Principal Component Analysis - PCA

TUTORIAL REVIEW - Principal component analysis

By Rasmus Bro & Age K. Smilde

Christian G. Holm kxm508

Emil H.C. Henningsen tzs820

09-03-2023

UNIVERSITY OF COPENHAGEN

Introduction

Analytical Methods

ROYAL SOCIETY OF CHEMISTRY

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

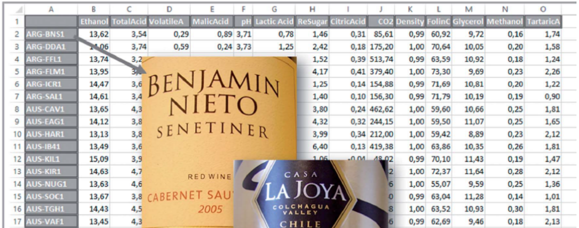
DOI: 10.1039/c3ay41907j

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,

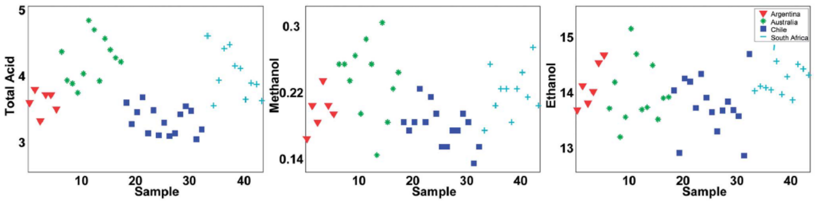


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Ethanol	TotalAcid	VolatileA	MalicAcid	pH	Lactic Acid	ReSugar	CitricAcid	CO2	Density	FolinC	Glycerol	Methanol	TartaricA
2	ARG-BNS1	13,62	3,54	0,29	0,89	3,71	0,78	1,46	0,31	85,61	0,99	60,92	9,72	0,16	1,74
3	ARG-DDA1	[illegible]	3,74	0,59	0,24	3,73	1,25	2,42	0,18	175,20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

Unaltered data

Objective:

- Reduce the dimensionality of a dataset
- Simplify complex datasets and make them more amenable to analysis.
- PCA captures the essential information in the data while removing redundant or noisy information



How to do it 1

1. Standardize the data - *Autoscaling*
2. Compute the covariance matrix
3. Compute the eigenvectors and eigenvalues
4. Select the principal components
5. Create the loading (variables) and scores vector (samples)
6. Cross-validation
7. Project the data onto the principal components
8. Interpreting results

How to do it (2) - Now with math

$$\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n = 0 \tag{1}$$

$$\begin{aligned}
\bar{x} = \frac{1}{N}\sum_{n=1}^{N} w^T y_n = 0 \quad \Rightarrow \quad \sigma_x^2 &= \frac{1}{N}\sum_{n=1}^{N} x_n^2 = \frac{1}{N}\sum_{n=1}^{N} (w^T y_n)^2 = \frac{1}{N}\sum_{n=1}^{N} w^T y_n w y_n^T \\
&= w^T \left(\frac{1}{N}\sum_{n=1}^{N} y_n y_n^T \right) w \\
&= w^T C w
\end{aligned} \tag{2}$$

$$\Rightarrow \sigma^2 w = Cw, \quad \textit{Eigenvalue problem} \tag{3}$$

Selecting the relevant principal components

- Selecting components for visualization - $ND \longrightarrow 2D/3D$, in this case: $14D \longrightarrow 2D$.
- Unless data trends are known beforehand, there may not be a precise way to determine, what components to take a closer look into.
- Select the components with the greatest variance, as greater variance is a sign of grouping trends in the data.

Broken-stick distribution

- One way is to look at components with variance greater than 1.
- Another way is to look at all components above the so-called "broken-stick" distribution:

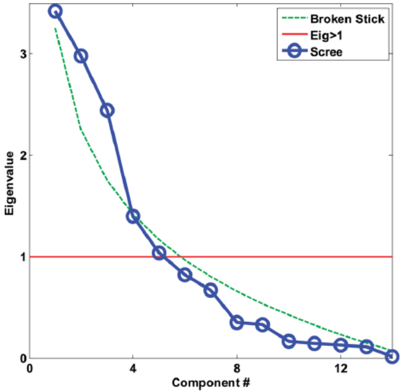
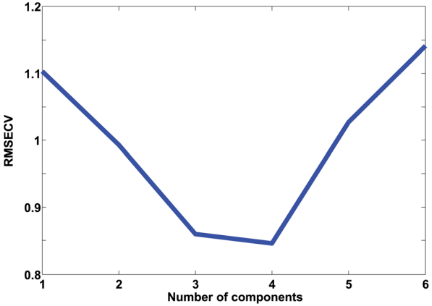


Figure: The eigenvalues ranging from greatest to lowest and two acceptance parameters.

Using cross-validation to select # of components

- When the before-mentioned approaches seem too ad hoc, other practices can be used.
- One other is the use of cross-validation and least-squares.
- Leave out k samples and fit them to the resulting principal components.
- Compute the sum of squared residues.
- Repeating the process



Projecting the data along the chosen principal components axes

- The wine samples plotted along the 4 principal component axes with greatest variance.
- The labels correspond to the region of origin
- A few conclusions can already be made from this projection of the data:
- Wine samples from Chile score exclusively negative values in the second principal component.
- Argentinian wine samples score exclusively positive values in the fourth principal components.

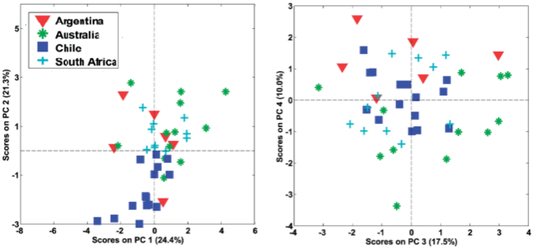


Figure: The wine data plotted along the axes of the 4 greatest principal components.

What do the principal components actually mean?

To conclude anything quantitatively from the PCA analysis itself, the features must be investigated.

- Are the data features themselves well-defined? How certain are the values of the individual sample features?
- If good, then the weights of the respective features making up a principal component directly translates the principal component values.
- A principal component could e.g. describe the ratio of methanol and ethanol.
- If very little is known about the features, it is much more difficult.

Further investigation of data after PCA

- Other Machine Learning practices
- Classification algorithms

“

Prediction is very difficult, especially if it’s about the future.

Niels Bohr