Analytical Methods

TUTORIAL REVIEW

View Article Online
View Journal | View Issue

Principal component analysis

Rasmus Bro[a] and Age K. Smilde[ab]

Cite this: *Anal. Methods*, 2014, **6**, 2812

Received 28th October 2013
Accepted 17th February 2014

DOI: 10.1039/c3ay41907j

www.rsc.org/methods

Principal component analysis is one of the most important and powerful methods in chemometrics as well as in a wealth of other areas. This paper provides a description of how to understand, use, and interpret principal component analysis. The paper focuses on the use of principal component analysis in typical chemometric areas but the results are generally applicable.

Introductory example

To set the stage for this paper, we will start with a small example where principal component analysis (PCA) can be useful. Red wines, 44 samples, produced from the same grape (*Cabernet sauvignon*) were collected. Six of these were from Argentina, fifteen from Chile, twelve from Australia and eleven from South Africa. A Foss WineScan instrument was used to measure 14 characteristic parameters of the wines such as the ethanol content, pH, *etc.* (Table 1).

Hence, a dataset is obtained which consists of 44 samples and 14 variables. The actual measurements can be arranged in a table or a matrix of size 44×14. A portion of this table is shown in Fig. 1.

With 44 samples and 14 columns, it is quite complicated to get an overview of what kind of information is available in the data. A good starting point is to plot individual variables or samples. Three of the variables are shown in Fig. 2. It can be seen that total acid as well as methanol tends to be higher in

[a]*Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958, Frederiksberg C, Denmark*

[b]*Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

Rasmus Bro studied mathematics and analytical chemistry at the Technical University of Denmark and received his M.Sc. degree in 1994. In 1998 he obtained his Ph.D. in multi-way analysis from the University of Amsterdam, The Netherlands with Age K. Smilde as one of two supervisors. He is currently Professor of chemometrics at the University of Copenhagen, Denmark. His work focusses on developing tools for analyzing data in process analytical technology, metabonomics and analytical chemistry. Rasmus Bro has published 150 papers as well as several books and has received a number of prizes and awards over the years. He heads the ten year old industrial consortium ODIN which is a networking facility that aims to make chemometrics as used and useful as possible in Danish industries.

Age K. Smilde is a full professor of Biosystems Data Analysis at the Swammerdam Institute for Life Sciences at the University of Amsterdam and he also works part-time at the Academic Medical Centre of the same university. As of July 1, 2013 he holds a part-time position as professor of Computational Systems Biology at the University of Copenhagen. His research interest focuses on two topics: data fusion and network inference. Data fusion concerns integrating functional genomics data and fusing data with prior biological knowledge. The network topic encompasses network reconstruction, deriving network properties from time-resolved data and computational network analysis. He has published more than 200 peer-reviewed papers and has been the Editor-Europe of the Journal of Chemometrics during the period 1994–2002. He chaired the 1996 Gordon Research Conference on Statistics in Chemistry & Chemical Engineering. In 2006, he received the Eastern Analytical Symposium Award for Achievements in Chemometrics.

Table 1 Chemical parameters determined on the wine samples (data from http://www.models.life.ku.dk/Wine_GCMS_FTIR [February, 2014][1,2])

Ethanol (vol%)
Total acid (g L^{-1})
Volatile acid (g L^{-1})
Malic acid (g L^{-1})
pH
Lactic acid (g L^{-1})
Rest sugar (Glu + Fru) (g L^{-1})
Citric acid (mg L^{-1})
CO_2 (g L^{-1})
Density (g mL^{-1})
Total polyphenol index
Glycerol (g L^{-1})
Methanol (vol%)
Tartaric acid (g L^{-1})

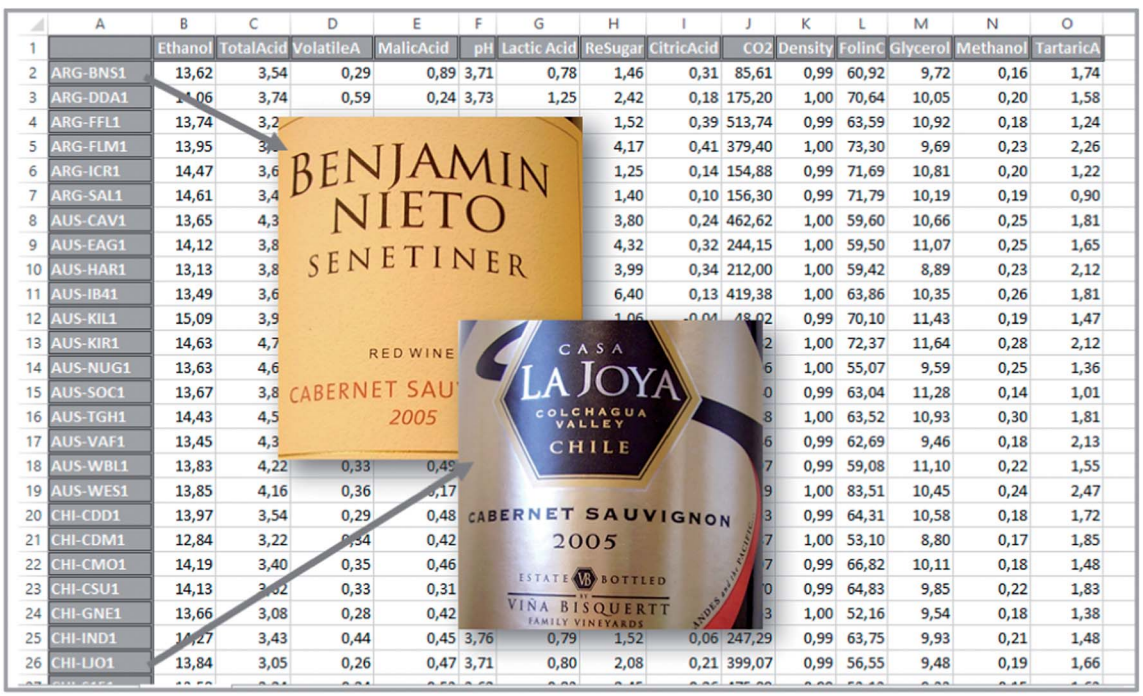
samples from Australia and South Africa whereas there are less pronounced regional differences in the ethanol content.

Even though Fig. 2 may suggest that there is little relevant regional information in ethanol, it is dangerous to rely too much on univariate analysis. In univariate analysis, any co-variation with other variables is explicitly neglected and this may lead to important features being ignored. For example, plotting ethanol *versus* glycerol (see Fig. 3) shows an interesting correlation between the two. This is difficult to deduce from plots of the individual variables. If glycerol and ethanol were completely correlated, it would, in fact, be possible to simply use *e.g.* the average or the sum of the two as one new variable that could replace the two original ones. No information would be lost as it would always be possible to go from *e.g.* the average to the two original variables.

This concept of using suitable linear combinations of the original variables will turn out to be essential in PCA and is explained in a bit more detail and a slightly unusual way here. The new variable, say, the average of the two original ones, can be defined as a weighted average of all 14 variables; only the other variables will have weight zero. These 14 weights are shown in Fig. 4. Rather than having the weights of ethanol and glycerol to be 0.5 as they would in an ordinary average, they are chosen as 0.7 to make the whole 14-vector of weights scaled to be a unit vector. When the original variables ethanol and glycerol are taken to be of length one (unit length) then it is convenient to also have the linear combination of those to be of length one. This then defines the unit on the combined variable. To achieve this it is necessary to take 0.7 ($\sqrt{2}/2$ to be exact) of ethanol and 0.7 of glycerol, as simple Pythagorean geometry shows in Fig. 5. This also carries over to more than two variables.

Using a unit weight vector has certain advantages. The most important one is that the unit vector preserves the size of the variation. Imagine there are ten variables rather than two that are being averaged. Assume, for simplicity that all ten have the value five.

Regardless of whether the average is calculated from two or ten variables, the average remains five. Using the unit vector, though, will provide a measure of the number of variables showing variation. In fact, the variance of the original variables and this newly calculated one will be the same, if the original variables are all correlated. Thus, using the unit vector preserves the variation in the data and this is an attractive property. One



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Ethanol	TotalAcid	VolatileA	MalicAcid	pH	Lactic Acid	ReSugar	CitricAcid	CO2	Density	FolinC	Glycerol	Methanol	TartaricA
2	ARG-BNS1	13,62	3,54	0,29	0,89	3,71	0,78	1,46	0,31	85,61	0,99	60,92	9,72	0,16	1,74
3	ARG-DDA1	14,06	3,74	0,59	0,24	3,73	1,25	2,42	0,18	175,20	1,00	70,64	10,05	0,20	1,58
4	ARG-FFL1	13,74	3,2					1,52	0,39	513,74	0,99	63,59	10,92	0,18	1,24
5	ARG-FLM1	13,95	3,					4,17	0,41	379,40	1,00	73,30	9,69	0,23	2,26
6	ARG-ICR1	14,47	3,6					1,25	0,14	154,88	0,99	71,69	10,81	0,20	1,22
7	ARG-SAL1	14,61	3,4					1,40	0,10	156,30	0,99	71,79	10,19	0,19	0,90
8	AUS-CAV1	13,65	4,3					3,80	0,24	462,62	1,00	59,60	10,66	0,25	1,81
9	AUS-EAG1	14,12	3,8					4,32	0,32	244,15	1,00	59,50	11,07	0,25	1,65
10	AUS-HAR1	13,13	3,8					3,99	0,34	212,00	1,00	59,42	8,89	0,23	2,12
11	AUS-IB41	13,49	3,6					6,40	0,13	419,38	1,00	63,86	10,35	0,26	1,81
12	AUS-KIL1	15,09	3,9					1,06	-0,01	48,02	0,99	70,10	11,43	0,19	1,47
13	AUS-KIR1	14,63	4,7							2	1,00	72,37	11,64	0,28	2,12
14	AUS-NUG1	13,63	4,6							6	1,00	55,07	9,59	0,25	1,36
15	AUS-SOC1	13,67	3,8							0	0,99	63,04	11,28	0,14	1,01
16	AUS-TGH1	14,43	4,5							8	1,00	63,52	10,93	0,30	1,81
17	AUS-VAF1	13,45	4,3							6	0,99	62,69	9,46	0,18	2,13
18	AUS-WBL1	13,83	4,22	0,33	0,49					7	0,99	59,08	11,10	0,22	1,55
19	AUS-WES1	13,85	4,16	0,36	0,17					9	1,00	83,51	10,45	0,24	2,47
20	CHI-CDD1	13,97	3,54	0,29	0,48					3	0,99	64,31	10,58	0,18	1,72
21	CHI-CDM1	12,84	3,22	0,34	0,42					7	1,00	53,10	8,80	0,17	1,85
22	CHI-CMO1	14,19	3,40	0,35	0,46					7	0,99	66,82	10,11	0,18	1,48
23	CHI-CSU1	14,13	3,62	0,33	0,31					0	0,99	64,83	9,85	0,22	1,83
24	CHI-GNE1	13,66	3,08	0,28	0,42					3	1,00	52,16	9,54	0,18	1,38
25	CHI-IND1	14,27	3,43	0,44	0,45	3,76	0,79	1,52	0,06	247,29	0,99	63,75	9,93	0,21	1,48
26	CHI-LJO1	13,84	3,05	0,26	0,47	3,71	0,80	2,08	0,21	399,07	0,99	56,55	9,48	0,19	1,66

Fig. 1 A subset of the wine dataset.

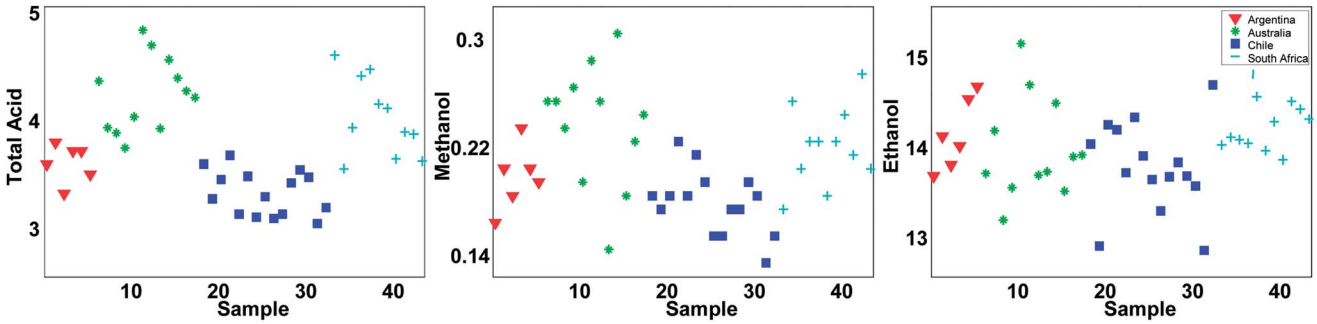


Fig. 2 Three variables coloured according to the region.

