An Introduction to PCA

Bryan A. Hanson[a]

[a]Prof. Emeritus, Dept. of Chemistry & Biochemistry, DePauw University; hanson@depauw.edu

This version was compiled on July 25, 2019

The audience for this document is chemists, spectroscopists and people in allied fields who are using spectroscopy to analyze their data.

We will use two data sets here: one is a set of elemental analyses of glass artifacts; we use this relatively small, non-spectroscopic data set to help understand PCA fundamentals. The second data set is a collection of IR spectra of plant oils.

Conceptual Introduction to PCA

PCA is conducted on data sets composed of:

- Samples, typically in rows.
- Variables which were measured for each sample.

The purpose of PCA is *data reduction*. This term refers to the goal of:

- Reducing the size of the data set by identifying variables that are not informative. Such variables are also described as "noisy", in that they don't add anything to the study. Such variables arise naturally in many situations. Example: a survey about food preferences may include questions about political party. The answers about political party may not be informative.
- Collapsing correlating variables. Several of the variables measured in a study may actually be measures of the same underlying reality. This is not to say they are noisy, but rather they may be redundant. Example: a survey asks participants if they eat kale, and separately, if they eat quinoa. Individuals may answer yes to both questions or no to both questions. The answers may reflect the individuals preference for a healthy diet. Either question alone may be sufficient. PCA will collapse these correlating variables into one variable.

What does one get from PCA?

- An indication of how many principal components (PC) are needed to describe the data, generally presented as a *scree plot*.
- Scores, generally presented as one or more *score plots*.
- Loadings, generally presented as one or more *loading plots*.

These plots will be explained further in the next section. Other things to know about PCA before going further:

Table 1. A portion of the archaeological glass data set. Values are percentages.

Na2O	MgO	Al2O3	SiO2	P2O5	SO3	Cl	K2O
13.904	2.244	1.312	67.752	0.884	0.052	0.936	3.044
14.194	2.184	1.310	67.076	0.938	0.024	0.966	3.396
14.668	3.034	1.362	63.254	0.988	0.064	0.886	2.828
14.800	2.455	1.385	63.790	1.200	0.115	0.988	2.878
14.078	2.480	1.072	68.768	0.682	0.070	0.966	2.402

- PCA is *principal* not *principle* components analysis!
- PCA is the "mother" of a number of other related techniques, so if you plan further study it is critical to understand PCA to the greatest degree possible.
- That said, it takes most people a long time to fully grasp what PCA does, especially from the mathematical perspective. Don't expect to get all the nuances on the first pass!
- *And the problem* . . . The results of PCA, scores and loadings, exist in a so-called abstract space. This space is only distantly and indirectly related to the space in which the original samples reside. Therefore, the results of PCA are frequently difficult to interepret in concrete terms. See previous point.

PCA Results Illustrated, No Code, No Math

This section is designed to illustrate the concepts of PCA, and how to interpret the plots that arise from PCA.

We'll use a data set which reports chemical analyses on archaeological glass artifacts that was designed to determine the origin of the artifacts. Table 1 gives a little bit of the data set.[1]

There are 180 glass artifacts (the samples) in this data set (hence 180 rows), and the elements analyzed were Na_2O, MgO, Al_2O_3, SiO_2, P_2O_5, SO_3, Cl, K_2O, CaO, MnO, Fe_2O_3, BaO, and PbO .[2]

We'll perform PCA on the glass data set, show the three plots and then discuss them in turn. Figure 1 shows the scree plot, Figure 2 shows the score plot and Figure 3 shows the first loadings.

Figure 1, the scree plot, shows the amount of variance in the data set explained by each principal component (PCs are along the x axis, from 1 to 10).[3] For now, think of variance as the variation or variability in the data set, or better, think of it as the hidden signal in the data. To interpret

[1]This is the `glass` data set in package `chemometrics`.

[2]The elements are reported as their oxides in the form of weight percent.

[3]Because there are 13 variables, the most PCs one could have is 13. In theory, keeping all 13 PCs perfectly reproduces the original data set.

