Original Research Article

THE PRINCIPAL COMPONENT ANALYSIS BIPLOT PREDICTIONS VERSUS THE ORDINARY LEAST SQUARES REGRESSION PREDICTIONS: THE ANTHROPOMETRIC CASE STUDY

Abstract

An indicative feature of a principal component analysis (PCA) variant to the multivariate data set is the ability to transform correlated linearly dependent variables to linearly independent principal components. Back-transforming these components with the samples and variables approximated on a single calibrated plot gives rise to the PCA Biplots. In this work, the predictive property of the PCA biplot was augmented in the visualization of anthropometric measurements namely; weight (kg), height (cm), skinfold (cm), arm muscle circumference AMC (cm), mid upper arm circumference MUAC (cm) collected from the students of School of Nursing and Midwifery, Federal Medical Center (FMC), Umuahia, Nigeria. The adequacy and quality of the PCA Biplot was calculated and the predicted samples are then compared with the ordinary least square (OLS) regression predictions since both predictions makes use of an indicative minimization of the error sum of squares. The result suggests that the PCA biplot prediction merits further consideration when handling correlated multivariate data sets as its predictions with mean square error (MSE) of 0.00149 seems to be better when compared to the OLS regression predictions with MSE of 29.452.

Keywords: Principal Component Analysis, Biplot, Prediction, OLS Regression

1. Introduction

In wide sense statistics is defined as the enterprise dealing with the collection of data sets, analyzing, extraction and presentation of the facts they contain". In the light of this definition it is clear that graphical presentations of a data set form an integral part of any statistical analysis. According to Chambers *et al.* (1983) "there is no single statistical tool that is as powerful as a well-chosen graph", graphical displays not only presentthe information contained in the data but can also be used to extract information that is difficult or even impossible to extract by means of traditional parametric multivariate analyses. In the words of Everitt (1994) "there are many patterns and relationships that are easier to detect in graphical displays than by any other data analysis method".

Applying biplot technique to any multidimensional scaling configuration enhances the informativeness of the lower-dimensional graphical display by adding information regarding the measured variables**.** Gabriel (1971) introduced biplots were, he also coined the name. A biplot is a joint map of the samples and variables of a data set. Gower *et al.* (2011) noted that the 'bi' in 'biplot' refers to the fact that two modes, namely samples and variables, are represented simultaneously and not to the dimension of the display space. The biplot proposed by Gabriel is known as the traditional (or classical) biplot. In the traditional biplot each row (sample) and column (variable) of the data matrix under consideration is represented by a vector emanating from the origin. The differences between the traditional PCA biplot proposed by Gabriel (1971) and the PCA biplot proposed by Gower and Hand (1996) are evident upon comparison. The main weakness of the traditional biplot is that inner products are difficult to visualize. Gower and Hand (1996) addressed this problem by proposing that the (continuous) variables be represented by axes, called biplot axes, which are calibrated such that the approximations to the elements of the data matrix of interest can be read off from the biplot axes by means of orthogonal projection onto the calibrated axes, as is done in the case of ordinary scatter plots. Gower (2003) outlined the fundamental geometry that underlies all biplots of a data-matrix **X** of n cases and p variables, with the cases represented by n points and variables by a reference system. This reference system for quantitative

variables may be orthogonal Cartesian axes, other linear axes or nonlinear trajectories. Greenacre (2012) proposed a new scaling of the display, showing visually the important contributors and thus facilitating the biplot interpretation and often simplifying the graphical representation considerably. Gower *et al.* (2013) studied the underlying theory and quality measures in PCA and CVA biplot with the primary focus on the quality measures associated with these biplots. Thus, this work compares the predictive capabilities of the calibrated PCA biplot to that of the multiple regression technique using anthropometric measurements as a case study. For an enhanced graphics display, the R statistical programming software will be utilized.

2. Methodology

This study is aimed at using the PCA Biplot to study the multidimensional relationship between the variables and also visualizing with predictive capabilities in the multidimensional configurations. Thus, a summary look of the PCA variants, the biplots, and the PCA Biplot is expected. In addition, since the PCA biplot predictions are to be compared with the ordinary least squares a regression prediction, a brief discussion on the OLS regression technique summary is inevitable. These procedures are already in many literary books and thus, a case study application procedure will be used.