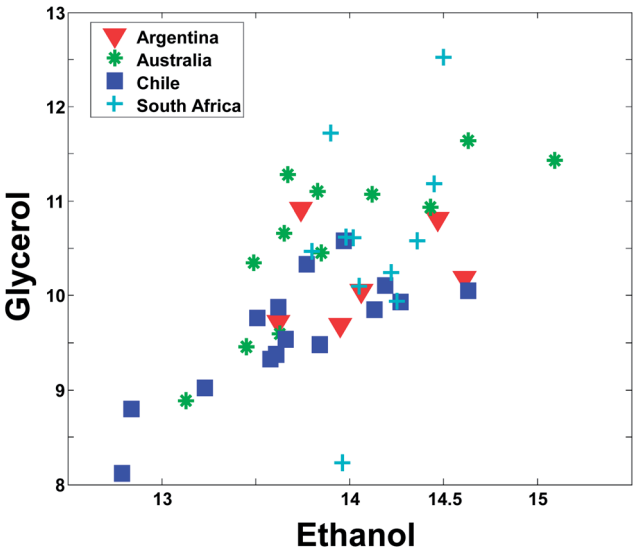


Fig. 3 A plot of ethanol *versus* glycerol.

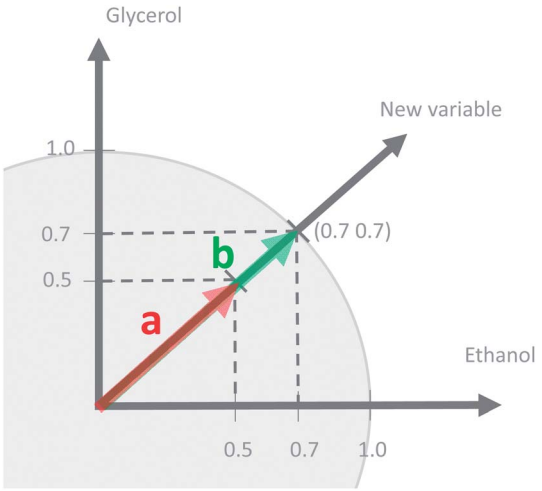


Fig. 5 The concept of a unit vector.

of the reasons is that it allows for going back and forth between the space of the original variables (say glycerol–ethanol) and the new variable. With this definition of weights, it is now possible to calculate the new variable, the 'average', for any sample, as indicated in Fig. 6.

As mentioned above, it is possible to go back and forth between the original two variables and the new variable. Multiplying the new variable with the weights provides an estimation of the original variables (Fig. 7).

This is a powerful property; that it is possible to use weights to condense several variables into one and *vice versa*. To generalize this, notice that the current concept only works perfectly when the two variables are completely correlated. Think of an average grade in a school system. Many particular grades can lead to the same average grade, so it is not in general possible to go back and forth. To make an intelligent new variable, it is natural to ask for a new variable that will actually provide a nice model of the data. That is, a new variable which, when multiplied with the weights, will describe as much as

Fig. 4 Defining the weights for a variable that includes only ethanol and glycerol information.

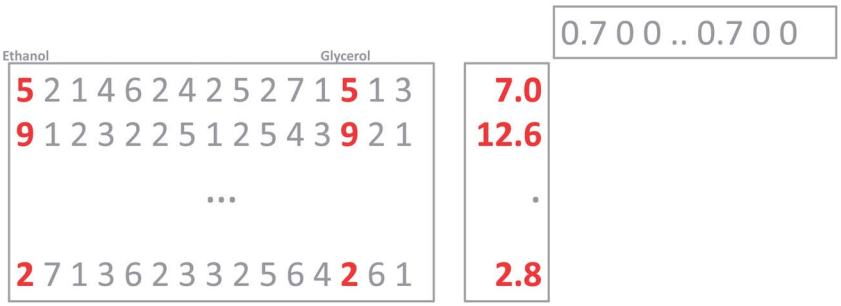


Fig. 6 Using defined weights to calculate a new variable that is a scaled average of ethanol and glycerol (arbitrary numbers used here). The average is calculated as the inner product of the 14 measurements of a sample and the weight vector. Some didactical rounding has been used in the example.

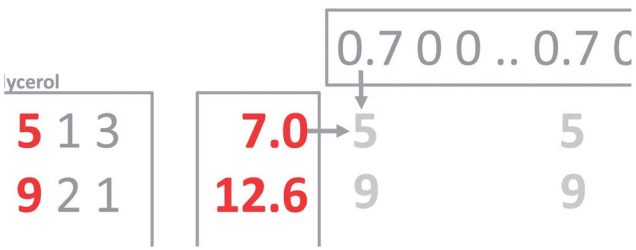


Fig. 7 Using the new variable and the weights to estimate the old original variables.

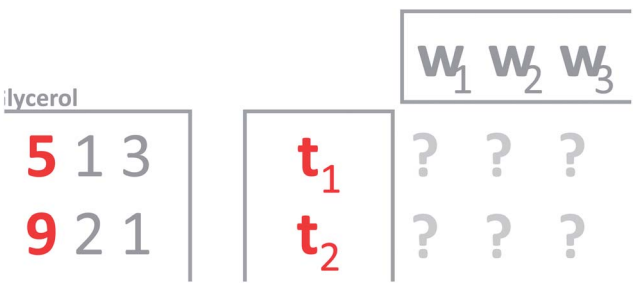


Fig. 8 Defining weights (w's) that will give a new variable which leads to a good model of the data.

possible the whole matrix (Fig. 8). Such a variable will be an optimal representative of the whole data in the sense that no other weighted average simultaneously describes as much of the information in the matrix.

It turns out that PCA provides a solution to this problem. Principal component analysis provides the weights needed to get the new variable that best explains the variation in the whole dataset in a certain sense. This new variable including the defining weights, is called the first principal component.

To find the first principal component of the actual wine data, it is necessary to jump ahead a little bit and preprocess the data first. Looking at the data (Fig. 1) it is seen, that some variables such as CO_2 are measured in numbers that are much larger than *e.g.* methanol. For example, for sample three, CO_2 is 513.74 [g L^{-1}] whereas methanol is 0.18 [vol%]. If this difference in scale and possibly offset is not handled, then the PCA model will only focus on variables measured in large numbers. It is desired to model all variables, and there is a preprocessing tool called autoscaling which will make each column have the same 'size' so that all variables have an equal opportunity of being modelled. Autoscaling means that from each variable, the mean value is subtracted and then the variable is divided by its standard deviation. Autoscaling will be described in more detail, but for now, it is just important to note that each variable is transformed to equal size and in the process, each variable will have negative as well as positive values because the mean of it has been subtracted. Note that an average sample now corresponds to all zeroes. Hence, zero is no longer absence of a 'signal' but instead indicates an average 'signal'.

With this pre-processing of the data, PCA can be performed. The technical details of how to do that will follow, but the first principal component is shown in Fig. 9. In the lower plot, the weights are shown. Instead of the quite sparse weights in Fig. 4, these weights are non-zero for all variables. This first component does not explain all the variation, but it does explain 25% of what is happening in the data. As there are 14 variables, it would be expected that if every variable showed variation independent of the other, then each original variable would explain 100%/14 = 7% of the variation. Hence, this first component is wrapping up information, which can be said to correspond to approximately 3–4 variables.

Just like the average of ethanol and glycerol or the average school grade, the new variable can be interpreted as "just a variable". The weights define how the variable is determined and how many scores each sample has of this linear combination. For example, it is seen that most of the South African samples have positive scores and hence, will have fairly high values on variables that have positive weights such as for example methanol. This is confirmed in Fig. 2.

Principal component analysis

Taking linear combinations

It is time to introduce some more formal notation and nomenclature. The weighted average as mentioned above is more formally called a linear combination: it is a way of combining the original variables in a linear way. It is also

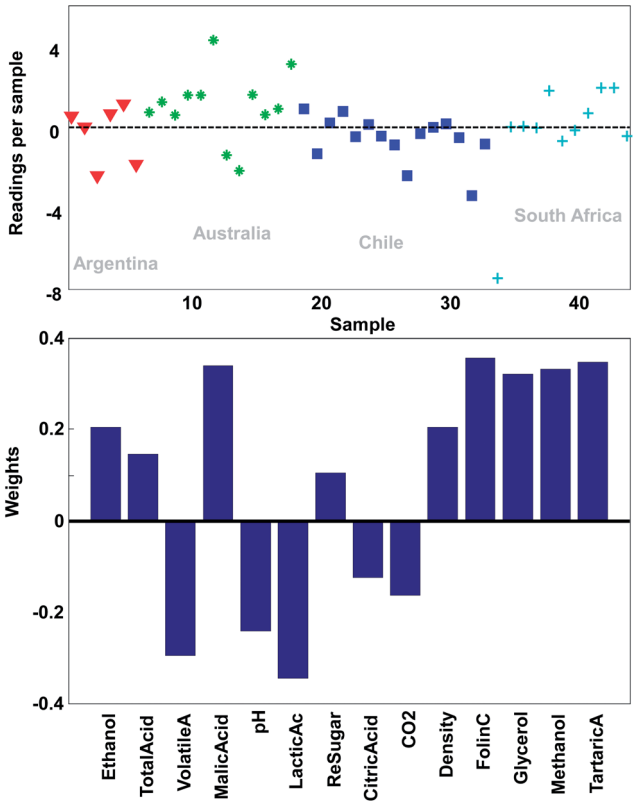


Fig. 9 The first principal component of the wine data. The lower plot shows the weights and the upper plot shows the weighted averages obtained with those weights.

sometimes called a latent variable where, in contrast, the original variables are manifest.

The data are collected in a matrix $\mathbf{X}$ with I rows ($i = 1, \ldots, I$; usually samples/objects) and J columns ($j = 1, \ldots, J$; usually variables), hence of size $I \times J$. The individual variables (columns) of $\mathbf{X}$ are denoted by $\mathbf{x}_j$ ($j = 1, \ldots, J$) and are all vectors in the I-dimensional space. A linear combination of those $\mathbf{x}$ variables can be written as $\mathbf{t} = w_1 \times \mathbf{x}_1 + \ldots + w_J \times \mathbf{x}_J$, where $\mathbf{t}$ is now a new vector in the same space as the $\mathbf{x}$ variables (because it is a linear combination of these). In matrix notation, this becomes $\mathbf{t} = \mathbf{Xw}$, with $\mathbf{w}$ being the vector with elements w_j ($j = 1, \ldots, J$). Since the matrix $\mathbf{X}$ contains variation relevant to the problem, it seems reasonable to have as much as possible of that variation also in $\mathbf{t}$. If this amount of variation in $\mathbf{t}$ is appreciable, then it can serve as a good summary of the $\mathbf{x}$ variables. Hence, the fourteen variables of $\mathbf{X}$ can then be replaced by only one variable $\mathbf{t}$ retaining most of the relevant information.

The variation in $\mathbf{t}$ can be measured by its variance, var($\mathbf{t}$), defined in the usual way in statistics. Then the problem translates to maximizing this variance choosing optimal weights $w_1, \ldots, w_J$. There is one caveat, however, since multiplying an optimal $\mathbf{w}$ with an arbitrary large number will make the variance of $\mathbf{t}$ also arbitrary large. Hence, to have a proper problem, the weights have to be normalized. This is done by requiring that their norm, *i.e.* the sum-of-squared values, is one (see Fig. 5). Throughout we will use the symbol $||.||^2$ to indicate the squared Frobenius norm (sum-of-squares). Thus, the formal problem becomes

$$\underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \operatorname{var}(\mathbf{t}) \tag{1}$$

which should be read as the problem of finding the $\mathbf{w}$ of length one that maximizes the variance of $\mathbf{t}$ (note that $||\mathbf{w}|| = 1$ is the same as requiring $||\mathbf{w}||^2 = 1$). The function argmax is the mathematical notation for returning the argument $\mathbf{w}$ of the maximization function. This can be made more explicit by using the fact that $\mathbf{t} = \mathbf{Xw}$:

$$\underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \left(\mathbf{t}^{\mathrm{T}}\mathbf{t}\right) = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}} \left(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}\right) \tag{2}$$

where it is assumed that the matrix $\mathbf{X}$ is mean-centered (then all linear combinations are also mean-centered). The latter problem is a standard problem in linear algebra and the optimal $\mathbf{w}$ is the (standardized) first eigenvector (*i.e.* the eigenvector with the largest value) of the covariance matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}/(n-1)$ or the corresponding cross-product matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

Explained variation

The variance of $\mathbf{t}$ can now be calculated but a more meaningful assessment of the summarizing capability of $\mathbf{t}$ is obtained by calculating how representative $\mathbf{t}$ is in terms of replacing $\mathbf{X}$. This can be done by projecting the columns of $\mathbf{X}$ on $\mathbf{t}$ and calculating the residuals of that projection. This is performed by regressing all variables of $\mathbf{X}$ on $\mathbf{t}$ using the ordinary regression equation

$$\mathbf{X} = \mathbf{t}\mathbf{p}^{\mathrm{T}} + \mathbf{E} \tag{3}$$

where $\mathbf{p}$ is the vector of regression coefficients and $\mathbf{E}$ is the matrix of residuals. Interestingly, $\mathbf{p}$ equals $\mathbf{w}$ and the whole machinery of regression can be used to judge the quality of the summarizer $\mathbf{t}$. Traditionally, this is done by calculating

$$\frac{\|\mathbf{X}\|^2 - \|\mathbf{E}\|^2}{\|\mathbf{X}\|^2} 100\% \tag{4}$$

which is referred to as the percentage of explained variation of $\mathbf{t}$.