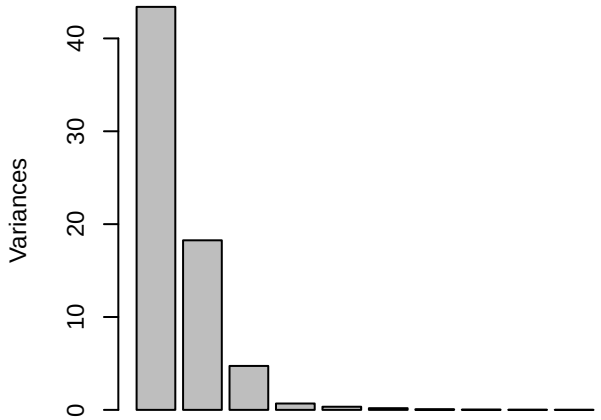

Fig. 1. Scree plot from PCA on the glass data set.

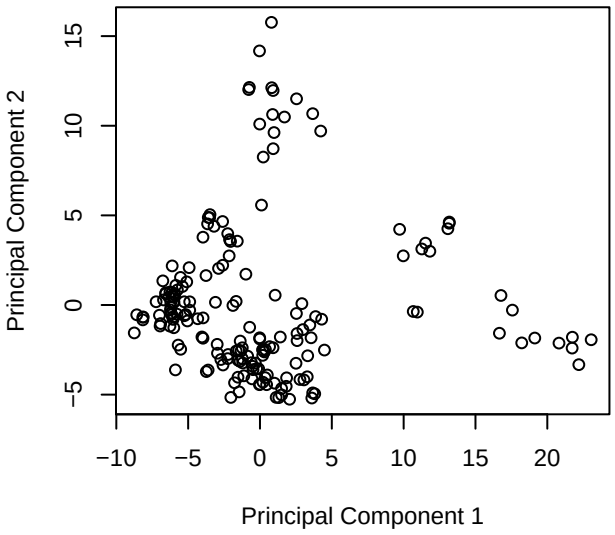

Fig. 2. Score plot from PCA on the glass data set.

this plot, we look for the point at which the height of the bars suddenly levels off. In this case, the first three PCs drop steadily downward, but from PC four and onward there is little additional variation (signal) that can be explained. We would say that three PCs are enough to explain this data set. In other words, the original 13 variables have been reduced to three, which is a great simplification.

In Figure 2 one sees the scores for PC 1 plotted against the scores for PC 2. There are 180 points in this plot because there is one point per sample (put another way, every sample has a score value for PC 1 and for PC 2). This plot is interpreted by looking for clustering of samples, as well as for samples that are outliers, off by themselves. Depending upon your eye, there are 3 to 5 clusters here. Later we'll discuss how we can explore this further.

We could also plot PC 1 against PC 3, or PC 2 against PC 3. These might show different clustering and separation of samples, but are not shown here. There wouldn't be much point in plotting PC 4 or higher, as PCs 4 and higher are mostly noise, not signal, as established by the scree plot (Figure 1).

The loadings plot, Figure 3, shows how much each measured variable contributes to the component and hence the separation of samples (in this case the loadings for PC 1). We see that three elements have large loadings, and the other elements contribute little to the separation. We would say separation along PC 1 is driven largely and *collectively* by the results for Na_2O, SiO_2 and CaO.[4] The first PC should be interpreted as a composite of these variables – these variables

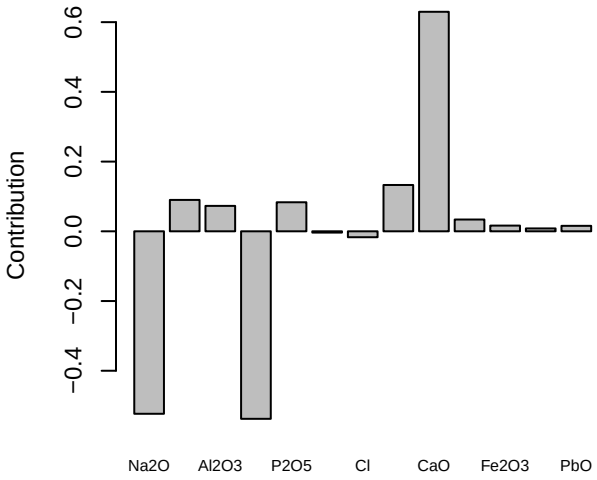

Fig. 3. Loadings plot for the first PC from PCA on the glass data set.

