ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Multivariate Data Visualization: Correspondence Analysis, Classical and Robust Singular Value Decomposition and Depth based Approach

***Nishith Kumar[1], and Mohammed Nasser[2]**

1. Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj.
2. Department of Statistics, Rajshahi University, Rajshahi.

Manuscript Info

Manuscript History:

Received: 18 November 2015
Final Accepted: 22 December 2015
Published Online: January 2016

Key words:

Correspondence Analysis, Singular value decomposition, Robust singular value decomposition, Principal component analysis and Depth function

Corresponding Author

Nishith Kumar

Abstract

Now a day's Data Mining is one of the challenging area in statistics as well as in computer science. Data visualization is one of the most important parts in Data Mining. We can only visualize two or three dimensional data but for Data Mining much more than three dimensions is usual rather than exception. So data reduction is very important for visualization of multivariate data. For this reason, in this article I would like to introduce four well known, effective and sophisticated scientific data reduction methods, correspondence analysis, singular value decomposition (SVD), robust singular value decomposition (RSVD) and depth for data visualization, pattern recognition and outlier detection. Since these techniques are used for data reduction technique so by using these techniques we can visualize data taking only two or three dimensions that maximize the total variation of data. In many cases two or three singular values or eigen values cannot explain most (greater than 80%) of the variation of data. In that case we can use L1 depth, half space depth and kernel based depth for visualizing data. . In this paper we have used four well known real dataset (fisher's iris data, Wisconsin breast cancer data, Glass identification data and Seeds data) for visualization and also rigorously explained the results on the motion of the aforementioned techniques.

INTRODUCTION

Multivariate data visualization is an exciting and recent research area of statisticians, engineers and those who are involved in data mining. Multivariate data visualization, as a specific type of information visualization. We can not visualize more than three dimensional data. Now a days most of the data are high dimensional. To know the pattern and structure of data, visualization is very important. But for these high dimensional data pattern recognition and structure visualizations are more complicated. So for visualizing such type of data (more than three dimensions), data reduction is very important. If we want to reduce dimension then we need some statistical tools. The most popular statistical technique for data reduction are principal component analysis (PCA) and Singular value decomposition (SVD). PCA and SVD (classical and Robust) is recently using for graphical clustering (Nishith et al. 2011). It has a fascinating data reduction capacity in both ways. Also SVD based PCA is more numerically stable (Terry Speed, 2003). If first two or three singular vectors explain most of the variation of data then visualization using SVD based PC gives good result. If first two or three vectors do not account most of the variation of data then we can use depth. After using depth in multivariate data, the resulted data come to a univariate data so then we can easily draw the index plot and visualize the structure of the data. Though in 1993 and 1994 Cleaveland discussed clearly about the visualization of multivariate data but he did not mention these techniques clearly. Beautifully

executed and fascinating data visualizations were given in the book of Tufte (1983, 1990, 1993,) and Wainer (1997). Visualization of categorical data was discussed in the monograph by Blasiusand Greenacre (1998). The books of Everitt (1978) and Toit et al. (1986) are now dated but they still provide us with readable accounts of some classical techniques for multivariate visualization. Muruz´abal, J. and Mu˜noz (1997) visualized data by using self organizing map (SOM). Nishith et al. (2010 and 2012) visualized multivariate data using only biplot, correlation scatter plot and SVD. Winne Wing Yi Chan (2006) used different techniques like scatterplot, projection plot, hyper slice, hyper box, radial coordinate visualization, star coordinates, space filling curve pixel bar chart, hierarchical axis, dimensional stacking etc for data visualization. Multidimensional scaling is one of the most important method for data visualization (Andreas Buja,2002). For visualization, In this article we have used four most popular and powerful data reduction techniques that are PCA, SVD, Robust SVD depth and a sophisticated technique corresponding analysis (CA) and also proposed a new kernel based depth. Making visualization more than two dimensions are difficult. Correspondence analysis is an ideal technique to analyze multivariate form of data because of its ability to extract the most important dimensions, allowing simplification of data matrix. Recently CA is widely used for two way clustering (Ciampi A. (2005)) and in psychological research (Doey L. and Kurta J (2011)). CA is an exploratory data technique used to analyze contingency table type data but we have used CA in this paper for positive real multivariate data and also have compared this technique with SVD (classical and robust) and depth.

Nature and Sources of Data

In this article we have used four well known real multivariate dataset (fisher's iris data, Wisconsin breast cancer data glass identification data and seeds data) for visualization.

Fisher's Iris Data

"The Use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, 7, Part II, (Fisher, 1936). The data were collected by Edgar Anderson, "The irises of the Gaspe Peninsula", Bulletin of the American Iris Society, 59, 1935, pp. 2-5. This data contain 4 measurements on 50 flowers from each of 3 species of iris. Sepal length and width, and petal length and width are measured in centimeters. Species are Setosa, Versicolor, and Virginica.Hence we combine three species so we get 150 observations.