2.1 *Data Collection*

The data to be used for this study were collected from a secondary source. A total of sixty (60) students of School of Nursing and Midwifery, Federal Medical Center Umuahia, Nigeria being the number of nurses taken in a single section of the hospital for a clinical experience in 2014 were included in this study. This suggests a total sampling of the data. The data were made up of anthropometric measurements such as weight (kg), height (ft), skinfold (cm), arm muscle circumference AMC (cm), mid upper arm circumference MUAC (cm). For statistical purposes, there were no medical assumptions attached to these measurements. A multivariate plot as packaged by Revelle (2016) of the data shows the scatter, correlation and distributions as displayed in Figure 1.

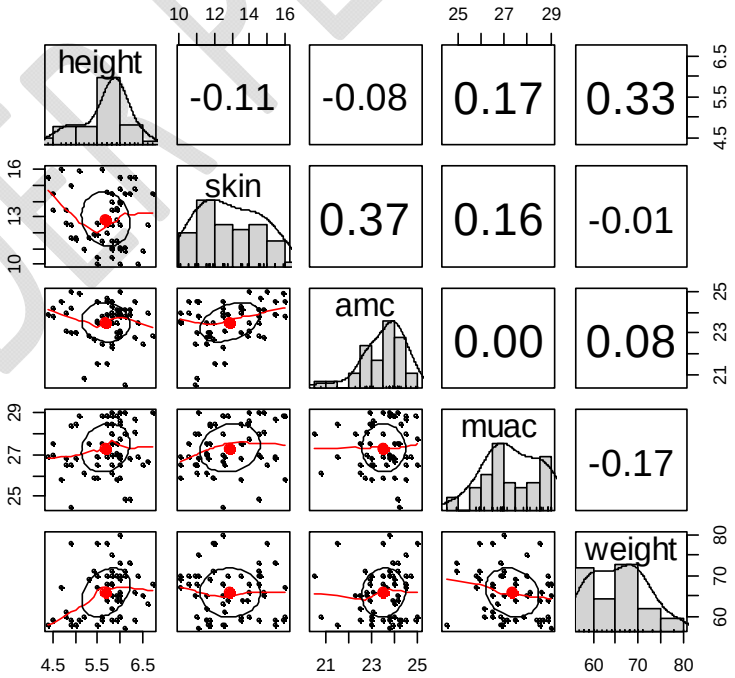

Figure 1.0: *Multivariate scatter plot with correlation, histogram and density lines for the FMC Anthropometric data set using the Psych package by* Revelle (2016).

It is evident in Figure 1.0 that there seems to be very weak correlations among the variables except that of the height vs weight and skin vs muac variables that displayed a relatively positive correlation.

2.2 *The Principal Component Analysis*

The Principal Component Analysis (PCA) is essentially directed 'to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set (Johnson & Wichern, 2007). Pearson (1901) and Hotelling (1933), independently of each other arrived at PCA following two different routes. While Pearson searched to find the straight line or hyperplane which is best fitting to a higher dimensional configuration of points, Hotelling aimed to summarize the total sample variance associated with the set of variables by means of a few uncorrelated linear combinations of the variables. Given a centered matrix $\mathbf{X}: n \times p$ with *p* variables and *n* observations where the *n* observations are denoted by the vector $\boldsymbol{x}_i : i = 1, 2, \ldots n$. Then sample covariance matrix of $\mathbf{X}$ is proportional to $(n-1)\mathbf{S} = \mathbf{X}'\mathbf{X}$ (where $\mathbf{S}$ is the sample covariance matrix) and this can be represented by applying the singular value decomposition (SVD) as :

