A Robust Principal Component Biplot using ROBPCA

T.A. Sajesh[1], M.R. Srinivasan[2]
[1] Department of Statistics, St. Thomas College (Autonomous), Thrissur
[2] Department of Statistics, University of Madras, Chennai
Corresponding author's email address: sajesh.t.aabraham@gmail.com

ABSTRACT

The principal component biplot (PCA biplot) is a graphical tool to simultaneously visualise the scores and loadings of the principal components obtained from the classical principal component analysis. The plot is widely using in the areas of plant breeding, genetics, manufacturing industry, agriculture, etc. Unfortunately, the least-square principal component analysis is not robust to the presence of outliers in the data set and hence the principal component biplot too. The extreme observations may unduly influence the form of the first few principal components and change the actual structure of the biplot. Consequently, the inference based on this plot will be misleading when the data contain outliers. This paper introduces a robust principal component biplot based on ROBPCA method proposed by Hubert, Rousseeuw and Vanden Branden (2005). The length of a vector representing a variable is then approximately proportional to its robust standard deviation while the cosine of the angle between two variables is approximately equal to their robust correlation.

Keywords: Principal component analysis, Robust, Biplot,

1. INTRODUCTION

A biplot is a graphical display of rows and columns of a rectangular $n \times p$ data matrix **X**, where the rows are often subjects or sample units, and the columns are variables. The biplot, introduced by Gabriel (1971), is a joint representation of the n individuals and of the p variables. It is constructed by factoring a $n \times p$ matrix **Z**, by singular value decomposition (SVD), which is a rank - r approximation of **X**, as

$$\mathbf{Z} = \mathbf{G}\mathbf{H}^{\mathbf{T}},$$

where **G** and **H** are $n \times r$ and $p \times r$ matrices respectively, with r usually equal to 2 or 3. Gabriel suggests using "biplot" only for $r = 2$ and using "bimodel" for cases when $r = 3$. The rows of **G** contain the coordinates of the points representing the n individuals. The plot of these n points is called a g – plot. The coordinates of the variables appear in the matrix **H**. Their graphical representation is called the h – plot. A biplot is made of two entities, a h – plot and a g – plot (Daigle and Rivest, 1992). This provides a useful tool of data analysis and allows the visual appraisal of the structure of large data matrices (Gabriel, 1971).

The principal component biplot (PCA biplot) is a graphical tool to simultaneously visualize the scores and loadings of the principal components obtained from the classical principal component analysis. Principal component analysis (PCA) is a popular statistical method, which tries to explain the covariance structure of the data by means of a small number of components. These components are linear combinations of the original variables, and often allow for an interpretation and a better understanding of the different sources of variation. In the classical approach, the first component corresponds to the direction in which the projected observations have the largest variance. The second component is then orthogonal to the first and again maximizes the variance of the projected data points. Continuing in this way produces all the principal components, which correspond to the eigen vectors of the empirical covariance matrix. The scores and loadings (eigen vectors) obtained from the PCA are using for g –plot and h-plot respectively.

The PCA biplot is useful to plant breeders who are conducting large-scale trials to investigate the performance of large numbers of genotypes in several environments with the aim of selecting the "best" genotypes for the purpose of further improvements of crops (Kroonenberg, 1995). This plot can be used for multiplicative models for analyzing G x E interaction. The biplot displays points for varieties and environments from the multiplicative model on the same graph so that the expected response may be derived from visual inspection of the positions on the plot. This is also useful in the areas of manufacturing industry, mining industry, agriculture, finance, archaeology, etc.

Unfortunately, most of the datasets contain anomalous observations or outliers, which are the observations with a unique combination of characteristics and are deviate from the pattern suggested by rest of the data. When the data contain nasty outliers, typically two things happen:

- the multivariate estimates differ substantially from the "right" answer, defined here as the estimates we would have obtained without outliers;
- the resulting fitted model does not allow to detect the outliers by mean of their residual, Mahalanobis distances or the widely used "leave-one-out" diagnostics

(Hubert, Rousseeuw and Van Aelst, 2008). The presence of outliers changes the actual structure of the biplot and hence the inference based on this will be useless or misleading. In this paper we propose a robust principal component biplot based on ROBPCA method which is much less influenced by the outliers.