Wisconsin Breast Cancer Data

This data is collected by Dr. WIlliam H. Wolberg (physician) University of Wisconsin Hospitals Madison, Wisconsin USA (1992) . One can easily get the total information of data from the following website http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ . Breast cancer data can be found form http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data and we can get the variable information from the following website http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/. Wisconsin breast cancer data contain 699 observations and 10 variables. In this data the variables are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, Class: (2 for benign, 4 for malignant) .

Glass Identification Data

This dataset has been collected from USA Forensic Science Service and the creator of this dataset is B. German, Central Research Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berkshire RG7 4PN. The donor of this dataset is Vina Spiehler, Ph.D.,DABFT, Diagnostic Products Corporation, (213) 776-0180 (ext 3014). This dataset contains 6 types of glass; defined in terms of their oxide content (i.e. Na, Fe, K, etc). The dimension of this dataset is 214×10; i.e. there are 10 variables in this dataset. The variables are RI(refractive index), NA, Mg, Al, Si, K, Ca, Ba, Fe, Type of glass. This dataset can be found from the following website, http://archive.ics.uci.edu/ml/datasets/Glass+Identification.

Seed Data

This dataset contains measurements of geometrical properties of kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements of each randomly selected for the experiment. High quality

visualization of the internal kernel structure was detected using a soft X-ray technique. This images were recorded on 13×18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agro physics of the Polish Academy of Science in Lublin. This dataset contain 210 observations. To construct the data, seven geometric parameters of wheat kernels were measured that are area, perimeter, compactness, length of kernel, asymmetry coefficient and length of kernel groove. M. Charytanowicz, J.Niewczas, P Kulczycki, P.A. Kowalski, S. Lukasik and S. Zak first used this dataset in their paper in 2010. One can easily find this dataset from the following website http://archive.ics.uci.edu/ml/datasets/Seeds.

Correspondence Analysis

Correspondence analysis is a statistical technique used to analyze categorical data (Benzecri, 1992) and provides a graphical representati**o**n of cross tabulations or contingency tables. Correspondence analysis (CA) can be viewed as a generalized principal component analysis tailored for the analysis of qualitative data. Although CA was originally created to analyze cross tabulation but CA is so multipurpose that it is used with a lot of other numerical data table types. It is formally applicable to any data matrix with nonnegative entries. The main objectives of CA are to transform a dataset into two factor scores (rows and columns) that give the best representation of the similarity structure of the rows and columns of the table. Correspondence analysis is used to reduce the dimension of a data matrix as in principal component analysis. So using CA we can visualize the data two or three dimensionally.

CA and its geometric interpretation comes from France in 1960 and associated with the France school of "Data analysis" and developed successfully under the supervision of Jean-Paul Benezecri (1973). The solution of Correspondence analysis was shown by Greenacre (1984) using singular value decomposition. We can summarize the theory of CA by the following way, suppose the table of data, with all values on the same scale, is denoted by *X*, first divide *X* by its grand total *n* to obtain so called correspondence matrix $P = \frac{X}{n}$. Let the vectors ***r*** and ***c*** be the row and column marginal totals of P respectively and D_r and D_c be the diagonal matrices of these matrices. To obtain coordinates, the computational algorithm of the row and column profiles with respect to principle axes, using SVD is $D_r^{-1/2}(P - rc^T)D_c^{-1/2} = U\Lambda V^T$ where $U^TU = V^TV = I$. The principal coordinates of rows: $F = D_r^{-1/2}U\Lambda$ and the principal coordinates of columns $G = D_c^{-1/2}V\Lambda$. Standard row and column coordinates are $D_r^{-1/2}U$ and $D_c^{-1/2}V$ respectively.

Singular Value Decomposition (Classical and Robust)

Singular value decomposition (SVD) is a very old technique in mathematics. Eugenio Beltrami and Camille Jordan originally developed the SVD in the mid to late 1800's. For final developments of the SVD several other mathematicians took part, including James Joseph Sylvester, Erhard Schmidt and Hermann Weyl who studied the SVD into the mid-1900. The most important properties of SVD is low rank approximation developed by C. Eckart and G. Young (1936). SVD has a magnificent data reduction capacity in both row and column mode. SVD can be viewed as the extension of the eigen value decomposition for the case of nonsquare matrices. By using SVD we can reduce both variables as well as observations. So taking only two or three singular vectors we can easily draw two or three dimensional graph that shows the structure or pattern of the data.