In Fig. 10, it is illustrated how the explained variation is calculated as also explained around eqn (4).

Note, that the measures above are called variations rather than variances. In order to talk about variances, it is necessary to correct for the degrees of freedom consumed by the model and this is not a simple task. Due to the non-linear nature of the PCA model, degrees of freedom are not as simple to define as for linear models such as in linear regression or analysis of variance. Hence, throughout this paper, the magnitude of variation will simply be expressed in terms of sums of squares. For more information on this, refer to the literature.[3,4]

PCA as a model

Eqn (3) highlights an important interpretation of PCA: it can be seen as a modelling activity. By rewriting eqn (3) as

$$\mathbf{X} = \mathbf{t}\mathbf{p}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E}, \tag{5}$$

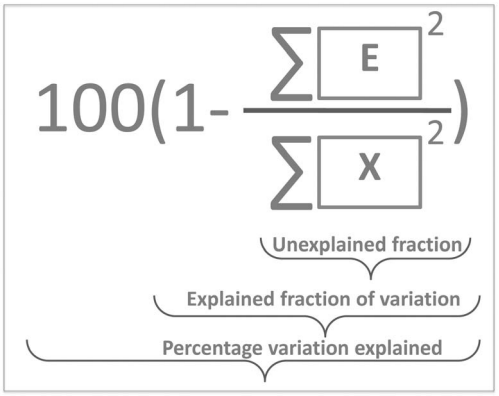


Fig. 10 Exemplifying how explained variation is calculated using the data and the residuals.

shows that the (outer-) product $\mathbf{tp}^T$ serves as a model of $\mathbf{X}$ (indicated with a hat). In this equation, vector $\mathbf{t}$ was a fixed regressor and vector $\mathbf{p}$ the regression coefficient to be found. It can be shown that actually both $\mathbf{t}$ and $\mathbf{p}$ can be established from such an equation[5] by solving

$$\underset{\mathbf{t},\mathbf{p}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{tp}^T\|^2 \quad (6)$$

which is also a standard problem in linear algebra and has the same solution as eqn (2). Note that the solution does not change if $\mathbf{t}$ is premultiplied by $\alpha \neq 0$ and simultaneously $\mathbf{p}$ is divided by that same value. This property is called the scaling ambiguity[6] and it can be solved in different ways. In chemometrics, the vector $\mathbf{p}$ is normalized to length one ($||\mathbf{p}|| = 1$) and in psychometrics, $\mathbf{t}$ is normalized to length one. The vector $\mathbf{t}$ is usually referred to as the score vector (or scores in shorthand) and the vector $\mathbf{p}$ is called the loading vector (or loadings in shorthand). The term 'principal component' is not clearly defined and can mean either the score vector or the loading vector or the combination. Since the score and loading vectors are closely tied together it seems logical to reserve the term principal component for the pair $\mathbf{t}$ and $\mathbf{p}$.

Taking more components

If the percentage of explained variation of eqn (4) is too small, then the $\mathbf{t}$, $\mathbf{p}$ combination is not a sufficiently good summarizer of the data. Eqn (5) suggests an extension by writing

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1^T + \ldots + \mathbf{t}_R\mathbf{p}_R^T = \hat{\mathbf{X}} + \mathbf{E} \quad (7)$$

where $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_R]$ ($I \times R$) and $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_R]$ ($J \times R$) are now matrices containing, respectively, R score vectors and R loading vectors. If R is (much) smaller than J, then $\mathbf{T}$ and $\mathbf{P}$ still amount to a considerably more parsimonious description of the variation in $\mathbf{X}$. To identify the solution, $\mathbf{P}$ can be taken such that $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ and $\mathbf{T}$ can be taken such that $\mathbf{T}^T\mathbf{T}$ is a diagonal matrix. This corresponds to the normalisation of the loadings mentioned above. Each loading vector, thus has norm one and is orthogonal to other loading vectors (an orthogonal basis). The constraint on $\mathbf{T}$ implies that the score vectors are orthogonal to each other. This is the usual way to perform PCA in chemometrics. Due to the orthogonality in $\mathbf{P}$, the R components have independent contributions to the overall explained variation

$$\|\mathbf{X}\|^2 = \|\mathbf{t}_1\mathbf{p}_1^T\|^2 + \ldots\|\mathbf{t}_R\mathbf{p}_R^T\|^2 + \|\mathbf{E}\|^2 \quad (8)$$

and the term 'explained variation per component' can be used, similarly as in eqn (4).

History of PCA

PCA has been (re-)invented several times. The earliest presentation was in terms of eqn (6).[7] This interpretation stresses the modelling properties of PCA and is very much rooted in regression-thinking: variation explained by the principal components (Pearson's view). Later, in the thirties, the idea of taking linear combinations of variables was introduced[8] and the variation of the principal components was stressed (eqn (1); Hotelling's view). This is a more multivariate statistical approach. Later, it was realized that the two approaches were very similar.

Similar, but not the same. There is a fundamental conceptual difference between the two approaches, which is important to understand. In the Hotelling approach, the principal components are taken seriously in their specific direction. The first component explains the most variation, the second component the second most, *etc.* This is called the principal axis property: the principal components define new axes which should be taken seriously and have a meaning. PCA finds these principal axes. In contrast, in the Pearson approach it is the subspace, which is important, not the axes as such. The axes merely serve as a basis for this subspace. In the Hotelling approach, rotating the principal components destroys the interpretation of these components whereas in the Pearson conceptual model rotations merely generate a different basis for the (optimal) subspace.[9]

Visualization and interpretation

It is now time to discuss how a PCA model can be visualized. There are four parts of a PCA model: the data, the scores, the loadings and the residuals. Visualization of the actual data is often dependent on the type of data and the traditions of a given field. For continuous data such as time-series and spectra, it is

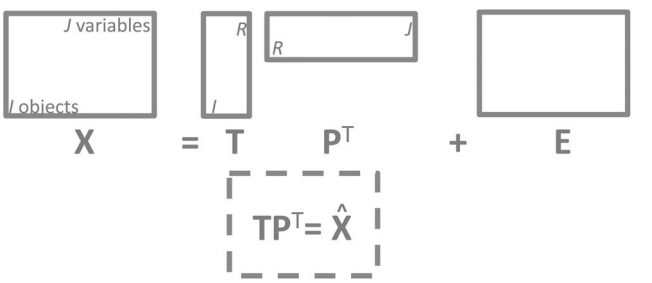


Fig. 11 The structure of a PCA model. Note that residuals ($\mathbf{E}$) have the same structure as the data and so does the model approximation of the data ($\mathbf{TP}^T$).

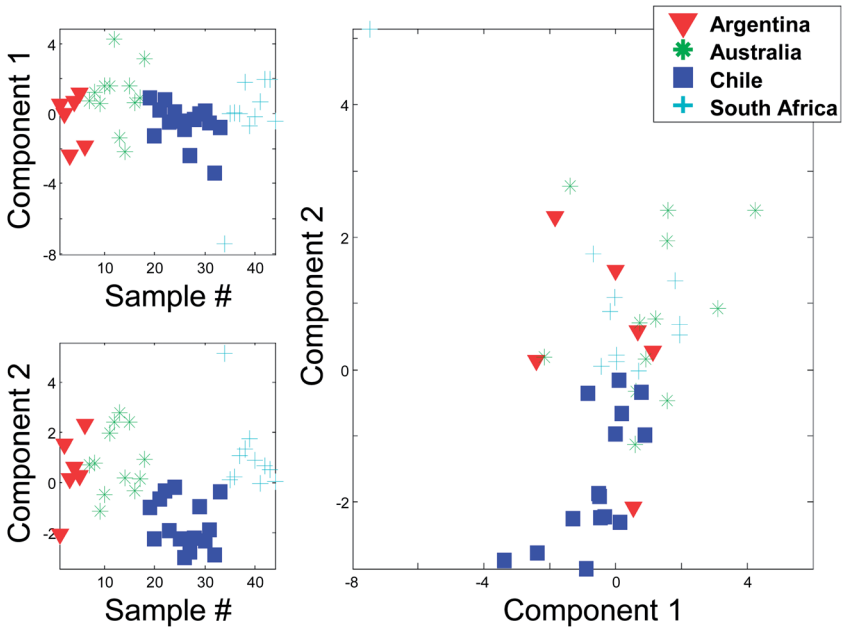


Fig. 12 Score 1 and 2 from a PCA on the autoscaled wine data. Upper left is a line plot of the 44 score values in component 1 and lower left the 44 score values of component 2. In the right plot, the two scores are plotted against each other.

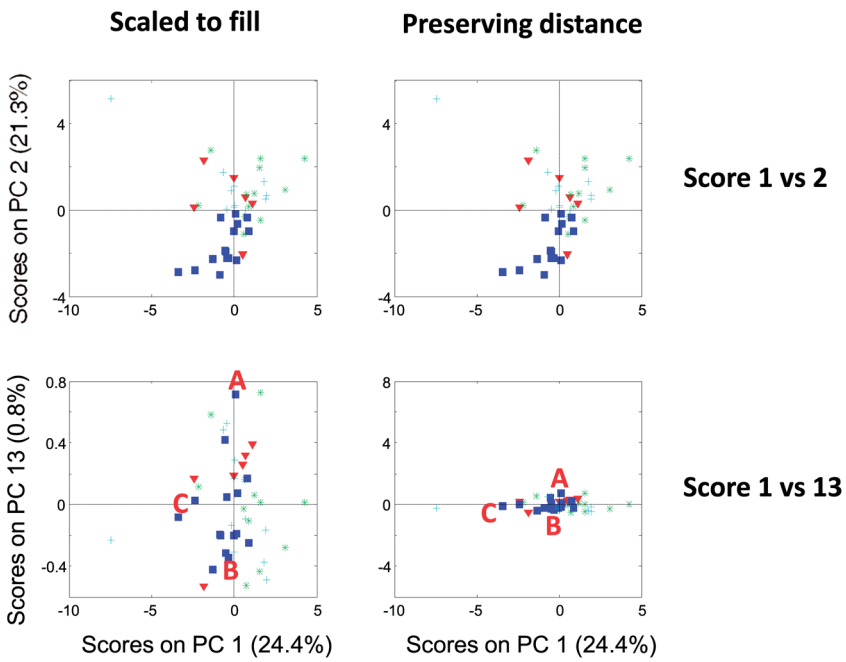


Fig. 13 Score plot of component 1 *versus* 2 (top) and 1 *versus* 13 (bottom). To the left, the plots are shown filling out the squares and to the right, they are shown preserving distances.

often feasible to plot the data as curves whereas more discrete data are often plotted in other ways such as bar plots.

Visualizing and interpreting residuals. Whatever visualization applies to the data would often also be useful for *e.g.* the residuals (Fig. 11). The residuals have the same structure and for example for spectral data, the residuals would literally correspond to the residual spectra and therefore provide important chemical information as to what spectral variation has not been explained (see also Fig. 23). In short, any visualization that is useful for the data will also be useful for the residuals.

Residuals can also be plotted as histograms or *e.g.* normal probability plots in order to see if the residuals are normally distributed. Alternatively, the residuals can be used for

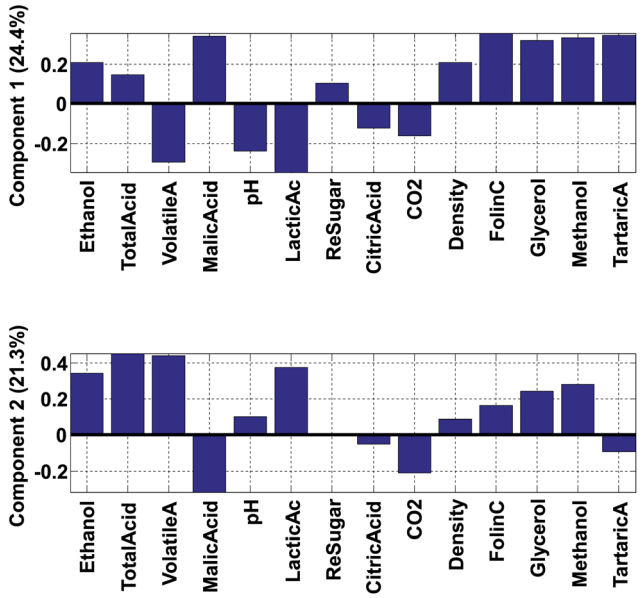


Fig. 14 Loading one (top) and loading two (bottom).