2. PRINCIPAL COMPONENT BIPLOT (PC BIPLOT)

The principal component biplot is a graphical tool to simultaneously visualise the scores and loadings of the principal components. Let **X** be a *n* x *p* matrix of observations of *n* units on *p* variables. Define **Y** as a *n* x *p* matrix in which the mean of each variable in **X** matrix has been subtracted out, i.e. **Y** = **X** – **M**, where **M** is the mean vector. Then

$$\mathbf{S} = (1/n)\ \mathbf{Y'Y} \qquad (2)$$

is the corresponding *p*-variate estimated variance – covariance matrix. A standardized measure of the distance between the ith and eth units is given by

$$d^2_{i,e} = (y_i - y_e)'\ \mathbf{S}^{-1}((y_i - y_e) \qquad (3)$$

(Seal, 1964, pp.126-127).

Principal component analysis consists of singular decomposition of such a matrix **Y** (Whittle, 1952). This is,

$$\mathbf{Y} = \sum_{\alpha=1}^{r} \lambda_\alpha \mathbf{p}_\alpha\, \mathbf{q}'_\alpha \qquad (4)$$

Where *r* is the rank, λ_α is the singular value, $\mathbf{p}_\alpha$ is the singular column and $\mathbf{q}_\alpha$’ is the singular row. Here λ_α, $\mathbf{p}_\alpha$, and $\mathbf{q}_\alpha$ are chosen to satisfy

$$\mathbf{p}_\alpha'\mathbf{Y} = \lambda_\alpha\, \mathbf{q}_\alpha \qquad (5)$$

$$\mathbf{Y}\mathbf{q}_\alpha = \lambda_\alpha\, \mathbf{p}_\alpha \qquad (6)$$

$$\mathbf{YY'}\mathbf{p}_\alpha = \lambda_\alpha^2\, \mathbf{p}_\alpha \qquad (7)$$

$$\mathbf{Y'Y}\mathbf{q}_\alpha = \lambda_\alpha^2\, \mathbf{q}_\alpha$$

$$\Rightarrow n\mathbf{S}\mathbf{q}_\alpha = \lambda_\alpha^2\, \mathbf{q}_\alpha \qquad (8)$$

$$\lambda_1 \geq, \ldots, \geq \lambda r > 0$$

Any solution to the pair of equations (5) and (6), (5) and (7), and (6) and (8) will satisfy the remaining equations. The method of least squares then provides

$$\mathbf{Y}_{(s)} = \sum_{\alpha=1}^{s} \lambda_\alpha \mathbf{p}_\alpha\, \mathbf{q}'_\alpha \qquad (9)$$

as the rank *s* approximation to **Y** (Householder and Young, 1938).

Now, for biplotting the matrix **Y**, consider the rank 2 approximation $\mathbf{Y}_{(2)}$ of **Y**,

$$\mathbf{Y}_{(2)} = \mathbf{GH'}$$

where $\mathbf{G} = (\mathbf{p}_1, \mathbf{p}_2)\sqrt{n}$ and $\mathbf{H} = (1/\sqrt{n})(\lambda_1\mathbf{q}_1, \lambda_2\mathbf{q}_2)$ (Gabriel, 1971). In biplot the loadings (**H** matrix) represented by arrows and the scores (**G** matrix) represented by data points or sample identifiers. Gabriel (Gabriel, 1971) explains the properties of this decomposition and the plot. The relationships between variables (via loadings) and observations (via scores) can be analysed through this plot. The lengths of the arrows in the plot are directly proportional to the variability included in the two components (PC1 and PC2) displayed, and the angle between any two arrows is a measure of the correlation between those variables.