If *X* is a data matrix of $m \times n$ with rank $k \le \min(m,n)$ then by using SVD we can write it as $X = U\Lambda V^T$; where U is a column orthonormal matrix, V is the row orthogonal matrix and Λ is a diagonal matrix that contain the singular values. U and V are the eigen vector of XX^T and X^TX respectively. If we approximate *X* by $\tilde{X}$ with rank r then we can write $\tilde{X}$ as $\tilde{X} = \lambda_1 u_1 v_1^T + \lambda_2 u_2 v_2^T + \cdots + \lambda_r u_r v_r^T$. In matrix form we can write it as $\tilde{x} = U_r \Lambda_r V_r^T$. Now $\tilde{X}V_r = U_r\Lambda_r$, its first column represents the first principal component (PC), 2nd column represents second PC and so on. If the variables are correlated then first two or three PC contains most of the variation of data, so if draw the two or three dimensional scatter plot using two PC or three PC then this visualization represent the pattern or structure of the data. The classical SVD is very much affected by outlier and Gabriel and Zamir (1979) addressed that the calculation of classical SVD cannot be performed, if the data matrix *X* has any missing element. Hawkins, liu and young (2001) proposed a method "Robust Singular Value Decomposition(RSVD)" on the basis of alternating L1

regression approach that can solve the outlier problem as well as handle missing element. The algorithm of alternating L1 regression approach for calculating robust singular value decomposition are given below,

Algorithm:

(i) Initialize the leading left singular vector u_i. There is no obvious choice of the initial values of u_i.

(ii) Fit the L_1 regression coefficient by minimizing $\sum_{i=1}^{n} | x_{ij} - c_j u_{i1} | \quad ; \ j = 1, 2, \cdots, p.$

(iii) Calculate right singular vector $v_1 = c / \| c \|$, where $\|.\|$ refers to the Euclidean norm.

(iv) Again fit the L_1 regression coefficient d_i by minimizing $\sum_{j=1}^{p} | x_{ij} - d_i v_{j1} |; \quad i = 1, 2, \cdots, n.$

(v) Calculate the left singular vector $u_i = d / \| d \|$

Iterate the process (ii) to (v) until it converge.

For getting 2nd and subsequent of SVD, we can replace X by an another matrix obtained by subtracting the most recently found them in the SVD $X \leftarrow X - \lambda_k u_k v_k^T$.

Depth Function

A data depth can be used to measure the ”depth" or “outlyingness" of a given multivariate sample with respect to its underlying distribution or a given data cloud. Firstly Tukey 1975 introduced the concept of depth function. After that many other mathematicians took part for the development of depth function. Barnet (1976) proposed convex hull peeling depth using the concept of convex hull. Oja (1983) and Liu(1990) introduced Oja depth and simplical depth respectively using the concept of simplex method. Liu and Singh 1993 proposed Mahalanobis depth using mahalanobis distance. Lastly in 2003 Yonghong Gao proposed spatial rank depth using concept of rank.A data depth is a way of measuring how deep or (central) a given point $x \in R^d$ is, with respect to F or with respect to a given data cloud ($(X_1, X_2, \cdots X_n)$. For a cdf F on R^d, a depth function D(x,F) provides an associated center-outward ordering of points $x \in R^d$.

Half-space depth function [Tukey (1975)]

The half-space depth at x with respect to (w.r.t) F is defined by

$HD(F;x) = \underset{H}{Inf}\{P(H):$ H is a closed half-space in R^d *and* $x \in H\}$. The sample version of $HD(F,x)$ is $HD(F_n,x)$, where F_n denotes the empirical distribution of the sample $(X_1, X_2, \cdots X_n)$.

Oja Depth (OD) [Oja (1983)]

Oja depth at x w.r.t F is defined by

$OD(F;x) = [1 + E_F(volume(S(x, X_1, \cdots, X_{d+1})))]^{-1}$,

where $S(x, X_1, \cdots, X_{d+1})$ is the closed simplex with vertices x and d random observations $X_1, X_2, \cdots, X_d$ from F.

Mahalanobis Depth (M_hD)[Liu and Singh (1993)]

Mahalanobis depth is an extension of Mahalanobis distance (Mahalanobis (1936)). We know Mahalanobis distance is $[(x-\mu_F)\Sigma_F^{-1}(x-\mu_F)']^{1/2}$. The Mahalanobis depth at x w.r.t F is defined by $M_hD(F;x) = [1 + (x-\mu_F)\Sigma_F^{-1}(x-\mu_F)']^{-1/2}$; where μ_F and Σ_F are the mean vector and dispersion matrix of F respectively. The sample version of M_hD is obtained by replacing μ_F and Σ_F with their sample estimates.

L_1 Depth [Vardi and Zhang (2000)]

The L_1 depth (L_1 D) of a point x with respect to a dataset $\{X_1, X_2, \cdots X_n\}$ in R^d is , $L_1D = 1 - \|\bar{e}(x)\|$; where $e_i(x) = \frac{x - X_i}{\|x - X_i\|}$ and $\bar{e}(x) = \frac{\sum_{i=1}^{n} \eta_i e_i(x)}{\sum_j \eta_j}$; η_i is the weight assigned to observation X_i and $\|x - X_i\|$ is the Euclidian

distance between x and X_i. In this article we have used L_1 depth. We also propose a new kernel based depth technique to visualize multivariate data.