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{U}'\mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{V}' = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}' \qquad (1)$$

where $\mathbf{V}$ and $\mathbf{U}$ are the left and right singular vectors. The notation $\boldsymbol{\Lambda}: p \times p$ is a diagonal matrix containing the ordered eigenvalues of $\mathbf{X}'\mathbf{X}$ with the eigenvalues $\boldsymbol{\lambda}_i : i = 1, 2, \ldots p$ ordered from the largest to the smallest while the notation $\mathbf{V}: p \times p$ is a matrix containing the orthonormal eigenvectors of $\mathbf{X}'\mathbf{X}$ as it's column vectors, ordered accordingly. The sample principal components (Sample PCs) are given by the *p* column vectors of $\mathbf{V}$ with $\mathbf{v}_i : i = 1, 2, \ldots p$. Thus, the coordinates of the sample PCs in the *p*-dimensional space is the matrix of principal component scores (PC scores), $\mathbf{Z}: n \times p$, given by

$$\mathbf{Z} = \mathbf{X}\mathbf{V} \qquad (2)$$

From the aspect of dimension reduction property of the PCA, the dimensionality of the data matrix $\mathbf{X}$ is reduced to *r*-dimensional space. Let the first *r*-column vectors to be extracted be denoted as $\mathbf{V}_r : p \times r$. Cox and Cox, (2001) showed that $\mathbf{V}_r$ is a matrix containing the first *r* eigenvectors $\mathbf{v}_i : i = 1, 2, \ldots r$ corresponding the *r* largest eigenvalues giving the principal component approximation of $\mathbf{X}$ as

$$\hat{\mathbf{Z}}: n \times p = \mathbf{X}\mathbf{V}_r\mathbf{V}_r' \qquad (3)$$

This PCA approximation uses a least-squares criterion as the basis of approximation to produce the least squared residuals between the original observations in *p*-dimensional spaces and its projection in *r*-dimensional space. This Eckart & Young (1936) minimization is shown to be a minimization problem given by:

$$minimize \quad tr\left\{(\mathbf{X} - \hat{\mathbf{Z}})'(\mathbf{X} - \hat{\mathbf{Z}})\right\} = \left\|(\mathbf{X} - \hat{\mathbf{Z}})\right\|^2 \qquad (4)$$

Note that following the PCA from another route by applying the SVD directly to $\mathbf{X}$ will yield the same equation as (2). This is readily seen as the SVD of $\mathbf{X}$ given by $\mathbf{X} = \mathbf{U}\Omega\mathbf{V}'$ produces the PC scores

$$\mathbf{X}\mathbf{V} = \mathbf{U}\boldsymbol{\Omega} = \mathbf{Z} \qquad (5)$$

since $\mathbf{V}$ is orthonormal, multiplying by $\mathbf{V}$ on the right of $\mathbf{X} = \mathbf{U}\Omega\mathbf{V}'$ with the Ω notation equivalent to $\boldsymbol{\Lambda}$. Following this approach, the *r*-dimensional subspace with the largest *r*-singular values of $\boldsymbol{\Omega}$ extracted and denoted as $\boldsymbol{\Omega}_\mathbf{r}$ could be obtained. Denoting $\mathbf{U}_r$ and $\mathbf{V}_r$ to be the first *r* columns of $\mathbf{U}$ and $\mathbf{V}$, respectively with the best *r*-dimensional approximation of the data matrix $\mathbf{X}$ obtained as

$$\mathbf{X} \approx \mathbf{X}\mathbf{V}_r\mathbf{V}_r' = \mathbf{U}_r\Omega_r\mathbf{V}_r' \qquad (6)$$

Thus, given a multivariate data set **X** with n samples and p variables the fundamental problem of PCA is to approximate **X** by *r* dimensions or, equivalently, of rank *r*.

2.3 *The Biplot*

Biplots are plots, usually with two or more dimensions, the *bi* signifying two modes, the observations or samples and the variables or axis. This two modes are displayed in two-dimensional (2D) spaces popularly known as the 2D biplots, which are the most common (Displays). Other biplot displays that can be constructed are the one dimensional (1D) and the three-dimensional (3D) space biplots. In general, one can visualize the 2D configuration as a display on a surface, the 3D as a display on a sphere, and the 1D display as that on a line. A brief summary underlying the construction of a Biplot is showcased in the book Gower *et al.* (2011).

2.4 *The Principal Component Analysis (PCA) biplot*

A Principal Components Analysis Biplot (PCA Biplot) is a two-dimensional chart that represents the relationship between the samples and variables of a given in the same plot in which they are being transformed with aid of PCA. The PCA biplot provides linear axes for points placed by PCA.

