Chapter 10
Comparison of Methods to Display Principal Component Analysis, Focusing on Biplots and the Selection of Biplot Axes

Carla Barbosa
REQUINTE-LAQV, Instituto Politécnico de Viana do Castelo, Portugal

M. Rui Alves
REQUINTE-LAQV, Instituto Politécnico de Viana do Castelo, Portugal

Beatriz Oliveira
REQUINTE-LAQV, Faculdade de Farmácia, Universidade do Porto, Portugal

ABSTRACT

Principal components analysis (PCA) is probably the most important multivariate statistical technique, being used to model complex problems or just for data mining, in almost all areas of science. Although being well known by researchers and available in most statistical packages, it is often misunderstood and poses problems when applied by inexperienced users. A biplot is a way of concentrating all information related to sample units and variables in a single display, in an attempt to help interpretations and avoid overestimations. This chapter covers the main mathematical aspects of PCA, as well as the form and covariance biplots developed by Gabriel and the predictive and interpolative biplots devised by Gower and coworkers. New developments are also presented, involving techniques to automate the production of biplots, with a controlled output in terms of axes predictivities and interpolative accuracies, supported by the AutoBiplot.PCA function developed in R. A practical case is used for illustrations and discussions.

1. INTRODUCTION

Principal components analysis, which will be referred to as PCA, is one of the most important multivariate analysis, being used in profusion by researchers throughout the world and in many fields of science. Its origins may be traced back to Pearson (1901) and mainly to Hotelling (1933a, 1933b) who gave it the current approach through the eigenvalue decomposition (or spectral decomposition) of a covariance

DOI: 10.4018/978-1-4666-8823-0.ch010

matrix. With the advent of powerful computers and software, PCA is still the basis for many developments in multivariate statistics. Among the many references available on PCA, the review by Dunteman (1989) is recommended for a quick overview, as well as the textbook by Jolliffe (2002), which is a comprehensive approach to PCA, including its history, mathematical developments, examples of applications and relationships with other multivariate techniques. The importance of PCA is due to the fact that it tries to simplify complex data matrices by reducing the number of variables necessary for interpretation of a given problem, in a process known as parsimonious description of the data, reduction in dimension, or data compression. This simplification is achieved without losing relevant information.

However, although being a well-known statistical technique, PCA encloses some problems that are seldom forgotten and sometimes not well understood. One of the main problems of PCA is that final solutions usually require interpretations subject to some individual judgement, which may easily lead non-statisticians to erroneous conclusions, among which overestimations are the most common problem. Moreover, with the use of modern computers and sophisticated statistical software, these problems may assume larger proportions, unless some mathematical work is carried out in order to control and automate PCA outputs.

After analyzing the advantages, problems and pitfalls of PCA, biplots will be presented, starting with a special reference to the pioneer work of Gabriel (1971, 1972, 1981). The approaches presented in the book by Greenacre (2010), with reviews on Gabriel's biplots and extensions to almost all multivariate analyses, will be followed. These biplots are a way of displaying PCA results in a single graph containing both the information on variables and on sample units. In this way, it is possible to interpret PCA results by relating sample units directly to initial measurement variables, removing the need for the intermediate step of principal component interpretation, thus reducing to some extent the randomness in judgements. As this type of biplot developed by Gabriel (1971) is now-a-days available in many statistical packages, e.g., Statistica (Statsoft, 2014) and SPSS (IBM, 2014) it deserves some attention, and its variants, advantages and pitfalls will be highlighted. In order to produce Gabriel'sbiplots, a function was built by the authors in the R language, called *Gabriel.PCA*, which produces two types of Gabriel's biplots. This function is available to interested users.

Although Gabriel and later mathematicians referred the possibility to calibrate biplot axes, it happens that biplots following Gabriel's methodology are not usually calibrated. As it will be discussed in this chapter, calibration of biplot axes is a usual procedure in relation to the predictive and interpolative biplots developed by Gower and Hand (1996). Predictive biplots with calibrated axes refer to the procedure by which variables are represent in biplots as axes with convenient scales for measurement. These biplots are an important step forward since they enable to relate the position of sample points in principal component displays directly to initial variables and also initial values, making interpretations less cumbersome. Moreover, it becomes easier to check the precision, or validity of the PCA solutions. On the other hand, interpolative biplots are another important step forward since they may be used as print-outs enabling an easy utilization of PCA outputs in current laboratory or field works. Gower and coworkers applied predictive and interpolative biplots to a huge number of multivariate analysis, and refined them to an outstanding level using the facilities of the R statistical project (R development core team, 2010), as it is shown in their recent book (Gower et al., 2011).

After discussing PCA and presenting Gabriel's and Gower's biplots, two innovations are presented in this chapter, in relation to the control of the output of PCA predictive and interpolative biplots, following the work of Alves (2012).It will be shown that it is possible to estimate the size of the error that