Proposed Kernel based Depth

In this article we have used our new proposed kernel based depth technique (using a radial basis kernel function) to visualize multivariate data.

$$\text{Kernel Depth} = \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} K(x_i, x_j); \quad \text{for } i = 1,2,\cdots,n.$$

Application of Different Visualization Techniques in Different Data

Iris data

If we apply correspondence analysis, SVD, L1 depth, kernel depth and robust SVD in iris data then we get the following figure

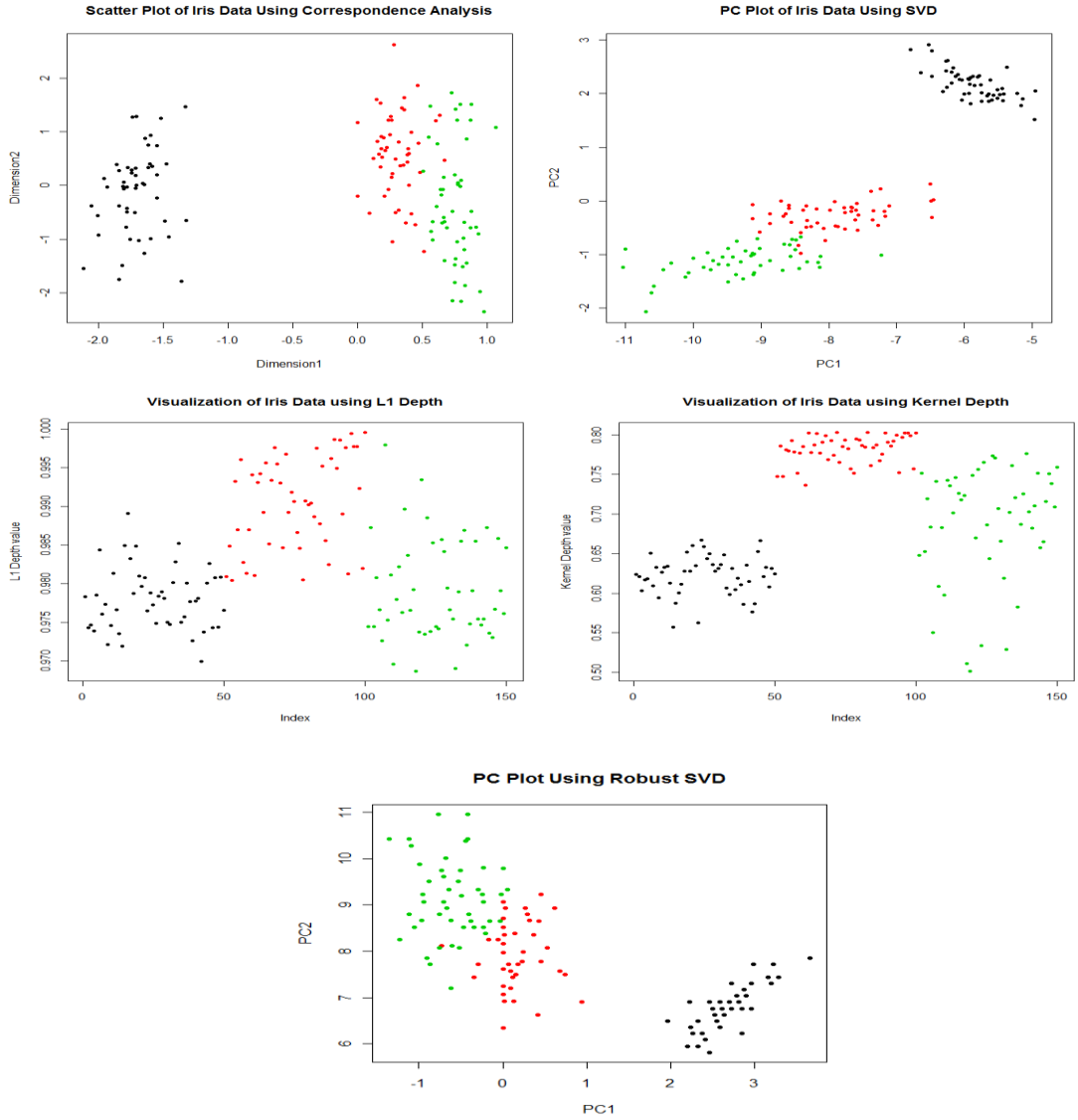

Figure 1. Iris Data Visualization Using CA, SVD, L_1 Depth, Kernel Depth and Robust SVD.

From figure 1 we can say that all these techniques have showed two or three types of observations may present in the data. Two groups are clearly identified by CA, SVD, L1 depth and Robust SVD analysis but kernel based depth indicates three groups may present in the Iris Data. From the data description of Iris data it is known that there were three types of flowers information's in the data.