Section 2.2 shows that the best *r*-dimensional subspace to represent observations from a *p*-dimension space is determined by $\mathbf{V}_r$. A set of orthogonal coordinate axes in the *r*-dimensional space is provided by $\mathbf{V}_r$ and this is called the principal axes (PA). The PA is also known as the scaffolding axes since they are only used for representing the biplot observations.

2.5 Biplot Interpolation

Gower *et al.* (2011) noted that the biplot observations are determined as projections from the principal axes and are given by,

$$\mathbf{Z}_r = \mathbf{X}\mathbf{V}_r \qquad (7)$$

In (7), the rows of $\mathbf{Z}_r$ are the PC scores of the first *r* sample PCs given by $\boldsymbol{z}_i : i = 1, 2, \ldots n$, of which the first *r* samples makes the difference between (7) and (2).

From the interpolation point of view, a new p-variable observation $\boldsymbol{x}^* : p \times 1$ needs to be projected to an observation in the *r*-dimensional space as $\boldsymbol{z}^* : p \times 1$. Analogously using the new $\boldsymbol{x}^*$ and $\boldsymbol{z}^*$ on (7), this *r*-dimension projection produces

$$\boldsymbol{z}^{*\prime} = \boldsymbol{x}^{*\prime}\mathbf{V}_r \qquad (8)$$

2.6 Biplot Prediction

In prediction, the original *p*-variable observation must be approximated as $\hat{\boldsymbol{x}}^* : p \times 1$ from the coordinates in the *r*-dimensional space $\boldsymbol{z}^*$. Gardner (2001) summarizes this using (8) as

$$\hat{\boldsymbol{x}}^{*\prime} = \mathbf{z}^{*\prime}\mathbf{V}\mathbf{V}_r' \qquad (9)$$

Note that focus will be placed on (9) as this case study will compare the predicted samples with that of the predicted samples obtained from the OLS approach.

2.7 Calibration of Biplot in 2D

An important step in the construction of the biplot axes is the plotting the axes that correspond to the *p* variables of the data. Although the axes for prediction and interpolation will differ in terms of the position of the axes markers, the different axes markers are determined by some value of θ, with $-\infty < \theta < \infty$. Let $\boldsymbol{w}_k : r \times 1$ represent a unit vector with the kth element equal to one and all other elements equal to zero. Gower *et al.* (2011) showed that each observation x_i with coordinates $\left(x_{i,1}, x_{i,2}, \ldots, x_{i,p}\right)$ can be expressed as $\boldsymbol{x}_i = \sum_{k=1}^{p} x_{i,k}\boldsymbol{e}_k$ giving the interpolation matrix as $\boldsymbol{z}_i' = \boldsymbol{x}_i'\mathbf{V}_r = \sum_{k=1}^{p} x_{i,k}\boldsymbol{w}_k'\mathbf{V}_r$. Interestingly, since the kth interpolation biplot axis markers is determined by $\mu\boldsymbol{w}_k\mathbf{V}_r$, it can be shown that the corresponding kth prediction biplot axis markers as θ varies is given by

$$\frac{\theta\boldsymbol{w}_k\mathbf{V}_r}{\boldsymbol{w}_k\mathbf{V}_r\mathbf{V}'\boldsymbol{w}_k'} \tag{13}$$

2.8 *Prediction in Ordinary Least Squares (OLS) Regression*

Given the general linear model as shown in Equation (6) with the notations y_i as the independent, x_i independent variables, β_i parameters and ε_i residuals.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{14}$$

The matrix notation of the linear regression model in (14) is given in Equation (15).