It is well known that least-square principal component analysis is not robust to the presence of outliers in the data set. The extreme observations may unduly influence the form of the first few principal components, making the *g* - plot a display of the opposition between outliers and the bulk of the data and sometimes suggest relationships (or lack of relationships) among variables which are not representative of the main structure of the data set (Daigle and Rivest, 1992). So, it is always better to look for an alternative biplot which is resistant to the influence of outliers.

3. ROBPCA BIPLOT

A robust principal component biplot is obtained through a robust principal component analysis (RPCA). The goal of RPCA methods is to obtain principal components that are not influenced much by outliers. A first group of methods is obtained by replacing the classical covariance matrix by a robust covariance estimator. This group includes the methods proposed by Maronna (1976), Campbell (1980), Croux and Haesbroeck (2000), Salibian-Barrera, van Aelst and Willems (2006). All these methods are affine equivariant but limited to small to moderate dimensions.

A second approach to robust PCA uses *projection pursuit* (PP) techniques. These methods maximize robust measure of spread to obtain consecutive directions on which the data points are projected. This group of methods includes the methods proposed by Hubert, Rousseeuw and Verboven (2002), Li and Chen (1985) and Croux and Ruiz-Gazen (1996, 2005).

Hubert, Rousseeuw and Vanden Branden (2005) proposed a robust PCA method, called ROBPCA, which combines the ideas of both projection pursuit and robust covariance estimation. The PP part is used for the initial dimension reduction. Some ideas based on the Minimum Covariance Determinant (MCD) estimator then applied to this lower dimensional data space.

Simulations in Hubert, Rousseeuw and Vanden Branden (2005) have shown that this combined approach yields more accurate estimates than the raw PP algorithm. ROBPCA can be computed rapidly, and is able to detect exact-fit situations.

In ROBPCA method the original data are stored in a n x p data matrix $\mathbf{X} = \mathbf{X}_{n,p}$, where n denotes the number of objects and p denotes the original number of variables. The ROBPCA method then proceeds in three major steps. First, the data are preprocessed such that the transformed data are lying in a subspace whose dimension is at most n-1. Next, a preliminary covariance matrix $\mathbf{S}_0$ is constructed and used for selecting the number of components k that will be retained in the sequel, yielding a k-dimensional subspace that fits the data well. Then the data points are projected on this subspace where their location and scatter matrix are robustly estimated, from which its k nonzero eigen values $l_1,\ldots, l_k$ are computed. The corresponding eigenvectors are the k robust principal components.

In the original space of dimension p, these k components span a k-dimensional subspace. Formally, writing the (column) eigen vectors next to one another yields the p x k matrix $\mathbf{P}_{p,k}$ matrix with orthogonal columns. The location estimate is denoted by the p-variate column vector $\hat{\boldsymbol{\mu}}$ and called the robust center. The scores are the entries of the n x k matrix

$$\mathbf{T}_{n,k} = (\mathbf{X}_{n,p} - \mathbf{1}_n\hat{\boldsymbol{\mu}}')\mathbf{P}_{p,k}, \qquad (10)$$

where $\mathbf{1}_n$ is the column vector with all n components equal to 1. Moreover the robust principal components generate a p x p robust scatter matrix $\mathbf{S}$ of rank k given by

$$\mathbf{S} = \mathbf{P}_{p,k}\mathbf{L}_{k,k}\mathbf{P}_{p,k}', \qquad (11)$$

where $\mathbf{L}_{k,k}$ is the diagonal matrix with the eigen values $l_1,\ldots, l_k$. Like classical PCA, the ROBPCA method is location and orthogonal equivariant. Hence the scores do not change under location or orthogonal transformation (Hubert, Rousseeuw and Vanden Branden, 2005). Hubert, Rousseeuw and Vanden Branden (2005) have given the complete description of this method along with an algorithm for ROBPCA.