Wisconsin breast cancer data

In Wisconsin breast cancer data If we apply correspondence analysis, SVD, L1 depth, kernel depth and robust SVD then we will get the following figure

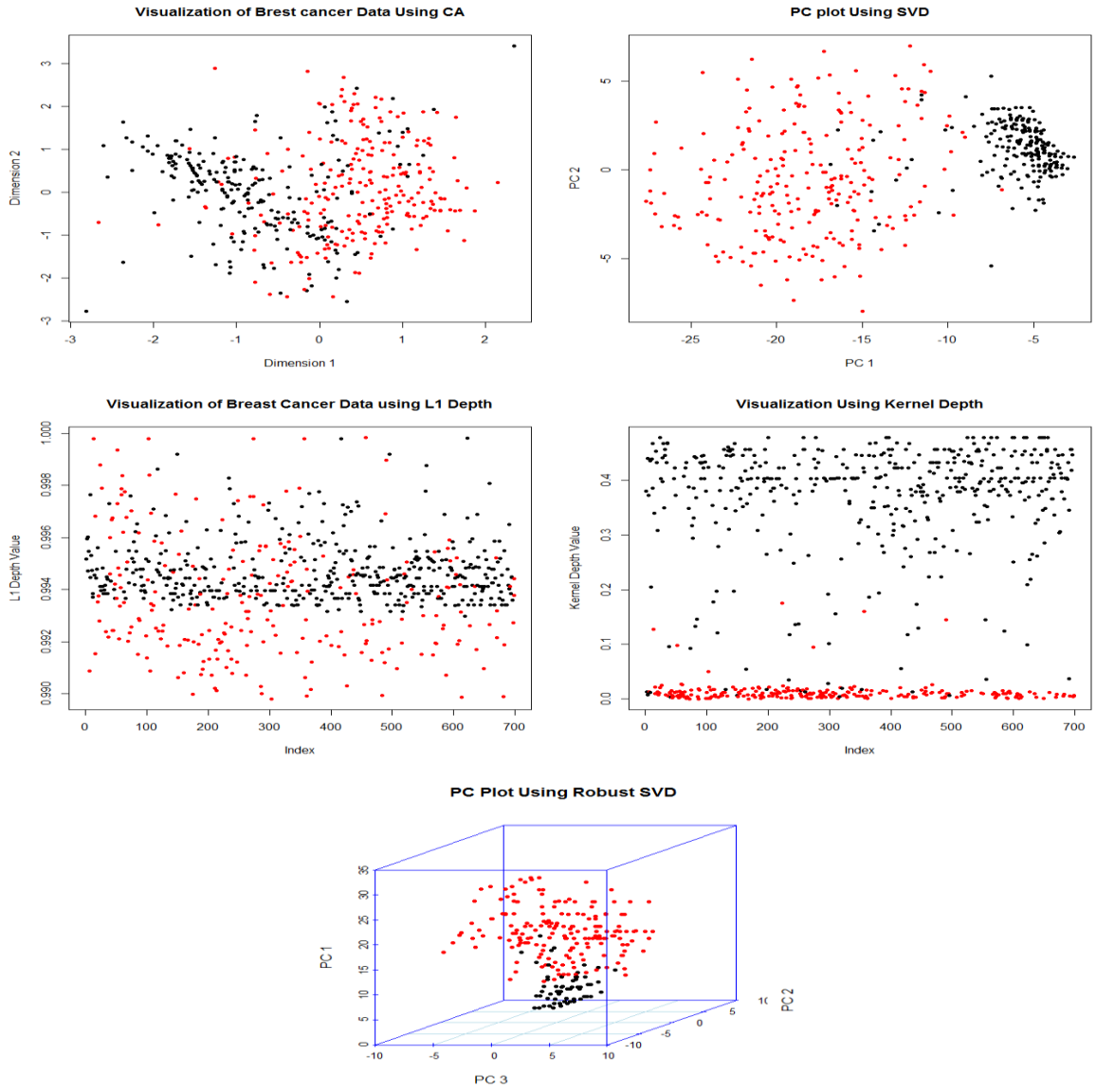

Fig-2: Wisconsin Breast Cancer Data Visualization Using CA, SVD, L_1 depth, Kernel Depth and Robust SVD.

From figure 2 we can say that all these techniques without L1 depth have showed two types of observations may present in the data. Two groups are clearly identified by CA, SVD, Robust SVD and Kernel Depth analysis and from the data description of Wisconsin Breast Cancer data it is known that there were two types of tumor information in the data.