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{15}$$

The condensed matrix notation of Equation (15) will yield

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{16}$$

From equation (16), the ordinary least squares (OLS) estimates seeks for a function of the estimates $\boldsymbol{\beta} = \left[\beta_1, \beta_2, \ldots, \beta_{ik}\right]'$ of the fitted coefficients that minimizes the error sum of squares

$$f(\boldsymbol{\beta}) = \sum \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \tag{17}$$

upon minimization, (17) yields $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$ with the vector of the OLS predicted or fitted values gives

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{18}$$

3. Comparative Results based on the PCA biplot and the OLS

There is a basis for comparing the PCA biplot predictions and the OLS prediction. Notice that the two methods are based on the minimization of the error sum of squares as indicated in (3) and (17). Figure 2 showcases the predictive biplot of the FMC anthropometric data set introduced in Section 2.1. The quality of the display in 2D configuration is 94.65% and this is reasonable enough to rely on information or interpretations that could be made from the biplot. The singular value decomposition (SVD) of the **X** matrix shows that the overall PCA Quality (94.65%), the Adequacy (0.002, 0.908, 0.066, 0.027, 0.998) and Predictivity (0.118, 0.975, 0.238, 0.093, 1.000) for the respective variables (height, skin, amc, muac, weight).

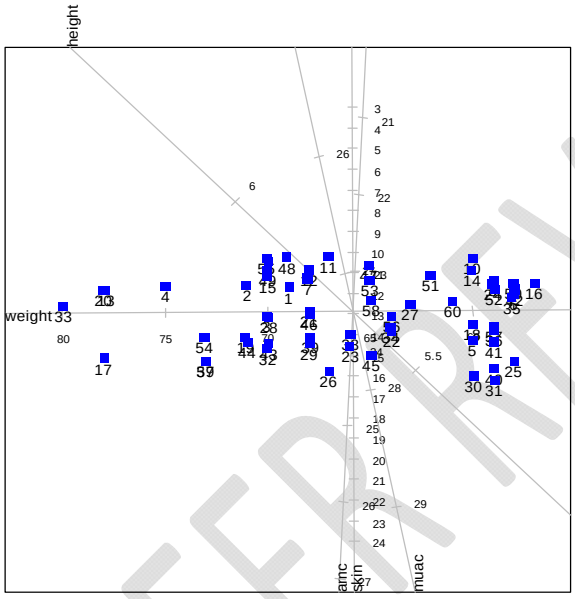

Figure 2: *A Predictive PCA Biplot display of the FMC data set with Quality of Display = 94.65%*

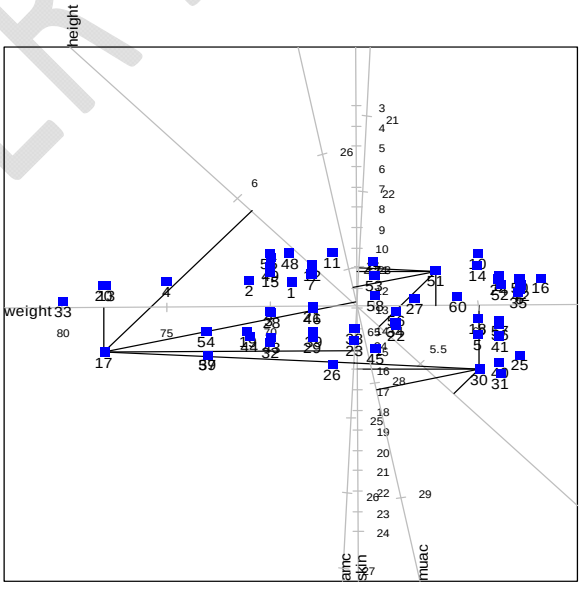

Figure 3: *A Predictive PCA Biplot display of the FMC data set with predictions of samples 17, 30 and 51.*

Our concern is to predict each of the sample points in the biplot and minus its value from the original data values to form the vector of residuals for each of the variables. For instance, Figure 3 showcases a scenario where this is done for only three samples with results given in Table 1. This process was used to predict the entire samples as shown in Figure 4. From the predictions recorded, the vector of residuals is calculated for each of the variables.

Table 1: *Predictions of Samples 17, 30 and 51 using the PCA biplots.*

	s17		s30		s51	
	Actual	**Predicted**	**Actual**	**Predicted**	**Actual**	**Predicted**
height	4.9000	5.9617	5.6000	5.4090	5.8000	5.6140
skin	15.0000	15.0919	16.0000	16.0082	11.5000	11.1757
amc	25.0000	24.2767	23.8000	24.2828	22.5000	23.0136
muac	26.5000	27.2886	28.9000	28.0557	26.1000	27.1647
weight	78.0000	78.0054	60.0000	59.9705	62.0000	62.0347