[4] If you knew this would be the result ahead of time, you probably would not have taken the time and expense to analyze the uninformative elements. However, we haven't looked at PC 2 or PC 3 so this conclusion is premature.

Table 2. Variance (signal) accounted for by PCs. Values in percent.

component	variance	cumulative
PC 1	64	64
PC 2	27	91
PC 3	7	98
PC 4	1	99
PC 5	0	100
PC 6	0	100
PC 7	0	100
PC 8	0	100
PC 9	0	100
PC 10	0	100
PC 11	0	100
PC 12	0	100
PC 13	0	100

have been collapsed into one new variable, PC 1.

Refinements 1. Rather than relying on a scree plot to determine the number of PCs that are important, we can present the same information in a table, see Table 2. A general rule of thumb says to keep enough PCs to account for 95% of the variation (signal). The table tells us the same as the plot: keep three PCs.

Refinements 2. The mathematics of PCA do not take into account anything about the samples other than the measured variables. However, the researcher may well know something about the samples, for instance, they may fall into groups based on their origin. If this is the case, the points on the score plot can be colored according to the group. This may aid significantly in the interpretation. Lucky for us, we can do this for the glass data set. The samples fell into four groups. We'll re-do the score plot with colors corresponding to the known groups (Figure 4).

With this figure, we can see that the large group in the lower left corner (in black), which to the eye might have been two groups, is composed of related samples.

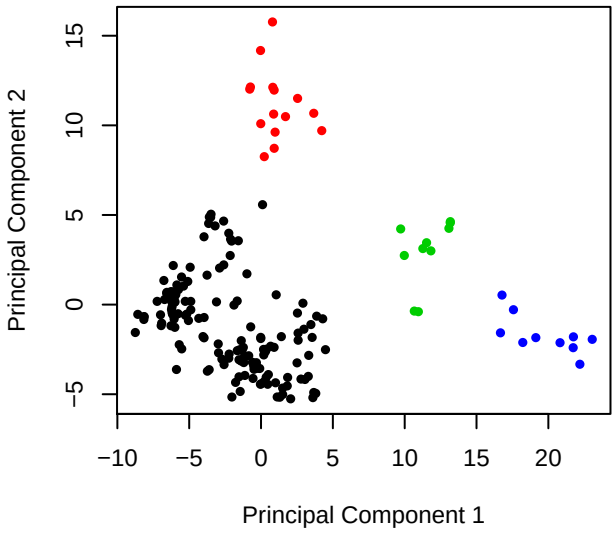

Fig. 4. Score plot from PCA on the glass data set, with groups color-coded.

A Spectroscopic Data Set

The archaeological data set has the advantage of only having a few variables, the percentages of each element in the glass artifacts. If we move to a spectroscopic data set, the number of variables goes up dramatically. A UV-Vis data set typically would have a few hundred to a thousand wavelength variables, an IR data set perhaps a few thousand data points, and a 1D NMR data set would typically have 16K or more data points. As far as PCA is concerned, in these cases the scree plot and score plot do not change in appearance or interpretation.

However, the loading plot changes appearance dramatically. This is because with hundreds to thousands of variables, one would not create a loading plot based on a bar chart (Figure 3 is a bar chart). Instead, the loading plot with many variables looks like a spectrum! While the appearance is different, the interpretation is the same as for when there are only a few variables.

