A New Graphical Multivariate Outlier Detection Technique Using Singular Value Decomposition

Nishith Kumar[1] , Mohammed Nasser[2]

[1]Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh;
[2]Department of Statistics, Rajshahi University, Rajshahi, Bangladesh.

Abstract

It is well recognized that even a high quality data set tends to contain a remarkable percent of unusual observations. There are various techniques to detect multivariate outliers. But they are highly depends on mathematics. We propose a new graphical technique to detect multivariate outliers which is easy to understand without hard mathematics, it can be applied to data for both supervised and unsupervised learning, it can be directly applied to separate extreme outliers from general outliers.

Keywords: Outliers, Singular Value Decomposition, Principal Component Analysis.

1. Introduction

Outliers detection problem is as old as statistics. Outliers present in both supervised and unsupervised learning of multivariate data set. Outlier can destroy our analysis. So outliers detection should be the first target of a statistician or a researcher. In this paper we mainly proposed a method for outliers detection in a multivariate data set and apply this in several well known data sets.

In multivariate statistics we detect outliers by Mahalalanobis Distance but it is not suitable for outlier detection since it is dependent on nonrobust mean and covariance matrix. To detect outliers, Rousseeuw and Leroy (1987) proposed robust distances, which are robustified versions of the Mahalanobis distances

$$RD_i = \sqrt{(x_i - t_n)' C_n^{-1} (x_i - t_n)}$$

with (t_n, c_n) robust estimates of location and scatter. Observations with *RDi* bigger than the critical value $\sqrt{\chi_k^2, 0.975}$ can be considered as potential outliers (Rousseeuw and Van Zomeren, 1990).

Regression analysis is one of the techniques of multivariate statistics. By regression analysis we can analyze business, economics and social science data. There are some existing techniques to detect outliers by using Regression analysis. In regression analysis statisticians mainly follow two ways. (i) After fitting classical least square lines they detect outliers in Y direction by standardized residuals, studentized residuals (Srikantan; 1961), deletion studentized residual (Ellenberg; 1976) and outliers in X direction by high leverage values. (ii) The robust techniques that are commonly used in the identification of multiple outliers are least median of squares (LMS) (Rousseeuw; 1984), least trimmed squares (LTS)(Rousseeuw; 1984) and reweighted least squares (RLS)(Rousseeuw and Leroy; 1987) etc. In logistic regression outliers are detected by generalized standardized pearson residual (GSPR) (Hadi and Simonoff, 1993; Atkinson, 1994; Munier, 1999 and Imon, 2005).

2. Singular Value Decomposition

The singular value decomposition (SVD) can be viewed as the extension of the eigenvalue

decomposition for the case of nonsquare matrices. It shows that any real matrix can be diagonalized by using two orthogonal matrices. The eigen value decomposition, instead, works only on square matrices and uses only one matrix (and its inverse) to achieve diagonalization. If the matrix is square and symmetric, then the two orthogonal matrices of SVD become equal, and eigen value decomposition and SVD become one and same thing. Because the SVD is much more general than the eigen value decomposition and intimately related to the matrix rank and reduced rank least square approximations, it is a very important and useful tool in matrix theory, statistics and signal analysis. It can be used as a data reduction technique.

Singular value decomposition, specially its low rank approximation property is an elegant part of modern matrix theory. After its inception (1936)[10] its two ways fascinating data reduction capacity remained unnoticed till the last quarter of last century. Since then statisticians have been showing increasing interest to SVD for principal component analysis (PCA), canonical correlation analysis (CCA) and cluster analysis. Principal component analysis (PCA), often performed by singular value decomposition (SVD), is a popular analysis method that has recently been explored as a method for analyzing large-scale expression data (Raychaudhuri et al., 2000; Alter et al., 2000)[11,12]. Additionally SVD/PCA has been used to identify high-amplitude modes of fluctuations in macromolecular dynamics simulations (Garcia, 1992; Romo et al., 1995)[13,14], and identify structural intermediates in lysozyme folding using small-angle scattering experiments (Chen et al., 1996)[15]. One of the challenges of bioinformatics is to develop effective ways to analyze global gene expression data. A rigorous approach to gene expression analysis must involve an up-front characterization of the structure of the data. In addition to a broader utility in analysis methods, singular value decomposition (SVD) and principal component analysis (PCA) can be valuable tools in obtaining such a characterization. SVD and PCA are common techniques for analysis of multivariate data, and gene expression data are well suited to analysis using SVD/PCA. A single microarray experiment can generate measurements for thousands, or even tens of thousands of genes. Gene expression data are currently rather noisy, and SVD can detect and extract small signals from noisy data. Since SVD can reduce data in both ways–columns (generally indicates variables) and rows (generally indicates cases), and is more numerically stable, and moreover, PCA can be undertaken as a by product of SVD, in modern research it is being used more frequently in place of classical PCA for data compression (Diamantaras and Kung,1996;)[16] , clustering (Murtagh,2002;)[17] and multivariate outliers detection (Penny and Jolliffe,2001;)[18].

