The Covariance Biplot to Reveal Relationships between Different Sets of Variables

Opeoluwa F. Oyedele[1]*, Sugnet Gardner-Lubbe[1]

[1]Department of Statistical Sciences

University of Cape Town, South Africa

Received: 10th March, 2014. Accepted: 24th April, 2014.

Abstract

The biplot has many advantages, including demonstrating the association between samples and (or) variables of a data set graphically. In this paper, the covariance biplot is used as a visual tool for exploring the relationships between different sets of variables.

Keywords: Biplot, Covariance matrix

ISTJN 2014; 3(1): 107-113.

1 Introduction

Often, in statistics, relationships between different sets of variables are of interest. In this situation, various statistical techniques can be used as tools for analysing these relationships. Among them is the *covariance matrix*.

Consider a set of variables $\mathbf{X} : N \times P$ and a set of variables $\mathbf{Y} : N \times M$. The covariance between $\mathbf{X}$ and $\mathbf{Y}$ is defined by the $(P \times M)$ covariance matrix

$$cov(\mathbf{X},\mathbf{Y}) = \frac{1}{(N-1)}(\mathbf{X}^T\mathbf{Y}). \tag{1}$$

*Corresponding author - E-mail: OpeoluwaOyedele@gmail.com

However, when only one set of variables is under consideration, the covariance matrix is referred to as the *variance-covariance matrix*

$$cov(\mathbf{Y},\mathbf{Y}) = \frac{1}{(N-1)}(\mathbf{Y}^T\mathbf{Y}). \tag{2}$$

This is also written as $cov(\mathbf{Y},\mathbf{Y}) = var(\mathbf{Y},\mathbf{Y}) = cov(\mathbf{Y})$. Here, the variances of $\mathbf{Y}$ are given on the diagonal of (2), while the covariances are shown off-diagonal.

The relationships between different sets of variables can be explored using some form of graphical display such as the biplot, which is a joint graphical display of all rows and columns of a data matrix. Since biplots are useful graphical tools for exploring the relationships between variables, in this paper, the biplot is employed in the form of the *covariance biplot*. In addition, a demonstration, with graphical illustration, is presented on how the covariance biplot can help reveal variable and inter-variable relationships.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the biplot, before its employment in the form of the covariance biplot is discussed. This is followed by an application with a sensory data to investigate the relationships between the sensory panel descriptors and the chemical quality measurements of olive oils. Finally, some concluding remarks are presented in Section 3.

2 The Biplot

By definition, the biplot is a joint graphical display of rows and columns of a data matrix $\mathbf{D} : C \times W$ by means of markers $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_C$ for its rows and markers $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_W$ for its columns (Barnett, 1981). Each marker is chosen in such a way that the inner product $\mathbf{a}_i^T\mathbf{b}_j$ represents d_{ij}, the $(i,j)^{th}$ element of the data matrix $\mathbf{D}$. That is,

$$\hat{\mathbf{D}} = \mathbf{A}\mathbf{B}^T. \tag{3}$$

Matrices $\mathbf{A} = (\mathbf{a}_1^T \ \mathbf{a}_2^T \ \cdots \ \mathbf{a}_C^T)^T$ and $\mathbf{B} = (\mathbf{b}_1^T \ \mathbf{b}_2^T \ \cdots \ \mathbf{b}_W^T)^T$ in (3) are called the $(C \times r)$ row markers and $(W \times r)$ column markers matrices respectively. Generally, the number of columns in $\mathbf{A}$ and $\mathbf{B}$ are determined by the rank r approximation of $\mathbf{D}$. In practice, $r = 2$ is usually preferred for a convenient biplot display.

