Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 15 Issue 6 Version 1.0 Year 2015
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Discriminant Analysis by Projection Pursuit

By Okeke, Joseph Uchenna, Onyeagu, Sidney & Okonkwo, Evelyn Nkiruka
Nnamdi Azikiwe University, Nigeria

Abstract- A non-parametric discriminant analysis (projection pursuit by principal component analysis) is discussed and used to compare three robust linear discriminant functions that are based on high breakdown point (of location and covariance matrix) estimators. The major part of this paper deals with practical application of projection pursuit by principal component. In this study 10 simulated data sets that are binomially distributed and a real data set on the yield of two different progenies of palm tree were used for comparisons. From the findings we concluded that the non-parametric procedure (projection pursuit by principal component) have the highest predictive power among other procedures we considered. S-estimator performed better than the other two estimators when real data is considered, while MCD estimator performed better than MWCD estimator.

Keywords: *discriminant analysis (DA), principal component analysis (PCA), projection pursuit, minimum covariance determinant (MCD), minimum within covariance determinant (MWCD), and s-estimator.*

GJSFR-F Classification : *FOR Code : 010499*

Strictly as per the compliance and regulations of :

Notes

Discriminant Analysis by Projection Pursuit

Okeke, Joseph Uchenna [α], Okeke, Evelyn Nkiruka [σ], and Onyeagu, Sidney [ρ]

Abstract- A non-parametric discriminant analysis (projection pursuit by principal component analysis) is discussed and used to compare three robust linear discriminant functions that are based on high breakdown point (of location and covariance matrix) estimators. The major part of this paper deals with practical application of projection pursuit by principal component. In this study 10 simulated data sets that are binomially distributed and a real data set on the yield of two different progenies of palm tree were used for comparisons. From the findings we concluded that the non-parametric procedure (projection pursuit by principal component) have the highest predictive power among other procedures we considered. S-estimator performed better than the other two estimators when real data is considered, while MCD estimator performed better than MWCD estimator.

Keywords: discriminant analysis (DA), principal component analysis (PCA), projection pursuit, minimum covariance determinant (MCD), minimum within covariance determinant (MWCD), and s-estimator.

I. Introduction

The problem of discriminant analysis arises when one wants to assign an individual into one of k groups on the basis of a p-dimensional characteristic vector. In classical discriminant analysis, the populations under study are, assumed to be, normally distributed with equal covariance matrices. But in some practical situations, problems with data that are categorical, mixed, sparse, and contaminated abound. In these situations, classical discriminant analysis will not be optimal in classifying objects into the existing populations. The need for robust methods that will be optimal in classifying objects becomes necessary.

Robust methods for discriminant analysis have been proposed by many authors; Todorov et al (1990) replaced the classical estimates of linear and quadratic discriminant functions by MCD estimates, Chork and Rousseeuw (1992) used MVE instead, Hawkins and Mclachlan (1997) defined MWCD especially for the case of linear discriminant analysis, He and Fung (2000) and Croux and Dehon (2001) used S estimates, Hubert and Van Driessen (2004) applied the MCD estimates computed by the FAST MCD algorithm, Todorov and Pires (2007) used M-iteration described by Woodruff and Rocke (1996). Most of these works concentrated on replacing the classical mean vectors and covariance matrices by their robust counterpart. 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. Polzehl (1993) studied discriminant analysis based on projection pursuit density estimation and chose his projection to minimize estimates of the expected overall loss in each projection pursuit

Author α: Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria.
Author σ ρ: Department of Statistics, Anambra State University, Uli, Nigeria. e-mail: evelyn70ng@yahoo.com

stage. According to him cross-validation techniques are used to avoid overfitting effect and at last he concluded that his procedure competes favorably with other classification methods in situations where parametric approaches are not flexible enough and when sample sizes are too small to use fully non-parametric procedure.