Glass Identification Data

After applying Correspondence analysis (CA), SVD, L1 depth, kernel depth and robust SVD in Glass identification data then we get the following figure

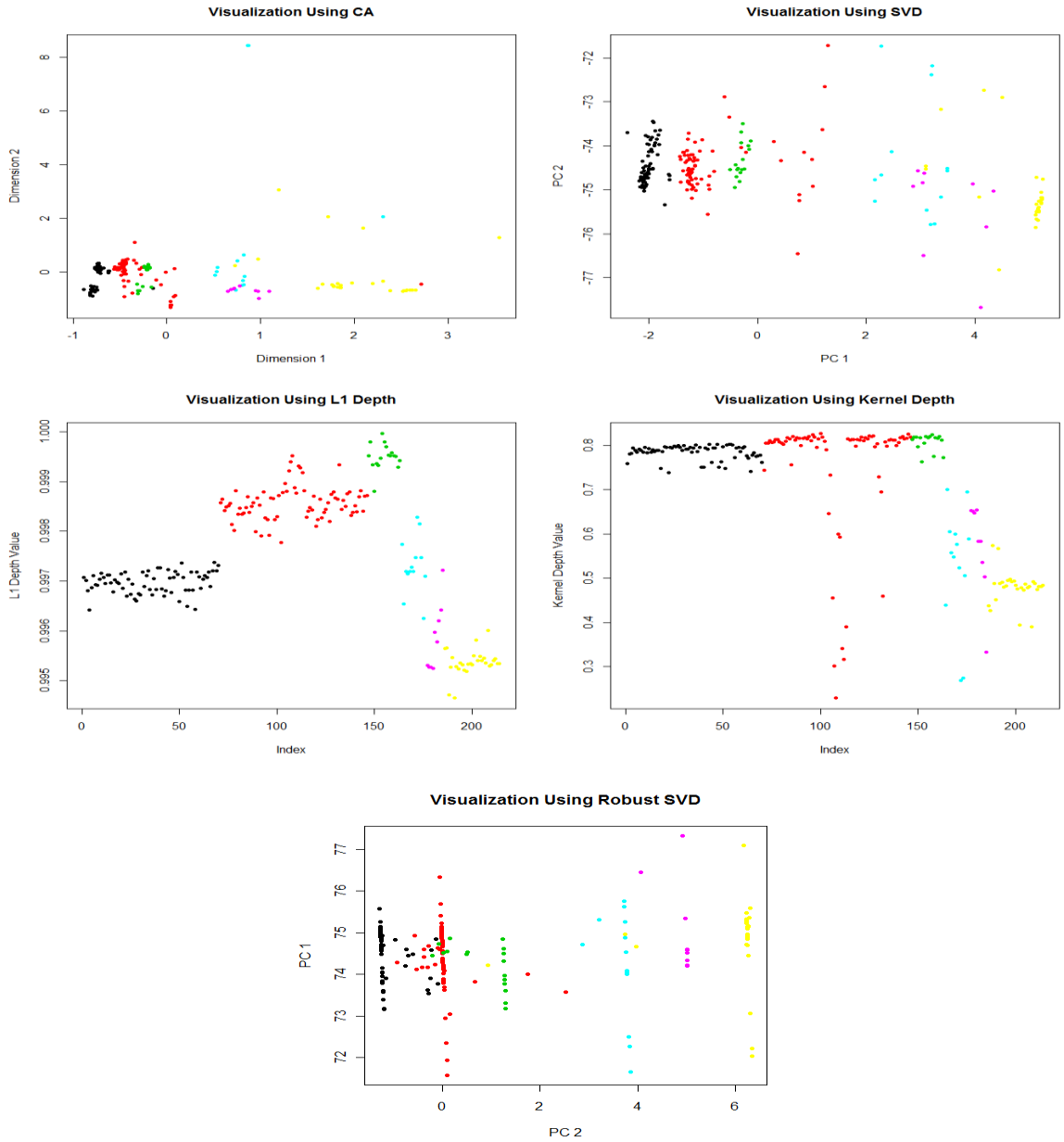

Fig-3: Glass Identification Data Visualization Using CA, SVD, L_1 depth, Kernel Depth and Robust SVD.

From figure 3 we can say that all these techniques have showed six types of observations may present in the data. six groups are clearly identified by CA, SVD, Robust SVD L_1 depth and Kernel Depth analysis and from the data description of Glass identification data it is known that there were six types of glass information in the data.

Seeds data

If we apply Correspondence analysis (CA), SVD, L1 depth, kernel depth and robust SVD in Seeds data then we get the figure-3.