In the biplot display, the rows of a data matrix are represented by points, while the columns are represented by vectors or axes. Traditionally, columns are represented by vectors (Gabriel, 1971), but Gower and Hand (1996) introduced axes to make the biplot similar to a scatter plot. This was done by extending the vectors, which represent the columns, through the biplot space to become axes. Thus, the biplot points will be defined by the row markers of the data matrix, whereas the biplot axes will be defined by the column markers. To be precise, for the biplot of a data matrix **D**, C rows of **A** will serve as the biplot points, while W rows of **B** will be used in calculating the directions of the biplot axes.

2.1 Covariance Biplot

In general, there are two kinds of features displayed in the biplot. These features can be specified as two sets of variables, or as a set of variables and samples, as in the case of the principal component analysis (PCA) biplot (Gower et al., 2011).

As only variables are represented in the covariance and variance-covariance matrices in (1) and (2) respectively, the biplot can be used as a graphical tool to explore their relationships. Consider the $(P \times M)$ covariance matrix between **X** and **Y** defined in (1). Let $\mathbf{S}_{XY} = cov(\mathbf{X}, \mathbf{Y})$. By the singular value decomposition (SVD), $\mathbf{S}_{XY} = \mathbf{U}\Lambda\mathbf{V}^T$, for $\mathbf{U}(P \times M)$, $\Lambda(M \times M)$ and $\mathbf{V}(M \times M)$. The matrix $\mathbf{S}_{XY} = \mathbf{U}\Lambda\mathbf{V}^T$ can be written as

$$\mathbf{S}_{XY} \equiv \hat{\mathbf{S}}_{XY} = \mathbf{U}\Lambda\mathbf{J}\mathbf{V}^T = \mathbf{U}\mathbf{J}\Lambda\mathbf{V}^T = \mathbf{U}\mathbf{J}\Lambda\mathbf{J}\mathbf{V}^T = \mathbf{G}\mathbf{H}^T \tag{4}$$

where $\mathbf{G} = \mathbf{U}\Lambda^{\alpha}\mathbf{J}_r$ and $\mathbf{H} = \mathbf{V}\Lambda^{(1-\alpha)}\mathbf{J}_r$, for any value of $\alpha \in (0,1)$.

In (4), the matrix **J** has dimension $(M \times M)$, while the matrix $\mathbf{J}_r$ has dimension $(M \times r)$. The matrix $\mathbf{G} : P \times r$ contains the information about the X-variables, while $\mathbf{H} : M \times r$ contains the information about the Y-variables.

The covariance matrix expression $\hat{\mathbf{S}}_{XY} = \mathbf{G}\mathbf{H}^T$ has an arrangement similar to (3). Now, focusing on revealing the relationships between these two sets of variables **X** and **Y**, only axes will be present in the resulting biplot. However, two sets of axes are needed, a set for the X-variables and a set for the Y-variables. From $\hat{\mathbf{S}}_{XY} = \mathbf{G}\mathbf{H}^T$, the directions of the axes representing the X-variables are calculated using the P rows of **G**, while M rows of **H** are used to calculate the directions of the axes representing the Y-variables. This biplot, called the covariance biplot, reveals the relationships between the two sets of variables as well as within each set.

When $\alpha = 0$, the covariance biplot display only approximates optimally the covariance between the Y-variables. Conversely, when $\alpha = 1$, the covariance biplot display only approximates optimally the covariance between the X-variables. However, choosing the symmetric choice of $\alpha = 1/2 = 0.5$, the covariance between the X-variables and Y-variables is both equally approximated in the covariance biplot display, although not optimal (Oyedele, 2013).

2.2 An Illustration

The following illustration of the covariance biplot is performed using the olive oil data from Mevik & Wehrens (2007). This data shows the sensory and chemical quality evaluations of sixteen olive oil samples. There were five chemical quality measurements (`Acidity, Peroxide, K232, K270` and `DK`) taken, and six sensory panel descriptors (`Yellow, Green, Brown, Glossy, Transparent` and `Syrup`) were used in this evaluation. The sixteen olive oils are assigned as samples, while the chemical quality measurements and sensory panel descriptors are assigned as the X- and Y-variables respectively. Thus, $\mathbf{X} : 16 \times 5$ and $\mathbf{Y} : 16 \times 6$ in this analysis.

