2019

Interpreting and Reporting Principal Component Analysis in Food Science Analysis and Beyond

D. Cozzolino

A. Power

J. Chapman

Food Analytical Methods (2019) 12:2469–2473
https://doi.org/10.1007/s12161-019-01605-5

Interpreting and Reporting Principal Component Analysis in Food Science Analysis and Beyond

D. Cozzolino[1] · A. Power[2] · J. Chapman[1]

Received: 10 July 2019 / Accepted: 15 July 2019 / Published online: 24 July 2019

Abstract
Principal component analysis (PCA) is one of the most widely used data mining techniques in sciences and applied to a wide type of datasets (e.g. sensory, instrumental methods, chemical data). However, several questions and doubts on how to interpret and report the results are still asked every day from students and researchers. This brief communication is inspired in relation to those questions asked by colleagues and students. Please note that this article is a focus on the practical aspects, use and interpretation of the PCA to analyse multiple or varied data sets. In summary, the application of the PCA provides with two main elements, namely the scores and loadings. The scores provide with a location of the sample where the loadings indicate which variables are the most important to explain the trends in the grouping of samples.

Keywords Principal components · Scores · Loadings · Data sets

Introduction

The application or use of multivariate data analysis often starts out with data involving a substantial number of correlated variables (e.g. wavelengths, retention times, peaks, sensory scores) where different algorithms and techniques are used to analyse and interpret such types of data sets (Brereton 2000, 2008, 2015; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Mutihac and Mutihac 2008; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014; Kumar et al. 2014). Thus, principal component analysis (PCA) appears to be one of the most frequently used multivariate data analysis methods in exploratory data analysis and data mining (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014).

✉ D. Cozzolino
daniel.cozzolino@rmit.edu.au

[1] School of Science, RMIT University, GPO Box 2476, Melbourne, Victoria 3001, Australia

[2] Centre for Research in Engineering and Surface Technology (CREST), FOCAS Institute, Technological University Dublin, City Campus, Kevin Street, Dublin D08 NF82, Ireland

The PCA method aims to extract the main orthogonal contributors (principal components) which explain most of the variance of the data matrix analysed (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014; Bevilacqua et al. 2014). Depending on the software or application, the rotated variables can be also quantitatively interpreted as possible sources of variation (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014).

The PCA is also considered a dimension-reduction technique that can be used to reduce a large set of variables to a small set that still contains most of the information derived from the original set of variables used to analyse the sample (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014). In this way, complex data sets can be easily analysed. Some authors also highlighted that the interpretability of PCA can be enhanced by the so-called VARIMAX rotation, a variation of coordinates that maximises the sum of the variance of the loading vectors (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014). In general terms, the PCA reduces the information which originated from a larger number

of variables to a smaller number of factors or components. These factors or components are defined as "non-dependent" (e.g. orthogonal) (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014).

This brief communication is inspired by those questions raised by colleagues and students about the use, interpretation and reporting of the results derived from the use of the PCA. Please note that this article is a focus on the practical aspects of the use and interpretation of the PCA to analyse different types of data sets (e.g. sensory, instrumental data).

A Theoretical Framework

The principal component analysis (PCA) is used as a tool able to provide with an overview of the complexity and interrelationships that exist in multivariate data sets (Bro and Smilde 2014). This method is generally used for revealing relations between variables and between samples (e.g. patterns), detecting outliers, finding and quantifying patterns and trends, extracting and compressing multivariate data sets, among other applications (Naes et al. 2002; Brereton 2008, 2015; Cozzolino 2012; Bro and Smilde 2014). Although this technique is extensively used and reported by several authors in many applications, the PCA cannot be considered as a classification method (Brereton 2009, 2015; Bro and Smilde 2014). It is important to highlight this point as many papers and reports in the literature stated the PCA as a classification technique (Naes et al. 2002; Brereton 2009, 2015; Cozzolino 2012; Bro and Smilde 2014).

The PCA employs a mathematical procedure that transforms a set of possibly correlated response variables into a new set of non-correlated variables, called principal components (Bro and Smilde 2014). The PCA can be performed on either a data matrix or a correlation matrix depending on the type of variables being measured (Naes et al. 2002; Bro and Smilde 2014). However, in a case where the original variables are nearly non-correlated, nothing can be gained by using a PCA analysis comparing with the use of classical statistics methods. Bro and Smilde (2014) provided with a compressive tutorial on the use of the PCA as well as discussing some other practical aspects on the implementation of the PCA (e.g. validation, pre-processing, the definition of the optimal number of components, data interpretation, and outlier detection) that are beyond the objective of this communication (Bro and Smilde 2014; Brereton 2009, 2015). Figure 1 illustrates a schematic representation of how the PCA works adapted from the literature (Bro and Smilde 2014; Brereton 2009, 2015).