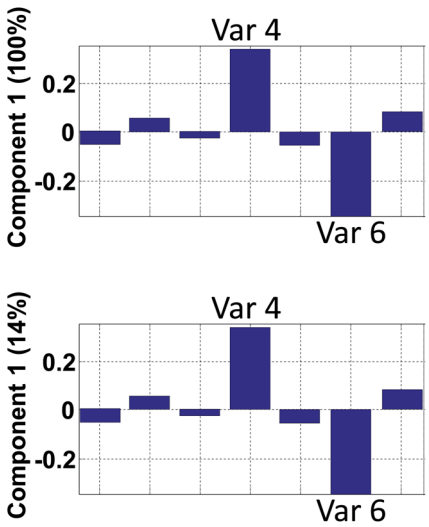


Fig. 15 Hypothetical loading vector from a model that explains 100% in component 1 (top) and 14% in component 1 (bottom).

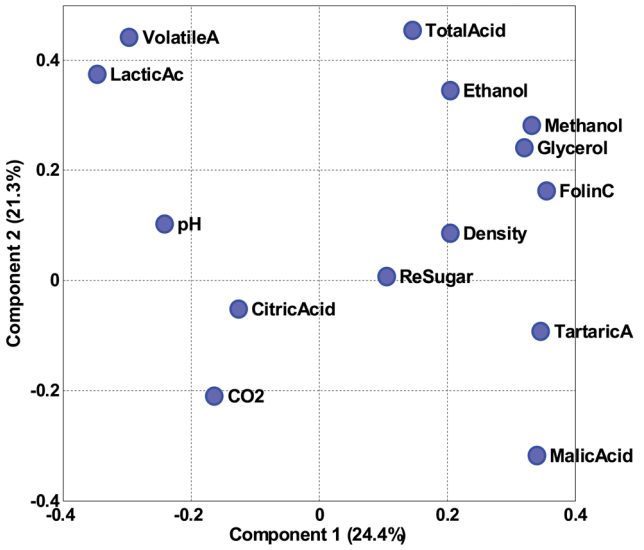


Fig. 16 Scatter plot of loading 1 *versus* loading 2.

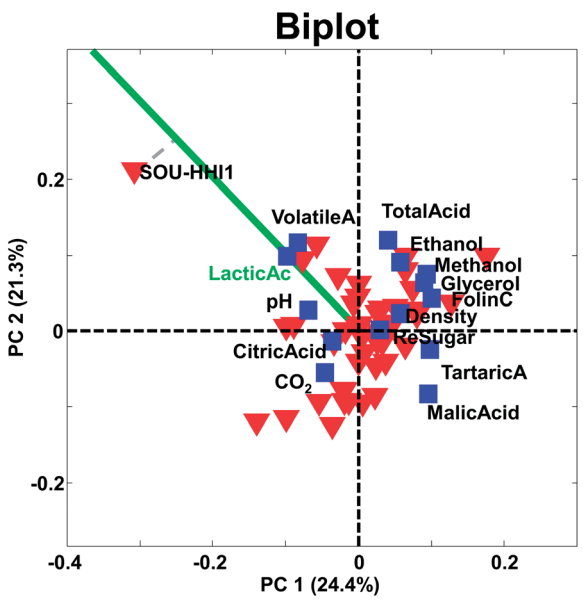


Fig. 17 A biplot of the first two components of a PCA model of the wine data.

calculating the explained or unexplained variation as explained earlier.

Visualizing and interpreting scores. It is well known that the readings of a variable can be plotted. Imagine that pH is measured on 20 samples. These 20 values can be plotted in a multitude of ways. Scores are readings in exactly the same way as any other variable and can hence be plotted and interpreted in many different ways. In Fig. 12, some visualizations of the first two components of the PCA model of the wine data are shown. If desired, they can be plotted as line plots as shown in the left in the figure. This plot of, for example, score 1, shows that the dark blue scores tend to have negative scores. This means that wines from Chile have relatively less of what this first component represents, which will be described by the loadings (see below).

Instead of plotting the scores in line plots, it is also possible to plot them in scatter plots. In Fig. 12 (right), such a scatter plot is shown and from the scatter plot it is more readily seen that there seem to be certain groupings in the data. For example, the Australian and Chilean wines seem to be almost distinctly different in this score plot. This suggests that it is possible to classify a wine using these measured variables. If a new sample ends up in the middle of the Chilean samples, it is probably not an Australian wine and *vice versa*. This possibility of using PCA for classification forms the basis for the classification method called SIMCA (Soft Independent Modelling of Class

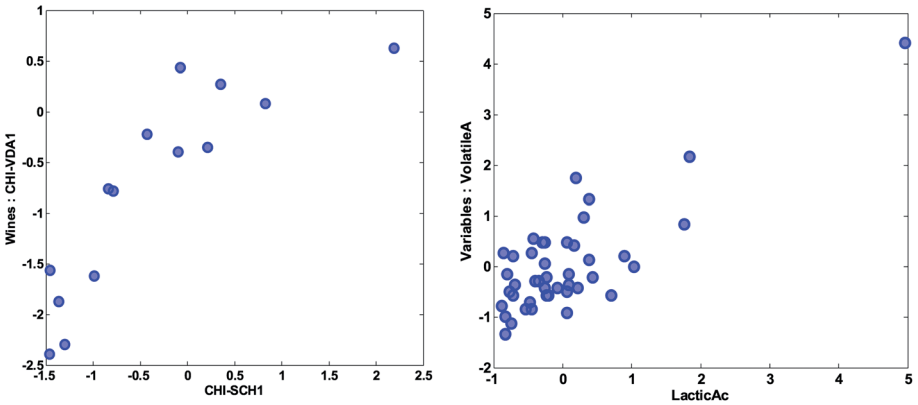


Fig. 18 Scatter plot of the preprocessed values of the variables of two wines (left) and two variables (right).

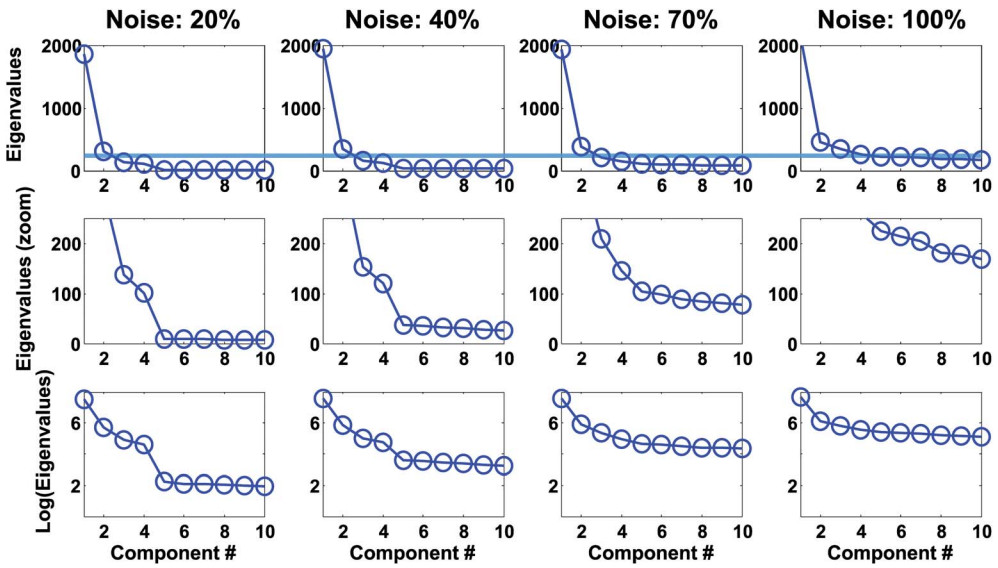


Fig. 19 Scree plots for simulated rank four data with various levels of noise. Top plots show eigenvalues. Middle plots show the same but zoomed in on the *y*-axis to the line indicated in the top plot. Lower plots show the logarithm of the eigenvalues.

Analogies).[10,11] The scatter plot can be interpreted in the same way that scatter plots are normally interpreted. For example, a plot of glycerol *versus* ethanol (Fig. 3) is simple to interpret. Samples that are close have similar glycerol and ethanol. Likewise, for a scatter plot of component 1 and 2. Samples that are close are similar in terms of what the components represent which is defined by the loading vectors. Also, if (and only if) the two components represent all or almost all of the variation in the data, then *e.g.* two closely lying samples are similar with respect to the actual data.

It is possible to assess similarities and differences among samples in terms of the raw data. If two components explain all or most of the variation in the data, then a score scatter plot will reflect distances in terms of the data directly if the scores are shown on the same scale. That is, the plot must be shown as original scores where the basis is the loading vector. As the loading vectors are unit vectors, they reflect the original data and if the two axes in the plot use the same scale, then distances can be read from the plots directly. If on the other hand the plots are not shown using the same scale on both axis, then assessing distances is not possible.

Compare the two versions of the two score plots in Fig. 13. The lower left plot has widely different scales on the two axes (because one component has much larger values numerically than the other). Henceforth, it is similar to plotting *e.g.* kilometres on one axis and meters on another. A map with such axes does not preserve distance. Consider, for example, the wine sample marked A. It seems to be closer to sample C than B in the lower left plot. The plot to the lower right preserves distances and here it is readily verified that sample A is, in fact, the closest to B in the space spanned by the two components.

There are several points worth mentioning in relation to this. Score plots are only indicative of the specific fraction of variance they explain. For example, scores that explain three percent do not imply much with respect to the raw data. To assess relative positions such as distances in a score plot, the plot needs to

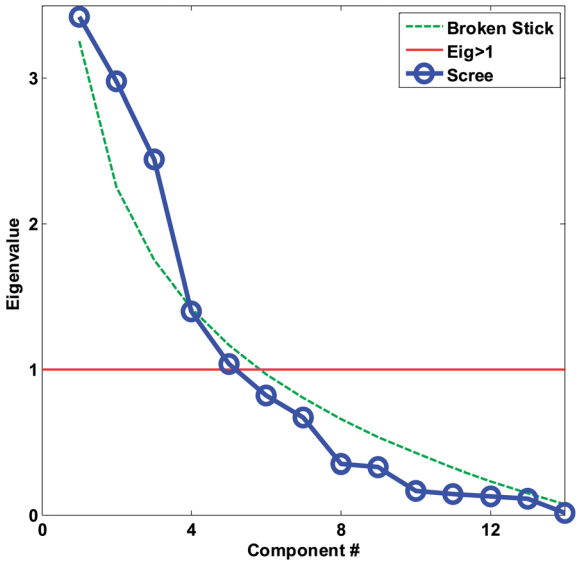


Fig. 20 Scree plot for the autoscaled wine data. The decision lines for having eigenvalues larger than one and the broken stick is also shown.

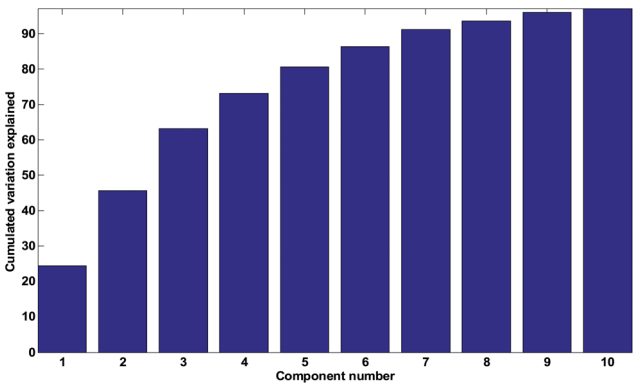


Fig. 21 Cumulated percentage variation explained.

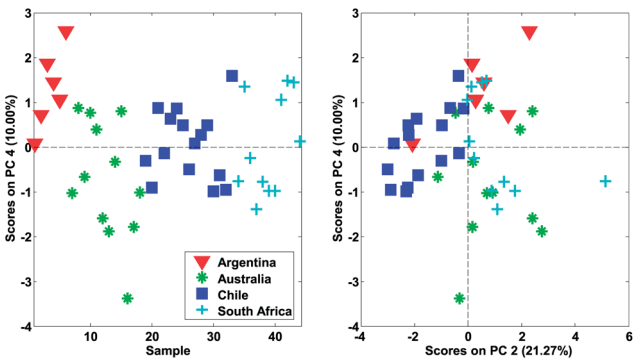


Fig. 22 Left: score number four of wine data. Right: score two *versus* score four.

preserve distances. This is mostly a problem in practice, when the magnitude of the two components is widely different. The score plot that does not preserve distances is still useful. For example, the lower left score plot in Fig. 13 is much better for discerning groupings and detecting patterns than the one to the lower right.

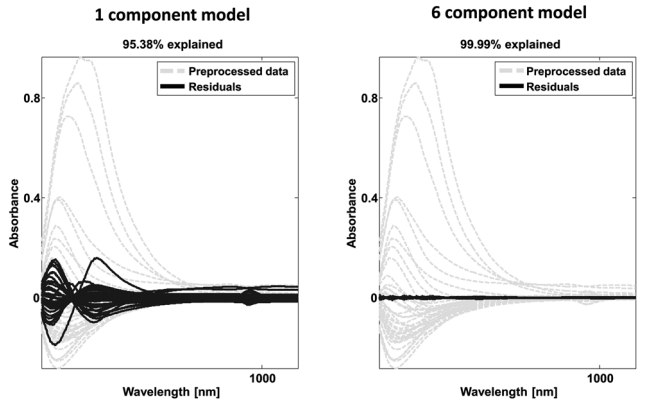


Fig. 23 Example of spectral data (grey) and residual spectral information after one (left) and six (right) components.