3. Low Rank Approximation of SVD

Low rank approximation (C. Eckart and G. Young, 1936)[10] is an important properties of SVD. It has an wonderful data reduction capacity with minimum recovery error. We can reduce variables as well as observations by using SVD. If X is $m \times n$ matrix of rank $k \leq min(m,n)$. Then by singular value decomposition we can write,

$$X = U\Lambda V^T \qquad (1)$$

where U is the column orthonormal matrix whose columns are the eigen vectors of XX^T, Λ is the diagonal matrix contain the singular values of X and V is the orthogonal matrix whose columns are the eigen vectors of $X^T X$.

From (1) we can write

$$X = \lambda_1 u_1 v_1^T + \lambda_2 u_2 v_2^T + \cdots + \lambda_k u_k v_k^T .$$

Suppose we approximate X by $\tilde{X}$ whose rank is $r < k \leq min(m,n)$.

$$\tilde{X} = \lambda_1 u_1 v_1^T + \lambda_2 u_2 v_2^T + \cdots + \lambda_r u_r v_r^T .$$

$$\Rightarrow \tilde{X} = U_r \Lambda_r V_r^T$$

where U_r is $m \times r$, Λ_r is a diagonal matrix of order r and V_r is $n \times r$. Now post multiply

V_r in both side we have,

$$\tilde{X}V_r = U_r\Lambda_r$$

its first column represents the first PC, second column represents the second PC and so on. Hence we see that X is a $m\times n$ matrix but $\tilde{X}V_r$ is a $m\times r$. Generally n represents no. of variables, so it reduces data by minimizing no. of variables.

4. SVD Based Outlier Detection Method

We develop a graphical method of outliers detection using SVD. It is suitable for both general multivariate data and regression data. For this we construct the scatter plots of first two PC's, and first PC and third PC. We also make a box in the scatter plot whose range lies *median*($1^{st}PC$) ± 3 × *mad*($1^{st}PC$) in the X-axis and *median*($2^{nd}PC/3^{rd}PC$) ± 3 × *mad*($2^{nd}PC/3^{rd}PC$) in the Y-axis. Where *mad* = median absolute deviation. The points that are outside the box can be considered as extreme outliers. The points outside one side of the box is termed as outliers. Along with the box we may construct another smaller box bounded by 2.5/2 MAD line.

5. Example

In this section we will use our method in some real data sets. These datasets are well known in multivariate analysis and regression analysis.

5.1 Car Data

Our first example is the low-dimensional car data set which is available in S-PLUS as the data frame *cu.dimensions*. For *n=111* cars, *p=11* characteristics were measured such as the length, the width and the height of the car. Using our method we get the **Figure-1**. From Figure-1(a) we see that observations 25, 30, 32, 34, 36, 102, 104, 107, 108, 110 and 111 are outside the box by the two groups. Also from Figure-1(b) we see that the observations 6, 102 and 104-111 are outside the box. So in our above graph we can say that 6, 25, 30, 32, 34, 36, 102, 104 -111 are unusual observations. Hubert, Rousseeuw and Branden(2005)[19] declared observations 25, 30, 32, 34, 36, 102-111 as outliers by using ROBPCA.

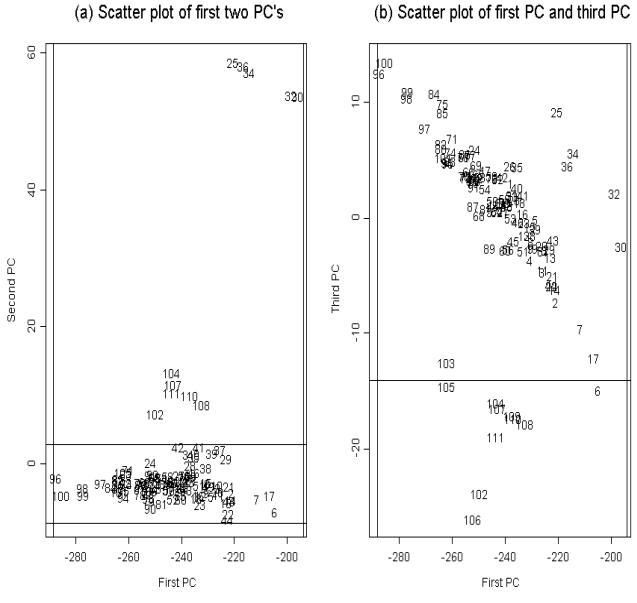

Figure 1. Scatter plot of car data (a) scatter plot of first two PC's and (b) scatter plot of first and third PC.

