ranking-based nonparametric techniques like Mann-Whitney test, Kruskal-Wallis test, Friedman test and others do not depend on restrictive distributional assumption and hence are robust to deviation from these assumptions. For higher dimension, and alternative to projection pursuit is the idea of data depth which is a multivariate version of rank [2,3]. Data depth are used to measure the “centrality” of a given multivariate sample point with respect to its underlying distribution (examples [4], [5], and [6]). In particular, a depth function assigns higher values to point that are more central with respect to a data cloud. This naturally gives a center-outward ranking of the sample point. Popular depth functions available in the literature include:

- Mahalanobis depth as in [7,8]
- Half-space depth as in [9]
- Simplicial depth as in [10]
- Majority depth as in [11]
- Projection depth as in [12]
- Spatial or L$_1$ depth as in [13,1]

Definitions of some of the more Popular Depth functions are

1. The Spatial or L$_1$ depth is given by

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

Where X~F_X, and $\|.\|$ is the Euclidean norm.

2. The Mahalanobis depth function is given by

$$MD(X,F_X) = \left[1 + \left(x-\mu_{F_X}\right)'\textstyle\sum_x^{-1}\left(x-\mu_{F_X}\right)\right]^{-1} \qquad (2)$$

Where μ_{F_X} and Σ_x are the mean vector and covariance matrix of F_X respectively. The sample version of MD is obtained by replacing μ_{F_X} and Σ_x^{-1} with $\bar{X}$ and S_x^{-1}.

3. The Half-space depth function is given by

$$HD(X;F_X) = {}^{inf}_{\ H}\{P(H): H\ is\ a\ close\ half-space\ in\ R^p, X \in H\} \qquad (3)$$

It turns out that Turkey depth $T_X(c)$ is the half-space depth of c in one dimension with respect to the population F_X, that is, $T_X(c) = HD(c,F)$. Half-space depth is sometimes referred to as Tukey depth.

$0 \le DF \le 1$,where DF is depth 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 functionDF. Let f be the class of distributions on the Borel sets of R^p : a statistical depth function is a bounded, nonnegative mapping $D: R^p \times f \to R$.

There are certain properties that are desired of depth functions [5,14]:

- Affine invariance: the depth of a point $X \in R^p$ should not depend on the underlying coordinate system or in particular, on the scale of the underlying measurements.

$$DF(AX+b;F_{AX+b}) = DF(X;F_X) \qquad (4)$$

- Maximality at center: for a distribution having a uniquely defined center (e.g; the point of symmetry with respect to some notion symmetry) , the depth function should attain maximum value at this center. If μ is the center of F, then

$$DF(\mu, F_X) = {}^{sup}_{X\epsilon R^p} DF(X;F_X) \qquad (5)$$

- Monotonicity relative to deepest point: as a point $X \in R^p$ moves away “from the deeper point” (the point at which the depth function attains maximum value; in particular, for a symmetric distribution, the center) along any fixed ray through the center, the depth at X should decrease monotonically.
- Vanishing at infinity: the depth of a point X should approach zero as $\|X\|$ approaches infinity

$$DF(X;F_X\,) \to 0 \; as \; \|X\| \to \infty$$

The interested reader may find an extensive list of depth function along with their definition in [8,5,15]. Among the numerous depth functions that are in existence, Mahalanobis depth and Spatial or L$_1$ depth are two of the most attractive ones due to their ease in computation. They can be computed exactly for any dimension. The computation of many other depth functions may require algorithms that provide only approximations. This is especially true for higher dimensional data. For example, one usually has to construct very complicated approximation algorithms to compute the half-space depth of points in three or higher dimensions.