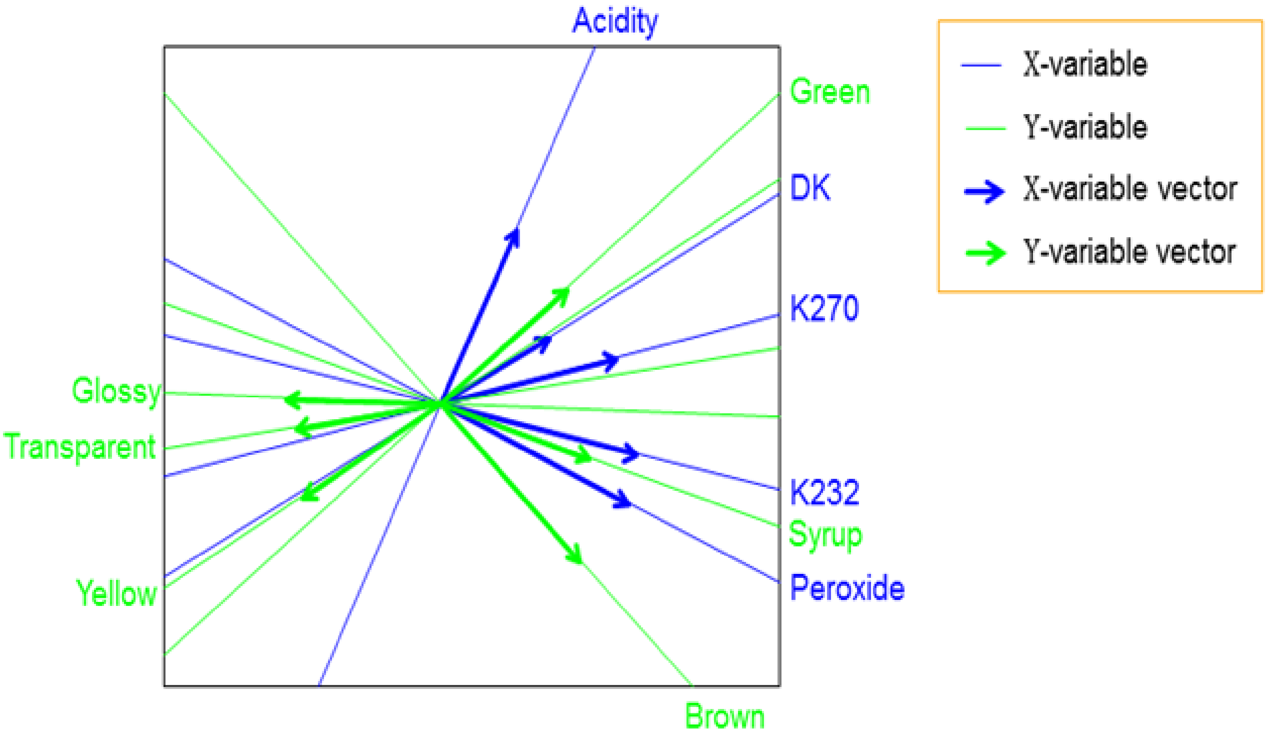