Let's illustrate with an IR data set. We'll use a data set included with the `ChemoSpec` package. This is a set of IR spectra of plant oils which are mixtures of triglycerides (triacylglyerols, which are esters), and free fatty acids. Figure 5 shows a typical spectrum from the data set.[5]

Next, we'll carry out PCA as before, and show the scree plot (Figure 6) and the score plot (Figure 7). Note that these appear like the corresponding plots for the glass data set, and are interpreted in the same manner. However, the loadings plot, Figure 8, looks a lot like a spectrum, because it has 1868 data points, and is plotted as a connected scatter plot and not as a bar chart (which would be difficult to read).

Let's look at the carbonyl region of the loadings plot in detail. Figure 9 shows the original spectrum in red, for reference, and the loadings in black. One can see that the ester carbonyl contributes positively to the first loading, while the carboxylic acid carbonyl contributes negatively.

Finally, to make the point that the loading plot for many variables is essentially the same as the loading plot for just a few variables, Figure 10 shows the carbonyl region as a kind of bar plot. If one connects the tips of the bars together, one gets the previous plot.

[5] Plots in this vignette are deliberately made rather plain to focus on the data and to be consistent for ease-of-comparison. Package `ChemoSpec` makes much more polished plots.

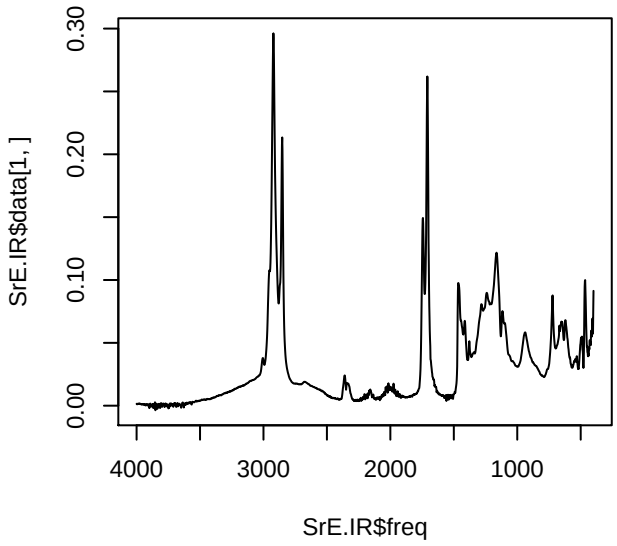

Fig. 5. Spectrum 1 from the IR data set.

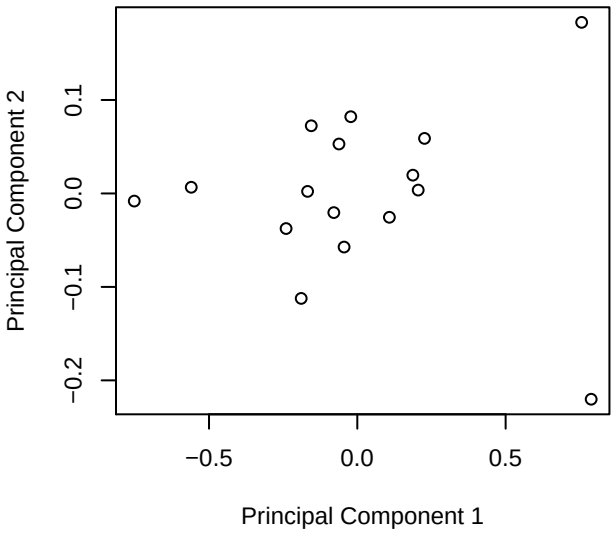

Fig. 7. Score plot from PCA on the IR data set.

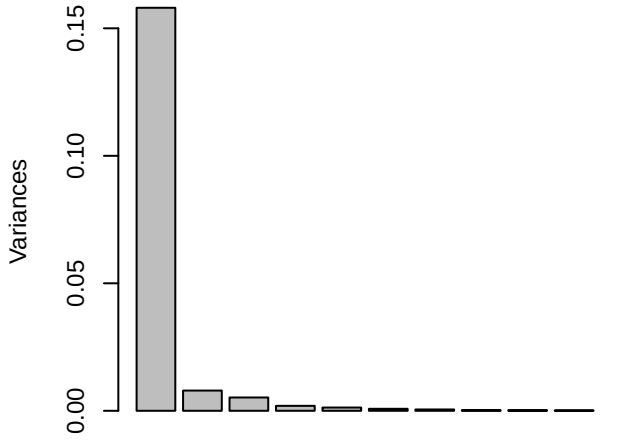

Fig. 6. Scree plot from PCA on the IR data set.