Pires and Branco (2010) studied projection pursuit estimator of the normalized discriminant vector induced by robust estimators of location, T, and the univariate estimator of scale, S and discovered that under contaminated data their method performed well and is strong competitor of other methods they studied.

Gunduz and Fokoue (2015) in their work explored and compared the predictive performance of robust classification and robust principal component analysis and applied it to variety of large small data sets. They also explored the performance of random forest by way of comparing and contrasting the differences of single model methods and ensemble method. Their work revealed that random forest although not robust to outlier substantially outperforms the existing techniques specifically design to achieve robustness.

In this paper we proposed projection pursuit by method of principal component because it allows prior standardization that is important for badly scaled data. This method was compared with the robust linear estimates: MCD and MWCD estimates obtained using Mahalanobis distance of the data points on 10 simulated data sets that are distributed binomially with a view of coming out with classifier with the highest predictive power.

II. Principal Component and Projection Pursuit (Nonparametric DA)

In many fields of research, principal component analysis (PCA) is used as an efficient tool for providing an informative and low-dimensional representation of multivariate data in which features in the data such as clustering, skewness and outliers can be easily detected (Bolton and Krzanowski 1999). PCA does not necessarily afford the 'best" view of the data structures and it may miss other interesting characteristics of the data. With this in mind, much research have been done in recent years on approaches to identifying projections that display, particularly, "interesting" features of the data. These techniques go under the generic name "projection pursuit" (Friedman and Tukey 1974). In projection pursuit the dimension that will give the required projection and the criterion that will find projections of the desired structure when optimized are studied.

Local optimization of the criterion over all projections of the required dimensionality yields "interesting" projections of the data that can be graphically displayed. The projection is usually chosen to be one, two, or three dimensional for convenience.

Projection pursuit indices are diverse but the construction of most of them is motivated by consideration of central limit theorem results. Diaconis and Freedman (1984) showed that projected subspaces of high-dimensional data converged, weakly in probability, to normality. Consequently, most projection pursuit indices are developed from the standpoint that normality represents the notion of "uninterestingness" (Huber 1985). These indices are thus optimized to find projections showing departures from normality.

PCA can be viewed as a particular case of projection pursuit in which the index of "interest" is the variance of the data, which is maximized over all unit length projections. In this paper, the first principal component was used to transform the *p*-

dimensional data space to one, so that we can apply the statistical tool for one dimensional data space.

a) Procedures of PCA

- Taking the whole dataset ignoring the class labels
- Computing the p-dimensional mean vector
- Computing the Covariance Matrix
- Computing eigenvectors and corresponding eigenvalues
- Sorting the eigenvectors by decreasing eigenvalues
- Choosing g eigenvectors with the largest eigenvalues
- Transforming the samples onto the new g subspace(s)

$$Y_i = a'X$$

Where $i = 1, \ldots, g$ with $g \leq p$; a is the selected eigenvectors

Note: If the p variables of the original data are measured with different scales correlation matrix is used in place of covariance matrix.

b) Allocation Based on Point-Group Transvariation

This is a nonparametric allocation option suggested by Montanari (2004) which is based on the ranking of new observations among two samples used for classification. This utilizes the point group transvariation defined by Gini (1916) between the projected new observation and the projected X and Y. Allocate a new observation $Z = z$ into Π_x if $T_x(z)$ ¿ $T_y(z)$; otherwise, it is assigned to Π_y where

$$T_x(z) = \frac{1}{n}\sum_{i=1}^{n} I\{(\hat{u}'_{opt} X_i - \hat{u}'_{opt} z)\}[m_x(\hat{u}'_{opt}) - \hat{u}'_{opt} z] < o\} \quad 2.1$$

The formula for $T_y(z)$ will be obtained when you replace all x in (2.1) with y.

$$\hat{u}_{opt} = \arg\min_{\|u\|=1} \left\{ \frac{1}{mn}\sum_{i=1}^{m} \sum_{j=1}^{n} I\{(u'x_i - u'y_j)(m_x(u) - m_y(u)\ ) < o \right\} \quad 2.2$$

where $m_x(u)$ and $m_y(u)$ are the locations of the two projected X and Y samples $u'x$ and $u'y$ respectively.