In this proposed ROBPCA biplot the loadings and scores obtained from the ROBPCA method have been used. To draw the biplot, value of k is fixing to 2 as the plot is two dimensional. Then the p x 2 matrix $\mathbf{P}_{p,2}$ contains the eigenvectors corresponding to the first two components and

$$\mathbf{T}_{n,2} = (\mathbf{X}_{n,p} - \mathbf{1}_n\hat{\boldsymbol{\mu}}')\mathbf{P}_{p,2},$$

consists of the scores corresponding to the first and second principal components. The first two robust principal components generate a p x p robust scatter matrix $\mathbf{S}$ of rank 2 given by

$$\mathbf{S} = \mathbf{P}_{p,2}\mathbf{L}_{2,2}\mathbf{P}_{p,2}', \qquad (12)$$

where $\mathbf{L}_{2,2}$ is the diagonal matrix with the eigen values $l_1$and l_2. The matrices $\mathbf{P}_{p,2}$ and $\mathbf{T}_{n,2}$ have been used to plot the robust principal component biplot. The length of a vector representing a variable is then approximately proportional to its robust standard deviation while the cosine of the angle between two variables is approximately equal to their robust correlation.

4. CRIME IN INDIA DATA - 2012

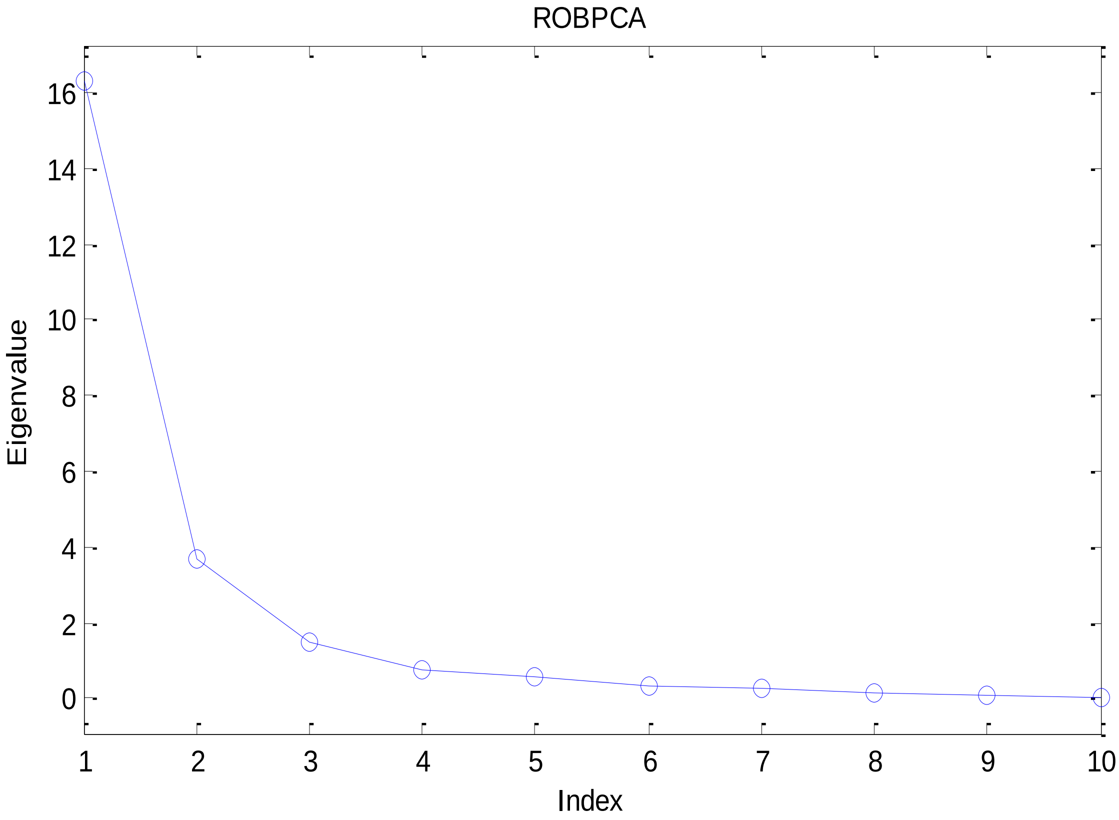
The example is concerned with the crime in India data for the year 2012 from the publication of National Crime Records Bureau (NCRB) and the population data from Census - 2011 report.