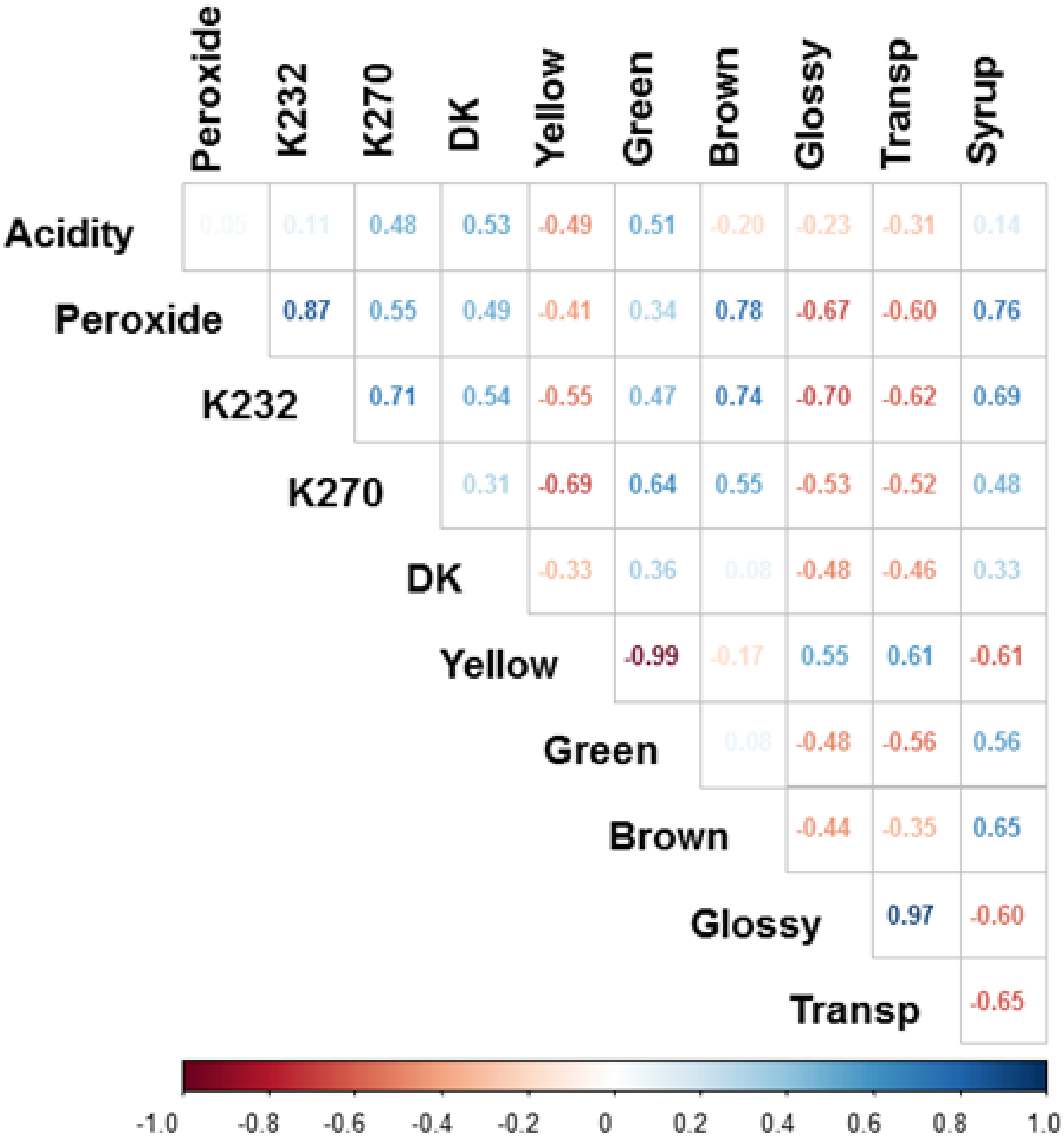