How Does It Work?

As defined in the previous section, the PCA is a dimension-reduction tool that can be used to condense a large set of variables to a small set that still contains most of the information in the large set (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014). However, there is no guarantee that the dimensions are always interpretable. Therefore, other statistical methods should be evaluated depending of the type and structure of the data set.

The results of a PCA are usually interpreted in terms of component, sometimes called factors (the transformed variable values corresponding to a data point), scores (original samples) and loadings (the original variables) (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014).

Figure 2 presents with a simple case study of the use of the PCA. In this example, beer samples sourced from three different regions were analysed using the UV-VIS spectroscopy. The data set containing the samples with their corresponding UV-VIS spectra was analysed using the PCA. The first three principal components defined by the model are presented in a diagram (Fig. 2a).

In this example, the score plot involves the projection of the data (beer samples from three regions) into the principal components (PCs) (three components) (Fig. 2a) (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014). The score plot contains dot points that represent the original samples (the beer samples from the three regions analysed using the UV-VIS spectroscopy) projected into the selected PCs. Please note that an outlier sample was also added (see red dot).

The loadings (UV-VIS wavelengths) are used to identify which regions in the data set (variables = wavelengths) have the largest effect on each component that contributed to the separation between thee beer samples sourced from the different regions (Fig. 2b). It is well known that loadings can range from − 1 to 1 where loadings close to − 1 or 1 indicated that such variable strongly influences that principal component. On the other hand, loadings close to zero indicated that the variable has a weak influence on that principal component (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014).

In summary, the loadings are the weights derived from the original variables; therefore, if you have analysed the sample using GC-MS, the loadings are the retention times; while if infrared spectroscopy was used, the loadings will represent the

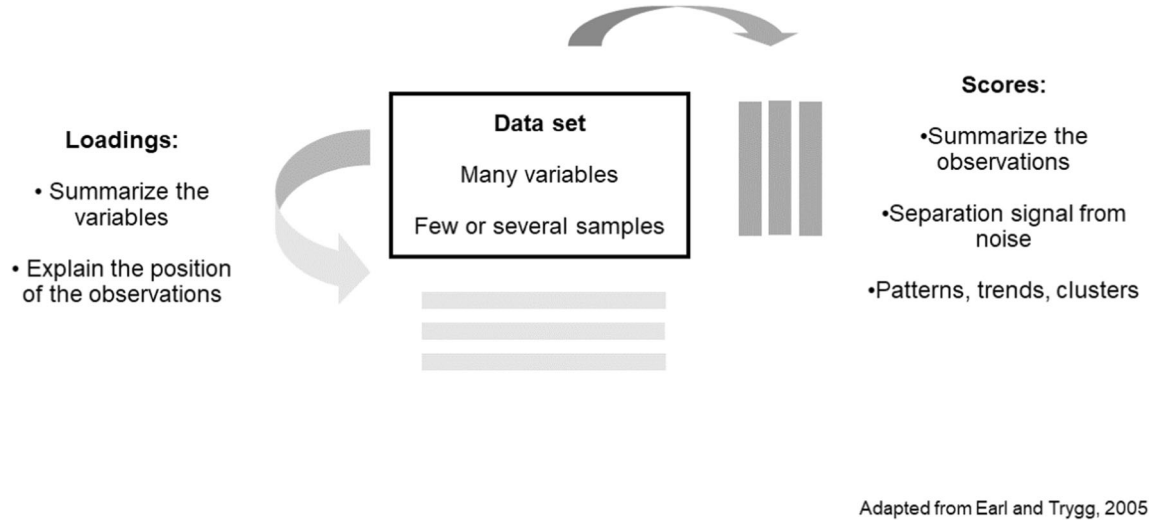

Fig. 1 Schematic representation of principal component analysis

wavelengths or frequencies, etc. Please note that if the variables are pre-processed (e.g. bias correction, standardisation, derivatives, smoothing), this will be reflected in the shape of the loadings.

One of the main objectives of the PCA is also to define the optimal weights. Where "optimal" means if the model can capture as much information in the original variables as possible, based on the correlations among those variables. If all the variables in a component are positively correlated with each other, all the loadings will be positive. Nevertheless, if there are some negative correlations among the variables, some of the loadings will be negative too (Brereton 2000, 2008; Cozzolino et al. 2009, 2011; Esbensen 2002; Martens and Martens 2001; Munck et al. 1998; Naes et al. 2002; Otto 1999; Skov et al. 2014; Bro and Smilde 2014).