Table 1: Variables

VC	Violent crimes
CAW	Crime against women
CAC	Crime against children
CASC	Crime against S.C.
EO	Economic offences
JC	Juveniles in conflict with Law
REC	Recidivism
CC	Cyber Crimes
CCAPP	Complaints/Cases Registered Against Police Personnel
VPS	Value Of Property Stolen
PPKI	Police Personnel Killed Or Injured On Duty
POP	Population
AREA	Area

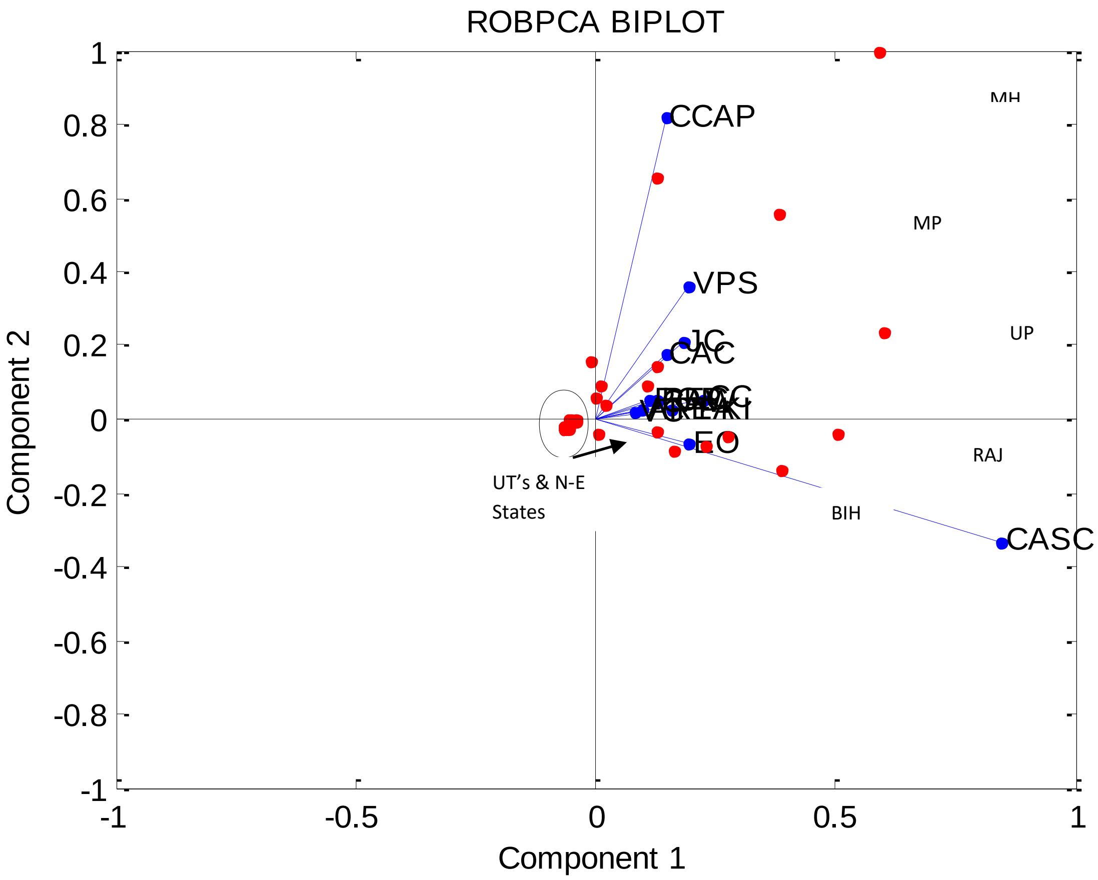
Robust pca biplot explains the correlated structure of the crime in India data. It explains the inter-relationship between states, crimes and states and crimes. Since the variables are measured in different scales, data matrix has been standardized by dividing by their robust standard deviations. MATLAB software has been used to construct the plots. Figure 1 provides the scree plot of Eigen values of the data and it reveals that the first two components are enough to explain the structure of the data. The first two components are explaining 84.17% of the variation of the data.

