Principal Component Analysis

Principal Component Analysis (PCA, [24, 25]) is a technique which, quite literally, takes a different viewpoint of multivariate data. In fact, PCA defines new variables, consisting of linear combinations of the original ones, in such a way that the first axis is in the direction containing most variation. Every subsequent new variable is orthogonal to previous variables, but again in the direction containing most of the remaining variation. The new variables are examples of what often is called latent variables (LVs), and in the context of PCA they are also termed principal components (PCs).

The central idea is that more often than not high-dimensional data are not of full rank, implying that many of the variables are superfluous. If we look at high-resolution spectra, for example, it is immediately obvious that neighbouring wavelengths are highly correlated and contain similar information. Of course, one can try to pick only those wavelengths that appear to be informative, or at least differ from the other wavelengths in the selected set. This could, e.g., be based on clustering the variables, and selecting for each cluster one “representative”. However, this approach is quite elaborate and will lead to different results when using different clustering methods and cutting criteria. Another approach is to use variable selection, given some criterion – one example is to select the limited set of variables leading to a matrix with maximal rank. Variable selection is notoriously difficult, especially in high-dimensional cases. In practice, many more or less equivalent solutions exist, which makes the interpretation quite difficult. We will come back to variable selection methods in Chapter 10.

PCA is an alternative. It provides a direct mapping of high-dimensional data into a lower-dimensional space containing most of the information in the original data. The coordinates of the samples in the new space are called scores, often indicated with the symbol $T$. The new dimensions are linear combinations of the original variables, and are called loadings (symbol $P$). The term Principal Component (PC) can refer to both scores and loadings; which is meant is usually clear from the context. Thus, one can speak of
sample coordinates in the space spanned by PC 1 and 2, but also of variables contributing greatly to PC 1.

The matrix multiplication of scores and loadings leads to an approximation \( \tilde{X} \) of the original data \( X \):

\[
\tilde{X} = T_a P_a^T \tag{4.1}
\]

Superscript \( T \), as usual, indicates the transpose of a matrix. The subscript \( a \) indicates how many components are taken into account: the largest possible number of PCs is the minimum of the number of rows and columns of the matrix:

\[
a_{\text{max}} = \min(n, p) \tag{4.2}
\]

If \( a = a_{\text{max}} \), the approximation is perfect and \( \tilde{X} = X \).

The PCs are orthogonal combinations of variables defined in such a way that [25]:

- the variances of the scores are maximal;
- the sum of the Euclidean distances between the scores is maximal;
- the reconstruction of \( X \) is as close as possible to the original: \( \|X - \tilde{X}\| \)

is minimal.

These three criteria are equivalent [24]; the next section will show how to find the PCs.

PCA has many advantages: it is simple, has a unique analytical solution optimizing a clear and unambiguous criterion, and often leads to a more easily interpretable data representation. The price we have to pay is that we do not have a small set of wavelengths carrying the information but a small set of principal components, in which all wavelengths are represented. Note that the underlying assumption is that variation equals information. Intuitively, this makes sense: one can not learn much from a constant number.

Once PCA has defined the latent variables, one can plot all samples in the data set while ignoring all higher-order PCs. Usually, only a few PCs are needed to capture a large fraction of the variance in the data set (although this is highly dependent on the type of data). That means that a plot of (the scores of) PC 1 versus PC 2 can already be highly informative. Equally useful is a plot of the contributions of the (original) variables to the important PCs. These visualizations of high-dimensional data perhaps form the most important application of PCA. Later, we will see that the scores can also be used in regression and classification problems.

4.1 The Machinery

Currently, PCA is implemented even in low-level numerical software such as spreadsheets. Nevertheless, it is good to know the basics behind the computations. In almost all cases, the algorithm used to calculate the PCs is \textit{Singular}