Visualizing and interpreting loadings. Loadings define what a principal component represents. Just as the weight in Fig. 4 defined the latent variable to represent a mixture of glycerol and ethanol, the loading vector of a PCA model does exactly the same. It defines what linear combination of the variables a particular component represents.

Fig. 14 shows the loadings of the two first components. With these, it is possible to explain what the scores of the model represent. For example, wines from Chile have low (negative) scores for component 2 (Fig. 12). This implies that they have a lot of the opposite of the phenomenon represented in loading 2. Hence, these samples have variation where ethanol, total, volatile, and lactic acids are low at the same time (relatively) while *e.g.* malic acid is high. Also, and this is an important point, certain variables that have low loadings close to zero, such as *e.g.* citric acid, do not follow this trend. Hence, the loading tells about what the trend is and also which variables are not part of the trend.

The phenomenon reflected in the principal component is also expected to be visually apparent in the raw data, but only with respect to how much variation of the data this component describes. The first component is seen in the label in Fig. 14 to explain 24.4% of the variation whereas the second one explains 21.3%. Together that means that 45.7% of the variation is explained by these two components. If the two components had explained 100%, all information would be contained in these two components, but for this particular model, half the variation is still retained in other components, so we should remain cautious not to claim that observations from the components are fully indicative of variations in the data.

An example on the importance of this is indicated in Fig. 15. The model reflected in the top plot shows that variables 4 and 6 are perfectly oppositely correlated. The model reflected in the bottom plot does not indicate that. In contrast, the low percentage explained, indicates that there are many other phenomena in the data so the correlation between variable 4 and 6 needs not be close to minus one as it will be in the first model.

Instead of looking at the loadings in line plots, it is also feasible to make scatter plots (Fig. 16). The scatter plot is often helpful for finding patterns of variation. For example, it is apparent in the plot that volatile acid and lactic acid are generally correlated in approximately 50% of the variation reflected in the two components. Residual sugar seems to be only moderately described in these two components as it is close to zero in both components. As the variables have been auto-scaled, a position close to zero implies that this particular variable does not co-vary with the variation that component 1 and 2 is reflecting.

As for the score scatter plot, distances are only preserved in the loading scatter plot, if the two loadings are plotted on the same scale. The basis for the loadings are the scores and these are generally not unit vectors as they carry the variance of the components. To correct for that, it is possible to simply normalize the scores and multiply the corresponding loading vectors by the inverse normalization factor. In essence, just moving the variance from the score vector to the loading vector.

Visualizing and interpreting loadings and scores together – biplots. It is possible and obvious to link the score and the loading plot. That way, it is possible to explain why *e.g.* a certain grouping is observed in a score plot. As hinted above, it is difficult to find a suitable base to plot on when combining scores and loadings, especially if preserving distances is desired. The biplot aims to solve this problem, or rather, presents a suitable set of compromises to choose from. Biplots were originally developed by K. R. Gabriel, but J. C. Gower has also contributed. The reader is urged to refer to the original literature for more in depth information.[12–14]

The principle behind biplots can be explained by representing the PCA model using

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} = \mathbf{T}^{(\mathrm{norm})}\mathbf{S}\mathbf{P}^{\mathrm{T}} \tag{9}$$

where $\mathbf{T}^{(\mathrm{norm})}$ is the score matrix with each column scaled to norm one just like the loadings are. The diagonal matrix $\mathbf{S}$ contains the norms of $\mathbf{T}$ on the diagonal. Above, no residuals are assumed for simplicity. Normally the scores are taken as $\mathbf{T}^{(\mathrm{norm})}\mathbf{S}$ ($=\mathbf{T}$) but if a distance preserving plot of the loadings is desired, it is more reasonable to set the loadings to $\mathbf{P}\mathbf{S}^{\mathrm{T}}$ and thus, have the scores be a normalized and orthogonal basis to base the plots on. Re-writing, the PCA model as

$$\mathbf{X} = \mathbf{T}^{(\mathrm{norm})}\mathbf{S}\mathbf{P}^{\mathrm{T}} = \mathbf{T}^{(\mathrm{norm})}\mathbf{S}^{\alpha}\mathbf{S}^{(1-\alpha)}\mathbf{P}^{\mathrm{T}} \tag{10}$$

It is possible to obtain the two solutions by either setting α equal to one or to zero. In fact, the most common biplot, takes α equal to 0.5 in order to produce a compromise plot where distances in both spaces can be approximately assessed. Hence, $\alpha = 0$ represents distances for variables (loadings) preserved, $\alpha = 1$ represents distances for samples (scores) preserved and $\alpha = 0.5$ represents distances for both samples and variables are (only) approximately preserved.

In addition to this scaling of the variance, there is often also a more trivial scaling of either the whole score matrix or the whole loading matrix to ensure that *e.g.* the score values are not so small compared to the loadings that they are not visible in a plot.

There are many interesting aspects of biplots and scatterplots but only a few important interpretational issues will be described here.

Two objects that are close and far from the origin have similar response (with respect to the variation explained by the components). For example, the two samples CHI-VDA1 and CHI-SCH1 are far from the origin and close together. Hence they are expected to be correlated, but only with respect to the approximately 50% that these two components describe. The two samples are plotted against each other in Fig. 18 (left). Note, that it is the preprocessed data that the PCA model reflects and hence, that interpretations can be made about.

Likewise, two variables that are close and far from the origin are correlated (with respect to the variation explained by the components). An example is given in Fig. 18 (right). Note, that the high correlation is apparently governed by an extreme sample – a potential outlier which will be discussed later.

The center of the plot represents the average sample – not zero – in case the data have been centered. Hence, samples with very negative scores have low values relative to the other samples and samples with high positive scores are the opposite. Again, with respect to the variation explained by the components.

The larger projection a sample has on the vector defined by a given variable, the more that sample deviates from the average on that particular variable (see *e.g.* how sample SOU-HHI1 projects to the axis defined by the variable lactic acid in Fig. 17).

It is often overlooked, that the above considerations for biplots apply equally well on loading plots or on score plots. Just like above, when for example, loadings are plotted without considering the magnitude of the scores, distances may be impossible to judge.

Practical aspects

Assumptions

In its most basic form, PCA can be seen as a basis for transformation. Instead of using the basis vectors $\mathbf{u}_j = (0, \ldots, 0, 1, 0, \ldots 0)'$ (with the one at place j) the basis given by $\mathbf{p}_1, \ldots, \mathbf{p}_J$ is used. For this transformation, no assumptions are needed. Considering PCA in the form of eqn (5) and (7), where a model is assumed and least squares fitting is chosen to estimate the parameters $\mathbf{T}$ and $\mathbf{P}$, it is not unreasonable to make some assumptions regarding the residuals as collected in $\mathbf{E}$. The mildest assumption is that one of these residuals being independently and identically distributed (iid), without specifying more than that this distribution is symmetrical around zero. Hence, there are no systematic errors and the error is homoscedastic.

When the errors are heteroscedastic and there is a model for the error, then eqn (7) can be fitted under this error model by using maximum likelihood or weighted least squares approaches.[15–17] Although this solves the problem of heteroscedasticity, certain implementations of maximum likelihood

fitting remove various aspects of the simplicity of PCA (orthogonal scores, nestedness of solutions, *etc.*).

Inference/validation

Since the PCA model parameters are used for interpretation and exploration, it is reasonable to ask how stable the results are. This calls for statistical inference tools. There are different routes to take in this respect. Upon assuming multivariate normality of the *x*-variables, statistical inference for the scores and loadings are available (see *e.g.* Anderson,[18] pp. 468). Multivariate normality cannot always be assumed, but approximate normality of the scores – they are linear combinations – envoking the Central Limit Theorem can sometimes be done. For a distribution-free approach, resampling methods can be used, *e.g.*, bootstrapping. This is, however, not trivial and several alternatives exist.[19,20]

Preprocessing

Often a PCA performed on the raw data is not very meaningful. In regression analysis, often an intercept or offset is included since it is the deviation from such an offset, which represents the interesting variation. In terms of the prototypical example, the absolute levels of the pH is not that interesting but the variation in pH of the different Cabernets is relevant. For PCA to focus on this type of variation it is necessary to mean-center the data. This is simply performed by subtracting from every variable in $\mathbf{X}$ the corresponding mean-level.

Sometimes it is also necessary to think about the scales of the data. In the wine example, there were measurements of concentrations and of pH. These are not on the same scales (not even in the same units) and to make the variables more comparable, the variables are scaled by dividing them by the corresponding standard deviations. The combined process of centering and scaling in this way is often called autoscaling. For a more detailed account of centering and scaling, see the literature.[21,22]

Centering and scaling are the two most common types of preprocessing and they normally always have to be decided upon. There are many other types of preprocessing methods available though. The appropriate preprocessing typically depends on the nature of the data investigated.[23–27]

Choosing the number of components

A basic rationale in PCA is that the informative rank of the data is less than the number of original variables. Hence, it is possible to replace the original J variables with R ($R \ll J$) components and gain a number of benefits. The influence of noise is minimized as the original variables are replaced with weighted averages,[28] and the interpretation and visualization is greatly aided by having a simpler (fewer variables) view to all the variations. Furthermore, the compression of the variation into fewer components can yield statistical benefits in further modelling with the data. Hence, there are many good reasons to use PCA. In order to use PCA, though, it is necessary to be able to decide on how many components to use. The answer to that problem depends a little bit on the purpose of the analysis, which is why the following three sections will provide different answers to that question.

Exploratory studies. In exploratory studies, there is no quantitatively well-defined purpose with the analysis. Rather, the aim is often to just 'have a look at the data'. The short answer to how many components to use then is: "just use the first few components". A slightly more involved answer is that in exploratory studies, it is quite common not to fix the number of components very accurately. Often, the interest is in looking at the main variation and per definition, the first components provide information on that. As *e.g.* component one and three do not change regardless of whether component six or seven is included, it is often not too critical to establish the exact number of components. Components are looked at and interpreted from the first component and downwards. Each extra component is less and less interesting as the variation explained is smaller and smaller, so often a gradual decline of interest is attached to components. Note that this approach for assessing the importance of components is not to be taken too literally. There may well be reasons why smaller variations are important for a specific dataset.[29]

If outliers are to be diagnosed with appropriate statistics (see next section), then, however, it is more important to establish the number of components to use. For example, the residual will change depending on how many components are used, so in order to be able to assess residuals, a reasonable number of components must be used. There are several *ad hoc* approaches that can be used to determine the number of components. A selection of methods is offered below, but note that these methods seldom provide clear-cut and definitive answers. Instead, they are often used in a combined way to get an impression on the effective rank of the data.

Eigenvalues and their relation to PCA. Before the methods are described, it is necessary to explain the relation between PCA and eigenvalues. An eigenvector of a (square) matrix $\mathbf{A}$ is defined as the nonzero vector $\mathbf{z}$ with the following property:

$$\mathbf{A}\mathbf{z} = \lambda\mathbf{z} \tag{11}$$

Where $\mathbf{z}$ is called the eigenvector. If matrix $\mathbf{A}$ is symmetric (semi-) positive definite, then the full eigenvalue decomposition of $\mathbf{A}$ becomes:

$$\mathbf{A} = \mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^{\mathrm{T}} \tag{12}$$

Where $\mathbf{Z}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a nonzero diagonal matrix. In chemometrics, it is customary to work with covariance or correlation matrices and these are symmetric (semi-) positive definite. Hence, eqn (12) describes their eigenvalue decomposition. Since all eigenvalues of such matrices are nonnegative, it is customary to order them from high to low; and refer to the first eigenvalue as the largest one.

The singular value decomposition of $\mathbf{X}$ ($I \times J$) is given by

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \tag{13}$$

Where $\mathbf{U}$ is an ($I \times J$) orthogonal matrix ($\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$); $\mathbf{S}$ ($J \times J$) is a diagonal matrix with the nonzero singular values on its

diagonal and $\mathbf{V}$ is an $(J \times J)$ orthogonal matrix ($\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$). This is for the case of $I > J$, but the other cases follow similarly. Considering $\mathbf{X}^T\mathbf{X}$ and upon using eqn (12) and (13) it follows

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{V}\mathbf{S}^2\mathbf{V}^T = \mathbf{Z}\Lambda\mathbf{Z}^T. \quad (14)$$

This shows the relationship between the singular values and the eigenvalues. The eigenvalue corresponding to a component is the same as the squared singular value which again is the variation of the particular component.