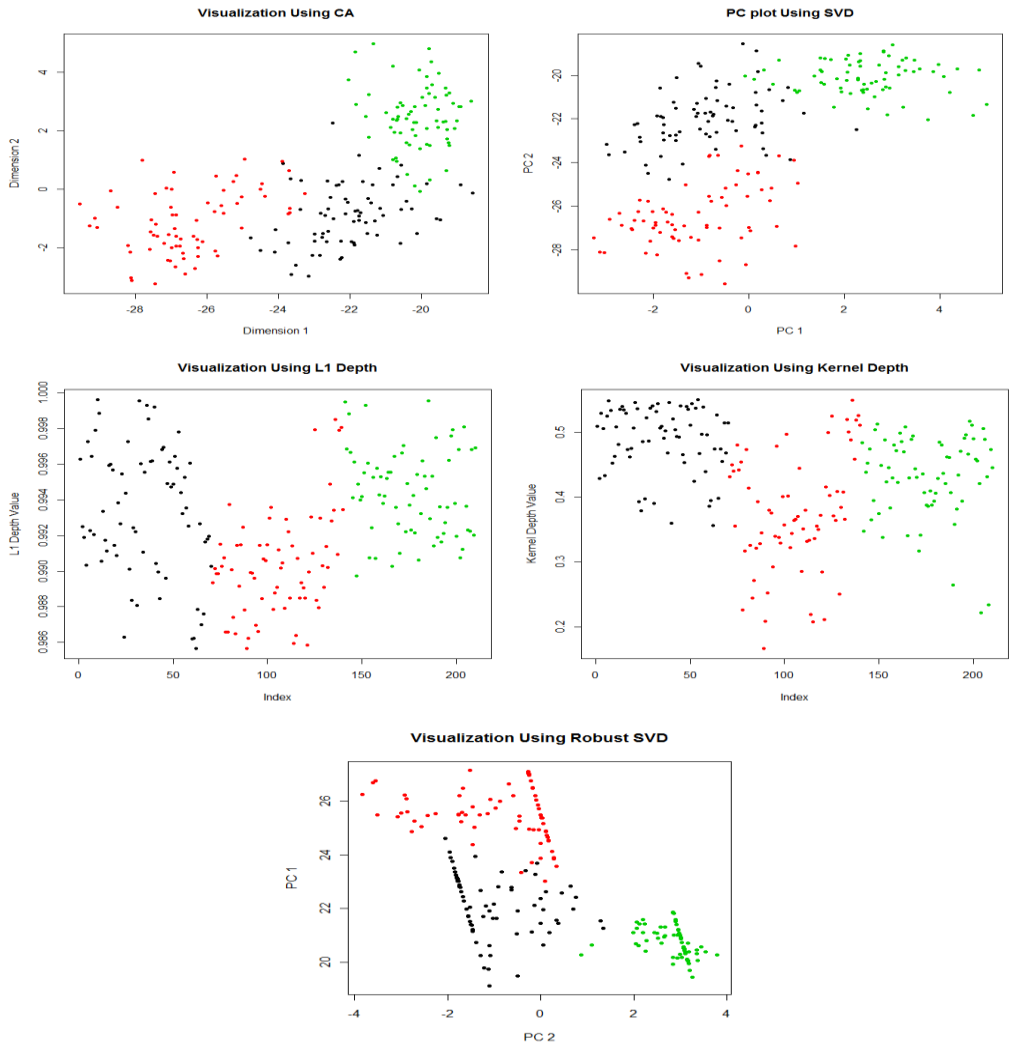

Fig-4: Seeds Data Visualization Using CA, SVD, L_1 depth, Kernel Depth and Robust SVD.

From figure 4 we have seen all these techniques are indicating there may be groups in the data. Three groups are clearly identified by Robust SVD and from the data description of Seeds data it is known that there were different varieties of wheat information in the data.

Table 1. Comparison Table

Data	CA	SVD	L_1 Depth	Kernel Depth	Robust SVD
Iris data	2 Clusters clearly identified	2 Clusters clearly identified	2 Clusters clearly identified	**3 Clusters identified**	2 Cluster clearly identified
Wisconsin breast cancer data	Cluster is not Clear	**2 Clusters clearly identified**	Cluster is not Clear	**2 Clusters clearly identified**	**2 Clusters clearly identified**
Glass Identification Data	6 Clusters identified but not clear	6 Clusters identified but not clear	**6 Clusters clearly identified**	**6 Clusters clearly identified**	**6 Clusters clearly identified**
Seeds data	3 Clusters may present in the data	3 Clusters may present in the data	3 Clusters may present in the data	**3 Clusters may present in the data**	**3 Clusters may present in the data**

Advantages of Visualization Using These Methods

The advantage of these methods from other existing techniques are given below

- It is easy to understand without hard mathematics, only need some matrix knowledge.
- We can apply these techniques for both supervised and unsupervised technique.
- It is directly applied to visualize the pattern of the data.
- It can be applied in extremely complicated data sets without any extraneous assumption. Only for correspondence analysis data values should be positive.

Conclusion

After getting a data it is very much difficult to choose appropriate statistical tools for analyzing the data to a researcher. To see the structure of the data we may take a proper decision which method is the best for analyzing the data. So it is wise to see the structure of the data using CA, SVD, Robust SVD and Kernel Depth before analyzing data. On the basis of our graph we can say Kernel Depth is better performer among them. So we can recommend it as a visualization technique to visualize multivariate data. For analyzing data we have used only R software in this article.

References

Barnett, V (1976). "The ordering of multivariate data (with discussion)", *Journal of the Royal Statistical Society. Series A. General, 139*, 318-352.