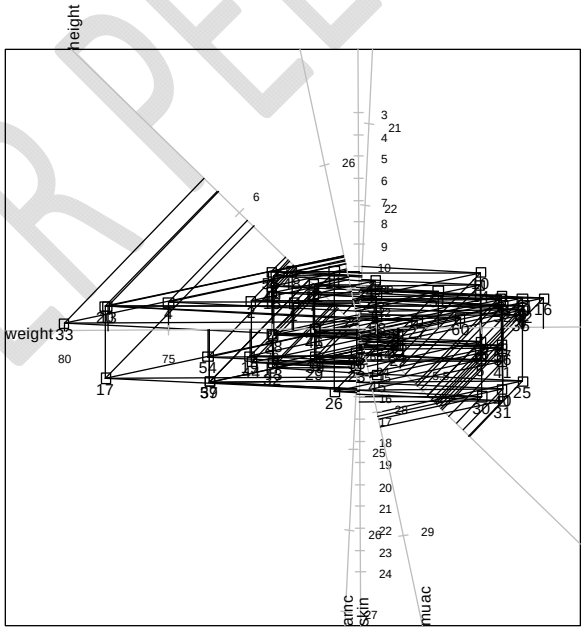

Figure 4: *Predictive PCA Biplot display of the FMC data set with predictions for all the samples (1 – 60).*

Taking the *weight* variable as a case study (weight variable is the response variable), the residuals of the PCA biplots prediction for this variable (for the full samples 1 - 60) are obtained by:

$$\textbf{PCA Biplots Residuals}_{\text{weight}} = \textbf{actual data samples}_{\text{weight}} - \textbf{PCA Biplots predicted results}_{\text{weight}} \quad (19)$$

The results obtained from Equation (11) are compared using the Mean Square Error (MSE) and the Standard Error (SE) of the residuals from the OLS residuals results obtained from (12). Note that the OLS result is simply modeled in the R software (R Core Team, 2017) using the code:

$$\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \text{lm(weight} \sim ., \text{ data = fmc.data)\$residuals} \quad (20)$$

Table 2: *MSE and SE results of the OLS Residuals and PCA Biplot Residuals (response variable=weight).*

	OLS Prediction Residuals	**PCA Biplot Prediction Residuals**
MSE	29.4535	0.001491
SE	0.7065	0.005027

4. Conclusion

The results obtained from Table 2 suggests that the PCA biplot merits further consideration when handling correlated multivariate data sets as its predictions with mean square error (MSE) of 0.00149 seems to be better when compared to the OLS regression prediction MSE of 29.452. One indicative reason is the PCA Biplots ability to transform a set of linearly dependent and correlated data matrix to linearly independent principal components, which are uncorrelated. Thus, predictivity with the PCA Biplots are envisaged to perform better than the OLS regression predictions especially in cases where all the OLS assumptions like multicolinearity is not eliminated.

REFERENCES

Chambers,J.,Cleveland, W., Kleiner,B. and Tukey, P. (1983). *Graphical Methods for Data Analysis.* Belmont, California.: Wadsworth International Group.

Cox,T and Cox,M. (2001). *Multidimensional Scaling.* Boca Raton,FL: Chapman & Hall/CRC.

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

Everitt, B. (1994). Exploring multivariate data graphically: A briefly review with examples. *journal of Applied Statistics*, vol.21 (3),pp. 63-92.

Gabriel, K. R. . (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 453-467.

Gower, J. C. (2003). Unified Biplot Geometry. *Developments in Applied Statistics.*

Gower, J.C. and Hand, D.J . (1996). *Biplots.* London,UK: Chapman & Hall.

Gower, J.C., Lubbe, S. and le Roux N . (2010). *Understanding Biplots.* Wiley.

Greenacre, M. (2010). *Biplots in Pratice.* Madrid,Spain: BBVA Foundation.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, vol. 24,pp. 417-441,4498-520.

Johnson, R. and Wichern, D. (2002). *Applied Multivariate Statistical Analysis.* Upper Sadle River: New Jersey: 5th edn. Prentice-Hall.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophicxal Magazine*, vol. 2, pp. 559-572.

R Core Team (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Revelle, W. (2016) psych: *Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.6.12.