Figure 1: Scree plot



The ROBPCA biplot in figure 2 reveals the fact that there is a strong correlation between the variables Juveniles in conflict with Law and Crime against children. Violent crimes, Crime against women, Recidivism, Cyber crimes and Police personnel killed or Injured on duty are also correlated. Lengths of the rays give information about the standard deviations of variables and from the figure it is clear that complaints/cases registered against Police personnel and Crime committed against SC has higher standard deviations than others. It is also clear that Maharashtra (MH), Madhya Pradesh (MP), Uttar Pradesh (UP) and Rajasthan (RAJ) are lying far away from other states with respect to its crime behavior.

Figure 2: ROBPCA Biplot



Biplot also provides information about the relationships between variables and observations. For example, the above figure reveals that there is a high relationship between Complaints/Cases Registered against Police Personnel (CCAP) and the capital territory Delhi. From the data table we can see that number of CCAP is very high in Delhi. Also Crime against scheduled caste is very well connected with Bihar. There is a well-developed cluster is visible in the figure which is formed by the Union Territories other than Delhi and North-East states. It means that, these states have similar crime behavior.

5. CONCLUSION

It is well-known that the classical principal component analysis is not robust against the presence of outliers in the data set and hence the principal component biplot too. Consequently, the inference based on this plot will be misleading when the data contain outliers. This paper introduces a robust principal component biplot based on ROBPCA method proposed by Hubert, Rousseeuw and Vanden Branden (2005). It is much less influenced by the outliers and thus can be used for a better study. The variables are represented by the rays of the biplot and the length of the ray is approximately proportional to its robust standard deviation while the cosine of the angle between two rays is approximately equal to their robust correlation.

REFERENCES

1. Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data*, John Wiley & Sons, Chichester, England.
2. Caroni, C. (2000). Outlier detection by robust principal component analysis, *Communications in Statistics: Simulation and Computation*, 29(1), 139-151.
3. Campbell, N.A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation, *Applied Statistics*, 29, 231-237.
4. Croux, C. and Haesbroeck, G. (2000). Principal components analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies, *Biometrika*, 87, 603-618.
5. Croux , C., and Ruiz-Gazen, A. (1996). A fast algorithm for robust principal components based on projection pursuit, *In COMPSTAT 1996* 211-217, Physica, Heidelberg.
6. Croux, C., and Ruiz-Gazen, A. (2005). High break-down estimators for principal components: The projection pursuit approach revisited, *Journal of Multivariate Analysis*, 95, 206-226.
7. Daigle, G. and Rivest L.P. (1992). A robust biplot, *The Canadian Journal of Statistics*, 20, 241-255.
8. Gabriel, K.R (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, 12, 453-467.
9. Gilbert Strang. (2003). *Introduction to Linear Algebra, Third Edition.* Wellesley-Cambridge Press, USA.
10. Hubert, M., Rousseeuw, P.J. and Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis, *Technometrics*, 47, 64-79.
11. Hubert, M., Rousseeuw, P.J. and Van Aelst, S. (2008), High-break down robust multivariate methods, *Statistical Science*, 23, 92-119.

12. Hubert, M., Rousseeuw, P.J., and Verboven, S. (2002). A fast robust method for principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60, 101-111.
13. Joliffe, I.T. (1986). *Principal Component Analysis*. Springler-Verlag, New York.
14. Kroonenberg, P.M. (1995). Introduction to biplots for *G* x *E* tables, *Research Report 51*, Centre for Statistics, The University of Queensland, Brisbane, Australia.
15. Li, G. and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo, *Journal of American Statistical Association*, 80, 759-766.
16. Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter, *Annals of Statistics*, 4, 51-67.
17. Salibian-Barrera, M.., Van Aelst, S. and Willems, G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap, *Journal of American Statistical Association*, 101, 1198-1211.