Taking advantage of this notion of ordering multivariate data in a center-outward manner, [1] proposed the maximum L$_1$ depth classifier that uses the discriminant function

$$S(z, F_X, F_Y) = D(z;F_Y\,) - D(z;F_X\,) \qquad (6)$$

$$= \left\| E_{F_Y}\left\{\frac{z-Y}{\|z-Y\|}\right\}\right\| - \left\| E_{F_X}\left\{\frac{z-X}{\|z-X\|}\right\}\right\|$$

$$= \left\| \int_{R^p} \frac{z-Y}{\|z-Y\|} dF_Y(y) \right\| = \left\| \int_{R^p} \frac{z-X}{\|z-X\|} dF_X(x) \right\| \qquad (7)$$

The new observation $Z = z$ is then classified in π_x if $S(z, F_X F_Y) > 0$ and in π_y otherwise. Despite its computational ease, a major drawback of this classifier is that it lacks affine invariance because L$_1$ depth is not affine invariant. However, it can be made affine invariant by taking $\sum_x^{-\frac{1}{2}}(z-X)$ and $\sum_x^{-\frac{1}{2}}(z-Y)$ in place of $z-X$ and $z-Y$, respectively, in equation (7) as in [13]; [15]. Note that one can use any affine equivariant estimator of $\sum_x$ and $\sum_Y$ when computing the discriminant function. An alternative method of obtaining affine invariance is to scale the data along its principal component direction (PC-scaling) as given in [16]. One could use robust principal component (e.g. robust PCA given by [17]) or scale the data with the robust estimate of covariance structure which will make the L$_1$ depth function affine invariant in addition to making it robust against deviation.

For practical purposes, given two independent training samples $x_1, x_2, \dots, x_m$ and $y_1, y_2, \dots, y_n$ from π_x and π_y, respectively, defined on $R^p (p \geq 1)$ the sample version of $D(z; F_X)$ and $D(z; F_Y)$ given in (7) can be found by replacing the empirical cdf in pace of F_X and F_Y resulting in the sample version of $S(z, F_X, F_Y)$ given by

$$S(z, F_{xm}, F_{yn}) = \int_{R^p} \frac{z - y_j}{\|z - y_j\|} dF_{yn}(y) - \int_{R^p} \frac{z - x_j}{\|z - x_j\|} dF_{xm}(x)$$

$$= \left\| \frac{1}{n} \sum_{j=1}^{n} \frac{z - y_j}{\|z - y_j\|} dF_y(y) \right\| - \left\| \frac{1}{m} \sum_{i=1}^{m} \frac{z - x_i}{\|z - x_i\|} dF_x(x) \right\| \qquad (8)$$

It must be noted that the maximum L$_1$ depth classifier is in the class of classifier known as maximum depth classifiers [15] in that any depth function DF can be used in place of the L$_1$ depth function. The optimality of the classifier is dependent on the choice of the depth function. The choice of depth function could be based on various properties like robustness. QDF is obtained if Mahalanobis depth is used in place of the L$_1$ depth in (8).One would assign the new observation $Z = z$ in π_x if $S(z, F_{xm}, F_{yn}) > 0$ and in π_y otherwise.

3. Data Presentation

To evaluate the discriminant procedure discussed in section two, two data sets are studied in this paper; simulated and real life data.

3.1 Simulation Data

The discriminant procedure by multivariate ranking considered is evaluated on 6 simulated data sets. The estimators is then evaluated on data sets generated from a variety of settings 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 sizes and probability of successes of the groups. The various combinations 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	100	5	25	40	0.5, …,0.5	0.7,…,0.7	0.3300
2	80	4	50	80	0.6,…,0.6	0.3,…,0.3	0.5883
5	60	2	40	50	0.5,…,0.5	0.5,…,0.5	0.5000
6	50	7	30	60	0.8,…,0.8	0.6,…,0.6	0.698
9	20	3	20	30	0.4,…,0.4	0.6,…,0.6	0.352
10	10	6	25	30	0.3,…,0.3	0.6,…,0.6	0.365

3.2 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 [18], 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 40.