Scree test. The scree test was developed by R. B. Cattell in 1966.[30] It is based on the assumption that relevant information is larger than random noise and that the magnitude of the variation of random noise seems to level off quite linearly with the number of components. Traditionally, the eigenvalues of the cross-product of the preprocessed data, are plotted as a function of the number of components, and when only noise is modelled, it is assumed that the eigenvalues are small and decline gradually. In practice, it may be difficult to see this in the plot of eigenvalues due to the huge eigenvalues and often the logarithm of the eigenvalues is plotted instead. Both are shown in Fig. 19 for a simulated dataset of rank four and with various amounts of noise added. It is seen that the eigenvalues level off after four components, but the details are difficult to see in the raw eigenvalues unless zoomed in. It is also seen, that the distinction between 'real' and noise eigenvalues are difficult to discern at high noise levels.

For real data, the plots may even be more difficult to use as also exemplified in the original publication of Cattell as well as in many others.[31–33] Cattell himself admitted that: "*Even a test as simple as this requires the acquisition of some art in administering it*". This, in fact, is not particular to the scree test but goes for all methods for selecting the number of components.

For the wine data, it is not easy to firmly assess the number of components based on the scree test (Fig. 20). One may argue that seven or maybe nine components seem feasible, but this would imply incorporating components that explain very little variation. A more obvious choice would probably be to assess three components as suitable based on the scree plot and then be aware that further components may also contain useful information.

Eigenvalue below one. If the data is autoscaled, each variable has a variance of one. If all variables are orthogonal to each other, then every component in a PCA model would have an eigenvalue of one since the preprocessed cross-product matrix (the correlation matrix) is identity. It is then fair to say, that if a component has an eigenvalue larger than one, it explains variation of more than one variable. This has led to the rule of selecting all components with eigenvalues exceeding one (see the red line in Fig. 20). It is sometimes also referred to as the Kaisers' rule or Kaiser–Guttmans' rule and many additional arguments have been provided for this method.[34–36] While it remains a very *ad hoc* approach, it is nevertheless a useful rule-of-thumb to get an idea about the complexity of a dataset. For the wine data (Fig. 20), the rule suggests that around four or five components are reasonable. Note, that for very precise data, it is perfectly possible that even components with eigenvalues far below one can be real and significant. Real phenomena can be small in variation, yet accurate.

Broken stick. A more realistic cut off for the eigenvalues is obtained with the so called broken stick rule.[37] A line is added to the scree plot that shows the eigenvalues that would be expected for random data (the green line in Fig. 22). This line is calculated assuming that random data will follow a so-called broken stick distribution. The broken stick distribution hypothesizes how random variation will partition and uses the analogy of how the lengths of pieces of a stick will be distributed when broken at random places into J pieces.[38] It can be shown that for autoscaled data, this theoretical distribution can be calculated as

$$b_r = \sum_{j=r}^{J} \frac{1}{j}. \quad (15)$$

As seen in Fig. 20, the broken stick would seem to indicate that three to four components are reasonable.

High fraction of variation explained. If the data measured has *e.g.* one percent noise, it is expected that PCA will describe all the variation down to around one percent. Hence, if a two-component model describes only 50% of the variation and is otherwise sound, it is probable that more components are needed. On the other hand, if the data are very noisy coming *e.g.* from process monitoring or consumer preference mapping and has an expected noise fraction of maybe 40%, then an otherwise sound model fitting 90% of the variation would imply overfitting and fewer components should be used. Having knowledge on the quality of the data can help in assessing the number of components. In Fig. 21, the variation explained is shown. The plot is equivalent to the eigenvalue plot except it is cumulative and on a different scale. For the wine data, the uncertainty is different for each variable, and varies from approximately 5 and even up to 50% relative to the variation in the data. This is quite variable and makes it difficult to estimate how much variation should be explained, but most certainly less than 50% would mean that all is not explained and explaining more than, say 90–95% of the variation would be meaningless and just modelling of noise. Therefore, based on variation explained, it is likely that there is more than two but less than, say, seven components.

Valid interpretation. As indicated by the results, the different rules above seldom agree. This is not as big a problem as it might seem. Quite often, the only thing needed is to know the neighbourhood of how many components are needed. Using the above methods 'informally' and critically, will often provide that answer. Furthermore, one of the most important strategies for selecting the number of components is to supplement such methods with interpretations of the model. For the current data, it may be questioned whether *e.g.* three or four components should be used.

In Fig. 22, it is shown, that there is distinct structure in the scores of component four. For example, the wines from Argentina all have positive scores. Such a structure or grouping will not happen accidentally unless unfortunate confounding has occurred. Hence, as long as Argentinian wines were not measured separately on a different system or something

similar, the mere fact that component four (either scores or loadings) shows distinct behaviour is an argument in favour of including that component. This holds regardless of what other measures might indicate.

The loadings may also provide similar validation by highlighting correlations expected from *a priori* knowledge. In the case of continuous data such as time series or spectral data, it is also instructive to look at the shape of the residuals. An example is provided in Fig. 23. A dataset consisting of visual and near-infrared spectra of 40 beer samples is shown in grey. After one component, the residuals are still fairly big and quite structured from a spectral point of view. After six components, there is very little information left indicating that most of the systematic variation has been modelled. Note from the title of the plot, that 95% of the variation explained is quite low for this dataset whereas that would be critically high for the wine data as discussed above.

Cross-validation. In certain cases, it is necessary to establish the appropriate number of components more firmly than in the exploratory or casual use of PCA. For example, a PCA model may be needed to verify if the data of a new patient indicate that this patient is similar to diseased persons. This may be accomplished by checking if the sample is an outlier when projected into a PCA model (see next section on outliers). Because the outlier diagnostics depend on the number of components chosen, it is necessary to establish the number of components before the model can be used for its purpose. There are several ways do to this including the above-mentioned methods. Oftentimes, though, they are considered too *ad hoc* and other approaches are used. One of the more popular approaches is cross-validation. S. Wold was the first to introduce cross-validation of PCA models[39] and several slightly different approaches have been developed subsequently. Only a brief description of cross-validation will be given here, but details can be found in the literature.[40,41]

The idea in cross-validation is to leave out part of the data and then estimate the left-out part. If this is done wisely, the prediction of the left-out part is independent of the actual left-out part. Hence, overfitting leading to too optimistic models is not possible. Conceptually, a single element (typically more than one element) of the data matrix is left out. A PCA model handling missing data,[42–46] can then be fitted to the dataset and based on this PCA model, an estimate of the left out element can be obtained. Hence, a set of residuals is obtained where there are no problems with overfitting. Taking the sum of squares of these yields the so-called Predicted REsidual Sums of Squares (PRESS)

$$\mathrm{PRESS}_r = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij}^{(r)}\right)^2 \quad (16)$$

where $x_{ij}^{(r)}$ is the residual of sample i and variable j after r components. From the PRESS the Root Mean Squared Error of Cross-Validation (RMSECV) is obtained as

$$\mathrm{RMSECV}_r = \sqrt{\frac{\mathrm{PRESS}_r}{IJ}} \quad (17)$$

In Fig. 24, the results of cross-validation are shown. As shown in Fig. 21 the fit to data will trivially improve with the number of components but the RMSECV gets worse after four components, indicating that no more than four components should be used. In fact, the improvement going from three to four components is so small, that three is likely a more feasible choice from that perspective.

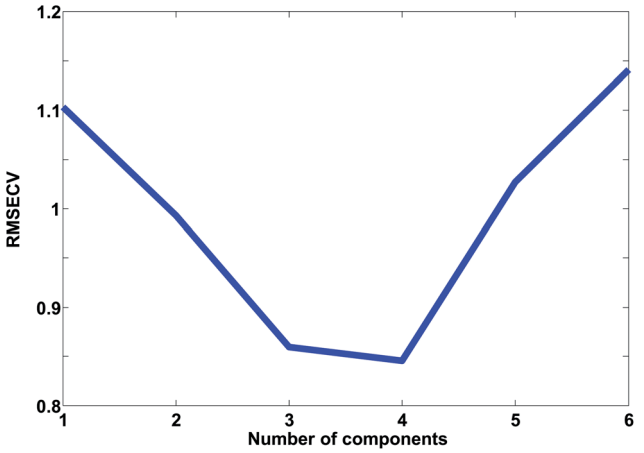


Fig. 24 A plot of RMSECV for PCA models with different number of components.

The cross-validated error, RMSECV, can be compared to the fitted error, the Root Mean Squared Error of Calibration, RMSEC. In order for the two to be comparable though, the fitted residuals must be corrected for the degrees of freedom consumed by the model. Calculating these degrees of freedom is not a trivial subject as mentioned earlier.[3,4,47]

When using PCA for other purposes. It is quite common to use PCA as a preprocessing step in order to get a nicely compact representation of a dataset. Instead of the original many (J) variables, the dataset can be expressed in terms of the few (R) principal components. These components can then in turn be used for many different purposes (Fig. 25).

It is common practice to use, for example, cross-validation for determining the number of components and then use that number of components in further modelling. For example, the scores may be used for building a classification model using linear discriminant analysis. While this approach to selecting components is both feasible and reasonable there is a risk that components that could help improve classification would be left out. For example, cross-validation may indicate that five components are valid, but it turns out that component seven

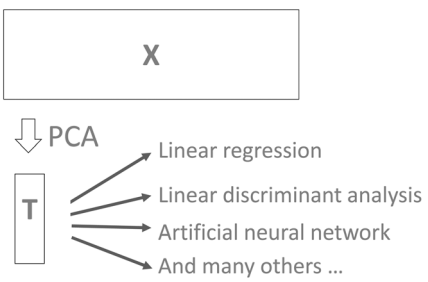


Fig. 25 Using the scores of PCA for further modelling.

can reliably improve classification. In order to be certain that useful information is retained in the PCA model, it is generally advised to validate the number of components in terms of the actual goal. Instead of validating the number of components that best describe **X** in some sense (PCA cross-validation), it will often make more sense to use the number of components that provides the best classification results if PCA is used in conjunction with discriminant analysis.

Detecting outliers

Outliers are samples that are somehow disturbing or unusual. Often, outliers are downright wrong samples. For example, in determining the height of persons, five samples are obtained ([1.78, 1.92, 1.83, 167, 1.87]). The values are in meters but accidentally, the fourth sample has been measured in centimeters. If the sample is not either corrected or removed, the subsequent analysis is going to be detrimentally disturbed by this outlier. Outlier detection is about identifying and handling such samples. An alternative or supplement to outlier handling is the use of robust methods, which will however, not be treated in detail here. The reader is referred to the literature for more details on robust methods.[48–59]

This section is mainly going to focus on identifying outliers, but understanding the outliers is really the critical aspect. Often outliers are mistakenly taken to mean 'wrong samples' and nothing could be more wrong! Outliers can be absolutely right, but *e.g.* just badly represented. In such a case, the solution is not to remove the outlier, but to supplement the data with more of the same type. The bottom line is that it is imperative to understand why a sample is an outlier. This section will give the tools to identify the samples and see in what way they differ. It is then up to the data analyst to decide how the outliers should be handled.

Data inspection. An often forgotten, but important, first step in data analysis is to inspect the raw data. Depending on the type of data, many kinds of plots can be relevant as already mentioned. For spectral data, line plots may be nice. For discrete data, histograms, normal probability plots, or scatter plots could be feasible. In short, any kind of visualization that will help elucidate aspects of the data can be useful. Several such plots have already been shown throughout this paper. It is also important, and frequently forgotten, to look at the preprocessed data. While the raw data are important, they actually never enter the modeling. It is the preprocessed data that will be modeled and there can be big differences in the interpretations of the raw and the preprocessed data.

Score plots. While raw and preprocessed data should always be investigated, some types of outliers will be difficult to identify from there. The PCA model itself can provide further information. There are two places where outlying behavior will show up most evidently: in the scores and in the residuals. It is appropriate to go through all selected scores and look for samples that have strange behaviour. Often, it is only component one and two that are investigated but it is necessary to look at all the relevant components.

As for the data, it is a good idea to plot the scores in many ways, using different combinations of scatter plots, line plots, histograms, *etc.* Also, it is often useful to go through the same plot but coloured by all the various types of additional information available. This could be any kind of information such as temperature, storage time of sample, operator or any other kind of either qualitative or quantitative information available. For the wine data model, it is seen (Fig. 26) that one sample is behaving differently from the others in score plot one *versus* two (upper left corner).

Looking at the loading plot (Fig. 16) indicates that the sample must be (relatively) high in volatile and lactic acid and low in malic acid. This should then be verified in the raw data. After removing this sample, the model is rebuilt and reevaluated. No more extreme samples are observed in the scores.

Before deciding on what to do with an outlier, it is necessary to look at how important the component is. Imagine a sample that is doing an 'excellent job' in the first seven components, but in the eighth has an outlying behaviour. If that eighth component is very small in terms of variation explained and not the most important for the overall use of the model; then it is probably not urgent to remove such a sample.