5.2. Hawkins-Bradu-Kass (1984) Data

Hawkins, Bradu and kass (Hawkins et al., 1984)[20] constructed an artificial three-predictor data set containing 75 observations with 14 influential observations. Among them there are ten high leverage outliers (cases 1-10) and for high leverage points (cases 11-14) -Imon (2005)[21]. If we apply our method in this data then we get the **Figure-2**. From **Figure-2** we see that observations 1-14 are outside our box so observations 1-14 are unusual observations. Also we see that three clusters are present in the data. Observations 1-10 make 1st cluster, observations 11-14 make second cluster and the rest observations make third cluster in figure-2(a). Index plot of standardized residuals obtained from LMS (Rousseeuw and Leroy, 1987)[5] is out performed by our plot.

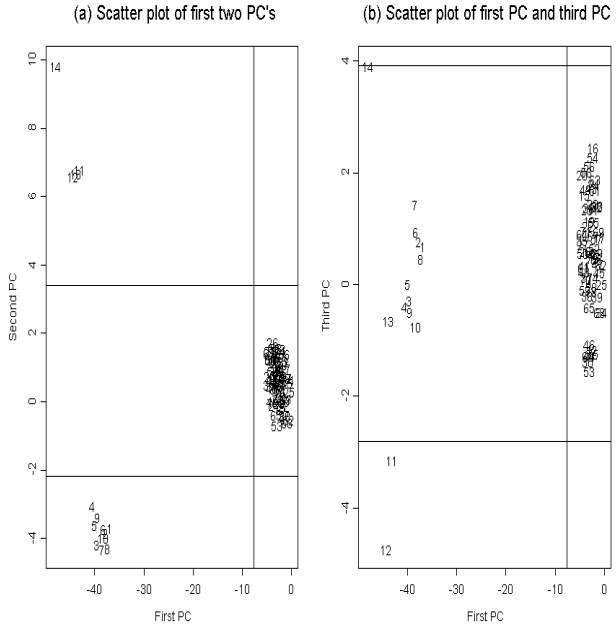

Figure 2. Scatter plot of Hawkins, Bradu and kass data (a) scatter plot of first two PC's and (b) scatter plot of first and third PC.

5.3. Modified Brown Data

We first consider the data set given by Brown (1980)[22]. Here the main objective was to see whether an elevated level of acid phosphates (A.P.) in the blood serum together would be of value for predicting whether or not prostate cancer patients also had lymph node involvement (L.N.I). Ryan (1997)[23] pointed out that the original data on the 53 patients which contains 1 outlier (observation number 24). Imon and Hadi(2005)[9] modified this data set by putting two more outliers as cases 54 and 55. Also they showed that observations 24, 54 and 55 are outliers by using generalized standardized Pearson residual (GSPR) (Hadi and Simonoff,1993; Atkinson, 1994; Munier, 1999; Imon, 2005) [6,7,8,9]. Now we apply our method in this data. Applying our method we get the figure 3. From figure-3 we see that observations 24, 54, 55, 53 and 25 are detected as outliers by our method.

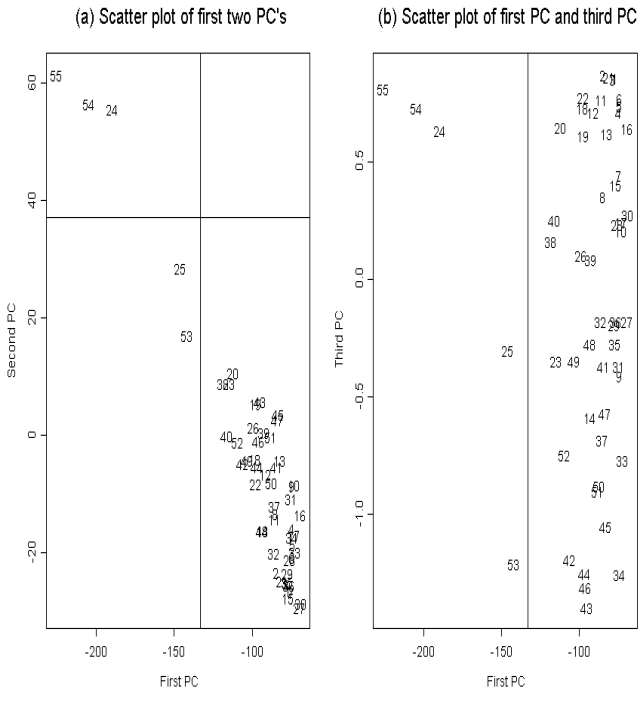

Figure-3. Scatter plot of modified Brown data (a) scatter plot of first two PC's and (b) scatter plot of first and third PC.