Benzecri, J. P. (1973). *L'analyse des donnees* (2 vol.). Paris: Dunod

Benzecri, J.P. (1992). *Correspondence Analysis Handbook*, NewYork: Marcel Dekker.

Blasius, J. and Greenacre, M. (1998). *Visualization of Categorical Data*, Academic Press, NewYork.

Buja, A. (2002). "Visualization Methodology for Multidimensional Scalling", *Journal of Classification* 19: 7-43 (2002) doi:10.1007/s00357-001-0031-0.

Chan, W. W. Y., (2006). " *A Survey on Multivariate Data Visualization*" Technical Report, Department of Computer Science and Engineering, Hong Kong University of Science and Technology Clear water Bay, Kowloon, Hong Kong.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Lukasik S., and Zak S., (2010),"A Complete Gradient Clustering Algorithm for Feature Analysis of X-ray Images", in : Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010,pp.15-24.

Ciampi, A., Markos, A. G. and Limas, C. M. (2005). "Correspondence analysis and two way clustering", *Institut d'Estadistica de catalunya, SORT* 29 (1) pp. 27-42. ISSN:1696-2281.

Cleveland, W.S. (1993). *Visualizing Data*, Hobart Press, Summit.

Cleveland, W.S. (1994). *The Elements of Graphing Data*, Hobart Press, Summit.

Doea ,L. and Kurta, J. (2011). "Correspondence Analysis applied to psychological research",Tutorials in Quantitative Methods for Psychology, Vol. 7 (1) pp.5-14.

Eckart, C. and Young, G. (1936). "The approximation of one matrix by another of lower rank", *Psychometrika*, 1, 211-218.

Everitt, B. (1987). *Graphical Techniques for Multivariate Data*, North-Holand, New York.

Fisher, R.A (1936). "The use of Multiple Measurement in Taxonomic Problems", Annals of Eugenics, &, PartII, pp. 179-188.

Gabriel, K. R., Zamir, S., (1979). "Lower rank approximation of matrices by least squares with any choice of weights", *Technometrics*, 21 489-498.

Greenacre, M. J. (1984*).* *Theory and Application of Correspondence analysis*. London: Academic Press.

Hawkins, D. M., Liu, L. and Young, S. S. (2001)."Robust Singular Value Decomposition. Technical Report 122, *National Institute of Statistical Sciences.* [Online] *htpp://www.niss.org/sites/default/files/pdfs/technicalreports/tr122.pdf*

Liu, R. (1990). "On a notion of data depth based on random simplices', *Ann. Statist.* 18 pp. 405-414.

Liu, R. and Singh, K. (1993). "A quality index based on data depth and multivariate rank tests', *J. Amer. Statist. Assoc.* 88 pp. 257-260.

Muruz´abal, J. and Mu˜noz A. (1997). "On the visualization of outliers via self-organizing maps", *Journal of Computation and Graphical Statistics* 6, 355-382.

Nishith, K. and Doulah, M. S. (2010). "Visualization of Biological Data Using Singular Value Decomposition", *International Journal of Engineering and Technology*, Vol.-7, Issue 4, 741-746.

Nishith, K. and Nasser, M. (2012). "A New Graphical Multivariate Outlier Detection Technique Using Singular Value Decomposition", *International Journal of Engineering Research and Technology*, Vol. 1 (06) ISSN-2271-0181.

Nishith, K. Nasser, M. and Sarker, S. C. (2011). "Anew singular Value Decomposition Based Graphical Clustering Technique and Its application in Climatic Data" Journal of Geography and Geology, Canadian Center of Science and Education, Vol-3, No. 1, 227-238. doi:10.5539/jgg.v3n1p227

Oja, H. (193). "Descriptive statistics for multivariate distribution", *Statistics and Probability Letters* (1) pp. 327-333.

Terry Speed, (2003). *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall/ CRC, 2003, pp. 190-200.

Toit, S. H., Steyn, A.G., and stumpf, A. G. W. (1986). *Graphical Exploratory Data Analysis*, Springer, New York.

Tufte, E. (1983). *The Visual Display of Quantitative Information,* Graphics Press, Cheshire.

Tufte, E. (1990). *Envisioning Information,* Graphics Press, Cheshire.

Tufte, E. (1993). *Visual Explanations:Images and Quantities, Evidence and Narrative,* Graphics Press, Cheshire.

Tukey, J. (1975). "Mathematics and Picturing Data", *In Proceedings of the 1975 International Congress of Mathematics* (2) pp. 523-531.

Vardi, V., Zhang C.H., (2000). "The multivariate L_1-median and associated data depth", *Nat. Acad. Sci. USA* 97, pp. 1423-1426.

Wainer H. (1997). *Visual revelations: graphical tales of fate and deception from Napoleon Bonaparte to Ross perot*, Copernicus, New York.

Yonghong Goa (2003). "Data depth based on spatial rank", *Statistics and Probability Letters* 65 pp. 217-225.