Whenever in doubt as to whether to remove an outlier or not, it is often instructive to compare the models before and after removal. If the interpretation or intended use changes dramatically, it indicates that the sample has an extreme

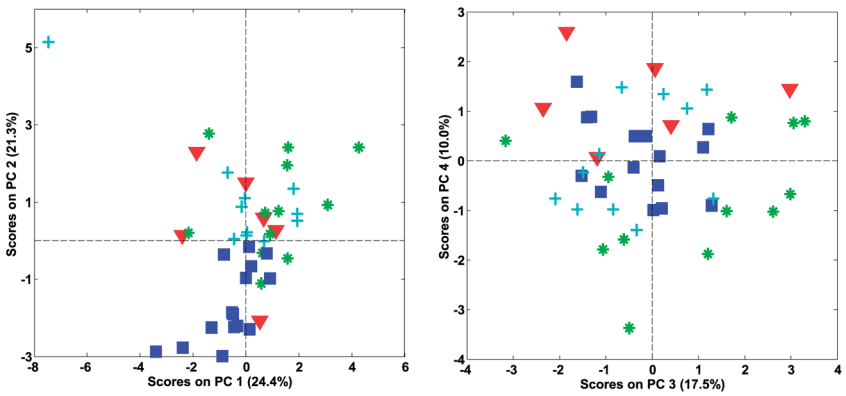


Fig. 26 Score plot of a four component PCA model of the wine data.

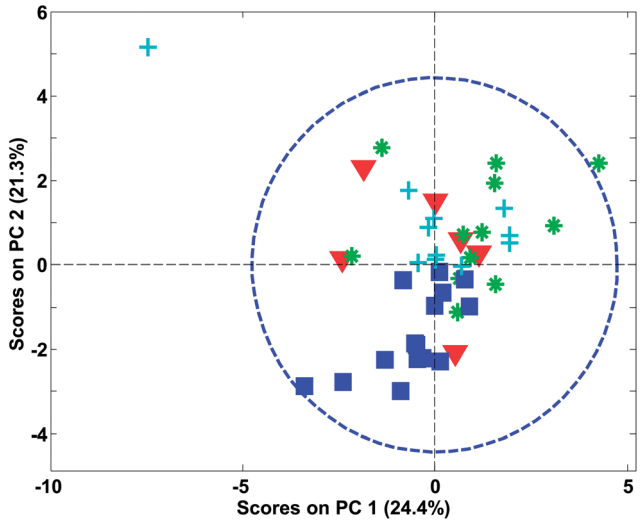


Fig. 27 PCA score plot similar to Fig. 26 (left) but now with a 95% confidence limit shown.

behaviour that needs to be handled whereas the opposite indicates that it is of little importance whether the sample is removed.

Hotelling's T^2. Looking at scores is helpful, but it is only possible to look at few components at a time. If the model has many components, it can be laborious and the risk of accidentally missing something increases. In addition, in some cases, outlier detection has to be automated in order to function *e.g.* in an on-line process monitoring system. There are ways to do so, and a common way is to use the so-called Hotelling's T^2 which was introduced in 1931.[60] This diagnostic can be seen as an extension of the *t*-test and can also be applied to the scores of a PCA model.[61] It is calculated as

$$T_i^2 = \frac{\mathbf{t}_i^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\mathbf{t}_i}{I-1} \quad (18)$$

Where $\mathbf{T}$ is the matrix of scores ($I \times R$) from all the calibration samples and $\mathbf{t}_i$ is an $R \times 1$ vector holding the R scores of the ith sample. Assuming that the scores are normally distributed, then confidence limits for T_i^2 can be assigned as

$$T_{i\,(I,R)}^2 = \frac{R(I-1)}{I-R}F_{R,I-R,\alpha} \quad (19)$$

In Fig. 27, an example of the 95% confidence limits is shown. This plot illustrates the somewhat deceiving effect such limits can have. Two samples are outside the confidence limit leading the inexperienced user to suggest leaving out both. However, first of all, samples should not be left out without understanding why they are 'wrong' and more importantly, there is nothing in what we know about the data thus far, that suggests the scores would follow a multivariate normal distribution. Hence, the limit is rather arbitrary and for this particular dataset, the plot in Fig. 26 is definitely to be preferred when assessing if samples behave reasonably. In some cases, when enough samples are available and those samples really do come from the same population, the scores are approximately normally distributed. This goes back to the central limit theorem.[62] Examples are, *e.g.* in the multivariate process control area.[63] In those cases Hotelling's T^2 is a particularly useful statistic.

The limits provided by Hotelling's T^2 can be quite misleading for grouped data. As an example, Fig. 28 shows the score plot of a dataset, where the samples fall in four distinct groups (based on the geological background). The sample in the middle called "outlier?" is by no means extreme with respect to Hotelling's T^2 even though the sample is relatively far from all other samples.

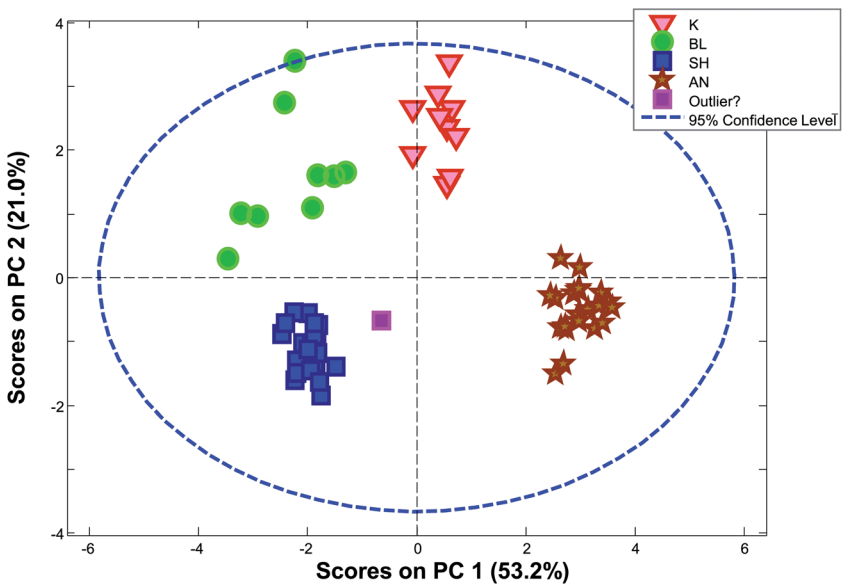


Fig. 28 PCA scores plot (1 *vs.* 2) for a dataset consisting of ten concentrations of trace elements in obsidian samples from four specific quarries – data from a study by Kowalski *et al.*[64]

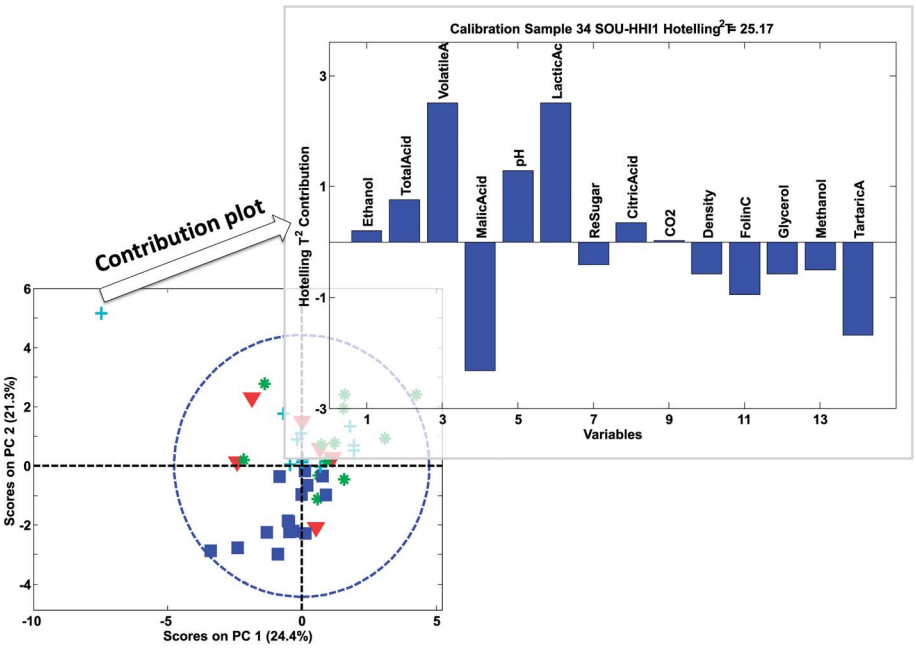


Fig. 29 Contribution plot for sample 34 in the wine data.

Score contribution plots. When a sample has been detected as being an outlier, it is often interesting to try to investigate the reason. Extreme scores indicate that the sample has high levels of whatever, the specific component reflects in its corresponding loading vector. Sometimes, it is difficult to verify directly what is going on and the so-called contribution plot can help. There are several different implementations of contribution plots[65] but one common version was originally developed by Nomikos.[66] The contribution for a given sample indicates what variables caused that sample to get an extreme set of scores. For a given set of components (*e.g.* component one and two in Fig. 29), this contribution can be calculated as

$$c_j^D = \sum_{r=1}^{R} \frac{t_r^{\text{new}} x_j^{\text{new}} p_{jr}}{\mathbf{t}_r^{\mathrm{T}} \mathbf{t}_r / (I-1)} \qquad (20)$$

The vector $\mathbf{t}_r$ is rth score vector from the calibration model, I the number of samples in the calibration set and t_r^{new} is the score of the sample in question. It can come from the calibration set or be a new sample. x_j^{new} is the data of the sample in question for variable j and p_{jr} is the corresponding loading element. In this case, R components are considered, but fewer components can also be considered by adjusting the summation in eqn (20).

The contribution plot indicates what variables make the selected sample have an extreme Hotelling's T^2 and in Fig. 29, the most influential variables are also the ones that that are visible in the raw data (not shown). Eqn (20) explains the simplest case of contribution plots with orthogonal **P** matrices. Generalized contributions are available for non-orthogonal cases.[65] Note that if x_j^{new} is a part of the calibration set, it influences the model. A more objective measure of whether x_j^{new} fits the model can be obtained by removing it from the data and then afterwards projecting it onto the model thereby obtaining more objective scores and residuals.

Lonely wolfs. Imagine a situation where the samples are constituted by distinct groups rather than one distribution as also exemplified in Fig. 28. Hotelling's T^2 is not the most obvious choice for detecting samples that are unusually positioned but not far from the center. A way to detect such samples, is to measure the distance of the sample to the nearest neighbor. This can also be generalized *e.g.* to the average distance to the k nearest neighbors and various distance measures can be used if so desired.

In Fig. 30, it is seen that colouring the scores by the distance to the nearest neighbour, highlights that there are, in fact, several samples that are not very close to other samples. When the samples are no longer coloured by class as shown in Fig. 28, it is much less obvious that the green 'K' class is indeed a well-defined class.

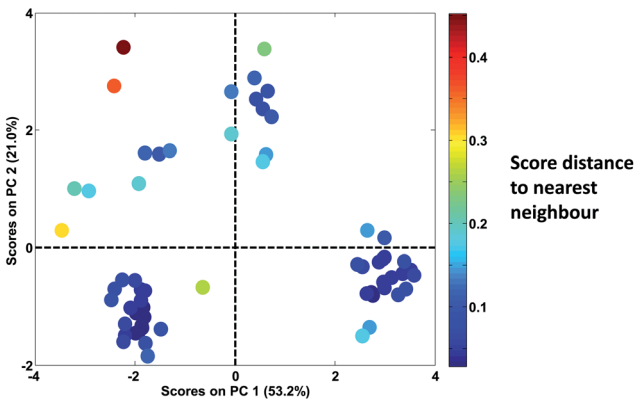


Fig. 30 Score plot of Fig. 28. Samples are coloured according to the distance of the sample to the nearest neighbour.

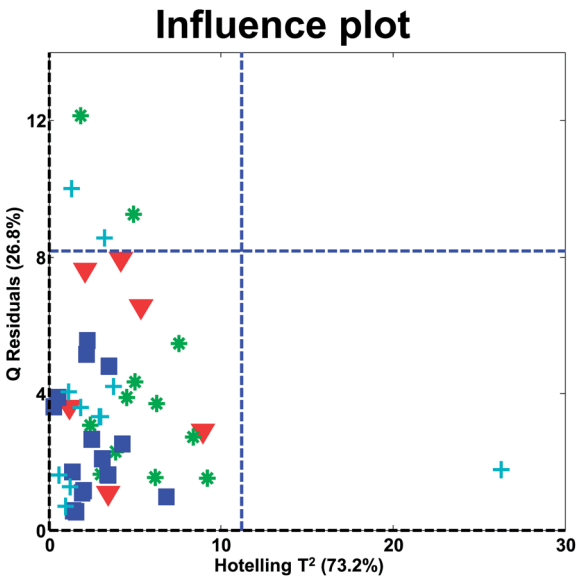


Fig. 31 Influence plot of wine data with a four component PCA model.

Residuals. The use of residuals has already been described in detail. For outlier detection, it is common to use the sum squared residuals, often called the Q-statistics, of each sample to look for samples that are not well-described by the PCA model. When Q is plotted against T^2, it is often referred to as an influence plot. Note, that both residuals and T^2 will change with the number of components, so if the number of components are not firmly defined, it may be necessary to go back and forth a bit between different numbers of components.