6. Advantages Our Method

Our method has the following advantages over other competitive methods;

- It is easy to understand without hard mathematics.
- It can be applied for both regression data and other type of multivariate data.
- It is directly applied to separate extreme outliers from general outliers.
- It can detect several clusters that other outliers detection methods fail to pinpoint.
- It can single out not only outlying observations but also outlying variables.

7. Conclusion

Form our above discussion we conclude that the proposed method is very much helpful for multivariate outlier detection. By using this method we can also see the structure of multivariate data graphically. Though there are several existing methods for detecting

multivariate outliers but SVD based technique is better than those methods.

8. Reference

[1] Rousseeuw P.J. and Leroy A. (1987). *Robust Regression and Outlier Detection*. New York: Wiley. doi:10.1002/0471725382

[2] Rousseeuw P.J., Van Zomeren B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*. Vol. 85(411), pp. 633-651.

[3] Srikantan K. S.(1961). Testing for the single outlier in a regression model. *Shankhya, vol-23, Series A*, 251-260.

[4] Ellenberg J.H.(1976). Testing for single outlier from a general regression. *Biometrics,* 32, 637- 645. doi:10.2307/2529752

[5] Rousseeuw P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871 - 880.

[6] Hadi A.S. and Simonoff J.S. (1993). Procedure for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88, 1264 - 1272. doi: 10.1080/01621459.1993.10476407

[7] Atkinson A.C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association 89;* 1329 - 1339. doi:10.1080/01621459.1994.10476872

[8] Munier S. (1999). Multiple outlier detection in logistic regression. *Student 3*, 117 -126.

[9] Imon A.H.M.R. and Hadi Ali S. (2005). Identification of Multiple Outliers in Logistic Regression. *International Statistics Conference on Statistics in the Tecnological Age, Institute of Mathematical Sciences,* University of Malaya, Kualalumpur, Malaysia, December 2005. doi:10.1080/03610920701826161

[10] Eckart C. and Young G. (1936). The approximation of one Matrix by another of lower Rank. *Psychometrika*,1,211-218. doi: 10.1007/BF02288367

[11] Raychaudhuri S., Stuart J.M. and Altman R.B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput 2000:*455-66.

[12] Alter O., Brown P.O. and Botstein D. (2000). Singular value decomposition for genome- wide expression data processing and modeling. *Proc Natl Acad Sci USA 2000;* 97:10101-06. doi:10.1073/pnas.97.18.10101

[13] Garcia A.E. (1992). Large-Amplitude Nonlinear Motions in Proteins. *Phys Rev Lett 1992;* 68:2696-99. doi: 10.1103/PhysRevLett.68.2696

[14] Romo T.D., Clarage J.B., Sorensen D.C. and Phillips G.N. Jr., (1995). Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. *Proteins 1995;* 22:311-21. doi:10.1002/prot.340220403

[15] Chen L., Hodgson K.O. and Doniach S. (1996). A lysozyme folding intermediate revealed by solution X-ray scattering. *J Mol Biol 1996;* 261:658-71. doi:10.1006/jmbi.1996.0491

[16] Diamantaras K.I. and Kung S.Y. (1996). *Principal Components Neural Netwarks: Theory And Applications*, John wiley & sons, Inc. N.Y 45-46.

[17] Murtagh F.(2002). Clustering in High-dimensional Data Spaces. *Classification, Clustering and Data Analysis*, eds. Jajuga,K., Sokolowski,A. and Bock,H., Springer-Verlag, Berlin, pp.89-96.

[18] Penny K. I. and Jolliffe I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. *Royal Statistical Society*, part 3 , PP.295-308. doi:10.1111/1467-9884.00279

[19] Hubert M., Rousseeuw P.J. and Branden K.V. (2005). ROBPCA: a New Approach to Robust Principal Components Analysis. *Technometrics, 47, 64-79.* doi:10.1198/004017004000000563

[20] Hawkins D. M., Bradu D. and Kass G.V.(1984),Location of several outliers in multiple regression data using elemental sets. *Technometrics*, 20, 197-208. doi:10.1080/00401706.1984.10487956

[21] Imon A. H. M. R. (2005). Identifying multiple influential observations in linear Regression. *Journal of Applied Statistics* 32, 73 - 90. doi:10.1080/02664760500163599

[22] Brown B.W., Jr. (1980). Prediction analysis for binary data. in *Biostatistics Casebook*, R.G. Miller, Jr., B. Efron, B. W. Brown, Jr., L.E. Moses (Eds.), New York: Wiley.

[23] Ryan T.P. (1997). *Modern Regression Methods*, Wiley, New York.