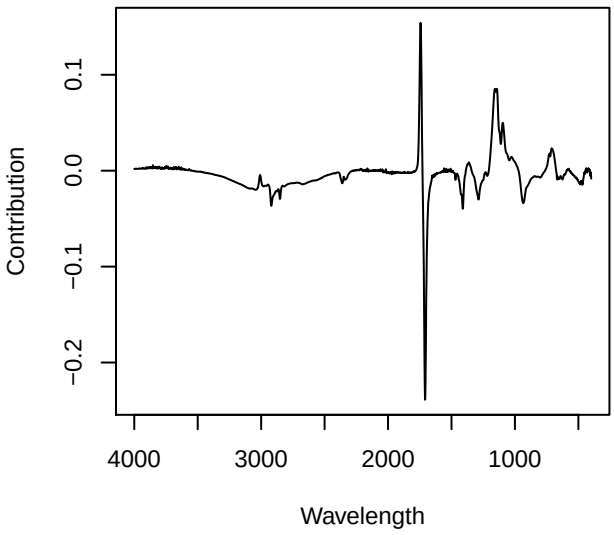

Fig. 8. Loadings plot for the first PC from PCA on the IR data set.

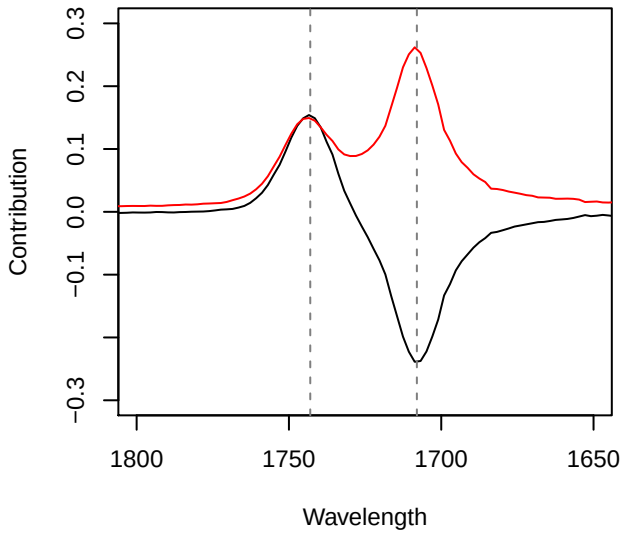

Fig. 9. Loadings plot for the first PC from PCA on the IR data set, carbonyl region. Reference spectrum shown in red.

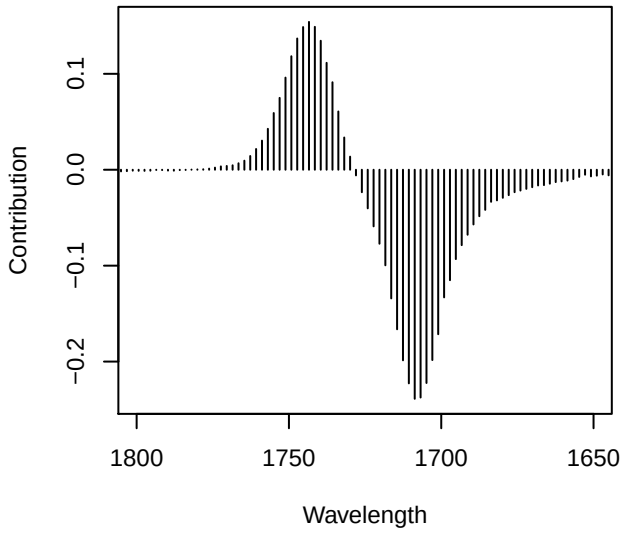

Fig. 10. Loadings plot for the first PC from PCA on the IR data set, carbonyl region, shown as a bar plot.