In the influence plot in Fig. 31, it is clear that one sample stands out with a high Hotelling's T^2 in the PCA model and no samples have extraordinarily large residuals. It will hence, be reasonable to check the T^2 contribution plot of that sample, to see if an explanation for the extreme behavior can be obtained. The two blue lines are 95% confidence levels. Such lines are often given in software but should not normally be the focus of attention as also described above for score plots.

Residual contribution plots. Just as contribution plots for scores can be defined, contribution plots for residual variation can be determined as well. These are simpler to define, as the contributing factor to a high residual is simply the squared residual vector itself. Hence, if a sample shows an extraordinary residual variation, the residual contribution plot (the residuals of the sample) can indicate why the sample has high residual variation. The squared residuals do not reveal the sign of the deviation and sometimes, the raw residuals are preferred to the squared ones to allow the sign to be visible.[67]

Conclusion

Principal component analysis is a powerful and versatile method capable of providing an overview of complex multivariate data. PCA can be used *e.g.* for revealing relations between variables and relations between samples (*e.g.* clustering), detecting outliers, finding and quantifying patterns, generating new hypotheses as well as many other things. This tutorial has provided a description of the basic concepts of how to use PCA critically.

References

1 T. Skov, D. Ballabio and R. Bro, Multiblock variance partitioning: a new approach for comparing variation in multiple data blocks, *Anal. Chim. Acta*, 2008, **615**, 18–29.
2 D. Ballabio, T. Skov, R. Leardi and R. Bro, Classification of GC-MS measurements of wines by combining data dimension reduction and variable selection techniques, *J. Chemom.*, 2008, **22**, 457–463.
3 K. Faber, Degrees of freedom for the residuals of a principal component analysis — A clarification, *Chemometrics and Chemoinformatics*, 2008, vol. 93, pp. 80–86.
4 H. Martens, S. Hassani, E. M. Qannari and A. Kohler, Degrees of freedom estimation in Principal Component Analysis and Consensus Principal Component Analysis, *Chemom. Intell. Lab. Syst.*, 2012, **118**, 246–259.
5 J. M. F. ten Berge, *Least squares optimization in multivariate analysis*, DSWO Press, Leiden, 1993.
6 A. K. Smilde, H. C. J. Hoefsloot, H. A. L. Kiers, S. Bijlsma and H. F. M. Boelens, Sufficient conditions for unique solutions within a certain class of curve resolution models, *J. Chemom.*, 2001, **15**(4), 405–411.
7 K. Pearson, On lines and planes of closest fit to points in space, *Philos. Mag.*, 1901, **2**, 559–572.
8 H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, 1933, **24**, 417–441.
9 J. M. F. ten Berge and H. A. L. Kiers, Are all varieties of PCA the same? A reply to Cadima & Jolliffe, *Br. J. Math. Stat. Psychol.*, 1997, **50**(2), 367–368.
10 S. Wold, C. Albano, W. J. Dunn, III, U. Edlund, K. H. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and M. Sjöström, Multivariate data analysis in chemistry, in *Chemometrics. Mathematics and Statistics in Chemistry*, ed. B. R. Kowalski, D. Reidel Publishing Company, Dordrecht, 1984, pp. 17–95.
11 I. E. Frank and S. Lanteri, Classification models: discriminant analysis, SIMCA, CART, *Chemom. Intell. Lab. Syst.*, 1989, **5**, 247–256.
12 J. C. Gower, A general theory of Biplots, in *Recent Advances in Descriptive Multivariate Analysis*, ed. W. J. Krzanowski, Clarendon Press, Oxford, 1995, pp. 283–303.
13 A. Carlier and P. M. Kroonenberg, Decompositions and biplots in three-way correspondence analysis, *Psychometrika*, 1996, **61**, 355–373.
14 K. R. Gabriel, The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, 1971, **58**, 453–467.
15 R. Bro, N. D. Sidiropoulos and A. K. Smilde, Maximum likelihood fitting using simple least squares algorithms, *J. Chemom.*, 2002, **16**(8–10), 387–400.
16 D. T. Andrews and P. D. Wentzell, Applications of maximum likelihood principal component analysis: incomplete data

sets and calibration transfer, *Anal. Chim. Acta*, 1997, **350**(3), 341–352.
17 P. D. Wentzell, D. T. Andrews, D. C. Hamilton, N. M. Faber and B. R. Kowalski, Maximum likelihood principal component analysis, *J. Chemom.*, 1997, **11**(4), 339–366.
18 T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, 2nd edn, 1984.
19 M. E. Timmerman, H. A. L. Kiers and A. K. Smilde, Estimating confidence intervals for principal component loadings: a comparison between the bootstrap and asymptotic results, *Br. J. Math. Stat. Psychol.*, 2007, **60**(2), 295–314.
20 H. Babamoradi and F. van den Berg, Rinnan, Å. Bootstrap based Confidence Limits in Principal Component Analysis–a case study, *Chemom. Intell. Lab. Syst.*, 2013, **120**, 97–105.
21 R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde and M. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, *BMC Genomics*, 2006, **7**(142).
22 R. Bro and A. K. Smilde, Centering and scaling in component analysis, *J. Chemom.*, 2003, **17**(1), 16–33.
23 N. K. Afseth, V. H. Segtnan and J. P. Wold, Raman spectra of biological samples: a study of preprocessing methods, *Appl. Spectrosc.*, 2006, **60**(12), 1358–1367.
24 C. D. Brown, L. Vega-Montoto and P. D. Wentzell, Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration, *Appl. Spectrosc.*, 2000, **54**(7), 1055–1068.
25 S. N. Deming, J. A. Palasota and J. M. Nocerino, The geometry of multivariate object preprocessing, *J. Chemom.*, 1993, **7**, 393–425.
26 H. Martens and E. Stark, Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy, *J. Pharm. Biomed. Anal.*, 1991, **9**(8), 625–635.
27 M. Pardo, G. Niederjaufner, G. Benussi, E. Comini, G. Faglia, G. Sberveglieri, M. Holmberg and I. Lundstrom, Data preprocessing enhances the classification of different brands of Espresso coffee with an electronic nose, *Sens. Actuators, B*, 2000, **69**(3), 397–403.
28 R. Bro, Multivariate calibration - What is in chemometrics for the analytical chemist?, *Anal. Chim. Acta*, 2003, **500**(1–2), 185–194.
29 O. E. de Noord, The influence of data preprocessing on the robustness and parsimony of multivariate calibration models, *Chemom. Intell. Lab. Syst.*, 1994, **23**, 65–70.
30 R. B. Cattell, The scree test for the number of factors, *Multivariate Behav. Res.*, 1966, **1**, 245–276.
31 P. M. Bentler and K. H. Yuan, Test of linear trend in eigenvalues of a covariance matrix with application to data analysis, *Br. J. Math. Stat. Psychol.*, 1996, **49**(2), 299–312.
32 P. M. Bentler and K. H. Yuan, Tests for linear trend in the smallest eigenvalues of the correlation matrix, *Psychometrika*, 1998, **63**, 131–144.
33 R. C. Henry, E. S. Park and C. Spiegelman, Comparing a new algorithm with the classic methods for estimating the number of factors, *Chemom. Intell. Lab. Syst.*, 1999, **48**, 91–97.
34 H. F. Kaiser, The Application of Electronic Computers to Factor Analysis, *Educ. Psychol. Meas.*, 1960, **20**, 141–151.
35 N. Cliff, The Eigenvalues-Greater-Than-One Rule and the Reliability of Components, *Psychol. Bull.*, 1988, **103**(2), 276–279.
36 L. Guttman, Some necessary conditions for common-factor analysis, *Psychometrika*, 1954, **19**(2), 149–161.
37 S. Frontier, Étude de la decroissance des valeurs propers dans une analyze en composantes principales: comparison avec le modèle de baton brisé, *J. Exp. Mar. Biol. Ecol.*, 1976, **25**, 67–75.
38 R. H. MacArthur, On the relative abundance of bird species, *Proc. Natl. Acad. Sci. U. S. A.*, 1957, **43**(3), 293–295.
39 S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, *Technometrics*, 1978, **20**, 397–405.
40 W. J. Krzanowski and P. Kline, Cross-validation for choosing the number of important components in principal component analysis, *Multivariate Behav. Res.*, 1995, **30**(2), 149–165.
41 R. Bro, K. Kjeldahl, A. K. Smilde and H. A. L. Kiers, Cross-validation of component models: a critical look at current methods, *Anal. Bioanal. Chem.*, 2008, **390**, 1241–1251.
42 T. C. Gleason and R. Staelin, A proposal for handling missing data, *Psychometrika*, 1975, **40**, 229–252.
43 P. R. C. Nelson, P. A. Taylor and J. F. MacGregor, Missing data methods in PCA and PLS: score calculations with incomplete observations, *Chemom. Intell. Lab. Syst.*, 1996, **35**(1), 45–65.
44 B. Grung and R. Manne, Missing values in principal component analysis, *Chemom. Intell. Lab. Syst.*, 1998, **42**, 125–139.
45 B. Walczak and D. L. Massart, Dealing with missing data Part I, *Chemom. Intell. Lab. Syst.*, 2001, **58**(1), 15–27.
46 E. Adams, B. Walczak, C. Vervaet, P. G. Risha and D. L. Massart, Principal component analysis of dissolution data with missing elements, *Int. J. Pharm.*, 2002, **234**(1–2), 169–178.
47 J. Mandel, The partitioning of interaction in analysis of variance, *J. Res. Natl. Bur. Stand., Sect. B*, 1969, **73B**, 309–328.
48 S. J. Devlin, R. Gnanadesikan and J. R. Kettenring, Robust estimation of dispersion matrices and principal components, *J. Am. Stat. Assoc.*, 1981, **76**(375), 354–362.
49 O. S. Borgen and Å. Ukkelborg, Outlier detection by robust alternating regression, *Anal. Chim. Acta*, 1993, **277**, 489–494.
50 Y. Xie, J. Wang, Y. Liang, L. Sun, X. Song and R. Yu, Robust principal component analysis by projection pursuit, *J. Chemom.*, 1993, **7**, 527–541.
51 H. Hove, Y. Liang and O. M. Kvalheim, Trimmed object projections: a nonparametric robust latent-structure decomposition method, *Chemom. Intell. Lab. Syst.*, 1995, **27**, 33–40.
52 J. G. Chen, J. A. Bandoni and J. A. Romagnoli, Robust PCA and normal region in multivariate statistical process monitoring, *AIChE J.*, 1996, **42**(12), 3563–3566.

53 W. C. Chen, H. Cui and Y. Z. Liang, A new principal component analysis method based on robust diagnosis, *Anal. Lett.*, 1996, **29**, 1647–1667.
54 A. Singh, Outliers and robust procedures in some chemometric applications, *Chemom. Intell. Lab. Syst.*, 1996, **33**, 75–100.
55 E. V. Thomas and N. X. Ge, Development of robust multivariate calibration models, *Technometrics*, 2000, **42**(2), 168–177.
56 K. A. Hoo, K. J. Tvarlapati, M. J. Piovoso and R. Hajare, A method of robust multivariate outlier replacement, *Comput. Chem. Eng.*, 2002, **26**(1), 17–39.
57 M. Hubert, P. J. Rousseeuw and K. Vanden Branden, ROBPCA: a new approach to robust principal component analysis, *Technometrics*, 2005, **47**(1), 64–79.
58 S. F. Møller, J. V. F. Frese and R. Bro, Robust methods for multivariate data analysis, *J. Chemom.*, 2005, **19**, 549–563.
59 P. J. Rousseeuw, M. Debruyne, S. Engelen and M. Hubert, Robustness and outlier detection in chemometrics, *Crit. Rev. Anal. Chem.*, 2006, **36**(3–4), 221–242.
60 H. Hotelling, The generalization of Student's ratio, *Ann. Math. Stat.*, 1931, **2**, 360–378.
61 J. E. Jackson, Principal components and factor analysis: part I - principal components, *J. Qual. Tech.*, 1980, **12**, 201–213.
62 A. M. Mood, F. R. Graybill and D. C. Boes, *Introduction to the Theory of Statistics*, McGraw-Hill, 3rd edn, 1974.
63 P. Nomikos and J. F. MacGregor, Multivariate SPC charts for monitoring batch processes, *Technometrics*, 1995, **37**, 41–59.
64 B. R. Kowalski, T. F. Schatzki and F. H. Stross, Classification of archaeological artifacts by applying pattern recognition to trace element data, *Anal. Chem.*, 1972, **44**(13), 2176–2180.
65 J. A. Westerhuis, S. P. Gurden and A. K. Smilde, Generalized contribution plots in multivariate statistical process monitoring, *Chemom. Intell. Lab. Syst.*, 2000, **51**(1), 95–114.
66 P. Nomikos, Detection and diagnosis of abnormal batch operations based on multiway principal component analysis, *ISA Trans.*, 1996, **35**, 259–266.
67 B. M. Wise, N. L. Ricker and D. Veltkamp, Upset and Sensor Failure Detection in Multivariate Processes, *AIChE 1989 Annual Meeting*, Nov. 1989.