Figure 1: The covariance biplot of the olive oil data.

The covariance biplot for the data is shown in Figure 1, with $\alpha = 0.5$ and $r = 2$. From this biplot, the standard deviation of chemical `DK` is smaller compared to the other chemicals. This is apparent from the length of the thicker arrow (vector) on the `DK` axis. Likewise, descriptor `Brown` can be seen to have a larger standard deviations compared to the other descriptors.

Furthermore, the positions of the biplot axes give an indication of the correlations between the variables. Axes forming small angles are said to be strongly correlated - either positively



	Peroxide	K232	K270	DK	Yellow	Green	Brown	Glossy	Transp	Syrup
Acidity	0.05	0.11	0.48	0.53	-0.49	0.51	-0.20	-0.23	-0.31	0.14
Peroxide		0.87	0.55	0.49	-0.41	0.34	0.78	-0.67	-0.60	0.76
K232			0.71	0.54	-0.55	0.47	0.74	-0.70	-0.62	0.69
K270				0.31	-0.69	0.64	0.55	-0.53	-0.52	0.48
DK					-0.33	0.36	0.08	-0.48	-0.46	0.33
Yellow						-0.99	-0.17	0.55	0.61	-0.61
Green							0.08	-0.48	-0.56	0.56
Brown								-0.44	-0.35	0.65
Glossy									0.97	-0.60
Transp										-0.65

Figure 2: The correlation values of the olive oil data. Nearly empty cells have a value very close to zero.

or negatively. Axes are positively correlated when they lie in the same direction, while negatively correlated axes lie in opposite directions. In addition, axes that are close to forming right angles are said to be uncorrelated. From Figure 1, looking at the angles between the blue vectors, all the chemicals measurements can be said to be positive related with each other. Similarly, descriptors `Glossy` and `Transparent` can be said to be positively related, while descriptors `Green` and `Yellow` are said to be negatively related. Descriptors `Brown` and `Syrup` can be said to be (some-what) positively related. The actual correlation values of this data is shown in Figure 2.

Moreover, various inter-variable relationships can be observed in Figure 1, such as a relation between the chemical `K270` and the descriptors `Glossy` and `Transparent`. Observing their (actual) correlation values (-0.53, and -0.52), shown in Figure 2, indicates a fair relationship between them. Also, a relation between descriptor `Syrup` and chemicals `K232` and `Peroxide` (0.69 and 0.76); and between chemical `DK` and descriptors `Green` and `Yellow` (0.36 and -0.33) can be noted. The latter relation is not quite as fair as the former. `Acidity` and `Brown` can be seen to have no clear relation with the others.

To illustrate how a covariance biplot can help to reveal relationships within one set of variables, consider only the green axes in Figure 1. Descriptors `Glossy, Transparent` and `Syrup` can be said to be related. Also, descriptors `Yellow` and `Green` can be said to be related. However, descriptor `Brown` can be seen to have no clear relation with the others.

3 Conclusion

One of the means used in analysing the relationships between different sets of variables is the covariance matrix. It describes how much two sets of variables change together. The covariance matrix of two sets of variables can be visualized graphically using the biplot. The resulting biplot is termed the covariance biplot. This biplot provides a single graphical display for revealing and exploring the relationships between two sets of variables as well as within each set.

Software

A collection of functions has been developed in the R language (R Core Team, 2013) to produce the covariance biplot. This is available electronically at `https://www.dropbox.com/s/mhmehp6tlyrzkzq/The_covariance_biplot_function.r`.

Acknowlegements
The National Research Foundation of South Africa and the University of Cape Town Postgraduate Centre and Funding Office are acknowledged.

References

[1] Barnett V. Interpreting Multivariate Data. Wiley Series in Probability and Mathematical Statistics. Wiley: New York, (1981).

[2] Gabriel K.R. The Biplot Graphic Display of Matrices with Application to Principal Component Analysis. Biometrika, 58, 453-467, (1971).

[3] Gower J.C., Hand D.J. Biplots. Chapman & Hall: London, (1996).

[4] Gower J.C., Lubbe S. and Le Roux N.J. Understanding Biplots. John Wiley & Sons: Chicester, (2011).

[5] Mevik B.H., Wehrens R. The pls Package: Principal Component and Partial Least Squares Regression in R. Journal of Statistical Software, 2(18), 1-24, (2007).

[6] Oyedele O.F. The Construction of a Partial Least Squares Biplot. Unpublished Ph.D. Thesis, University of Cape Town, South Africa, (2013).

[7] R Core Team (2013) R: A Language and Environment for Statistical Computing, the R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.