Validation

Previous reports and reviews highlighted the importance of validation when multivariate data analysis methods and techniques are used (Badertscher and Pretsch 2006; Berrueta et al. 2007; Brereton 2006, 2009, 2015; Westad and Marini 2015; Doyle et al. 2016). In all the applications of multivariate data analysis (including the PCA), the validation of any given model requires that an independent set of samples must be

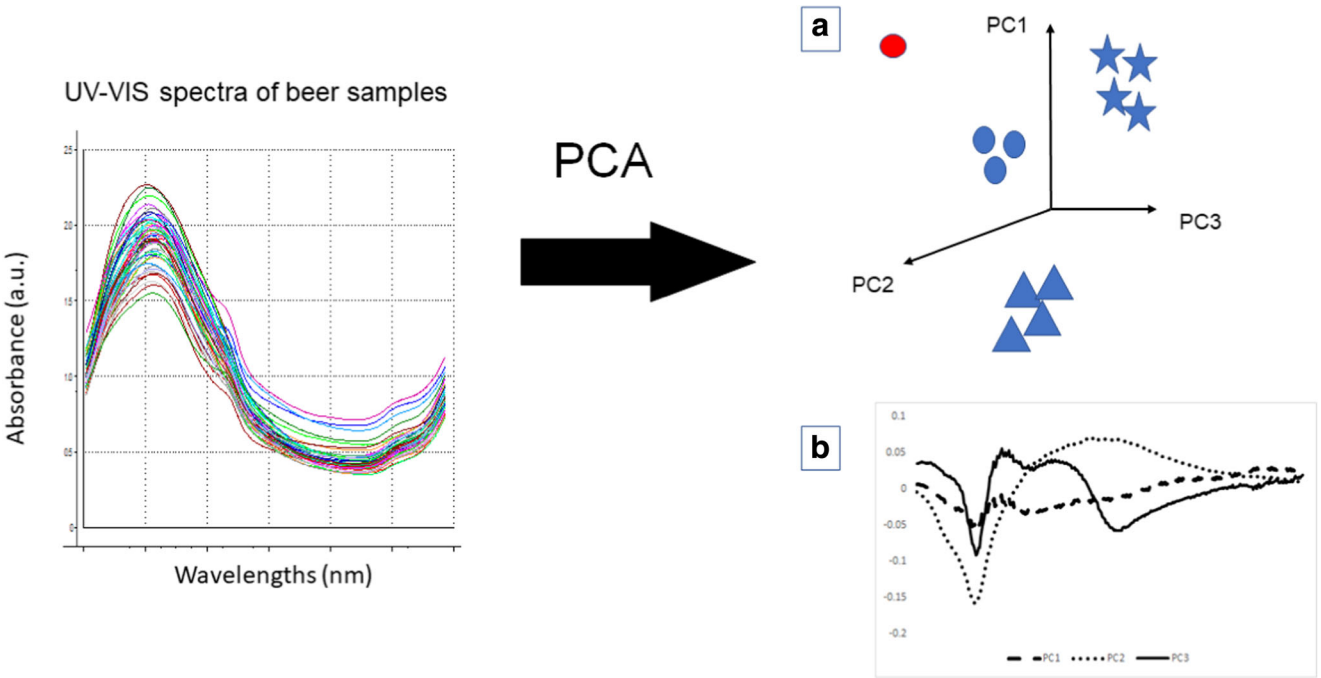

Fig. 2 Example of the application of PCA to a UV-VIS analysis of beer samples sourced from three regions. (a) Score plot. (b) Loadings

Table 1 Minimal information required to report a PCA results

	Minimal information for reporting
Scores (samples)	Score plot indicating the source such as sample identification (e.g. name, symbols)
Loadings	*x*-axis indicating the variable and units, *y*-axis label as loadings
PCA score plot	Axis label as principal component (one, two, etc.) and per cent of variance explained by each principal component in brackets

used to test the ability of the model to predict a set of unknown samples (Badertscher and Pretsch 2006; Berrueta et al. 2007; Brereton 2006, 2015; Westad and Marini 2015). Although, using an external data set cannot be achieved in some of the real-life situations due to different issues (e.g. cost of the analysis, number of samples available), cross-validation has been suggested as the most widely method to overcome some of these issues. Several authors have also improved this important step using or combining other pre-processing techniques such as K-fold and repeated K-fold cross validation (leave one out, random, etc.), jack-knife, among other techniques (Hawkins 2004; Kjeldhal and Bro 2010; Berrueta et al. 2007; Gonzalez 2007; Westad and Marini 2015).