$\hat{u}_{opt}$ is found using projection pursuit

The allocation scheme is based on ranks. It gets rid of the non optimality problem that transvariation distance (TD) has when skewed distributions are considered. Although the allocation scheme makes this method completely non parametric and works better than TD for skewed distributions, it does not perform as well for data with unequal sample sizes. This is due to the fact that an equal prior restriction is imposed by counting and we neglect group two (one) when we find the ranking of the new point in group one (two). So the priors are not necessarily taken into account and the effect shows in the misclassification error rate especially when the sample sizes are unequal.

III. Methodology

The discriminant procedures considered would be evaluated on 10 simulated data sets of two groups with different specifications distributed binomially and a real life

data. The real life data we used were obtained from Ph.D. seminar paper presented at the Department of Statistics, Nnamdi Azikiwe University, Awka, by Ekezie (2010). The data were from Nigeria Institute for Oil Palm Research and is on the characteristics and yield of two different progenies of palm tree. Table 1 below shows 10 simulated data sets and their specifications and optimal probability of misclassification.

Table 1 : Data Specifications and their Optimal Probability of Misclassification P(MC)

S/N	Sample Size	No. of variables	No. of trials		Probability of success		P(MC)
			Group X	Group Y	Group X	Group Y	
1	50	5	30	40	0.5, …,0.5	0.7,…,0.7	0.3300
2	45	3	60	70	0.4,…,0.4	0.5,…,0.5	0.426
3	40	4	50	100	0.4,…,0.4	0.3,…,0.3	0.5883
4	35	6	50	36	0.5,…,0.5	0.5,…,0.5	0.5000
5	30	6	30	50	0.5,…,0.5	0.5,…,0.5	0.5000
6	25	7	30	60	0.8,…,0.8	0.6,…,0.6	0.698
7	20	4	30	20	0.4,…,0.4	0.6,…,0.6	0.352
8	15	6	30	50	0.5,…,0.5	0.5,…,.5	0.5000
9	10	4	20	30	0.4,…,0.4	0.6,…,0.6	0.352
10	5	2	25	30	0.3,…,0.3	0.6,…,0.6	0.365

The steps we follow in computing the projection pursuit discriminant procedure are explained in detail in section 3.1 and 3.2.

a) Projection Pursuit (by Method of Principal Component)

We started by pooling the two samples of X and Y. The pooled data was centered. The principal component analysis of pooled data was computed using Minitab computer package. From the computed result the first principal component was chosen for the final analysis. The coefficients of the variables $\hat{u}_{opt}$, which are the projection direction that maximizes the separation of the data between two groups were made to be orthogonal. The first principal component with orthogonal coefficient was used to sweep the p-dimensional data space R^p to one dimension R data space. With reduced data space, point-group transvariation probability that is univariate statistical tool was then used to cross validate the training samples.

b) Probability of classification

In order to evaluate the performance of this method in classification of future observations we estimate the overall probability of misclassification. A number of methods to estimate this probability exist in the literature but in this study we used apparent error rate (known also as resubstitution error rate or reclassification error rate). This is a straightforward estimator of the actual (true) error rate in discriminant analysis and is calculated by applying the classification criterion to the same data set from which it was derived and counting the number of misclassified observations. If there are plenty of observations in each class the error rate can be estimated by splitting the data into training and validation sets. The first one is used to estimate the discriminant rules and the second to estimate the misclassified error.

IV. Results

The non-parametric discriminant method was evaluated with regard to its performance assessed by misclassification probabilities using 11 data sets. This method

competes favorably with robust linear estimators: MCD estimator, MWCD estimator, and S-estimator with the following misclassification probabilities.