4. Illustration and the Result of the Study

Because of the nature of our data and its computational ease, spatial or L_1 dept classifier is used in this work to find the depth of each data Point. We started by dividing the data set into two equal parts to have training and the validated data for testing the performance of our classifier. Using the training data we first centered all the data points: This is done by finding the deviation of each data point from their means using Eviews 3.1 statistical package. The Euclidean norm of each data point is calculated to reduce the p-dimensional X an Y data sets to a unit ball. For the real life date, $\sum_x^{-\frac{1}{2}}(z-X)$ and $\sum_x^{-\frac{1}{2}}(z-Y)$ were used in place of $z-X$ and $z-Y$ to make L_1 depth classifier affine invariant. With the centered data and the calculated norm the depth of the data points in the X and Y sample are computed by the combination of Eviews and SPSS computer packages. The calculations revealed that all the depths in X vector and Y vector spaces obtained belong to R^+ space. With the data depths, L_1 depth classifier is used to classify the validated data sets and the following results are obtained.

Table 2: estimated probability of misclassification according to sample size

Sample size (validated data)	50	40	30	25	10	5	Life Data
PMC	0.0000	0.0125	0.0167	0.02	0.0500	0.0000	0.2500

5. Summary and Conclusion

The L_1 depth classifier was evaluated in a prediction context. The performance of the classifier was evaluated by the misclassification probabilities obtained using apparent error rate of the validated data sets. From table 3 it is clear that the classifier in optimal in classifying data with number of variables more than the sample size. Apart from real life data where there is replacement in the formula, the PMC is not large enough in all other cases to conclude that the classifier is not optimal.

Acknowledgement

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

References

[1] R. Jornsten. Clustering and classification based on the L_1data depth. *Journal of Multivariate Analysis*, 90 (1), Pp 67-89, 2004.

[2] W. F. Eddy. Ordering of multivariate data. *In Computer Science and Statistics*: The Interface (L. Billard, ed). North-Holland: Amsterdam. 1985, Pp 25-30.

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

[4] R. Y. Liu, J. M. Parelius, and K. Singh. “Multivariate analysis by data depth:descriptive statistics, graphics and inferences.” With discussion and a rejoinder by Liu and Singh .*Annals of Statistics,* 27(3), Pp 783-858. 1999.

[5] Y. Zuo, and R. Serfling.” General notion of statistical depth function.” *Annals of Statistics*, 28(2), Pp 461-482, 2000

[6] K. Mosler. Multivariate dispersion, central regions and depth, volume 165 of Lecture Notes in Statistics. Springer-Verlag, Berlin. The lift zonoid approach. 2002.

[7] P. C. Mahalanobis. “On the generalized distance in statistics.” Proceeding of the National Institute of Science of India.1936, Pp 49-55.

[8] R. Y. Liu, and K. Singh. “A quality index based on data depth and multivariate rank tests.” *Journal of American Statistical Association*. 88(421), Pp 252-260, 1993.

[9] J. Tukey. Address to international congress of mathematics. Vancouver. 1974.

[10] R. Y. Liu.” On a notion of data depth based on random simplices.” *Annals of Statistics*, 18(1), Pp 405- 414, 1990.

[11] K.Singh. A notion of majority depth. Technical Report. Department of Statistics, Rurtgers University. 1991.

[12] D. Donoho. Breakdown properties of multivariate location estimators. PhD Quality paper. Department of Statistics, Harvard University. 1982.

[13] Y. Vardi, and C. H. Zhang . “The multivariate L_1-median and associated data depth.” *Proc. Natl. Academy of Science.* USA, 97(4), Pp 1423-1426 (electronic), 2000.

[14] R. Hoberg “Cluster analyses.” In *Klassifikation und Datentiefe*. Eul, Lohmar. 2003.

[15] A. K. Ghosh, and P. Chaudhuri. “On maximum depth, and related classifiers.” *Scand. J. Statist.,* 32(2), Pp 327-350, 2005.

[16] J. Hugg, R. Rafalin, K.Seyboth, and D. Souvaine. An experimental study of old and new depth measures. Springer-Verlag Lecture Notes in Computer Science, New York, Pp 51-64, 2006.

[17] C. Croux, P. Filzmoser, and M. R.Oliveira. “Algorithms for projection pursuit:robust principal component analysis .” *Chemometrics and intelligent Laboratory Systems,* 87(2), Pp 218-225, 2007.

[18] 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, 2010.