Reporting PCA Results

Table 1 provides some of the minimal information required to report the results derived from the PCA. Please note that this is just a recommendation, and some of the information available will depend on the type of analysis and software used. However, as minimum in any scientific publication, the score plot (including the amount of variance explained by each PC) and the loadings must be reported and interpreted. In addition, any comment or discussion about outliers, number of principal components and pre-processing (e.g. instrumental methods) should be also added into the discussion and interpretation of the PCA.

Take Home Message

An increasing number of publications and reports in food sciences are targeting issues related to authenticity, contamination, fraud, origin and traceability of foods, as well as the increase use of instrumental methods (e.g. GC-MS, HPLC, electronic noses and tongues, sensors, sensory data) indicated that the PCA is the most widely used method in data mining and interpretation. In summary, the application of the PCA provides with two main elements—the scores and loadings. The scores provide with a location of the sample where the loadings indicated which variables are important to explain the trends in the grouping of samples.

Acknowledgements The authors thank the support of our colleagues and friends that encouraged writing this article.

Compliance with Ethical Standards

Conflict of Interest Dr. Daniel Cozzolino declares that he has no conflict of interest. Dr. Aoife Power declares that she has no conflict of interest. Dr. James Chapman declares that he has no conflict of interest.

Ethical Approval This article does not contain any studies with human or animal subjects.

Informed Consent (In case humans are involved) Informed consent was obtained from all individual participants included in the study. (If not applicable on the study) Not applicable.

References

Badertscher M, Pretsch E (2006) Bad results from good data. Trends Anal Chem 25:1131–1138

Berrueta LA, Alonso-Salces RM, Herberger K (2007) Supervised pattern recognition in food analysis. J Chromatogr A 1158:196–214

Bevilacqua M, Necatelli R, Bucci R, Magri AD, Magri SL, Marini F (2014) Chemometric classification techniques as tool for solving problems in analytical chemistry. J AOAC Int 97:19–27

Brereton RG (2000) Introduction to multivariate calibration in analytical chemistry. Analyst 125:2125–2154

Brereton RG (2006) Consequences of sample size, variable selection, and model validation and optimization, for predicting classification ability from analytical data. Trends in Analytical Chemistry 25, 1103–1111

Brereton RG (2008) Applied chemometrics for scientist. Wiley, Chichester

Brereton RG (2015) Pattern recognition in chemometrics. Chemom Intell Lab Syst 149(2015):90–96

Bro R, Smilde AK (2014) Principal component analysis: a tutorial review. Anal Methods 6:2812–2831

Cozzolino D, Cynkar WU, Dambergs RG, Shah N, Smith P (2009) Multivariate methods in grape and wine analysis. Int J Wine Res 1:123–130

Cozzolino D, Shah N, Cynkar W, Smith P (2011) A practical overview of multivariate data analysis applied to spectroscopy. Food Res Int 44: 1888–1896

Cozzolino D (2012) Recent trends on the use of infrared spectroscopy to trace and authenticate natural and agricultural food products. Applied Spectroscopy Reviews 47: 518–530

Doyle N, Roberts JJ, Swain D, Cozzolino D (2016) The use of qualitative analysis in food research and technology: considerations and reflections from an applied point of view. Food Anal Methods 10:964–969

Esbensen KH (2002) Multivariate data analysis in practice. CAMO Process AS, Oslo

Gonzalez GA (2007) Use and misuse of supervised pattern recognition methods for interpreting compositional data. J Chromatogr A 1158: 215–225

Hawkins DM (2004) The problem of overfitting. J Chem Inf Comput Sci 44:1–12

Kjeldhal K, Bro R (2010) Some common misunderstanding in chemometrics. J Chemom 24:558–564

Kumar N, Bansal A, Sarma GS, Rawal RK (2014) Chemometrics tools used in analytical chemistry: an overview. Talanta 123:186–199

Martens H, Martens M (2001) Multivariate analysis of quality. An introduction. Wiley, Chichester

Munck L, Norgaard L, Engelsen SB, Bro R, Andersson CA (1998) Chemometrics in food science: a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance. Chemom Intell Lab Syst 44:31–60

Mutihac L, Mutihac R (2008) Mining in chemometrics. Anal Chim Acta 612:1–18

Naes T, Isaksson T, Fearn T, Davies T (2002) A user-friendly guide to multivariate calibration and classification. NIR Publications, Chichester 420 p

Otto M (1999) Chemometrics: statistics and computer application in analytical chemistry. Wiley-VCH 314 p

Skov T, Honore AH, Jensen HM, Naes T, Engelsen SB (2014) Chemometrics in foodomics: handling data structures from multiple analytical platforms. Trends Anal Chem 60:71–79

Westad F, Marini F (2015) Validation of chemometric models: a tutorial. Anal Chim Acta 893:14–23