Table 2.below contain the results of the 10 simulated data in terms of their misclassification probabilities,

Table 2 : Estimated Probability of Misclassification According to Sample Size

Sample size	MCD Estimator	MWCD Estimator	S-estimator	PP (PCA)
50	0.0000	0.0000	0.0000	0.0000
45	0.0111	0.0111	0.0111	0.0000
40	0.0000	0.0000	0.0000	0.0000
35	0.0142	0.0142	0.0142	0.0000
30	0.0000	0.0000	0.0000	0.0000
25	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000	0.0000
15	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000
Life Data	0.0750	0.1000	0.0500	0.0125

The result of real life data showed that projection pursuit has the highest predictive power with P(MC) of 0.0125, followed by S-estimator with P(MC) of 0.05, MCD with P(MC) of 0.075 and then, MWCD estimator which have P(MC) of 0.1.

Considering the computational ease and P(MC), projection pursuit (by PC) performed better than the other three procedure. S-estimator performed better than the other two estimators when real data is considered, while MCD estimator performed better than MWCD estimator. Although the quality of the estimates (for robust linear estimators) is important since it entirely determines the robustness of the discriminant rule towards outlier. In our study we only concentrated on the predictive power of the procedures leaving the other aspects for further work.

V. Conclusion

Based on our observations during iteration and our findings after the analysis, we conclude that nonparametric classification procedure (projection pursuit by principal component) has highest predictive power among other procedures we considered.

References Références Referencias

1. Bolton, R. J., and Krzanowski, W.J. (1999). "A characterization of principal components for projection pursuit", The American Statistician, 53(2):108-109.
2. Chork, C. and Rousseeuw, P. J. (1992). "Integrating a high breakdown option into discriminant analysis in exploration geochemistry", Journal of Geochemical Exploration, 43, 191-203.
3. Croux, C., and Dehon, C. (2001). "Robust linear discriminant analysis using S-estimators", Canadian Journal of Statistics, 29(3):473-493.
4. Diaconis, P., and Freedman, D. (1984). "Asymptotic of graphical projection pursuit", The Annals of Statistics, 12:793-815.

5. Friedman, J. H., and Turkey, J. W. (1974). "Projection pursuit for exploratory data analysis", IEEE Transactions on Computers, 23(9): 881-890.
6. Gunduz, N, and Fokoue, E. (2015). Robust classification of high dimension low sample size data. arxiv.org e-print archive.
7. Hawkins, D. M., and McLachlan, G. J. (1997). "High-breakdown linear discriminant analysis" .Journal of American Statistical Association, 92(437): 136-143.
8. He, X., and Fung, W. K.(2000). "High breakdown estimation for multiple populations with application to discriminant analysis", Journal of Multivariate Analysis, 72(2):151-162.
9. Huber, P. J. (1985). "Projection pursuit", Annals of Statistics, 13920: 435-525.
10. Hubert, M. and Van Driessen, K. (2004). "Fast and robust discriminant analysis", Comput. Statist. Data Anal., 45(2):301-320.
11. Pires, A. M. and Branco, J. A. (2010). "Projection pursuit approach to robust linear discriminant analysis", Journal of Multivatiate Analysis, 101(10): 2464-2485.
12. Polzehl, J., (1993). Projection pursuit discriminant analysis, Center for Operations Research and Economics.
13. Todorov, V. and Pires, A. M. (2007). "Comparative performance of several robust linear discriminant analysis methods", REVSTAT-Statistical Journal, 5(1):63-83.
14. Todorov, V, Neykov, N., and Neytchev, P. (1990). "Robust selection of variable in the discriminant analysis based on MVE and MCD estimators", In: "Proceedings in Computational Statistics', COMPSTAT", Physica Verlag, Heidelberg.
15. Woodruff, D. L. and Rocke, D.M. (1994). "Computable robust estimation of multivariate location and shape in high dimension using compound estimators", Journal of the American Statistical Association, 89, 888-896.

Notes